Metagenomic Analysis of Shrew Enteric Virome Reveals Novel Viruses Related to Human Stool-Associated Viruses

Michihito Sasaki¹, Yasuko Orba¹, Keisuke Ueno², Akihiro Ishii³, Ladslav Moonga⁴, Bernard M. Hang’ombe⁴, Aaron S. Mweene⁵, Kimihito Ito² and Hirofumi Sawa¹,⁶*

¹Division of Molecular Pathobiology, Research Center for Zoonosis Control, Hokkaido University, N20, W10, Kita-ku, Sapporo 001-0020, Japan
²Division of Bioinformatics, Research Center for Zoonosis Control, Hokkaido University, N20, W10, Kita-ku, Sapporo 001-0020, Japan
³Hokudai Center for Zoonosis Control in Zambia, PO Box 32379, Lusaka, Zambia
⁴Department of Paraclinical Studies, School of Veterinary and Medicine, University of Zambia, PO Box 32379, Lusaka, Zambia
⁵Department of Disease Control, School of Veterinary and Medicine, University of Zambia, PO Box 32379, Lusaka, Zambia
⁶Global Institution for Collaborative Research and Education, Hokkaido University, N20, W10, Kita-ku, Sapporo 001-0020, Japan

*Corresponding author:
Hirofumi Sawa
Division of Molecular Pathobiology, Research Center for Zoonosis Control, Hokkaido University, N20, W10, Kita-ku, Sapporo 001-0020, Japan
Tel: +81-11-706-5185, Fax: +81-11-706-7370, E-mail: h-sawa@czc.hokudai.ac.jp
Running title: Shrew Enteric Virome

Contents Category: Other viruses

Word count of the summary: 181 words

Word count of the main text: 5360 words

Number of tables and figures: 3 tables and 4 figures

Number of supplementary figures: 1 supplementary figure

Footnote: The GenBank/EMBL/DDBJ accession numbers for the viral sequences reported in this paper are AB937980-AB937987 and AB937989-AB937991. DRA002561 is the GenBank/EMBL/DDBJ accession number for the raw sequence reads from the metagenomic library.
Summary

Shrews are small insectivorous mammals that are distributed worldwide. Similar to rodents, shrews live on the ground and are commonly found near human residences. In this study, we investigated the enteric virome of wild shrews in the genus *Crocidurinae* using a sequence-independent viral metagenomics approach. A large portion of the shrew enteric virome was composed of insect viruses, while novel viruses including cyclovirus, picornavirus and picorna-like virus were also identified. Several cycloviruses, including variants of human cycloviruses detected in cerebrospinal fluid (CSF) and stool, were detected in wild shrews at a high prevalence rate. The identified picornavirus is distantly related to human parechovirus, inferring the presence of a new genus in this family. The identified picorna-like viruses were characterized as different species of calhevirus 1, which was previously discovered in human stool. Complete or nearly complete genome sequences of these novel viruses were determined in this study and then were subjected to further genetic characterization. Our study provides an initial view of the diversity and distinctiveness of the shrew enteric virome and highlights unique novel viruses related to human stool-associated viruses.
51 **Introduction**
52
53 Shrews are small, mole-like insectivorous mammals in the family *Soricidae*, order *Soricomorpha*. Members of this family comprise at least 385 species with a nearly global distribution (Wilson & Reeder, 2011). In Africa, *Crocidura* spp. are distributed throughout the continent, including in Zambia (Dubey *et al.*, 2007). Despite their similarity in size and appearance to rodents (the order *Rodentia*), they are genetically closer to bats (the order *Chiroptera*) than to rodents (Guo *et al.*, 2013). Similar to rodents, shrews live on the ground and are commonly found near human residences. Although rodents and bats are well-known reservoirs of a number of zoonotic infectious diseases (Meerburg *et al.*, 2009; Smith & Wang, 2013), the knowledge of pathogens harbored by shrews is comparatively limited. Shrews are reservoirs of borna disease virus and a number of hantavirus species (Dürrwald *et al.*, 2014; Hilbe *et al.*, 2006; Witkowski *et al.*, 2014; Yanagihara *et al.*, 2014). Paramyxoviruses closely related to henipavirus have also been found in shrews (Sasaki *et al.*, 2014). Collectively, these studies suggest the unique viral diversity of shrews carries potential risk for public health.
54
55 Mammals are speculated to harbor at least 320,000 undiscovered viruses (Anthony *et al.*, 2013). Most emerging diseases in humans are caused by unexpected transmission by the microbial flora of wildlife and domestic animals. Therefore, to predict and manage future outbreaks, it is helpful to investigate the baseline level of viruses in animals and virus-host relationships (Mokili *et al.*, 2012; Morse *et al.*, 2012). The advent of high-throughput sequencing technology has enabled comprehensive approaches for the simultaneous detection of many viral genomes and the identification of unknown viral genomes without viral isolation (Firth & Lipkin, 2013). Using high-throughput sequencing, viral metagenomics approaches have elucidated shown enteric viromes, resulting in the
discovery of unknown viruses in a variety of mammals, including nonhuman primates, bats, pigs, rodents, cats, sea lions, martens, badgers, foxes, ferrets and pigeons (Baker et al., 2013; Bodewes et al., 2013; Dacheux et al., 2014; Donaldson et al., 2010; Ge et al., 2012; Handley et al., 2012; Li et al., 2011b; Li et al., 2010b; Ng et al., 2014; Phan et al., 2011; Phan et al., 2013a; Shan et al., 2011; Smits et al., 2013a; van den Brand et al., 2012; Wu et al., 2012). Further molecular characterization has revealed high nucleotide sequence diversity and unique genome organization of novel viruses (Boros et al., 2013; Boros et al., 2012; Li et al., 2010a; Phan et al., 2013b; Sauvage et al., 2012). To the best of our knowledge, no report of the shrew enteric virome has been described.

In the present study, we aimed to investigate the enteric viral flora of wild shrews living in close proximity to human habitation. Using a viral metagenomics approach, we identified novel viruses related to human stool-associated viruses. These viruses were subjected to further genetic characterization.
Results

Sequence data overview

Viral nucleic acids were isolated from a combined suspension of intestinal contents from 22 *Crocidura hirta* and 1 *Crocidura luna* captured at Mpolungu in the Northern Province of Zambia. Before the extraction of nucleic acids, the intestinal contents suspension was filtered and treated with nucleases to reduce incorporation of nucleic acids derived from the host and/or bacteria. For the sequencing of RNA viruses, cDNA was prepared from the isolated viral nucleic acids by reverse transcription (RT). High-throughput sequencing generated a total of 6,243,181 reads with an average length of 266 bp and a range of lengths from 8 to 624 bp. Sequence reads were compared with the NCBI nucleotide database (nt) by using BLASTN. The taxonomic content of the sequences was computed by the lowest common ancestor method in MEGAN (Huson *et al.*, 2007). As a result, 893,430 reads in total were assigned to taxonomic groups and 726,286 reads (81.2% of all the assigned sequence reads) were assigned as virus-related sequences, while 124,699 reads (14.0%) and 22,999 reads (2.6%) were related to bacteria and eukaryota, respectively (Fig. 1a), indicating that viral nucleic acids were enriched by the filtration and nuclease treatment.

Among virus-related sequence reads, 50.0% of the reads were grouped into single-stranded RNA viruses, most of which were assigned to the family *Dicistroviridae* (Fig. 1b), a family of invertebrate viruses. In addition, 15.7% of the reads were double-stranded DNA viruses, the majority of which belonged to the bacteriophage families *Siphoviridae* (8.7%), *Myoviridae* (4.0%), and *Podoviridae* (2.7%). A total of 32% of the reads were single-stranded DNA viruses from the family *Parvoviridae* (18.6%) and *Circoviridae* (13.8%). More than 99% of the reads assigned to the *Parvoviridae* family
were densoviruses, insect and crustacean parvoviruses. Very few sequences (<0.1%) corresponding to Drosophila A virus, which is a known species of double-stranded RNA virus, were detected in this experiment. A large proportion of the obtained sequences obtained were invertebrate viruses (67.9%, Fig. 1c).

No viruses with relatively high sequence identity (>90%) to known mammalian viruses were identified in this analysis. Sequences with relatively low similarity to known mammalian viruses such as human cyclovirus (Circoviridae), human parechovirus (Picornaviridae) and calhevirus 1 (CHV1, unclassified picorna-like virus) were detected and attributed to novel mammalian viruses. These sequence reads were assembled into several contigs by de novo assembly, but these contigs did not cover the overall genome sequence. Therefore, we bridged these sequence reads and contigs by conventional PCR or RT-PCR and confirmed the sequences by Sanger sequencing. The full or nearly full genome sequences of these viruses were determined and further characterized.

**Identification of novel cycloviruses**

Cycloviruses are members of the newly proposed genus *Cyclovirus* within the family *Circoviridae* (King et al., 2012; Li et al., 2010a). They have a circular single-stranded DNA genome encoding a capsid protein (Cap) gene on the virion sense and a rolling circle replication initiator protein (Rep) gene on the complementary sense (Rosario et al., 2012). On the basis of the cyclovirus-related sequence reads detected by our metagenomic analysis, we designed primers and amplified the complete genome of cycloviruses by inverse PCR. Notably, cyclovirus genomes were identified in the intestinal content of 91% (21/23) of the sampled shrews (Table 1), and dual detection of different cycloviruses was observed in three samples. Based on sequence identity, the cycloviruses we identified can be grouped
into the following five types: ZM01, ZM41, ZM36a, ZM50a and ZM62; each of these consisted of isolates sharing more than 95% nucleotide identity with the representative isolates.

The genome organization of the representative isolates is shown in Fig. 2a. All identified cyclovirus genomes, which ranged from 1,851 nucleotides to 1,865 nucleotides, contained ORFs encoding Cap, Rep and hypothetical proteins. The potential splice acceptor sequence (TTG↓GT) and donor sequence (CAG↓CA) was observed in the Rep coding region of all identified cycloviruses. Similar to known cycloviruses, the Rep proteins of the identified viruses had three conserved rolling circle amplification (RCA) motifs; RCA I (WTLNN), RCA II (HLQGFCNL) and RCA III (YCSKGGD). They also had three conserved superfamily 3 helicase motifs; the Walker A (GCTGTGKS), B (VVIDDFYGW) and C (ITSE) motifs (Dayaram et al., 2013). A putative stem-loop structure containing a highly conserved nonamer sequence (TAGTATTAC) is considered the origin of replication and was identified at the intergenic region between the 5′ ends of the Cap and Rep ORFs (Figs. 2a and 2b).

Phylogenetic analysis was performed on the basis of the amino acid sequences of full-length Rep (Fig. 2c). All identified cycloviruses clustered phylogenetically with human cyclovirus CyCV-VN and human cyclovirus VS5700009, which were initially identified in cerebrospinal fluid (CSF) from patients with suspected central nervous system (CNS) infections or unexplained paraplegia (Smits et al., 2013b; Tan et al., 2013). CyCV-VN has also been detected in the stools of healthy children (Tan et al., 2013). The International Committee on Taxonomy of Viruses (ICTV) suggests criteria for circovirus species demarcation of genome nucleotide identities of less than 75% and Cap protein amino acid identities of less than 70% (King et al., 2012). Accordingly, all isolates except for ZM36a
may be variants of CyCV-VN (Table 2). Although the isolate ZM36a exhibits relatively low sequence identity with known cyclovirus species, our phylogenetic study showed that ZM36a fell inside a cluster containing other cycloviruses we identified. Therefore, ZM36a would be considered the same species of cyclovirus as the other viruses in this cluster.

Identification of a novel picornavirus

*Parechovirus* is a genus in the family *Picornaviridae* that comprises the following 3 species: human parechovirus, Ljungan virus and Sebokele virus. Human parechovirus is a common enteric pathogen associated with gastroenteritis, respiratory illness and, rarely, more severe diseases such as myocarditis, encephalitis, pneumonia, meningitis and flaccid paralysis (Esposito et al., 2014). Ljungan virus and Sebokele virus were isolated from rodents in Sweden (Niklasson et al., 1999) and the Central African Republic (Joffret et al., 2013), respectively. We identified sequence reads distantly related to known parechoviruses and temporarily named this *Crocidura hirta*-derived picornavirus “Crohivirus 1 (CroV1)”.

In general, picornviruses have single-stranded RNA genomes encoding a single polyprotein downstream of an internal ribosome entry site (IRES) element. A nearly complete genome sequence of CroV1, 7,321 nucleotides in length, was determined, but the 5′-end sequence was not obtained by 5′ Rapid amplification of cDNA end (RACE) experiments (Fig. 3a). A single large ORF encoding a putative polyprotein of 2,170 amino acids and the initiation codon (AUG) was located in the Kozak consensus sequence (AAGAUGG) in the CroV1 genome. Potential polyprotein cleavage sites were predicted by the NetPicoRNA program (Blom et al., 1996) and multiple alignment with members of the genus *Parechovirus* (Fig. 3a). Comparison analyses of amino acid sequences identified several distinctive motifs conserved across the different genera in the family
Picornaviridae (Le Gall et al., 2008). The P1 region of CroV1 contains putative capsid proteins with the characteristic motif KxKxxRxK (x = all amino acid residues), which is conserved in human parechoviruses and Ljungan viruses but not Sebokele virus (Williams et al., 2009). The P2 and P3 regions contain non-structural proteins involved in protein processing and genome replication. The ribosomal skipping 2A sequence (DxExNPGP) was identified in the N-terminal P2 region of CroV1 as well as Ljungan virus and Sebokele virus (Luke et al., 2008). The picornavirus 2C protein belongs to the superfamily 3 helicases, and the 2C protein of CroV1 contains the conserved walker motifs (GxxGxGKS and DD) critical for the ATPase activity of the 2C protein (Sweeney et al., 2010). Consistent with all other picornaviruses, the tyrosine residue (Y) was present at position 3 of the predicted N-terminus of the 3B protein (Vpg, viral genome-linked protein); this residue is responsible for the covalent linkage of Vpg to the 5’ end of the viral RNA genome (Goodfellow, 2011). The 3C protease harbors the catalytic triad H-D-C and the conserved protease active sites GxCG and GxH (Gorbalenya et al., 1989). The 3D RNA-dependent RNA polymerase (RdRp) also harbors the highly conserved KDELR, GxPSG, YGDD and FLKR motifs (Kamer & Argos, 1984). The positions of these identified motifs are mapped on the diagram of the CroV1 genome organization shown in Fig. 3a.

SimPlot sliding window analysis revealed that 3D RdRp is relatively conserved between CroV1 and parechovirus species, while a high degree of amino acid divergence was observed in the P1 region (Fig. 3b). The pairwise amino acid identities of the P1, P2 and P3 regions of CroV1 and those of its closest relatives were as follows: 33.9% identity of the P1 region with Human parechovirus type 3, 38.2% identity of the P2 region with Sebokele virus 1, and 39.7% identity of the P3 region with Ljungan virus strain M1146 (Table 3). According to the taxonomy guidelines of the ICTV Picornaviridae Study Group
209 (http://www.picornastudygroup.com/definitions/genus_definition.htm), members of a
210 picornavirus genus share greater than 40%, 40% and 50% amino acid identity in the P1, P2
211 and P3 regions, respectively. Therefore, CroV1 is not assigned to any genus and may be a
212 member of a new picornavirus genus. Phylogenetic analysis of 3D RdRp revealed a clear
213 phylogenetic division between CroV1 and parechoviruses (Fig. 3c). CroV1 was also
214 distinct from ferret parechovirus, a recently discovered picornavirus distantly related to
215 parechoviruses (Smits et al., 2013a), and clustered with Swine pasivirus 1 and PLV-CHN,
216 which are new picornaviruses identified in piglets (Sauvage et al., 2012; Yu et al., 2013).
217
218 **Identification of novel picorna-like viruses**
219
220 In the analysis of sequence reads from high-throughput sequencing, we identified two
221 similar picorna-like virus sequences related to CHV1, a recently identified unclassified
222 picorna-like virus identified in the feces of a patient with acute flaccid paralysis (Kapoor et
223 al., 2010). We named the viruses calhevirus 2a (CHV2a, 9,837 nucleotides) and calhevirus
224 2b (CHV2b, 8,899 nucleotides). We determined large proportions of the viral genome
225 organization, including a partial ORF1 encoding a putative nonstructural polyprotein, an
226 intergenic region, an ORF2 encoding a putative structural protein, a putative ORF3 with
227 unknown function and the 3’ untranslated region (UTR) (Fig. 4a). Consistent with CHV1,
228 the putative nonstructural protein had the following characteristic motifs: a Walker A motif
229 (GxxGxGKS) and B motif (DD) for helicase activity, an H-D-S motif for protease activity
230 and highly conserved RdRp motifs (KDELR, YGDD, FLKR) (Le Gall et al., 2008). Similar
231 genome organizations have been observed in dicistroviruses, which are pathogenic
232 picorna-like insect viruses (Bonning & Miller, 2010). In the dicistrovirus genome, the IRES
233 element is present in the intergenic region between two ORFs and is characterized by
multiple stem-loops and pseudoknots (Nakashima & Uchiumi, 2009). Although a relatively longer intergenic region was present in the genomes of CHV2a and CHV2b, none of the conserved motifs of the dicistrovirus IRES element were observed.

A BLASTP search revealed that only the putative RdRp regions of CHV2a and CHV2b shared low amino acid sequence identity with members of the order Picornavirales. Therefore, phylogenetic analysis was conducted based on the RdRp region. CHV2a and CHV2b were closely related to CHV1 but distinct from all other known picorna-like viruses (Fig. 4b).

To infer the possible host(s) for CHV2a and CHV2b, we performed nucleotide composition analysis and subsequent canonical discriminant analysis (Kapoor et al., 2010; Shan et al., 2011). Analysis of the mononucleotide and dinucleotide frequencies of the viral genomes suggested that CHV2a and CHV2b, as well as CHV1, originated from arthropod hosts (Fig. S1, available in the online Supplementary Material).

Molecular screening of the viruses identified in intestinal contents and tissue samples

In addition to the aforementioned cyclovirus screening of shrew intestinal contents, we performed RT-PCR screening of the same samples to identify RNA viruses. The results are summarized in Table 1. CroV1, CHV2a and CHV2b were detected in 4-17% of the intestinal contents from the individual shrews. We further evaluated the presence of each virus in the lung, liver, spleen and kidney tissues of shrews showing a positive result in the screening test on the intestinal contents. CroV1 were detected in the liver and spleen as well as the intestinal contents. None of the other viruses were detected in tissue samples.

We then applied the viral screening test to rodent samples obtained at the same sampling occasion. Only CHV2b and various types of cycloviruses were detected in the
intestinal contents but not in rodent tissues (Table 1). Dual detection of different
cycloviruses was observed in the intestinal contents of two rodents. Of the 20 cyclovirus
sequences obtained from the intestinal contents of 18 rodents, 18 sequences corresponded
to cycloviruses ZM01, ZM41, ZM36a, ZM50a and ZM62, which were identified in shrew
intestinal contents in this study and described above. The remaining two identical
sequences, named cyclovirus ZM32, shared 92% nucleotide sequence identity with ZM50a
(Table 2) and were included in the phylogenetic analysis of cycloviruses (Fig. 2c).

Commercial columns and reagents can be unexpectedly contaminated with nucleic
acids, including circovirus-like sequences (Lysholm et al., 2012). To exclude the possibility
of false detection of viruses via contamination, the sample lysis buffers from each of the
processed nucleic acid extraction kits were used as negative control specimens. No positive
signal was detected from these controls in our molecular screening experiments.
Insect viruses constituted a large proportion of the shrew enteric virome and mainly included members of *Dicistroviridae* and *Densovirusae*. This result reflects the diets of shrews as well as insectivorous bats (Donaldson *et al.*, 2010; Ge *et al.*, 2012; Li *et al.*, 2010b). Although it remains unclear whether the novel viruses described in this study infect shrews, the detection of CroV1 from some tissues supports the hypothesis of replication in the organs of shrews (Delwart, 2013).

Cycloviruses ZM01, ZM41, ZM36a, ZM50a and ZM62 were detected in the intestinal contents of both shrews and rodents, suggesting circulation of these cycloviruses between the shrew and rodent populations. By contrast, CroV1 and CHV2a were identified in the intestinal contents of shrews but not rodents. Nevertheless, considering the influence of some biases such as the small size of the population and limited geography, the host ranges of the identified viruses remain to be determined. In this study, we also cannot exclude the possibility that the cycloviruses we identified came from common prey. Further epidemiological studies are necessary to understand the distribution and host specificity of these viruses.

Recent metagenomic studies have identified a number of cycloviruses from the feces, respiratory tract, CSF and sera of humans, bat feces, chimpanzee feces, muscle tissues of chickens, cows and goats, and insect abdomens (Dayaram *et al.*, 2013; Ge *et al.*, 2011; Li *et al.*, 2010a; Li *et al.*, 2011a; Li *et al.*, 2010b; Padilla-Rodriguez *et al.*, 2013; Phan *et al.*, 2014; Rosario *et al.*, 2011; Smits *et al.*, 2013b; Tan *et al.*, 2013). Consistent with the wide range of host animals, we found a high incidence of cycloviruses in shrews. Our phylogenetic analysis revealed that all identified sequences are closely related to cycloviruses, which were initially identified in CSF from human patients with CNS
manifestations (Smits et al., 2013b; Tan et al., 2013), raising the possibility of cross-species transmission between humans and shrews or rodents. Close sequence identity of cyclovirus species CyCV-VN has been observed between humans and domestic animals (Tan et al., 2013). Although circovirus infection causes various clinical manifestations in birds and pigs, the pathogenicity of cycloviruses remains to be determined (Delwart & Li, 2012). Therefore, it is difficult to estimate the current risk of endemic cyclovirus in shrews and rodents.

The family Picornaviridae is a highly diverse virus family comprising 26 genera (Adams et al. 2013) (http://www.picornaviridae.com), and the continuous discovery of new species has further expanded the diversity of this family (Boros et al., 2013; Boros et al., 2012; Honkavuori et al., 2011; Kapoor et al., 2008a; Kapoor et al., 2008b; Li et al., 2009; Lim et al., 2014; Ng et al., 2012; Reuter et al., 2012; Sauvage et al., 2012; Woo et al., 2012; Woo et al., 2010). A number of picornaviruses have been detected in mammalian feces, but picornavirus has not been reported in shrews. Here, we identified the novel shrew picornavirus CroV1, which is distantly related to members of Parechovirus. Our findings broaden the current knowledge of genetic diversity of Picornaviridae. Unfortunately, the complete sequence of the 5′ UTR was unavailable; therefore, the IRES element in the CroV1 genome was not characterized.

CHV1, CHV2a and CHV2b have dicistronic genomes consisting of two nonoverlapping large ORFs encoding nonstructural and structural polyproteins. A similar genome organization has been observed in some picorna-like viruses, such as members of the family Dicistroviridae, the genera Bacillarnavirus and Labyrinthavirus, and picalivirus A (Bonning & Miller, 2010; Ng et al., 2012; Shirai et al., 2006; Takao et al., 2006). However, CHV1, CHV2a and CHV2b are phylogenetically quite distinct from these viruses and
picornaviruses. CHV1 is an unclassified picorna-like virus identified in human stool.

Nucleotide composition analysis suggested that CHV1 belongs to the insect host virus group, and the detection of CHV1 in human stools was assumed to reflect insect-contaminated food intake (Kapoor et al., 2010). Interestingly, the closely related viruses CHV2a and CHV2b were identified in the intestinal contents from shrews and rodents. Given the insectivorous habit of shrews, these viruses might also reflect insect consumption. Our nucleotide composition analysis also inferred an arthropod origin for CHV2a and CHV2b. However, for rodents, the transmission route is difficult to estimate and might be similar to the human case. A subsequent survey of calheviruses or related viruses will provide insights into the distribution and host tropism of calheviruses.

In the present study, a combination strategy of viral nucleic acid enrichment and subsequent high-throughput sequencing analysis revealed the enteric virome of Crocidura spp. This initial description of the shrew enteric virome resulted in the discovery of novel viruses. Subsequent analyses yielded complete or almost complete genome sequences of these viruses and provided deep phylogenies. Consequently, these viruses can be considered novel viral species. Our study provides an initial comprehensive view of the diversity and distinctiveness of the shrew enteric virome, and also increases our understanding of the viral diversity in mammals.
Materials and Methods

Ethics Statement

Samples were collected from wild shrews and rodents with permission from the Zambia Wildlife Authority (Act No.12 of 1998). All rodents and shrews were euthanized by inhalation of diethyl ether prior to dissection.

Sample information

We captured 24 shrews and 48 rodents around houses and fields using Sherman traps and cage traps in Mpopungu, the northern province of Zambia, in 2012. After euthanasia, lung, liver, spleen, kidney and intestinal contents were collected. Species were verified based on the nucleotide sequence of the mitochondrial cytochrome b gene (Sasaki et al., 2014).

Enrichment and isolation of viral nucleic acids from shrew intestinal contents

Viral nucleic acids were isolated and enriched for high-throughput sequencing as described previously (Donaldson et al., 2010; Phan et al., 2011; Wu et al., 2012) with some modifications. In brief, aliquots of 700 μl of Hank's Balanced Salt Solution were added to the intestinal contents of each shrew (100-200 mg). The suspensions were vortexed until well-blended and were centrifuged at 10,000 × g for 3 min. Aliquots of 250 μl of each clarified supernatant were pooled and filtered through a Minisart 0.45-μm syringe filter (Sartorius) to remove unpelleted bacterial-size substances. The filtrate was concentrated and buffer-exchanged into 800 μl of fresh Hank's Balanced Salt Solution using Amicon Ultra-15 Centrifugal Filter Units with Ultracel-50 membranes (Merck Millipore). The concentrated filtrate was treated with a cocktail of nuclease enzymes consisting of 10 μl of TURBO DNase (20 U, Ambion; Life Technologies), 0.5 μl of benzonase (125 U,
Sigma-Aldrich) and 8 μl of 10 mg/ml RNase A (Roche Diagnostics) in 1× TURBO DNase buffer (Ambion) at 37 °C for 1 h to digest naked nucleic acids. Viral nucleic acids within viral capsids are resistant to nuclease digestion (Allander et al., 2001). Then, 400 μl of the sample solution was processed using a High Pure Viral Nucleic Acid kit (Roche Diagnostics) to extract nucleic acids from DNA viruses according to the manufacturer’s protocol, with the exception that 10 μg of linear polyacrylamide (Sigma) was used as the carrier instead of the carrier RNA supplied with the kit (Malboeuf et al., 2013). The remaining 420 μl of the sample solution was processed using a QIAamp Viral RNA Mini kit to extract nucleic acids from RNA viruses (Qiagen) according to the manufacturer’s protocol, with the exception that 15 μg of linear polyacrylamide was used as the carrier.

**cDNA synthesis and sequence-independent amplification**

Double-stranded cDNA was synthesized using the sequence-tagged random hexamer (5′-cgctcttccgatctNNNNN-3′) (Yozwiak et al., 2010) using the cDNA Synthesis kit (TAKARA BIO) according to the manufacturer’s protocol and then purified using the Agencourt AMPure XP kit (Beckman Coulter). Sequence-independent amplification was performed with a tag sequence primer (5′-cgctcttccgatct-3′) and Ex Taq Hot Start Version (TAKARA BIO). The PCR cycling was performed as follows: 94 °C for 1 min, followed by 30 cycles of 98 °C for 10 s, 40 °C for 30 s and 72 °C for 1 min, with a final extension at 72 °C for 5 min.

**Library preparation and high-throughput sequencing on the Ion-PGM system**

Library preparation and high-throughput sequencing were performed according to the manufacturer’s protocols provided by Ion Torrent (Life Technologies). In brief, the
extracted viral DNA sample and total amplified cDNA sample were pooled and sheared using a Covaris S2 focused-ultrasonicator (Covaris) following the 400 bp protocol. From this fragmented sample, a 400-base-read library was prepared using the Ion Plus Fragment Library kit (Ion Torrent) and E-Gel SizeSelect 2% Agarose Gels (Invitrogen; Life Technologies). Emulsion PCR was performed using the diluted library (13 pM) with the Ion PGM Template OT2 400 kit (Ion Torrent). Sequencing was performed using the Ion PGM Sequencing 400 kit, the Ion 318 Chip V2 and the Ion PGM sequencer (Ion Torrent). The raw sequence data from the metagenomic analysis have been deposited in the Sequence Read Archive of GenBank/EMBL/DDBJ (accession number DRA002561).

Taxonomic assignment
Unassembled sequence reads were compared with NCBI nucleotide database (nt) by using BLASTN (version 2.2.26+). Results with an E-value ≤ 0.0001 were selected and used for taxonomic classification by MEGAN (version 4.62.5) (Huson et al., 2007). The lowest common ancestor algorithm with parameters of minimum support = 5, minimum score = 25, top percent = 10, and win score = 0 was used to compute the taxonomic content of the sequences.

Genome sequencing of novel cycloviruses
The complete genome sequences of novel shrew and rodent cycloviruses were amplified using nucleic acids from each individual shrew or rodent sample by inverse PCR with Tks Gflex DNA polymerase (TAKARA BIO) and a primer set targeting Rep, (5’-GAGTCCCTGTCAAAGGAGGATATGA-3’) and (5’-TCKRTAAGGRTATCKGTCGAGATCTTTG-3’). Amplicons were purified using the
QIAquick Gel Extraction kit (Qiagen), cloned into the pCR4Blunt-TOPO vector (Invitrogen) and then sequenced by Sanger sequencing with primer walking. The sequence region recognized by the primer set targeting Rep was amplified by viral species-specific primers to confirm the true sequence.

**Genome sequencing of novel linear viruses**

To determine the genome sequence of CroV1, CHV2a and CHV2b in intestinal contents, the overlapping large fragments were amplified by RT-PCR using primers designed from the high-throughput sequencing reads. RACE was performed to obtain the 5′ and 3′ UTR sequences using the SMARTer RACE cDNA Amplification kit (Clontech) or an alternative strategy using the DT88 adaptor as reported previously (Li et al., 2005). All amplified fragments were sequenced by Sanger sequencing with primer walking and assembled manually using GENETYX software ver. 10 (GENETYX).

**Genetic characterization and phylogenetic analysis**

Stem-loop structures of cycloviruses with the nonamer sequence were predicted by the Mfold webserver (Zuker, 2003). SimPlot sliding window analysis was performed by SimPlot software ver. 3.5.1 with a window size of 200 amino acids and a step size of 5 amino acids (Lole et al., 1999). For phylogenetic analysis, reference sequences were obtained from the GenBank database, and multiple sequence alignments were constructed using the ClustalW and MEGA 6 packages (Tamura et al., 2013; Thompson et al., 1994). Bayesian phylogenetic analysis was performed using MrBayes software version 3.2.2 (Ronquist et al., 2012) with the WAG amino acid substitution model. The obtained trees were visualized with FigTree software, version 1.4.
Nucleotide composition analysis and canonical discriminant analysis

Nucleotide composition analysis was performed as described previously (Kapoor et al., 2010). Mononucleotide and dinucleotide frequencies for each viral sequence were obtained using the composition scan program in the SSE package (Simmonds, 2012). Canonical discriminant analysis was performed using the RAfisher2cda program (Trujillo-Ortiz et al., 2004). The genome sequences of 112 vertebrate-derived viruses, 64 arthropod-derived viruses and 171 plant-derived viruses classified as picorna-like viruses that were used as reference sequences for the analysis are listed in Shan et al., 2011.

PCR/RT-PCR screening

To screen for the identified viruses, DNA and RNA were extracted from intestinal contents suspensions using the High Pure Viral Nucleic Acid kit and High Pure Viral RNA kit (Roche Diagnostics), respectively. Tissue DNA and RNA were extracted using the QIAamp DNA Mini kit (Qiagen), a combination of TRIzol reagent and the PureLink RNA Mini kit (Ambion), or the AllPrep DNA/RNA Mini kit (Qiagen). PCR screening for novel cycloviruses was performed using the Tks Gflex DNA polymerase. The PCR cycling was performed as follows: 94 °C for 1 min, followed by 35 cycles of 98 °C for 10 s, 65 °C for 15 s and 68 °C for 1 min, with a final extension at 68 °C for 5 min. RT-PCR screening for CroV1, CHV2a and CHV2b was performed using the SuperScript III One-Step RT-PCR System with Platinum Taq (Invitrogen). The one-step RT-PCR cycling was performed as follows: 60 °C for 1 min, 50 °C for 30 min, and 94 °C for 2 min, followed by 40 cycles of 94 °C for 15 s, 56 °C for 30 s and 68 °C for 1 min, with a final extension at 68 °C for 5 min. The following primers were used in this screening experiment: (5′-
GAGTCCCTGTCAAGGAGGATATGA -3’) and (5’-
TCKRTAAGGRTATCKGTCGAGATCTTG -3’) for cyclovirus screening; (5’-
CACACTGGAAATATCGATTGAGGAAG -3’) and (5’-
CAACACAGTTGTACAAGGATCCA -3’) for CroV1 screening; (5’-
GATTGCTGCCTTAAAGTCGCTAGA -3’) and (5’-
AAATCGCCGCTTGAGAAACGTGA -3’) for CHV2a screening; and (5’-
CTCGGATGTCTTTGGAAGTGACTG -3’) and (5’-
AAGCTGCGTGTACACTTCCTCAAG -3’) for CHV2b screening. All positive
PCR/RT-PCR signals were confirmed by direct sequencing.
Acknowledgments

We thank the Zambia Wildlife Authority for their support of this research. This study was supported by the Japan Initiative for Global Research Network of Infectious Diseases (J-GRID) from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT), Japan, and Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 24405043.
References


Bodewes, R., van der Giessen, J., Haagmans, B. L., Osterhaus, A. D. & Smits, S. L.


Ng, T. F., Mesquita, J. R., Nascimento, M. S., Konodov, N. O., Wong, W., Reuter, G.,


Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S.,


**Figure Legends**

**Fig. 1. Taxonomic classification of sequence reads from shrew intestinal contents.**

The proportions of whole-sequence reads (A), viral-sequence reads (B) and the type of host predicted to be associated with the virus from which the sequence reads were derived (C) are shown in the charts. The numbers in parentheses indicate the percentage of sequence reads related to members of each taxon. dsDNA, double-stranded DNA; dsRNA, double-stranded RNA; ssDNA, single-stranded DNA; ssRNA, single-stranded RNA.

**Fig. 2. Genome organization and phylogenetic relationship of the cycloviruses identified.**

(A) Diagrams of the predicted genome organization of the identified cycloviruses. Black arrows indicate ORFs encoding rolling circle replication initiator protein (Rep). White arrows indicate ORFs encoding capsid protein (Cap). Gray arrows indicate hypothetical ORFs with unknown functions. The positions of the conserved rolling circle amplification (RCA) motifs RCA I (WTLNN), RCA II (HLQGFCNL) and RCA III (YCSKGGD) and the helicase motifs Walker A (GCTGTGKS), B (VVIDDFYGW) and C (ITSE) are indicated by black and white arrowheads, respectively. (B) Predicted stem-loop structure of cyclovirus CyCV/ZM01. The highly conserved nonamer sequence is highlighted in grey. (C) Phylogenetic analysis of the full-length Rep of representative cycloviruses and cycloviruses identified in this study. The respective accession numbers of the viral sequences are shown in parentheses. Bayesian posterior probabilities are indicated at each tree root. The scale bar represents a distance of 0.2 substitutions per site.

**Fig. 3. Genome organization and phylogenetic relationships of crohivirus 1 (CroV1).**
(A) Diagram of the predicted genome organization of CroV1. The P1 region consists of structural proteins. The P2 and P3 regions consist of nonstructural proteins. The positions of the cleavage sites in the polyprotein are indicated by white arrowheads with the nucleotide numbers. The characteristic motifs mapped in the diagram are as follows: the parechovirus-conserved motif (KxKxxRxK), the ribosomal skipping 2A motif (DxExNPGP), the helicase motifs Walker A (GxxGxGKS) and Walker B (DD), the picornavirus-conserved tyrosine residue (Y), the protease catalytic triad residues (H-D-C), the protease active motifs (GxCG and GxH) and the 3D polymerase motifs (KDELR, GxPSG, YGDD and FLKR).

(B) SimPlot sliding window analysis of CroV1 compared with human parechovirus 1 (red line), Ljungan virus 87-012 (green line) and Sebokele virus 1 (blue line). A window size of 200 amino acids and a step size of 5 amino acids were used.

(C) Phylogenetic analysis of the full-length 3D polymerase of representative picornaviruses and CroV1. The accession numbers of the picornavirus sequences are shown in parentheses. Bayesian posterior probabilities are indicated at each tree root. The scale bar represents a distance of 0.2 substitutions per site.

Fig. 4. Genome organization and phylogenetic relationships of the calhevirus identified. (A) Diagram of the predicted genome organization of calhevirus 2a. Black and white boxes show the putative nonstructural polyprotein and structural polyprotein, respectively. The gray box shows a hypothetical ORF with unknown function. The positions of the helicase Walker A (GxxGxGKS) and B motifs (DD), the protease catalytic triad residues (H-D-S) and the highly conserved RNA-dependent RNA polymerase motifs (KDELR, YGDD and FLKR) are shown. (B) Phylogenetic analysis of the predicted RNA-dependent RNA
polymerase-encoding region of calhevirus 2a, calhevirus 2b, picorna-like viruses, picornaviruses, and caliciviruses. The accession numbers of the viral sequences are shown in parentheses. Bayesian posterior probabilities are indicated at each tree root. The scale bar represents a distance of 0.2 substitutions per site.
**Table 1 - PCR/RT-PCR screening results**

The results are presented as the number of PCR or RT-PCR-positive individuals per number of shrews or rodents tested.

<table>
<thead>
<tr>
<th></th>
<th>CyCVs</th>
<th>CroV1</th>
<th>CHV2a</th>
<th>CHV2b</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Shrew samples</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intestinal content</td>
<td>21/23</td>
<td>3/23</td>
<td>1/23</td>
<td>4/23</td>
</tr>
<tr>
<td>Lung</td>
<td>0/8</td>
<td>0/3</td>
<td>0/1</td>
<td>0/4</td>
</tr>
<tr>
<td>Liver</td>
<td>0/8</td>
<td>3/3</td>
<td>0/1</td>
<td>0/4</td>
</tr>
<tr>
<td>Spleen</td>
<td>0/8</td>
<td>1/3</td>
<td>0/1</td>
<td>0/4</td>
</tr>
<tr>
<td>Kidney</td>
<td>0/8</td>
<td>0/3</td>
<td>0/1</td>
<td>0/4</td>
</tr>
<tr>
<td><strong>Rodent samples</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intestinal content</td>
<td>18/48</td>
<td>0/48</td>
<td>0/48</td>
<td>3/48</td>
</tr>
<tr>
<td>Lung</td>
<td>0/8</td>
<td>-</td>
<td>-</td>
<td>0/3</td>
</tr>
<tr>
<td>Liver</td>
<td>0/8</td>
<td>-</td>
<td>-</td>
<td>0/3</td>
</tr>
<tr>
<td>Spleen</td>
<td>0/8</td>
<td>-</td>
<td>-</td>
<td>0/3</td>
</tr>
<tr>
<td>Kidney</td>
<td>0/8</td>
<td>-</td>
<td>-</td>
<td>0/3</td>
</tr>
</tbody>
</table>
### Table 2 - Pairwise genomic nucleotide sequence identities between different cycloviruses.

<table>
<thead>
<tr>
<th>Isolates</th>
<th>Percentage of nucleotide sequence identity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZM01</td>
</tr>
<tr>
<td>CyCV/ZM01 (AB937981)</td>
<td>81.4</td>
</tr>
<tr>
<td>CyCV/ZM41 (AB937984)</td>
<td>81.4</td>
</tr>
<tr>
<td>CyCV/ZM36a (AB937982)</td>
<td>76.2</td>
</tr>
<tr>
<td>CyCV/ZM62 (AB937987)</td>
<td>78.4</td>
</tr>
<tr>
<td>CyCV/ZM50a (AB937985)</td>
<td>78.1</td>
</tr>
<tr>
<td>CyCV/ZM32 (AB937980)</td>
<td>78.9</td>
</tr>
<tr>
<td>CyCV/CyCV-VN (KFO31465)</td>
<td>78.5</td>
</tr>
<tr>
<td>CyCV/VS5700009 (NC_021568)</td>
<td>64.0</td>
</tr>
<tr>
<td>CyCV/TN18 (GQ404858)</td>
<td>67.0</td>
</tr>
</tbody>
</table>
Table 3 - Pairwise amino acid identities in the P1, P2 and P3 regions between Crohivirus 1 and related members of the family *Picornavirus*

<table>
<thead>
<tr>
<th>Genus</th>
<th>Species</th>
<th>Amino acid identities (%) with Crohivirus 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P1</td>
</tr>
<tr>
<td><em>Parechovirus</em></td>
<td>Ljungan virus 64-7855 (EU854568)</td>
<td>27.2</td>
</tr>
<tr>
<td></td>
<td>Ljungan virus 87-012 (NC_003976)</td>
<td>31.0</td>
</tr>
<tr>
<td></td>
<td>Ljungan virus M1146 (AF538689)</td>
<td>27.2</td>
</tr>
<tr>
<td></td>
<td>Ljungan virus 145SL (FJ384560)</td>
<td>31.3</td>
</tr>
<tr>
<td></td>
<td>Human parechovirus 1 (EF051629)</td>
<td>33.5</td>
</tr>
<tr>
<td></td>
<td>Human parechovirus 2 (NC_001897)</td>
<td>31.7</td>
</tr>
<tr>
<td></td>
<td>Human parechovirus 3 (GQ183028)</td>
<td>33.9</td>
</tr>
<tr>
<td></td>
<td>Human parechovirus 4 (AB433629)</td>
<td>27.8</td>
</tr>
<tr>
<td></td>
<td>Human parechovirus 5 (HQ696576)</td>
<td>26.9</td>
</tr>
<tr>
<td></td>
<td>Human parechovirus 6 (EU077518)</td>
<td>32.6</td>
</tr>
<tr>
<td></td>
<td>Human parechovirus 7 (EU556224)</td>
<td>31.9</td>
</tr>
<tr>
<td></td>
<td>Human parechovirus 8 (EU716175)</td>
<td>31.6</td>
</tr>
<tr>
<td></td>
<td>Sebokele virus 1 (NC_021482)</td>
<td>31.4</td>
</tr>
<tr>
<td><em>Parechovirus-related</em></td>
<td>Ferret parechovirus (KF006989)</td>
<td>33.8</td>
</tr>
<tr>
<td><em>Pasivirus</em></td>
<td>Swine pasivirus 1 (NC_018226)</td>
<td>32.4</td>
</tr>
<tr>
<td><em>Avihepatovirus</em></td>
<td>Duck hepatitis A virus 1 (NC_008250)</td>
<td>19.0</td>
</tr>
<tr>
<td><em>Aquamavirus</em></td>
<td>Seal picornavirus type 1 (NC_009891)</td>
<td>12.4</td>
</tr>
</tbody>
</table>
a) Partial 5'-UTR

b) Graph showing similarity between amino acid sequences

Croivirus 1 (AB937989) vs

- Human parechovirus 1 (L02971)
- Ljungan virus 87-012 (NC_003976)
- Sebokele virus 1 (NC_021482)

C) Phylogenetic tree

- Parechovirus
  - Aquamavirus
  - Avisivirus
  - Avihepatovirus
- Pasivirus
  - Hepatovirus
  - Enterovirus
  - Senecavirus
- Encephalomyocarditis virus
- Bovine hepatitis virus
- Porcine teschovirus
- Mosavirus
- Foot and mouth disease virus
- Salmonid alphavirus
- Equine rhinitis B virus
- Bovine coronavirus
- Porcine coronavirus
- Canine picobovirus
- Turkey gatovirus
- Turkey nucleovirus
- Alchi virus
- Salivirus

Nodes with branch lengths in millions of years.