Original Article

Next-generation sequencing and bioinformatic approaches to detect and analyze influenza virus in ferrets

Zhen Lin[1,2,3], Amber Farooqui[1,2], Guishuang Li[1,2], Gane Ka-Shu Wong[3,4], Andrew L. Mason[4], David Banner[5], Alyson A. Kelvin[6], David J. Kelvin[1,2,5,7], Alberto J. León[1,2,5]

[1] *Division of Immunology, International Institute of Infection and Immunity, Shantou University Medical College, Shantou, Guangdong, China*
[2] *Guangdong Provincial Key Laboratory of Infectious Diseases and Molecular Immunopathology, Shantou, Guangdong, China*
[3] *Department of Medicine, University of Alberta, Edmonton, Alberta, Canada*
[4] *Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada*
[5] *Division of Experimental Therapeutics, University Health Network, Toronto, Ontario, Canada*
[6] *Immune Diagnostics and Research (IDR), Toronto, Ontario, Canada*
[7] *Dipartimento di Scienze Biomediche, Università di Sassari, Sassari, Italy*

Abstract

Introduction: Conventional methods used to detect and characterize influenza viruses in biological samples face multiple challenges due to the diversity of subtypes and high dissimilarity of emerging strains. Next-generation sequencing (NGS) is a powerful technique that can facilitate the detection and characterization of influenza, however, the sequencing strategy and the procedures of data analysis possess different aspects that require careful consideration.

Methodology: The RNA from the lungs of ferrets infected with influenza A/California/07/2009 was analyzed by next-generation sequencing (NGS) without using specific PCR amplification of the viral sequences. Several bioinformatic approaches were used to resolve the viral genes and detect viral quasispecies.

Results: The genomic sequences of influenza virus were characterized to a high level of detail when analyzing the short-reads with either the fast aligner Bowtie2, the general purpose aligner BLASTn or *de novo* assembly with Abyss. Moreover, when using distant viral sequences as reference, these methods were still able to resolve the viral sequences of a biological sample. Finally, direct sequencing of RNA samples did not provide sufficient coverage of the viral genome to study viral quasispecies, and, therefore, prior amplification of the viral segments by PCR would be required to perform this type of analysis.

Conclusions: the introduction of NGS for virus research allows routine full characterization of viral isolates; however, careful design of the sequencing strategy and the procedures for data analysis are still of critical importance.

Key words: ferret; influenza; next generation sequencing; deep sequencing; virus

J Infect Dev Ctries 2014; 8(4):498-509. doi:10.3855/jidc.3749

(Received 01 May 2013 – Accepted 18 June 2013)

Introduction

Influenza virus is responsible for a major burden of disease and still represents a major concern in public health [1]. Surveillance and diagnostics of influenza virus by PCR-based methods face challenges derived from influenza's high mutation rates and frequent reassortments [2]; therefore, it is not unusual for surveillance studies to show a fraction of samples which are influenza A positive but unsubtypable [3]. Microarray-based approaches to viral detection, such as ViroChip [4], constitute a good alternative for viral screening; however, the rapidly evolving nature of many viruses makes it difficult for microarrays to deliver the same level of detail as sequencing-based approaches [5]. Next-generation sequencing (NGS) allows the detection of pathogens when only little prior knowledge of their genomes is available and without the need for target-specific PCR primers. Additionally, NGS technologies deliver a large amount of genomic information that allows the study of additional aspects such as development of resistance to antivirals, variety of quasi-species and determinants of adaptation to different host species [6, 7] .

The common process behind most NGS approaches begins with random fragmentation of the

template DNA chains and binding to a solid substrate, followed with parallel PCR amplification that results in spatially separated clonal populations of DNA which can be sequenced independently [8]. Interpretation of NGS data presents bioinformatic challenges due to the large size and complexity of the sequencing data [9]. The initial approach to the analysis of NGS data can be done using three different types of tools: short-read aligners, *de novo* assemblers and general-purpose aligners. In those scenarios intended to confirm the presence, quantify or study minor sequence variations of known viruses, short-read aligners such as Bowtie [10], Burrows-Wheeler Aligner (BWA) [11] or Short Oligonucleotide Analysis Package 2 (SOAP2) [12], provide a well-established and rich framework when used in combination with other downstream analysis tools. For viral discovery studies or when a significant dissimilarity is expected between the short-read sequences and the viral reference, *de novo* assemblers such as ABySS [13], Velvet [14] or SOAPdenovo [15] can be used to generate longer sequence contigs; subsequent BLAST [16] analysis by carefully adapting the parameters to the necessities of each scenario allows the detection of viral sequences in a collection of contigs by aligning them with a database of known viral sequences. Another approach to scenarios with high sequence dissimilarity is to use the general purpose aligner BLAST to interrogate directly the short-read libraries against the viral reference sequences followed by assembly of the consensus sequence. Direct BLAST analysis of short-reads can be more sensitive than *de novo* assembly, provided that the short-reads are of sufficient length so that the analysis can "absorb" a number of indels and mismatches without causing a dramatic decrease in the similarity score.

In this study, we explore different existing paths intended to analyze the data generated by NGS in the context of viral research. Using short-read sequences generated from lung tissue of ferrets experimentally infected with influenza A/California/07/2009 (H1N1), we illustrate in detail the bioinformatic process to classify those short-reads matching influenza sequences and the subsequent generation of the consensus sequences. We simulated the characterization of an "unknown" influenza virus and explored the viral variants or quasispecies within a sample. Finally, we evaluate different options that must be considered during the design of any NGS-based strategy for viral detection, such as NGS platform and sequencing length and depths, which can cause a dramatic impact in the study results.

Methodology

Ferret virus infection and sample collection

Ferrets were experimentally infected with 1×10^6 50% egg infectious doses (EID50) of influenza A/California/07/2009 (H1N1); the animals were euthanized at different time-points post-infection and lung tissue was collected and stored in RNALater at -80°C. Detailed information about the infection procedures, clinical data and the results of the microarray and NGS analysis were previously published by our group [17]. A total of four lung tissue samples from days 1, 3, 5 and 14 post-infection, respectively, were selected for analysis by deep sequencing. The virus was not detectable in the lung sample day 14 post-infection and it was included in this study as negative control.

RNA isolation, sample preparation and deep sequencing

RNA was extracted from the lung tissue samples using the TriPure reagent (Roche, Indianapolis, IN, USA). The quality of the RNA was verified in an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, California, USA) ensuring an RNA Integrity Number (RIN) $\geq$ 8.5. cDNA library construction and deep sequencing were performed at BGI (Shenzhen, Guangdong, China) according to previously published procedures [18]. Briefly, the mRNA isolated and fragmented, double-stranded cDNA was synthesized followed by adaptor ligation; DNA fragments were selected by excising the 200±25bp band in an agarose gel electrophoresis followed by PCR for library enrichment. Paired-end 90bp sequencing was performed using an Illumina Genome Analyzer IIx sequencer. Adaptor sequences were removed and those reads with more than 10% Q<20 bases were filtered out. The resulting short-reads were uploaded to the Sequence Read Archive (accession# SRA048986) and they are publicly available [17].

In a separate analysis, RNA purified from the lung tissue of a ferret infected with influenza A /California/07/2009 (H1N1), 5 days post-infection, was submitted for sequencing in a Roche 454 GS FLX system to the Plate-forme d'Analyses Génomiques, l'Université Laval (Laval, Quebec, Canada).

Detection of influenza virus-matching reads using Bowtie and downstream analysis

Bowtie2 (v2.0.2, Linux 64 version) was downloaded (http://bowtie-bio.sourceforge.net/index.shtml) and it was executed under Linux Ubuntu desktop 11.04. Nucleotide sequences of all the viral segments of A/California/07/2009 [19] were retrieved from Genbank: FJ966976 (polymerase PB2 subunit), FJ966978 (polymerase PB1 subunit), FJ966977 (polymerase PA subunit), FJ966974 (hemagglutinin, HA), FJ969536 (nucleocapsid protein, NP), FJ984386 (neuraminidase, NA), FJ966975 (matrix proteins, MP) and FJ969528 (non-structural genes, NS). A Bowtie2 index containing the sequences of the viral segments was generated with the bowtie2-build program. The analysis of the short-reads was performed by using Bowtie in paired-end mode with the -S option to obtain the output in SAM format (http://samtools.sourceforge.net/SAM1.pdf) [20]. The SAM Tools-0.1.12 (Linux 64 version) package was downloaded (http://sourceforge.net/projects/samtools/files/samtools/0.1.12/) and used to process the sequence alignment files in SAM format sequentially using the SAM Tools commands *view*, *sort* and *pileup*. The resulting output is in pileup format and it describes the base-pair information at each position (http://samtools.sourceforge.net/pileup.shtml). Next, the consensus sequence for each viral segment was generated by running the script "*samtools.pl pileup2fq*" with a minimum coverage per-base of 3. Additionally, SAM files were imported in the assembly visualization tool Tablet v1.11.08.29 (http://bioinf.scri.ac.uk/tablet/) [21] and the number of times that each position had been covered by the aligned short-reads was determined. Finally, to study the variations present in the influenza-matching short-read sequences of each sample, the previously generated *pileup* files were analyzed with VarScan-2.2.5 software (http://varscan.sourceforge.net/) [22] by executing the program with the *pileup2snp* command.

De novo assembly with Abyss and annotation of the generated contigs

To reduce the level of complexity and the computational requirements, short-reads matching the ferret genome were subtracted from the short-read libraries; the 1,871 genomic scaffolds that comprise the ferret genome assembly MusPutFur1.0 were downloaded from GenBank (accessions GL896898-GL898768). A Bowtie index containing these genomic scaffolds was built. The short-read libraries were analyzed with Bowtie (v0.12.7) and the unaligned reads were stored on separate files. The *de novo* assembler ABySS-1.2.7 (http://www.bcgsc.ca/platform/bioinfo/software/abyss) was run on a Linux environment (Ubuntu desktop 11.04) with 12Gb of RAM, using the parameter k=32 and paired-end mode. Next, the resulting contigs were subjected to BLAST analysis to select those contigs showing high degree of similarity with the influenza sequences.

Pre-selection of influenza-matching short-reads with BLAST and assembly of the consensus sequences with Iliad Assembler

The software BLAST 2.2.25+ (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/) was installed and run locally under Windows 7. BLAST databases containing the short-read sequences of the different samples were constructed with the makeblastdb program included in the BLAST package. The nucleotide sequences of all the viral segments of A/California/07/2009 (accession numbers shown above) were used as "query" for the BLAST analysis. Different combinations of settings were tested to optimize the BLAST analysis. Iliad Assembler is a software tool developed by our group which falls in the category of guided assemblers and it relies on BLAST to perform the alignments (http://www.ferretscience.org/2012/02/iliad-assembler.html). The program generates the consensus sequence by using a reference sequence together with a set of pre-selected short-reads; additionally, it finds the correct position for the contigs even when unresolved areas are present. Given the flexibility of BLAST alignments, the program offers a good performance in situations when high dissimilarity between the reference and the reads are present (manuscript under preparation).

Results

Overview of the Illumina sequencing data output

RNA was purified from the lung tissues collected on days 1, 3, 5 and 14 post infection; for each of those time-points, one sample was subjected to NGS analysis at BGI, Shenzhen, China. The sequencing analysis produced 20 million paired-end reads per sample (totally 40 million reads per sample), 90 base-pairs long. The vast majority of the sequences correspond to the ferret mRNA and only a small fraction to influenza virus (determined by BLAST

analysis, details are shown below). An overview of the data analysis workflow is provided in Figure 1.

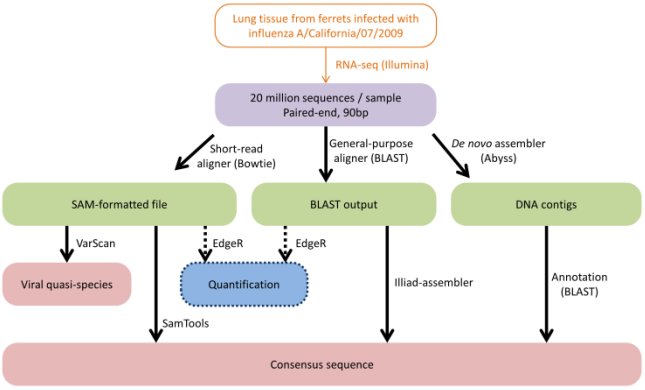
Detection of influenza virus by using the fast aligner Bowtie and SAM Tools

A Bowtie index was generated using the eight viral segments of A/California/07/2009. Later, the paired-end reads from each sample, which were contained in their respective fastq format files, were aligned with Bowtie and the resulting output files were generated in SAM format. Next, the consensus sequences for all the viral genes were generated with the *pileup* command of SAM Tools; as expected, the resulting sequences were well formed and they showed a high degree of similarity with respect to the reference sequences. The alignments were loaded in the visualization tool Tablet obtaining the number of short-read alignments for each viral gene (Table 1) and the sequencing coverage at every nucleotide position (Figure 2).

Detection of influenza virus by BLAST analysis and Iliad Assembler

BLAST was used to pre-select the short-reads that match the viral genes and the generation of the consensus sequence was performed by guided assembly with Iliad Assembler. First, we generated BLAST databases containing the reads from each sample; the short-read sequences were subsequently used as the BLAST "subject" during the analysis. Since this method does not allow the processing of paired-end reads, all the reads were treated as single-end. Next, each viral segment of A/California/07/2009 was independently used as the BLAST "query" using the following BLASTn parameters: *word_size* 12, *evalue* 1E-12, *reward* 2 and *penalty* -3; the results are shown in Table 1. The number of short-reads mapped to influenza genes and the percentage of flu-matching reads per sample were as follows: day 1: 1,811 reads (0.005%), day 3: 9,580 reads (0.024%), day 5: 22,497 reads (0.056%) and on day 14 no influenza sequences were detected. The differences in the number of reads among time-points are in accordance with the evolution of the viral loads previously described for this infection model [23]. Finally, the pre-selected short-reads were processed with Iliad Assembler to generate the consensus sequence and to calculate the coverage percentage (Table 1). BLAST analysis of the final assembly of the hemagglutinin gene (day 5 post-infection data) showed 1,694 out of 1,695 identities and zero gaps with respect to the reference sequence.

Figure 1: Next-generation sequencing data analysis for detection and characterization of viruses. The first step of the bioinformatic process can be performed by three different types of programs: short-reads aligners, general-purpose aligners and *de novo* assemblers. Afterwards, biological interpretation requires coupling with specific tools to generate the consensus sequence, quantification of viral genes and SNP calling for detection of viral quasispecies. The chart is comprised by the following elements: biological samples and sequencing (orange colour); data (coloured boxes); bioinformatic analysis (black arrows) which were performed (continuous lines) and not performed in this study but of relevance in the field (dotted lines).



Detection of influenza virus by de novo assembly

We achieved a reduction in the size of the short-read libraries of around 50% (Table 2) by subtracting those reads matching the ferret genomic DNA; this led to a significant reduction of the computing workload during the *de novo* assembly process. After performing several preliminary runs to optimize the program settings, *de novo* assembly was performed using the subtracted libraries and the ABySS option k = 32 and paired-end mode. Contigs that significantly matched influenza sequences were identified with BLAST and Iliad Assembler was used to calculate % length of the assembly with respect to each reference sequence (Table 2). Even when the number of available reads was low, the results were comparable to those obtained with Bowtie2 or direct BLAST analysis (Table 1).

Simulation of virus discovery using BLAST

We aimed to use a realistic scenario to explore the challenges involved in the characterization of new viruses when using other previously known viruses with high degrees of dissimilarity as reference for the sequence alignments. We focused this simulation on the hemagglutinin gene because this gene presents the highest degree of sequence variability among strains, and therefore, it is the most challenging gene to resolve in newly isolated viruses.

Figure 2. Sequencing coverage at every nucleotide position for the genomic segments of influenza virus. Short-reads from day 5 post-infection were aligned to the sequences from A/California/07/2009 using Bowtie2; the resulting SAM file was loaded in the visualization tool Tablet to generate coverage summaries, and later, these were plotted with Microsoft Excel.

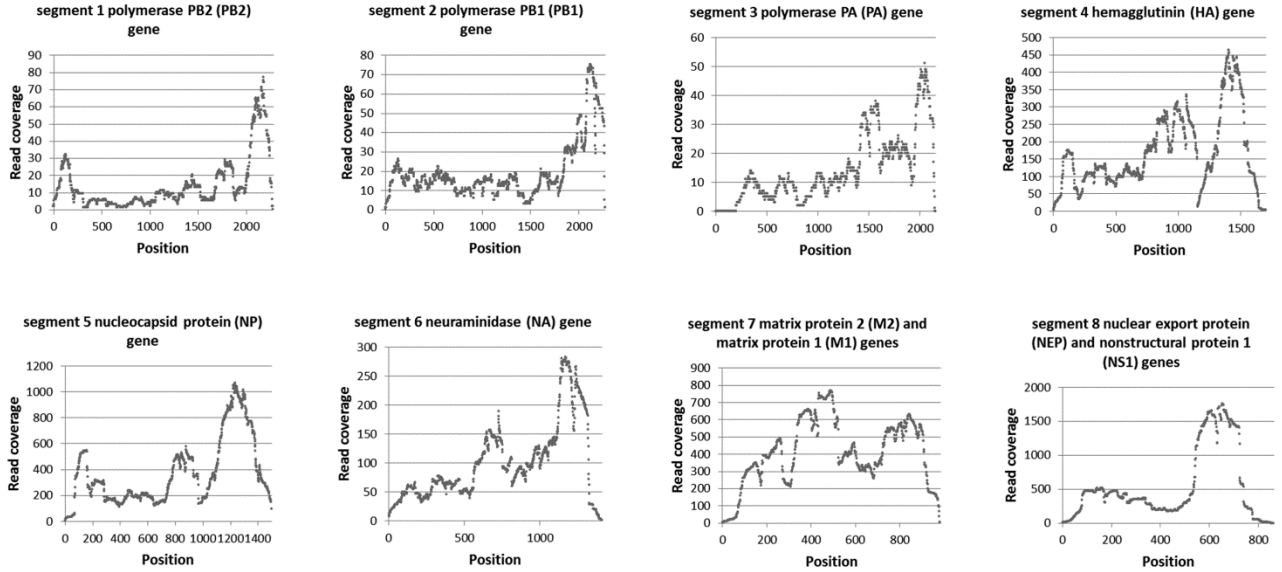


Table 1. Analysis of next-generation sequencing data with Bowtie2 and BLASTn to characterize the genomic segments of influenza virus. The table shows the number of reads that match the influenza segments and the % length of the consensus sequence with respect to each reference sequence at different times post-infection (PI).

Segment[a]	Ferret day1 PI	Ferret day3 PI	Ferret day5 PI	Ferret day14 PI
Bowtie2 and SAM Tools[b]				
PB2	50 (24.0)	227 (72.9)	350 (91.3)	0 (0)
PB1	156 (79.0)	345 (93.0)	482 (99.5)	0 (0)
PA	27 (11.1)	125 (60.9)	340 (87.2)	0 (0)
HA	257 (95.2)	1,461 (99.4)	3,154 (99.9)	0 (0)
NP	494 (96.8)	2,678 (99.7)	6,172 (100)	0 (0)
NA	119 (85.4)	742 (97.4)	1,553 (99.5)	0 (0)
MP	321 (92.3)	1,730 (97.3)	4,421 (99.8)	0 (0)
NS	334 (85.1)	1,969 (96.1)	5,092 (98.6)	0 (0)
Total influenza	1,758	9,277	21,564	0
% library	0.0043	0.0231	0.0539	0
BLASTn and Iliad Assembler				
PB2	50 (24.9)	232 (91.4)	381 (96.4)	0 (0)
PB1	157 (91.3)	386 (95.2)	505 (99.3)	0 (0)
PA	29 (17.8)	132 (76.3)	358 (90.2)	0 (0)
HA	274 (98.1)	1,508 (99.5)	3,363 (99.6)	0 (0)
NP	329 (99.6)	1,752 (99.6)	4,489 (99.1)	0 (0)
NA	121 (91.2)	759 (99.4)	1,578 (99.9)	0 (0)
MP	511 (98.2)	2756 (99.6)	6,461 (99.9)	0 (0)
NS	340 (94.4)	2055 (97.7)	5,362 (98.4)	0 (0)
Total influenza	1,811	9,580	22,497	0
% library	0.0045	0.0239	0.0562	0

[a] The nucleotide sequences of influenza A/California/07/2009 (H1N1) were used as reference. The GenBank accession numbers were as follows: FJ966976 (PB2), FJ966978 (PB1), FJ966977 (PA), FJ966974 (HA), FJ969536 (NP), FJ984386 (NA), FJ966975 (MP) and FJ969528 (NS).
[b] Only positions with a minimum coverage of 3 where considered for calculating the % length of the consensus sequence.

Table 2. Summary of *de novo* assembly with ABySS and subsequent identification of influenza-matching contigs by BLAST analysis at different times post-infection (PI).

	Ferret Day 1 PI		Ferret Day 3 PI		Ferret Day 5 PI	
Short-read sequences library size [a]						
Total short-reads	40 million		40 million		40 million	
After subtraction	19.0 million		18.2 million		20.3 million	
Number of contigs assembled by ABySS [b]						
Total (≥50bp)	442,693		433,748		398,298	
≥200bp	40,713		36,089		37,711	
≥500bp	13,344		11,846		13,755	
≥1000bp	4,412		3,797		4,802	
Influenza-matching contigs [c]	Contigs	% coverage	Contigs	% coverage	Contigs	% coverage
Segment 1 (PB2)	3	24.8	9	82.6	2	96.3
Segment 2 (PB1)	6	77.3	4	97.0	1	100.0
Segment 3 (PA)	5	23.7	5	68.3	3	90.7
Segment 4 (HA)	3	99.2	1	99.9	11	100.0
Segment 5 (NP)	1	100.0	3	99.5	21	99.6
Segment 6 (NA)	3	89.0	1	99.4	2	99.8
Segment 7 (MP)	1	92.5	5	99.0	24	99.9
Segment 8 (NS)	4	94.4	7	92.4	20	96.8

[a] Reads matching ferret sequences (MusPutFur1.0) were subtracted with Bowtie (v0.12.7).
[b] ABySS was run in paired-end mode and k=32
[c] Sequences from A/California/07/2009 (H1N1) were used as reference. % coverage was calculated with Iliad Assembler.

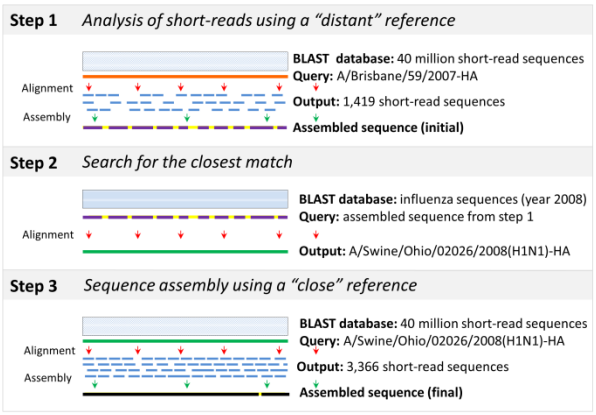


Figure 3. Simulation of virus discovery using direct BLAST analysis of NGS data. Step 1: a BLAST database containing the sequencing data from day 5 post-infection was screened by BLAST using a "distant" reference (orange line) from a virus of a lineage that was circulating at that moment (A/Brisbane/59/2007-HA); a number of short-reads were selected (short blue lines) and they were used to generate the consensus sequence (initial) by Iliad Assembler (purple-yellow dashed line). Step2: BLAST analysis of the assembled sequence using as reference influenza isolates from 2008 revealed that the closest match was a strain of swine origin, A/Swine/Ohio/02026/2008 (H1N1)-HA. Step 3: the database was searched using A/Swine/Ohio/02026/2008-HA as the "close" reference (green line) and the consensus sequence (final) was assembled (black line with a yellow dash). BLASTn settings were *word_size* 12, *reward* 2 and *penalty* -3. Yellow dashes indicate uncharacterized areas of the assembled sequences. Red arrows: BLAST alignment. Green arrows: sequence assembly.

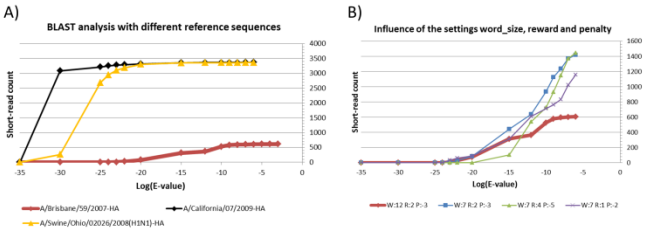


Figure 4. Factors influencing the output of BLASTn analysis when detecting sequences from influenza A/California/07/2009 in the short-read library from 5 days post-infection. (**A**) Counts of aligned short-reads obtained when using the sequences of several influenza strains as reference and different BLAST Expect value (E-value) thresholds. BLASTn settings were *word_size* 12, *reward* 2 and *penalty* -3. (**B**) Effect of different BLASTn settings in the short-read counts when using A/Brisbane/59/2007-hemagglutinin (HA) as reference.

It was assumed that our 90bp short-reads from day 5 post-infection would contain an "unknown" influenza A virus from 2009, and only hemagglutinin sequences from 2008 isolates deposited in GenBank were used as reference (Figure 3). In a preliminary stage, the sequence from hemagglutinin of A/Brisbane/59/2007 was used as reference and several BLAST analyses with different levels of stringency were performed. We found that *word_size* was the most relevant parameter when trying to detect influenza sequences with high degrees of dissimilarity; the use of an *evalue* of 10E-6 was stringent enough to discriminate influenza sequences from those belonging to the host species (Figure 4). Using the optimal BLAST parameters, *word_size* 7 and *evalue* 10E-6, 1,419 reads were pre-selected and a consensus sequence was generated with 71.7% length coverage (Figure 3). This sequence was queried against a BLAST database containing all the influenza sequences published in 2008; the closest match was A/swine/Ohio/02026/2008(H1N1)-HA (GenBank accession CY09915), showing 90.3% homology between them. Finally, BLAST analysis was performed using the sequence from A/swine/Ohio/02026/2008(H1N1)-HA as reference; after assembling the HA-matching reads, the resulting consensus sequence showed 1,698/1,701 identities with the A/California/07-HA sequence.

Detection of virus subpopulations with Varscan

To investigate the capacity of deep sequencing to detect virus subpopulation or quasispecies within a biological sample, the alignment files in SAM format previously generated by Bowtie2 were analyzed using VarScan software. The quality thresholds to discriminate allele variants from sequencing errors were empirically adjusted by using ferret mRNA beta actin as reference (data not shown). VarScan analysis was run with a minimum base quality of 50, and only those allele variants with more than two supporting reads in both plus and minus strands were considered. Our sequencing strategy was based on direct RNA sequencing without prior PCR amplification. Consequently, the vast majority of the viral genome did not have sufficient coverage to allow the detection of variants with low frequency (Figure 5). Three

Figure 5. Overview of regions of the influenza genome in which nucleotide variants can be called using the NGS data from our study at different times post-infection (PI). A sufficient number of reads from both forward and reverse complementary strands need to support the presence of a nucleotide variant; for each position, the read count from the strand with the lowest coverage was plotted. Frequency thresholds were set with a minimum requirement of more than 2 supporting reads in both the plus and minus strands.

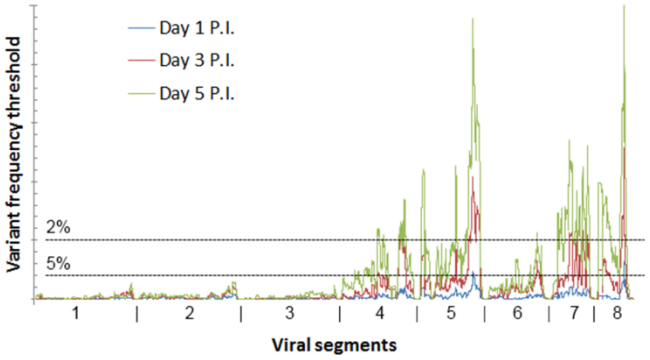


Table 3. Variants detected in the influenza sequences by VarScan analysis at different times post-infection (PI)

Viral segment	Nucleotide change	Aminoacid change	Variant Frequency	Supporting Reads Reference (Plus/Minus)	Supporting Reads Variant (Plus/Minus)	Average Quality of Reads (Reference/Variant)
Ferret day 1 PI						
NP(5)	159 T/G	D53E	100%	0 / 0	15 / 20	- / 68
NP(5)	365 T/A	L122Q	100%	0 / 0	4 / 7	- / 66
Ferret day 3 PI						
HA(4)	598 T/C	S200P	100%	0 / 0	26 / 13	- / 67
NP(5)	159 T/G	D53E	100%	0 / 0	76 / 77	- / 67
NP(5)	365 T/A	L122Q	100%	0 / 0	27 / 37	- / 65
Ferret day 5 PI						
HA(4)	598 T/C	S200P	100%	0 / 0	69 / 41	- / 68
HA(4)	1546 G/A	E516K	18.72%	59 / 93	16 / 19	66 / 67
NP(5)	159 T/G	D53E	100%	0 / 0	220 / 319	- / 67
NP(5)	365 T/A	L122Q	100%	0 / 0	92 / 61	- / 64
NS(8)	291 A/G	-	3.78%	134 / 196	6 / 7	65 / 67
NS(8)	588 A/G	-	2.97%	1260 / 212	37 / 8	66 / 67

coding variants with 100% frequency were shared between the samples from days 3 and 5 post-infection (Table 3), which indicates that these changes were introduced before the inoculation of ferrets, possibly during the viral expansion in eggs. Additionally, three more variants with low frequency were found in the RNA sample from day 5 post-infection, one of which was a coding mutation in the hemagglutinin gene, suggesting the presence of viral quasispecies.

Overview of the Roche 454 GS FLX sequencing data output

The sequencing run produced 265,484 reads of an average length of 386bp, which is in the range of a successful analysis according to the manufacturer's standards. BLAST analysis revealed that the number of sequences matching each viral segment was PB2: 0, PB1: 1, PA: 0, HA: 12, NP: 6, NA: 6, MP: 9 and NS: 16.

Discussion

After having performed different studies focused on the host immune responses during respiratory viral infections involving microarray analysis [23-25], our group decided to use RNA-seq to better characterize transcriptional variations during experimental influenza infections in ferrets. As part of this work, we also evaluated the presence of influenza virus in our samples. Although the detection of sequences matching the influenza genome can be regarded as a technically simple task, we found that a robust data analysis requires careful consideration of different aspects. Hence this paper is intended to explore the complexities of sequencing data analysis in the context of viral detection and discovery.

Some NGS studies reported the use of prior PCR amplification of the viral segments [26,27] or enrichment of viral sequences by using probe-capture methods [28]. Our results (Table 1) and also previously published studies [6,7] show that direct RNA sequencing can provide very high coverage of the viral genome; however, there can be clinical samples where the number of virus-matching reads is low [29] and pre-amplification is still an option that may need consideration. The selection of the sequencing platform will determine the number of reads that can be obtained. Roche 454 FLX GS was the first NGS platform commercially available; however, its technical capacities were surpassed shortly after by the Illumina sequencers [30]. We sequenced one RNA sample using both platforms, and although the experimental design was not intended to make direct comparisons between technologies, we were able to conclude that direct RNA sequencing in the 454 platform can resolve only the most highly expressed viral genes (as shown in results), a scenario that highly resembles the results from a previously published paper [7]; therefore, viral analysis by 454 sequencing requires prior enrichment of the viral segments by PCR amplification or sequence-specific capture. Meanwhile, Illumina sequencing obtained much higher coverage and succeeded in delivering information from all the viral segments (Table 1). Unless otherwise indicated, the results and the discussion refer to the data obtained by Illumina GAIIx sequencing.

Bowtie2 is a fast aligner widely used in different NGS applications, such as re-sequencing of mammal genomes and study of gene splicing variants [31]. Unlike other aligners such as BWA [11] and SOAP2 [12] that only search for ungapped alignments, Bowtie2 is capable gapped-read alignment, which results in an increase in the number of correct alignments. To perform successful alignments, these tools require a high degree of similarity between the reference and the experimentally obtained short-reads. Here, Bowtie2 was able to align the short-reads to the sequences from A/California/07/2009 used as reference (Table 1) and the subsequent consensus sequence was obtained through SAM Tools. The quality thresholds must be set in accordance with the degree of confidence demanded by each application. For example, when obtaining the consensus sequence from the PB2 segment in day 1 post-infection, we found that when using different minimum depths of 3, 5 or 8, the resulting percentage coverage was 23.9, 17.6 and 7.9, respectively. It should be noted that the way in which quality thresholds are implemented varies among methods of analysis; therefore, the percentage coverage alone is not suitable to establish direct performance comparisons. As expected, our simulation shows that the length of the reads and the number of reads mapped to a certain gene have a dramatic impact in the coverage (Table 4).

Direct BLAST analysis of short-reads is one of the key approaches that should be considered when little or no sequence information is available, or when a significant degree of dissimilarity is expected between a new virus and the previously known strains.

Table 4. The resulting % coverage varies with the sequencing platform, length and depth

NGS platform[a]	Length of reads (bp)	Reads matching California/07-HA	% coverage
Illumina Genome Analyzer IIx[b]	90	1,000	95.8
		500	95.1
		250	92.7
		125	78.7
	60	1,000	94.6
		500	93.7
		250	83.6
		125	65.1
	30	1,000	94.6
		500	77.0
		250	60.1
		125	36.2
Roche 454 GS FLX [c]	287 (average)	13	67.3

[a] Total RNA from the lung tissue of one ferret infected with A/California/07/2009, 5 days post-infection, was analyzed using two different sequencing platforms.
[b] The sequencing run generated $20x10^6$ paired-end reads, 90bp. In order to study variations in the % length coverage, a number influenza-matching sequences were randomly selected and trimmed to the desired length. Alignments were performed with Bowtie and the % coverage was calculated with Samtools.
[c] The sequencing run using the Roche GS FLX 454 platform generated 265,454 reads with an average length of 386bp. Out of these, 13 reads matched the sequence from California/07-HA, showing and average length of 287bp. The % coverage was calculated with Iliad Assembler.

Table 5. Overview of the analysis techniques used in this study and their performance

Type of analysis	Performance
Illumina GAIIx sequecing	
Sequencing of RNA without prior amplification	Good for viral detection Good for characterization of viral segments
Fast-aligner Bowtie2 + Samtools General purpose aligner BLAST + Iliad Assembler *De novo* assembler Abyss	Near complete characterization of viral segments in samples with high viral loads (ferret day 5 post-infection).
SNP calling with VarScan	Insufficient coverage, prior enrichment of viral sequences is required
Roche 454 GS FLX sequencing	
Sequencing of RNA without prior amplification	Sufficient coverage for viral detection Insufficient coverage for characterization of viral segments

Because of the sustained increase in the read length that upcoming NGS platforms can deliver, direct BLAST analysis of short-reads will probably be embraced more widely. Influenza genes have a low degree of homology with respect to the genes of mammal hosts, making their identification easy within libraries where the majority of the sequences correspond to the host species. On the other hand, the great flexibility that BLAST offers through careful selection of the alignment parameters makes it a tool of great value in a variety of studies. After pre-selecting the reads that match the reference genome, they must be assembled to generate the consensus sequence; we used Iliad Assembler to perform this task, a flexible tool written by our group that is well suited for the assembly of complex transcripts (manuscript under preparation). Nonetheless, other tools can be used to perform the assembly of the pre-selected reads such as SSAKE (Short Sequence Assembly by K-mer search and 3' read Extension) [32]; alternatively, this task can be performed by de novo assemblers.

De novo assembler programs are designed to find overlaps in the short-reads to generate longer contig sequences; these need to be later identified by using a general-purpose aligner. This approach has been previously used to resolve genomic sequences of influenza virus [6]; for example, Greninger *et al.* reported that *de novo* assembled contigs had 90.3% coverage using 60bp short-reads [5], which is in accordance with our results (Table 2). Given the short length of the influenza genome and the structural simplicity of their genes, as compared with most mammal genes, the number of reads covering the target sequence is the most important factor for the success of this approach rather than the election of the

assembler. Next, we tried to simulate the conditions of the analysis that occur during outbreaks of new influenza strains in which only "distant" viral sequences are available. We found that direct BLAST analysis of the short-reads is a viable option (Figure 3); when using the sequence from A/Brisbane/59/2007-HA as reference, we were able to obtain 72% coverage of the "new" virus, and 99% coverage was obtained when using A/Swine/Ohio/02026/2008 as the intermediate reference. On the other hand, when BLAST is used for either direct short-read analysis or identification of *de novo* assembled contigs, the selection of the database of reference sequences is of critical importance; the analysis needs to biologically rich while keeping the computing requirements at reasonable levels.

The sequencing data allowed us to obtain the consensus sequences of all the viral genes, and they were almost identical to the previously published sequences of A/California/07/2009. For example, the consensus sequence of hemagglutinin showed only one mismatch with respect to the reference sequence, and the fact that this variation was present in all three virus-containing samples suggests that the introduction of this mutation possibly occurred during viral expansion in embryonated eggs prior to ferret infection. Also, we found that the number of reads matching influenza sequences increased gradually from day 1 to day 5 after infection, and none was found on day 14 (Table 1). This trend correlates well with the viral titers that were previously observed in the lung tissue from which those samples were retrieved [23]. However, given the lack of biological replicates, we were unable to obtain any statistically significant conclusions regarding differences in the quantity of virus among the different experimental groups. The methods for estimating differential expression levels using NGS data are still an active area of research. RPKM-based methods [33] are widely used to determine differential gene expression; however, they rely on information from both the transcripts and the genomic DNA to make certain statistical assumptions and therefore they may not be well suited to study levels of virus expression. Other methods such as edgeR [34] and DEGSeq [35] rely on only the number of short-reads per transcript; therefore, they can be used to study the relative expression of viral genes.

SNP calling is a valuable tool that can help to track the changes that viral segments undergo during different adaptation processes [2]. To characterize the variants or quasispecies of influenza virus that are present in the lungs of infected ferrets the sequencing data was analyzed with VarScan [22] (Table 3). Unfortunately, we found that direct sequencing of RNA samples does not provide sufficient coverage of the viral genome to study these sub-populations (Figure 5); therefore, the analysis of viral quasispecies requires preliminary amplification of the viral segments by PCR and subsequent deep sequencing.

In conclusion, the combination of NGS technology with an adequate strategy of data analysis (Table 5) constitutes a major leap forward in surveillance and diagnostics of influenza virus. The increased capacity in the acquisition of sequencing data means that nearly full characterization of the viral genomes can now be performed routinely. Also, increased coverage allows influenza to be approached as populations rather than just isolates, which will boost the characterization of determinants of pathogenicity and drug resistance.

Acknowledgements

This study was supported by grants from the Li Ka Shing Foundation, the Guangdong Provincial Key Laboratory of Infectious Diseases and Molecular Immunopathology, the Canadian Institute of Health Research, and Immune Diagnostics & Research. We would like to thank Thomas Rowe for technical assistance. We would like to thank Ted Ross for valuable insight into performing the ferret studies. We would like to thank Longsi Ran and Luoling Xu for processing the samples and BGI-Shenzhen for the construction of the cDNA libraries and sequencing. Z.L. is enrolled in the Li Ka Shing Sino-Canadian Exchange Program.

References

1. Suk JE, Semenza JC (2011) Future infectious disease threats to Europe. Am J Public Health 101: 2068-79.
2. McHardy AC, Adams B (2009) The role of genomics in tracking the evolution of influenza A virus. PLoS Pathog 5: e1000566.
3. Farooqui A, Lei Y, Wang P, Huang J, Lin J, Li G, Leon AJ, Zhao Z, Kelvin DJ (2011) Genetic and clinical assessment of 2009 pandemic influenza in southern China. J Infect Dev Ctries 5: 700-10. doi:10.3855/jidc.2251
4. Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, Ganem D, DeRisi JL (2002) Microarray-based detection and genotyping of viral pathogens. Proc Natl Acad Sci U S A 99: 15687-92.
5. Greninger AL, Chen EC, Sittler T, Scheinerman A, Roubinian N, Yu G, Kim E, Pillai DR, Guyard C, Mazzulli T, Isa P, Arias CF, Hackett J, Schochetman G, Miller S, Tang P, Chiu CY (2010) A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. PLoS One 5: e13381.

6. Kuroda M, Katano H, Nakajima N, Tobiume M, Ainai A, Sekizuka T, Hasegawa H, Tashiro M, Sasaki Y, Arakawa Y, Hata S, Watanabe M, Sata T (2010) Characterization of quasispecies of pandemic 2009 influenza A virus (A/H1N1/2009) by de novo sequencing using a next-generation DNA sequencer. PLoS One 5: e10256.
7. Ghedin E, Laplante J, DePasse J, Wentworth DE, Santos RP, Lepow ML, Porter J, Stellrecht K, Lin X, Operario D, Griesemer S, Fitch A, Halpin RA, Stockwell TB, D.J. Spiro, E.C. Holmes, and K. St George (2011) Deep sequencing reveals mixed infection with 2009 pandemic influenza A (H1N1) virus strains and the emergence of oseltamivir resistance. J Infect Dis 203: 168-74.
8. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11: 31-46.
9. Nowrousian M (2010) Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. Eukaryot Cell 9: 1300-10.
10. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.
11. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-95.
12. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966-7.
13. Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, irst M, Schein JE, Horsman DE, Connors JM, Gascoyne RD, Marra MA, Jones SJ (2009) De novo transcriptome assembly with ABySS. Bioinformatics 25: 2872-7.
14. Zerbino DR (2010) Using the Velvet de novo assembler for short-read sequencing technologies. Curr Protoc Bioinformatics Chapter 11: Unit 11 5.
15. Li Y, Hu Y, Bolund L, Wang J (2010) State of the art de novo assembly of human genomes from massively parallel sequencing data. Hum Genomics 4: 271-7.
16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-10.
17. Leon AJ, Banner D, Xu L, Ran L, Peng Z, Yi K, Chen C, Xu F, Huang J, Zhao Z, Lin Z, Huang SH, Fang Y, Kelvin AA, Ross TM, Farooqui A, Kelvin DJ (2013) Sequencing, annotation, and characterization of the influenza ferret infectome. J Virol 87: 1957-66.
18. Wang Z, Fang B, Chen J, Zhang X, Luo Z, Huang L, Chen X, Li Y (2010) De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweet potato (Ipomoea batatas). BMC Genomics 11: 726.
19. Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A, Sessions WM, Xu X, Skepner E, Deyde V, Okomo-Adhiambo M, Gubareva L, Barnes J, Smith CB, Emery SL, Hillman MJ, Rivailler P, Smagala J, de Graaf M, Burke DF, Fouchier RA, Pappas C, Alpuche-Aranda CM, Lopez-Gatell H, Olivera H, Lopez I, Myers CA, Faix D, Blair PJ, Yu C, Keene KM, Dotson PD Jr, Boxrud D, Sambol AR, Abid SH, St George K, Bannerman T, Moore AL, Stringer DJ, Blevins P, Demmler-Harrison GJ, Ginsberg M, Kriner P, Waterman S, Smole S, Guevara HF, Belongia EA, Clark PA, Beatrice ST, Donis R, Katz J, Finelli L, Bridges CB, Shaw M, Jernigan DB, Uyeki TM, Smith DJ, Klimov AI, Cox NJ (2009) Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. Science 325: 197-201.
20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-9.
21. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D (2010) Tablet-next generation sequence assembly visualization. Bioinformatics 26: 401-2.
22. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22: 568-76.
23. Rowe T, Leon AJ, Crevar CJ, Carter DM, Xu L, Ran L, Fang Y, C.M. Cameron, M.J. Cameron, D. Banner, Ng DC, Ran R, Weirback HK, Wiley CA, Kelvin DJ, Ross TM (2010) Modeling host responses in ferrets during A/California/07/2009 influenza infection. Virology 401: 257-65.
24. Danesh A, Cameron CM, Leon AJ, Ran L, Xu L, Fang Y, Kelvin AA, Rowe T, Chen H, Guan Y, Jonsson CB, Cameron MJ, Kelvin DJ (2011) Early gene expression events in ferrets in response to SARS coronavirus infection versus direct interferon-alpha2b stimulation. Virology 409: 102-12.
25. Cameron CM, Cameron MJ, Bermejo-Martin JF, Ran L, Xu L, Turner PV, Ran R, Danesh A, Fang Y, Chan PK, Mytle N, Sullivan TJ, Collins TL, Johnson MG, Medina JC, Rowe T, Kelvin DJ (2008) Gene expression analysis of host innate immune responses during Lethal H5N1 infection in ferrets. J Virol 82: 11308-17.
26. Hoper D, Hoffmann B, Beer M (2011) A comprehensive deep sequencing strategy for full-length genomes of influenza A. PLoS One 6: e19075.
27. Nakamura S, Yang CS, Sakon N, Ueda M, Tougan T, Yamashita A, Goto N, Takahashi K, Yasunaga T, Ikuta K, Mizutani T, Okamoto Y, Tagami M, Morita R, N Maeda, Kawai J, Hayashizaki Y, Nagai Y, Horii T, Iida T, Nakaya T (2009) Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. PLoS One 4: e4219.
28. Ramakrishnan MA, Tu ZJ, Singh S, Chockalingam AK, Gramer MR, Wang P, Goyal SM, Yang M, Halvorson DA, Sreevatsan S (2009) The feasibility of using high resolution genome sequencing of influenza A viruses to detect mixed infections and quasispecies. PLoS One 4: e7105.
29. Yongfeng H, Fan Y, Jie D, Jian Y, Ting Z, Lilian S, Jin Q (2011) Direct pathogen detection from swab samples using a new high-throughput sequencing technology. Clin Microbiol Infect 17: 241-4.
30. Cheval J, Sauvage V, Frangeul L, Dacheux L, Guigon G, Dumey N, Pariente K, Rousseaux C, Dorange F, Berthet N, Brisse S, Moszer I, Bourhy H, Manuguerra CJ, Lecuit M, Burguiere A, Caro V, Eloit M (2011) Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. J Clin Microbiol 49: 3268-75.
31. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-9.
32. Warren RL, Sutton GG, Jones SJ, Holt RA (2007) Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23: 500-1.

33. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621-8.
34. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139-40.
35. Wang L, Feng Z, Wang X, Zhang X (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26: 136-8.

Corresponding author
David J. Kelvin
Division of Immunology, International Institute of Infection and Immunity,
Shantou University Medical College,
22 Xinling Road, Shantou,
Guangdong 515041,
People's Republic of China.
Phone and fax: (86)-754-88573991.
E-mail: dkelvin@jidc.org

Conflict of interests: No conflict of interests is declared.