Title: Rapid, sensitive, full genome sequencing of Severe Acute Respiratory Syndrome Virus Coronavirus 2 (SARS-CoV-2)

Authors: Clinton R Paden[1], Ying Tao[1], Krista Queen, Jing Zhang, Yan Li, Anna Uehara, Suxiang Tong[2]

[1]These first authors contributed equally to this manuscript

[2]Corresponding author: sot1@cdc.gov

Affiliations:

Centers for Disease Control and Prevention, Atlanta, GA (all authors)

Keywords: COVID-19, SARS-CoV-2, coronavirus, genomics, High-Throughput Nucleotide Sequencing, Whole Genome Sequencing

Abstract

SARS-CoV-2 recently emerged, resulting a global pandemic. Rapid genomic information is critical to understanding transmission and pathogenesis. Here, we describe validated protocols for generating high-quality full-length genomes from primary samples. The first employs multiplex RT-PCR followed by MinION or MiSeq sequencing. The second uses singleplex, nested RT-PCR and Sanger sequencing.

In December 2019, SARS-CoV-2, the etiological agent of Coronavirus Disease 2019 (COVID-19), emerged in Wuhan, China. Since then it has rapidly spread to the rest of the world (1-3). As of April 16, 2020, there have been 1,991,562 confirmed cases, including 130,885 deaths, in 185 countries or regions (4).

Initial sequencing of SARS-CoV-2 showed limited genetic variation between cases, but did document specific changes that may be useful for understanding the source and transmission chains (5-8). Because SARS-CoV-2 has shown the capacity to spread rapidly and lead to a range of presentations in infected persons, from asymptomatic infection to mild, severe, or fatal disease, it is important to identify genetic variants in order to understand any changes in transmissibility, tropism, and pathogenicity. Sequence data can be used to inform decisions to better manage the spread of disease.

In this report, we describe the design and use of two PCR-based methods for sequencing SARS-CoV-2 clinical specimens. The first is a multiplex PCR panel followed by sequencing on either the Oxford Nanopore MinION or Illumina MiSeq. When coupled with MinION sequencing, the protocol can be implemented outside a traditional laboratory and can be completed in a single workday, similar to previous mobile genomic surveillance of Ebola and Zika virus outbreaks (9, 10). Additionally, we provide a complementary singleplex, nested PCR strategy, which improves sensitivity for samples with lower viral load and is compatible with Sanger sequencing.

The Study

On January 10, 2020, the first SARS-CoV-2 genome sequence was released online (11). That day, we designed two complementary panels of primers to amplify the virus genome for

sequencing. For one panel, we used the PRIMAL primer design tool (9) to design multiplex PCRs to amplify the genome in using only a few PCR reactions (Appendix). The final design consists of 6 pools of primers, targeting amplicon sizes of 550 base pairs (bp) with 100bp overlaps, to allow for sequencing on either the ONT MinION or Illumina MiSeq. For the second panel, we designed sets of primers to generate nested, tiling amplicons across the SARS-CoV-2 genome (Appendix), for enhancing sensitivity in samples with lower viral loads. Each amplicon is 322-1030bp with an average overlap of 80bp. They are designed to be amplified and sequenced individually on Sanger instruments but may also be pooled for sequencing on next-generation sequencing platforms.

To determine the sensitivity of each sequencing strategy, we generated a set of six ten-fold serial dilutions of a SARS-CoV-2 isolate (12). Virus RNA was diluted into a constant background of A549 human cell line total nucleic acid (RNaseP C_T 29). We quantitated each dilution using the CDC SARS-CoV-2 rRT-PCR for the N2 gene (13) (data not shown). The six dilutions spanned C_T values from 22-37, corresponding to ca. 2 to 1.8 x 10^5 copies. We amplified triplicate samples at each dilution using the multiplex PCR pools. Next, we pooled, barcoded, and made libraries from each sample's amplicons using the ligation-based kit and PCR barcode expansion kit (methods in Appendix). MinION sequencing was performed on an R9.4.1 or R10.3 flow cell until we obtained >1-2M raw reads. From those, 50-60% of reads could be demultiplexed (data not shown). Additionally, we sequenced these amplicons using the Illumina MiSeq for comparison (methods in Appendix).

For MinION sequencing, the reads were basecalled and analyzed using an in-house read mapping pipeline (detailed in Appendix). For samples with $C_T \leq 29$, we obtained >99% SARS-CoV-2 reads and >99% genome coverage at 20X depth, decreasing to an average of 93%

63 genome coverage at C_T 33.2 and 48% at C_T 35 (Figure 1A and 1B). Further, we were able to
64 obtain full >20X genomes within the first 40-60 minutes of sequencing (Figure 1C).

65 Consensus accuracy, including SNPs and indels, is critical for determining coronavirus
66 lineage and transmission networks. For high consensus level accuracy, we filtered reads based on
67 length, mapped them to the reference sequence (RefSeq NC_045512), trimmed primers based on
68 position, and called variants with Medaka (https://github.com/nanoporetech/medaka) (details in
69 Appendix). Each Medaka variant was filtered by coverage depth (>20X) and by the Medaka
70 model-derived variant quality (>40). Here we used the variant quality score as a heuristic to filter
71 remaining noise from the Medaka variants, compared to Sanger-derived sequences. After these
72 steps, the data approaches 100% consensus accuracy (Table 1). Identical results were found
73 using the R9.4.1 pore, through the C_T 33.2 samples (data not shown). We noted larger deletions
74 in some of the C_T 33.2+ samples which likely reflect biases from limited copy numbers.

75 In the MiSeq data, we observed a similar trend in percent genome coverage at 100X
76 depth, and a slightly lower percent mapped reads, compared to Nanopore data (Figure 1A and
77 B). Increased read depth using the MiSeq potentially allows increased sample throughput,
78 however the number of available dual unique barcodes limits actual throughput.

79 For the nested, singleplex PCR panel, we amplified the same serial dilutions with each
80 nested primer set (methods in Appendix). The endpoint dilution for full genome coverage is
81 approximately C_T 35 (Figure 1B). At the C_T 37 dilution, we observed significant amplicon
82 dropout—at this dilution, there are <10 copies of the genome on average per reaction.

83 These protocols enabled rapid sequencing of the initial clinical cases of SARS-CoV-2 in
84 the United States. For these cases, we amplified the virus genome using the singleplex PCR

amplicons, sequencing them with both MinION and Sanger instruments to validate MinION consensus accuracy. The MinION produced full-length genomes in <20 minutes of sequencing, while Sanger data was available the following day.

We used the multiplex PCR strategy in subsequent SARS-CoV-2 clinical cases (n=167), ranging in C_T values from 15.7 to 40 (mean 28.8, median 29.1). In cases below C_T 33, we observed an average of 99.02% specific reads and 99.2% genome coverage at >20X depth (Figures 2A and 2B). Between C_T 30-33, genome coverage varied by sample, and declined dramatically at higher C_T values, analogous to the isolate validation data. For these samples, we multiplexed 20-40 barcoded samples per flowcell. Enough data is obtained with 60 minutes of MinION sequencing for most samples, though for higher titer samples 10-20 minutes of sequencing is sufficient (Figure 2C).

Up-to-date primer sequences, protocols, and analysis scripts are found at https://github.com/CDCgov/SARS-CoV-2_Sequencing/tree/master/protocols/CDC-Comprehensive. Data from this study is deposited in the NBCI SRA (BioProjects PRJNA622817 and PRJNA610248).

Conclusions

Full genome sequencing is an indispensable tool in understanding emerging viruses. Here we present two validated PCR target-enrichment strategies that can be used with MinION, MiSeq, and Sanger platforms for sequencing SARS-CoV-2 clinical specimens. This ensures that most labs have access to one or more strategies.

The multiplex PCR strategy is effective at generating full genome sequences up to C_T 33. The singleplex, nested PCR is effective up to C_T 35, varying based on sample quality. The

turnaround time for the multiplex PCR MinION protocol is about 8 hours from nucleic acid to consensus sequence, compared to Sanger sequencing at about 14-18 hours (Table 2). Importantly, the multiplex PCR protocols offer an efficient, cost-effective, scalable system, and add little time and complexity as sample numbers increase (Table 2). The results from this study suggest multiplex PCR may be used effectively for routine sequencing, complemented by singleplex, nested PCR for low virus-titer samples and confirmation sequencing.

Acknowledgments

We would like to acknowledge the efforts of those in the Respiratory Viruses Branch at CDC who helped in organizing samples for this study, including Azaibi Tamin, Jennifer Harcourt, Natalie Thornburg, Shifaq Kamili, Xiaoyan Lu, and Stephen Lindstrom.

Disclaimers

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Author Bio (first author only, unless there are only 2 authors)

Clinton Paden is a virologist and bioinformatician in the CDC Pathogen Discovery Team, within the Division of Viral Diseases. His work is in identifying and characterizing novel and emerging pathogens.

References

1. Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, et al. First Case of 2019 Novel Coronavirus in the United States. The New England Journal of Medicine. 2020;382(10):929-36.
2. Patel A, Jernigan DB, Abdirizak F, Abedi G, Aggarwal S, Albina D, et al. Initial Public Health Response and Interim Clinical Guidance for the 2019 Novel Coronavirus Outbreak — United States, December 31, 2019–February 4, 2020. Morbidity and Mortality Weekly Report. 2020;69(5):140-6.
3. Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. Lancet (London, England). 2020;395(10223):470-3.
4. World Health Organization. Coronavirus disease 2019 (COVID-19) Situation Report 87. 2020 16 April 2020 [cited; Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
5. Andersen K. Clock and TMRCA based on 27 genomes. 2020 25 January 2020 [cited; Available from: http://virological.org/t/clock-and-tmrca-based-on-27-genomes/347
6. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nature Medicine. 2020 2020/03/17.
7. Bedford T, Greninger AL, Roychoudhury P, Starita LM, Famulare M, Huang M-L, et al. Cryptic transmission of SARS-CoV-2 in Washington State. 2020:2020.04.02.20051417.
8. Deng X, Gu W, Federman S, Du Plessis L, Pybus O, Faria N, et al. A Genomic Survey of SARS-CoV-2 Reveals Multiple Introductions into Northern California without a Predominant Lineage. 2020:2020.03.27.20044925.

9. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nature Protocols. 2017;12(6):1261-76.
10. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530(7589):228-32.
11. Holmes EC, Y Z. Novel 2019 coronavirus genome. 2020 [cited 2020 5 Apr]; Available from: http://virological.org/t/novel-2019-coronavirus-genome/319
12. Harcourt J, Tamin A, Lu X, Kamili S, Kumar Sakthivel S, Wang L, et al. Isolation and characterization of SARS-CoV-2 from the first US COVID-19 patient. 2020:2020.03.02.972935.
13. Kujawski SA, Wong KK, Collins JP, Epstein L, Killerby ME, Midgley CM, et al. First 12 patients with coronavirus disease 2019 (COVID-19) in the United States. 2020:2020.03.09.20032896.

Address for correspondence: Suxiang Tong, Division of Viral Diseases, Centers for Diseases Control and Prevention, 1600 Clifton Rd NE, Mailstop H18-6, Atlanta, GA 30329; email: sot1@cdc.gov

165 Table 1. Genome consensus accuracy

Virus tier (C_T)	%Coverage (20X)[a]	Indels	Indel bases	SNPs	%Identity[b]
22.3	99.659	0	0	0	100
	99.722	0	0	0	100
	99.635	0	0	0	100
25.7	99.635	0	0	0	100
	99.615	0	0	0	100
	99.642	0	0	0	100
29.2	99.508	0	0	0	100
	99.635	0	0	0	100
	99.615	0	0	0	100
33.2	93.024	1	1	0	100
	93.603	2	35	0	100
	87.894	0	0	0	100
35.6	41.653	1	1	0	100
	51.266	0	0	1	99.993
	50.821	1	15	2	99.987
37.6	14.634	0	0	1	99.977
	9.317	0	0	0	100
	12.363	0	0	0	100

166 [a] The 5’ and 3’ ends are primer sequence, so 100% coverage is not possible

167 [b] Percent of covered bases identical to reference sequence, excludes indels and low-coverage bases

168

169 Table 2. Comparison of input, time, and cost requirements for sequencing one or 96 specimens

		Single sample		96 Samples	
Method	**Input**[a]	**Turnaround time**	**Approx. cost per sample**[c]	**Turnaround time**	**Approx. cost per sample**[c]
Multiplex/MinION	10 uL	6-8 hours	$528.70	8-10 hours	$35.88
Multiplex/MiSeq	10 uL	30-68 hours[b]	$1443.29	30-68 hours[b]	$57.87
Singleplex/Sanger	190uL	16-18 hours	$354.40	17-19 days	$354.40

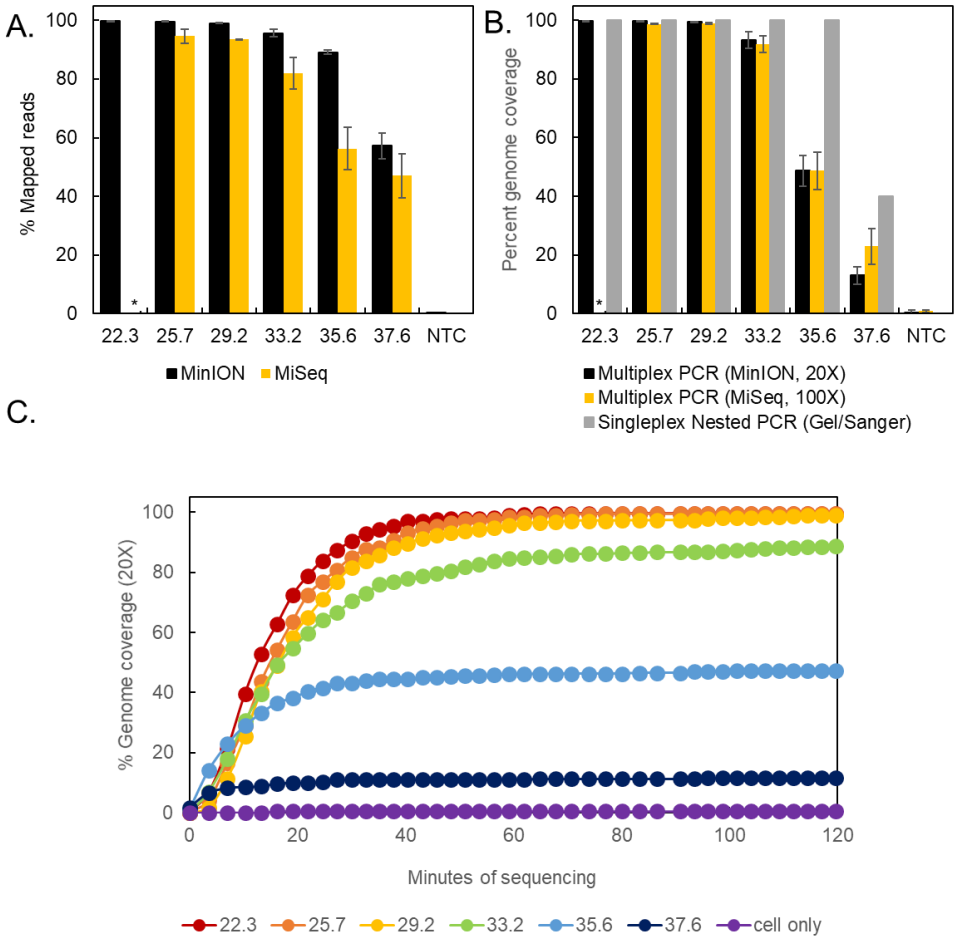
170 [a] Assumes a process of: 200uL resuspended respiratory specimen, extracted and eluted into 100uL

171 [b] Varies based on sequencing kit used

172 [c] Includes specific enzyme and reagent costs, excludes common laboratory supplies and labor costs

173

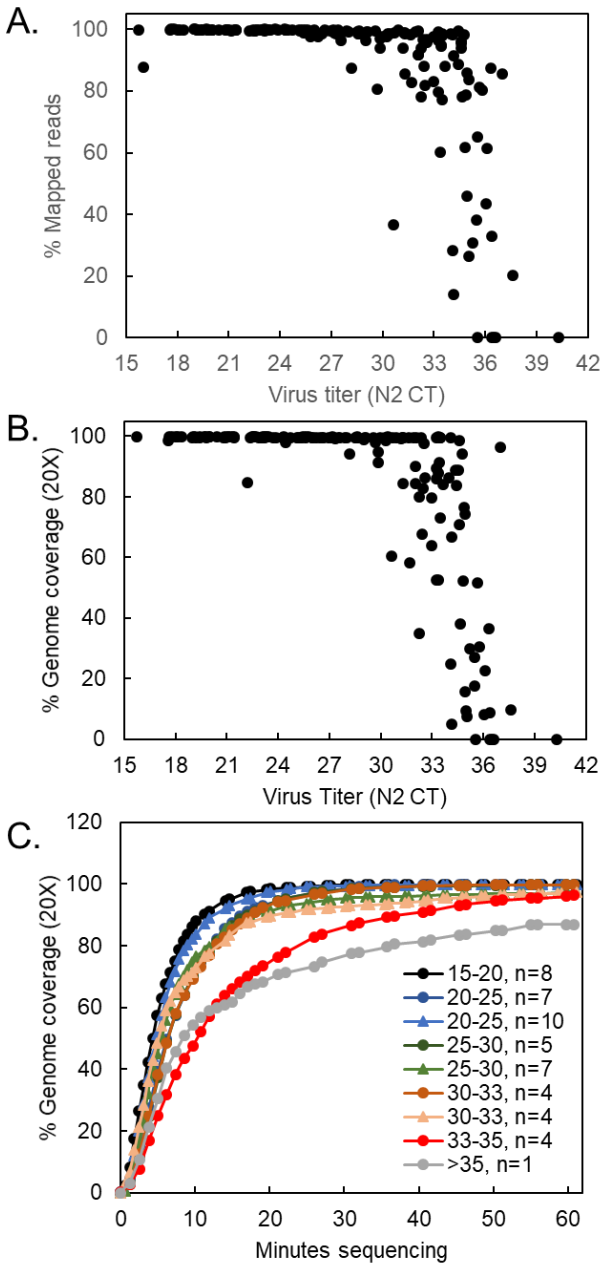
174 **Figure 1. Limits of detection.**



175

176 Triplicate serial dilutions of SARS-CoV-2 isolate A12 (12) amplified using the singleplex or
177 multiplex primer set. The multiplex amplicons were barcoded, library-prepped, and sequenced
178 on a MinION or MiSeq. (A) Percent of reads that map to the virus genome for each sample. (B)
179 Percent of virus genome that is covered at >20X depth by the multiplex amplicons on the
180 MinION (black) or >100X depth on the MiSeq (orange), or covered by the nested, singleplex
181 amplicons (grey) (measured by presence or absence on a gel). (C) Real-time analysis of MinION
182 sequencing data. Each data point represents the average 20X genome coverage of three
183 replicates.

184 **Figure 2. Sequencing SARS-CoV-2 clinical samples.**



185

186 (A) Percent mapped and (B) percent genome coverage for 167 clinical SARS-CoV-2 samples,

187 amplified with multiplex PCR strategy and sequenced on the MinION. (C) Time-lapse of 20X

188 genome coverage obtained by MinION for clinical specimens at the indicated C_T values. Data

189 points represent the average coverage for the indicated number of samples