Published in final edited form as:
Lett Appl Microbiol. 2011 April ; 52(4): 386–392. doi:10.1111/j.1472-765X.2011.03014.x.

Viral Metagenome Analysis to Guide Human Pathogen Monitoring in Environmental Samples

Kyle Bibby, **Emily Viau**, and **Jordan Peccia**[*]
Department of Chemical and Environmental Engineering, Yale University, New Haven, CT, 06520, USA

Abstract

Aims—The aim of this study was to develop and demonstrate an approach for describing the diversity of human pathogenic viruses in an environmentally isolated viral metagenome.

Methods and Results—*In silico* bioinformatic experiments were used to select an optimum annotation strategy for discovering human viruses in virome datasets, and applied to annotate a class B biosolids virome. Results from the *in silico* study indicated that less than 1% errors in virus identification could be achieved when nucleotide-based search programs (BLASTn or tBLASTx), viral genome only databases, and sequence reads greater than 200 nt were considered. Within the 51,925 annotated sequences, 94 DNA and 19 RNA sequences were identified as human viruses. Virus diversity included environmentally transmitted agents such as parechovirus, coronavirus, adenovirus, and aichi virus, as well as viruses associated with chronic human infections such as human herpes and hepatitis C viruses.

Conclusions—This study provided a bioinformatic approach for identifying pathogens in a virome dataset, and demonstrated the human virus diversity in a relevant environmental sample.

Significance and Impact of Study—As the costs of next generation sequencing decrease, the pathogen diversity described by virus metagenomes will provide an unbiased guide for subsequent cell-culture and quantitative pathogen analyses, and ensures that highly enriched and relevant pathogens are not neglected in exposure and risk assessments.

Keywords

virus; bioinformatics; biosolids; next generation DNA sequencing; viral metagenome; pathogen; virome

Introduction

Next generation DNA sequencing has recently been applied to study viral metagenomes (viromes) in varied environmental matrices including fresh water (Djikeng et al. 2009; Lopez-Bueno et al. 2009), oceans (Angly et al. 2006), and reused wastewater (Rosario et al. 2009). Most of these studies are interested in describing gene diversity, but a potential application within virome studies is to determine which human viral pathogens are most prevalent in a given environmental matrix. The major limitation to method extension toward human viral identifications is that post-sequencing bioinformatics protocols for analyzing viromes are in nascent stages of development and require careful consideration to produce the high quality virome annotations required for pathogen identification. Viruses do not

[*]Corresponding Author: Mailing Address: Department of Chemical and Environmental Engineering, Yale University, New Haven, Connecticut 06520. Phone: (203) 432-4385. Fax (203) 432-4387., jordan.peccia@yale.edu.

contain ubiquitous genetic elements such as the 16s rRNA encoding genes (Rohwer and Edwards 2002), hence, virome studies must sort and assemble sequences that could come from any location on the viral genome. Unresolved virome construction and annotation concerns include uncertainty about the optimal sequence read length for identification, as well as appropriate use of databases and database search programs. An additional limitation is the unknown sequencing depth required to reach the rare human viruses amidst the ubiquity of bacteriophages in the environment (Breitbart and Rohwer 2005).

The goal of this study was to develop and test a method for describing the diversity of human pathogenic viruses in an environmentally-isolated virome. To improve sequence analysis methods, we conducted an *in silico* study of 10 known human viral pathogen genomes with the aim of decreasing errors in annotating next generation sequencing reads as pathogenic viral nucleic acids. As a demonstration, viral DNA and RNA (cDNA) extracted from sewage sludge residuals resulting from municipal wastewater treatment (termed biosolids) were sequenced using 454 Life Sciences pyrosequencing technology. We then applied the optimal annotation schemes identified by the *in silico* study to describe the diversity and abundance of viral pathogens, and to determine the sequencing depth required to portray this viral pathogen diversity. Biosolids are ideal for demonstrating virome pathogen recovery as this waste stream originates from the solids residuals of wastewater treatment plants serving up to one million people, their pathogen content is not well documented (Gerba et al. 2002; Viau and Peccia 2009), and growing public opposition to the land application of biosolids as a soil conditioning product has initiated an expressed desire for comprehensive viral pathogen surveys in biosolids (NRC 2002).

Materials and methods

Bioinformatic experiments

An *in silico* study was conducted by parsing the genomes of ten environmentally relevant viruses into short sequences and determining the sequence length, BLAST program, and viral databases that resulted in the highest confidence of correct annotation. Human virus genotypes were chosen to represent common environmental viral diseases caused by inhalation and ingestion exposure routes (Table 1). Artificial reads were produced at every location along the genome and lengths were set to represent common read lengths produced by next generation sequencing platforms: 100 nucleotide (nt) reads from Illumina Genome Analyzer, 200 nt paired-end reads from Illumina HiSeq2000, and 400 nt reads from 454 Life Sciences GS FLX sequencer with Titanium chemistry.

To choose optimal sequence search and alignment programs, the following NCBI BLASTALL programs were compared: BLASTn nucleotide to nucleotide searches, BLASTx translated nucleotide to amino acid searches, and tBLASTx translated nucleotide to translated nucleotide searches. Each BLASTALL program was applied to two search databases including the full NCBI database (nt, non-redundant nucleotide database and nr, non-redundant amino acid database) and the NCBI viral databases (vnt, nucleotide database and vnr, amino acid database).

When the top hit, as determined by lowest E-value, matched the human pathogen strain that was searched for and there were no ambiguous classifications (i.e. same virus but different host), the read was listed as correct. In the case of multiple hits with equivalent E-values, the highest bit score was used for annotation. Reads were classified as missing if they contained no hits at or below the 10^{-3} E-value threshold. The sum of ambiguous and missing sequences were grouped and reported as total classification errors. Classifications were only made when an Evalue of 10^{-3} or less was observed. This E-value threshold was based on precedent set in prior virome studies (Zhang et al. 2005; Lopez-Bueno et al. 2009; Rosario et al. 2009), and also from an evaluation of annotating 100 nt adenovirus sequence segments using an E-value of either 10^{-3} or 10^{-5}. This evaluation revealed that by excluding a greater number of correct sequence reads, an average 18% increase in error was observed when the E-value threshold was set at 10^{-5} instead of 10^{-3} (Table S1).

Biosolids sample preparation and sequencing

Class B biosolids were sampled from an anonymous U.S. wastewater treatment facility that collected solid residuals by primary sedimentation and secondary activated sludge clarification, and treated by mesophilic anaerobic digestion (35–37 °C, 15 d solids retention time). Digested sludges were dewatered by belt pressing to 17% solids content. Previous class B biosolids indicator and pathogen monitoring from this plant revealed fecal coliform concentrations of 5.1×10^4 colony forming units/dry g, male-specific coliphage concentrations of 2.7×10^4 plaque forming units/dry g, and adenovirus concentrations of 3×10^6 genomic units/dry g (Viau and Peccia 2009).

Five 100 g grab samples were collected in accordance with U.S. EPA method 1680 (USEPA 2006), and shipped on ice overnight to the laboratory. Within 24 hours of collection, biosolids samples were recombined to form a composite sample and viruses were eluted and concentrated following a U.S. EPA method for the recovery of viruses from sludge (USEPA 1999a). The concentrated viral solution was passed through a 0.45 μm filter to remove any remaining bacterial and eukaryotic cells and DNase/RNase-digested with OmniCleave endonuclease (Epicentre Biotechnologies, Madison, WI) to remove any naked nucleic acids. Purified viral extracts were stored at −80°C.

Both DNA and RNA were recovered from the viral concentrate. Three DNA extractions were performed each with 0.6 ml of viral concentrate using the MoBio PowerSoil DNA kit (MoBio Laboratories, Carlsbad, CA) and modifications described elsewhere (Viau and Peccia 2009). Triplicate RNA extractions were performed with 2 ml of the viral concentrate each using the MoBio PowerSoil RNA kit (MoBio) followed by DNA digestion. Viral RNA was converted to cDNA with a Multiscribe high-capacity cDNA reverse transcription kit (Life Technologies™ AB, Carlsbad, CA).

Samples were combined and a total of 5 μg of DNA and cDNA each were sent to the Yale Center for Excellence in Genome Science for shotgun pyrosequencing on a 454 GS-FLX sequencer using Titanium Chemistry (Roche Diagnostics Corporation, Indianapolis, IN). One quarter of a microwell plate was used for this analysis. Prior to sequencing, DNA was fragmented by nebulization into 300 to 800 nt sequences.

Virome annotation

To remove artificial replicates, a known artifact of 454 pyrosequencing, the 454 replicate filter with default settings was used (Gomez-Alvarez et al. 2009). Filtered sequence reads were assembled with the Newbler runAssembly program from the 454 Life Sciences Data Analysis 2.3 package (Branford, CT). Sequence assembly settings utilized a minimum overlap of 40 bp and a minimum identity of 90%, while all other settings were default. Unassembled sequences (singletons) were then extracted and combined with assembled contiguous sequences (contigs) for annotation. The virome data was annotated by tBLASTx searches within the NCBI viral database from January 2010. Annotation used the previously described E-score selection criteria. Sequences are available from the NCBI Sequence Read Archive under accession SRX016659.

Results

Annotation accuracy

For the in silico experiments, the percentage of erroneous reads (Figure 1) suggests that the most appropriate annotation strategy for viral pathogen identification will be a nucleotide-based search (BLASTn or tBLASTx) with a virus only database, and read lengths of 200 nt and greater. Average error rate in classification using these methods was 0.1% with a 1.2% to 0% range among the 10 viruses.

Four other important trends emerged for annotating human viruses from short virome sequences. First, smaller, more focused viral databases resulted in less incorrect classifications than the larger databases (Figure 1). Viral databases resulted in less total error in 87 of 90 search scenarios conducted. Secondly, the amino acid-based search program, BLASTx, produced a greater amount of total errors than the nucleotide-based search programs, BLASTn and tBLASTx, when comparing both the complete and virus only databases. For example, when using 200 nt artificial reads and the virus database, BLASTx percent errors averaged 355 times greater than BLASTn errors and 35 times greater than tBLASTx errors. Results from BLASTn and tBLASTx were statistically indistinguishable for total errors using the viral databases at read lengths of 200 nt (P= 0.367). Third, increasing read length either maintained or reduced the number of errors in all scenarios considered. When read length was increased from 100 nt to 200 nt, the tBLASTx overall error decreased five times to 0.13%, while average error for 400 nt reads was decreased to 0.0015%. Finally, the number of errors was extremely dependent on the type of virus (Table S2). BLASTx-nr total errors ranged from 7.4% for norovirus to 94.5% for rotavirus for 100 nt reads. Total error rates in rotavirus were high end outliers to the other genome sequences due to the high similarity of genome sequences between human and animal rotaviruses.

Human viruses in class B biosolids

After replicate filtering, sequencing provided 123,893 raw sequences. Reads were assembled into 1,028 contigs that averaged 874 nt and 46,153 singletons that averaged 260.7 nt. Through tBLASTx comparison with the NCBI viral-nt database, 51,925 total sequences were annotated and classified as being of viral origin (215 contigs comprising 48,831 sequences, and 3,094 singletons) Within these viral classifications, ten different human pathogen viruses (16 strains) representing 113 sequences were identified and included 94 DNA virus sequences and 19 RNA virus sequences (Table 2). Only three sequences identified as human pathogens were ambiguous and were excluded from these results. Through comparisons with the Greengeens core set rDNA database, less than 0.2% of all sequences were annotated as bacteria (10^{-30} E-value threshold) (DeSantis et al. 2006).

When using shotgun sequencing techniques, it is recognized that the likelihood of a viral fragment being identified is a function of both the virus' abundance and genome size (Angly et al. 2009). Figure 2 shows the potential number of virus genomes relative to adenovirus content after correction for virus genome size. These results indicate that the RNA viruses parechovirus and coronavirus and the DNA virus herpes virus were the most abundant human viruses in the biosolids sample tested here. Overall, annotated viral sequences consisted of 33.8% eukaryotic viruses and 66.2% bacteriophages, while human pathogenic viruses comprised less than 0.1% of total sequences (Figure 2 *inset*). Table S3 provides a complete list of eukaryotic viruses identified in this study.

Discussion

Bioinformatic approaches to improve annotation certainty

Use of the viral database with either BLASTn or tBLASTx search programs and read lengths greater than 200 nt is recommended for annotating human pathogen diversity in environmental virome sequences. This recommendation is, however, specific for the goal of pathogen identification and differs from the common practice of using BLASTx for annotating functional genes and nonpathogenic viruses in previous metagenome studies (Breitbart et al. 2002; Angly et al. 2006; Vega Thurber et al. 2008; Djikeng et al. 2009; Lopez-Bueno et al. 2009; Coetzee et al. 2010). Here, the BLAST search program tBLASTx was used to annotate human pathogens from the biosolids virome. Higher error rates in the translated nucleotide to amino acid BLASTx searches are likely due to the presence of non-coding viral genome regions in queries. Searches conducted by tBLASTx include non-protein encoding regions that are left out of BLASTx searchers. Although the percent errors associated with the nucleotide to nucleotide BLASTn searches were statistically indistinguishable to those associated with the translated nucleotide to translated nucleotide tBLASTx searches, the latter offers advantages associated with amino acid conservation that are not included in BLASTn searches. This amino acid conservation advantage is demonstrated in the biosolids virome sequencing effort where a BLASTn search of the viral nucleotide database yielded only 8,726 sequence identifications compared to the total of 51,925 sequences identifications using tBLASTx. Finally, and in addition to decreased computational time, advantages of the focused virus database are that it does not include similar sequences from non-target organisms that may be deposited into full databases and thus results in lower ambiguity than the full NCBI database. Physical separation of virus sized particles and destruction of free DNA and RNA during sample preparation ensure that the gene sequences produced were of viral origin and obviate the need for annotation using databases that contain additional, nonviral, nucleic acid sequences.

Viral pathogen diversity in class B biosolids

A major public health concern surrounding the land application of biosolids is risk of infection from viruses that are aerosolized when spread onto land or viruses that enter into ground or surface water supplies (Westrell et al. 2004; Brooks et al. 2005; Eisenberg et al. 2008). To date, viruses previously found in biosolids by PCR-based methods or culturing include enterovirus (Gerba et al. 2002; Wong et al. 2010), polyomavirus (Bofill-Mas et al. 2006), reovirus (Gallagher and Margolin 2007), hepatitis A virus (Straub et al. 1994), norovirus (Wong et al. 2010) and adenovirus (Viau and Peccia 2009; Wong et al. 2010). The resulting data from these different studies suggests that adenoviruses are the most abundant human virus in class B biosolids (Viau and Peccia 2009; Schlindwein 2010; Wong et al. 2010). These previous efforts, however, were limited by a requirement that investigators must choose the viruses that will be searched for. By contrast, the production of a viral metagenome produces a list of viruses that is based on abundance and is independent of researcher bias.

Of the viruses described in Figure 2, their high prevalence within the general population further improves confidence in their identification in biosolids. Viruses found in this study with known environmental routes of exposure and causing respiratory and gastroenteritis infections include adenovirus (Crabtree et al. 1997), parechovirus (Baumgarte et al. 2008), aichi virus (Le Guyader et al. 2008), torque teno virus (TTV) (Griffin et al. 2008), and coronavirus (Yu et al. 2004). Coronavirus is recognized as a major cause of the common cold (Falsey et al. 2002) and 95% of humans are infected with parechovirus within two to five years of age (Joki-Korpela and Hyypiä 2001). Commonly used enterovirus qPCR primers do not include parechovirus, thus this virus’s presence has not been reflected in previous qPCR enterovirus monitoring (Wong et al. 2010). Although commonly enumerated in class B biosolids (Gerba et al. 2002), enteroviruses were not detected by this study. TTV also circulates in healthy individuals with an estimated worldwide prevalence of 80%, and researchers have suggested its use as a fecal indicator (Bendinelli et al. 2001;Griffin et al. 2008). Both Aichi virus and parechovirus have been identified by a sequencing efforts in reused wastewater (Rosario et al. 2009), providing support for their presence in wastewater residuals. Among viral agents described that do not have environmental exposure routes, herpes virus may be carried by as much as 90% of the population (Arbuckle et al. 2010).

The identifications generated by this viral metagenome sequencing effort are intended to direct quantitative pathogen monitoring efforts, not replace them. Obtaining a more unbiased view of virus diversity through virome production is labor intensive and costly. Given limited database size and the inherent genetic similarity between host specific viruses, some level of classification error may always be present. An example of this is the unexpected identifications by this sequencing effort of Variola virus (Table S3). Since smallpox (Variola virus) has been eradicated worldwide since 1980 it is likely that these identifications are from some other member of the family *Poxviridae*. While the results indicate that some form of *Poxviridae* are present, these known ambiguities highlight the need to accompany virus metagenome-based pathogen identifications with more in depth, confirmatory analysis. The ability to distinguish host specificity was demonstrated in our *in silico* study, however, this may not extend to all viruses. These limitations suggest that rather than applying massively parallel sequencing as the only form of virus detection, a more appropriate approach for using virus metagenome information is to describe the viral pathogen diversity of a class of environmental samples (e.g. class B biosolids) in order to efficiently guide quantitative and confirmatory analysis of selected agents of interest.

Conclusions

Through an *in silico* study of simulated viral pathogens reads and an initial biosolids virome sequencing effort, this work has demonstrated the utility of next generation DNA sequencing for identifying human viruses in environmental samples of concern. An annotation approach specific for pathogen identification is described that delineates appropriate BLAST programs (tBLASTx, BLASTn), databases (viral-only database) and required sequence lengths (>200nt) to achieve less than 1% error in viral pathogen classification. Several viruses not previously identified in biosolids, including coronavirus, herpes virus, TTV, and parechovirus, were identified and ranked as highly abundant compared to adenoviruses. These results indicate the importance of obtaining an unbiased view of viral pathogen diversity as a guide for subsequent cell culture and specific quantitative PCR investigations required to fully understand biosolids pathogen content.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Science Foundation grant BES0348455. Computing was carried out at the Yale University Center for High Performance Computation in Biology and Biomedicine (HPC) which is supported by NIH grant RR19895. The authors were grateful for the advice and assistance of Nicholas Carriero and Rob Bjornsen at the high performance computation center. KB was supported by a fellowship from the Environmental Education and Research Foundation and the U.S. EPA.

References

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F. The Marine Viromes of Four Oceanic Regions. PLoS Biol 2006;4:e368. [PubMed: 17090214]

Angly FE, Willner D, Prieto-Davó A, Edwards RA, Schmieder R, Vega-Thurber R, Antonopoulos DA, Barott K, Cottrell MT, Desnues C, Dinsdale EA, Furlan M, Haynes M, Henn MR, Hu Y, Kirchman DL, McDole T, McPherson JD, Meyer F, Miller RM, Mundt E, Naviaux RK, Rodriguez-Mueller B, Stevens R, Wegley L, Zhang L, Zhu B, Rohwer F. The GAAS Metagenomic Tool and Its Estimations of Viral and Microbial Average Genome Size in Four Major Biomes. PLoS Comput Biol 2009;5:e1000593. [PubMed: 20011103]

Arbuckle JH, Medveczky MM, Luka J, Hadley SH, Luegmayr A, Ablashi D, Lund TC, Tolar J, De Meirleir K, Montoya JG, Komaroff AL, Ambros PF, Medveczky PG. The latent human herpesvirus-6A genome specifically integrates in telomeres of human chromosomes in vivo and in vitro. Proc Natl Acad Sci USA 2010;107:5563–5568. [PubMed: 20212114]

Baumgarte S, de Souza Luna LK, Grywna K, Panning M, Drexler JF, Karsten C, Huppertz HI, Drosten C. Prevalence, Types, and RNA Concentrations of Human Parechoviruses, Including a Sixth Parechovirus Type, in Stool Samples from Patients with Acute Enteritis. J Clin Microbiol 2008;46:242–248. [PubMed: 18057123]

Bendinelli M, Pistello M, Maggi F, Fornai C, Freer G, Vatteroni ML. Molecular Properties, Biology, and Clinical Implications of TT Virus, a Recently Identified Widespread Infectious Agent of Humans. Clin Microbiol Rev 2001;14:98–113. [PubMed: 11148004]

Bofill-Mas S, Albinana-Gimenez N, Clemente-Casares P, Hundesa A, Rodriguez-Manzano J, Allard A, Calvo M, Girones R. Quantification and Stability of Human Adenoviruses and Polyomavirus JCPyV in Wastewater Matrices. Appl Environ Microbiol 2006;72:7894–7896. [PubMed: 17028225]

Breitbart M, Rohwer F. Here a virus, there a virus, everywhere the same virus? Trends Microbiol 2005;13:278–284. [PubMed: 15936660]

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F. Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci USA 2002;99:14250–14255. [PubMed: 12384570]

Brooks JR, Tanner BD, Josephson KL, Gerba C, Haas CN, Pepper I. A national survey on the residential impact of biological aerosols from the land application of biosolids. J Appl Microbiol 2005;99:310–322. [PubMed: 16033462]

Coetzee B, Freeborough MJ, Maree HJ, Celton JM, Rees DJG, Burger JT. Deep sequencing analysis of viruses infecting grapevines: Virome of a vineyard. Virology 2010;400:157–163. [PubMed: 20172578]

Crabtree KD, Gerba CP, Rose JB, Haas CN. Waterborne adenovirus: a risk assessment. Water Sci Technol 1997;35:1–6.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol 2006;72:5069–5072. [PubMed: 16820507]

Djikeng A, Kuzmickas R, Anderson NG, Spiro DJ. Metagenomic Analysis of RNA Viruses in a Fresh Water Lake. PLoS ONE 2009;4:e7264. [PubMed: 19787045]

Eisenberg JNS, Moore K, Soller JA, Eisenberg DM, Colford JM Jr. Microbial risk assessment framework for exposure to amended sludge projects. Environ Health Persp 2008;116:727–733.

Falsey AR, Walsh EE, Hayden FG. Rhinovirus and Coronavirus Infection-Associated Hospitalizations among Older Adults. J Infect Dis 2002;185:1338–1341. [PubMed: 12001053]

Gallagher EM, Margolin AB. Development of an integrated cell culture--Real-time RT-PCR assay for detection of reovirus in biosolids. J Virol Methods 2007;139:195–202. [PubMed: 17161876]

Gerba C, Pepper I, Whitehead L. A risk assessment of emerging pathogens of concern in the land application of biosolids. Water Sci Technol 2002;46:225–230. [PubMed: 12479475]

Gomez-Alvarez V, Teal TK, Schmidt TN. Systematic artifacts in metagenomes from complex microbial communities. ISME J 2009;3:1314–1317. [PubMed: 19587772]

Griffin J, Plummer J, Long S. Torque teno virus: an improved indicator for viral pathogens in drinking waters. Virol J 2008;5:112. [PubMed: 18834517]

Joki-Korpela P, Hyypiä T. Parechoviruses, a novel group of human picornaviruses. Ann Med 2001;33:466–471. [PubMed: 11680794]

Le Guyader FS, Le Saux JC, Ambert-Balay K, Krol J, Serais O, Parnaudeau S, Giraudon H, Delmas G, Pommepuy M, Pothier P, Atmar RL. Aichi Virus, Norovirus, Astrovirus, Enterovirus, and Rotavirus Involved in Clinical Cases from a French Oyster-Related Gastroenteritis Outbreak. J Clin Microbiol 2008;46:4011–4017. [PubMed: 18842942]

Lopez-Bueno A, Tamames J, Velazquez D, Moya A, Quesada A, Alcami A. High Diversity of the Viral Community from an Antarctic Lake. Science 2009;326:858–861. [PubMed: 19892985]

NRC. Biosolids applied to land: advancing standards and practices. Washington D.C: National Research Council of the National Academies; 2002.

Rohwer F, Edwards R. The Phage Proteomic Tree: a Genome-Based Taxonomy for Phage. J Bacteriol 2002;184:4529–4535. [PubMed: 12142423]

Rosario K, Nilsson C, Lim YW, Ruan Y, Breitbart M. Metagenomic analysis of viruses in reclaimed wastewater. Environ Microbiol 2009;11:2806–2820. [PubMed: 19555373]

Schlindwein AD, Rigotto C, Simoes CMO, Barardi CRM. Detection of enteric viruses in sewage sludge and treated wastewater effluent. Water Sci Technol 2010;61:537–544. [PubMed: 20107281]

Straub TM, Pepper IL, Gerba CP. Detection of naturally occurring enteroviruses and hepatitis A virus in undigested and anaerobically digested sludge using the polymerase chain reaction. Can J Microbiol 1994;40:884–888. [PubMed: 7528092]

USEPA. Environmental Regulations and Technology: Control of Pathogens and Vector Attraction in Sewage Sludge. Washington DC: Office of Research and Development, US Environmental Progection Agency; 1999.

USEPA. Method 1680: Fecal Coliforms in Sewage Sludge (Biosolids) by Multiple-Tube Fermentation using Lauryl Tryptose Broth (LTB) and EC Medium. 2006.

Vega Thurber RL, Barott KL, Hall D, Liu H, Rodriguez-Mueller B, Desnues C, Edwards RA, Haynes M, Angly FE, Wegley L, Rohwer FL. Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral Porites compressa. Proc Natl Acad Sci USA 2008;105:18413–18418. [PubMed: 19017800]

Viau E, Peccia J. Survey of Wastewater Indicators and Human Pathogen Genomes in Biosolids Produced by Class A and Class B Stabilization Treatments. Appl Environ Microbiol 2009;75:164–174. [PubMed: 18997022]

Westrell T, Schonning C, Stenstrom TA, Ashbolt NJ. QMRA (quantitative microbial risk assessment) and HACCP (hazard analysis and critical control points) for management of pathogens in wastewater and sewage sludge treatment and reuse. Water Sci Technol 2004;50:23–30. [PubMed: 15344769]

Wong K, Onan BM, Xagoraraki I. Quantification of Enteric Viruses, Pathogen Indicators, and Salmonella Bacteria in Class B Anaerobically Digested Biosolids by Culture and Molecular Methods. Appl Environ Microbiol 2010;76:6441–6448. [PubMed: 20693452]

Yu ITS, Li Y, Wong TW, Tam W, Chan AT, Lee JHW, Leung DYC, Ho T. Evidence of Airborne Transmission of the Severe Acute Respiratory Syndrome Virus. New Engl J Med 2004;350:1731–1739. [PubMed: 15102999]

Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, Soh SWL, Hibberd ML, Liu ET, Rohwer F, Ruan Y. RNA Viral Community in Human Feces: Prevalence of Plant Pathogenic Viruses. PLoS Biol 2005;4:e3. [PubMed: 16336043]

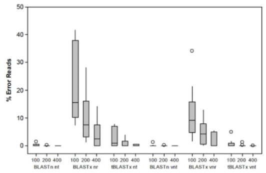

Figure 1.

Box plot of total classification errors for 10 human viruses according to read length and annotation method. Total errors include both ambiguous and missing identifications. Groupings are by the read length (100, 200, 400 nt), the BLAST search program (BLASTn, BLASTx, tBLASTx), and the database where “nt” represents nucleotide database, “nr” represents amino acid database, and “v” represents virus only database. In each box the centerline represents the median, the top and bottom of the box represent the 25th and 75th error percentiles, and the lines represent the data spread. Outliers, marked by circles, were outside three standard deviations of the median. Outliers for Rotavirus were greater than 80% error for the BLASTx nr and tBLASTx nt cases and were excluded from the graph. Complete results for each individual virus are listed in Supplementary Material Table S2.

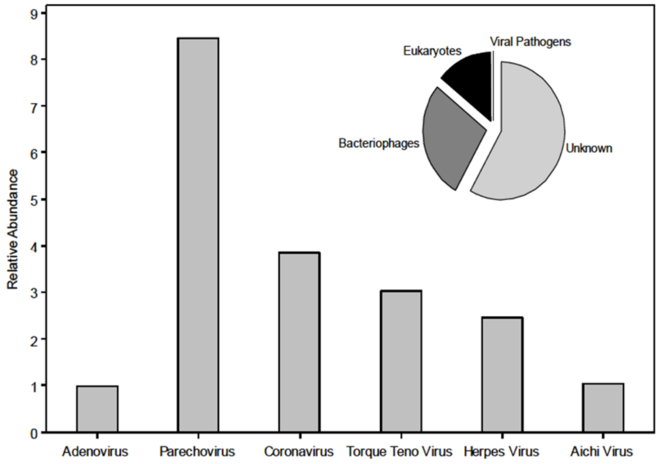

Figure 2.

Relative abundance of pathogenic viruses in biosolids virome normalized by genome size and the abundance of adenovirus. *Inset:* Pie chart of sequence identifications (n=51,000). Human viral pathogens represent less than 0.1% of total sequences.

Table 1

Human viral pathogens included in the *in silico* study

Virus	Nucleic acid	Genome size	Accession number
Adenovirus	dsDNA	34,794 nt	AC_000019
Astrovirus	ssRNA	6,813 nt	NC_001943
Coronavirus	ssRNA	27,317 nt	NC_002645
Hepatitis A virus	ssRNA	7,478 nt	NC_001489
Norovirus	ssRNA	7,654 nt	NC_001959
Parechovirus	ssRNA	7, 348 nt	NC_001897
Polyomavirus JC Respiratory	dsDNA	5,130 nt	NC_001699
Syncytial virus	ssRNA	15,225 nt	NC_001781
Rhinovirus	ssRNA	7,152 nt	NC_001617
Rotavirus	dsRNA	17,448 nt	NC_011507[a]

[a] Segment 1, successive segments also included.

Table 2

Human pathogenic viruses identified in the class B biosolids virome

Virus	Nucleic Acid	Genome Length	Number of Sequences Identified
Human herpesvirus 2	dsDNA	154,746 nt	46
Human herpesvirus 8 type P	dsDNA	137,868 nt	12
Human herpesvirus 1	dsDNA	152,261 nt	10
Human herpesvirus 4	dsDNA	171,823 nt	3
Human herpesvirus 6A	dsDNA	159,322 nt	1
Human coronavirus 229E	ssRNA	27,317 nt	9
Human coronavirus HKU1	ssRNA	29,926 nt	1
Tanapox virus	dsDNA	144,565 nt	9
Orf virus	dsDNA	139,962 nt	8
Human parechovirus	ssRNA	7,348 nt	7
Human adenovirus D	dsDNA	35,083 nt	2
Human adenovirus E	dsDNA	35,994 nt	1
Human adenovirus type 1	dsDNA	36,001 nt	1
Aichi virus	ssRNA	8,521 nt	1
Hepatitis C virus genotype 1	ssRNA	9646 nt	1
Torque Teno Virus-like minivirus	ssDNA	2916 nt	1