AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology

Unbiased Detection of Respiratory Viruses by Use of RNA Sequencing-Based Metagenomics: a Systematic Comparison to a Commercial PCR Panel

Erin H. Graf,[a]* Keith E. Simmon,[b,c] Keith D. Tardif,[c] Weston Hymas,[c] Steven Flygare,[d] Karen Eilbeck,[b,e] Mark Yandell,[d,e] Robert Schlaberg[a,c]

University of Utah School of Medicine, Department of Pathology, Salt Lake City, Utah, USA[a]; University of Utah School of Medicine, Department of Biomedical Informatics, Salt Lake City, Utah, USA[b]; ARUP Institute for Clinical and Experimental Pathology, Salt Lake City, Utah, USA[c]; Department of Human Genetics, University of Utah, Salt Lake City, Utah, USA[d]; USTAR Center for Genetic Discovery, Salt Lake City, Utah, USA[e]

Current infectious disease molecular tests are largely pathogen specific, requiring test selection based on the patient's symptoms. For many syndromes caused by a large number of viral, bacterial, or fungal pathogens, such as respiratory tract infections, this necessitates large panels of tests and has limited yield. In contrast, next-generation sequencing-based metagenomics can be used for unbiased detection of any expected or unexpected pathogen. However, barriers for its diagnostic implementation include incomplete understanding of analytical performance and complexity of sequence data analysis. We compared detection of known respiratory virus-positive ($n = 42$) and unselected ($n = 67$) pediatric nasopharyngeal swabs using an RNA sequencing (RNA-seq)-based metagenomics approach and Taxonomer, an ultrarapid, interactive, web-based metagenomics data analysis tool, with an FDA-cleared respiratory virus panel (RVP; GenMark eSensor). Untargeted metagenomics detected 86% of known respiratory virus infections, and additional PCR testing confirmed RVP results for only 2 (33%) of the discordant samples. In unselected samples, untargeted metagenomics had excellent agreement with the RVP (93%). In addition, untargeted metagenomics detected an additional 12 viruses that were either not targeted by the RVP or missed due to highly divergent genome sequences. Normalized viral read counts for untargeted metagenomics correlated with viral burden determined by quantitative PCR and showed high intrarun and interrun reproducibility. Partial or full-length viral genome sequences were generated in 86% of RNA-seq-positive samples, allowing assessment of antiviral resistance, strain-level typing, and phylogenetic relatedness. Overall, untargeted metagenomics had high agreement with a sensitive RVP, detected viruses not targeted by the RVP, and yielded epidemiologically and clinically valuable sequence information.

Laboratory diagnosis of infectious diseases has historically taken a syndrome-based approach. Culture of appropriate specimens on a combination of relevant media or cell lines enables detection of certain common bacterial, viral, and fungal pathogens. However, culture requires experienced personnel, requires several days to weeks to yield a definitive answer, depends on viability and appropriate culture conditions, and has limited sensitivity. Molecular tests have superior turnaround times, sensitivity, and taxonomic resolution. However, only targeted pathogens can be detected, and differentiation of clinically or epidemiologically relevant strains or genotypes is limited. Moreover, molecular tests need to be updated when new species or strains are recognized to ensure that newly identified genetic variants can be detected.

In contrast, next-generation sequencing-based metagenomic testing combines and extends many advantages of molecular tests and culture-based methods. Host- and pathogen-derived nucleic acids are sequenced without *a priori* knowledge of expected pathogens, allowing simultaneous detection of a virtually unlimited number of microorganisms, the only requirement being that they possess sequence homology with reference sequences.

Metagenomics-based pathogen detection is especially powerful when many diverse pathogens cause overlapping symptoms and when molecular markers for drug resistance are known. One such application is the detection of respiratory pathogens. Even with state-of-the-art, multiplex molecular tests, identifying the etiology of respiratory tract infections is often unsuccessful; e.g., respiratory pathogens are detected in only ~40 to 80% of patients with community-acquired pneumonia (CAP) using standard testing approaches (1–5). In addition, respiratory viruses of unclear pathogenicity (e.g., rhinovirus) are often found as the sole pathogen in many respiratory samples. These facts suggest that the true etiology (6–9) of many cases remains unknown. In these scenarios, metagenomics-based detection methods have great diagnostic potential as alternative causes can be identified or excluded with greater confidence compared to panel-based approaches. Moreover, metagenomics-based testing enables genotyping, assessment

Received 18 November 2015 **Returned for modification** 14 December 2015
Accepted 20 January 2016

Accepted manuscript posted online 27 January 2016

Citation Graf EH, Simmon KE, Tardif KD, Hymas W, Flygare S, Eilbeck K, Yandell M, Schlaberg R. 2016. Unbiased detection of respiratory viruses by use of RNA sequencing-based metagenomics: a systematic comparison to a commercial PCR panel. J Clin Microbiol 54:1000–1007. doi:10.1128/JCM.03060-15.

Editor: A. M. Caliendo

Address correspondence to Robert Schlaberg, robert.schlaberg@path.utah.edu.

* Present address: Erin H. Graf, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA, and Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/JCM.03060-15.

of molecular markers for drug resistance, and molecular epidemiologic studies.

While several recent studies have demonstrated the power of next-generation sequencing-based metagenomics for pathogen detection (10–18), its performance compared to that of commercially available molecular tests is incompletely understood. Equally important, it remains to be demonstrated whether metagenomics approaches can be implemented in diagnostic laboratories and employed within a clinically meaningful time frame using computational resources and data analysis expertise available in diagnostic laboratories. Complexities of laboratory workflow, speed of sequence analysis, and expertise required for analysis and interpretation are chief concerns.

We evaluated the analytical performance of metagenomics for detection of respiratory viruses using kit-based RNA sequencing (RNA-seq) analysis of total RNA extracted from pediatric nasopharyngeal (NP) swabs. Resulting sequence data were analyzed with a rapid, interactive, web-based data analysis tool, Taxonomer, eliminating the need for expensive computational hardware and bioinformatics expertise (S. Flygare, K. E. Simmon, C. Miller, Y. Qiao, B. Kennedy, T. Di Sera, E. H. Graf, K. D. Tardif, A. Kapusta, S. Rynearson, C. Stockmann, K. Queen, S. Tong, K. V. Voelkerding, A. Blaschke, C. L. Byington, S. Jain, A. Pavia, K. Ampofo, K. Eilbeck, G. Marth, M. Yandell, R. Schlaberg, submitted for publication). We compared results to those of an FDA-cleared, multiplex PCR-based test, the GenMark eSensor respiratory virus panel (RVP). Overall RNA-seq-based pathogen detection was highly concordant with the GenMark eSensor RVP, detected additional viruses not targeted by the RVP, and yielded epidemiologically and clinically valuable sequence information.

MATERIALS AND METHODS

Samples. All nasopharyngeal (NP) swabs from children less than 5 years old tested by the GenMark eSensor respiratory virus panel (RVP, GenMark Dx, Carlsbad, CA) between April 2013 and March 2014 were de-identified using standard institutional procedures (University of Utah IRB number 56504) and stored at −80°C. Specimens positive for RNA viruses tested by the GenMark assay were retrospectively collected, with preference given to dual infections (human metapneumovirus [HMPV], $n = 5$; human rhinovirus [HRV], $n = 10$; influenza A virus, $n = 5$; influenza B virus, $n = 5$; parainfluenza 1 virus [PIV-1], $n = 5$; PIV-2, $n = 1$; PIV-3 virus, $n = 4$; respiratory syncytial virus [RSV], $n = 7$ [see Table S1 in the supplemental material]). In addition, 67 samples were selected at random for inclusion in a direct side-by-side comparison.

GenMark eSensor respiratory virus panel. Nucleic acid was extracted from 200 μl of sample, plus 10 μl of internal control, on the NucliSENS easyMAG (bioMérieux, Durham, NC) and eluted into 60 μl, 5 μl of which was reverse transcribed and amplified with the eSensor respiratory virus panel reagents by following the manufacturer's instructions (GenMark). The following 14 viral targets are reported in the eSensor XT-8 system (GenMark): adenovirus B/E, adenovirus C, influenza A virus, influenza A H1 virus, influenza A H3 virus, influenza A 2009 H1N1 virus, influenza B virus, RSV subtype A (RSV-A), RSV-B, PIV-1, PIV-2, PIV-3, HMPV, and HRV.

Library preparation and RNA sequencing. NP swabs were thawed and vortexed, and 160 μl of the transport medium was transferred for extraction with the QIAamp viral RNA minikit by following the manufacturer's instructions (Qiagen, Valencia, CA). Eluted RNA was vacuum dried and stored at −80°C overnight. RNA-seq libraries were prepared with the TruSeq RNA sample prep kit by following the manufacturer's instructions (Illumina, San Diego, CA). Libraries were quantified with the Illumina universal library quantification kit (Kapa Biosystems, Inc., Wilmington, MA). Library quality was assessed with a high-sensitivity DNA analysis kit on a 2100 bioanalyzer (Agilent Technologies, Santa Clara, CA). Libraries from 24 samples were combined in equimolar ratios for a final concentration of 9.6 nM and sequenced in batches of 24 samples per lane on a HiSeq 2500 instrument (Illumina, San Diego, CA).

Data analysis. RNA-seq data were analyzed with Taxonomer, a kmer-based, rapid, interactive metagenomic sequence analysis tool accessed through a web interface on the iobio framework at http://taxonomer.iobio.io (19; Flygare et al., submitted). Taxonomer classifies each read to the highest taxonomic rank possible given a comprehensive sequence database (Fig. 1A and B). For the purposes of this study and to validate Taxonomer results, relevant human pathogens detected by Taxonomer were manually confirmed using Geneious (Biomatters, Ltd., Auckland, New Zealand) by mapping reads against a manually curated list of full-length viral reference sequences downloaded from the NCBI. Instead of using an absolute or relative read count threshold for respiratory virus detection, all positive results were manually reviewed to ensure the accuracy of viral classifications and to reduce false-positive calls. To minimize errors due to sample-to-sample contamination or demultiplexing errors, we generated consensus sequences for all viral strains with sufficient coverage. For all virus-positive samples with low read counts, sequencing reads were aligned to the viral consensus sequence of samples with high read counts processed in the same batch and sequenced on the same HiSeq lane. When sequencing reads from samples with low read counts showed no polymorphisms compared to the consensus sequence from samples with high read counts, results were excluded as likely contaminants. Sequence-based typing of viral strains was performed by manual alignment to a reference genome and BLAST analysis of the largest contig of the appropriate genomic segment (e.g., VP1/3 for rhinovirus) or whole viral genome, if possible. Strains were considered typed if the references with the highest sequence identity over the entire contig all belonged to the same genotypes (e.g., RSV-B was the highest match with no RSV-A at the same percent identity).

qPCR for respiratory viruses. Total nucleic acid was extracted on the Chemagic MSM I platform (PerkinElmer) using 200 μl of the NP swab transport medium. Nucleic acid was eluted into 80 μl, and 10 μl of the elution was used for amplification on an ABI 7900 instrument (Life Technologies, Foster City, CA). Quantitative PCR (qPCR) assays for human metapneumovirus, respiratory syncytial virus, influenza A and B viruses, parainfluenza virus types 1 to 4, and enterovirus validated for diagnostic testing at ARUP Laboratories were used for comparison (see Table S2 in the supplemental material; all primers and probes were obtained from EliTech) (20, 21). For human rhinovirus and coronavirus species (HKU1, NL63, and OC43), research tests were used (22). qPCR detection of enterovirus, HMPV, HRV, influenza A and B viruses, and RSV was performed using QuantiTect reverse transcription-PCR (RT-PCR) master mix (Qiagen) on an ABI 7900 real-time PCR system (Applied Biosystems) with the following amplification conditions: 50°C for 30 min, 95°C for 15 min, and 50 cycles of 95°C for 15 s, 56°C for 30 s, and 76°C for 30 s. qPCR detection of parainfluenza viruses was performed using the Rotor-Gene multiplex RT-PCR master mix (Qiagen) on a Rotor-Gene Q (Qiagen) with the following amplification conditions: 50°C for 15 min, 95°C for 5 min, and 50 cycles of 95°C for 5 s, 56°C for 20 s, and 76°C for 20 s. qPCR detection of human coronaviruses was performed using the AgPath-ID One-Step RT-PCR kit (ThermoFisher) on an ABI 7900 real-time PCR system with the following amplification conditions: 45°C for 10 min, 95°C for 10 min, and 45 cycles of 95°C for 15 s and 55°C for 1 min. RNA standards specific for each virus were used to generate standard curves. qPCR was performed for each of the viruses detected by the RVP on the respective samples.

Statistics. Linear and Spearman's rank correlations were performed with Prism (version 5.04, GraphPad Software, Inc., La Jolla, CA), and P values less than 0.05 were considered significant.

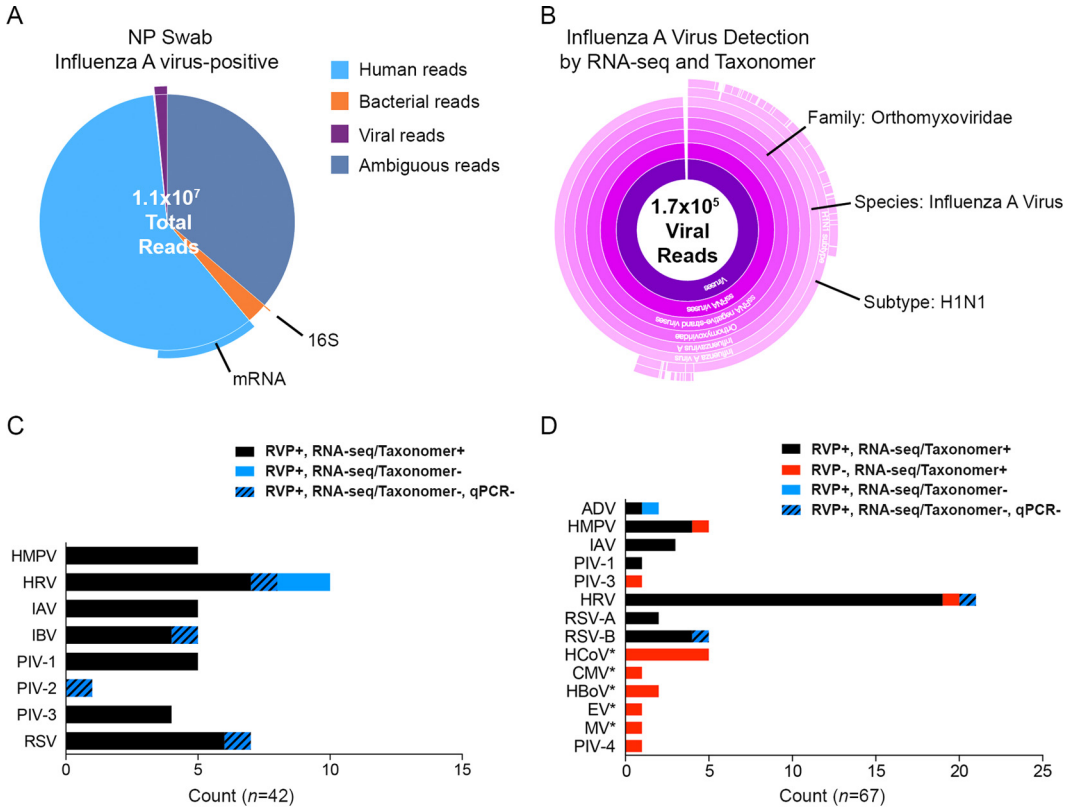

FIG 1 Respiratory virus detection by untargeted metagenomics and RVP. (A) Fractional abundance of human, bacterial, and viral sequences in untargeted metagenomics data from an influenza A virus-positive NP swab of a female infant determined by Taxonomer (19; Flygare et al., submitted). Approximately 1.5% of reads were of viral and 2.7% of bacterial origin. (B) Of 1.74×10^5 viral reads, 1.73×10^5 (99.4%) could be classified to the species level (influenza A virus) and 7.9×10^4 (45.3%) to the H1N1 subtype. (C) Untargeted metagenomics identified 36 of 42 (86%) respiratory viruses detected by the RVP. Four of the 6 viruses missed by untargeted metagenomics (blue bars) could not be detected by qPCR (hashed bars), resulting in detection of 36 of 38 viruses (95%) that were consistently detected by targeted methods. (D) Untargeted metagenomics detected more viral infections ($n = 48$, including 11 not targeted by the RVP [asterisk]) than the RVP ($n = 37$) in unselected NP swabs ($n = 67$) collected during a 12-month period. ADV, adenovirus; HMPV, human metapneumovirus; IAV, influenza A virus; PIV, parainfluenza virus; HRV, human rhinovirus; RSV, respiratory syncytial virus; HCoV, human coronavirus; CMV, cytomegalovirus; HBoV, human bocavirus; EV, enterovirus; MV, measles virus.

Nucleotide sequence accession number. The complete genome sequence of HRV-C strain UT-I was deposited in GenBank under accession no. KU695562.

RESULTS

Metagenomic analysis of RVP-positive samples. To assess the accuracy of untargeted metagenomics, archived nasopharyngeal swabs ($n = 42$) positive for one or more viruses by the RVP were retrospectively selected. Preference was given to samples with co-detection of >1 virus. Agreement between RNA viruses detected by the RVP and untargeted metagenomics was 86% (Fig. 1C). Six respiratory viruses detected by the RVP were not detected by untargeted metagenomics (blue bars). Four of these (one each of rhinovirus, influenza B virus, parainfluenza type 2, and respiratory syncytial virus) were also not detected by qPCR (hatched blue bars). Considering these by RVP, adjusted positive agreement between the RVP and metagenomics was 95%. Both of the remaining RVP-positive, untargeted metagenomics-negative samples were low positive for rhinovirus, with qPCR threshold cycle (C_T) values of 33 and 35.

Untargeted metagenomic and RVP analysis of unselected samples. Between April 2013 and March 2014, all NP swabs from children less than 5 years of age submitted for the RVP were banked after testing. From these, 67 were selected at random for testing by untargeted metagenomics. Of these, 36 samples (55.2%) were positive by RVP for one or more respiratory viruses (adenovirus, $n = 2$; HMPV, $n = 4$; influenza A virus, $n = 3$; PIV-1, $n = 1$; HRV, $n = 20$; and RSV, $n = 7$). Of the 37 swabs for which respiratory virus was detected, 34 (91.9%) also had virus detected by untargeted metagenomics (Fig. 1D). For two of the remaining three samples, there was sufficient sample to attempt qPCR confirmation of RVP results. In both cases, qPCR results agreed with those of untargeted metagenomics. The overall positivity rate for untargeted metagenomics was 63%. Untargeted metagenomics detected 12 additional respiratory viruses, 3 of which were targeted by the RVP. Seven of the 12 additional viruses (58.3%) were confirmed by qPCR, 3 were qPCR negative, and 2 could not be tested due to due to limited sample volumes. The 3 viruses detected by untargeted metagenomics but not qPCR had low viral read counts (4 to 31 reads). However, these reads were located across unique regions of viral genomes and sequences differed by several nucleotides from those of other viral strains detected at higher read counts within the batch, suggesting that they were not misclassifications and unlikely to be contaminants. Rhinovirus was the pathogen most fre-

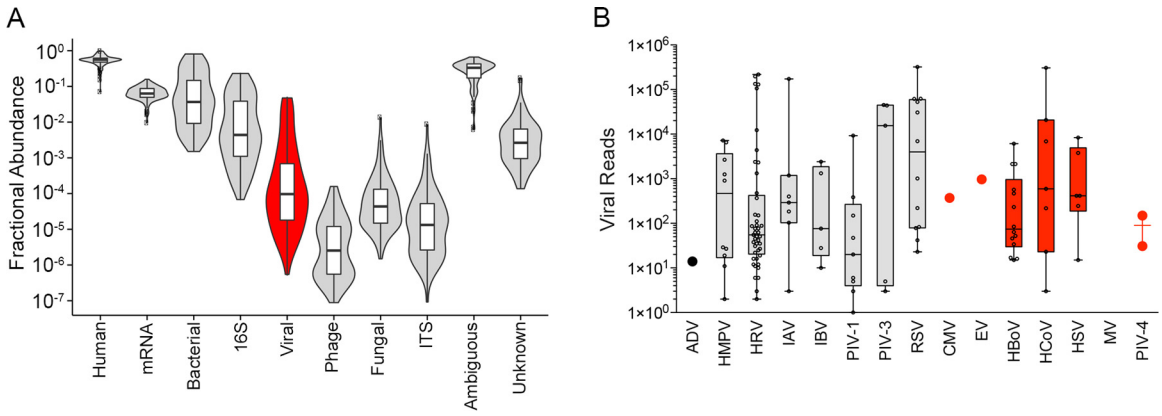

FIG 2 Overall taxonomic composition of RNA-seq reads and numbers of viral reads by respiratory virus. (A) Fractional abundance of reads binned as human, human mRNA (referred to as "mRNA"), any bacterial sequence (referred to as "bacterial"), bacterial 16S only (referred to as "16S"), viral, phage, any fungal sequence (referred to as "fungal"), fungal ITS only (referred to as "ITS"), ambiguous, and unknown is shown as median and interquartile range (box plots) and as violin plots. Only reads identified as viral (red, median ~1:10^{-4} reads) were used for this analysis. (B) Viral read counts differed across 5 orders of magnitude. Viruses not targeted by the RVP are shown in red. ADV, adenovirus; HMPV, human metapneumovirus; IAV, influenza A virus; PIV, parainfluenza virus; HRV, human rhinovirus; RSV, respiratory syncytial virus; HCoV, human coronavirus; CMV, cytomegalovirus; HBoV, human bocavirus; EV, enterovirus; MV, measles virus; ITS, internal transcribed spacer.

quently detected by either method ($n = 22$); RSV was second ($n = 6$), followed by coronavirus (not included on the RVP; $n = 5$) and HMPV ($n = 5$). Manual confirmation of Taxonomer results showed 100% qualitative agreement for detection of respiratory viruses.

Codetection of ≥2 respiratory viruses was more common with untargeted metagenomics (14%) than with the RVP (3%). The majority of codetections involved rhinoviruses and human bocavirus (HBoV). When possible, these codetections were confirmed with qPCR. There were several samples that were positive by both the RVP and untargeted metagenomics but negative by qPCR, suggesting that untargeted metagenomics has a sensitivity at least comparable to those of the RVP and qPCR.

Correlation of viral read counts with viral loads determined by qPCR. In some studies, semiquantitative detection of respiratory viruses has been shown to correlate with disease severity (23–25). To assess the use of viral read counts to estimate viral burden, we compared viral loads by qPCR with normalized viral reads counts using the normalization scheme described in reference 26. Briefly, numbers of viral reads in 68 positive samples were divided by the number of total reads and the size of the respective viral genome in kilobases and then multiplied by 1 million to generate an RPKM (reads per kilobase of reference sequence per million total sequencing reads) value. Our untargeted metagenomics approach generated a median of 15 million reads per sample (interquartile range [IQR], 8 to 19 million), of which a median of 0.01% was of viral origin (IQR, 0.002 to 0.07%) (Fig. 2A). The number of viral reads spanned >5 orders of magnitude (2×10^5 to 3.2×10^5 reads) in RVP- and monoplex qPCR-positive samples (Fig. 2B). Correlation of RPKM and viral copies per milliliter was highly significant, with a P value of <0.0001 across viral taxa (Fig. 3A). This suggests that normalized viral read counts can be used for semiquantitative measurement of the viral burden in clinical samples.

Reproducibility. Three samples at different positivity levels by untargeted metagenomics (fractional abundance of viral reads) were selected to evaluate within-run and between-run variability. Two of the viruses were also detected by the RVP; the third one was a sample positive by qPCR for coronavirus. Each sample was processed from start to finish (extraction to analysis) a total of 5 (HRV and HMPV) or 14 (HCoV) times. Libraries were sequenced on the same (within run) and on different (between-run) HiSeq lanes (Fig. 3B). Fractional abundance (viral reads as a proportion of total reads) is graphed for each repeat, and the coefficient of variation was calculated from these values. Given the complexity of the workflow, untargeted metagenomics demonstrated excellent reproducibility, with coefficients of variation of 65% (HMPV, lowest fractional abundance), 16% (HCoV), and 47% (HRV, greatest fractional abundance).

Sequence-based characterization of viral strains and antiviral drug resistance determination. As metagenomics provides sequence information in addition to mere determination of presence or absence of pathogens, we studied available viral sequences to demonstrate utility. Even though viral reads were a very small proportion of the total reads, sufficient sequence was obtained for 84% of positive specimens to enable high-resolution, sequence-based genotyping. Consistent with the RVP results, all of the influenza A virus-positive specimens were typed as 2009 H1N1 strains. By untargeted metagenomics, we were able to examine the oseltamivir resistance mutation at amino acid position 275 (H275Y) of the neuraminidase gene in 6 out of 8 positive specimens (1.6- to 200-fold median coverage), none of which showed the H275Y amino acid substitution. Sequence coverage in the remaining 2 specimens was too low. RSV-B was far more common than RSV-A (9 versus 3 of 12 RSV-positive samples). These results were consistent with RVP-based typing. Most rhinoviruses belonged to rhinovirus species C (62%), with only 21% belonging to rhinovirus species A and 3% to rhinovirus B (Fig. 4). Fourteen percent of rhinoviruses were nontypeable. For 14 (52%) of the rhinovirus-positive samples, coverage of the viral genome was sufficient to generate full-length viral consensus sequences. Genetic diversity was greatest for strains that belonged to rhinovirus species C. Strains B and N, which clustered closely together, were collected during the same month from patients from the same state. The most divergent sample from any full-length sequence in the NCBI nucleotide database was sample I, which had only 75%

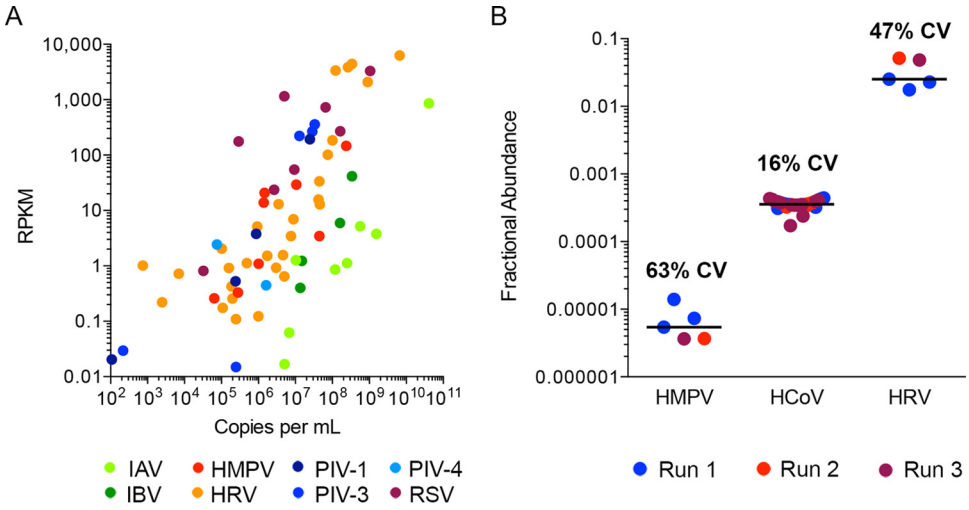

FIG 3 Correlation of normalized read counts with viral burden and precision of viral read abundance within and between sequencing runs. (A) The correlation between viral copies per milliliter of viral transport medium determined by qPCR and normalized viral reads (viral reads per kilobase of viral genome size per million total reads [RPKM]) detected by untargeted metagenomics was assessed by a Spearman correlation test (rho = 0.7 $P < 0.0001$). (B) Reproducibility was evaluated by extracting and sequencing the same sample 5 (human rhinovirus [HRV] and human metapneumovirus [HMPV]) or 14 (human coronavirus [HCoV]) times. Replicate libraries were prepared independently and sequenced on the same lane (within run) or different lanes (between runs). Fractional abundance (viral reads per total reads) is shown for within-run replicates (same color) and between-run replicates (different colors). Precision is shown as percent coefficient of variation (CV).

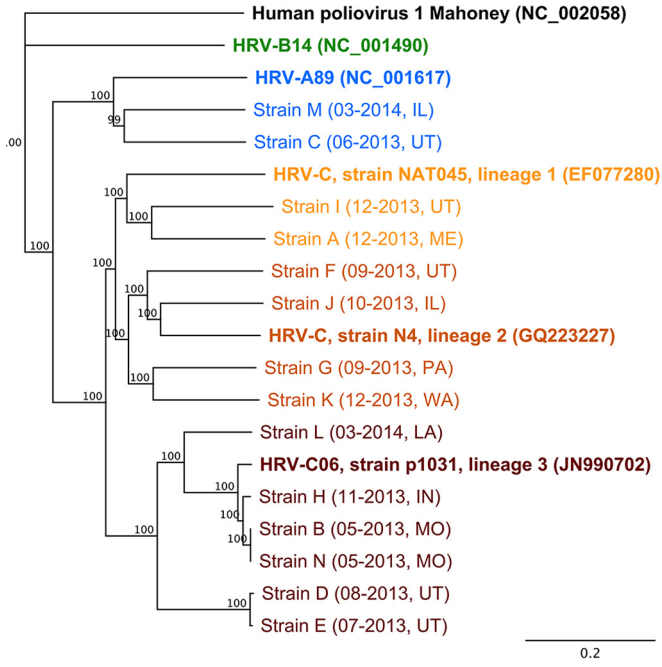

FIG 4 High-resolution, sequence-based typing of 14 human rhinovirus strains based on RNA-seq directly from NP swabs. Most strains belonged to rhinovirus species C ($n = 12$; 86%), with 2 strains (14%) belonging to lineage 1, 4 strains (29%) belonging to lineage 2, and 6 strains (43%) belonging to lineage 3; 2 strains belonged to rhinovirus species A, and no rhinovirus species B strains were detected. Near full-length sequences of 14 human rhinovirus strains (strains A through N) were aligned (MUSCLE); a neighbor-joining consensus tree (1,000 replicates) is shown. Full-length reference sequences for rhinovirus A (HRV-A89), rhinovirus B (HRV-B14), and representative full-length genome sequences from each of the rhinovirus C lineages (GenBank accession numbers EF077280, GQ223227, and JN990702 [40]) were included for comparison. Poliovirus 1 was used as the outgroup. For strains sequenced as part of this study, month, year, and state of sample collection are indicated in parentheses. Colors represent species- and lineage-level clades.

sequence identity with the closest match, HRV-C3 (strain HRV-QPM; GenBank accession no. GQ223228). This sample was missed by the RVP but tested positive by the monoplex qPCR, with a threshold cycle value of 20. The one enterovirus sequence was most similar to coxsackievirus B4 strain E2 (NCBI accession number AF311939; 84% overall nucleotide identity). The alpha coronavirus NL63 was detected in 3 samples, and beta coronaviruses HKU1 and OC43 were detected in 2 samples each. All human bocaviruses detected ($n = 4$) belonged to genotype 1.

Detection of RNA from DNA viruses. Untargeted metagenomics was able to detect only 1 of 2 adenovirus-positive samples (Fig. 1B). Only a very few adenovirus reads were generated in the 2 untargeted metagenomics-positive samples (Fig. 2B). However, human bocavirus RNA was detected at high read counts in four samples (Fig. 2B). Additionally, high levels of RNA reads from a number of nonrespiratory DNA viruses were detected by metagenomics, including herpes simplex virus 1 (HSV-1), cytomegalovirus (CMV), Epstein-Barr virus (EBV), and anellovirus (data not shown). Optimized nucleic acid extraction methods or simultaneous preparation of cDNA and DNA libraries may enable more complete characterization of the DNA virome in clinical samples.

Reagent contamination. Contamination from reagents employed during extraction, library preparation, and sequencing has been previously described (27). To assess the contamination generated by our approach, we extracted and sequenced 3 molecular-grade water samples alongside clinical samples. The reads generated by these samples were largely bacterial. No respiratory viruses or known human-pathogenic viruses were detected (data not shown).

DISCUSSION

We showed that untargeted metagenomic analysis can attain accuracies and sensitivities that compare favorably with those of a commercial RVP, even though different extraction protocols were

used. The unbiased nature of RNA-seq allowed us to query a theoretically unlimited number of pathogens in parallel, resulting in detection of more human viruses and a higher positivity rate. These included well-known respiratory viruses with clinical relevance when detected in the upper respiratory tract, as well as potential pathogens that may only be relevant in the appropriate (e.g., immunocompromised) host and when detected from the lower respiratory tract (e.g., HSV and CMV). Interestingly, even though we used RNA-seq and included a DNase treatment step, DNA viruses were detected in some, but not all, RVP-positive samples. It is possible that detection of mRNA from DNA viruses may serve as a marker of active replication. This is of relevance, as several DNA viral respiratory pathogens can become latent (e.g., HSV and CMV) or persist for extended periods (e.g., human bocavirus [HBoV]), so detection of their genomic DNA may not be a sufficient indication for acute infections (28).

The sensitivity of untargeted metagenomics is a function of sample composition and sequencing depth. When sequenced to the same depth, samples with an abundance of nonpathogen RNA (e.g., highly cellular samples or samples with abundant normal flora) result in lower analytical sensitivity than samples in which the pathogen RNA is more abundant (e.g., less cellular samples, higher pathogen load, or absence of normal flora). As rRNA represents a large proportion of host RNA, rRNA depletion strategies have been used to mitigate this effect. We decided not to use this approach, as it may have off-target effects (e.g., depleting microbial rRNA or other sequences with sufficient homology), which limits the unbiased nature of metagenomics. In addition, rRNA depletion or target enrichment steps add complexity and cost to the workflow. Samples were sequenced to a depth of 5×10^6 to 10×10^6 reads/sample to limit sequencing costs. This sequencing depth resulted in comparable positivity rates and agreement with RVP and qPCR of >90%. When clinically relevant, samples can be sequenced more deeply, resulting in proportional increases of analytical sensitivity.

When approaching the limit of detection, small numbers of viral reads pose challenges to result interpretation, as they can represent true-positive, low-level detections or artifacts. False-positive detections can be due to contamination during library preparation (e.g., sample-to-sample contamination prior to indexing), may be a result of sequencing artifacts (e.g., run-to-run carryover or demultiplexing errors), or may be caused by erroneous classification during data analysis (e.g., due to highly homologous or low-complexity regions) (29). Thus, the confidence of viral detection depends on the number of viral reads and evenness of coverage. Given the testing complexity, read counts may vary between analytical replicates. To determine the within-run and between-run variability, we tested multiple aliquots of 3 virus-positive samples from sample extraction through data analysis. Respiratory viral read counts across a wide range of fractional abundance were highly reproducible within and between runs (coefficients of variation [CV] ranging from 16% to 63%). Collectively, our results demonstrate that untargeted metagenomics can be used to supplement current PCR-based tests and that metagenomics data can be rapidly and effectively analyzed using recently published, ultrafast read-classification tools such as Taxonomer, SURPI (30), and Kraken (31).

Another distinct advantage of metagenomics-based pathogen detection is the ability to determine the molecular subtype of a particular virus and query it quickly for genotypic markers of drug resistance or pathogenicity. In our study, molecular typing was possible for 84% of all viral strains. Relevant information derived from typing included the following: (i) almost two-thirds of rhinoviruses belonged to the more pathogenic species, C, including one highly divergent strain missed by the RVP (32, 33); (ii) all influenza A viruses were 2009 H1N1 strains, but none contained the H275Y mutation conferring oseltamivir resistance; (iii) RSV-B was 3 times more prevalent than RSV-A, which may be relevant, as strain-specific differences in pathogenicity have been suggested; (iv) high-resolution typing of an enterovirus as coxsackievirus B4; (v) typing of 7 coronaviruses as NL63, HKU1, and OC43; and (vi) genotyping of 4 bocavirus strains as HBoV-1. As genotype-phenotype correlations become better understood, genotypic strain characterization will gain importance. This will also facilitate epidemiologic investigations or studies of vaccine effectiveness. Particularly in the case of influenza, real-time sequence information will improve surveillance studies, enable early detection of antiviral drug resistance, and inform vaccination strategies.

Respiratory viral burden correlates with disease severity and may help differentiate asymptomatic shedding from active infection (23–25, 34). Published studies correlating viral read counts with qPCR had limited sample sizes (13–15). Thus, we tested whether normalized read counts could be used for quantification of the viral burden by comparison to viral loads determined by pathogen-specific, laboratory-developed qPCRs. While viral reads always represented a small fraction of total reads (Fig. 2A), normalized counts correlated highly significantly with viral loads. Untargeted metagenomics could therefore also be used to measure viral burden.

While we demonstrated analytical performance comparable to the RVP, there are several barriers for routine diagnostic deployment of metagenomics-based testing. These include lengthy turnaround times, costs, and complexity of data analysis. First, the library preparation method used in this study required ~14 h. We performed sequencing on an Illumina HiSeq 2500 instrument in high-output run mode, which took an additional ~11 days. At the time of this writing, partially automated solutions for RNA-seq library preparation within ~8 h and sequencing within ≤1 day for comparable per-base costs have become available (11, 35). These advances are starting to enable diagnostic laboratories to provide results in a clinically meaningful time frame and with a workflow that can be implemented in diagnostic laboratories. However, for wide adoption, rapid, automated, closed-system library preparation methods and quicker sample-to-data times are needed.

Second, cost is a great concern regarding the use of next-generation sequencing in infectious disease diagnostics. For the present study, RNA-seq reagent costs per sample were within $10 to $20 of reagent costs for RVP. This was in part due to multiplexing 24 samples per sequencing lane. At the time of this writing, cheaper library preparation kits and sequencing platforms have further decreased costs, quickly eliminating the cost differential. Enrichment of viral sequences and depletion of uninformative host RNA can reduce sequencing costs by increasing coverage but introduces complexity and costs of library preparations (36, 37). We analyzed untargeted metagenomics data solely for the presence of respiratory viruses, ignoring bacterial respiratory pathogens simultaneously identified by Taxonomer, as most of those can be part of the normal upper respiratory tract flora and only NP swabs were tested. However, when used with lower respiratory tract samples, untargeted metagenomics has the potential to also

replace a large number of commonly performed culture- and PCR-based tests.

Finally, data analysis needs to be rapid, user-friendly, and reliable enough that it can be implemented without large investments in highly trained personnel and computational infrastructure. We used our recently published metagenomics data analysis tool, Taxonomer, to screen for the presence of respiratory viruses (Flygare et al., submitted). Taxonomer analyzed $\sim 1 \times 10^6$ reads/minute, requiring <10 min per sample. For diagnostic applications, data analysis solutions are needed that minimize the time users spend reviewing results. We confirmed all respiratory virus detections manually to ensure accuracy. However, this was only informative for samples with low viral read counts given concern of false-positive results due to misclassification or sequencing artifacts (see above). Several RVP and qPCR-positive samples produced only <10 reads for that virus, making detections unreliable at this low end (Fig. 2B). Deeper sequencing or target enrichment depletion approaches could alleviate the problem but increase costs and/or workflow complexity. For highly variable viruses (e.g., *Picornaviridae*), suspicious reads can be mapped back to viral consensus sequences of source strains to identify reads that likely represent artifacts. For diagnostic adoption, interpretive criteria similar to those being established for genomics laboratories will need to be developed and incorporated in diagnostic data analysis tools to enable consistent and rapid analyses (38, 39).

In summary, we showed that metagenomics-based detection of respiratory viruses holds promise as a diagnostics tool enabling unbiased pathogen detection, molecular tying, and genotypic assessment of drug resistance or pathogenicity. Barriers to adoption, including turnaround time, cost, and complex data analysis, are rapidly being removed. Adoption may be for testing of immunocompromised or otherwise predisposed patients, when routine therapeutic approaches fail, during clusters of infections of unknown etiology, or when molecular characterization of pathogens is sought. As highlighted by a diverse HRV-A strain missed by the RVP, the unbiased nature of metagenomics can also assist with detection of novel viruses or variant strains.

ACKNOWLEDGMENTS

We thank Tatum Lunt for her assistance with sample collection. S.F., K.E.S., K.E., M.Y., and R.S. have a patent application pending for Taxonomer, which was licensed by IDbyDNA. M.Y. and R.S. own equity in and consult for IDbyDNA.

FUNDING INFORMATION

This work was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health (awards 1KL2TR001065 and UL1TR001067), by the Primary Children's Hospital Foundation, by a grant from the Department of Pathology, University of Utah School of Medicine, and by the ARUP Institute for Clinical and Experimental Pathology.

REFERENCES

1. **Choi SH, Hong SB, Ko GB, Lee Y, Park HJ, Park SY, Moon SM, Cho OH, Park KH, Chong YP, Kim SH, Huh JW, Sung H, Do KH, Lee SO, Kim MN, Jeong JY, Lim CM, Kim YS, Woo JH, Koh Y.** 2012. Viral infection in patients with severe pneumonia requiring intensive care unit admission. Am J Respir Crit Care Med **186:**325–332. http://dx.doi.org/10.1164/rccm.201112-2240OC.
2. **Karhu J, Ala-Kokko TI, Vuorinen T, Ohtonen P, Syrjala H.** 11 April 2014. Lower respiratory tract virus findings in mechanically ventilated patients with severe community-acquired pneumonia. Clin Infect Dis http://dx.doi.org/10.1093/cid/ciu237.
3. **Honkinen M, Lahti E, Osterback R, Ruuskanen O, Waris M.** 2012. Viruses and bacteria in sputum samples of children with community-acquired pneumonia. Clin Microbiol Infect **18:**300–307. http://dx.doi.org/10.1111/j.1469-0691.2011.03603.x.
4. **Jain S, Williams DJ, Arnold SR, Ampofo K, Bramley AM, Reed C, Stockmann C, Anderson EJ, Grijalva CG, Self WH, Zhu Y, Patel A, Hymas W, Chappell JD, Kaufman RA, Kan JH, Dansie D, Lenny N, Hillyard DR, Haynes LM, Levine M, Lindstrom S, Winchell JM, Katz JM, Erdman D, Schneider E, Hicks LA, Wunderink RG, Edwards KM, Pavia AT, McCullers JA, Finelli L,** CDC **EPIC Study Team.** 2015. Community-acquired pneumonia requiring hospitalization among U.S. children. N Engl J Med **372:**835–845. http://dx.doi.org/10.1056/NEJMoa1405870.
5. **Jain S, Self WH, Wunderink RG, Fakhran S, Balk R, Bramley AM, Reed C, Grijalva CG, Anderson EJ, Courtney DM, Chappell JD, Qi C, Hart EM, Carroll F, Trabue C, Donnelly HK, Williams DJ, Zhu Y, Arnold SR, Ampofo K, Waterer GW, Levine M, Lindstrom S, Winchell JM, Katz JM, Erdman D, Schneider E, Hicks LA, McCullers JA, Pavia AT, Edwards KM, Finelli L,** CDC **EPIC Study Team.** 2015. Community-acquired pneumonia requiring hospitalization among U.S. adults. N Engl J Med **373:**415–427.
6. **Pavia AT.** 2011. Viral infections of the lower respiratory tract: old viruses, new viruses, and the role of diagnosis. Clin Infect Dis **52**(Suppl 4)**:**S284–S289.
7. **Ruuskanen O, Jarvinen A.** 2014. What is the real role of respiratory viruses in severe community-acquired pneumonia? Clin Infect Dis http://dx.doi.org/10.1093/cid/ciu242.
8. **Ruuskanen O, Lahti E, Jennings LC, Murdoch DR.** 2011. Viral pneumonia. Lancet **377:**1264–1275. http://dx.doi.org/10.1016/S0140-6736(10)61459-6.
9. **Self WH, Williams DJ, Zhu Y, Ampofo K, Pavia AT, Chappell JD, Hymas WC, Stockmann C, Bramley AM, Schneider E, Erdman D, Finelli L, Jain S, Edwards KM, Grijalva CG.** 2015. Respiratory viral detection in children and adults: comparing asymptomatic controls and patients with community-acquired pneumonia. J Infect Dis http://dx.doi.org/10.1093/infdis/jiv323.
10. **Grard G, Fair JN, Lee D, Slikas E, Steffen I, Muyembe JJ, Sittler T, Veeraraghavan N, Ruby JG, Wang C, Makuwa M, Mulembakani P, Tesh RB, Mazet J, Rimoin AW, Taylor T, Schneider BS, Simmons G, Delwart E, Wolfe ND, Chiu CY, Leroy EM.** 2012. A novel rhabdovirus associated with acute hemorrhagic fever in central Africa. PLoS Pathog **8:**e1002924. http://dx.doi.org/10.1371/journal.ppat.1002924.
11. **Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, Salamat SM, Somasekar S, Federman S, Miller S, Sokolic R, Garabedian E, Candotti F, Buckley RH, Reed KD, Meyer TL, Seroogy CM, Galloway R, Henderson SL, Gern JE, DeRisi JL, Chiu CY.** 2014. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. N Engl J Med **370:**2408–2417. http://dx.doi.org/10.1056/NEJMoa1401268.
12. **Wylie KM, Mihindukulasuriya KA, Sodergren E, Weinstock GM, Storch GA.** 2012. Sequence analysis of the human virome in febrile and afebrile children. PLoS One **7:**e27735. http://dx.doi.org/10.1371/journal.pone.0027735.
13. **Prachayangprecha S, Schapendonk CM, Koopmans MP, Osterhaus AD, Schurch AC, Pas SD, van der Eijk AA, Poovorawan Y, Haagmans BL, Smits SL.** 2014. Exploring the potential of next-generation sequencing in detection of respiratory viruses. J Clin Microbiol **52:**3722–3730. http://dx.doi.org/10.1128/JCM.01641-14.
14. **Yang J, Yang F, Ren L, Xiong Z, Wu Z, Dong J, Sun L, Zhang T, Hu Y, Du J, Wang J, Jin Q.** 2011. Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. J Clin Microbiol **49:**3463–3469. http://dx.doi.org/10.1128/JCM.00273-11.
15. **Fischer N, Indenbirken D, Meyer T, Lutgehetmann M, Lellek H, Spohn M, Aepfelbacher M, Alawi M, Grundhoff A.** 2015. Evaluation of unbiased next-generation sequencing of RNA (RNA-seq) as a diagnostic method in influenza virus-positive respiratory samples. J Clin Microbiol **53:**2238–2250. http://dx.doi.org/10.1128/JCM.02495-14.
16. **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, Maeda N, 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. http://dx.doi.org/10.1371/journal.pone.0004219.
17. **Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Moller N, Aarestrup FM.** 2014. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. J Clin Microbiol **52:**139–146. http://dx.doi.org/10.1128/JCM.02452-13.
18. **Seo S, Renaud C, Kuypers JM, Chiu CY, Huang ML, Samayoa E, Xie H, Yu G, Fisher CE, Gooley TA, Miller S, Hackman RC, Myerson D, Sedlak RH, Kim YJ, Fukuda T, Fredricks DN, Madtes DK, Jerome KR, Boeckh M.** 2015. Idiopathic pneumonia syndrome after hematopoietic cell transplantation: evidence of occult infectious etiologies. Blood **125:**3789–3797. http://dx.doi.org/10.1182/blood-2014-12-617035.
19. **Miller CA, Qiao Y, DiSera T, D'Astous B, Marth GT.** 2014. bam.iobio: a web-based, real-time, sequence alignment file inspector. Nat Methods **11:**1189. http://dx.doi.org/10.1038/nmeth.3174.
20. **Hymas WC, Aldous WK, Taggart EW, Stevenson JB, Hillyard DR.** 2008. Description and validation of a novel real-time RT-PCR enterovirus assay. Clin Chem **54:**406–413. http://dx.doi.org/10.1373/clinchem.2007.095414.
21. **Hymas WC, Mills A, Ferguson S, Langer J, She RC, Mahoney W, Hillyard DR.** 2010. Development of a multiplex real-time RT-PCR assay for detection of influenza A, influenza B, RSV and typing of the 2009-H1N1 influenza virus. J Virol Methods **167:**113–118. http://dx.doi.org/10.1016/j.jviromet.2010.03.020.
22. **Dare RK, Fry AM, Chittaganpitch M, Sawanpanyalert P, Olsen SJ, Erdman DD.** 2007. Human coronavirus infections in rural Thailand: a comprehensive study using real-time reverse-transcription polymerase chain reaction assays. J Infect Dis **196:**1321–1328. http://dx.doi.org/10.1086/521308.
23. **El Saleeby CM, Bush AJ, Harrison LM, Aitken JA, Devincenzo JP.** 2011. Respiratory syncytial virus load, viral dynamics, and disease severity in previously healthy naturally infected children. J Infect Dis **204:**996–1002. http://dx.doi.org/10.1093/infdis/jir494.
24. **Houben ML, Coenjaerts FE, Rossen JW, Belderbos ME, Hofland RW, Kimpen JL, Bont L.** 2010. Disease severity and viral load are correlated in infants with primary respiratory syncytial virus infection in the community. J Med Virol **82:**1266–1271. http://dx.doi.org/10.1002/jmv.21771.
25. **Zhao B, Yu X, Wang C, Teng Z, Wang C, Shen J, Gao Y, Zhu Z, Wang J, Yuan Z, Wu F, Zhang X, Ghildyal R.** 2013. High human bocavirus viral load is associated with disease severity in children under five years of age. PLoS One **8:**e62318. http://dx.doi.org/10.1371/journal.pone.0062318.
26. **Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B.** 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods **5:**621–628. http://dx.doi.org/10.1038/nmeth.1226.
27. **Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW.** 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol **12:**87. http://dx.doi.org/10.1186/s12915-014-0087-z.
28. **Martin ET, Fairchok MP, Kuypers J, Magaret A, Zerr DM, Wald A, Englund JA.** 2010. Frequent and prolonged shedding of bocavirus in young children attending daycare. J Infect Dis **201:**1625–1632. http://dx.doi.org/10.1086/652405.
29. **Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J.** 2014. Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. PLoS One **9:**e94249. http://dx.doi.org/10.1371/journal.pone.0094249.
30. **Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, Bouquet J, Greninger AL, Luk KC, Enge B, Wadford DA, Messenger SL, Genrich GL, Pellegrino K, Grard G, Leroy E, Schneider BS, Fair JN, Martinez MA, Isa P, Crump JA, DeRisi JL, Sittler T, Hackett J, Jr, Miller S, Chiu CY.** 2014. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res **24:**1180–1192. http://dx.doi.org/10.1101/gr.171934.113.
31. **Wood DE, Salzberg SL.** 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol **15:**R46. http://dx.doi.org/10.1186/gb-2014-15-3-r46.
32. **Lau SK, Yip CC, Lin AW, Lee RA, So LY, Lau YL, Chan KH, Woo PC, Yuen KY.** 2009. Clinical and molecular epidemiology of human rhinovirus C in children and adults in Hong Kong reveals a possible distinct human rhinovirus C subgroup. J Infect Dis **200:**1096–1103. http://dx.doi.org/10.1086/605697.
33. **Bochkov YA, Gern JE.** 2012. Clinical and molecular features of human rhinovirus C. Microbes Infect **14:**485–494. http://dx.doi.org/10.1016/j.micinf.2011.12.011.
34. **Christensen A, Nordbo SA, Krokstad S, Rognlien AG, Dollner H.** 2010. Human bocavirus in children: mono-detection, high viral load and viraemia are associated with respiratory tract infection. J Clin Virol **49:**158–162. http://dx.doi.org/10.1016/j.jcv.2010.07.016.
35. **Lefterova MI, Suarez CJ, Banaei N, Pinsky BA.** 2015. Next-generation sequencing for infectious disease diagnosis and management: a report of the Association for Molecular Pathology. J Mol Diagn **17:**623–634.
36. **Wylie TN, Wylie KM, Herter BN, Storch GA.** 2015. Enhanced virome sequencing through solution-based capture enrichment. Genome Res **25:**1910–1920.
37. **Briese T, Kapoor A, Mishra N, Jain K, Kumar A, Jabado OJ, Lipkin WI.** 2015. Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. mBio **6:**e01491-15. http://dx.doi.org/10.1128/mBio.01491-15.
38. **Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, Lu F, Lyon E, Voelkerding KV, Zehnbauer BA, Agarwala R, Bennett SF, Chen B, Chin EL, Compton JG, Das S, Farkas DH, Ferber MJ, Funke BH, Furtado MR, Ganova-Raeva LM, Geigenmuller U, Gunselman SJ, Hegde MR, Johnson PL, Kasarskis A, Kulkarni S, Lenk T, Liu CS, Manion M, Manolio TA, Mardis ER, Merker JD, Rajeevan MS, Reese MG, Rehm HL, Simen BB, Yeakley JM, Zook JM, Lubin IM.** 2012. Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol **30:**1033–1036. http://dx.doi.org/10.1038/nbt.2403.
39. **Gargis AS, Kalman L, Bick DP, da Silva C, Dimmock DP, Funke BH, Gowrisankar S, Hegde MR, Kulkarni S, Mason CE, Nagarajan R, Voelkerding KV, Worthey EA, Aziz N, Barnes J, Bennett SF, Bisht H, Church DM, Dimitrova Z, Gargis SR, Hafez N, Hambuch T, Hyland FC, Luna RA, MacCannell D, Mann T, McCluskey MR, McDaniel TK, Ganova-Raeva LM, Rehm HL, Reid J, Campo DS, Resnick RB, Ridge PG, Salit ML, Skums P, Wong LJ, Zehnbauer BA, Zook JM, Lubin IM.** 2015. Good laboratory practice for clinical next-generation sequencing informatics pipelines. Nat Biotechnol **33:**689–693. http://dx.doi.org/10.1038/nbt.3237.
40. **Kuroda M, Niwa S, Sekizuka T, Tsukagoshi H, Yokoyama M, Ryo A, Sato H, Kiyota N, Noda M, Kozawa K, Shirabe K, Kusaka T, Shimojo N, Hasegawa S, Sugai K, Obuchi M, Tashiro M, Oishi K, Ishii H, Kimura H.** 2015. Molecular evolution of the VP1, VP2, and VP3 genes in human rhinovirus species C. Sci Rep **5:**8185. http://dx.doi.org/10.1038/srep08185.