Advances in diagnostics of respiratory viruses and insight in clinical implications of rhinovirus infections

Rijn-Klink, A.L. van

APPLICATION AND ADDED VALUE OF ADVANCED RESPIRATORY VIRAL DIAGNOSTIC METHODS

PART I

CHAPTER 2

Retrospective validation of a metagenomic sequencing protocol for combined detection of RNA and DNA viruses using respiratory samples from paediatric patients

S. van Boheemen[a#], A. L. van Rijn[a#], N. Pappas[b], E.C. Carbo[a], R.H.P. Vorderman[b], I. Sidorov[a], P.J. van`t Hof[b], H. Mei[b], E.C.J. Claas[a], A.C.M. Kroes[a*], J.J.C. de Vries[*$]

[a] Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands
[b] Sequencing Analysis Support Core, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands
[#] These authors contributed equally to this work
[*] These authors contributed equally to this work

Journal of molecular diagnostics 2020 22 (2): 196-207

ABSTRACT

Viruses are the main cause of respiratory tract infections. Metagenomic next-generation sequencing (mNGS) enables unbiased detection of all potential pathogens. To apply mNGS in viral diagnostics, sensitive and simultaneous detection of RNA and DNA viruses is needed. Herein, were studied the performance of an in-house mNGS protocol for routine diagnostics of viral respiratory infections with potential for automated pan-pathogen detection. The sequencing protocol and bioinformatics analysis were designed and optimized, including exogenous internal controls. Subsequently, the protocol was retrospectively validated using 25 clinical respiratory samples. The developed protocol using Illumina NextSeq 500 sequencing showed high repeatability. Use of the National Center for Biotechnology Information's RefSeq database as opposed to the National Center for Biotechnology Information's nucleotide database led to enhanced specificity of classification of viral pathogens. A correlation was established between read counts and PCR cycle threshold value. Sensitivity of mNGS, compared with PCR, varied up to 83%, with specificity of 94%, dependent on the cutoff for defining positive mNGS results. Viral pathogens only detected by mNGS, not present in the routine diagnostic workflow, were influenza C, KI polyomavirus, cytomegalovirus, and enterovirus. Sensitivity and analytical specificity of this mNGS protocol were comparable to PCR and higher when considering off-PCR target viral pathogens. One single test detected all potential viral pathogens and simultaneously obtained detailed information on detected viruses.

INTRODUCTION

Respiratory tract infections pose a great burden on public health, causing extensive morbidity and mortality among patients worldwide[1-3]. The majority of acute respiratory infections is caused by viruses, such as rhinovirus (RV), influenza (INF) A and B viruses, metapneumovirus (MPV), and respiratory syncytial virus (RSV)[4]. However, in 20-62% of the patients, no pathogen is detected[4-6]. This might be the result of diagnostic failures or even infection by unknown pathogens, such as the Middle East respiratory syndrome coronavirus (MERS-CoV), in 2012[7].

Rapid identification of the respiratory pathogen is critical to determine downstream decision-making such as isolation measures or treatment, including cessation of antibiotic therapy. Current diagnostic amplification methods as real-time polymerase chain reaction (qPCR) are very sensitive and specific, but are only targeting predefined virus species or types. Genetic diversity within the virus genome and the sheer number of potential pathogens in many clinical conditions pose limitations to predefined primer and probe based approaches, leading to false negative results[8]. These limitations, combined with the potential emergence of new or unusual pathogens highlight the need for less restricted approaches that could improve the diagnosis and subsequent outbreak management of infectious diseases.

Metagenomics relates to the study of the complete genomic content in a complex mixture of (micro)organisms[9]. Unlike bacteria, viruses do not display a common gene in all virus families, and therefore pan-virus detection relies on catch-all analytic methods. Metagenomics or untargeted next-generation sequencing (mNGS) offers a culture and nucleotide-sequence-independent method that eliminates the need to define the targets for diagnosis beforehand. Besides primary detection, mNGS immediately offers additional information, on virulence markers, epidemiology, genotyping, and evolution of pathogens[7,10-12]. Furthermore, quantitative assessment of the presence of virus copies in the sample is enabled by the number reads[8].

While original mNGS studies typically aim at analysis of (shifts in) population diversity of abundant DNA microbes, detection of viral pathogens in patient samples requires a different technical approach because of 1) the usually very low abundance of viral pathogens (<1%) in clinical samples and 2) the requisite of detecting both DNA and RNA viruses. Hence, a low limit of detection for RNA and DNA in one single assay is essential for implementation of mNGS for routine pathogen detection in clinical diagnostic laboratories. Current viral mNGS protocols are optimized for either RNA or DNA detection[11,13-15]. Consequently, detection of both RNA and DNA viruses requires parallel work-up of both RNA and DNA pre-treatment methods. Additionally, to increase the relative concentration of viral sequences, viral particle enrichment techniques are often applied[8,12]. These techniques are laborious and not easily automated for routine clinical diagnostic use. Moreover, during enrichment directed at viral particles, intracellular viral nucleic acids as genomes and mRNAs are being discarded. Following sequencing, the bioinformatic classification and interpretation of the results remain a major challenge. Bioinformatic classifiers are often developed for usage in either microbiome studies or classification of high abundant reads whereas extensive validation for clinical diagnostic usage in settings of very low abundance is very limited. After bioinformatics classification, the challenge remains to discriminate between viruses that play a role in disease aetiology and non-pathogenic

viruses[16]. Before considering mNGS in routine diagnostics, there is a need for critical evaluation and validation of every step in the procedure.

In this study, we evaluated a metagenomic protocol for NGS-based pathogen detection with sample pre-treatment for DNA and RNA in a single tube. The method was validated using a selection of 25 respiratory paediatric samples with in total 29 positive and 346 negative viral PCR results. The main study objective was to define a sensitive and specific method for mNGS to be used as a broad diagnostic tool for viral respiratory diseases with the potential for automated pan-pathogen detection.

MATERIAL AND METHODS

Sample selection

Twenty-five stored clinical respiratory samples (-80 ºC) from paediatric patients, sent to the microbiological laboratory for routine viral diagnostics in 2016, were selected from the laboratory database (general laboratory information management system, MIPS, Ghent, Belgium) at the Leiden University Medical Center (LUMC). Based on previous PCR test results, a variety of 21 positive and four negative respiratory virus samples with a wide range of quantification cycle (Cq) values were included. The sample types represented routine diagnostic samples from paediatric patients that had been sent to our laboratory: 19 nasopharyngeal washings , two sputa, two broncho-alveolar lavages (BAL), one bronchial washing and one throat swab (in viral transport medium). The patient selection (age range 1.2 months – 15 years) represented the paediatric population with respiratory diagnostics in our university hospital in terms of (underlying) illness.

Sample pre-treatment

Total nucleid acids (NA) were extracted directly from 200 µL of clinical material using the MagNAPure 96 DNA and Viral NA Small Volume Kit (Roche Diagnostics, Almere, the Netherlands) with 100 µL output eluate.

Internal controls

Clinical material was spiked with equine arteritis virus (EAV) and phocine herpesvirus 1 (PhHV1, kindly provided by prof. dr. H.G.M. Niesters, UMC Groningen, the Netherlands), as internal controls for RNA detection[17] and DNA detection respectively[18]. To determine the optimal concentration of the internal controls a ten-fold dilution series of PhHV1/EAV was added to a mix of two pooled influenza A positive throat swabs (Cq value 25) and read count and Cq values were compared. Concentration was based on the number of mNGS reads.

Quality control

Before sequencing the DNA input concentration was measured with the Qubit (ThermoFisher Scientific, Waltham, USA), to determine whether there was sufficient DNA in the sample to obtain sequencing results. The range of DNA input for library preparation was 0.5 ng/µl for throat swabs (see reproducibility experiment) up to 300 ng/µl for bronchoalveolar lavages and sputa.

Fragmentation

To compare the effect of different DNA fragmentation techniques, six PCR positive (containing one to three viruses) and three PCR negative samples were 1) chemically fragmented using zinc (10 min.) as part of the NEB (New England Biolabs) Library Prep Kit protocol as described below (see library preparation) and 2) physically fragmented using sonication with the Bioruptor® pico (Diagenode, Seraing, Belgium, on/off time: 18/30s, 5 cycli)[19]. Three samples were also tested with the 3) high intensity settings of the Bioruptor® pico (on/off time: 30/40s, 14 cycli).

Library preparation

Libraries were constructed with 7µL extracted nucleic acids using the NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina® (New England Biolabs, IpsWich, USA) using single, unique adaptors. This kit has been developed for transcriptome analyses. We made several adaptations to the manufacturers protocol in order to enable simultaneous detection of both DNA and RNA viruses: the following steps were omitted: Poly A mRNA capture isolation (Instruction manual NEB #E7420S/L, version 8.0, Chapter 1), rRNA depletion and DNase step (Chapter 2.1-2.4, 2.5B, 2.11A).

The size of fragments in the library was 300-700 bp. Adaptors were diluted 30 fold given the low RNA/DNA input and 21 PCR cycli were run post-adaptor ligation.

Nucleotide Sequence Analysis

Sequencing was performed on Illumina HiSeq 4000 and NextSeq 500 sequencing systems (Illumina, San Diego, CA, USA), obtaining 10 million 150 bp paired-end reads per sample.

Detection limit

To determine the detection limit of mNGS, serial dilutions (undiluted, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}) of an influenza A positive sample was tested with both mNGS and lab developed real-time PCR . Based on run-off transcript experiments the typical limit of detection of our real-time RNA PCRs was estimated to be 10-50 copies/reaction (data not shown).

Repeatability (within run precision)

To estimate the reproducibility of metagenomic sequencing an influenza A positive clinical sample (throat swab) was divided into four aliquots, nucleic acids were extracted, library preparation and subsequent sequence analysis on the Illumina HiSeq 4000 was performed in one run.

Bioinformatics: taxonomic classification

All FASTQ files were processed using the BIOPET Gears pipeline version 0.9.0 developed at the LUMC (*http://biopet-docs.readthedocs.io/en/stable/* accessed 9-12-2018). This pipeline performs FASTQ pre-processing (including quality control, quality trimming and adapter clipping) and taxonomic classification of sequencing reads. In this project, FastQC version 0.11.2 (*https://www.bioinformatics.babraham.ac.uk/projects/fastqc/* accessed 9-12-2018*)* was used for checking the quality of the raw reads. Low quality read trimming was done using Sickle[20] version 1.33 with default settings. Adapter clipping was performed using Cutadapt[21] version 1.10 with default settings. Taxonomic classification of reads was performed with Centrifuge[22] version 1.0.1-beta. The pre-built NT index, which contains all sequences from NCBI's nucleotide database, provided by the Centrifuge developers was used (*ftp://ftp.ccb.jhu.edu/pub/infphilo/centrifuge/data/old-indices* accessed 16-11-2017) as the reference database. An overview of the bioinformatic process is shown in Figure 1.

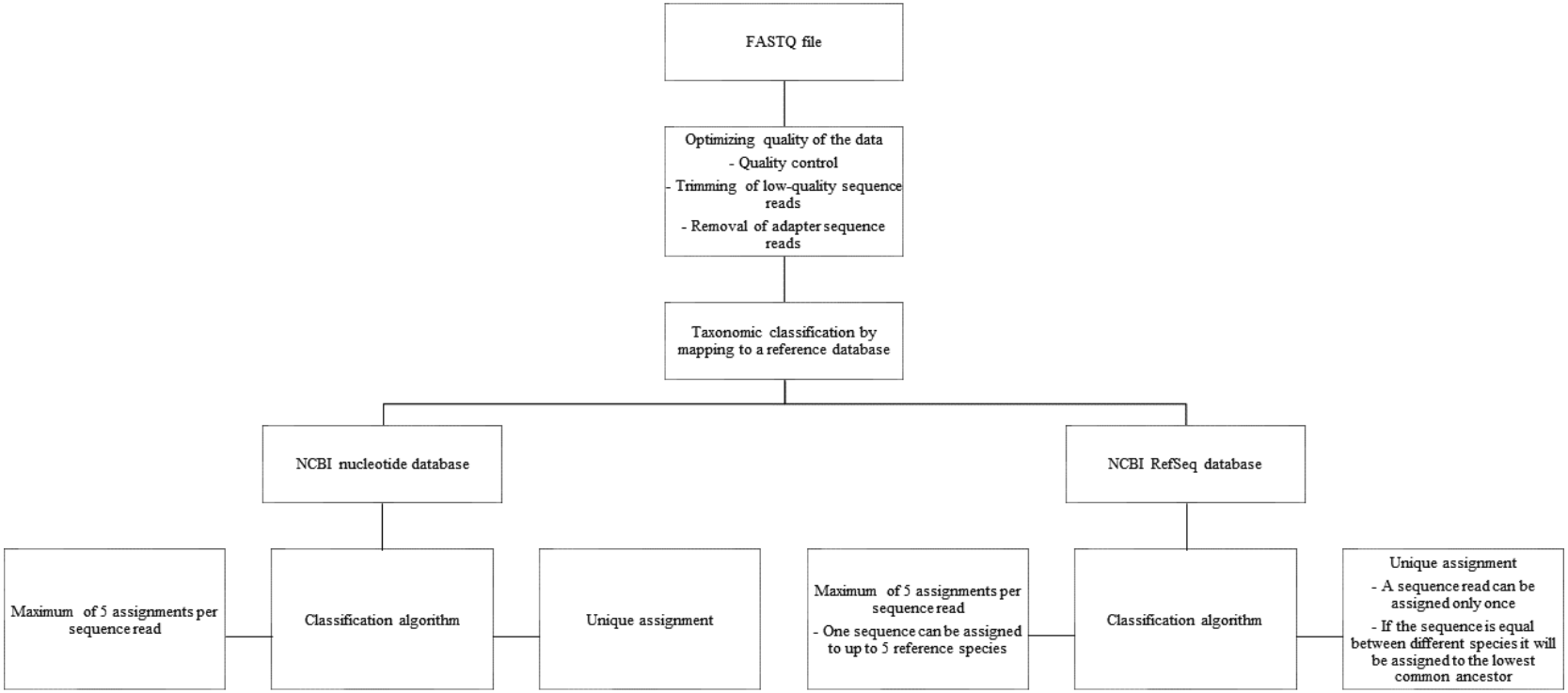

Figure 1. **The bioinformatic workflow of the mNGS protocol studied.**

In addition, a customized reference centrifuge index with sequence information obtained from the NCBI's RefSeq[23] (accessed February 2019) database was built. RefSeq genomic sequences for the domains of bacteria, viruses, archaea, fungi, protozoa, as well as the human reference, along with the taxonomy identifiers, were downloaded with the Centrifuge-download utility and were used as input for Centrifuge-build.

Centrifuge settings were evaluated to increase the sensitivity and specificity. The default setting, with which a read can be assigned to up to five different taxonomic categories, was compared to one unique assignment per read[22] where a read is assigned to a single taxonomic category, corresponding to the lowest common ancestor of all matching species.

Kraken-style reports with taxonomical information were produced by the Centrifuge-kreport utility for all (default) options. Both unique and non-unique assignments can be reported, and these settings were compared. The resulting tree-like structured, Kraken-style reports were visualized with Krona[24] version 2.0.

Horizontal coverage (%) was determined using GenomeDetective website[25] version 1.111 (*https://www.genomedetective.com/*, accessed 5-4- 2019).

In silico simulated EAV reads were analysed in different databases (NCBI's nucleotide vs RefSeq), classification algorithms (max 5 labels per sequence, vs unique, lowest(common ancestor) and reporting (non-unique vs unique) to determine the most sensitive an specific bioinformatic analyses using Centrifuge.

To determine the amount of reads needed, results of one and 10 million reads were compared. A total of one million reads were randomly selected of the 10 million reads of one FASTQ file and analysed. The random selection was performed with the FastqSplitter (*https://github.com/biopet/biopet/blob/v0.9.0/docs/tools/FastqSplitter.md* accessed 9-12-2018), which cuts a FASTQ file of 10 million reads is into 10 pieces, of which one was selected. Read counts were normalized by the total read count and target virus genome size.

Bioinformatics: assembly of PhHV1 sequences

Since NCBI's databases were lacking a complete PhHV1 genome sequence, PhHV1 was sequenced and based on the gained sequence reads the genome was build using SPAdes[26]. Assembly of PhHV1 was done using the biowdl virus-assembly pipeline 0.1 *(https://github.com/biowdl/virus-assembly* accessed 9-12-2018*)*. The QC part of the biowdl pipeline determines which adapters need to be clipped by using FastQC version 0.11.7 (*https://www.bioinformatics.babraham.ac.uk/projects/fastqc/* accessed 9-12-2018*)* and cutadapt version 1.16[21], with minimum length setting "1''. The resulting reads were downsampled within bowdl to 250 000 reads using seqtk 1.2 *(https://github.com/lh3/seqtk* accessed 9-12-2018*)* after which SPADES version 3.11.1[26] was run to get the first proposed genome contigs.

To retrieve longer assembly contigs a reiterative assembly approach was used by processing the proposed contigs by the biowdl reAssembly pipeline 0.1. This preassembly pipeline aligns reads to

contigs of a previous assembly, then selects the aligned reads, downsamples them and runs a new assembly using SPADES. Subtools used for this consisted of BWA 0.7.17[27] for indexing and mapping, SAMtools 1.6[28] for creating bam files, SAMtools view (version 1.7) for filtering out unmapped reads using the setting "-G 12", Picard SamToFastq (version 2.18.4) and seqtk for creating FASTQ files with 250 000 reads. The contigs from the reAssembly pipeline were then processed for a second using SPADES, with setting the 'cov-cutoff' to 5. The resulting contigs were then processed with the reAssembly pipeline for the third and last time setting the 'cov-cutoff' in SPADES to 20.

The contigs from the last reAssembly step were then run against the blast NT database using blastn 2.7.1[29] Out of 23 contigs only 5 contigs, that showed the lowest % in identity matches with any other possible non herpes virus species, were selected. The final 5 contigs contained sequence lengths of 97893, 8170 3710, 3294 and1279 nucleotides, the average coverage was 206, 131, 211, 285 and 154, respectively. The proposed almost complete genome of PhHV1 was added to NCBI's GenBank database (https://www.ncbi.nlm.nih.gov/genbank/, accession number GenBank MH509440, release date 4 Dec 2018).

Retrospective validation

Clinical sensitivity was analysed using the optimized procedure, which in short consisted of total NA extraction including internal controls (1:100 dilution),the adapted NEB Next library preparation protocol including fragmentation with zinc, for combined RNA and DNA detection (see library preparation), and sequencing of 10 million reads(Illumina NextSeq 500). Bioinformatic analyses was performed using Centrifuge with NCBI's RefSeq database and unique assignment of the sequence reads.

Sensitivity and specificity of the metagenomic NGS procedure was compared to a published updated version of our lab developed multiplex qPCR[30]. The routine multiplex PCR panel consisted of 15 respiratory target pathogens: influenza virus A/ B, respiratory syncytial virus (RSV), metapneumovirus (MPV), adenovirus (ADV), human bocavirus (HBoV), parainfluenza viruses (PIV) 1/ 2/ 3/ 4, rhinovirus (RV), and the coronaviruses HKU1, NL63, 227E and OC43. Thus, in total 375 PCR results were available (15 targets x 25 samples) of which 29 PCR positive and 346 PCR negative for comparison with mNGS.

Ethical approval of patient studies

The study design was approved by the medical ethics review committee of the Leiden University Medical Center (reference B16.004).

RESULTS

Internal controls

Serial dilutions of EAV and PhHV1 were added to an influenza A PCR positive sample. Serial dilution 1:10,000 detected EAV with a substantial read count in the presence of a viral infection and without a significant decline in target virus family reads (Table 1). Based on these results we determined the concentration of internal controls for further experiments.

The EAV Cq value of the dilutions correlated with the number of EAV reads from the Centrifuge analysis.

Fragmentation

The comparison of fragmentation methods was done using a selection of samples with relevant target reads and performed on the Illumina Nextseq 500 As shown in Figure 2, the total reads were comparable among the three protocols. The protocol with Zinc fragmentation had higher yield in target virus reads for all RNA viruses tested and adenovirus.

Table 1. Internal controls EAV/PhHV-1: serial dilutions against a clinical sample background and within-run precision (INFA)

Sample EAV/PhHV-1 dilution	INFA Cq	EAV Cq	PhHV-1 Cq	INFA reads (log) Centrifuge	EAV reads (log) Centrifuge	PhHV-1 reads (log) Centrifuge
1:100	24.52	21.59	23.52	4438 (3.6)	12925 (4.1)	347 (2.5)
1:1,000	24.67	24.91	26.83	3742 (3.6)	1202 (3.1)	49 (1.7)
1:10,000	24.76	28.45	30.33	4628 (3.7)	95 (2.0)	14 (1.1)
1:100,000	24.79	30.85	32.55	4093 (3.6)	18 (1.3)	14 (1.1)

Abbreviations:, Cq: quantification cycle value, INFA: influenza A, EAV: equine arteritis virus, PhHV-1 phocine herpesvirus 1.

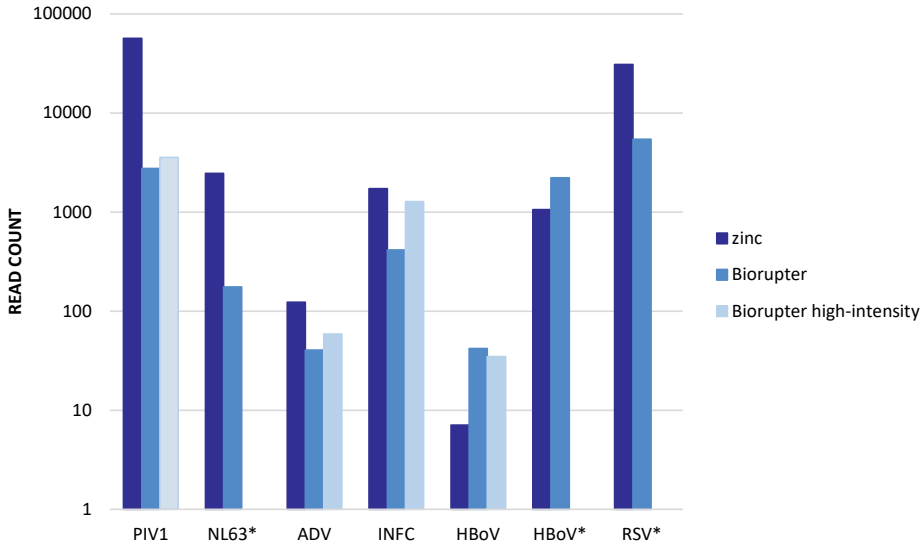

Figure 2. Comparison of fragmentation methods on target reads (species level, log scale).
*Not tested with Bioruptor setting high intensity.
PIV parainfluenza, NL63: coronavirus NL63, ADV: adenovirus, INFC: influenza C , hBoV: human bocavirus, RSV: respiratory syncytial virus

Detection limit

The detection threshold of our NGS limit, deduced from serial dilutions of influenza A (Figure 3) and EAV (table 1) was comparable with a real time PCR Cq value of >35, corresponding to, approximately <50-250 copies/reaction.

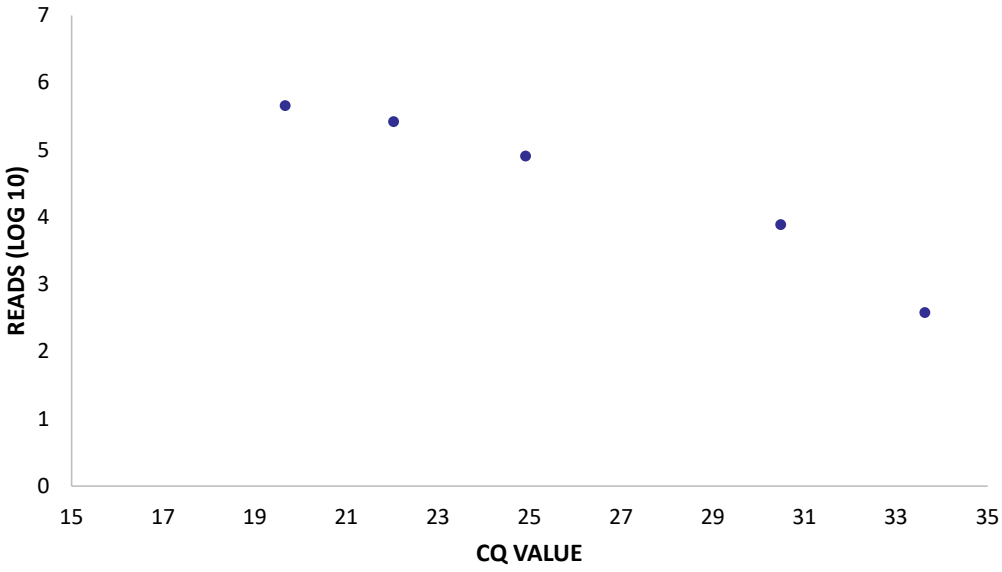

Figure 3. Serial dilutions of an influenza A positive clinical sample.

Repeatability: within run precision

The mNGS results of an influenza A positive sample tested in quadruple could be reproduced with only minor differences (table 1): coefficient of variation of 1.1%: 0.04 log SD/ 3.6 log average.

Bioinformatics: taxonomic classification

The Centrifuge default settings, with NCBI's nucleotide database and assignment of sequence reads to a maximum 5 labels per sequence, resulted in various spurious classifications (Figure 4), for example Lassa virus (Figure 5), evidently highly unlikely to be present in patient samples from the Netherlands with respiratory complaints. The specificity could be increased by using NCBI's RefSeq database instead of NCBI's nucleotide database. The classification was further improved by changing the Centrifuge tool settings to limit the assignment of homologous reads to the lowest common ancestor (maximum 1 label per sequence).

The Centrifuge reporting of shared sequences between different organisms/ subtypes differs dependent of the classification and reporting algorithm. The default classification will assign a shared read to a maximum of 5 organisms (one read will be assigned 5 times) and with the lowest common ancestor classification setting this read will only be assigned once, namely to the lowest ancestor these organisms/ subtypes have in common. Classification with maximum 5 labels per read resulted in two different outcomes using the report with all mappings and the report with unique mappings, with the latter not reporting the reads assigned to multiple organisms.

Comparison of classification using these different settings shows the highest sensitivity and specificity using NCBI's RefSeq database with one label (lowest common ancestor) assignment, both with *in silico* prepared datasets containing solely EAV sequence fragments (Figure 4) and with clinical datasets (with highly abundant background) (Figure 5).

To determine the effect of the total number of sequencing reads obtained per sample on sensitivity, one million and 10 million total reads were compared by *in silico* analysis (Table 2). One million total reads resulted in an approximate tenfold decrease in target virus read count as compared to 10 million total reads, implicating a reduction of sensitivity.

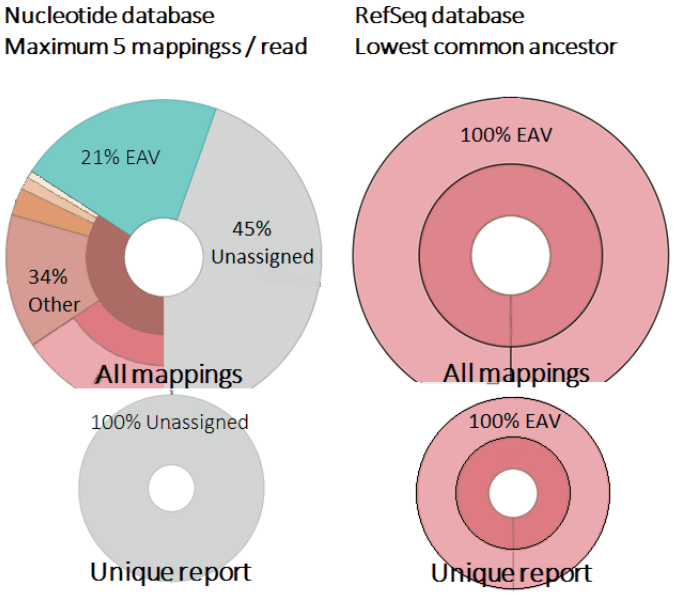

Figure 4. Analysis of *in silico* simulated EAV reads with the different bioinformatic settings of the Centrifuge pipeline.

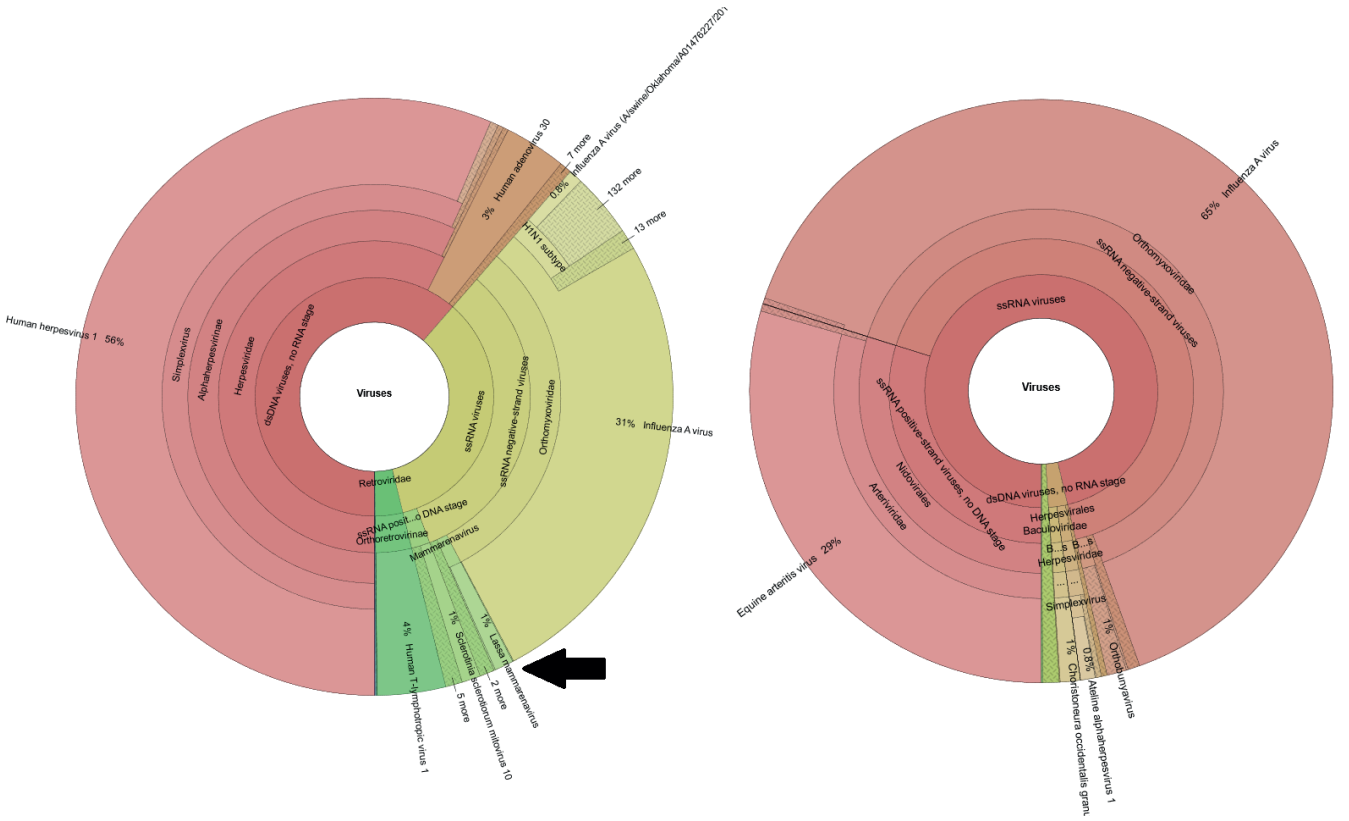

Figure 5. Spurious Lassa virus reads detected using NCBI's Nucleotide (NT) database, versus NCBI's RefSeq database.
Black arrow points to the spurious Lassa virus reads.

Table 2. **Comparison of analysis of 1 million vs 10 million reads.**

virus	virus family	Cq value	10 million reads Total reads	virus family reads	% of total	% of viral	1 million reads Total reads	virus family reads	% of total	% of viral
RV	*Picornaviridae*	37.7	8203894	8941	0.06	84.37	822218	889	0.07	86.11
PIV4	*Paramyxoviridae*	24.9	10886798	2136	0.04	41.90	1088067	199	0.08	40.73
CMV	*Herpesviridae*	34.5	15889428	22	00.01	10.88	1588922	2	0.04	11.87
ADV	*Adenoviridae*	30.2	11146488	0	0	0	1115135	0	0	0
RSV	*Pneumoviridae*	27.3	10191995	1477	0.02	53.29	1019415	163	0.04	59.25
INFB	*Orthomyxoviridae*	30	8535672	652	0.01	48.67	853149	61	0.02	46.58
NL63	*Coronaviridae*	36.2	10386928	0	0	0	1038469	0	0	0
INFA	*Orthomyxoviridae*	27.5	10981601	8403	0.11	70.28	1097872	855	0.17	69.84
MPV	*Pneumoviridae*	34.1	12972626	2	0	0.10	1297151	0	0	0
HBOV	*Parvoviridae*	32.2	11819805	0	0	0	1181738	0	0	0
RV	*Picornaviridae*	23.1	11819805	58695	0.42	84.27	1183738	5754	0.49	84.25

Abbreviations: Cq: quantification cycle value, % of total: percentage of total reads, % of viral: percentage of all viral reads , RV: rhinovirus, PIV4: parainfluenza 4, CMV: cytomegalovirus, ADV: adenovirus, RSV: respiratory syncytial virus, INF: influenza, NL63: coronavirus NL63, MPV: metapneumovirus, hBoV: human bocavirus

Retrospective validation

Clinical sensitivity based on PCR target pathogens

Clinical sensitivity was analysed using the optimized mNGS procedure. The sample collection consisted of 21 clinical specimens positive for at least one of the following PCR target viruses: rhinovirus, influenza A&B, parainfluenza 1 &4 (PIV), metapneumovirus, respiratory syncytial virus, coronaviruses NL63 and HKU1 (CoV), human bocavirus (hBoV), and adenovirus (ADV). Fourteen samples were positive for one virus, six samples for two and one sample for three viruses with the lab-developed respiratory multiplex qPCR. Cq values ranged from Cq 17 to Cq 35, with a median of 23.

With mNGS 24 of the 29 viruses demonstrated in routine diagnostics were detected (Table 3), resulting in a sensitivity of 83% for PCR targets. If a cut-off of 15 reads was handled, sensitivity declined to 66% (19/29) (Table 4). A Receiver-operating Characteristic (ROC) curve for mNGS detection of PCR target viruses, depending on the cut-off level of the number of mapped sequence reads for defining a positive result, is shown in Figure 6.

mNGS target read count (log value) showed a correlation (Pearson correlation coefficient -0.582, p=0.003), with the Cq values of the qPCR (Figure 7).

Table 3. **Detection of qPCR viruses positive respiratory samples with mNGS**

	Routine diagnostics		Metagenomic NGS			
Material	**PCR positive**	**Cq values**	**Virus genus**	**Genus reads***	**Virus species**	**Species reads***
NP wash	RV	30.7	*Enterovirus*	0	*Rhinovirus*	0
	PIV1	17.1	*Respirovirus*	58619	*Human respirovirus 1*	56407
	ADV	33.6	*Mastadenovirus*	0	*Human mastadenovirus C*	0
NP wash	MPV	24	*Metapneumovirus*	127	*Human metapneumovirus*	123
BAL	NL63	24.4	*Alphacoronavirus*	1999	*Human coronavirus NL63*	2176
	HKU1	28.2	*Betacoronavirus*	1	*Human coronavirus HKU1*	1
Sputum	RV	32	*Enterovirus*	2326	*Rhinovirus C*	2204
NP wash	INFA	22.2	*Alphainfluenzavirus*	1490	*Influenza A virus* (A/California/07/2009 (H1N1))	1490
NP wash	MPV	33.4	*Metapneumovirus*	1	*Human metapneumovirus*	3
	ADV	19.3	*Mastadenovirus*	125	*Human mastadenovirus C*	123
Sputum	PIV4	21	*Orthorubulavirus*	7729	*Human rubulaviurs 4* (subtype a)	6798
NP wash	HBoV	22.3	*Bocaparvovirus*	7	Human bocavirus	7
NP wash	MPV	22.2	*Metapneumovirus*	139	*Human metapneumovirus*	312
NP wash	INFB	16.5	*Betainfluenzavirus*	4971	*Influenza B virus* (B/Lee/1940)	4971
NP wash	RV	25.4	*Enterovirus*	8	*Rhinovirus A*	6
	RSV	30.7	*Orthopneumovirus*	32	*Human orthopneumovirus*	32
NP wash	INFB	21.4	*Betainfluenzavirus*	2686	*Influenza B virus* (B/Lee/1940)	2686
NP wash	RSV	17.8	*Orthopneumovirus*	29900	*Human orthopneumovirus*	22483
NP wash	RV	34.4	*Enterovirus*	0	*Rhinovirus*	0
	INFB	22.6	*Betainfluenzavirus*	68972	*Influenza B virus* (B/Lee/1940)	68972
BAL	INFB	34.8	*Betainfluenzavirus*	0	*Influenza B virus*	0
	HBoV	34.1	*Bocaparvovirus*	0	Human bocavirus	0

NP wash	HKU1	24.3	*Betacoronavirus*	534	*Human coronavirus HKU1*	535
NP wash	RV	16.8	*Enterovirus*	3877	*Rhinovirus A*	1721
NP wash	RV	27.4	*Enterovirus*	1	*Rhinovirus B*	2
	HBoV	19	*Bocaparvovirus*	1014	Human bocavirus	1064
NP wash	INFA	22.1	*Alphainfluenzavirus*	657	*Influenza A virus* (A/California/07/2009 (H1N1))	657
NP wash	RSV	17.2	*Orthopneumovirus*	31179	*Human orthopneumovirus*	72
NP wash	RV	17.7	*Enterovirus*	50642	*Rhinovirus A*	29293

Abbreviations: NGS: next-generation sequencing, nr: number, Cq: quantification cycle value, NP wash: nasopharyngeal wash, BAL: bronchoalveolar lavagae, RV: rhinovirus, PIV parainfluenza, ADV: adenovirus, MPV: metapneumovirus, NL63: coronavirus NL63, HKU1: coronavirus HKU1, INF: influenza , hBoV: human bocavirus, RSV: respiratory syncytial virus

*number of reads assigned to the genus or species of the target virus

Table 4. Sensitivity and specificity of the mNGS protocol tested, based on PCR target viruses, with different sequence read cut-off levels for defining a positive result.

	All reads	≥15 sequence reads	≥50 sequence reads
Sensitivity	83 (24/29)	66 (19/29)	62 (18/29)
Specificity	94 (325/346)	100 (345/346)	100 (346/346)

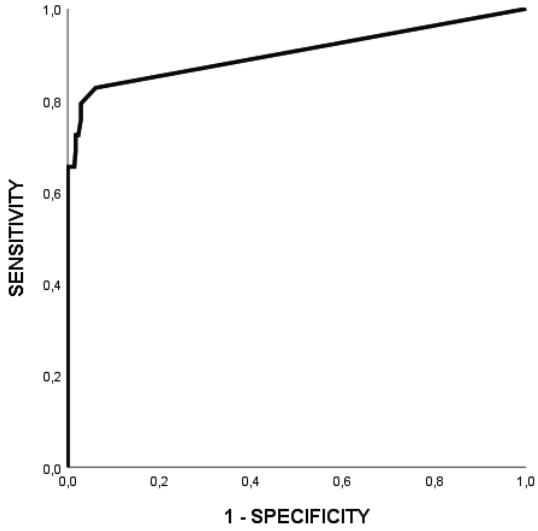

Figure 6. Receiver-operating characteristic (ROC) curve for mNGS detection of PCR target viruses depending on the cut-off level of the number of mapped sequence reads for defining a positive result

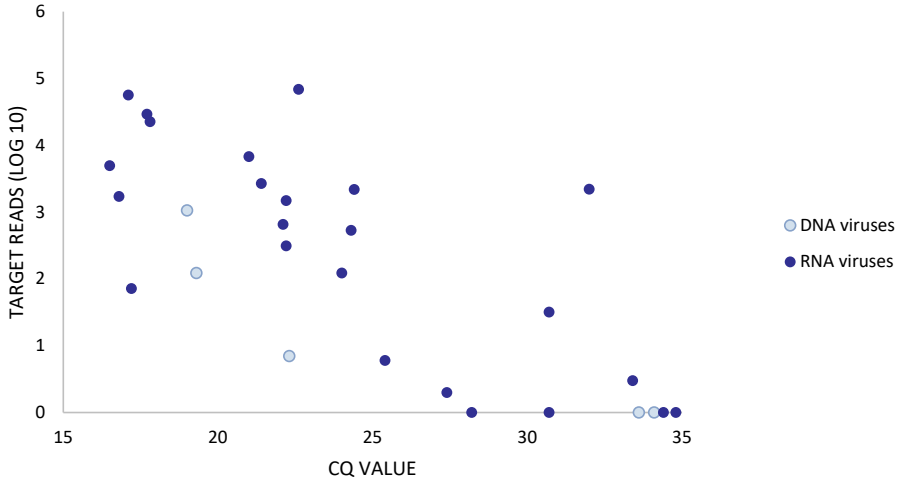

Figure 7. Semi-quantification of the mNGS assay for target virus detection in clinical samples with qPCR confirmed human respiratory viruses.

Detection of additional viral pathogens by mNGS: off-PCR target viruses

Next to the viral pathogens tested by PCR, mNGS also detected other pathogenic viruses, indicating additional viral sequences uncovered by mNGS but not included in the routine diagnostics, with influenza C virus being the most prominent. A high amount,2221 reads (99% horizontal coverage), of influenza virus C reads (58% of all viral reads and 0.02 of the total reads) was found in one sample, confirmatory PCR was not routinely available. Other potential respiratory pathogens detected by mNGS and not included in PCR analysis were KI polyomavirus (2 samples: 262 and 46 reads respectively, retrospective in-house PCR Cq 25 (1:10 dilution) and 26 respectively), cytomegalovirus (human betaherpesvirus 5) (55 and 3 reads, retrospective in-house PCR Cq 22 and 27 respectively) and enterovirus (10073 reads, retrospective in-house PCR rhinovirus/ enterovirus Cq 18). All of these viruses are not included routinely in the diagnostic multiplex qPCRs.

Internal controls

The spiked-in internal controls were detected by mNGS in all samples. EAV sequence reads ranged from 14 - 19894 (median 362) and PhHV1 ranged from 41 - 1206 (median 121).

Analytical specificity based on PCR target viruses

In total 25 paediatric respiratory samples were available to evaluate the analytical specificity of mNGS: 4 samples were negative for all 15 viral pathogens in the multiplex PCR panel (influenza A/B,

RSV, HMPV, ADV, HBoV, PIV1/2/3/4, RV, HKU1, NL63, 227E, OC43) and 21 samples were negative for 12-14 of these PCR target pathogens.

Out of in total 346 negative target PCR results of these 25 samples, 325 results corresponded with the finding of 0 target specific reads by mNGS. If a cut-off of 15 reads was used 345 of the 346 negative PCR targets were negative with mNGS. The sample positive by mNGS and negative by PCR was human parainfluenzavirus 3 (18 reads). Though no conclusive proof for neither true or false positive mNGS results could be found, specificity of mNGS was 94% (325/346) when encountering all reads and ≥99% (345/346) with a 15 reads cut-off (Table 4, ROC curve in Figure 6).

Antiviral susceptibility

Additional to subtyping (Table 3), using the metagenomic sequence data we analysed the nucleotide positions that conferred resistance to either oseltamivir or zanamivir. Sequence data of amino acids I117, E119, D198, I222, H274, R292, N294 and I314 showed susceptibility to oseltamivir and V116, R118, E119, Q136, D151, R152, R224, E276, R292 and R371 revealed susceptibility to zanamivir[31,32].

Data access

The raw sequence data of the samples, after removal of human reads have been deposited to Sequence Read Archive database (https://www.ncbi.nlm.nih.gov/sra; accession number SRX6715205-SRX6715229).

DISCUSSION

Metagenomic sequencing has not yet been implemented as routine tool in clinical diagnostics of viral infections. Such application would require the careful definition and validation of several parameters to enable the accurate assessment of a clinical sample with regard to the presence or absence of a pathogen, in order to fulfil current accreditation guidelines. For this purpose, this study has initiated the optimization of several steps throughout the pre- and post-sequencing workflow, which are considered essential for sensitive and specific mNGS based virus detection. Many virus discovery or virus diagnostic protocols have focussed on the enrichment of viral particles[33] with the intention to increase the relative amount of virus reads. However, these methods are laborious and intrinsically exclude viral nucleic acid located in host cells. Here, a sample pre-treatment protocol was designed with potential for: 1) automation, 2) pan-pathogen detection and 3) detection of intracellular viral nucleic acids. Consequently, any type of viral enrichment was excluded (filtration, centrifugation, nucleases, rRNA removal). The current protocol enabled high throughput sample pre-treatment by means of automated NA extraction and without depletion of bacterial nor human genome, with potential for pan-pathogen detection. Several adaptations in the bioinformatic script resulted in more accurate reporting of the classification output.

Addition of an internal control to a PCR reaction is commonly used for quality control in qPCR[34]. While the addition of internal controls in mNGS is not yet an accepted standard procedure, we employed EAV and PhHV1 as an RNA and DNA control, respectively, to monitor the workflow in this diagnostic application. The amount of internal control reads and target virus reads have been reported to be dependent of the amount of background reads (negative correlation)[35]. In our protocol, the internal controls were used as qualitative controls but may be used as indicator of the amount of background. PhHV1 showed less linearity in the dilution series, as compared to EAV, which may be indicative for a potential relative difference in efficiency of amplification of PhHV1 viral sequences. Since NCBI's databases were lacking a complete PhHV1 genome, the Centrifuge index building and classification was limited to classification on a higher taxonomic rank. In order to achieve classification of PhHV1 at species level, the whole genome of PhHV1 was sequenced, and based on the gained sequence reads the genome was built[26]. The proposed nearly complete genome of PhHV1 was submitted to NCBI's GenBank database.

Sensitivity of the mNGS protocol was maximum 83% based on PCR target viruses and depended on the cut-off level of reads for defining a positive result. Five viruses, that were not recovered by mNGS had high Cq values, over 30, i.e. a relatively low viral load. This may be a drawback of the retrospective nature of this clinical evaluation as RNA viruses may be degraded due to storage and freeze-thaw steps, resulting in lower sensitivity of mNGS. A correlation was found between read counts and PCR Cq value, demonstrating the quantitative nature of viral detection by mNGS. Discrepancies between the Cq values and the number of mNGS reads may be explained by 1) unrepresentative Cq values, e.g. by primer mismatch for highly divergent viruses like rhino/enteroviruses and 2) differences in sensitivity of mNGS for several groups of viruses, as has been reported by others[36]. Additionally, viral pathogens were detected that were not targeted by the routine PCR assays, including influenza C virus, which is typical of the unbiased nature of the method. In addition, though not within the scope of this study, bacterial pathogens, including *Bordetella pertussis* (qPCR confirmed), were also detected. In the current study only viruses were targeted since these could be well compared to qPCR results, bacterial targets remain to be studied in clinical sample types as sputum or broncheo-alveolar lavages that are more suitable for bacterial detection. The analytical specificity of mNGS appeared to be high, especially with a cut-off of 15 reads. However, the clinical specificity, the relevance of the lower read numbers, still needs further investigation in clinical studies.

Sequencing using Illumina HiSeq 4000 with single, unique indexes resulted in rhinovirus-C sequences (55-909 reads) in all samples run on one lane, which appeared to be identical sequences. Retesting of the samples with Illumina Nextseq 500 resulted in disappearance of these reads. This problem could be attributed to 'index hopping' (index misassignment) as described earlier[37]. Due to the chemistry, essential for the increased speed, the HiSeq 4000 is more prone to index hopping between neighbouring samples. Although the percentage of reads which contributed to the index hopping was very low, this is critical for clinical viral diagnostics, as this is aimed specifically at low abundance targets[37,38].

Bioinformatics classification of metagenomic sequence data with the pipeline Centrifuge required identification of the optimal parameters in order to minimize misclassified and unclassified reads.

Default settings of this pipeline resulted in higher rates of both false positive and false negative results. NCBI's nucleotide database includes a wide variety of unannotated viral sequences, such as partial sequences and (chimeric) constructs, in contrast to the curated and well-annotated sequences in NCBI's Refseq database, which resulted in a higher specificity. In addition to the database, settings for the assignment algorithm were adapted as well. The assignment settings were adjusted to unique assignment in the case of homology to the lowest common ancestor. This modification resulted in higher sensitivity and specificity than the default settings, however the ability to further subtyping diminished. This is likely to be attributed to the limited representation/availability of strain types within NCBI's RefSeq database. In consequence, this leads to a more accurate estimation of the common ancestor for particular viruses, but limited typing results in case of highly variable ones. To obtain optimal typing results, additional annotated sequences may be added or a new database should be built, with a high variety of well-defined and frequently updated virus strain types.

To conclude, this study contributes to the increasing evidence that metagenomic NGS can effectively be used for a wide variety of diagnostic assays in virology, such as unbiased virus detection, resistance mutations, virulence markers, and epidemiology, as shown by the ability to detect SNPs in influenza virus.

These findings support the feasibility of moving this promising field forward to a role in the routine detection of pathogens by the use of mNGS. Further optimization should include the parallel evaluation of adult samples, the inclusion of additional annotated strain sequences to the database, and further elaboration of the classification algorithm and reporting for clinical diagnostics. The importance of both negative non-template control samples[39] and healthy control cases may support the critical discrimination of contaminants and viral 'colonization' from clinically relevant pathogens.

Conclusions

Optimal sample preparation and bioinformatics analysis are essential for sensitive and specific mNGS based virus detection.

Using a high-throughput genome extraction method without viral enrichment, both RNA and DNA viruses could be detected with a sensitivity comparable to PCR.

Using mNGS, all potential pathogens can be detected in one single test, while simultaneously obtaining additional detailed information on detected viruses. Interpretation of clinical relevance is an important issue but essentially not different from the use of PCR based assays and supported by the available information on typing and relative quantities. These findings support the feasibility of a role of mNGS in the routine detection of pathogens.

ACKNOWLEDGEMENTS

We thank our project partners Floyd Wittink, Wouter Suring (Hogeschool Leiden), Danny Duijsings (BaseClear) and Christiaan Henkel (Leiden University). We also like to thank Tom Vreeswijk, Lopje Höcker and Mario van Bussel clinical microbiology laboratory, LUMC) for help with the pre-sequencing experiments, Jeroen Laros (Human Genetics, LUMC) for help with the bioinformatics.

REFERENCES

1. Lozano R, Naghavi M, Foreman K, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet (London, England)* 2012; **380**(9859): 2095-128.
2. Nair H, Simoes EA, Rudan I, et al. Global and regional burden of hospital admissions for severe acute lower respiratory infections in young children in 2010: a systematic analysis. *Lancet (London, England)* 2013; **381**(9875): 1380-90.
3. Bates M, Mudenda V, Mwaba P, Zumla A. Deaths due to respiratory tract infections in Africa: a review of autopsy studies. *Current opinion in pulmonary medicine* 2013; **19**(3): 229-37.
4. Jain S, Self WH, Wunderink RG, et al. Community-Acquired Pneumonia Requiring Hospitalization among U.S. Adults. *N Engl J Med* 2015; **373**(5): 415-27.
5. Heikkinen T, Jarvinen A. The common cold. *Lancet (London, England)* 2003; **361**(9351): 51-9.
6. Ieven M, Coenen S, Loens K, et al. Aetiology of lower respiratory tract infection in adults in primary care: a prospective study in 11 European countries. *Clinical microbiology and infection* 2018; **24**(11): 1158-63.
7. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med* 2012; **367**(19): 1814-20.
8. Prachayangprecha S, Schapendonk CM, Koopmans MP, et al. Exploring the potential of next-generation sequencing in detection of respiratory viruses. *Journal of clinical microbiology* 2014; **52**(10): 3722-30.
9. Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. *PLoS computational biology* 2010; **6**(2): e1000667.
10. Hoffmann B, Scheuch M, Hoper D, et al. Novel orthobunyavirus in Cattle, Europe, 2011. *Emerging infectious diseases* 2012; **18**(3): 469-72.
11. Mongkolrattanothai K, Naccache SN, Bender JM, et al. Neurobrucellosis: Unexpected Answer From Metagenomic Next-Generation Sequencing. *Journal of the Pediatric Infectious Diseases Society* 2017; **6**(4): 393-8.
12. van Boheemen S, de Graaf M, Lauber C, et al. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. *mBio* 2012; **3**(6).
13. Kohl C, Brinkmann A, Dabrowski PW, Radonic A, Nitsche A, Kurth A. Protocol for metagenomic virus detection in clinical specimens. *Emerging infectious diseases* 2015; **21**(1): 48-57.
14. Parker J, Chen J. Application of next generation sequencing for the detection of human viral pathogens in clinical specimens. *Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology* 2017; **86**: 20-6.
15. Zou X, Tang G, Zhao X, et al. Simultaneous virus identification and characterization of severe unexplained pneumonia cases using a metagenomics sequencing technique. *Science China Life sciences* 2017; **60**(3): 279-86.
16. Wylie KM, Mihindukulasuriya KA, Sodergren E, Weinstock GM, Storch GA. Sequence analysis of the human virome in febrile and afebrile children. *PloS one* 2012; **7**(6): e27735.
17. Scheltinga SA, Templeton KE, Beersma MF, Claas EC. Diagnosis of human metapneumovirus and rhinovirus in patients with respiratory tract infections by an internally controlled multiplex real-time RNA PCR. *Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology* 2005; **33**(4): 306-11.
18. Kalpoe JS, Kroes AC, de Jong MD, et al. Validation of clinical application of cytomegalovirus plasma DNA load measurement and definition of treatment criteria by analysis of correlation to antigen detection. *Journal of clinical microbiology* 2004; **42**(4): 1498-504.
19. Wery M, Descrimes M, Thermes C, Gautheret D, Morillon A. Zinc-mediated RNA fragmentation allows robust transcript reassembly upon whole transcriptome RNA-Seq. *Methods (San Diego, Calif)* 2013; **63**(1): 25-31.
20. Joshi NA, Fass JN. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. 2011.
21. Martin M. Cutadept removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 2011; **17**(10).
22. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome research* 2016; **26**(12): 1721-9.
23. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* 2016; **44**(D1): D733-45.
24. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. *BMC bioinformatics* 2011; **12**: 385.
25. Vilsker M, Moosa Y, Nooij S, et al. Genome Detective: An Automated System for Virus Identification from High-throughput sequencing data. *Bioinformatics* 2018.

26. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology* 2012; **19**(5): 455-77.
27. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM; 2013.
28. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* 2009; **25**(16): 2078-9.
29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of molecular biology* 1990; **215**(3): 403-10.
30. Loens K, van Loon AM, Coenjaerts F, et al. Performance of different mono- and multiplex nucleic acid amplification tests on a multipathogen external quality assessment panel. *Journal of clinical microbiology* 2012; **50**(3): 977-87.
31. Orozovic G, Orozovic K, Lennerstrand J, Olsen B. Detection of resistance mutations to antivirals oseltamivir and zanamivir in avian influenza A viruses isolated from wild birds. *PloS one* 2011; **6**(1): e16028.
32. Hsieh NH, Lin YJ, Yang YF, Liao CM. Assessing the oseltamivir-induced resistance risk and implications for influenza infection control strategies. *Infection and drug resistance* 2017; **10**: 215-26.
33. Hasan MR, Rawat A, Tang P, et al. Depletion of Human DNA in Spiked Clinical Specimens for Improvement of Sensitivity of Pathogen Detection by Next-Generation Sequencing. *Journal of clinical microbiology* 2016; **54**(4): 919-27.
34. Ninove L, Nougairede A, Gazin C, et al. RNA and DNA bacteriophages as molecular diagnosis controls in clinical virology: a comprehensive study of more than 45,000 routine PCR tests. *PloS one* 2011; **6**(2): e16142.
35. Schlaberg R, Chiu CY, Miller S, Procop GW, Weinstock G. Validation of Metagenomic Next-Generation Sequencing Tests for Universal Pathogen Detection. *Archives of pathology & laboratory medicine* 2017; **141**(6): 776-86.
36. Bal A, Pichon M, Picard C, et al. Quality control implementation for universal characterization of DNA and RNA viruses in clinical respiratory samples using single metagenomic next-generation sequencing workflow. *BMC infectious diseases* 2018; **18**(1): 537.
37. Sinha R, Stanley G, Gulati GS, et al. Index Switching Causes "Spreading-Of-Signal" Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing. *bioRxiv* 2017.
38. van der Valk T, Vezzi F, Ormestad M, Dalen L, Guschanski K. Estimating the rate of index hopping on the Illumina HiSeq X platform. *bioRxiv* 2018.
39. Naccache SN, Hackett J, Jr., Delwart EL, Chiu CY. Concerns over the origin of NIH-CQV, a novel virus discovered in Chinese patients with seronegative hepatitis. *Proceedings of the National Academy of Sciences of the United States of America* 2014; **111**(11): E976.