Vol. 81: 19–35, 2018
https://doi.org/10.3354/ame01856

AQUATIC MICROBIAL ECOLOGY
Aquat Microb Ecol

Published online January 17

Quantifying aquatic viral community change associated with stormwater runoff in a wet retention pond using metagenomic time series data

Jasmin C. Green[1], Faraz Rahman[1], Matthew A. Saxton[2,3], Kurt E. Williamson[1,*]

[1]Department of Biology, College of William & Mary, Williamsburg, VA 23185, USA
[2]Environmental Science and Policy Program, College of William & Mary, Williamsburg, VA 23185, USA
[3]*Present address:* Department of Marine Sciences, University of Georgia, Athens, GA 30602, USA

ABSTRACT: In the United States, wet retention ponds are a common best management practice for controlling peak stormwater flows from impervious surfaces. Currently, studies evaluating the microbial component of these landscape features exclude community response to natural disturbances. We conducted a series of metagenomic analyses to quantify viral community change in a wet retention pond over the course of a storm event, and to explore 2 potential sources and reservoirs of water column viral diversity: soil in the surrounding pond watershed and pond sediments. Our results indicate that the viral communities of different sample types (soil, sediment, and water) are distinct, and that watershed soils and sediments do not appear to be primary sources of novel viral genotypes detected in the pond water column following a storm event. Though it is apparent that rapid and dramatic change in viral community composition occurs in response to the input of stormwater runoff — as evidenced by a sharp rise in richness, an increase in functional diversity, and increasing dissimilarity of water column viromes over time — the source of this change could not be determined. This study demonstrates the need for additional time-series metagenomic explorations to gain a complete picture of viral community dynamics in aquatic environments.

KEY WORDS: Virome · Time series · Best management practice · Microbe · Mitigation · Freshwater · Viral community composition

INTRODUCTION

Wet retention ponds are constructed basins commonly employed as a best management practice (BMP) in urban and suburban locales to mitigate peak stormwater flows. By channeling runoff, wet retention ponds slow the flow of water and allow pollutants to settle out passively over time before the remaining supernatant is displaced from the pond and carried to downstream bodies of water (US EPA 1999). The increase in impervious surfaces (e.g. asphalt, concrete) from human land development results in higher peak flows and greater runoff volumes (Leopold 1968). Runoff can cause erosion and disperse pollutants, making stormwater runoff an important driver of environmental change.

Although wet retention ponds are human-designed features, they become part of the natural landscape to be used by humans, pets, migrating birds, amphibians, emergent vegetation, and, importantly, microbes. In addition to physical and chemical inputs (e.g. sediments and pesticides) to downstream bodies of water, stormwater runoff contains an important but poorly characterized microbial component. The first forays into understanding the microbial community contained within wet retention ponds featured the detection of fecal indicator bacteria as a benchmark for public health and a proxy for other harmful pathogens (US EPA 2006). However, very little is known regarding the microbial communities, particularly viral communities, that develop in these ponds (Saxton et al. 2016).

[*] Corresponding author: kewilliamson@wm.edu

Viruses are widely acknowledged to be the most abundant biological agents in the world, outnumbering bacteria by an estimated 10-fold in aquatic ecosystems (Suttle 2007). Viruses influence host community dynamics by a combination of top-down and bottom-up control: viral lysis controls the abundance of host organisms that bloom to excess and also releases macromolecules into the system, creating competition over nutrient availability among host populations (Thingstad 2000). In this way, viruses influence microbial community composition, which has downstream effects on the food web and ecosystem function. Despite the widespread use of wet retention ponds, the viral communities that develop in such ponds remain largely uncharacterized, and their impacts on pond microbial community structure and function are unknown.

Thus far, investigations into the influence of runoff on aquatic viral communities have used qPCR and RAPD-PCR (Hewson et al. 2012, Williamson et al. 2014). Stormwater runoff drives change in the viral community of the receiving wet retention pond (Williamson et al. 2014), and the water column abundance of viral genotypes associated with watershed soil corresponds with rainfall and storm events in a given catchment (Hewson et al. 2012). One advantage of PCR methods is that a large number of samples can be analyzed so that a finer temporal or spatial gradient can be assessed over the course of the study. A major disadvantage, however, is that the results provide a limited view of viral taxonomy: RAPD-PCR provides a broad overview of changes to the community without indicating what genotypes may be represented by a given band, while qPCR can only provide information about a specifically targeted viral genotype. Metagenomics is an increasingly popular method for studying viral ecology because of its ability to provide a high level of taxonomic detail (e.g. Angly et al. 2006, López-Bueno et al. 2009, Bolduc et al. 2015). However, library construction can be expensive (although costs are decreasing), and analyzing the resulting, often large, datasets is computationally intensive. As a result, the majority of published viromes tend to represent single time point samples from a given environment. These 'snapshot' views, while providing an important in-depth view of the extant viral community, are unable to provide insights into community change over time. In this study, we provide viral metagenomic time series to obtain a fine-grained view of the taxonomic and metabolic community-level changes that occur within a wet retention pond over the course of a storm event.

Previous work has indicated that storm events can drive significant changes in the viral community composition of wet retention ponds (Williamson et al. 2014). Here, we tested 2 non-exclusive hypotheses that could explain the source of novel viral genotypes that were detected in the water column over the course of a storm: (1) viruses are transported from watershed soils into the pond; and/or (2) viruses are re-suspended from pond sediments into the water column. Understanding the magnitude of changes that occur in the viral community over the course of a storm event and identifying possible sources of transported microbes will allow us to better understand and predict potential ecological changes (e.g. community composition and function) and public health hazards (e.g. introduction and exposure to pathogens) associated with wet retention ponds.

MATERIALS AND METHODS

Sample collection

The sampling site for this study was the primary wet retention pond for the Longhill Grove apartment complex, located in Williamsburg, Virginia, USA (37.315° N, 76.787° W; Fig. 1). This pond was constructed in 2003 to control storm water runoff from the surrounding impervious surfaces in the complex. The pond is situated in Emporia loamy fine sand in the Virginia Coastal Plain, Chesapeake Bay watershed. The pond's maximum depth is 2.23 m with 263 m^2 surface area, and it drains a 6.8 ha area with a runoff coefficient (ratio of total site runoff volume to total rainfall volume) of 0.44 (Hancock et al. 2010). Surface water, sediment, and soil samples were collected on 6 June 2013 ($t = 0$ h, 'Initial'), approximately 24 h prior to Tropical Storm Andrea, which produced 150.6 mm rainfall over the next 55 h. Surface water samples were collected again on 7 June (t = 24 h, 'Early'), 8 June ($t = 49$ h 'Late'), and 10 June ($t = 154$ h, 'Post'). Water samples were collected by hand from several points along the pond perimeter and from the spillway in acid-washed 2 l polycarbonate bottles. Soil samples were collected from the surrounding watershed using an ethanol-sterilized hand trowel. A composite 1 kg sample was generated by combining multiple individual samples excavated with the trowel into a single, sterile Whirl-pak bag. Sediment samples were collected from multiple locations along the perimeter of the pond using an ethanol-sterilized plastic scoop and combined into a 1 kg composite sample in a sterile Whirl-pak bag. All samples were

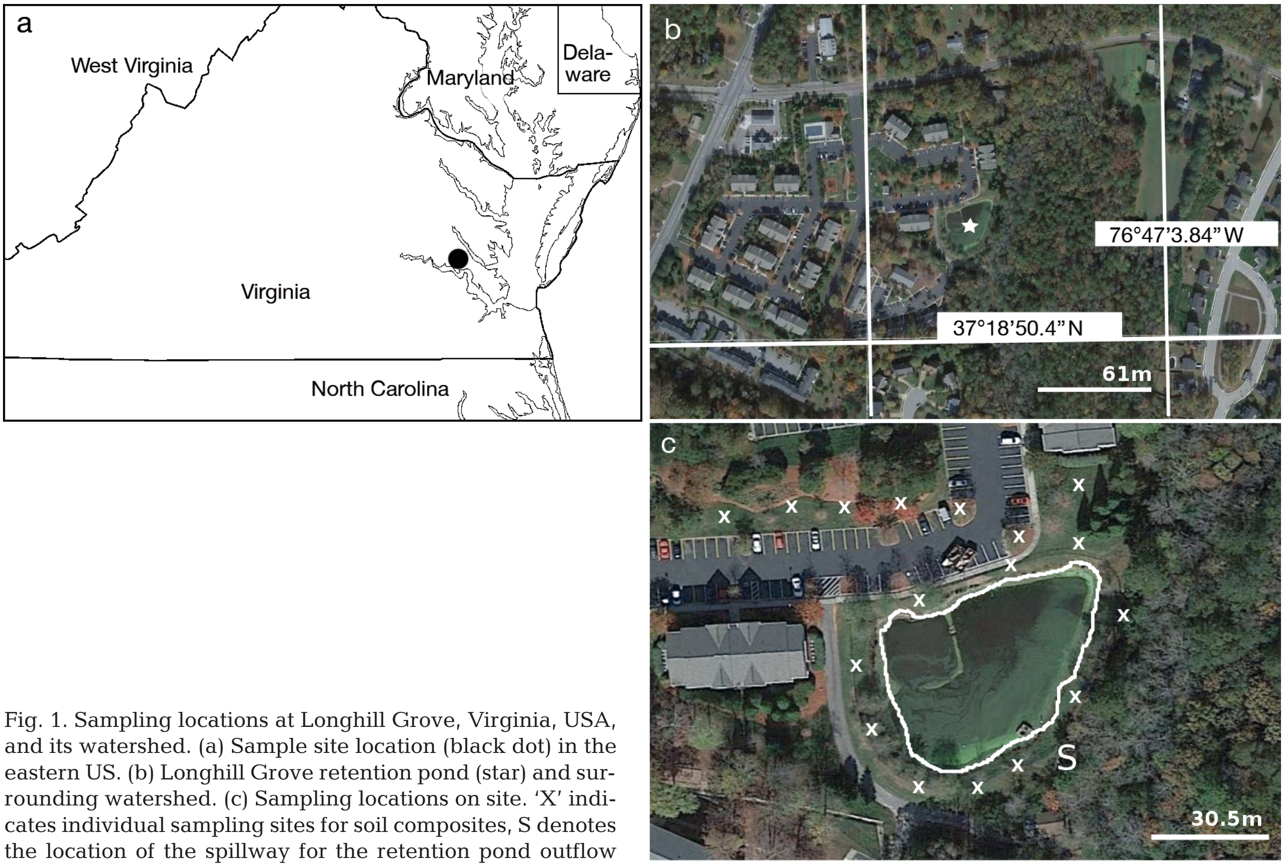

Fig. 1. Sampling locations at Longhill Grove, Virginia, USA, and its watershed. (a) Sample site location (black dot) in the eastern US. (b) Longhill Grove retention pond (star) and surrounding watershed. (c) Sampling locations on site. ‘X’ indicates individual sampling sites for soil composites, S denotes the location of the spillway for the retention pond outflow

transported to the lab on ice (approximately 20 min travel time) and stored at −80°C until processing.

Environmental metadata

Viral and bacterial abundance in water samples was determined using epifluorescence microscopy from triplicate subsamples (Noble & Fuhrman 1998, Hardbower et al. 2012). Viral and bacterial abundances in soil and sediment samples were determined using triplicate sub-samples as previously described (Williamson et al. 2013). Virus to bacterium ratios were calculated based on averages of both viral and bacterial abundances for each sample. A YSI-63 hand-held multimeter was used to measure temperature and conductivity and a YSI-55 hand-held probe was used to measure dissolved oxygen in the field. Water pH was measured in the laboratory using an UltraBasic pH probe (Denver Instruments). Nutrient levels ($NO_2 + NO_3$, NH_4 and inorganic phosphorus) were determined using colorimetric assays with water filtered through Whatman glass fiber filters (GF/F) (Parsons et al. 1984). Composite soil and sediment samples were air-dried, homogenized, and sieved to 2 mm prior to analysis. Chemical and physical analyses were performed by A&L Eastern Laboratories (Richmond, VA). Sample metadata are summarized in Table 1.

Viral metagenome construction

Water samples were processed as previously described (Thurber et al. 2009). Briefly, 4 l samples were passed through 0.22 µm bottle-top filters (Steri-top, Millipore), and viral particles in the filtrate were precipitated using polyethylene glycol (PEG-8000, 10% w/v) and NaCl (1 M), incubating at 4°C overnight. Samples were then centrifuged at 8000 × *g* (30 min at 4°C). Viral particles from the PEG pellet were purified using CsCl density gradient ultracentrifugation (1.70, 1.50, and 1.35 g ml^{-1} layers in SM buffer; viruses recovered from 1.50–1.35 g ml^{-1} interface) and treated with RQ1 DNase (Promega). Nucleic acids for the DNA libraries were extracted using the formamide procedure (Thurber et al. 2009). Nucleic acids for cDNA libraries were extracted using the QIAgen RNEasy Mini Kit according to the manufacturer’s instructions. Extracted RNA was treated with 2.5 U ml^{-1}

Table 1. Environmental metadata for metagenomic libraries. VA: viral abundance (SD), BA: bacterial abundance (SD), DO: dissolved oxygen, TSS: total suspended solids, DW: dry weight, OM: organic matter (determined by loss on ignition), CEC: cation exchange capacity

Sample	Date (2013)	Time (h)	VA (10^7) ml^{-1}	BA (10^5) ml^{-1}	pH	Temp (°C)	DO (mg l^{-1})	NO_2+NO_3 (µM)	NH_4 (µM)	PO_4 (µM)	TSS (mg l^{-1})		
Initial	6 June	14:00	1.27 (0.04)	5.1 (0.91)	8.62	27.5	10.94	0.2	4.4	2.6	10.2		
Early	7 June	14:00	1.17 (0.03)	6.5 (0.39)	6.56	22.8	6.45	0.7	1.0	2.2	19.6		
Late	8 June	15:00	0.56 (0.001)	0.69 (0.13)	6.15	27.9	6.31	1.6	1.5	2.5	20.3		
Post	12 June	15:00	0.52 (0.002)	1.45 (0.29)	6.10	27.3	8.07	0.2	0.7	6.8	19.3		
Sample	Date (2013)	Time (h)	VA (10^8) g^{-1} DW	BA (10^6) g^{-1} DW	pH	OM (%)	K (mg kg^{-1})	Mg (mg kg^{-1})	Ca (mg kg^{-1})	CEC (meq 100g^{-1})	% Sand	% Silt	% Clay
Soil	6 June	14:30	6.5 (0.03)	3.79 (0.56)	5.4	2.1	40	56	573	4.8	64.8	23.6	11.6
Sediment	6 June	14:45	15 (0.22)	4.06 (0.12)	6.2	6.3	123	100	1865	11.9	64.8	25.6	9.6

RQ1 DNase at 37°C for 1 h. DNase was then destroyed using the stop solution provided by the manufacturer before converting the extracted RNA to cDNA using the Epicenter MMLV Reverse Transcriptase First-Strand cDNA Synthesis Kit. Although the original intent of these procedures was to generate RNA virus metagenomes, the resulting libraries did not appear to achieve this goal. Due to the manner in which they were constructed, these libraries are referred to as 'cDNA libraries' from this point forward.

All viral nucleic acids were amplified using the Illustra Genomiphi V2 DNA Amplification Kit (GE Healthcare Life Sciences); details on sample preparation and nucleic acid yields can be found in Table S1 in the Supplement at www.int-res.com/articles/suppl/a081p019_supp.pdf. Duplicate reactions were pooled and then purified using the QIAgen DNeasy Blood and Tissue Kit before sequencing on a Roche Applied Sciences GS-FLX+ platform (454 Life Sciences). Library construction and sequencing were performed at Selah Genomics, Greenville, SC, USA.

For soil and sediment samples, viral particles were extracted using established methods (Williamson et al. 2013). Briefly, composite soil and sediment samples were sieved to 2 mm, and viral particles were extracted by blending samples in 1% potassium citrate buffer (100 g sample:300 ml of buffer). The resulting slurries were centrifuged and the supernatants were passed through 0.22 µm bottle-top filters (SteriTop, Millipore). Filtrates were processed as described above to generate soil and sediment viral metagenomes.

Taxonomic and functional annotations

All libraries were dereplicated prior to analysis. All reads passing quality control (QC) were annotated using MG-RAST v3.3.6 (Meyer et al. 2008) and Metavir v2.0 (Roux et al. 2014) with an E-value cutoff of 10^{-5}. MG-RAST generates taxonomic assignments based on BLASTx searches against the M5NR database (which includes SEED, KEGG, NCBI nr, Phantome, GO, EBI, JGI, UniProt, VBI, and eggNOG), and functional assignments based on BLASTx searches against the SEED-Subsystem database. Metavir generates taxonomic assignments of virus-affiliated sequences based on BLASTx searches against the RefSeq Virus database.

Contig assembly

Viral metagenomes were assembled using the GS De Novo Assembler v2.8 (Roche Diagnostics). Open reading frames (ORFs) were annotated in Metavir, which predicts ORFs for each contig through MetaGeneAnnotator and compares them to the Refseq Virus protein database by BLASTp (E-value cutoff of 10^{-3}) and by HMMScan (Bit score cutoff of 30) to the PFAM database (Roux et al. 2014).

Viral community structure and diversity

All reads passing QC were used in the following analyses. Contig spectra were generated by Circonspect using default parameters (Angly et al. 2006). Community structure and α-diversity were modeled in Phage Communities from Contig Spectrum (PHACCS; Angly et al. 2005) using the contig spectra from Circonspect and average genome size determined by the genome relative abundance and average size (GAAS) tool (Angly et al. 2009). Power law, exponential, logarithmic, and log-normal rank-abundance models were tested, and the best model was selected based on lowest error values. Reference-independent comparative metagenomics and change in viral community structure over time were estimated using cross-assembly of contigs based on Wootter's distance metric (Dutilh et al. 2012). Cluster trees were generated using the pvclust package in R (distance method = correlation; cluster method = average; bootstrap number = 10 000), based on distance matrices generated from BLASTx comparisons of the libraries.

RESULTS

Taxonomic distribution

A total of 421.9 Mbp were generated from the 6 DNA libraries, corresponding to 812 687 individual reads with an average read length of 513 bp. The number of sequences failing QC (based on duplicate read inferred sequencing error estimation [Keegan et al. 2012], k-mer profiles, and nucleotide bias within reads) varied from 13.2–32 % for DNA libraries (Fig. 2a). A total of 611.3 Mbp were generated from the 6 cDNA libraries, corresponding to 1 189 865 reads with an average read length of 330 bp. The cDNA libraries had a much higher overall proportion of sequences that failed QC as compared to the DNA libraries, varying from 31.1–80 % of reads within a given library (Fig. 2c). All sequences failing QC were removed from subsequent analysis.

Annotation of reads using MG-RAST indicated that 26.1–41.2 % of reads in the DNA libraries had homology to known sequences (Fig. 2a). Of the reads with known homology, the majority of sequences within any given library were affiliated with either viruses (31–78 %) or *Bacteria* (17–61 %, Fig. 2b). Of the sequences affiliated with *Bacteria*, the majority (31–69 %) were affiliated with *Proteobacteria* (Table S2 in the Supplement). The representation of reads affiliated with eukaryotic sequences was generally low (5–10 %), with the exception of the Late library, in which nearly a third (29 %) of sequences were affiliated with Eukaryota (Fig. 2b). Of these eukaryote-affiliated sequences in the Late library, 42 % were affiliated with Nematoda.

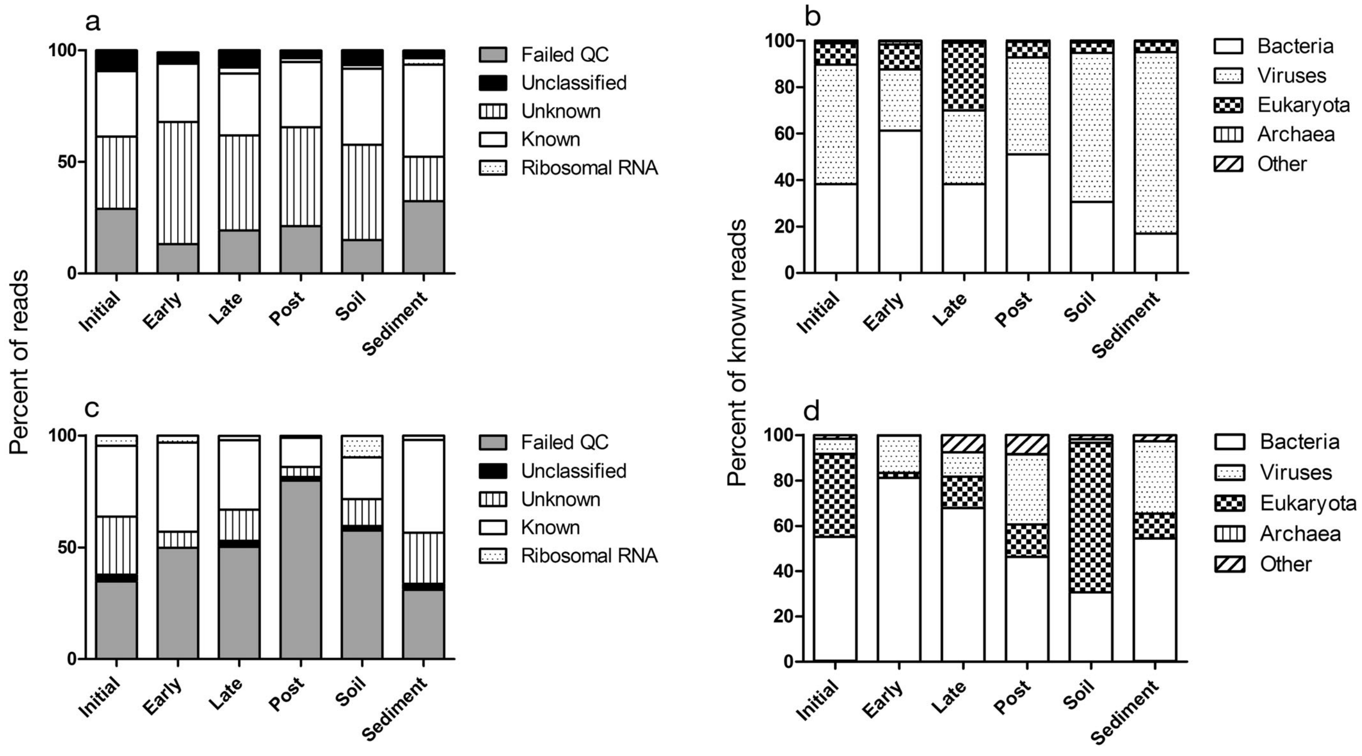

Fig. 2. (a,c) Classification of reads for (a) DNA and (c) cDNA libraries (Failed QC: did not meet minimum length/quality control parameters; Known: predicted protein of known function; Unknown: predicted protein of unknown function; Unclassified: sequence contained no predicted open reading frame). (b,d) Domain of known reads according to best BLAST hit using MG-RAST (E-value cutoff 10^{-5}) for (b) DNA and (d) cDNA libraries. Libraries: Initial (*t* = 0 h), Early (*t* = 24 h), Late (*t* = 49 h), Post (*t* = 154 h),), Soil (*t* = 0 h), Sediment (*t* = 0 h)

Annotation of reads in cDNA libraries indicated that 12.9–41.4 % of reads had homology to sequences in existing databases (Fig. 2c). Of these, reads affiliated with *Bacteria* were highly represented in all libraries (30–80 %), while reads affiliated with viruses (1.5–32 %) and Eukaryota (2–65 %) were much more variable by library (Fig. 2d). Of the sequences affiliated with *Bacteria* in the cDNA libraries, the majority (41.8–78.9 %) were affiliated with *Proteobacteria*, with the exception of the Early library, in which the majority (41.6 %) were affiliated with *Bacteroidetes* (Table S3). The Post and Sediment libraries contained the highest proportion of sequences affiliated with viruses (~30 %), while the Soil library contained the highest proportion of sequences affiliated with Eukaryota (~66 %, Fig. 2d). Of these eukaryote-affiliated sequences in the Soil library, 50.7 % were affiliated with Nematoda.

Viral DNA and cDNA used in library construction was screened for the presence of potential bacterial contamination by PCR amplification using 27f and 519r universal primers. Although all libraries were prepared at the same time using the same methods, 2 of the 12 libraries produced faint bands, suggesting the presence of bacterial DNA (Late and Post DNA libraries). When compared with the other DNA libraries, however, these 2 libraries did not contain a significantly higher proportion of reads affiliated with bacterial sequences (Mann-Whitney *U*-tests, $p > 0.05$). In spite of the higher proportion of reads affiliated with ribosomal RNA sequences in the cDNA libraries, none of the cDNA libraries produced amplicons in 16S rRNA-PCR. Sequences affiliated with bacteria often appear in viral metagenomes because bacterial genomes are littered with prophage sequences (Canchaya et al. 2003, Srividhya et al. 2007). Alternatively, the sequences in our libraries that were affiliated with *Bacteria* might come from genes of bacterial origin that were transferred to phages (Del Casale et al. 2011).

Reads were also assessed using Metavir for an in-depth analysis of library taxonomic composition using the GAAS tool (Angly et al. 2009). GAAS normalizes the number of hits according to genome size, generating a more accurate estimate of species abundances within a metagenome. According to Metavir annotation of the DNA libraries, the majority (79.19–93.56 %) of reads annotated as virus-affiliated belonged to ssDNA viruses. Of the ssDNA viruses, *Circoviridae* and *Microviridae* were the most represented viral families (Table 2). The most abundant viral taxon varied across libraries, as well as the percentage of the library comprised by that particular taxon (Table S4). Similar viral taxa were observed in each library, but at varying abundances. Of the viruses with bacterial hosts, phages that infect *Chlamydia*, phages that infect *Bdellovibrio*, and marine gokushovirus were common to all libraries. The most commonly observed viruses with eukaryotic hosts were those infecting fish, birds, and plants (Table S4).

For the cDNA libraries, MG-RAST annotation indicated that 0.57–13.92 % of reads had homology to known virus sequences (Fig. 2c). The Post cDNA library had the lowest proportion of reads matching known viral sequences (0.57 %) as well as the highest percentage of reads failing QC (~80 %). While the low proportion of reads matching known viral sequences may be related to the smaller number of reads passing QC and therefore available for analysis, the Post cDNA library still contained 47 758 reads (for comparison, the other cDNA libraries contained 56 000–150 000 post-QC reads). Based on these observations, it appears that the Post cDNA library simply contained fewer sequences affiliated with known viral genomes. In spite of DNase treatment prior to first strand synthesis, and lack of PCR amplification of DNA from RNA starting material used in construction of cDNA libraries, the majority of reads that could be classified as viruses in the cDNA libraries (Metavir annotation) were affiliated with ssDNA viruses (59.1–86.8 %), with very few reads affiliated with known RNA viruses (0–0.4 %; Table 3). Of these ssDNA viruses, *Circoviridae* and *Microviridae* were the most represented viral families (Table 3). The most abundant viral taxon varied greatly across the cDNA libraries, from nepavirus-affiliated reads representing 30 % of the known viral reads in the Soil library to *Labidocera aestiva* circovirus-affiliated reads representing 4.6 % of the known viral reads in the Early library (Table S5). In general, a greater proportion of viral reads affiliated with viruses known to infect eukaryotic hosts was found in the cDNA libraries as compared to the DNA libraries (Tables S4 & S5).

Functional annotation

Metabolic subsystems were annotated using MG-RAST based on the best BLASTx hit. For the DNA libraries, the subsystem with the greatest global representation of ORFs was for Phages, Prophages, Transposable Elements, and Plasmids, ranging from 50 to 90 % of the annotated reads (Fig. S1). The Early and Late libraries showed the widest range of repre-

Table 2. Classification of reads in DNA libraries with homology to viral sequences. Comparisons of the taxonomic compositions were computed via the GAAS tool (Angly et al. 2009), from a BLAST comparison with NCBI RefSeq complete viral genome proteins using BLASTx (E-value < 10^{-5}). dsDNA: double-stranded DNA, ssDNA: single-stranded DNA, ssRNA: single-stranded RNA

Group	Order	Family	Initial	Early	Late	Post	Soil	Sediment
dsDNA viruses, no RNA stage	–	*Adenoviridae*	0	0	0.001	0	0.001	0
	–	*Ampullaviridae*	0	0	0	0.001	0	0
	–	*Ascoviridae*	0.021	0.062	0.003	0.018	0.003	0
	–	*Baculoviridae*	0.002	0.006	0.002	0.002	0.001	0
	Caudovirales	*Myoviridae*	0.496	2.021	0.118	1.031	0.083	0.007
		Podoviridae	1.017	3.577	0.177	2.227	0.109	0.006
		Siphoviridae	0.837	3.111	0.176	1.671	0.27	0.015
		Unclassified *Caudovirales*	0.137	0.513	0.02	0.292	0.012	0
	–	*Corticoviridae*	0	0.004	0	0	0	0
	Herpesvirales	–	0.003	0.01	0.003	0.004	0.002	0
	–	*Iridoviridae*	0.011	0.026	0.004	0.006		
	Ligamenvirales	–	0	0.001	0.001	0.001	0	0
	–	*Marseilleviridae*	0.003	0.01	0	0.003	0.001	0
	–	*Mimiviridae*	0.022	0.069	0.009	0.019	0.004	0.001
	–	*Nudiviridae*	0	0.001	0.003	0.001	0.001	0
	–	*Papillomaviridae*	0	0	0	0	0.001	0
	–	*Phycodnaviridae*	0.047	0.173	0.014	0.159	0.015	0.003
	–	*Polydnaviridae*	0	0.001	0.096	0.006	0	0
	–	*Polyomaviridae*	0	0	0	0	0.002	0
	–	*Poxviridae*	0.001	0.003	0.002	0.003	0.001	0.001
	–	*Tectiviridae*	0.007	0.037	0.01	0.033	0.005	0
		Unclassified dsDNA phages	0.288	0.862	0.045	0.676	0.03	0.002
		Unclassified dsDNA viruses	0.066	0.125	0.006	0.112	0.007	0.001
ssDNA viruses	–	*Circoviridae*	34.093	27.654	16.667	17.3	29.975	22.926
	–	*Geminiviridae*	1.452	2.419	3.719	1.698	2.02	0.811
	–	*Inoviridae*	0.754	1.373	0.429	0.191	0.053	0.033
	–	*Microviridae*	28.344	23.686	41.543	49.376	41.878	22.011
	–	*Nanoviridae*	0.235	0.113	0.138	0.073	0.14	0.226
	–	*Parvoviridae*	0.016	0.01	0.02	0	0.009	0.004
		Unclassified ssDNA viruses	26.476	27.847	27.852	19.631	19.242	48.62
dsRNA viruses			0.003	0	0	0.008	0	0
ssRNA viruses			1.295	1.331	0.991	0.616	0.123	0.308
Unassigned viruses			0.019	0.021	0.013	0.007	0	0.001
Retro-transcribing viruses			0	0	0.013	0.006	0.002	0.002
Satellites			3.303	1.697	1.714	2.747	2.484	4.801
Unclassified phages			0.061	0.29	0.018	0.131	0.022	0.001
Unclassified virophages			0.018	0.034	0	0.022	0.001	0
Unclassified viruses			0.703	2.606	5.782	1.698	3.461	0.213

sented metabolic subsystems. The results of pairwise comparisons of the potential metabolic profiles of each library using Mann-Whitney *U*-tests are presented in Table S6. Most differences between relative abundance of metabolic subsystems occurred when comparing the Initial to other libraries, particularly the Late library. These significant differences occurred most often because there was no representation of a particular subsystem in the Initial library compared to a higher level of representation in the other libraries. Reads affiliated with the Carbohydrates, Membrane Transport, and Amino Acid Derivatives subsystems were present in the Initial library, but at a significantly lower relative abundance than the other libraries. In addition, the Early library had a significantly greater relative abundance of predicted ORFs in the Dormancy and Sporulation, and Nucleosides and Nucleotides subsystems than the Sediment library (Table S6).

Within the cDNA libraries, the Sediment library had highest abundance of ORFs with predicted functions in Stress Response, Secondary Metabolism, Respiration, Regulation and Cell Signaling, Protein Metabolism, Nitrogen Metabolism, Metabolism of Aromatic Compounds, Membrane Transport, Iron Acquisition and Metabolism, Fatty Acids, Lipids and Isoprenoids, Cell Wall and Capsule Synthesis, Carbohydrates, and Amino Acids and Deriv-

Table 3. Classification of reads in cDNA libraries with homology to viral sequences

Group	Order	Family	Initial	Early	Late	Post	Soil	Sediment
dsDNA viruses, no RNA stage			21.303	19.249	1.155	0.835	1.066	2.267
		Ascoviridae	0.017	0.001	0.012	0.004	0.001	0.006
		Asfarviridae	0	0	0	0	0	0
		Baculoviridae	0.007	0	0	0	0.001	0.001
	Caudovirales		18.563	17.534	1.036	0.712	0.407	2.031
		Myoviridae	3.331	0.775	0.216	0.066	0.178	0.209
		Podoviridae	4.517	3.774	0.276	0.139	0.068	0.55
		Siphoviridae	10.184	12.93	0.521	0.502	0.159	1.233
		Unclassified *Caudovirales*	0.531	0.055	0.023	0.006	0.002	0.039
	Herpesvirales		0.001	0	0	0.005	0.015	0.001
		Iridoviridae	0.007	0.001	0.015	0.003	0.01	0.005
		Marseilleviridae	0.003	0	0.003	0	0	0
		Mimiviridae	0.036	0.022	0.004	0.002	0.013	0.007
		Nimaviridae	0	0	0	0	0	0
		Nudiviridae	0.002	0	0	0	0.003	0
		Phycodnaviridae	0.055	0.013	0.015	0.022	0.007	0.018
		Plasmaviridae	0	0	0.006	0	0	0
		Polydnaviridae	0.079	0	0.003	0.002	0.597	0.01
		Poxviridae	0.007	0.002	0.002	0.001	0.005	0.002
		Tectiviridae	0.008	0	0	0	0	0.007
		Turriviridae	0.014	0	0	0	0	0
		Unclassified dsDNA phages	2.373	1.666	0.055	0.082	0.003	0.167
		Unclassified dsDNA viruses	0.13	0.01	0.003	0.001	0.004	0.013
		Environmental samples	7.615	4.539	5.776	6.11	2.522	10.274
ssDNA viruses			59.177	65.644	73.375	81.724	80.314	86.887
		Circoviridae	24.623	18.781	29.312	19.881	5.946	22.459
		Geminiviridae	5.156	7.312	4.428	11.116	31.536	2.662
		Inoviridae	1.176	0.072	0.009	0.193	0.124	0.154
		Microviridae	4.385	14.207	2.761	15.997	0.31	20.489
		Nanoviridae	0.61	0.134	0.063	0.218	0.005	0.162
		Parvoviridae	0.08	0.016	0.012	0.039	0.006	0.007
		Unclassified ssDNA viruses	23.147	25.122	36.79	34.28	42.386	40.954
ssRNA negative-strand viruses			0	0	0	0	0.003	0
ssRNA positive-strand viruses			0.03	0.116	0	0.04	0	0.401
		Picornavirales	0.02	0	0	0	0	0.002
		Tombusviridae	0	0	0	0	0	0.175
		Unclassified (+) ssRNA viruses	0.011	0.116	0	0.04	0	0.224
Unassigned viruses			0	0	0	0	0	0.013
Retro-transcribing viruses			0.043	0	0	0	0.13	0.008
Satellites			8.452	6.743	13.768	10.094	0.259	6.14
Unclassified phages			0.487	1.105	0.041	0.019	0.022	0.108
Unclassified virophages			0	0	0	0	0	0.002
Unclassified viruses			10.507	7.121	11.661	7.288	18.207	4.096

atives (Fig. S2). Pairwise comparisons of the metabolic profiles of each cDNA library using Mann-Whitney *U*-tests indicated that the representation of ORFs in DNA metabolism was significantly higher in the Initial compared to the Post library; that representation of ORFs in RNA Metabolism was significantly higher in the Late library than in the Post library; and that representation of ORFs in Nucleosides and Nucleotides was significantly higher in the Early library compared to the Initial library (Table S6).

Contig assembly and mapping of reference genomes

In total, 6580 contigs were assembled, representing anywhere from 3.56 to 81.34 % of the reads. The average contig size was 1384 bp, with the largest contigs ranging from 5453 to 38 563 bp long, depending on the library (Table S7). The Initial and Sediment libraries presented a few outliers to the data set, as the percentage of reads assembled in the Initial library was drastically lower (3.46 %) compared to the average (47.17%), while the Sediment library had a much higher percentage of reads assembled into contigs (81.34%). Reads were also mapped against the most abundant reference genome using the analysis tools available through Metavir. For all but the Early and Post libraries, the most abundant reference genomes were circular ssDNA viruses. The most abundant reference genome for Early and Post libraries were the linear dsDNA genomes of the *Phaeocystis globosa* virus. Reads mapped to 15.65–23.59% of the larger dsDNA genomes and 88.32–100% of the smaller ssDNA genomes (Table S8).

For cDNA libraries, 28 444 contigs were assembled, representing 1.03–51.33% of the reads in any given library. The average contig size was 1353 bp, with the largest contigs ranging from 1756–30 025 bp long, depending on the library (Table S7). The percentage of reads assembled in the Sediment library was notably lower (1.03%) compared to the other cDNA libraries (30–50%). Based on mapping of reads, the most abundant reference genome in both the Initial and Early cDNA libraries was the dsDNA *Flavobacterium* phage 11b (reads mapping to 22 and 20% of the reference genome, respectively; Table S8). The most abundant reference genomes in the Late, Post, Soil, and Sediment libraries were ssDNA viruses, with reads mapping to 34–100% of these reference genomes (Table S8).

The original intent with the libraries we have designated 'cDNA libraries' was to capture RNA virus sequences. Because the most abundant reference genomes in these libraries belonged to DNA viruses, however, we conducted a targeted search, mapping reads from the cDNA libraries to the most abundant RNA virus genomes (Table S9). The percentage of reads mapping to these reference genomes was low, ranging from 0% (no reads from the Late library could be matched to an RNA virus reference genome) to a maximum of 16.98%.

Community structure and diversity

The average genome size estimated using GAAS was used in the PHACCS analysis of community structure providing estimates of richness (R), evenness (E), and Shannon-Weiner diversity index (*H'*) (Angly et al. 2005). All of the DNA rank abundance curves were best modeled by the power law. For the cDNA libraries, all of the rank abundance curves were also best modeled by the power law, except for the Early storm sample point, which was best modeled by an exponential curve (Table 4).

The estimated richness of the DNA library time series started out low (Initial R = 4800) and increased by over an order of magnitude at the next time point (Early R = 60 495; Table 4). The estimated richness remained elevated for the duration of the storm and even to a few days afterward (Late R = 27 068; Post R = 27 078). The richness of the Soil library was almost 10-fold higher than that of the Sediment library (Soil R = 30 000; Sediment R = 4800). Evenness remained consistent across the DNA libraries, indicating that the community did not become dominated by one specific viral genotype, even though the estimated total number of viral taxa fluctuated (Table 4). Similar trends in richness estimates were observed in the cDNA libraries (Table 4). A less dramatic, but relatively large increase in richness occurred between the Initial and

Table 4. Diversity statistics for DNA and cDNA libraries

	Initial	Early	Late	Post	Soil	Sediments
DNA library						
Mean (SD) genome size, kbp	8.93 (2.03)	9.23 (1.35)	4.22 (0.56)	6.42 (0.80)	4.04 (0.03)	4.76 (0.26)
Rank-abundance model	Power law	Power law	Power law	Power law	Power law	Power law
Richness	4800	60 495	27 068	27 078	30 000	4800
Evenness	0.87	0.88	0.86	0.84	0.88	0.89
% most abundant	3.92	2.05	2.91	3.51	2.16	3.26
H'	7.42	9.69	8.83	8.64	9.15	7.56
cDNA library						
Mean (SD) genome size, kbp	12.46 (5.4)	5.57 (2.2)	9.25 (5.8)	6.00 (2.33)	23.42 (3.5)	2.87 (0.04)
Rank-abundance model	Power law	Exponential	Power law	Power law	Power law	Power law
Richness	5912	20 000	2619	2000	10 183	4612
Evenness	0.91	0.67	0.88	0.87	0.924	0.889
% most abundant	2.37	0.34	3.99	5.23	1.65	3.38
H'	7.92	6.69	6.99	6.61	8.53	7.51

Early water samples (Initial R = 5912; Early R = 20 000). Instead of remaining elevated, the number of viral genotypes decreased in the Late and Post cDNA libraries to values lower than Initial estimates, but similar to each other (Late R = 2619; Post R = 2000; Table 4). The Soil cDNA viral community had a higher estimated richness than that of the Sediment sample (Soil R = 10183; Sediment R = 4612), mirroring the trend seen in the DNA libraries. Evenness also remained consistent across sample points for the cDNA libraries. The only exception was the Early cDNA library, which had a notably lower estimated evenness (Early E = 0.67). For both the DNA and cDNA libraries, *H′* tracked with richness except for the Early cDNA sample, which is likely due to the lower evenness observed for that sample (Table 4).

Table 5. Dissimilarity between libraries based on cross-assembly of reads. Values represent Wootter's distances, wherein 0 indicates all reads were shared between 2 libraries, and 1.0 indicates that no reads were shared between 2 libraries

	Initial	Early	Late	Post	Sediment	Soil
DNA library						
Initial	0	0.408	0.658	0.721	0.796	0.981
Early		0	0.629	0.640	0.888	0.974
Late			0	0.597	0.896	0.949
Post				0	0.897	0.972
Sediment					0	0.976
Soil						0
cDNA library						
Initial	0	1.0	0.733	0.660	0.718	0.967
Early		0	0.955	1.0	1.0	1.0
Late			0	0.583	0.583	0.964
Post				0	1.0	0.953
Sediment					0	0.914
Soil						0

Comparisons across libraries

Reference-independent cross-contig assembly was performed using crAss (Dutilh et al. 2012) in order to quantify the degree of change in the aquatic library time series, as well as to potentially identify the source(s) of previously undetected viral genotypes in the water column, which we hypothesized had originated from either watershed soils or pond sediments. The crAss program assembles reads from 2 libraries into contigs and estimates the dissimilarity of the input libraries using the Wootter's distance algorithm, with a score of 0 indicating 100% overlap (all reads shared) and a score of 1.0 indicating 0% overlap (no reads shared). The advantage of this method is that it does not rely on database matches to attain an estimate of similarity between libraries and therefore can use all reads within a given data set. According to the crAss results, both DNA and cDNA libraries showed a high degree of dissimilarity among sample time points (Table 5). For the DNA libraries, the Initial and Early water samples were most similar (*d* = 0.4081), but became more dissimilar to other samples over time. The Soil and the Sediment libraries were highly dissimilar from the water samples. The Initial library was most similar to the Sediment library, but the water column libraries grew more dissimilar to the Sediment library over time (Table 5). The Sediment library also had the greatest dissimilarity with Soil. The Soil library was least similar to the Initial library (*d* = 0.9817) and most similar to the Late library (*d* = 0.9499; Table 5). This temporal trend of increasing dissimilarity over time was not observed between the Soil library and any of the water column libraries, as was the case for the Sediment and water column samples.

For the cDNA libraries, it is notable that the Early library was most dissimilar from the other samples, with a 1.0 score for all comparisons except between the Early and Late libraries (Table 5). The greatest similarity was observed between the Late and Post, and Late and Sediment libraries, but there was no overlap between the Post and Sediment libraries. As observed with the DNA libraries, the cDNA Soil library was highly dissimilar from the other cDNA libraries, with no distance score below 0.9. In contrast with DNA library time series, no temporal trend was observed among the cDNA water column libraries. Additionally, while the greatest overlap occurred between the Initial and Early DNA water samples, there was no similarity between those time points in the cDNA libraries (Table 5).

Hierarchical cluster trees were also generated to graphically display differences amongst the samples based on BLASTx comparisons (Fig. 3). The trends observed from the cluster plot for the DNA libraries (BLAST-based comparison) generally agreed with the cross-contig comparison generated via crAss. The Initial and Early samples clustered closely together, and also had the smallest reported dissimilarity. The Late and Post samples also clustered together and had the second-lowest reported dissimilarity. In both the BLAST-based cluster tree analysis and the cross-contig analysis, the Sediment library appeared more similar to the water samples, while the Soil library was much further removed (Fig. 3a, Table 5).

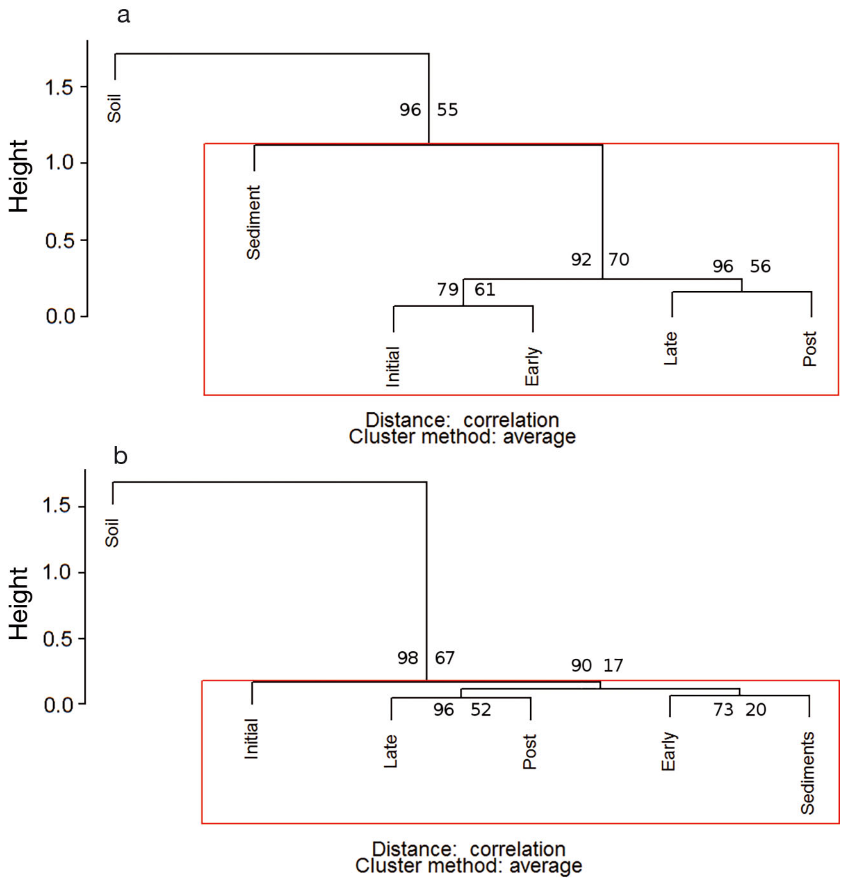

Fig. 3. Hierarchical cluster tree based on tBLASTx comparisons across (a) DNA and (b) cDNA libraries from the present study. Numerical values to the left of each branch point indicate approximately unbiased p-values; values to the right indicate boot-strap probability; red boxes indicate groupings in which the approximate $p < 0.05$

The cDNA libraries generated a similar clustering pattern as the DNA libraries, but with some important exceptions. While the Initial and Early libraries clustered together, and the Late and Post libraries also clustered together, the Sediment and Late libraries appeared to be more similar to each other than the Late and Post libraries were to each other (Fig. 3b). As observed with the DNA libraries, the cDNA Soil library was the least similar to any of the other cDNA libraries. That the Initial and Early libraries clustered together contradicts the results from the cross-assembly analysis (Table 5). This is likely due to the fact that different relationships are identified through BLAST-based similarity as opposed to forming cross-contigs.

Comparison with other viral metagenomes

The 6 DNA libraries generated from this study were compared to other DNA viromes via BLASTx using a hierarchical cluster tree. This comparison included 3 additional wet retention ponds (Saxton et al. 2016), 3 human fecal samples and 2 human gut samples (Kim et al. 2011), 3 desert gueltas (Fancello et al. 2013), 2 French lakes (Roux et al. 2012), an Antarctic lake (López-Bueno et al. 2009), 4 aquaculture ponds and 7 saltern samples (Rodriguez-Brito et al. 2010), 2 samples from a temperate lake (Green et al. 2015), and 4 marine samples (Rodriguez-Brito et al. 2010). Clusters (groupings with approximately unbiased $p \geq 0.05$) formed based on sample type, with clear groupings of human-associated, high-salinity, low-salinity, and freshwater viromes (Fig. 4). A fifth cluster included the DNA libraries from the present study and viromes of Lake Limnopolar (Antarctica) that represent 2 different time points (Fig. 4). Additional cluster plots were generated with the samples from the present study in isolation to determine if each time point resembled other viromes due to changes in community composition. In these comparisons, the Initial, Early, and Post water viromes tended to cluster with other wet retention pond samples while the Late storm, Soil, and Sediment viromes from this study clustered with the Antarctic lake samples (data not shown).

DISCUSSION

Metagenomics has been a powerful approach for investigating environmental viral assemblages (viromes) for well over a decade. While our understanding of aquatic viral diversity and ecology has been driven largely by marine studies (e.g. Breitbart et al. 2002, Angly et al. 2006, Wawrzynczak 2007, Williamson et al. 2012, Hurwitz & Sullivan 2013), freshwater viromes, particularly those of temperate, eutrophic systems, remain poorly characterized and understood (Green et al. 2015). Wet retention ponds represent an important subset of freshwater ecosystems, as they are the most commonly employed BMP to capture and treat stormwater runoff, and their installation is typically paired with anthropogenic im-

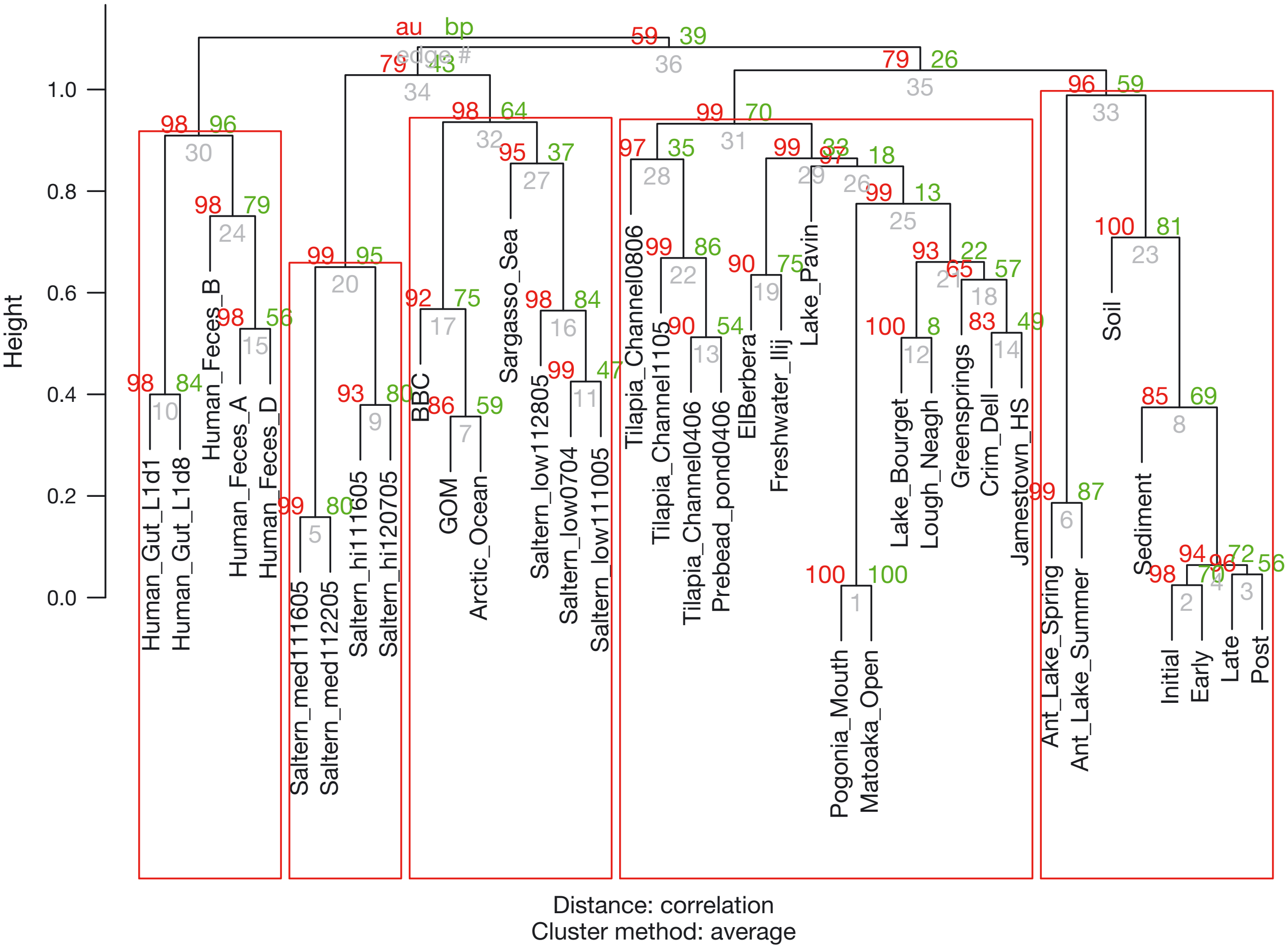

Fig. 4. Hierarchical cluster tree with au/bp values (%) based on tBLASTx comparison of the 6 DNA libraries from the present study (Initial, Early, Late, Post, Soil, Sediment) and other DNA viromes, including: 3 additional wet retention ponds (Greensprings, Crim Dell, Jamestown High School; Saxton et al. 2016), an Antarctic Lake (Antarctic Lake Spring/Summer; López-Bueno et al. 2009), a freshwater lake in Ireland (Lough Neagh; Skvortsov et al. 2016), temperate French lakes (Lake Bourget, Lake Pavin; Roux et al. 2012), a temperate lake in Virginia (Matoaka Open, Pogonia Mouth; Green et al. 2015), desert gueltas (Ilij, El Berbera; Fancello et al. 2013); tilapia aquaculture (Tilapia Channel and Prebead pond; Rodriguez-Brito et al. 2010), solar salterns (Rodriguez-Brito et al. 2010), marine water column (BBC: Bay of British Columbia, GOM: Gulf of Mexico, Arctic Ocean, Sargasso Sea; Angly et al. 2006), and human gastroenteric samples (human gut, human feces; Kim et al. 2011). Tree parameters as in Fig. 3

pacts on watersheds (Collins et al. 2010). Determining the factors that affect the viral community associated with wet retention ponds is important in understanding these structures in the greater context of human and environmental health. To our knowledge, this study is the first to use metagenomic time series data to investigate changes in aquatic viral community composition due to natural disturbance.

Consideration of methods and limitations

The vast majority of reads that could be matched with known sequences in the NCBI RefSeq Virus database were affiliated with ssDNA viruses (Tables 1 & 2). As with many virome studies, we used GenomiPhi (phi-29 polymerase-based multiple displacement amplification) to generate enough viral nucleic acid for library construction and sequencing. GenomiPhi has been shown to preferentially amplify small ssDNA templates (Angly et al. 2006, Roux et al. 2012, López-Bueno et al. 2015), and could have contributed the observed dominance of reads affiliated with known ssDNA viruses in our libraries. Use of 0.22 µm filtration may have removed nucleocytoplasmic large DNA viruses (NCLDV) from our sample sets (Hingamp et al. 2013), but the use of filtration to remove bacteria and other cells is difficult to avoid.

With specific regard to the cDNA libraries, the high proportion of reads failing QCs is likely due to the additional steps involved in generating these libraries (e.g. DNase treatment, generation of cDNA using random primers, GenomiPhi amplification of cDNA products), as large proportions of low-quality reads have been reported in other studies that used similar approaches to generate cDNA or RNA viromes (Miranda et al. 2016). The reliance on random reverse transcription to generate cDNA for sequencing also increases the likelihood of detecting contaminant sequences in addition to viral sequences of interest (Djikeng et al. 2008, Rosseel et al. 2012, Hall et al. 2014) and reads affiliated with ssDNA and dsDNA viruses have also been observed in other RNA viromes that were generated using similar methods (Djikeng et al. 2008, Rosseel et al. 2012). However, even if some contaminating viral DNA remained, we would still expect our libraries to have been enriched in RNA viruses. It is also possible that there was a low proportion of RNA viruses to DNA viruses in the pond, or that the RNA virus community of the pond contained few previously sequenced representatives.

It is clear based on both rarefaction and analysis of contigs that none of the viromes described in this study were exhaustively sampled. Because of this, observed differences between any pair of viromes may arise due to differences in the rank-abundance of taxa rather than strict changes in the presence/absence of taxa: a rare taxon may be present but below the limit of detection and so not observed in one library, while the same taxon may be high enough in abundance to be detected in another library. In comparing libraries across time and space, however, we can safely assume that at equal sequencing depth, 2 viral communities with similar rank-abundances would share more overlap with each other than 2 viral communities with different rank-abundances. Thus, even at imperfect coverage, we may still draw conclusions about storm-induced changes in aquatic virus community composition.

Finally, we acknowledge that the comparisons performed in this study are based on single libraries representing each time point. While replicate libraries for each time point would undoubtedly help to constrain within-sample variability and lead to more robust comparisons across time points, budgetary limitations prevented such an approach in the present work.

Change in taxonomic composition over time

Water samples were collected from the Longhill Grove retention pond in Williamsburg, Virginia, USA, over the course of Tropical Storm Andrea (Initial, Early, Late, and Post). Additionally, soil and sediment samples were taken at the same time as the Initial water sample to test the hypotheses that new detectable viral genotypes could be introduced to the water column during storms due to runoff from watershed soils (Williamson et al. 2014), resuspension of sediments, or both. Similar to other metagenomic studies, the majority of reads (58.8–73.9%; Fig. 1) had no known sequence homology. This result emphasizes the ongoing problem that viral diversity, particularly in freshwaters and soils, is under-explored and highlights the importance of bolstering databases with a wider representation of appropriate reference genomes. Although the majority of affiliated sequences were classified as viral or bacterial, there was also a notably high representation of eukaryote-affiliated sequences in our libraries. Additionally, over half (58–70%) of the most abundant viral taxa (i.e. those with reads representing >1% of the total library) infect eukaryotic hosts. The high representation of these viruses could be due to the fact that wet retention ponds are man-made ecosystems. Wet retention ponds are built to manage peak stormflow caused by impermeable surfaces due to land development; thus, the most abundant host organisms in these developed watersheds would reflect transient or opportunistic populations such as emergent vegetation, insects, migratory birds, and pets. The Late library, which coincided with peak storm flow, returned the highest percentage of reads affiliated with eukaryotes (Fig. 2b). These trends have been observed in other stormwater retention ponds (Saxton et al. 2016), and challenge the assumption that all aquatic viral communities are dominated by phages. It is possible that many of the viruses of eukaryotes observed in our libraries have been washed into the pond and, lacking local hosts, are not part of the active viral community.

Change in representation of metabolic subsystems over time

The results of pairwise comparisons of the functional annotation of DNA libraries indicated that a majority of significant differences in represented subsystems occurred between the Initial library and other samples from later in the storm. Since sequence analysis only demonstrates metabolic potential rather than actual gene expression, the most likely explanation for the rapid changes in the representation of metabolic subsystems across libraries is a rapid increase in the abundance of novel (i.e. previously undetected) virus genes or genotypes during the storm. That the Initial library had low representation of subsystems compared to Sediment and Soil libraries (which were sampled at the same time) suggests that the community in the water column started out functionally distinct from the surrounding watershed and the settled sediments.

The majority of differences between the cDNA libraries occurred due to a significantly higher representation of metabolic subsystems from the Sediment compared to other libraries. In the DNA samples, differences occurred because one library (Initial) regularly lacked representation of a subsystem compared to the other libraries. Comparatively, in the cDNA libraries, this difference in representation occurred because one library (Sediment) contained a higher relative abundance of ORFs of a particular functional group compared to the other libraries. While in the DNA samples, this shift could be explained by the emergence of newly detectable viral genotypes in the water column due to storm runoff, the high abundance of ORFs for specific metabolic subsystems in the RNA Sediment sample reflects this environment pre-disturbance. However, there were no significant differences between the Sediment and Late storm RNA libraries, which may suggest mixing between pond sediments and the water column at the peak of the storm.

Combined, the results of the functional annotation of both DNA and cDNA libraries suggest that each type of sample (sediment, soil, and water) has a distinct metabolic profile. It also suggests that the water column is more responsive to metabolic shifts over time, potentially resulting from the mixing and homogenization of viral communities in the pond.

Change in viral community structure over time

According to PHACCS, watershed soils contained much higher viral richness than pond sediments. This suggests that the population of viruses that are produced within pond sediments (or perhaps that have settled out from the water column over time) represents a less diverse subset of the population in the surrounding watershed. One could also posit that the spike in viral richness from the Initial to the Early water samples is more likely due to the addition of viral species through runoff collecting in the retention pond as opposed to the addition of viral species through sediment resuspension.

Cross-contig analysis indicated that snapshots of water column DNA virus communities grew more dissimilar over time, suggesting a community shift as the disturbance from the rain event increased (Table 5). The Soil and Sediment libraries showed a high degree of dissimilarity from the water samples, suggesting that characteristic viral assemblages are housed within each of these compartments (soil, water, sediments). This idea is supported by another study of aquatic viruses in stormwater retention ponds that showed particle-associated viral communities were completely different from the planktonic viral community (Williamson et al. 2014). If the libraries collected after storm inputs were to show similarity to the soil or sediment samples, this would suggest the potential source(s) of these viral genotypes that were not previously detected in the water column. The results of cross-assembly analysis did not allow us to differentiate between these 2 potential sources. BLASTx comparisons of the DNA libraries revealed that the Sediment library had much higher homology with the water column libraries than did the Soil library (Fig. 3a). This could implicate pond sediments as an important source of novel viral genotypes observed in the pond water column over the course of the storm, but because of the relatively low viral richness of the Sediment library, such mixing alone cannot explain the large jump in viral richness in the Early library (Table 4).

For the cDNA libraries, the results of cross-contig analysis did not follow any clear trends. Rapid change in the viral community due to the influx of water could account for the stark dissimilarity observed between the Early storm time point and almost all other samples. As with the DNA libraries, the Soil cDNA library had higher richness than the Sediment library. It was posited that runoff from virus-rich watershed soils could account for the spike in richness for the early storm, but in the cDNA crAss comparison, there is no observed overlap between the early reads and the soil metagenome (Table 5). BLASTx comparisons of the cDNA libraries indicated a clear grouping of the Sediment library within the other water column samples, clustering with the Early library (Fig. 3b). This grouping suggests that for the viruses contained in the cDNA libraries, pond sediments rather than soil were more likely the source of novel genotypes that were detected in the pond water column during the early phase of the storm. However, as with the DNA virus community, the increase in viral richness in the Early library is much higher than can be accounted for by the addition of sediment viruses alone.

Overall, our analyses of the DNA and cDNA virome time series indicate that the community composition and metabolic capabilities of freshwater viromes can change drastically in a matter of hours due to storm disturbances. The spike in richness at the beginning of the storm indicates that novel viruses are detected, but our sampling methods have not conclusively identified the source of the novel viral genotypes. One potential source could be the induction of prophages due to the environmental disturbance caused by the storm. However, RNA viruses are not known to engage in lysogeny, and even DNA prophage induction would not likely account for the large increase in richness observed in our Early libraries. The very low prevalence of reads affiliated with prophage integrases (0–12 affiliated sequences in any given library) does not support the idea that observed increases in the number of viral genotypes detected in the water column were due to prophage induction. More probable is that we have missed an environmental reservoir of viruses, such as storm drains that collect runoff from more distant locations and channel it into the pond through underground pipes. The Longhill Grove pond does have a single main drainpipe that collects water from storm grates in the parking lot and funnels runoff into the pond. Atmospheric viruses deposited by rainfall could also be a source of novel viral genotypes (Whon et al. 2012). Further studies targeting each of these reservoirs (especially estimates of viral richness and community overlap) could be helpful for narrowing down the potential sources of viral diversity within a pond environment.

Comparison with other viral communities

In this project, hits to known sequences were dominated by ssDNA viruses. Viral metagenomes prepared from 4 other local stormwater retention ponds (Saxton et al. 2016) and a freshwater lake (Green et al. 2015) showed a variable pattern in dominance of dsDNA vs. small ssDNA viruses — despite having been prepared using the same methods. The lake and 2 of the 4 ponds were dominated by dsDNA phages, while the remaining 2 ponds were dominated by ssDNA viruses. The variable dominance of viral genome types across other freshwater viromes is striking: the French Lake Pavin had ~20% ssDNA viruses, whereas nearby Lake Bourget had ~33% ssDNA viruses (Roux et al. 2012); in Antarctic Lake Limnopolar, a spring sample had >75% ssDNA viruses, while the same lake in the summer consisted of >80% dsDNA viruses; in 4 Saharan gueltas, the vast majority of viruses (>90%) were dsDNA phages (Fancello et al. 2013). In each of these studies, viral DNA was amplified using GenomiPhi, which has been shown to preferentially amplify ssDNA, as mentioned above. In order to separate methodological artifact from natural variation, we would avoid dependence on this amplification step to generate sufficient starting material for library construction in any future projects.

A series of hierarchal cluster trees were generated to compare the viromes of the present study to other viromes of interest, using methods detailed in the 'Results' section. When compared to other aquatic and human-impacted viromes, The Longhill Grove viromes clustered with the Antarctic Lake (Limnopolar) libraries, although bootstrap values were relatively weak (Fig. 4). This result was intriguing because these 2 water bodies are more than 11 000 km apart, in entirely different biomes; however, both of these data sets include time series data. The remaining viromes in the cluster tree grouped according to sample type (freshwater, low saline, high saline, human-associated), and each terminal node represents a single time point. However, the last grouping consisted only of viromes that included repeated sampling over time. While we do not believe this alone explains the clustering pattern, it is a noteworthy coincidence. While most of the viromes clustered according to sample type, surprisingly, our libraries did not cluster with other human-impacted freshwater structures (i.e. other wet retention ponds, aquaculture ponds, and lakes). However, when each water column library (time point) was compared individually with other viromes, the Initial, Early, and Post libraries clustered with other wet retention ponds (data not shown), while when analyzed together, the Late, Sediment, and Soil libraries clustered weakly with the Antarctic Lake (Limnopolar) samples. This implies that temporal changes may account for just as much, if not more, difference observed between viromes than environment type.

The cDNA libraries of the present study were compared to other extant RNA viromes, because these libraries were originally designed to target RNA viruses. However, due to the lack of RNA viromes available for comparison, relatively few insights could be gleaned. A cluster tree generated using BLAST-based comparisons indicated that libraries clustered according to study (Fig. S3). Further metagenomic studies targeting RNA viruses can only improve our collective understanding of freshwater virus communities. Importantly, methods to generate RNA virus libraries have improved since the work in the present study was undertaken (Manso et al. 2017).

CONCLUSIONS

Overall, our results show that the viral community of the Longhill Grove wet retention pond is similar to other wet retention ponds, but the community present after peak stormwater input may be drastically different than the initial or pre-storm community. The observed similarity in viral community composition across wet retention ponds could be due to the impacted nature of the pond watersheds, as we observed an uncharacteristically large number of reads affiliated with eukaryotic homologues, as well as a large number of reads affiliated with known viral taxa that infect eukaryotic hosts. Observed changes in taxonomic composition, metabolic profiles, and estimated viral richness over time support findings from previous studies showing that storm events drive viral community change (Williamson et al. 2014). However, for this particular storm event, the observed changes in the Longhill Grove pond viral community could not be explained by the introduction of viral species from watershed soils, nor resuspension of viruses in pond sediments over the course of the storm. Rather, it seems more likely that novel viral genotypes detected in the water column during and after storm perturbations were derived from multiple origins, including watershed soil, pond sediments, and runoff water transported via storm drains from more distant sources. The differences in viral richness that occurred during and following storm perturbation would have been overlooked had we not sampled over the course of the storm event, which clearly demonstrates the benefit and need for time-series data in concert with metagenomics approaches.

Acknowledgements. This work was supported by a grant from the Virginia Environmental Endowment to KEW. This work was performed in part using computing facilities at the College of William and Mary which were provided by contributions from the National Science Foundation, the Commonwealth of Virginia Equipment Trust Fund, and the Office of Naval Research. In particular, we thank Eric Walter in the High Performance Computing cluster at William & Mary; Dr. Randolph Chambers and Adrianna Gorsky in the Keck Environmental Field Laboratory at William & Mary for their assistance with water chemistry and nutrient analysis; and Joe Jones, Director of Genomic Services at Selah Genomics, for assistance with library construction, sequencing, and troubleshooting.

LITERATURE CITED

Angly F, Rodriguez-Brito B, Bangor D, McNairnie P and others (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 6: 41

Angly FE, Felts B, Breitbart M, Salamon P and others (2006) The marine viromes of four oceanic regions. PLOS Biol 4: e368

Angly FE, Willner D, Prieto-Davó A, Edwards RA and others (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLOS Comput Biol 5:e1000593

Bolduc B, Wirth JF, Mazurie A, Young MJ (2015) Viral assemblage composition in Yellowstone acidic hot springs assessed by network analysis. ISME J 9:2162–2177

Breitbart M, Salamon P, Andresen B, Mahaffy JM and others (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci USA 99:14250–14255

Canchaya C, Proux C, Fournous G, Bruttin A, Brüssow H (2003) Prophage genomics. Microbiol Mol Biol Rev 67: 238–276

Collins KA, Lawrence TJ, Stander EK, Jontos RJ and others (2010) Opportunities and challenges for managing nitrogen in urban stormwater: a review and synthesis. Ecol Eng 36:1507–1519

Del Casale A, Flanagan PV, Larkin MJ, Allen CCR, Kulakov LA (2011) Extent and variation of phage-borne bacterial 16S rRNA gene sequences in wastewater environments. Appl Environ Microbiol 77:5529–5532

Djikeng A, Halpin R, Kuzmickas R, DePasse J and others (2008) Viral genome sequencing by random priming methods. BMC Genomics 9:5

Dutilh BE, Schmieder R, Nulton J, Felts B, Salamon P, Edwards RA, Mokili JL (2012) Reference-independent comparative metagenomics using cross-assembly: crAss. Bioinformatics 28:3225–3231

Fancello L, Trape S, Robert C, Boyer M, Popgeorgiev N, Raoult D, Desnues C (2013) Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. ISME J 7: 359–369

Green JC, Rahman F, Saxton MA, Williamson KE (2015) Metagenomic assessment of viral diversity in Lake Matoaka, a temperate, eutrophic freshwater lake in southeastern Virginia, USA. Aquat Microb Ecol 75:117–128

Hall RJ, Wang J, Todd AK, Bissielo AB and others (2014) Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. J Virol Methods 195:194–204

Hancock GS, Holley JW, Chambers RM (2010) A field-based evaluation of wet retention ponds: How effective are ponds at water quantity control? J Am Water Resour Assoc 46:1145–1158

Hardbower DM, Dolman JL, Glasner DR, Kendra JA, Williamson KE (2012) Optimization of viral profiling approaches reveals strong links between viral and bacterial communities in a eutrophic freshwater lake. Aquat Microb Ecol 67:59–76

Hewson I, Barbosa JG, Brown JM, Donelan RP, Eaglesham JB, Eggleston EM, LaBarre BA (2012) Temporal dynamics and decay of putatively allochthonous and autochthonous viral genotypes in contrasting freshwater lakes. Appl Environ Microbiol 78:6583–6591

Hingamp P, Grimsley N, Acinas SG, Clerissi C and others (2013) Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. ISME J 7: 1678–1695

Hurwitz BL, Sullivan MB (2013) The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. PLOS ONE 8:e57355

Keegan KP, Trimble WL, Wilkening J, Wilke A, Harrison T, D'Souza M, Meyer F (2012) A platform-independent method for detecting errors in metagenomic sequencing data: DRISEE. PLOS Comput Biol 8:e1002541

Kim MS, Park EJ, Roh SW, Bae JW (2011) Diversity and abundance of single-stranded DNA viruses in human feces. Appl Environ Microbiol 77:8062–8070

Leopold LB (1968) Hydrology for urban land planning. A guidebook on the hydrologic effects of urban land use. US Geological Survey, Reston, VA

López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcamí A (2009) High diversity of the viral community from an Antarctic lake. Science 326:858–861

López-Bueno A, Rastrojo A, Peiró R, Arenas M, Alcamí A (2015) Ecological connectivity shapes quasispecies structure of RNA viruses in an Antarctic lake. Mol Ecol 24: 4812–4825

Manso CF, Bibby DF, Mbisa JL (2017) Efficient and unbiased metagenomic recovery of RNA virus genomes from human plasma samples. Sci Rep 7:4173

Meyer F, Paarmann D, D'Souza M, Olson R and others (2008) The metagenomics RAST server — a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9:386

Miranda JA, Culley AI, Schvarcz CR, Steward GF (2016) RNA viruses as major contributors to Antarctic virioplankton: RNA viruses in the Antarctic. Environ Microbiol 18:3714–3727

Noble RT, Fuhrman JA (1998) Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. Aquat Microb Ecol 14:113–118

Parsons TR, Maita Y, Lalli CM (1984) A manual of chemical and biological methods for seawater analysis. Pergamon Press, Oxford and New York, NY

Rodriguez-Brito B, Li L, Wegley L, Furlan M and others (2010) Viral and microbial community dynamics in four aquatic environments. ISME J 4:739–751

Rosseel T, Scheuch M, Höper D, De Regge N, Caij AB, Vandenbussche F, Van Borm S (2012) DNase SISPA-next generation sequencing confirms Schmallenberg virus in Belgian field samples and identifies genetic variation in Europe. PLOS ONE 7:e41967

Roux S, Enault F, Robin A, Ravet V and others (2012) Assessing the diversity and specificity of two freshwater viral communities through metagenomics. PLOS ONE 7: e33641

Roux S, Tournayre J, Mahul A, Debroas D, Enault F (2014) Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. BMC Bioinformatics 15: 76

Saxton MA, Naqvi NS, Rahman F, Thompson CP, Chambers RM, Kaste JM, Williamson KE (2016) Site-specific environmental factors control bacterial and viral diversity in stormwater retention ponds. Aquat Microb Ecol 77:23–36

Skvortsov T, de Leeuwe C, Quinn JP, McGrath JW and others (2016) Metagenomic characterisation of the viral community of Lough Neagh, the largest freshwater lake in Ireland. PLOS ONE 11:e0150361

Srividhya KV, Alaguraj V, Poornima G, Kumar D and others (2007) Identification of prophages in bacterial genomes by dinucleotide relative abundance difference. PLOS ONE 2:e1193

Suttle CA (2007) Marine viruses — major players in the global ecosystem. Nat Rev Microbiol 5:801–812

Thingstad TF (2000) Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. Limnol Oceanogr 45:1320–1328

Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. Nat Protoc 4:470–483

US EPA (United States Environmental Protection Agency) (1999) Storm water technology fact sheet: wet detention ponds. EPA 832-F-99-048. US EPA, Office of Water, Washington, DC

US EPA (2006) Performance of stormwater retention ponds and constructed wetlands in reducing microbial concentrations. EPA/600/R-06/102. US EPA, Office of Research and Development, Washington, DC

Wawrzynczak E (2007) A global marine viral metagenome. Nat Rev Microbiol 5:6–7

Whon TW, Kim MS, Roh SW, Shin NR, Lee HW, Bae JW (2012) Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. J Virol 86:8221–8231

Williamson KE, Corzo KA, Drissi CL, Buckingham JM, Thompson CP, Helton RR (2013) Estimates of viral abundance in soils are strongly influenced by extraction and enumeration methods. Biol Fertil Soils 49:857–869

Williamson KE, Harris JV, Green JC, Rahman F, Chambers RM (2014) Stormwater runoff drives viral community composition changes in inland freshwaters. Front Microbiol 5:105

Williamson SJ, Allen LZ, Lorenzi HA, Fadrosh DW and others (2012) Metagenomic Exploration of viruses throughout the Indian Ocean. PLOS ONE 7:e42047

Editorial responsibility: Mya Breitbart,
St. Petersburg, Florida, USA

Submitted: March 29, 2017; Accepted: September 26, 2017
Proofs received from author(s): December 31, 2017