A diverse group of small circular ssDNA viral genomes in human and non-human primate stools

Terry Fei Fan Ng,1,2,†,‡ Wen Zhang,1,3 Jana Sachsenröder,1,4 Nikola O. Kondov,1 Antonio Charlys da Costa,1,5 Everardo Vega,6 Lori R. Holtz,7 Guang Wu,8 David Wang,9 Colin O. Stine,10 Martin Antonio,11 Usha S. Mulvaney,1 Marcus O. Muench,1,2,8 Xutao Deng,1,2 Katia Ambert-Balay,12 Pierre Pothier,12 Jan Vinjé,13 and Eric Delwart1,2,*,**

1Blood Systems Research Institute, San Francisco, 270 Masonic Ave, San Francisco, CA 94118, USA, 2Department of laboratory Medicine, University of California at San Francisco, San Francisco, CA, USA, 3Department of Microbiology, School of Medicine, Jiangsu University, Jiangsu, Zhenjiang, China, 4Federal Institute for Risk Assessment, Berlin, Germany, 5Institute of Tropical Medicine, University of Sao Paulo, Sao Paulo, Brazil, 6NCIRD, Polio and Picornavirus Laboratory Branch, Centers for Disease Control and Prevention, Atlanta, GA, USA, 7Department of Pediatrics 8Department of Molecular Microbiology, and 9Departments of Molecular Microbiology and Pathology & Immunology, Washington University in St. Louis, St. Louis, MO, USA, 10University of Maryland School of Medicine, Baltimore, MD, USA, 11Medical Research Council Unit, Banjul, The Gambia, 12National Reference Centre for enteric viruses, Dijon University Hospital, Dijon, France and 13NCIRD, National Calicivirus Laboratory, Centers for Disease Control and Prevention, Atlanta, GA, USA

*Corresponding author: E-mail: delwarte@medicine.ucsf.edu
†Present address: Division of Viral Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, GA, USA
‡http://orcid.org/0000-0002-4815-8697
§http://orcid.org/0000-0001-8946-6605
**http://orcid.org/0000-0002-6296-4484

Abstract

Viral metagenomics sequencing of fecal samples from outbreaks of acute gastroenteritis from the US revealed the presence of small circular ssDNA viral genomes encoding a replication initiator protein (Rep). Viral genomes were ~2.5 kb in length, with bi-directionally oriented Rep and capsid (Cap) encoding genes and a stem loop structure downstream of Rep. Several genomes showed evidence of recombination. By digital screening of an in-house virome database (1.04 billion reads) using BLAST, we identified closely related sequences from cases of unexplained diarrhea in France. Deep sequencing and PCR detected such genomes in 7 of 25 US (28 percent) and 14 of 21 French outbreaks (67 percent). One of eighty-five sporadic diarrhea cases in the Gambia was positive by PCR. Twenty-two complete genomes were characterized showing that viruses from patients in the same outbreaks were closely related suggesting common origins. Similar genomes were also characterized from the stools of captive chimpanzees, a gorilla, a black howler monkey, and a lemur that were more diverse than the human stool-associated genomes. The name smacovirus is proposed for this monophyletic viral clade. Possible tropism include mammalian enteric cells or ingested food components such as infected plants. No evidence of viral amplification was

© The Author 2015. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
found in immunodeficient mice orally inoculated with smacovirus-positive stool supernatants. A role for smacoviruses in diarrhea, if any, remains to be demonstrated.

**Key words:** ssDNA virus; metagenomics; deep sequencing; smacovirus; digital screening; acute gastroenteritis.

1. Introduction

Small circular, Rep-encoding, ssDNA (CRESS-DNA) genomes encode a large and rapidly expanding collection of diverse viruses (Delwart and Li 2012; Rosario, Duffy, and Breitbart 2012) that infect a wide range of hosts including vertebrates (Circoviridae), plants (Geminiviridae and Nanoviridae), crustaceans (Dunlap et al. 2013; Ng et al. 2013a), and fungi (SaHADV) (Yu et al. 2010). CRESS-DNA genomes have also frequently been detected in complex environmental samples including aquatic settings (Zawar-Reza et al. 2014), insects (Rosario et al. 2012; Dayaram et al. 2013, 2014), and animal stools (Blinkova et al. 2010; Kim et al. 2012; Sachsenröder et al. 2012; Cheung et al. 2013, 2014a, b; Sikorski et al. 2013; Smits et al. 2013; Reuter et al. 2014). Collectively these reports indicate that CRESS-DNA viruses can infect a diverse range of hosts. The genomes of most CRESS-DNA viruses are 2-3 kb in length. Most genomes are monopartite; however, nonviruses are multipartite with 6–8 segments, and some geminiviruses, such as the begomoviruses, contain bipartite genomes. CRESS-DNA genomes typically encode both a replication initiator protein (Rep) and a capsid protein (Cap), and contain a DNA stem loop structure required for the initiation of DNA replication (Stenger et al. 1991; Fontes et al. 1994). Based on the orientation of the Rep and Cap genes and the location of the stem loop, these viral genomes can be classified into 8 different genome types (Rosario, Duffy, and Breitbart 2012). Substitution rates for ssDNA genomes can be as high as 3 substitutions/site/year, approaching that of RNA viruses (Duffy, Shackelton, and Holmes 2008) and are also prone to recombination (Lefevre et al. 2009). The hosts of most CRESS-DNA viruses characterized in fecal or environmental samples remain elusive.

A group of CRESS-DNA viruses initially identified as ‘stool-associated circular virus (SCV)’ were detected in stools of chimpanzees as well as stools of pigs, turkeys, cows, and rats (Blinkova et al. 2010; Kim et al. 2012; Sachsenröder et al. 2012, 2014; Cheung et al. 2013, 2014; Sikorski et al. 2013; Reuter et al. 2014). SCV-like genomes were reported in stools of both diarrheic and healthy animals. In this report, we describe genomes of related viruses we named smacoviruses, in human and non-human primate fecal samples. Attempts to replicate the smacoviruses by inoculation of immunodeficient mice are described. The cellular tropism and role of smacoviruses in enteric diseases remain unknown.

2. Methods

2.1. Sample collection and viral metagenomics

A total of 70 human stool samples from 25 US outbreaks of unexplained diarrheal disease with typical gastroenteritis epidemiology (Kaplan et al. 1982) were analyzed by viral metagenomics. Human samples were submitted to the Centers for Disease Control and Prevention (CDC) with the corresponding dates (Table 1). Fifty-five samples from 21 French outbreaks were similarly pooled and analyzed by metagenomics (Table 1). Stools from non-human primates were collected in December 2009, from the San Francisco Zoo, including 4 samples from aye-aye (Daubentonia madagascariensis), 3 from bare-face tamarin (Saguinus bicolor), 3 from black howler monkey (Alouatta caraya), 4 from chimpanzee (Pan troglodytes), 3 from emperor tamarin (Saguinus imperator), 5 from gorilla (Gorilla gorilla), 7 from ring-tailed lemur (Lemur catta), 2 from lion-tailed macaques (Macaca silenus), 3 from mandrill (Mandrillus sphinx), 4 from patas monkeys (Erythrocebus patas), 2 from siamang (Symphalangus syndactylus), and 3 from squirrel monkey (Saimiri sp.).

Deep sequencing using the Illumina MiSeq platform (human samples) and the 454 Genome Sequencer FLX platform (primate samples) was performed on enriched viral particles according to previously described protocols (Ng et al. 2012, 2013b). Sequence data were analyzed by a customized NGS pipeline as described previously (Deng et al. 2015). Specifically, human host reads and bacterial reads were subtracted by mapping the reads to human reference genome hg19 and bacterial RefSeq genomes release 66 using bowtie2. Remaining reads were considered duplicates if position 5 to 55 from 5’ prime end were identical. One random copy of duplicates was kept. Low-sequence quality tails were trimmed using Phred quality score 10 as the threshold. Adaptor and primer sequences were trimmed using the default parameters of VecScreen (NCBI). The cleaned reads were de novo assembled using EnsembleAssembler (Deng et al. 2015). The assembled contigs, along with singlets were aligned to an in-house viral proteome database using BLASTx and E-value cutoff 0.01. The matches to viral sequences were then aligned to an in-house non-virus non-redundant (NVNR) universal proteome database using BLASTx. Hits with more significant adjusted E-value than to virus were removed. To digitally screen for smacovirus-related sequences in our in-house virome, the available 19 al DNA genomes were compared with 1.04 billion sequences using BLASTn and E-value cutoff of 0.0001. Resulting hits were analyzed manually by sequence alignment and phylogenetic analysis.

2.2. Whole genome sequencing

Since smacoviruses are encoded by small circular DNA genomes, a combination of regular and inverse PCR primers were designed to amplify the entire viral genomes in US diarrheal samples (primer sequences are provided in Supplementary Table 1). First, nucleic acids were extracted from stool samples using QIAamp Viral RNA Mini Kit (Qiagen). PCRs were performed using LA Taq (Clontech) with reagent concentrations according to the manufacturer’s instructions. PCR reactions were carried out with a ‘universal touch-down PCR’ suitable for the melting temperatures of all primers (Ng et al. 2013c), as follows: 95°C for 5 min, 45 cycles of [94°C for 1 min, 58°C–0.2°C per cycle for 1 min, 72°C for 1–3 min depending on amplicons size], followed by 72°C for 10 min.

To obtain complete genomes from NHP stool-associated smacoviruses, the extracted DNA were first randomly amplified by rolling circle amplification using Phi29 polymerase for 18 h (Illustra GenomiPhi V2 DNA Amplification Kit, GE Healthcare Biosciences). Inverse PCR was performed using abutting primers targeting individual genomes (primer sequences available upon request). The PCR products were sequenced by Sanger sequencing and primer walking.
Table 1. Virome analysis and PCR prevalence of smacovirus in 25 unexplained acute gastroenteritis outbreaks in the US and 21 outbreaks in France, showing the outbreak locations, sample size, and next generation sequencing result.

| Collection info | No. of sample | NGS info | Deep sequencing/viral metagenomics | PCR Prevalence |
|-----------------|---------------|----------|-----------------------------------|----------------|
|                 | City          | State    | Date                | Available NGS Raw reads | Virus (normalized to reads per million raw reads) | Enteric pathogens | Picobirnavirus | Other human viruses | Likely dietary | Smacovirus |
| USA             | Beachwood     | Ohio     | 2/2/2011            | 1 1 401,668             | Smacovirus (139) Norovirus (8) PBV (1) Betapapillomavirus 2 (1) GyV4 (1) | 3/6               |
|                 | St. Ignace    | Michigan | 4/5/2011            | 2 2 713,544             | HIV (4) | – |
|                 | Rockingham    | Virginia | 4/6/2011            | 3 3 585,730             | – | – |
|                 | Cottage Grove | Oregon   | 6/6/2011            | 6 4 1,605,266           | Smacovirus (139) Norovirus (8) PBV (1) | – |
|                 | Forest Grove  | Oregon   | 7/9/2011            | 2 2 499,184             | Smacovirus (1) | – |
|                 | Portland      | Oregon   | 8/10/2011           | 13 5 1,624,684          | Smacovirus (139) Norovirus (8) PBV (1) | – |
|                 | Fairfax       | Virginia | 10/17/2011          | 2 2 2,093,400           | CAV (4) | – |
|                 | Virginia Beach| Virginia | 10/22/2011          | 2 2 328,454             | – | – |
|                 | Hampton       | Virginia | 11/16/2011          | 4 2 2,400,570           | PBV (9) | – |
|                 | Henrico       | Virginia | 12/9/2011           | 2 2 279,912             | – | – |
|                 | Accomack      | Virginia | 12/19/2011          | 2 2 2,729,952           | – | Smacovirus (1) |
|                 | Albemarle     | Virginia | 12/23/2011          | 3 2 2,335,410           | Smacovirus (12) PBV (12) | – |
|                 | Cleveland     | Ohio     | 1/13/2012           | 2 2 529,962             | – | – |
|                 | Mecklenburg   | Virginia | 1/16/2012           | 5 4 1,265,224           | Smacovirus (843) | – |
|                 | Hampton       | Virginia | 1/17/2012           | 3 3 3,619,476           | PBV (2) Feline papillomavirus 2 (1) GyV4 (1) | – |
|                 | Orange        | Virginia | 1/17/2012           | 1 1 118,076             | – | CAV (17) |
|                 | York          | Virginia | 1/24/2012           | 3 3 209,234             | – | AGV2 (10) |
|                 | Albemarle     | Virginia | 2/18/2012           | 3 3 453,002             | – | 1/3 |
|                 | Chesapeake    | Virginia | 2/27/2012           | 2 2 1,496,366           | Smacovirus (33,154) PBV (351) Merkel cell polyomavirus (2) | 1/2 |
|                 | Middlesex     | Virginia | 2/27/2012           | 2 2 1,310,932           | Smacovirus (1,035) | – |
|                 | NA            | South Carolina | 3/10/2012 | 1 1 585,730 | Norovirus (4) | – |
|                 | cruise ship   | NA       | 4/4/2012            | 3 3 1,952,158           | Norovirus (6), AichiV (2) | – |
|                 | Sioux Falls   | South Dakota | 8/17/2012 | 11 11 4,524,754 | Human polyomavirus 6 (1) | – |
|                 | Cartersville  | Georgia  | 9/7/2012            | 2 2 5,712               | – | – |
|                 | Fresno        | California | 4/14/2013 | 4 4 146,772 | – | – |
| France          | Bain de       | Bretagne | Feb 2008           | 11 5 1,952,326           | Smacovirus (11241) | – |
|                 | Montluçon     | Auvergne | Jan 2009           | 5 0 –                             | – | – |
|                 | Plan de Cuques| Provence-Alpes | Apr 2009 | 4 0 –                             | – | – |
|                 | Figeac        | Midi-Pyrénées | Feb 2007 | 3 3 313,902 | – |
|                 | Saint-Etienne | Rhône-Alpes | Jun 2007 | 5 5 102,930 | – |
|                 | Dijon         | Bourgogne | Sept 2007          | 4 4 399,002             | – | Smacovirus (994) |
|                 | Sevran        | Ile de France | Feb 2008 | 4 3 233,370 | Smacovirus (254) |
|                 | Bergeserion   | Bourgogne | March 2008         | 9 4 93,098             | Smacovirus (1095) | – |
|                 | Dieuze        | Lorraine | Dec 2008           | 2 2 306,082             | Smacovirus (254) | – |
2.3. PCR screening for smacoviruses in human fecal samples

DNA was extracted from 86 US stool supernatants (70 samples were analyzed by deep sequencing and 16 samples analyzed by PCR only) using QIAamp Viral RNA Mini Kit (Qiagen). To validate that smacovirus DNA was not the result of possible contamination from DNA extraction columns (Naccache et al. 2013), we confirmed the presence of the viral DNA by PCR after extraction with a different method lacking silica (MyTaq Extract-PCR Kit, Bioline, Gilroy) for all 121 samples from France (70 samples were sequenced by NGS and an additional 51 samples by PCR only). A single round of PCR was performed using primer pairs (ScvXF and ScvXR, ScvIF and ScvIR, FscvAF and FscvAR, FscvEF and FscvER; Supplementary Table 1) targeting conserved regions of the human feces-associated smacovirus clusters. Negative water controls were included in every PCR experiment.

2.4. Case–control study

The case control study involved 85 fecal samples from cases of moderate–severe diarrhea and 89 fecal samples from controls aged 2–58 months participating in the Global Enteric Multicenter Study (GEMS) from The Gambia (Kotloff et al. 2013). The samples were PCR screened using primer pairs (ScvXF and ScvXR). Demographics of the subjects are summarized in Supplementary Table 2. Samples were collected from The Gambia between March 2008 and August 2008. The ethics committees at the University of Maryland, Baltimore, MD, USA, and at the field site, approved the study. Cases of moderate–severe diarrhea were defined as acute (onset within 7 days) with ≥3 loose stools within 24 h and 1 of the following criteria: sunken eyes, loss of skin turgor, intravenous hydration prescribed, dysentery, or admission to hospital. Controls reported no diarrhea in the prior 7 days.

2.5. Genome and phylogenetic analysis

To generate phylogenetic trees, the protein sequences were aligned using Mafft with the E-INS-I alignment strategy (Katoh et al. 2005). Bayesian inference trees were constructed using MrBayes (Huelsenbeck and Ronquist 2001). The Markov chain was run for a maximum of 1 million generations, in which every 50 generations were sampled and the first 25 percent of Markov chain Monte Carlo samples were discarded as burn-in.

Rolling circle replication (RCR) motifs were analyzed from the sequence alignment with reference genomes of circoviruses and geminiviruses. Pairwise protein identities of replicase and capsid protein sequences were calculated using the species demarcation tool software (Muhire, Varsani, and Martin 2014). Stem–loop structure was analyzed using mfold with default settings (Zuker 2003).

2.6. Inoculation of immunodeficient NGS mice

Infectivity studies of human smacovirus in immunodeficient mice were conducted with approval of the Institutional Animal Care and Use Committee at ISIS Services LLC (San Carlos, CA). Breeding pairs of immunodeficient NOD.Cg-Prkdscid Il2rgtm1Wjl/SzJ (NSG) were purchased from Jackson Laboratories (Sacramento, CA) and bred at our institute. In a barrier facility free of standard recognized murine pathogens detected using a sentinel-mouse screening program. In the first experiment, smacovirus-positive human stool samples were resuspended in sterile PBS. Host and
bacteria cells were removed using 0.2-μm filter (Millipore). An aliquot of 100 μl of the filtrate containing smacovirus was inoculated into the gastrointestinal tract of a 19-week-old immunodeficient NSG mouse by gavage feeding under light anesthesia. Fecal samples were collected and tested by PCR every day until Day 14 post-inoculation (PI). One cage change was performed at Day 3 PI.

In a second experiment, 4 mice were inoculated with similar inoculum, but 1 mouse per day was sacrificed at Days 1, 2, 3, and 4 PI, with the remaining mice’s cage changed every day. Their stool, stomach, duodenum, ileum, cecum, colon, liver, spleen, kidney, heart, lung, brain, and mesenteric lymph nodes were collected. The DNA was extracted from these samples using QIAamp Viral RNA Mini Kit, which also extracts DNA, and tested for smacovirus using the PCR assay described above.

3. Results

3.1. Deep sequencing reveals smacovirus in human diarrheal outbreaks in the US

A total of 70 samples from 25 outbreaks in the US that tested negative for norovirus (GI, GII, GIV), sapovirus, rotavirus, adenovirus gP1, astrovirus (types 1-8), and enterovirus were selected for viral metagenomics (Victoria et al. 2008; Li et al. 2010; Grard et al. 2012; Ng et al. 2012). Fecal samples were first pooled by outbreaks (Table 1) and were then processed by filtration of fecal supernatant followed by nuclease treatment of the filtrate to enrich for virus particle-associated nucleic acids. Extracted nucleic acids were then randomly amplified and prepared for Illumina sequencing. On average, 1.2 million sequence reads were generated for each outbreak (Table 1).

The viral sequences detected could be classified into 5 groups (Table 1): (I) Sequences related to a group of CRESS-DNA viruses previously reported in stools from wild chimpanzees, domesticated animals including pigs, cows, a turkey, and urban rats (Blinkova et al. 2010; Kim et al. 2012; Sachsenröder et al. 2012, 2014; Cheung et al. 2013, 2014; Sikorski et al. 2013; Reuter et al. 2014). (II) Sequences from known enteric viral pathogens, norovirus, Aichivirus, and astrovirus, presumably missed by prior testing due to low viral load or sequence divergence relative to the screening assay used. (III) Picobirnaviruses. A highly diverse virus group not associated with human diarrhea. (IV) Viral sequences not known to be associated with gastroenteritis, including HIV, polyomavirus, papillomavirus, and anellovirus. (V) Viral sequences likely originating from diet, such as chicken anemia virus and related gyroviruses. Here we investigated the genomic features and genetic diversity of the CRESS-DNA viruses in group I, here tentatively named smacoviruses (small circular genome virus), to differentiate them from other CRESS-DNA viruses.

3.2. Detection of smacovirus by digital and PCR screening of diarrheal outbreaks in France

Through the use of digital screening, we detected smacovirus sequences in previously analyzed stool samples from France. Specifically, using the US stool-associated smacovirus genomes as queries, we used BLASTn to search for closely related sequences in our in-house virome sequence database. At time of testing, the in-house virome database consisted of over 1.04 billion sequence reads generated from 40 runs of 454 pyrosequencing and 130 runs of Illumina MiSeq from fecal, respiratory, blood, and tissue samples collected from humans and animals. Smacovirus-related sequences were exclusively identified in fecal specimens from gastroenteritis outbreaks collected in France during 2005–2009. Based on this finding, we performed additional deep sequencing and/or PCR analysis on samples from 21 outbreaks from France. Samples from 14 of these outbreaks tested positive for smacovirus (Table 1 and summarized below).

3.3. Prevalence and genetic diversity of smacoviruses in AGE outbreaks

A total of 86 diarrhea samples from the US and 121 diarrhea samples from France were screened for smacovirus DNA by PCR. Out of the investigated outbreaks, smacovirus DNA was detected in samples from 7 (28 percent) of the 25 unexplained outbreaks in the US and in 14 (67 percent) of the 21 unexplained outbreaks in France (Table 1). Smacovirus DNA was detected in 1–3 samples per outbreak (Table 1. Column PCR prevalence).

A subset of 17 smacovirus PCR positive stool samples were then individually analyzed (not in pools) by deep sequencing (Table 2). Four samples also contained picobirnaviruses, a recently described virus frequently reported in stools of both healthy and diarrheic subjects that has not been associated with diarrhea (van Leeuwen et al. 2010; Ng et al. 2014). Astrovirus MLB1, a virus recently associated with diarrhea in a Kenyan pediatric population (Meyer et al. 2015), was detected in 3 subjects. Seven samples contained viruses of likely dietary origins such as chicken anemia virus, pig circovirus, and ungulate protoparvovirus 1 from pig. Seven of these 17 samples contained no other vertebrate virus. Smacovirus DNA were the only recognizable viral sequences detected in 7 samples. In some samples, smacovirus accounted for as much as 7 percent of all sequence reads indicating a high concentration of smacovirus (Sample J23 and E2623 in Table 2).

We generated 22 full human stool-associated smacovirus genomes. The sequence identity among them ranged from 64 to 100 percent and 64 to 100 percent (nt and a.a., respectively) in the Rep gene and 54–100 and 48–100 percent (nt and a.a., respectively) in the Cap gene (nt identities are shown in Fig. 1B), indicating that Cap genes were slightly more diverse than Rep genes.

Phylogenetic and pairwise analysis of the Cap gene revealed 3 major clusters (Fig. 1A). Cluster A was comprised of smacoviruses in fecal samples from France and USA. Cluster B was comprised of only French stool-associated viruses from two 2009 outbreaks. Cluster C included mostly outbreaks from the USA during 2011–2012, but also viruses from 2 French outbreaks collected as early as 2008. A single stool sample (Dieuze/3454) contained 2 smacoviruses, 1 from cluster A and 1 from cluster C. Notably, individual sequences from the same outbreak often clustered together (Fig. 1A), showing that similar smacovirus sequences were shared by different patients in the same outbreak presumably viruses originating from a common source.

The Rep gene shared similar tree topology with the Cap gene, except for 4 strains (4191, I22, H19, and 3454b)—whose Rep and Cap genes exhibited dissimilar phylogenetic clustering, suggesting possible recombination (Fig. 1A). Because I22, H19, and 3454b genomes were nearly identical, only H19 together with 4191 were analyzed for recombination.

3.4. Recombination in human stool-associated smacovirus genome

Recombination analysis (Simplot) as well as sequence alignment analysis showed that smacoviruses strain H19 is a recombinant with cluster A contributing Cap sequence, and cluster C contributing Rep sequence (Simplot and alignment analysis, Supplementary Fig. 1). A likely recombination event was also detected using RDP4 by RDP (P-value = 5.20 × 10⁻²⁵), GENECONV (P-value = 7.12 × 10⁻²⁹), Maxchi (P-value = 2.03 × 10⁻²⁹), Chimaera
| City            | State         | Date       | Sample Name | Raw reads | Sm acovirus | Enteric pathogen | Picobirnavirus | Other human viruses | Likely dietary |
|-----------------|---------------|------------|-------------|-----------|-------------|------------------|----------------|---------------------|----------------|
| Cottage Grove   | Oregon        | 6/6/2011   | B3          | 938,086   | Sm acovirus (179) |                 |                |                     |                |
| Albemarle       | Virginia      | 2/18/2012  | G16         | 674,106   | Sm acovirus (16)  |                 | PBV (2820)     |                     |                |
| Mecklenburg     | Virginia      | 1/16/2012  | H19         | 552,958   | Sm acovirus (6988)|                 |                |                     |                |
| Middlesex       | Virginia      | 2/27/2012  | J22         | 221,478   | Sm acovirus (4,700)|                 |                |                     |                |
| Chesaapeake     | Virginia      | 2/27/2012  | J23         | 1,195,568 | Sm acovirus (27,743)| PBV (323)      |                 |                     |                |
| USA             |               |            |             |           |             |                  |                |                     |                |
| Figeac          | Midi-Pyrénées | Feb 2007   | E1668       | 547,428   | Sm acovirus (18)  |                 | Anellovirus (0) |                     |                |
| Saint-Etienne   | Rhone-Alpes   | June 2007  | E2013       | 792,956   | Sm acovirus (114)| Astrovirus-MLB1(3)|             |                     |                |
| Dijon           | Bourgogne     | Sept 2007  | E2074       | 642,246   | Sm acovirus (1674)|                 |                |                     |                |
| Sevran          | Ile de France | Feb 2008   | E2548       | 1,215,354 | Sm acovirus (21669)|                 |                |                     |                |
| Bergeserin      | Bourgogne     | March 2008 | E2623       | 308,786   | Sm acovirus (375) |                 |                | Human polyomavirus (133)| CAV (42), GyV1 (7)|
| Dieuze          | Lorraine      | Dec 2008   | E3454       | 817,894   | Sm acovirus (2)  |                 |                |                     |                |
| Lutterbach      | Alsace        | Jan 2009   | E3804       | 1,032,400 | Sm acovirus (289)| Astrovirus-MLB1(6815)| PBV (4) |                     | CAV (4), Ungulate protoparvovirus 1(2) |
| Drulingen       | Alsace        | Feb 2009   | E3953       | 1,450,828 | Sm acovirus (2273)|                 |                |                     | PCV (2)        |
| Maisons         | Ile de France | March 2009 | E4191       | 754,000   | Sm acovirus (2)  |                 |                | Human polyomavirus (28)| Po-Circo-like virus (28),  |
|                 |               |            |             |           |             |                  |                |                     | Ungulate protoparvovirus 1 (7) |
|                 |               |            |             |           |             |                  |                |                     | Duck circovirus (7) |

Abbreviations are the same as Table 1. ‡ Smacovirus DNA detected by PCR only.
3.5. Smacovirus DNA rare in African human stool samples

Stool samples from 85 Gambian children with diarrhea and 89 samples from age-matched healthy Gambian children were tested by PCR. Samples were previously screened for 17 pathogens (Kotloff et al. 2013; Meyer et al. 2015). Only 1 sample from a 15-month-old boy with diarrhea was PCR positive. Based on prior microbiological testing, this sample was also positive for norovirus GII and rotavirus.

3.6. Highly diverse smacovirus genomes in non-human primate stools

In order to determine the extent of genetic diversity of smacoviruses in the stools of other primates, we collected 43 stool samples from 12 non-human primate (NHP) species without clinical diarrhea for metagenomic sequencing and PCR detection. Smacovirus genomes were detected in 4 of the 12 NHP species tested from the San Francisco Zoo.

3.7. Genomic characteristics of the smacovirus clade

Smacovirus genomes contained a set of conserved features. Each contained 2 major bi-directionally transcribed ORFs, encoding a Rep and a Cap (Fig. 2). All genomes but 1 are of similar length. Each contained 2 major bi-directionally transcribed ORFs, including one encoding a Rep and another encoding a Cap. All genomes but 1 are of similar length.
Figure 2. Schematic genome organization of smacoviruses. Representative genomes were selected from each genogroup for their genome size, ORFs (Rep, replication initiator protein; Cap, capsid protein) intergenic regions, LIR, long intergenic region; SIR, small intergenic region), and predicted stem loop structures. Stem loop structures were predicted by mfold (Zuker 2003). Three categories of stem loop configurations, stem loop type L, S, and SL, were detected in these smacovirus genomes (for example: type L, genogroup D; type S, genogroup E; type SL, genogroup C). Nucleotide motif sequence NANNNTTAC in the stem-loop in stem loop type L, homologous to those involved in the initiation of DNA replication, were labeled with black dots.
type IV organization (Rosario, Duffy, and Breitbart 2012), with the predicted stem loop located at the 3’ end of the Rep gene. The single exception was the PigSCV (JQ023166) from pig stools, which atypically belongs to type V organization since the two major ORFs were in the same orientation. The human and NHP stool-associated smacovirus genomes showed a tight genome size distribution (2452–2651 bases), as previously described for other animal-associated smacoviruses (2403–2639 bases) (Blinkova et al. 2010; Kim et al. 2012; Sachsenröder et al. 2012, 2014; Cheung et al. 2013, 2014; Sikorski et al. 2013; Reuter et al. 2014). All Rep of the NHP smacoviruses were encoded by a single continuous ORF when aligned with other Rep proteins of CRESS-DNA virus except for the black howler monkey smacovirus whose ORF was interrupted by an intron.

Phylogenetic analysis of the complete smacovirus Rep proteins showed them to be distinct from other CRESS-DNA viral families such as Circoviridae, Geminiviridae, and Nanoviridae (Fig. 3A). Human and NHP stool-associated smacovirus genomes were clustered with the previously reported “SCV” from chimpanzees, pigs, turkeys, cows, and rats (Blinkova et al. 2010; Kim et al. 2012; Sachsenröder et al. 2012, 2014; Cheung et al. 2013, 2014; Sikorski et al. 2013; Reuter et al. 2014). By Bayesian inference analysis, we determined that these smacovirus Rep sequences clustered into a monophyletic group (Fig. 3A) distinct from the other described CRESS-DNA from animals, plants, or environmental samples available in GenBank (Rosario, Duffy, and Breitbart 2012). In addition to the monophyletic nature of the smacovirus Rep gene, these viruses also shared similar genome organization and location of their stem loops (Fig. 2).

Pairwise comparison of all smacovirus Rep proteins showed sequence identities as low as 27 percent (Fig. 3B), reflecting a high level of genetic diversity. After plotting the frequency of pairwise identities of Rep sequences (Fig. 3B), we chose an arbitrary cut-off value of 60 percent identity to group Rep sequences into 11 smacovirus genogroups (A through K, Fig. 2). Viruses in most genogroups originated from the stools of a single host species (Fig. 3A) with 3 exceptions: genogroup K contained smacoviruses from stools of chimpanzee, black howler monkey, pig, and turkey; genogroup H were found in stools of pigs and rats; and genogroup G in stools of pigs and gorilla. Smacoviruses from pig stools have been the most extensively characterized (Sachsenröder et al. 2012; Cheung et al. 2013, 2014) and were represented in 8 of the 11 genogroups; whereas chimpanzee stool-associated smacoviruses were represented in 2 genogroups. All human stool-associated smacoviruses analyzed here fell in a single genogroup (Genogroup A, Fig. 2 and 3A).

CRESS-DNA viruses, such as members of the Circoviridae and Geminiviridae families, typically contain highly conserved RCR and NTPase/helicase motifs in their Rep. All 6 motifs were detected in most smacoviruses (Fig. 3C), indicating their genomes likely replicate via rolling circle amplification. There were only a few exceptions to the presence of all 6 complete motifs (ChiSCV-GT306, genome of 1198 nt; ChiSCV-DP152,
3.8. Inoculation with immunodeficient mice

We attempted to establish a small animal model for human stool-associated smacoviruses by inoculating fecal filtrates into the gastrointestinal tract of an adult immunodeficient NOD.Cg-Pkd<sup>−/−</sup>Il2rg<sup>−/−</sup> [2Drm<sup>1</sup>Wjl] (NSG) mouse by gavage feeding. The viral DNA was detected in the stools by PCR up to Day 2 PI, but was not detected at Days 3–10 PI.

Four NSG mice were similarly inoculated orally and were sacrificed at Days 1, 2, 3, and 4 PI. When their internal organs were PCR tested, only the cecum was PCR positive for the virus on Days 1 and 2 only (Supplementary Table 3). These results suggest that the human-stool-associated smacovirus either underwent low level transient replication or simply transited the gut without replication before being passively shed in the stools.

4. Discussion

We detected smacovirus DNA in 7 of the 25 (28 percent) unexplained US diarrhea outbreaks using deep sequencing and PCR. We subsequently detected smacovirus DNA in 14 of 21 (67 percent) unexplained acute gastroenteritis outbreaks in France. No other known viral enteric pathogens were detected in 14/17 of individual smacovirus-positive samples using metagenomics sequencing (3 samples also contained astrovirus ML81 which has been associated with diarrhea in Gambian children) (Meyer et al. 2015) (Table 2). When 85 stools from children with diarrhea and 89 from healthy controls from The Gambia were PCR tested, a single diarrhea case was smacovirus PCR positive—the same patient was coinfected with norovirus GI2 and rotavirus.

Smacovirus is distantly related to other CRESS-DNA viruses, such as Circoviridae, Geminivirusidae, and Nanoviridae, which include known pathogens of animals and plants. Smacovirus DNA (named SCV or small circular genome viruses in prior reports) has been reported in the stools of chimpanzees, pigs, turkeys, cows, rats, both with and without diarrhea and in sewage (Blinkova et al. 2010; Kim et al. 2012; Sachsenroder et al. 2012, 2014; Cheung et al. 2013, 2014; Sikorski et al. 2013; Reuter et al. 2014; Kraberger et al. 2015). Here, we describe smacovirus genomes in stool samples from people with diarrhea and from 4 species of captive primates.

We also demonstrate the utility of digitally screening a large collection of pre-existing metagenomics data from animal and human biological samples. Here this screening was applied in our in-house metagenomics database. Using similar approaches in regional or global networks could shorten the time required to detect wider presence of newly described viral genomes which may have been unrecognized or unreported in previous analyses.

Analysis of 22 complete smacovirus genomes in human stool samples collected from 2008 to 2012 showed that these viruses showed a high level of genetic variations in both Cap and Rep proteins while retaining overall genome organization and genome size. Plant CRESS-DNA viruses such as geminiviruses are known to generate sequence diversity via recombination (Padidam, Sawyer, and Fauquet 1999). Based on the analysis of smacovirus genomes, it is likely that recombination also plays a role in their evolution.

When smacoviruses were detected in more than one patient from the same outbreak, their genomes clustered phylogenetically. This observation supports the possibility that smacoviruses, or possibly their cellular host, were ingested from a shared contaminated source by different individuals in the same outbreak, or were spread following person-to-person transmission. Although smacoviruses were shed by people with diarrhea, it remains possible that smacoviruses infect organisms present in the gastroenteric tract or originate from ingested food including plants. Since all animal and human-associated smacovirus genomes reported have been detected in stool samples, their cellular hosts remain uncertain, as stools may also contain viruses in the gut such as bacteria, protozoas, or worms (Zhang et al. 2006; Victoria et al. 2009; Dutilh et al. 2014). The detection of an intron region in the Rep gene of the black howler monkey stool-associated smacovirus provides evidence for a eukaryotic host. An indirect line of evidence against a dietary origin of smacovirus comes from a study failing to detect bovine smacovirus DNA in animal feed, even though the virus was detected in the cows’ stools (Kim et al. 2012). Another study weighing against a common dietary origin of smacovirus comes from a controlled feeding trial of pigs where, despite being fed the same diet, only a subset of the pigs shed smacovirus in their stools (Sachsenroder et al. 2012). It is therefore conceivable, yet remains to be shown, that smacoviruses in feces originate from infections of enteric cells. While our initial attempts to infect laboratory mice with human stool-associated smacovirus were not successful, these results may reflect either its inability to cross species barrier, or its origin from a non-vertebrate host, such as plant. Further investigations will be required to resolve the tropism of smacoviruses and their possible role in gastroenteritis outbreaks in humans.

Data availability

Data are available under GenBank accession numbers KP233174–KP233194 and KP264964–KP264969.
Supplementary data

Supplementary data are available at Virus Evolution online.

Acknowledgements

We thank the San Francisco Zoo for the collection of non-human primate fecal samples. This study was supported by NHLBI R01 HL105770 grant and BSRI funds to E.D. and the Deutsche Forschungsgesellschaft to J.S. A.C. was supported by Sao Paulo Research Foundation (FAPESP), grant numbers: 2014/05211-2 and 2012/03417-7. This study was supported in part by the Children’s Discovery Institute (MD-FR-2013-292), and the Bill & Melinda Gates Foundation (OPP1016839). D.W. holds an Investigators in the Pathogenesis of Infectious Disease award from the Burroughs Welcome Fund.

Conflict of interest. None declared.

References

Blinkova, O. et al. (2010) ‘Novel Circular DNA Viruses in Stool Samples of Wild-Living Chimpanzees’, Journal of General Virology, 91: 74–86.
Bon, M. F., Posada, D., and Feldman, M. W. (2007) ‘An Exact Nonparametric Method for Inferring Mosaic Structure in Sequence Triplets’, Genetics, 176: 1035–47.
Cheung, A. K. et al. (2013) ‘A Divergent Clade of Circular Single-Stranded DNA Viruses from Pig Feces’, Archives of Virology, 158: 2157–62.
—— et al. (2014a) ‘Unique Circovirus-Like Genome Detected in Pig Feces’, Genome Announcements 2.
—— et al. (2014b) ‘Identification of Several Clades of Novel Single-Stranded Circular DNA Viruses with Conserved Stem-Loop Structures in Pig Feces’, Archives of Virology, 160: 353–8.
Dayaram, A. et al. (2013) ‘High Global Diversity of Cytoviruses Amongst Dragonflies’, The Journal of General Virology, 94: 1827–40.
—— et al. (2014) ‘Novel Circular DNA Viruses Identified in Procorcula grayi and Xanthochernes zealandica Larvae Using Metagenomic Approaches’, Infection, Genetics and Evolution, 22: 134–41.
Delwart, E., and Li, L. (2012) ‘Rapidly Expanding Genetic Diversity and Host Range of the Circoviridae Viral Family and Other Rep Encoding Small Circular ssDNA Genomes’, Virus Research, 164: 114–21.
Deng, X. et al. (2015) ‘An Ensemble Strategy that Significantly Improves De Novo Assembly of Microbial Genomes from Metagenomic Next-Generation Sequencing Data’, Nucleic Acids Research, 43: e46.
Duffy, S., Shackelton, L. A., and Holmes, E. C. (2008) ‘Rates of Evolutionary Change in Viruses: Patterns and Determinants’, Nature Reviews. Genetics, 9: 267–76.
Dunlap, D. S. et al. (2013) ‘Molecular and Microscopic Evidence of Viruses in Marine Copepods’, Proceedings of the National Academy of Sciences of the United States of America, 110: 1375–80.
Dutilh, B. E. et al. (2014) ‘A Highly Abundant Bacteriophage Discovered in the Unknown Sequences of Human Faecal Metagenomes’, Nature Communications, 5: 4498.
Fontes, E. P. et al. (1994) ‘Geminivirus Replication Origins have a Modular Organization’, The Plant Cell, 6: 405–16.
Grard, G. et al. (2012) ‘A Novel Rhabdovirus Associated with Acute Hemorrhagic Fever in Central Africa’, PloS Pathogens, 8: e1002924.
Huelsenbeck, J. P., and Ronquist, F. (2001) ‘MRBAYES: Bayesian Inference of Phylogenetic Trees’, Bioinformatics, 17: 754–5.
Kaplan, J. E., et al. (1982) ‘Epidemiology of Norwalk Gastroenteritis and the Role of Norwalk Virus in Outbreaks of Acute Nonbacterial Gastroenteritis’, Annals of Internal Medicine, 96: 756–61.
Katoh, K. et al. (2005) ‘MAFFT Version 5: Improvement in Accuracy of Multiple Sequence Alignment’, Nucleic Acids Research, 33: 511–8.
Kim, H. K. et al. (2012), ‘Identification of a Novel Single-Stranded, Circular DNA Virus from Bovine Stool’, The Journal of General Virology, 93: 635–9.
Kotloff, K. L. et al. (2013) ‘Burden and Aetiology of Diarrhoeal Disease in Infants and Young Children in Developing Countries (the Global Enteric Multicenter Study, GEMS): a Prospective, Case-Control Study’, Lancet (London, England), 382: 209–22.
Kraberger, S. et al. (2015) ‘Characterisation of a Diverse Range of Circular Rep-Encoding DNA Viruses Recovered from a Sewage Treatment Oxidation Pond’, Infection, Genetics and Evolution, 31: 73–86.
Lefevre, P. et al. (2009) ‘Widely Conserved Recombination Patterns Among Single-Stranded DNA Viruses’, Journal of Virology, 83: 2697–707.
Li, L. et al. (2010) ‘Multiple Diverse Circoviruses Infect Farm Animals and are Commonly Found in Human and Chimpanzee Feces’, Journal of Virology, 91: 74–86.
Martin, D. P. et al. (2005) ‘A Modified Bootstrap Algorithm for Automated Identification of Recombinant Sequences and Recombination Breakpoints’, AIDS Research and Human Retroviruses, 21: 98–102.
—— et al. (2011) ‘Recombination in Eukaryotic Single Stranded DNA Viruses’, Viruses, 3: 1699–738.
—— et al. (2015) ‘RPD4: Detection and Analysis of Recombination Patterns in Virus Genomes’ Virus Evolution, 1: evv003.
Meyer, C. T. et al. (2015) ‘Prevalence of Classic, MLB-Clade and VA-Clade Astroviruses in Kenya and The Gambia’, Virology Journal, 12: 78.
Muhire, B. M., Varsani, A., and Martin, D. P. (2014) ‘SDT: a Virus Classification Tool Based on Pairwise Sequence Alignment and Identity Calculation’, PLoS One 9: e108277.
Naccache, S. N. et al. (2013) ‘The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns’, Journal of Virology, 87: 11966–77.
Ng, T. F. et al. (2012) ‘High Variety of Known and New RNA and DNA Viruses of Diverse Origins in Untreated Sewage’, Journal of Virology, 86: 12161–75.
—— et al. (2013a) ‘Metagenomic Identification of a Nodavirus and a Circular ssDNA Virus in Semi-Purified Viral Nucleic Acids from the Hepatopancreas of Healthy Farfantepenaeus Duorarum Shrimp’, Diseases of Aquatic Organisms, 105: 237–42.
—— et al. (2013b) ‘Identification of an Astrovirus Commonly Infecting Laboratory Mice in the US and Japan’, PLoS One, 8: e66937.
—— et al. (2013c) ‘Distinct Lineage of Vesiculovirus from Big Brown Bats, United States’, Emerging Infectious Diseases, 19: 1978–80.
—— et al. (2014) ‘Divergent Picobirnaviruses in Human Feces’, Genome Announcements 2.
Padidam, M., Sawyer, S., and Fauquet, C. M. (1999) ‘Possible Emergence of New Geminiviruses by Frequent Recombination’, Virology, 265: 218–25.
Posada, D., and Crandall, K. A. (2001) ‘Evaluation of Methods for Detecting Recombination from DNA Sequences: Computer
Simulations’, Proceedings of the National Academy of Sciences of the United States of America, 98: 13757–62.
Reuter, G. et al. (2014) ‘Novel Circular Single-Stranded DNA Virus from Turkey Faeces’, Archives of Virology, 159: 2161–4.
Rosario, K., Duffy, S., and Breitbart, M. (2012) ‘A Field Guide to Eukaryotic Circular Single-Stranded DNA Viruses: Insights Gained from Metagenomics’, Archives of Virology, 157: 1851–71.
—— et al. (2012) ‘Diverse Circular ssDNA Viruses Discovered in Dragonflies (Odonata: Epiprocta)’, The Journal of General Virology, 93: 2668–81.
Sachsenröder, J. et al. (2012) ‘Simultaneous Identification of DNA and RNA Viruses Present in Pig Faeces Using Process-Controlled Deep Sequencing’, PLoS One, 7: e34631.
—— et al. (2014) ‘Metagenomic Identification of Novel Enteric Viruses in Urban Wild Rats and Genome Characterization of a Group A Rotavirus’, The Journal of General Virology, 95: 2734–47.
Sikorski, A. et al. (2013), ‘Novel Myco-Like DNA Viruses Discovered in the Faecal Matter of Various Animals’, Virus Research, 177: 209–16.
Smith, J. M. (1992) ‘Analyzing the Mosaic Structure of Genes’, Journal of Molecular Evolution, 34: 126–9.
Smits, S. L. et al. (2013) ‘Metagenomic Analysis of the Ferret Fecal Viral Flora’, PLoS One, 8: e71595.
Stenger, D. C. et al. (1991) ‘Repliconational Release of Geminivirus Genomes from Tandemly Repeated Copies: Evidence for Rolling-Circle Replication of a Plant Viral DNA’, Proceedings of the National Academy of Sciences of the United States of America, 88: 8029–33.
van Leeuwen, M. et al. (2010) ‘Human Picobirnaviruses Identified by Molecular Screening of Diarrhea Samples’, Journal of Clinical Microbiology, 48: 1787–94.
Victoria, J. G. et al. (2008) ‘Rapid Identification of Known and New RNA Viruses from Animal Tissues’, PLoS Pathogens, 4: e1000163.
—— et al. (2009) ‘Metagenomic Analyses of Viruses in Stool Samples from Children with Acute Flaccid Paralysis’, Journal of Virology, 83: 4642–51.
Yu, X. et al. (2010) ‘A Geminivirus-Related DNA Mycovirus that Confers Hypovirulence to a Plant Pathogenic Fungus’, Proceedings of the National Academy of Sciences of the United States of America, 107: 8387–92.
Zawar-Reza, P. et al. (2014) ‘Diverse Small Circular Single-Stranded DNA Viruses Identified in a Freshwater Pond on the McMurdo Ice Shelf (Antarctica)’, Infection, Genetics and Evolution, 26: 132–8.
Zhang, T. et al. (2006) ‘RNA Viral Community in Human Feces: Prevalence of Plant Pathogenic Viruses’, PLoS Biology, 4: 108–18.
Zuker, M. (2003) ‘Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction’, Nucleic Acids Research, 31: 3406–15.