illumina®

Comprehensive workflow for detecting coronavirus using Illumina benchtop systems

A shotgun metagenomics sequencing workflow for effective detection and characterization of coronavirus strains.

Introduction

Viral infections are a major global health concern, and new infectious diseases continue to emerge that affect public health. The 2019 outbreak of novel coronavirus (SARS-CoV-2) that began in Wuhan, China and has quickly spread to multiple countries is a particularly concerning example. 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). SARS-CoV-2 is a new strain that has not been previously identified in humans. With tens of thousands of confirmed cases around the world and a death toll that has surpassed the SARS epidemic of 2003, the World Health Organization (WHO) has declared the disease associated with SARS-CoV-2 (COVID-19) a public health emergency of international concern.[1]

Next-generation sequencing (NGS) provides an effective, novel way to screen samples and detect viruses without previous knowledge of the infectious agent.[2] Shotgun metagenomics is a sequencing method that comprehensively examines all organisms present in a given complex sample. It enables detection of viral pathogens without concerns of highly mutagenic regions that can be difficult for amplicon-based assays such as qPCR.[3] Beyond merely detecting viral particles, NGS provides a detailed view of the viral genome, enabling valuable insights into viral function and biology. The ability to have near-complete sequence data of viral genomes allows for implementation of effective viral surveillance strategies to prevent further transmission and infection.

This application note highlights a streamlined workflow for detection of coronavirus in control samples using the TruSeq™ Stranded Total RNA Library Preparation kit, proven Illumina sequencing, and simplified data analysis (Figure 1).

Methods

Sample preparation

To demonstrate the performance of a shotgun metagenomic workflow for detecting coronavirus, commercially available coronavirus samples, including coronavirus strains OC43 and 229E from Microbiologics (QC Sets and Panels: Helix Elite; Cat no. 8217) were used in this study. Each coronavirus was run as two different sample types: first, as extracted material and second as a 5% by mass spike into a mock human RNA background with Universal Human Reference (UHR) RNA (Agilent Technologies, PN 74000) (Table 1). The purified viral samples are referred to as CoVOC43CtrlCult and CoV229ECtrlCult to mimic the results of a viral culture, while the spiked viral samples in human background are referred to as CoVOC43CtrlPat and CoV229ECtrlPat to mimic samples with host background.

Library preparation

After RNA extraction (QIAGEN QIAmp Viral Mini Kit, PN 52904), 200 ng of input RNA from each sample was taken through the Ribo-Zero Gold rRNA depletion protocol to remove human cytoplasmic and mitochondrial rRNA (Illumina, 48 samples, Cat no. 20020598, 96 samples, 20020599).

Sequencing-ready libraries were prepared from depleted RNA using the TruSeq Stranded Total RNA Library Prep Gold kit (Illumina, Cat no. 20020599) along with the IDT for Illumina TruSeq RNA UD

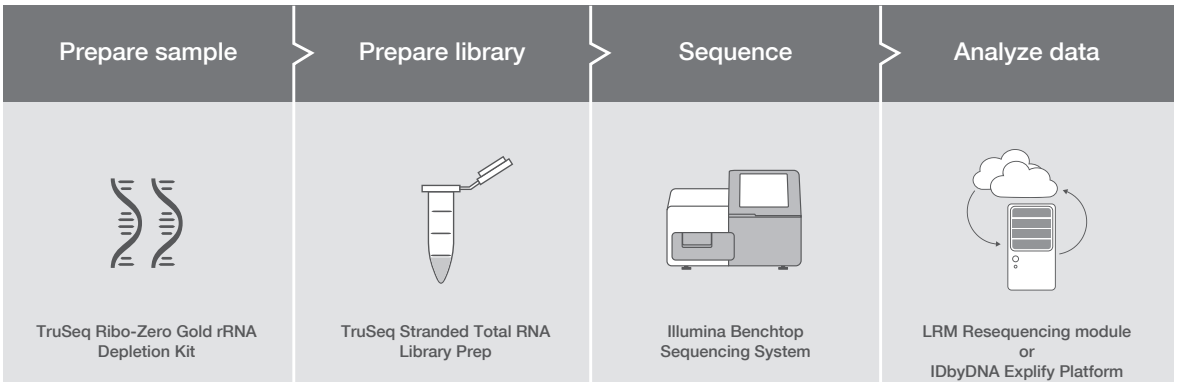

Figure 1: Shotgun metagenomics workflow for coronavirus detection—The streamlined NGS workflow for coronavirus detection integrates sample preparation, library preparation, sequencing, and data analysis.

Table 1: Composition of input samples

Sample	Composition	Reference genome
CoVOC43CtrlPat	10 ng coronavirus OC43 RNA and 190 ng UHR RNA	AY391777.1 Human coronavirus OC43
CoV229ECtrlPat	10 ng coronavirus 229E RNA and 190 ng UHR RNA	NC_002645.1 Human coronavirus 229E
CoVOC43CtrlCult	200 ng coronavirus OC43 RNA	AY391777.1 Human coronavirus OC43
CoV229ECtrlCult	200 ng coronavirus 229E RNA	NC_002645.1 Human coronavirus 229E

Indexes (96 indexes, 96 samples) (Illumina, Cat no. 20022371). RNA underwent fragmentation, first- and second-strand cDNA synthesis, adenylation, adapter ligation, and amplification, according to the TruSeq Stranded Total RNA protocol.[4] After amplification, the prepared libraries were quantified, pooled, and loaded onto the MiSeq™ System for sequencing.

Sequencing

Shotgun metagenomic sequencing can be performed on any Illumina instrument, but samples prepared directly from swabs or similar matrices are most suited for the NextSeq™ Series of Systems due to the recommended 10M reads per sample. It is possible to run libraries prepared from those samples on other benchtop instruments, but there may be lower depth of coverage of the target as a result. Libraries prepared from viral culture using the same workflow are particularly well suited for the benchtop iSeq™ 100, MiniSeq™, and MiSeq Systems due to the lower recommended read count of 500,000 reads per sample.

Libraries prepared from the coronavirus control samples were sequenced on the MiSeq System at 2 × 76 bp read length. Two separate MiSeq runs were used for the two sample types, due to the higher read count necessary for the 5% by mass spiked viruses in the UHR background.

Virus titer, efficiency of human rRNA depletion, and the number of reads per sample impact the number of virus-specific reads obtained and coverage of the viral genome. A general guideline includes 10M reads for direct-from-patient samples and 0.5M reads for positive virus culture, but these numbers can be variable and are only a recommended starting point.

Data analysis

Local data analysis can be performed using the Illumina Local Run Manager (LRM) Resequencing Module for any desired reference genome. For this application note, the coronavirus reference genomes for strains 229E and OC43 were used (NC_002645.1 Human coronavirus 229E and AY391777.1 Human coronavirus OC43). The LRM resequencing module provides alignment to the reference genome, depth of coverage for the reference genome, a summary of the identified small variants, including single nucleotide variants (SNVs) and insertions and deletions (indels), a summary of the fragment length observed, and duplicate information for library diversity. Additional commercial tools are available if cloud-based data analysis is not possible, but they will need to be evaluated by the end user (Figure 2).

For cloud-based, in-depth analysis, the IDbyDNA Explify Platform enables comprehensive identification of more than 35,000 viruses using a proprietary database of curated DNA and RNA reference sequences.

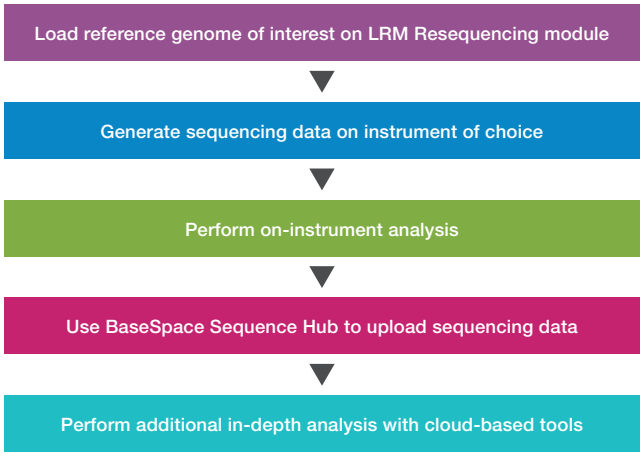

Figure 2: Shotgun metagenomics analysis workflow—Data analysis follows a simplified workflow using LRM and third-party applications.

Results

LRM Resequencing Module

After library preparation and sequencing, the LRM Resequencing Module (v2.5.56.11) was used to analyze the reference coronavirus genomes for each sample, as described in the Local Run Manager Resequencing Analysis Module Workflow Guide (Document # 1000000002705 v01). The samples showed percent alignment of reads to their respective reference genomes of 78.8% and 71.5% for the purified viral RNA, and 5.5% and 6.1% for the viral RNA in human background (Table 2). Consistent SNV calls were made between samples with and without the host background (Table 2). Additionally, the mean coverage provided an overview of the depth of coverage for each base in the reference genome, identifying any regions that may not have been sequenced. All samples showed alignment to, and coverage of, the reference genomes, indicating successful identification of the viral target of interest (Figure 3). As expected, the purified viral RNA showed significantly higher alignment and coverage compared to the spiked samples due to the host background.

Table 2: Basic metrics using LRM Resequencing Module

Sample	Aligned reads	Mean coverage	Small variants
CoVOC43CtrlPat	5.5%	1786.8	47
CoV229ECtrlPat	6.1%	2266.5	24
CoVOC43CtrlCult	78.8%	25,971.5	47
CoV229ECtrlCult	71.5%	33,833.3	24

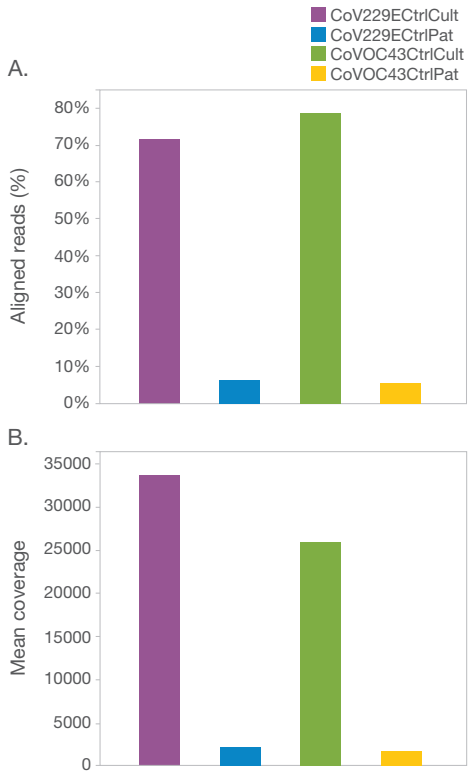

Figure 3: Identification of coronavirus 229E and OC43 spiked into human background—The control CoV samples were successfully identified, indicated by (A) aligned reads and (B) mean coverage.

IDbyDNA Explify

The Explify Platform provides 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 Explify Platform identified the spiked coronaviruses, and it provided viral consensus genome sequences, coverage plots, and the ability to detect co-infections with other viruses (eg, HPV18 contained in UHR), bacteria, fungi, or parasites (Figure 4). A simulated data set demonstrated detection of SARS-CoV-2 (2019-nCoV) by the Explify Platform.

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 TruSeq Stranded Total RNA Library Prep for viral shotgun metagenomics sequencing and analysis, researchers are able to obtain genomic data that can confirm the presence of a viral pathogen without prior knowledge and continue with further 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 including proven Illumina sequencing enables detection and characterization of pathogen outbreaks such as the novel SARS-CoV-2.

Learn more

Learn more about viral sequencing methods at www.illumina.com/areas-of-interest/microbiology/infectious-disease-surveillance.html

Learn more about LRM analysis at www.illumina.com/products/by-type/informatics-products/local-run-manager.html

Learn more about the IDbyDNA Explify analysis platform at www.idbydna.com/explify-platform/

References

1. World Health Organization. WHO Director-General's statement on IHR Emergency Committee on Novel Coronavirus (2019-nCoV). www.who.int/dg/speeches/detail/who-director-general-s-statement-on-ihr-emergency-committee-on-novel-coronavirus-(2019-ncov). 30 January 2020.
2. Bulcha B. Review on viral metagenomics and its future perspective in zoonotic and arboviral disease surveillance. *J Biol Agr Healthc*. 2017;7(21):35–41.
3. Duncavage EJ, Magrini V, Becker N, et al. Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue. *J Mol Diagn*. 2011;13(3):325–333.
4. Illumina (2017). TruSeq Stranded Total RNA Reference Guide. Accessed February 13, 2020.

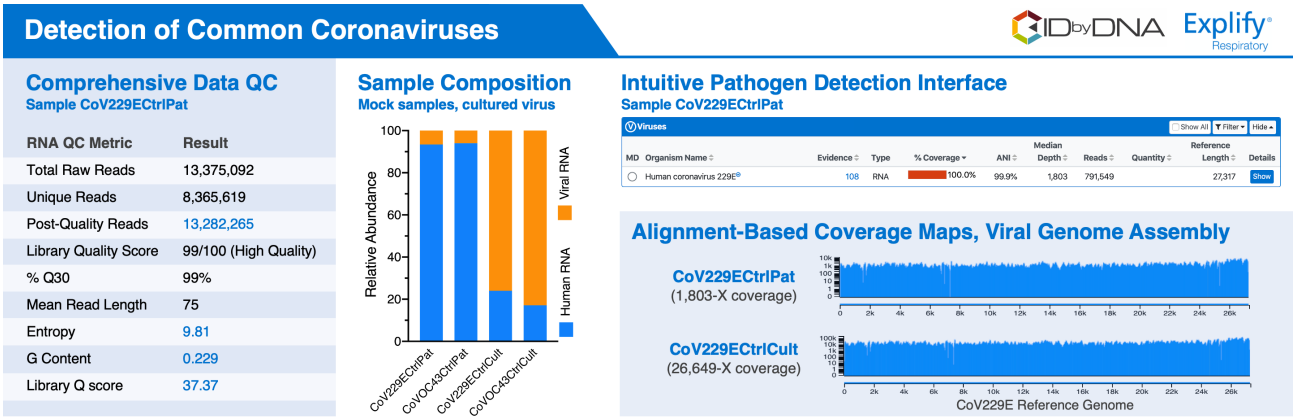

Figure 4: IDbyDNA Explify Platform—The Explify platform provides viral consensus genome sequences, coverage plots, and the ability to detect co-infections with other viruses.

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