A proposal of alternative primers for the ARTIC Network's multiplex PCR to improve coverage of SARS-CoV-2 genome sequencing

Kentaro Itokawa*, Tsuyoshi Sekizuka, Masanori Hashino, Rina Tanaka, Makoto Kuroda

Pathogen Genomics Center, National Institute of Infectious Diseases, Tokyo, Japan
Toyama 1-23-1, Shinjuku-ku, Tokyo, Japan

To whom correspondence should be addressed*: itokawa@nih.go.jp

Abstract

A group of biologists, ARTIC Network, has proposed a multiplexed PCR primer set for whole-genome analysis of the novel coronavirus, SARS-CoV-2, soon after the start of COVID-19 epidemics was realized. The primer set was adapted by many researchers worldwide and has already contributed to the high-quality and prompt genome epidemiology of this rapidly spreading viral disease. We have also seen the great performance of their primer set and protocol; the primer set amplifies all desired 98 PCR amplicons with fairly small amplification bias from clinical samples with relatively high viral load. However, we also observed an acute drop of reads derived from some amplicons especially amplicon 18 and 76 in "pool 2" as a sample's viral load decreases. We suspected this low coverage issue was due to dimer formation between primers targeting those two amplicons. Indeed, replacement of just one of those primers, nCoV-2019_76_RIGHT, to a newly designed primer resulted in a drastic improvement of coverages at both regions targeted by the amplicons 18 and 76. Given this result, we further replaced four primers in "pool 1" with each respective alternative. These modifications also improved coverage in eight amplicons particularly in samples with low viral load. The results of our experiments clearly indicate that primer dimer formation is one critical cause of coverage bias in ARTIC protocol. Importantly, some of the problematic primers are detectable by observing primer dimers in raw NGS sequence reads and replacing them with alternatives as shown in this study. We expect a continuous improvement of the ARTIC primer set will extend the limit for completion of SARS-CoV-2 genomes to samples with lower viral load, that supports better genomic epidemiology and mitigation of spread of this pathogen.

Background

The spreading of the novel corona virus, SARS-CoV-2, which is responsible for the respiratory illness, COVID-19, starting from December 2019 has become a huge concern in

the medical community around the world. In modern epidemiology, it is important to capture variations in genome sequence among isolates of such outbreaking pathogens for monitoring pathogen's evolution or tracking epidemiological chains in local to even global scale. Relatively large genome of the corona virus (approx. 30 kb), however, makes it challenging to reconstruct whole genome of the virus from samples with various viral loads in cost-effective manner. Recently, a group of molecular biologists which is called ARTIC Network (https://artic.network/) proposed 98 multiplexing PCR primer pairs (hereafter ARTIC primer set: https://github.com/artic-network/artic-ncov2019/tree/master/primer_schemes/nCoV-2019/V1). The ARTIC primer set are designed based on a published reference SARS-CoV-2 genome MN908947.3 and are tilled almost whole region of the genomic region. Those 98 primer pairs are actually divided into two separate subsets (pool_1 and pool2) so that each PCR fragment does not overlap to each other in the same PCR.

We have already tested the ARTIC primer set and their protocol for several clinical samples, mostly RNA from pharyngeal swab, and observed a great performance of their primer set. Actually, the primer set works quite well for samples with relatively high viral load (Ct < 25 in clinical qPCR test). For such high viral load samples, all designated amplicons are amplified within acceptable level of coverage bias for subsequent NGS analysis. The approach of ARTIC Network is expected to save resources and scale to a number of samples with a given NGS sequencing capacity, which will be crucial for large epidemic situation being concerned in many countries. As sample's viral load decreases, however, gradual increase of the entire coverage bias was observed with their protocol. Although such phenomenon is normally expected in multiplexed PCR for low-copy templates, we observed that coverages for the two particular PCR amplicons, 18 and 76, which correspond to the genomic regions coding for the nsp3 in ORF1a and S protein, respectively, decays far more rapidly than other targets. In our experience so far, low to absolute-zero depth for those two amplicons tends to be most frequent bottle neck for completion of all targeted genomic regions from samples with middle to low viral load (Ct > 27). This low coverage issue at the amplicons 18 and 76 are also seen in data published from other groups (e.g. https://cadde.s3.climb.ac.uk/covid-19/BR1.sorted.bam).

Results

In ARTIC primes set, the PCR amplicons, 18 and 76, are amplified by the primer pairs nCoV-2019_18_LEFT and nCoV-2019_RIGHT and nCoV-2019_76_LEFT and nCoV-2019_RIGHT, respectively, which are included in the same multiplexed reaction "pool_2". We noticed that two of those primers, nCoV-2019_18_LEFT and nCoV-2019_76_RIGHT were perfectly complement to each other by their 10-nt sequence at the 3'-end (Fig 1). From this

observation, we reasoned that the rapid decrease of the amplification efficiencies of those amplicons was due to a primer dimer formation between nCoV-2019_18_LEFT and nCoV-2019_76_RIGHT, that could compete to the amplification of desired targets. Indeed, we observed many NGS reads derived from the predicted dimer in raw FASTQ data (data not shown).

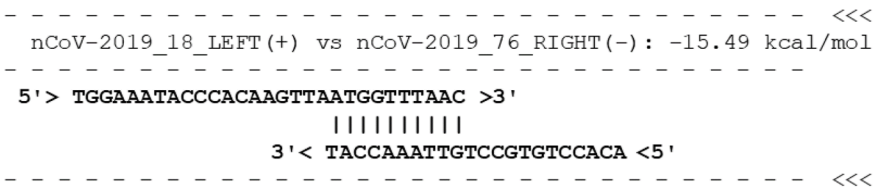

Fig 1 Predicted primer dimer formed by nCoV-2019_18_LEFT and nCoV-2019_76_RIGHT by PrimerROC (Johnston et al., 2019).

Then, we replaced one of those 'unlucky' primer pair, nCoV-2019_76_RIGHT, in the pool_2 to a newly designed nCoV-2019_76_RIGHTv2 (Table 1) which locates at 48-nt downstream to nCoV-2019_76_RIGHT.

Table 1 The original and alternative primers

Primer name	Location (1-base)	Sequence
nCoV-2019_76_RIGHT	23193..23214	5′-**ACACCTGTGCCTGTTAAACCAT**-3′
nCoV-2019_76_RIGHTv2	23241..23265	5′-**TCTCTGCCAAATTGTTGGAAAGGCA**-3′

We conducted amplification of SARS-CoV-2 genome from eight clinical samples with various Ct-values in clinical qPCR test which range from 25.0 to 29.6 by the ARTIC protocol. The number PCR cycle was set to 30 for all samples. The amplified PCR products in pool1 and pool 2 were combined, purified and fed to Illumina library preparation, and then sequenced in illumina iSeq100. Comparison of genome coverage with original and new primer set for a clinical sample with Ct-values 27.5 is shown in Fig 2A (see Fig S1 for results of all eight samples). The replacement of single primer certainly added notable improvement in read depths at the regions covered by the amplicons 18 and 76 (Fig 2A and Fig S1, highlighted by green strips). There was no notable adverse effect observed in other PCR products by this replacement of a primer.

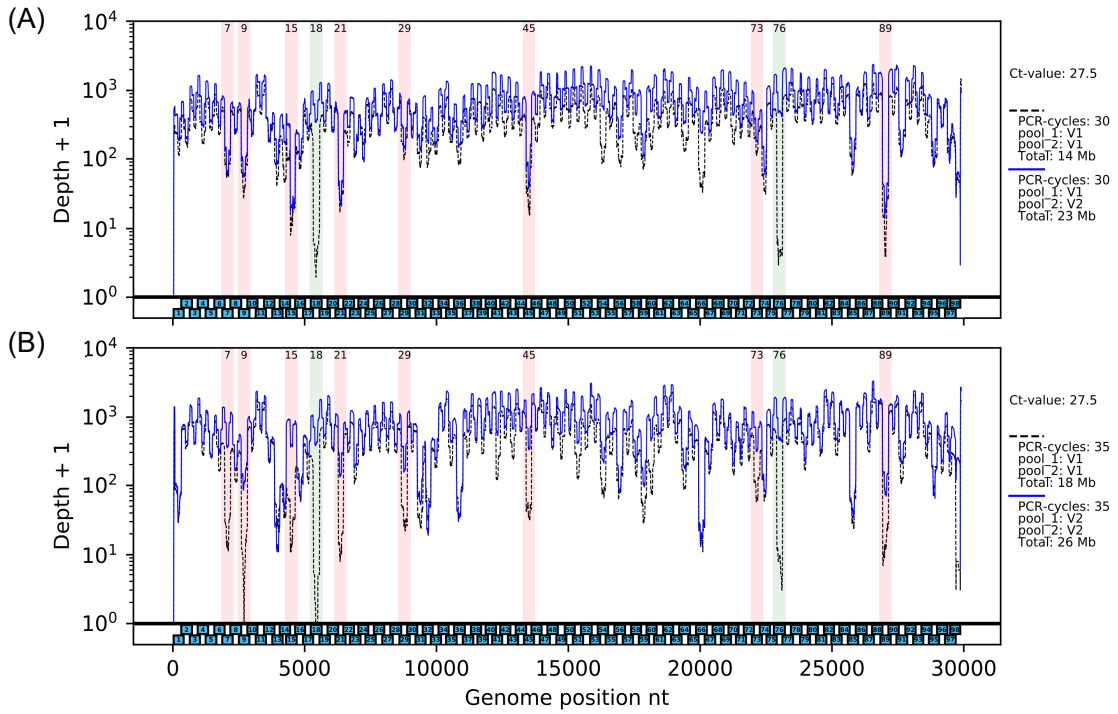

Fig 2 Depth plot on the SARS-CoV2 genome for mapped NGS reads obtained from ARTIC protocol. (A) With the original or alternative pool_2 primer set. (B) With the original or alternative pool_1 and pool_2 primer sets. Both plots are results for one identical clinical sample with moderate viral loads (Ct=27.5). It should be noted that the experiment was conducted for 4-fold diluted cDNA sample than our usual protocol. The numbered blue boxes depicted on the bottom area correspond to locations of each primer's targets. The green and red vertical strips highlight regions of amplicons potentially affected by dimer formation among primers in pool_1 or pool_2, respectively.

The above result indicated that formation of primer dimers plays critical role in coverage bias in samples with low viral load. Given this observation, we detected additional six primer dimers (Fig 3) from raw NGS read data. The primers involved in these dimers were all included in the pool 1. Interestingly, the eight amplicons related to those primers (7, 9, 15, 21, 29, 45, 73 and 89) consistently showed relative low depth (Fig 2A and Fig S1, highlighted by red strips).

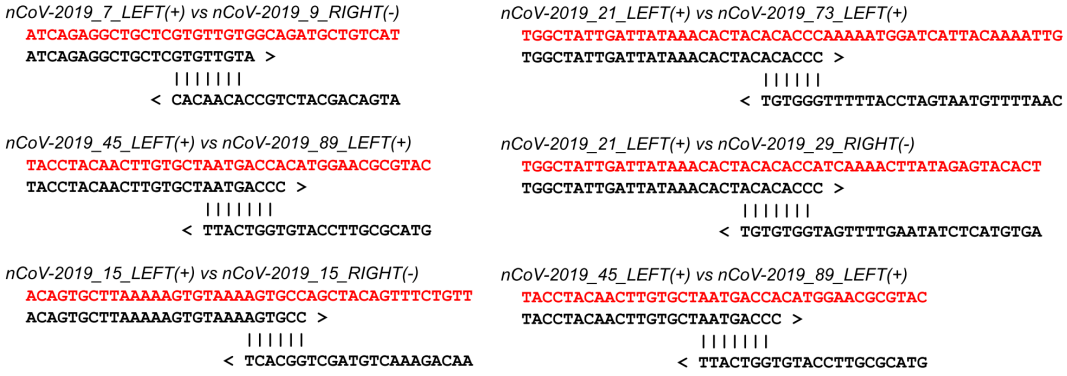

Fig 3 The six primer dimers observed in raw NGS reads. Observed dimer sequences are shown in red fonts.

We chose four primers, 2019_9_RIGHT, 2019_15_RIGHT, 2019_21_LEFT and 2019_89_LEFT in the pool_1 and replaced them with alternatives (Table 2) to prevent formation of all eight primer dimers depicted in Fig 3. We compared performances of original and the new primer set (both pool_1 and 2) for the same samples to Fig 2 but 35 PCR cycles was used instead of 30. The results of eight clinical samples with various Ct values are shown in supportive figure Fig S1. The result for a sample with moderate Ct value (same to the sample shown in Fig 2A) is shown in Fig 2B. Although using 35 PCR cycles resulted in better balanced yield of PCR product among samples with various viral loads compared to that with 30 cycles (Fig 4), which eases multiplexing in NGS library construction, this increase in number of PCR cycles enlarged coverage bias overall regardless when original or new primer sets was used (Fig 2AB and Fig S1). Nevertheless, improvement of coverages in the eight amplicons were seen with modified primer set (Fig 2B and Fig S1, highlighted by red strips).

Table 2 Modification of the four primers in pool_1 (V2)

add / remove	Primer name (nCoV-2019_XXX)	Location (1-base)	Direction	Sequence
◀remove	9_RIGHT	2883..2904	minus	5′-**ATGACAGCATCTGCCACAACAC**-3′
add▶	9_RIGHTv2	2886..2907	minus	5′-**TTTATGACAGCATCTGCCACAA**-3′
◀remove	15_RIGHT	4675..4696	minus	5′-**AACAGAAACTGTAGCTGGCACT** -3′
add▶	15_RIGHTv2	4678..4699	minus	5′-**AGAAACAGAAACTGTAGCTGGC**-3′
◀remove	21_LEFT	6168..6196	plus	5′-**TGGCTATTGATTATAAACACTACACACCC**-3′
add▶	21_LEFTv2	6165..6193	plus	5′-**TGGTGGCTATTGATTATAAACACTACACA**-3′
◀remove	89_LEFT	26836..26857	plus	5′-**GTACGCGTTCCATGTGGTCATT**-3′
add▶	89_LEFTv2	26833..26854	plus	5′-**CGCGTACGCGTTCCATGTGGTC**-3′

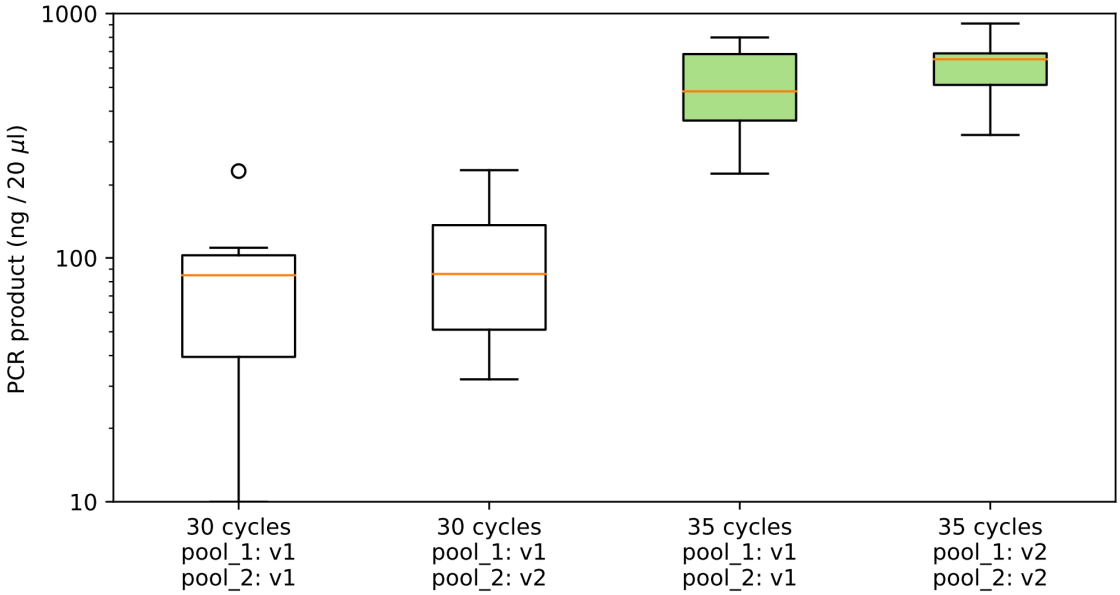

Fig 4 Yields of multiplex PCR product for eight clinical samples after pooling pool_1 and _2 reactions and purification by AmpureXP in different numbers of PCR cycles and primer sets.

Discussion

Since we published the first version of this manuscript which described about the modification of nCoV-2019_76_RIGHT primer, the ARTIC Network group has updated their primer set to 'V2' to cope with the dropout of the amplicon 18 and 76 (https://github.com/artic-network/artic-ncov2019/tree/master/primer_schemes/nCoV-2019/V2). Although the modification was done on the primer 2019_18_LEFT instead of 2019_76_RIGHT, their V2 primer set is likewise expected to improve the coverage of amplicon 18 and 76 as seen this study.

In this version of manuscript, we added result of further modification in four primers contained in pool_1 to prevent six dimer formations. This modification entangled interference among batch of primers which involved in as many as eight PCR targets. With this modified primer set, one could recover more genomic region particularly from samples with low viral loads (>30) with smaller sequencing effort.

The result of these experiments clearly indicates that primer dimer formation is one critical cause of coverage bias in ARTIC protocol. Importantly, some of those problems can be fixed easily by observing primer dimers in raw NGS sequence reads and replacing them with alternatives as shown in this study.

Materials and Methods

RNA extracted from clinical specimens (pharyngeal swabs) are reverse transcribed as

described in protocol published by ARTIC Network (Quick, 2020) but scaled downed to 1/4. The Ct values in clinical qPCR test for those samples ranged from 25 to 30. The cDNA was diluted to 10-fold by H_2O, and 1 µl of the diluted cDNA was used for 10 µl reaction Q5 Hot START DNA Polymerase (NEB) (2 µl of 5x buffer, 0.8 µl of 2.5 mM dNTPs, 0.1 µl of polymerase and 0.29 µl of 50 µM primer mix volumed-up by milli-Q water). It should be noted that this amount of cDNA template per PCR reaction was 4-fold less than that in our usual protocol because we were intended to save those clinical samples. The thermal program was identical to the original ARTIC protocol, and the numbers of CPR cycles were 30 for experiment in Fig 2 and 35 for experiment in Fig 4. The PCR products in pool_1 and pool_2 reactions for same clinical samples were combined and purified by 1x concentration of AmpureXP. The purified PCR product was subjected to illumina library prep using QIAseq FX library kit (Qiagen) in 1/4 scale and using 6 min fragmentation time. After the ligation of barcoded adaptor, libraries were heated to 65 ºC for 20 min to inactivate ligase, and then, all libraries were pooled in a 1.5 ml tube without balancing DNA concentrations. The pooled library was first purified by AmpureXP at x0.8 concentration, and then again at x1.2 concentration. The purified library was sequenced for 151 cycles at both paired-ends in Illumina iSeq100 along with other samples which were not involved in this study.

Obtained reads were mapped to the reference genome of SARS-CoV-2 MN908947.3 (Wu et al., 2020) by using *bwa mem* (Li and Durbin, 2009). To estimate the coverage of each PCR products, we counted depth of genomic parts only specific to each PCR product (Fig 4) using *samtools depth* function (Li and Handsaker et al., 2009) with '-a' option. The depth counts were summarized and visualized using the python3.6 and *matplotlib* library (Hunter 2007).

Supportive figure legend

Fig S1 Depth plots on the SARS-CoV2 genome for mapped NGS reads obtained from ARTIC protocol for all eight clinical samples. Both results with the original or alternative pool_2 primer set or with the original or alternative pool_1 and pool_2 primer sets are shown. It should be noted that the experiment was conducted for 4-fold diluted cDNA sample than our usual protocol. The green and red vertical strips highlight regions of amplicons potentially affected by dimer formation among primers in pool_1 or pool_2, respectively.

References

Hunter, J. D. 2007. Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

Johnston, A.D., Lu, J., Ru, K. et al. PrimerROC: accurate condition-independent dimer prediction using ROC analysis. Sci Rep 9, 209 (2019). https://doi.org/10.1038/s41598-

018-36612-9

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25: 1754–1760.

Li, H., B. Handsaker, A. Wysoker, et al., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25: 2078–2079.

Quick, J, 2020. nCoV-2019 sequencing protocol, protocol.io dx.doi.org/10.17504/protocols.io.bbmuik6w

Wu, F., S. Zhao, B. Yu, et al., 2020. A New Coronavirus Associated With Human Respiratory Disease in China, Nature, 579 (7798), 265-269