A rapid, low cost, and highly sensitive SARS-CoV-2 diagnostic based on whole genome sequencing

Authors: Brian Glenn St Hilaire[1,2,3], Neva C. Durand[1,2,3], Namita Mitra[1,2,3], Saul Godinez Pulido[1,2,3], Ragini Mahajan[1,2], Alyssa Blackburn[1,2,3], Zane L. Colaric[1,2,3], Joshua W. M. Theisen,[1,3], David Weisz[1,2,3], Olga Dudchenko[1,2,3], Andreas Gnirke[1,4], Suhas S.P. Rao[1,2,3,5], Parwinder Kaur[7], Erez Lieberman Aiden[1,2,3,6,8,#], Aviva Presser Aiden[1,2,3,9,10,#]

[1]The Center for Genome Architecture, Baylor College of Medicine, Houston, TX 77030, USA
[2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX
[3]Center for Theoretical Biological Physics, Rice University, Houston, TX 77030, USA
[4]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA
[5]Department of Structural Biology, Stanford University School of Medicine, Stanford, CA 94305, USA
[6]Departments of Computer Science and Computational and Applied Mathematics, Rice University, Houston, TX 77030, USA
[7]UWA School of Agriculture and Environment, Faculty of Science, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia
[8]Shanghai Institute for Advanced Immunochemical Studies, Shanghai Tech University, Shanghai, China
[9]Department of Bioengineering, Rice University, Houston, TX, USA.
[10]Department of Pediatrics, Texas Children's Hospital, Houston, TX, USA
Co-corresponding authors

Abstract

Early detection of infection with SARS-CoV-2 is key to managing the current global pandemic, as evidence shows the virus is most contagious on or before symptom onset. Here, we introduce a low-cost, high-throughput method for diagnosis of SARS-CoV-2 infection, dubbed Pathogen-Oriented Low-Cost Assembly & Re-Sequencing (POLAR), that enhances sensitivity by aiming to amplify the entire SARS-CoV-2 genome rather than targeting particular viral loci, as in typical RT-PCR assays. To achieve this goal, we combine a SARS-CoV-2 enrichment method developed by the ARTIC Network (https://artic.network/) with short-read DNA sequencing and de novo genome assembly. We are able to reliably (>95% accuracy) detect SARS-CoV-2 at concentrations of 84 genome equivalents per milliliter, better than the reported limits of detection of almost all diagnostic methods currently approved by the US Food and Drug Administration. At higher concentrations, we are able to reliably assemble the SARS-CoV-2 genome in the sample, often with no gaps and perfect accuracy. Such genome assemblies enable the spread of the disease to be analyzed much more effectively than would be possible with an ordinary yes/no diagnostic, and can help identify vaccine and drug targets. Finally, we show that POLAR diagnoses on 10 of 10 clinical nasopharyngeal swab samples (half positive, half negative) match those obtained in a CLIA-certified lab using the Center for Disease Control's 2019-Novel Coronavirus test. Using POLAR, a single person can process 192 samples over the course of an 8-hour experiment, at a cost of ~$30/patient, enabling a 24-hour turnaround with sequencing and data analysis time included. Further testing and refinement will likely enable greater enhancements in the sensitivity of the above approach.

Introduction

There have been over 4 million cases of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection to date (as of 5/8/2020), claiming over 270,000 lives worldwide[3].

Knowing who is infected is a key first step towards pandemic containment. When a virus has a relatively high basic reproductive ratio (R_0) and evidence of asymptomatic transmission, early identification of infected individuals is critical[1,2]. High sensitivity (i.e. a low limit of detection, or LoD) could facilitate detection of early infections.

Most SARS-CoV-2 diagnostic assays approved by the US Food and Drug Administration (FDA) are based on viral nucleic acid detection via amplification of a small number of specific viral target loci via Real-Time Polymerase Chain Reaction (RT-PCR). Although RT-PCR reactions can be extraordinarily specific, they suffer from key limitations. First, since RT-PCR assays amplify specific target loci, the assays will report a negative result if the particular target locus is not present in the sample. Consequently, RT-PCR will often produce an incorrect result when the sample is positive, but contains less than one genome equivalent in the initial reaction volume. Second, RT-PCR does not provide any genotypic information about a patient's infection beyond the causal organism. Such data can provide insight into the specific infecting strain and aid in tracing transmission within communities. Furthermore, the capacity to quickly and efficiently generate new viral genome data could expedite the generation of new diagnostics, vaccines and precise antivirals.

In principle, whole-genome DNA sequencing of SARS-CoV-2 has the potential to overcome these limitations. DNA sequencing can detect genome fragments even when a complete genome is not present in the sample. It can also extract extensive genotypic information about the viral genomes and genome fragments that are present. Notably, the SARS-CoV-2 genome is free of repeats, making it amenable to complete characterization using short DNA reads.

To exploit this possibility, we have developed Pathogen-Oriented Low-cost Assembly & Re-sequencing (POLAR), which combines: (i) enrichment of SARS-CoV-2 sequence using a PCR primer library designed by the ARTIC Network[3]; (ii) tagmentation-mediated library preparation for multiplex sequencing on an Illumina platform; and (iii) SARS-CoV-2 genome assembly (Figure 1). We show that POLAR is a reliable, inexpensive, and high-throughput SARS-CoV-2 diagnostic. Specifically, POLAR makes it possible for a single person to process 192 patient samples in an 8-hour workday day at a cost of $31 per sample. Including time for sequencing and data analysis, POLAR still enables a 24-hour turnaround time. POLAR also achieves very high sensitivity, with a limit of detection of 84 genome equivalents per milliliter, outperforming nearly all diagnostic tests currently approved by the US Food and Drug Administration (FDA).

In POLAR, total RNA from a clinical sample is reverse transcribed into DNA. This is followed by a multiplex polymerase chain reaction using a SARS-CoV-2 specific primer library to generate 400bp amplicons that tile the viral genome with ~200bp overlap, enriching the library for SARS-CoV-2 sequence. The amplicons are then fragmented and ligated to adapters using a rapid tagmentation mediated library preparation, and barcoded to enable multiplex sequencing. Finally, the data is analyzed using a one-click analysis software package that we have created, dubbed POLAR BEAR (POLAR Bioinformatics Evaluation of Assembly and Resequencing).

As part of this analysis, we determine whether a sample is infected by aligning the sequenced reads against a set of coronavirus reference genomes. Positive samples are identified as ones in which the sequenced reads cover more than 5% of the SARS-CoV-2 reference genome after primer sequences are filtered out. This diagnostic approach achieves a limit of detection of 84 genome equivalents per milliliter, making it more sensitive than nearly all methods currently approved by the FDA. When the viral concentration is higher (though still lower than, for instance, the limits of detection for the CDC SARS-CoV-2 tests), the data is also used to assemble an end-to-end, error-free SARS-CoV-2 genome from the sample, *de novo*. POLAR diagnoses have been validated by comparison to the results of the Center for Disease Control's 2019-Novel Coronavirus test in 10 clinical samples (nasopharyngeal swabs), yielding an exact match in 10 of 10 cases (5 positive, 5 negative).

Results

Whole-genome sequencing of SARS-CoV-2 yields a highly sensitive diagnostic.

We began by evaluating the suitability of the POLAR protocol as a potential diagnostic methodology.

To do so, we created 5 successive 10-fold serial dilutions of a quantified SARS-CoV-2 genomic RNA sample obtained from the American Tissue Culture Society (ATCC), which is widely used as a reference standard for diagnostic development. Specifically, we prepared positive controls containing 840,000 genome equivalents/mL, 84,000 genome equivalents per milliliter, 8,400 genome equivalents per milliliter, 840 genome equivalents per milliliter and 84 genome equivalents per milliliter. We performed 20 replicates at each concentration.

We also prepared a series of negative controls: 2 replicates of nuclease-free water, processed separately from the positive samples; 2 replicates of HeLa RNA extract, and 2 replicates of K562 RNA extract. We additionally included 20 replicates of nuclease-free water, prepared side-by-side with the positive samples, to serve as cross-contamination controls. The side-by-side cross contamination controls were included in order to ensure that our method was not susceptible to

false positives due to cross-contamination, a common error modality that is not well regulated in current FDA guidelines for diagnostic test development. In total, we performed the POLAR protocol on 26 different negative controls. Note that the above reflects the totality of experiments performed with our assay; we did not censor completed replicate experiments for any reason.

Each of the above 126 samples was processed using the POLAR protocol, and sequenced on a NextSeq550 Mid-Output Flow-cell. Note that, although a single technician can perform 192 experiments using the above workflow in an 8-hour shift, we did not perform all 192 experiments in the initial test. For these samples we generated 20 million paired-end 75bp reads of preliminary data.

To classify samples as positive or negative, we down sampled the data to 500 reads (2.5x coverage) per sample and checked to see if the breadth of coverage (the percentage of the target genome covered by at least 1 read, once primers are filtered out) was larger than 5% for each of the above samples (Figure 2A).

Of the 100 true positives, we accurately classified 99 (99%), with a single false negative at the most dilute concentration, 84 genome equivalents per milliliter. All 80 higher-concentration samples (840 genome equivalents/mL or more) were accurately identified as positive with an average breadth of coverage of 71%; 95% of the samples at 84 genome equivalents/mL were accurately classified (19 of 20), with an average breadth of coverage of 19%. All but 1 of 26 true negatives were accurately classified as negative, with an average breadth of coverage of 1%; the single misclassification was one of the cross-contamination controls.

Taken together, these data highlight the accuracy of the diagnostic test even when the amount of sequence data generated is negligible. Furthermore, they establish that the limit of detection of our assay, using the FDA definition, is 84 genome equivalents per milliliter (see Figure 2A).

The POLAR protocol for whole-genome sequencing of SARS-CoV-2 is more sensitive than nearly all diagnostics currently approved by the US FDA.

To compare POLAR to available diagnostic tests, we examined the 57 emergency use authorization summaries describing each of the 57 molecular SARS-CoV-2 diagnostic tests approved by the US FDA. For 44 tests, a limit of detection was clearly reported to the FDA in genome equivalents/milliliter (or, alternatively, genome copies/mL). For 41 of these 44 tests, the limit of detection was >= 100 genome equivalents/milliliter[4–50]. (Note that the LoD for the more sensitive of the two tests developed by the Center for Disease Control is 1000 genome equivalents per milliliter.) Thus, our test was significantly more sensitive than nearly all of the available tests.

We believe that this enhanced LoD is likely due to the fact that our method amplifies the entire viral genome, whereas RT-PCR only targets a handful of loci (Figure 2B). For instance, when examining the 21 different publicly available SARS-CoV-2 RT-PCR primer sets from the UCSC Genome Browser, we see that, even in aggregate, these primers amplify only 6.86% of the SARS-CoV-2 genome. At low starting concentrations of SARS-CoV-2, a sample can contain fragments of the viral genome that are detectable via whole genome sequencing, but which may not include the specific locus targeted by a particular RT-PCR assay.

The POLAR protocol for whole-genome sequencing of SARS-CoV-2 enables assembly of an end-to-end SARS-CoV-2 genome even from low-concentration samples.

Next, we sought to determine if the sequencing data resulting from the POLAR protocol could be used to assemble *de novo* the SARS-CoV-2 viral genome.

To explore this question, we took 150,000 75–base pair paired-end Illumina reads (2 x 75bp) from each of 18 libraries, comprising 3 different dilution series and corresponding negative controls. For each library, we generated a *de novo* assembly using the memory efficient assembly algorithm MEGAHIT[51] with default parameters. We first assessed the accuracy of these *de novo* assemblies by comparing them to the SARS-CoV-2 reference genome using a rescaled genome dot plot (Figure 3). The *de novo* assemblies showed very good correspondence with the SARS-CoV-2 reference, including the samples that contained only 84 genome equivalents per milliliter. We then quantified the accuracy and quality of these *de novo* SARS-CoV-2 assemblies. For the *de novo* assemblies that contained ≥8,400 equivalents per milliliter, 86.67% of assemblies consisted of a singular contig comprising 99.77% of the SARS-CoV-2 genome (Table 1). The remaining 0.23% of the SARS-CoV-2 genome corresponds to short regions at both ends of the genome, which are not amplified by the ARTIC primer set. While the *de novo* assemblies created from samples with 840 genome equivalents/mL and 84 genome equivalents/mL are less contiguous, we are able to recover on average 84.63% and 34.16% of the viral genome, respectively. Remarkably, 100% of the bases in 16 of these 20 *de novo* assemblies match the corresponding bases in the SARS-CoV-2 reference genome. Three of the remaining four *de novo* assemblies have only a single base pair difference as compared to the SARS-CoV-2 reference genome. Collectively, these data demonstrate that our method provides *de novo* SARS-CoV-2 genome assemblies at viral concentrations at or below the CDC RT-PCR limit of detection. Furthermore, at most of the concentrations examined, the SARS-CoV-2 genome assemblies produced by POLAR are gapless, and completely free of errors.

POLAR accurately assembles other coronaviruses, while distinguishing them from SARS-CoV-2.

SARS-CoV-2 is one of many coronaviruses that commonly infect humans. We therefore sought to determine whether POLAR (which uses SARS-CoV-2 specific primers) could accurately distinguish between SARS-CoV-2 and other coronaviruses. To do so, we applied POLAR to samples containing genomic RNA from the following coronaviruses: Human Coronavirus NL63, Human Coronavirus strain 229E, Porcine Respiratory Coronavirus strain ISU-1 and Avian Coronavirus. (Genomic RNA was obtained from ATCC.)

Notably, for Porcine Respiratory Coronavirus strain ISU-1, Human Coronavirus strain 229E and Avian Coronavirus, our automated pipeline assembled the entire viral genome with no gaps (Figure 4). For Human Coronavirus NL63, there was a single gap. These assemblies covered at least 96.01% of their respective reference genome assembly, with a base accuracy of >99.9% (Table 2).

At the same time, like our other SARS-CoV-2 negative controls, the data from these experiments had a breadth of coverage of <0.5% when the sequenced reads were aligned back to the SARS-CoV-2 reference genome. Thus, in all four cases, our pipeline accurately determined that these true negatives did not contain SARS-CoV-2. This highlights the potential of our approach for diagnosing other coronaviruses, including cases of co-infection by multiple coronaviruses including, but not limited to, SARS-CoV-2.

POLAR BEAR is a fully automated analysis pipeline for transforming POLAR sequence data into genome assemblies and a diagnostic report.

To aid in the analysis of data produced by POLAR, we also developed a one-click open-source pipeline, dubbed the POLAR Bioinformatics Evaluation of Assembly and Resequencing (BEAR). POLAR BEAR takes the DNA reads produced from a sample and performs all of the above analyses, generating a document containing breadth of coverage statistics, a genome dot plot, and a test result (positive or negative), based on whether the breadth of SARS-CoV-2 coverage for a sample is greater than or equal to 5% (positive) or is less (negative). See Figure 5. The pipeline also reports the resulting SARS-CoV-2 reference genome, as well as any other genome assemblies that were generated. We confirmed that the pipeline can be run efficiently on a wide range of high-performance computing platforms and that the computational cost per test is negligible (<1¢). The pipeline, including documentation and test set, is publicly available at https://github.com/aidenlab/POLAR-BEAR.

We used POLAR to classify 5 positive and 5 negative clinical samples in a blinded experiment, exhibiting 100% agreement with the CDC 2019-Novel Coronavirus test.

Next, we applied the POLAR protocol on ten clinical samples, five negative and five positive, in a blinded experiment.

We obtained 10 unique mid-turbinate nasal swab samples that were collected in viral transport media. These samples had previously been tested using the CDC 2019-Novel Coronavirus (2019-nCOV) Real-time PCR diagnostic panel by the Respiratory Virus Diagnostic Laboratory (RVDL), a CLIA-certified laboratory at the Baylor College of Medicine. Five of the samples had tested positive, and five had tested negative.

Although the authors of the present manuscript were aware of the facts in the preceding paragraph, the authors were otherwise blinded as to whether each sample was positive or negative. For instance, the labelling and ordering of the samples was randomized. The authors remained blinded throughout our experimentation, analysis, classification, and assembly procedure.

Briefly, each of the 10 clinical samples were processed using the POLAR protocol and sequenced on a NextSeq550 Mid-Output Flow-cell, as described above. For each of these samples we generated 150,000 75–base pair paired-end Illumina reads (2 x 75bp). We then ran POLAR BEAR on each sample, as described above.

Five clinical samples were classified by POLAR BEAR as positive (i.e., the breadth of SARS-CoV-2 coverage was greater than or equal to 5%) and five were negative (Figure 6A). The differences were unambiguous: 5 positive clinical samples had an average breadth of coverage of 99.65% while the 5 negative clinical samples had an average breadth of coverage of 0.65%.

Four of the five samples that POLAR BEAR classified as positive yielded a *de novo* assembly of the SARS-CoV-2 viral genome consisting of a single contig spanning >99.74% of the SARS-CoV-2 genome (Table 4). The remaining positive sample yielded a SARS-CoV-2 assembly comprising 2 contigs spanning 99.25% of the SARS-CoV-2 genome. Of course, the five samples POLAR BEAR classified as negative did not yield an assembly spanning a significant portion of SARS-CoV-2. See Figure 6B.

Finally, the authors were unblinded, and compared the POLAR BEAR classification to the results of the CDC 2019-Novel Coronavirus test performed by RVDL. The diagnosis of positive or negative matched in all 10 cases (100%).

Collectively, these data demonstrate that our method accurately classifies clinical samples and provides a complete and accurate SARS-CoV-2 genome assembly for each and every affected patient.

Discussion

Given the current need for SARS-CoV-2 testing, we developed POLAR, a reliable, inexpensive, and high-throughput SARS-CoV-2 diagnostic based on whole genome sequencing. Our method builds off those developed by ARTIC Network for in-field viral sequencing in order to generate real-time epidemiological information during viral outbreaks[52,53]. We have demonstrated that this approach is sensitive, scalable, reproducible, produces diagnoses on clinical samples that match those of the CDC's standard RT-PCR based diagnostic, and is consistent with US FDA guidelines governing diagnostic testing for SARS-CoV-2.

The POLAR protocol has two key advantages over RT-PCR-based diagnostics.

First, it is highly sensitive, achieving a LoD of 84 genome equivalents per milliliter, which exceeds the reported LoD of all but three diagnostic tests[54–56] approved by the US FDA. We believe that further refinements of the protocol will likely allow this to be further improved. By enhancing sensitivity, it may be possible to detect infection earlier in the course of infection – ideally, before a person is contagious – and to detect infection from a wider variety of sample types.

Second, it produces far more extensive genotype data than targeted, RT-PCR based diagnostics, including an end-to-end SARS-CoV-2 genome at concentrations that are beyond the limit of detection of many other assays. Having whole viral genomes from all diagnosed individuals enables the creation of viral phylogenies to better understand the spread of the virus in community and health care settings. It will further yield valuable understanding of the different strains and patterns of mutations of the disease. Finally, it can enable the discovery of additional testing, vaccine, and drug targets.

At the same time, the approach we describe also has several limitations as compared to other diagnostic tests. For example, our method does not provide any information regarding the SARS-CoV-2 viral load of a patient. This might be addressed by adding a synthetic RNA molecule with a known concentration into each patient sample in order to estimate viral load using relative coverage.

Another limitation is that our method is slower than other approaches, in the sense that it requires 24 hours from acquisition of a patient sample to a diagnostic result. By contrast, Abbott Labs (one example of many new diagnostic technologies developed in the past few months) has developed a diagnostic test capable of returning results in as little as 5 minutes for a positive result and 13 minutes for a negative result[57]. However, it is worth noting that the maximum number of diagnostic results an Abbot device could complete running 24 hours a day is roughly between 111 tests and 126 tests depending on the number of positive results[58,59].

Beyond diagnosis of individual patients, POLAR can also be applied to SARS-CoV-2 surveillance, in settings such as municipal wastewater treatment plants[60]. In principle, such approaches could identify and characterize infection in a neighborhood or city very inexpensively, even for a large population, informing public policy decisions.

We note that multiple groups have been developing methods for sequencing whole SARS-CoV-2 genomes, and in some cases sharing the protocols ahead of publication on protocols.io (https://www.protocols.io/). Like POLAR, these methods often use the ARTIC primer set, with some of these approaches relying on long-read DNA sequencing[61]. Although long reads enable more contiguous genome assemblies when the underlying genome contains complex repeats, we find that such reads are not necessary for gapless assembly of SARS-CoV-2. As such, the use of long reads, which is costly, produces less accurate base calls, and hampers multiplexing, may be less necessary in the context of SARS-CoV-2 sequencing. At the same time, long read technologies such as Oxford Nanopore may offer other advantages, such as the potential to sequence in real-time. This capability could be valuable for the development of point-of-care diagnostics.

Other methods use short read DNA sequencing[62–65]. Most of these approaches partition individual samples into multiple wells, making them difficult to perform in a highly multiplex fashion[63,64]. One method enables extensive multiplexing, but does not include a random fragmentation step. As such, it requires 2 x 150bp paired-end reads in order to produce gapless assemblies, making a 24-hour turnaround impossible on extant Illumina devices[65]. Because POLAR includes a random fragmentation step requiring only 75bp reads, it can be performed on more rapid instruments such as NextSeq 550. However, at least one of these methods appears to enable both rapid turnaround and 9-fold multiplexing[62]. Although prior studies have not explored the sensitivity and specificity of these protocols when deployed as a diagnostic, emerging work from many laboratories make it clear that whole-genome sequencing of SARS-CoV-2 is likely to be a promising modality that is well-suited for clinical use.

Acknowledgments

This work was supported by a Thrasher Research Fund Early Career Award (#14801) to A.P.A., a Howard Hughes Medical Institute Gilliam Fellowship (#GT11533) to B.G.S., an Israel Binational Science Foundation Grant (#2017086) and an NSF Physics Frontier Center Grant (#PHY-1427654).

We thank Dr. Gary Schroth, Dr. Linda Ray, Dr. Feng Chen, Dr. Erich Jaeger, Dr. Steph Craig and Dr. Mehdi Keddache of Illumina for providing flow cells, reagents, and constructive feedback on our project. We also thank Dr. Joseph Petrosino of Baylor College of Medicine for fruitful discussions related to our limit of detection. We are grateful for access to clinical samples for validating our method provided by Dr. Pedro Piedra and Dr. Vasanthi Avadhanula, in addition to fruitful conversations.

We thank Terry Leatherland, Grace Liu, Loic Fura and Victoria Nwobodo for access to a high RAM IBM E880 server where the majority of our computational analysis and the POLAR BEAR pipeline construction was done. The POLAR BEAR pipeline benchmarking work was supported by resources provided by the University of Western Australia and the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia. We also gratefully acknowledge Microsoft and the WA technology company DUG for testing and benchmarking the pipeline on their systems.

We thank Dr. Christophe Herman for providing flow cells and use of the Herman lab's NextSeq550. We thank Dr. Joshua Quick, Dr. Clavia Ruth Wooton-Kee, Dr. David Cunningham, Dr. Ellen Busschers, and Dr. Dmitriy Khodakov for providing reagents at the very start of the project when resources were limited to due to reagent shortages.

Methods & Materials

Collection of SARS-CoV-2

The quantified sample material used for the limit of detection was genomic RNA (gRNA) extracted from a cell line (Vero E6, ATCC® CRL-1586™) infected with SARS-related coronavirus 2 (SARS-CoV-2, isolate USA-WA1/2020, Lot: 70033700), using QIAamp® Viral RNA Mini Kit (Qiagen 52904) deposited by American Type Culture Collection (ATCC) and obtained from Biodefense and Emerging Infections Research Resources Repository (BEI Resources). The amount of viral genome RNA molecules per volume of total RNA including cellular nucleic acid and carrier RNA for the lot we received of SARS-CoV-2 gRNA was quantified as 5.5 x 10^4 genome equivalents/µL using a BioRad QX200 Droplet Digital PCR (ddPCR™) System.

Negative control RNA extraction

Approximately 1 million K562 cells and 1 million HeLa cells cultured in our lab were used as the starting material for RNA extraction using columns provided in the RNeasy Mini Kit (Cat no: 74104). The final elute was collected in 30µl of RNA-free waterThis elution was then split into 10µL aliquots and concentration was measured by using the GE Nanovue plus.

Clinical Sample RNA extraction

Approximately 100µl of mid-turbinate nasal swab sample in viral transport media was used as the starting material for RNA extraction using columns provided in the Quick-RNA Viral Kit (Cat no: R1034). The final elute was collected in 15µl of RNA-free water.

Performing the SARS-CoV LoD

The limit of detection (LoD) was determined by making five 10-fold serial dilutions using the SARS-CoV-2 gRNA from ATCC as a stock and nuclease-free water as a diluent. Each dilution was tested with two biological replicates each of which further had 10 technical replicates, so in total 20 replicates from each dilution of stock SARS-CoV-2 gRNA. For experiment,1µl of each dilution was spiked into a mix of 4.5µl of nuclease-free water, 0.5µl of 10mM dNTPs Mix (NEB, N0447L) and 0.5 µl of 50µM Random Hexamers (ThermoFisher, N8080127) to serve as the starting material, “RNA Extract”, for the protocol. The RNA, hexamers, and dNTPs mixture was incubated at 65°C for 5 minutes followed by a 1 minute incubation at 4°C in order to anneal hexamers to RNA. In order to reverse transcribe RNA into cDNA, we added 2µl of 5X SuperScript™ IV Reverse Buffer (ThermoFisher, 18090050), 0.5µl of SuperScript™ IV Reverse Transcriptase (200 U/µL) (ThermoFisher, 18090050), 0.5µl of 100mM DTT (ThermoFisher, 18090050), 0.5µl of RNaseOUT Recombinant Ribonuclease Inhibitor (ThermoFisher, 10777-019) to the hexamer annealed RNA. The reaction was then incubated at 42°C for 50 minutes followed by an incubation at 70°C for 10 minutes before holding at 4°C. For amplification of cDNA, we used SARS-CoV-2-specific version 3 primer set (total 218 primers) designed by Josh Quick from the ARTIC Network. Primers were purchased at LabReady concentration of 100µM in IDTE buffer (pH 8.0) from Integrated DNA Technologies (IDT). Multiplex-polymerase chain reaction (PCR) was performed in two separate reaction mixes prepared by combining 5µl of 5X Q5 Reaction Buffer (NEB, M0493S), 0.5µl of 10 mM dNTPs (NEB, N0447L), 0.25µl Q5 Hot Start DNA Polymerase (NEB, M0493S)with either 12.7µl Nuclease-free water (Qiagen, 129114) and 4.05µl of 10µM “Primer Pool #1” or, 12.7µl Nuclease-free water (Qiagen, 129114) and 3.98uL for 10µM “Primer Pool #2”. The final concentration of each primer in the reaction mix was 0.015µM in the

PCR mix. 22.5µl of the corresponding mastermix (Pool #1 or Pool #2) was combined with 2.5µl of the reverse transcribed cDNA. The reaction was then incubated at 98°C for 30 seconds for 1 cycle followed by 25 cycles at 98°C for 15 seconds and 65°C for 5 minutes before holding at 4°C. Pool #1 or Pool #2 amplicons from each replicate were then mixed together and cleaned by adding 1:1 volume of sparQ PureMag beads (QuantaBio, 95196-060) and incubating at room temperature for 5 minutes. The beads were separated using a magnet and the supernatant was discarded, followed by two 200µl washes of freshly made 80% ethanol. Each sample was eluted in 11µl of 10mM Tris-HCl (pH 8.0) and incubated for 2 minutes at 37°C followed by separation on a magnet. The DNA was then quantified using a Qubit® High Sensitivity Kit (ThermoFisher, Q32851) as per manufacturer's instructions and the concentrations were used to ensure 1ng of amplicon DNA in 4µl was carried per sample into library preparation.

Library preparation was performed using the Nextera XT DNA Library Preparation Kit (Illumina, FC-131-1096) and Nextera XT Index Kit v2 (Illumina, FC-131-2001/2002). 4µl of 1ng amplicon DNA was combined with a mix containing 1µl of Amplicon Tagment Mix (Illumina, FC-131-1096) and 5µl of Tagment DNA Buffer (Illumina, FC-131-1096) and incubated at 55°C for 5 minutes. Temperature was then lowered to 10°C followed by addition of 2.5µl of Neutralize Tagment Buffer immediately after the cooling started, mixed by pipetting, and incubated at room temperature for 5 minutes. After 5 minutes, the reaction was centrifuged at 280xG for 1 minute and the next reaction was set-up during centrifugation. 12.5µl of a mastermix containing 7.5µl of Nextera PCR Master Mix (Illumina, FC-131-1096) and 2.5µl of each Index primer i7 (Illumina, FC-131-2001/2002) and Index primer i5 (Illumina, FC-131-2001/2002) was combined with 12.5µl of the tagmented amplicon DNA. The reaction was then incubated on a thermal cycler with the following parameters: 1 cycle at 72°C for 3 minutes and 95°C for 30 seconds, 18 cycles at 55°C for 10 seconds, 72°C for 30 seconds, 72°C for 5 minutes followed by a 4°C hold. Post PCR clean up was done using 1:1.8 volume (45µL beads in 25uL reaction) of sparQ PureMag beads (QuantaBio, 95196-060), washed twice with 80% ethanol, eluted in 20µL of 10mM Tris-HCl (pH 8.0) followed by an incubation at 37°C for 2 minutes and separated on a magnetic plate. 10µl from each well of the plate was then transferred onto the corresponding well on a new midi plate. A Library Normalization (LN) (Illumina, FC-131-1096) master mix was created by combining two reagents in a 15µl conical tube. The reagents were multiplied by the number of samples being processed: 23µl of LNA1 and 4µl of LNB1. The mixture was then mixed by pipetting 10 times and then poured into a trough. Using a p200 multichannel pipette, 22.5µl of LN master mix was placed into each sample well. To mix, we sealed the plate, and vortexed using a plate shaker at 1800rpm for 30 minutes. The plate was then placed on a magnetic stand to separate the beads. Once the liquid on the plate was clear, without disturbing the beads, we discarded the supernatant. The beads were then washed twice by adding 22.5µl of LNW1 to each well, sealing the plate, using the plate shaker at 1800rpm for 5 minutes, then separating the beads on a magnetic plate and discarding the supernatant. After the washes, 15µl of 0.1N NaOH was added to each well. The plate was then sealed and vortexed at 1800rpm to mix the sample for 5 minutes. During the 5-minute mixing, 15µl of LNS1 was added to each well of a new 96-well PCR plate that was labeled as SGP. After the 5-minute elution step, the plate was placed on a magnetic stand, and 15µl of the supernatant was transferred to the corresponding well of the SGP plate. The plate was then sealed and spun at 1000xG for 1 minute.

Preparation of Illumina sequencing run

To prepare for the sequencing run, a "Mid-Output Kit" reagent cartridge (Illumina, 20024904) was removed from the -20°C freezer to thaw in a secondary container filled with room temperature deionized water about 1/3rd of its height. The cartridge was left in this water bath for 1 hour to

completely thaw. 30 minutes into the thawing, the “Mid-Output Kit” flow cell (Illumina, 20024904) was removed from the 4°C refrigerator to warm up to room temperature for 30 minutes.

A tube of Hybridization Buffer (HT1) (Illumina, 20015892) was collected from the -20°C and thawed at room temperature. Once thawed, it was placed on ice. A thermomixer was pre-heated to 98°C. The libraries were pooled equally by using 2µl of all the normalized samples and the pool was then vortexed and centrifuged down to ensure proper mixing of the samples. 5µl volume from this pool was transferred into a new tube to which 995µl of ice-cold Hybridization Buffer (HT1) (Illumina, 20015892) was then added. The tube was quickly vortexed and centrifuged at 300xG for 1 minute. From this tube, 750µl was transferred to a new tube. To this new tube 750µl ice-cold Hybridization Buffer (HT1) (Illumina, 20015892) was added to further dilute the sample molarity. The tube was quickly vortexed, centrifuged and was placed at 98°C in the thermomixer for 2 minutes. Immediately after 2 minutes, the tube was placed on ice for 5 minutes. In a fresh 1.5mL tube, to bring the final concentration in the pool to 1.5pM, 97µl of the previous dilution and 1203µl of ice-Cold Hybridization buffer (HT1) (Illumina, 20015892) were mixed together. The final denatured library pool was then placed on ice until it was ready to load onto the reagent cartridge. Finally, the loaded reagent cartridge, along with the flow cell and buffer pack were inserted into the NextSeq500 for sequencing.

SARS-CoV-2 Coverage Analysis

To compare SARS-CoV-2 coverage across starting concentrations, FASTQs were aligned to the SARS-CoV-2 reference genome (NCBI Reference Sequence: NC_045512.2) using BWA[66] with default parameters. Samtools[67] was then used to sort, fixmates, dedup, and calculate depth per base. To filter out primer reads, we first discarded all depths per base below a threshold of >1, and then removed islands that had 50 or fewer consecutive positions covered. Remaining values are plotted by base position in the coverage track.

The rescaled dot plot below is generated by plotting contig alignment generate by Minimap2[68] to the reference genome. Contigs are sorted and non-mapped contigs have been removed, leaving all remaining aligning contigs lying along the diagonal.

RT-PCR primers regions were created by downloading RT-PCR primers from the UCSC genome browser[69]. Forward and reverse primers were paired to generate RT-PCR target regions for each pair. Bedtools[70] was then used to merge these individual RT-PCR target regions into a single track in order to collapse the overlapping RT-PCR target regions.

Breadth of Coverage Scatter Plot

Breadth of coverage was determined after filtering out primer reads (described above). The number of positions with coverage after primer filtering was divided by the total length of the reference. This value is stored in the “stats.csv” file produced by the POLAR pipeline for all coronaviruses and used in the final report to create the bar charts and determine the result (positive or negative).

To create the scatter plot in Figure 2, data was plotted in R Studio using ggplot2[71], dplyr[72] and forcats libraries. A position jitter was used to allow for better visualization of data points, which at high concentrations of SARS-CoV-2 often overlapped. The jitter parameters were calibrated to allow for optimal visualization of data points without changing the relative position of each data point.

Comparative Assembly Statistics

In order to determine the accuracy of our *de novo* assemblies, we compared our SARS-CoV-2 *de novo* assembly to the SARS-CoV-2 reference assembly (NCBI Reference Sequence: NC_045512.2), our Human coronavirus 229E *de novo* assembly to the Human coronavirus 229E reference assembly (NCBI Reference Sequence: NC_002645.1), our Avian Coronavirus to the Avian Coronavirus Massachusetts (formerly Avian Infectious Bronchitis Virus) (GenBank: GQ504724.1), our Human Coronavirus NL63 *de novo* assembly to the Human Coronavirus NL63 (GenBank: AY567487.2) reference assembly and our Porcine Respiratory Virus to the PRCV ISU-1 (GenBank: DQ811787.1) reference genome using MetaQuast[73].

References

1. He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. (2020).
2. To, K. K.-W. *et al.* Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: an observational cohort study. *Lancet Infect. Dis.* **0**, (2020).
3. Artic Network. https://artic.network/ncov-2019.
4. Boston Children's & Hospital. Childrens-Altona-SARS-CoV-2 Assay EUA Summary. (2020).
5. Quest Diagnostics. *SARS-CoV-2 RNA, Qualitative Real-Time RT-PCR* . https://www.fda.gov/media/136231/download (2020).
6. Abbott Molecular. *Abbott RealTime SARS-CoV-2*. https://www.fda.gov/media/136258/download (2020).
7. Thermo Fisher Scientific. *TaqPath™ COVID-19 Combo Kit*. https://www.fda.gov/media/136112/download (2020).
8. BGI Genomics Co. *Real-Time Fluorescent RT-PCR Kit for Detecting SARS-2019-nCoV*. https://www.fda.gov/media/136472/download (2020).
9. Hologic. *SARS-CoV-2 Assay (Panther Fusion ®System) Panther Fusion ®SARS-CoV-2*. https://www.fda.gov/media/136156/download.
10. Luminex. *ARIES ®SARS-CoV-2 Assay*. https://www.fda.gov/media/136693/download (2020).
11. Maccura. SARS-CoV-2 Fluorescent PCR Kit. (2020).
12. ORIG3N, I. N. C. *Orig3n 2019 Novel Coronavirus (COVID-19) Test*. https://www.fda.gov/media/136873/download (2020).
13. Centers for Disease Control and Prevention. *CDC 2019-Novel Coronavirus (2019-nCoV) Real-Time RT-PCR Diagnostic Panel* . https://www.fda.gov/media/134922/download (2020).
14. Ipsum Diagnostics, L. L. C. *The COV-19 IDx assay*. https://www.fda.gov/media/136621/download (2020).
15. Baptist Health Care System. *Baptist Hospital Pathology/Laboratory Medicine, Miami, FL COVID-19 RT-PCR test*. https://www.fda.gov/media/136944/download (2020).
16. Wadsworth Center. *New York SARS-CoV-2 Real-time Reverse Transcriptase (RT)- PCR Diagnostic Panel*. https://www.fda.gov/media/135847/download (2020).
17. Luminex. *NxTAG ®CoV Extended Panel Assay Package Insert*. https://www.fda.gov/media/136500/download (2020).
18. Mesa Biotech. *SARS-CoV-2 Test*. https://www.fda.gov/media/136355/download (2020).
19. Yale New Haven Hospital & Stanford Health Care Clinical Virology Laboratory. SARS-CoV2 RT-PCR Assay . (2020).
20. Cobas. *cobas®SARS-CoV-2*. https://www.fda.gov/media/136049/download (2020).

21. HACKENSACK UNIVERSITY MEDICAL CENTER, M. & LABORATORY, P. CDI Enhanced COVID-19 Test. (2020).
22. Massachusetts General Hospital. Massachusetts General Hospital SARS-CoV-2 Test EUA Summary. (2020).
23. GenoSensor. GenoSensor GS TM COVID-19 Real-Time PCR Kit. (2020).
24. Children's Hospital of Philadelphia. CHOP SARS-CoV-2 test EUA Summary. (2020).
25. BioFire. *BioFire ®COVID-19 Test*. https://www.fda.gov/media/136353/download (2020).
26. Molecular, N. *NeuMoDx™ SARS-CoV-2 Assay*. https://www.fda.gov/media/136565/download (2020).
27. ScienCell. *ScienCell™ SARS-CoV-2 Coronavirus Real-time RT-PCR (RT-qPCR) Detection Kit*. https://www.fda.gov/media/136691/download (2020).
28. University of North Carolina Medical Center. UNC Health SARS-CoV-2 real-time RT-PCR test. (2020).
29. Exact Sciences Laboratory. SARS-CoV-2 (E, N and RdRP gene detection) Test. (2020).
30. ATILA BioSystems. iAMP ®COVID-19 Detection Kit iAMP-COVID19-100. (2020).
31. SDI Laboratories. SDI SARS-CoV-2 Assay. (2020).
32. Curative-Korva, K. I. C. L. *CURATIVE-KORVA SARS-COV-2 ASSAY*. https://www.fda.gov/media/137089/download (2020).
33. Fisher Scientific, T. TaqPath ™ COVID-19 Combo Kit INSTRUCTIONS FOR USE Multiplex real-time RT-PCR test intended for the qualitative detection of nucleic acid from SARS-CoV-2 Catalog Numbers A47813 and A47814 Publication Number MAN0019181 Revision B.0. (2020).
34. GeneXpert. *Xpert®Xpress SARS-CoV-2*. https://www.fda.gov/media/136314/download (2020).
35. Fosun Pharma USA Inc. *Fosun COVID-19 RT-PCR Detection Kit Rx Only For*. https://www.fda.gov/media/137120/download (2020).
36. Co-Diagnostics Inc. *Logix Smart™ Coronavirus Disease 2019 (COVID-19) Kit*. https://www.fda.gov/media/136687/download (2019).
37. DiaSorin Molecular LLC. *Simplexa™ COVID-19 Direct*. https://www.fda.gov/media/136286/download (2020).
38. Primerdesign Ltd. *Primerdesign Ltd COVID-19 genesig®Real-Time PCR assay*. https://www.fda.gov/media/136309/download (2020).
39. DiaCarta Inc. QuantiVirus™ SARS-CoV-2 Test Kit Rx . (2020).
40. GenMark Dx. *PI1109-A ePlex ®SARS-CoV-2 Test Assay Manual*. https://www.fda.gov/media/136282/download (2020).
41. Integrity Laboratories. Integrity Laboratories SARS-CoV-2 Test EUA Summary. (2020).
42. LabCorp. *LabCorp COVID-19 RT-PCR test*. https://www.fda.gov/media/136151/download (2020).
43. Gnomegen. *COVID-19 RT-Digital PCR Detection Kit*. https://www.fda.gov/media/136738/download (2020).
44. AVELLINO LAB USA. *AvellinoCoV2 test*. https://www.fda.gov/media/136453/download (2020).
45. OSANG Healthcare. GeneFinder™ COVID-19 Plus RealAmp Kit. (2020).
46. Quidel Corporation. *Lyra®SARS-CoV-2 Assay* . https://www.fda.gov/media/136227/download (2020).
47. NORTHWESTERN MEDICINE. Northwestern Medicine SARS-CoV-2 Assay EUA Summary. (2020).
48. QIAGEN. *QIAstat-Dx ®Respiratory SARS-CoV-2 Panel Instructions for Use (Handbook)*. https://www.fda.gov/media/136571/download (2020).
49. InBios. *Smart Detect TM SARS-CoV-2 rRT-PCR Kit* . https://www.fda.gov/media/136786/download (2020).

50. CirrusDx Laboratory. CirrusDx SARS-CoV-2 assay. (2020).
51. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* vol. 102 3–11 (2016).
52. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).
53. Quick, J. *et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* **12**, 1261–1266 (2017).
54. PerkinElmer. *Instructions for PerkinElmer® New Coronavirus Nucleic Acid Detection Kit v 1.0*. (2020).
55. Becton, D. & C. *SARS-CoV-2 Reagents for BD MAX™ System*. https://www.fda.gov/media/136816/download (2020).
56. Viracor Eurofins Clinical Diagnostics. Viracor SARS-CoV-2 RT-PCR TEST. (2020).
57. Detect COVID-19 in as Little as 5 Minutes | Abbott Newsroom. https://www.abbott.com/corpnewsroom/product-and-innovation/detect-covid-19-in-as-little-as-5-minutes.html.
58. The COVID Tracking Project. US Historical Data. https://covidtracking.com/data/us-daily (2020).
59. Gudbjartsson, D. F. *et al.* Spread of SARS-CoV-2 in the Icelandic Population. *N. Engl. J. Med.* (2020) doi:10.1056/NEJMoa2006100.
60. Medema, G., Heijnen, L., Elsinga, G., Italiaander, R. & Brouwer, A. Presence of SARS-Coronavirus-2 in sewage. *medRxiv* (2020) doi:10.1101/2020.03.29.20045880.
61. nCoV-2019 sequencing protocol for illumina. https://www.protocols.io/view/ncov-2019-sequencing-protocol-for-illumina-betejeje.
62. Itokawa, K., Sekizuka, T., Hashino, M., Tanaka, R. & Kuroda, M. A proposal of alternative primers for the ARTIC Network's multiplex PCR to improve coverage of SARS-CoV-2 genome sequencing. *bioRxiv* (2020) doi:10.1101/2020.03.10.985150.
63. SARS-CoV-2 Genome Sequencing Using Long Pooled Amplicons on Illumina Platforms. https://www.protocols.io/view/sars-cov-2-genome-sequencing-using-long-pooled-amp-befyjbpw.
64. RTPCR Amplification of SARS-CoV2 Whole Genome for Illumina NGS. https://www.protocols.io/view/rtpcr-amplification-of-sars-cov2-whole-genome-for-bew8jfhw.
65. COVID-19 ARTIC v3 Illumina library construction and sequencing protocol. https://www.protocols.io/view/covid-19-artic-v3-illumina-library-construction-an-be3wjgpe.
66. Li, H. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. *Oxford* vol. 34 1401–1403 (Oxford Academic, 2013).
67. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
68. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
69. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. doi:10.1093/bib/bbs038.
70. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* (2010) doi:10.1093/bioinformatics/btq033.
71. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).
72. Wickham, H., François, R., Henry, L. & Müller, K. *Package 'dplyr'*. *Cran* (2018).
73. Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: Evaluation of metagenome assemblies. *Bioinformatics* (2016) doi:10.1093/bioinformatics/btv697.

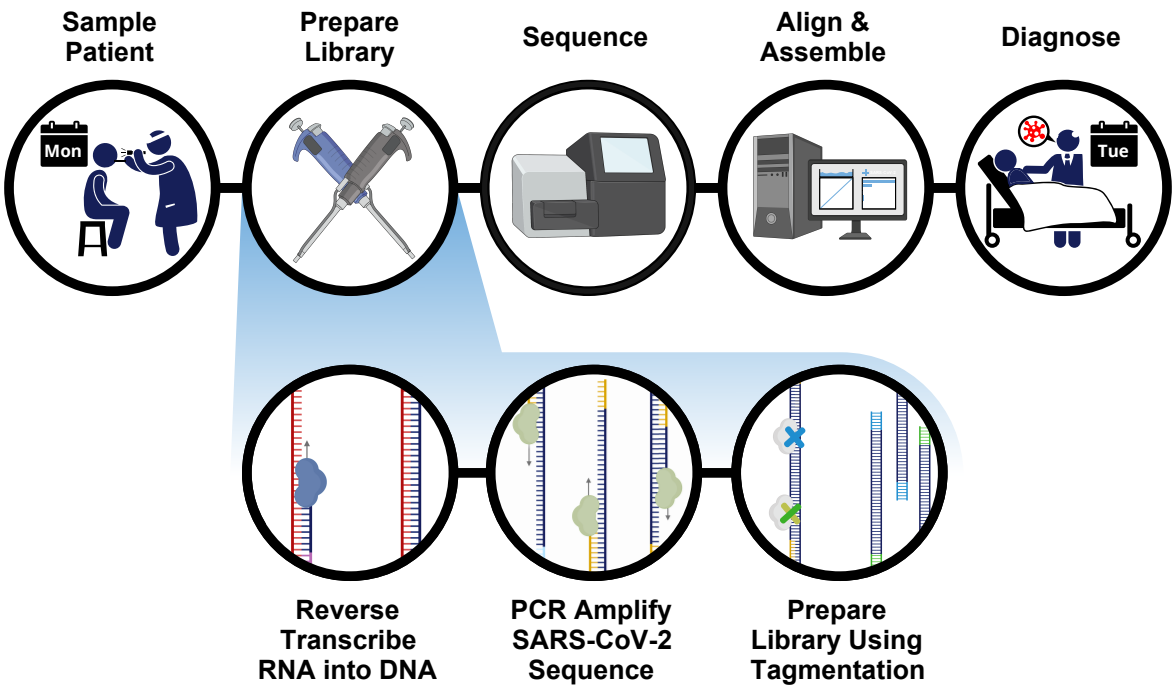


Figure 1. Pathogen-Oriented Low-cost Assembly & Re-sequencing (POLAR) Method Overview. Patient is sampled in the clinic and total RNA from this sample is extracted and reverse transcribed into DNA. The sample is then enriched for SARS-CoV-2 sequence using a SARS-CoV-2 specific primer library. The amplicons then undergo a rapid tagmentation mediated library preparation. Data is then analyzed and used to report patient result the next day.

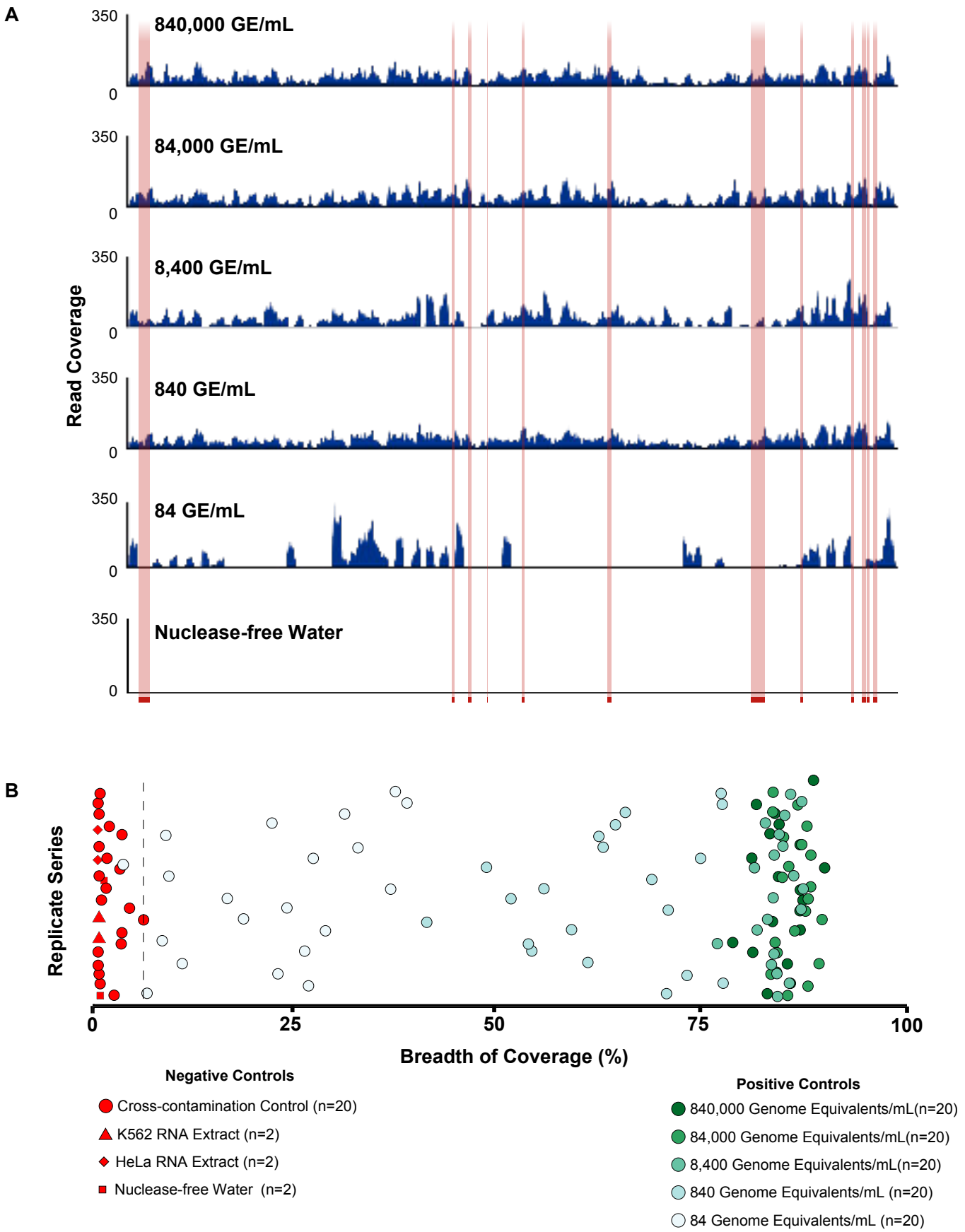
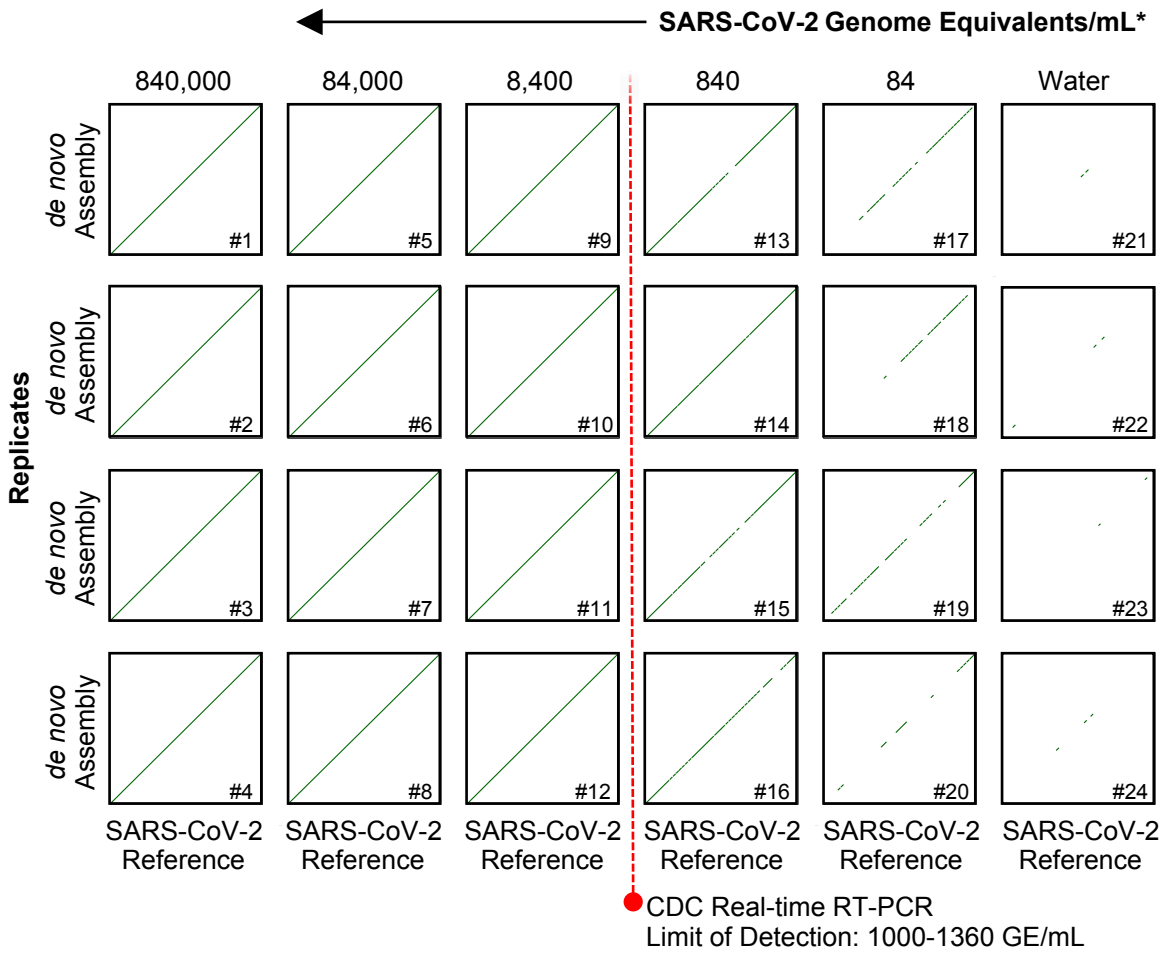

A
840,000 GE/mL
84,000 GE/mL
8,400 GE/mL
840 GE/mL
84 GE/mL
Nuclease-free Water
350
0
Read Coverage
B
Replicate Series
0
25
50
75
100
Breadth of Coverage (%)
Negative Controls
Cross-contamination Control (n=20)
K562 RNA Extract (n=2)
HeLa RNA Extract (n=2)
Nuclease-free Water (n=2)
Positive Controls
840,000 Genome Equivalents/mL(n=20)
84,000 Genome Equivalents/mL(n=20)
8,400 Genome Equivalents/mL(n=20)
840 Genome Equivalents/mL (n=20)
84 Genome Equivalents/mL (n=20)

Figure 2. Polar Protocol Detects SARS-CoV-2 in Dilute Samples. (A) Coverage tracks demonstrate sequencing depth across the SARS-CoV-2 genome produced by our protocol from samples with a range of starting SARS-CoV-2 genome concentrations. Red-highlighted regions represent virus sequence detected by qPCR-based COVID-19 diagnostics in use or development. Each coverage track was generated using 150,000 75-PE reads. (B) Scatter plot shows breath of coverage for samples from lower replicate dilution series and negative controls. Dashed red line represents the empirically determined breadth of coverage threshold for positive samples. The breadth of coverage of each library was calculated using 500 75-PE reads.



* total starting sample volume = 6.5µl

Figure 3. Diagnosis-by-sequencing Generates *de novo* Viral Genome Assemblies. Each rescaled genome dot plot (black boxes numbered 1 to 24) compares a *de novo* SARS-CoV-2 assembly (Y-axes) to the SARS-CoV-2 reference genome (X-axes). Columns contain replicate assemblies at a given SARS-CoV-2 concentration. The de novo assemblies displayed on the Y-axes have been ordered and oriented to match the reference viral genome in order to facilitate comparison. Each line segment represents the position of an individual contig from the de novo assembly that aligned to the reference genome. Dotted red line represents the limit of detection for the Center for Disease Control qPCR tests currently used to detect SARS-CoV-2. For rescaled dot plots, contigs were sorted and unmapped contigs have been removed, leaving all remaining aligning contigs lying along the diagonal. Each *de novo* assembly was generated using 150,000 75-PE reads.

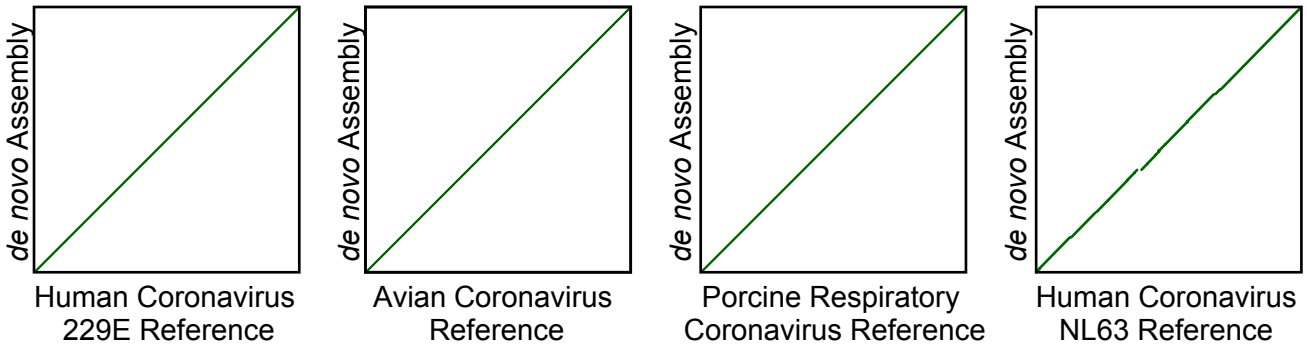
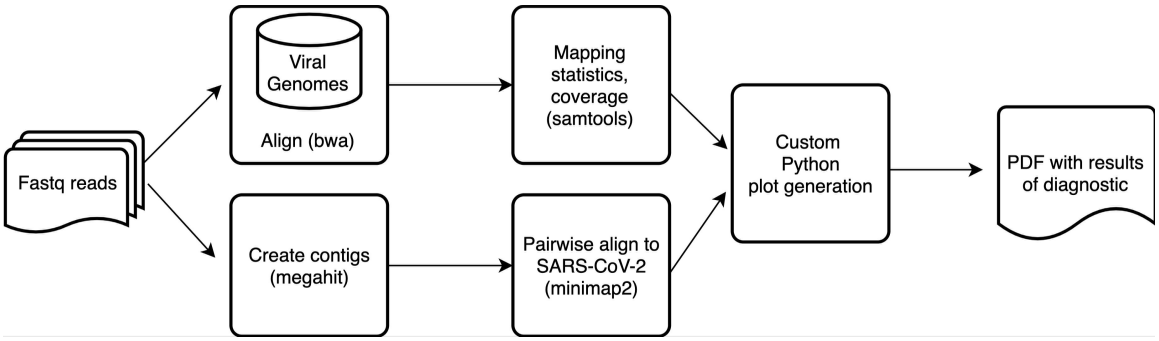
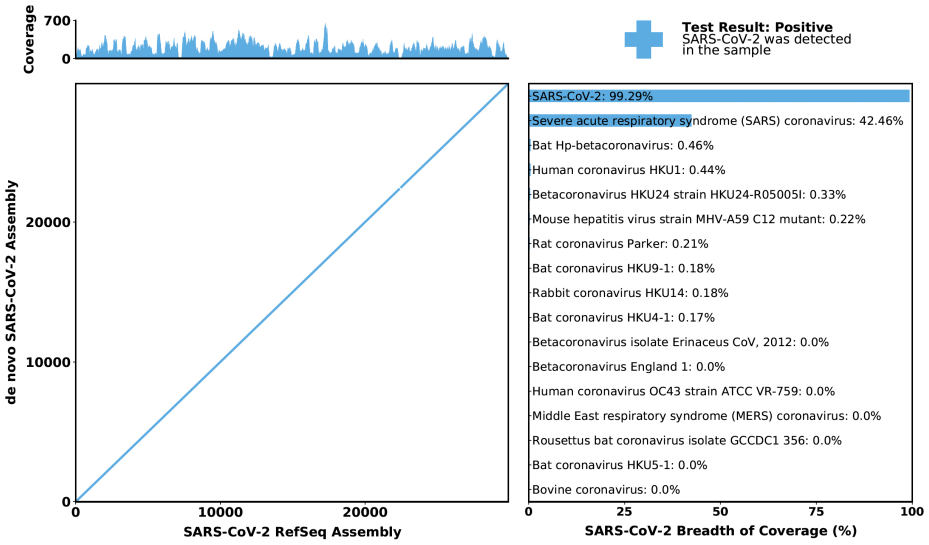


Figure 4. Non-SARS-CoV-2 Viruses Are Identified and Assembled without Compromising SARS-CoV-2 Detection. Genome dot plots comparing de novo assemblies and reference genomes for test samples spiked with non-SARS-CoV-2: Human Coronavirus strain 229E, Avian Coronavirus, Porcine Respiratory Coronavirus, and Human Coronavirus NL63. De novo assemblies on Y-axes, species matched reference genomes on X-axes. The de novo assemblies displayed on the Y-axes have been ordered and oriented to match the reference viral genomes in order to facilitate comparison. Each *de novo* assembly was generated using 130,000-200,000 75-PE reads.

A


Fastq reads
Viral Genomes
Align (bwa)
Create contigs (megahit)
Mapping statistics, coverage (samtools)
Pairwise align to SARS-CoV-2 (minimap2)
Custom Python plot generation
PDF with results of diagnostic

B


Coverage
700
0
Test Result: Positive
SARS-CoV-2 was detected in the sample
de novo SARS-CoV-2 Assembly
0
10000
20000
SARS-CoV-2 RefSeq Assembly
SARS-CoV-2: 99.29%
Severe acute respiratory syndrome (SARS) coronavirus: 42.46%
Bat Hp-betacoronavirus: 0.46%
Human coronavirus HKU1: 0.44%
Betacoronavirus HKU24 strain HKU24-R05005I: 0.33%
Mouse hepatitis virus strain MHV-A59 C12 mutant: 0.22%
Rat coronavirus Parker: 0.21%
Bat coronavirus HKU9-1: 0.18%
Rabbit coronavirus HKU14: 0.18%
Bat coronavirus HKU4-1: 0.17%
Betacoronavirus isolate Erinaceus CoV, 2012: 0.0%
Betacoronavirus England 1: 0.0%
Human coronavirus OC43 strain ATCC VR-759: 0.0%
Middle East respiratory syndrome (MERS) coronavirus: 0.0%
Rousettus bat coronavirus isolate GCCDC1 356: 0.0%
Bat coronavirus HKU5-1: 0.0%
Bovine coronavirus: 0.0%
0
25
50
75
100
SARS-CoV-2 Breadth of Coverage (%)

C

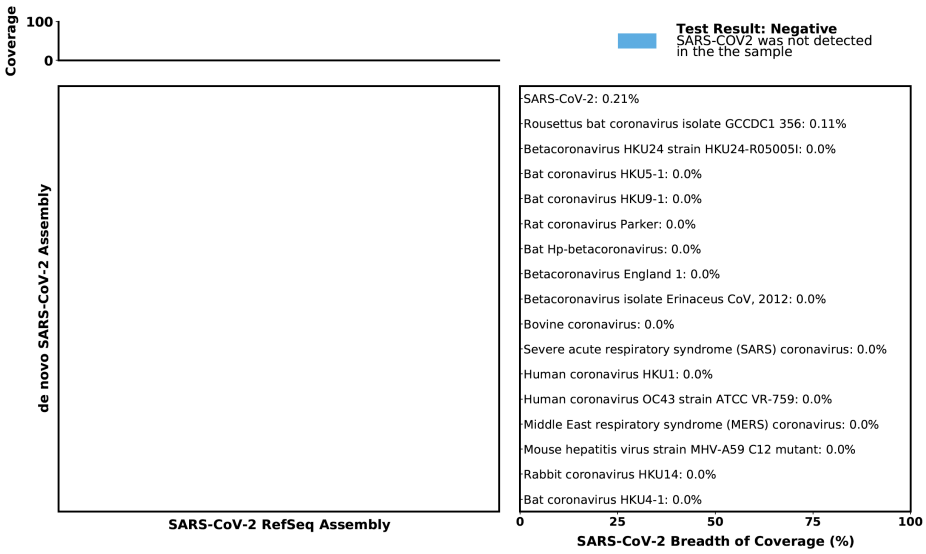

Coverage
100
0
Test Result: Negative
SARS-COV2 was not detected in the the sample
de novo SARS-CoV-2 Assembly
SARS-CoV-2 RefSeq Assembly
SARS-CoV-2: 0.21%
Rousettus bat coronavirus isolate GCCDC1 356: 0.11%
Betacoronavirus HKU24 strain HKU24-R05005I: 0.0%
Bat coronavirus HKU5-1: 0.0%
Bat coronavirus HKU9-1: 0.0%
Rat coronavirus Parker: 0.0%
Bat Hp-betacoronavirus: 0.0%
Betacoronavirus England 1: 0.0%
Betacoronavirus isolate Erinaceus CoV, 2012: 0.0%
Bovine coronavirus: 0.0%
Severe acute respiratory syndrome (SARS) coronavirus: 0.0%
Human coronavirus HKU1: 0.0%
Human coronavirus OC43 strain ATCC VR-759: 0.0%
Middle East respiratory syndrome (MERS) coronavirus: 0.0%
Mouse hepatitis virus strain MHV-A59 C12 mutant: 0.0%
Rabbit coronavirus HKU14: 0.0%
Bat coronavirus HKU4-1: 0.0%
0
25
50
75
100
SARS-CoV-2 Breadth of Coverage (%)

Figure 5. POLAR BEAR Pipeline Analyzes and Visualizes Test Results with a Single Click. (A) Workflow diagram describing the one-click analysis pipeline. The pipeline aligns the sequenced reads to a database of coronaviruses; if run on a cluster, this is done in parallel. Separately, the pipeline creates contigs from the sequenced reads. The resulting *de novo* assembly is then pairwise aligned to the SARS-CoV-2 reference genome. A custom python script then analyzes these data to determine the test result and compiles dot plots and alignment percentages into a single PDF. (B & C) Each report includes a genome dot plot of the de novo assembly against the SARS-CoV-2 reference genome, with a coverage track of sequenced reads aligned to the SARS-CoV-2 reference genome above the dot plot. The report also includes the breadth of coverage of sequenced reads aligned to 17 different coronaviruses. The diagnostic answer is given in the form of a “+” or “-” symbol and “Positive” or “Negative” for SARS-CoV-2 coronavirus in the top right corner of the report.

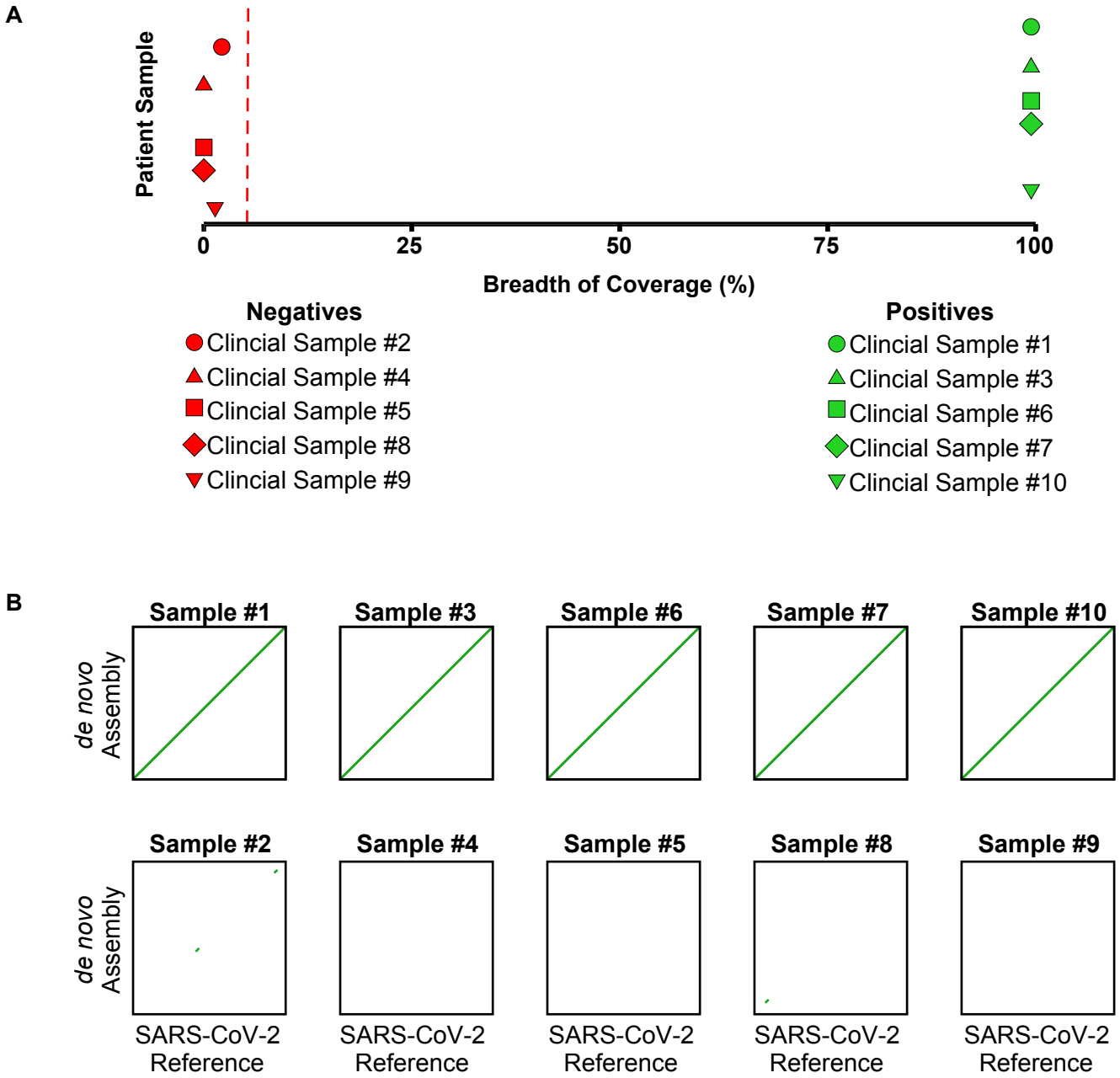


Figure 6. POLAR Protocol Generates *de novo* SARS-CoV-2 Genome Assemblies from Clinical Samples. (A) Scatter plot shows breadth of coverage for all ten clinical samples. Dashed red line represents the breadth of coverage threshold for positive samples. The breadth of coverage of each library was calculated using 150, 000 75-PE reads. (B) Each rescaled genome dot plot compares the *de novo* SARS-CoV-2 assembly (Y-axes) created directly from a clinical sample to the SARS-CoV-2 reference genome (X-axes). The *de novo* assemblies displayed on the Y-axes have been ordered and oriented to match the reference viral genome in order to facilitate comparison. Each line segment represents the position of an individual contig from the *de novo* assembly that aligned to the reference genome. For rescaled dot plots, contigs were sorted and unmapped contigs have been removed, leaving all remaining aligning contigs lying along the diagonal. Each *de novo* assembly was generated using 150,000 75-PE reads.

Genome Equivalents/mL	*de novo* Assembly (#)	Sequenced Read Pairs	Assembly Length (bp)	Number of Contigs	Number of Gaps	Contig N50 (bp)	Breadth Of Assembly	Base Accuracy
840,000	1	150,000	29,844	1	0	29,844	99.78%	100%
	2	150,000	29,852	1	0	29,852	99.82%	100%
	3	150,000	29,832	1	0	29,832	99.76%	100%
	4	150,000	29,835	1	0	29,835	99.82%	100%
84,000	5	150,000	29,843	1	0	29,843	99.78%	100%
	6	150,000	29,797	1	0	29,797	99.64%	100%
	7	150,000	29,835	1	0	29,835	99.82%	100%
	8	150,000	29,888	2	1	22,485	99.79%	100%
8,400	9	150,000	29,837	1	0	29,837	99.76%	100%
	10	150,000	29,825	1	0	29,825	99.74%	100%
	11	150,000	29,824	1	0	29,824	99.74%	100%
	12	150,000	29,887	2	1	21,252	99.79%	100%
840	13	150,000	27,548	11	10	6,250	89.59%	100%
	14	150,000	29,138	7	6	6,004	95.77%	100%
	15	150,000	24,836	18	17	5,316	81.13%	100%
	16	150,000	22,482	28	27	1,705	72.03%	99.99%
84	17	150,000	15,414	33	32	429	38.63%	99.99%
	18	150,000	129,301	404	403	311	29.27%	100%
	19	150,000	22,335	50	49	414	47.00%	100%
	20	150,000	8,356	13	12	972	21.75%	99.98%
Water	21	150,000	229,679	720	719	311	2.70%	100%
	22	150,000	136,720	444	443	298	3.81%	100%
	23	150,000	149,830	491	490	292	1.82%	100%
	24	150,000	271,999	871	870	302	3.79%	100%

Table 1. *De novo* Assembly Statistics of SARS-CoV-2 Genome Using Range of Starting SARS-CoV-2 Genome Concentrations.

Virus	Assembly Length (bp)	Number of Contigs	Number of Gaps	Contig N50 (bp)	Breadth of Assembly	Base Accuracy
Human Coronavirus strain 229E	27,327	1	0	27,327	99.79%	99.93%
Human Coronavirus NL63	26,903	15	14	2,823	96.01%	99.97%
Porcine Respiratory Coronavirus (PRCV), ISU-1	27,618	1	0	27,618	99.75%	99.98%
Avian Coronavirus	27,554	1	0	27,554	99.97%	99.95%

Table 2. Assembly Statistics of Other Coronavirus Genomes *de novo*.

Test Name, Manufacturer/Laboratory	Limit of Detection (Genomes/mL)
PerkinElmer New Coronavirus Nucleic Acid Detection Kit, PerkinElmer, Inc.	30
BD SARS-CoV-2Reagents for BD MAX System, Becton, Dickinson & Company	40
Viracor SARS-CoV-2 assay, Viracor Eurofins Clinical Diagnostics	73
POLAR: SARS-CoV-2 assay, The Center for Genome Architecture (TCGA)	84
Abbott RealTime SARS-CoV-2 assay, Abbott Molecular	100
Real-Time Fluorescent RT-PCR Kit for Detecting SARS-2019-nCoV, BGI Genomics Co. Ltd	100
QuantiVirus SARS-CoV-2 Test kit, DiaCarta, Inc	100
ID NOW COVID-19, Abbott Diagnostics Scarborough, Inc.	125
Quest SARS-CoV-2 rRT-PCR, Quest Diagnostics Infectious Disease, Inc.	136
NeuMoDx SARS-CoV-2 Assay, NeuMoDx Molecular, Inc.	150
Curative-Korva SARS-Cov-2 Assay, KorvaLabs Inc.	200
Xpert® Xpress SARS-CoV-2, Cepheid	250
Fosun COVID-19 RT-PCR Detection Kit, Fosun Pharma USA Inc.	300
BioFire COVID-19 Test, BioFire Defense, LLC	330
COVID-19 genesig Real-Time PCR assay, Primerdesign Ltd.	330
Simplexa COVID-19 Direct, DiaSorin Molecular LLC	500
GeneFinder COVID-19 Plus RealAmp Kit, OSANG Healthcare	500
QIAstat-Dx Respiratory SARS-CoV-2 Panel, QIAGEN GmbH	500
SDI SARS-CoV-2 Assay, Specialty Diagnostic (SDI) Laboratories	500
Lyra SARS-CoV-2 Assay, Quidel Corporation	800
CDC 2019-nCoV Real-Time RT-PCR Diagnostic Panel (CDC), Centers for Disease Control and Prevention's (CDC)	1,000
SARS-CoV-2 Fluorescent PCR Kit, Maccura Biotechnology (USA) LLC	1,000
Stanford SARS-CoV-2 assay, Stanford Health Care Clinical Virology Laboratory	1,000
UNC Health SARS-CoV-2 real-time RT-PCR test, University of North Carolina Medical Center	1,000
Smart Detect SARS-CoV-2 rRT-PCR Kit, InBios International, Inc	1,100
COVID-19 RT-PCR Test, Pathology/Laboratory Medicine Lab of Baptist Hospital Miami	2,000
SARS-CoV-2 PCR test, Yale New Haven Hospital,Clinical Virology Laboratory	2,000
Childrens-Altona-SARS-CoV-2 Assay, Infectious Diseases Diagnostics Laboratory (IDDL), Boston Children's Hospital	2,300
ScienCell SARS-CoV-2 Coronavirus Real-time RT-PCR, ScienCell Research Laboratories	3,160
iAMP COVID-19 Detection Kit, Atila BioSystems, Inc.	4,000
CDI Enhanced COVID-19 Test, Hackensack University Medical Center (HUMC) Molecular Pathology Laboratory	4,000
Logix Smart Coronavirus Disease 2019 (COVID-19) Kit, Co-Diagnostics, Inc.	4,290
NxTAG CoV Extended Panel Assay, Luminex Molecular Diagnostics, Inc.	5,000
MGH COVID-19 qPCR assay, Massachusetts General Hospital	5,000
Orig3n 2019 Novel Coronavirus (COVID-19) Test, Orig3n, Inc.	5,000
GS™ COVID-19 RT-PCR KIT, GenoSensor, LLC	6,250
COVID-19 RT-PCR Test, Laboratory Corporation of America (LabCorp)	6,250
COV-19 IDx assay, Ipsum Diagnostics, LLC	8,500
SARS-CoV-2 Assay, Integrity Laboratories	10,000
SARS-CoV-2 RT-PCR test, Infectious Disease Diagnostics Laboratory - Children's Hospital of Philadelphia	20,747
AvellinoCoV2 test, Avellino Lab USA, Inc.	55,000
ARIES SARS-CoV-2 Assay, Luminex Corporation	75,000
SARS-Cov-2 Assay, Diagnostic Molecular Laboratory – Northwestern Medicine	100,000
ePlex SARS-CoV-2 Test, GenMark Diagnostics, Inc.	100,000

Table 3. Compilation of The Limit of Detection of FDA Approved SARS-CoV-2 Diagnostic Tests.

Clinical Sample	Breadth of Coverage	Assembly Length (bp)	Number of Contigs	Number of Gaps	Contig N50 (bp)	Breadth of Assembly
#1	99.72	29,824	1	0	29,824	99.74%
#2	2.47	686	2	1	319	2.29%
#3	99.73	29,824	1	0	29,824	99.74%
#4	0	-	0	0	-	0.00%
#5	0	-	0	0	-	0.00%
#6	99.84	29,830	1	0	29,830	99.76%
#7	99.7	29,678	2	1	22,339	99.25%
#8	0	-	0	0	-	0.00%
#9	1.29	286	1	0	286	0.96%
#10	99.72	29,837	1	0	29,837	99.78%

Table 4. *De novo* Assembly Statistics for SARS-CoV-2 Genome created directly from clinical samples.

Reagent	Company	Catalog #	Reagent Price	Reagent Volume (µl)	Amount/Sample	Cost/Sample
sparQ PureMag Beads, Quantabio	Quantabio	76302-830	$3,950.79	4.50E+05	9.50E+01	$0.83
Quick-RNA Viral 96 Kit	Zymo	R1041	$716.00	3.84E+02	1.00E+00	$1.86
50uM random hexamers	ThermoFisher	N8080127	$91.00	1.00E+02	5.00E-01	$0.46
10mM dNTPs mix	NEB	N0447L	$250.00	4.00E+03	1.50E+00	$0.09
SuperScript™ IV Reverse Transcriptase	ThermoFisher	18090050	$383.00	5.00E+01	5.00E-01	$3.83
RNaseOUT Recombinant Ribonuclease Inhibitor	ThermoFisher	10777-019	$188.00	1.25E+02	5.00E-01	$0.75
Q5 Hot Start DNA polymerase	NEB	M0493S	$137.00	5.00E+01	5.00E-01	$1.37
Nuclease-Free Water	Qiagen	129114	$131.00	5.00E+05	2.55E+01	$0.01
Nextera XT DNA Library Preparation Kit	Illumina	FC-131-1096	$3,368.00	96 rxns	0.5 rxns	$17.54
Nextera XT Index Kit v2	Illumina	FC-131-2001	$1,051.00	384 rxns	0.5 rxns	$1.37
NextSeq550 (Mid-Output Flow Cell)	Illumina	20024904	$1,179.00	1.50E+07	3.91E+05	$3.07
					Total =	$31.18

Table S1. Per Sample Cost Breakdown of Reagents Needed to Perform the POLAR Method.

pair#	#bin	matches	strand	qName	qSize	qStart	qEnd	tName	tSize	tStart	tEnd	blockCount	blockSizes	qStarts	tStarts
1	585	21	+	Seq1_NIID_WH-1_F501	21	0	21	NC_045512v2	29903	483	504	1	21	0	483
	585	23	-	Seq1_NIID_WH-1_R913	23	0	23	NC_045512v2	29903	873	896	1	23	0	873
2	585	19	+	Seq1_NIID_WH-1_F509	19	0	19	NC_045512v2	29903	491	510	1	19	0	491
	585	22	-	Seq1_NIID_WH-1_R854	22	0	22	NC_045512v2	29903	815	837	1	22	0	815
3	585	20	+	Seq1_NIID_WH-1_Seq_F519	20	0	20	NC_045512v2	29903	501	521	1	20	0	501
	585	19	-	Seq1_NIID_WH-1_Seq_R840	19	0	19	NC_045512v2	29903	804	823	1	19	0	804
4	585	18	+	FR-Pasteur_nCoV_IP2-12669Fw	18	0	18	NC_045512v2	29903	12689	12707	1	18	0	12689
	585	18	-	FR-Pasteur_nCoV_IP2-12759Rv	18	0	18	NC_045512v2	29903	12779	12797	1	18	0	12779
5	585	19	+	FR-Pasteur_nCoV_IP4-14059Fw	19	0	19	NC_045512v2	29903	14079	14098	1	19	0	14079
	585	19	+	FR-Pasteur_nCoV_IP4-14146Rv	19	0	19	NC_045512v2	29903	14079	14098	1	19	0	14079
6	585	20	+	US-CDC_2019-nCoV_N1-F	20	0	20	NC_045512v2	29903	28286	28306	1	20	0	28286
	585	24	-	US-CDC_2019-nCoV_N1-R	24	0	24	NC_045512v2	29903	28334	28358	1	24	0	28334
7	585	26	+	EU-Drosten_E_Sarbeco_F	26	0	26	NC_045512v2	29903	26268	26294	1	26	0	26268
	585	22	-	EU-Drosten_E_Sarbeco_R	22	0	22	NC_045512v2	29903	26359	26381	1	22	0	26359
8	585	20	+	WH-NIC-N-F-	20	0	20	NC_045512v2	29903	28319	28339	1	20	0	28319
	585	19	-	WH-NIC-N-R-	19	0	19	NC_045512v2	29903	28357	28376	1	19	0	28357
9	585	21	+	EU-Drosten_RdRP_SARSr-F	22	0	22	NC_045512v2	29903	15430	15452	1	22	0	15430
	585	24	-	EU-Drosten_RdRP_SARSr-R	26	0	26	NC_045512v2	29903	15504	15530	1	26	0	15504
10	585	20	+	US-CDC_2019-nCoV_N2-F	20	0	20	NC_045512v2	29903	29163	29183	1	20	0	29163
	585	18	-	US-CDC_2019-nCoV_N2-R	18	0	18	NC_045512v2	29903	29212	29230	1	18	0	29212
11	585	22	+	HKU-NF	22	0	22	NC_045512v2	29903	29144	29166	1	22	0	29144
	585	19	-	HKU-NR	19	0	19	NC_045512v2	29903	29235	29254	1	19	0	29235
12	585	18	+	HKU-ORF1b-nsp14F	20	0	20	NC_045512v2	29903	18777	18797	1	20	0	18777
	585	20	-	HKU--ORF1b-nsp14R	21	0	21	NC_045512v2	29903	18888	18909	1	21	0	18888
13	585	21	+	Seq2-NIID_9_2nd_NIID_WH-1_F24381	21	0	21	NC_045512v2	29903	24363	24384	1	21	0	24363
	585	23	-	Seq2-NIID_10_2nd_NIID_WH-1_R24873	23	0	23	NC_045512v2	29903	24833	24856	1	23	0	24833
14	585	21	+	Seq2-NIID_11_Seq_NIID_WH-1_Seq_F24383	21	0	21	NC_045512v2	29903	24365	24386	1	21	0	24365
	585	19	-	Seq2-NIID_12_Seq_NIID_WH-1_Seq_R24865	19	0	19	NC_045512v2	29903	24829	24848	1	19	0	24829
15	585	21	+	Seq2-NIID_11_Seq_NIID_WH-1_Seq_F24383	21	0	21	NC_045512v2	29903	24365	24386	1	21	0	24365
	585	19	-	Seq2-NIID_12_Seq_NIID_WH-1_Seq_R24865	19	0	19	NC_045512v2	29903	24829	24848	1	19	0	24829
16	585	22	+	US-CDC-EXCL_2019-nCoV_N3-F	22	0	22	NC_045512v2	29903	28680	28702	1	22	0	28680
	585	21	-	US-CDC-EXCL_2019-nCoV_N3-R	21	0	21	NC_045512v2	29903	28731	28752	1	21	0	28731
17	585	19	+	EU-Drosten-Unused_N_Sarbeco_F	19	0	19	NC_045512v2	29903	28705	28724	1	19	0	28705
	585	20	-	EU-Drosten-Unused_N_Sarbeco_R	20	0	20	NC_045512v2	29903	28813	28833	1	20	0	28813
18	585	20	+	NIID_2019-nCOV_N_F2	20	0	20	NC_045512v2	29903	29124	29144	1	20	0	29124
	585	19	-	NIID_2019-nCOV_N_R2	20	0	20	NC_045512v2	29903	29262	29282	1	20	0	29262
19	585	24	+	Seq2_WuhanCoV-spk1-f	24	0	24	NC_045512v2	29903	24353	24377	1	24	0	24353
	585	25	-	Seq2-NIID_8_1st-WuhanCoV-spk2-r	25	0	25	NC_045512v2	29903	24875	24900	1	25	0	24875
20	585	22	+	CN-CDC_primer4	22	0	22	NC_045512v2	29903	28880	28902	1	22	0	28880
	585	22	-	CN-CDC_primer5	22	0	22	NC_045512v2	29903	28957	28979	1	22	0	28957
21	585	21	+	CN-CDC_primer1	21	0	21	NC_045512v2	29903	13341	13362	1	21	0	13341
	585	19	-	CN-CDC_primer2	19	0	19	NC_045512v2	29903	13441	13460	1	19	0	13441

Table S2. List of SARS-CoV-2 RT-PCR Primer Regions.

Systems Beta-Tested	Resource Type	Processor (for each node or 'virtual 'instance')	Cores available (per instance/node)	Runtime
DUG KNL	HPC	Intel Xeon Phi 7250 @ 1.60GHz	68 cores per node	109s
DUG HighPerf	HPC	Dual Intel Xeon Platinum 9242 @ 2.3 - 3.8 GHz	96 cores per node	36s
Pawsey Zeus	HPC	Intel Xeon E5-2680 v4 2.4 GHz	28 cores per node	64s
Pawsey Nimbus	Cloud	AMD EPYC Processor x86_64 2.34GHz	16 vCPUS: n3.16c64r	78s
Microsoft Azure	Cloud	Intel Xeon Platinum 8168 2.7GHz base; 3.4 GHz - 3.7 GHz max	2 vCPUs: F2S_v2	75s
Docker	HPC	Intel(R) Xeon(R) CPU E5-2690 v3 @ 2.60GHz	24	59 s
Docker	HPC	Intel(R) Xeon(R) Gold 6126 CPU @ 2.60GHz	48	41 s
Docker	HPC	Intel(R) Xeon(R) CPU X5660 @ 2.80GHz	48	46 s

Table S3. Benchmarking parameters for "Pipeline for POLAR BEAR: viral diagnostic for SARS-CoV-2"