Letter to the Editor

Full-Length Genome Sequencing of SARS-CoV-2 Directly from Clinical and Environmental Samples Based on the Multiplex Polymerase Chain Reaction Method*

NIU Pei Hua, ZHAO Xiang, LU Rou Jian, ZHAO Li, HUANG Bao Ying, YE Fei, WANG Da Yan, and TAN Wen Jie#

The new coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is spreading worldwide with the number of confirmed cases increasing dramatically[1,2]. Coronaviruses have the largest genome among RNA viruses and have caused two major outbreaks[3,4]. The new coronavirus (SARS-CoV-2) is the third coronavirus spread globally in the past 17 years. According to the World Health Organization, SARS-CoV-2 has affected more than 155.6 millon patients in 222 countries as of 6 May 2021 and has become a major global health concern (https://www.who.int/emergencies/diseases/novel-coronavirus-2019).

Rapid genome sequencing of the virus directly from clinical and environmental specimens is important for real-time genomic surveillance in managing virus outbreaks[5]. Genomic sequencing directly from clinical and environmental samples is challenging owing to low sensitivity. Therefore, we designed a multiplex PCR method for virus genome sequencing of SARS-CoV-2 directly from clinical and environmental samples with reference to the whole-genome sequencing method for Zika virus[6]. SARS-CoV-2 can be detected in the bronchoalveolar lavage fluid, sputum, nasal swab, throat swab, feces, blood, and saliva[7,8]. Hence, in this study, we validated the method with different types of these samples.

The complete protocol is schematically represented in Figure 1. The first step of the protocol was to design overlapping primer pairs to cover the entire genome (Supplementary Table S1 available in www.besjournal.com). The multiplex PCR comprised a set of 102 oligonucleotide primer pairs and the amplicons generated by the primer pairs spanned the target genome (Figure 1A). The subsequent step was the amplification and sequencing of fragments that were performed with Nanopore GridION X5 to obtain reads of 400 bp or Illumina paired-end library protocol, allowing reads of approximately 300 bp. During genome sequencing, some amplicons were over-sequenced, while others were under-sequenced (Figure 2). This occurs owing to differences in the amplification process and genome content.

We evaluated the sensitivity of the multiplex PCR sequencing method using a 10-fold gradient dilution of the reference strain (EPI_ISL_402119). The sequencing results for Nanopore GridION X5 showed that the mapping reads of six gradient dilution samples against the SARS-CoV-2 reference (EPI_ISL_402119) were 3,733.734 (99.71%), 3,639.845 (99.73%), 3,819.822 (99.65%), 2,934.107 (99.27%), 682.764 (91.12%), and 55.617 (44.7%), respectively (Figure 2B). We identified that as the cycle threshold (*Ct*) value increased, the percentage of SARS-CoV-2 reads and the percentage of SARS-CoV-2 genome coverage showed a downward trend (Figure 2A). Furthermore, when the Ct value was < 37, the coverage of the SARS-CoV-2 genome was > 95%. Interestingly, when the Ct value was > 40, the coverage of the SARS-CoV-2 genome was < 50%. The median depth and interquartile range (IQR) of coverage of all samples are listed in Figure 2B, whereas the results on the Illumina MiSeq platform are not presented.

In this study, we used 14 different types of clinical and environmental specimens, including bronchoalveolar lavage fluid, sputum, nasal swabs, throat swabs, and feces samples, from patients infected with the new coronavirus and environmental samples. An in-depth summary, including detailed information on sequence reads, depth distributions, and genome coverage per

doi: 10.3967/bes2021.100

*This work was supported by the National Key Research and Development Program of China [2016YFD0500301, 2020YFC0840900].

National Institute for Viral Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China

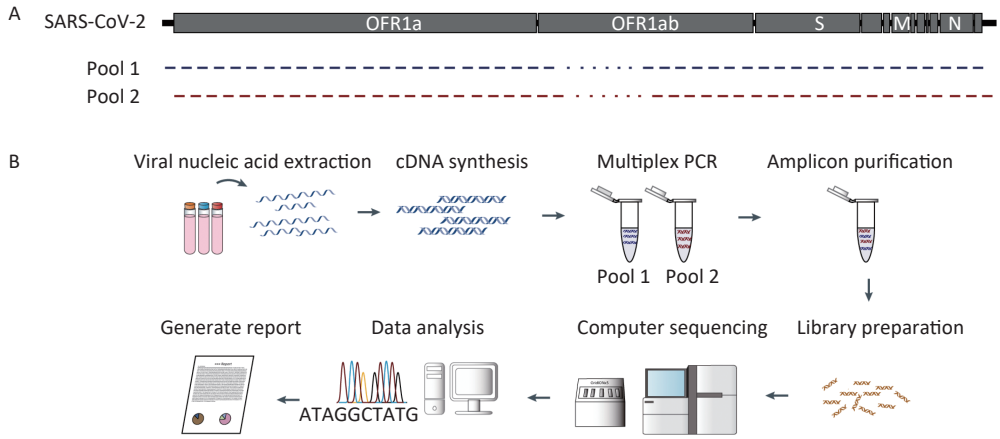
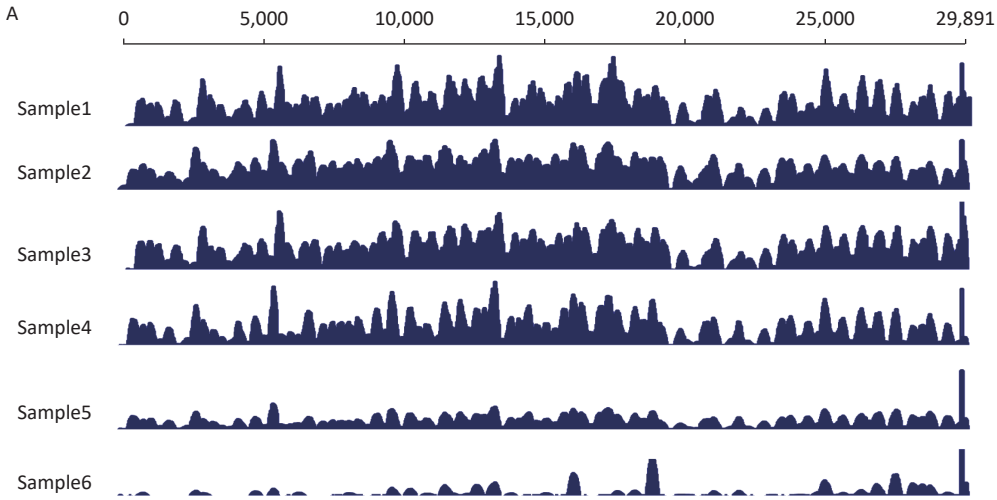


Figure 1. Sequencing schemes employed in the study. (A) Schematic showing expected amplicon products for each pool of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genome. We designed and optimized polymerase chain reaction (PCR) primers to generate amplicons that would span the SARS-CoV-2 genome. We designed 102 primer pairs, which covered 100% of the SARS-CoV-2 genome. The predicted forward primers (blue arrows) and reverse primers (red arrows) scaled according to the SARS-CoV-2 virus coordinates. (B) Workflow of the multiplex PCR method for sequencing SARS-CoV-2 directly from clinical specimens.



B Results and metrics of the 10-fold gradient dilution of the reference strain (EPI_ISL_402,119) using the multiplex PCR sequencing method.

Name	*Ct* value	Total reads	SARS-CoV-2 reads	SARS-CoV-2 (%)	Coverage (%)	Mesian depth [IQP]*
Sample1	21.47	3,744,706	3,733,734	99.71	99.82	4,443 [12,423]
Sample2	24.38	3,649,652	3,639,845	99.73	98.75	2,345 [10,928]
Sample3	28.28	3,833,356	3,819,822	99.65	99.35	1,277 [6,914]
Sample4	32.28	2,955,718	2,934,107	99.27	96.29	644 [3,910]
Sample5	37.38	749,276	682,764	91.12	95.21	172 [748]
Sample6	40.12	124,418	55,617	44.70	47.00	0 [6]

*Note.**Interquartile range.

Figure 2. Study of the six diluted strains using the multiplex polymerase chain reaction sequencing method on a Nanopore GridION X5. (A) Analysis of the severe acute respiratory syndrome coronavirus 2 genome assembly at various read depths. The longest contiguous sequence produced at each read depth as a fraction of the full genome length of six diluted samples is shown. (B) Summary of statistical analysis for the sequencing results of six strains.

sample, of the outputs from all samples is presented in Table 1. In comparison with previous studies and despite the long genome of the virus, sequencing of the SARS-CoV-2 genome was very specific. The number of reads that mapped against the SARS-CoV-2 sequence was very high (> 96% for 10 of the samples), revealing that the primer design and the biological protocol led to some noise in the sequencing data. A total of 11 full-length or near-full-length SARS-CoV-2 genomes (> 99% genome coverage) were obtained from 14 libraries.

Sequencing depth was not uniform among the amplicons along the genome; some amplicons were over-sequenced (median depth: 8,832×), while others were only sequenced a few times or not at all. Quantitative results of samples showed that all samples with relatively low Ct values (high viral load) led to a complete or nearly complete assembly. Three of the samples, hCoV-19_Sample3, hCoV-19_Sample10, and hCoV-19_Sample14, with higher Ct values (lower viral load), led to a final assembly covering only 77.15%, 96.57%, and 98.99% of the genome sequence, respectively. The significant benefit of utilizing a multiplex PCR method for virus genome sequencing over a Sanger sequencing or an unbiased metagenomic approach was the substantial increase in the number of reads specific to the viral genome[6,9]. The number of merged reads was between 150754 and 2628756, providing sufficient sequencing data for reliable analysis of the 14 samples. We constructed a phylogenetic tree based on the full-length genome sequences of SARS-CoV-2 derived from sequencing (Supplementary Figure S1 available in www.besjournal.com).

A limitation of this study is that our method is not suitable for identifying novel viruses as primers are SARS-CoV-2 specific. Amplicon sequencing may result in incomplete genome coverage, especially when lower abundance viral genomes are present, and the loss of both 5′ and 3′ regions. To obtain complete genomes, it may be necessary to replace the problematic primers or adjust their concentration according to other primers.

In summary, our method showed advantageous prospects for generating new coronavirus sequences directly from clinical and environmental samples; however, further studies are required to confirm this finding. On a long-term basis, this method can potentially be used as a routine laboratory test to aid in treatment, vaccine design and deployment, infection control strategies, and surveillance.

[#]Correspondence should be addressed to TAN Wen Jie, Tel/Fax: 86-10-58900878, E-mail: tanwj28@163.com

Biographical note of the first author: NIU Pei Hua, female, born in 1985, PhD, majoring in medical virology research.

Received: March 5, 2021;

Accepted: July 27, 2021

Table 1. Results of amplicon scheme sequencing on fourteen SARS-CoV-2 positive clinical and environmental samples in China using the multiplex PCR sequencing method on Nanopore GridION X5

Sample name	Sample type	*Ct* value	Total reads	SARS-CoV-2 reads	SARS-CoV-2 (%)	Coverage (%)	Median depth [IQR][a]
hCoV-19_Sample1	Alveolar lavage fluid	27.43	1,272,772	1,255,249	98.62	99.85	5,298 [4,701]
hCoV-19_Sample2	Alveolar lavage fluid	31.71	1,789,508	1,735,201	96.97	99.83	5,951 [6,161]
hCoV-19_Sample3	Sputum	38.00	598,376	399,343	66.74	77.15	30 [1,041]
hCoV-19_Sample4	Sputum	19.06	1,865,190	1,837,254	98.50	99.81	5,492 [9,965]
hCoV-19_Sample5	Sputum	25.14	2,678,180	2,628,756	98.15	99.87	8,832 [6,412]
hCoV-19_Sample6	Nasal swabs	33.25	1,279,020	1,233,595	96.45	99.82	4,563 [5,926]
hCoV-19_Sample7	Throat swabs	30.68	2,031,002	1,982,592	97.62	99.83	7,609 [8,081]
hCoV-19_Sample8	Throat swabs	22.77	1,912,014	1,889,140	98.80	99.81	6,607 [4,459]
hCoV-19_Sample9	Throat swabs	22.17	1,724,740	1,706,097	98.92	99.84	6,398 [3,722]
hCoV-19_Sample10	Feces	36.44	312,000	150,863	48.35	96.57	751 [1,435]
hCoV-19_Sample11	Feces	25.95	1,022,978	1,008,660	98.60	99.82	3,768 [2,382]
hCoV-19_Sample12	Feces	26.06	1,771,730	1,753,368	98.96	99.83	7,078 [4,750]
hCoV-19_Sample13	Environmental samples	36.06	268,000	159,279	59.43	99.69	874 [1,411]
hCoV-19_Sample14	Environmental samples	36.54	312,150	150,754	48.30	98.99	732 [1,390]

Note. [a]Interquartile range.

REFERENCES

1. Lu R, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet, 2020; 395, 565−74.
2. Zhu N, Zhang D, Wang W, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med, 2020; 382, 727−33.
3. Ksiazek TG, Erdman D, Goldsmith CS, et al. A novel coronavirus associated with severe acute respiratory syndrome. N Engl J Med, 2003; 348, 1953−66.
4. Zaki AM, van Boheemen S, Bestebroer TM, et al. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med, 2012; 367, 1814−20.
5. Gardy J, Loman NJ, Rambaut A. Real-time digital pathogen surveillance - the time is now. Genome Biol, 2015; 16, 155.
6. Quick J, Grubaugh ND, Pullan ST, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc, 2017; 12, 1261−76.
7. Wang W, Xu Y, Gao R, et al. Detection of SARS-CoV-2 in Different Types of Clinical Specimens. JAMA, 2020; 323, 1843−4.
8. To KK, Tsang OT, Yip CC, et al. Consistent Detection of 2019 Novel Coronavirus in Saliva. Clin Infect Dis, 2020; 71, 841−3.
9. Faria NR, Sabino EC, Nunes MR, et al. Mobile real-time surveillance of Zika virus in Brazil. Genome Med, 2016; 8, 97.