Raw Sewage Harbors Diverse Viral Populations

Paul G. Cantalupo,[a] Byron Calgua,[b] Guoyan Zhao,[c] Ayalkibet Hundesa,[b] Adam D. Wier,[a] Josh P. Katz,[a] Michael Grabe,[a] Roger W. Hendrix,[a] Rosina Girones,[b] David Wang,[c] and James M. Pipas[a]

Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania, USA[a]; Department of Microbiology, Faculty of Biology, University of Barcelona, Barcelona, Spain[b]; and Departments of Molecular Microbiology and of Pathology and Immunology, Washington University School of Medicine, St. Louis, Missouri, USA[c]

ABSTRACT **At this time, about 3,000 different viruses are recognized, but metagenomic studies suggest that these viruses are a small fraction of the viruses that exist in nature. We have explored viral diversity by deep sequencing nucleic acids obtained from virion populations enriched from raw sewage. We identified 234 known viruses, including 17 that infect humans. Plant, insect, and algal viruses as well as bacteriophages were also present. These viruses represented 26 taxonomic families and included viruses with single-stranded DNA (ssDNA), double-stranded DNA (dsDNA), positive-sense ssRNA [ssRNA(+)], and dsRNA genomes. Novel viruses that could be placed in specific taxa represented 51 different families, making untreated wastewater the most diverse viral metagenome (genetic material recovered directly from environmental samples) examined thus far. However, the vast majority of sequence reads bore little or no sequence relation to known viruses and thus could not be placed into specific taxa. These results show that the vast majority of the viruses on Earth have not yet been characterized. Untreated wastewater provides a rich matrix for identifying novel viruses and for studying virus diversity.**

IMPORTANCE **At this time, virology is focused on the study of a relatively small number of viral species. Specific viruses are studied either because they are easily propagated in the laboratory or because they are associated with disease. The lack of knowledge of the size and characteristics of the viral universe and the diversity of viral genomes is a roadblock to understanding important issues, such as the origin of emerging pathogens and the extent of gene exchange among viruses. Untreated wastewater is an ideal system for assessing viral diversity because virion populations from large numbers of individuals are deposited and because raw sewage itself provides a rich environment for the growth of diverse host species and thus their viruses. These studies suggest that the viral universe is far more vast and diverse than previously suspected.**

Received 10 August 2011 **Accepted** 18 August 2011 **Published** 4 October 2011

Citation Cantalupo PG, et al. 2011. Raw sewage harbors diverse viral populations. mBio 2(5):e00180-11. doi:10.1128/mBio.00180-11.

Editor Michael Imperiale, University of Michigan Medical School

Address correspondence to James M. Pipas, pipas@pitt.edu.

Viruses are everywhere. On Earth, every species of bacteria, archaea, fungi, plants, worms, insects, and animals is likely to harbor numerous viruses. The presence of viruses is not limited to sites within cellular organisms; extracellular virions are also found in the environment. Oceans, rivers, lakes, and air all contain virions released from infected hosts. Every time we touch another human or pet, and often when we have contact with a contaminated environment, we are exposed to microbes, including viruses. Metagenomic studies of the oceans (1–6), arctic lakes (7), stool samples (8–14), and other environments (15–19) suggest that known viruses are found in unsuspected locations and that a large number of uncharacterized viruses exist in nature.

How big is the viral universe and how many types of viruses exist? Current views of viral diversity are shaped by the analysis of about 3,000 fully sequenced viral genomes representing 84 viral families (20). Recently, powerful metagenomic strategies in which all viruses present in an environmental or clinical sample are detected by sequencing virion-associated nucleic acids have been developed (21). Metagenomic approaches allow simultaneous comparisons of many genomes from multiple taxa, including those viruses that cannot be cultured. We are using metagenomics to explore the virus populations in diverse biomes and unique niches throughout the world. For our initial studies, we sought an environment, raw sewage (untreated wastewater), that we hypothesized would harbor a high diversity of viruses.

Raw sewage represents the effluence of society. Human waste from thousands of individuals is deposited into collection systems that terminate at a common point, the wastewater treatment plant. Pathogens excreted into urban sewage reflect the infections that have been transmitted in the population (22) and would include the viral pathogens that are transmitted through fecally contaminated water or food (23, 24). The implementation of current regulations on wastewater treatments has significantly reduced the levels of microbiological contamination. However, human viruses are still widely disseminated in water and the environment through discharges of untreated and treated sewage (25, 26) to river catchments and to coastal water, water reuse in food irrigation, and shellfish production (27). This mixture of water, human and animal wastes, and plant material forms a special ecosystem supporting insect, rodent, and plant populations as well as both prokaryotic and eukaryotic microorganisms. Viruses are associated with the biological wastes deposited into sewage as well as with all the species growing in sewage, making untreated wastewater an ideal environment for exploring viral diversity. In fact,

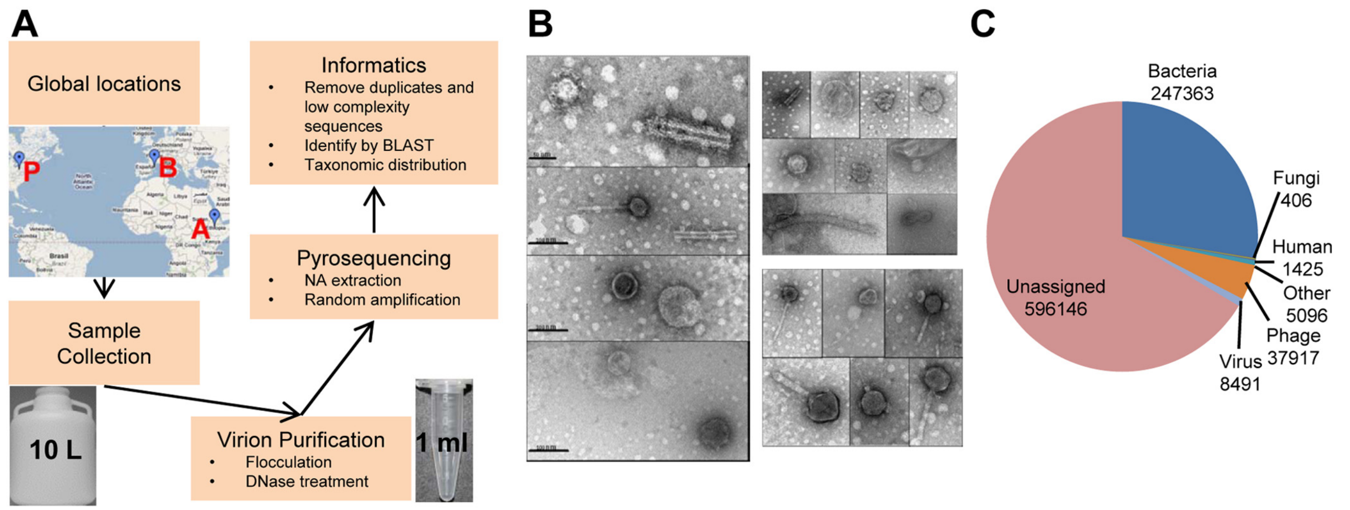

FIG 1 Raw sewage contains diverse viruses. (A) Raw sewage was obtained from three cities (P, Pittsburgh, Pennsylvania, United States; B, Barcelona, Spain; A, Addis Ababa, Ethiopia) on three different continents. Virion populations were concentrated by organic flocculation (31). Raw sewage metagenomes were obtained through pyrosequencing, and the sequences are classified by subsequent bioinformatic methods. 10 L, 10 liters; NA, nucleic acid. (Reproduced from Google—Map data ©2011 Geocentre Consulting, MapLink, Tele Atlas.) (B) Examination of raw sewage by electron microscopy reveals a diversity of virion morphologies. All black bars represent 100 nm, except the top bar, which represents 50 nm. (C) Total nucleic acid (DNA and reverse-transcribed RNA) was sequenced and binned according to taxa based on BLAST searches. Most sequences found within virions do not match the sequences in public databases.

many studies have shown that multiple types of viruses can be found in raw sewage (28–30). Here we report the results of a metagenomic survey of viruses present in raw sewage.

RESULTS

Untreated wastewater was collected from three different locations: (i) Pittsburgh, Pennsylvania, United States; (ii) Barcelona, Spain; and (iii) Addis Ababa, Ethiopia (Fig. 1A). Electron microscopy confirmed the presence of numerous different virion morphologies in the samples (Fig. 1B). Virions were concentrated and purified by organic flocculation and DNase treatment (31). In order to capture the genomes of both DNA and RNA viruses, total nucleic acids were isolated from each sample and reverse transcribed followed by deep sequencing. This resulted in a total of 897,647 high-quality reads (approximately 278 megabases) from all three samples (see Table S1 in the supplemental material). Each individual read was then compared to databases by a series of BLAST searches and binned according to taxa (Fig. 1C). A total of 8,491 sequence reads were most closely related to eukaryotic viruses, while 37,917 were most closely related to bacteriophages. About 27% of the sequence reads (247,363) were identified as bacterial, a number consistent with other metagenomic studies. Since these sequences were obtained from a purification scheme designed to enrich for virions, the putative bacterial sequences most likely represent either prophage genes misannotated as bacterial or gene transfer agents (GTAs) (8, 15, 19, 32, 33). Most sequences (596,146) showed no sequence relation to any known sequences in the databases and thus are most likely to be derived from novel, uncharacterized viruses. Further analysis of the bacterial and unassigned sequences is described in the supplementary material.

Raw sewage contains many known viruses from a diversity of hosts. We further partitioned the individual reads that have a significant BLAST hit (see Materials and Methods) to eukaryotic viruses or phages into two categories: known and novel. We arbitrarily defined known sequences as those that are related to a viral genome listed in the NCBI taxonomy database with ≥80% sequence identity over ≥95% of the length of the sequence read. By these criteria, 3,027 reads were deemed to be derived from known viruses. The remaining sequences were binned as novel viruses and are discussed below. Analysis of the sequences identified as known viruses demonstrates that our methods detected diverse types of viruses. We detected 234 known viruses. Members of 26 different families, including those with double-stranded DNA (dsDNA), single-stranded DNA (ssDNA), positive-sense ssRNA [ssRNA(+)], and dsRNA genomes, and those with either enveloped or nonenveloped virions were found, making raw sewage the most diverse viral biome examined thus far (Fig. 2A; see Table S2 in the supplemental material).

Like other biomes that have been studied, the virome of raw sewage is dominated by bacteriophages. Of the 46,408 high-quality reads that matched viruses in the databases at this time, 37,917 (~80%) were related to bacteriophages. These viruses included members of 13 virus families, but members of five families dominated the population. The five families were the *Microviridae* (37%), *Siphoviridae* (24%), *Myoviridae* (17%), *Podoviridae* (14%), and *Inoviridae* (3%). These bacteriophage families are associated with 24 bacterial host species, but over half of the reads are related to bacteriophages that infect enterobacteria or lactococci (Fig. 2C). The bacteriophage sequences binned as novel viruses outnumbered those that matched bacteriophage genomes in GenBank databases by 30:1.

Most of the known eukaryotic virus reads (90.9%) found in raw sewage were derived from plant viruses (Fig. 2B). This is not surprising, given that plant viruses dominate the viral communities present in human stool samples and that they have been detected in a number of aquatic biomes (13, 29). Roughly 85% of the sequence reads classified as known viruses were derived from 18 different species of the family *Virgaviridae*. Many other types of plant viruses were found; they included members of the *Alphaflexiviridae*, *Betaflexiviridae*, *Bromoviridae*, *Closteroviridae*, sobe-

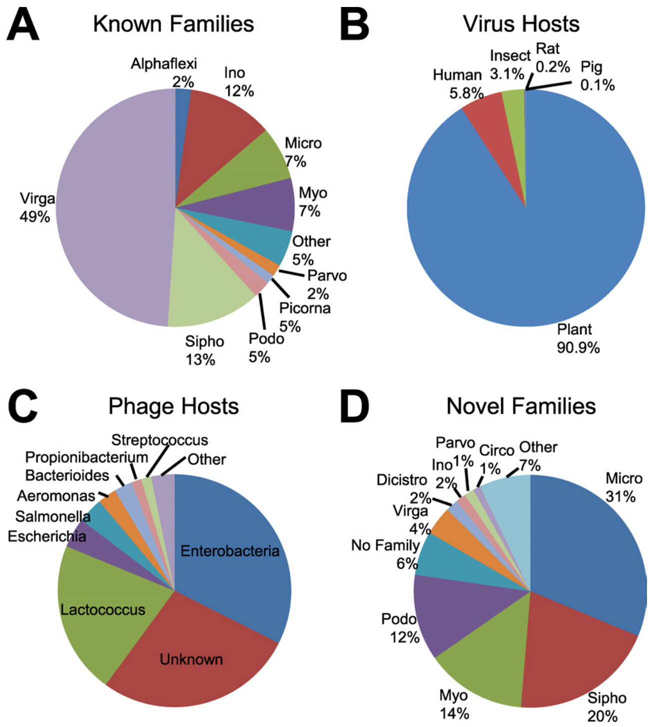

FIG 2 Raw sewage contains many known and novel viruses. (A) Known sequences ($n = 3{,}027$) identified by BLAST are related to many different viral families. Families with <1% abundance were collapsed into the "Other" category. Only the prefixes of family names are shown (e.g., Virga for *Virgaviridae*). (B) Distribution of the hosts of the known eukaryotic virus reads ($n = 1{,}748$). Plant, human, and insect viruses are abundant in raw sewage. (C) Distribution of the hosts of the known bacteriophage reads ($n = 1{,}279$). (D) Novel sequences ($n = 43{,}381$) identified by BLAST are related to many different virus families. Families with <1% abundance were collapsed into the "Other" category. See Table S6 for a list of families and hosts in the "other" category.

TABLE 1 Human viruses present in raw sewage

Family	Species	Genome
Adenoviridae	Human adenovirus 41	dsDNA
Astroviridae	Astrovirus MLB1	ssRNA(+)
	Human astrovirus 1	ssRNA(+)
Caliciviridae	Norwalk virus	ssRNA(+)
	Sapporo virus	ssRNA(+)
Papillomaviridae	Human papillomavirus 112	dsDNA
Parvoviridae	Adeno-associated virus	ssDNA
	Human bocavirus 2	ssDNA
	Human bocavirus 3	ssDNA
Picobirnaviridae	Human picobirnavirus	dsRNA
Picornaviridae	Aichi virus	ssRNA(+)
	Human klassevirus 1/Salivirus NG-J1	ssRNA(+)
	Human parechovirus 1	ssRNA(+)
	Human parechovirus 3	ssRNA(+)
	Human parechovirus 4	ssRNA(+)
	Human parechovirus 7	ssRNA(+)
Polyomaviridae	Polyomavirus HPyV6	dsDNA

movirus, *Tombusviridae*, and *Tymoviridae*. A large number of insect virus reads (3.1%), including those that infect cockroaches, flies, and mosquitoes, were present in all three samples. Insect viruses most likely are present because some insect viruses also infect plants and because wastewater transmission lines can harbor large insect populations. These viruses included members of the *Dicistroviridae*, *Iridoviridae*, *Nodaviridae*, and *Parvoviridae* families. We also identified several viruses of rodents, including strains closely related to a newly identified rat hepatitis E virus (see Fig. S1 and Table S3 in the supplemental material) (34).

We detected 17 viruses known to infect humans in the three sewage samples (Table 1). These viruses included human adenovirus, a well-studied indicator of human fecal contamination (35, 36), as well as a number of known human pathogens, including astroviruses, Norwalk virus, and members of the family *Picornaviridae*, such as Aichi virus and parechoviruses. We also detected the newly discovered klassevirus (37). The relatively newly characterized human bocavirus and picobirnaviruses were also present. We also detected human papillomavirus 112 (data not shown) and the newly discovered human polyomavirus 6 (see Fig. S1 and Table S3 in the supplemental material) (38). Both of these viruses are tropic for skin, suggesting that viruses from human skin as well as stools find their way into sewage, possibly through excretion in urine as is the case for human polyomaviruses. Despite the large number of viruses detected, the current depth of sequencing was not sufficient to detect all viruses known experimentally to be present in the samples. For example, no sequences related to the human polyomavirus JC virus (JCV) were found, even though its presence in the samples was established by PCR (Table 2).

Raw sewage contains many novel viruses. Next, we examined the 43,381 sequence reads that represent novel viruses according to our criteria (see Table S4 in the supplemental material). Figure 3 shows the distribution of these sequences by identity to known viruses in the GenBank databases. The outer ring represents the group of sequences with >90% identity to the reference genome in the top BLAST hit. The internal rings indicate sequences with decreasing identity, binned by 10% intervals. The area of each colored circle is proportional to the number of sequence reads that match the reference genome at a given percent identity in that location for that virus family. The color of the circle indicates the location from which the sequence was obtained. For some virus families, such as the *Virgaviridae*, nearly all of the sequence reads matched known viruses. In other cases, such as the *Picornaviridae* and *Parvoviridae*, some of the sequences matched recognized viruses, but the majority were only distantly related to known members of these families. In most cases (i.e., *Circoviridae*, *Phycodnaviridae*, *Microviridae*, and *Siphoviridae*), nearly all of the sequence reads were derived from putatively novel viruses. Thus, greater than 90% of the sequence reads that could be aligned to known viruses represent sequences from novel viruses that have not been described previously. The novel viruses in the samples show enormous diversity, falling into 51 different viral families (Fig. 2D; see Table S4 in the supplemental material).

Next, we assembled the virus sequence reads and aligned them to a common GenBank reference genome (Fig. 4). In these fragment recruitment plots, assembled sequences belonging to a particular virus family were aligned to GenBank reference genomes for that virus family. Then, the sequence relations of the common regions were compared to each other and to known members of the viral taxon using standard phylogenetic methods.

For example, Fig. 4A shows the four assembled sequences that align to the same region of the human bocavirus genome. Phylogenetic analysis of these sequences suggests that they each repre-

TABLE 2 Detection of classical and emerging viruses in urban sewage by PCR assays

		Virus detected[b] in urban sewage sample from:		
			Addis Ababa, Ethiopia	
Virus analyzed[a]	PCR type	Barcelona, Spain	Sample 1	Sample 2
Human adenovirus	Real time	10,100 GC/ml	10.3 GC/ml	802 GC/ml
JC polyomavirus	Real time	18.3 GC/ml	178 GC/ml	734 GC/ml
Human hepatitis E virus	Nested	−	+	−
Human hepatitis A virus	Nested	−	+	+
Klassevirus 1	Nested	+	+	+
Asfarvirus like-virus	Nested	−	+	−

[a] See Materials and Methods for references for each PCR.

[b] GC, genome copies; −, not detected; +, detected. The volume of sample analyzed in 10 μl of extracted nucleic acid was 33.33 ml for the sewage sample from Barcelona, Spain, and the volume was 43.75 ml for the samples from Addis Ababa, Ethiopia.

sent a different novel bocavirus. Figure 4C shows a similar analysis of 11 sequences that align to the human picobirnavirus genome. Picobirnaviruses are dsRNA viruses whose genome consists of two segments. Five assembled sequences aligned to a common region of genomic segment 1, while six aligned to segment 2. Again, phylogenetic analysis suggests the presence of 5 or 6 novel picobirnaviruses. Fragment recruitment plots also suggested the presence of at least three different novel viruses related to the human pathogen Aichi virus (Fig. 4B), and multiple novel viruses related to the dicistroviruses (see Fig. S2 in the supplemental material). In addition, a large number of novel circovirus-like genomes were identified (see Table S4 in the supplemental material; also data not shown). Circoviruses are a family of viruses with a single-stranded circular DNA genome that have been shown to be present in animal, bird, and human feces as well as raw sewage (29, 39).

A novel member of the *Inoviridae* is abundant in raw sewage worldwide. The initial assignment of sequence reads to viral taxa was accomplished by BLAST searches. We performed two additional computational steps to confirm our conclusions regarding virus diversity in raw sewage. First, we subjected selected assembled sequences to genetic signature analysis (GSA), a manual sequence analysis procedure in which the sequence reads and open reading frames (ORFs), contained within the reads, were examined for the presence of eukaryotic and prokaryotic genetic signatures such as promoters, factor binding sites, polyadenylation and splice signals, and ribosome binding sites. GSA also included a close examination of the sequence alignments that led to the taxonomic assignment of each sequence. These steps led to the reassignment of some of the sequences to different taxa. The most striking example of a misassignment uncovered by GSA is that of non-A, non-B hepatitis virus. The large number of sequence reads related to this virus suggested that it was among the most abundant eukaryotic viruses present in raw sewage, a result confirmed by PCR (Fig. 5C). This virus was originally isolated from stool samples from hepatitis patients and thus potentially was of great interest (40, 41).

We assembled 794 reads that had at least 80% identity to non-A, non-B hepatitis virus (GenBank accession no. X53411) with phrap (http://www.phrap.org), using the default parameters. The assembly produced a 4,818-bp contig (named WW-nAnB). Initial alignments with the sequence deposited in GenBank under accession no. X53411 (X53411 sequence) showed that WW-nAnB assembled as a circle. After we edited the contig to put it in the same orientation as in the X53411 sequence, we aligned it to the X53411 sequence with BLASTN, and the resulting dot matrix plot is shown in Fig. 5A. At the 5′ end of the contig, there were two small insertions of 20 and 38 bp with respect to the X53411 sequence. There was also a 250-bp deletion in WW-nAnB with respect to the X53411 sequence at nucleotide position 1542. We identified homologs of the four ORFs in the X53411 sequence. We discovered several positions in the sequence of WW-nAnB that disrupted the reading frame of three ORFs compared to homologous ORFs in the X53411 sequence. Additionally, there was an ambiguous base in one position. To determine the correct nucleotide sequence of these positions, targeted regions of the genome were resequenced by Sanger sequencing of PCR products, and appropriate corrections were made to the WW-nAnB sequence. Sanger sequencing gave unambiguous resolution to the uncertainties in the original sequence, in particular correcting all the apparent frameshift errors, which brought the ORF structures of the X53411 sequence and WW-nAnB into agreement (Fig. 5B). We further confirmed the presence of non-A, non-B hepatitis virus in the virion preparations using specific PCR primers targeted to the open reading frame 4 (ORF4) sequence of the X53411 sequence. Forty-five cycles of PCR were performed on different virion preparations from five samples of raw sewage. The expected 373-bp PCR product appeared in all virion preparations (Fig. 5C). Sequencing of the PCR products revealed some nucleotide variation, suggesting the presence of different variants of non-A, non-B hepatitis virus in raw sewage. Also, the phylogenetic relationships among the sequences revealed that they are more similar to each other than to the X53411 sequence.

There are no reports on the properties of non-A, non-B hepatitis virus beyond the original report of the genomic sequence (41), and it has not been classified into any formal taxonomic group. Our attempts to identify some of the sequence signals typically found in a virus infecting eukaryotic hosts, such as promoters and poly(A) addition sequences, were not successful. However, we did find strong evidence of prokaryotic transcription and translation signals, including sigma70-like promoters and Shine-Dalgarno (SD) translation initiation sequences. We found convincing SD sequences appropriately positioned at the beginnings of three of the four ORFs annotated in the X53411 sequence. For the fourth ORF (ORF2), there is no SD sequence upstream from the AUG start codon annotated in the X53411 sequence. However, there is an excellent SD sequence upstream of that position, appropriately located for an initiation codon 90 bases upstream from the annotated start codon in the ORF2 reading frame in both genomes. Therefore, we suggest that this is the correct start site for translation of this gene. This initiation codon is AUG in the WW-

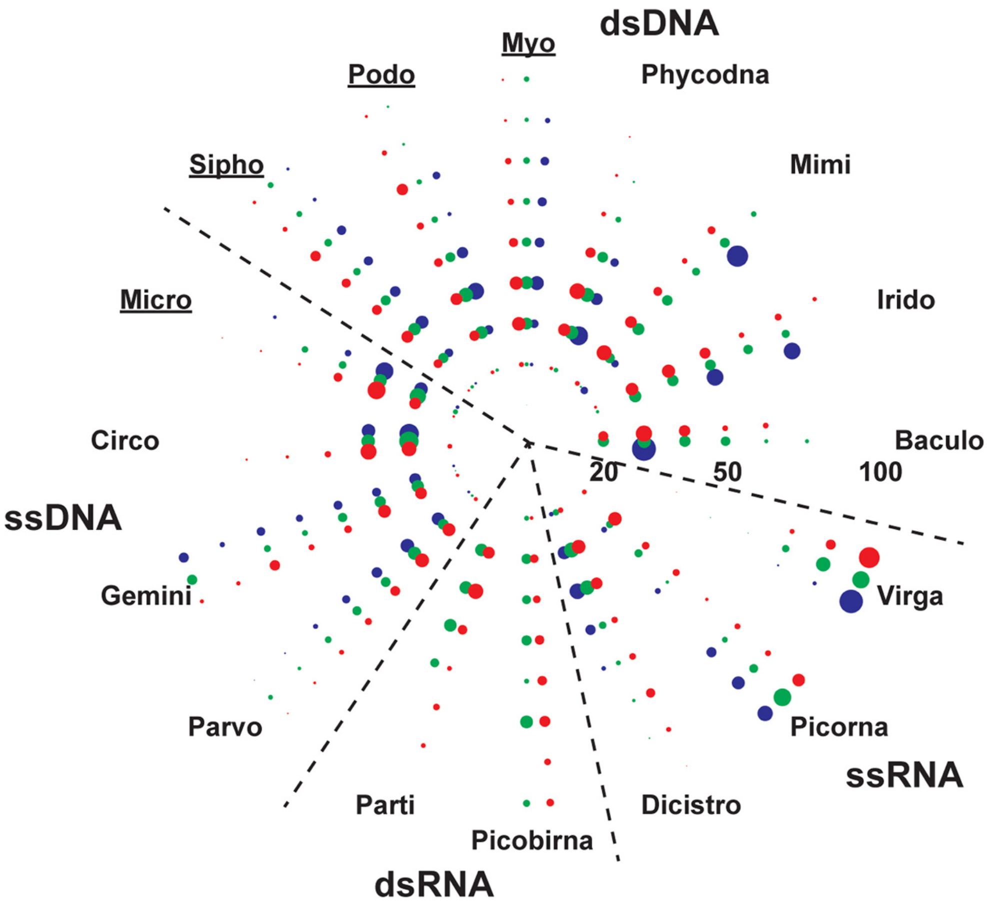

FIG 3 Most virus-related pyrosequencing reads found in raw sewage represent previously unknown viruses. Diversity plot of selected viral families (only the prefix of a family name is shown) are organized by genomic content (dsDNA, ssDNA, ssRNA, and dsRNA). The four phage families are underlined. The rings of the plot represent bins of increasing percent identity (20%, 50%, and 100% are marked for orientation) relative to the GenBank reference sequence as identified by the top BLAST hit. The area of each circle is proportional to the number of virus family reads in that location. The geographic locations from which the sequence reads were obtained are indicated by color: blue, Addis Ababa, Ethiopia; green, Barcelona, Spain; red, Pittsburgh, PA, USA.

nAnB sequence but GUG in the X53411 sequence. GUG start codons are found rather commonly in prokaryotic sequences but virtually never in eukaryotic sequences. In addition to the four large ORFs, we have identified four additional small putative genes located in the spaces between the larger ORFs, based on appropriately positioned SD sequences and good coding potential (Fig. 5D).

We probed the public databases with the predicted protein sequences from WW-nAnB, and the results are reported in the supplemental material. On the basis of the size of the genome, the sequence matches obtained, and other features of the sequence described in the supplemental material, we believe that WW-nAnB (and the non-A non-B hepatitis virus with GenBank accession no. X53411) are members of the *Inoviridae* family of bacteriophages. The *Inoviridae* family contains the filamentous phages, of which the best-characterized examples are the *Escherichia coli* phages f1, fd, and M13. Figure 5D compares the genome map of WW-nAnB to those of 3 well-characterized filamentous phages.

Deep sequencing of virion-associated nucleic acids suggests the presence of large numbers of uncharacterized viruses. Most of our analysis has focused on the 46,408 sequence reads that could be assigned to one of the existing 84 viral taxa. However, over 247,000 reads were binned as bacteria, and nearly 600,000 reads were not related to sequences in genomic databases (Fig. 1C). The bacterial sequences in the samples could represent bacteria that escaped the virion enrichment methods, gene transfer agents (33), or prophage genes (8, 15, 19, 32). Microscopic examination of the virion preparations used for deep sequencing did not reveal any bacterial contamination. Still, we cannot rule out the possibility that a small amount of bacterial DNA remains in the virion preparations. Furthermore, the amount of sequences binned as bacteria in our study is consistent with the results of several other metagenomic studies (1, 8, 12, 18, 19, 29). It is likely that these sequences either represent GTAs or bacterial genes present in bacteriophage transducing particles or they are in fact bacteriophage genes. Thus, novel bacteriophages are likely included among these bacterial sequences.

A majority of the high-quality sequence reads obtained in this study were binned as "unassigned" because they did not significantly match sequences present in the current databases. These sequences most likely represent uncharacterized viruses that are not related to or are very distantly related to the 3,000 or so known viruses. Examination of some of the assembled unassigned sequences revealed ORF patterns consistent with members of the

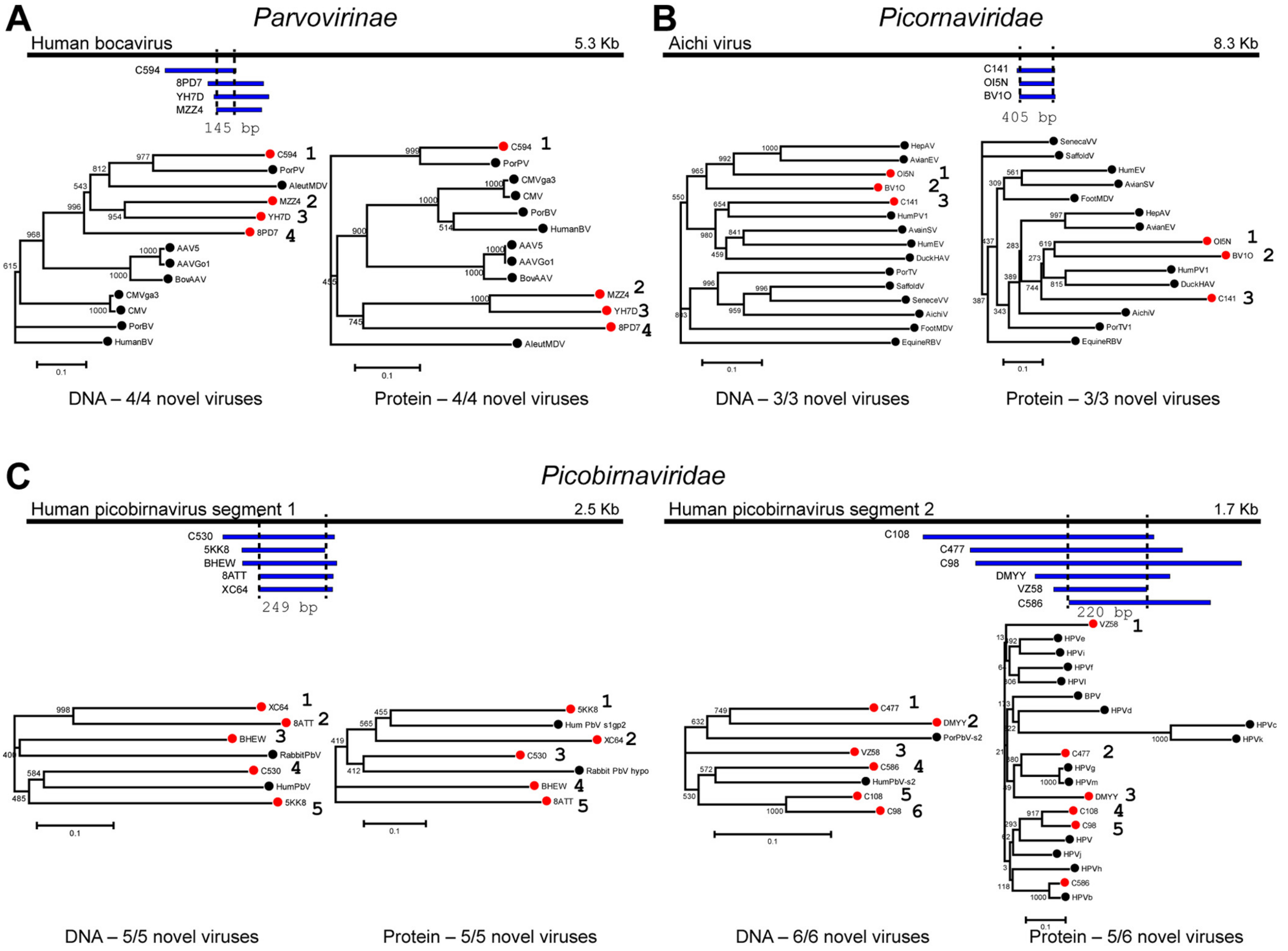

FIG 4 Novel virus analysis from selected virus families. (A to C) A selected set of novel assembled sequences from three different virus families that overlapped each other on a representative genome from each family was aligned with ClustalW2. The nucleotide alignment is shown graphically in the fragment recruitment plot (top) with vertical black broken lines marking the common alignment region against a selected reference genome from the virus family. Each assembled sequence was translated, and the resulting ORFs were aligned with ClustalW2. DNA and protein neighbor-joining (NJ) phylogenetic trees were constructed from homologous positions without any gaps. Metagenomic sequences (red circles) and GenBank sequences (black circles) are indicated. Metagenomic sequences that are labeled with a number represent different novel virus species in the raw sewage. (A) For the *Parvovirinae*, 4 novel assembled sequences were aligned with 9 selected reference *Parvovirinae* genomes, and the nonstructural (NS) gene from each genome was used for the protein alignment. (B) For the *Picornaviridae*, 3 novel assembled sequences were aligned with 12 selected reference *Picornaviridae* genomes, and the polyprotein from each genome was used for the protein alignment. Alignment is in the P-loop NTPase domain of the 2C ATPase mature peptide of the polyprotein. (C) For the *Picobirnaviridae*, for segment 1 (left), 5 novel assembled sequences were aligned with the 2 reference segment 1 sequences in GenBank (human and rabbit), and the segment 1 ORF from each genome was used for the protein alignment. For segment 2 (right), 6 novel assembled sequences were aligned with the 2 reference segment 2 sequences in GenBank (human and porcine) and 14 RdRp ORFs (13 human and 1 bovine) was used for the protein alignment. See Fig. S3 for the ORF alignments. Virus abbreviations: PorPV, porcine parvovirus; AleutMDV, aleutian mink disease virus; AAV5, adeno-associated virus 5; BovAAV, bovine AAV; CMV, canine minute virus; PorBV, porcine bocavirus; HepAV, hepatitis A virus; AvianEV, avian encephalomyelitis virus; HumPV1, human parechovirus 1; DuckHAV, duck hepatitis A virus; PorTV, porcine teschovirus; SaffoldV, Saffold virus; AichiV, Aichi virus; FootMDV, foot-and-mouth disease virus; EquineRBV, equine rhinitis B virus; SenecaVV, Seneca Valley virus; Rabbit PbV, rabbit picobirnavirus; HumPbV, human picobirnavirus; BPV, bovine picobirnavirus.

Microviridae and other bacteriophage taxa (data not shown). Furthermore, approximately 355,000 metagenomic reads did not assemble into multiread contigs, suggesting a high degree of sequence diversity. If we assume that all individual sequence reads binned as unassigned represent novel viruses, then novel viruses (596,146 + 43,381 = 639,527) outnumber those binned as known viruses (3,027) by a ratio of over 200:1. On the other hand, if none of the unassigned sequences represent novel viruses but rather are derived from other taxa (bacteria, etc.), then the ratio (43,381/3,027) of novel to known viral sequence reads is approximately 10:1. In any event, our data demonstrate that known viruses represent a small fraction of the viral universe.

Finally, we compared the high-quality sequence reads from our experiment with sequences detected in other metagenomic studies, including reclaimed wastewater (29), human feces (8, 11, 14), and three marine environments (1, 2, 19). Since several of the metagenomes consisted of individual reads, we used CD-HIT (using the same parameters as performed on the raw sewage metagenome) to remove duplicate reads. For this comparison, we performed a BLASTN search using the 897,647 high-quality raw

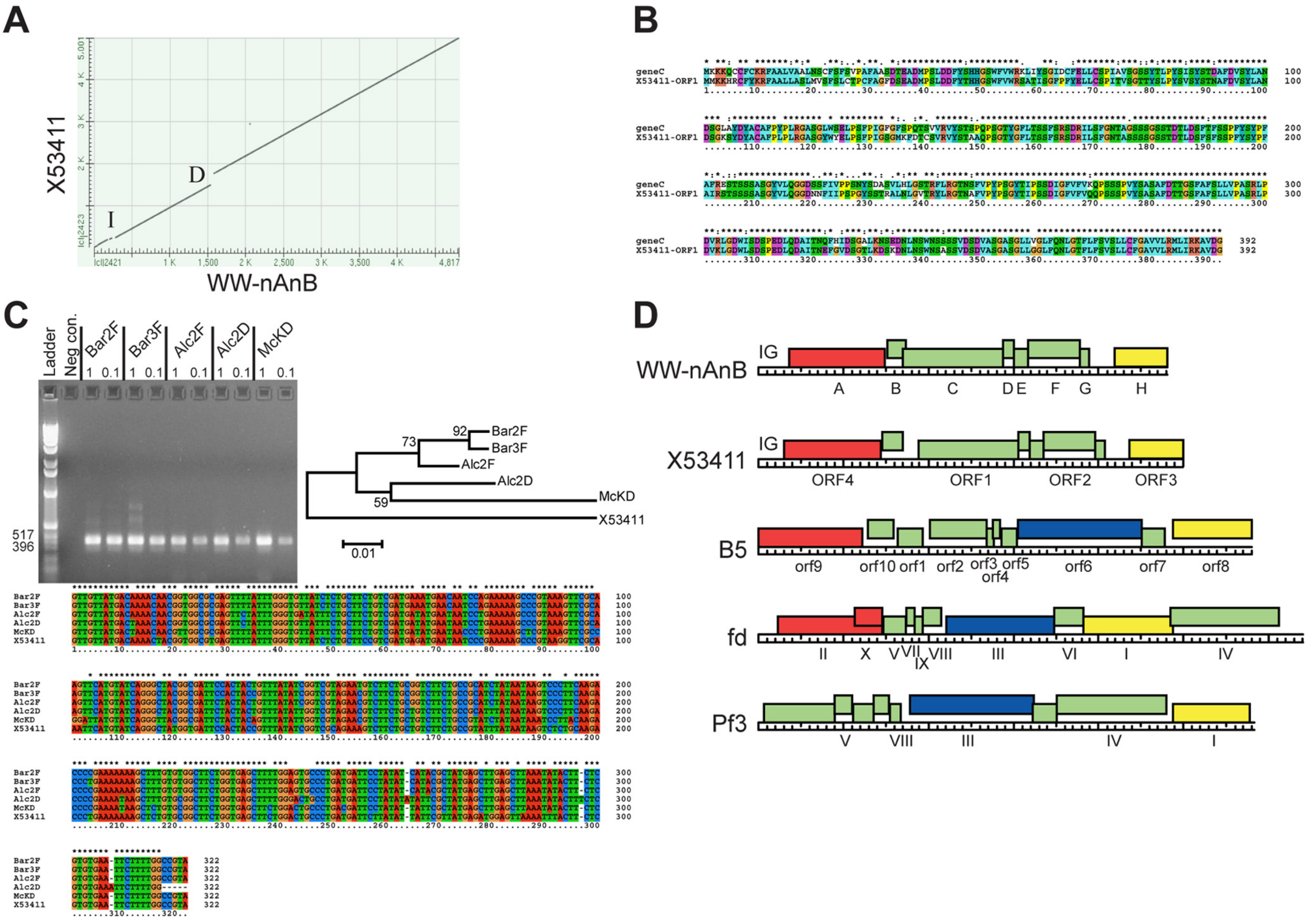

FIG 5 An assembled genome of non-A, non-B hepatitis virus from raw sewage shows that it belongs to the *Inoviridae* family. (A) BLASTN alignment of WW-nAnB and the non-A, non-B hepatitis virus with GenBank accession no. X53411 (X53411 sequence) is displayed as a dot matrix plot. The WW-nAnB sequence and the X53411 sequence run 5′ to 3′ on the *x* axis and *y* axis, respectively. The positions of the insertions (I) and deletion (D) are labeled. (B) Protein alignment of the X53411 ORF1 with the corrected gene C sequence from WW-nAnB. Identical amino acids (*), highly similar amino acids (:), and amino acids with low similarity (.) are indicated. (C) Forty-five cycles of PCR were performed with 0.1 and 1 μl of five different virion preparations from raw sewage. Shinola was used as a negative control (Neg. Con.). PCR products were visualized by EtBr on a 1.5% agarose gel. DNA ladder sizes are indicated in base pairs. The specific PCR product bands (373 bp) were excised and sequenced. The resulting nucleotide sequences were aligned (shown at the bottom of the panel), and a bootstrapped phylogenetic tree was generated based on the alignment (top right corner of panel). (D) WW-nAnB belongs to the *Inoviridae* family of bacteriophages. The genomic organization of WW-nAnB compared to non-A, non-B hepatitis virus (GenBank accession no. X53411), *Propionibacterium* phage phiB5 (B5) (GenBank accession no. AF428260), enterobacterial phage fd (GenBank accession no. J02451), and bacteriophage Pf3 (GenBank accession no. M19377) is shown. Unlabeled X53411 ORFs are homologous to the similarly located ORFs in WW-nAnB. DNA replication initiation proteins are shown in red, assembly proteins with an ATPase domain are shown in yellow, absorption proteins are shown in blue, and all other identified ORFs are shown in green. Each tick mark in the ruler below each genome represents 100 bp. IG, noncoding intergenic region.

sewage reads as the query sequences against each metagenome. We applied an E-value cutoff of 1e − 5 to score a significant match. We found that only a small number of sequences detected in each of these metagenomes were significantly related (see Table S5 in the supplemental material). The metagenome most closely related to raw sewage is the monozygotic twin feces metagenome (11). A total of 486,392 unique sequences were obtained in the twin study of which 40,594 (8.3%) showed a significant match to 17.3% (155,083) of the raw sewage sequence reads. Similarly, about 12.2% and 9.9% of the sequences we identified in raw sewage were similar to sequences from the human gut microbiome and reclaimed water, respectively. Other metagenomes harbored fewer viral sequences similar to those found in raw sewage. In total, these observations emphasize the vastness of viral diversity among different biomes.

DISCUSSION

The International Union for Conservation of Nature lists nearly 1.8 million species of living organisms on Earth. Each of these species is likely to harbor multiple types of viruses uniquely adapted to proliferate in the cellular environment they provide. However, only about 3,000 viruses have been identified thus far, suggesting that our knowledge of the viral universe is limited to a tiny fraction of the viruses that exist. Pioneering studies in viral metagenomics have led to advances in methods for capturing virus particles, sequencing their nucleic acids, and in the computational analysis of metagenomic data (21, 42). The results of metagenomic studies of the viromes present in oceans, lakes, human gut and stool samples, and reclaimed wastewater are consistent with the notion that large numbers of uncharacterized viruses exist in nature.

We performed a metagenomic survey of the viruses present in three samples of untreated wastewater obtained from three different continents. After steps to remove bacteria and other relatively large particles, virus particles were concentrated by organic flocculation and treated with DNase. Virion-associated nucleic acids were extracted and reverse transcribed so as to include both RNA and DNA genomes in the subsequent deep sequencing steps. Although each of the three samples was sequenced separately, we pooled these data for the purposes of this study. Computational methods were then used to assign each sequence read to specific taxa and to determine whether the sequence represented a previously characterized (known) virus recorded in the GenBank databases. This approach detected 234 known viruses. However, the vast majority of genomes present in the samples represent novel viruses. Representatives of 51 viral families were detected, making raw sewage the most diverse viral biome examined thus far.

Despite the large number of known and novel viruses detected, not all viruses present in the samples were detected by our methods. For example, JC virus (JCV), a human polyomavirus frequently associated with fecal/urine contamination was not detected by deep sequencing, although PCR experiments indicated its presence. This suggests that our data underestimate the number of viruses present in the samples. One reason viruses present in the sample could fail to be detected is that their abundance is below the resolution of sequencing. For example, JC polyomavirus is present in samples of raw sewage from Barcelona, Spain, at 18 genome copies (GC)/ml, but human adenovirus, which is represented by 20 sequencing reads in the raw sewage metagenome, is present at 10,100 GC/ml (Table 2). In this case, deeper sequencing of the sample will reveal additional viruses.

The probability of detecting a particular virus in a complex environmental sample such as untreated wastewater is directly proportional to the number of observable virions of species *i* in the sample (N_i^{obs}). This value changes in time according to the differential equation shown below, with the right hand side being a function of five time-dependent variables.

$$\frac{dN_i^{obs}}{dt} = \left(\underbrace{\phi_i}_{deposition} + \underbrace{\kappa_i}_{production} - \underbrace{\delta_i}_{decay}\right)\underbrace{\varepsilon_i}_{recovery}\quad\underbrace{\beta_i}_{detection}$$

First is the rate with which virus particles are deposited in the sample. In the case of raw sewage, virus particles enter the sample in the form of human and animal feces and urine, plant material from domestic and agricultural areas, as well as from insects and rodents found in the sewer system (ϕ_i). Second, new virus particles are created by the infection of host species growing in the sewage (κ_i). Raw sewage provides a rich environment for the growth of bacteria, rotifers, amoeba, and fungi, and as these organisms become infected, the resulting progeny viruses will be shed into the sample. The accumulation of virus particles in sewage via deposition and infection is balanced by the physical decay of virions (δ_i). All three of these parameters are dependent on time and thus will vary during different times of day, in different seasons, and in different climates. Finally, the probability of detection is a function of both the efficiency of virion recovery (ε_i) from the sample and the efficiency of detection (β_i). For example, the use of CsCl gradients to purify virions eliminates certain types of viruses either because they do not band in the selected density range or because they are disrupted by CsCl. Similarly, the methods used to isolate and amplify viral nucleic acids can eliminate or favor certain genome types. No one method efficiently recovers and detects all types of virions, and thus, a complete survey of viral diversity will require a combination of approaches.

A key step in metagenomic analysis is the assignment of individual sequence reads or assembled sequences to viral taxa. Each individual read or assembled sequence should represent the nucleic acid present in an individual virion, and thus, a single viral species. Generally, this taxon assignment is accomplished by a BLAST search with the E value being the arbiter of taxon assignment with most metagenomic studies using the top BLAST hit to identify and classify sequence reads. In this study, we divided the taxonomic classification of sequence reads into three steps. First, the broad binning of sequences into those related to viruses, bacteria, or other major taxa was based on BLAST scores. Second, known viruses were identified on the basis of nucleotide identity through the entire sequence read with a viral genome listed in the GenBank database. However, it is still possible that some novel viruses might be classified as a known virus. For example, bacteriophages exhibit high levels of horizontal gene transfer generating a mosaic of genome types (43–45). Since metagenomic studies seldom yield enough sequence data to assemble an entire genome, it is possible that some of the viruses classified as known are actually chimeras where only a portion of the genome matches the GenBank reference sequence. Finally, the remaining sequences representing potentially novel viruses were manually examined to confirm their taxonomic assignment. This manual analysis revealed numerous ambiguities and in some cases errors in taxon assignments. Some errors in taxon assignments resulted from misannotations of databases. In other cases, the correct viral taxon could not be ascertained because homologs of viral genes exist in multiple viral taxa.

We are using metagenomics to explore viral diversity in a number of different biomes. To begin these studies, we wanted to examine environments where viral concentrations and diversity are relatively high. In this regard, we hypothesize that the highest concentrations of viruses will be found where there is a high density of host species and that viral diversity will correspond to the biodiversity of host species. Urban sewage has been selected as a unique example of a matrix with high concentrations of highly diverse viruses. Urban sewage is a virus-rich matrix because humans excrete waste materials from the diverse food consumed, especially plants that are known to be very rich in viruses, and the bacterial and viral members of the human microbiota and common viral infections. The matrix we analyze includes the excreted virome plus the external input from insects, rodents, and other inhabitants of the urban sewerage system as well as bacteria growing in the wastewater. We have not attempted to measure the relative numbers of different viral species present in the sample. Nor have we sampled sewage in different seasons or in different climates or performed an extensive study of different geographic locations, all of which are likely to influence the dynamics of viral populations. These issues await future studies.

Finally, we point out that while untreated wastewater is a rich source of novel viruses, it is still a limited one. The diversity of host species that occupy this ecosystem is limited by its unique chemical composition. Earth is rich with many disparate biomes, each harboring a multitude of host species and their viruses. The exploration of the viral universe has only just begun.

MATERIALS AND METHODS

Sample collection sites. Untreated wastewater was obtained from three locations: (i) Pittsburgh, Pennsylvania, United States; (ii) Barcelona, Spain; and (iii) Addis Ababa, Ethiopia. The Pittsburgh wastewater treatment plant (WWTP) provides services to approximately 1 million people in the city and many surrounding communities. The Barcelona WWTP is located on the south coast of Spain. The Barcelona WWTP receives wastewater from six towns with an approximate total population of 172,000 inhabitants. The WWTP treats the raw wastewater from domestic origin as well as treated wastewater from industries. The Addis Ababa WWTP services a city that contains approximately 3 million inhabitants. Data on the volume of raw sewage that is treated by the WWTPs are not available.

Enrichment of virion populations from untreated wastewater. Untreated wastewater (5 liters) was collected from the WWTP in Pittsburgh, PA, in December 2009 and was stored at 4°C for 2 h prior to processing. Similarly, 10 liters of untreated wastewater was collected from the WWTP in Barcelona, Spain, in September 2008 and stored for 2 h at 4°C before processing. Two samples (10 liters each) were collected from the WWTP in Addis Ababa, Ethiopia, in June 2009 and processed on-site. In this case, the virion concentrates were stored frozen prior to viral nucleic acid isolation.

Virions were concentrated from wastewater samples by organic flocculation based on the procedure previously described (31). Briefly, 100 ml preflocculated skim milk solution (pH 3.5) was added to 10 liters acidified raw sewage (pH 3.5) and mixed for 8 h. Flocculants were allowed to settle and then centrifuged. The flocculated viral concentrate was resuspended in 15 ml phosphate buffer (1:2 [vol/vol] mixture of 0.2 M Na_2HPO_4 and 0.2 M NaH_2PO_4) and then eluted in 30 ml of 0.25 M glycine (pH 9.5) for 45 min at 4°C by slow agitation with vortexing. Suspended solids were separated by low-speed centrifugation at 7,500 × g for 30 min at 4°C, and the high pH of the supernatant was stabilized by adding 20 ml of 2× phosphate buffer. Virions present in the supernatant were concentrated by ultracentrifugation at 100,000 × *g* for 1 h at 4°C and resuspended in phosphate buffer.

Nucleic acid preparation and 454 sequencing. Aliquots (100 μl) of the virion concentrates from Addis Ababa, Ethiopia, Pittsburgh, Pennsylvania, and Barcelona, Spain, were treated with DNase to remove non-virion-associated DNA. One thousand units (10 μl) of DNase (catalog no. EN0523; Fermentas) and 10 μl of the supplied 10× reaction buffer were added to each sample and incubated at 37°C for 1 h. Virion nucleic acid was purified from the DNase-treated samples and 100 μl of untreated Barcelona virus preparation using the Qiagen DNeasy blood and tissue kit (catalog no. 69504) using the manufacturer's protocol (46) except that elution was performed with 30 μl of distilled H_2O (dH_2O).

To enable subsequent detection of both RNA and DNA viruses, total virion-associated nucleic acid from each sample was reverse transcribed and amplified as previously described (47, 48). Briefly, RNA templates were reverse transcribed using PrimerA (5′-GTTTCCCAGTCACGATANNNNNNNNN) containing a 17-nucleotide specific sequence followed by 9 random nucleotides for random priming. Sequenase (United States Biochemical) was used for second-strand cDNA synthesis and for random-primed amplification of DNA templates using PrimerA. Each sample was then subjected to 40 cycles of PCR amplification using PrimerB with a bar code (5′-XXXXXXGTTTCCCAGTCACGATA) for the Barcelona samples or PrimerB without the bar code for the Pittsburgh and Addis Ababa samples using the following program: 30 s at 94°C, 30 s at 40°C, 30 s at 50°C, and 60 s at 72°C. The bar code is a unique 6-nucleotide sequence (indicated by "X") at the 5′ end of PrimerB. PrimerB is complementary to the 17-nucleotide sequence that was incorporated by PrimerA. The amplified material was visualized on an agarose gel as a final quality control step and was sequenced at the Washington University Genome Sequencing Center on the 454 GS FLX titanium platform (454 Life Sciences) according to the manufacturer's instructions.

Sequence annotation. Raw sequence reads were trimmed to remove bar codes and PrimerB sequences. CD-HIT (49) was used to remove redundant sequences. Sequences were clustered on the basis of 95% identity over 95% sequence length, and the longest sequence from each cluster was picked as the representative sequence. Then, unique sequences were masked by RepeatMasker (http://www.repeatmasker.org). If a sequence did not contain a stretch of at least 50 consecutive non-"N" nucleotides or if greater than 40% of the total length of the sequence is masked, it was removed from further analysis (i.e., "filtered"). These preprocessing steps resulted in 897,647 high-quality sequences which were sequentially compared against (i) the human genome using BLASTN; (ii) GenBank nt database using BLASTN; (iii) GenBank nr database using BLASTX; and (iv) the NCBI viral genome database (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/) using TBLASTX. The nt and nr databases were downloaded on 29 May 2009, and the viral genome database was downloaded on 12 August 2010. Minimal E-value cutoffs of 1e − 10 for BLASTN and 1e − 5 for BLASTX or TBLASTX were applied. Sequences were phylotyped as human, mouse, fungal, bacterial, phage, viral, or other based on the identity of the top BLAST hit. Sequences without any significant hit to any of the databases were placed in the "unassigned" category. All virus and phage sequences were further classified into families using the taxonomic information from the top BLAST hit.

A second annotation analysis (Bar-v1) was performed with the Barcelona raw sequence reads only. The reads were trimmed to remove any bar code and PrimerB sequences. CD-HIT was used to remove redundant sequences. Sequences were clustered on the basis of 98% identity over 98% sequence length, and the longest sequence from each cluster was picked as the representative sequence. Then, unique sequences were masked using RepeatMasker and processed as described above to generate a high-quality set of reads. The high-quality Barcelona sequences (n = 680,295) were sequentially compared against (i) the human genome using BLASTN, (ii) GenBank nt database using BLASTN and TBLASTX, and (iii) the NCBI viral genome database using TBLASTX. Minimal E-value cutoffs of 1e − 10 for BLASTN and 1e − 5 for TBLASTX were applied. Sequences were phylotyped and classified as described above.

Sequence assembly. Using the high-quality Pittsburgh, Addis Ababa, and Barcelona reads (from Bar-v1 annotation analysis), sequences identified as eukaryotic viruses regardless of the source of isolation were separately assembled into contigs using phrap (version 1.090518; http://www.phrap.org) at 95% nucleotide identity by using the command line option "-penalty -19." The phrap singlets and contig files were merged to create an assembled set of virus sequences (n = 2,782). The assembled sequences were sequentially annotated by (i) BLASTN and then by TBLASTX versus the GenBank nt database and (ii) TBLASTX against the viral genome database using an E-value cutoff of 1e − 5. Sequences with no significant hit were classified as "unassigned." Sequences were binned into families using the taxonomic information from the top BLAST hit.

A full assembly of the 897,647 high-quality Pittsburgh, Addis Ababa, and Barcelona reads and quality scores was done with phrap at 95% nucleotide identity. The phrap singlets and contig files were merged to create a set of assembled sequences (n = 476,960).

Sequence alignments. Nucleotide and protein sequences were aligned with ClustalW2 using default parameters. Bootstrap neighbor-joining (NJ) trees (1,000 iterations) were constructed using homologous positions that do not contain any gaps.

Electron microscopy. Samples were observed with a transmission electron microscope Tecnai SPIRIT (FEI Company, Eindhoven, The Netherlands) working at an acceleration voltage of 120 kV. Images were acquired with a MegaviewIII camera and digitized with the iTEM program, both from Soft Imaging System (SIS).

Wastewater non-A non-B hepatitis virus analysis. The Pittsburgh, Addis Ababa, and Barcelona reads (from Bar-v1 annotation analysis) that had at least 80% identity to non-A, non-B hepatitis virus (n = 794) were assembled using phrap with default parameters. Virions were purified from five different samples of raw sewage. PCR was performed with 0.1

and 1 μl of virion preparations using GoTaq (Promega) under the following conditions: initial denaturation, 5 min at 94°C; 45 cycles, with 1 cycle consisting of 1 min at 94°C, 1 min at 54°C, and 75 s at 72°C; a final extension step of 7 min at 72°C. The primers (forward [5′-GATGCAGGAAGGTCACGAAT] and reverse [5′-ACGGCCAAAAGAATTCACAC]) were designed to ORF4 of non-A, non-B hepatitis virus (GenBank accession no. X53411). PCR products were resolved on a 1.5% agarose gel and stained with ethidium bromide. PCR bands were excised and sequenced using the forward primer. Sequences were aligned with ClustalW2 and a bootstrapped NJ tree was constructed using MEGA4.

Molecular detection of viruses in wastewater by PCR. Extractions of viral nucleic acids from the Addis Ababa and Barcelona samples used in the present metagenomic study were analyzed to detect classical and emerging viruses (Table 2) by nested PCR (nPCR) and quantitative PCR (qPCR) TaqMan assays. The viruses analyzed were human strains of hepatitis E viruses (HEV), hepatitis A (HAV), klassevirus I (KV) (37), asfarvirus-like virus (ASFLV) (50), human adenoviruses (HAdV), and JC polyomavirus. The protocols used are based on previous studies (22, 51–53; B. Calgua et al., submitted for publication).

For the detection of HPyV6 polyomavirus and rat HELV (see Table S3 in the supplemental material), urban sewage samples were collected in Barcelona, Spain. Viruses from 42 ml of each untreated wastewater sample were concentrated in 100 μl of PBS by applying a virus concentration procedure based on ultracentrifugation and elution with glycine-alkaline buffer as described previously (36). Nucleic acids from the viral concentrates were extracted using the QIAamp viral RNA minikit (catalog no. 522906; Qiagen). Nested primers for the VP1 region of HPyV6 were designed for nested PCR (nPCR) assays based on the NCBI reference sequence with accession no. NC_014406. For the detection of rat HELV, a nested set of primers for the ORF1 region was designed on the basis of the sequence obtained in the present metagenomic study (6AIF). For reverse transcription, a Qiagen OneStep RT-PCR kit (catalog no. 210212) was used according to the manufacturer's instructions. The first and second round of enzymatic amplification for both viruses (DNA/RNA) were performed as follows. In the first round of enzymatic amplification, 10 μl of the undiluted and a 10-fold dilution of the extracted nucleic acids was analyzed. The amplification mixture (40 μl) contained 1× PCR buffer, 1.5 mM $MgCl_2$, 250 μM each deoxynucleoside triphosphate (dNTP), 0.5 μM of each specific primer for each virus, and 4 U of TaqGold DNA polymerase (Applied Biosystems). In the second round of enzymatic amplification, 2 μl of the product obtained in the first round was added to 48 μl of amplification mix, containing a set of specific primers for each virus and the same reagent composition described above. The PCR conditions for the first and second rounds were as follows: 10 min at 95°C; 30 cycles, with 1 cycle consisting of 60 s at 94°C, 60 s at 52°C for HPyV6 or 60 s at 56°C for rat HELV, and 60 s at 72°C; a final extension step of 7 min at 72°C.

Accession numbers. The sequence of WW-nAnB was submitted to GenBank (JN402401), and the raw sewage metagenome was deposited in the Sequence Read Archive (SRA040148).

ACKNOWLEDGMENTS

We thank Alcosan for proving samples of untreated wastewater from Pittsburgh, PA. We thank Justin Brown and W. Mark Pipas for providing Shinola.

We are grateful to the University of Pittsburgh School of Arts and Sciences for providing seed funds for this project.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00180-11/-/DCSupplemental.

Text S1, DOCX file, 0.1 MB.
Figure S1, PDF file, 0.3 MB.
Figure S2, PDF file, 0.4 MB.
Figure S3, PDF file, 0.4 MB.
Table S1, XLSX file, 0.1 MB.
Table S2, XLSX file, 0.1 MB.
Table S3, XLSX file, 0.1 MB.
Table S4, XLSX file, 8.6 MB.
Table S5, XLSX file, 0.1 MB.
Table S6, XLSX file, 0.1 MB.

REFERENCES

1. **Angly FE, et al.** 2006. The marine viromes of four oceanic regions. PLoS Biol. **4**:e368.
2. **Breitbart M, et al.** 2002. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. U. S. A. **99**:14250–14255.
3. **Culley AI, Lang AS, Suttle CA.** 2006. Metagenomic analysis of coastal RNA virus communities. Science **312**:1795–1798.
4. **Monier A, Claverie JM, Ogata H.** 2008. Taxonomic distribution of large DNA viruses in the sea. Genome Biol. **9**:R106.
5. **Rusch DB, et al.** 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. **5**:e77.
6. **Williamson SJ, et al.** 2008. The Sorcerer II Global Ocean Sampling expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS One **3**:e1456.
7. **López-Bueno A, et al.** 2009. High diversity of the viral community from an Antarctic lake. Science **326**:858–861.
8. **Breitbart M, et al.** 2003. Metagenomic analyses of an uncultured viral community from human feces. J. Bacteriol. **185**:6220–6223.
9. **Finkbeiner SR, et al.** 2008. Metagenomic analysis of human diarrhea: viral detection and discovery. PLoS Pathog. **4**:e1000011.
10. **Holtz LR, Finkbeiner SR, Kirkwood CD, Wang D.** 2008. Identification of a novel picornavirus related to cosaviruses in a child with acute diarrhea. Virol. J. **5**:159.
11. **Reyes A, et al.** 2010. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature **466**:334–338.
12. **Victoria JG, et al.** 2009. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. J. Virol. **83**:4642–4651.
13. **Zhang T, et al.** 2006. RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol. **4**:e3.
14. **Kurokawa K, et al.** 2007. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Res. **14**:169–181.
15. **Bench SR, et al.** 2007. Metagenomic characterization of Chesapeake Bay virioplankton. Appl. Environ. Microbiol. **73**:7629–7641.
16. **Djikeng A, Kuzmickas R, Anderson NG, Spiro DJ.** 2009. Metagenomic analysis of RNA viruses in a fresh water lake. PLoS One **4**:e7264.
17. **Li L, et al.** 2010. Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses. J. Virol. **84**:6955–6965.
18. **Schoenfeld T, et al.** 2008. Assembly of viral metagenomes from Yellowstone hot springs. Appl. Environ. Microbiol. **74**:4164–4174.
19. **Vega Thurber RL, et al.** 2008. Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral *Porites compressa*. Proc. Natl. Acad. Sci. U. S. A. **105**:18413–18418.
20. **Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA.** 2005. Virus taxonomy: classification and nomenclature of viruses. Elsevier Academic, San Diego, CA.
21. **Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F.** 2009. Laboratory procedures to generate viral metagenomes. Nat. Protoc. **4**:470–483.
22. **Pina S, et al.** 2001. Genetic analysis of hepatitis A virus strains recovered from the environment and from patients with acute hepatitis. J. Gen. Virol. **82**:2955–2963.
23. **Rodriguez-Manzano J, et al.** 2010. Analysis of the evolution in the circulation of HAV and HEV in eastern Spain by testing urban sewage samples. J. Water Health **8**:346–354.
24. **Puig M, et al.** 1994. Detection of adenoviruses and enteroviruses in polluted waters by nested PCR amplification. Appl. Environ. Microbiol. **60**:2963–2970.
25. **Bofill-Mas S, et al.** 2006. Quantification and stability of human adenoviruses and polyomavirus JCPyV in wastewater matrices. Appl. Environ. Microbiol. **72**:7894–7896.
26. **Bofill-Mas S, et al.** 2010. Newly described human polyomaviruses Merkel cell, KI and WU are present in urban sewage and may represent potential environmental contaminants. Virol. J. **7**:141.
27. **Formiga-Cruz M, et al.** 2002. Distribution of human virus contamination

in shellfish from different growing areas in Greece, Spain, Sweden, and the United Kingdom. Appl. Environ. Microbiol. **68**:5990–5998.
28. **Blinkova O, et al.** 2009. Frequent detection of highly diverse variants of cardiovirus, cosavirus, bocavirus, and circovirus in sewage samples collected in the United States. J. Clin. Microbiol. **47**:3507–3513.
29. **Rosario K, Nilsson C, Lim YW, Ruan Y, Breitbart M.** 2009. Metagenomic analysis of viruses in reclaimed water. Environ. Microbiol. **11**: 2806–2820.
30. **Symonds EM, Griffin DW, Breitbart M.** 2009. Eukaryotic viruses in wastewater samples from the United States. Appl. Environ. Microbiol. **75**:1402–1409.
31. **Calgua B, et al.** 2008. Development and application of a one-step low cost procedure to concentrate viruses from seawater samples. J. Virol. Methods **153**:79–83.
32. **Kristensen DM, Mushegian AR, Dolja VV, Koonin EV.** 2010. New dimensions of the virus world discovered through metagenomics. Trends Microbiol. **18**:11–19.
33. **Casjens S.** 2003. Prophages and bacterial genomics: what have we learned so far? Mol. Microbiol. **49**:277–300.
34. **Johne R, et al.** 2010. Novel hepatitis E virus genotype in Norway rats, Germany. Emerg. Infect. Dis. **16**:1452–1455.
35. **Albinana-Gimenez N, et al.** 2009. Analysis of adenoviruses and polyomaviruses quantified by qPCR as indicators of water quality in source and drinking-water treatment plants. Water Res. **43**:2011–2019.
36. **Pina S, Puig M, Lucena F, Jofre J, Girones R.** 1998. Viral pollution in the environment and in shellfish: human adenovirus detection by PCR as an index of human viruses. Appl. Environ. Microbiol. **64**:3376–3382.
37. **Holtz LR, et al.** 2009. Klassevirus 1, a previously undescribed member of the family Picornaviridae, is globally widespread. Virol. J. **6**:86.
38. **Schowalter RM, Pastrana DV, Pumphrey KA, Moyer AL, Buck CB.** 2010. Merkel cell polyomavirus and two previously unknown polyomaviruses are chronically shed from human skin. Cell Host Microbe **7**:509–515.
39. **Li L, et al.** 2010. Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. J. Virol. **84**: 1674–1682.
40. **Seelig R, et al.** 1988. Hepatitis non-A, non-B-associated substance in stool from patients with posttransfusion and sporadic hepatitis. Immun. Infekt. **16**:85–90. (In German.)
41. **Burckhardt J, Seelig R, Calvo-Riera F, Seelig HP.** 1988. A hepatitis non-A, non-B-associated substance in the feces—identification and cloning of a partially double-stranded circular DNA. Immun. Infekt. **16**:91–96. (In German.)
42. **Wooley JC, Godzik A, Friedberg I.** 2010. A primer on metagenomics. PLoS Comput. Biol. **6**:e1000667.
43. **Lawrence JG, Hatfull GF, Hendrix RW.** 2002. Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. J. Bacteriol. **184**:4891–4905.
44. **Hendrix RW, Hatfull GF, Smith MC.** 2003. Bacteriophages with tails: chasing their origins and evolution. Res. Microbiol. **154**:253–257.
45. **Juhala RJ, et al.** 2000. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. J. Mol. Biol. **299**:27–51.
46. **Qiagen.** 2006. DNeasy blood & tissue handbook, p. 25–27. Qiagen, Hilden, Germany.
47. **Wang D, et al.** 2002. Microarray-based detection and genotyping of viral pathogens. Proc. Natl. Acad. Sci. U. S. A. **99**:15687–15692.
48. **Wang D, et al.** 2003. Viral discovery and sequence recovery using DNA microarrays. PLoS Biol. **1**:E2.
49. **Li W, Godzik A.** 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22**: 1658–1659.
50. **Loh J, et al.** 2009. Detection of novel sequences related to African swine fever virus in human serum and sewage. J. Virol. **83**:13019–13025.
51. **Erker JC, Desai SM, Mushahwar IK.** 1999. Rapid detection of hepatitis E virus RNA by reverse transcription-polymerase chain reaction using universal oligonucleotide primers. J. Virol. Methods **81**:109–113.
52. **Hernroth BE, Conden-Hansson AC, Rehnstam-Holm AS, Girones R, Allard AK.** 2002. Environmental factors influencing human viral pathogens and their potential indicator organisms in the blue mussel, *Mytilus edulis*: the first Scandinavian report. Appl. Environ. Microbiol. **68**: 4523–4533.
53. **Pal A, Sirota L, Maudru T, Peden K, Lewis AM, Jr..** 2006. Real-time, quantitative PCR assays for the detection of virus-specific DNA in samples with mixed populations of polyomaviruses. J. Virol. Methods **135**:32–42.