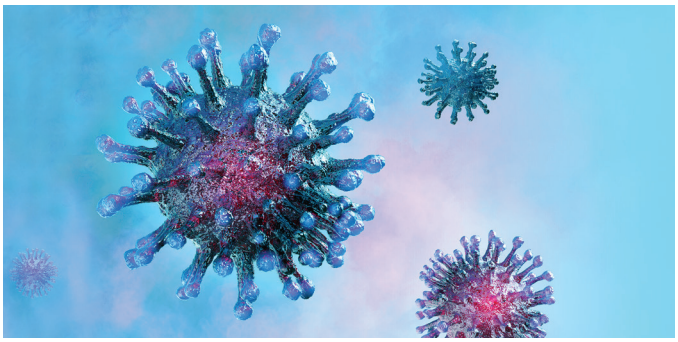
GenapSys™ Sequencing Platform:

SARS-CoV-2 Genome Sequencing using a Hybrid Capture Enrichment Panel

- **Decentralized and low cost NGS solution for surveillance and testing for the COVID-19 pandemic**
- **Accurate detection of SARS-CoV-2 variants, with uniform genome coverage using a hybrid capture panel**
- **Detection of a wide range of viral titers (5-orders of magnitude) and good performance at low read depths**
- **Multiplex a large number of samples to enable higher throughput and a lower price per sample**



Introduction

The ongoing outbreak of the novel coronavirus (SARS-CoV-2) that began in late 2019, has now spread globally and become a critical health issue for many countries. This novel virus is from a large family of viruses, Coronaviruses (CoV), which are responsible for causing a range of disorders from the common cold to more serious disease outbreaks such as the Severe Acute Respiratory Syndrome (SARS-CoV) in 2003 and the Middle East Respiratory Syndrome (MERS-CoV) in 2012. With more than two million confirmed cases around the world at the time of this writing, the World Health Organization has declared the disease associated with SARS-CoV-2 (COVID-19) a pandemic.

SARS-CoV-2, is a single-stranded RNA virus. Detection of the virus in a sample is typically done with Reverse Transcription PCR (RT-PCR) which utilizes primers to amplify specific limited regions of interest from the SARS-CoV-2 genome and provide a rapid response. In addition to performing rapid testing, it is equally important to track the spread and mutational evolution of the viral genome in an accurate and distributed manner.

Next-Generation Sequencing (NGS) provides an effective and comprehensive method to interrogate the full genome of the virus and its mutations vs. the limited focus of current testing methodologies. Epidemiological studies require more information than just the presence or absence of the virus, and a complete view of the genome across a large number of patient samples can provide the ability to monitor transmission and viral evolution. NGS is a powerful tool that provides a comprehensive view of the viral genome and allows the ability to correlate the host genotype and immune response to the virus. Thus, it enables essential insights into viral function and evolution and can help better direct therapeutic research and vaccine development efforts.

Technology

The GenapSys™ Sequencer employs a novel electrical detection method that is capable of generating highly accurate DNA sequence information. With a CMOS based detector, simple fluidics, and low computational requirements, the GenapSys instrument is small, affordable, and accessible even to novice genomic scientists. Inside

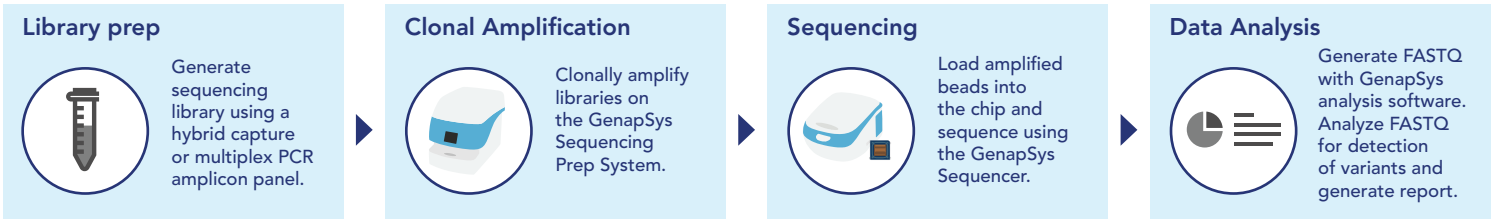


Fig. 1 The GenapSys workflow for the SARS-CoV-2 assay starts with RNA extracted from patient samples, followed by library preparation using a hybrid capture panel or a multiplex PCR amplicon panel. The libraries are clonally amplified on the GenapSys Sequencing Prep System and are sequenced using the GenapSys Sequencer. The result consist of FASTQ files for each sample and variant calls using the GenapSys data analysis pipeline, in the form of a summary report.

the sequencing chip are millions of individual sensors, each loaded with a single bead coated in thousands of clonal copies of an individual DNA template. Individual nucleotides are flowed across the chip in succession and successful incorporation is detected by changes in electrical impedance as the complementary DNA strand grows. These impedance changes are converted into high accuracy base calls and the results are output as FASTQ files.

Methods

GenapSys has demonstrated high accuracy sequencing of the full SARS-CoV-2 genome using both hybrid capture panels and multiplex amplicon PCR panels. Here, we highlight the performance of the GenapSys workflow using a hybrid capture panel for SARS-CoV-2, over a wide range of viral loads (5 orders of magnitude) and reads per sample (3 order of magnitude). Briefly, two synthetic SARS-CoV-2 RNA controls that differ by 4 variants were spiked into Universal Human Reference RNA (UHRR) at different viral copy numbers. The RNA samples were converted to cDNA, followed by genomic library construction, and hybrid capture enrichment of the full viral genome. The libraries were clonally amplified on the GenapSys Sequencing Prep System and sequenced using the GenapSys Sequencer, as shown in Fig 1. The FASTQ files generated were analyzed for genomic coverage performance and detection of variants at different reads per sample.

We used synthetic RNA SARS-CoV-2 control standards generated by Twist Bioscience, which were based on the reference genome of MN908947.3 (ncbi.nlm.nih.gov/nuccore/MN908947) and MT007544.1 (ncbi.nlm.nih.gov/nuccore/MT007544) (Catalog # 102024 and 102019 respectively). These controls are split into six non-overlapping fragments, and quantitated to 1 million viral genome copies/µL. To demonstrate assay performance across a wide range of viral titers, samples were created by spiking in as many as 1 million copies and as low as 10 copies of each standard into 10 ng of UHRR (Agilent Part # 740000), and a No Template Control (NTC) with just 10 ng of UHRR. The RNA mix was converted to cDNA, using the NEBNext® Ultra™ II RNA First Strand Synthesis (FSS) Module (E7771L) and NEBNext® Ultra II Directional RNA Second Strand Synthesis Module (E7550L). These kits were used according to the manufacturer's recommendations, with FSS module parameters resulting in fragment inserts of 200 bp length. This was followed by End Repair, A-tailing, Adapter Ligation and Dual Indexed PCR using the NEBNext® Ultra™ II DNA Library Prep Kit (E7645L) and NEBNext® Multiplex Oligos (E6440L, E7600, or E7780). Unique Dual Indexes are recommended for library preparation. Alternatively, one may use the NEB ProtoScript® II First Strand cDNA Synthesis Kit (E6560L) and NEBNext® Ultra II Non-Directional RNA Second Strand Synthesis Module (E6111L), and an enzymatic shearing based library preparation approach, such as the NEBNext® Ultra™ II FS DNA Library Prep (E6177L), or the Twist Bioscience Library Preparation Kit using Enzymatic Fragmentation (PN 100401) and Unique Dual Indexes (PN 101307). Fragmentation times and bead-based size selection should be optimized for the enzymatic fragmentation protocol to generate a library with a median insert size of 200 bp and a tight size distribution.

Sample libraries are typically pooled together before hybrid capture enrichment of the viral reads. Hybrid capture was performed using the Twist Bioscience SARS-CoV-2 Research Panel (PN 102017) using 750 ng of library in single-plex capture reactions and 1500 ng for a multiplexed sample. Washes were performed with the Twist Bioscience Fast Hybridization Kit (PN 101174) with an 8 hr. hybridization time. Shorter hybridization times are possible with the kit, but were not tested for this study. It is recommended that hybrid capture of multiplex libraries should be performed with samples that have similar viral loads. This is because widely varying viral loads will result in selective enrichment and follow-up PCR amplification of samples with higher viral loads, resulting in most reads being assigned to those samples. Samples can be batched into similar viral loads based on orthogonal tests such as RT-qPCR. Alternatively, sample library inputs before the hybrid capture can be adjusted based on expected viral loads from orthogonal tests such as RT-qPCR. Single sample or multiplexed libraries are clonally amplified using the GenapSys Sequencing Prep instrument. Amplified libraries are sequenced on the GenapSys Sequencer, using the supplied protocol.

BWA MEM is used to align reads to the SARS-CoV-2 genome sequence and the human reference genome (hg38). Sequencing runs with multiple samples are demultiplexed, and a FASTQ file as well as a SARS-CoV-2 whole genome sequencing report are generated per sample. The report specifies the number of viral reads, depth of coverage and variant calls with reference to the Genbank (MN908947.3) / RefSeq (NC_045512) sequence. Variant calling was performed using GenapSys' proprietary variant calling pipeline.

Genome coverage of SARS-CoV-2 Controls & Identification of Variants

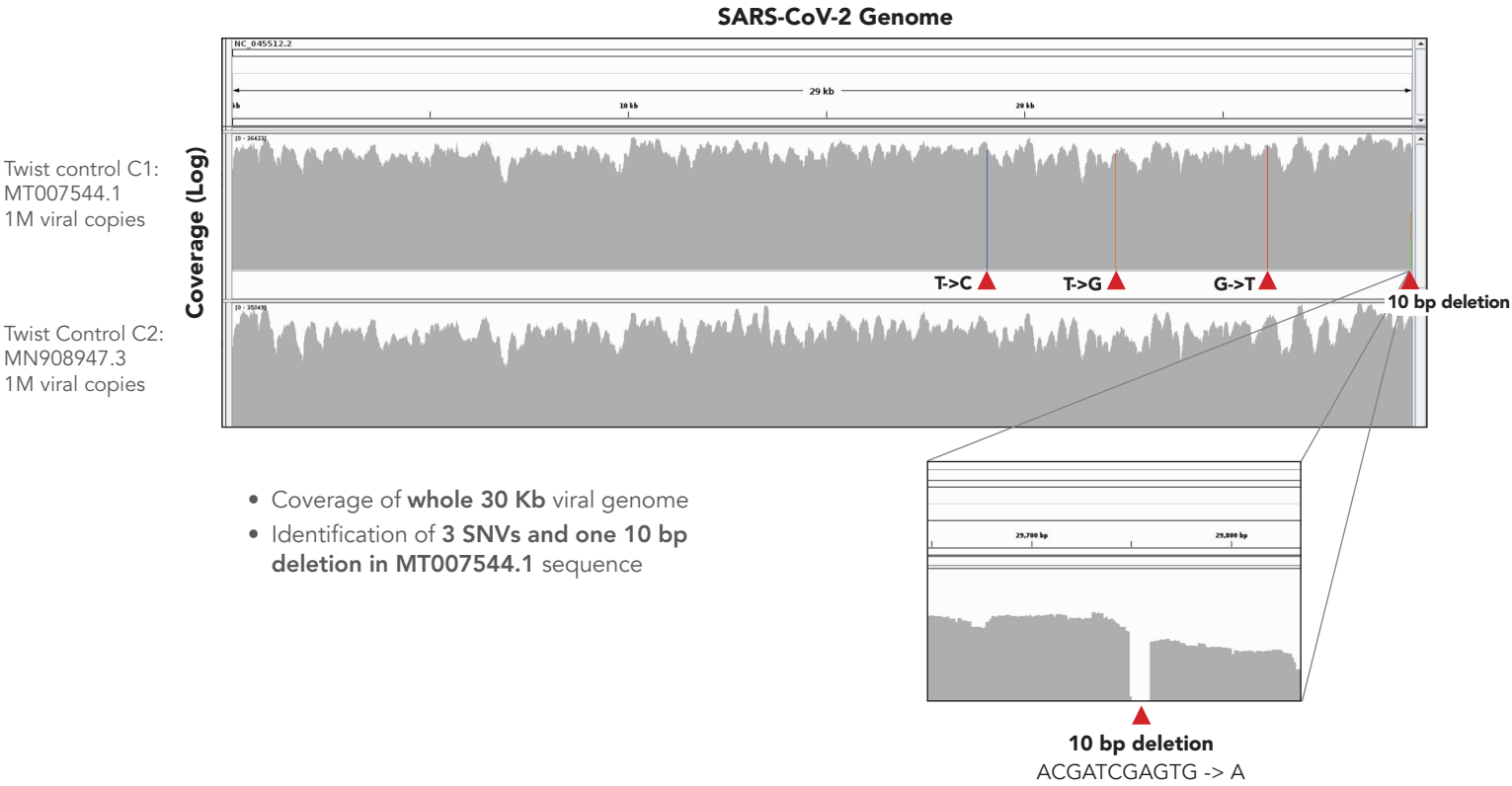


Fig. 2 The GenapSys SARS-CoV-2 assay generates 100% coverage of the whole 30 Kb viral genome for the MT007544.1 and MN908947.3 control standards from Twist Bioscience (1M viral copy spike in samples, respectively). Additionally, the MT007544.1 strain shows the presence of 4 variants: 3 SNVs (T -> C at position 19065, T -> G at position 22303, G ->T at position 26144) and a 10 bp deletion (ACGATCGAGTG -> A) at position 29749. The bottom panel shows a zoomed in view of the 10 bp deletion.

Results

We first demonstrated the genomic coverage of the whole ~30 Kb SARS-CoV-2 viral genome using this workflow. As shown in Fig. 2, we sequenced samples with 1M viral copies of the SARS-CoV2-MT007544.1 and SARS-CoV2-MN908947.3 respectively, and observed 100% coverage of the genome, both at 1X and 30X coverage. Additionally, the results confirm the presence of the 4 variants: 3 SNVs and 1 deletion in the MT007544.1 standard as compared to the reference genome. Further analysis of the sample with 1M viral copies of the SARS-CoV2-MT007544.1 genome demonstrates the high genome coverage and accurate variant calling using the GenapSys SARS-CoV-2 assay across different reads per sample (see Table 1). These metrics were calculated for a sequencing run which was downsampled to 10M, 1M, 250K, 100K, 25K and 10K reads. The fraction of viral reads out of the aligned reads was > 99% for all read counts, indicating high enrichment efficiency. The genome coverage at ⩾ 1X was 100% for all read depths (even as low as 10K reads), and genome coverage at ⩾ 30X was 99% at 100K reads. Importantly, all 4 variants were detected at all the read counts in this analysis. The ability to achieve high genome coverage and variant calling at low read counts shows the potential for multiplexing large numbers of samples.

We next tested the performance of the SARS-CoV-2 assay across a wide range of viral loads. Table 2 demonstrates the detection of SARS-CoV-2 virus with high genome coverage across 5 orders of magnitude change in viral loads and 100-fold variation in reads/sample. Samples were generated with 1M, 1K, 100 and 10 viral genome copies of the MT007544.1 control sample, and 1M, 1K and 10 viral genome copies of the MN908947.3 control sample, spiked into 10 ng of UHRR respectively. 100% of the viral genome was covered at ⩾ 1X for samples with viral loads of 1,000 copies or higher, across a wide range of reads per sample, even as low as 25K reads. Viral copies as low as 100 or even 10 copies were detected

by this assay, showcasing its detection ability. Fig. 3 shows that the GenapSys SARS-CoV-2 workflow has the potential for high sample multiplexing. The theoretical sample multiplexing calculations are based on the genome coverage (at 1x) and number of reads per sample from Table 2, given 12M reads per sequencing run (G3 Sequencing Chip). Note that multiplexing throughput in practice depends on the viral titers of patient samples, genomic coverage desired per sample, and the ability to multiplex several samples in a hybrid capture reaction by balancing for library input based on viral titers.

Finally, we also calculated the hybrid capture-based enrichment with the GenapSys SARS-CoV-2 workflow, shown in Table 3. Viral Fraction Pre-Capture is calculated as the ratio of the mass of the viral genome and the UHRR weight. The Viral Fraction Post-Capture is calculated as the ratio of reads aligned to the viral genome to the total number of reads aligned to the human hg38 & viral genomes. Fold enrichment is calculated as a ratio of the viral fraction post capture to pre capture. The results demonstrate high enrichment of 5000-fold for the 1M viral copy sample and up to 3 million-fold enrichment for the 10 viral copy sample. The negative control sample (NTC) showed a small fraction of viral reads (less than the fraction for the 10 copy sample), likely due to contamination during parallel handling of samples with high viral loads.

The GenapSys workflow for SARS-CoV-2 whole genome sequencing produces a FASTQ file as well as a report for each sample (see Fig. 4). The report references the number of reads mapped to the SARS-CoV-2 genome and the Human reference genome (hg 38), the mean read depth, the genome coverage at ⩾1X, and the variants detected. The report lists the variants with reference to the RefSeq (NC_045512) sequence, and the depth and frequency of the variant calls.

Conclusion

The GenapSys SARS-CoV-2 workflow solution provides a decentralized, low-cost and accurate NGS solution for whole genome sequencing of the virus and detection of variants along with high sample multiplexing. The ability to have rapid access to such genome-wide data across different geographies can provide public health experts with the ability to track the viral evolution, perform surveillance, and develop solutions to manage the current SARS-CoV-2 pandemic and monitor future viral outbreaks. Additionally, the GenapSys SARS-CoV-2 workflow can be combined with sequencing of patient gene variants and immune response genes, to highlight the underlying causes behind the wide variation in severity of COVID-19 response in patients. The GenapSys platform can enable a broad range of academic and biopharma labs who are performing vaccine research and developing therapeutic solutions.

Genome Coverage and Variant Calling : C1- MT007544.1 sample : 1M viral copies						
Number of Reads	10K	25K	100K	250K	1M	10M
% Fraction of Viral Reads in Aligned Reads	99.1	99.1	99.19	99.21	99.2	99.2
% of genome represented at 1X	100	100	100	100	100	100
% of genome represented at 30X	64	93	99	100	100	100
Number of Variants called	4/4	4/4	4/4	4/4	4/4	4/4

Table 1: High genome coverage and accurate variant calling at low read depth using the GenapSys SARS-CoV-2 assay. The table shows sequencing metrics of the MT007544.1 control sample, with 1 M viral genome copies spiked into 10 ng of UHRR. The fraction of viral reads out of the aligned reads was > 99% for all read depths. Genome coverage at ⩾1X was 100% at all read depths, and genome coverage at ⩾30X was 99% at 100K reads. Additionally, all 4 variants were detected at a low number of reads per sample.

To learn more about the GenapSys Sequencing Platform, visit GenapSys.com

Sample		Percent of Genome covered at 1X				
Viral Copy #	Control Type	10K Reads	25K Reads	100K Reads	250K Reads	1M Reads
1M	C1: MT007544.1	100	100	100	100	100
1M	C2: MN908947.3	99	100	100	100	100
1000	C1: MT007544.1	98	100	100	100	100
1000	C2: MN908947.3	99	100	100	100	100
100	C1: MT007544.1	26	44	69	78	83
10	C1: MT007544.1	5	10	19	23	32
10	C2: MN908947.3	11	17	25	30	53

Table 2: Detection of SARS-CoV-2 virus with high genome coverage across 5 orders of viral loads and 100-fold variation in reads per sample. Samples were generated with 1M, 1K, 100 and 10 viral genome copies of the MT007544.1 control sample, and 1M, 1K and 10 viral genome copies of the MN908947.3 control sample, spiked into 10 ng of UHRR respectively. 100% of the viral genome was covered at ⩾1X for samples with 1,000 viral copies, even at 25K reads. Viral copies as low as 10 copies were detected by this assay.

Viral Copy #	Control Type	UHRR (ng)	# Reads	Viral Fraction Pre-Capture %	Viral Fraction Post-Capture %	Fold Enrichment
1M	C1: MT007544.1	10	1M	0.016	99.18	6199
1M	C2: MN908947.3	10	1M	0.016	90.67	5667
1K	C1: MT007544.1	10	1M	0.000016	21.83	1364375
1K	C2: MN908947.3	10	1M	0.000016	36.01	2250625
100	C1: MT007544.1	10	1M	0.0000016	0.93	581250
10	C1: MT007544.1	10	1M	0.00000016	0.2	1250000
10	C2: MN908947.3	10	1M	0.00000016	0.48	3000000

Table 3: High viral enrichment with the SARS-CoV-2 hybrid capture workflow. Samples with viral copies ranging from 10 copies to 1M copies in 10 ng of UHRR, were enriched using the Twist Bioscience SARS-CoV-2 panel and the GenapSys workflow described in the Methods section. Fold enrichment ranges from 5000-fold to 3 million-fold depending on the sample.

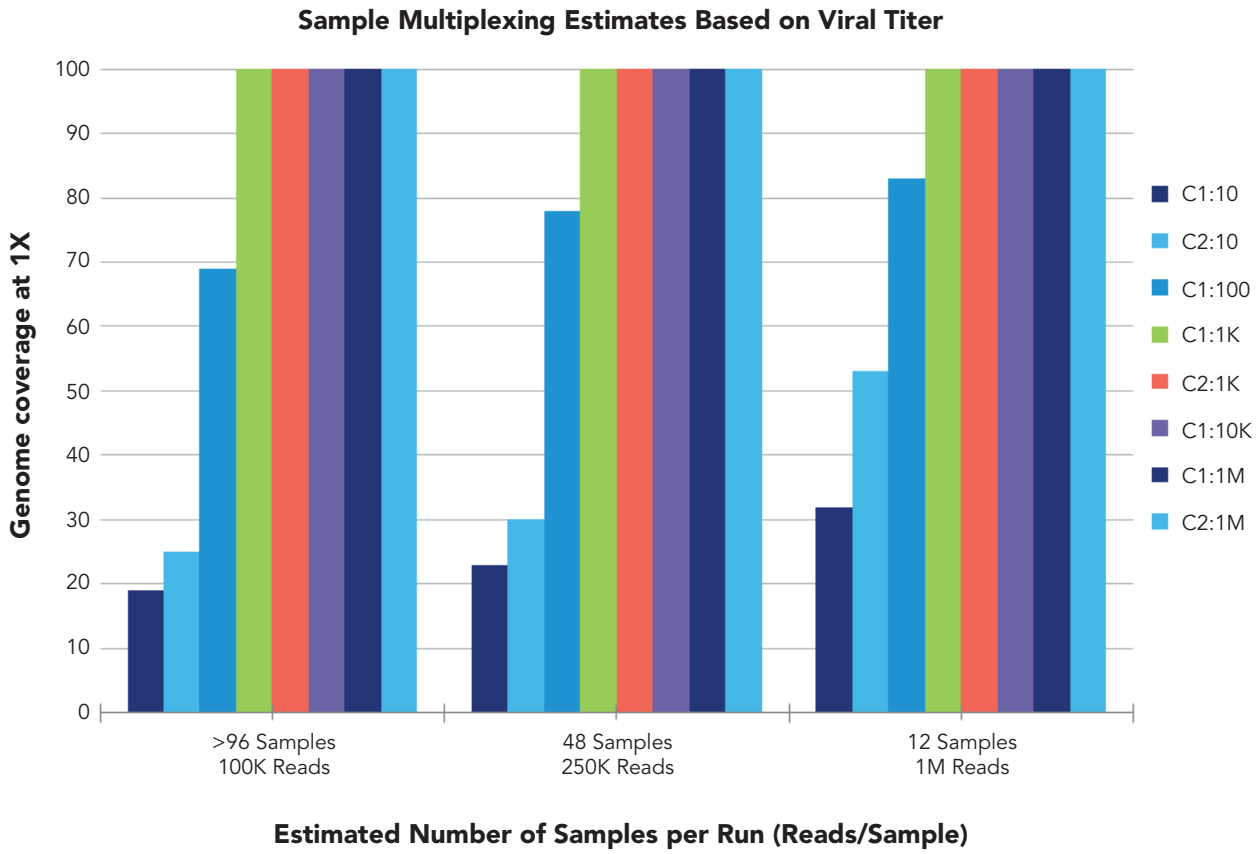
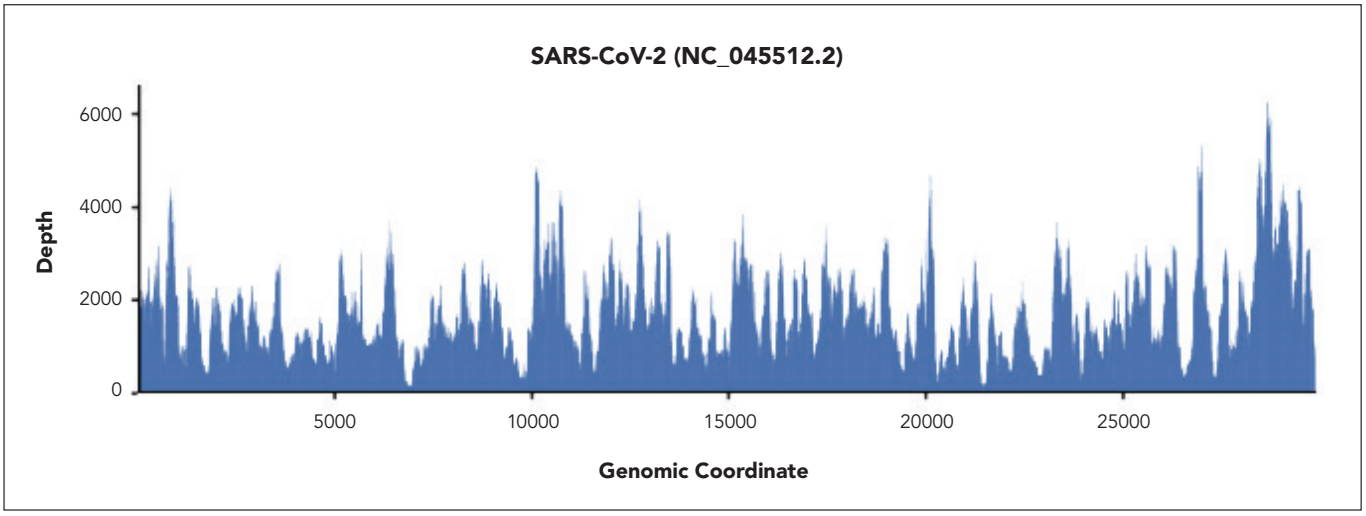


Fig. 3 Sample multiplexing projections with the SARS-CoV-2 whole genome sequencing assay. The estimated number of samples per sequencing run was calculated, based on the genomic coverage at 1x and reads per sample listed in Table 2, and is based on 12M reads (typically 10-13M) reads obtained with the G3 sequencing chip.

To learn more about the GenapSys Sequencing Platform, visit GenapSys.com

SARS-CoV-2 (COVID-19) Sequencing Report (sample_1)

The sample contained 3,457 reads that mapped to the human genome (HG38) and 406,827 reads that mapped to the SARS-CoV-2 genome. With a mean depth of 1,768x, reads cover 100% of the SARS-CoV-2 genome. A total of 4 variants were detected.



Position	Variant	Depth	Frequency
19065	T -> C	1892	99%
22303	T -> G	1543	97%
26144	G -> T	2350	99%
29749	ACGATCGAGTG -> A	2507	98%

Fig. 4 An example of the SARS-CoV-2 whole genome sequencing report for each sample

GENAPSYS, INC.
200 Cardinal Way, 3rd Floor, Redwood City, CA 94063 • (800) 796-7710 • GenapSys.com

GenapSys
Sequencing without compromise

100-2020-004-A