A quantitative metagenomic approach to determine population concentrations with examination of quantitative limitations

Kathryn Langenfeld[1], Bridget Hegarty[1], Santiago Vidaurri[1], Emily Crossette[1], Melissa Duhaime[2,+], Krista Wigginton[1,+]

[1]Department of Civil and Environmental Engineering, University of Michigan
1351 Beal Ave, EWRE
Ann Arbor, MI 48109

[2]Department of Ecology and Evolutionary Biology, University of Michigan
1105 N University Ave
Ann Arbor, MI 48109

[+]Co-corresponding authors
Melissa Duhaime: duhaimem@umich.edu
Krista Wigginton: kwigg@umich.edu

ABSTRACT
Quantitative metagenomic methods are maturing but continue to lack clearly defined analytical limits and have not been applied to viral metagenomes. We developed QuantMeta, a quantitative metagenomic tool to determine population concentrations from adding standards before sequencing. We developed an approach to assess quantitative limitations for each population by applying entropy-based detection limits that incorporate both read distribution and coverage and setting read depth variability thresholds to detect and correct quantification errors caused by non-specific mapping and assembly errors. We applied the approach to quantify viruses in wastewater viral metagenomes, including pathogens and bacteriophages. Concentrations of total DNA viruses in wastewater influent and effluent were greater than 10^{10} gc/mL. Human-associated DNA viruses were detected and quantifiable with QuantMeta thresholds, including polyomavirus, papillomavirus, adenovirus, and crAss-like phages at concentrations similar to previous reports. Reads mapped to the genomes of several additional human viruses but these failed to exceed detection thresholds despite a sequencing depth of 200 million reads per sample. Our results highlight the high detection thresholds of quantitative metagenomics (500 gc/µL) compared to PCR-based quantification (~10 gc/µL). The QuantMeta approach, applicable to both viral and cellular metagenomes, advances quantitative metagenomics by improving the accuracy of measured target concentrations.

KEYWORDS
Quantitative metagenomics, synthetic DNA standards, viromes, quantitative limitations

INTRODUCTION
Metagenomics has allowed unprecedented insight into the diversity, structure, and function of microbial communities, from natural and engineered systems to human bodies (1). Metagenomic data is inherently relative, however, and this complicates comparisons of population and gene abundances between samples, especially under changing conditions. A comprehensive quantitative approach is needed to resolve issues with sample comparisons to

expand conclusions from metagenomes by accurately quantifying a multitude of targets simultaneously.

One approach to make metagenomic data quantitative is by converting relative abundances to absolute abundances with factors that include total cell estimates based on flow cytometry or 16S rRNA or housekeeping gene concentrations measured by quantitative PCR (2-5). Both of these approaches depend on the accuracy of time- and resource-consuming ancillary measurements that can be error prone and the latter is not suitable for quantitative viral metagenomes (viromes), as viruses lack conserved genes. Alternatively, DNA standards have been added to samples at known concentrations to facilitate absolute abundance measurements in metagenomes (6-15). With this approach, target gene or genome counts in the metagenome are quantified using the number of reads that map to the standards and the known standard concentrations (6,8,11-13,15).

There are several approaches to spike-in standards used in metagenomics, including the addition of foreign genomes (12,13), mock microbial communities (15), and synthetic DNA sequences (6,8,11). Of these, the synthetic standards have a number of benefits that contribute to their rising popularity. Synthetic DNA standards avoid non-specific mapping, as the nonsense sequences used are unique from known organisms (6,8), they can be designed to reflect the range of characteristics expected in a sampled microbiome, including input DNA lengths, GC contents, and concentrations, and they are readily available in premade mixtures (16). They have been used to assess the temporal and spatial variability of microbial populations in healthcare environments (9), antibiotic resistance genes and pathogens in wastewater (10) and manure (11), and microbial community structure and function in saltmarshes (7). Synthetic DNA standards have also been used as positive controls in metagenomics, such as to generate false positive thresholds (17). As in any quantitative method, the next step is to establish well-defined limits of detection and quantification to meet quality targets for bias, imprecision, and total error (18,19).

Despite the growing application of spike-in standards for quantification, the approach lacks well-defined limits of detection and quantification (20). Our team and others have considered as little as one read sufficient for detection and quantification (6,11,12,21); this approach, however, does not incorporate target length biases in the detection limit (i.e., larger targets have lower detection limits). Ideally, detection limits would be established dynamically relative to a given dataset and target. Beyond detection, accurate target quantification relies on accurate read mapping to sequences of interest. Previous studies implementing quantitative metagenomics determined target concentrations by mapping reads onto targets from databases (i.e., reference-based quantification) (7,9-11). Targets can also be quantified in metagenomes by mapping reads onto contigs assembled from the same sample (i.e., contig-based quantification). Ultimately, the average number of reads mapped along a sequence (read depth) is used to quantify a target; consequently, read mapping errors caused by non-specific mapping (i.e., reads mapping to a target other than its true origin, (22)) or assembly errors (23) may reduce quantification accuracy. An approach to detect and correct read mapping errors is needed to improve the accuracy of quantitative metagenomics.

Here, we implemented synthetic DNA standards to develop a dynamic approach for evaluating whether a target is detectable and quantifiable in a metagenomic dataset. QuantMeta was developed to determine concentrations from quantitative metagenomes and incorporate our detection thresholds and our methods to detect and correct assembly and mapping errors that affect quantification accuracy. We then applied our new approach to quantify ssDNA and dsDNA viruses in municipal wastewater samples. We anticipate that this approach will be useful

in future quantitative metagenomics studies for not only viromes, but also whole metagenomes. Utilizing QuantMeta enables the expansion of genes or populations quantified in future studies as compared with conventional qPCR-based approaches.

MATERIALS AND METHODS

Sample Collection and Processing

Two grab samples of secondary effluent (20 L) or raw influent (10 L) were collected in autoclaved carboys from automatic samplers at the Ann Arbor Wastewater Treatment Plant (Ann Arbor, MI, USA) daily from December 19 to 24, 2020. Samples were transported to the lab on ice within 2 h of collection where they immediately underwent concentration and purification steps based on a previously described ultrafiltration method (24). Briefly, secondary effluent and raw influent were pre-filtered through 100-µm sized pores (Long-Life polyester felt filter bag, McMaster-Carr, Cat. No. 6835K58) then 0.45-µm sized pores (Express PLUS PES filters, MilliporeSigma™, Cat. No. HPWP14250) and concentrated with tangential ultrafiltration (approximately 50-fold and 25-fold, respectively, using 30 kDa MWCO dialysis filters, Asahi Kosei Medical Co., Ltd, Cat. No. 6292966). The concentrated samples were treated with 1 mL chloroform to lyse any remaining cells, filtered through 0.45-µm filters, further concentrated with dead-end ultrafiltration (approximately 20-fold, using 100 kDa MWCO Amicon™ Ultra Centrifugal filter units, MilliporeSigma™, Cat. No. UFC510096). The extra-viral nucleic acids were then degraded with 100 U/mL DNase (Roche, Cat. No. 10104159001) for 1 h on the bench top. The DNase enzymatic reaction was ended by adding 100 mM EDTA and 100 mM EGTA. DNA was then extracted immediately with QIAamp UltraSens Virus Kit (QIAGEN, Cat. No. 53706). The manufacturer's instructions were followed with the first six steps replaced by combining 140 µL of sample with 5.6 µL carrier RNA and vortexing briefly. DNA extracts were stored at -20˚C. On December 26, 2020, a 20 L deionized water sample was processed with the same viral concentration and purification process to account for contamination during sample processing.

Sample Characterization

Each raw influent and secondary effluent sample was analyzed for pH, turbidity, solids content, fold concentration, and viral recovery (Table S1). Sample volumes were determined by weighing samples and assuming a density of 1 g/mL. To more accurately determine virus recovery through concentration and purification, a replicate grab sample was always collected and processed in parallel with the grab samples intended for sequencing. The sample-specific viral recovery was determined by adding *Enterobacteria* phage T3 (GenBank accession no. NC_003298, ATCC® BAA-1025-B1™) at a concentration of 10^5 gene copies/µL. The concentration of phage T3 remaining after the concentration and purification steps was determined with a previously developed ddPCR probe assay (24) (Table S2). The sample-specific viral recoveries ranged 22-45%.

Sequencing Standard Spike Additions and ssDNA Standard Development

A set of 91 dsDNA and ssDNA standards were spiked into each DNA extract prior to library preparations. Sequins metagenome mix A dsDNA standards with lengths varying 981 to 9,120 bp and GC content varying 20 to 71% (6) were spiked into each sample. The dsDNA concentration of the sequin metagenome mix A was measured with the Qubit™ dsDNA HS Assay (ThermoFisher Scientific, Cat. No. Q32851) with 2 µL of DNA template added to the 200 µL assay. The set of standards was expanded with five ssDNA standards. A similar approach to the sequin metagenome mix was employed by inverting ssDNA viral genomes in the RefSeq virus database (downloaded on 3/19/2019) and dividing the genomes into 1-kb long fragments.

The fragments were mapped to the NCBI nr database (downloaded on 8/15/2019) with bowtie2 (v2.3.5). Inverted genome fragments with no alignments were selected as potential ssDNA standard candidates. A random selection of five sequences from the potential candidates were selected with GC contents of 31.7, 40, 45, 50, and 60% and made into Megamer® ssDNA fragments without the complementary strand (IDT, Coralville, IA, Table S3). The ssDNA fragments were resuspended in molecular biology grade ddH_2O (Fisher Scientific, Cat. No. BP28191) to an approximate concentration of 7.5 ng/µL, aliquoted into 10 µL increments, and stored at -20˚C. Standards underwent a maximum of one freeze-thaw cycle. Immediately prior to use, the concentration of the ssDNA aliquots were measured with the Qubit™ ssDNA Assay (ThermoFisher Scientific, Cat. No. Q10212). A mix of the ssDNA standards was then prepared to match the range of concentrations in the sequins metagenome mix A, namely 10^7, 10^4, 10^6, 10^8, 10^5 gc/µL of the 31.7, 40, 45, 50, and 60% GC content ssDNA standards, respectively. The sequins dsDNA metagenome mix A and the ssDNA standard mix were spiked into each sample at 10 gc/ng DNA extract for the standards at the lowest abundance in the mixes based on Qubit measurements. We predicted that the standards spiked at 10 gc/ng DNA extract would be near the detection threshold at our sequencing depth (approximately 200 million reads per sample). The spike-in concentrations were confirmed with ddPCR assays designed for one dsDNA standard and one ssDNA standard (see section below for details, Table S2).

Foreign Marine Phage HM1 Genome Spike Addition

To further examine quantification accuracy of low abundance viruses at our sequencing depth, we spiked in marine phage HM1 (Genbank accession no. KF302034.1) that is foreign to our wastewater samples. The dsDNA HM1 genome (129,401 bps) has an average GC content of 35.7%. HM1 DNA was extracted using the QIAamp UltraSens Virus Kit with the modified protocol described above. Then, 15 µL of DNA extract with 1 µL of loading dye was run on a 0.3% agarose gel with 1 µL of 10,000x SYBR™ Gold nucleic acid gel stain (Invitrogen™, Cat. No. S11494) per 10 mL of gel at 3 V cm^{-1} for 90 minutes. GeneRuler High Range DNA ladder (Thermo Scientific™, Cat. No. FERSM1351) was run according to the manufacturer instructions. HM1 genomes were extracted from the gel with the QIAEX II Gel Extraction Kit (QIAGEN, Cat. No. 20021). The concentration of the purified HM1 genomes was measured with the Qubit dsDNA HS Assay. HM1 was spiked into all samples at approximately 50 gc/ng DNA and the concentration was subsequently checked with ddPCR (see section below for details, Table S2).

Spike-in ddPCR Assays

Spike-in standards and HM1 concentrations were checked with singlet ddPCR reactions performed with the QX200 AutoDG Droplet Digital PCR System (Bio-Rad Laboratories, Inc., Hercules, CA). For each plate, at least two ddH_2O negative controls and two positive controls from spike-in stocks were included. Specific primers and probes were developed for T3, HM1, dsDNA standard S1106_MG_020_A, and the ssDNA standard with 45% GC content (Table S3). 22 µL reactions were prepared with 11 µL of 2x ddPCR™ Supermix for Probes (No dUTP) (Bio-Rad Laboratories, Inc., Cat. No. 1863023), 0.4 µM of all probes and primers, and 3 µL of template. Droplets were generated using the automated droplet generation oil for probes (Bio-Rad Laboratories, Inc., Cat. No. 1864110) to a 20 µL volume, then PCR was performed on the C1000 Touch™ Thermal Cycler (Bio-Rad Laboratories, Inc., Hercules, CA) immediately after droplet generation. The ssDNA assays consisted of 40 cycles of denaturation for 30 seconds at 95˚C, annealing for 1 minute at 56˚C, and extension for 2 minutes at 72˚C, then enzyme deactivation for 5 minutes at 4˚C and 5 minutes at 95˚C, and a final hold at 4˚C. The same PCR

reaction for dsDNA assays was performed with an initial denaturation step at 95˚C for 10 minutes. Plates were run on the droplet reader within 1 hour of PCR completion. For each ddPCR reaction, thresholds were set using a previously defined method that uses kernel density to categorize droplets as positive, negative, or rain (25). Reactions with more than 2 populations, more than 2.5% of droplets classified as rain, or less than 30% compartmentalization were rerun. Inhibition was checked by running 10 and 100-fold dilutions on an influent and effluent sample and was not found to significantly alter the concentration; therefore, 10-fold dilutions were used for all dsDNA and ssDNA standard reactions and 1- or 2-fold dilutions were used for HM1 reactions. ddH_2O negative controls infrequently resulted in positive droplets, and the corresponding concentrations were significantly lower than samples. Specifically, for T3 dsDNA extracts, dsDNA standards, ssDNA standards, and HM1 dsDNA extracts, the ddH_2O negative controls resulted in 20.8 gc/µL template (1/6 reactions), 17.3 gc/µL template (1/4 reactions), 18.5-36.1 gc/µL template (2/4 reactions), and 55.6 gc/µL template (1/4 reactions), respectively.

Illumina NovaSeq Sequencing

Libraries were prepared with the Accel-NGS® 1S Plus DNA Library Kit (Swift Biosciences, Cat. No. 10024) using 50 ng DNA. Samples were sequenced on the Illumina NovaSeq 600 with five samples sequenced per paired-end 500 cycle SP flow cell yielding 251-bp long reads. Library preparations and sequencing were conducted by the Advanced Genomics Core at the University of Michigan. Quality control was performed by trimming Illumina adaptors and an additional 15 bp from the rightmost and leftmost of each read to remove adaptors from the Accel-NGS® 1S Plus DNA Library Kit, and reads were decontaminated of PhiX174 with BBDuk (BBTools, v37.64) (Table S4).

Oxford Nanopore Sequencing

We combined short and long read sequencing technologies to improve viral assemblies because viral genomes commonly have repetitive regions, high mutation rates, and contain host genome fragments (26-28). DNA extracts from each experimental replicate without standards or phage HM1 were cleaned up prior to library preparations with the Zymo Genomic DNA Clean and Concentrate-10 kit (Cat. No. D4011, Zymo Research Corporation). Long read libraries were prepared with the Ligation Sequencing Kit (Cat. No. SQK-LSK109, Oxford Nanopore Technologies) and barcoded with the Native Barcoding Expansion 1-12 (Cat. No. EXP-NBD104, Oxford Nanopore Technologies). The six samples were sequenced on two flow cells (R9.4.1, Cat. No. FLO-MIN106, Oxford Nanopore Technologies) (Table S4). Basecalling was performed using Guppy (v4.2.3) and called reads were classified as either pass or fail depending on their mean quality score (≥ 7). Library preparations, sequencing, and basecalling were conducted by the Advanced Genomics Core at the University of Michigan.

Assemblies: Long Read Only and Hybrid Co-assemblies

Two assemblies were run with the reads from each sample. Hybrid co-assemblies were performed with long reads and short reads from each sample separately with metaSPAdes (v3.15.2) using kmer sizes of 21, 33, 55, 77, 89, and 127. Long read only assemblies were performed with Flye (v2.8.3) for each sample separately followed by several polishing steps including four rounds of Racon (v1.4.10), one round of medaka (v1.3.2), and short-read error correction with pilon (v1.24) (26). Following both assemblies, the contigs for each sample were pooled and contigs less than 1,000-bp were removed. The remaining contigs were assessed for likelihood of viral or proviral origin. Five viral detection methods were run on the contigs. VirSorter (v1.0.5) (29) with the virome flag, VirSorter2 (v2.2.2) (30), VIBRANT (v1.2.1) (31) with the virome flag, VirFinder (v1.1) (32), and CheckV (v0.7.0) (33) end-to-end were run to

assess likelihood of each contig being viral or proviral. Potential viral or proviral contigs were sorted into high and low confidence categories. High confidence contigs were any contigs where at least one viral detection method indicated a contig was highly likely to be viral or proviral. The high confidence cut-offs were a VirSorter score less than 2 or 4, VirSorter2 score greater than 0.9, VirFinder score greater than 0.9, VIBRANT score less than 3 or included in the "prophage" list, and included in the CheckV "prophage" list. Low confidence contigs were any contigs where at least one viral detection method indicated some confidence in the contigs to be viral or proviral. The low confidence cut-offs were a VirSorter score of 3 or greater than 4, VirSorter score between 0.5 and 0.9, VirFinder score between 0.7 and 0.9, and a VIBRANT score of 4. The low confidence viral contigs were assessed with CheckV run end-to-end and any contig containing one or more viral genes or no host genes were assigned as viral and included with the high confidence viral contigs. Of the contigs greater than 1000 bp, 75.5-78.0% of contigs were classified as viral. The viral contigs were dereplicated by clustering contigs with clustergenomes.pl (v5.1) (34) with 95% ANI similarity and sharing at least 70% coverage, then retaining the longest contig in each cluster as a representative of the clustered viral population. After clustering, 16.5-23.3% of viral contigs were removed. The total viral population count ranges 91,596 - 273,910 in our viromes.

Downsampling and Creation of Negative Control Dataset

Downsampling was performed by randomly sampling 1% and 20% of reads with seqtk (v1.3) from each sample with spiked-in standards to expand the concentration range of the standards dataset. The set of negative controls was created by modeling single nucleotide polymorphisms (SNPs) for each standard with a Mutation Simulator (35) (SI Section 4) and any mutated sequence with a coverage less than 10% or expected read distribution less than 0.3 was retained in the model dataset to capture the variability in mapping to targets that are not confidently detected.

Mapping

Read mapping with short reads was performed using Bowtie2 (v2.4.2) using the default mapping parameters with deinterleaved fastq formatted files of QC reads. Bowtie2 indexes were built with default parameters for the NCBI viral database as whole genomes and broken into separate genes (downloaded 9/6/2021), VirSorter curated database (downloaded 5/5/2021), dsDNA and ssDNA standard sequences, mutated standard sequences, HM1 sequence, and RefSeq crAss-like phage database (downloaded 10/4/2021). Primer blast to the NCBI nr database with the crAss-like phage CPQ056 primer set (36) was performed to determine the list of NCBI accessions captured by the primer set (performed on 10/26/21).

For read mapping to DNA virus pathogens, complete genomes of herpesviruses (n = 1099) and poxviruses (n = 395) from ViPR (cite, downloaded on 7/22/2022) and adenoviruses or mastadenoviruses (n = 15), papillomaviruses (n = 64), polyomaviruses (n = 6), parvoviruses (n = 2), Torque Teno Viruses (n = 18), and bocaviruses (n = 3) from RefSeq (downloaded on 7/22/2022) were compiled, then genomes were clustered if there was at least 90% ANI similarity and shared at least 70% coverage with CheckV (33). The clustering produced 118 virus pathogen clusters (Table S6) and one virus per cluster was selected as a representative to build Bowtie2 indexes with default parameters.

Viral populations from each sample were indexed using the "large-index" parameter. The Bowtie2 sam file output was converted to a tabular format with the number of reads mapping to each basepair of a target using depth (samtools, v1.11). Minimap2 (v2.17) was used to map contigs onto standard sequences and genes or genomes from databases using default parameters.

Assembly quality of contigs derived from standards were assessed by performing Blastn (v2.9.0) with a custom database made from the standard sequences using default settings.

Statistical Analysis

All statistical analysis was performed in R (v4.0.3). Linear and logistic regression analyses and student's t-tests were performed with the R stats package (v4.0.3). Paired and unpaired t-tests were performed with 0.95 confidence levels with *p*-values less than 0.05 considered significant. ROC curves and optimal cutpoints for logistic regressions were performed with the cutpointr package (v1.1.1). Graphs were created with ggplot2 (v3.3.5).

RESULTS

One-metric Detection Threshold Improves Relative-to-Absolute Relationship

In order to predict the absolute abundances of unknown targets in our samples, we developed a relationship between the standard relative abundances and known absolute abundances in a set of wastewater viral metagenomes (viromes) used to develop the QuantMeta approach (see SI Section 2 and equation S1). Three wastewater influent and three secondary effluent samples were spiked with 86 dsDNA standards (6) and newly designed 5 ssDNA standards at concentrations varying from 10^3 to 10^8 gc/µL of DNA extract. To expand the dynamic range of the standards beyond what was achieved with our spike-in strategy, reads from each sample were downsampled 20% and 1% to generate datasets where spike-in concentrations were as low as 10 gc/µL of DNA extract (Table S7).

The relationship between expected and predicted concentrations of the dsDNA and ssDNA standards was evaluated with a linear regression; the residuals of the ssDNA and dsDNA standard sets were significantly different (see SI Section 3 and Figure S1). Reasons for the observed differences are unknown but may have been due to different behaviors of the DNA molecule types during library preparation or sequencing biases that preferentially selected for ssDNA. Although significant, the differences were small, with the standard concentrations predicted by the separate dsDNA and ssDNA standard curves differing by between 6.5% and 12.8% from 10^7 to 10^3 gc/µL standard concentrations, respectively. We, therefore, combined the ssDNA and dsDNA regressions in a single standard curve for the remainder of the study.

Similar to earlier reports using the synthetic DNA standards approach (6,8,11), we observed a strong linear relationship between predicted and expected concentrations of the standards ($R^2 = 0.94$, Figure 1A) with a slope of 0.895 using a single read as the detection threshold. The poorer correlations between expected and predicted concentrations at the low end of the standard concentration range (Figure 1A) suggests a single read detection threshold may not always be appropriate. Previous studies have used minimum read counts (6,11,12) or a combination of read count, coverage, and distribution (37,38) to determine detection thresholds. The "one read" threshold does not account for the fact that each measured read comprises only a small fraction of a target gene or genome. Considering only read counts to establish a detection limit results in target length biases, given that a single read covers a larger proportion of a shorter target than a longer target. Such a case would therefore require higher concentrations to detect short targets. We do not include a minimum read count in the detection threshold to avoid these associated length biases (see SI Section 4 and Figure S2).

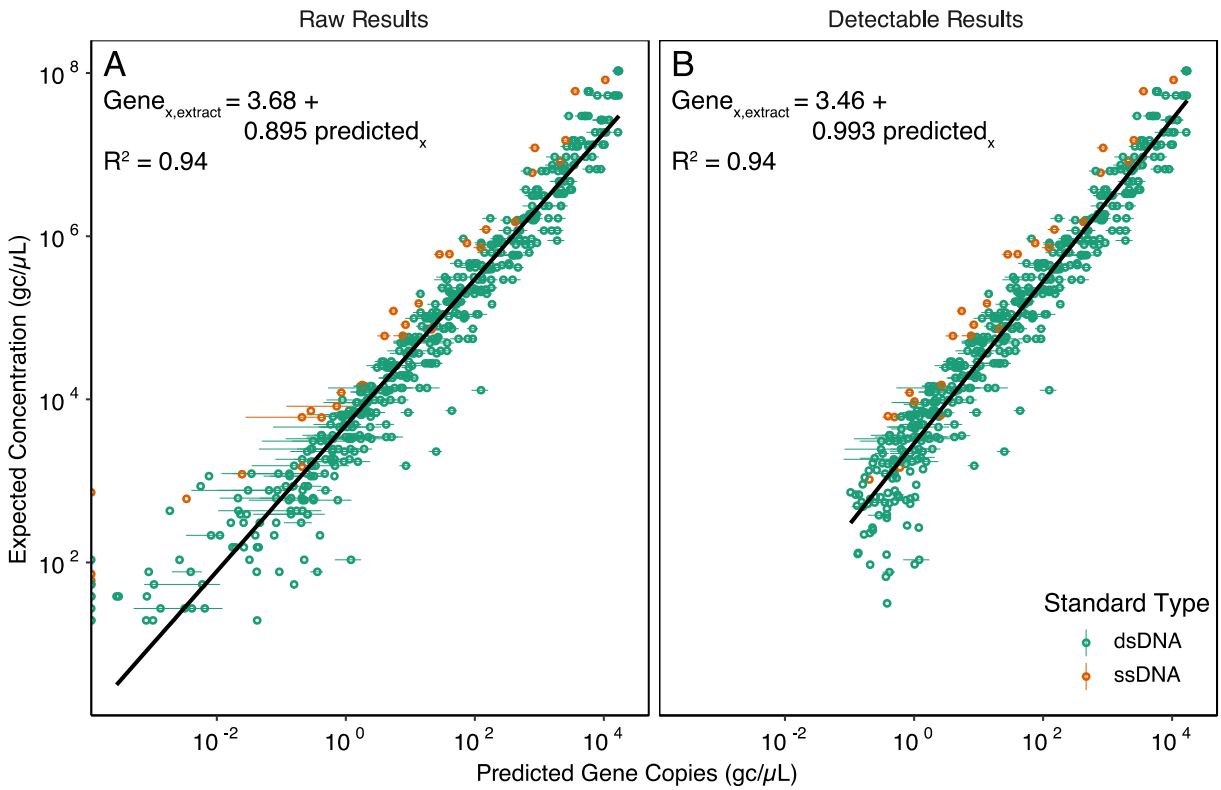

Figure 1 Application of the entropy-based detection threshold. The relationships between expected and predicted concentrations of the standards (A) before and (B) after the detection threshold was applied and standards that failed to meet the detection threshold were removed. The predicted gene copies per µL of DNA extract were determined with equation S1 for all the standards across all samples, including the 20% and 1% downsampled results (results of replicate downsampled sets are reported as single values with standard deviations). The black line represents the linear regressions; green and orange points represent values of the dsDNA and ssDNA standards, respectively.

Instead, we developed a mapping entropy (i.e., measure of randomness) approach that incorporates both read distribution and coverage into a single detection threshold parameter (E_{detect}), similar to how entropy is used to summarize community diversity and richness in the Shannon's Diversity Index (39,40). We used standards and negative controls to determine the entropy that meets coverage and distribution requirements for detection (E_{detect}). Negative control sequences were generated *in silico* by creating single nucleotide polymorphisms in the standard sequences. We created a binary logistic regression model to generate probabilities of relative entropy (E_{rel}) meeting read coverage and distribution requirements (Equation S7, details provided in SI Section 4). Specifically, a standard was assigned as confidently detected in the binary logistic regression if it had greater than 10% coverage and a ratio of expected Poisson read distribution to observed read distribution greater than 0.3; as empirically determined previously (37). The binary logistic regression model was then tested with sample virome reads mapped to gene or whole genome targets from databases. Optimal entropy cutpoints were set to maximize the sum of sensitivity and specificity (details provided in SI Section 4) and varied with target length in our test dataset (Figure S3). Together, this resulted in a length-dependent entropy threshold, E_{detect} (Equation 1).

$$E_{detect} = 0.894 + 0.732\, log^3(L) \qquad (1)$$

With this, a target with a measured entropy (E_{rel}) that is less than E_{detect} is not considered detected, as it does not achieve the minimum coverage and read distribution requirements.

The standards provided an opportunity to identify detection limits in our wastewater viromes. Of 2,730 standards across all samples with 20%- and 1%-downsampling, 2,347 were detected and exceeded the length-specific detection threshold. With this approach, the detection limit was approximately 500 gc/µL of DNA extract (Figure 1B). Removing the standards below the threshold substantially improved the relationship between the expected and predicted standard concentrations. Without the threshold, the regression slope was 0.895 (R^2 = 0.94) and with the threshold applied, the regression slope was 0.993 (R^2 = 0.94). The entropy detection thresholds reported in Equation 1 are applicable to other metagenomes, with or without standards added.

Incorporating Mapping Correction Improves the Accuracy of Quantification

Having established a strategy to set detection thresholds, we next used the standards to develop a method for detecting and correcting quantification errors due to non-specific mapping and assembly errors. When we quantified using reads mapped to *contigs* assembled from the sequenced standards, we observed poorer correlations between the expected and predicted concentrations than when reads were mapped to the known standard reference sequences (R^2 = 0.67, Figure 2A). We attributed this trend to assembly errors in the sequenced standards that interfered with standard read mapping (SI Section 5, Table S10).

In the absence of nonspecific mapping and assembly errors, the variability in the number of reads mapping across a target sequence is related to the local GC content and the read depth of the target sequence (41,42). Because the synthetic standards do not share DNA sequence homology with any known organism (6) we assumed the absence of non-specific mapping to the standards. We could therefore use the relationship between local GC content and local read depth of the standard sequences (Table S8) to establish maximum read depth variability thresholds for targets, where read depth variability is represented by root mean square error (RMSE) (Table S9, SI Section 5, Figure 2B). Targets with read depth variability RMSE above the maximum thresholds were identified as having either nonspecific mapping or assembly errors. They then passed through an iterative correction process (Figure S5): (1) identify outlier regions along the target (Figure 2C), (2) correct outlier regions (Figure 2C), (3) recalculate the read depth variability RMSE of the target (Figure 2C), and (4) iterate until the RMSE of the whole target is less than the threshold or until 20 iterations were completed. At the end of the iteration process, if targets had more than 20% of 49-bp sliding windows altered or an RMSE still greater than the threshold, they were identified as non-quantifiable. Our reasonings for these cut-offs are described in the Supplementary Information (SI Section 5, Figure S6).

To assess our ability to identify assembly errors with this approach, we evaluated the RMSE of reads mapping to the subset of contigs originating from the standards (Figure S7A). Of 910 standards in all samples, 837 were assembled into *de novo* contigs. Of these, 103 had read depth variability RMSE that was too high based on our thresholds and 15 of these were correctable. Correcting the standard contigs and removing the non-quantifiable standard contigs significantly reduced the differences between predicted and expected concentrations (p-value = 0.0063). By evaluating the quality of the standard assemblies, we confirmed that the targets that exceeded our established read depth variability threshold indeed had assembly errors and that correction improved the quantification of those targets (SI Section 5).

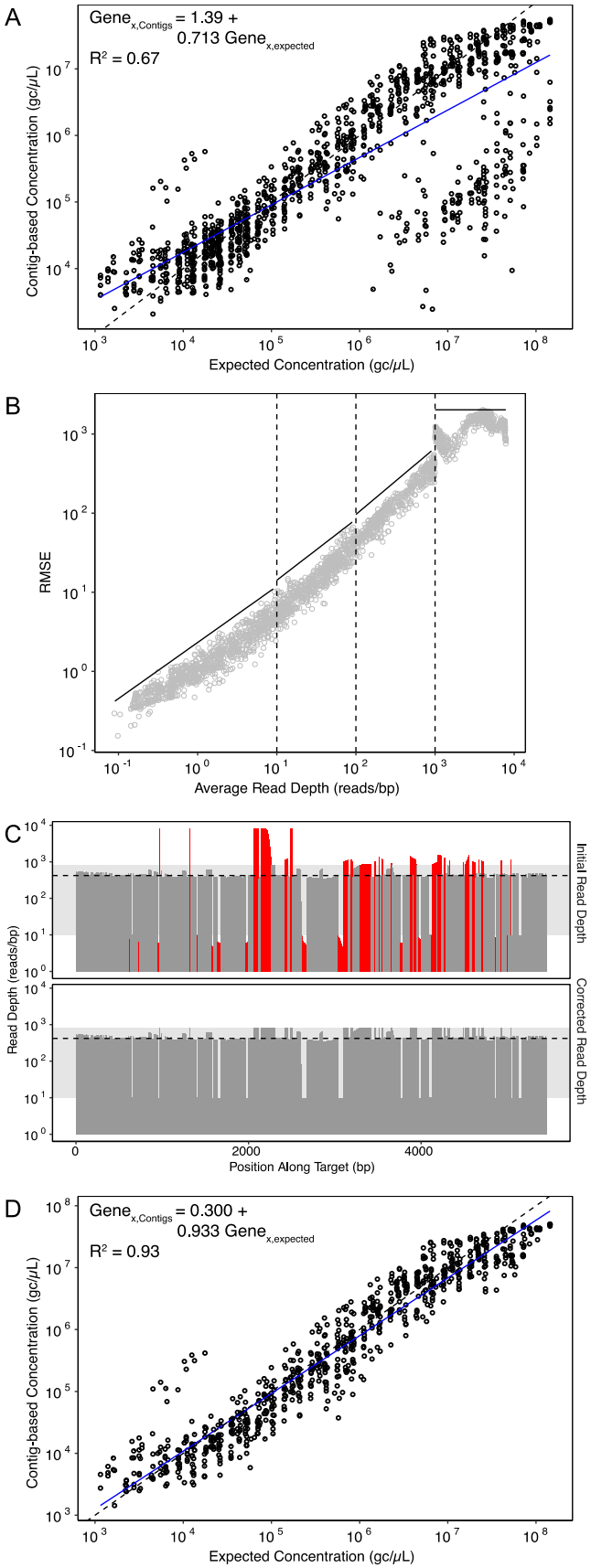

Figure 2 Quantitative accuracy improves by applying quantification correction. (A) Comparison of expected and contig-based concentrations of standards-derived contigs; each point represents a single standard-derived contig in a sample. The black, dashed line indicates the ideal 1:1 relationship between expected and contig-based quantification; the blue line indicates the observed linear regression. (B) RMSE thresholds (black lines) used to identify read mapping issues (e.g., non-specific mapping or assembly errors) based on RMSE of the reads mapped to standard references divided into bins at $\log_{10}$ intervals (dashed vertical lines, Table S3). (C) An example of a contig with nonspecific mapping or assembly errors before (top) and after (bottom) correction. Regions in red indicate areas outside 1.5 standard deviations from the mean read depth (gray shaded area). The corrected read depth is shown with the black, dashed line. (D) Comparison of expected and contig-based concentrations of standards-derived contigs after quantification correction.

Quantification Errors Were Rare but Significantly Altered Target Concentrations

We applied the full detection threshold and quantification correction approach to viral sequences of interest to assess the impacts on calculated concentrations. Quantification correction was performed on reads that were mapped to sequences from the NCBI RefSeq viral genome database (whole genomes and individual genes) and the VirSorter curated database (29), as well as to representative sequences of viral populations assembled *de novo* from the viromes. Quantification correction was required for 1.4% of targets regardless of if reads were mapped to database sequences or populations. Of the 95,351 genes and genomes from reference

databases across all samples that exceeded the detection threshold, high read depth variability RMSE was observed for 1.4% of targets (n = 1,290) with 0.74% of targets (n = 702) identified as not quantifiable (Figure 3A). The false positive rate (i.e., percent of reads mapped to non-quantifiable targets in each database) were 3.8%, 15%, and 56% for the RefSeq viral genes, RefSeq viral genomes, and curated VirSorter genomes, respectively. These observations were consistent with previous read mapping to gene-centric databases that found false positive rates of 12-24% (43). When quantification correction was performed on *de novo* assembled viral populations, the mean percentage of populations with high read depth variability RMSE was 1.5% across our samples (Figure 3B). On average across the samples, 0.20% of populations underwent quantification correction with the remaining 1.3% of populations non-quantifiable.

Quantification correction impacted the concentrations of targets from database reference targets (genes and genomes) and *de novo* assembled viral contigs differently. For database targets, quantification correction significantly reduced the observed target concentration (p-value $< 2.2\text{x}10^{-16}$) by 67% (95% CI: 59 - 75%). The relative concentrations of *de novo* assembled viral populations were also significantly altered through the quantification correction process (p-value $< 2.2\text{x}10^{-16}$), increasing or decreasing by 29% (95% CI: 28 - 30%) on average across the samples. For example, *Aeromonas salmonicida* phage 56 (NCBI accession NC_019527) (44) concentrations before correction were an average of 1.7-fold higher than the corrected concentration in the four of six samples identified as having read mapping errors along the phage sequence. The concentration of *Salmonella enterica* phage LSPA1 (NCBI accession NC_026017, (45)) changed by an average of 1.5-fold following the correction process for the two samples flagged as having read mapping errors to the phage sequence. The targets that required quantification correction spanned all concentrations in our viromes.

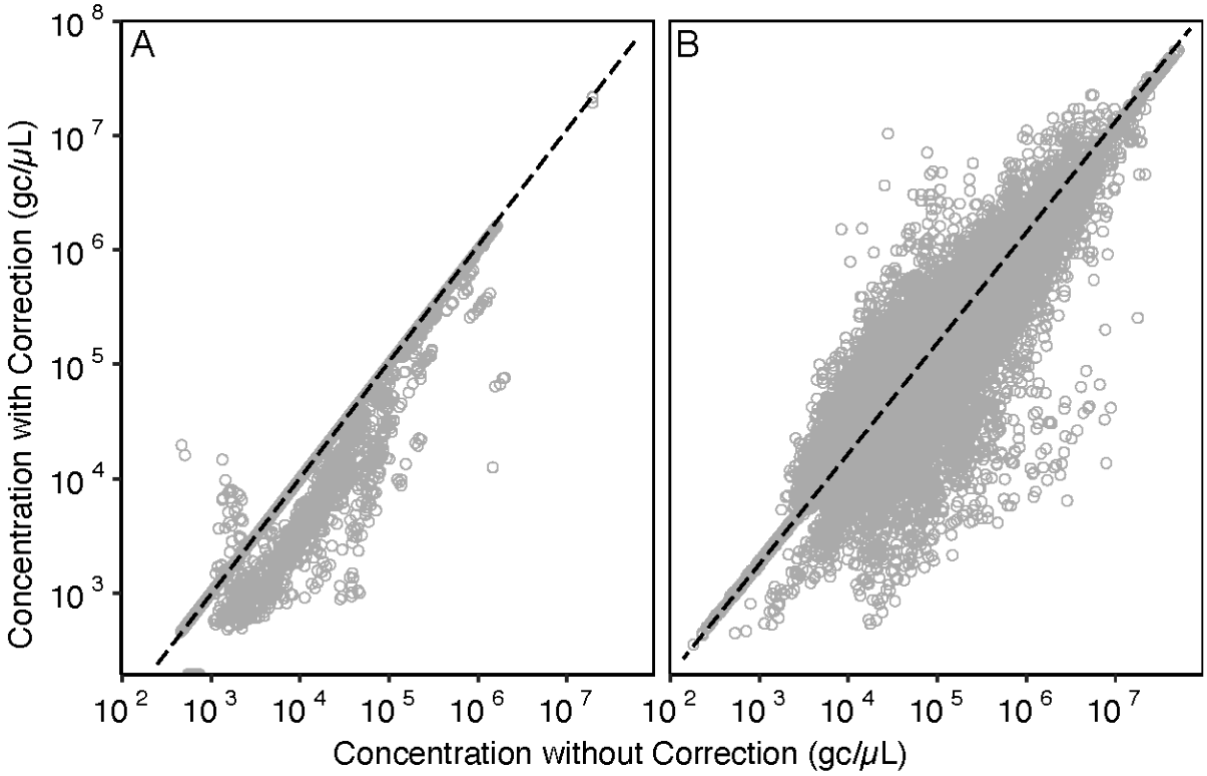

Figure 3 Impact of quantification correction on target concentrations. Relationship between target concentrations without (x-axis) and with (y-axis) quantification correction for reads mapped to viral sequences from reference databases (A) and *de novo* assembled viral populations (B). The databases included genomes and genes from the RefSeq viral and VirSorter curated databases. Targets that did not require quantification correction fall on the black, dashed line.

Viral Abundances Through Wastewater Treatment

To test our quantitative approach against a standard molecular quantification method, we quantified a spiked-in marine phage (*Pseudoalteromonas* phage PSA-HM1) in the three wastewater influent and the three wastewater effluent viromes with both quantitative metagenomics and ddPCR. We tested quantification using reads mapped to the HM1 reference genome (reference-based quantification), as well as to *de novo* assembled HM1 contigs (contig-based quantification). The mean HM1 concentration in the DNA extracts based on ddPCR measurements was $1.1x10^3$ gc/µL (range: $7.2x10^2$-$1.6x10^3$ gc/µL, Table S11). HM1 was above the detection thresholds for reference-based detection in all samples and quantifiable in four of six samples. Reference-based concentrations using the QuantMeta approach were found to be 1% lower and 8.4% higher than those measured with ddPCR in influent and effluent samples, respectively. One *de novo* assembled HM1 contig was identified in one influent sample with a 1,100 bp length. Contig-based concentrations of HM1 in this sample were higher than the ddPCR and reference-based measurements.

We next applied the quantitative metagenomics approach to quantify the absolute abundances of DNA virus populations found naturally in the wastewater samples. The absolute abundances of all DNA viruses in wastewater were quantified by summing the concentrations of the identified *de novo* assembled viral populations. Total virus concentrations were slightly higher in wastewater effluent than influent samples (p-value = 0.37), with mean concentrations of $7.4x10^{10}$ gc/mL (s.d. = $2.4x10^{10}$ gc/mL) in effluent and $5.6x10^{10}$ gc/mL (s.d. = $1.8x10^{10}$ gc/mL) in influent (Figure S8).

We next quantified select human-specific dsDNA viruses that are common in wastewater, namely crAss-like phage, which infects *Bacteroides* in the human gut (46,47). In every case, crAss-like phage concentrations were higher in the influent samples than the effluent samples (Table 1). Mean reference-based crAss-like phage measurements in influent and effluent were $2.7x10^7$ gc/mL (3/3 samples) and $1.1x10^6$ gc/mL (3/3 samples), respectively, with 2.0% of detected crAss-like phage sequences requiring quantification correction across all samples. The contig-based approach resulted in higher crAss-like phage concentrations, namely 120% higher than reference-based quantification (p-value = 0.045). CrAss-like phage is commonly quantified in wastewater with qPCR using the CPQ056 primer set (36). However, primers may underestimate concentrations of crAss-like phage by limiting quantification to specific crAss-like phage sequences amplified with the CPQ056 primers. When *in silico* crAss-like phage quantification was limited to those targeted by CPQ056 primer sequences, the crAss-like phage concentrations were reduced by 80-83% with reference-based quantification and 15-83% with contig-based quantification.

Table 1 QuantMeta-derived concentrations of crAss-like phages (total crAss-like phages and CPQ056 primer-specific crAss-like phages) were measured in each sample by mapping contigs and reads to genomes from RefSeq. Concentrations are reported as gc/mL of wastewater.

Sample	**crAss-like phage/CPQ056 Concentration (log_{10} gc/mL wastewater)**	
	Contig-based	**Reference-based**
12/19/20 Influent	$9.77x10^7/1.86x10^7$	$3.24x10^7/5.75x10^6$
12/21/20 Influent	$5.62x10^7/1.38x10^7$	$2.29x10^7/4.17x10^6$
12/23/20 Influent	$8.13x10^7/1.35x10^7$	$2.57x10^7/4.47x10^6$
12/20/20 Effluent	$7.94x10^6/6.76x10^6$	$1.15x10^6/2.09x10^5$
12/22/20 Effluent	$1.05x10^7/8.13x10^6$	$1.32x10^6/2.63x10^5$
12/24/20 Effluent	$4.17x10^6/3.24x10^6$	$6.46x10^5/1.29x10^5$

We measured the concentrations of DNA virus pathogens in each sample (Figure 4). Reads mapped onto reference genomes from all types of virus pathogens, however, almost all of these instances did not meet our detection thresholds. Only polyomaviruses, papillomaviruses, bocavirus, and adenovirus were quantified in our wastewater viromes when accounting for quantitative limitations and none required quantification correction (Table S12). JC and BK polyomaviruses were quantifiable in all influent samples with mean JC and BK polyomavirus concentrations of $5.6x10^4$ and $6.3x10^4$ gc/mL, respectively. Most of the viruses failing detection were not near our threshold. Only 16 of 683 viruses below detection were within 10% of their respective detection threshold and these included adenovirus, gammaherpesvirus, and papillomaviruses.

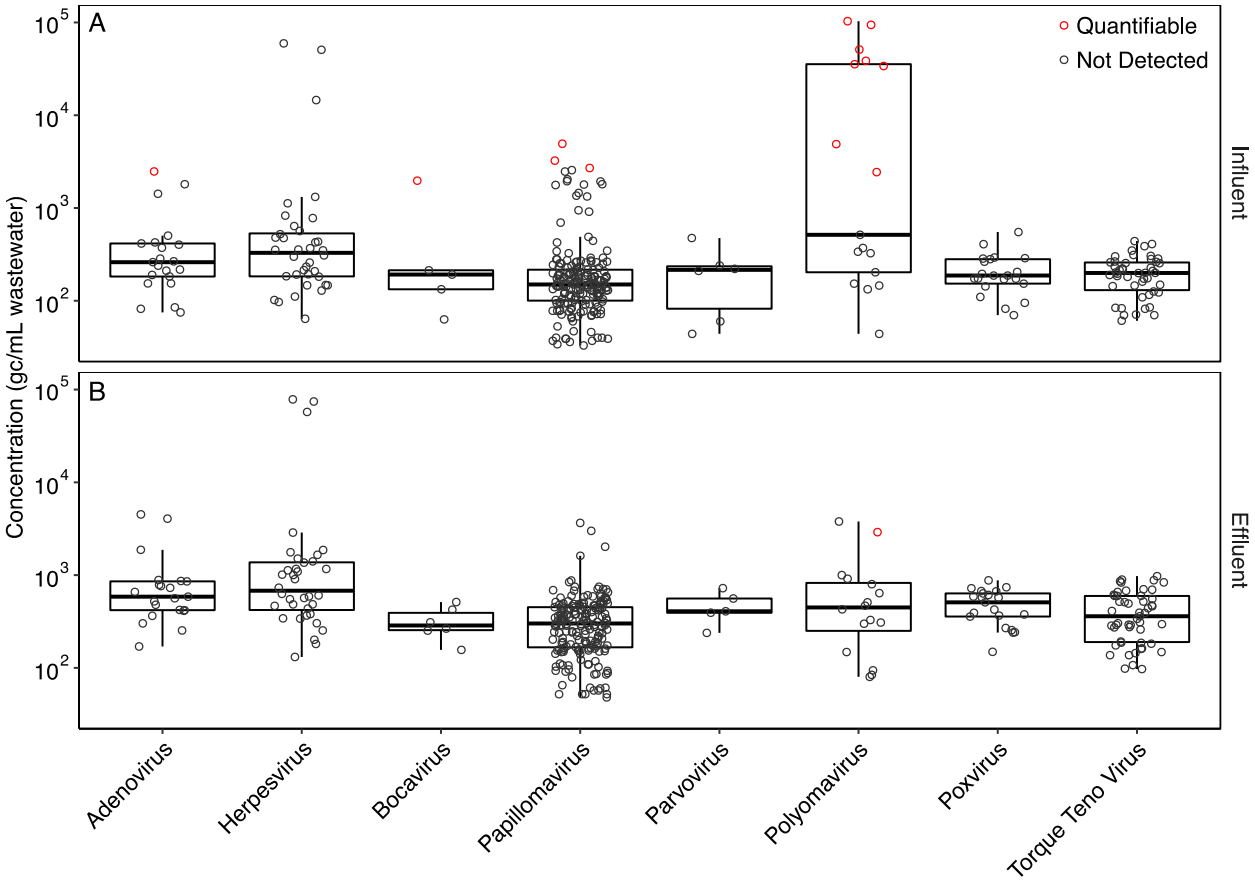

Figure 4 Reference-based concentrations of DNA virus pathogens were measured in influent (A) and effluent (B) samples by mapping reads onto representative genomes from clustering virus pathogen genomes in RefSeq and ViPR (see methods section for specifics). Each point represents the concentration of a virus cluster in a single sample with the status of passing or failing detection thresholds indicated in red and gray, respectively. All concentrations are provided in Table S12.

DISCUSSION

Spike-in standards facilitate the quantification of microbial targets in metagenomes at accuracies similar to quantitative PCR methods (6-8,11,12). The major advantage of quantitative metagenomics is the ability to simultaneously quantify thousands of targets. Methods proposed thus far, however, lack detection thresholds and quantification correction approaches, and this limits the trustworthiness of the resulting data. Through the application of spike-in standards, we developed an approach to identify quantitative limitations in metagenomes and correct quantification errors. Incorporation of new quality thresholds and quantification correction improved the relationships between the expected and measured concentrations of spike-in standards (Figures 1 and 2).

We assessed two methods of quantifying targets—reference-based, by mapping reads to database reference sequences, and contig-based, where reads were mapped to *de novo* assembled contigs. To compare detection thresholds with these two approaches, we quantified the spike-in standards and a bacteriophage that was spiked into samples at low concentrations. The detection threshold of the standards with the reference-based method was approximately 500 gc/µL, although detection thresholds varied with target length. This is approximately 100-fold lower

than the limits of detection reported in manure metagenomes (11,12), despite the previous studies having less stringent limits of quantification (e.g., one read). This discrepancy is likely due to the four- to five-fold lower sequencing effort and increased complexity of the communities studied previously. The detection threshold of the standards with the contig-based method was approximately 1000 gc/µL, which is twice as high as the reference-based. This trend is expected given that contig formation depends on input concentrations. In addition to comparing reference-based and contig-based quantification with standards, we compared quantification with a marine phage HM1 genome that was spiked into the six samples at low abundances. HM1 was detectable and quantifiable in four of six samples using the reference-based approach as opposed to one of six samples with the contig-based approach. These results further demonstrate that *de novo* contig formation required a higher input concentration than the detection threshold based on read coverage and distribution.

Our quantitative metagenomics detection thresholds are higher than those of quantitative PCR methods, such as qPCR or ddPCR, which can be less than 10 gc/µL (48). Quantitative PCR therefore remains a more appropriate method for quantifying low abundance targets. As sequencing costs continue to decline, however, greater sequencing efforts and lower sequencing error rates will result in lower detection thresholds for quantitative metagenomics. When applying quantitative metagenomics methods, researchers should consider community complexity and anticipated target concentrations when determining sequencing depths. Lower quantification thresholds are expected with lower community complexity and increased sequencing depth. Community complexity can be reduced with target enrichment steps, such as the virus enrichment performed on our wastewater samples, that increased the virus-to-cellular genome ratio (24).

To evaluate the influence of the reference-based and contig-based approaches on target quantification, we quantified the spike-in standards and viral targets of interest in our sample viromes. For standards that were quantifiable with both reference-based and contig-based approaches, the mean difference between reference-based and contig-based concentrations was only 1.2% (95% CI: 0.99 - 1.4%). However, when both approaches were applied to quantify crAss-like phages, the contig-based quantification resulted in significantly higher concentrations of targets, with differences ranging from 84-160%. Our reference-based quantification relied on mapping reads to database sequences that represent a previously sampled and sequenced population variant archived in a public sequence database. In order to quantify crAss-like phages using the reference-based approach, reads closely matching the crAss-like phage genomes in RefSeq are required. Contig-based quantification, however, maps reads to *de novo* contigs created from population variants present in a sample. The contig-based method captures population variants that may be absent from databases and may not pass detection or quantification thresholds with the reference-based method. Future applications of QuantMeta can incorporate different levels of population variation when defining references (e.g., contigs, metagenome assembled genomes, genomes or genes from databases) to quantify targets. The extent of population variation considered depends on the study goals. For example, quantifying the concentration of a strain-level variant would require mapping reads to a specific reference genome or to populations operationally defined to include minimal genome variation. On the other hand, the concentration of a broader group of targets (e.g., all crAss-like phage) can be quantified using a broader reference population that encompasses greater population-level variation.

One of the most powerful applications of this approach is the ability to quantify total viruses in a sample. This has not been possible with quantitative metagenomic methods that rely on normalizing by viral counts, e.g., based on flow cytometry or microscopy quantification (given the inability to quantitatively confirm viral particles) or qPCR quantification (given the lack of conserved viral genes). We have overcome this obstacle with our use of spike-in synthetic standards, that included both single- and double-stranded DNA, added to wastewater viromes prepared to maximize viral recovery and purity (24). The measured total DNA virus concentrations in wastewater influent and effluent samples were greater than $1x10^{10}$ gc/mL, with effluent samples containing slightly more DNA viruses than influent samples. Previous studies using epifluorescent microscopy or flow cytometry to measure virus-like particle (VLP) concentrations in influent and effluent have reported lower concentrations, ranging from $1.0x10^{8}$-$7.1x10^{8}$ VLP/mL and $1.0x10^{8}$-$4.0x10^{8}$ VLP/mL in influent and effluent, respectively (49-51). The 1-2 orders of magnitude higher total viral concentrations measured with quantitative metagenomics may be due to the fact that epifluorescent microscopy and flow cytometry rely on the detection of intercalating dyes that bind with genomes. These techniques underestimate counts of viruses with smaller genomes, including ssDNA viruses (52). Previous work suggested that epifluorescent microscopy detects 1- and 2-log fewer dsDNA and ssDNA viruses than target-specific qPCR measurements, respectively (52). Because the quantitative metagenomics approach depends on our ability to distinguish viral sequences from cellular genomes *in silico*, the accuracy of total virus concentrations with this approach will continue to improve as our ability to accurately identify viral sequences in metagenomic datasets improves (1).

Further, the QuantMeta method facilitates the quantification of many different viruses simultaneously without introducing biases from PCR primers, which has important implications for detection of human viruses in wastewater epidemiology. We measured mean crAss-like phage concentrations of $2.5x10^{7}$ and $1.0x10^{6}$ gc/mL in influent and effluent samples, whereas previous studies measured a crAss-like phage marker gene to be $6.9x10^{1}$-$1.1x10^{9}$ gc/mL and $5.6x10^{2}$-$1.0x10^{6}$ gc/mL in influent and effluent samples, respectively (53-59). We hypothesized that our levels of crAss-like phage were on the higher end of previously reported ranges because the commonly used CPQ056 primer set, designed in 2017 based on one of the now 129 available crAss-like phage genomes, may not capture the full genomic diversity of the crAss-like phage population in our samples. When the *in silico* quantification of crAss-like phage was limited to those targeted by the CPQ056 primer set, the average crAss-like phage concentration dropped 81% and 50% for reference-based and contig-based quantification methods, respectively (*p*-values = $9.1x10^{-9}$ and 0.033). This highlights an important advantage of quantitative metagenomics compared to qPCR or ddPCR, namely that it is not biased by primers that fail to capture all of the targets of interest, and supports similar findings when quantitative metagenomics has been used to quantify antibiotic resistance genes in agriculture systems (12).

When we measured DNA virus pathogen concentrations, reads mapped onto all types of human DNA virus pathogens (Figure 4). However, only a few pathogens met our quantitative limitations including polyomaviruses, papillomaviruses, bocavirus, and adenovirus. Few additional pathogens were within 10% of achieving the detection threshold. Some of DNA virus pathogens failing detection thresholds are likely present in the wastewater, but another method with a lower limit of detection, such as sequencing amplified products of specific targets, or increasing sequencing depth is needed to confidently confirm their presence. Our quantitative metagenomics-based JC and BK polyomavirus concentrations were consistent with previously

reported concentrations in influent. We measured polyomavirus concentrations of $3.4x10^4$-$1.0x10^5$ gc/mL, while previous studies have reported ranges of $8.9x10^0$-$2.0x10^5$ gc/mL for influent, respectively (53-56,60). Additional sequencing effort is needed to confidently determine presence and quantify other virus pathogens in the wastewater samples despite sequencing approximately 200 million reads per sample. The high detection thresholds of quantitative metagenomics relative to quantitative PCR methods impedes implementing the method to monitor pathogens in environmental matrices.

Our quantitative metagenomic approach is publicly available as QuantMeta (github.com/klangenf/QuantMeta) and provides R notebooks for users to alter detection and quantification thresholds for their specific applications. More or less stringent entropy detection thresholds can be applied based on coverage and read distribution specifications provided by the investigator. For example, Castro et al. (2018) suggested parameters specific to bacterial population detection. Some results in this publication may be specific to our sequencing depth, Illumina NovaSeq SP flowcells with 251-bp paired-end reads, and our selected library preparation methods. Future work should address how the read depth variability RMSE thresholds alter depending on the selected methods and sequencing depth.

DATA AVAILABILITY

All sequencing data are available in the NCBI BioProject database under the accession PRJNA853368. All relevant code is compiled as QuantMeta and available on Github at github.com/klangenf/QuantMeta.

FUNDING

This research was financially supported by the NSF PIRE award number 1545756 and partially supported by a University of Michigan Integrated Training in Microbial Systems (ITiMS) Mini-Grant to K.L., University of Michigan research funding to M.B.D., and a University of Michigan Rackham graduate student research grant to K.L. K.L. and E.C. were funded by NSF GRFPs and ITiMS fellowships. The ITiMS program at the University of Michigan is supported by the Burroughs Wellcome Fund.

CONFLICT OF INTEREST DISCLOSURE

The authors declare no conflicts of interest.

ACKNOWLEDGEMENTS

We thank Jacob Evans for improving the efficiency of the script listing the average GC content and read depths of 49-bp sliding windows and Eric Bastian for providing the viral RefSeq database divided into open reading frames. We thank the lead investigator of the NSF PIRE award, Peter Vikesland, for his leadership of the research study.

REFERENCES

1. Tas, N., de Jong, A.E., Li, Y., Trubl, G., Xue, Y. and Dove, N.C. (2021) Metagenomic tools in microbial ecology research. *Curr Opin Biotechnol*, **67**, 184-191.
2. Yang, Y., Jiang, X., Chai, B., Ma, L., Li, B., Zhang, A., Cole, J.R., Tiedje, J.M. and Zhang, T. (2016) ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database. *Bioinformatics*, **32**, 2346-2351.

3. Vandeputte, D., Kathagen, G., D'Hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., Wang, J., Tito, R.Y., De Commer, L., Darzi, Y. *et al.* (2017) Quantitative microbiome profiling links gut community variation to microbial load. *Nature*, **551**, 507-511.
4. Jian, C., Luukkonen, P., Yki-Jarvinen, H., Salonen, A. and Korpela, K. (2020) Quantitative PCR provides a simple and accessible method for quantitative microbiota profiling. *PLoS One*, **15**, e0227285.
5. Majeed, H.J., Riquelme, M.V., Davis, B.C., Gupta, S., Angeles, L., Aga, D.S., Garner, E., Pruden, A. and Vikesland, P.J. (2021) Evaluation of Metagenomic-Enabled Antibiotic Resistance Surveillance at a Conventional Wastewater Treatment Plant. *Front Microbiol*, **12**, 657954.
6. Hardwick, S.A., Chen, W.Y., Wong, T., Kanakamedala, B.S., Deveson, I.W., Ongley, S.E., Santini, N.S., Marcellin, E., Smith, M.A., Nielsen, L.K. *et al.* (2018) Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. *Nat Commun*, **9**, 3096.
7. Santini, N.S., Lovelock, C.E., Hua, Q., Zawadzki, A., Mazumder, D., Mercer, T.R., Muñoz-Rojas, M., Hardwick, S.A., Madala, B.S., Cornwell, W. *et al.* (2019) Natural and Regenerated Saltmarshes Exhibit Similar Soil and Belowground Organic Carbon Stocks, Root Production and Soil Respiration. *Ecosystems*, **22**, 1803-1822.
8. Reis, A.L.M., Deveson, I.W., Wong, T., Madala, B.S., Barker, C., Blackburn, J., Marcellin, E. and Mercer, T.R. (2020) A universal and independent synthetic DNA ladder for the quantitative measurement of genomic features. *Nat Commun*, **11**, 3609.
9. Shen, J., McFarland, A., Blaustein, R., Rose, L., Perry-Dow, K., Hayden, M., Young, V. and Hartmann, E. (2022) An improved workflow for accurate and robust healthcare environmental surveillance using metagenomics. *Research Square*.
10. Yang, Y., Che, Y., Liu, L., Wang, C., Yin, X., Deng, Y., Yang, C. and Zhang, T. (2022) Rapid absolute quantification of pathogens and ARGs by nanopore sequencing. *Sci Total Environ*, **809**, 152190.
11. Li, B., Li, X. and Yan, T. (2021) A Quantitative Metagenomic Sequencing Approach for High-Throughput Gene Quantification and Demonstration with Antibiotic Resistance Genes. *Appl Environ Microbiol*, **87**, e0087121.
12. Crossette, E., Gumm, J., Langenfeld, K., Raskin, L., Duhaime, M.B. and Wigginton, K.R. (2021) Enhancing metagenomic quantification of genes in environmental samples with internal standards. *mBio*, **in press**.
13. Satinsky, B.M., Gifford, S.M., Crump, B.C. and Moran, M.A. (2013) Use of internal standards for quantitative metatranscriptome and metagenome analysis. *Methods Enzymol*, **531**, 237-250.
14. Stammler, F., Glasner, J., Hiergeist, A., Holler, E., Weber, D., Oefner, P.J., Gessner, A. and Spang, R. (2016) Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome*, **4**, 28.
15. Lin, Y., Gifford, S., Ducklow, H., Schefield, O. and Cassar, N. (2019) Towards Quantitative Microbiome Community Profiling Using Internal Standards. *Applied and Environmental Microbiology*, **85**, e02634-02618.
16. Mercer, T.R. (2020). Garvan Institute of Medical Research, Vol. 2022.
17. Weinroth, M.D., Thomas, K.M., Doster, E., Vikram, A., Schmidt, J.W., Arthur, T.M., Wheeler, T.L., Parker, J.K., Hanes, A.S., Alekoza, N. *et al.* (2022) Resistomes and

microbiome of meat trimmings and colon content from culled cows raised in conventional and organic production systems. *Anim Microbiome*, **4**, 21.
18. Armbruster, D. and Pry, T. (2008) Limit of Blank, Limit of Detection and Limit of Quantitation. *Clinical Biochemistry Review*, **29**.
19. Forootan, A., Sjoback, R., Bjorkman, J., Sjogreen, B., Linz, L. and Kubista, M. (2017) Methods to determine limit of detection and limit of quantification in quantitative real-time PCR (qPCR). *Biomol Detect Quantif*, **12**, 1-6.
20. Shen, J., McFarland, A.G., Young, V.B., Hayden, M.K. and Hartmann, E.M. (2021) Toward Accurate and Robust Environmental Surveillance Using Metagenomics. *Front Genet*, **12**, 600111.
21. Bharucha, T., Oeser, C., Balloux, F., Brown, J.R., Carbo, E.C., Charlett, A., Chiu, C.Y., Claas, E.C.J., de Goffau, M.C., de Vries, J.J.C. *et al.* (2020) STROBE-metagenomics: a STROBE extension statement to guide the reporting of metagenomics studies. *The Lancet Infectious Diseases*, **20**, e251-e260.
22. Zhou, Z., Luhmann, N., Alikhan, N.-F., Quince, C. and Achtman, M. (2017) Accurate Reconstruction of Microbial Strains Using Representative Reference Genomes. *bioRxiv*.
23. Hsieh, P.H., Oyang, Y.J. and Chen, C.Y. (2019) Effect of de novo transcriptome assembly on transcript quantification. *Sci Rep*, **9**, 8304.
24. Langenfeld, K., Chin, K., Roy, A., Wigginton, K. and Duhaime, M.B. (2021) Comparison of ultrafiltration and iron chloride flocculation in the preparation of aquatic viromes from contrasting sample types. *PeerJ*, **9**, e11111.
25. Lievens, A., Jacchia, S., Kagkli, D., Savini, C. and Querci, M. (2016) Measuring Digital PCR Quality: Performance Parameters and Their Optimization. *PLoS One*, **11**, e0153317.
26. Warwick-Dugdale, J., Solonenko, N., Moore, K., Chittick, L., Gregory, A.C., Allen, M.J., Sullivan, M.B. and Temperton, B. (2019) Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ*, **7**, e6800.
27. Giordano, F., Aigrain, L., Quail, M.A., Coupland, P., Bonfield, J.K., Davies, R.M., Tischler, G., Jackson, D.K., Keane, T.M., Li, J. *et al.* (2017) De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Sci Rep*, **7**, 3935.
28. Brown, B.L., Watson, M., Minot, S.S., Rivera, M.C. and Franklin, R.B. (2017) MinION nanopore sequencing of environmental metagenomes: a synthetic approach. *Gigascience*, **6**, 1-10.
29. Roux, S., Enault, F., Hurwitz, B.L. and Sullivan, M.B. (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ*, **3**, e985.
30. Guo, J., Bolduc, B., Zayed, A.A., Varsani, A., Dominguez-Huerta, G., Delmont, T.O., Pratama, A.A., Gazitua, M.C., Vik, D., Sullivan, M.B. *et al.* (2021) VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*, **9**, 37.
31. Kieft, K., Zhou, Z. and Anantharaman, K. (2020) VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, **8**, 90.
32. Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A. and Sun, F. (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, **5**, 69.

33. Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S. and Kyrpides, N.C. (2021) CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol*, **39**, 578-585.
34. Roux, S. (2015). GitHub.
35. Kuhl, M.A., Stich, B. and Ries, D.C. (2021) Mutation-Simulator: fine-grained simulation of random mutations in any genome. *Bioinformatics*, **37**, 568-569.
36. Stachler, E., Kelty, C., Sivaganesan, M., Li, X., Bibby, K. and Shanks, O.C. (2017) Quantitative CrAssphage PCR Assays for Human Fecal Pollution Measurement. *Environ Sci Technol*, **51**, 9146-9154.
37. Tithi, S.S., Aylward, F.O., Jensen, R.V. and Zhang, L. (2018) FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data. *PeerJ*, **6**, e4227.
38. Castro, J.C., Rodriguez, R.L., Harvey, W.T., Weigand, M.R., Hatt, J.K., Carter, M.Q. and Konstantinidis, K.T. (2018) imGLAD: accurate detection and quantification of target organisms in metagenomes. *PeerJ*, **6**, e5882.
39. Shannon, C. and Weaver, W. (1963) The mathematical theory of communication. *Urbana*, **3**.
40. Ludwig, J. and Reynolds, J. (1988) *Statistical ecology: a primer in methods and computing*. John Wiley & Sons.
41. Ekblom, R., Smeds, L. and Ellegren, H. (2014) Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria. *BMC Genomics*, **15**.
42. Browne, P.D., Nielsen, T.K., Kot, W., Aggerholm, A., Gilbert, M.T.P., Puetz, L., Rasmussen, M., Zervas, A. and Hansen, L.H. (2020) GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. *Gigascience*, **9**.
43. Commichaux, S., Shah, N., Ghurye, J., Stoppel, A., Goodheart, J., Luque, G., Cummings, M. and Pop, M. (2021) A critical assessment of gene catalogs for metagenomic analysis. *Bioinformatics*, **37**, 2848-2857.
44. Comeau, A.M., Tremblay, D., Moineau, S., Rattei, T., Kushkina, A.I., Tovkach, F.I., Krisch, H.M. and Ackermann, H.W. (2012) Phage morphology recapitulates phylogeny: the comparative genomics of a new group of myoviruses. *PLoS One*, **7**, e40102.
45. Zeng, W., Mao, P., Hong, Y., Feng, M., Xu, Z., Huang, F. and Jing, S. (2015) Complete Genome Sequence of the Salmonella enterica Serovar Paratyphi A Bacteriophage LSPA1 Isolated in China. *Genome Announc*, **3**.
46. Shkoporov, A.N., Khokhlova, E.V., Fitzgerald, C.B., Stockdale, S.R., Draper, L.A., Ross, R.P. and Hill, C. (2018) PhiCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nat Commun*, **9**, 4781.
47. Koonin, E.V. and Yutin, N. (2020) The crAss-like Phage Group: How Metagenomics Reshaped the Human Virome. *Trends Microbiol*, **28**, 349-359.
48. Wang, Y., Zheng, G., Wang, D. and Zhou, L. (2022) Occurrence of bacterial and viral fecal markers in municipal sewage sludge and their removal during sludge conditioning processes. *J Environ Manage*, **310**, 114802.
49. Ma, L., Mao, G., Liu, J., Yu, H., Gao, G. and Wang, Y. (2013) Rapid quantification of bacteria and viruses in influent, settled water, activated sludge and effluent from a wastewater treatment plant using flow cytometry. *Water Sci Technol*, **68**, 1763-1769.
50. Rosario, K., Nilsson, C., Lim, Y.W., Ruan, Y. and Breitbart, M. (2009) Metagenomic analysis of viruses in reclaimed water. *Environ Microbiol*, **11**, 2806-2820.

51. Wu, Q. and Liu, W.T. (2009) Determination of virus abundance, diversity and distribution in a municipal wastewater treatment plant. *Water Res*, **43**, 1101-1109.
52. Kaletta, J., Pickl, C., Griebler, C., Klingl, A., Kurmayer, R. and Deng, L. (2020) A rigorous assessment and comparison of enumeration methods for environmental viruses. *Sci Rep*, **10**, 18625.
53. Tandukar, S., Sherchan, S.P. and Haramoto, E. (2020) Applicability of crAssphage, pepper mild mottle virus, and tobacco mosaic virus as indicators of reduction of enteric viruses during wastewater treatment. *Sci Rep*, **10**, 3616.
54. Malla, B., Ghaju Shrestha, R., Tandukar, S., Sherchand, J.B. and Haramoto, E. (2019) Performance Evaluation of Human-Specific Viral Markers and Application of Pepper Mild Mottle Virus and CrAssphage to Environmental Water Samples as Fecal Pollution Markers in the Kathmandu Valley, Nepal. *Food Environ Virol*, **11**, 274-287.
55. Crank, K., Li, X., North, D., Ferraro, G.B., Iaconelli, M., Mancini, P., La Rosa, G. and Bibby, K. (2020) CrAssphage abundance and correlation with molecular viral markers in Italian wastewater. *Water Res*, **184**, 116161.
56. Wu, Z., Greaves, J., Arp, L., Stone, D. and Bibby, K. (2020) Comparative fate of CrAssphage with culturable and molecular fecal pollution indicators during activated sludge wastewater treatment. *Environ Int*, **136**, 105452.
57. Ahmed, W., Lobos, A., Senkbeil, J., Peraud, J., Gallard, J. and Harwood, V.J. (2018) Evaluation of the novel crAssphage marker for sewage pollution tracking in storm drain outfalls in Tampa, Florida. *Water Res*, **131**, 142-150.
58. Ahmed, W., Payyappat, S., Cassidy, M., Besley, C. and Power, K. (2018) Novel crAssphage marker genes ascertain sewage pollution in a recreational lake receiving urban stormwater runoff. *Water Res*, **145**, 769-778.
59. Garcia-Aljaro, C., Balleste, E., Muniesa, M. and Jofre, J. (2017) Determination of crAssphage in water samples and applicability for tracking human faecal pollution. *Microb Biotechnol*, **10**, 1775-1780.
60. McQuaig, S.M., Scott, T.M., Lukasik, J.O., Paul, J.H. and Harwood, V.J. (2009) Quantification of human polyomaviruses JC Virus and BK Virus by TaqMan quantitative PCR and comparison to other water quality indicators in water and fecal samples. *Appl Environ Microbiol*, **75**, 3379-3388.