Virus Evolution, 2016, 2(1): vew005

doi: 10.1093/ve/vew005
Rapid Communication

A Novel Astrovirus-Like RNA Virus Detected in Human Stool

Bas B. Oude Munnink,[1] Matthew Cotten,[2] Marta Canuti,[1,†,‡] Martin Deijs,[1] Maarten F. Jebbink,[1] Formijn J. van Hemert,[1] My V. T. Phan,[2] Margreet Bakker,[1] Seyed Mohammad Jazaeri Farsani,[1] Paul Kellam,[2,3,§] and Lia van der Hoek[1,*,**]

[1]Laboratory of Experimental Virology, Department of Medical Microbiology, Center for Infection and Immunity Amsterdam (CINIMA), Academic Medical Center of the University of Amsterdam, 1105 AZ Amsterdam, the Netherlands, [2]Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK and [3]Division of Infection and Immunity, University College London, WC1E 6BT London, UK

*Corresponding author: E-mail: c.m.vanderhoek@amc.uva.nl

†Present address: Department of Biology, Memorial University of Newfoundland, St John's, NL, Canada.

‡http://orcid.org/0000-0002-9959-128X

§http://orcid.org/0000-0003-3166-4734

**http://orcid.org/0000-0003-2803-642X

Abstract

Several novel clades of astroviruses have recently been identified in human faecal samples. Here, we describe a novel astrovirus-like RNA virus detected in human stools, which we have tentatively named bastrovirus. The genome of this novel virus consists of 6,300 nucleotides organized in three open reading frames. Several sequence divergent strains were detected sharing 67–93 per cent nucleotide identity. Bastrovirus encodes a putative structural protein that is homologous to the capsid protein found in members of the *Astroviridae* family (45% amino acid identity). The virus also encodes a putative non-structural protein that is genetically distant from astroviruses but shares some homology to the non-structural protein encoded by members of the *Hepeviridae* family (28% amino acid identity). This novel bastrovirus is present in 8.7 per cent (35/400) of faecal samples collected from 300 HIV-1-positive and 100 HIV-1-negative individuals suggesting common occurrence of the virus. However, whether the source of the virus is infected human cells or other, for example, dietary, remains to be determined.

Key words: astrovirus; hepatitis E virus; virus discovery; next generation sequencing; novel virus; bastrovirus.

1 Introduction

Human astrovirus infections account for up to 10 per cent of the sporadic diarrhoea cases in children globally (Nguyen et al. 2008; Soares et al. 2008). Human astroviruses belong to the unassigned family of *Astroviridae*, genus *Mamastrovirus* and are single-stranded positive-sense RNA viruses. The 6–8 kb virus genome is organized in three open reading frames (ORFs) encoding a serine protease (ORF1a), an RNA-dependent RNA polymerase (RdRp) (ORF1b), and a structural protein (ORF2) (Mendez 2007). Within the *Astroviridae,* the 3'-domain of the gene encoding the structural protein is conserved, while the 5'-end is more variable (Wang et al. 2001). Astroviruses translate ORF1b via ribosomal frame shifting, and their structural proteins lack a helicase domain (Jiang et al. 1993). Recombination among astroviruses may occur at a breakpoint hotspot at the junction between ORF1b and ORF2 (Walter et al. 2001).

Following the first identification of human astrovirus in 1975 (Madeley and Cosgrove 1975), eight serotypes of human astrovirus have been defined (human astroviruses 1–8, Kjeldsberg 1994). Two new clades of astroviruses have been described in stool samples from patients with diarrhoea (Finkbeiner, Kirkwood, and Wang 2008; Finkbeiner et al. 2009b). The first novel clade of astroviruses consists of the human, mink, and ovine-like astrovirus (HMOAstV) species A–C and the VA1–VA3 species. VA1 was identified in an outbreak of acute gastroenteritis in the USA (Finkbeiner et al. 2009b), and the VA2 and VA3 species were found in stool samples from children with diarrhoea in India (Finkbeiner et al. 2009a). The HMOAstV-A, -B, and -C were identified in stools from Nigeria, Pakistan, and Nepal and were found in adults and children with and without diarrhoea (Kapoor et al. 2009). This clade is not confined to human astroviruses, animal astroviruses also cluster within this clade (Kapoor et al. 2009). The second novel clade consists of the MLB1 and MLB2 astroviruses identified in stool samples of paediatric diarrhoea patients from Australia and India, respectively (Finkbeiner, Kirkwood, and Wang 2008; Finkbeiner et al. 2009a). Recently, astrovirus MLB3 and VA4 were discovered in stool samples from children in India (Jiang et al. 2013) and astrovirus VA5 was discovered in Gambia (Meyer et al. 2015), further expanding the two novel clades.

Human astrovirus 1–8 infection is typically associated with diarrhoea. Kurtz et al. (1979) showed that a filtrate from faecal material with astrovirus can lead to diarrhoeal illness and shedding of large amounts of astrovirus in faeces (Kurtz et al. 1979). For the other clades of human astroviruses, the association between the virus and diarrhoeal illness is less clear. MLB3 was found in four diarrhoea cases and in one asymptomatic person, and VA4 was found in two diarrhoeal patients, but no significant association with diarrhoea was found in a cohort consisting of 400 diarrhoeal patients and 400 healthy children (Jiang et al. 2013). In another study in Kenya and Gambia, MLB1 was found to be significantly associated with diarrhoea, while MLB3 and human astrovirus 1–8 were not associated with diarrhoea (Meyer et al. 2015). Animal astrovirus infections generally manifest as a relatively mild enteric disease (Moser and Schultz-Cherry 2005), but astrovirus infection in ducks may lead to fatal hepatitis (Fu et al. 2009).

Hepatitis E viruses are single-stranded, positive sense RNA viruses that can cause liver disease. Hepatitis E virus is part of the unassigned family of *Hepeviridae* (Pringle 1998) and has a genome length of around 7.5 kb, which is organized in three ORFs which code for non-structural proteins (ORF1), structural proteins (ORF2), and regulatory proteins (ORF3) (Tam et al. 1991; Chandra et al. 2011). Hepatitis E virus is transmitted via the faecal-oral route (Balayan et al. 1983) and is estimated to infect 3.4 million individuals per year in developing countries (Kamar et al. 2014). Four genera of *Hepeviridae* have recently been proposed: *Orthohepevirus*, *Chiropteranhepevirus*, *Avihepevirus,* and *Piscihepevirus* (Meng 2013). Only members of the genera *Orthohepevirus*, consisting of four genotypes, have been detected in humans. Genotypes 1 and 2 are exclusively found in humans, are mainly prevalent in developing countries, and are spread via faecal-oral contact of contaminated water. Genotypes 3 and 4 are mainly found in developed countries, are of zoonotic origin, and are spread via faecal-oral contact of infected pig meat, direct exposure, or contaminated water (Cao and Meng 2012).

In this study, a search for novel viruses using next-generation sequencing data obtained from stools of predominantly healthy individuals is described and several novel astrovirus-like RNA viruses, tentatively named bastroviruses (basal astrovirus) are reported.

2 Materials and methods

2.1 Sample collection

Two hundred faecal samples collected in 1984 and 1985 from participants of the Amsterdam Cohort Studies on HIV-1 infection and AIDS were included in this study. HIV-1-positive ($n = 100$) and HIV-1-negative ($n = 100$) men having sex with men were part of the study. Furthermore, 200 samples from HIV-1-infected patients visiting the outpatient clinic at the Academic Medical Centre in 1994 and 1995 were included. Informed written consent was obtained from all participants of this study, and the studies were approved by the Medical Ethical Committee of the Academic Medical Center, Amsterdam. All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2000. At the time of collection, faecal samples were diluted 1:3 in broth (containing nutrient broth no. 2 supplied by Oxoid, 500 IU penicillin per ml, 500 μg streptomycin, and 3 μg amphotericin B per ml) and stored at –80 °C.

2.2 Sequencing of stool samples

Samples were deep sequenced using the Viseq method (Cotten et al. 2014). Briefly, 150 μl of faecal suspension was centrifuged for 10 min at 10,000 × *g*, and a DNase treatment (20 U TURBO DNase, Ambion) was performed on the supernatant. Nucleic acids were extracted using the Boom method (Boom et al. 1990), followed by reverse transcription with Superscript II (Invitrogen) and Endoh primers (Endoh et al. 2005). Second strand synthesis was performed using Klenow polymerase (5U, Invitrogen) followed by phenol/chloroform/isoamylalcohol extraction and ethanol precipitation.

Illumina sequencing was performed by the standard methods for the paired-end Illumina MiSeq library preparation. Each sample was sheared and fractionated to an average length of 400–500 bp after which adaptors with sample-specific barcodes were ligated. Samples were polymerase chain reaction (PCR) amplified and sequenced with an Illumina MiSeq instrument (Cotten et al. 2014) to generate 149 nt paired end reads.

2.3 Sequence analysis and statistical analysis

Adapter sequences were removed, and sequence reads were trimmed to a median Phred score of thirty and minimum length of 125 nucleotides using QUASR (Watson et al. 2013). Reads that passed quality control filter were *de novo* assembled using SPAdes version 3.5.0 (Bankevich et al. 2012) followed by improve_assembly (2015). The resulting contigs were compared to all entries of the GenBank non-redundant database (Benson et al. 2010) using the local NCBI BLAST tool from NCBI (Altschul et al. 1990). The following settings were used: expect threshold 1,000, Match/Mismatch Scores 1/-1, Gap Costs: Existence 2 and Extension 1. Among the contigs several bastrovirus genomes were present with two overlapping ORFs. The overlap between the ORF1/ORF2 junction was confirmed by specific PCRs (for each of seven viruses) aiming at this region and Sanger sequencing.

Conserved functional domains in the bastroviruses encoded proteins were determined via Pfam analysis (Finn et al. 2014) and with a search against the NCBI conserved domain database (Marchler-Bauer et al. 2015). Simplot analysis (Lole et al. 1999) was performed to visualize the nucleotide identity of bastroviruses along the genome. The GC content of bastroviruses, human astroviruses, and hepatitis E viruses was determined using Geneious 8.1.7 software (Kearse et al. 2012). Statistical analysis was performed with the Mid-P exact test, using the two by two table from OpenEpi (Sullivan, Dean, and Soe 2009).

2.4 Phylogenetic analysis and antigenic epitope prediction

All complete structural and non-structural refseq and Swiss-prot sequences from the *Hepeviridae* and the *Astroviridae* from GenBank (retrieved on 20 April 2015) were aligned using Cobalt software from the NCBI (Papadopoulos and Agarwala 2007). Phylogenetic maximum-likelihood trees were constructed with MEGA software version 6.06 (Tamura et al. 2013) using the model that was predicted to be the best model fitting after performing a model selection test. To test robustness of the evolutionary analysis, a bootstrap analysis of 1,000 replicates was performed. Identity plots for the bastroviruses were made with BioEdit software version 7.2.5 (Hall 1999). Antigenic epitopes were predicted using the SVMTrip tool (Yao et al. 2012). Only antigenic epitopes larger than ten amino acids present in at least three bastroviruses were reported.

2.5 Codon usage

Condon usage was characterized by means of plotting the effective number of codons (ENC values) of bastrovirus, astrovirus, and hepatitis E virus genes versus their GC content at the third codon positions (GC3-values): the ‘Nc-plot’ (Wright 1990). A continuous line indicates theoretical ENC values with random codon usage as a function of GC3. Deviation from this line in the direction of lower ENC values points to translational selection acting in favour of a preferred set of codons, as has been described for highly expressed genes in yeast (Bennetzen and Hall 1987) and *Escherichia coli* (Sharp and Li 1987). Calculations were performed in Excel.

2.6 Bastrovirus diagnostics

To screen samples for the presence of bastrovirus, nucleic acids were isolated via Boom isolation, and reverse transcription was performed with MMLV-Reverse Transcriptase, as previously described (de Vries et al. 2011). A nested PCR was developed targeting the conserved 5’-end of the bastrovirus genome. Amplifications were performed using Dreamtaq DNA polymerase (Thermo Scientific). The first amplification was performed using primers BV_1FW (5’-TCCGGGTTCTCMVTGAYCTC-3’) and BV_1RE (3’-GGYCKGGGSTCRATCTGG-5’), following thermal cycling profile of 5 min at 95 °C, 40 cycles of 95 °C for 30 s, 55 °C for 30 s, and 72 °C for 1 min followed by a final elongation step of 7 min at 72 °C. A nested PCR was performed using primers BV_1FW_NE (5’-CGGCBTGGYACCTRYTGTC-3’) and BV_1RE_NE (3’-ATCTGGATGGTGTAGAACCA-5’) and with the cycling profile: 5 min at 95 °C, 25 cycles of 95 °C for 30 s, 55 °C for 30 s, and 72 °C for 1 min followed by a final elongation step of 7 min at 72 °C. PCR products were visualized via agarose gel electrophoreses, and sequencing of PCR amplicons was performed using the BigDye Terminator v1.1 protocol (ABI life science). To exclude that extraction reagents were the source of the virus, seven bastrovirus-positive samples were also isolated with MagNA Pure nucleic acid isolation (Roche).

2.7 Accession numbers

The genomic sequences of the bastroviruses have been deposited in the GenBank database under the accession numbers KU318315–KU318321.

3 Results

A search for astroviruses was performed on 200 randomly chosen faecal samples of predominantly healthy HIV-1-positive ($n = 100$) and HIV-1-negative ($n = 100$) men having sex with men (Amsterdam Cohort studies on HIV-1 infection and AIDS, de Wolf et al. 1988). The RNA from these stool samples was converted to DNA with random priming, and the resulting material was sequenced to high coverage with Illumina MiSeq. Short read sequences were *de novo* assembled, and resulting contigs were examined for homology to known viruses. Several novel astrovirus-like virus genomes were identified that showed partial homology to both astrovirus and hepatitis E virus reference genomes. The agent, subsequently named bastrovirus, was only detectable in faecal samples after reverse transcription and was still detectable in faecal samples after passage through a 0.2 μm filter, indicating that bastrovirus is virus-like in size and is encoded by an RNA genome (data not shown). Since another recombinant virus has recently been detected in silica that is used to isolate nucleic acid (Naccache et al. 2013; Xu et al. 2013; Smuts et al. 2014), seven bastrovirus-positive samples were also extracted using the MagNA Pure method followed by PCR and sequencing. In all cases, the patient-specific virus was detected, thereby excluding that bastrovirus originates from treatment ingredients.

3.1 Complete genome sequence analysis

The complete genome of bastrovirus assembled from the short-read data encodes three predicted ORFs. The largest ORF (ORF1, 1,280 amino acids) most likely encodes for non-structural proteins (see below), whereas the second ORF (ORF2, 666 amino acids) most likely encodes for a protein containing homology to a conserved astrovirus capsid protein precursor domain (Fig. 1 and Table 1). A third small ORF (ORF3, 137 amino acids) of unknown function is located at the 3’-end of the genome.

Complete coding regions for the putative structural and non-structural protein of bastrovirus were obtained from seven independent faecal samples (see Table 1 for virus characteristics). All genomes shared the same genomic organization and displayed 67–93 per cent identity at amino acid level in the putative non-structural ORF1 protein and 73–98 per cent amino acid identity in the putative structural ORF2 protein (Table 2). All bastroviruses genomes encoded similar conserved amino acid domains, but regions flanking these conserved domains showed more variability (Supplementary Fig. S1). The first 40 N-terminal residues and the last 242 C-terminal residues of the capsid protein of bastroviruses are highly variable (Supplementary Fig. S2). Several antigenic epitopes with a length over ten amino acids could be identified and those were particularly concentrated at the C-terminal side of the putative

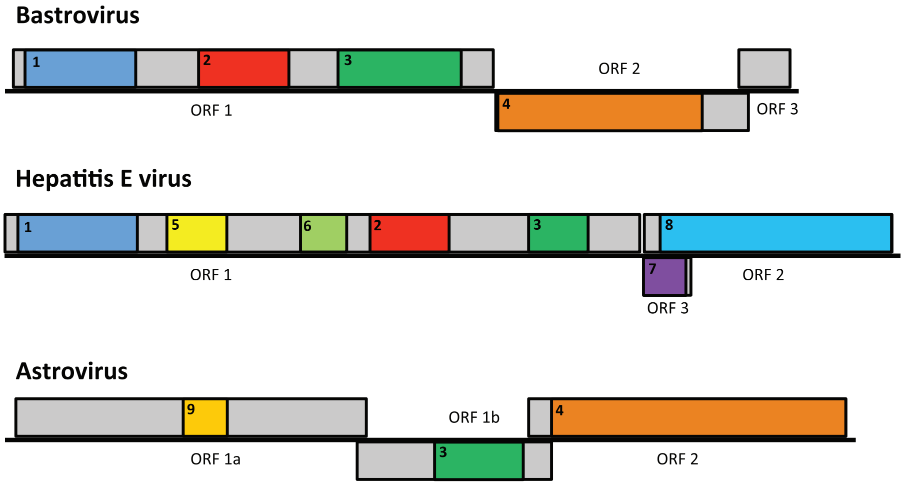

Figure 1. Genome organization and position of conserved domains of bastrovirus, hepatitis E virus (NC_001434.1), and astrovirus (NC_001943.1). Conserved domains were determined using a NCBI conserved domain search in combination with a Pfam conserved domain search (Finn et al. 2014; Marchler-Bauer et al. 2015). Dark blue (1) indicates the viral methyltransferase, red (2) the viral (superfamily 1) RNA helicase, dark green (3) the RdRp, orange (4) the astrovirus capsid protein precursor, light yellow (5) the Y-domain, light green (6) the Hepatitis E papain-like cysteine protease, purple (7) the Hepatitis E virus putative capsid domain, light blue (8) the Hepatitis E virus structural protein 2, and dark yellow (9) the trypsin-like peptidase domain.

Table 1. Characteristics of the different bastroviruses and the amount of sequence reads derived from bastrovirus.

Virus	Length (nt)	ORF1 (AA)	ORF2 (AA)	ORF3 (AA)	Amount of sequence reads derived from bastrovirus	Genome coverage
Bastrovirus 1	6,207	1,280	666	146	778	18.80x
Bastrovirus 2	6,087	1,280	666	96	2,205	54.34x
Bastrovirus 3	6,336	1,279	674	138	3,869	91.60x
Bastrovirus 4	6,300	1,279	670	137	2,819	67.11x
Bastrovirus 5	6,040	1,279	671	76	560	13.91x
Bastrovirus 6	6,017	1,282	666	70	811	20.22x
Bastrovirus 7	6,339	1,279	670	137	636	15.05x

Table 2. Amino acid identities between the coding regions among the identified bastroviruses (BV) proteins.

Putative non-structural protein coding region	BV-1	BV-2	BV-3	BV-4	BV-5	BV-6	BV-7
Bastrovirus 1	x	0.939	0.636	0.746	0.755	0.756	0.745
Bastrovirus 2		x	0.642	0.746	0.761	0.752	0.749
Bastrovirus 3			x	0.642	0.653	0.647	0.647
Bastrovirus 4				x	0.935	0.853	0.974
Bastrovirus 5					x	0.852	0.938
Bastrovirus 6						x	0.853
Bastrovirus 7							x
Putative structural protein coding region							
Bastrovirus 1	x	0.898	0.768	0.835	0.845	0.840	0.838
Bastrovirus 2		x	0.727	0.787	0.792	0.787	0.787
Bastrovirus 3			x	0.758	0.760	0.756	0.757
Bastrovirus 4				x	0.931	0.868	0.977
Bastrovirus 5					x	0.876	0.936
Bastrovirus 6						x	0.868
Bastrovirus 7							x

capsid protein, whereas no antigenic epitopes were identified at the N-terminal part (Supplementary Fig. S2).

3.2 Prevalence

A sensitive nested bastrovirus PCR targeting the conserved 5’-region of the genome revealed that 32 out of 200 HIV-1-positive ($n = 100$) and HIV-1-negative ($n = 100$) individuals in this cohort contained the virus in their stool. There were no associations of bastrovirus content with diarrhoea or other reported clinical symptoms or disease (Table 3). However, because the number of diarrhoeal cases was low in this cohort ($n = 3$), the survey was expanded to include faecal samples from 200 additional patients collected in 1994–5 of which twenty-nine had diarrhoea

(Oude Munnink et al. 2014). Bastrovirus sequences were detected in three individuals of the 200, all diarrhoea-free.

3.3. Genetic relatedness to other known viruses

The genome organization of bastrovirus differs from members of the *Astroviridae* family. Astrovirus genomes contain ORF1a, ORF1b, and ORF2, with partial overlap between ORF1a and ORF1b (both encoding non-structural proteins). The second ORF (ORF1b) in astroviruses is translated via ribosomal frameshifting (Jiang et al. 1993). However, ORF1 of bastrovirus is predicted to be translated as one large precursor protein. The predicted domains of ORF1, conserved among the bastroviruses, were not found in members of the *Astroviridae* and an identity search revealed that these domains showed homology to members of the *Hepeviridae* (ranging between 25% and 35% amino acid identity, Table 4). Both bastrovirus and hepatitis E virus encode putative viral methyltransferase, putative viral RNA helicase, and putative RdRp domains, but bastrovirus ORF1 seems to lack the Y-domain and papain-like cysteine protease domain that are present in ORF1 of hepatitis E virus (Fig. 1). To exclude that the genome organization of bastrovirus was the result of artificial *de novo* assembly, the ORF1/ORF2 connection was confirmed by PCR and Sanger sequencing for all seven bastroviruses.

The putative structural protein encoded by the bastrovirus ORF2 was 666 and 674 amino acids long and showed 44–45 per cent amino acid identity to astrovirus ORF2 protein. Phylogenetic analysis based on the amino acid sequence of this putative structural protein, focusing on the more conserved amino acid residues 1–424, showed clustering of bastroviruses within the *Astroviridae*. The bastrovirus ORF2 clearly groups with other human astroviruses with the closest relatives being the recently discovered human MLB astroviruses (Fig. 2).

Table 3. Clinical data recorded for the cohort members and the number of cases in which bastrovirus was identified.

Symptoms/disease	Bastrovirus + ($n = 32$)	Bastrovirus − ($n = 168$)	*P*
HIV infection	44% (14)	51% (86)	0.45
Tiredness	0% (0)	3.0% (5)	0.41
Fever	3.1% (1)	1.8% (3)	0.63
Night sweats	6.3% (2)	11% (18)	0.48
Nausea	0% (0)	1.2% (2)	0.70
Diarrhoea	3.1% (1)	1.2% (2)	0.48
Weight loss	3.1% (1)	1.2% (2)	0.48
Cough	0% (0)	13% (22)	0.02
Wheeze	13% (4)	6.0% (10)	0.97
Skin disorders	13% (4)	4.2% (7)	0.1
Hepatitis	22% (7)	14% (24)	0.29
Cold sore	6.3% (2)	14% (23)	0.26
Human herpesvirus 3	6.3% (2)	9.5% (16)	0.60
Fungal infection	6.3% (2)	3.6% (6)	0.41
Human herpesvirus 2	3.1% (1)	15% (26)	0.05

Phylogenetic analysis based on amino acid sequences of the putative RdRp protein, considered the most conserved part of the putative non-structural protein, showed that bastroviruses are distinct but related to astrovirus and hepatitis E virus (Fig. 3). In support of this, phylogenetic analyses based on the amino acid sequences of the individual conserved domains (the viral methyltransferase and the viral helicase) consistently show segregation of bastroviruses and members of the *Hepeviridae* into separate groups (Supplementary Figs S3 and S4).

It was not possible to perform a recombination analysis to verify whether a recombination event between ancient viruses was at the origin of bastroviruses since bastroviruses, astroviruses, and hepatitis E viruses are too divergent from each other. As an alternative, the nucleotide composition of bastrovirus was determined and compared to the nucleotide composition of astrovirus and hepatitis E virus. Nucleotide compositional analysis (Supplementary Table S1) revealed a preference for C&G nucleotides at the expense of A&U in the two ORFs of bastrovirus. In each ORF, the ENC can be plotted versus the G+C proportion at the third codon position in bastrovirus, astrovirus, and hepatitis E virus, to determine if the viral characteristics may confirm our hypothesis of recombination (Bennetzen and Hall 1987; Sharp and Li 1987); however, the GC3 content and the nucleotide composition in both bastrovirus ORFs are nearly identical (Fig. 4), making it unlikely that bastrovirus is the result of a recent recombination event of astrovirus and hepatitis E virus.

4 Discussion

Similarities between *Astroviridae* and *Hepeviridae* have been described in literature (Dryden et al. 2012). Originally, hepatitis E virus was classified as a member of the *Caliciviridae* based on morphological similarities but the virus was reclassified to the currently unassigned family *Hepeviridae* (Pringle 1998). Recently, the three-dimensional structure of the astrovirus capsid has revealed similarities with the capsid shells and the dimeric spikes of hepatitis E viruses capsids (Dryden et al. 2012). Therefore, it has been suggested that these two viral families are related and may share common capsid assembly and activation mechanisms (Dryden et al. 2012) suggesting a distant common ancestor. The novel bastrovirus described here may be related to this common ancestor. The putative bastrovirus ORF2 capsid protein shares limited amino acid identity with the *Astroviridae,* while the putative non-structural ORF1 protein has no recognizable similarity to this family. Instead the amino acid sequence of the ORF1 non-structural protein in bastroviruses is predicted to contain functional domains that are found in members of the *Hepeviridae*. The similarities of bastrovirus with astrovirus and hepatitis E virus suggest an exchange of sequences early in bastroviral evolution. Subsequently, bastrovirus has evolved during

Table 4. Conserved domains identified in the amino acid sequence of the bastrovirus structural and non-structural proteins.

ORF	Identified domain	Description	Start (AA)	End (AA)	AA identity (HepE)	AA identity (Astro)
ORF1	Vmethyltranf	Viral methyltransferase	37	333	25%	NA
ORF1	Viral_helicase1	Viral (superfamily 1) RNA helicase	503	740	35%	NA
ORF1	RdRP_2	RNA-dependent RNA polymerase	956	1.237	30%	NA
ORF2	Astro_Capsid	Astrovirus capsid protein precursor	5	504	NA	45%

NA, not applicable.

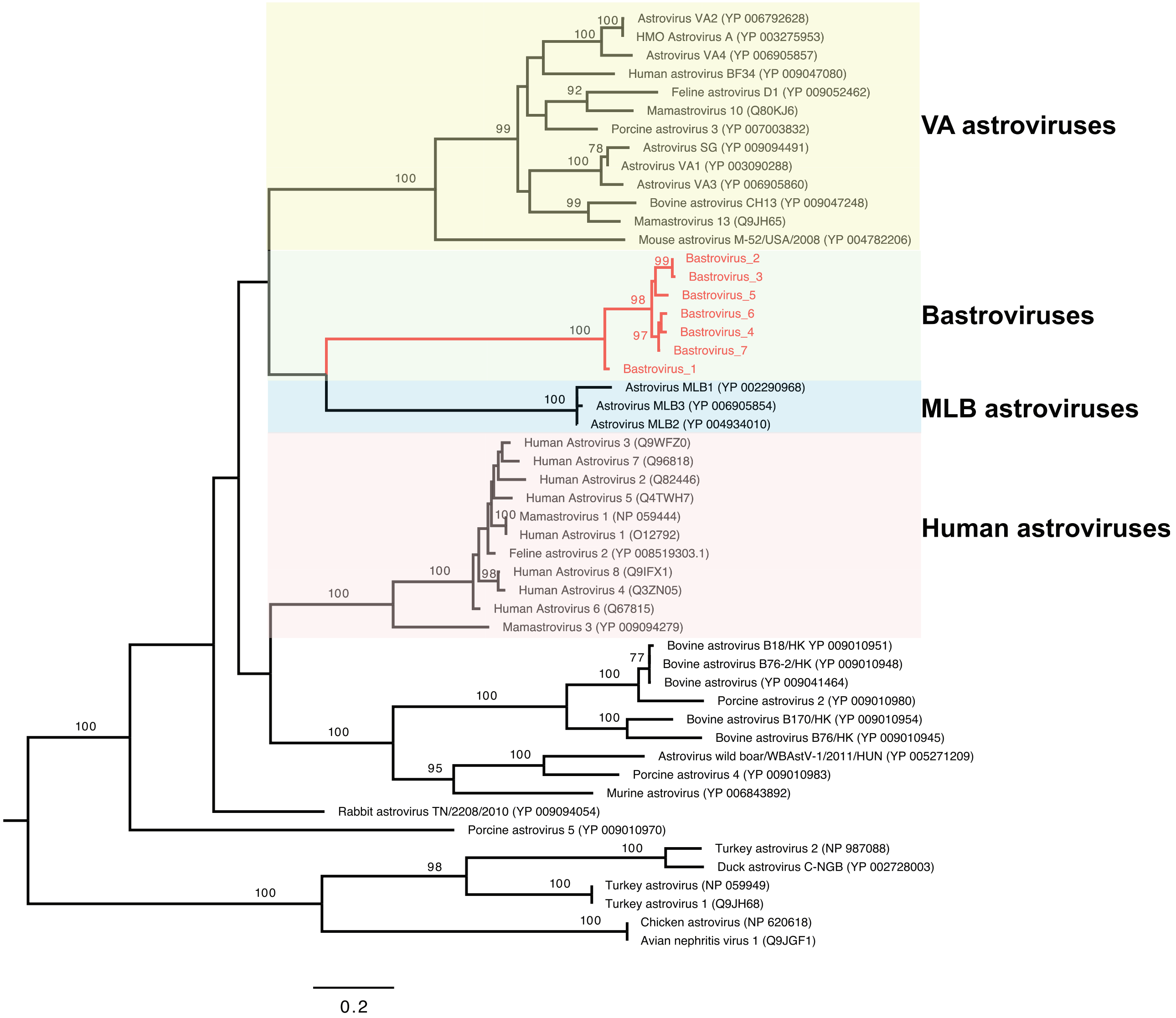
100
Astrovirus VA2 (YP 006792628)
100
HMO Astrovirus A (YP 003275953)
Astrovirus VA4 (YP 006905857)
Human astrovirus BF34 (YP 009047080)
92
Feline astrovirus D1 (YP 009052462)
Mamastrovirus 10 (Q80KJ6)
99
Porcine astrovirus 3 (YP 007003832)
78
Astrovirus SG (YP 009094491)
100
Astrovirus VA1 (YP 003090288)
100
Astrovirus VA3 (YP 006905860)
99
Bovine astrovirus CH13 (YP 009047248)
Mamastrovirus 13 (Q9JH65)
Mouse astrovirus M-52/USA/2008 (YP 004782206)
VA astroviruses
99
Bastrovirus_2
Bastrovirus_3
98
Bastrovirus_5
Bastrovirus_6
100
97
Bastrovirus_4
Bastrovirus_7
Bastrovirus_1
Bastroviruses
Astrovirus MLB1 (YP 002290968)
100
Astrovirus MLB3 (YP 006905854)
Astrovirus MLB2 (YP 004934010)
MLB astroviruses
Human Astrovirus 3 (Q9WFZ0)
Human Astrovirus 7 (Q96818)
Human Astrovirus 2 (Q82446)
Human Astrovirus 5 (Q4TWH7)
100
Mamastrovirus 1 (NP 059444)
Human Astrovirus 1 (O12792)
Feline astrovirus 2 (YP 008519303.1)
Human astroviruses
100
98
Human Astrovirus 8 (Q9IFX1)
Human Astrovirus 4 (Q3ZN05)
100
Human Astrovirus 6 (Q67815)
Mamastrovirus 3 (YP 009094279)
Bovine astrovirus B18/HK YP 009010951)
77
Bovine astrovirus B76-2/HK (YP 009010948)
100
Bovine astrovirus (YP 009041464)
100
Porcine astrovirus 2 (YP 009010980)
100
Bovine astrovirus B170/HK (YP 009010954)
100
100
Bovine astrovirus B76/HK (YP 009010945)
100
Astrovirus wild boar/WBAstV-1/2011/HUN (YP 005271209)
95
Porcine astrovirus 4 (YP 009010983)
Murine astrovirus (YP 006843892)
Rabbit astrovirus TN/2208/2010 (YP 009094054)
Porcine astrovirus 5 (YP 009010970)
100
Turkey astrovirus 2 (NP 987088)
98
Duck astrovirus C-NGB (YP 002728003)
100
100
Turkey astrovirus (NP 059949)
Turkey astrovirus 1 (Q9JH68)
100
Chicken astrovirus (NP 620618)
Avian nephritis virus 1 (Q9JGF1)
0.2

Figure 2. Phylogenetic tree of the amino acids sequences encoding the putative ORF2 capsid protein of bastrovirus and members of the *Astroviridae* family. Maximum likelihood tree was constructed using the LG+G amino acid model, as the best-fit model determined by MEGA (Tamura et al. 2013), with 1,000 bootstrap replications. Tree was mid-point rooted for clarity, and only bootstrap values >75% are shown. The scale bar indicates the number of substitutions per site.

a prolonged period of time towards the nucleotide composition and codon usage characteristic for the currently circulating species.

Seven complete coding sequences of bastroviruses were obtained showing 67–93 per cent nucleotide identity, corresponding to 63–98 per cent amino acid identity to each other. This diversity across a small cohort suggests that if bastrovirus does infect humans, the virus is not recently introduced and has most likely circulated in humans for some time. Alternatively the observed diversity could have been generated in another host or hosts and was introduced to humans as a food contaminant or zoonosis from pets, livestock, or wild animals.

A bastrovirus-specific PCR revealed that the virus is present in thirty-five of the 400 (8.7%) stool samples examined. Bastrovirus was found in thirty-two of the 200 faecal samples from 1984 to 1985 and in three of the 200 faecal samples from 1994 to 1995, indicating sustained presence of bastrovirus. The lower prevalence of bastrovirus in the 1994–5 cohort could be reduced prevalence or due to reduced sensitivity of the primers due to virus evolution in the later cohort.

All bastroviruses detected share a similar genome organization and contain the same predicted conserved domains in the non-structural ORF1 (viral methyltransferase, viral helicase, and RdRp). The domains share identity with domains from members of the *Hepeviridae*. However, hepatitis E virus encodes a putative Y-domain and a putative papain-like cysteine protease domain that were not found in bastroviruses. These additional domains are either not necessary for bastrovirus replication or are replaced by alternate proteins. The structural protein of bastrovirus shares the highest sequence identity with members of the *Astroviridae*.

In the two cohorts studied here, the presence of bastrovirus was not correlated with diarrhoea, hepatitis, or any of the monitored symptom or disease. Future screening for the presence of bastrovirus in unexplained cases of diarrhoea or hepatitis are needed to determine the pathogenicity of this novel agent. The

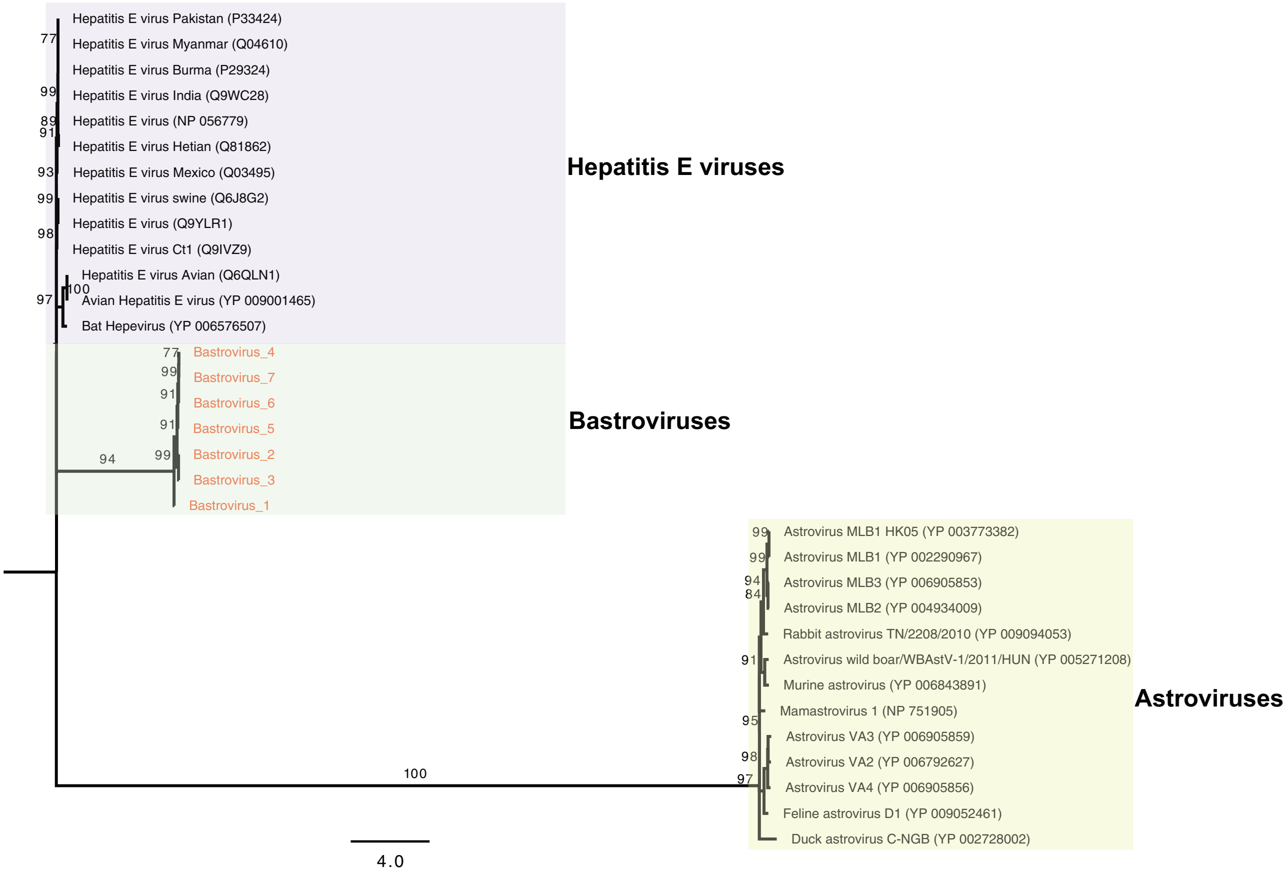
Hepatitis E virus Pakistan (P33424)
77
Hepatitis E virus Myanmar (Q04610)
Hepatitis E virus Burma (P29324)
99
Hepatitis E virus India (Q9WC28)
89
91
Hepatitis E virus (NP 056779)
Hepatitis E virus Hetian (Q81862)
93
Hepatitis E virus Mexico (Q03495)
99
Hepatitis E virus swine (Q6J8G2)
98
Hepatitis E virus (Q9YLR1)
Hepatitis E virus Ct1 (Q9IVZ9)
Hepatitis E virus Avian (Q6QLN1)
100
97
Avian Hepatitis E virus (YP 009001465)
Bat Hepevirus (YP 006576507)
Hepatitis E viruses
77
Bastrovirus_4
99
Bastrovirus_7
91
Bastrovirus_6
91
Bastrovirus_5
94
99
Bastrovirus_2
Bastrovirus_3
Bastrovirus_1
Bastroviruses
99
Astrovirus MLB1 HK05 (YP 003773382)
99
Astrovirus MLB1 (YP 002290967)
94
84
Astrovirus MLB3 (YP 006905853)
Astrovirus MLB2 (YP 004934009)
Rabbit astrovirus TN/2208/2010 (YP 009094053)
91
Astrovirus wild boar/WBAstV-1/2011/HUN (YP 005271208)
Murine astrovirus (YP 006843891)
95
Mamastrovirus 1 (NP 751905)
Astrovirus VA3 (YP 006905859)
98
Astrovirus VA2 (YP 006792627)
97
Astrovirus VA4 (YP 006905856)
Feline astrovirus D1 (YP 009052461)
Duck astrovirus C-NGB (YP 002728002)
Astroviruses
100
4.0

Figure 3. Phylogenetic tree of the amino acids sequences encoding the ORF1 RDRP domain of bastroviruses, astroviruses, and hepatitis E viruses. Maximum-likelihood tree was constructed employing the LG+G+I amino acid model as a best-fit model determined by MEGA (Tamura et al. 2013). Tree was mid-point rooted for clarity, and only bootstrap values >75% are shown. The scale bar indicates the number of substitutions per site.

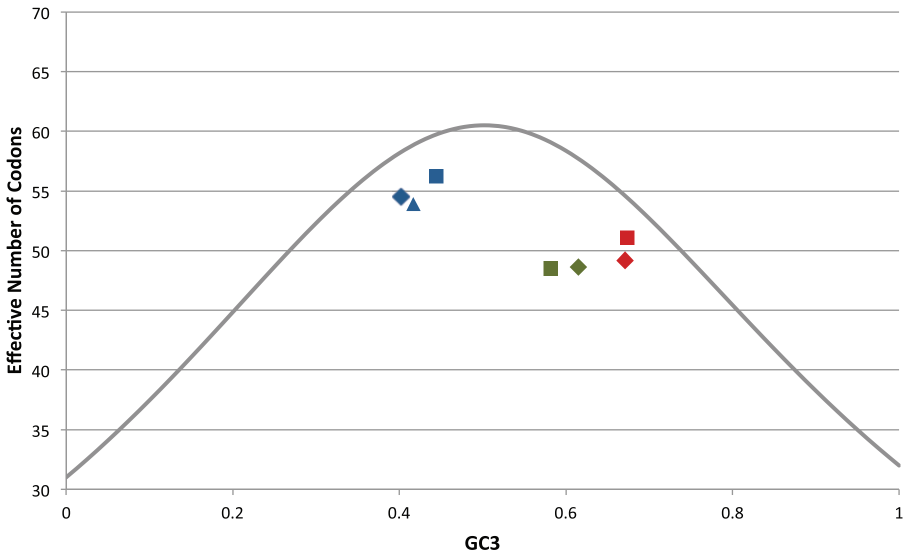

Figure 4. Nc plot of bastrovirus, hepatitis E virus (NC_001434.1), and astrovirus (NC_001943.1). The proportion of G+C at the third codon position (GC3, X-axis) is plotted versus the ENC values (Y-axis). The grey bell-shaped curve indicates the ENC values at corresponding values of GC3. Bastroviral, astroviral, and hepatitis E viral values are in red, blue, and green, respectively. The values for ORFs 1 and 2 are in diamond and square format, respectively. The blue triangle indicates the ORF1b of astrovirus.

multiple bastrovirus full genomes reported here will provide useful references for future virus discovery efforts.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

Acknowledgements

The Amsterdam Cohort Studies on HIV infection and AIDS, a collaboration between the Amsterdam Health Service, the Academic Medical Center of the University of Amsterdam, Sanquin Blood Supply Foundation, Medical Center Jan van Goyen, and the HIV Focus Center of the DC-Clinics are part of the Netherlands HIV Monitoring Foundation and supported by the Center for Infectious Disease Control of the Netherlands National Institute for Public Health and the Environment. This study was supported by the European Community's Seventh Framework Programme (FP7/2007–2013) under the project EMPERIE, EC grant agreement number 223498. We thank SURFsara (www.surfsara.nl) for the support in using the HPC cloud.

References

Altschul, S. F. et al. (1990) 'Basic Local Alignment Search Tool', *Journal of Molecular Biology*, 215: 403–10.

Balayan, M. S. et al. (1983) 'Evidence for a Virus in Non-A, Non-B Hepatitis Transmitted Via the Fecal-Oral Route', *Intervirology*, 20: 23–31.

Bankevich, A.et al. (2012) 'SPAdes: A New Genome Assembly Algorithm and its Applications to Single-Cell Sequencing', *Journal of Computational Biology*, 19: 455–77.

Bennetzen, J. L., and Hall, B. D. (1987) 'Codon Selection in Yeast', *Journal of Biological Chemistry*, 257: 3026–31.

Benson, D. A. et al. (2010) 'GenBank', *Nucleic Acids Research*, 38(Database issue): D46–51.

Boom, R. et al. (1990) 'Rapid and Simple Method for Purification of Nucleic Acids', *Journal of Clinical Microbiology*, 28: 495–503.

Cao, D., and Meng, X. J. (2012) 'Molecular Biology and Replication of Hepatitis E Virus', *Emerging Microbes and Infections*, 1: e17.

Chandra, V. et al. (2011) 'The Hepatitis E Virus ORF3 Protein Regulates the Expression of Liver-Specific Genes by Modulating Localization of Hepatocyte Nuclear Factor 4', *PLoS One*, 6: e22412.

Cotten, M. et al. (2014) 'Full Genome Virus Detection in Fecal Samples Using Sensitive Nucleic Acid Preparation, Deep Sequencing, and a Novel Iterative Sequence Classification Algorithm', *PLoS One*, 9: e93269.

de Vries, M. et al. (2011) 'A Sensitive Assay for Virus Discovery in Respiratory Clinical Samples', *PLoS One*, 6: e16118.

de Wolf, F. et al. (1988) 'Numbers of CD4+ Cells and the Levels of Core Antigens of and Antibodies to the Human Immunodeficiency Virus as Predictors of AIDS among Seropositive Homosexual Men', *Journal of Infectious Diseases*, 158: 615–22.

Dryden, K. A. et al. (2012) 'Immature and Mature Human Astrovirus: Structure, Conformational Changes, and Similarities to Hepatitis E Virus', *Journal of Molecular Biology*, 422: 650–8.

Endoh, D. et al. (2005) 'Species-Independent Detection of RNA Virus by Representational Difference Analysis Using Non-Ribosomal Hexanucleotides for Reverse Transcription', *Nucleic Acids Research*, 33: e65.

Finkbeiner, S. R., Kirkwood, C. D., and Wang, D. (2008) 'Complete Genome Sequence of a Highly Divergent Astrovirus Isolated from a Child with Acute Diarrhea', *Virology Journal*, 5: 117.

——— et al. (2009a) 'Human Stool Contains a Previously Unrecognized Diversity of Novel Astroviruses', *Virology Journal*, 6: 161.

——— et al. (2009b) 'Identification of a Novel Astrovirus (Astrovirus Va1) Associated with an Outbreak of Acute Gastroenteritis', *Journal of Virology*, 83: 10836–9.

Finn, R. D. et al. (2014) 'Pfam: The Protein Families Database', *Nucleic Acids Research*, 42(Database issue): 222–30.

Fu, Y. et al. (2009) 'Complete Sequence of a Duck Astrovirus Associated with Fatal Hepatitis in Ducklings', *Journal of General Virology*, 90(Pt 5): 1104–8.

Hall, T. (1999) 'BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/NT', *Nucleic Acids Symposium Series*,: 95–8.

Improve_assembly (2012) [http://search.cpan.org/∼ajpage/Bio_AssemblyImprovement-1.140300/bin/improve_assembly] accessed 16 Nov 2015.

Jiang, B. et al. (1993) 'RNA Sequence of Astrovirus: Distinctive Genomic Organization and a Putative Retrovirus-Like Ribosomal Frameshifting Signal That Directs the Viral Replicase Synthesis', *Proceedings of the National Academy of Sciences of the United States of America*, 90: 10539–43.

Jiang, H. et al. (2013) 'Comparison of Novel MLB-Clade, VA-Clade and Classic Human Astroviruses Highlights Constrained Evolution of the Classic Human Astrovirus Nonstructural Genes', *Virology*, 436: 8–14.

Kamar, N. et al. (2014) 'Hepatitis E Virus Infection', *Clinical Microbiology Reviews*, 27: 116–38.

Kapoor, A. et al. (2009) 'Multiple Novel Astrovirus Species in Human Stool', *Journal of General Virology*, 90(Pt 12): 2965–72.

Kearse, M. et al. (2012) 'Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data', *Bioinformatics*, 28: 1647–9.

Kjeldsberg, E. (1994) 'Serotyping of Human Astrovirus Strains by Immunogold Staining Electron Microscopy', *Journal of Virological Methods*, 50: 137–44.

Kurtz, J. B. et al. (1979) 'Astrovirus Infection in Volunteers', *Journal of Medical Virology*, 3: 221–30.

Lole, K. S. et al. (1999) 'Full-Length Human Immunodeficiency Virus Type 1 Genomes from Subtype C-Infected Seroconverters in India, with Evidence of Intersubtype Recombination', *Journal of Virology*, 73: 152–60.

Madeley, C. R., and Cosgrove, B. P. (1975) 'Letter: 28 nm Particles in Faeces in Infantile Gastroenteritis', *Lancet*, 2: 451–2.

Marchler-Bauer, A. et al. (2015) 'CDD: NCBI's Conserved Domain Database', *Nucleic Acids Research*, 43(Database issue): 222–6.

Mendez, E., and Arias, C. F. (2007) 'Astroviruses', in Howley, P. M., Knipe, D. M, Griffin, D. E., Lamb, R. A., Martin, M. A., Roizman, B. and Straus, S. E. (eds) *Fields Virology, Volume I* edn, pp. 981–1000. Philadelphia: Lippincott Williams & Wilkins.

Meng, X. J. (2013) 'Zoonotic and Foodborne Transmission of Hepatitis E Virus', *Seminars in Liver Disease*, 33: 41–9.

Meyer, C. T. et al. (2015) 'Prevalence of Classic, MLB-Clade and VA-Clade Astroviruses in Kenya and the Gambia', *Virology Journal*, 12: 78.

Moser, L. A., and Schultz-Cherry, S. (2005) 'Pathogenesis of Astrovirus Infection', *Viral Immunology*, 18: 4–10.

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.

Nguyen, T. A. et al. (2008) 'Identification of Human Astrovirus Infections among Children with Acute Gastroenteritis in the Southern Part of Vietnam During 2005-2006', *Journal of Medical Virology*, 80: 298–305.

Oude Munnink, B. B. et al. (2014) 'Unexplained Diarrhoea in HIV-1 Infected Individuals', *BMC Infectious Diseases*, 14: 22.

Papadopoulos, J. S., and Agarwala, R. (2007) 'COBALT: Constraint-Based Alignment Tool for Multiple Protein Sequences', *Bioinformatics*, 23: 1073–9.

Pringle, C. R. (1998) 'Virus Taxonomy—San Diego 1998', *Archieves of Virology*, 143: 1449–59.

Sharp, P. M., and Li, W. H. (1987) 'The Codon Adaptation Index—A Measure of Directional Synonymous Codon Usage Bias, and its Potential Applications', *Nucleic Acids Research*, 15: 1281–95.

Smuts, H. et al. (2014) 'Novel Hybrid Parvovirus-Like Virus, NIH-CQV/PHV, Contaminants in Silica Column-Based Nucleic Acid Extraction Kits', *Journal of Virology*, 88: 1398.

Soares, C. C. et al. (2008) 'Astrovirus Detection in Sporadic Cases of Diarrhea among Hospitalized and Non-Hospitalized Children in Rio De Janeiro, Brazil, from 1998 to 2004', *Journal of Medical Virology*, 80: 113–7.

Sullivan, K. M., Dean, A., and Soe, M. M. (2009) 'OpenEpi: A Web-Based Epidemiologic and Statistical Calculator for Public Health', *Public Health Report*, 124: 471–4.

Tam, A. W. et al. (1991) 'Hepatitis E Virus (HEV): Molecular Cloning and Sequencing of the Full-Length Viral Genome', *Virology*, 185: 120–31.

Tamura, K. et al. (2013) 'MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0', *Molecular Biology and Evolution*, 30: 2725–9.

Walter, J. E. et al. (2001) 'Molecular Characterization of a Novel Recombinant Strain of Human Astrovirus Associated with Gastroenteritis in Children', *Archives of Virology*, 146: 2357–67.

Wang, Q. H. et al. (2001) 'Genetic Analysis of the Capsid Region of Astroviruses', *Journal of Medical Virology*, 64: 245–55.

Watson, S. J. et al. (2013) 'Viral Population Analysis and Minority-Variant Detection Using Short Read Next-Generation Sequencing', *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, 368: 20120205.

Wright, F. (1990) 'The 'Effective Number of Codons' Used in a Gene', *Gene*, 87: 23–9.

Xu, B. et al. (2013) 'Hybrid DNA Virus in Chinese Patients with Seronegative Hepatitis Discovered by Deep Sequencing', *Proceedings of the National Academy of Sciences of the United States of America*, 110: 10264–9.

Yao, B. et al. (2012) 'SVMTriP: A Method to Predict Antigenic Epitopes Using Support Vector Machine to Integrate Tri-Peptide Similarity and Propensity', *PLoS One*, 7: e45152.