APPLICATION NOTE

Enrichment workflow for detecting coronavirus using Illumina NGS systems

Highly sensitive detection and characterization of common respiratory viruses, including coronavirus strains

illumina®

Viral surveillance with NGS

Viral infections are a major global health concern, with new infectious disease variants continually emerging. The 2019 outbreak of novel coronavirus (SARS-CoV-2) that began in Wuhan, China and quickly spread to multiple countries is an important example of the potential risks of global viral pandemics. Coronaviruses (CoV) are a large family of viruses that can infect humans, causing respiratory illnesses ranging from the common cold to more severe diseases, such as Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV). The World Health Organization (WHO) designated the illness associated with SARS-CoV-2 (COVID-19) a public health emergency of international concern, highlighting the need for rapid, accurate viral detection.[1]

Next-generation sequencing (NGS) provides an effective, novel way to screen samples and detect viruses without previous knowledge of the infectious agent.[2] Target enrichment is a sequencing method that captures genomic regions of interest by hybridization of nucleic acids to target-specific biotinylated probes. Hybrid–capture target enrichment allows for highly sensitive detection, without requiring the high read depth needed for shotgun metagenomics sequencing. Additionally, the target enrichment NGS workflow allows for near-complete sequence data of targets and opens up applications such as variant analysis for viral evolution or viral surveillance.[3] Compared to other targeted sequencing methods, such as amplicon sequencing, enrichment through hybrid capture allows for dramatically larger probe panels with more comprehensive profiling of the target regions. Additionally, the oligo probes used for hybrid–capture protocols remain effective, even within highly mutagenic regions, allowing targeting of rapidly evolving viruses, such as RNA viruses.

Viral enrichment workflow

This application note highlights a streamlined workflow for detecting and analyzing coronavirus using the Illumina DNA Prep with Enrichment library preparation kit, viral targeting panels, proven Illumina sequencing, and simplified data analysis (Figure 1). This workflow is intended to enrich viral DNA and RNA targets from total nucleic acid extraction. The protocol begins with extraction of nucleic acid samples and subsequent reverse transcription of extracted RNA into a double-stranded cDNA. Library preparation is performed using Illumina DNA Prep with Enrichment with IDT for Illumina DNA/RNA UD Indexes. In this protocol, DNA and cDNA undergo tagmentation, clean-up, and pre-enrichment amplification. After amplification, up to 12 samples can be pooled for one enrichment reaction using a panel of oligos that target viral sequences. Probe hybridization is followed by probe capture, enrichment amplification, quantification, and sequencing.

Methods

Sample preparation

To demonstrate the performance of the viral target enrichment panel, a strain of deactivated coronavirus sample, CoV strain OC43 from Microbiologics (QC Sets and Panels: Helix Elite, Catalog no. 8217), was analyzed (Table 1). The coronavirus viral culture sample was

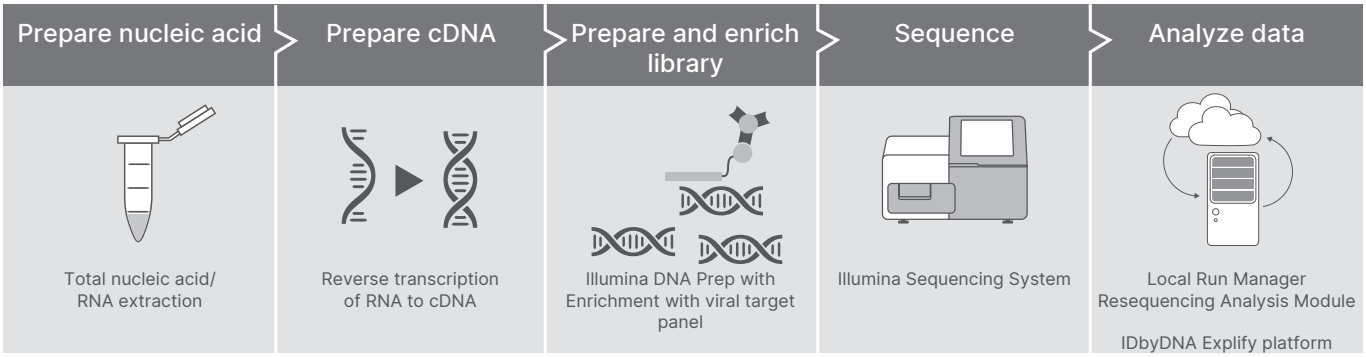

Figure 1: Enrichment workflow for coronavirus detection—The streamlined NGS workflow for coronavirus detection integrates sample preparation, library preparation, target enrichment, sequencing, and data analysis.

extracted using the QIAGEN QIAmp Viral Mini Kit (QIAGEN, Catalog no. 52904) in a BSL2 laboratory environment. 150 ng of extracted RNA was reversed transcribed into cDNA using two different workflows: one derived from Illumina TruSeq RNA reagents and the other using Thermo Fisher Scientific Maxima H Minus Double-Stranded cDNA Synthesis Kit (Thermo Fisher Scientific, Catalog no. K2561). To mimic sequencing results from patient samples, the viral sample was also spiked into 95% Universal Human Reference (UHR) background RNA (Agilent Technologies, Catalog no. 74000) and reverse transcribed into cDNA using the methods described above (Table 1).

Additionally, a respiratory viral pool containing four RNA viruses and one DNA virus (Table 2) was extracted using the QIAGEN AllPrep PowerViral DNA/RNA Kit (QIAGEN, Catalog no. 28000-50) in a BSL2 laboratory environment. The maximum volume of total nucleic acid was reversed transcribed into cDNA (without DNase treatment) using a workflow derived from Illumina TruSeq RNA reagents. To mimic sequencing results from patient samples with multiple pathogens present, the viral pool sample was also spiked into 95% Universal Human Reference (UHR) background RNA (Agilent Technologies, Catalog no. 74000) and reverse transcribed into cDNA (Table 1).

Table 1: Composition of viral samples for analysis

Sample	Composition[a]	Reference genome
CoVOC43	150 ng CoV OC43 RNA	AY391777.1 Human coronavirus OC43
CoVOC43_95UHR	7.5 ng CoV OC43 RNA and 142.5 ng UHR RNA	AY391777.1 Human coronavirus OC43
ViralPool	Max volume viral pool total nucleic acid (8.5 µl)	See Table 2
ViralPool_95UHR	5% viral pool total nucleic acid and 95% UHR RNA by volume	See Table 2

a. The recommended minimum RNA/total nucleic acid input for reverse transcription is 10 ng. For best results, reverse transcription should be performed on freshly extracted nucleic acid samples.

Table 2: Composition of viral pool

Virus	Nucleic acid type	Reference genome
Influenza A virus (H1N1)	RNA	Influenza A/ Michigan/45/2015
Influenza B virus	RNA	Influenza B/ Colorado/06/2017
Human Parainfluenza virus 3	RNA	NC_001796.2
RSV B9320	RNA	AY353550.1
Adenovirus 7	DNA	AY594255.1

Library preparation

Sequencing-ready libraries were prepared with Illumina DNA Prep with Enrichment (Illumina, Catalog no. 20025524) and IDT for Illumina DNA UD Indexes (Illumina, Catalog no. 20027213)* using cDNA from the CoV sample (CoVOC43), the viral pool sample (ViralPool), and both spike-in samples (CoVOC43_95UHR and ViralPool_95UHR). The total DNA input recommended for tagmentation is 10–1000 ng.

After amplification, the samples were split into separate pools for enrichment based on sample type (CoVOC43, CoVOC43_95UHR, ViralPool, and ViralPool_95HR). Enrichment reactions were performed with the Respiratory Virus Oligos Panel (Illumina, Catalog no. 20042472), which features ~7800 probes designed to detect respiratory viruses, recent flu strains, and SARS-CoV-2, as well as human probes to act as positive controls (Table 3). After enrichment (Figure 2), the prepared libraries were quantified, pooled, and loaded onto the MiSeq™ System for sequencing.

* Catalog no. 20027213 is now obsolete, replaced by Illumina DNA/RNA UD Indexes Set A, Tagmentation (96 indexes, 96 samples), Catalog no. 20091654.

Table 3: Viruses targeted by the Respiratory Virus Oligos Panel

Human coronavirus 229E	Human parainfluenza virus 1
Human coronavirus NL63	Human parainfluenza virus 2
Human coronavirus OC43	Human parainfluenza virus 3
Human coronavirus HKU1	Human parainfluenza virus 4a
SARS-CoV-2	Human metapneumovirus (CAN97-83)
Human adenovirus B1	Respiratory syncytial virus (type A)
Human adenovirus C2	Human Respiratory syncytial virus 9320 (type B)
Human adenovirus E4	Influenza A virus (A/Puerto Rico/8/1934(H1N1))
Human bocavirus 1 (Primate bocaparvovirus 1 isolate st2)	Influenza A virus (A/Korea/426/1968(H2N2))
Human bocavirus 2c PK isolate PK-5510	Influenza A virus (A/New York/392/2004(H3N2))
Human bocavirus 3	Influenza A virus (A/goose/Guangdong/1/1996(H5N1))
Human bocavirus 4 NI strain HBoV4-NI-385	Influenza A virus (A/Zhejiang/DTID-ZJU01/2013(H7N9))
KI polyomavirus Stockholm 60	Influenza A virus (A/Hong Kong/1073/99(H9N2))
WU Polyomavirus	Influenza A virus (A/Texas/50/2012(H3N2))
Human parechovirus type 1	PicoBank/HPeV1/a Influenza A virus (A/Michigan/45/2015(H1N1))
Human parechovirus 6	Influenza B virus (B/Lee/1940)
Human rhinovirus A89	Influenza B virus (B/Wisconsin/01/2010)
Human rhinovirus C (strain 024)	Influenza B virus (B/Brisbane/60/2008)
Human rhinovirus B14	Influenza B virus (B/Colorado/06/2017)
Human enterovirus C104 strain: AK11	Influenza B virus (B/Washington/02/2019)
Human enterovirus C109 isolate NICA08-4327	Human control genes

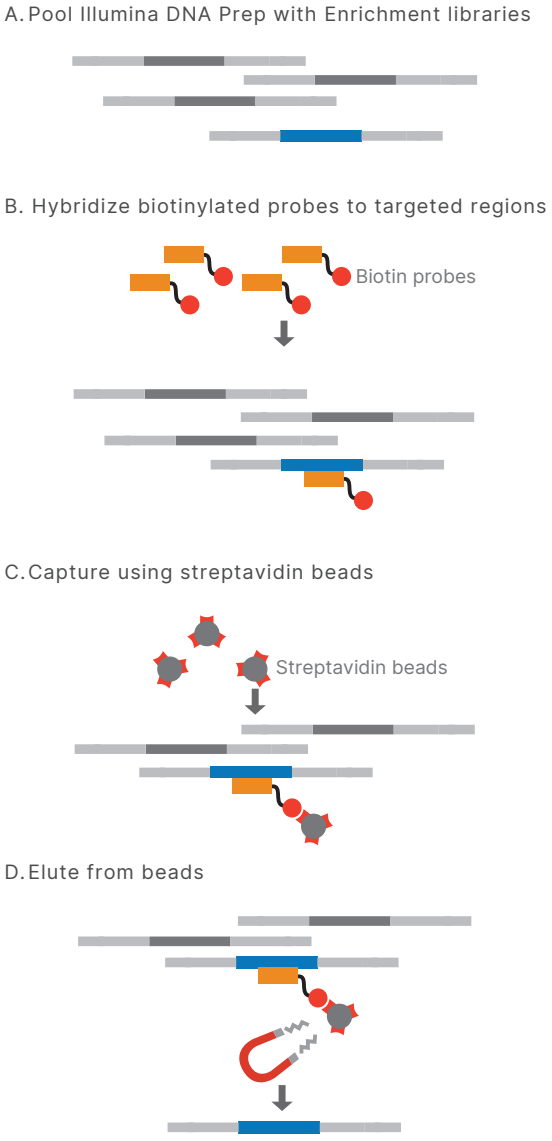

Figure 2: Enrichment chemistry—Prepared libraries undergo enrichment with a viral target panel, followed by amplification, quantification, and sequencing.

Sequencing

Prepared libraries can be sequenced on any Illumina instrument. The benchtop iSeq™ 100, MiniSeq™, and MiSeq Systems are well suited due to the low read requirements for these samples. In this study, libraries prepared from the viral samples were denatured and diluted to a final loading concentration of 10 pM, according to the MiSeq System

Denature and Dilute Libraries Guide (Document no. 15039740 v10) and sequenced on the MiSeq System at 2 × 151 bp read length using MiSeq v3 reagents (Illumina, Catalog no. MS-102-3001). Aliquots of all libraries were also sequenced prior to probe hybridization to determine the fold change of the enrichment reaction.

Virus titer, RNA quality, and the number of reads per sample impact the number of virus-specific reads obtained and coverage of the viral genome. As a general guideline, the read recommendation for this workflow is 500K reads per sample but these numbers can be variable and this is only a recommended starting point.

Data analysis

For automatic on-instrument analysis, sequencing runs were set up in Local Run Manager using the Resequencing Analysis Module. This module allows for input of all run information and reference genome selection for subsequent sequence alignment. Users can upload reference genomes directly to the instrument, allowing for easy customization. Analysis is kicked off automatically after sequencing is complete so users can interpret results as quickly as possible. The Resequencing module provides alignment, coverage, and small variant data as well as FASTQ, BAM, and VCF files for use in other data analysis pipelines, if desired (Figure 3).

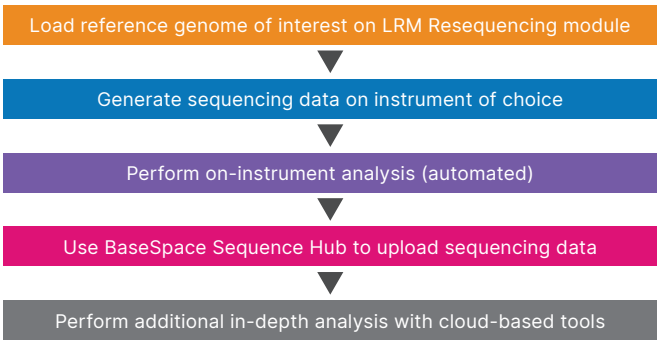

Figure 3: Enrichment analysis workflow—Sequencing runs are set up in Local Run Manager with data analysis automatically proceeding after the run is complete. Additional analysis with third-party applications is available, if desired.

The IDbyDNA Explify Platform is an easy-to-use solution for in-depth data analysis that features robust data quality control (QC), standardized result interpretation, carefully curated databases, and custom report generation. Data analysis is based on k-mers and alignment steps, including protein-level detection of viruses, which increases the ability to identify novel and highly divergent viruses. The platform can be accessed in BaseSpace™ Sequence Hub.

Results

After library preparation and sequencing, the Local Run Manager Resequencing Analysis Module (v2.5.56.11) was used to align each sample to the coronavirus OC43 reference genome, as described in the Local Run Manager Resequencing Analysis Module Workflow Guide.

Reverse transcription and detection of CoV with the Local Run Manager Resequencing Module

In this study, two different reverse transcription kits were used to generate cDNA for library preparation—a commercially available kit from Thermo Fisher Scientific and a workflow derived from Illumina TruSeq RNA reagents (see Methods for details). For CoVOC43 samples, results showed that percent alignment to the reference genome was > 99% for the Illumina reverse transcription method and > 97% for the Thermo Fisher Scientific method (Figure 4). Percent alignment of the spike-in samples were > 64% and > 19% for the Respiratory Virus Oligos Panel, regardless of the reverse transcription method (Figure 4).

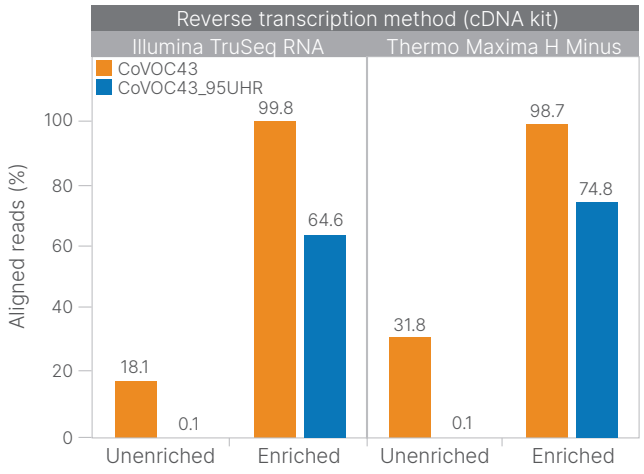

Figure 4: Equivalent performance with different cDNA kits—The percent aligned reads for viral samples reverse transcribed with the two different cDNA kits (see Methods section) are plotted before (unenriched) and after (enriched) enrichment with the Respiratory Virus Oligos Panel.

As a measure of the success of the target capture and hybridization reaction, the percent aligned reads from samples pre- and postenrichment were compared to calculate the fold change of enrichment. Results showed significant increases in fold enrichment for the CoV samples analyzed regardless of reverse transcription method. This was especially true for CoVOC43_95UHR samples, which showed > 640-fold enrichment for the Respiratory Virus Oligos Panel (Table 4).

The mean coverage provides an overview of the depth of coverage for each base in the reference genome, identifying any regions that may not have been sequenced. Enrichment with viral target probes greatly increased the mean coverage of the viral sequence for all samples, regardless of reverse transcription method (Table 5).

Table 4: Respiratory Virus Oligo Panel enrichment metrics

	Illumina TruSeq RNA		Thermo Maxima H Minus	
Sample	CoVOC43	CoVOC43_95UHR	CoVOC43	CoVOC43_95UHR
Fold enrichment	5.5×	646×	3.1×	748×

Table 5: Mean coverage with Respiratory Virus Oligos Panel[a]

cDNA kit	CoVOC43		CoVOC43_95UHR	
	Unenriched	Enriched	Unenriched	Enriched
Illumina TruSeq RNA	427.1	3283.5	2.6	2437.3
Thermo Maxima H Minus	987.0	3861.4	4.2	3035.0

a. Mean coverage metrics have been normalized to 1M reads per sample.

Analysis with IDbyDNA Explify platform

The Explify platform identified the spiked coronavirus (Figure 5) and all five viruses in the viral pool (Figure 6). The Explify platform provided viral consensus genome sequences, coverage plots, and the demonstrated ability to detect coinfections with other viruses, bacteria, fungi, or parasites (Figure 5 and Figure 6).

The Explify platform also demonstrated that CoVOC43_95UHR enrichment data from the Respiratory Virus Oligos Panel showed equivalent coverage to shotgun metagenomics sequencing data[4] (Figure 7), demonstrating that enrichment with viral target probe pools does not introduce coverage bias compared to the shotgun method.

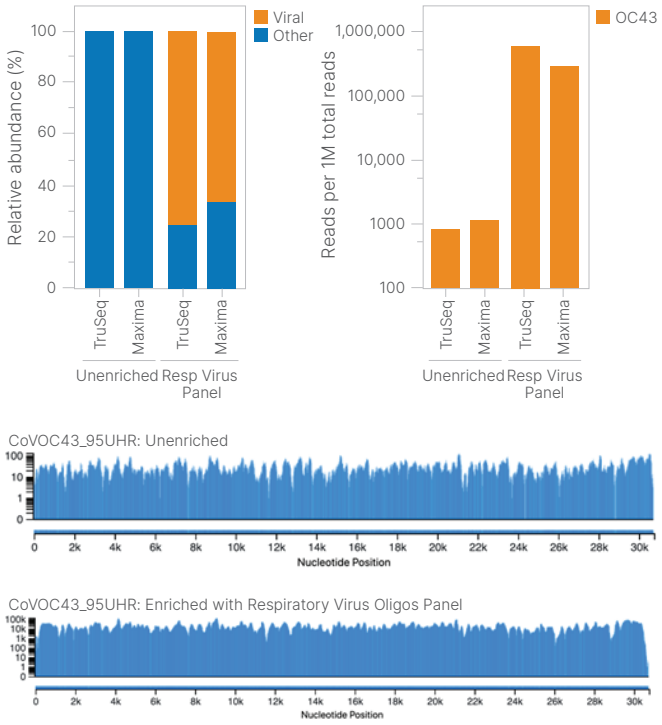

Figure 5: Identification of OC43—OC43 was identified, regardless of cDNA kit, with high genomic coverage. Data was downsampled to 1M reads per sample.

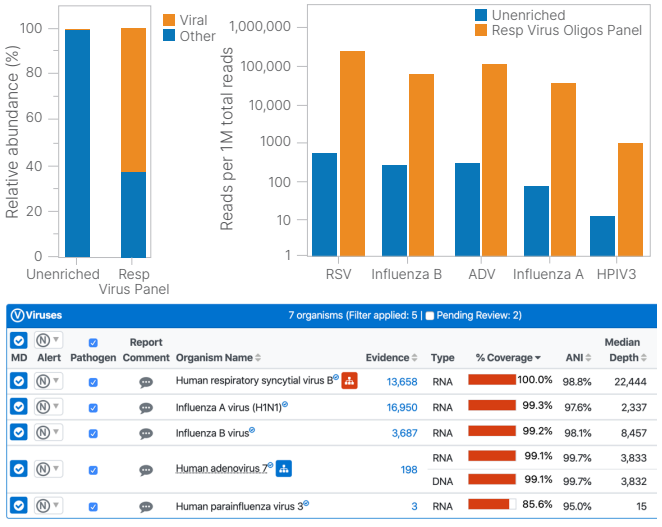

Figure 6: Characterization of viral pools—All RNA and DNA viral strains were identified in the viral pool. The viral pool was not verified with orthogonal techniques, so there is no guarantee of equivalent representation of each virus. Data was downsampled to 1M reads per sample.

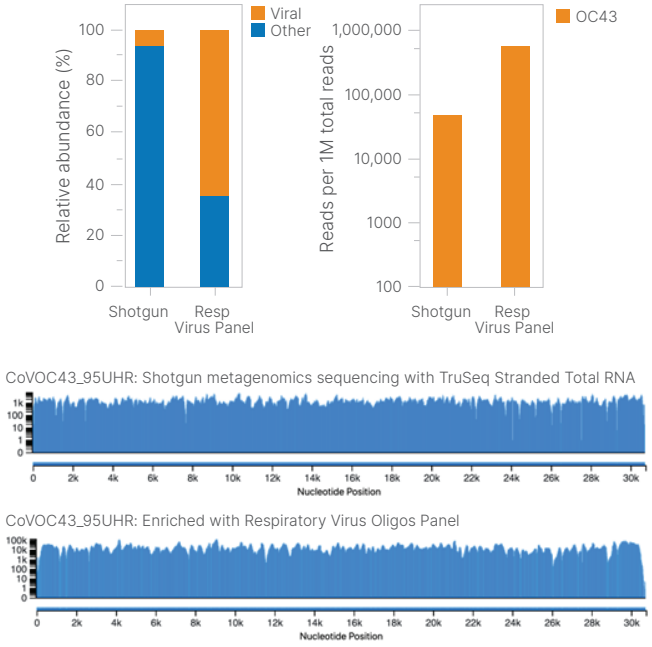

Figure 7: Equivalent performance with shotgun method—Enrichment showed equivalent performance and coverage to shotgun metagenomics sequencing.

Summary

The identification and characterization of emerging viruses is central to improving public health. In these situations, NGS is a powerful method for broad-range detection to identify known and emerging viruses. Using Illumina DNA Prep with Enrichment with panels that target viral pathogens enables researchers to obtain genomic data that can confirm the presence of CoV and advance analyses such as genotyping and variant analysis. The agnostic design allows for widespread identification of pathogenic viruses across all sample types of interest and the use of unique dual indexes reduces the risk of any indexing crossover from multiplexing samples. This easy-to-follow workflow enables detection and characterization of pathogen outbreaks, such as the novel SARS-CoV-2.

Learn more

Viral sequencing methods

References

1. World Health Organization. COVID 19 Public Health Emergency of International Concern (PHEIC) Global research and innovation forum: towards a research roadmap. Published 12 February, 2020. Accessed 9 September 2023.
2. Bulcha B. Review on viral metagenomics and its future perspective in zoonotic and arboviral disease surveillance. *J Biol Agric Healthcare*. 2017;7(21):35–41.
3. Gaudin M, Desnues C. Hybrid Capture-Based Next Generation Sequencing and Its Application to Human Infectious Diseases. *Front Microbiol*. 2018;9:2924. Published 2018 Nov 27. doi:10.3389/fmicb.2018.02924
4. Illumina. Comprehensive workflow for detecting coronavirus using Illumina benchtop systems. Published 2020. Accessed August 29, 2023.

illumina®

1.800.809.4566 toll-free (US) | +1.858.202.4566 tel
techsupport@illumina.com | www.illumina.com

M-GL-02146 v1.0