Arch Virol (2017) 162:2003–2012
DOI 10.1007/s00705-017-3367-4

ORIGINAL ARTICLE

A metagenomics study for the identification of respiratory viruses in mixed clinical specimens: an application of the iterative mapping approach

Yu-Nong Gong[1,2] · Shu-Li Yang[1,3] · Guang-Wu Chen[1,2,4] · Yu-Wen Chen[5] · Yhu-Chering Huang[5,6] · Hsiao-Chen Ning[1,3] · Kuo-Chien Tsao[1,2,3]

Received: 25 November 2016 / Accepted: 9 March 2017 / Published online: 19 April 2017

Abstract Metagenomic approaches to detect viral genomes and variants in clinical samples have various challenges, including low viral titers and bacterial and human genome contamination. To address these limitations, we examined a next-generation sequencing (NGS) and iterative mapping approach for virus detection in clinical samples. We analyzed 40 clinical specimens from hospitalized children diagnosed with acute bronchiolitis, croup, or respiratory tract infections in which virus identification by viral culture or polymerase chain reaction (PCR) was unsuccessful. For our NGS data analysis pipeline, clinical samples were pooled into two NGS groups to reduce sequencing costs, and the depth and coverage of assembled contigs were effectively increased using an iterative mapping approach. PCR was individually performed for each specimen according to the NGS-predicted viral type. We successfully detected previously unidentified respiratory viruses in 26 of 40 specimens using our proposed NGS pipeline. Two dominant populations within the detected viruses were human rhinoviruses (HRVs; n = 14) and human coronavirus NL63 (n = 8), followed by human parainfluenza virus (HPIV), human parechovirus, influenza A virus, respiratory syncytial virus (RSV), and human metapneumovirus. This is the first study reporting the complete genome sequences of HRV-A101, HRV-C3, HPIV-4a, and RSV, as well as an analysis of their genetic variants, in Taiwan. These results demonstrate that this NGS pipeline allows to detect viruses which were not identified by routine diagnostic assays, directly from clinical samples.

Electronic supplementary material The online version of this article (doi:10.1007/s00705-017-3367-4) contains supplementary material, which is available to authorized users.

✉ Kuo-Chien Tsao
kctsao@cgmh.org.tw

[1] Department of Laboratory Medicine, Linkou Chang Gung Memorial Hospital, Taoyuan, Taiwan

[2] Research Center for Emerging Viral Infections, College of Medicine, Chang Gung University, Taoyuan, Taiwan

[3] Department of Medical Biotechnology and Laboratory Science, College of Medicine, Chang Gung University, Taoyuan, Taiwan

[4] Department of Computer Science and Information Engineering, School of Electrical and Computer Engineering, College of Engineering, Chang Gung University, Taoyuan, Taiwan

[5] Department of Pediatrics, Linkou Chang Gung Memorial Hospital, Taoyuan, Taiwan

[6] College of Medicine, Chang Gung University, Taoyuan, Taiwan

Introduction

Virus identification using traditional polymerase chain reaction (PCR)-based methods is challenging [1, 2], particularly when viral loads are low. Genetic diversity of viruses could also lead to mismatches between probes and primer sequences, resulting in incorrect PCR results [3]. To address these challenges, unbiased next-generation sequencing (NGS) techniques have been developed to improve viral discovery. These methods have been shown to be effective for the identification and genomic characterization of influenza A viruses (IAVs) [4], hepatitis C virus [5], and other respiratory viruses [6]. Additionally, not all viruses can be cultured using common cell lines, e.g. human rhinovirus (HRV) type C [7], human bocavirus (HBoV) [8], and the human

coronaviruses (HCoV) NL63 and HKU1 [3, 9]. Therefore, NGS techniques allow to successfully detect viruses which can be difficult to culture and may lead to erroneous PCR results.

In metagenomics, modern genomic techniques are applied to characterize communities of microbial organisms directly from their natural environments, without the isolation and cultivation of individual species [10]. NGS technology has been applied to metagenomes to detect the presence of viral pathogens from single non-cultured specimens [11], including influenza virus identification and whole-genome sequencing from swab specimens [12–14] and respiratory virus identification from nasopharyngeal aspirate specimens [15]. However, the direct recovery of viral genomes from clinical specimens using NGS methods has challenges, including noise from host or microbiota cells and the limited viral RNA quantities [16]. To separate viruses from host or background flora, a novel enrichment technique called NetoVir has been developed for high-throughput sample preparation [17]. Notably, viral etiologies were unknown in ~25% of 120 clinical specimens collected from children with bronchiolitis [18], ~45% of 336 hospitalized children (<5 years old) with lower respiratory tract infections (RTIs) [19], and ~60% of 2259 clinical specimens from community acquired pneumonia patients with undetected viral pathogens [20]. Although NGS pipelines [21, 22] have been proposed to identify previously unidentified viruses from clinical specimens, sequencing costs remain an issue. In our previous study, to reduce sequencing costs, we applied NGS to pooled samples [6]. An NGS data analysis pipeline was proposed for the detection of unidentified viruses, including human parechovirus (HPeV), using an iterative mapping approach to obtain genome sequences.

In this study, clinical specimens suspected of harboring viruses were collected from hospitalized patients in Taiwan. Our NGS data analysis method was applied to identify unknown viral pathogens directly from these clinical specimens and to obtain their genome sequences by iterative mapping. Furthermore, we explored genetic variation within the samples to clarify the molecular evolution of the observed Taiwanese strains.

Materials and methods

Ethics statement

This study was approved by the Institutional Review Board of Chang Gung Medical Foundation, Linkou Medical Center, Taoyuan, Taiwan (approval no. 100-4378B).

Specimen collection, pretreatment, and RNA extraction

Children (592) hospitalized between 2008 and 2010 with respiratory symptoms were recruited from Chang Gung Memorial Hospital, Taiwan. Among the patients, 105, 124, and 363 were diagnosed with croup, acute bronchiolitis, and RTIs, respectively. Specimens from these patients were examined by routine culture for further viral/bacterial identification and by real-time PCR for viral identification. Negative results were obtained for 20 croup, 29 acute bronchiolitis, and 138 RTIs cases according to routine culture and real-time PCR. Thus, these patients were suspected to have infections caused by unknown/untested viruses. Forty specimens were randomly collected from these samples for further NGS analyses, including ten, nine, and 21 samples from patients diagnosed with croup, acute bronchiolitis, and RTIs, respectively. Among these, 34, five, and one clinical specimens were collected via throat swabs, nasopharyngeal swabs, and sputum. Specimens (0.5 mL each) were divided into two groups; for each group, 20 samples were pooled in a single tube, and the mixed specimens were filtered through 0.22-μm filters to improve analytical sensitivity and reduce background contamination prior to subsequent analyses. For viral particle enrichment, filtered specimens underwent ultracentrifugation using a SW41/Ti rotor at 28,800 × *g*, 4 °C overnight. Viral RNA extraction from 140 μL of the specimens was performed using the QIAamp Viral RNA Mini Kit (Qiagen, Hilden, Germany). To extract nucleic acids, carrier RNA was replaced using linear acrylamide (Life Technologies, Carlsbad, CA, USA) as a precipitation reagent, included in the QIAamp Viral RNA Mini Kit, according to previously published methods [11]. Extracted RNA was eluted using 30 μL of elution buffer and stored at −70 °C.

NGS data analysis pipeline

Forty specimens were pooled into two NGS groups and cDNA as well as libraries of mixed samples were synthesized using Ovation RNA-Seq System V2 and Ovation Ultralow Library Systems (NuGEN, San Carlos, CA, USA), according to previously published methods [11]. The Illumina MiSeq system was used to obtain paired-end reads (2 × 250 bp). Approximately 4 GB of raw data were generated for each of NGS 1 and 2 (an average throughput of 0.2 GB for each sample), providing 5,996,746 and 8,140,721 paired-end reads, respectively. All of raw reads were deposited to the Sequence Read Archive (SRA) of NCBI with accession number SRP100814. The NGS data analysis pipeline [6] was used to identify unknown pathogens, assemble whole gen-

omes, and investigate viral diversity from mixed clinical samples. Briefly, the pipeline consisted of the following steps: 1) the Illumina system was used to collect data from clinical samples; 2) data were preprocessed to generate NGS reads and determine homologs for the assembled contigs using BLASTN (default setting with an E-value threshold of 10 and word size of 20) and then BLASTX (default settings with an E-value threshold of 10); 3) PCR validation was performed; 4) iterative mapping was performed to obtain viral genomes; 5) reference mapping was performed to identify genetic variants. Six genomes obtained in this study were deposited in GenBank, including HPeV-1 (accession number KY460513), human parainfluenza virus 4a (HPIV-4a, KY460514), HRV-A101 (KY460515), and HRV-C3 (KY460516) detected in NGS 1, and HPIV-4a (KY460517) and RSV (KY460519) detected in NGS 2. To investigate the read depth and detect viral genomic variants, reference mapping was performed with genomic templates using Bowtie2 version 2.2.5 [23] with default settings. Position-specific read counts and nucleotide compositions obtained after read mapping were examined. A genetic variant was identified when mutant nucleotides were greater than 25% of the total reads and read depth was at least 20 [24, 25]. In addition, viral genomes of different genotypes from the same species usually share high sequence homology. Consequently, one NGS read might redundantly map to different templates if each of them was mapped independently. To avoid multiple mappings of single reads in pooled samples, individual templates from different viral genotypes (e.g. three HRV genomes G1, G2, and G3) were concatenated into a longer template (G1 + G2 + G3) which served as an initial template in our iterative mapping approach.

PCR validation of clinical specimens

Each of the 40 clinical specimens was tested by PCR to confirm the results of the NGS data analysis. Virus-specific primers and probes for enterovirus (EV), HCoV-229E, -NL63 and -OC43, human metapneumovirus type 2 (hMPV-2), HPIV-3 and -4, HPeV, HRV, IAV, RSV, and rotavirus (Rota) were used for viral RNA amplification and detection, as listed in Supplementary File 1. HRV and enterovirus (EV) were genotyped using BLASTN searches against the NCBI nucleotide database.

Results

Detection of potential viral types in metagenomes

Forty clinical specimens were collected and pooled into two NGS groups; 21,333 and 39,458 contigs (defined as a set of overlapping reads, representing a consensus sequence of a partial or whole genome) were assembled from NGS experiments 1 and 2, respectively. These contigs were primarily catalogued as viruses, bacteria, and eukaryotes based on BLASTN results. Viral contigs identified by BLASTN are summarized in Supplementary File 2, and Figure 1 shows the reads classification according to the BLASTN results for the assembled contigs. The dominant populations in NGS 1 and 2 were viral (86%) and bacterial (69%) reads, respectively. Table 1 shows 16 potential viral types identified by BLASTN, including: HCoV-NL63, HPeV-1, HPIV-4a, HRV-A101, -C3, -C4, and -C40, and IAV from NGS 1, as well as HCoV-NL63, HPeV-1, HRV-B92, -C26, and -C40, HPIV-4a, IAV, and RSV from NGS 2. We then calculated the average depths and genome coverages (%) for the mapped contigs. As shown in Table 1, in NGS 1, the genomic coverage ranged from 48.4% (from HPIV-4a) to 100% [HCoV-NL63, HRV-C40, and the polymerase acidic (PA), neuraminidase (NA), and matrix protein (MP) genes of IAV] and the contig depths ranged from 6.4 (the NA gene of IAV) to 690.0 (HRV-C40). In NGS 2, the genomic coverage ranged from 4.0% (HPeV-1) to 64.3% (HRV-C40) and the contig depth ranged from 2.2 (HPeV-1) to 167.0 (HPIV-4a). Contig coverages presented in Table 1 were estimated as the covered region of contig(s) divided by the length of the coding sequence. Contig depths were calculated as the average number of times each base in the contig was sequenced from different reads. Only 255 and 552 contigs in NGS 1 and 2, respectively, did not show any matches with database sequences; these were further examined using BLASTX searches. Among them, 153 NGS 1 and 380 NGS 2 contigs matched to sequences in the database, and only two NGS 1 and four NGS 2 contigs belonged to the virus group (unclassified RNA viruses or bacteriophages, but not respiratory viruses), as shown in Supplementary File 2.

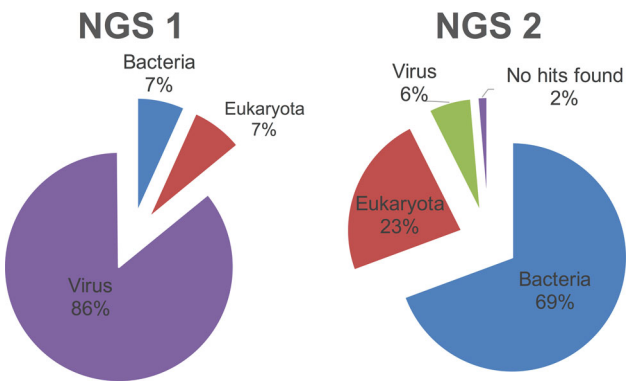

Fig. 1 Read classification by BLASTN searches

Table 1 Genomic depths and coverages of viral contigs obtained by NGS using an iterative mapping approach. Sixteen viruses were identified in two NGS pools using the proposed NGS pipeline. Genomic depths of assembled contigs for each virus type were averaged. Genomic coverages were estimated as the covered regions of contig(s) divided by the length of the coding sequence (CDS). Covered regions of assembled contigs were defined as the sum of aligned regions of contigs in BLASTN searches. The CDS lengths for HCoV-NL63, HPIV-4a, and RSV were approximately 27,000, 14,450, and 13,600 bp, respectively, and for HRV and HPeV they were approximately 6450 bp. The CDS lengths for the PB2, PB1, PA, HA, NP, NA, MP, and NS segments were approximately 2280, 2274, 2151, 1701, 1497, 692, 409, and 614 bp, respectively. In iterative mapping, genomics depths were examined by the average of position-specific read counts, and genomic coverages were defined as the sum of aligned regions of read mapping

NGS pool	*De novo* assembly				Iterative mapping		
	NGS-predicted candidate	Contig count	Genomic coverage (%)	Average depth	Iterations	Genomic coverage (%)	Average depth
1	HCoV-NL63	123	100	570.0	1	100	7001.1
	HPeV-1	3	97.0	376.9	3	100	644.0
	HPIV-4a	17	48.4	17.1	6	99.8	65.7
	HRV-A101	5	95.0	46.2	3	100	81.1
	HRV-C3	3	96.0	174.0	3	99.9	275.2
	HRV-C4	9	56.3	10.3	3	95.2	16.9
	HRV-C40	1	100	690.0	3	100	743.7
	IAV*	19	66.3, 69.1, 100, 99.5, 87.3, 100, 100, and 85.8	5.8, 21.1, 13.6, 11.6, 16.4, 6.4, 6.5, and 13.7	4	95.1, 77.5, 100, 98.9, 88.2, 84.7, 100, and 86.4	11.4, 19.3, 18.3, 18.5, 27.5, 7.1, 8.3, and 10.0
2	HCoV-NL63	59	22.0	149.0	1	100	773.8
	HPeV-1	1	4.0	2.2	2	19.4	1.6
	HPIV-4a	22	44.8	167.0	2	100	341.6
	HRV-B92	4	64.3	23.5	5	83.1	42.0
	HRV-C6	5	43.0	29.5	5	42.1	34.6
	HRV-C40	6	44.3	28.5	5	85.0	38.1
	IAV	1	17.4 (PA)	6.1 (PA)	1	3.5 (PB1), 37.2 (PA)	1.0 (PB1), 4.0 (PA)
	RSV	33	33.6	45.6	6	100	180.2

* For IAV, eight depth or coverage estimates for the PB2, PB1, PA, HA, NP, NA, MP, and NS genes are presented in order

Increasing genomic depths and coverages via iterative mapping

Although a few long contigs might completely or nearly span the full-length genome, many short contigs were mapped only partially to detected viruses (one HRV-C40 and nine HRV-C4 contigs encompassed 100% and 56.3% of the genomes, respectively). An iterative mapping approach was used to include as many NGS reads as possible for reference mapping. Briefly, a consensus sequence was generated from one or more short contigs and reference genomes. The reference genome was used to assemble the missing regions in order to generate an initial template because the short contigs might only cover the partial viral genome. Subsequently, the iterative process was initiated based on this combined genome. Table 1 compares the average depth and coverage using the iterative mapping approach (for various iterations) and *de novo* assembly. Following three iterations, the genomic coverage of the aforementioned HRV-C4 genome increased from 56.3% to 95.2%, with genomic depths from 10.3 to 16.9. The coverages of detected viruses in NGS 1 included nearly complete genomes, except for HRV-C4, with 95.1% coverage, and each of the eight IAV genes, with >77.0% coverage. Furthermore, the contig depths of viral genomes generated by iterative mapping were generally greater than the original contig depths. These results demonstrated that, by using the iterative mapping process, the depths and genomic content of the terminal templates in NGS 1 and 2 were increased compared with those of their respective initial contigs. Read-coverage distributions along the reported viral genomes in NGS 1 (Fig. 2A–2O) and 2 (Fig. 2P–2W) are shown in Figure 2. NGS 1 showed depth distributions with >77.0% coverage and average depths from 7.1 to 7001.1. NGS 2 showed depth

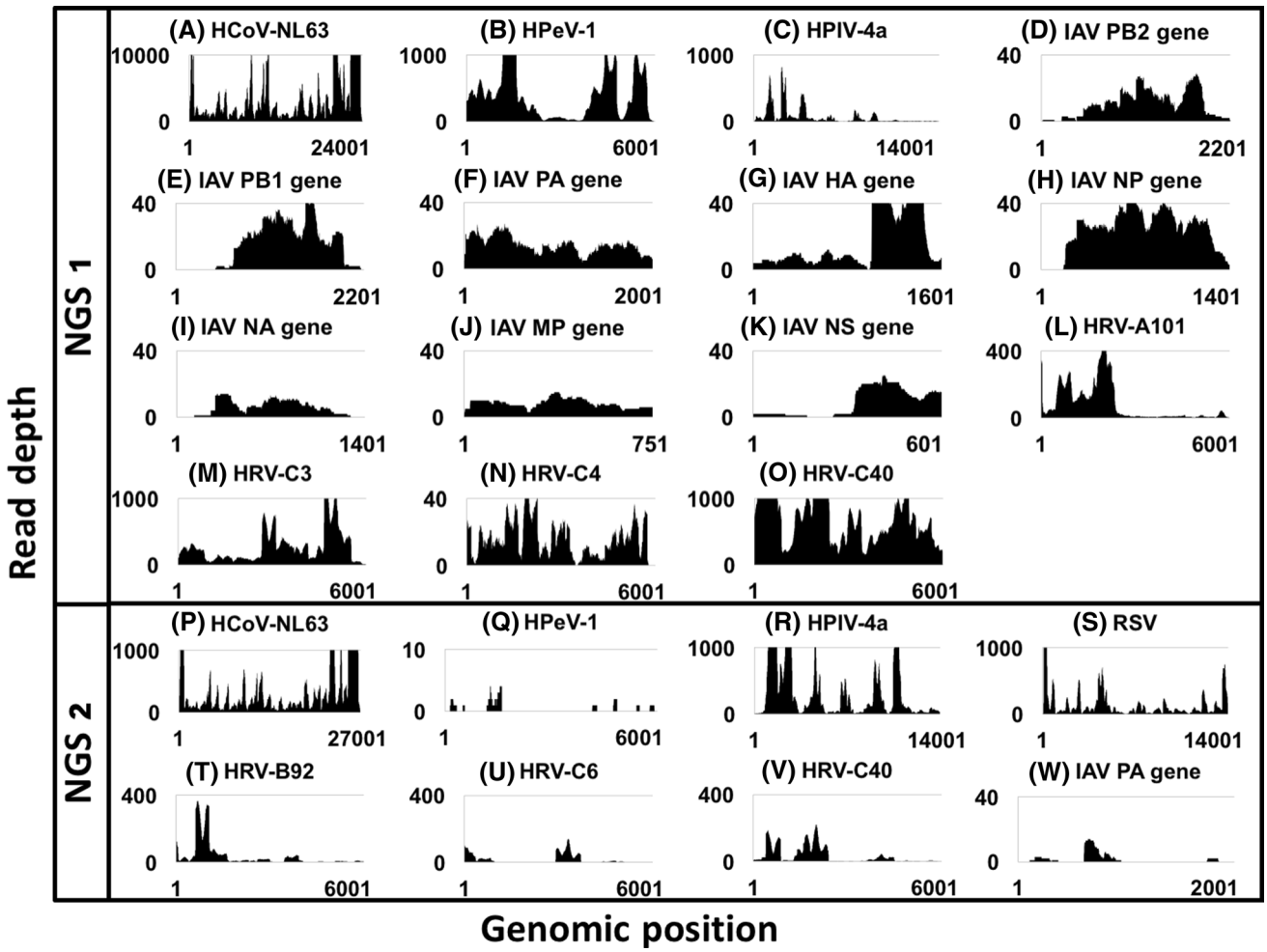

Fig. 2 Read distributions of viral genomes. Read distributions of the viral genomes revealed by read mapping: (A) HCoV-NL63, (B) HPeV-1, (C) HPIV-4a, (D-K) from PB2 to NS gene of IAV, (L) HRV-A101, (M) HRV-C3, (N) HRV-C4, and (O) HRV-C40 in NGS 1; (P) HCoV-NL63, (Q) HPeV-1, (R) HPIV-4a, (S) RSV, (T) HRV-B92, (U) HRV-C6, (V) HRV-C40, and (W) PA gene of IAV in NGS 2. Read distributions in NGS 1 showed >77.0% coverage and average depths from 7.1 to 7001.1, and in NGS 2 showed >83.0% coverage and average depths from 38.1 to 773.8, except for HPeV-1, HRV-C6, and IAV. The read distribution of IAV PB1 gene in NGS 2 was not shown, due to only one read mapping to this gene

distributions with >83.0% coverage and average depths from 38.1 to 773.8, except for HPeV-1, HRV-C6, and IAV.

PCR results for each clinical specimen

Following the NGS analysis of mixed samples, virus-specific molecular diagnoses for each of the 40 specimens were determined using the predicted viruses shown in Table 1. PCR was performed to validate the presence of viruses, including EV, HCoV-229E, -NL63, and -OC43, hMPV-2, HPIV-3 and -4, HPeV, HRV, IAV, RSV, and Rota, using specific PCR primers (Table 2 and Supplementary File 1). PCR results agreed with NGS predictions for NGS 1; however, for sample IDs 40, 22, and 35, PCR analysis detected hMPV, HRV-C3, and HRV-C (without a specific type), respectively, but these could not be identified by NGS. By contrast, HPeV-1 was detected by NGS, with a genomic depth and coverage of 1.6 and 19.4, respectively, but was not identified by PCR.

In terms of clinical symptoms, the 40 specimens used in this study were grouped into ten croup, nine acute bronchiolitis, and 21 RTIs cases. The viruses detected in the hospitalized children with acute bronchiolitis included four HRVs of type A101, C3, C4, and C40, followed by two each of HCoV-NL63 and IAV and one each of HPIV-4a and hMPV. The viruses detected in the hospitalized children with acute bronchiolitis included six HCoV-NL63, followed by one each of HRV-B92, HRV-C40, and HPeV-1. The viruses detected in the hospitalized children with RTIs included seven HRV-Cs, followed by one each of HRV-B92, HPIV-4a, and RSV. Notably, the negative

Table 2 PCR confirmation of the 40 clinical samples. Virus-specific PCR assays for each of the 40 isolates were performed to confirm NGS-predicted viruses and to validate their presence. PCR results agreed with NGS predictions in NGS 1; however, hMPV, HRV-C3, and an unclassified type of HRV-C detected by PCR and HPeV-1 detected by NGS were inconsistent in NGS 2 and are marked in bold

NGS pool: prediction	Sample ID	PCR-validated candidates
1: HCoV-NL63, HPeV-1, HPIV-4a, HRV-A101, -C3, -C4, and -C40, and IAV	1	IAV
	2	HRV-C3
	3	HCoV-NL63
	4	HRV-C4
	5	HCoV-NL63
	6	HRV-A101
	7	HPIV-4a
	8	—
	9	HRV-C40
	10	HCoV-NL63, HPeV-1
	11	HRV-C40
	12	—
	13	HCoV-NL63
	14	—
	15	HCoV-NL63
	16	—
	17	—
	18	HRV-C40
	19	**—**
	20	HCoV-NL63
2: HCoV-NL63, **HPeV-1**, HPIV-4a, HRV-B92, -C26, and -C40, IAV, and RSV	21	HRV-C40
	22	**HRV-C3**
	23	HRV-C6
	24	—
	25	HCoV-NL63
	26	RSV
	27	HRV-B92
	28	—
	29	IAV, HPIV-4a
	30	—
	31	—
	32	—
	33	HRV-C40
	34	HRV-C6
	35	**HRV-C (without a specific type)**
	36	HRV-B92
	37	—
	38	—
	39	—
	40	HCoV-NL63, **hMPV**

detection rate of 52.4% (11 of 21) for hospitalized children with RTIs was much greater than the 11% and 20% rates observed for acute bronchiolitis and croup cases, respectively.

Read depths and genetic variants of viral genomes

A limitation of this study was the inability to distinguish NGS reads from different samples with the same viral type;

however, HRV-C3, HRV-A101, HPIV-4a, and HPeV-1 in NGS 1 and RSV and HPIV-4a in NGS 2 showed single viral types with complete coverages in their mixed samples. Read mapping was performed to assess the overall depth distribution along the genome and to identify genetic variants exhibiting heterogeneous nucleotide compositions. A genetic variant was defined as including nucleotide positions where the major nucleotide appeared in <75% of the mapped reads with read depths of at least 20. For the aforementioned viruses, ten, one, three, and five nonsynonymous mutations for HRV-C3, HRV-A101, HPIV-4a, and HPeV-1 of NGS 1, respectively, and 17 and three for RSV and HPIV-4a of NGS 2, respectively (Table 3), were detected. These six genomes were generated by iterative mapping. In addition to genetic variants, we observed two insertions located at positions 904–906 (codon ATA) and 988–990 (codon CCA) in the VP2 gene in our Taiwanese HRV-A101 strain. Compared with three genomes (accession numbers GQ415051, JQ245965, and GQ415052) downloaded from GenBank in October 2016, the first insertion was absent in GQ415051 and JQ245965, and the second insertion was absent in GQ415052.

Discussion

More than 25% of clinical samples harbour undetected viral pathogens [18–20]; therefore, new diagnostic tests, such as NGS-based analyses, are required for viral identification. Here, we collected 40 clinical samples for which viruses were not detected by routine viral culture or PCR methods. After applying an NGS data analysis pipeline, we successfully identified viral pathogens from 26 clinical samples (including three co-infections in sample IDs ten, 29, and 40) in two NGS groups. Detected viruses included HCoV-NL63, HPeV-1, HPIV-4a, HRV-A101, -C3, -C4, and -C40, and IAV from NGS 1 and HCoV-NL63, hMPV, HPeV-1, HPIV-4a, HRV-B92, -C26, and -C40, IAV, and RSV from NGS 2. The two dominant viral populations were HRV (n = 13) and HCoV-NL63 (n = 8).

Reads of assembled contigs were categorized as viruses, bacteria, and eukaryotes (Fig. 1). The two pooled samples exhibited different taxonomic compositions; for example, 86% of total reads belonged to the virus group in NGS 1, but only 6% in NGS 2. This difference might be explained by differences in viral copies in mixed samples. Nevertheless, a detection rate of 60% in NGS 2 was obtained using our NGS data analysis pipeline, approaching 70% in NGS 1.

In addition to low viral copies, challenges to the recovery of viral genomes from clinical samples using NGS platforms include bacterial and human genome contamination. In this study, we applied an NGS data analysis pipeline using iterative mapping to improve the recovery of virus-derived reads in mixed clinical samples, in order to identify their genomic sequences. Genomic depths and coverages in two NGS samples were effectively increased using this process; however, these increases were not evident for viral isolates [6]. A decrease in contamination with human and bacterial genomes increases the depth and coverage of viral contigs from viral isolates. In this study, viral genomes were identified following ≤6 iterations (as reference mappings) using this approach.

NGS has various advantages over PCR-based methods. Using this technology, whole genomes can be obtained for viral detection, genotyping, and diversity analyses without designing primers or prior knowledge of viral presence [1, 6]. Correlations between NGS reads and PCR cycle threshold values have been observed, suggesting that NGS reads are suitable indicators of viral copy number [2, 26]. In this study, NGS applied to pooled samples offered advantages, including reduced sequencing costs and the ability to screen multiple epidemiological samples for potential outbreaks. However, the use of mixed samples had some limitations, including the inability to assign viral candidates to each clinical sample or to calculate the association between NGS reads and Ct values when multiple viruses of the same type were pooled in a single NGS group.

In this study, 40 specimens were collected from 10, nine, and 21 hospitalized children diagnosed with croup, acute bronchiolitis, and RTIs, respectively. The dominant viral population in the acute bronchiolitis cases was HRV (including types A101, C3, C4, and C40), which is inconsistent with previous studies indicating that RSV is the dominant population in patients with bronchiolitis [18, 27]. HRV type C was first reported in 2007 and could not be grown by standard cell culture methods [28]. A systematic approach has been proposed to culture the virus in differentiated epithelial cells of the human airway at the air–liquid interface [29]; however, this method is not suitable for routine viral diagnosis in clinical laboratories. Accordingly, the prevalence of HRV-C in acute bronchiolitis cases has likely been underestimated. HCoV-NL63, a global respiratory tract pathogen, is a primary cause of respiratory disease in children admitted to hospital in Taiwan, specifically in patients diagnosed with croup [30]. This agrees with our finding showing that six out of ten croup cases were diagnosed as NL63 infections. One study demonstrated that newly discovered viruses might fail to be amplified by RT-PCR assays designed for known viruses, due to primer sequence mismatches [3]. Therefore, NGS analysis, which does not require specific primers and probes, is superior to these diagnostic assays for the detection of viruses harboring novel mutations in PCR primer binding sites. Viruses detected in patients

Table 3 Genetic variants in heterogeneous populations identified by read mapping

NGS pool	Sample ID	Viral type	Gene	Nonsynonymous mutation	Variant nucleotide position (total depth: nucleotide composition/frequency)
1	2	HRV-C3	VP2	L307I	919 (135: T/100, A/27, C/7, G/1)
			VP1	D622Y	1864 (79: G/58, T/15, C/5, A/1)
				N623K	1869 (69: C/50, A/19)
			2C	Q1486K	4456 (234: C/164, A/46, T/19, G/5)
				P1487Q/T/K	4459 (194: C/120, A/59, T/15); 4460 (186: C/131, A/49, T/5 G/1)
				V1489I	4465 (133: G/93, A/26, C/7, T/7)
			3D	I1976N/K/L/H/Q	5926 (372: A/240, C/90, T/29, G/13); 5927 (328: T/201, A/68, C/52, G/7); 5928 (271: T/131, A/63, C/48, G/29)
				F1977I	5929 (218: T/129 A/58 C/26 G/5)
				I1980N/K	5939 (82: T/58, A/19, G/5); 5940 (80: T/56, A/14, G/7, C/3)
				I1981L	5941 (76: A/55, T/10, G/9, C/2)
	6	HRV-A101	VP4	T15K	44 (90: C/66, A/24)
	7	HPIV-4a	M	C201R	604 (88: T/44, A/40, G/4)
				L202M	609 (117: A/81, G/21, T/15)
			L	R3595G	754 (36: A/24, G/12)
	10	HPeV-1	VP1	N589K	1767 (1054: T/766, A/161, G/103, C/24)
			2A	K918N	2754 (21: A/15, C/3, G/2, T/1)
				G919V	2756 (21: G/15, T/4, A/2)
			3D	Y1770N	5308 (302: T/221, A/53, C/28)
				F2126L	6376 (430: T/320, C/75, A/34, G/1)
2	26	RSV	NS1	S99P	295 (740: T/530, C/124, A/76, G/10)
			NS2	D2N	4 (75: G/53, A/21, C/1)
			N	M257L	769 (94: A/70, T/14, G/6, C/4)
				L258F	774 (106: A/75, T/22, G/6, C/3)
				E388D	1164 (21: G/10, T/10, C/1)
			F	R49G	145 (50: A/36, G/7, T/6, C/1)
				T50S	148 (66: A/46, T/15, G/5)
			M2	I173T	518 (71: T/50, C/16, A/5); 519 (68: C/49, T/15, A/4)
				N174H	520 (62: A/45, C/15, T/2)
			L	L196S	587 (30: T/16, C/14)
				Q198R	593 (39: A/26, G/9, T/4)
				Y724S	2171 (26: A /17, C/7, T/2)
				S767G	2299 (52: A/34, G/13, T/3, C/2)
				S910T	2729 (54: G/38, C/15, T/1)
				N969D	2905 (35: A/20, G/15)
				N991D	2971 (57: A/42, G/8, C/7)
				R1256G	3766 (52: A/29, G/22, T/1)
	29	HPIV-4a	NP	R216S	648 (25: A/18, C/5, G/2)
				A268E	803 (128: C/92, A/32, T/3, G/1)
				P345L	1034 (1801: C/1265, T/451, A/59, G/26)

diagnosed with RTIs included seven HRV type C, and one each of HRV type B, RSV, and HPIV-4a. The negative detection rate for patients with RTIs was 52.4% (11 of 21), which was greater than that for acute bronchiolitis cases (11.1%; one of nine) and croup cases (20%; two of ten). Further investigations are needed to identify currently unknown human viruses in patients diagnosed with RTIs.

HRVs, including 11 type C, two B, and one A, were the dominant viruses detected in this study. HRV-C causes more severe illnesses than HRV-A and -B [31, 32]; however, limited HRV-C genomes have been published due to the lack of availability of culture systems in clinical laboratory settings [33]. More detailed viral typing information, for viruses such as HRV, HEV, hMPV, and RSV, can be obtained by NGS than by PCR-based methods [26]. Our results demonstrated that NGS effectively characterizes respiratory viral infections, specifically those involving HRV; however, two HRVs (one C3 and the other un-typed) and one hMPV were not identified by NGS in NGS 2 of this study. It was difficult to discriminate between all HRV genotypes in this mixed NGS sample because NGS 2 contained eight HRVs that shared similar genetic segments. Additionally, hMPV might exhibit a low viral load in sample ID 40 co-infected with HCoV-NL63. In other words, most of the reads in NGS 2 were derived from viruses with high titers. However, one HPeV-1 in NGS 2 was not identified by PCR. This HPeV exhibited a genomic coverage of 19.4% in our NGS analysis. Further investigations are needed to explain this low coverage (e.g. low viral titers or bleed-through contamination from the HPeV-1 in NGS 1 [34]).

In this metagenomics study, an NGS data analysis pipeline was used to identify viral pathogens from 40 hospitalized patients. No viruses were detected in these clinical samples using routine culture and PCR methods; however, our NGS pipeline with PCR confirmation resulted in a detection rate of 65% (26 of 40 samples). With respect to nucleotide diversity in the detected viruses, 39 polymorphic genetic variants were detected from six whole genomes using mapped NGS reads. Further analyses are needed to determine the biological consequences of this genetic variation. Our results indicated that the NGS method was able to detect viruses that cannot be directly identified by routine culture or PCR methods targeting divergent or other viruses. Furthermore, this NGS method could be used to reveal genetic sequences and detect diversity from mixed samples using an iterative mapping approach. Our results provide the first complete genomes for HRV-A101, HRV-C3, HPIV-4a, and RSV in Taiwan. In future work, we will apply this method to the rapid identification of emerging viruses directly from clinical samples during potential outbreaks as well as to the detection of genetic variants, to further explore the associations between viral genotypes and disease severity.

Acknowledgements The authors thank the staff of the Virology Laboratory, Linkou Chang Gung Memorial Hospital, Taoyuan, Taiwan for helpful discussions.

Compliance with ethical standards

Funding This study was partially supported by two grants from Linkou Chang Gung Memorial Hospital, Taoyuan, Taiwan to KCT (No. CLRPG3B0045 and CMRPG3F1881; http://www.cgmh.org.tw). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This study was approved by the Institutional Review Board of Linkou Chang Gung Memorial Hospital, Taoyuan, Taiwan (approval no. 100-4378B). All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Lecuit M, Eloit M (2014) The diagnosis of infectious diseases by whole genome next generation sequencing: a new era is opening. Front Cell Infect Microbiol 4:25. doi:10.3389/fcimb.2014.00025
2. Prachayangprecha S, Schapendonk CME, Koopmans MP et al (2014) Exploring the potential of next-generation sequencing in detection of respiratory viruses. J Clin Microbiol 52:3722–3730. doi:10.1128/JCM.01641-14
3. van der Hoek L, Pyrc K, Jebbink MF et al (2004) Identification of a new human coronavirus. Nat Med 10:368–373. doi:10.1038/nm1024
4. Baillie GJ, Galiano M, Agapow P-M et al (2012) Evolutionary dynamics of local pandemic H1N1/2009 influenza virus lineages revealed by whole-genome analysis. J Virol 86:11–18. doi:10.1128/JVI.05347-11
5. Ninomiya M, Ueno Y, Funayama R et al (2012) Use of Illumina deep sequencing technology to differentiate hepatitis C virus variants. J Clin Microbiol 50:857–866. doi:10.1128/JCM.05715-11
6. Gong Y-N, Chen G-W, Yang S-L et al (2016) A next-generation sequencing data analysis pipeline for detecting unknown pathogens from mixed clinical samples and revealing their genetic diversity. PLoS One 11:e0151495. doi:10.1371/journal.pone.0151495
7. Bochkov YA, Palmenberg AC, Lee W-M et al (2011) Molecular modeling, organ culture and reverse genetics for a newly identified human rhinovirus C. Nat Med 17:627–632. doi:10.1038/nm.2358
8. Allander T (2008) Human bocavirus. J Clin Virol 41:29–33. doi:10.1016/j.jcv.2007.10.026
9. Woo PCY, Lau SKP, Chu C et al (2005) Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. J Virol 79:884–895. doi:10.1128/JVI.79.2.884-895.2005
10. Chen K, Pachter L (2005) Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comput Biol 1:e24. doi:10.1371/journal.pcbi.0010024
11. Cheval J, Sauvage V, Frangeul L et al (2011) Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. J Clin Microbiol 49:3268–3275. doi:10.1128/JCM.00850-11

12. Kuroda M, Katano H, Nakajima N et al (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. doi:10.1371/journal.pone.0010256
13. Yongfeng H, Fan Y, Jie D et al (2011) Direct pathogen detection from swab samples using a new high-throughput sequencing technology. Clin Microbiol Infect 17:241–244. doi:10.1111/j.1469-0691.2010.03246.x
14. Greninger AL, Chen EC, Sittler T et al (2010) A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. PLoS One 5:e13381. doi:10.1371/journal.pone.0013381
15. Yang J, Yang F, Ren L et al (2011) Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. J Clin Microbiol 49:3463–3469. doi:10.1128/JCM.00273-11
16. Li D, Li Z, Zhou Z et al (2016) Direct next-generation sequencing of virus-human mixed samples without pretreatment is favorable to recover virus genome. Biol Direct 11:3. doi:10.1186/s13062-016-0105-x
17. Conceição-Neto N, Zeller M, Lefrère H et al (2015) Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. Sci Rep 5:16532. doi:10.1038/srep16532
18. Chen Y-W, Huang Y-C, Ho T-H et al (2014) Viral etiology of bronchiolitis among pediatric inpatients in northern Taiwan with emphasis on newly identified respiratory viruses. J Microbiol Immunol Infect 47:116–121. doi:10.1016/j.jmii.2012.08.012
19. Thomazelli LM, Vieira S, Leal AL et al (2007) Surveillance of eight respiratory viruses in clinical samples of pediatric patients in southeast Brazil. J Pediatr (Rio J) 83:422–428. doi:10.2223/JPED.1694
20. Jain S, Self WH, Wunderink RG et al (2015) Community-acquired pneumonia requiring hospitalization among U.S. adults. N Engl J Med 373:415–427. doi:10.1056/NEJMoa1500245
21. Wang Q, Jia P, Zhao Z (2013) VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. PLoS One 8:e64465. doi:10.1371/journal.pone.0064465
22. Ho T, Tzanetakis IE (2014) Development of a virus detection and discovery pipeline using next generation sequencing. Virology 471–473:54–60. doi:10.1016/j.virol.2014.09.019
23. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. doi:10.1038/nmeth.1923
24. Koboldt DC, Zhang Q, Larson DE et al (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22:568–576. doi:10.1101/gr.129684.111
25. Deng X, Naccache SN, Ng T et al (2015) An ensemble strategy that significantly improves de novo assembly of microbial genomes from metagenomic next-generation sequencing data. Nucleic Acids Res 43:e46. doi:10.1093/nar/gkv002
26. Thorburn F, Bennett S, Modha S et al (2015) The use of next generation sequencing in the diagnosis and typing of respiratory infections. J Clin Virol 69:96–100. doi:10.1016/j.jcv.2015.06.082
27. Sung C-C, Chi H, Chiu N-C et al (2011) Viral etiology of acute lower respiratory tract infections in hospitalized young children in Northern Taiwan. J Microbiol Immunol Infect 44:184–190. doi:10.1016/j.jmii.2011.01.025
28. Lau SKP, Yip CCY, Tsoi H-W et al (2007) Clinical features and complete genome characterization of a distinct human rhinovirus (HRV) genetic cluster, probably representing a previously undetected HRV species, HRV-C, associated with acute respiratory illness in children. J Clin Microbiol 45:3655–3664. doi:10.1128/JCM.01254-07
29. Ashraf S, Brockman-Schneider R, Gern JE (2015) Propagation of rhinovirus-C strains in human airway epithelial cells differentiated at air–liquid interface. Methods Mol Biol 1221:63–70. doi:10.1007/978-1-4939-1571-2_6
30. Wu P-S, Chang L-Y, Berkhout B et al (2008) Clinical manifestations of human coronavirus NL63 infection in children in Taiwan. Eur J Pediatr 167:75–80. doi:10.1007/s00431-007-0429-8
31. McErlean P, Shackelton LA, Lambert SB et al (2007) Characterisation of a newly identified human rhinovirus, HRV-QPM, discovered in infants with bronchiolitis. J Clin Virol 39:67–75. doi:10.1016/j.jcv.2007.03.012
32. Miller EK, Khuri-Bulos N, Williams JV et al (2009) Human rhinovirus C associated with wheezing in hospitalised children in the Middle East. J Clin Virol 46:85–89. doi:10.1016/j.jcv.2009.06.007
33. Khaw YS, Chan YF, Jafar FL et al (2016) Comparative genetic analyses of human rhinovirus C (HRV-C) complete genome from Malaysia. Front Microbiol 7:543. doi:10.3389/fmicb.2016.00543
34. Wilson MR, Fedewa G, Stenglein MD et al (2016) Multiplexed metagenomic deep sequencing to analyze the composition of high-priority pathogen reagents. mSystems. doi:10.1128/mSystems.00058-16