


SHORT COMMUNICATION

High-coverage SARS-CoV-2 genome sequences acquired by target capture sequencing

Shaoqing Wen^{1,2}  | Chang Sun² | Huanying Zheng³ | Lingxiang Wang² | Huan Zhang³ | Lirong Zou³ | Zhe Liu³ | Panxin Du² | Xuding Xu² | Lijun Liang³ | Xiaofang Peng³ | Wei Zhang³ | Jie Wu³ | Jiyuan Yang² | Bo Lei² | Guangyi Zeng⁴ | Changwen Ke³ | Fang Chen⁴ | Xiao Zhang^{1,5}

¹Guangzhou Regenerative Medicine and Health Guangdong Laboratory, Guangzhou, China

²Institute of Archaeological Science, School of Life Sciences, Fudan University, Shanghai, China

³Guangdong Provincial Center for Disease Control and Prevention, Guangzhou, China

⁴MGI Shenzhen, BGI Shenzhen, Shenzhen, China

⁵CAS Key Laboratory of Regenerative Biology, Joint School of Life Sciences, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou, China

Correspondence

Shaoqing Wen and Xiao Zhang, Guangzhou Regenerative Medicine and Health Guangdong Laboratory, Guangzhou 510700, China.
Email: wenshaoqing@fudan.edu.cn (SW) and zhang_xiao@gibh.ac.cn (XZ)

Abstract

In this study, we designed a set of SARS-CoV-2 enrichment probes to increase the capacity for sequence-based virus detection and obtain the comprehensive genome sequence at the same time. This universal SARS-CoV-2 enrichment probe set contains 502 120 nt single-stranded DNA biotin-labeled probes designed based on all available SARS-CoV-2 viral sequences and it can be used to enrich for SARS-CoV-2 sequences without prior knowledge of type or subtype. Following the CDC health and safety guidelines, marked enrichment was demonstrated in a virus strain sample from cell culture, three nasopharyngeal swab samples (cycle threshold [C_t] values: 32.36, 36.72, and 38.44) from patients diagnosed with COVID-19 (positive control) and four throat swab samples from patients without COVID-19 (negative controls), respectively. Moreover, based on these high-quality sequences, we discuss the heterozygosity and viral expression during coronavirus replication and its phylogenetic relationship with other selected high-quality samples from the Genome Variation Map. Therefore, this universal SARS-CoV-2 enrichment probe system can capture and enrich SARS-CoV-2 viral sequences selectively and effectively in different samples, especially clinical swab samples with a relatively low concentration of viral particles.

KEYWORDS

gene expression, genetic networks, genetic variability, mutation, SARS coronavirus

1 | INTRODUCTION

The outbreak of the novel coronavirus (SARS-CoV-2) disease has become a global and ongoing health concern. Since a patient with pneumonia of unknown etiology was first reported in the city of Wuhan on 30 December 2019, epidemiological, clinical, radiological, laboratory and genomic findings of this virus were gradually discovered by Chinese and international experts.¹ At the current stage of

research, however, two crucial topics must be addressed. First, according to the latest diagnostic criteria, reverse-transcriptase polymerase chain reaction (RT-PCR) assays are recommended as the standard diagnosis of SARS-CoV-2-infection. However, present studies found that some patients have typical imaging findings, including ground-glass opacity, but negative RT-PCR results.² The false-negative RT-PCR results can be caused by many factors, especially the insufficient detection sensitivity in a low viral load scenario.²

Shaoqing Wen, Chang Sun, and Huanying Zheng contributed equally to this study.

Second, more work must be done to monitor the virus mutation, and these mutations influence of disease severity and progression. Necessitating the full-length of the SARS-CoV-2 genome, metagenome sequencing technology is the latest and most comprehensive approach³⁻⁶ but still costly. Moreover, in the metagenome sequencing library, there are significant amounts of host (human) nucleic acid contamination and carrier RNA contamination introduced in commercial RNA extraction kits, both of which impair the amount of viral sequence readout.

In this context, we developed a set of SARS-CoV-2 enrichment probes by using hybridization capture technology to increase the sensitivity of sequence-based virus detection and characterization. This method was first used to enrich sequence targets from the human genome⁷ and then from vertebrate virome.⁸ The enrichment probe set contains 502 single-stranded DNA biotin-labeled probes at 2× tiling designed based on all available SARS-CoV-2 viral sequences, downloaded from the Global Initiative on Sharing All Influenza Data (GISAID; <https://www.gisaid.org/>) on 1 February 2020, and it can be used to enrich for SARS-CoV-2 sequences without prior knowledge of type or subtype. In addition, the probes for human housekeeping genes (GAPDH, PCBP1, EIF3L, POLR2A, EIF3A, TGOLN2, TCEB3, CDK12, and BTBD7) were spiked in the probe set as internal controls for studying viral expression.

2 | MATERIALS AND METHODS

To evaluate the sensitivity and specificity, we tested the enrichment probe set by using a virus strain sample derived from cell culture, three nasopharyngeal swab samples collected from patients diagnosed COVID-19 (positive controls), and four throat swab samples were taken from patients without COVID-19 (negative controls), respectively. Blank control is RNase free water.

The SARS-CoV-2 virus isolation and culturing were reported previously,⁹ which followed the CDC guidelines and good practice in laboratory health and safety requirements. Experiments were performed with the approval of the W96-027B framework. The RT-PCR tests were performed on all samples following a previously described method.¹⁰ The RT-PCR test kits (Bioperfectus) were officially approved by China's National Medical Products Administration. The C_t values for all samples are listed in Table 1. Notably, the sample GDFS2020329 showed weakly positive RT-PCR results, and the C_t value was adjacent to the cut-off value (40) for positivity.

We divided the total RNA sample of the SARS-CoV-2 virus strain (20SF014) into six samples (with slightly different experimental conditions) (Table 1). Six virus strain samples, three positive samples, four negative samples, and one blank control were reverse-transcribed into complementary DNA, respectively, followed by the second-strand synthesis. Using the synthetic double-stranded DNA, all DNA libraries were constructed through DNA-fragmentation, end-repair, adaptor-ligation, and PCR amplification. Subsequently, library hybridization capture was performed by using the SARS-CoV-2 enrichment probe set. The enriched libraries were qualified with Agilent

2100 Bioanalyzer using Agilent High Sensitivity DNA Kit and equivalent double-stranded DNA libraries were pooled and transformed into a single-stranded circular DNA library through DNA-denaturation and circularization. DNA nanoballs were generated from single-stranded circular DNA by rolling circle amplification, then qualified with Invitrogen Qubit 2.0 Fluorometer (Thermo Fisher Scientific, Foster City, CA) and loaded onto the flow cell and sequenced with PE100 on the MGISEQ-2000 platform (MGI, Shenzhen, China). Detailed experimental protocol in the Chinese and English version is presented in the Supporting Information Doc S1.

The Cutadapt (version 2.7) and trimmomatic (version 0.38) software were used for clipping adaptors and trimming low-quality reads. After removing the adaptor, low-quality, and low-complexity reads, high-quality reads were first filtered against the human reference genome (hg 38) using Burrows-Wheeler Alignment (MEM). The remaining nonhuman reads were then realigned to the SARS-CoV-2 reference (MN908947.3, <https://www.ncbi.nlm.nih.gov/nucleotide/MN908947>) using bowtie2 (version 2.3.4.1) and filtered reads according to mapping quality ($-q$ 30) by SAMtools (version 1.10). The variant was called by SAMtools and VarScan (version 2.3.9, parameter: $-\text{strand-filter } 0 -\text{min-avg-qual } 30 -\text{min-reads2 } 15 -\text{min-coverage } 15$). Finally, the sample consensus sequence was created by SAMtools and BCFtools (version 1.9) according to the variants called above.

3 | RESULTS AND DISCUSSION

The summary statistics for each enrichment library are described in Table 1. For virus strain sample (library 1-6) and three positive samples, 4 797 881-199 421 414 and 64 347 810-86 471 821 unique reads were obtained, of which 4 412 665-192 545 532 and 4750-756 973 reads (reflecting the viral RNA copy number [inversely related to C_t value]) were mapped to SARS-CoV-2 reference sequence (MN908947.3), respectively. For four negative and one control samples, none reads were mapped to the SARS-CoV-2 reference sequence. The fraction of SARS-CoV-2 endogenous DNA from virus strain enrichment libraries were found to be between 90.07% and 96.58%, demonstrating that the numbers of mapped reads to SARS-CoV-2 reference sequence significantly increased compared with metagenomic sequencing technology. The library complexity is evaluated by cluster factor, which is defined by "the number of raw reads divided by the number of reads after removing duplicates." In all enrichment libraries, the clustering factor is less than 1.5, with 1 being the best value for library construction. Notably, when adding the PCR cycle numbers of library amplification from 15 to 17, the library quality improves. Moreover, by merging the data from six virus strain enrichment libraries, we obtained a total of 371 981 580 unique reads, among which 358 112 573 reads were mapped to SARS-CoV-2 reference. Using these unique SARS-CoV-2 fragments from the virus strain sample, we reconstructed six SARS-CoV-2 genomes (mean depth being 186 869× and minimum coverage 13 816×). Only the merged sequence (coverage 1 121 217×) was

TABLE 1 Summary statistics of the enrichment libraries in this study

Sample ID	Sample type	RT-PCR (C _t value)	Volume, µL	PCR circles of library amplification	Total_reads_raw	Total_reads_rmdup	Map_SARS-CoV- 2_Reads_Reads	SARS-CoV- 2_Endo_Ratio	Cluster_ Factor	Mean_depth
20200217A_1	Virus strain	33	9	15	5 841 044	5 036 902	4 813 734	0.95569	1.15965	15 070.33
20200217A_2	Virus strain	33	9	15	5 738 560	4 797 881	4 412 665	0.91971	1.19606	13 816.72
20200217A_3	Virus strain	33	7	15	6 125 219	5 150 484	4 638 866	0.90067	1.18925	14 526.37
20200217A_4	Virus strain	33	7	15	8 837 914	7 432 367	6 694 515	0.90072	1.18911	20 966.49
20200217A_5	Virus strain	33	6	17	170 189 070	150 142 532	145 007 261	0.96580	1.13352	454 007.04
20200217A_6	Virus strain	33	7	17	224 071 046	199 421 414	192 545 532	0.96552	1.12361	602 829.61
GDFS2020309	Nasopharyngeal swabs	36.72	10	15	90 343 603	86 471 821	31 821	0.00037	1.04478	98.92
GDFS2020329	Nasopharyngeal swabs	38.44	10	15	97 086 902	79 596 774	4750	0.00006	1.21973	14.76
GDFS2020336	Nasopharyngeal swabs	32.36	10	15	97 608 772	64 347 810	756 973	0.01176	1.51689	2370.64
MG1056Z14D	Throat swab	...	20	13	13 062 316	12 121 570	0	0	1.07761	0
MG1057Z15A	Throat swab	...	20	13	18 048 427	15 685 239	0	0	1.15066	0
MG1066Z17B	Throat swab	...	20	13	19 511 564	15 694 790	0	0	1.24319	0
MG1076Z19D	Throat swab	...	20	13	36 170 265	31 850 222	0	0	1.13564	0
Blank control	RNase free water	...	20	13	105 835	91 275	0	0	1.15952	0

Abbreviation: RT-PCR, reverse-transcriptase polymerase chain reaction.

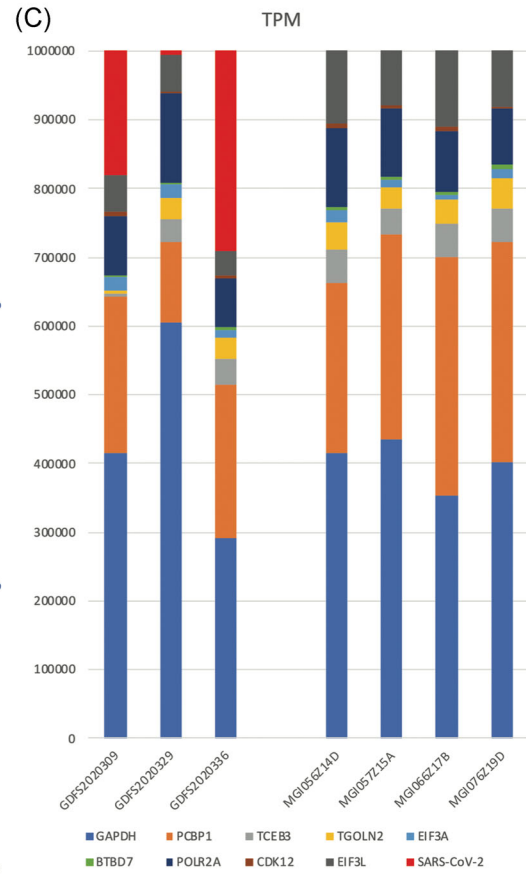
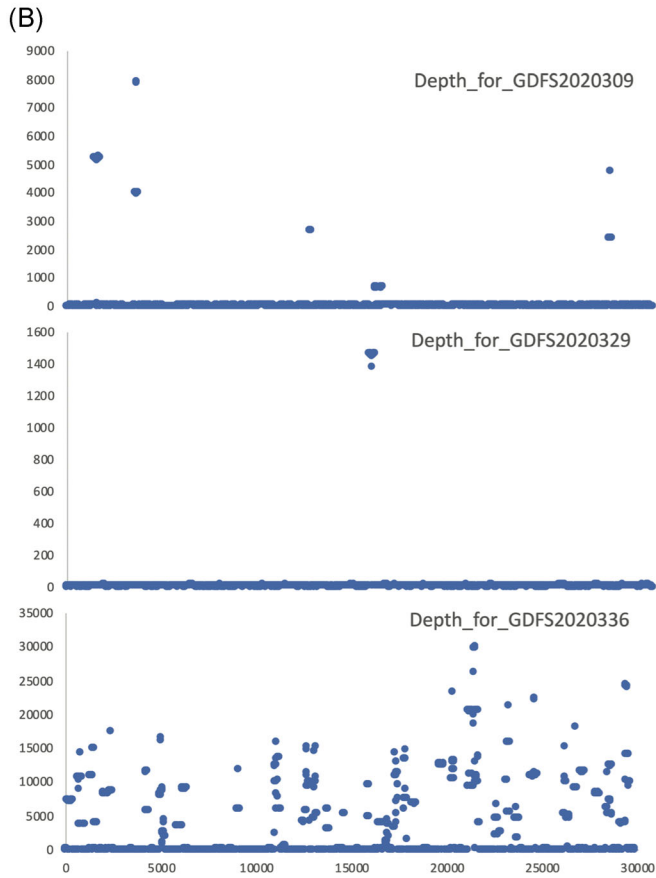
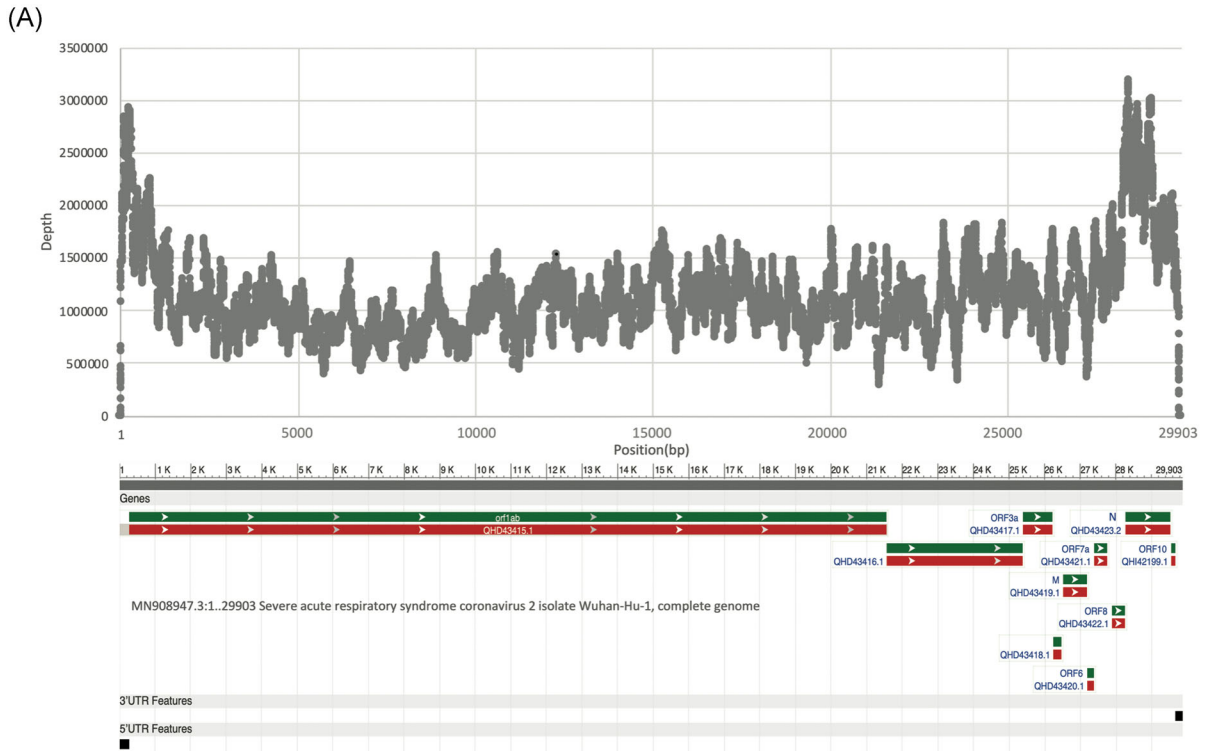


FIGURE 1 Sequencing depth statistics and transcripts per million (TPM) statistics. A, Sequencing depth of the virus strain sample, corresponding to SARS-CoV-2 genome reference (MN908947.3). B, Sequencing depths of three positive samples. C, TPM statistics of the positive and negative samples

used for further analysis. For three positive samples, we also reconstructed three SARS-CoV-2 genomes (mean depth 98.92 \times , 14.76 \times , and 2370.64 \times , respectively). Their C_t values are 36.72, 38.44, and 32.36, accordingly. Finally, for the virus strain sample, there are five variants called from merged data, including one homozygous variant at SNP (T23569C), and four heterozygous variants (three SNPs: C4534T, A5522T, C23525T, and one deletion: CT16779C). For three positive samples, GDFS2020309 has two homozygous variants: C23525T, CT27791C, and heterozygous variants T23569C; GDFS2020336 has two homozygous variants: C635T and C29303T; GDFS2020329 has no variant. The phenomenon of heterozygosity had been reported in previous studies,^{6,11} we propose that this heterogeneity could be caused by the mutations that occur during viral replication or the infection by multistrain of coronavirus.

We collected the variations information (gff3 files) of high-quality samples from the Genome Variation Map (<ftp://download.big.ac.cn/GVM/Coronavirus/gff3/>) (on 22 March 2020). According to the quality criteria for 2019-nCoV delivered by National Genomics Data Center (2019nCoV; <https://bigd.big.ac.cn/ncov>),¹² we enrolled 601 samples with 45 SNVs at first and second levels (with MAF > 0.01 and no dense variation regions; see <https://bigd.big.ac.cn/ncov/variation/annotation>) in the following analysis. The information of raw variations in the gff3 file is recoded into the binary format as an input file for Network analysis (Network version 5; www.fluxus-engineering.com; Table S1). Five clades could be identified and labeled, corresponding to the full genome tree delivered by GISAID (see Figure S1). Except for three main larger clades (named: S:ORF8-L84S [defined by SNP: 28144], G: S-D614G [SNP: 23403], V:NS3-G251V [SNP: 26144]), we defined a new clade, clade I: orf1ab-V378I (segregating at position 1397). The haplotype of the reference genome (MN908947) is in the central clade (yellow circle), and our samples (20200217A, GDFS2020309, GDFS2020329, and GDFS2020336) are also in this clade.

In Figure 1A, we found two peaks in genome sequencing depths, one covering the 5'-UTR region (MN908947.3:1-256) and another covering the N region (MN908947.3:28274-29533), which may be associated with the high expression in these two regions during replication of coronavirus.^{13,14} For high sequencing depths in 5'-UTR region, a reasonable explanation is that 5'-UTRs before ORF1a is necessary for the discontinuous synthesis of subgenomic RNAs in the beta coronaviruses and contains the cis-acting sequences necessary for viral replication.¹² Clinically, N gene RT-PCR assay was found to be more sensitive than other genes in SARS-CoV-2 detection, which is consistent with our finding of high sequencing depths in N region. This can be explained as the structural composition of coronavirus, also the difference in expression regulation in the host cells regarding subgenomic mRNA.¹⁴⁻¹⁶ In Figure 1B, however, there were no typical depth peaks found in the 5'-UTR region and N region in positive samples. We suggest that a larger sample size is needed to evaluate the divergent expression pattern in the future.

In general, in our selected human housekeeping genes (GAPDH, PCBP1, EIF3L, POLR2A, EIF3A, TGOLN2, TCEB3, CDK12, and BTBD7), GAPDH exhibited a relatively high expression level, PCBP1,

EIF3L and POLR2A showed a moderate expression level, and the rest genes had a relatively low expression level. This gene expression pattern was clearly shown in all positive and negative samples (see Figure 1C). Importantly, according to the transcripts per million statistics, we found that positive samples (GDFS2020336 [C_t value: 32.36], GDFS2020309 [C_t value: 36.72], and GDFS2020329 [C_t value: 38.44]) exhibited the high, moderate, and low expression level (red bar), respectively, which was nearly equivalent to that of gene GAPDH, PCBP1, and BTBD7 (Figure 1C).

In the current study, we, based on the available SARS-CoV-2 virus sequences, designed a set of SARS-CoV-2 enrichment probes. We made six enrichment libraries from one cultured SARS-CoV-2 virus strain and seven enrichment libraries from three positive samples (especially a weakly positive sample) and four negative samples to test the enrichment effects and sequenced them on MGISEQ-2000 platform. Overall, the SARS-CoV-2 enrichment probe described in this study showed significant, SARS-CoV-2-specific enrichment and should be a useful tool for the SARS-CoV-2 research community for detecting SARS-CoV-2 RNA in low amounts and for monitoring the future mutations.

ACKNOWLEDGMENTS

The authors sincerely thank those who are on the front lines battling the SARS-CoV-2 virus. We also thank the technical support provided by Guangzhou Koalson Bio-Technique Co Ltd. Groups interested in testing this protocol can request guidance by emailing wenshaoqing@fudan.edu.cn, and a limited number of our SARS-CoV-2 enrichment probe set are available on request.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

AUTHOR CONTRIBUTIONS

SQW and XZ were involved in designing the study and preparing the manuscript. HYZ, HZ, LRZ, ZL, LJL, XFP, WZ, JW, JYY, BL, and GYZ performed most of the experiments. SQW, CS, LXW, PXD, and XDX analyzed the data. CWK, FC, and XZ contributed to the critical revision of the manuscript. The corresponding authors were responsible for all aspects of the study and ensured that issues related to the accuracy or integrity of any part of the work were investigated and resolved. All authors reviewed and approved the final version of the manuscript.

ORCID

Shaoqing Wen  <http://orcid.org/0000-0003-1223-4720>

REFERENCES

- Chan JF, Yuan S, Kok KH, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet*. 2020;395(10223):514-523.
- Ai T, Yang Z, Hou H, et al. Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases [published online ahead of print February 26, 2020]. *Radiology*. 2020:200642. <https://doi.org/10.1148/radiol.2020200642>

3. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579(7798):265-269.
4. Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020; 579(7798):270-273.
5. Lu R, Zhao X, Li J, et al. Genomic characterization and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*. 2020;395(10224):565-574.
6. Tang XL, Wu CC, Li X, et al. On the origin and continuing evolution of SARS-CoV-2. *Natl Sci Rev*. 2020:nwaa036. <https://doi.org/10.1093/nsr/nwaa036>
7. Tewhey R, Nakano M, Wang X, et al. Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biol*. 2009;10(10):R116.
8. Briese T, Kapoor A, Mishra N, et al. Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. *mBio*. 2015;6(5):e01491-15.
9. Li HC, Ma J, Zhang H, et al. First Isolation and Identification of SARS-CoV-2 in Guangdong province, China. *Chinese J Virol*. 2020;43: 396-400. <https://doi.org/10.13242/j.cnki.bingduxuebao.003657>
10. Zou L, Ruan F, Huang M, et al. SARS-CoV-2 viral load in upper respiratory specimens of infected patients. *N Engl J Med*. 2020;382(12):1177-1179.
11. Chen L, Liu W, Zhang Q, et al. RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. *Emerg Microbes Infect*. 2020;9(1):313-319.
12. Zhao WM, Song SH, Chen ML, et al. The 2019 novel coronavirus resource. *Hereditas*. 2020;42:212-221.
13. Yang D, Leibowitz JL. The structure and functions of coronavirus genomic 3' and 5' ends. *Virus Res*. 2015;206:120-133.
14. Chu DKW, Pan Y, Cheng SMS, et al. Molecular diagnosis of a novel coronavirus (2019-nCoV) causing an outbreak of pneumonia. *Clin Chem*. 2020;66:549-555.
15. Shen Z, Xiao Y, Kang L, et al. Genomic diversity of SARS-CoV-2 in coronavirus disease 2019 patients [published online ahead of print March 4, 2020]. *Clin Infect Dis*. 2020:ciaa203. <https://doi.org/10.1093/cid/ciaa203>
16. Wang C, Liu Z, Chen Z, et al. The establishment of reference sequence for SARS-CoV-2 and variation analysis. *J Med Virol*. 2020;92(6): 667-674.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Wen S, Sun C, Zheng H, et al. High-coverage SARS-CoV-2 genome sequences acquired by target capture sequencing. *J Med Virol*. 2020;92:2221-2226. <https://doi.org/10.1002/jmv.26116>