Journal of Virological Methods 283 (2020) 113916

Contents lists available at ScienceDirect

Journal of Virological Methods

journal homepage: www.elsevier.com/locate/jviromet

Evaluation of a commercial exogenous internal process control for diagnostic RNA virus metagenomics from different animal clinical samples

Steven Van Borm[a,*], Qiang Fu[b], Raf Winand[b], Kevin Vanneste[b], Mikhayil Hakhverdyan[c], Dirk Höper[d], Frank Vandenbussche[a]

[a] Department of Animal Infectious Diseases, Sciensano, Groeselenbergstraat 99, 1180, Brussels, Belgium
[b] Transversal Activities in Applied Genomics, Sciensano, Rue Juliette Wytsmanstraat 14, 1050, Brussels, Belgium
[c] SVA, National Veterinary Institute, Ulls väg 2B, 751 89, Uppsala, Sweden
[d] FLI, Friedrich Löffler Institut, Südufer 10, 17493 Greifswald, Germany

ARTICLE INFO

Keywords:
Diagnostic metagenomics
Virology
Next-generation sequencing
Quality control
Mengovirus

ABSTRACT

Metagenomic next generation sequencing (mNGS) is increasingly recognized as an important complementary tool to targeted human and animal infectious disease diagnostics. It is, however, sensitive to biases and errors that are currently not systematically evaluated by the implementation of quality controls (QC) for the diagnostic use of mNGS. We evaluated a commercial reagent (*Mengovirus* extraction control kit, CeraamTools, bioMérieux) as an exogenous internal control for mNGS. It validates the integrity of reagents and workflow, the efficient isolation of viral nucleic acids and the absence of inhibitors in individual samples (verified using a specific qRT-PCR). Moreover, it validates the efficient generation of viral sequence data in individual samples (verified by normalized mengoviral read counts in the metagenomic analysis). We show that when using a completely random metagenomics workflow: (1) *Mengovirus* RNA can be reproducibly detected in different animal sample types (swine feces and sera, wild bird cloacal swabs), except for tissue samples (swine lung); (2) the *Mengovirus* control kit does not contain other contaminating viruses that may affect metagenomic experiments (using a cutoff of minimum 1 Kraken classified read per million (RPM)); (3) the addition of 2.17×10^6 *Mengovirus* copies/mL of sample does not affect the virome composition of pig fecal samples or wild bird cloacal swab samples; (4) *Mengovirus* Cq values (using as cutoff the upper limit of the 99 % confidence interval of Cq values for a given sample matrix) allow the identification of samples with poor viral RNA extraction or high inhibitor load; (5) *Mengovirus* normalized read counts (cutoff RPM > 1) allow the identification of samples where the viral sequences are outcompeted by host or bacterial target sequences in the random metagenomic workflow. The implementation of two QC testing points, a first one after RNA extraction (Mengoviral qRT-PCR) and a second one after metagenomic data analysis provide valuable information for the validation of individual samples and results. Their implementation in addition to external controls validating runs or experiments should be carefully considered for a given sample type and workflow.

1. Introduction

The use of metagenomic approaches, here defined as the unbiased sequencing of the complete genomic content of a sample, has increased dramatically with the wider availability of Next-Generation Sequencing (NGS) platforms. The power of its sequence-independent and hypothesis-free characterization is highly recognized as an important complementary tool to targeted assays in research and diagnostic virology laboratories (Chiu and Miller, 2019; Höper et al., 2017). It has led to important discoveries in animal and human virology including (1) the characterization of multifactorial diseases (Blomström et al., 2016; Qin et al., 2018); (2) the characterization of uncultivable pathogens (Taylor-Brown et al., 2017); and (3) the identification and characterization of novel or unexpected pathogens (Epstein and Anthony, 2017; Hoffmann et al., 2012; Zhu et al., 2020). However, factors related to sample processing, sequencing and data analysis have been documented to affect the specificity and/or sensitivity of metagenomic NGS (mNGS) workflows and may lead to misinterpretation of the resulting datasets in a clinical or epidemiological context (Greninger, 2018). Sample processing related issues include inhibition, competition (e.g. for host nucleic acids), reagent contamination (Naccache et al., 2013; Rosseel et al., 2014), and background (Bukowska-Ośko et al., 2016) and

* Corresponding author.
E-mail address: Steven.Vanborm@sciensano.be (S. Van Borm).

https://doi.org/10.1016/j.jviromet.2020.113916
Received 15 April 2020; Received in revised form 9 June 2020; Accepted 9 June 2020
Available online 20 June 2020

laboratory (Salter et al., 2014) contamination. Sequencing related issues including carry-over contamination between runs and sample crosstalk due to index bleedthrough in multiplex Illumina sequencing (Kircher et al., 2012; Wright and Vetsigian, 2016) have been documented. Data analysis biases may be related to the metagenomic classification algorithms or the databases used.

In addition to proper sampling and laboratory practice, and adhering to 'best practices' for bioinformatics (ten Hoopen et al., 2017), the implementation of quality controls (QC) can improve the interpretation of mNGS data. Despite the increasing use of mNGS in diagnostic laboratories, QC are not systematically implemented. Control strategies commonly used in targeted molecular diagnostic assays (like realtime PCR) include the parallel processing of external positive and negative controls validating the run; and the inclusion of exogenous (spike-in) or endogenous (inherent part of a sample such as host genes) internal controls validating the processing of an individual sample (e.g. Vandenbussche et al., 2010). These concepts have previously been extrapolated to mNGS virus detection (Bal et al., 2018; van Boheemen et al., 2020; Lewandowski et al., 2019), allowing run-level as well as sample-level validation of the workflow using either qRT-PCR assays to measure successful nucleic acid isolation of the spike-in control and/or evaluation of the spike-in sequence reads generated in the workflow.

The objective of this study was to evaluate a commercially available RNA virus control reagent as internal exogenous control for use in a random metagenomic protocol (without virus enrichment) and to assess its suitability and reproducibility in different clinical veterinary sample types.

2. Methods

2.1. Samples

We selected four sample matrices commonly used in the surveillance of priority animal health viruses. These included cloacal swabs from 28 live captured healthy wild birds (12 different species, Supplementary Table 1) that had tested negative for avian influenza and Newcastle disease in active wild bird screening (October-November 2018). Swabs were immersed in viral transport medium and transported to the laboratory cooled, where they were kept at −80 °C until further testing. Twenty-two swine serum samples that had tested negative for porcine reproductive and respiratory syndrome virus were taken from ongoing serosurveillance efforts. Five swine lung tissue samples were taken from a biobank representing negative control animals from previous animal experiments. Twenty-one fecal samples were taken rectally in sterile containers from sows in three production farms without health issues, two additional samples were sourced from a negative control group of piglets in Sciensano's experimental farm. One pig fecal sample with known *Astrovirus* positive status was sourced from a proficiency test panel (OneHealth EJP internal project METASTAVA). All samples were sourced from existing sample collections (no animal experiments were organized for the present study).

2.2. Quality control implementation

Internal Quality control (IQC): 2.17×10^6 copies (based on the lot specific quantification report supplied by the manufacturer) of *Mengovirus* strain vMC0 (Mengo Extraction control kit, CeraamTools, bioMérieux) was added to 1 mL of liquid sample prior to extraction (serum and swab supernatant) or to the solid sample prior to homogenization in 1 mL of PBS (feces, lung). Negative Extraction Controls (NEC) consisted of 1 mL of molecular biology grade water spiked with 2.17×10^6 copies of *Mengovirus*. *Mengovirus* qRT-PCR detection in purified RNA was performed for a total of 5 lung samples, 22 fecal samples, 28 swabs, and 22 sera. To characterize the inherent virome of the *Mengovirus* control reagent, and for use as a negative process control, 2.17×10^6 copies of *Mengovirus* control was added to 1 mL of molecular biology grade water and processed as described for liquid (serum) samples (negative extraction control, NEC; two tested replicates). Two sample level QC measurements were implemented: (1) IQC-RNA consisting of the realtime RT-PCR detection of mengoviral RNA using the accessory reagents in the *Mengovirus* control kit according to the manufacturer's instruction (bioMérieux); (2) IQC-SEQ consisting of the evaluation of the sequencing performance as the normalized number of reads classified by Kraken as *Mengovirus* (*Mengovirus* reads per million of reads in the raw dataset, RPM). Sample level acceptance criteria for IQC-RNA were set at the upper limit of the 99 % confidence interval (CI) of the Cq values for a given sample matrix (assuming a normal distribution of Cq values), while sample level acceptance criteria for IQC-SEQ were arbitrarily set at > 1 RPM. A comparison between unspiked and *Mengovirus* spiked replicates was done for three wild bird swabs and a single swine fecal sample, the latter including two unspiked and one spiked replicate(s).

2.3. Metagenomic sequencing workflow

Solid samples (lung, feces) were homogenized (10 % wt/vol) in sterile precooled (4 °C) 1x phosphate-buffered saline for 2 × 3 min at 30 Hz using a TissueLyser II (Qiagen) and pre-cooled (4 °C) aluminum sample blocks. Solid sample homogenates and bird swab samples were centrifuged for 2 min at 10,000 *g* in a precooled (4 °C) centrifuge to separate the supernatants. Total RNA was extracted from 250 µL of supernatants or serum sample using a combination of TRIzol LS reagent (ThermoFisher) and RNeasy Mini Kit as previously described (Wylezich et al., 2018). cDNA was synthesized from 8 µL of purified RNA using Superscript IV reverse transcriptase (ThermoFisher), followed by double stranding using the NEBNext® Ultra II Non-Directional RNA Second Strand Synthesis Module (New England Biolabs) and cDNA cleanup using Zymo DNA clean & concentrator-5 (Zymo Research), all according to manufacturer's instructions. Sequencing libraries were prepared using the Nextera XT DNA Library Preparation Kit (Illumina) and standard Nextera XT index primers according to the manufacturer's instructions using 1 ng (or the maximum available amount) of double stranded cDNA. The fragment length distributions of the libraries were verified on a Bioanalyzer system (Agilent Technologies), and libraries were quantified using a library quantification kit (Kapa Biosystems). Sequencing was performed using a MiSeq reagent kit version 3 (Illumina) with 2 × 300-bp paired-end sequencing on a MiSeq system (Illumina), aiming for a minimum of 3.0×10^6 read pairs per library. Metagenomic NGS data was generated for a subset of 12 swabs, 9 fecal samples, 8 sera, 2 lung samples, and two negative extraction control (NEC) samples. The resulting fastq raw datasets are publicly available in the Sequence Read Archive (SRA) under BioProject accession number PRJNA615303.

2.4. Bioinformatic analysis

Raw reads for all samples were trimmed and quality controlled, followed by taxonomic classification and read mapping to the *Mengovirus* reference genome (to confirm the validity of using Kraken classification for *Mengovirus* QC implementation).

2.5. QC & trimming

A two-step trimming strategy including Trimmomatic v0.38 (Bolger et al., 2014) to remove adapter sequences and low quality bases (setting the 'ILLUMINACLIP 2:30:10′, 'LEADING:5′, 'TRAILING:5′, 'SLIDINGWINDOW:4:10′, and 'MINLEN:20′ options), followed by low-complexity sequence information removal using PRINSEQ v0.20.4 (Schmieder and Edwards, 2011) (setting the '-trim_tail_right 10′ and '-trim_tail_left 10′ options), was implemented. Only paired reads were retained for further analysis.

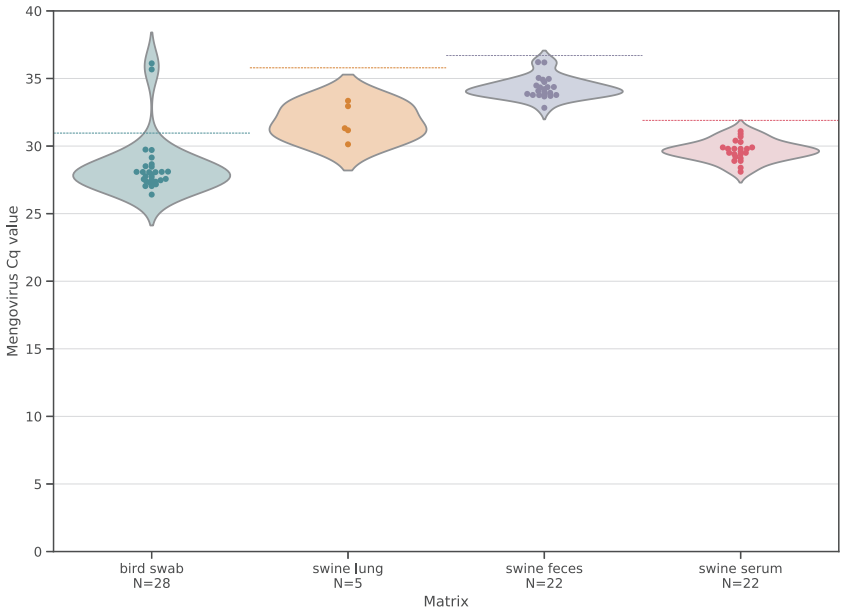

Fig. 1. Violin plot with superimposed swarm plot showing the distribution of *Mengovirus* qRT-PCR Cq values for each sample matrix. The Cq acceptance criterion (upper limit of the 99 % CI) is plotted as a dotted line for each sample matrix.

2.6. Taxonomic analysis

Classification of trimmed reads was performed with Kraken v1.1 (Wood and Salzberg, 2014). A customized Kraken database was built using all available RefSeq "Complete Genome" sequences of six targeted taxonomic groups (archaea, bacteria, fungi, human, protozoa, and viral) downloaded from RefSeq Genome (O'Leary et al., 2016) (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/) on 18/02/2019. An overview of the number of sequences sampled per order can be found in Supplementary Table 2. Read counts per taxonomic level classified by Kraken were further normalized as reads per million (RPM) total (trimmed) reads based on the following formula to remove technical bias introduced by sequencing depth variation between samples:

$$RPM = \frac{Total\ number\ of\ reads\ classified \times 1\ million}{Total\ number\ of\ trimmed\ reads}$$

Of note, *Mengovirus* is represented in the NCBI RefSeq Genomes database by *Encephalomyocarditis virus* (EMCV, closest RefSeq complete genome and parent taxon of *Mengovirus* based on the NCBI taxonomy database, NCBI:txid12104) for which Kraken hits were counted in this study. For computational reasons, only the human genome was included as vertebrate species in the database. Data representing bacteriophage species hits were always regrouped at the Order level *(Caudovirales)*. An arbitrary acceptance criterion of RPM > 1 for a significant mNGS finding was used.

2.6.1. Read mapping with Bowtie2 on the Mengovirus genome

Because different candidate entries for the genome sequence of *Mengovirus* existed in the NCBI nucleotide database (L22089.1, DQ294633.1, KX231802.1, KU955338.1), the sequence most similar to *Mengovirus* strain vMC0 (Mengo Extraction control kit, CeraamTools, bioMérieux) was identified as follows. Low-complexity regions of the candidate genomes were identified using sdust v0.1-r2 (https://github.com/lh3/sdust) and masked with bedtools v2.27.1 using the 'maskfasta' function (Quinlan and Hall, 2010) before read mapping to avoid generating spurious alignments. Reads of a NEC sample (molecular biology grade water spiked with 2.17×10^6 copies of *Mengovirus* control) were subsequently aligned to all masked genomes using Bowtie2 v2.3.4.3 (setting the '–very-sensitive-local' option) (Langmead and Salzberg, 2012). Alignment results were visually verified in IGV (Robinson et al., 2011) indicating complete and consistent coverage across the genome for references L22089.1 and DQ294633.1 (Supplementary Fig. 1). Next, variants were identified using samtools v1.9 and bcftools v1.9 (Li, 2011; Li et al., 2009) with the options '–multiallelic-caller', '–variants-only', '–ploidy 1′, and '–prior 0.0011′. Candidate genome DQ294633.1 was selected as the *Mengovirus* reference genome with the lowest number of identified variants. For each sample, trimmed reads were therefore always aligned to DQ294633.1 using Bowtie2 (setting the '–very-sensitive-local' option). Only high-quality alignments with a mapping quality score (MAPQ, a transformation of the probability that a read is aligned wrong) of at least 15 were retained with samtools (setting the '-q 15′ option). Each read pair was counted only once and normalized to RPM to allow comparison with Kraken (where counts are only based on read pairs).

3. Results

3.1. QC implementation

We implemented the Mengo Extraction control kit (CeraamTools, bioMérieux) as a quality control reagent for metagenomics. It consists of a titrated stock of *Mengovirus* strain *vMC0* and an accessory qRT-PCR assay for *Mengovirus* quantification. After an initial titration experiment (0.5×10^6 vs 1×10^6 vs 2.17×10^6 copies/mL pig fecal material homogenate 10 % wt/vol sample, data not shown), 2.17×10^6 *Mengovirus* copies/mL was selected as the lowest spiking level with reproducible qRT-PCR and Kraken RPM results. This concentration corresponds to the manufacturer's guidelines for external control addition for targeted PCR based detection of foodborne viruses. Given the genome size of 7765 nt and the molecular weight of 78,033 KDa (based on *Mengovirus* DQ294633.1), this corresponded to approximately 2.83 ng of *Mengovirus* RNA added to 1 mL of liquid sample or homogenate. Taking the sample matrix specific range of total RNA extraction yields of our protocol into account (swine serum 0.5–10 ng; bird swab 10–60 ng; swine feces 30–115 ng; swine lung 6–9.5 µg), this resulted in approximate *Mengovirus* RNA/total RNA ratios of 0.3–55 for swine sera, $0.05-0.3$ for bird swabs, $0.025-0.10$ for swine feces and $0.0003-0.0005$ for swine lung samples. IQC-RNA: *Mengovirus* RNA can reproducibly be detected in all matrix types with minimal variation (Fig. 1). We propose the upper limit of the 99 % CI of *Mengovirus* Cq

values of a given sample matrix (spiked at 2.17×10^6/mL) as acceptance criterion for successful viral RNA extraction, resulting in the following matrix-specific cutoff values: lung Cq < 35.79; feces Cq < 36.70; swabs Cq < 30.96; serum Cq < 31.90). Two outlier Cq values for swab samples (Cq > 35) were excluded from the dataset to calculate the swab acceptance criterion.

For all sample types where Kraken could consistently detect *Mengovirus* reads (bird swabs, swine feces, swine serum) Kraken *Mengovirus* RPM counts correlated to Bowtie2 RPM counts (R^2 between 0.9831 and 0.9996) but Kraken RPM counts were consistently an order of magnitude lower (Supplementary Fig. 2). Kraken *Mengovirus* RPM counts were therefore selected as standard IQC-SEQ as it is consistent with the metagenomics read classification workflow used for other taxa and represents a conservative measure of *Mengovirus* control reads in a sample dataset.

3.2. Characterization of sample matrix metagenomes

For each of the sequencing libraries resulting from clinical samples, an average of 3.35×10^6 read pairs was generated (range 1.32×10^6 - 4.46×10^6, excluding one weak library from a swab producing < 0.5 $\times 10^6$ read pairs corresponding to an outlier *Mengovirus* Cq value). Negative control samples (water spiked with *Mengovirus*) generated considerably lower number of reads (average 0.67×10^6 read pairs).

As expected when using a random mNGS protocol without viral enrichment, low viral read percentages were observed (Fig. 2; serum: max 0.5 %; swabs: max 0.3 %; fecal: max 4%; lung: max 0.00.3 %). The variability in viral reads in swine fecal samples is related to high loads of pig Astroviruses and *Posavirus 1* in two respective samples. The percentages of unclassified (i.e. no hits) Bacterial and Eukaryotic reads varied substantially between and within sample matrices. In wild bird swabs, the majority of reads is either eukaryotic or bacterial with pronounced variability in both categories. In swine feces, on the other hand, consistently the majority of reads is bacterial. In swine lung tissue virtually all reads are Eukaryotic, for which additional analysis using Kraken against selected mammalian genomes indicated that these can be attributed to the swine genome or RNA it encodes (see Supplementary Table 3). Finally, in swine serum samples, the majority of reads is Eukaryotic (attributable to the swine genome), with the exception of two sera that were bacterially contaminated (respectively with *Pseudomonas* and *Acinetobacter* as dominant taxon).

3.3. Mengovirus control reagent is exempt of extraneous viral sequences influencing mNGS applications

Mengovirus extraction control, when added to molecular biology grade water in the same concentration as to samples, did not contain animal or human viruses besides *Mengovirus* when a QC cutoff of RPM > 1 was applied (Fig. 3). The detected background sequences in these spiked negative control samples were mostly bacterial or eukaryotic. Only read counts of bacteriophages (*Caudovirales*), plant viruses (Citrus endogenous pararetrovirus), and herbivore insect viruses (*Choristoneura fumiferana* granulovirus) were detected above RPM > 1. Individual reads were detected in one of both replicate negative control samples for Influenza A virus, a mammal Papillomavirus, and an algae associated member of the *Phycodnaviridae*, while two human *Mastadenovirus* reads were detected in a single replicate of the negative controls. As our metagenomic interpretation criterion requires RPM > 1, these reads were not considered. Inspection of the Kraken read classification (specificity and distribution of k-mers over reads), as well as Blastn verification of the corresponding reads, for viral taxa detected with RPM > 1 confirmed the correct classification of the detected Caudovirales and the Citrus endogenous pararetrovirus. However, the reads classified by Kraken as *Choristoneura fumiferana* granulovirus (*Baculoviridae*) are most likely misclassifications of fungal 28S sequences (based on a random subsample of 3 out of 18 reads). The classification of 2 reads as human *Mastadenovirus* and one read as Influenza A virus was confirmed in a detailed inspection of the Kraken results and a Blastn of the corresponding reads. The single read classified as *Acanthocystis turfacea* chlorella virus 1 (*Phycodnaviridae*) was a Kraken misclassification of plant chromosomal DNA or corresponding RNA transcript. The bacterial community in both NEC replicates was highly reproducible and dominated by Gammaproteobacteria (39 % and 38 % of bacterial sequences in NEC-1 and NEC-2 respectively; of which 82 % and 76 % represent a single species, *Moraxella osloensis*), Acinetobacteria (27 % and 26 %) and Alphaproteobacteria (23 % and 17 %; of which 59 % and 60 % represent a single genus, *Paracoccus*).

3.4. The addition of Mengovirus does not change the virome reconstruction of a pig fecal sample

The normalized read counts of detected taxa in a *Mengovirus* spiked replicate was within the variation of two replicates of the unspiked replicates of the sample (Fig. 4). All taxa detected with RPM > 1 were confirmed in the *Mengovirus* spiked dataset at similar RPM levels, with the exception of *Picobirnavirus* (RPM > 2 in both unspiked replicates) and Unclassified *Astroviridae* (RPM 1.28 in one of both unspiked replicates) where RPM dropped below 1 in the spiked replicate. Moreover, two additional viral taxa with RPM > 1 were detected in the spiked sample that remained either undetected or below a RPM threshold of 1 in the unspiked samples. These included evidently *Mengovirus* (29 reads, RPM 5.62), as well as *Pasivirus* A1 (8 reads, RPM 1.55) and *Bocaparvovirus* (7 reads, RPM 1.36). The detection of taxa with RPM < 1 in the spiked sample (representing mostly individual reads, max 2 reads) was within the variation of the two unspiked replicates.

3.5. The addition of Mengovirus does not change the virome reconstruction of bird cloacal swab samples

Using RPM > 1 as a cutoff for mNGS result, the detected viral taxa had even higher RPM counts when *Mengovirus* was added (Fig. 5). Only taxa detected as individual reads (or rarely up to three reads) in unspiked samples were not detected after spiking but these taxa were not within our acceptance criteria of RPM > 1 for a significant mNGS result. One bacteriophage, Sk1 virus (*Siphoviridae*), was only significantly (RPM 2.90) detected in one spiked swab.

3.6. Reproducibility of IQC-RNA and IQC-SEQ in different sample types

A poor correlation between *Mengovirus* Cq values and *Mengovirus* normalized read counts was observed for all sample matrix types tested. *Mengovirus* Cq values were generally reproducible per sample type, while *Mengovirus* normalized read counts reproducibility varied per sample matrix (Fig. 6).

Due to the marginal viral read counts observed in initial experiments, only 2 lung samples were analysed. These had consistent *Mengovirus* Cq values (31–32) but only in one replicate could mengoviral reads be identified with the Kraken based mNGS analysis (two reads, < 1RPM). Bowtie2 mapping confirmed that in both replicates, only low numbers of reads mapping to *Mengovirus* genome DQ294633 could be identified (4 and 20 reads).

RNA extracted from wild bird swabs showed a narrow, reproducible Cq range but a wide variability in normalized *Mengovirus* read counts, reflecting the variable impact of competing nucleic acids from the host and bacteriome. Two swabs had outlier Cq values well exceeding the 99 % CI of the normal Cq range for swab samples. One of these samples for which the library produced only low read numbers ($< 0.5 \times 10^6$ read pairs) had among the lowest normalized *Mengovirus* read counts (38.67). Interestingly, the other swab having an outlier Cq value produced the highest normalized *Mengovirus* read count (2581).

Swine serum samples showed reproducible *Mengovirus* Cq values in

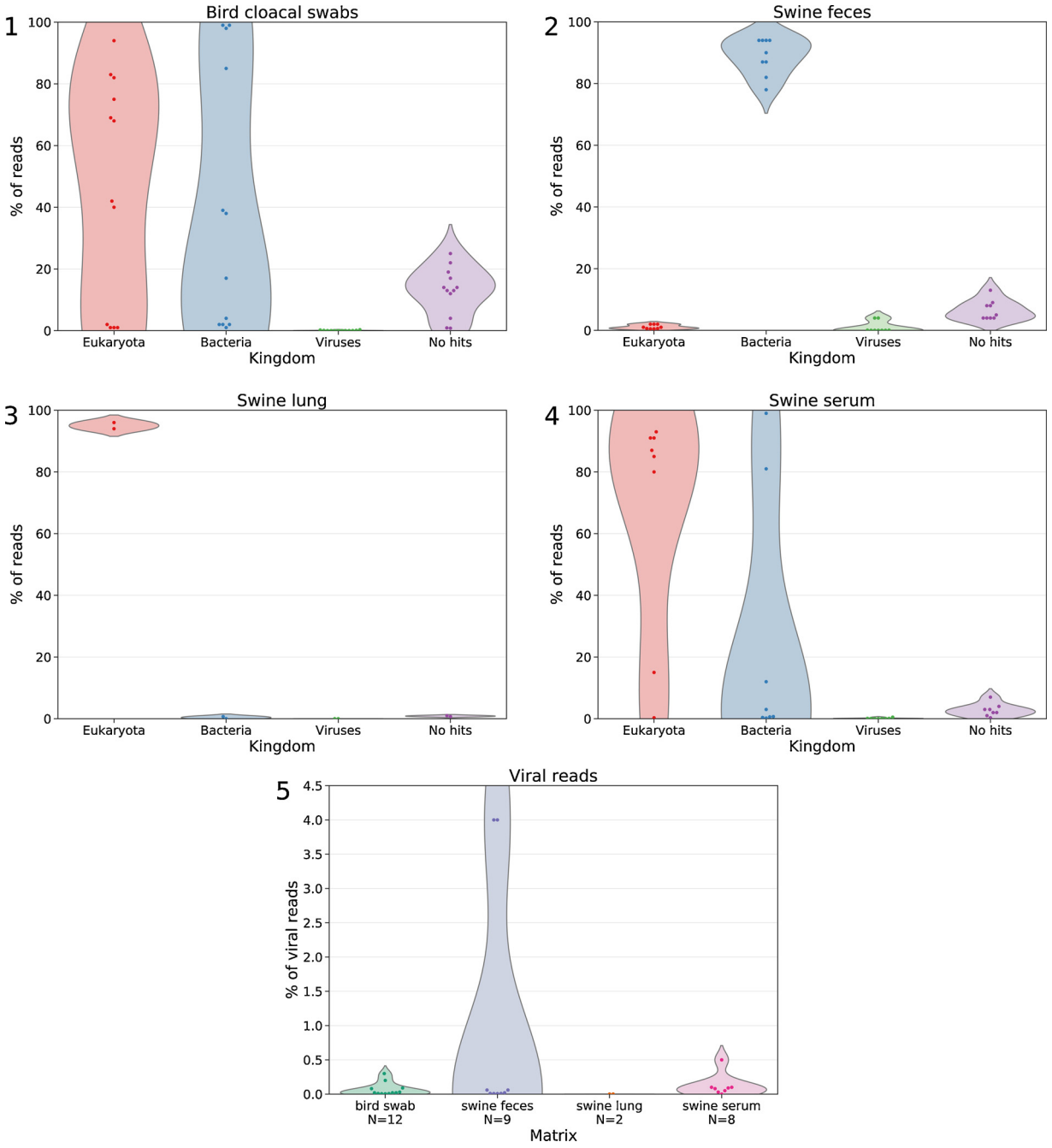

Fig. 2. Distribution of mNGS reads over Superkingdoms (Eukaryota, Bacteria, Viruses) and unclassified reads for different sample matrices (panel 1-4). Category 'No hits': no similarities found in database. Panel 5: Detail comparing the percentage of viral reads detected in each of the sample matrices.

a narrow range (confirming successful viral RNA extraction in all samples) and a lower variability in *Mengovirus* normalized read counts than other sample types. One serum sample showed a *Mengovirus* RPM < 1, while its Cq was well within the 99 % CI of *Mengovirus* Cq values for sera. Further analysis revealed that this serum had a bacterial contamination with 84 % of the reads indicating *Pseudomonas sp*.

Swine fecal samples resulted in highly reproducible *Mengovirus* Cq values that were consistently higher than in other samples, while producing highly variable normalized *Mengovirus* read counts that did not correlate to the Cq values.

4. Discussion

Although mNGS approaches are increasingly being implemented in research and diagnostic labs, QC strategies are not consistently used. Besides the importance of good laboratory practice in metagenomics investigation, awareness is growing that the analysis can be biased due to reagent contamination (Naccache et al., 2013; Rosseel et al., 2014) and database limitations both in terms of content and curation (Lu and Salzberg, 2018). However, the implementation of sample level QC metrics contributes to the avoidance of false negative mNGS results.

In the present study, we evaluated the suitability of a commercial kit as internal quality control (IQC) material for the validation of virus detection by mNGS in individual samples. We purposely selected a completely random RNAseq approach to evaluate its suitability, aiming to produce conservative read count measures, as workflows employing virus enrichment steps are expected to produce higher viral read numbers. As expected with a completely random mNGS approach, only

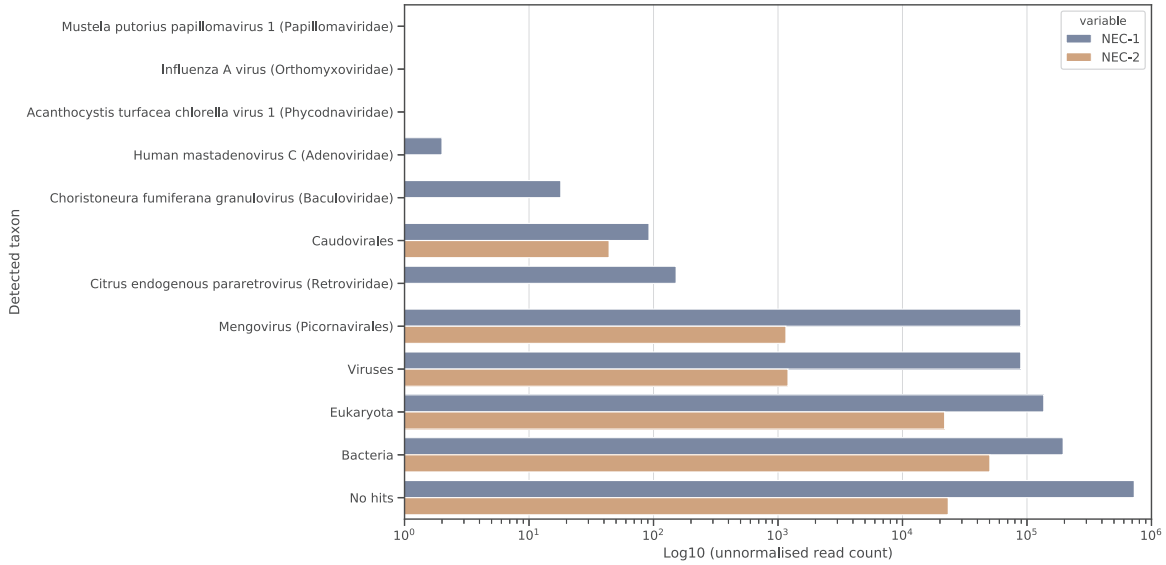

Fig. 3. Viral taxa detected and distribution of reads across Superkingdoms in two *Mengovirus* spiked NEC replicates (log10 unnormalized Kraken counts). The top 3 taxa have only single reads in one of both replicates and are consequently not displayed after log10 transformation. Data representing bacteriophage species hits were regrouped at the Order level (Caudovirales).

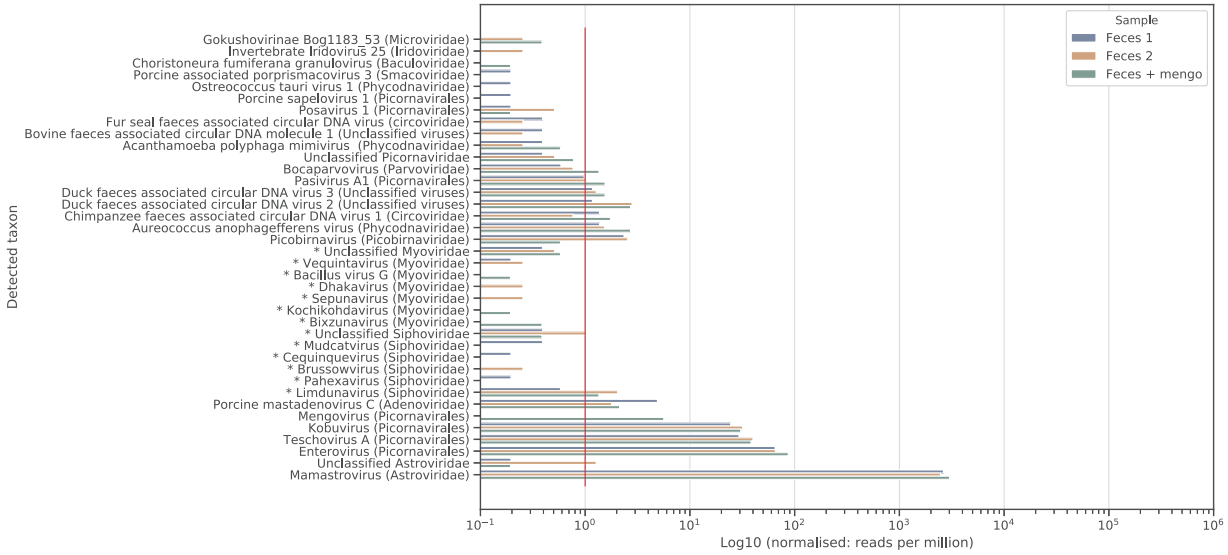

Fig. 4. Viral taxa detected in spiked vs. unspiked replicates of a swine fecal sample (log10 normalized Kraken counts). Bacteriophage species or genera (order Caudovirales) are indicated using an asterisk (*). The employed RPM threshold of one is indicated in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

relatively low viral read percentages were obtained. Reads were mostly Eukaryotic, Bacterial, or unclassified, with a varying frequency of either category depending on the sample matrix.

Although the implemented control is an RNA virus, the random nature of the sequencing workflow should allow to detect RNA from all replicating microorganisms (Wylezich et al., 2018).

Unlike the previously documented use of bacteriophages as QC reagents for mNGS in human respiratory samples (Bal et al., 2018), or veterinary RNA and DNA viruses for human respiratory samples (van Boheemen et al., 2020), we selected *Mengovirus* for four main reasons: (1) *Mengovirus* is widely used as an IQC for foodborne virus targeted molecular diagnostics (Hennechart-Collette et al., 2015; Londoñe-Bailon and Sánchez-Robinet, 2018); (2) the use of an RNA virus with similar structure, genome size, and genome organization to economically important animal viruses ensures control of the entire laboratory workflow including sample homogenization, RNA purification, reverse transcription and second strand synthesis; (3) *Mengovirus* has a fully characterized genome facilitating its detection in large datasets; and (4) its availability as a commercial quantified reagent with accessory qRT-PCR assay facilitates standardization between laboratories. However, it should be noted that the taxon comprising *Mengovirus*, i.e., EMCV, is represented by a single genome (EMCV strain Ruckert, NC_001479.1) in the RefSeq Genome database that we used for metagenomic read classification. Mengovirus and EMCV are antigenically indistinguishable from each other (Martin et al., 2000) and are serologically related to *Cardiovirus* (*Picornaviridae*) (Carocci and Bakkali-Kassimi, 2012). Besides murine EMCV (including *Mengovirus*), this taxon also contains swine EMCV genomes. In the event of higher than expected normalized *Mengovirus* read counts for a given sample matrix (from swine or murine hosts), it should be investigated whether other EMCV were intrinsically present in the sample. In the present study, Kraken *Mengovirus* read counts (thus based on EMCV NC_001479.1) highly correlated to specific read mapping counts using Bowtie2 (thus based on *Mengovirus* DQ294633.1).

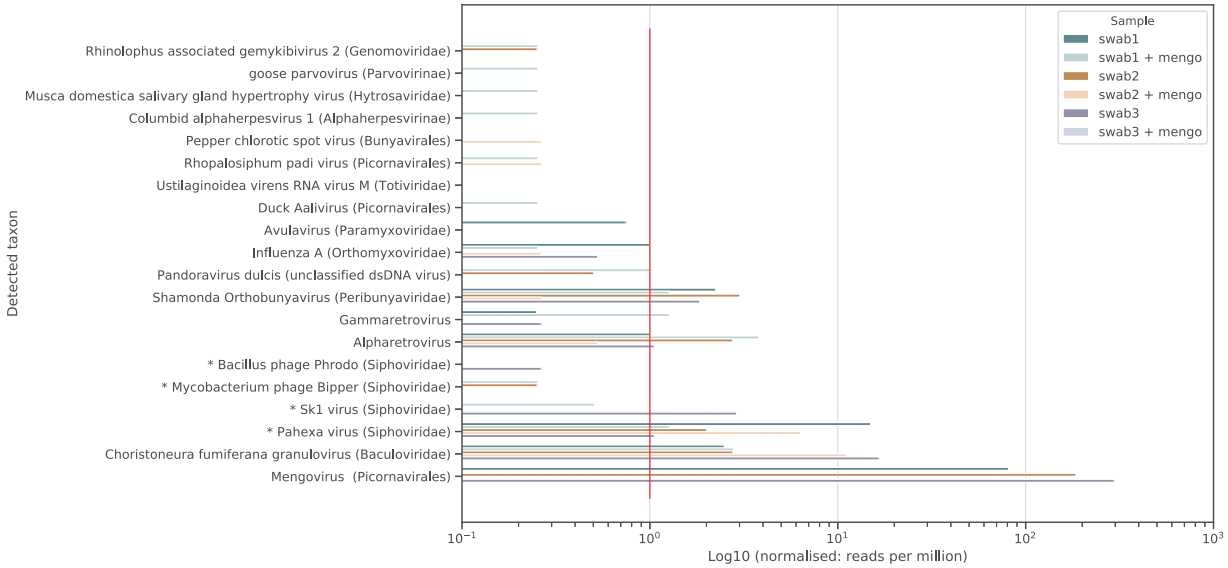

Fig. 5. Viral taxa detected in spiked vs. unspiked wild bird cloacal swabs (log10 normalised Kraken counts). Bacteriophage species or genera (order Caudovirales) are indicated using an asterisk (*). The employed RPM threshold of one is indicated in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

The striking similarity between the bacterial communities in both NEC replicates (water spiked with *Mengovirus* control), processed several months apart, suggests that these contaminants are rather part of the *Mengovirus* control reagent used than laboratory contaminations. They may represent environmental or reagent contaminants of any part of the production chain of this commercial reagent. *Mengovirus* control reagent is originally intended for use in targeted PCR based assays, and to our knowledge, this study presents the first metagenomic characterization of its viral and bacterial community. We have shown that the *Mengovirus* control kit does not contain animal or human viral sequences (other than *Mengovirus*) that would invalidate its suitability as a sample-level control for mNGS on human or animal samples. Given the low amount of data produced from the NEC samples in comparison to other samples (1.15 and 0.19 $\times$ 10^6 reads), stochastic variation may have pushed some of the detected taxa above the RPM threshold of 1. In addition to *Mengovirus*, only two viral taxa were detected with RPM $>$ 1. Among these, bacteriophages (*Caudovirales*) are expected due to the observed bacterial contamination of NEC samples, while we showed that 18 reads misclassified by Kraken as an insect virus (*Choristoneura fumiferana granulovirus*, *Baculoviridae*) most likely represent misclassified fungal 28S ribosomal RNA sequences. This finding, together with the Kraken misclassification of an individual plant chromosomal sequence read as a member of the *Phycodnaviridae*, stresses the importance of verifying results based on low numbers of classified reads. The strict adherence to a normalized read count cutoff, and follow-up investigation of results based on low read numbers (e.g. BLAST alignment to database) are essential in the assessment of the diagnostic relevance of such findings. Such investigations are essential to avoid false positive results e.g. due to poorly annotated sequences in the used databases. Only sporadic and low read counts (one to maximum three raw reads, RPM always $<$ 1) of mammal or human viruses were present in our *Mengovirus* spiked negative controls, and they were not consistently detected in all negative control samples. As we confirmed the specificity of the Kraken classification, these individual reads most likely represent background contamination that may have occurred at any point from reagent production over sample processing to NGS data generation. At any rate, the observed read counts are all below our acceptance criterion for a meaningful mNGS result of

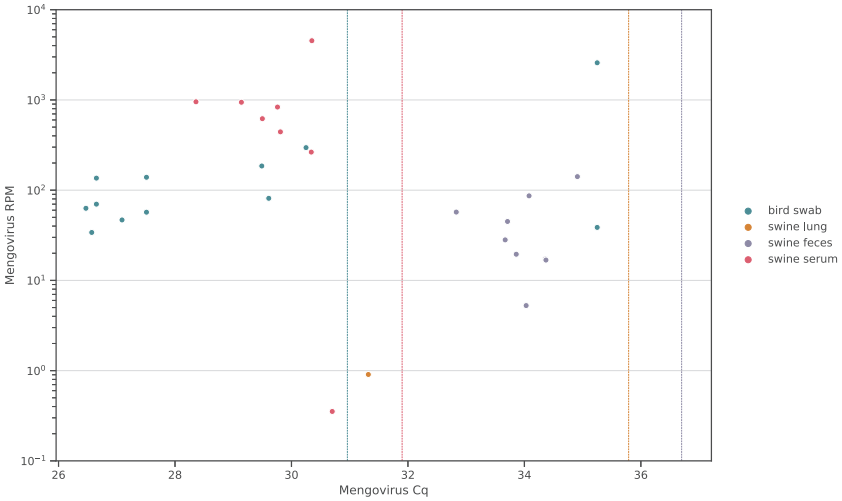

Fig. 6. Scatterplot showing the correspondence between normalized Kraken *Mengovirus* read counts (RPM) and *Mengovirus* qRT-PCR Cq values. Only a single lung sample is plotted as no *Mengovirus* reads were detected in the other lung sample.

RPM > 1.

As several metagenomic libraries were multiplexed on sequencing runs in the present study, we cannot exclude that low read count classification results may represent artefacts due to index swapping (Costello et al., 2018). This phenomenon, based on incorporation of free indexing primers results in the mis-assignment of sequencing reads to a sample. It has been documented to impact up to 0.25 % of sequencing reads in non-patterned flow cells when using PCR amplification steps during the library preparation (Illumina Inc., 2017; Wright and Vetsigian, 2016). Although we did not use unique dual indexes, our library preparation workflow included multiple SPRI (Solid Phase Reversible Immobilization) paramagnetic bead purification steps, reducing the risk of free index primers. We did not observe any bleed through of highly prevalent viral taxa to other libraries. In addition, the majority of reads in the samples in our study being Eukaryotic or Bacterial, it is unlikely that this phenomenon may have resulted in bleed through of viral reads. However, laboratories aiming to implement diagnostic metagenomic applications, should adhere to library preparation best practices and consider using unique dual indices (Costello et al., 2018; Illumina Inc., 2017).

We implemented the control reagent with two readout points. Firstly, IQC-RNA consisted of a semi-quantitative *Mengovirus* qRT-PCR assay to validate efficient viral RNA extraction of individual samples, which can be used as a 'go-no-go' decision point in the mNGS workflow. For each sample matrix, the acceptance criterion for successful RNA extraction was set as the upper limit of the 99 % CI of Cq values for the given sample matrix. The values of this acceptance criterion may be periodically updated when sample matrix *Mengovirus* Cq values are tracked for routinely implemented mNGS testing. Secondly, IQC-SEQ consisted of normalized Kraken *Mengovirus* reads (expressed as RPM) in a given dataset, which can be used to evaluate the consistent generation of viral reads of the entire mNGS data generation and analysis workflow. The acceptance criterion for successful generation of viral mNGS data was set as RPM > 1 for *Mengovirus*, invalidating individual (or a few) absolute reads as we aimed to have metagenomic datasets of minimum 3×10^6 read pairs per sample. Expression of this acceptance criterion as number of reads normalized per million input reads allows independent comparison of samples with different total absolute read numbers. Although RPM numbers can be sensitive to variation as the total number of absolute reads decrease due to stochasticity, this is not expected to be an issue in our data as every sample had high absolute count numbers (average 3.35×10^6 read pairs, excluding NEC samples that resulted in smaller datasets, and a single bird swab sample with an outlier *Mengovirus* Cq value). IQC-SEQ has the power to identify samples where viral nucleic acids are present in the sample (as validated by IQC-RNA), but are not detectable by the mNGS workflow (e.g. due to excessive non-viral nucleic acids competing for sequencing). In addition, we recommend the inclusion of a negative control sample (spiked with *Mengovirus*) to identify excessive contamination issues in the mNGS run.

The differences in *Mengovirus* RPM between the different sample matrices reflects the differences in *Mengovirus* RNA/total RNA ratios, resulting in a trend of decreasing RPM in the order sera > swabs > feces > lung, although sample type RPM overlap due to the variation in each sample type. The negligible percentage of viral reads, and the low *Mengovirus* read counts in lung tissue samples show that although viral RNA is present in the samples (as validated by IQC-RNA), it is outcompeted by host nucleic acids during the downstream steps towards NGS data. This observation is not unexpected given the low *Mengovirus* RNA/total RNA ratio ($0.0003 - 0.0005$) and confirms previous observations (e.g. Rosseel et al., 2015) that unbiased mNGS workflows result in extremely low viral read counts. Consequently, sensitive mNGS studies on animal tissue samples require either extreme sequencing efforts or enrichment procedures focusing sequencing effort on viral targets (e.g. Conceição-Neto et al., 2018; Rosseel et al., 2015). However, virus enrichment strategies may come with inevitable biases such as the cost of losing sensitivity for bacterial and eukaryotic pathogens, while the scope of the methodology in this study was to detect all replicating microorganisms.

Unlike the bacteriophage IQC implemented for mNGS diagnostics in human respiratory samples by Bal and colleagues (Bal et al., 2018), we could not show an overall correlation between *Mengovirus* qRT-PCR results (IQC-RNA) and *Mengovirus* normalized read counts (IQC-SEQ). This may be partly related to the inherent complexity and diversity of sample matrices tested here, with a variable proportion of reads classifying as Eukaryotic (host), Bacterial, or unclassified. The validity and added value of IQC-RNA and IQC-SEQ as separate QC points in a mNGS workflow is clearly illustrated in several sample matrices. Two swabs with mengoviral Cq values beyond the IQC-RNA acceptance criterion indicated problems with viral RNA extraction, which was confirmed in one swab by lower *Mengovirus* normalized read counts compared to other swab samples, while the other IQC-RNA non-compliant swab produced extremely high *Mengovirus* RPM. Our recommendation would be to exclude such samples at the intermediate IQC-RNA control point, avoiding library preparation and sequencing costs and to repeat nucleic acid extraction procedures, if enough input material is still available and adequately preserved.

One serum sample showed consistent viral RNA extraction (IQC-RNA within acceptance criterion), while showing an unacceptable IQC-SEQ (Fig. 6, sole serum sample with RPM < 1), indicating viral nucleic acids were potentially being outcompeted by other sequencing targets. This serum accordingly exhibited a bacterial contamination with 84 % of reads in the dataset being classified as *Pseudomonas sp.*, clearly illustrating the value of IQC-SEQ to detect false negative virus identification results due to template competition.

Importantly, we have shown that the addition of *Mengovirus* at a concentration of 2.17×10^6 copies/mL of sample, or sample homogenate, does not affect the detection of viral taxa in animal samples. However, it should be noted that our direct comparison of *Mengovirus*-spiked vs. unspiked samples is based on a limited number of samples (3 swabs and 2 fecal samples). In bird swabs, the same taxa are detected in spiked and unspiked replicates of a sample, as long as our significance criterion for mNGS findings (RPM > 1) is applied. In swine fecal samples, a single viral species (*Bocaparvovirus*) drops just below the RPM threshold in the spiked compared to unspiked replicate (RPM = 1.36), while two species (*Pasivirus* A1 and *Picobirnavirus*) that are just above the RPM threshold in the spiked sample, remain just below the threshold in unspiked samples. Sporadically, low absolute read counts (one to maximum three reads) are detected either in unspiked or spiked samples alone. This is consistent with low level background contamination that is frequently documented in mNGS workflows (Asplund et al., 2019). Of note, the presence of high normalized read counts for porcine viruses belonging to the genera *Mamastrovirus*, *Enterovirus*, *Teschovirus*, and *Kobuvirus* did not inhibit the detection of *Mengovirus* reads in the expected normalized read count range for that sample matrix (5.26–141.32).

The described QC metrics allow validation of successful viral RNA extraction, avoiding unnecessary expensive library preparation and sequencing (IQC-RNA, using a quantitative Cq base criterion), as well as validating the entire mNGS (data generation + analysis) for individual samples (IQC-SEQ, using normalized read counts). However, as with the implementation of QC's in multiplexed targeted molecular assays (e.g. Vandenbussche et al., 2010), carefully considered interpretation criteria are essential. Especially the interpretation of normalized *Mengovirus* read counts deserves careful consideration given the variable interference of background nucleic acids. For instance, a dataset with poor *Mengovirus* normalized read counts (IQC-SEQ RPM < 1) but high RPM for another viral taxon may be accepted because the abundant generation of non-*Mengovirus* sequence data is expected to hamper the detection of *Mengovirus* RNA (Lewandowski et al., 2019). It should be highlighted that any IQC-SEQ interpretation criterion (we arbitrarily used RPM > 1 based on Kraken and the complete NCBI RefSeq

Microbial Genomes catalogue) should be carefully evaluated as fit-for-purpose by the user based on critical parameters such as the read classification or mapping approach, especially when using different tools where other RPM thresholds may need to be enforced, but also the sample matrix and available validation data.

Once both an experiment (negative controls, sequencing workflow QC metrics) and results from individual samples (IQC-RNA, IQC-SEQ) are validated, a careful interpretation and follow-up of mNGS results should be enforced. The biological significance of detected sequences should be evaluated by looking at the total number of sequence hits, coverage breadth and depth of the detected pathogen's genome, annotation information (presence of functional open reading frames), etc. In all cases, follow-up investigations including targeted molecular assays, confirmation screening, virus isolation and *in vitro/in vivo* characterization should be implemented to confirm or reject the initial mNGS results.

Carefully implemented sample-level internal controls, as the ones implemented in the present study, may contribute to the ongoing transfer of mNGS technologies from research to diagnostic laboratories, hereby providing powerful generic tools complementary to targeted assays for the surveillance and detection of infectious disease agents.

Funding

This work was supported by funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 773830: One Health European Joint Programme as internal joint research project METASTAVA.

CRediT authorship contribution statement

Steven Van Borm: Conceptualization, Methodology, Investigation, Writing - original draft, Supervision, Funding acquisition, Project administration. **Qiang Fu:** Formal analysis, Writing - review & editing. **Raf Winand:** Software, Visualization, Formal analysis. **Kevin Vanneste:** Formal analysis, Software, Conceptualization, Writing - review & editing. **Mikhayil Hakhverdyan:** Resources, Writing - review & editing. **Dirk Höper:** Methodology, Resources, Writing - review & editing. **Frank Vandenbussche:** Conceptualization, Investigation, Writing - review & editing.

Declaration of Competing Interest

None.

Acknowledgements

The authors thank M. Steensels, B. Lambrecht, L. Mostin, D. Maes, and A.B. Cay for providing samples, and L. Weckx for excellent technical assistance.

Sequencing data generation was professionally handled by Transversal activities in Applied Genomics (Sciensano), Laboratory for NGS and microarray diagnostics (Friedrich Loeffler Institut, Germany), Genomics Core Leuven, and VIB Nucleomics Core.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jviromet.2020.113916.

References

Asplund, M., Kjartansdóttir, K.R., Mollerup, S., Vinner, L., Fridholm, H., Herrera, Ja.R., Friis-Nielsen, J., Hansen, T.A., Jensen, R.H., Nielsen, I.B., Richter, S.R., Rey-Iglesia, A., Matey-Hernandez, M.L., Alquezar-Planas, D.E., Olsen, P.V.S., Sicheritz-Pontén, T., Willerslev, E., Lund, O., Brunak, S., Mourier, T., Nielsen, L.P., Izarzugaza, J.M.G., Hansen, A.J., 2019. Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis. 25, 1277–1285. https://doi.org/10.1016/j.cmi.2019.04.028.

Bal, A., Pichon, M., Picard, C., Casalegno, J.S., Valette, M., Schuffenecker, I., Billard, L., Vallet, S., Vilchez, G., Cheynet, V., Oriol, G., Trouillet-Assant, S., Gillet, Y., Lina, B., Brengel-Pesce, K., Morfin, F., Josset, L., 2018. Quality control implementation for universal characterization of DNA and RNA viruses in clinical respiratory samples using single metagenomic next-generation sequencing workflow. BMC Infect. Dis. 18, 537. https://doi.org/10.1186/s12879-018-3446-5.

Blomström, A.-L., Fossum, C., Wallgren, P., Berg, M., 2016. Viral metagenomic analysis displays the Co-infection situation in healthy and PMWS affected pigs. PLoS One 11, e0166863. https://doi.org/10.1371/journal.pone.0166863.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma. Oxf. Engl. 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Bukowska-Ośko, I., Perlejewski, K., Nakamura, S., Motooka, D., Stokowy, T., Kosińska, J., Popiel, M., Płoski, R., Horban, A., Lipowski, D., Caraballo Cortés, K., Pawełczyk, A., Demkow, U., Stępień, A., Radkowski, M., Laskus, T., 2016. Sensitivity of next-generation sequencing metagenomic analysis for detection of RNA and DNA viruses in cerebrospinal fluid: the confounding effect of background contamination. Adv. Exp. Med. Biol. https://doi.org/10.1007/5584_2016_42.

Carocci, M., Bakkali-Kassimi, L., 2012. The encephalomyocarditis virus. Virulence 3, 351–367. https://doi.org/10.4161/viru.20573.

Chiu, C.Y., Miller, S.A., 2019. Clinical metagenomics. Nat. Rev. Genet. 20, 341–355. https://doi.org/10.1038/s41576-019-0113-7.

Conceição-Neto, N., Yinda, K.C., Van Ranst, M., Matthijnssens, J., 2018. NetoVIR: modular approach to customize sample preparation procedures for viral metagenomics. Methods Mol. Biol. Clifton NJ 1838, 85–95. https://doi.org/10.1007/978-1-4939-8682-8_7.

Costello, M., Fleharty, M., Abreu, J., Farjoun, Y., Ferriera, S., Holmes, L., Granger, B., Green, L., Howd, T., Mason, T., Vicente, G., Dasilva, M., Brodeur, W., DeSmet, T., Dodge, S., Lennon, N.J., Gabriel, S., 2018. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genomics 19. https://doi.org/10.1186/s12864-018-4703-0.

Epstein, J.H., Anthony, S.J., 2017. Viral discovery as a tool for pandemic preparedness. Rev. Sci. Tech. Int. Off. Epizoot. 36, 499–512. https://doi.org/10.20506/rst.36.2.2669.

Greninger, A.L., 2018. The challenge of diagnostic metagenomics. Expert Rev. Mol. Diagn. 18, 605–615. https://doi.org/10.1080/14737159.2018.1487292.

Hennechart-Collette, C., Martin-Latil, S., Guillier, L., Perelle, S., 2015. Determination of which virus to use as a process control when testing for the presence of hepatitis A virus and norovirus in food and water. Int. J. Food Microbiol. 202, 57–65. https://doi.org/10.1016/j.ijfoodmicro.2015.02.029.

Hoffmann, B., Scheuch, M., Höper, D., Jungblut, R., Holsteg, M., Schirrmeier, H., Eschbaumer, M., Goller, K.V., Wernike, K., Fischer, M., Breithaupt, A., Mettenleiter, T.C., Beer, M., 2012. Novel orthobunyavirus in Cattle, Europe, 2011. Emerg. Infect. Dis. 18, 469–472. https://doi.org/10.3201/eid1803.111905.

Höper, D., Wylezich, C., Beer, M., 2017. Loeffler 4.0: diagnostic metagenomics. Adv. Virus Res. 99, 17–37. https://doi.org/10.1016/bs.aivir.2017.08.001.

Illumina Inc, 2017. Effects of Index Misassignment on Multiplexing and Downstream Analysis.

Kircher, M., Sawyer, S., Meyer, M., 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 40, e3. https://doi.org/10.1093/nar/gkr771.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923.

Lewandowski, K., Xu, Y., Pullan, S.T., Lumley, S.F., Foster, D., Sanderson, N., Vaughan, A., Morgan, M., Bright, N., Kavanagh, J., Vipond, R., Carroll, M., Marriott, A.C., Gooch, K.E., Andersson, M., Jeffery, K., Peto, T.E.A., Crook, D.W., Walker, A.S., Matthews, P.C., 2019. Metagenomic nanopore sequencing of influenza virus direct from clinical respiratory samples. J. Clin. Microbiol. 58. https://doi.org/10.1128/JCM.00963-19.

Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinforma. Oxf. Engl. 27, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Londoño-Bailon, P., Sánchez-Robinet, C., 2018. Efficiency evaluation of the process control virus "Mengovirus" in real time RT-PCR viral detection in the bivalve mollusc Donax sp. J. Virol. Methods 262, 20–25. https://doi.org/10.1016/j.jviromet.2018.09.006.

Lu, J., Salzberg, S.L., 2018. Removing contaminants from databases of draft genomes. PLoS Comput. Biol. 14, e1006277. https://doi.org/10.1371/journal.pcbi.1006277.

Martin, L.R., Neal, Z.C., McBride, M.S., Palmenberg, A.C., 2000. Mengovirus and encephalomyocarditis virus poly(C) tract lengths can affect virus growth in murine cell culture. J. Virol. 74, 3074–3081. https://doi.org/10.1128/jvi.74.7.3074-3081.2000.

Naccache, S.N., Greninger, A.L., Lee, D., Coffey, L.L., Phan, T., Rein-Weston, A., Aronsohn, A., Hackett, J., Delwart, E.L., Chiu, C.Y., 2013. The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. J. Virol. 87, 11966–11977. https://doi.org/10.1128/JVI.02323-13.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733–745. https://doi.org/10.1093/nar/gkv1189.

Qin, S., Ruan, W., Yue, H., Tang, C., Zhou, K., Zhang, B., 2018. Viral communities associated with porcine respiratory disease complex in intensive commercial farms in Sichuan province. China. Sci. Rep. 8, 13341. https://doi.org/10.1038/s41598-018-31554-8.

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma. Oxf. Engl. 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P., 2011. Integrative genomics viewer. Nat. Biotechnol. 29, 24–26. https://doi.org/10.1038/nbt.1754.

Rosseel, T., Pardon, B., De Clercq, K., Ozhelvaci, O., Van Borm, S., 2014. False-positive results in metagenomic virus discovery: a strong case for follow-up diagnosis. Transbound. Emerg. Dis. 61, 293–299. https://doi.org/10.1111/tbed.12251.

Rosseel, T., Ozhelvaci, O., Freimanis, G., Van Borm, S., 2015. Evaluation of convenient pretreatment protocols for RNA virus metagenomics in serum and tissue samples. J. Virol. Methods 222, 72–80. https://doi.org/10.1016/j.jviromet.2015.05.010.

Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J., Walker, A.W., 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 12, 87. https://doi.org/10.1186/s12915-014-0087-z.

Schmieder, R., Edwards, R., 2011. Quality control and preprocessing of metagenomic datasets. Bioinforma. Oxf. Engl. 27, 863–864. https://doi.org/10.1093/bioinformatics/btr026.

Taylor-Brown, A., Spang, L., Borel, N., Polkinghorne, A., 2017. Culture-independent metagenomics supports discovery of uncultivable bacteria within the genus Chlamydia. Sci. Rep. 7, 10661. https://doi.org/10.1038/s41598-017-10757-5.

ten Hoopen, P., Finn, R.D., Bongo, L.A., Corre, E., Fosso, B., Meyer, F., Mitchell, A., Pelletier, E., Pesole, G., Santamaria, M., Willassen, N.P., Cochrane, G., 2017. The metagenomic data life-cycle: standards and best practices. GigaScience 6. https://doi.org/10.1093/gigascience/gix047.

van Boheemen, S., van Rijn, A.L., Pappas, N., Carbo, E.C., Vorderman, R.H.P., Sidorov, I., van `t Hof, P.J., Mei, H., Claas, E.C.J., Kroes, A.C.M., de Vries, J.J.C., 2020. Retrospective validation of a metagenomic sequencing protocol for combined detection of RNA and DNA viruses using respiratory samples from pediatric patients. J. Mol. Diagn. 22, 196–207. https://doi.org/10.1016/j.jmoldx.2019.10.007.

Vandenbussche, F., Vandemeulebroucke, E., De Clercq, K., 2010. Simultaneous detection of bluetongue virus RNA, internal control GAPDH mRNA, and external control synthetic RNA by multiplex real-time PCR. Methods Mol. Biol. Clifton NJ 630, 97–108. https://doi.org/10.1007/978-1-60761-629-0_7.

Wood, D.E., Salzberg, S.L., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15, R46. https://doi.org/10.1186/gb-2014-15-3-r46.

Wright, E.S., Vetsigian, K.H., 2016. Quality filtering of Illumina index reads mitigates sample cross-talk. BMC Genomics 17, 876. https://doi.org/10.1186/s12864-016-3217-x.

Wylezich, C., Papa, A., Beer, M., Höper, D., 2018. A versatile sample processing workflow for metagenomic pathogen detection. Sci. Rep. 8, 13108. https://doi.org/10.1038/s41598-018-31496-1.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G.F., Tan, W., China Novel Coronavirus Investigating and Research Team, 2020. A novel coronavirus from patients with pneumonia in China, 2019. N. Engl. J. Med. https://doi.org/10.1056/NEJMoa2001017.