High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage

JVI Journals.ASM.org

High Variety of Known and New RNA and DNA Viruses of Diverse Origins in Untreated Sewage

Terry Fei Fan Ng,[a,b] Rachel Marine,[c] Chunlin Wang,[d] Peter Simmonds,[e] Beatrix Kapusinszky,[a,b] Ladaporn Bodhidatta,[f] Bamidele Soji Oderinde,[g] K. Eric Wommack,[c] and Eric Delwart[a,b]

Blood Systems Research Institute, San Francisco, California, USA[a]; Department of Laboratory Medicine, University of California, San Francisco, California, USA[b]; Departments of Biological Sciences and Plant & Soil Sciences, Delaware Biotechnology Institute, University of Delaware, Newark, Delaware, USA[c]; Stanford Genome Technology Center, Stanford University, Stanford, California, USA[d]; Centre for Immunity, Infection and Evolution, University of Edinburgh, Edinburgh, United Kingdom[e]; Department of Enteric Diseases, Armed Forces Research Institute of Medical Sciences, Bangkok, Thailand[f]; and WHO National Polio Laboratory, University of Maiduguri Teaching Hospital, Borno State, Nigeria[g]

Deep sequencing of untreated sewage provides an opportunity to monitor enteric infections in large populations and for high-throughput viral discovery. A metagenomics analysis of purified viral particles in untreated sewage from the United States (San Francisco, CA), Nigeria (Maiduguri), Thailand (Bangkok), and Nepal (Kathmandu) revealed sequences related to 29 eukaryotic viral families infecting vertebrates, invertebrates, and plants (BLASTx E score, $<10^{-4}$), including known pathogens (>90% protein identities) in numerous viral families infecting humans (*Adenoviridae*, *Astroviridae*, *Caliciviridae*, *Hepeviridae*, *Parvoviridae*, *Picornaviridae*, *Picobirnaviridae*, and *Reoviridae*), plants (*Alphaflexiviridae*, *Betaflexiviridae*, *Partitiviridae*, *Sobemovirus*, *Secoviridae*, *Tombusviridae*, *Tymoviridae*, *Virgaviridae*), and insects (*Dicistroviridae*, *Nodaviridae*, and *Parvoviridae*). The full and partial genomes of a novel kobuvirus, salivirus, and sapovirus are described. A novel astrovirus (casa astrovirus) basal to those infecting mammals and birds, potentially representing a third astrovirus genus, was partially characterized. Potential new genera and families of viruses distantly related to members of the single-stranded RNA picorna-like virus superfamily were genetically characterized and named *Picalivirus*, *Secalivirus*, *Hepelivirus*, *Nedicistrovirus*, *Cadicistrovirus*, and *Niflavirus*. Phylogenetic analysis placed these highly divergent genomes near the root of the picorna-like virus superfamily, with possible vertebrate, plant, or arthropod hosts inferred from nucleotide composition analysis. Circular DNA genomes distantly related to the plant-infecting *Geminiviridae* family were named *Baminivirus*, *Nimivirus*, and *Niminivirus*. These results highlight the utility of analyzing sewage to monitor shedding of viral pathogens and the high viral diversity found in this common pollutant and provide genetic information to facilitate future studies of these newly characterized viruses.

The characterization of previously unknown viral genomes has accelerated following the introduction of high-throughput sequencing technologies. We reasoned that untreated sewage would provide a rich source of both known and previously uncharacterized viruses with which to expand the reach of the existing viral taxonomy and also provide candidate viral genomes for future disease association studies, particularly of idiopathic human enteric diseases. Enteric infections cause more than 2 million deaths each year (93), primarily among infants in developing countries (59, 93). In the United States, greater than 40% of cases of diarrhea are caused by unknown agents (33). Viruses in sewage, which reflect in part ongoing enteric infections in the sampled human population (79), may therefore include still unknown human pathogens. Studies of sewage have frequently been used to monitor known viral pathogens, most frequently, polioviruses (6–9, 36, 88, 94). Viruses in untreated sewage may disseminate through rivers, plant irrigation, and fish and shellfish production and affect downstream human, animal, and plant health (27).

Sewage includes feces and urine from humans as well as from domesticated and wild animals, such as pets and rodents. Numerous bacteria, archaea, unicellular eukaryotes, plants, and insects and their associated viruses are also expected within sewage systems. The large number of potential host species contributing to sewage provides an opportunity to sample the viral diversity infecting cellular organisms from all kingdoms of life.

Recent studies have shown that besides human viruses, untreated sewage also contains a wide diversity of other animal, plant, insect, and bacterial viruses (11, 76, 82, 89). We performed a metagenomic deep-sequencing analysis of viral particles in sewage from four countries on three continents. We also acquired the full or partial genomes of several selected viruses and phylogenetically compared them to previously known viruses. We report here on the wide diversity of eukaryotic viruses found in sewage and on multiple novel viral genomes, significantly increasing the known diversity of viruses, especially those related to the single-stranded RNA (ssRNA) picorna-like virus superfamily.

MATERIALS AND METHODS

Sample collection. Untreated sewage waters were collected from a polluted canal in Khlong Maha Nak, Yommarat, Pom Prap Sattru Phai, Bangkok, Thailand, on 18 June 2009; from a junction of a main sewage line and river at the Kalimati Bridge in Kathmandu, Nepal, on 14 August 2009; from the city of Maiduguri (Gomboru Ward) in Nigeria on 15 April

Received 5 April 2012 Accepted 22 August 2012
Published ahead of print 29 August 2012
Address correspondence to Eric Delwart, delwarte@medicine.ucsf.edu.
Supplemental material for this article may be found at http://jvi.asm.org/.

doi:10.1128/JVI.00869-12

2008; and from the southeast water pollution control plant in San Francisco, CA, on 15 May 2009.

Viral particle purification and sequence-independent nucleic acid amplifications. Following collection, sewage samples were shipped on dry ice and stored at −80°C. Six hundred to 800 ml from each location was thawed in the dark at 4°C over 72 h. Sewage samples were centrifuged at 4,000 × *g* for 15 min at 4°C followed by a second centrifugation at 10,000 × *g* for 15 min in a Beckman SW55Ti rotor at 4°C to remove large particulates and bacteria. The resulting supernatant was subjected to 0.22-μm-pore-size tangential-flow filtration, and the collected viral fraction was concentrated to 12 ml using a 30-kDa tangential-flow filter (PXGVPPC50 and PXC030C50; Millipore). Viruses were enumerated using SYBR gold epifluorescence microscopy (73). Six milliliters of the viral concentrate was further concentrated by sucrose cushion centrifugation (38%, wt/vol) in a Beckman SW28 rotor. Pelleted viruses were resuspended in 500 μl SM buffer (100 mM NaCl, 8 mM $MgSO4 \cdot 7H_2O$, 50 mM Tris-HCl [pH 7.5]). The resuspended virus particles were treated with a cocktail of DNases (Turbo DNase from Ambion, Baseline-ZERO from Epicentre, and Benzonase from Novagen) and RNase (Fermentas) to digest unprotected nucleic acids (10, 69, 72, 92). Viral nucleic acids were then extracted using QIAamp spin columns (Qiagen).

Viral RNA and DNA were used to construct libraries by random PCR amplification as previously described (92). Twelve different primers (an arbitrarily designed 20-base oligonucleotide followed by a randomized octamer sequence at the 3′ end) were used in separate reverse transcription (RT) reactions for each sample (Invitrogen). After denaturation, the cDNA was then subjected to a round of DNA synthesis using Klenow fragment polymerase (New England BioLabs), followed by PCR amplification using a primer consisting of only the 20-base fixed portion of the random primer. The amplification from each primer was then quantified and pooled in an equal amount, and DNA libraries were prepared and sequenced using a 454 GS FLX titanium platform.

Bioinformatics. Pyrosequencing reads were trimmed and parsed according to their primer tags. For each sample, sequences sharing more than 95% nucleotide sequence identities over 35 bp were assembled into contigs. Metagenomic reads, assembled contigs, and nonassembled singlets were compared to the sequences in the GenBank nonredundant nucleotide and protein sequence databases using BLASTn and BLASTx and an E-value cutoff of 10^{-4}. On the basis of the best BLAST result, sequences were classified into their likely taxonomic groups of origin. Sequence identity distribution analysis was performed by parsing the BLASTx result for the taxonomy, host, and percent protein identities of the local pairwise alignment provided by BLASTx.

Acquisition of novel virus genomes. Total nucleic acid was extracted from the purified virus preparation by using QIAamp spin columns. In order to amplify the extremities of the viral genomes, both 3′ and 5′ rapid amplification of cDNA ends (RACE) amplification kits (Invitrogen and Clontech) were used according to the manufacturer's instructions. In short, three primers were designed for each virus genome. For 3′ RACE, the first primer was used in a reverse transcription reaction to produce cDNA from the poly(A) 3′ ends. For 5′ RACE, two additional steps were performed: cDNA launched from a primer complementary to the viral sequence was purified using spin columns, and poly(C) was added to the cDNA 5′ end using terminal deoxynucleotidyltransferase and dCTP. For the first round of PCR amplification, cDNA was amplified using a virus-specific primer and a RACE-specific primer ending with poly(G) or poly(A). The second round of PCR was performed using the second downstream primer, together with another RACE-specific primer without the poly(G) or poly(A), supplied by the RACE kit. Amplicons were analyzed using gel electrophoresis and Sanger sequenced by primer walking until the extremities were obtained. The genome sequence fragments originally derived from 454 pyrosequence data were confirmed using Sanger sequencing.

To complete the circular DNA genomes, we used rolling-circle and inverse PCR amplification as described previously (68, 69). Viral nucleic acids were first nonspecifically amplified using random hexamers by rolling-circle amplification (Genomiphi; GE Healthcare) and then further amplified by specific primers designed to amplify the whole circular genome using inverse PCR (70, 71). PCR amplicons were then Sanger sequenced.

Phylogenetic analysis. Phylogenetic analyses were performed using novel virus sequences, their closest BLAST hits, and other type species from related viral genera or families. Due to the divergent nature of the virus genomes, all sequence alignments and phylogenetic analyses were performed on the translated amino acid sequences. Multiple-amino-acid-sequence alignments were performed using the MUSCLE (version 3.8) (19) and MAFFT (50) programs, and the best overall alignments were selected for further analysis. Pairwise distance analyses and conserved amino acid analyses were performed over the alignments produced. Maximum likelihood (ML) trees were generated from translated protein sequences using RAxML and PROTGAMMA and Dayhoff similarity matrix parameters (87). These specify a general time-reversible model with a gamma distribution for rates over sites. All model parameters were estimated by RAxML. ML trees were run with 100 bootstrap replications; branches with 60% or greater bootstrap support are labeled. Resulting trees were examined for consistency with published phylogenetic trees. For trees without an outgroup, midpoint rooting was conducted using the program MEGA (90). Sliding-window analyses were performed using the same alignment, with protein identities between translated sequential fragments of 32 in-frame codons, incrementing by 8 codons, being calculated (66, 85).

NCA. Nucleotide composition analysis (NCA) was performed as previously described (46, 84) using sequences infecting mammals ($n = 117$), insects ($n = 63$), and plants ($n = 167$) for classification. The frequencies of each mononucleotide and dinucleotide were used for discriminant analysis to maximize discrimination between control sequences; these canonical factors were then used to infer the host origin of the RNA virus sequences obtained in the current study.

Nucleotide sequence accession numbers. All sequenced genomes were deposited in GenBank under accession numbers JQ898331 to JQ898345. Pyrosequences were deposited in GenBank under short-read archive accession number SRA054852. The GenBank accession numbers of all viral taxa used in the phylogenetic trees in Fig. 3 to 9 are listed (see Table S1 in the supplemental data).

RESULTS

Untreated raw sewage was collected from four locations: San Francisco (United States), Bangkok (Thailand), Kathmandu (Nepal), and Maiduguri (Nigeria) (see Materials and Methods). Epifluorescence microscopy showed 1.46×10^{10}, 9.26×10^{8}, 4×10^{9}, and 6.01×10^{9} virus particles per ml, respectively. Viral particles were purified using tangential flow filtration, sucrose cushion centrifugation, and nuclease treatment (see Materials and Methods). Nucleic acids were then extracted and amplified by random RT-PCR using primers with degenerate 3′ ends (92). The amplified nucleic acids were prepared into DNA libraries for pyrosequencing in two 454 FLX runs. A total of 304,498, 392,638, 217,383, and 161,798 sequence reads were generated from the U.S., Thai, Nepalese, and Nigerian libraries, respectively (Fig. 1). Overlapping sequence reads were assembled into contigs, and both contigs and singlets of >100 bp were analyzed for translated protein similarities to sequences in the nonredundant database of GenBank using the best BLASTx with a threshold E score of 10^{-4} (see Materials and Methods). The resulting classification (see Fig. S1 and S2 in the supplemental material) indicates that sequences other than those of viruses remained in our virus concentrates. Recognizable phage sequences were found at a frequency of 13.5% relative to the 11% of sequences showing similarity to eukaryotic viruses. A large

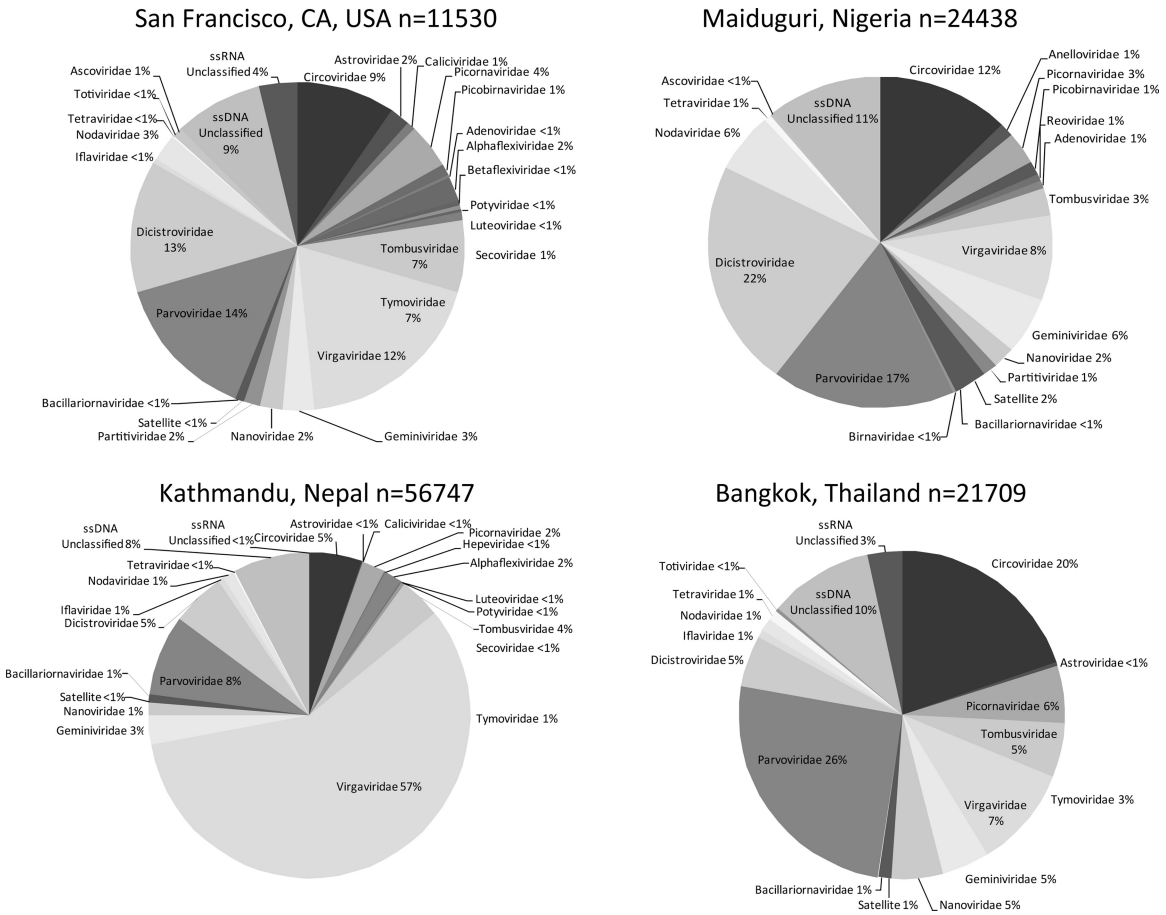

FIG 1 Taxonomic distributions of the eukaryotic virus-related sequences from four sewage samples. Assembled sequences were compared to the sequences in the nonredundant protein database using BLASTx (E score, $<10^{-4}$). The number of sequences with identities to eukaryotic viruses is shown. Twenty-nine eukaryotic virus families were detected from human, vertebrate, plant, and insect hosts (as well as unclassified viruses and satellites).

fraction (37%) of the sequences was unrecognizable on the basis of the BLASTx criterion of an E value of $<10^{-4}$. In contrast to other environmental metagenomes, these results indicated that in these sewage samples, eukaryotic viruses and phages were present in roughly equivalent numbers. The large number of unclassifiable sequences, within the range reported by other studies of purified viral particles, is likely to consist, at least in part, of highly divergent prokaryotic and eukaryotic viral sequences unrecognizable by BLASTx against the current viral database.

Twenty-nine eukaryotic viral families represented in sewage viromes. Approximately 110,000 sequences derived from the four sewage libraries exhibited sequence similarity to eukaryotic viruses (BLASTx E score, $<10^{-4}$), including matches to at least 29 different eukaryotic viral families (Fig. 1 and 2; Table S13 in the supplemental material contains the assembled sequences and sequence identity information). The most frequently amplified viral sequences belonged to families infecting invertebrates, followed by viruses found in diverse environmental samples but without defined hosts, plant viruses, viruses of vertebrates, and finally, human viruses at 6% of all identified eukaryotic viral sequences (see Fig. S1 and S2 in the supplemental material). When the threshold of sequence similarity was increased to >90% amino acid sequence identities, we detected members of 18 eukaryotic viral families (Table 1; see Tables S2 to S4 in the supplemental material). All viruses identified in sewage were nonenveloped viruses. Sequence matches to large-genome lipid-enveloped double-stranded DNA (dsDNA) viruses, including asfaviruses, poxviruses, and herpesviruses, were also detected but upon visual inspection of the alignments were excluded, as these consisted of repeat sequences that were possibly the result of amplification or sequencing artifacts.

Enteric human viral pathogens. The human viruses detected included members of seven distinct families (Table 1). Many recently described human viruses were detected (cosavirus [34, 49], cardiovirus [1, 12, 17, 42], salivirus/klassevirus [31, 35, 65], bocaviruses 2 and 3 [4, 47, 48]), underlining sewage's potential for further human virus discovery. Californian and Nepalese sewage contained the greatest diversity of human viruses. Astroviruses and caliciviruses (noroviruses and sapoviruses) were detected in California, while Nepal's human viruses consisted mostly of picornaviruses. Poliovirus type 2 vaccine strain (Sabin-2) was detected in the Nepalese sample as a pyrosequence with 100% nucleotide sequence identities to poliovirus 2 (reference genome positions 4329 to 4428). Human Saffold viruses (family *Picornaviridae*, genus *Cardiovirus*) were found in both the United States and Nigeria. Aichi viruses (family *Picornaviridae*, genus *Kobuvirus*) were found in Nepal, Thailand, and the United States. Bocaviruses 2 and 3 were found in Nepal and the United States, respec-

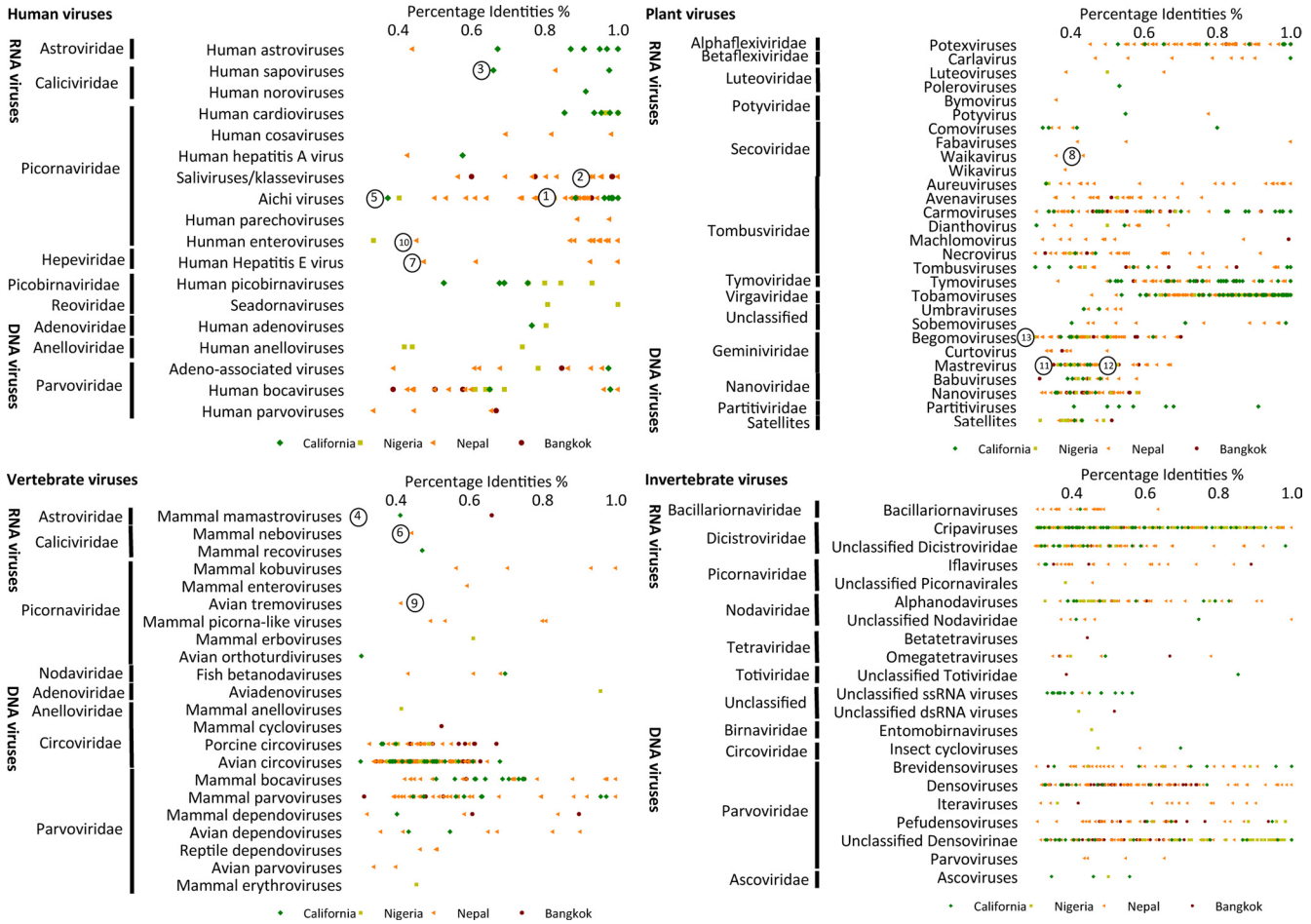

FIG 2 Sequence identity distribution analysis of eukaryotic viral sequences from the sewage metagenomes of four countries. Each dot represents an assembled sequence contig or singlet with the corresponding protein identity (best BLASTx match with an E score of $<10^{-4}$) to a human, vertebrate, plant, or insect virus in the GenBank nonredundant database. The original pyrosequences used as anchor points for partial for genomes extensions are indicated by numbers, as follows: 1, KoV-SewKTM; 2, SaliV-SewBKK; 3, SaV-SewSFO; 4, AstV-casa; 5, PicaV; 6, SecaliV; 7, HepeV; 8, NediV; 9, CadiV; 10, NiflaV; 11, BamiV; 12, NimiV; 13, NepaV.

tively. Human hepatitis E virus (HEV) was detected in the Nepalese sample only. In terms of sequence read numbers, the most strongly represented human viruses were all positive ssRNA viruses: Aichi viruses > saliviruses > astroviruses.

Overall, sequences related to human viruses by best BLASTx comparisons in U.S. sewage showed higher percent identities to database viruses (86%) than similar sequences from the other three countries (74% to 78%). This greater degree of similarity may reflect the larger contribution to GenBank of viral sequences from the United States and other developed countries relative to viral sequences from less developed countries.

High diversity of vertebrate, plant, and insect viruses in sewage viromes. Viruses infecting other eukaryotes besides humans were also abundant. Overall, at least 29 eukaryotic viral families were found in the sewage samples (Fig. 1 and 2).

Known vertebrate (nonhuman) viruses (>90% protein identities; see Table S2 in the supplemental material) consisted of adenoviruses from birds, parvoviruses from birds, pigs, cows, dogs, cats, and mice, as well as picornaviruses infecting pigs. Besides these viruses, the majority of animal virus sequences shared much lower levels of 30% to 90% protein identities to known animal viruses (Fig. 2). These divergent viral sequences showed sequence similarities to the *Astroviridae*, *Caliciviridae*, *Picornaviridae*, *Nodaviridae*, *Adenoviridae*, *Anelloviridae*, *Circoviridae*, and *Parvoviridae* families (Fig. 2), reflecting an abundance of novel animal viruses in sewage.

Plant viruses accounted for 21% of the total viral reads, including both known plant viruses and more divergent viral sequences (Fig. 2 and see Fig. S1 in the supplemental material). Known plant pathogens included alphaflexiviruses, betaflexiviruses, partitiviruses, secoviruses, sobemovirus, tombusviruses, tymoviruses, and virgaviruses (see Table S4 in the supplemental material). An even larger proportion of the plant virus-like sequences were from more divergent plant viruses (Fig. 2; <90% identities). Sewage therefore contained 12 out of the total 21 known plant viral families, dominated by major groups of nonenveloped plant viruses.

Known insect viruses, including dicistroviruses, nodaviruses, and densoviruses (>90% amino acid sequence identities; see Table S3 in the supplemental material) that infect bees, mosquitoes, and other insects, were found in sewage. The majority of the insect virus sequences belonged to more divergent insect viruses that

TABLE 1 Distribution and sequence identities of sequence reads in sewage with >90% amino acid identities to known human viruses

Viral family	Virus	% amino acid identity to known human viruses							
		Nigeria		Nepal		Bangkok		California	
		Minimum	Maximum	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum
Astroviridae	Human astrovirus							91	91
	Human astrovirus 1							100	100
	Human astrovirus 3							100	100
	Human astrovirus 4							95	95
	Human astrovirus 8							97	97
Caliciviridae	Norovirus Hu/GI/Otofuke/1979/JP							91	91
	Sapovirus SaKaeo-15/Thailand							98	98
Hepeviridae	Hepatitis E virus			93	100				
Parvoviridae	Adeno-associated virus			96	96			97	97
	Human bocavirus 2			96	100				
	Human bocavirus 3							98	98
Picobirnaviridae	Human picobirnavirus	93	93						
Picornaviridae	Aichi virus			90	100	93	98	96	100
	Human cosavirus D			98	98				
	Human coxsackievirus B2			93	93				
	Human coxsackievirus B6			93	93				
	Human echovirus 11			92	92				
	Human enterovirus 76			95	100				
	Human enterovirus 97			98	98				
	Human parechovirus 1			98	98				
	Human poliovirus 2			97	97				
	Saffold virus	96	100					94	100
	Salivirus NG-J1			92	96				
Reoviridae	Banna virus	100	100						

shared <90% amino acid similarities with known insect viruses (Fig. 2).

In summary, the sewage contained a very high diversity of known (Table 1; see Tables S2 to S4 in the supplemental material) and divergent (Fig. 2) viruses infecting multiple kingdoms of eukaryotic hosts. Several divergent viral sequences were selected for further directed genome sequencing (Fig. 2, numbered dots), resulting in the characterization of the novel viral genera and families described below. The large number of divergent sequences that were not extended indicates that numerous other novel viral genomes remained only minimally characterized.

Novel picornaviruses in the kobuvirus and salivirus genera. Using sequences showing similarities to human picornaviruses as starting points for genome extension, we acquired near complete genomes of new viruses using RT-PCR and 5′ and 3′ RACE amplifications, followed by Sanger sequencing.

From the Nepalese sample, a near complete genome of a novel kobuvirus in the *Picornaviridae* family was characterized and was named kobuvirus sewage Kathmandu (KoV-SewKTM) after the location of its discovery. Related human kobuviruses, the pathogenic Aichi viruses, have been associated with gastroenteritis (3, 30, 63, 77, 83, 95–98), while other kobuviruses have been detected worldwide in pigs and cows with diarrhea (51, 81) as well as in mouse and canine feces (45, 64, 78). KoV-SewKTM (GenBank accession number JQ898342) was 6,939 nucleotides (nt) long and shared ~82% nucleotide sequence identities to human Aichi virus, canine kobuvirus, and murine kobuvirus over its genome (see Fig. S3A in the supplemental material). Pairwise distance analysis indicated that KoV-SewKTM shared the highest amino acid sequence identities in the P1 region with mouse kobuvirus (87%) and was equidistant to other kobuviruses in the 2C plus 3CD region (86% to 88%) (see Table S5 in the supplemental material). Phylogenetic analysis of the P1 region confirmed these relationships (Fig. 3). Since members of each kobuvirus species share >70% amino acid sequence identities in P1 plus >80% amino acid sequence identities in the 2C plus 3CD region (54), KoV-SewKTM, together with murine and canine kobuviruses, all fit into the human Aichi virus species on the basis of the species demarcation criterion used for enteroviruses (54) and recently substantiated by a larger analysis of picornavirus diversity (60). Since the VP1 of KoV-SewKTM shared less than 84% amino acid sequence identities to the Aichi virus-related viruses, a threshold functionally determined to differentiate enterovirus genotypes corresponding to serotypes (74), KoV-SewKTM may qualify as a candidate for a novel kobuvirus genotype.

A near complete genome of a novel salivirus was also sequenced from the Bangkok sample and was named salivirus sewage Bangkok (SaliV-SewBKK; GenBank accession number JQ898343). The near complete genome of SaliV-SewBKK was 6,397 nt long, containing the full-length polyprotein except for only a partial VP0 and no 5′ untranslated region (UTR) at the 5′ end. Currently, human saliviruses (also called klasseviruses) have been associated with diarrhea and detected in feces from both

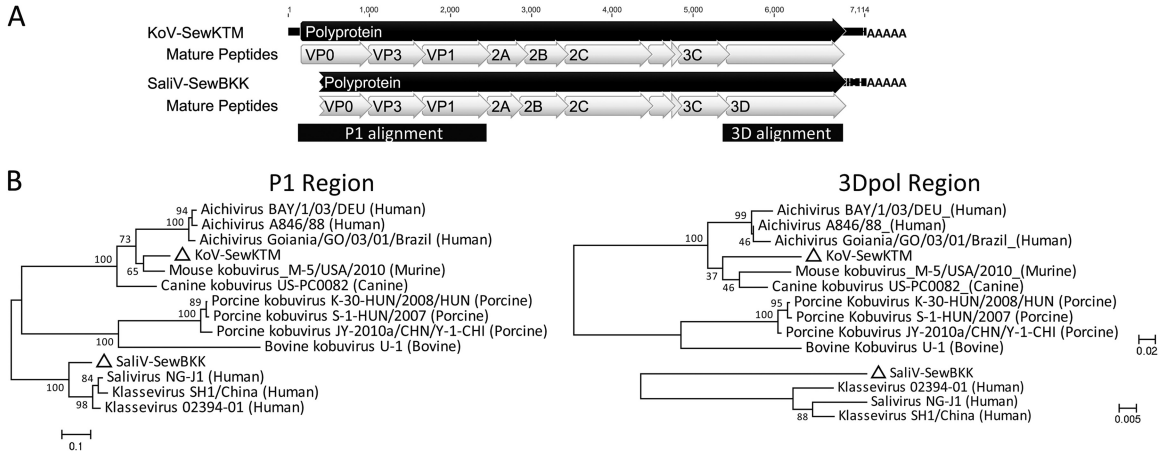

FIG 3 Genome organization and phylogenetic relationships of new kobuvirus and salivirus. (A) Genome organization of the kobuvirus sewage Kathmandu (KoV-SewKTM) and salivirus sewage Bangkok (SaliV-SewBKK). (B) Phylogenetic analysis of the two novel picornaviruses with known members of the *Kobuvirus* and *Salivirus* genera in the family *Picornaviridae*, using translated protein sequence alignments of P1 capsid and 3D polymerase regions and the maximum likelihood method. Aligned regions are shown by black bars.

gastroenteritis patients and healthy subjects from Nigeria, Tunisia, Nepal, Australia, and the United States, as well as in sewage from Spain, suggesting a widespread geographic distribution (31, 35, 65). Phylogenetically, the salivirus from sewage from Bangkok (SaliV-SewBKK) branches out at a basal position to the known saliviruses that have recently been proposed to form their own genus in the *Picornaviridae* family (54, 60). SaliV-SewBKK was equidistant to other saliviruses in the P1 region (85% to 86%), 2C plus 3CD region (90%), and VP1 region (82% to 85%), as shown in both similarity plot and pairwise distance analysis (see Tables S5A to C in the supplemental material). Comparing such percent identity with criteria for enterovirus serotypes, SaliV-SewBKK may be considered the second serotype of the salivirus species since the prior saliviruses/klasseviruses from human feces were all closely related (see Table S5 in the supplemental material).

New sapovirus in California. A new sapovirus, namely, sapovirus sewage/California/2009 (SaV-SewSFO; GenBank accession number JQ898338), was characterized from the California sewage sample. A fraction of its genome (1,385 nt long), including half of the capsid, was sequenced and phylogenetically compared to the 12 current genogroups of human and animal sapoviruses (see Fig. S4A and B and Table S6 in the supplemental material). On the basis of genetic distance criteria and clustering, it appears that this virus may be classified either as a divergent member of genogroup 2 or as a member of a new genogroup.

A divergent astrovirus related to mama- and avastroviruses. Nearly half of the genome of a highly divergent astrovirus was acquired from the California sewage sample and was tentatively named casa (for California sewage-associated) astrovirus (AstV-casa; GenBank accession number JQ898337). The positive single-stranded RNA (ssRNA) family *Astroviridae* consists of two genera, *Avastrovirus* and *Mamastrovirus*, known to infect avian and mammalian hosts, respectively. Human astroviruses are transmitted through the fecal-oral route and have been associated with gastroenteritis (13, 26, 29, 32, 39, 91). Several new human astroviruses have recently been described (24, 25, 44). Astroviruses in other mammalian and avian species have also been associated with diarrhea (40, 41, 55, 56). Consistent with the genome organization of *Astroviridae*, the partial genome of AstV-casa (3,206 nt) consists of ORF1b encoding RNA-dependent RNA polymerase (RdRP), followed by a second open reading frame (ORF) encoding a capsid (Fig. 4). The 3′ UTR of AstV-casa lacked the stem-loop-2-like motif described in a subset of known astroviruses (40).

Phylogenetic analyses indicated that AstV-casa was highly divergent from the other astroviruses, placing it at the root of the *Astroviridae* family (Fig. 4). Pairwise distance analysis showed that AstV-casa was equidistant with mamastrovirus and avastrovirus in both the RdRP (<20%) and capsid (<11%) regions (see Table S7 in the supplemental material). Its basal position makes it difficult to suggest either birds or mammals as likely hosts.

To classify AstV-casa, we compared the amino acid sequence identities throughout its coding region using sliding-window analysis. When AstV-casa was compared to known astroviruses, it showed greater divergence from them than the intragenus variations within mamastroviruses and avastroviruses (see Fig. S5 in the supplemental material). Together with the results of phylogenetic analysis and pairwise distance analysis (see Table S7 in the supplemental material), the sequence identities suggested that AstV-casa reflects the existence of a third *Astroviridae* genus, provisionally named *Casastrovirus*.

Picalivirus, a highly divergent virus in the picorna-like virus superfamily. Positive-strand RNA virus-related sequences showing even lower-level sequence similarities to both *Picornaviridae* and *Caliciviridae* families were abundant. We acquired the full-length genome of one such virus from Nepalese sewage (GenBank accession number JQ898334-6), and provisionally named it picalivirus (picornavirus- and calicivirus-related virus [PicaV]). The picalivirus RNA genome consisted of 8,996 nucleotides beginning with a 2,025-amino-acid-long ORF encoding a nonstructural polyprotein, followed by a second ORF encoding a capsid protein and a 3′ UTR ending in a poly(A) tail (Fig. 5). As characterized by 5′ RACE amplification, the 5′ end of the picalivirus genome contained a very short 5′ UTR in a manner reminiscent of caliciviruses. The genomic organization of picalivirus resembled most closely that of the *Lagovirus* and *Sapovirus* genera in the *Caliciviridae* family (15), with their very short 5′ UTR and a long nonstructural polyprotein containing the helicase and RdRP regions, followed by a second ORF encoding the capsid protein (Fig. 5A; see

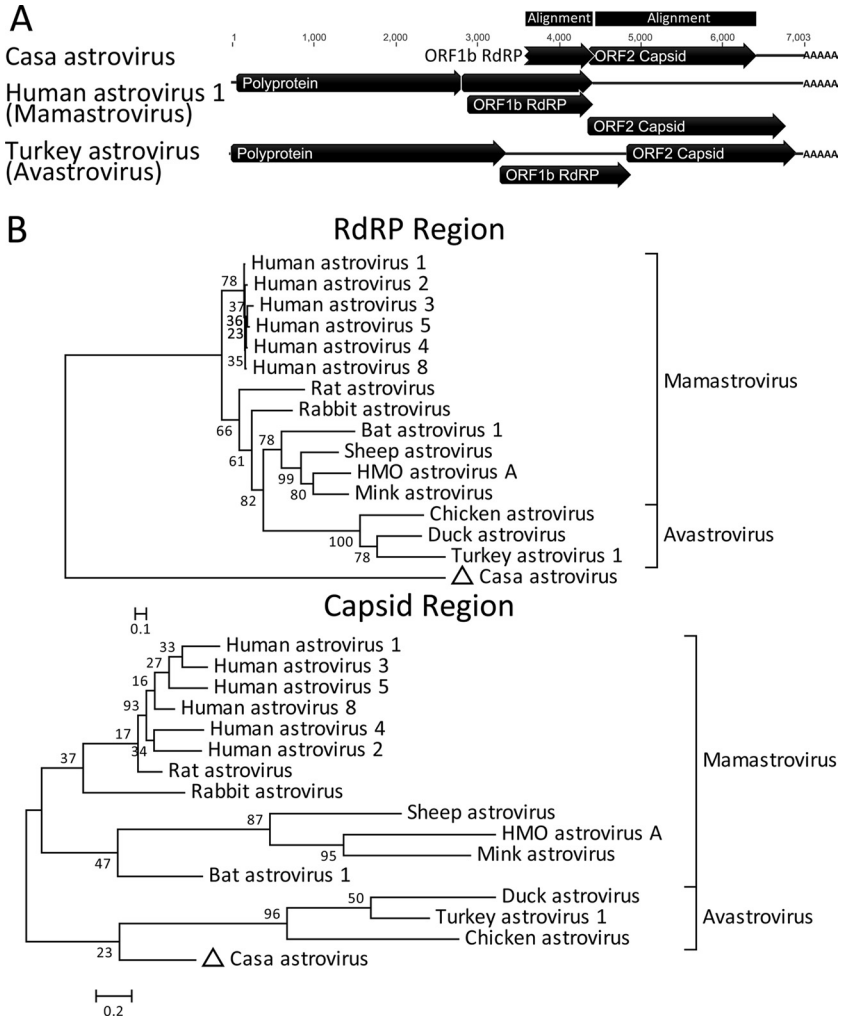

FIG 4 Genome organization and phylogenetic relationship of the casa astrovirus (AstV-casa). (A) Genome organization of the casa astrovirus compared with human and turkey astroviruses; (B) phylogenetic analysis of the translated partial RdRP and the full capsid protein sequences of AstV-casa and representatives of the *Mamastrovirus* and *Avastrovirus* of the *Astroviridae* family using the maximum likelihood method.

Fig. S6 in the supplemental material for a comparison of ORF organizations).

Amino acid sequence identities in the helicase and RdRP regions of picalivirus and members of the *Picornavirales* order (families *Picornaviridae*, *Iflaviridae*, *Dicistroviridae*, *Marnaviridae*, and *Secoviridae*) (61) and *Caliciviridae* were calculated through pairwise analysis (see Tables S8A and B in the supplemental material). The most closely related genome was that of a calicivirus with 25% and 20% amino acid sequence identities in the helicase and RdRP regions, respectively. Except for several short conserved motifs, including the helicase nucleotide-binding motif [GXXGXGK(T/S)] and several RdRP motifs (KDEX, YGDD, and FLRR) (43, 57, 61), the picalivirus polyprotein was highly divergent from the other RNA viruses (Fig. 5A and C; see Tables S8A and B in the supplemental material). Conserved motifs were slightly closer to those of the *Caliciviridae* than to other members of the *Picornavirales* order (Fig. 5C). Protein structure prediction (86) showed that the structures closest to the PicaV RdRP were the poliovirus three-dimensional (3D) RdRP (E value, $4e^{-103}$) and Norwalk virus RdRP (E value, $7e^{-96}$), while the structures closest to the PicaV capsid were the cripavirus capsid (E value, $7e^{-15}$) and the cardiovirus capsid (E value, $2e^{-8}$).

Phylogenetic analyses using the RdRP and the helicase regions placed picalivirus close to the root of the *Picornavirales* order, indicating the early divergence of this viral group and suggesting picalivirus as a candidate prototype of a novel viral family provisionally named *Picaliviridae* (Fig. 5B). Previous genetic analyses have described a large grouping of positive-strand RNA viruses which includes the *Picornavirales* order (58, 61), the *Caliciviridae*, plus other positive single-stranded RNA viral families into a picorna-like virus superfamily (58). The picorna-like virus superfamily, due to its wide distribution in all forms of nucleated cells, is theorized to have emerged early in the evolution of eukaryotes (58). A subset of the viruses in the picorna-like virus superfamily encodes a superfamily 3 helicase, including all members of the *Picornavirales* order plus the *Caliciviridae* (58). Picalivirus also contains a homolog of this enzyme. On the basis of its high level of divergence relative to other members of the picorna-like virus superfamily, picalivirus may therefore be a prototype of a new viral family related to the *Picornavirales* order.

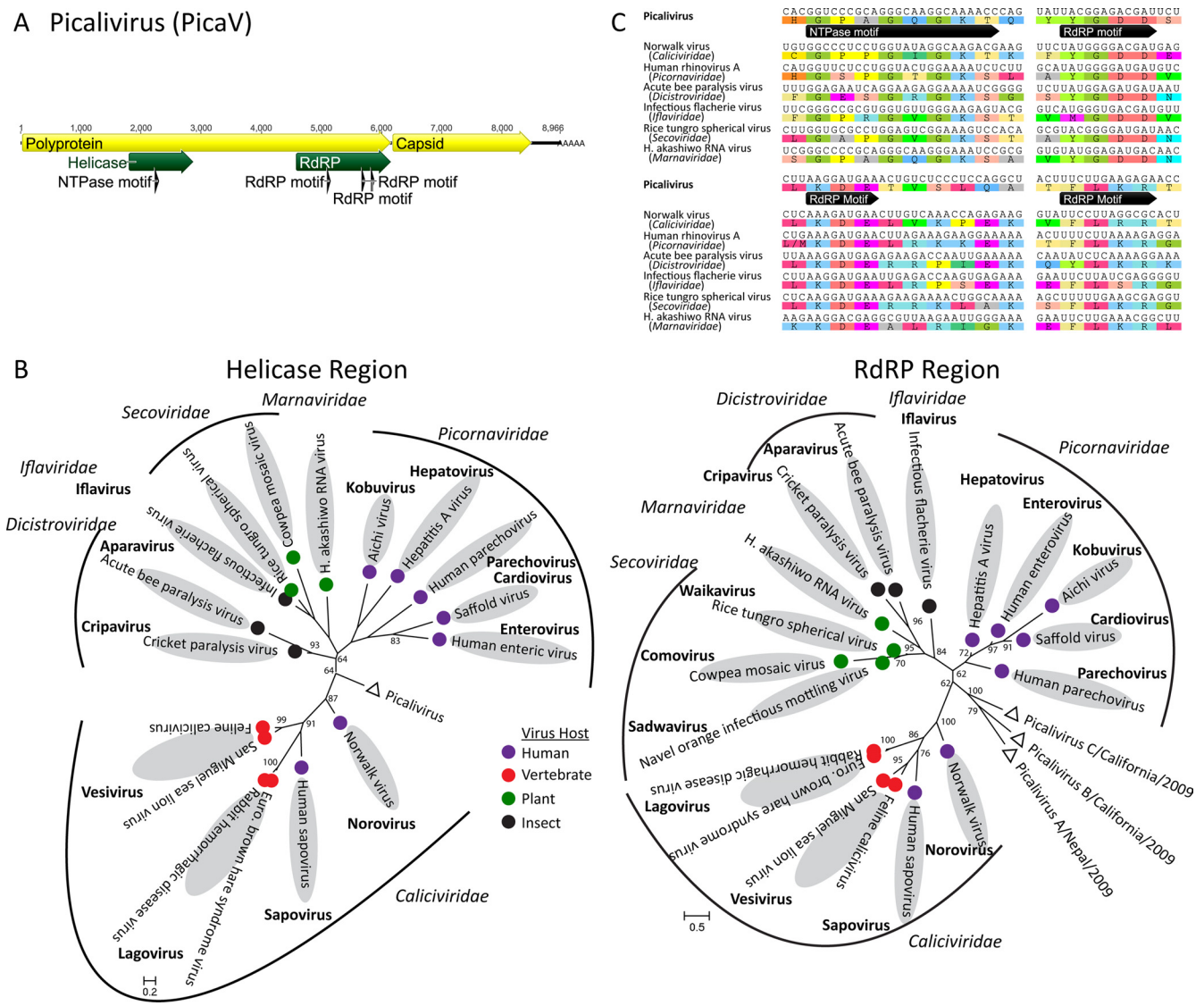

FIG 5 Genome organization, phylogenetic analyses, and conserved motifs of the picalivirus (PicaV) with members of the *Picornavirales* and *Caliciviridae*. (A) Genome organization and conserved helicase and RdRP motifs of the picalivirus; (B) phylogenetic analyses of picalivirus using the maximum likelihood method; (C) conserved helicase and RdRP motifs of PicaV compared with those of members of the *Picornavirales* and *Caliciviridae*.

High level of picalivirus diversity and geographic distribution. Metagenomics-derived sequences were analyzed for the presence of sequences closely related to the first picalivirus genome. Two other picalivirus genomes were identified and partially sequenced using RACE amplifications (GenBank accession numbers JQ898335 and JQ898336). Phylogenetic analysis of their RdRP regions showed that they shared a common root with the original picalivirus genome (Fig. 5B) with amino acid sequence identities of 30% to 34% (see Table S8 in the supplemental material), suggesting that these picaliviruses constitute prototypes of three *Picalividae* genera (using as criteria the genetic distances between genera within the *Picornaviridae* and *Caliciviridae* families) (52, 60). Picaliviruses were found in both San Francisco and Kathmandu, suggesting a wide geographic distribution.

Secalivirus, a virus distantly related to *Caliciviridae*. The family *Caliciviridae* consists of five genera, *Norovirus*, *Sapovirus*, *Vesivirus*, *Lagovirus*, and *Nebovirus*, as well as a recently described group, recovirus (14, 15, 20, 75). Caliciviruses are known to infect only vertebrate hosts, causing a wide range of diseases, including respiratory infections, vesicular lesions, gastroenteritis, and hemorrhagic disease (14, 16). The partial genome of a highly divergent calicivirus was acquired from the Nepal sewage sample and preliminarily named secalivirus (sewage-associated calici-like virus [SecaliV]; GenBank accession number JQ898339). The partial genome consisted of 4,068 nucleotides, including a partial capsid gene, followed by two ORFs of unknown function, and a 553-nt-long 3′ UTR ending with a poly(A) tail (Fig. 6). Pairwise identity analysis suggested that the SecaliV capsid protein was highly divergent from the capsid proteins of the *Caliciviridae* (10.5 to 17.5% identity) (see Table S9 and Fig. S7 in the supplemental material).

The capsid of SecaliV did contain a conserved capsid motif (calicivirus coat protein pfam00915), with certain amino acid positions being more like those of caliciviruses than other positive

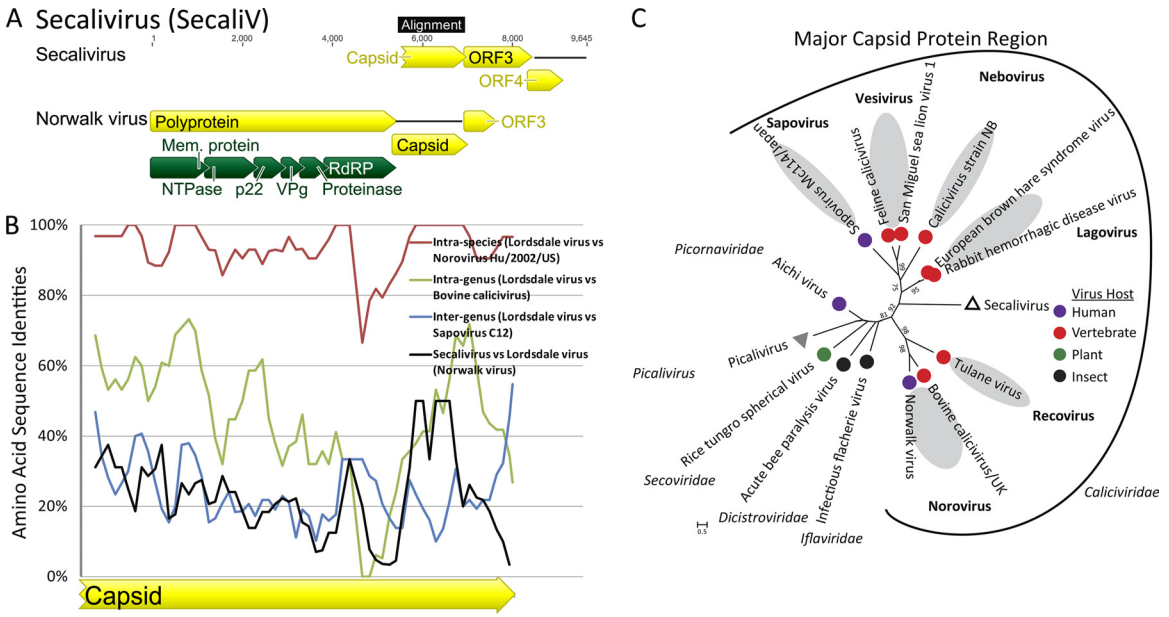

FIG 6 (A) Genome organization of secalivirus (SecaliV) compared with Norwalk virus (*Caliciviridae*). The partial genome of secalivirus consists of a partial capsid ORF, followed by two ORFs of unknown function. (B) Sequence identity comparisons of secalivirus and calicivirus diversity through sliding-window analyses of pairwise translated protein *p* distances of the capsid. (C) Phylogenetic analysis of the partial capsid region of secalivirus and representatives of the *Caliciviridae* and *Picornavirales* using the maximum likelihood method.

ssRNA viruses (see Fig. S7 in the supplemental material). Similarly, protein structure prediction (86) showed that the structure closest to the SecaliV capsid was the calicivirus coat protein (E value, $2e^{-57}$). Comparing the capsid protein of SecaliV, caliciviruses, and other members of the *Picornavirales*, the phylogenetic analysis also placed SecaliV near the root of the *Caliciviridae* (Fig. 6C). Lastly, sliding-window analysis showed that SecaliV shared a degree of divergence from known caliciviruses similar to the intergenus variations of known calicivirus genera (Fig. 6B). Collectively, these results indicate that secalivirus may be tentatively classified as the prototype of a new genus in the *Caliciviridae*.

A highly divergent HEV-like genome. A sequence from Nepal shared low-level similarity to HEV in the family *Hepeviridae*. This genome was tentatively named hepevirus-like virus (hepelivirus [HepeV]; GenBank accession number JQ898340). The partial genome of hepelivirus acquired was 2,721 nt long, consisted of half of ORF1 (RdRP), a complete ORF2 (capsid), as well as the 3′ UTR (Fig. 7). Nepal was the only sampled region where human HEV sequences were detected (i.e., >90% protein identity) (Table 1). HEV can cause self-limited or fulminant hepatitis in humans, as well as infect a range of mammalian species as a zoonotic agent. Recently, other HEV-related viral species have been characterized in rodent, avian, bat, and fish host species (2, 5, 18, 38, 52).

Pairwise distance analysis of the RdRP amino acid suggested that HepeV was equidistant to HEV, rodent HEV, avian HEV, and cutthroat trout virus (~20% identity), while showing weaker identity to members of other ssRNA families (7% to 12%) (see Table S10 in the supplemental material). HepeV contained conserved features of RdRP of the *Hepeviridae*. Like other HEVs, HepeV contained FKGDDS (underscore highlights a conserved RdRP motif) (5, 28) and DVXR motifs, compared to the XY GDDX and FL(K/R)R motifs common to members of the *Picornavirales* (see Fig. S8 in the supplemental material) (43, 57, 61). HepeV, like other genomes in *Hepeviridae*, lacked the KDEX motif that is consistently found among members of the order *Picornavirales* (see Fig. S8 in the supplemental material). Phylogenetic analysis of the RdRP region also showed HepeV's closer relationship to the family *Hepeviridae*, placing HepeV at the root of HEV-related viruses and suggesting an early divergence from other known *Hepeviridae* species (Fig. 7).

All three analyses, i.e., conserved motifs, pairwise distances, and phylogenetics, suggested that HepeV belongs to the *Hepeviridae* family. When HepeV was compared to other members of the *Hepeviridae*, it showed divergence greater than their interspecies variations (Fig. 7B). HepeV may therefore represent a prototype for a new genus within the family *Hepeviridae*.

Genomes of other highly divergent ssRNA viruses. We also characterized the RdRP regions together with their 3′ ends of three other highly divergent viruses, provisionally called nedicistrovirus (Nepal sewage dicistro-like virus [NediV]), cadicistrovirus (California sewage dicistro-like virus [CadiV]), and niflavirus (Nepal sewage ifla-like virus [NiflaV]) (GenBank accession numbers JQ898341, JQ898344, and JQ898345, respectively).

Phylogenetic analysis of these RdRP regions showed that NediV, CadiV, and NiflaV were highly divergent from known viruses (Fig. 8). The partial genome of NediV was 4,631 nt long, with a genome organization consistent with that of other dicistroviruse and an RdRP region located at the 3′ end of a polyprotein, followed by another ORF encoding a capsid and a long 3′ UTR. The genetic organization and phylogenetic and genetic distances of the RdRP region indicated that NediV likely belongs to the genus *Cripavirus* in the *Dicistroviridae* family (Fig. 8; see Fig. S9 and Table S11 in the supplemental material).

Dicistroviridae and *Iflaviridae* are the only families within the order *Picornavirales* that infect insects. The partial genome of CadiV was 2,603 nt long, containing a long polyprotein encoding

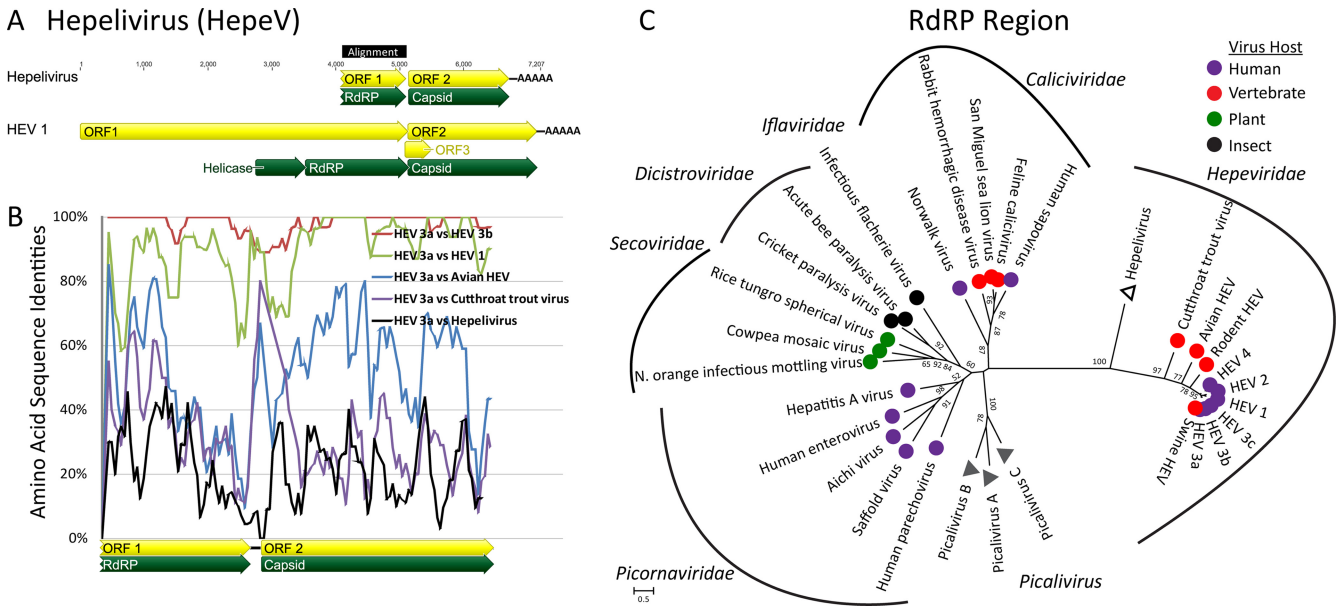

FIG 7 (A) Genome organization of hepelivirus (HepeV) compared with that of human hepatitis E virus. The partial genome of hepelivirus consists of a partial ORF1b encoding RdRP, followed by ORF2 encoding a capsid protein. (B) Sequence identity comparisons of hepelivirus and hepevirus diversity through sliding-window analyses of pairwise translated protein *p* distances of the RdRP and capsid. (C) Phylogenetic analysis of the partial RdRP region of hepelivirus and representatives of the *Hepeviridae* and other ssRNA viruses using the maximum likelihood method.

the RdRP, followed by the capsid protein in the same reading frame (see Fig. S9 in the supplemental material). NiflaV (partial genome of 1,614 nt) resembled iflaviruses in terms of a polyprotein containing RdRP, without a capsid-encoding gene at the 3′ end of the genome. Phylogenetic analysis of the RdRP region suggested that CadiV and NiflaV branched out near the root of the tree, suggesting an early divergence from the picorna-like virus superfamily (Fig. 8).

NediV, CadiV, and NiflaV are three examples of divergent ssRNA viruses among many likely present in the sewage viromes (Fig. 2). Given the number of sequences showing low-level similarities to known viruses, we postulate that many more divergent ssRNA genomes remain uncharacterized in these sewage samples.

Genomes distantly related to plant geminiviruses. Sewage viromes contained plant viral sequences with various degrees of divergence from known plant pathogens, especially from the single-stranded DNA (ssDNA) family *Geminiviridae*. The family *Geminiviridae* contains four genera, *Begomovirus*, *Curtovirus*, *Topocuvirus*, and *Mastrevirus*, which are responsible for various types of plant diseases (21, 23, 52). Geminiviruses often limit the production of tomato, pepper, squash, melon, and cotton in the subtropics and tropics (67, 80), causing famines in the developing world (62). We describe three highly divergent geminivirus-like genomes (GenBank accession number JQ898331-3).

Baminivirus (Bangkok gemini-like virus [BamiV]), niminivirus (Nigeria gemini-like virus [NimiV]), and nepavirus (Nepal gemini-like virus virus [NepaV]) genomes consisted of circular DNA 2.3 to 2.8 kb long. Similar to other geminiviruses, these genomes contained two major ORFs encoding Rep and capsid proteins in an opposite orientation (Fig. 9A). Notably, BamiV and NimiV also contained a stem-loop in the UTR region and the nonanucleotide TAATATTAC, which are highly conserved features among geminiviruses (22) (Fig. 9A). NepaV contained a stem loop with the 15-nt sequence CTATTATAACATTGC. Similar to geminiviruses, RCR motifs 1, 2, 3 and Walker A and B motifs were identified in the Rep protein of the three gemini-like viruses (Arguello-Astorga G., personal communication). ORFs related to movement protein (MP) were also identified in BamiV and NimiV, with ~35% protein identities to an array of geminivirus MP, suggesting that BamiV and NimiV genomes may be monopartite.

Phylogenetic analysis showed that BamiV, NimiV, and NepaV, while highly divergent from known geminiviruses, still clustered in the same clade as other ssDNA viral families (Fig. 9B). Baminivirus and niminivirus Rep proteins shared less than 30% pairwise amino acid sequence identities with other geminiviruses (see Table S12 in the supplemental material). Pairwise distance analysis of BamiV and NimiV with the Rep protein of other circular ssDNA genomes showed >19% Rep protein identities to Rep proteins in the *Geminiviridae* versus <15% identities to those in the *Circoviridae* and *Nanoviridae* (see Table S12 in the supplemental material). Sliding-window analysis of the replication gene amino acid sequence alignment showed that the identities between baminivirus and different geminiviruses were greater than the intragenus variations of geminiviruses (Fig. 9C), suggesting that baminivirus and niminivirus may be prototypes for two new genera within the *Geminiviridae*. Formal classification of these viruses in the *Geminiviridae* will require particle structure analysis and confirmation of their ability to infect plants.

Nepavirus lacked the exact stem-loop nonanucleotide signature of geminiviruses; however, it was still more closely related to *Geminiviridae* than to other known ssDNA viral families in pairwise distance analysis (>11 to 19% Rep protein identities to the Rep proteins of the *Geminiviridae* versus <7 to 12% identities to the Rep of a circovirus and nanovirus (see Table S12 in the supplemental material). Nepavirus is therefore related to the *Gemini-*

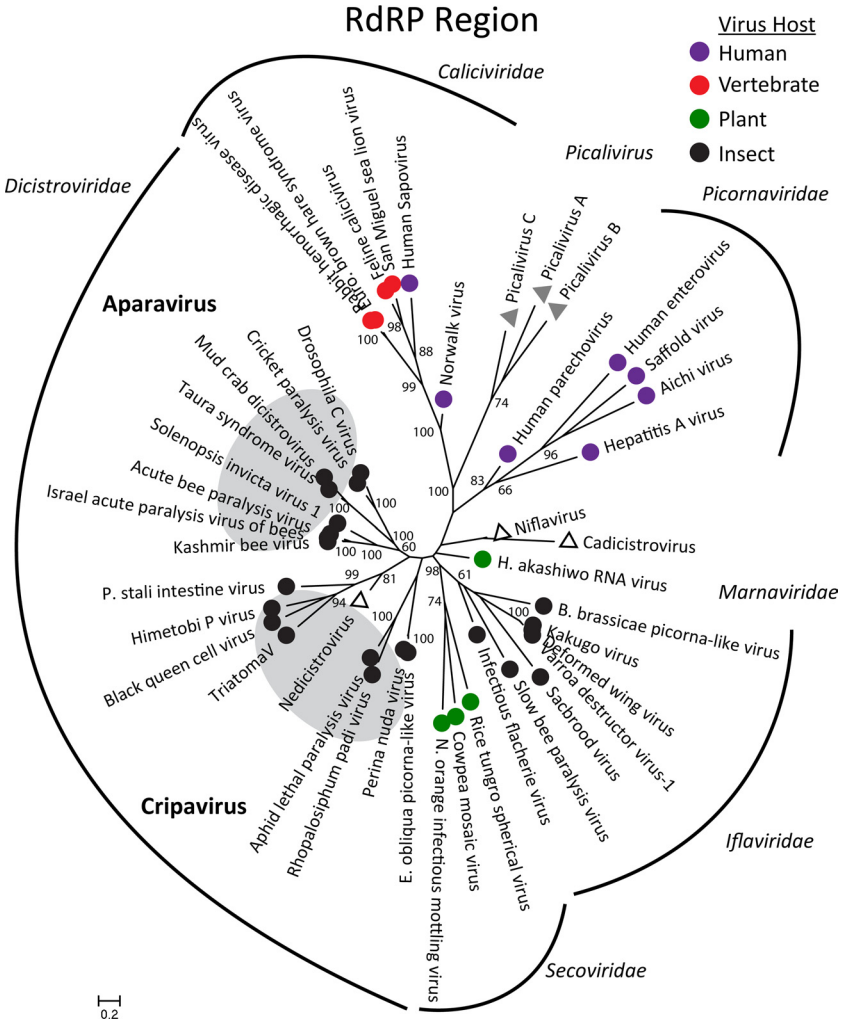

FIG 8 Phylogenetic relationships of nedicistrovirus, cadiscistrovirus, niflavirus, and other ssRNA viruses, based on the amino acid alignment of the RdRP region using the maximum likelihood method.

viridae family, but due to its sequence divergence and lack of conserved nonanucleotide, it may represent a new viral family.

Inference of cellular hosts using nucleotide composition analysis. We previously showed that ssRNA viruses from vertebrates, invertebrates, and plants could be broadly differentiated on the basis of a discriminant analysis of their di- and trinucleotide composition (46, 84). The complete and partial genome sequences generated here were therefore analyzed using nucleotide composition analysis (Fig. 10). The kobuvirus (KoV-SewKTM), salivirus (SaliV-SewBKK), and sapovirus (SaV-SewSFO) showed a nucleotide composition consistent with their expected vertebrate origins based on phylogenetic affinities described in previous sections (Fig. 3 and 4; see Fig. S4 in the supplemental material). NCA also inferred a vertebrate host for the new astrovirus AstV-casa.

The full or partial genomes of picaliviruses (PicaVs A, B, and C), secalivirus (SecaliV), hepelivirus (HepeV), and cadicistrovirus (CadiV) showed nucleotide composition properties comparable to those of invertebrate (insect/nematode)-infecting viruses. Currently, all known members of the *Hepeviridae* and *Caliciviridae* infect vertebrates. If the NCA predictions of invertebrate hosts for hepelivirus and secalivirus are correct, these two deep clades may represent insect-infecting viruses that diverged early in the evolution of *Hepeviridae* and *Caliciviridae*, respectively, expanding the host range for these two families. More experimental data addressing the tropism of these viruses are required to confirm this prediction.

The niflavirus (NiflaV) clustered with plant viruses, whereas the nedicistrovirus (NediV) did not group with any of these three kingdoms of eukaryotic hosts. Further discriminant analysis using more nucleotide composition components yielded the same inferred hosts for all viruses and a vertebrate host for NediV (data not shown).

DISCUSSION

Several trends were evident on the basis of our metagenomic analyses of viruses in untreated sewage from four countries. Each sample contained distinct subsets of human viral pathogens (Table 1). Numerous novel viral genotypes, species, genera, and, possibly, families of nonenveloped eukaryotic RNA and DNA viruses were identified. Sewage therefore harbors a very high viral diversity of known and previously uncharacterized viruses that allows for

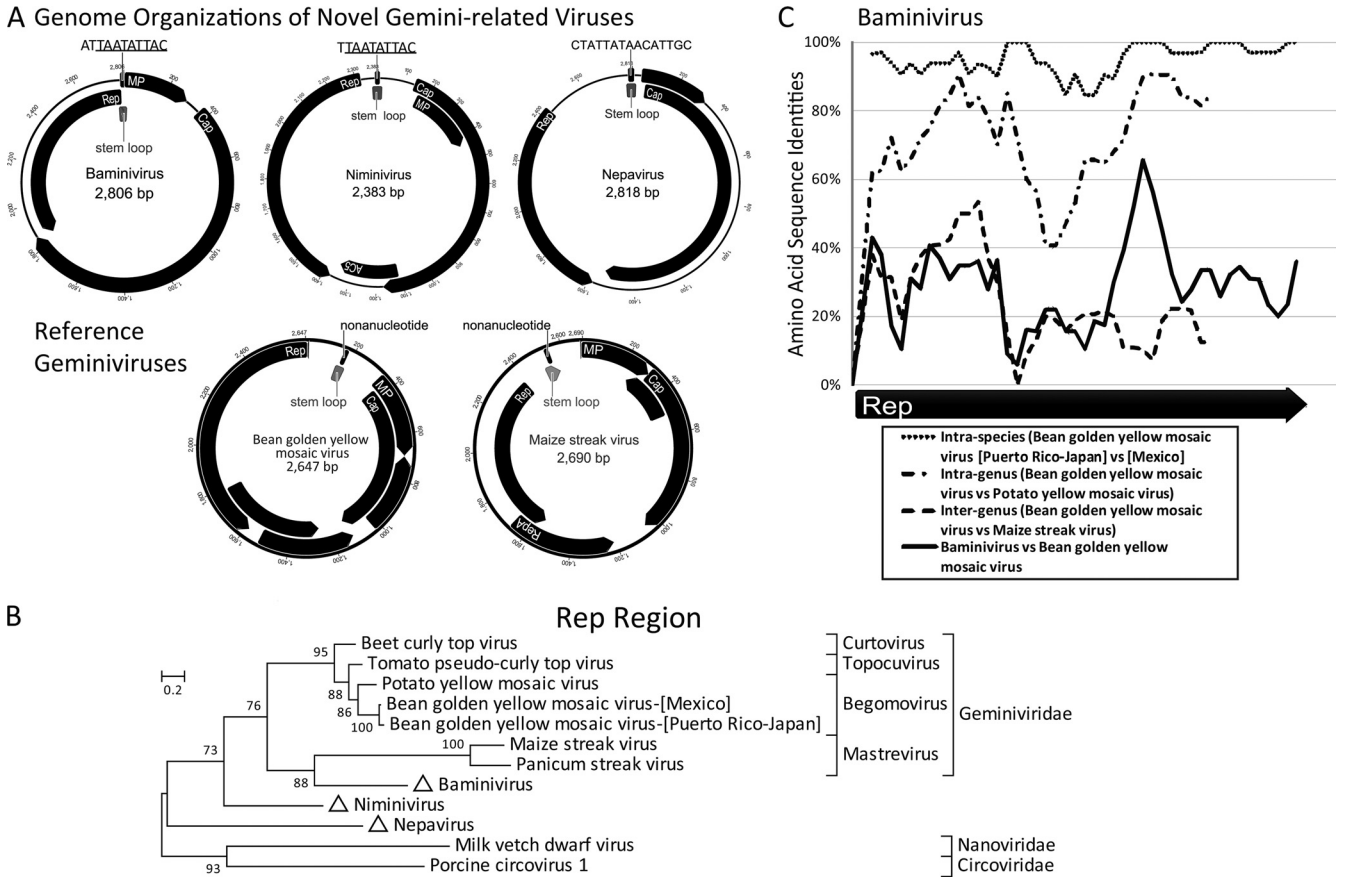

FIG 9 (A) Genome organization of the novel geminivirus-related viruses (baminivirus, niminivirus, and nepavirus) and two geminiviruses (reference species of begomoviruses and mastreviruses). Baminivirus and niminivirus both contained the nonanucleotide TAATATTAC within a stem-loop. (B) Phylogenetic analysis of the translated Rep protein sequences among niminivirus, baminivirus, and nepavirus members and representatives of *Geminiviridae* and other ssDNA viruses with circular ssDNA genomes using the maximum likelihood method. (C) Sequence identity comparisons of baminivirus and geminivirus diversity through sliding-window analysis of pairwise translated protein *p* distances in the Rep gene alignment. Niminivirus showed a similar result in sliding-window analysis (data not shown).

large-scale viral discovery and for monitoring of the presence of pathogens from humans and other cellular hosts.

While a large body of literature analyzing wastewater effluents for the presence of a specific virus(es) using PCR exists, unbiased studies of viruses in wastewaters using metagenomics approaches remain few. A metagenomics analysis of reclaimed water (for non-potable public uses) using pyrosequencing of viral particle-associated DNA indicated 98% prokaryotic versus 2% eukaryotic viral sequences, while the viral RNA population showed a more equally mixed distribution of viruses from a variety of inferred hosts more akin to that reported here (82). A metagenomics analysis of DNA viruses purified from activated sludge (a product of wastewater treatment) also showed a predominance (95%) of prokaryotic DNA viruses (76). A pyrosequencing analysis of the DNA viruses in the influent, activated sludge, effluent, and anaerobic digester of a wastewater treatment plant showed >90% of viral reads to be of likely bacterial origin (89). Lastly, a recent metagenomic analysis using pyrosequencing of DNA and RNA in viral particles from untreated wastewater from Pittsburgh, PA, Barcelona, Spain, and Addis Ababa, Ethiopia, found >80% of virus-like sequences likely originating from prokaryotes (11). The last study also revealed a great diversity of DNA and RNA viruses in which 85% of the eukaryotic virus-like sequences had a likely plant origin and generated the full viral genome of a phage-like genome previously thought to be a eukaryotic virus called non-A non-B hepatitis virus (11).

The sewage samples analyzed here were also rich in plant viruses, possibly reflecting the diversity of local plants as well as those plants consumed by local residents and animals. Some plant viruses (pepper mild mottle virus) are known to pass through the human gastrointestinal tract and remain infectious (99). Sewage may therefore provide another means for the dissemination of plant viruses if used untreated as fertilizer. Metagenomics studies of DNA viruses in insect vectors (whiteflies) also showed the majority of virus-like sequences to be closely related to known plant geminiviruses, including novel begomoviruses (68). Mosquito viromes also contained sequences related to *Circoviridae* and *Geminiviridae* (72).

Sequences with best BLASTx E score to human viruses showed higher percentage identities when derived from U.S. sewage than from that of other countries (average of 84% versus 72 to 74%, respectively). Such a result may reflect the greater contribution to GenBank of viral genomes from developed relative to more resource-constrained countries. The ease of recognition of diver-

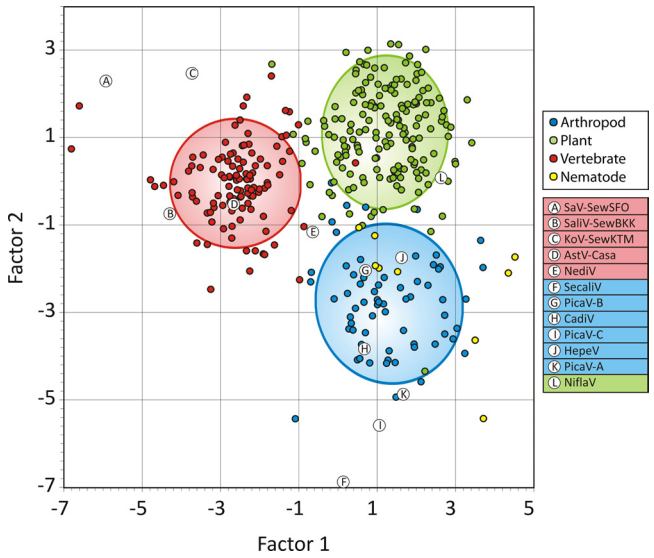

FIG 10 NCA of the genomes of novel viruses. Projections of the first two (most significant) canonical factors that differentiate host origins of the control sequences with known cellular hosts using mononucleotide and dinucleotide frequencies are shown. Points represent values for individual sequences, with 95% confidence ellipses positioned around the centroid of each group. Positions of the RNA virus sequences obtained in the current study are labeled A to L, and their provisional assignment by NCA is indicated in the color-coded key.

gent viruses, in the form of better E scores and higher percent identity, may therefore be improved as further studies contribute more viral genomes from less exhaustively sampled geographic regions.

The latest (ninth) release of the International Committee on Taxonomy of Viruses reports a total of 86 recognized viral families infecting prokaryotes and eukaryotes, plus other unassigned genera and numerous unclassified sequences (52). In this report, we identified, in highly variable proportions, sequences related to members of 29 eukaryotic viral families (BLASTx E score, $<10^{-4}$). Specific extension and genome sequencing of selected viral sequences led to the genetic characterization of multiple novel viruses, broadening the diversity of the known virosphere. The cellular hosts of many sewage-derived viruses remain unknown. On the basis of their genetic similarities to viruses with known tropism or their nucleotide composition, broad categories of hosts such as mammals, vertebrates, insects/nematodes, or plants may be inferred. Inoculation of a complex viral mixture into model animals or plants may help to determine the host range of such highly divergent viruses. The addition of highly divergent viral genomes to public databases will also facilitate the identification of previously unrecognizable viruses distantly related to these new genomes. It is also conceivable that the novel vertebrate viruses in the *Picornaviridae*, *Astroviridae*, and *Caliciviridae* families are of human origin. Knowledge of their genomes will facilitate further studies of their tropism, seroprevalence, and disease associations, as well as improve the design of consensus PCR-based assays for further viral discovery.

Sewage contains the fecal input of large numbers of local residents, some of which will be shedding diarrhea-causing viruses. Future metagenomic efforts may be conducted on larger scales, at greater frequencies, and with greater sequencing depths or targeting specific viral populations (such as polioviruses [37]) to expand the long tradition of analyzing sewage for enteric viruses. Metagenomic analysis combined with genomic characterization of viruses in sewage also has the potential to help monitor seasonal trends of enteric infections, better detect epidemic outbreaks, measure the extinction of wild-type polioviruses, as well as assist in optimizing vaccination strategies by characterizing recently replicated viral genotypes.

ACKNOWLEDGMENTS

We thank Carl J. Mason from the Armed Forces Research Institute of Medical Sciences in Bangkok, Thailand, for help with sample collection from Bangkok and Kathmandu and Rod Miller and Kenneth Lee from the San Francisco Public Utilities Commission with the San Francisco sample collection. We thank Gerardo Rafael Arguello Astorga for assistance in geminivirus sequence analyses and Gia Tung Phan, Jakk Wong, Yunhee Cha, and Shiquan Wu for laboratory assistance.

This work was supported by NIH grants R01HL083254 and R01HL105770 and funds from BSRI to E.D.

REFERENCES

1. **Abed Y, Boivin G.** 2008. New Saffold cardioviruses in 3 children, Canada. Emerg. Infect. Dis. **14**:834–836.
2. **Ahmad I, Holla RP, Jameel S.** 2011. Molecular virology of hepatitis E virus. Virus Res. **161**:47–58.
3. **Ambert-Balay K, et al.** 2008. Prevalence and genetic diversity of Aichi virus strains in stool samples from community and hospitalized patients. J. Clin. Microbiol. **46**:1252–1258.
4. **Arthur JL, Higgins GD, Davidson GP, Givney RC, Ratcliff RM.** 2009. A novel bocavirus associated with acute gastroenteritis in Australian children. PLoS Pathog. **5**:e1000391. doi:10.1371/journal.ppat.1000391.
5. **Batts W, Yun S, Hedrick R, Winton J.** 2011. A novel member of the family Hepeviridae from cutthroat trout (Oncorhynchus clarkii). Virus Res. **158**:116–123.
6. **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.
7. **Bofill-Mas S, et al.** 2006. Quantification and stability of human adenoviruses and polyomavirus JCPyV in wastewater matrices. Appl. Environ. Microbiol. **72**:7894–7896.
8. **Bofill-Mas S, Rodriguez-Manzano J, Calgua B, Carratala A, Girones R.** 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.
9. **Bosch A, Guix S, Sano D, Pintó RM.** 2008. New tools for the study and direct surveillance of viral pathogens in water. Curr. Opin. Biotechnol. **19**:295–301.
10. **Breitbart M, et al.** 2002. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. U. S. A. **99**:14250–14255.
11. **Cantalupo PG, et al.** 2011. Raw sewage harbors diverse viral populations. mBio **2**(5):e00180–11. doi:10.1128/mBio.00180-11.
12. **Chiu CY, et al.** 2008. Identification of cardioviruses related to Theiler's murine encephalomyelitis virus in human infections. Proc. Natl. Acad. Sci. U. S. A. **105**:14124–14129.
13. **Clark B, McKendrick M.** 2004. A review of viral gastroenteritis. Curr. Opin. Infect. Dis. **17**:461–469.
14. **Clarke IN, et al.** 2012. Caliciviridae, p 977–986. *In* King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (ed), Virus taxonomy: classification and nomenclature of viruses. Ninth report of the International Committee on Taxonomy of Viruses. Elsevier, San Diego, CA.
15. **Clarke IN, Lambden PR.** 1997. The molecular biology of caliciviruses. J. Gen. Virol. **78**:291–301.
16. **Clarke IN, Lambden PR, Caul EO.** 1998. Human enteric RNA viruses: caliciviruses and astroviruses, p 511–535. In Collier L, Balows A, Sussman M (ed.), Topley & Wilson's microbiology and microbial infections. Arnold, London, United Kingdom.
17. **Drexler JF, et al.** 2008. Circulation of 3 lineages of a novel Saffold cardiovirus in humans. Emerg. Infect. Dis. **14**:1398–1405.
18. **Drexler JF, et al.** 2012. Bats worldwide carry hepatitis E virus-related

viruses that form a putative novel genus within the family Hepeviridae. J. Virol. **86**:9134–9147.
19. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**:1792–1797.
20. **Farkas T, Sestak K, Wei C, Jiang X.** 2008. Characterization of a rhesus monkey calicivirus representing a new genus of Caliciviridae. J. Virol. **82**:5408–5416.
21. **Fauquet CM, et al.** 2008. Geminivirus strain demarcation and nomenclature. Arch. Virol. **153**:783–821.
22. **Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA (ed).** 2005. Virus taxonomy: classification and nomenclature of viruses. Elsevier Academic Press, San Diego, CA.
23. **Fauquet CM, Stanley J.** 2005. Revising the way we conceive and name viruses below the species level: a review of geminivirus taxonomy calls for new standardized isolate descriptors. Arch. Virol. **150**:2151–2179.
24. **Finkbeiner SR, Kirkwood CD, Wang D.** 2008. Complete genome sequence of a highly divergent astrovirus isolated from a child with acute diarrhea. Virol. J. **5**:117.
25. **Finkbeiner SR, et al.** 2009. Identification of a novel astrovirus (astrovirus VA1) associated with an outbreak of acute gastroenteritis. J. Virol. **83**:10836–10839.
26. **Fodha I, et al.** 2006. Identification of viral agents causing diarrhea among children in the eastern center of Tunisia. J. Med. Virol. **78**:1198–1203.
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. **Fry KE, et al.** 1992. Hepatitis E virus (HEV): strain variation in the nonstructural gene region encoding consensus motifs for an RNA-dependent RNA polymerase and an ATP/GTP binding site. Virus Genes **6**:173–185.
29. **Gabbay YB, et al.** 2007. First detection of a human astrovirus type 8 in a child with diarrhea in Belém, Brazil: comparison with other strains worldwide and identification of possible three lineages. Mem. Inst. Oswaldo Cruz **102**:531–534.
30. **Goyer M, Aho L-S, Bour J-B, Ambert-Balay K, Pothier P.** 2008. Seroprevalence distribution of Aichi virus among a French population in 2006-2007. Arch. Virol. **153**:1171–1174.
31. **Greninger AL, et al.** 2009. The complete genome of klassevirus—a novel picornavirus in pediatric stool. Virol. J. **6**:82.
32. **Guix S, et al.** 2002. Molecular epidemiology of astrovirus infection in Barcelona, Spain. J. Clin. Microbiol. **40**:133–139.
33. **Hedberg C.** 2011. Foodborne illness acquired in the United States. Emerg. Infect. Dis. **17**:1338.
34. **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.
35. **Holtz LR, et al.** 2009. Klassevirus 1, a previously undescribed member of the family Picornaviridae, is globally widespread. Virol. J. **6**:86.
36. **Hovi T, et al.** 2005. Environmental surveillance of wild poliovirus circulation in Egypt—balancing between detection sensitivity and workload. J. Virol. Methods **126**:127–134.
37. **Hovi T, et al.** 2012. Role of environmental poliovirus surveillance in global polio eradication and beyond. Epidemiol. Infect. **140**:1–13.
38. **Huang FF, et al.** 2004. Determination and analysis of the complete genomic sequence of avian hepatitis E virus (avian HEV) and attempts to infect rhesus monkeys with avian HEV. J. Gen. Virol. **85**:1609–1618.
39. **Jin Y, et al.** 2009. Viral agents associated with acute gastroenteritis in children hospitalized with diarrhea in Lanzhou, China. J. Clin. Virol. **44**:238–241.
40. **Jonassen CM, et al.** 2001. Comparison of capsid sequences from human and animal astroviruses. J. Gen. Virol. **82**:1061–1067.
41. **Jonassen CM, Jonassen TØ, Sveen TM, Grinde B.** 2003. Complete genomic sequences of astroviruses from sheep and turkey: comparison with related viruses. Virus Res. **91**:195–201.
42. **Jones MS, Lukashov VV, Ganac RD, Schnurr DP.** 2007. Discovery of a novel human picornavirus in a stool sample from a pediatric patient presenting with fever of unknown origin. J. Clin. Microbiol. **45**:2144–2150.
43. **Kamer G, Argos P.** 1984. Primary structural comparison of RNA-dependent polymerases from plant, animal and bacterial viruses. Nucleic Acids Res. **12**:7269–7282.
44. **Kapoor A, et al.** 2009. Multiple novel astrovirus species in human stool. J. Gen. Virol. **90**:2965–2972.
45. **Kapoor A, et al.** 2011. Characterization of a canine homolog of human Aichivirus. J. Virol. **85**:11520–11525.
46. **Kapoor A, Simmonds P, Lipkin WI, Zaidi S, Delwart E.** 2010. Use of nucleotide composition analysis to infer hosts for three novel picorna-like viruses. J. Virol. **84**:10322–10328.
47. **Kapoor A, et al.** 2010. Human bocaviruses are highly diverse, dispersed, recombination prone, and prevalent in enteric infections. J. Infect. Dis. **201**:1633–1643.
48. **Kapoor A, et al.** 2009. A newly identified bocavirus species in human stool. J. Infect. Dis. **199**:196–200.
49. **Kapoor A, et al.** 2008. A highly prevalent and genetically diversified *Picornaviridae* genus in South Asian children. Proc. Natl. Acad. Sci. U. S. A. **105**:20482–20487.
50. **Katoh K, Kuma K-i, Toh H, Miyata T.** 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. **33**:511–518.
51. **Khamrin P, et al.** 2008. Bovine kobuviruses from cattle with diarrhea. Emerg. Infect. Dis. **14**:985–986.
52. **King AM, Lefkowitz E, Adams MJ, Carstens EB (ed).** 2012. Virus taxonomy: classification and nomenclature of viruses. Ninth report of the International Committee on Taxonomy of Viruses. Elsevier, San Diego, CA.
53. Reference deleted.
54. **Knowles NJ, et al.** 2012. Picornaviridae, p 855–880. *In* King AMQ, Lefkowitz EJ, Adams MJ, Carstens EB (ed), Virus taxonomy: classification and nomenclature of viruses. Ninth report of the International Committee on Taxonomy of Viruses. Elsevier, San Diego, CA.
55. **Koci MD, et al.** 2003. Astrovirus induces diarrhea in the absence of inflammation and cell death. J. Virol. **77**:11798–11808.
56. **Koci MD, Seal BS, Schultz-Cherry S.** 2000. Molecular characterization of an avian astrovirus. J. Virol. **74**:6173–6177.
57. **Koonin EV, Dolja VV.** 1993. Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. Crit. Rev. Biochem. Mol. Biol. **28**:375–430.
58. **Koonin EV, Wolf YI, Nagasaki K, Dolja VV.** 2008. The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. Nat. Rev. Microbiol. **6**:925–939.
59. **Kosek M, Bern C, Guerrant RL.** 2003. The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000. Bull. World Health Organ. **81**:197–204.
60. **Lauber C, Gorbalenya AE.** 2012. Toward genetics-based virus taxonomy: comparative analysis of a genetics-based classification and the taxonomy of picornaviruses. J. Virol. **86**:3905–3915.
61. **Le Gall O, et al.** 2008. *Picornavirales*, a proposed order of positive-sense single-stranded RNA viruses with a pseudo-T=3 virion architecture. Arch. Virol. **153**:715–727.
62. **Legg JP, Fauquet CM.** 2004. Cassava mosaic geminiviruses in Africa. Plant Mol. Biol. **56**:585–599.
63. **Le Guyader FS, et al.** 2008. Aichi virus, norovirus, astrovirus, enterovirus, and rotavirus involved in clinical cases from a French oyster-related gastroenteritis outbreak. J. Clin. Microbiol. **46**:4011–4017.
64. **Li L, et al.** 2011. Viruses in diarrhoeic dogs include novel kobuviruses and sapoviruses. J. Gen. Virol. **92**:2534–2541.
65. **Li L, et al.** 2009. A novel picornavirus associated with gastroenteritis. J. Virol. **83**:12002–12006.
66. **Lole KS, et al.** 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol. **73**:152–160.
67. **Mansoor S, Briddon RW, Zafar Y, Stanley J.** 2003. Geminivirus disease complexes: an emerging threat. Trends Plant Sci. **8**:128–134.
68. **Ng TFF, et al.** 2011. Exploring the diversity of plant DNA viruses and their satellites using vector-enabled metagenomics on whiteflies. PLoS One **6**:e19050. doi:10.1371/journal.pone.0019050.
69. **Ng TFF, et al.** 2009. Discovery of a novel single-stranded DNA virus from a sea turtle fibropapilloma by using viral metagenomics. J. Virol. **83**:2500–2509.
70. **Ng TFF, Suedmeyer WK, Gulland F, Wheeler E, Breitbart M.** 2009. Novel anellovirus discovered from a mortality event of captive California sea lions. J. Gen. Virol. **90**:1256–1261.
71. **Ng TFF, et al.** 2011. Metagenomic identification of a novel anellovirus in Pacific harbor seal (Phoca vitulina richardsii) lung samples and its detection in samples from multiple years. J. Gen. Virol. **92**:1318–1323.
72. **Ng TFF, et al.** 2011. Broad surveys of DNA viral diversity obtained

through viral metagenomics of mosquitoes. PLoS One **6**:e20579. doi:10.1371/journal.pone.0020579.
73. **Noble RT, Fuhrman JA.** 1998. Use of SYBR green I for rapid epifluorescence counts of marine viruses and bacteria. Aquat. Microb. Ecol. **14**:113–118.
74. **Oberste MS, et al.** 1999. Typing of human enteroviruses by partial sequencing of VP1. J. Clin. Microbiol. **37**:1288–1293.
75. **Oliver SL, Asobayire E, Dastjerdi AM, Bridger JC.** 2006. Genomic characterization of the unclassified bovine enteric virus Newbury agent-1 (Newbury1) endorses a new genus in the family Caliciviridae. Virology **350**:240–250.
76. **Parsley LC, et al.** 2010. Census of the viral metagenome within an activated sludge microbial assemblage. Appl. Environ. Microbiol. **76**:2673–2677.
77. **Pham NTK, et al.** 2007. Isolation and molecular characterization of Aichi viruses from fecal specimens collected in Japan, Bangladesh, Thailand, and Vietnam. J. Clin. Microbiol. **45**:2287–2288.
78. **Phan TG, et al.** 2011. The fecal viral flora of wild rodents. PLoS Pathog. **7**:e1002218. doi:10.1371/journal.ppat.1002218.
79. **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.
80. **Polston JE, Anderson PK.** 1997. The emergence of whitefly-transmitted geminiviruses in tomato in the Western hemisphere. Plant Dis. **81**:1358–1369.
81. **Reuter G, Boldizsár Á, Kiss I, Pankovics P.** 2008. Candidate new species of kobuvirus in porcine hosts. Emerg. Infect. Dis. **14**:1968–1973.
82. **Rosario K, Nilsson C, Lim YW, Ruan YJ, Breitbart M.** 2009. Metagenomic analysis of viruses in reclaimed water. Environ. Microbiol. **11**:2806–2820.
83. **Sdiri-Loulizi K, et al.** 2008. Acute infantile gastroenteritis associated with human enteric viruses in Tunisia. J. Clin. Microbiol. **46**:1349–1355.
84. **Shan T, et al.** 2011. The fecal virome of pigs on a high-density farm. J. Virol. **85**:11697–11708.
85. **Simmonds P, Tuplin A, Evans DJ.** 2004. Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: implications for virus evolution and host persistence. RNA **10**:1337–1351.
86. **Söding J, Biegert A, Lupas AN.** 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. **33**:W244–W248.
87. **Stamatakis A.** 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**:2688–2690.
88. **Symonds EM, Griffin DW, Breitbart M.** 2009. Eukaryotic viruses in wastewater samples from the United States. Appl. Environ. Microbiol. **75**:1402–1409.
89. **Tamaki H, et al.** 2012. Metagenomic analysis of DNA viruses in a wastewater treatment plant in tropical climate. Environ. Microbiol. **14**:441–452.
90. **Tamura K, Dudley J, Nei M, Kumar S.** 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. **24**:1596–1599.
91. **Tayeb HT, Dela Cruz DM, Al-Qahtani A, Al-Ahdal MN, Carter MJ.** 2008. Enteric viruses in pediatric diarrhea in Saudi Arabia. J. Med. Virol. **80**:1919–1929.
92. **Victoria JG, et al.** 2009. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. J. Virol. **83**:4642–4651.
93. **WHO.** 2004. World health report. WHO, Geneva, Switzerland.
94. **Wilhelm OKG.** 2007. Overview of health-related water virology, p 1–25. *In* Bosch A (ed), Perspectives in medical virology, vol 17. Elsevier, Amsterdam, The Netherlands.
95. **Yamashita T, Ito M, Tsuzuki H, Sakae K.** 2001. Identification of Aichi virus infection by measurement of immunoglobulin responses in an enzyme-linked immunosorbent assay. J. Clin. Microbiol. **39**:4178–4180.
96. **Yamashita T, Sakae K, Ishihara Y, Isomura S, Utagawa E.** 1993. Prevalence of newly isolated, cytopathic small round virus (Aichi strain) in Japan. J. Clin. Microbiol. **31**:2938–2943.
97. **Yamashita T, et al.** 1998. complete nucleotide sequence and genetic organization of Aichi virus, a distinct member of the Picornaviridae associated with acute gastroenteritis in humans. J. Virol. **72**:8408–8412.
98. **Yamashita T, et al.** 2000. Application of a reverse transcription-PCR for identification and differentiation of Aichi virus, a new member of the picornavirus family associated with gastroenteritis in humans. J. Clin. Microbiol. **38**:2955–2961.
99. **Zhang T, et al.** 2006. RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol. **4**:108–118.