Identification of Viral Pathogen Diversity in Sewage Sludge by Metagenome Analysis

SUPPORTING INFORMATION

4 tables
1 figures
12 pages

KYLE BIBBY[1] and JORDAN PECCIA[1*]

[1] Department of Chemical and Environmental Engineering, Yale University, Mason Laboratory, 9 Hillhouse Avenue, P.O. Box 208286, New Haven, CT 06520

*Corresponding author tel: (203) 432-4385; fax: (203)432-4387; email: Jordan.Peccia@yale.edu

Additional Materials and Methods

Viral elution and coliphage culturing. Viruses were eluted from sewage sludge samples and concentrated following a procedure adapted from Monpoeho and co-workers[1]. For each composite sample, 250 ml of liquid sludge was mixed with 250 ml of 0.25M glycine (Sigma Aldrich, Missouri, USA), pH=9, and stirred at 200 rpm for 2 hours at 4°C. This mixture was centrifuged at 5000xg for 1 hour and the supernatant was collected and sequentially filtered through 5 μm (Pall, USA) and 0.45 μm sterile membrane (Millipore, Massachusetts, USA) to remove large particles, including eukaryotes and bacteria. Estimation of viral extraction efficiency was performed by spiking sterilized sewage sludge with a known amount of wild-type F+ coliphages and then quantifying the fraction of spiked coliphages that could be eluted. To quantify coliphage concentrations, elutions were serially diluted in sterile PBS, 1 mL sample was gently mixed with 1 mL host bacteria in tryptic soy broth (TSB) (BD Diagnostics, Maryland, USA) and 4 mL of 0.7% bacteriological agar and plated on tryptic soy agar (TSA) (BD Diagnostics, Maryland, USA). F+ coliphages were cultured using log-phase *E. coli* F_{amp} (ATCC 700891), applying an agar overlay amended with 0.15 mg/mL streptomycin sulfate salt (Sigma Aldrich, Missouri, USA) and anhydrous ampicillin (Sigma Aldrich, Missouri, USA). Somatic coliphages were cultured using log-phase *E. coli* CN-13 (ATCC 700609) and applying an agar overlay amended with 1 mg/mL nalidixic acid (Sigma Aldrich).

Nucleic Acid Extraction. RNA and DNA were recovered from the viral concentrate using a Qiagen Viral RNA extraction kit (Qiagen, California, USA) following manufacturer's instructions. To obtain a sufficient quantity of DNA and reverse transcribed RNA (cDNA) for sequencing, it was necessary to amplify the viral nucleic acids using a random transcription/amplification protocol as previously described[2, 3]. This nucleic acid kit and amplification method have previously been recognized to extract and amplify both genomic RNA and DNA[4, 5]. During reverse transcription, nucleic acids were incubated at 65°C for 5 minutes with 100 pmol primer A (5'-GTTTCCCAGTCACGATCNNNNNNNNN-3') before slowly cooling to room temperature to encourage primer annealing and inactivate any native RNAses. The degenerate N bases of the primer form a random priming site and anneal to the viral RNA, while the remaining primer bases create an artificial primer site for PCR amplification. Reverse transcription was then performed using an AffinityScript Multiple Temperature cDNA synthesis kit (Agilent, California, USA) following manufacturer instructions. Second strand synthesis was performed using T7 Sequenase (GE Healthcare, New Jersey, USA) following the manufacturer's instructions. This product (2 µl) was used in a 100 µl PCR reaction using PCR Master Mix (Roche, Indiana, USA) and 100 pmol primer B (5'-GTTTCCCAGTCACGATC-3') and 40 cycles of 30 seconds at 94°C, 30 seconds at 40°C, 30 seconds at 50°C, 60 seconds at 72°C. At least three PCR reactions were performed for each sample. The resulting PCR products were combined and purified using a PCR purification kit (MoBio, California, USA), eluted in sterile molecular grade water, and immediately delivered on ice for sequencing.

Detailed bioinformatic analyses. The overall bioinformatic strategy included the following: (i) trim and clean sequencing reads, (ii) generate a master assembly of all sequence data, (ii) BLAST assembled contiguous sequences (contigs) for annotations, and (iv) map sample specific reads onto the master assembly to determine sequence coverage (relative abundance). In total, 12 samples were sequenced; ten representing the influent and effluent of the five digesters (AI, AE, BI, BE, CI, CE, DI, DE, EI, EE) and an additional set of true biological replicates of the digester B samples (BI2 and BE2). Sequencing was also performed twice (technical replicates) for each sample preparation. Biological replicates are defined as different nucleic acid extracts prepared from the same sample and viral elution. Technical replicates are defined as replicate sequencing runs from the same nucleic acid extracts.

Prior to assembly and annotation of metagenomic data, raw reads derived from sequencing were trimmed to remove the "Primer B" amplification adaptor using Tagcleaner version 0.12[6]. High stringency quality trimming for raw sequencing reads has been shown to significantly improve assembly and annotation statistics for metagenomic projects using Illumina technology[7]. Reads were then trimmed from the 3' end for a minimum phred quality score of 15 using DynamicTrim from the solexaQA toolkit version 1.13[8]. Following both trimmings, reads shorter than 50 nt were excluded from further analysis using lengthsort, also a part of the solexaQA toolkit. Reads were assembled into contigs for each sample individually with Velvet version 1.1.06[9] using k-mer lengths of 27 and 47. Unused reads from these assemblies were collected and assembly was attempted again at a k-mer length of 47. All assemblies were then merged using the whole read assembler minimus2 in the amos software package version 3.1.0[10], with a minimum overlap of 20 nt and identity of 98%, previously shown to be appropriate for viral metagenome assembly[11]. The results of this assembly are shown in **Figure S1**.

Contigs were initially annotated using MG-RAST[12], a web-based annotation pipeline, to determine general assembly characteristics. Contigs were then subjected to a tBLASTx (translated nucleotide to translated nucleotide) search against an amended NCBI viral genome database updated to include whole annotated genomes of viral pathogens that were not yet in the database to identify pathogen-related sequences. This annotation approach has previously been shown to minimize annotation errors while maximizing pathogen identifications[13]. Contigs annotated as potential human pathogens were then subjected to tBLASTx searches against the NCBI nt (non-redundant nucleotide) and BLASTx searches against the NCBI nr (non-redundant protein) databases in order to investigate and verify initial annotations. The top hit for all BLAST searches was extracted, with a maximum E-value of 0.001, which has previously been shown to minimize false negatives in metagenome annotation for human viral pathogen identification[13]. In cases of multiple top hits with the same E-value, all top hits were extracted. To be included as a potential human pathogen identification, the requirements were that the reference pathogen have a record in the NCBI viral genome database and there exist published information on that virus infecting humans. Bacteriophages of bacterial human pathogens were not included, nor were animal pathogens with rare or undocumented cases of zoonotic transmission involving humans. Additionally, to assess sample similarity, the Sorensen similarity index for each sample was calculated using pathogen occurrence (presence/absence). Using this similarity index, principal component analysis was then completed using the QIIME toolbox, version[14].

To estimate sequence coverage, reads were mapped against the resulting assembly using BWA (Burrow-Weavers Aligner) version 0.5[15] and samtools version 0.1.18[15]. For inter-sample comparisons, these values were normalized by the total number of reads for the sample and the logarithm of the resulting value was taken. To assess biological reproducibility of metagenome sequencing and annotation for the BI and BE biological replicates (parallel nucleic acid extractions), contigs that were annotated by a tBLASTx comparison to the amended viral genome database were compared using the $\log_{10}$ of the number of reads mapped to contig/ number of reads produced for that sample (relative abundance). To assess technical reproducibility (same nucleic acid extraction, replicate sequencing), the relative abundances of contigs annotated by tBLASTx search against the NCBI viral database for replicates from all sequencing runs were compared. A linear line of best fit was then calculated and fitting parameters determined to assess reproducibility

PCR assay. PCR of selected human viruses was used as an independent validation to metagenome annotation results. Viruses targeted by specific PCR included human strains of *Adenovirus*, *Enterovirus*, and *Parechovirus*, and the *Norovirus* GII strain. All PCR assays were performed on the nucleic acid extracts from the viral elution. *Enterovirus* and *Norovirus* GII qPCR assays were performed as previously described in a sewage sludge study by Wong et. al[16]. The qPCR assay for *Parechovirus* was adapted from Benschop et. al[17]. Prior to amplification, viral RNA was reverse transcribed using AffinityScript Multiple Temperature cDNA synthesis kit (Agilent, California, USA) with random hexamers, following manufacturer instructions. Negative template controls were run for each reverse transcription. Primers and synthetic genes used in standardization for each reaction were synthesized at the Oligo Synthesis Lab at the Keck Center, Yale University and Taqman® Probes for each reaction were synthesized by Biosearch Technologies (California, USA). Primer and probe sequence information is summarized in **Table S1**. All primer and probe reaction concentrations were 10 nmol. All reactions underwent an initial denaturation period of 15 minutes at 95°C, followed by 45 cycles of 15 seconds denaturation (95°C) and 60 seconds annealing (60°C *Enterovirus*, 56°C *Norovirus* GII, 55°C *Parechovirus*). Negative template reverse transcriptase controls were confirmed negative by qPCR and calibration curves using synthesized gene target regions were run with all samples. Dilution of target cDNA showed no indication of PCR inhibition. Adenovirus PCR was performed as described previously[18], with 40 cycles of 95°C for 60 seconds, 55°C for 60 seconds, and 72°C for 90 seconds.

Statistics. Linear best fit equations for reads mapped to contigs were calculated using Microsoft Excel version 12.3.3 (Microsoft, Washington, USA). To assess statistical significance, unpaired t-tests were conducted using standard methods. Due to the method of assembly, the generation of a single master assembly using multiple assembly software, not all samples contributed to the generation of all contigs. To estimate each sample's contribution to the master assembly, the raw reads from that sample were mapped to the master assembly using the software BWA[15].

SI TABLES AND FIGURES

Table S1. Primer and Probe Sequences used in PCR Assays.

Component	Sequence (5'-3')
Adenovirus Forward	GACATGACTTTCGAGGTCGATCCCATGGA
Adenovirus Reverse	CCGGCTCAGAAGGGTGTGCGCAGGTA
Enterovirus Forward Primer	ACATGGTGTGAAGAGTCTATTGAGCT
Enterovirus Reverse Primer	CCAAAGTAGTCGGTTCCGC
Enterovirus Probe	FAM- TCCGGCCCCTGAATGCGGCTAAT-BHQ
Norovirus GII Forward Primer	CARGARBCNATGTTYAGRTGGATGAG
Norovirus GII Reverse Primer	TCGACGCCATCTTCATTCACA
Norovirus GII Probe	FAM-TGGGAGGGCGATCGCAATCT-BHQ
Parechovirus Forward Primer	CTGGGGCCAAAAGCCA
Parechovirus Reverse Primer	GGTACCTTCTGGGCATCCTTC
Parechovirus Probe	FAM-AAACACTAGTTGTAWGGCCC-BHQ

Table S2. Summary of Sample Contribution to Master Assembly

Sample	Input into Assembly	Contigs with Reads Mapped	Total Contigs from Assembly with Reads Mapped
	# sequences	# sequences	# sequences
AI	17,966,866	64,611	15.66%
AE	5,821,605	86,271	20.91%
BI	23,895,227	150,277	36.42%
BE	31,936,990	150,970	36.59%
BI2	28,822,410	155,914	37.78%
BE2	24,485,618	132,689	32.15%
CI	19,280,675	50,006	12.12%
CE	32,089,018	108,497	26.29%
DI	59,744,231	116,246	28.17%
DE	30,286,611	131,412	31.85%
EI	27,270,006	83,344	20.20%
EE	27,671,998	91,457	22.16%
TOTAL	**329,271,255**	**1,321,694**	

Table S3. Identified non-Herpesvirus Human DNA Viruses and Contig Lengths

Identification	Accession Number	Contig Length (bp)
Papillomavirus	NC_004104.1	799
Human Papillomavirus type 90 (candHPV90)		
Human Papillomavirus type 16	NC_001526.2	370
Human Papillomavirus type 53	NC_001593.1	515
Human Papillomavirus type 16	NC_001526.2	469
Human Papillomavirus type 10	NC_001576.1	551
Human Papillomavirus type 92	NC_004500.1	409
Human Papillomavirus type 53	NC_001593.1	274
Human Papillomavirus type 34	NC_001587.1	210
Human Papillomavirus type 53	NC_001593.1	233
Human Papillomavirus type 7	NC_001595.1	392
Human Papillomavirus type 63	NC_001458.1	505
Human Papillomavirus type 49	NC_001591.1	270
Human Papillomavirus type 18	NC_001357.1	418
Human Papillomavirus type 53	NC_001593.1	399
Human Papillomavirus type 129	NC_014953.1	384
Adenoviruses		
Human Adenovirus F	NC_001454.1	724
Human Adenovirus F	NC_001454.1	447
Human Adenovirus type 5 (C)	AC_000008.1	496
Human Adenovirus type 1 (C)	AC_000017.1	203
Human Adenovirus C	NC_001405.1	501
Human Adenovirus D	AC_000006.1	289
Human Adenovirus F	NC_001454.1	206
Human Adenovirus type 7 (B)	AC_000018.1	334
Parvovirus		
Human Parvovirus B19	NC_000883.2	444
Human Bocavirus 2	NC_012042.1	1659
Toque Teno Virus		
Torque Teno Virus type 6	NC_014094.1	261
Torque Teno Virus type 19	NC_014078.1	248
Torque Teno Virus type 15	NC_014096.1	242
Torque Teno Virus type 16	NC_014091.1	533
Torque Teno Virus type 16	NC_014091.1	630
Torque Teno midi Virus type 2	NC_014093.1	283
Torque Teno Virus type 7	NC_014080.1	229

Table S4. Identified Human RNA Viruses and Contig Lengths

Identification	Accession Number	Contig Length (bp)
Parechovirus		
Human Parechovirus type 1 (Echovirus 22)	EF051629.2	680
Human Parechovirus type 1 (Echovirus 22)	EF051629.2	639
Human Parechovirus type 2 (Echovirus 23)	AF055846.1	2054
Human Parechovirus type 1 (Echovirus 22)	EF051629.2	301
Human Parechovirus type 1 (Echovirus 22)	EF051629.2	245
Coronavirus		
Human Coronavirus HKU1	NC_006577.2	252
Human Coronavirus HKU1	NC_006577.2	279
Human Coronavirus 229E	NC_002645.1	672
Human Coronavirus HKU1	NC_006577.2	538
Other RNA human viruses		
Human Klassevirus type 1	NC_012986.1	1882
Human Klassevirus type 1	NC_012986.1	855
Human Astrovirus MLB2	NC_016155.1	973
Human Astrovirus MLB1	NC_011400.1	1270
Aichi virus	NC_001918.1	255
Aichi virus	NC_001918.1	218
Human immunodeficiency virus 1	NC_001802.1	234
Human immunodeficiency virus 1	NC_001802.1	221
Human Rotavirus A	NC_011506.2	493
Human Cosavirus B	NC_012801.1	1189
Human Coxsackievirus A16		357
Sapovirus Mc10	NC_010624.1	740
Human Rhinovirus B14	NC_001490.1	363
Human Hepatitis C genotype 2	NC_009823.1	427
Rubella virus	NC_001545.2	374
Human T-Lymphotrophic virus 1	NC_001436.1	200

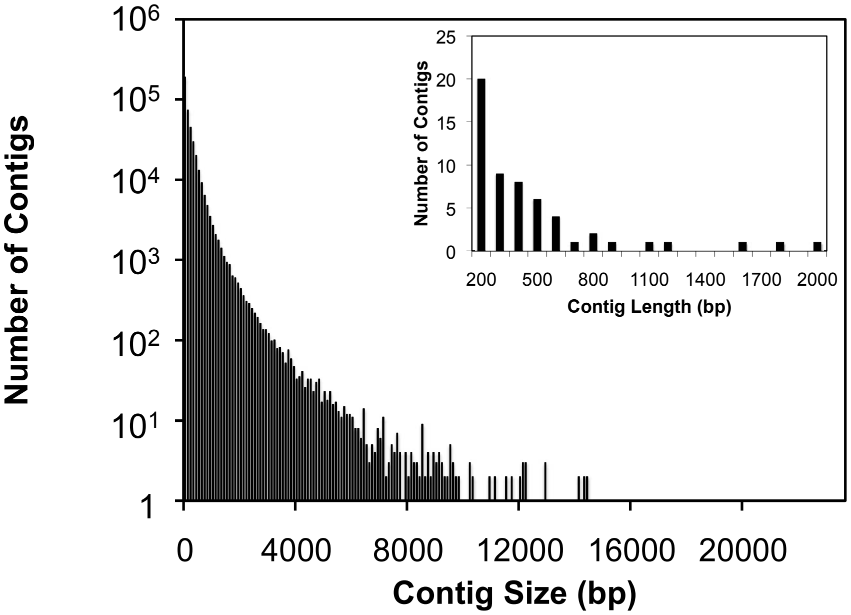

Figure S1. Histogram of assembled contig sizes. Contigs are grouped into 100 nt bins. The largest contig is 23,860 bp. *Inset.* Histogram of non-*Herpes* contigs annotated as viral pathogens.

REFERENCES

1. Monpoeho, S.; Maul, A.; Mignotte-Cadiergues, B.; Schwartzbrod, L.; Billaudel, S.; Ferre, V., Best viral elution method available for quantification of enteroviruses in sludge by both cell culture and reverse transcription-PCR. *Applied and Environmental Microbiology* **2001,** *67*, 2484-2488.
2. Wang, D.; Coscoy, L.; Zylberberg, M.; Avila, P. C.; Boushey, H. A.; Ganem, D.; DeRisi, J. L., Microarray-based detection and genotyping of viral pathogens. *Proceedings of the National Academy of Sciences* **2002,** *99*, 15687-15692.
3. Cantalupo, P. G.; Calgua, B.; Zhao, G.; Hundesa, A.; Wier, A. D.; Katz, J. P.; Grabe, M.; Hendrix, R. W.; Girones, R.; Wang, D.; Pipas, J. M., Raw sewage harbors diverse viral populations. *mBio* **2011,** *2*, (5).
4. Rosario, K.; Nilsson, C.; Lim, Y. W.; Ruan, Y.; Breitbart, M., Metagenomic analysis of viruses in reclaimed water. *Environmental Microbiology* **2009,** *11*, 2806-2820.
5. Wylie, K. M.; Mihindukulasuriya, K. A.; Sodergren, E.; Weinstock, G. M.; Storch, G. A., Sequence analysis of the human virome in febrile and afebrile children. *PLoS ONE* **2012,** *7*, e27735.

6. Schmieder, R.; Lim, Y.; Rohwer, F.; Edwards, R., TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics* **2010,** *11*, 341.

7. Mende, D. R.; Waller, A. S.; Sunagawa, S.; Jarvelin, A. I.; Chan, M. M.; Arumugam, M.; Raes, J.; Bork, P., Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS ONE* **2012,** *7*, e31386.

8. Cox, M.; Peterson, D.; Biggs, P., SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* **2010,** *11*, 485.

9. Zerbino, D. R.; Birney, E., Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research* **2008,** *18*, 821-829.

10. Sommer, D.; Delcher, A.; Salzberg, S.; Pop, M., Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* **2007,** *8*, 64.

11. Hoffmann, K. H.; Rodriguez-Brito, B.; Breitbart, M.; Bangor, D.; Angly, F.; Felts, B.; Nulton, J.; Rohwer, F.; Salamon, P., Power law rank–abundance models for marine phage communities. *FEMS Microbiology Letters* **2007,** *273*, 224-228.

12. Meyer, F.; Paarmann, D.; D'Souza, M.; Olson, R.; Glass, E.; Kubal, M.; Paczian, T.; Rodriguez, A.; Stevens, R.; Wilke, A.; Wilkening, J.; Edwards, R., The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **2008,** *9*, 386.

13. Bibby, K.; Viau, E.; Peccia, J., Viral metagenome analysis to guide human pathogen monitoring in environmental samples. *Letters in Applied Microbiology* **2011,** *52*, 386-392.

14. Caporaso, G.; Kuczynski, J.; Stombaugh, J.; Bittinger, K.; Bushman, F.; Costello, E.; Fierer, N.; Pena, A.; Goodrich, J.; Gordon, J.; Huttley, G.; Kelley, S.; Knights, D.; Koenig, J.; Ley, R.; Lozupone, C.; McDonald, D.; Muegge, B.; Pirrung, M.; Reeder, J.; Sevinsky, J.; Turnbaugh, P.; Walters, W.; Widmann, J.; Yatsunenko, T.; Zaneveld, J.; Knight, R., QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, **2010**, *7*, 335-336.

15. Blinkova, O.; Rosario, K.; Li, L.; Kapoor, A.; Slikas, B.; Bernardin, F.; Breitbart, M.; Delwart, E., Frequent detection of highly diverse variants of Cardiovirus, Cosavirus, Bocavirus, and Circovirus in sewage samples collected in the United States. *Journal of Clinical Microbiology* **2009,** *47*, 3507-3513.

16. Wong, K.; Onan, B. M.; Xagoraraki, I., Quantification of enteric viruses, pathogen indicators, and *Salmonella* bacteria in Class B anaerobically digested biosolids by culture and molecular methods. *Applied and Environmental Microbiology* **2010,** *76*, 6441-6448.

17. Benschop, K.; Molenkamp, R.; van der Ham, A.; Wolthers, K.; Beld, M., Rapid detection of human Parechoviruses in clinical samples by real-time PCR. *Journal of Clinical Virology* **2008,** *41*, 69-74.

18. Bibby, K.; Peccia, J., Prevalence of respiratory adenovirus species B and C in sewage sludge. *Environmental Science: Processes & Impacts* **2013**. **DOI:** 10.1039/C2EM30831B