bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.985150; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A proposal of an alternative primer 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, 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 corona virus, SARS-CoV-2, soon after the epidemics of this pathogen was revealed. The primer set seems to have been adapted already by many researchers worldwide and contributed for the high-quality and prompt genome epidemiology of this potential pandemic virus. We have also seen the great performance of their primer set and protocol; the primer set was able to amplify all desired PCR products with fairy small amplification bias from clinical samples with relatively high viral load. However, we observed acute drop of reads derived from two particular PCR products, 18 and 76, out of the 98 designated products as sample's viral load decreases. We suspected the reason for this low coverage issue was due to dimer formation between primers used to amplify those two PCR products. Here, we propose replacing just one of those primers, nCoV-2019_76_RIGHT(-), to a newly designed primer. The result of the replacement of primer showed improvement in coverage at both regions targeted by the products, 18 and 76. We expect this simple modification will extend the limit for whole SARS-CoV-2 genome analysis to samples with lower viral load and enhance genomic epidemiology 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.985150; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(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 products 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 multiplex PCR for low-copy templates, we observed that coverages for the two particular PCR products, 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 products (each upper half of Fig1A and B). In our experience so far, low to absolute-zero depth for those two fragments 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 products 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 products, 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 multiplex reaction "pool_2". We noticed that two of those primers, nCoV-2019_18_LEFT(+) and nCoV-2019_76_RIGHT(-) are perfectly complement to each other by their 10-nt sequence at the 3'-end (Fig 2). From this observation, we reasoned that the rapid decrease of the amplification efficiencies of those PCR products was due to a primer dimer formation between nCoV-2019_18_LEFT(+) and nCoV-2019_76_RIGHT(-), that could complete to the intended amplification reactions. Indeed, we observed many NGS reads derived from the predicted dimer in raw FASTQ data (data not shown).

Then, we replaced one of those 'unlucky' primer pair, nCoV-2019_76_RIGHT(-), in the pool_2

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.985150; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

to a newly designed 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
2019_76_RIGHT(-)	23193..23214	5′-ACACCTGTGCCTGTTAAACCAT-3′
2019_76_RIGHTv2(-)	23241..23565	5′-TCTCTGCCAAATTGTTGGAAAGGCA-3′

Comparison of original and new primer set for 8 clinical samples with Ct-values ranging from 25 to 30 are shown in Fig 3. The replacement certainly added notable improvement in read depths at the regions covered by the products 18 and 76 (Fig 3 and upper half of each Fig 1A and B as an example). There was no notable adverse effect observed in other PCR products by this replacement of a primer.

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 number of cycles in PCR was fixed to 30 using thermal program identical to the original protocol. 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 by using *bwa mem* (Li and Durbin, 2009). To estimate the coverage of each PCR products, we

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.985150; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

counted depth of genomic parts only specific to each PCR product (Fig 4) using *samtools depth* function (Li and Handsaker et al., 2009). The depth counts were summarized and visualized in R (R Core Team, 2009) using ggplot2 (Wickham, 2016).

Figure Legend

Fig 1 Depth view for one example clinical sample after mapping short reads at the regions around PCR product 48 (A) and 76 (B). The upper half of each figure represent result obtained using original ARTIC primer set, and the lower half represents result using primer set with replacement of nCoV-2019_76_RIGHT(-) to nCoV-2019_76_RIGHTv2(-). The low coverage region in the original primer set are highlighted.

Fig 2 Predicted primer dimer formed by nCoV-2019_18_LEFT(+) and nCoV-2019_76_RIGHT(-) by PrimerROC (Johnston et al., 2019).

Fig 3 Distributions of depth at each base within each PCR product specific region defined in Fig 4 for 8 clinical samples. The red boxes indicate distribution when original ARTIC primes set was used. The blue boxes indicate distribution when nCoV-2019_76_RIGHT(-) was replaced to nCoV-2019_76_RIGHTv2(-). Depth for the PCR products 18 and 76 are highlighted. The reported Ct-values for each sample are described at right side of each plot. It should be noted that the experiment was conducted with 4-fold diluted cDNA sample than our usual protocol in a purpose to save samples to test several conditions.

Fig 4 A schematic diagram for defined regions used for depth analysis shown in Fig 3. White and black arrows indicate primers belonging to the pool_1 and pool_2, respectively.

References

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

R Core Team, 2009. R: A language and environment for statistical computing. 2009 R

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.985150; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

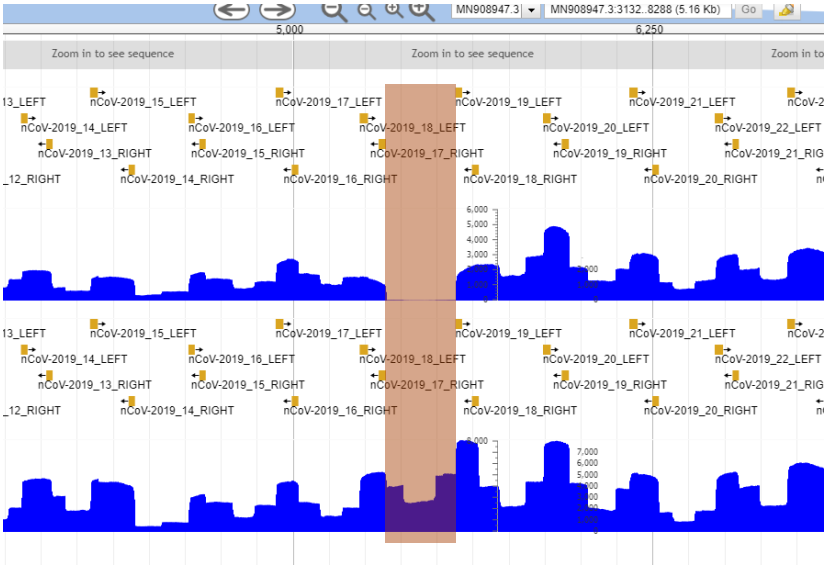
Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.985150; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(A)

Original primer set

With a new primer



(B)

Original primer set

With a new primer

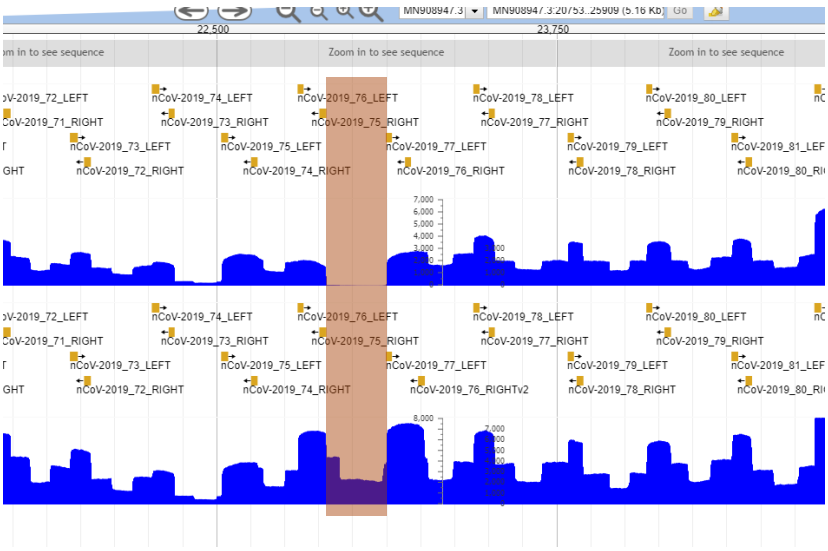
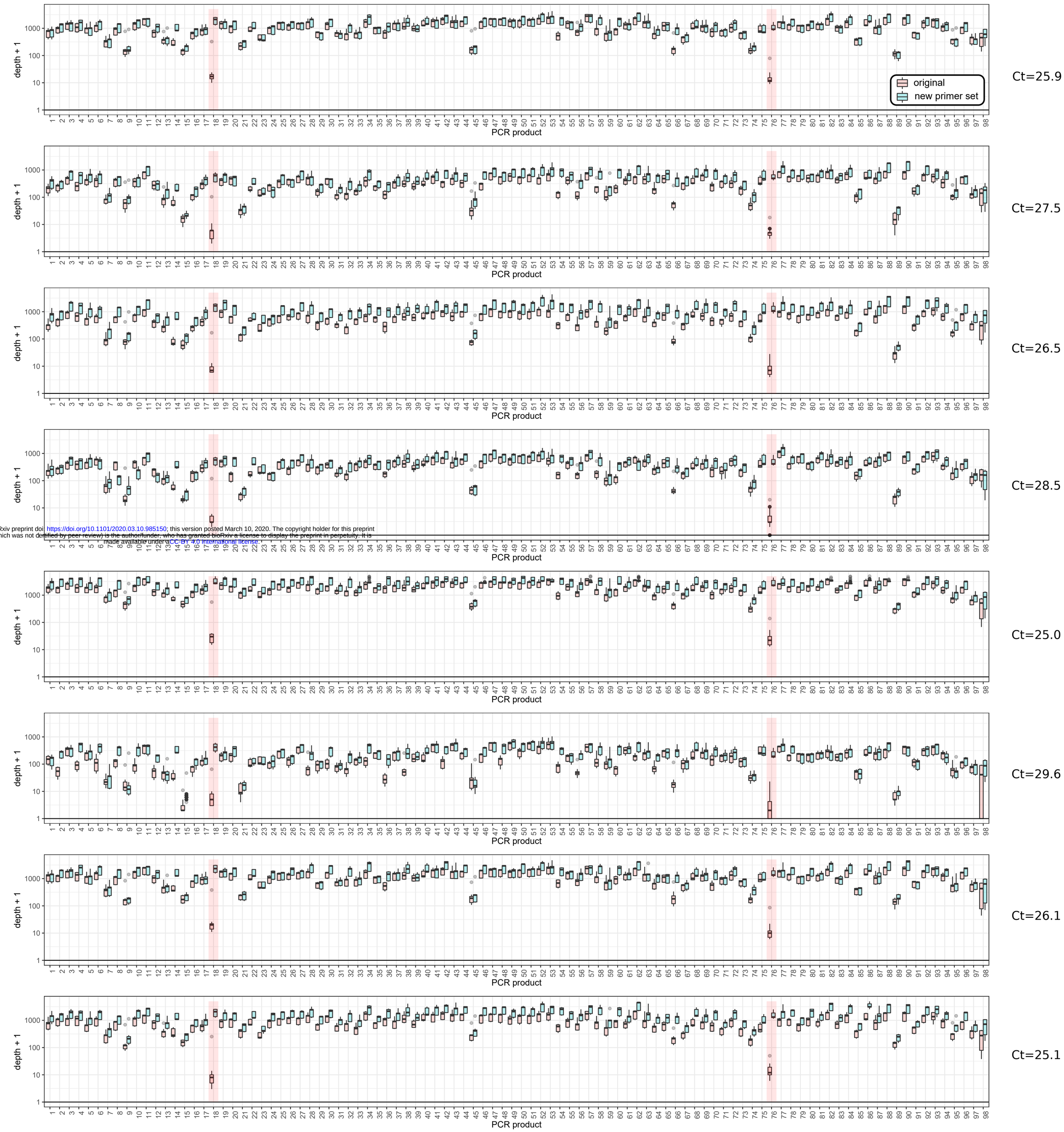


Fig. 1

```
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - <<<
  nCoV-2019_18_LEFT(+) vs nCoV-2019_76_RIGHT(-): -15.49 kcal/mol
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
 5'> TGGAAATACCCACAAGTTAATGGTTTAAC >3'
                          ||||||||||
                    3'< TACCAAATTGTCCGTGTCCACA <5'
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - <<<
```

Fig. 2



bioRxiv preprint doi https://doi.org/10.1101/2020.03.10.985150; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Fig 3

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.10.985150; this version posted March 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

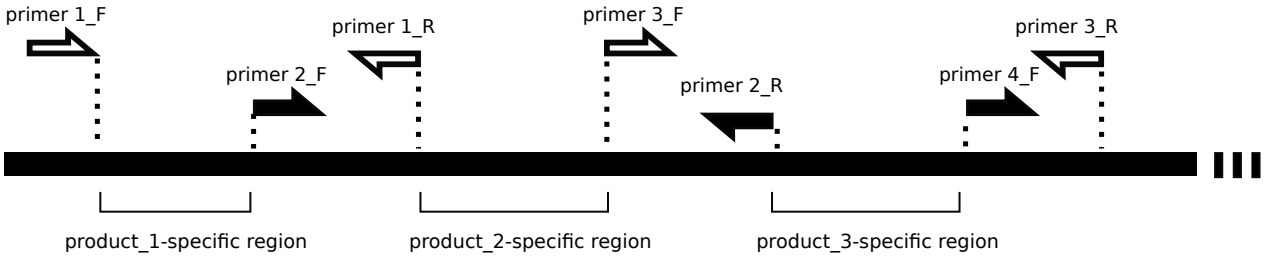


Fig. 4