illumina®

Culture-Free Detection and Identification of Unknown RNA Viruses

Using the MiSeq® system and TruSeq® technology to identify RNA viruses from clinical samples.

Introduction

Advances in massively parallel sequencing and informatics are redefining the sensitivity of assays designed to detect and identify pathogenic microorganisms. Conventional RT-PCR, sequencing by capillary electrophoresis, and array-hybridization approaches have several constraints. They are limited in plexity, rely on prior knowledge of DNA sequence context, and require constant refinement to cope with rapid mutation and hybridization rates. These methods are also dependent on the ability to isolate, culture, or enrich sufficient starting material. This application note outlines a culture-free approach for sequencing all RNA viruses present in a clinical sample. The method was developed at the Armed Forces Research Institute for Medical Sciences (AFRIMS) in Thailand, along with an informatics workflow that enables rapid identification of known pathogens. The workflow can be coupled with additional algorithms to identify and characterize new or unexpected virus species present in a sample. This method represents a powerful new tool for disease monitoring and potentially rapid clinical and diagnostic applications.

Methods

Library Preparation

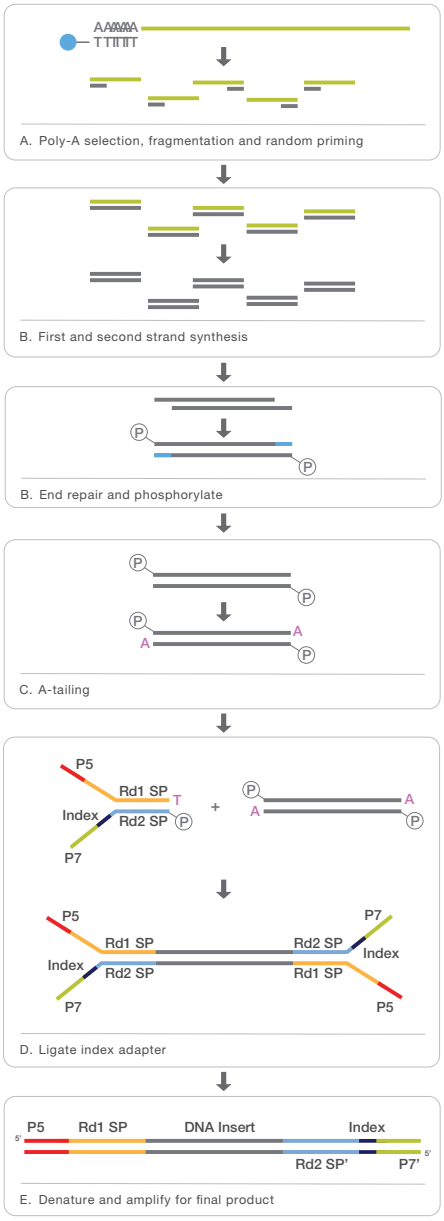
Total RNA was extracted using the QIAamp Viral RNA Mini Kit (QIAGEN) from a patient's nasopharyngeal swab and a cultured viral isolate from the same specimen[1]. The RNA was then prepared for sequencing using the TruSeq RNA v2 Library Preparation Kit (Figure 1). Indexed samples were pooled and sequenced on the MiSeq system with 2 × 150 bp paired-end reads.

Data Analysis

The generated data were analyzed in two stages (Figure 2). Primary analysis involved alignment of the sequencing reads to a customized viral database of known respiratory pathogens. Pathogen identification was inferred from the resulting frequency of aligned reads.

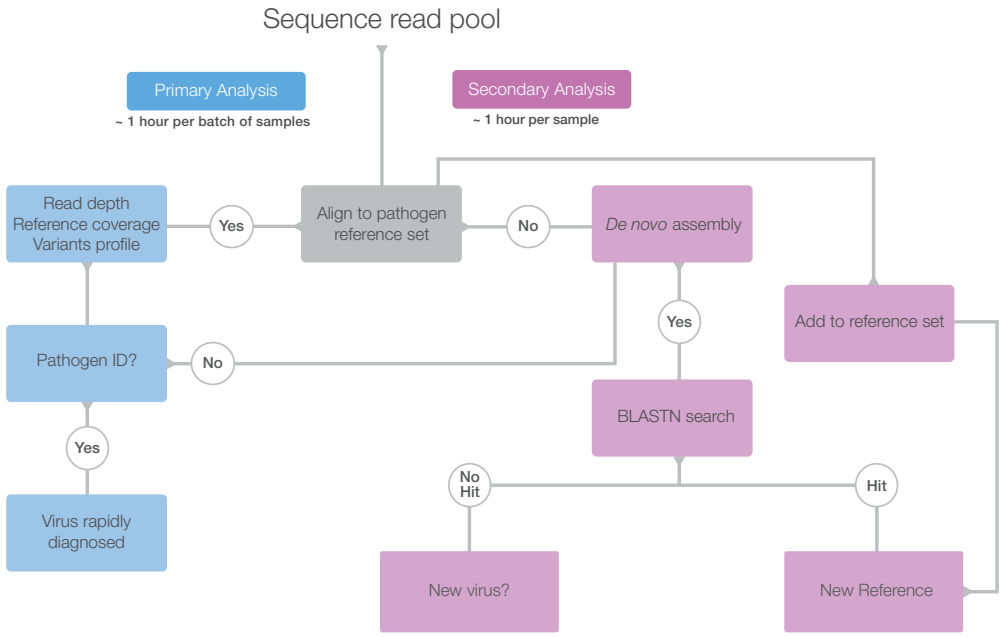
Secondary analysis included *de novo* alignment of the reads using the Trinity alignment algorithm[2] and subsequent Basic Local Alignment Tool[3] (BLASTN) analysis of resulting contigs to identify viruses not found in the database. This two-tiered approach demonstrates the rapid screening of viral sequences against a known reference database and the potential for *ex situ* identification of viruses not present in the reference database.

Figure 1: TruSeq RNA v2 Library Preparation Workflow



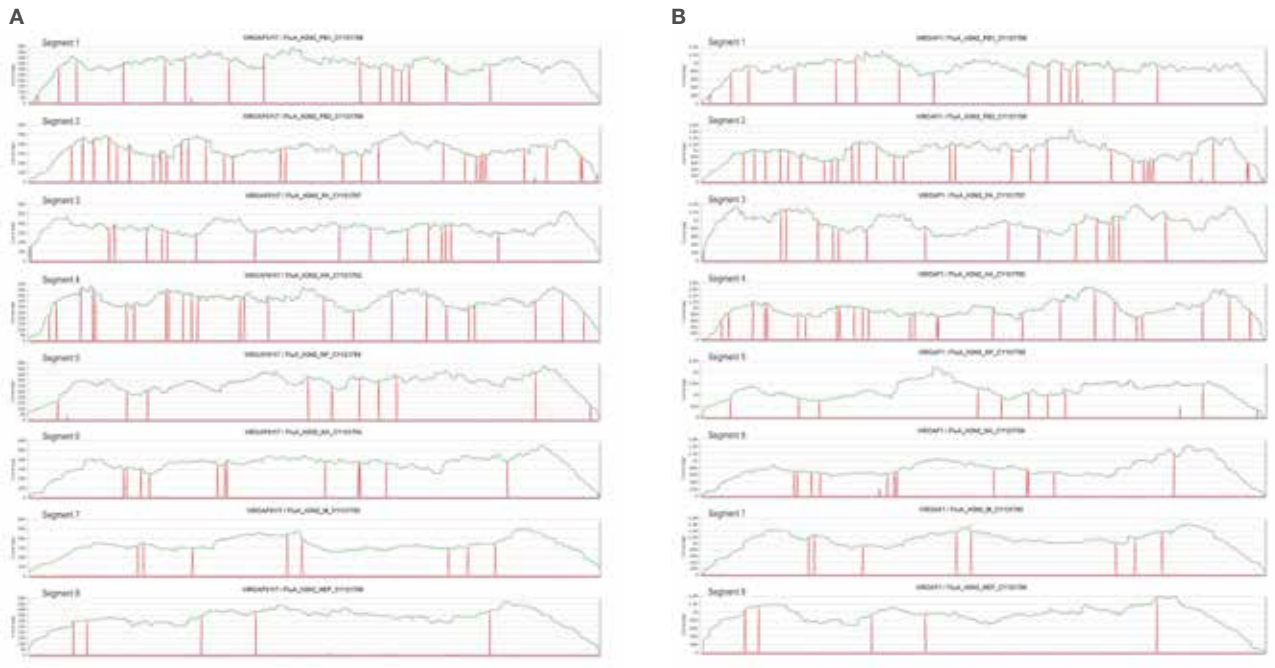
Sequencing libraries were prepared using the TruSeq RNA v2 Library Preparation Kit. Viral RNA was fragmented, reverse transcribed, and ligated to P5 and P7 adapters for sequencing with the MiSeq system.

Figure 2: Overview of Analysis Workflow



Primary analysis using the MiSeq Reporter Resequencing workflow is shown on the left. Secondary analysis involving *de novo* assembly of reads and BLASTN alignment is shown on the right.

Figure 3: Coverage Profiles for Clinical and Cultured Isolates



The reference coverage plots (green) and mismatches (red) were generated by the MiSeq Reporter Resequencing workflow. "Clinical" (A) refers to RNA purified directly from primary specimens and "cultured isolate" (B) refers to the *in vitro* viral culture from the same specimen. Coherent alignment across both samples provides strong evidence for the identification of the H3N2 virus in the specimen.

Table 1: BLASTN Search Results for *De Novo* Assembled Contigs

Specimen 1 (Clinical)

Contig	Contig Length (bp)	Read Depth	BLASTN Hit	Identity (%)	Coverage (%)
1	1,610	8,764	16S ribosomal RNA gene, partial sequence	92	100
2	2,328	5,915	Influenza A virus (H3N2), *PB1* gene	100	99
3	2,323	5,974	Influenza A virus (H3N2), *PB2* gene	100	99
4	2,237	5,710	Influenza A virus (H3N2), *PA* gene	100	99
5	1,770	4,552	Influenza A virus (H3N2), *HA* gene	99	99
6	1,549	3,819	Influenza A virus (H3N2), *NP* gene	100	99
7	1,484	3,735	Influenza A virus (H3N2), *NA* gene	100	99
8	1,011	2,449	Influenza A virus (H3N2), *M* gene	100	98
9	873	2,134	Influenza A virus (H3N2), *NEP* and *NS1* genes	100	98
10	5,386	739	*Helicobacter pylori* WGS sequence	100	92

Specimen 1 (Isolate)

Contig	Contig Length (bp)	Read Depth	BLASTN Hit	Identity (%)	Coverage (%)
1	869	5,622	Influenza A virus (H3N2), *NEP* & *NS1* genes	100	100
2	681	8	60S ribosomal protein L23a-like, mRNA	97	100
3	2,323	15,952	Influenza A virus (H3N2), *PB2* gene	100	99
4	2,404	14,361	Influenza A virus (H3N2), *PB1* gene	100	99
5	2,218	14,651	Influenza A virus (H3N2), *PA* gene	100	99
6	1,775	11,926	Influenza A virus (H3N2), *HA* gene	100	99
7	1,545	12,986	Influenza A virus (H3N2), *NP* gene	100	99
8	1,006	7,711	Influenza A virus (H3N2), *M* gene	100	98
9	1,505	7,672	Influenza A virus (H3N2), *NA* gene	100	97
10	994	106	Satellite DNA, clone D254YB6	92	97

De novo assembled contigs were searched against public reference databases. BLASTN hits were ordered by decreasing percentage length coverage against the reference genome. The 10 results with the highest coverage percentages when compared to the reference genome are shown here. The influenza viruses identified in these results are highlighted.

Results

Samples were sequenced on the MiSeq platform to generate approximately 1 million reads each. The influenza A H3N2 subtype was successfully identified through the primary alignment using MiSeq Reporter software in both clinical and isolate samples from the same specimen. Figure 3A shows the coverage profiles of the clinical sample across eight H3N2 viral segments with a minimum of 300× mean coverage of each segment. The high alignment coverage provides a good indication for the virus identity in the clinical specimen.

The same analysis was performed on the paired cultured viral isolate. Analysis took approximately 1 hour using the automated MiSeq Reporter Resequencing workflow. This time, the alignment resulted in 900× mean coverage across each reference segment (Figure 3B). The coherence in the alignment results for both clinical and cultured isolates provides strong evidence for the identification of H3N2 virus in the specimen. This information, together with the complete coverage plots obtained for all segments of the influenza H3N2 virus, demonstrates the ability of the MiSeq system to identify RNA viruses present in clinical specimens directly, without virus culture.

To characterize new or unexpected virus species present in clinical samples and cultured isolates, secondary analysis included *de novo* alignment of the reads. Because the customized viral references used during primary analysis represent a limited database of known respiratory pathogens, assembled contigs were searched using BLASTN for viral identity not found in the database. The alignment results were then filtered for hits to viral sequences. BLASTN results for these contigs revealed the closest reference in the public databases. *De novo* assembly generated contigs of all segments of H3N2 virus in both clinical and viral isolates of the specimen, as shown in Table 1. All generated viral contigs were aligned with high percentage identity and length coverage to the H3N2 viral reference segments. Reads were realigned to the contigs to assess coverage quality and read depth for each of the assembled viral segments.

High read depth for all the assembled viral contigs (Table 1) indicates a high level of confidence in the accuracy of the assembly. Concordance of high read depth and percentage length coverage of viral references demonstrates consistency with the alignment coverage plots generated during primary analysis with MiSeq Reporter software. In addition, the BLASTN search of the *de novo* assembled contigs (Table 1) did not report additional known respiratory pathogens. This result suggests high specificity and accuracy in identifying the H3N2 virus in both the clinical sample and paired viral isolate.

This two-tiered informatics approach demonstrates the feasibility of characterizing plausible identity indicative of the virus present in a clinical sample. *De novo* assembly and local BLASTN analysis were completed in 1 hour for each sample, using an Intel Xeon CPU X7350, 2.93 GHz with 16 CPU cores, and the Red Hat 4.1.2–52 operating system.

Conclusions

This method combines the proven TruSeq library preparation method with the high output of the MiSeq system to enable identification of known and unknown RNA viruses directly from clinical samples. It also provides a simple yet effective informatics workflow that uses MiSeq Reporter software for rapid identification and downstream pipelines for characterizing multiple and unknown viruses from the same samples. Further improvements to the workflow can be achieved with the Illumina Nextera® XT Library Preparation Kit[4]. With this kit, it is possible to complete the entire workflow (sample to analyzed data) in approximately 36 hours (Figure 4).

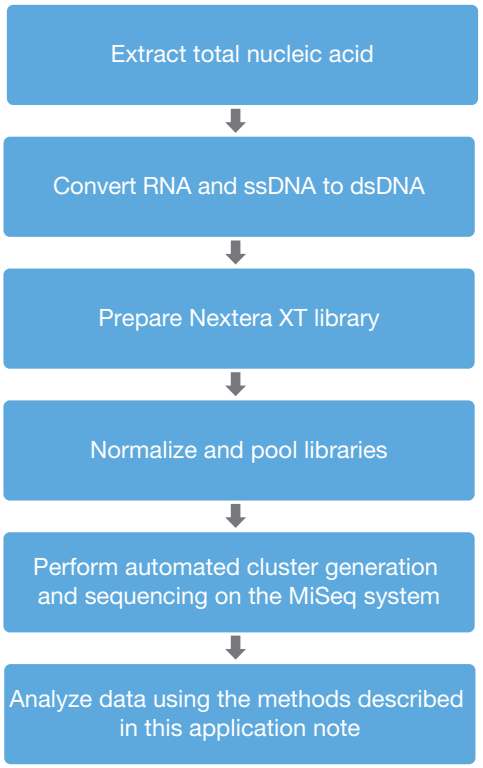
Learn More

Go to www.illumina.com/miseq to learn more about the next revolution in personal sequencing.

References

1. Rutvisuttinunt W, Chinnawirotpisan P, Simasathien S, Shrestha SK, Yoon IK, et al. Simultaneous and complete genome sequencing of influenza A and B with high coverage by Illumina MiSeq platform. J Virol Meth (2013) 193:394–404.
2. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644–652
3. Basic Local Alignment Search Tool, blast.ncbi.nlm.nih.gov
4. Nextera XT DNA Library Preparation Kit, www.illumina.com/products/nextera_xt_dna_sample_prep_kit.ilmn

Figure 4: Workflow Summary Using Nextera XT Library Preparation Kit



The use of the Nextera XT Library Preparation kit significantly reduces the hands-on time and effort required for preparation of sequencing libraries. This workflow enables quick sample turnaround when combined with automated analysis.

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

FOR RESEARCH USE ONLY

Pub. No. 1270-2013-007 Current as of 11 November 2014

illumina®