JOURNAL OF CLINICAL MICROBIOLOGY, Oct. 2011, p. 3463–3469 Vol. 49, No. 10
0095-1137/11/$12.00 doi:10.1128/JCM.00273-11

Unbiased Parallel Detection of Viral Pathogens in Clinical Samples by Use of a Metagenomic Approach▽‡

Jian Yang,[1]† Fan Yang,[1]† Lili Ren,[1,2]† Zhaohui Xiong,[1] Zhiqiang Wu,[1] Jie Dong,[1] Lilian Sun,[1] Ting Zhang,[1] Yongfeng Hu,[1] Jiang Du,[1] Jianwei Wang,[1,2]* and Qi Jin[1]*

State Key Laboratory for Molecular Virology and Genetic Engineering, Institute of Pathogen Biology (IPB), Chinese Academy Medical Sciences (CAMS) & Peking Union Medical College (PUMC), Beijing 100176, China,[1] and Christophe Mérieux Laboratory, IPB, CAMS-Fondation Mérieux, CAMS&PUMC, Beijing 100176, China[2]

Received 8 February 2011/Returned for modification 22 March 2011/Accepted 22 July 2011

Viral infectious diseases represent a major threat to public health and are among the greatest disease burdens worldwide. Rapid and accurate identification of viral agents is crucial for both outbreak control and estimating regional disease burdens. Recently developed metagenomic methods have proven to be powerful tools for simultaneous pathogen detection. Here, we performed a systematic study of the capability of the short-read-based metagenomic approach in the molecular detection of viral pathogens in nasopharyngeal aspirate samples from patients with acute lower respiratory tract infections ($n = 16$). Using the high-throughput capacity of ultradeep sequencing and a dedicated data interpretation method, we successfully identified seven species of known respiratory viral agents from 15 samples, a result that was consistent with results of conventional PCR assays. We also detected a coinfected case that was missed by regular PCR testing. Using the metagenomic data, 11 draft genomes of the abundantly detected viruses in the samples were reconstructed with 21.84% to 98.53% coverage. Our results show the power of the short-read-based metagenomic approach for accurate and parallel screening of viral pathogens. Although there are some inherent difficulties in applying this approach to clinical samples, including a lack of controls, limited specimen quantity, and high contamination rate, our work will facilitate further application of this unprecedented high-throughput method to clinical samples.

Despite great advances in modern medicine, infectious diseases are among the most common causes of human death and represent the greatest disease burdens worldwide, as estimated by the WHO (13). As we have experienced in the past decade, viral pathogens, especially respiratory viruses, pose significant threats to public health. Indeed, both the severe acute respiratory syndrome (SARS) outbreak and the H1N1 influenza A virus pandemic had serious impacts on public health as well as the global economy. Undoubtedly, the rapid identification of newly emerging pathogens using advanced molecular techniques has played a crucial role in controlling such infectious diseases worldwide. Moreover, accurate determination of regional agent-specific infectious diseases provides critical information for local authorities to optimize public health prevention strategies and maximize the limited available resources, particularly in developing countries.

Molecular assays are becoming increasingly popular for laboratory diagnosis because they outperform traditional viral diagnostic methods, such as cell-culture-based testing and antigen detection, in efficiency, sensitivity, and specificity. In addition, several multiplex PCR assays and DNA microarray testing methods have been proposed to screen dozens to hundreds of pathogens simultaneously (10, 11, 23). Metagenomics is a powerful tool for viral and microbial community characterization and new pathogen identification because genetic materials are isolated directly from environmental/clinical samples and sequenced; thus, no culturing, cloning, or *a priori* knowledge of what pathogens may be present is required.

The recent advent of next-generation sequencing (NGS) technologies, such as pyrosequencing (Roche/454), sequencing by synthesis (Solexa; Illumina), and sequencing by ligation (SOLiD; Applied Biosystems), has dramatically decreased sequencing cost and time (reviewed in reference 14). These methods have fueled studies of the shotgun metagenomic strategy, which targets the metagenomes of entire communities rather than just the 16S rRNA of bacteria, which was common in many former metagenomic studies. The shotgun metagenomics strategy has recently been applied to clinical samples, such as fecal (5, 16, 17, 22) and nasal (4, 17, 26, 27) specimens.

To date, most metagenomic studies of clinical samples have employed pyrosequencing technology, which produces longer reads but less output than other NGS platforms. Short read length was believed to be a barrier for the application of ultradeep sequencing (e.g., Solexa [Illumina] and SOLiD [Applied Biosystems]) in metagenomics for years. However, a recent study investigating functional gene categories in the human gut demonstrated the potential of short-read sequencing in characterizing complex microbiomes (19). Moreover, in a

* Corresponding author. Mailing address for Qi Jin: State Key Laboratory for Molecular Virology and Genetic Engineering, Institute of Pathogen Biology, 6 Rongjing Eastern Street, BDA, Beijing 100176, China. Phone: 86 10 67877732. Fax: 86 10 67877736. E-mail: zdsys@vip.sina.com. Mailing address for Jianwei Wang: Christophe Mérieux Laboratory, IPB, CAMS-Fondation Mérieux, CAMS & PUMC, 9 Dong Dan San Tiao, Dongcheng District, Beijing 100730, China. Phone and fax: 86 10 65105188. E-mail: wangjw28@163.com.

† These authors contributed equally to this work.

‡ Supplemental material for this article may be found at http://jcm.asm.org/.

▽ Published ahead of print on 3 August 2011.

TABLE 1. Summary of clinical samples and ultradeep sequencing information

Sample	Gender[a]	Age	No. of valid sequencing reads	
			DNA library	cDNA library
009A	M	2 yr 10 mo	1,308,914	1,034,291
086A	M	7 mo	1,927,751	2,006,094
135A	F	9 mo	3,825,472	3,904,105
141A	M	4 mo	2,045,227	2,494,505
182A	M	7 mo	5,246,621	4,917,550
183A	M	5 mo	4,032,561	4,573,611
190A	F	10 yr 3 mo	4,151,168	4,135,859
287A	M	2 mo	4,215,810	4,652,574
306A	M	9 yr 9 mo	3,747,175	5,014,404
362A	M	1 yr	1,811,167	5,848,517
365A	F	4 mo	1,567,051	3,809,508
593A	F	1 yr 10 mo	1,519,376	2,419,513
607A	F	7 yr	1,497,088	4,338,462
636A	M	11 mo	4,014,872	3,644,000
653A	M	4 mo	4,373,248	3,500,024
723A	M	7 mo	1,075,188	3,239,499

[a] F, female; M, male.

pilot study, we successfully identified influenza A viruses from swab specimens using a short-read-based metagenomic approach (27), and Greninger and colleagues constructed near full-length genomic segments of the influenza A (H1N1) virus by *de novo* assembly using short-read sequencing data (4).

To further investigate the potential of short-read-based metagenomic methods with simultaneous detection of various respiratory pathogens, we systematically analyzed a cohort of 16 clinical samples from children with acute lower respiratory tract infections (ALRTIs). By introducing a statistical index, we successfully identified various viral agents from the high-throughput data, demonstrating the power of the short-read-based metagenomic approach for accurate and parallel screening of respiratory pathogens. Additionally, we discuss the potential (advantages and disadvantages) for further application of different NGS platforms in the clinical screening of pathogenic viruses.

MATERIALS AND METHODS

Sample collection. Nasopharyngeal aspirates (NPAs) were collected from 16 children (11 males and 5 females) with ALRTIs upon admission to the Beijing Children's Hospital. The children ranged in age from 2 months to 10 years, with a median age of 8 months (Table 1). To optimize selection of patients with ALTRIs due to viral infections and exclude typical bacterial infections, patients enrolled in the study were selected by physicians according to the following previously used criteria (20): (i) respiratory symptoms, such as coughing or wheezing; (ii) acute fever (body temperature $\geq$ 38°C); and (iii) normal or low leukocyte count.

Nucleic acid preparation and conventional/nested PCR assays. To ensure the maximal retrieval of virome profiles, each NPA sample (1 ml) was divided into two parts (0.5 ml each) for DNA and RNA extraction. Total DNA was extracted using the Qiagen Genomic-tip 20/G and Genomic DNA Buffer set (Qiagen, Hilden, Germany). Total RNA was isolated using an RNeasy minikit (Qiagen, Hilden, Germany), and cDNA was then synthesized using a Moloney murine leukemia virus reverse transcriptase cDNA synthesis kit (Takara, Otsu, Japan).

Each specimen was independently screened for the presence of common respiratory viruses, including influenza viruses (IFVs), parainfluenza viruses (PIVs), human coronaviruses, human enteroviruses (HEVs), human rhinoviruses (HRVs), human metapneumoviruses, human adenoviruses (HAdVs), respiratory syncytial viruses (RSVs), and human bocaviruses (HBoVs), using PCR or reverse transcription-PCR (RT-PCR) as described previously (8, 20). Two primer pairs targeting the VP2- and VP3-encoding sequences of echovirus 11 (GenBank accession number AY167105) were designed and used for nested PCR in sample 723A, amplifying a 381-bp fragment. The external primers were the following: Echo11-F1 (5′-TGTTTAGTAAGACCGCGACCAATG-3′) and Echo11-R1 (5′-GCTCTGAACAGGAATGCGGAAAAT-3′). The internal primers were the following: Echo11-F2 (5′-AACCTTACCATATTCCCGCACCAG-3′) and Echo11-R2 (5′-CTCCATCAGATTCTTCACCTCACC-3′).

Real-time RT-PCR analysis. Five respiratory syncytial virus (RSV)-positive samples were used for real-time RT-PCR analysis. For production of viral RNAs, the viral templates were amplified using primers RSV-CF (5′-GATGGGGCAA ATATGGAAACA-3′) and RSV-CR (5′-GATTGCAAATCGTGTAGCTGT-3′). The amplicon was cloned into the pGEM-T Easy vector to construct the standard plasmid. We then generated the standard curve using the quantitative primers RSV-F (5′-TGGAAACATACGTGAACAARCTTCA-3′) and RSV-R (5′-GCACCCATATTGTWAGTGATGCA-3′), which have been previously described (7).

Single-stranded cDNA was synthesized from 300 ng of total RNA from each sample using a Superscript first-strand synthesis system (Invitrogen) in a 20-μl reaction mixture according to the manufacturer's instructions. Real-time PCR for RSV was conducted using an ABI Prism 7000 real-time PCR system (Applied Biosystems). Reactions were performed in a 50-μl volume containing 1 μl of cDNA, 1 μl of each quantitative primer, and 25 μl of Power SYBR green PCR master mix (Applied Biosystems). Conditions for real-time PCR were 50°C for 2 min and 95°C for 10 min, followed by 40 cycles at 95°C for 15 s and 60°C for 1 min. The absolute quantity of viral RNA was calculated using a standard curve, and a melting curve analysis was performed to verify the authenticity of the amplification.

Deep sequencing and initial bioinformatics processing. DNA and cDNA libraries were constructed according to the manufacturer's instructions (Illumina). Each library was sequenced with an Illumina GA II sequencer for a single read of 36 bp in length. To maximize the available length of sequencing reads and the total output of raw data, each library of each sample was sequenced in an individual lane without indexing. Initial image analysis and base calling were performed with the GAPipeline program (version 1.0; Illumina) using default parameters. A series of in-house Perl scripts were then employed for further quality control with the following screen criteria: (i) reads with no-call sites (i.e., containing N residues in sequences), (ii) reads containing >7 low-quality base calls (i.e., those with Phred-like quality scores of <20, which correspond to an error rate of >1%), (iii) reads with similarity to the sequencing adaptor, and (iv) duplicate reads. Only reads passing the quality control were considered valid sequences. We tried to skip the duplication remove step in data analysis and found that there were no essential differences between the two results (data not shown).

Host genetic materials (DNA or RNA) are intrinsic contaminants in complete metagenomic studies of human microbiomes (26). Hence, reads with similarity to sequences in the human reference genome (NCBI version 36) or the human expressed sequence tag (EST) data set (downloaded from the NCBI FTP server in September 2010) were filtered using the Bowtie ultrafast short sequence aligner with default parameters (9), allowing up to 2 mismatches in the leading 28 bp and up to 70 for the sum of the quality scores of all mismatches. All postscreened data sets are available from the website http://www.mgc.ac.cn/Resources/Yang_et_al_metagenomic_datasets.tar.bz2.

Taxonomic assignment. Sequence similarity-based taxonomic assignments were applied to the nonhost data sets. The Bowtie program was used again to efficiently align each short sequence to sequences in the nonredundant nucleotide (NT) database from GenBank (downloaded in September 2010). We specified the parameter "-a --best" to allow the program to report all equally best matches to the reference sequence with up to 4 mismatches. The results were then imported into MEGAN software (6) to assign each sequence to the lowest common ancestor (LCA) of the set of taxa that it hits. The "min support" (minimum support) parameter was changed to 1 to include all assigned sequences.

Data normalization. For the quantitative comparison of identified viruses among different samples, sequences assigned to each virus were normalized by individual viral genome size and the total number of valid sequences generated from the library of each sample (number of valid tags per million sequences per kb of genome [VTMK]) using the formula $10^6 C_p/NL$, where C_p is the number of sequences assigned to the p virus, N is the total number of valid sequences produced from the sample, and L is the average genome size of the p virus in kilobase pairs. The index was calculated individually for each library. The final VTMK index for each virus in each sample was based on the following criteria: (i) for RNA viruses that do not have a DNA phase, use only index values from the cDNA library, and (ii) for DNA viruses, use the average of the index values from both libraries. To avoid potential random sequence matches, an arbitrary enrolling cutoff VTMK of $\geq$0.5 was used for each virus.

TABLE 2. General ultradeep sequencing information for each sample

Sample	No. of valid sequences[a]	Origin (%)		No. of remaining sequences	No. (%) of reads derived from[b]:	
		Human	Unknown		Viruses	Bacteria
009A	2,343,205	93.65	5.73	14,531	873 (6.01)	5,130 (35.30)
086A	3,933,845	91.64	7.43	36,644	504 (1.38)	21,640 (59.05)
135A	7,729,577	94.76	4.49	57,400	1,687 (2.94)	31,490 (54.86)
141A	4,539,732	84.02	13.47	114,180	846 (0.74)	98,740 (86.48)
182A	10,164,171	95.49	3.81	70,494	268 (0.38)	31,913 (45.27)
183A	8,606,172	94.22	5.15	54,624	1,256 (2.30)	23,051 (42.20)
190A	8,287,027	93.02	5.19	148,100	2,146 (1.45)	114,238 (77.14)
287A	8,868,384	93.25	5.59	103,018	6,781 (6.58)	68,864 (66.85)
306A	8,761,579	93.21	5.17	141,542	1,507 (1.06)	100,792 (71.21)
362A	7,659,684	94.09	4.41	114,454	11,914 (10.41)	68,354 (59.72)
365A	5,376,559	90.23	4.22	298,194	3,377 (1.13)	276,623 (92.77)
593A	3,938,889	88.88	9.87	49,005	358 (0.73)	31,957 (65.21)
607A	5,835,550	76.82	19.48	215,781	339 (0.16)	165,289 (76.60)
636A	7,658,872	93.83	4.83	102,895	335 (0.33)	64,502 (62.69)
653A	7,873,272	95.07	3.95	77,250	3,143 (4.07)	41,991 (54.36)
723A	4,314,687	92.90	6.40	30,014	2,958 (9.86)	10,418 (34.71)

[a] Pooled data from both DNA and cDNA libraries.
[b] The percentages are relative to the number of remaining sequences.

Viral genome reassembly. On the basis of the analysis of the Bowtie aligner against the NT database, the known viral genome with the most sequence hits was used as the reference for each individual species. For abundantly detected (VTMK > 2) viruses in each sample, only the subset of nonhost sequences that showed similarities to sequences in the NT database were fed to the MAQ program to map the respective reference genome using default parameters (12). The "assemble" command of MAQ was then used to determine the consensus genome of the virus detected in the sample from sequence mapping information.

RESULTS

Ultradeep sequencing of metagenomics. We collected nasopharyngeal aspirate specimens from 16 children with ALRTIs. Total RNA (reverse transcribed to cDNA before sequencing) and total DNA were extracted from each sample. Ultradeep sequencing of cDNA and DNA libraries for each sample generated an average of 3.7 million and 2.9 million valid sequences, respectively (Table 1). However, 76.8% to 95.5% (average, 91.6%) of the valid sequences were derived from the host (human) genome or transcriptome (Table 2). Furthermore, an average of 6.8% (range, 3.8% to 19.5%) of the valid sequences did not have any significant match in known databases (Table 2). Therefore, only about 1.6% of the ultradeep sequencing data from each sample were utilized for further metagenomic analysis (Fig. 1).

Taxonomic analysis using the LCA algorithm assigned the majority (99.7%) of the remaining sequences on the basis of similarities to known data. An average of 61.5% (range, 34.7% to 92.8%) were assigned to the Bacteria kingdom, and only 3.1% were potentially virus derived (Fig. 1). The latter data were valuable for further analyzing the ALRTI-associated human respiratory virome. The remaining eukaryote-like reads included unscreened human-derived sequences and possible contamination as previously observed (17) and were therefore excluded from further analysis.

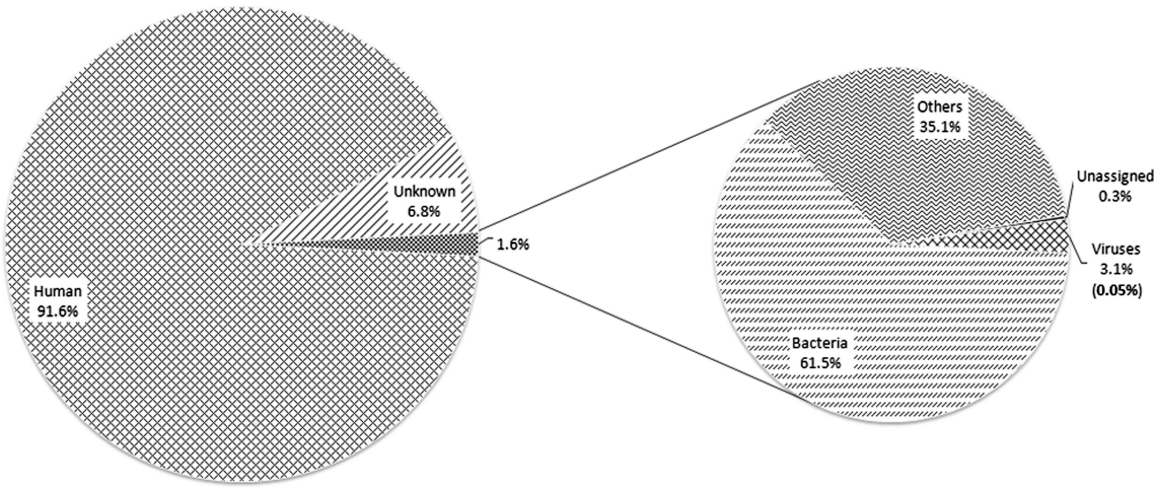

FIG. 1. Pie chart of taxonomic distribution of ultradeep sequencing reads from clinical samples. Data are average values from 16 NPA samples.

TABLE 3. VTMK profiles of known respiratory viruses and results from conventional PCR assay

Sample	RNA viruses						DNA viruses			Conventional PCR assay result
	Human respiratory syncytial virus (HRSV)	Human rhinovirus (HRV)	Influenza A virus (IFVA)	Influenza B virus (IFVB)	Human parainfluenza virus 3 (PIV3)	Human enterovirus B (HEVB)	Human adenovirus (HAdV)	Human bocavirus (HBoV)	Human herpesvirus 5[a]	
009A	0.0	0.0	0.0	0.0	0.0	0.0	**7.4**[b]	0.0	0.0	HAdV
086A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	
135A	0.0	**8.6**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	HRV
141A	0.0	0.2	0.0	0.0	0.0	0.0	0.0	**13.3**	**0.5**	HBoV
182A	0.0	0.0	0.0	0.0	**1.8**	0.2	0.0	0.0	0.0	PIV3
183A	0.0	**0.5**	0.0	0.0	0.0	0.1	**1.5**	0.0	0.0	HAdV, HRV
190A	0.0	**1.2**	0.0	0.0	0.0	0.1	0.0	0.0	**1.0**	HRV
287A	**9.4**	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	HRSV
306A	**2.6**	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	HRSV
362A	**12.1**	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	HRSV
365A	**2.3**	**61.4**	0.0	0.0	0.0	0.0	0.1	0.0	0.0	HRSV, HRV
593A	0.0	0.0	0.0	**5.9**	0.0	0.0	0.1	0.0	0.0	IFVB
607A	0.0	0.0	**0.5**	0.0	0.0	0.1	0.2	0.0	0.0	IFVA
636A	0.0	0.0	0.0	0.0	0.0	0.0	**0.5**	0.0	0.0	HAdV
653A	**6.2**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	HRSV
723A	0.0	**100.2**	0.0	0.0	0.0	**1.0**	0.1	0.0	0.2	HRV

[a] An opportunistic virus that is not directly associated with ALRTIs.
[b] Boldface data represent positive cases identified by the metagenomic method.

Parallel screening of respiratory viruses. Over 100 species of viruses (based on the NCBI taxonomy database) were identified with at least one specific sequence from the 16 samples, but the majority of them (89 species) belonged to bacteriophages and were therefore excluded from further analysis.

As previously observed (4), the raw yields of metagenomic sequencing for different samples vary due to the intrinsic instability in sample preparation, library generation, and sequencing. Therefore, we proposed to use the VTMK index (see Materials and Methods) rather than the absolute number of sequences for comparison between samples. On the basis of the customized criteria, many known respiratory viruses were identified, including human adenovirus (HAdV; 3/16 samples), human respiratory syncytial virus (HRSV; 5/16), human rhinovirus (HRV; 5/16), influenza virus (IFV; 2/16), parainfluenza virus (PIV; 1/16), human bocavirus (HBoV; 1/16), and human enterovirus (HEV; 1/16). Three of the 16 samples (183A, 365A, and 723A) were coinfected by two different viruses (Table 3). In addition, human herpesvirus 5, an opportunistic virus that is not directly associated with ALRTIs, was codetected in two samples, whereas one sample (086A) was negative for any known respiratory virus (Table 3).

Conventional/nested PCR assays and real-time RT-PCR analysis. To verify the consistency between traditional and advanced molecular methods, we independently analyzed the 16 samples by conventional PCR assays for common respiratory viruses. These PCR methods have been optimized for routine screening of clinical samples for respiratory viral pathogens (20). The target viruses cover all common respiratory viral agents (see Materials and Methods).

The results of the PCR assays were consistent with those of the metagenomic analysis, including the results for the virus-negative sample and two out of the three samples with copathogens (Table 3). However, in sample 723A, the PCR method only identified HRV. Given that the PCR assay employs consensus primers for the genus *Enterovirus*, to detect HRV and HEV in one gel on the basis of different band sizes only, the overwhelming amount of HRV relative to HEV in the sample may result in biased PCR amplification (Table 3). The HEV in sample 723A was confirmed by a nested PCR test using primers targeting the VP2 and VP3 genes (see Fig. S1 in the supplemental material).

To further investigate the sensitivity of this metagenomic approach, we estimated the number of viral RNA copies in the five HRSV-positive samples by real-time RT-PCR. Sample 306A, which had the lowest HRSV RNA copy number (29 copies in 1 ng of total RNA), had a VTMK index of 2.6 (Table 3). This indicates that the sensitivity of the metagenomic approach was comparable with that of the real-time PCR assay. In addition, the VTMK indices correlated exponentially with the viral RNA copy numbers determined by real-time RT-PCR (Fig. 2).

Reconstruction of genomes of abundantly detected viruses. The power of ultradeep sequencing enabled us to directly recover the genomes of viruses abundantly detected in the metagenomic analysis in each sample. To ensure a high degree of coverage, only viruses with a VTMK index greater than 2 were subjected to genome reassembly. From 10 samples, we retrieved 11 viral genomes, including those of HRSV, HRV, HBoV, HAdV, and IFVB. In general, 21.84% to 98.53% of the viral genomes were covered, depending on the VTMK index and the overall output of the individual sequencing run (see Table S1 in the supplemental material). The genome coverage largely correlated with sequencing depth, which is consistent with the Lander-Waterman theory of shotgun genome sequencing (see Fig. S2 in the supplemental material). Furthermore, the genomic data revealed that all five isolates of HRSV in our samples were closely related, whereas the HRV strains belonged to at least three different subgroups (data not shown).

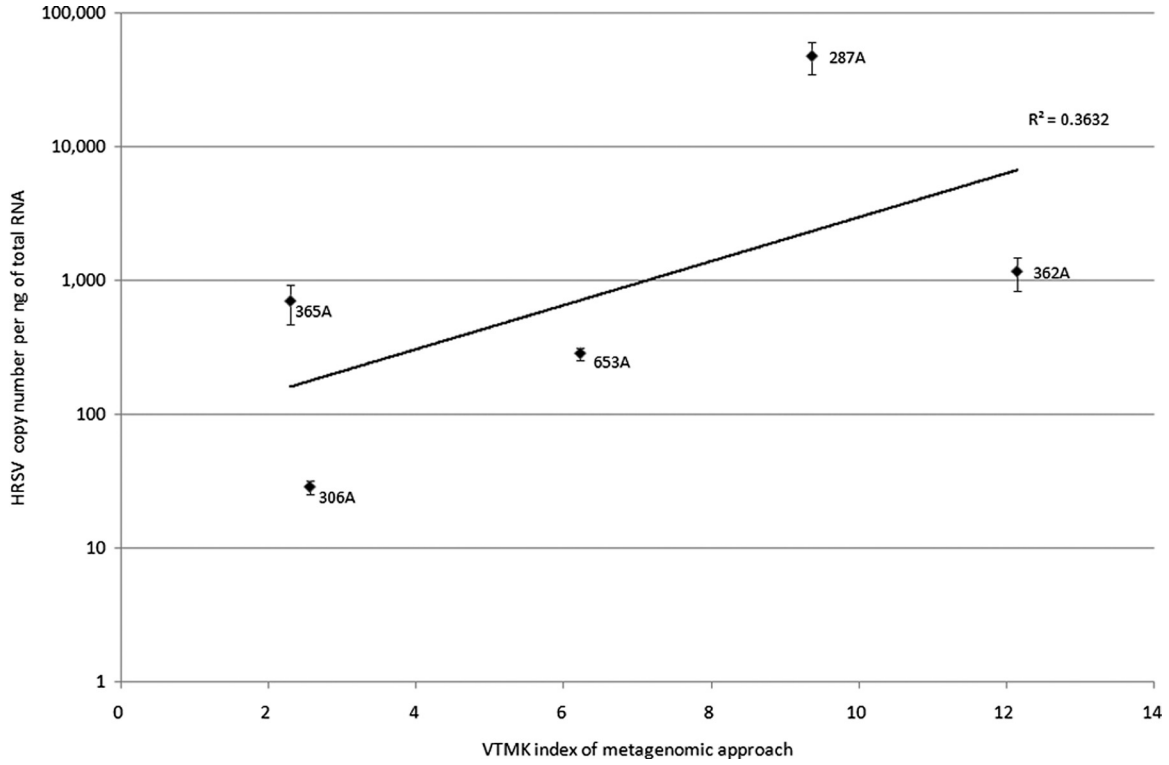

FIG. 2. Relationship between VTMK index of metagenomic approach and HRSV copy number per ng of total RNA extracted from the NPA samples determined by quantitative PCR. The standard deviations of three quantitative PCR repeats of each sample are given, and an exponential regression line is fitted to the data.

DISCUSSION

Application of metagenomics in clinical samples. From 16 total samples, our metagenomic approach identified seven species of known respiratory viral agents in 15 samples, and these results were in agreement with those of conventional PCR methods (Table 3). In addition, the coinfection of HEV with HRV in sample 723A was revealed by the metagenomic approach and later confirmed by the nested PCR test, but it was initially undetected by the conventional PCR assay. Our results represent the comprehensive and unbiased features of the metagenomic approach relative to conventional methods. Furthermore, the ultradeep sequencing enabled us to reconstruct the majority of genomes of the abundantly detected viruses, which will facilitate follow-up phylogenetic analyses and may reveal critical information such as potential antiviral resistances. Our studies indicate that the metagenomic method not only is powerful for virus identification but also is useful in viral genome analysis and further evolutionary study. However, the metagenomic approach may be biased to favor well-known and sequence-rich pathogens, as the results are primarily based on sequence similarities within public databases.

Clinical studies usually require a control set to produce statistically meaningful results. We did not have a control sample in this study because the collection of NPAs from healthy children is very difficult in China. This problem will be encountered by other metagenomic studies that use invasive sampling techniques to obtain specimens. In this study, by using the VTMK index, we not only overcame the intrinsic data variations between samples introduced by ultradeep sequencing but we also removed the majority of the background noise due to sequencing errors or potential contamination. Moreover, the VTMK indices of the metagenomic approach correlated with the number of viral copies in the total RNA extracted from NPA samples. Consequently, detection of different viruses in each sample was straightforward (Table 3), although the cutoff setting of the VTMK index was experimental. Those sequencing reads with VTMK indices under the cutoff are likely spurious due to sequencing errors or potential contaminations, or alternatively, they may be derived from viruses of very low titer that truly exist in the samples, which might be trivial clinically. Therefore, in studies in which control samples are available, statistical methods, such as MetaStat (24), will be helpful for estimating a more reasonable cutoff value by comparing different groups.

In our data set, only approximately 0.05% of the valid metagenomic sequences were useful for virus detection. A reason for this was the large amount (91.6% on average) of host-derived sequences (Fig. 1). Over 90% human-derived sequence contamination has been reported by other researchers when using nasal specimens for metagenomic studies (4, 17). Hence, a high percentage of human-derived sequences is likely an inherent problem when RNA/DNA is directly extracted from respiratory samples. Although the computational subtraction that is currently used is able to efficiently calculate out host contamination, additional sample-filtering steps before sequencing to eliminate host cells and enrich for microbial

genomes are undoubtedly required to detect microbial pathogens in a more efficient and cost-effective manner.

Previous metagenomic studies on viromes have employed a combination of filtration and density-dependent centrifugation to enrich for the majority of (but not all) viruses (1–3). Greninger et al. suggested using postextraction treatment with DNase (4), but this approach may introduce a bias and decrease the sensitivity for DNA viruses. Therefore, in metagenomic studies aimed at identifying comprehensive virome profiles, pretreatments should be kept to a minimum to avoid any factitious bias. Additionally, any efficient filtering steps will significantly decrease the quantity of valid DNA/RNA produced, which may hamper direct sequencing using the current NGS platforms. For example, from each 0.5-ml NPA sample we obtained a maximum of only approximately 1 μg of purified DNA/cDNA, and over 90% of it was human derived. Such a small amount of DNA is not sufficient for sequencing by the current Roche 454 GS FLX sequencer, which requires approximately 10 μg for initial library generation, and it just meets the minimal requirement for sequencing by the Illumina Solexa GA II sequencer. Therefore, we did not introduce any viral purification processes into sample preparation in this study. Recently proposed real-time PCR (15) or digital PCR (25) assays might be helpful in future studies to enable deep sequencing with only nanogram or even picogram amounts of sample DNA.

NGS platforms for clinical metagenomic samples. The Roche 454 and Illumina Solexa sequencers represent two different types of NGS technologies. The former generates longer reads (~400 bp) but less overall output (200 to 400 Mb) in one run; the latter produces a large number of shorter reads (36 to 150 bp), leading to dozens of Gb of sequences in a single run. In a recent study using the Roche 454 technology, Nakamura and coworkers identified an average of 0.76% virus-associated sequences from three NPA samples (17), which is a much higher percentage than we found in our results (0.05%). This might reflect the disadvantage of short reads (36 bp) in sequence similarity-based species classification. For maintenance of significant nucleotide sequence similarity, short sequences allow fewer variant sites (due to either mutations or sequencing errors) than longer sequences. This led to the 6.8% unknown sequences in our data (Fig. 1). Nevertheless, Nakamura et al. used the best-hit method instead of the LCA algorithm for species classification (16), so their results tended to have more reads assigned at the species level than our results. For reads with multiple matches in different species, the LCA algorithm traces backward in the taxonomic tree to assign the reads to the lowest common ancestor of these species, whereas the best-hit method directly assigns each read to the species of its most similar sequence in the database, ignoring other matches. Generally, in clinical virome studies, it is estimated that long-read-based technologies utilize over a 10-fold greater percentage of sequencing data than short-read-based platforms. In addition, studies have shown that long sequences enable the translation of nucleotide sequences to possible peptides, which is essential for identification of potential novel viruses or variants (5, 18, 22). These disadvantages may be the most critical factors that have hampered the broad application of short-read-based NGS technologies in previous metagenomic studies.

However, the throughput of currently available short-read-based NGS technologies, such as the Illumina Solexa and ABI SOLiD technologies, is usually more than 10-fold greater than that of the Roche 454 sequencer, which should largely compensate for one of the disadvantages described above. Therefore, the overall size of useful sequences in clinical virome analyses should be comparable between both types of NGS technologies. For example, Nakamura et al. covered 8.10% to 58.30% of the influenza A virus genome using Roche 454 data (17), whereas in this study, we reconstructed 21.84% of the influenza B virus genome in sample 593A (data not shown). In addition, using 67-bp reads, Greninger et al. successfully reconstructed over 90% of the influenza A (H1N1) virus genome by *de novo* assembly (4), indicating the power of ultradeep sequencing in metagenomic studies.

An inherent property of some clinical specimens is the limited amount of original sample available from each individual. For example, in this study, we had access to only 0.5 ml of each NPA sample for metagenomic analysis, which produced approximately 1 μg of purified DNA/cDNA without any filtering process. Although a random amplification step may be a solution as described previously (17), it may introduce artificial bias. Therefore, as mentioned above, the Illumina Solexa platform is capable of initiating deep sequencing with a relatively small amount of DNA (approximately 1 μg), which makes it attractive when working with small volumes of clinical samples for metagenomic analysis.

It is important to note that the above discussion of the advantages and disadvantages of NGS platforms is based on current methodology. However, all currently available NGS technologies are in rapid development. For example, just after we completed the sequencing work in this study, Illumina upgraded its sequencing kit to enable reads of over 100 bp. This will undoubtedly overcome the above-mentioned weakness in short-read-based metagenomic analyses and should therefore be applied in further metagenomic studies on clinical samples. Alternatively, as proposed recently, paired-end sequencing strategies may be used to facilitate *de novo* assembly (4) or produce longer sequences *in silico* (21). In addition, Roche recently released improved protocols to allow a minimum of 0.5 μg of initial DNA for sequencing on the 454 sequencer, which makes the Roche 454 platform competitive in clinical metagenomic studies. Therefore, due to these advances, both long-read-based (e.g., Roche 454) and short-read-based (e.g., Illumina Solexa) metagenomics may be used to analyze small volumes of clinical samples. The choice of technology in future studies might be multifactorial, including cost performance, run time, accessibility, and convenience.

ACKNOWLEDGMENTS

This work was supported by the National Major Science and Technology Project for Prevention and Treatment of AIDS and Viral Hepatitis and Other Major Infectious Diseases of China grants 2009ZX10004-102 and 2009ZX10004-206; intramural grant 2009IPB112 from the Institute of Pathogen Biology, CAMS; and Beijing Nova Program grant 2009A67 (to J.Y.).

We are grateful to Wei Wang and Lan Qin for their technical assistance and Chaochun Wei (Shanghai Center for Bioinformation Technology, Shanghai, China), Xiaopeng Zhu (Bioinformatics Laboratory, Institute of Biophysics, CAS, Beijing, China), and Guoguang Zhao and Yi Zhao (Bioinformatics Research Group, Institute of Com-

puting Technology, CAS, Beijing, China) for inspirational discussion and bioinformatics assistance in analysis of metagenomic data.

REFERENCES

1. **Angly, F. E., et al.** 2006. The marine viromes of four oceanic regions. PLoS Biol. **4:**e368.
2. **Breitbart, M., et al.** 2002. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. U. S. A. **99:**14250–14255.
3. **Dinsdale, E. A., et al.** 2008. Functional metagenomic profiling of nine biomes. Nature **452:**629–632.
4. **Greninger, A. L., et al.** 2010. A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. PLoS One **5:**e13381.
5. **Greninger, A. L., et al.** 2009. The complete genome of klassevirus—a novel picornavirus in pediatric stool. Virol. J. **6:**82.
6. **Huson, D. H., A. F. Auch, J. Qi, and S. C. Schuster.** 2007. MEGAN analysis of metagenomic data. Genome Res. **17:**377–386.
7. **Jokela, P., H. Piiparinen, K. Luiro, and M. Lappalainen.** 2010. Detection of human metapneumovirus and respiratory syncytial virus by duplex real-time RT-PCR assay in comparison with direct fluorescent assay. Clin. Microbiol. Infect. **16:**1568–1573.
8. **Kapoor, A., et al.** 2010. Human bocaviruses are highly diverse, dispersed, recombination prone, and prevalent in enteric infections. J. Infect. Dis. **201:**1633–1643.
9. **Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg.** 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10:**R25.
10. **Lee, W. M., et al.** 2007. High-throughput, sensitive, and accurate multiplex PCR-microsphere flow cytometry system for large-scale comprehensive detection of respiratory viruses. J. Clin. Microbiol. **45:**2626–2634.
11. **Li, H., et al.** 2007. Simultaneous detection and high-throughput identification of a panel of RNA viruses causing respiratory tract infections. J. Clin. Microbiol. **45:**2105–2109.
12. **Li, H., J. Ruan, and R. Durbin.** 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. **18:**1851–1858.
13. **Mathers, C., T. Boerma, and D. M. Fat.** 2008. The global burden of disease: 2004 update. World Health Organization, Geneva, Switzerland.
14. **Metzker, M. L.** 2010. Sequencing technologies—the next generation. Nat. Rev. Genet. **11:**31–46.
15. **Meyer, M., et al.** 2008. From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing. Nucleic Acids Res. **36:**e5.
16. **Nakamura, S., et al.** 2008. Metagenomic diagnosis of bacterial infections. Emerg. Infect. Dis. **14:**1784–1786.
17. **Nakamura, S., et al.** 2009. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. PLoS One **4:**e4219.
18. **Palacios, G., et al.** 2008. A new arenavirus in a cluster of fatal transplant-associated diseases. N. Engl. J. Med. **358:**991–998.
19. **Qin, J., et al.** 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature **464:**59–65.
20. **Ren, L., et al.** 2009. Prevalence of human respiratory viruses in adults with acute respiratory tract infections in Beijing, 2005-2007. Clin. Microbiol. Infect. **15:**1146–1153.
21. **Rodrigue, S., et al.** 2010. Unlocking short read sequencing for metagenomics. PLoS One **5:**e11840.
22. **Victoria, J. G., et al.** 2009. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. J. Virol. **83:**4642–4651.
23. **Wang, D., et al.** 2002. Microarray-based detection and genotyping of viral pathogens. Proc. Natl. Acad. Sci. U. S. A. **99:**15687–15692.
24. **White, J. R., N. Nagarajan, and M. Pop.** 2009. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput. Biol. **5:**e1000352.
25. **White, R. A., III, P. C. Blainey, H. C. Fan, and S. R. Quake.** 2009. Digital PCR provides sensitive and absolute calibration for high throughput sequencing. BMC Genomics **10:**116.
26. **Willner, D., et al.** 2009. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. PLoS One **4:**e7370.
27. **Yongfeng, H., et al.** 2011. Direct pathogen detection from swab samples using a new high-throughput sequencing technology. Clin. Microbiol. Infect. **17:** 241–244.