January 2009

Identification and Characterization of Novel Astroviruses

Stacy Finkbeiner
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Department of Molecular Microbiology

Dissertation Examination Committee:
Chair: David Wang, PhD
Michael Diamond, MD, PhD
Henry Huang, PhD
Gregory Storch, MD
Phillip Tarr, MD
Herbert 'Skip' Virgin IV, MD, PhD

IDENTIFICATION AND CHARACTERIZATION OF NOVEL ASTROVIRUSES

by

Stacy Renee Finkbeiner

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2009

Saint Louis, Missouri

ABSTRACT OF THE DISSERTATION

Identification and characterization of novel astroviruses

by

Stacy Renee Finkbeiner

Approximately 1.8 million children die from diarrhea annually, and millions more suffer multiple episodes of nonfatal diarrhea. Despite the availability of improved molecular diagnostics to detect the known viral agents, the etiology of a large proportion of diarrheal cases is unknown. In fact, it is estimated that no causative agent can be identified in up to 40% of sporadic cases or in gastroenteritis outbreaks. Detection of novel or unexpected viruses is the first step in identifying agents that could potentially close the diagnostic gap and pave the way for the development of more comprehensive preventative measures and better treatments.

This dissertation encompasses the first application of cutting edge mass sequencing approaches to the analysis of viruses present in fecal specimens from patients with diarrhea. Known enteric viruses as well as multiple sequences (with only limited sequence similarity to viruses in GenBank) from putatively novel viruses were detected in pediatric sporadic diarrhea specimens. One virus, Astrovirus MLB1 (AstV-MLB1), was fully sequenced and determined to be a highly divergent, novel astrovirus

based on phylogenetic analysis. AstV-MLB1 was further detected by RT-PCR in 4/254 fecal specimens collected at the St. Louis Children's hospital in 2008, indicating that AstV-MLB1 is currently circulating in North America.

A second highly divergent, novel astrovirus, Astrovirus VA1 (AstV-VA1), was identified in two specimens from a gastroenteritis outbreak at a child care center. Mass sequencing yielded nearly the entire genome of AstV-VA1 which appears to be most closely related to astroviruses found in mink and sheep. One additional sample also tested positive for AstV-VA1 by RT-PCR, resulting in detection of the virus in 3/5 specimens collected from the outbreak. This presents the possibility that further investigations might reveal that AstV-VA1 is a causative agent of gastroenteritis outbreaks.

The identification of two novel astroviruses in fecal specimens from children with diarrhea suggests that astroviruses may cause a larger fraction of diarrhea cases than previously recognized. Furthermore, the identification and characterization of novel astroviruses MLB1 and VA1 lays the foundation for future investigations into their potential roles as etiologic agents of diarrhea.

Acknowledgements

I would like to thank all of our collaborators and my thesis committee for their valuable contributions that resulted in me A.) actually having a project to work on and B.) having a project that turned into a really great story. They are: Carl D. Kirkwood, Phil Tarr, Skip Virgin, Greg Storch, Henry Huang, Mike Diamond, Jan Vinje, Theo Sloots, and Gagandeep Kang.

I would also like to thank members of the lab and the Pathogen Discovery Facility. They have been very helpful throughout the years and have made the lab environment a great place to work each and every day. They are: Adam Allred, Lori Holtz, Jade Le, Anne Gaynor, Guang Wu, Nang Nguyen, Tuya Wulan, Guoyan Zhao, Lindsay Droit, Collin Todd, Kathie Mihindukulasuriya.

I would like to acknowledge the valuable support I received from the Food Safety Research and Response Network – USDA, the ASV travel grant, the Berg-Morse Research Fellowship, and the Schlesinger Travel Grants.

Dave Wang deserves immense gratitude. Not only has he provided me with excellent research training, but he has also taught me how to be a good scientist in ways that extend beyond the bench. He has spared no effort in providing me with exceptional opportunities to help me develop as a young scientist and to allow me to network with people from all around the world. Furthermore, I think he demonstrates a quality of character and an understanding of social dynamics and human nature that make him a down-to-earth, very approachable person as well as a mentor. Dave has set the bar very high making it hard to imagine how any other research experience will ever possibly compare to the time I have spent in his lab over the last 5 years.

Lastly, I need to of course thank my parents, Randy & Susan Finkbeiner, because without their relentless love and support I never would have made it this far in my education. I am very lucky to have such wonderful parents that as an adult I can now consider my friends and greatest advocates.

Table of Contents

List of tables and figures

Chapter 1: Introduction

Portions of this chapter will be published in "New Virus Discovery in the 21st Century." *Molecular Microbiology: Diagnostic Principles and Practice. ASM Press*

Background on detection of novel viruses

The detection of novel viruses has traditionally relied heavily on the ability to culture viruses in a laboratory setting. This has limited our ability to detect a large number of viruses due to the fact that many viruses do not grow well, or at all, in cell culture. Of the ones that do grow in cell culture, it is often a tedious process to determine a permissive cell line(s) in combination with the optimal culturing conditions that allow a virus to be successfully propagated. Furthermore, the ability to culture a virus does not inherently lead to easy identification or characterization of the virus. In the past, the identification and characterization of unknown viruses grown in culture was based primarily on visual inspection of viral particle size and morphology using electron microscopy (EM) or serological cross reactivity with antibodies to known viruses. The ability to determine the viral family to which a given virus belongs to based on EM observations is severely hampered by the fact that many viral families exhibit similar morphologies, making it difficult to easily distinguish between them. Serological assays also have their limitations in that the unknown virus must cross react with a known virus for which there are serological reagents available. It also requires having some *a priori* knowledge about the possible identity of the

A. 20th Century Molecular Virus Discovery Methods

Method	Year	1st Discovery
cDNA Library Immunoscreening	1989	Hepatitis C
Degenerate PCR	1993	Sin nombre Hantavirus
Subtractive Hyridization (RDA)	1994	Kaposi's Sarcoma Associated Herpes

B. 21st Century Molecular Virus Discovery Methods

Method	Year	1st Discovery
DNase-SISPA	2001	Bovine Parvovirus
Pan-viral Microarray	2003	SARS
VIDISCA	2004	Coronavirus NL63
Mass Sequencing	2005	Human Bocavirus

Figure 1.1: Methods for virus discovery in the past 20 years

unknown virus, since screening has to be done individually for each viral family.

Since the advent of the polymerase chain reaction (PCR) in the 1980s, PCR has become a common method for detecting viral genomes. Once the genome sequence for a given virus is available, it is a

straightforward endeavor to design primers that detect the target virus. Application of PCR to highly conserved regions of a given virus family or taxa has enabled the identification of many novel viruses related to known viral families, much as PCR targeting 16S ribosomal genes in bacteria has resulted in the identification of novel bacteria. While the cases of viruses discovered in this fashion are too numerous to exhaustively list, some notable examples include sin nombre virus (Figure 1.1A) (1) and human coronavirus HKU1 (2). Various computational strategies for design of consensus PCR primers for novel virus identification have been described (3,4). However, one fundamental limitation of PCR assays, which is shared with serological approaches, is that the breadth of these assays is limited to interrogating a single taxa, typically a given viral family. Thus, one must have some idea of which viral family or families one wants to examine, which can be a daunting task given the abundance of possibilities.

As a consequence of these limitations, other methods not subject to this form of bias, such as representational difference analysis (RDA) (5), have been applied to virus discovery. This method is a form of subtractive hybridization involving the hybridization of nucleic acids between putatively infected samples and carefully chosen controls (uninfected samples). Unique sequences present only in the infected specimens are preferentially amplified during subsequent PCR cycles leading to

dramatic enrichment for the unique sequences. The most notable application of RDA to date resulted in the identification of human herpesvirus-8 in Kaposi's Sarcoma lesions (Figure 1.1A) (6). Following this landmark event in the use of molecular methods for virus discovery, many other efforts to use RDA to identify novel viruses ensued, but generally with only limited success. In one recent successful case, RDA was used to identify viral sequences from the novel murine norovirus 1 following serial passage of a putative infectious agent in mice lacking innate immunity, which effectively served to amplify the virus to much higher titers in vivo (7). A major technical limitation is that it is difficult to obtain a well matched negative control for the specimen of interest, and frequently the amplified differences between the samples reflect differences in host gene expression rather than foreign microbial sequences.

Another landmark discovery, that of hepatitis C, relied upon a unique strategy that combined both molecular cloning techniques with serological methods (8). In cDNA library immunoscreening (Figure 1.1A), nucleic acids were extracted from putatively infected specimens and used to generate bacterial cDNA clones. Serum from a patient suffering from non-A, non-B hepatitis was used to screen the cDNA clones to identify polypeptides that were bound by serum antibodies. One reactive clone was identified that contained an insert that was not present in control samples. Furthermore, the sequence of this clone did not hybridize

to host sequences, suggesting that the clone represented sequence unique to the infected sample. Sequencing of the clone yielded a new member of the Flavivirus family, hepatitis C virus. While this technique identified the etiologic agent of non-A, non-B hepatitis, its application to other diseases has been limited, and no other significant viruses have been identified to date using this approach.

The 1st decade of the 21st century: the Dawn of Sequence based Discovery Methods

Sequence-independent PCR strategies and new technologies such as microarrays and mass sequencing, while not without their own limitations, circumvent many of the limitations associated with traditional virus discovery methods and consequently have revolutionized the process of virus discovery. These approaches, which can be applied either alone or in concert with traditional methods such as tissue culture, have enabled unbiased and massively parallel analysis of the nucleic acid composition of a given specimen. It must be emphasized however, that these approaches complement and do not replace traditional virologic methods. For example, recent discoveries of viruses, such as Chapare hemorrhagic fever virus (9) and Melaka virus, a zoonotic bat virus associated with respiratory disease (10), continue to demonstrate the effectiveness of culture and serological methods. However, it is clear that

there are viruses that do not grow in culture or that are so highly divergent that they are not amenable to traditional methods but can be unveiled by these newer molecular methods. In this review, we will focus on those viruses where molecular methods have played a critical role in their discovery.

Sequence-independent PCR amplification strategies for virus detection

PCR screening for novel viruses using primers designed to conserved features of viral families offers a powerful approach to virus identification, provided there is adequate rationale to select a given candidate viral family to test. In the absence of a logical set of candidates to screen, it is typically not feasible to target every viral family individually, which would be prohibitively laborious, time-intensive, and reagent intensive. The development of methods in which nucleic acids could be amplified in the absence of primers designed to a specific target sequence has played a critical role in the latest methods for virus discovery. These sequence independent methods are capable of increasing the quantity of an unknown agent to detectable levels. Various post amplification detection strategies have been devised to 'visualize' the products of interest that include gel electrophoresis based differential display approaches (e.g. VIDISCA), hybridization to pan viral microarrays, and high throughput mass sequencing.

Identification of human metapneumovirus. The discovery of a novel human pneumovirus in 2001, called human metapneumovirus (hMPV), relied upon a sequence independent PCR approach in conjunction with viral culture to identify a previously unknown virus. (11). In this study, 28 serologically related, but otherwise unidentifiable viruses were isolated in cell culture from patients with respiratory tract infections. One of the virus isolates grown in cell culture was amplified using a technique called arbitrarily primed PCR (12,13). In this technique, primers such as the commonly used Universal M13, T7, and T3 primers are arbitrarily selected for use in amplification. The amplification scheme involves low stringency annealing in the early cycles so that mismatches will be accommodated during the annealing of the primers to many different sequences in the sample. In the later cycles, a higher stringency annealing temperature is used in which amplicons that can be efficiently amplified will predominate and create a fingerprint of the sample. The fingerprint, consisting of various sized amplicons, can then be compared to control samples in order to identify the unique PCR products to then be subsequently sequenced. In the case of hMPV, 10/20 PCR products that were sequenced had amino acid similarity to avian pneumovirus (APV). Upon further sequencing, it was determined that this novel virus encoded proteins that were 52-87% similar to avian pneumovirus at the amino acid

level. Preliminary animal studies showed that hMPV was able to replicate and cause mild upper respiratory symptoms in monkeys, but did not cause any symptoms in birds nor showed any signs of replication. Finally, the seroprevalence of hMPV was shown to be 100% by age 5. As a result of numerous follow up studies (14), it is now believed that hMPV is a significant cause of respiratory tract infections. It is interesting to note that the initial report of hMPV also noted that hMPV has been circulating in the human population for at least 50 years, but that it eluded detection because it did not grow well under typical cell culturing conditions, it had low nucleotide similarity to known viruses, and lacked apparent serological cross reactivity with antibodies generated against other viruses.

Viruses identified using DNase-SISPA. Another approach known as sequence-independent single primer amplification (SISPA) (15) involves the ligation of a common primer binding sequence to both ends of cDNA molecules, which can then be amplified with a cognate primer. A refinement of this approach called DNase-SISPA involves the treatment of samples with DNase prior to nucleic acid extraction (16). The theory behind DNase-SISPA is that the DNase should degrade host DNA, but viral nucleic acid should be protected by the protein capsid and for some viruses also by the lipid viral envelope. Therefore, nucleic acids that survive the DNase treatment step should be enriched for sequences of viral origin.

In the initial description of the method, the extracted nucleic acid was digested using restriction enzymes and the resulting restriction fragments were then subjected to SISPA. The SISPA products were visualized by gel electrophoresis and prominent bands were sequenced. Amplicons derived from novel bovine parvoviruses were identified in this initial proof of concept (Figure 1.1B). Subsequently, DNase-SISPA has been used to identify novel viruses from human plasma (parvovirus 4 (PARV4) and two TTV-like anelloviruses (17)). PARV4 has since been detected in human serum in multiple studies, but its role in human disease is not currently known. In addition, application of DNase-SISPA to tissue cultures inoculated with human stools has led to identification of the novel human adenovirus 52 (HAdV-52) (18) and a novel cardiovirus, Saffold virus (SAF-V) (19). Multiple other studies have since demonstrated that viruses related to Saffold are circulating in the human population (20-22). The fact that novel viruses from multiple different viral families have been identified with this approach highlights its utility as an unbiased means for viral discovery.

Viruses identified using VIDISCA. Another approach called VIDISCA, which stands for Virus-Discovery-cDNA-amplified restriction fragment-length polymorphism, was used in the discovery of a novel human coronavirus, NL63 (HCoV-NL63), in 2004 (Figure 1.1B) (23). Like DNase-SISPA, the VIDISCA method begins with a viral enrichment step that entails

centrifugation to remove residual cells and mitochondria and then DNase treatment to degrade non-enveloped or non-encapsidated DNA. Upon RNA extraction, double stranded cDNA is synthesized and then the cDNA is digested with restriction enzymes that are known to cut frequently occurring DNA sequences so as to generate smaller amplifiable sequences. Two different adapters are ligated onto the cDNA fragments so that only sequences which have a different adapter on each end can be subsequently amplified. To clearly visualize the products following gel electrophoresis and compare the products to those obtained from a control sample, another round of amplification reactions is carried out in which the adapter primers are modified to include one extra random nucleotide. Since one random nucleotide is added to each primer, there are a total of 4 variations of each of the two different adapter primers that can be used in 16 primer pair combinations to amplify subsets of the original pool of amplicons. Bands present in the experimental sample lanes that are absent in the controls are excised from the gel and sequenced.

The first application of VIDISCA to virus discovery resulted in the identification of human coronavirus NL63 (Figure 1.1B). An unidentified virus was propagated in tertiary monkey kidney cells after inoculation of a nasopharyngeal aspirate from a patient with respiratory infection. VIDSCA was performed on the supernatant from the infected cells and a

cell culture supernatant from uninfected cells was used as a negative control. PCR bands that were specific to the virus infected sample and not present in the uninfected control were then sequenced. The initial fragments shared limited sequence identity to known coronaviruses. Subsequently, the full genome of a novel coronavirus was sequenced that shared on average 65% identity to its closest relative HCoV-229E. Furthermore, HCoV-NL63 contained unique genomic features, such as an insertion in the spike protein, which is involved in receptor binding and is thought to determine the tropism of the virus. In the original study, HCoV-NL63 was also detected in up to 7% of patients with respiratory disease. A follow-up study suggested that NL63 is associated with croup (24). Shortly following the identification of NL63, yet another coronavirus, HKU1, was discovered using degenerate PCR screening strategy (2). The identification of these two human viruses immediately following the SARS-coronavirus outbreak of 2003 has greatly increased awareness of the role of coronaviruses in human disease, and suggested that the importance of coronaviruses as etiologic agents of respiratory infections and their diversity has been vastly underestimated. VIDISCA has also been used to identify a new strain of human parechovirus type 1 (25).

These examples used sequence-independent PCR assays followed by selective sequencing of a subset of "unique" PCR products and have highlighted how powerful unbiased approaches are in the discovery of

novel viruses. In many ways, these techniques provide the foundation for the higher-throughput techniques that will be discussed in the rest of this chapter.

Sequence independent PCR followed by microarray hybridization for virus discovery

DNA microarrays first emerged in the mid 1990s as powerful tools to measure gene expression or genomic content changes in various organisms (26-28). While many flavors of microarrays exist, the commonality to all microarrays is that a high density of DNA probes, numbering from thousands to hundreds of thousands, is attached to a surface or surfaces enabling the analysis of complex mixtures of input nucleic acid (26-28) in a single assay. Capitalizing on the inherently massively parallel nature of microarrays, Wang et al. described in 2002 the first microarray designed to detect large numbers of viruses (29). This prototype DNA microarray (the "ViroChip") harbored 1,600 70mer oligonucleotides derived from 140 different virus species with an average of ~ 10 oligonucleotides per virus species. Conventionally, microarrays had been designed primarily with probes from a single organism of interest. The rationale for development of this microarray was twofold: (1) the microarray would enable simultaneous screening for hundreds to thousands of known viruses in a single assay, thus making it an ideal

diagnostic tool; (2) by careful selection of highly conserved sequences, it was anticipated that novel viruses related to known virus families could be detected.

Experimentally, viral nucleic acids extracted from clinical samples are subject to random PCR amplification, fluorescently labeled and hybridized to the microarray. The resulting hybridization patterns observed between the nucleic acids in the sample and the oligonucleotides on the array can be analyzed to identify whether a known or novel virus is present in the sample. The performance of the ViroChip was validated on both cultured viruses as well as on clinical samples. It was demonstrated that a wide variety of viruses could be detected by the array, including representatives of both RNA and DNA viruses (29). With the development of this technology, it quickly became apparent that objective computational approaches for analysis and interpretation of the raw microarray data were needed. In 2005, the first algorithm for objectively interpreting microarray hybridization data to infer the presence of microbial species, E-predict, was described (30). In one follow-up study, the ViroChip detected known viruses in 53/82 (65%) nasal lavage samples whereas only 14 (17%) of these samples yielded viruses by culture (31). In another follow-up study, the ViroChip demonstrated sensitivity in the range of 85-90% and specificity of ≥99%, as compared to virus specific PCR/RT-PCR reactions for respiratory syncytial virus, influenza and rhinovirus

(32). These results demonstrated the feasibility of using sequence independent PCR amplification of clinical specimens in combination with microarrays to detect a broad range of known viruses. Critically, in contrast to traditional screening strategies, *a priori* assumptions regarding the types of virus present do not need to be made.

DNA microarrays using long (70mer) oligonucleotide probes have also proven to be more robust to viral mutations than conventional PCR. For example, in one case, ViroChip analysis indicated that a patient specimen contained human metapneumovirus even though multiple assays using metapneumovirus PCR primers were negative. Sequencing of the virus present in the sample ultimately revealed that mutations present in the primer binding sites were most likely the cause of the false negative PCR result (33). This case illustrates two features of microarray based diagnostics: (1) depending on the length of the microarray probe, varying degrees of mutation can be tolerated, enabling detection of variant species; (2) the presence of multiple oligonucleotide probes for each target virus (i.e. redundancy) provides greater opportunities to detect variant viruses.

The field of diagnostic microarrays has grown tremendously in the past few years, with multiple broad-range diagnostic microarrays described (34-37). These have utilized a multitude of different microarray platforms, probe design strategies, and oligonucleotides of varying

lengths, demonstrating the robustness of the general methodology. In parallel, there has been more limited progress in the development of algorithms for interpretation of diagnostic microarrays (38,39). The published applications of these diagnostic microarrays to date have focused exclusively on detection of known viruses. By contrast, the second goal in development of the ViroChip was to facilitate the discovery of novel viruses. To date, this approach has been used to identify a number of novel viruses, from humans and animals.

Identification of SARS-CoV. The World Health Organization (WHO) issued a global alert in 2003 regarding an illness described as "severe acute respiratory syndrome" (SARS) that was emerging in Southeast Asia with significant mortality. A WHO coordinated collaboration was established between labs around the world to try and identify the causative agent of the highly contagious illness that had rapidly spread from Asia to other parts of the world. This was the first case in which a DNA microarray was employed to detect a novel virus (Figure 1.1B). The then unknown virus was cultured in Vero cells (40). Extraction of total nucleic acids from the culture followed by random amplification and hybridization to the ViroChip (41) yielded a significant signal intensity from just a few oligonucleotides derived from the viral families *Coronaviridae* and *Astroviridae*. Given the limited number of oligonucleotides that were

hybridized to the sample, it appeared that there may have been two novel viruses present in the sample. However, further analysis revealed that all of the oligonucleotides derived from the *Astroviridae* family came from a genetic element that is found in the 3' untranslated region of the genomes of both viral families (42). Therefore, based on the nature of the hybridization pattern, it appeared that a single virus, likely a coronavirus, was present in the sample. Since only a small subset of the highly conserved probes from the family *Coronaviridae* were bound by the virus, it was reasoned that the virus was likely to be highly divergent. In order to rapidly sequence parts of the novel virus genome, cDNA fragments derived from the unknown virus were physically recovered from the surface of the microarray, PCR amplified, cloned, and sequenced. The largest clone contained a ~1.1kb fragment that had 33% amino acid identity to a protein derived from murine hepatitis virus, a mouse coronavirus, thus confirming that the SARS-CoV was a novel virus. The coalition of scientists collaborating on the SARS effort was then able to show that SARS-CoV specific antibodies could be detected in convalescent-phase serum from SARS patients and that SARS-CoV RNA could be detected in respiratory specimens. This strongly suggested that SARS-CoV was in fact the etiological agent of the SARS outbreak (40,43,44).

Identification of XMRV. The ViroChip has also been applied in the search for novel human tumor viruses, which lead to the 2007 discovery of a novel human retrovirus in a subset of prostate cancer patients carrying mutations in the RNase L gene (45). Germline mutations in RNase L have been reported to confer an increased susceptibility to prostate cancer (29,46-48). It was also known that such mutations were also linked with hereditary prostate cancer, which is defined by having 3 or more affected family members and is often recognized by the early onset of the disease(49). However, there were conflicting reports based on case-controlled studies regarding the involvement of the RNase L mutations in the development of prostate cancer. This suggested to some that there were perhaps other environmental or population differences that affected the contribution of RNase L mutations to prostate cancer susceptibility. Given that RNase L has a well established antiviral function, Urisman, et al. hypothesized whether the observed differences in susceptibility could be explained by the fact that mutations in RNase L actually led to a greater susceptibility to a viral agent (45). RNA derived from biopsied prostate tumors with and with out mutations in RNase L was hybridized to the ViroChip. A distinct gammaretrovirus was identified that was detected in 8/20 (40%) tumors homozygous for the mutation in *RNASEL*, while retroviral sequences were only detected in 1/66 (1.5%) tumors bearing at least one wild-type allele. This virus appeared to be

most closely related to xenotropic murine leukemia virus lending itself to the name Xenotropic Murine-like Retrovirus (XMRV) and represented the first of its kind to be detected in humans. *In situ* hybridization of prostatic tumors with XMRV derived probes and immunohistochemistry probing for viral retroviral proteins showed the presence of the virus in the tumors suggesting there could be a link between XMRV infection and prostate cancer. Later work has shown that XMRV is integrated in some tumor cells and that replication of the virus is susceptible to RNase L (50). It is currently unclear whether XMRV plays a causal role in prostate cancer or whether its presence is a secondary event.

Identification of HTCV-UC1. In an effort to identify novel respiratory pathogens, respiratory secretions collected from patients with influenza-like symptoms were screened by conventional assays for respiratory syncytial virus, influenza virus, parainfluenza virus, adenovirus, and picornaviruses (rhinoviruses and enteroviruses). Samples that remained negative after all diagnostic testing were then analyzed with the ViroChip. One array had a hybridization pattern suggesting that a cardiovirus was present in the sample (20). Conventional diagnostic testing panels for respiratory viruses do not include any members of the Cardiovirus genus in the family *Picornaviridae*. Primers were designed from the highest intensity oligonucleotides and from alignments of conserved regions of

the known cardiovirus genomes in order to try and amplify portions of the cardiovirus present in the sample. The full sequence of this cardiovirus, HTCV-UC1, was then obtained. Testing of 428 respiratory specimens failed to identify additional cases. However, similar viruses were identified in 6 out of 767 stool samples tested (20). HTCV-UC1 appeared to be closely related to Saffold-like cardioviruses, which were initially discovered using the DNase-SISPA strategy (19,21,22). These studies define a new group of cardioviruses which seemed to be the first human cardioviruses and have tropism for both the respiratory and gastrointestinal tract. The role of these viruses in human disease remains to be determined.

Identification of animal viruses. Two final examples demonstrate the universality of pan-viral microarrays designed using all available viral sequences. By including sequences from all viruses without restrictions based on either viral taxonomy or putative host species, not only human samples but plant and animal specimens can be analyzed. In one instance, the ViroChip was used to analyze liver tissue from a beluga whale that had died mysteriously in an animal park. A strong hybridization signature from conserved coronavirus probes was observed on the ViroChip, which ultimately led to the identification and complete sequence of a highly divergent, novel coronavirus (51).

The ViroChip was also used in the discovery of an avian bornavirus found in samples derived from cases of a disease called proventricular dilation disease (PDD) (52). PDD is an inflammatory disease affecting over 50 species of parrots as well as other orders of birds. It had been long thought that there was an infectious etiology for this disease and a number of viral agents had been implicated over the years. For example, electron microscopy analysis led some to believe that a paramyxovirus was responsible, but this was later ruled out (52). Other reports came out reporting the culturing of an unidentified virus, but this virus was never identified (52). ViroChip analysis of PDD cases versus controls suggested that a novel bornavirus, referred to as avian bornavirus (ABV), was present in 62.5% of the PDD samples and none of the controls (0/8). Mass sequencing (described in the next section) was used to obtain the full bornavirus genome from one of the samples. Further PCR screening of additional PDD samples revealed that the avian bornavirus was detected in 71% (5/7) of PDD cases and none (0/14) of the controls. Given the high correlation of ABV with PDD samples and not controls, it is highly suggestive that this virus could in fact be the etiologic agent of PDD. This case perfectly highlights the limitations of conventional approaches since for years the etiology of this disease was unknown and now the ViroChip has identified what could potentially be the causative agent of PDD.

Pan-viral microarrays have proven to be successful tools for the detection of known viruses and the identification of multiple novel viruses. In conjunction with sequence independent PCR, microarrays offer a robust readout to define the nature of viruses present in a specimen. However, one key limitation of pan-viral microarrays is that highly divergent viruses that have little to no nucleotide similarity with known viruses will not be detected since the technology depends on cross-hybridization between the unknown virus and probes derived from known viral genomes.

DNA sequencing methods

Dideoxy sequencing or Sanger sequencing, first described in 1977 (53), has been the dominant DNA sequencing technology used during the last ~30 years. In the dideoxy sequencing process, primers designed to bind to a template of interest are added with dideoxynucleotides to a polymerase extension reaction. Incorporation of a dideoxynucleotide prevents further extension of the primer-driven DNA synthesis reaction. In the automated form of Sanger sequencing, each of the 4 dideoxynucleotides is labeled with a distinct fluorescent dye. The truncated synthesis products are then run through a capillary containing a denaturing polymer that provides single base pair resolution separation. As each product passes through the end of the capillary it is excited by

light and the resulting fluorescence is recorded as one of four wavelengths, each corresponding to a different base. In this way, the sequence of the template can be deduced by determining which dideoxynucleotide was incorporated at each position of the complementary sequence. The typical sequence read length can reach 750-800 bp. This technology has been the cornerstone of all significant sequencing projects up to 2005, including the sequencing of the human genome. Efficiencies gained in the monumental effort to sequence the human genome have enabled the production costs of Sanger sequencing in large genome sequencing centers to fall to as little as $0.35 per read (as of this writing).

Starting in 2005, multiple new sequencing technology platforms ("NextGen") emerged that have far surpassed conventional Sanger sequencing in terms of increased total sequence production capacity and decreased cost. While there are many NextGen platforms, in this review, we will limit our discussion to those that have been effectively utilized to date in the efforts to identify novel viruses. These include 454 pyrosequencing (54), commercially available in the form of the Roche FLX instrument and Solexa technology, which is marketed as the Illumina Genome Analyzer.

454 pyrosequencing technology differs from Sanger sequencing in that natural nucleotides are used in the sequencing reaction. Rather

than monitor the incorporation of nucleotides directly, successful incorporation of a nucleotide is inferred by measuring the quantity of inorganic phosphate (PPi) produced. PPi is stoichiometrically converted to ATP, which provides energy for luciferase. The quantity of light that is produced by luciferase as a given nucleotide is added to the reaction is indicative of the number of bases incorporated at that step. By repeatedly adding dATP, dCTP, dGTP and dTTP into the reaction in series, up to 250 bp of contiguous sequence can be determined. Another key feature of the 454 technology is that hundreds of thousands of these sequencing reactions are carried out simultaneously on independent beads. This is accomplished by generating 1:1 DNA template:bead complexes. First the target DNA to be sequenced is randomly sheared into fragments and then ligated to adapters. Single stranded template DNA is isolated, mixed with the beads and then subject to emulsion PCR to clonally amplify the template on each bead. Next, the beads are distributed into a fiber optically constructed plate (PicoTiterPlate) that contains millions of tiny wells. Within each well that receives an individual bead, an isolated environment is created for the sequencing of each template, resulting in parallel sequencing of hundreds of thousand of templates at once. Currently, one 454 sequencing run produces ~400,000 sequence reads of an average read length of ~250bp with imminent expansion of its capacity to ~1,000,000 reads of ~ 400 bp (55).

Solexa sequencing also uses a synthesis approach, but differs from 454 pyrosequencing in a number of ways. Unlike the bead format of the 454 platform, templates to be sequenced via Solexa sequencing are affixed to a solid surface. Adapter sequences are ligated onto the templates just as with 454, however, in Solexa sequencing these adapters not only serve as priming sites for PCR and sequencing, but they are also critical for the affixation of complementary templates to the solid surface through a mechanism called bridge amplification. The bridge amplification generates clusters of amplified template on the solid surface, where each cluster represents a different template. Solexa uses a unique chemistry that incorporates fluorescently labeled reversible terminator nucleotides in the sequencing reaction. The nucleotides are fluorescently labeled with different colors so that all four can be added to the reactions simultaneously. Due to the termination property of the nucleotides, only one individual nucleotide can be incorporated into each sequence during one sequencing cycle. The color of the fluorescent label incorporated into the sequences of each cluster is recorded and discerns which nucleotide was incorporated. Removal of the terminator group on the nucleotide just added enables incorporation of the next complementary nucleotides and the cycle is repeated. The length of the read is dependent on the efficiency of removing the chain terminator

group at each step. Currently, Solexa sequencing generates ~40-50 million sequences of a read length of ~35bp (55).

To date, all three of these technologies, Sanger sequencing, 454 sequencing and Solexa sequencing, have been used to characterize novel viruses. A schematic for generalized strategies of virus detection using mass sequencing are outlined in Figure 1.2.

Sequence-independent amplification and sequencing strategies for virus detection

Meyerson and colleagues conducted the first proof of principle experiment to demonstrate that extensive sequencing of nucleic acids in a specimen in conjunction with computational subtraction could be used to detect pathogens (56). They examined a tissue specimen from a case of post-transplant lymphoproliferative disorder (PTLD), knowing that most cases of PTLD are caused by Epstein Barr virus (EBV). A cDNA library with 27,840 sequences was generated and each individual sequence was compared to the human genome in order to identify those sequences of human origin. These sequences were then eliminated from further consideration in a process termed "transcript filtering". The remaining sequences that did not match human genomic sequences were postulated to be derived from microbial organisms present in the sample. In fact, from this library, 10 sequences that did not match any human

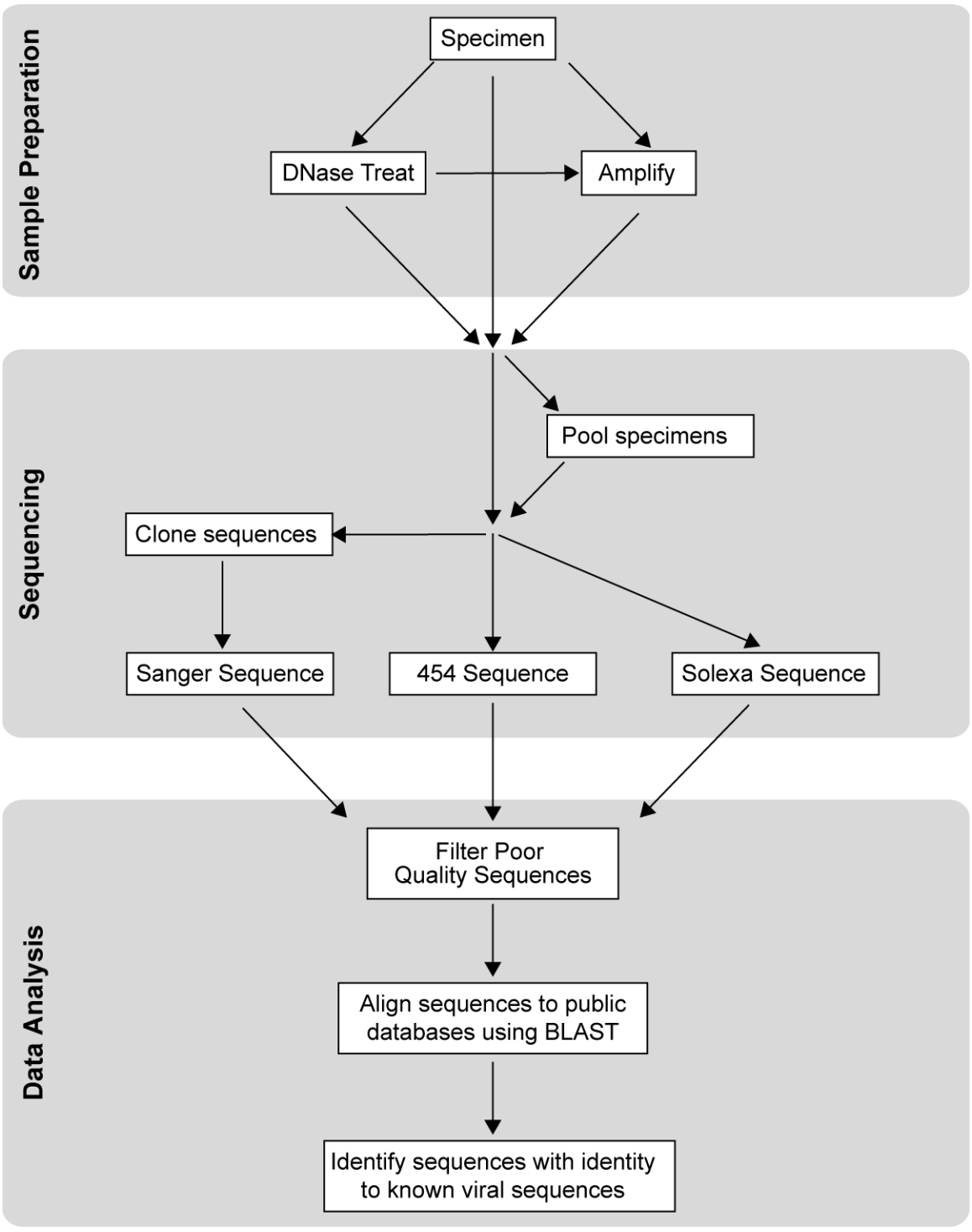


Figure 1.2: Schematic of mass sequencing strategies for detection of known and novel viruses

sequences but had similarity to EBV were identified. Furthermore, these sequences could only be amplified from EBV infected tissues and not from controls, thus showing the potential power of mass sequencing strategies.

Identification of human bocavirus. The discovery of human bocavirus (HBoV) marked the 1st example of discovery of a novel virus by direct high throughput sequencing of clinical specimens (Figure 1.1B) (57). In this study, pools of respiratory samples were ultracentrifuged, filtered, and DNase treated to produce samples for analysis that were highly enriched for intact viral particles (in which the genome is protected from DNase treatment) and to minimize the amount of other contaminating DNAs. In contrast to the discovery of human metapneumovirus, in which primers were arbitrarily chosen for amplification, a random PCR amplification strategy was used for amplification in the HBoV study. The amplified products were then cloned and ~300-500 clones were sequenced without any effort to identify specific amplicons. In this analysis, 20% of the sequences showed similarity to viral sequences, the majority of which were derived from known viruses. However, some of the sequences had only limited similarity to viruses in the *Parvovirdae* family, and more specifically the genus *Bocavirus*. The respiratory samples that were initially pooled together for sequencing analysis were independently screened by PCR for the presence of the novel bocavirus of which two

tested positive. Further PCR screening of other pediatric respiratory specimens revealed that HBoV could be detected in 3.1% of the samples. The potential role of human bocavirus in respiratory disease and gastroenteritis has been investigated in numerous subsequent studies. These epidemiologic and seroepidemiologic analyses have determined incidence rates of bocavirus between 2.7-19% (in respiratory secretions) and seroprevalence rates of 94.7-98.3% in healthy adults (14). As with human metapneumovirus, this high seroprevalence rate suggests that HBoV may be well established in humans and could be an important human pathogen.

Identification of novel polyomaviruses. For 36 years following the culture based discovery of JC virus (JCV) and BK virus (BKV) in 1971, it was assumed that these were the only two human polyomaviruses. That picture changed dramatically in a 12 month span in 2007-2008 as 3 novel human polyomaviruses were discovered, all by use of high throughput sequencing strategies. In 2007, sequencing of randomly amplified products generated from clinical respiratory samples resulted in the detection of 2 novel polyomaviruses, KI polyomavirus (KIV) and WU polyomavirus (WUV) (58,59). KIV was identified using the same experimental strategy that led to the discovery of human bocavirus (57). The discovery of WUV followed a parallel approach with the exception

that individual respiratory specimens rather than pooled specimens were analyzed. In both cases, sequences were identified with limited similarity (≤50%) at the amino acid level to the known primate polyomaviruses at the time.

The full genomes of KIV and WUV were sequenced and analyzed. Phylogenetic analysis demonstrated that KIV and WUV were actually most similar to each other (~65-70% amino acid identity) and appeared to represent a new subclass of polyomaviruses. Both of these viruses shared many similarities that distinguish them from JCV and BKV, including the lack of an Agno protein, the absence of a C-terminal extension domain of the Large T antigen, and altered origins of replication. In terms of tropism, neither WUV nor KIV have been detected to date in urine, in contrast to JCV and BKV which are both frequently detected in urine. Both WUV and KIV have been found to be prevalent in the respiratory tract, with WUV having a prevalence of up to 7% in respiratory infections with KIV's prevalence generally slightly lower (60-65). They are often found as part of a co-infection with other viruses and have also been found in respiratory secretions of asymptomatic individuals, so the pathogenesis of these viruses is presently unclear. Persistence of WUV in respiratory secretions of immunocompromised individuals has been described (66).

A third novel polyomavirus was discovered in 2008 in human merkel cell carcinomas (MCC) (67). Merkel cell carcinoma is an aggressive form

of skin cancer that has a 33% mortality rate which has been increasing in incidence in recent years (68). The rationale for pursuing a potential infectious etiology for MCC lies in the increased susceptibility among immunosuppressed patients (69). Feng et al. generated cDNA libraries from four MCC tumors and subjected them to high-throughput mass sequencing using 454 pyrosequencing. ~400,000 sequences were generated. Poor quality sequences and those that aligned to known human RNA, human chromosomes, mitochondria, or immunoglobulin sequences were removed from the final analysis. Of the remaining 2,395 sequences, one sequence was detected which had limited similarity to a primate polyomavirus. The full genome of the polyomavirus represented by this sequence, referred to as Merkel cell polyomavirus (MCV), was subsequently sequenced. Phylogenetic analysis determined that MCV was most closely related to African green monkey polyomaviruses and was only much more distantly related to JCV, BKV, WUV and KIV. Upon screening 10 other MCC tumors by PCR and Southern blotting, it was found that 80% of them were positive for MCV, whereas only 5/59 control tissues were positive. The virus was also found in 4/25 (16%) skin and non-MCC skin tumors. A striking observation was that the MCV DNA was clonally integrated into the MCC tumors, indicating that MCV integration is an early event during the transformation process and thus may contribute to tumorigenesis. Mechanistically, this may be the

consequence of the putative transforming ability of the MCV T-antigen or alternately it may be due to cellular alterations resulting from integration of the MCV genome, or a combination of the two. Follow up studies have generally corroborated the detection of MCV in the majority of Merkel tumor specimens (70-73). If future studies definitively implicate MCV in the etiology of Merkel cell carcinoma, it would be the first clear cut example of a polyomavirus playing a role in human cancer, after many years of debate over the role of polyomaviruses in human cancer (74). The discovery of Merkel polyomavirus was the first published report using 454 sequencing to discovery a novel virus. Of relevance for the continuing quest to identify potential infectious etiologies of human cancers, the fact that only 1 out of 400,000 sequence reads yielded a viral sequence suggests that as sequencing technologies continue to improve, it will be possible to detect viral nucleic acid sequences present at even lower abundances.

The identification of 3 novel polyomaviruses within a 12 month span by use of similar high throughput sequencing strategies underscores both the promise of such strategies as well as the fact that the diversity of viruses that constitute the human virome is vastly underestimated. Moreover, the identification of a novel human polyomavirus that appears to be strongly associated with a human tumor greatly broadens our paradigms of polyomaviruses. As the application of mass sequencing

strategies gain favor, we anticipate that most viral families will gain new members that may similarly broaden our understanding of those virus families. Even more tantalizing is the possibility that completely novel families of viruses may be identified through the use of such approaches.

Detection of an arenavirus in a fatal transplant cluster. Another example of the utility of 454 pyrosequencing in detecting viruses comes from identification of an arenavirus associated with a cluster of fatal transplant-associated diseases (75). Cerebrospinal fluid, serum, and tissue samples from two individuals who died 4-6 weeks after receiving organ transplants from the same donor were sequenced. 14 fragments out of 94,043 sequence reads had sequence similarity to LCMV. The results of this analysis suggested the specimens contained an arenavirus, which was then isolated in tissue culture from an infected kidney homogenate. Serum antibodies of both individuals reacted with the viral culture and were used to show immunostaining of viral antigens in the tissue specimens. While significant debate over whether this virus is indeed novel is ongoing (76), this case nonetheless illustrates two important points: (1) high throughput sequencing is a robust methodology for detection of viruses both known and novel; (2) as strictly sequence based methods gain prominence, viral taxonomy needs to be restructured to accommodate sequence based classification.

Avian bornavirus sequencing. The initial discovery of a novel avian bornavirus (as discussed previously) was based upon hybridization of specimens to a pan-viral microarray (52). In order to sequence the complete genome of the avian bornavirus, Kistler et al. utilized Solexa sequencing technology. A total of 1.4 million 33 bp reads were obtained from specimens positive for the virus. The sequences were filtered for read quality, sequence complexity, and the presence of inserts which reduced the number of usable reads to 600,000. Host sequences, approximately 50% of all the reads, were computationally subtracted from this pool by comparison to sequences of all available avian species. Comparison of the remaining sequence reads to all Borna Disease virus sequences identified 1,400 sequences that appeared to be derived from this novel virus. Conventional PCR-based methods were required to fill in the gaps of the genome where no sequences were identified in the mass sequencing. Retrospective analysis once the full genome of the avian bornavirus was obtained showed that there were twice as many sequences derived from the virus than initially identified. In total, sequences from the virus constituted 1% of all the sequence reads. This study was the first to use Solexa technology for *de novo* sequencing of a novel virus genome.

Identification of seal picornavirus. As demonstrated with some of the other methods, mass sequencing is not limited to the analysis of human samples. Application of mass sequencing to the supernatant of cells inoculated with a nasal swab from a seal yielded sequences that had 22-41% amino acid identity to known picornaviruses (77). The degree of divergence of this virus, called seal picornavirus 1 (SePV-1), to other picornaviruses suggests that this virus might represent a new genus in the family *Picornaviridae*. Moreover, this is the first example of a picornavirus infecting a marine mammal, which again underscores our relative ignorance regarding the diversity of viruses that surround us.

Mass sequencing of human stool. Recently, high throughput sequencing has been applied to the analysis of human stool samples collected from healthy individuals (78,79). Sequences from phage and plant viruses were dominant in the analysis of healthy stools. However, the question: "What is viral content of diarrheal stools?" remained a major question. Specifically, I wanted to know what viruses could be found in diarrheal stools, if I could find novel viruses, and if I found novel viruses, might they be linked to diarrhea or some other human disease? It therefore became my interest to apply cutting edge technologies like mass sequencing to examine which viruses are present in diarrhea

specimens and to search for novel viruses potentially associated with diarrhea.

Impact of diarrheal disease

Diarrhea is estimated to be the third leading cause of death due to an infectious disease (80). While data from socioeconomically developing countries is sparse, mathematical modeling based on the data that is available suggests that ~2 - 2.5 million children under the age of 5 die every year from diarrhea. The overall disease burden is much greater with an estimated 1.4 billion episodes of diarrhea occurring each year in developing countries and 9 million of these episodes requiring hospitalization (81). In developing countries, conditions such as poor sanitation, malnutrition, poor healthcare, and the rising rates of HIV infections in many regions all contribute to the burden of diarrheal disease (81,82). In developed countries, mortality rates have been drastically reduced by the use of oral rehydration therapy as a treatment for diarrhea. Despite this, there are still an estimated 211-375 million episodes of acute diarrhea that occur each year in the United States, with 1.8 million episodes resulting in hospital admissions (83).

There are over 20 known enteropathogens which include bacteria, viruses and parasites. A small percentage of the annual cases of diarrhea are attributable to parasite infections with most occurring in developing

countries. Viruses are responsible for causing the greatest number of the annual cases of diarrhea, while bacteria are responsible for a significant portion of the cases, but still less than those caused by viruses (84-86). Rotaviruses, astroviruses, caliciviruses, and adenoviruses are the major causes of viral diarrhea (87,88). These viruses were all identified as etiologic agents of diarrhea in the 1970s (89-92). While there are other viruses that have been suggested to be linked to diarrhea, there have been no major discoveries of new enteric viruses since the initial identification of the four major viral diarrhea pathogens (87,88). This becomes important in light of the fact that ~20-40% of cases of acute sporadic diarrhea are caused by unknown etiology (93). Likewise, it is estimated that up to ~12-40% of gastroenteritis outbreaks are also of unknown etiology even after extensive testing, suggesting that there is a diagnostic gap (94,95). In the United States alone, it is estimated that there are 5,000 yearly deaths which occur as a result of gastroenteritis of unknown etiology (GUE), accounting for 77% of the deaths in U.S. caused by diarrhea (96). Given that viruses play such a large role in the disease burden of diarrhea, that no new viral cause of diarrhea has been discovered in the last 30 years, and given that a number of viruses have been described morphologically upon observation of their presence in loose stools but not characterized further, there is good reason to believe

that there are other viral agents of diarrhea that have yet to be discovered (97-99).

Mass sequencing of pediatric diarrhea samples which were determined to be negative for known enteric agents by conventional assays such as PCR and enzyme immunoassays was carried and is described in Chapter 2. Use of this methodology in screening of acute sporadic diarrhea samples resulted in the identification of a novel astrovirus which will be referred to as Astrovirus MLB1 (AstV-MLB1), and which is described in Chapters 2-4. In addition, the use of mass sequencing to analyze diarrhea samples from a gastroenteritis outbreak resulted in the identification of an additional novel astrovirus referred to as Astrovirus VA1 (AstV-VA1), which is described in Chapter 5.

Astroviruses were first described in 1975 by electron microscopic examination of fecal extracts (100,101). Since this first discovery, astroviruses infecting cattle, sheep, cats, dogs, deer, chickens, turkeys, ducks, and bats have also been described in addition to the eight human serotypes that have been identified (102,103). The human astroviruses most frequently cause diarrhea in children under the age of 2, the elderly, and immunocompromised individuals (104). Typical symptoms are watery diarrhea, while vomiting, headache, fever, abdominal pains, and anorexia occurring to a lesser extent. Symptoms usually last 2-4 days. Astroviruses account for 10% of sporadic cases of non-bacterial diarrhea

and the magnitude of their impact on disease burden seems to be increasing with the development of more sensitive technologies for astrovirus detection and increased surveillance (105).

The name astrovirus comes from the observation that ~10% of the 28 nm particles have a star like morphology. Astroviruses are non-enveloped, single stranded, positive sense RNA viruses. The genomes, which range from 6.4 kb to 7.9 kb in length, are polyadenylated, contain 3 open reading frames (ORFs 1a, 1b, and 2), and have both 5' and 3' untranslated regions (Figure 1.3). Their genomic organization, which is similar to that of caliciviruses, is: (from 5' to 3') ORF 1a, which encodes a serine protease; ORF1b, which encodes the RNA dependent polymerase; and ORF 2, which encodes the structural proteins. A frameshift must occur during the translation of ORF 1a in order for ORF 1b to be translated. ORF 2, on the other hand, is translated from a sub-genomic RNA and produces a polyprotein which is cleaved by cellular proteases (106).

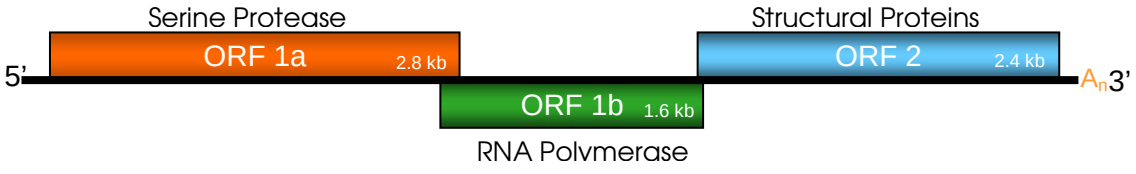


Figure 1.3: Schematic of genomic organization of astroviruses. Sizes of the open reading frames correspond to those of Human Astrovirus 1.

This dissertation manuscript describes work that was done to first identify which viruses can be found in diarrhea samples and which then led to the discovery of multiple novel astroviruses (AstV-MLB1 and AstV-VA1). The remaining body of this manuscript describes the efforts to further characterize these newly discovered astroviruses and to begin answering important questions raised by their discoveries. Characterization of these viruses may reveal that astroviruses contribute more to the global disease burden of gastroenteritis than previously recognized. Furthermore, the discovery of two novel astroviruses in such a short time frame indicates that there may be more astroviruses waiting to be discovered and that perhaps more energy should be devoted to studying astroviruses since the work described herein only begins to scratch the surface in terms of our understanding of astrovirus biology.

Chapter 2:
Metagenomic Analysis of Human Diarrhea: Viral Detection and Discovery

This work is published in *PLoS Pathogens* (2008), 4:e1000011

Authors: Stacy R. Finkbeiner[1†], Adam F. Allred[1†], Phillip I. Tarr[2], Eileen J. Klein[3], Carl D. Kirkwood[4], David Wang[1]
[†]These authors contributed equally to this work.

[1] Departments of Molecular Microbiology and Pathology & Immunology, Washington University School of Medicine, St. Louis, MO USA
[2]Department of Pediatrics, Washington University School of Medicine, St. Louis, MO USA
[3]Department of Emergency Medicine, Children's Hospital and Regional Medical Center, Seattle, Washington, USA
[4] Enteric Virus Research Group, Murdoch Childrens Research Institute, Royal Children's Hospital, Victoria, Australia.

ABSTRACT

Worldwide, approximately 1.8 million children die from diarrhea annually, and millions more suffer multiple episodes of nonfatal diarrhea. On average, in up to 40% of cases, no etiologic agent can be identified. The advent of metagenomic sequencing has enabled systematic and unbiased characterization of microbial populations; thus, metagenomic approaches have the potential to define the spectrum of viruses, including novel viruses, present in stool during episodes of acute diarrhea. The detection of novel or unexpected viruses would then enable investigations to assess whether these agents play a causal role in human diarrhea. In this study, we characterized the eukaryotic viral communities present in diarrhea specimens from 12 children by employing a strategy of 'micro-mass sequencing' that entails minimal starting sample quantity (<100 mg stool), minimal sample purification and limited sequencing (384 reads per sample). Using this methodology we detected known enteric viruses as well as multiple sequences from putatively novel viruses with only limited sequence similarity to viruses in Genbank.

INTRODUCTION

While traditional sequencing approaches are designed to characterize genomes of a single species of interest, metagenomic

approaches, such as mass sequencing, transcend species boundaries allowing one to explore the makeup of microbial communities. Such methods provide a holistic look at microbial diversity within a given sample, completely bypassing the need for culturing (107-111). Previous efforts in this field have explored the structure of virus communities in ecosystems as diverse as the ocean (107,112) and the human gut (78,79). To date, the reported metagenomic studies of human stool have been limited to analysis of 4 specimens collected from 3 healthy patients (78,79). To our knowledge, no metagenomic investigation of the viral diversity found in human diarrhea has previously been described. Human diarrhea is the third leading cause of infectious deaths worldwide and is responsible for ~ 1.8 million deaths in children under age five each year (80). Bacteria, protozoa and viruses have all been implicated as causal agents. Chief among the known etiologic agents are rotaviruses, noroviruses, astroviruses, and adenoviruses (113) However, it is estimated that on average up to 40% of diarrhea cases are of unknown etiology, suggesting that unrecognized infectious agents, including viruses, remain to be discovered (85,93,114-116). Mass sequencing affords an opportunity to explore the viral diversity (including both known and novel viruses) present in stool during acute episodes of diarrhea in a systematic and unbiased fashion, thereby laying the foundation for future studies aimed at

assessing whether any novel or unexpected viruses detected play a causal role in human diarrhea.

In this study, mass sequencing was applied to explore specifically the viral communities present in pediatric patients suffering from diarrhea. We anticipated that the viral communities would vary significantly from specimen to specimen and that it would be desirable to sample broadly from multiple patients to obtain an overall perspective on the diversity of viruses that might be present. Toward this end, a simple yet robust experimental strategy was developed that circumvented certain technical and economic limitations of conventional mass sequencing. In both previous viral metagenomic studies of the human gut, large quantities of fecal matter (~500g) were collected from adults and then extensively purified to enrich for viral particles (78,79). In contrast, pediatric samples provide considerably smaller volumes of stool; therefore, our strategy was designed to minimize the number of physical purification steps so that as little as 30 mg of archived fecal matter could be analyzed. Here we present data generated by performing what we refer to as `micro-mass sequencing' of several hundred sequence reads per sample from 12 different patients with acute diarrhea. This analysis provides evidence for the detection of known enteric viruses, viral co-infections, and novel viruses.

RESULTS

Aggregate library analysis. Metagenomic analysis was carried out on fecal samples collected from 12 distinct pediatric patients suffering from acute diarrhea. Patient characteristics are shown in Table 2.1. A sequence independent PCR strategy was employed to amplify the extracted nucleic acids from each sample (41). 384 clones were sequenced for each sample library. Because the goal of this project was to define the diversity of viruses present in the clinical specimens regardless of their relative abundance, nearly identical sequence reads were clustered to generate a set of non-redundant sequence reads. Unique, high quality sequence reads were then classified into broad taxonomic groups based on the taxonomy of the most frequent top scoring BLAST matches for each sequence. A total of 4,608 sequences were generated, of which 3,169 passed through a quality filter and 2,013 of those contained unique sequence information. Of the unique sequences passing through the filter, 1,457 (72%) could be identified by similarity to sequences in the Genbank nr database based on tBLASTx (E-value $\leq 10^{-5}$) alignments. The remaining 556 (28%) sequences had no significant similarity to any sequences in the nr database and were therefore categorized as being of 'unknown' origin. The 1,457 identifiable sequences were further classified into categories based on their proposed origin (Figure 2.1). 519 (35.6%) were most similar to eukaryotic viruses, 25

Table 2.1: Sample Information.

Sample	Year Collected	Age of Patient	# of high quality sequence reads	# of unique reads	Average unique read length (bp)
D01	2005	14 mo	365	166	526
D02	1998	10 mo	348	104	499
D03	1984	NA	302	281	506
D04	1984	4 mo	311	154	626
D05	1980	NA	243	168	563
D06	2003	11 mo	153	132	393
D07	1999	23 mo	352	186	617
D08	1999	35 mo	302	167	255
D09	1981	NA	302	294	491
D10	1983	20 mo	195	146	447
D11	1978	NA	253	103	367
D12	2005	8 mo	198	129	300

(1.7%) to phage, 857 (58.8%) to bacteria, 3 (0.2%) to fungi, and 20 (1.4%) to human sequences. The remaining 33 (2.3%) were most similar to sequences that did not fall into the other previous categories and were consequently labeled as 'other'. For example, some of the sequences had significant hits to mouse, fish, and plant genomes.

Individual library statistics. 384 clones were sequenced for each individual sample. The proportion of high quality sequences for each sample varied between 40% and 95% of the total clones (Table 2.1). The percentages of unique sequences per sample ranged from 41% to 97% of the high quality reads (Table 2.1). The average length of the unique, high quality sequences ranged from 255 to 626 bp. Viral sequences constituted

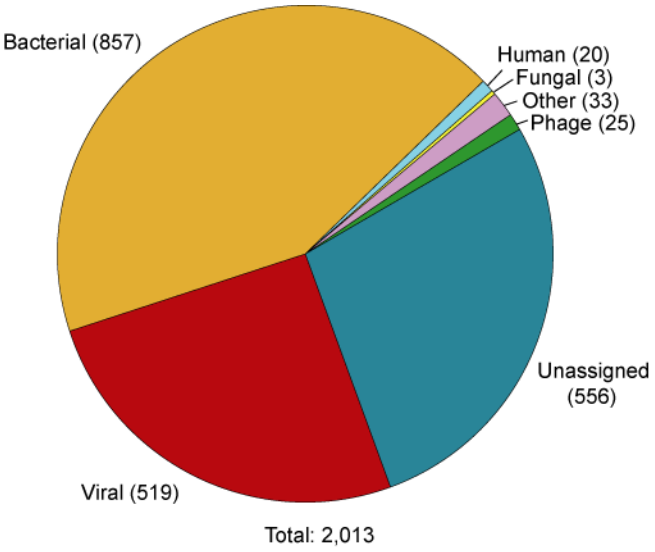


Figure 2.1: Composite analysis of all sequences
Sequences from all 12 libraries were categorized based on the best tBLASTX scores (E-value: $<10^{-5}$) as viral, phage, bacterial, human, fungal, other, or unassigned. Numbers in parenthesis represent the number of sequences in each category.

between 0-100% of the reads in each library (Figure 2.2). Some libraries (e.g., D01 and D05) were predominantly composed of viral sequences (64% and 95% respectively), whereas others consisted largely of bacterial (e.g., D08 and D12) or unassigned (e.g., D03 and D07) sequences. Based on the initial BLAST classification criteria, sequences with similarity to viruses from 7 different viral families and three unclassified genera (picobirnavirus, anellovirus and mimivirus) were detected in the 12 different samples (Fig. 2.2). Five of the samples (D03, D05, D06, D08, and D12) contained sequences from at least two different virus families known to infect humans.

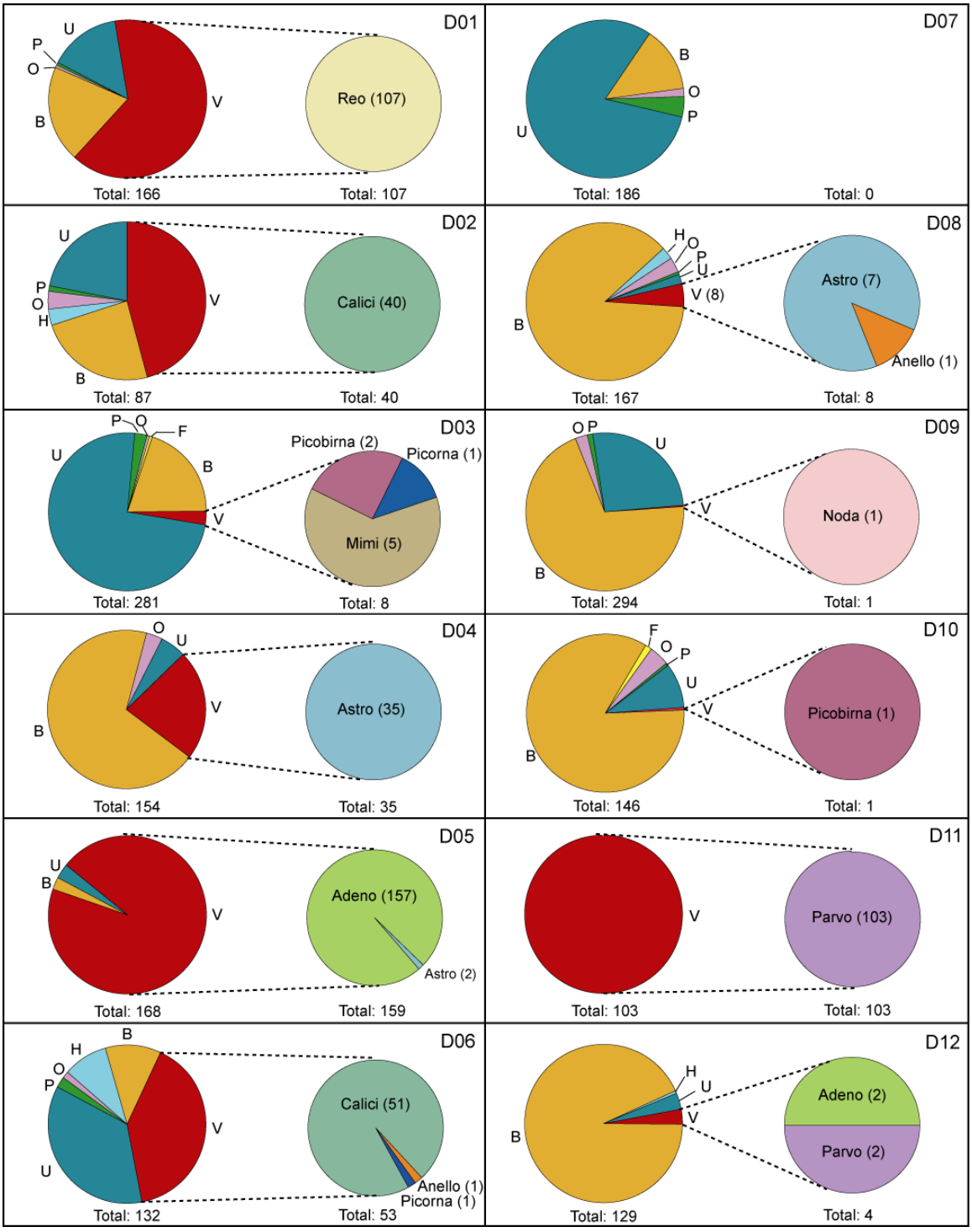


Figure 2.2: Categorization of sequence reads based on best tBLASTX scores (E-value: <10^{-5})

Pies on the left side of each box depict the categorization of sequences from individual samples by phylotype: viral (V); phage (P); bacterial (B); human (H); fungal (F); other (O); and unassigned (U). Pies on the right side of each box depict further characterization of viral sequences by viral families/taxa: *Reoviridae* (Reo); *Caliciviridae* (Calici); *Astroviridae* (Astro); anellovirus (Anello); picobirnavirus (Picobirna); *Picornaviridae* (Picorna); mimivirus

(Mimi); *Nodaviridae* (Noda); *Adenoviridae* (Adeno); *Parvoviridae* (Parvo). Numbers in parentheses indicate the number of sequence reads in each category.

Detection of known viruses. The first specimen analyzed was a positive control stool specimen that had tested positive for rotavirus (D01) by enzyme immunoassay. It was our expectation that this sample would yield sequences derived from the infecting rotavirus. In this library, 107 non-redundant sequence reads were identified as viral in origin, almost all of which possessed ≥ 90% amino acid (aa) BLAST identity to known rotavirus sequences in Genbank. The sequence data included cloned fragments from all 11 RNA segments of the rotavirus genome.

An additional 11 stool specimens were then selected that had tested negative in conventional PCR and enzyme immunoassays for the known diarrhea viruses (rotaviruses, caliciviruses, astroviruses, and adenoviruses). Despite such screening, sequences derived from the canonical enteric viruses were detected in a number of samples. For example, calicivirus sequences were detected in D02 and D06, astrovirus sequences in D04, and adenoviruses were detected in D05 and D12. Almost all individual sequence reads in these cases possessed >90% aa identity to existing viral sequences in Genbank.

Adeno-associated virus (AAV), a member of the *Parvoviridae* family, was detected in two samples, D11 and D12. These viruses are known to infect the gastrointestinal tract, but are not thought to be enteric

pathogens. For productive infections or reactivation from a latent state, AAV requires co-infection with a helper virus that is most commonly an adenovirus or less typically, a herpesvirus (117). In D12, adenovirus sequences were detected. No additional viruses were detected in D11.

Detection of novel virus sequences. In many of the libraries, individual sequence reads were detected that possessed ≤ 90% aa identity to their highest scoring BLAST hit (representative sequences are listed in Table 2.2) suggesting that these sequences might be derived from novel viruses. In part because BLAST alignments are based on local sequence comparisons, BLAST is not an optimal method for making taxonomic assignments. In order to more accurately and precisely assess the relationship of these sequences to known viruses, we generated phylogenetic trees using the maximum parsimony method (118). In cases where more than one sequence read hit the same region of a genome, only one representative sequence read is listed in Table 2.2 and phylogenetic trees are shown for only these representative sequences (Fig. 2.3, 2.4 and Fig. 2.S1-S4). Phylogenetic analysis revealed that many of the

Table 2.2: Selected sequence reads with limited BLAST identity to known viruses				
Sample	**Sequence Read Accession #**	**Identity to top hit**	**Top Hit (Accession #)**	**Virus Family/Taxa**
D03	ET065742	78%	Human picobirnavirus strain 1-CHN-97 (AF246939)	Picobirnavirus
D03	ET065743	90%	Human coxsackievirus A19 (AF499641)	Picornaviridae
D06	ET067042	74%	Human enterovirus 91 (AY697476)	Picornaviridae
D06	ET067045	66%	TTV-like mini virus (AB026931)	Anellovirus
D06	ET067040	79%	Snow Mountain virus (AY134748.1)	Caliciviridae
D06	ET067041	88%	Norovirus C14 (AY845056.1)	Caliciviridae
D08	ET065575	57%	Human astrovirus 4 (AY720891)	Astroviridae
D08	ET065582	67%	Human astrovirus 5 (DQ028633)	Astroviridae
D08	ET065578	45%	TT virus (AB041963)	Anellovirus
D09	ET066010	35%	Epinephelus septemfasciatus nervous necrosis virus (AM085331)	Nodaviridae
D10	ET066456	81%	Human picobirnavirus 2-GA-91 (AF245701)	Picobirnavirus

sequences were divergent from known sequences on the order that approximated a distinct subtype or genotype (Figures 2.S1-S4). This included two libraries with picobirnaviruses (D03, D10) (Figures 2.S1), two with picornaviruses (D03, D06) (Figures 2.S2), two with anelloviruses (D06, D08) (Figure 2.S3), and one with a norovirus (D06) (Fig. 2.S4).

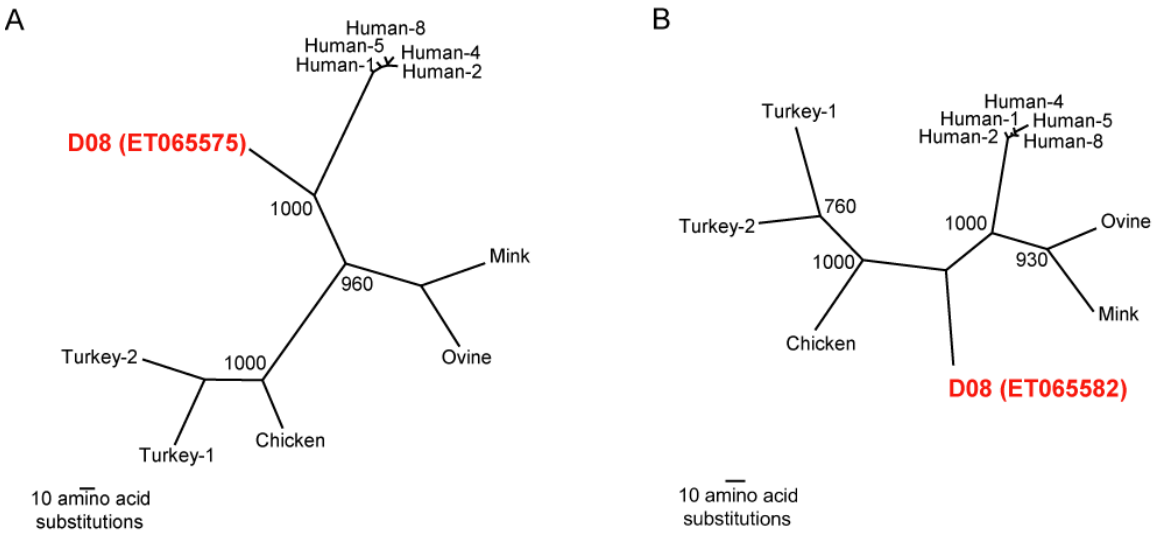


Figure 2.3: Phylogenetic analysis of highly divergent astrovirus-like sequence reads. Maximum parsimony phylogenetic trees were generated by comparing the translated amino acid sequence of individual sequence reads to the corresponding sequences from known astroviruses. 1,000 replicates were generated with bootstrap values over 700 shown. A) Representative sequence read mapping to astrovirus serine protease ORF (Accession number ET065575); B) Representative sequence read mapping to astrovirus RNA polymerase (Accession number ET065582).

In several instances, much more highly divergent sequences were detected that suggested that novel virus species might be present. The library generated for sample D08 included 7 unique sequence reads derived from two loci that displayed 52-67% aa identity to human astroviruses. Phylogenetic analysis of the individual sequence reads suggested that a novel astrovirus was present in D08 (Figure 2.3). These sequence reads were assembled into two contigs, one of ~800 bp that mapped to ORF1a and one of ~500 bp that mapped to ORF1b. RT-PCR and subsequent sequencing of the amplicon confirmed the presence of the contigs in the original RNA extract as well as the contig assemblies (data not shown). Phylogenetic analysis of the two contigs yielded trees

essentially identical to those generated from the individual sequence reads (data not shown).

In sample D09, we detected one sequence read which exhibited limited similarity to viruses in the family *Nodaviridae* (Table 2.2). RT-PCR of this sample using primers designed from the sequence read confirmed the presence of a 229 bp fragment in the original RNA extract (data not shown). Phylogenetic analysis of the sequence of the RT-PCR product demonstrated that the nodavirus in sample D09 was highly divergent from other known nodaviruses (Figure 2.4).

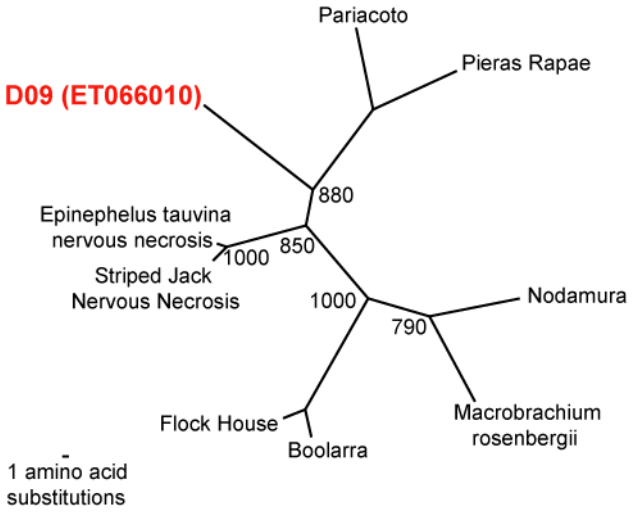


Figure 2.4: Phylogenetic analysis of a highly divergent nodavirus-like sequence read. Maximum parsimony phylogenetic trees were generated by comparing the translated amino acid sequence of one sequence read (Accession number ET066010) to the corresponding RNA polymerase sequences of nodaviruses. 1,000 replicates were generated with bootstrap values over 700 shown.

Finally, one sample, D03, contained five sequence reads that, based on the top tBLASTX hits, contained 47% to 52% aa identity to endonuclease genes in the amoeba-infecting virus *Acanthamoeba*

polyphaga mimivirus. These sequences also possessed approximately similar levels of sequence identity to a number of bacterial genomes and phage genomes containing putative endonuclease proteins. Phylogenetic analysis comparing the sequence reads to the top scoring BLAST hits (Figure 2.S5) did not conclusively clarify the origin of these sequences. Further experimentation will be required to unambiguously determine if these sequences are derived from a mimi-like virus, phage, or a bacterial species.

Unassigned sequences. Some sequences in the libraries had no significant hits to any sequences in the Genbank nr database. Samples D03 and D07 had a large abundance of these 'unassigned' reads. Relaxing E-value thresholds for designating various sequence categories resulted in the ability to classify a greater number of these unassigned sequences; however, many of these classifications likely represent artifactual alignments. Viral assignments remained largely unaffected, even when E-value thresholds as permissive as 10 were applied.

DISCUSSION

We examined the diversity of viral communities in stools from 12 children with diarrhea using a strategy we describe as 'micro-mass sequencing'. This strategy, which entails crude purification of fecal suspensions, nucleic acid purification, random PCR amplification, and

cloning and sequencing of several hundred colonies, effectively detected known enteric viruses, viral co-infections, and novel viruses. In most traditional metagenomic studies, large sample volumes are subjected to multiple stages of filtration and purification before sequencing. For example, in previous metagenomic studies of the gut, 500g of fecal samples were initially collected for the analyses. Because clinical pediatric diarrhea specimens are much more limited in volume, we chose to both minimally purify the samples and to employ a random PCR amplification strategy. These combined steps enabled us to rapidly generate sequencing libraries from small quantities of archived stools (30-100 mg). Furthermore, we wished to sample broadly from multiple patients because of the large number of viruses known or suspected to be associated with diarrhea. Therefore, rather than sequence few specimens in great depth as has been done previously (10,000 sequences per sample) (78), we focused on sequencing fewer clones (384 per sample) from more samples (12 specimens).

Our analysis detected viruses, bacteria, host, phage and other sequences (Figures 2.1 and 2.2). The presence of non-viral sequences in the libraries was not surprising as only minimal efforts were made to enrich for viral sequences. In fact, the goal of this strategy was to manipulate the specimens as little as possible in the interest of simplicity. Even so, in a few libraries, 100% of the sequence reads were of viral origin. Additional

processing, such as treating the specimens with DNase, reduced the background signal and increased the percentage of viral reads in some instances (data not shown).

Viral sequences were detected in all but one sample. Interestingly, a number of DNA viruses (bacteriophages, adenoviruses, and adeno-associated viruses) were detected in our analysis, despite our use of a methodology focused on purification of RNA. While it is possible that RNA transcripts from these viruses were purified (119), it is more likely that viral DNA was co-purified with RNA, as is common in other RNA purification methods (120). PCR analysis of samples D05 and D11 in the absence of reverse transcription, yielded positive results for adenovirus and adeno-associated virus, respectively, indicating that viral DNA was present in the RNA preparations (data not shown).

Analysis of this initial cohort of 12 specimens yielded a wealth of original findings. In contrast to previous metagenomic studies of stool (78), a number of known human viruses were detected in these clinical specimens. These included common enteric pathogens such as rotavirus, adenovirus, calicivirus, and astrovirus. In addition, putatively benign adeno-associated viruses (AAV) were also detected which are not generally associated with human diarrhea. Aside from one sample known to contain rotavirus, we intended to analyze the viral communities present in samples that were not infected by known enteric pathogens in order to

identify viruses that might be responsible for the unexplained cases of diarrhea. The fact that micro-mass sequencing detected these canonical viruses in some of the specimens, despite conventional diagnostic testing by EIA and PCR, underscores the sensitivity limits of conventional diagnostics.

Detection of novel viruses. Sequences were detected in this study from at least 9 putatively novel viruses. For 7 of these sequences, the degree of divergence observed based on phylogenetic analysis suggested that they might represent novel virus subtypes or genotypes of picobirnavirus, enterovirus, TT virus and norovirus (Figures 2.S1-S4). Picobirnaviruses belong to an unclassified genus of double stranded RNA viruses and have been detected in fecal matter from human and other animals both with and without diarrhea (121). Only a limited number of picobirnavirus sequences have been previously described in the literature and thus the identification of two novel picobirnaviruses significantly expands the known diversity of this taxonomic group, underscoring the unrecognized viral diversity inhabiting the human body.

Sequences representing a divergent norovirus were detected in sample D06 (Figure 2.S4). Phylogenetic analysis of individual sequence reads that mapped to the RNA polymerase and the NS4 regions of human norovirus suggested that these sequences were derived from a novel or

unsequenced member of norovirus genogroup 2. In the initial screening by conventional PCR, this sample tested negative for norovirus. Upon closer examination, four mutations were observed in one of the PCR primer binding sites, which plausibly hindered the PCR screening assay (85).

In two samples, much more highly divergent sequences were detected. In D08, phylogenetic analysis of 7 unique sequence reads strongly suggested that a novel astrovirus species was present (Figure 2.3). The observed sequence variation between these sequence reads and the known astrovirus genomes greatly exceeds the variation that exists between the 8 known serotypes of human astrovirus, suggesting that this virus is not simply another serotype of the known astroviruses. Astroviruses are non-enveloped, single stranded, positive sense RNA viruses that account for up to 10% of sporadic diarrhea cases (105). Infections with astroviruses most frequently cause watery diarrhea lasting 2-4 days, and, less commonly vomiting, headache, fever, abdominal pain, and anorexia in children under the age of 2, the elderly, and immunocompromised individuals (104). The detection of this genetically distinct astrovirus raises the question as to whether or not this is an authentic human virus, and if so, whether or not it is a causal agent of human diarrhea.

Another novel sequence detected appeared by phylogenetic analysis to belong to the family *Nodaviridae* (Figure 2.4). Nodaviruses are

small single-stranded, positive sense, bipartite RNA viruses, divided into two genera, the alphanodaviruses (insect viruses) and the betanodaviruses (fish viruses). Currently, none of the established family members are known to naturally infect mammals although experimental manipulation of the viral genome has enabled viral replication in a wide array of organisms including mammals (122). While it is tempting to speculate that this might represent the first instance of human infection with a nodavirus, further experimentation such as serological analysis is required to definitively answer this question. Another plausible explanation is that the virus may be present simply as a result of consumption of fish infected by the virus. A prior report describing the presence of plant virus RNAs in human stool has similarly been attributed to dietary exposure (78). Incidentally, some fish genomic sequences were detected in this particular sequence library (D09 "other" bin) supporting the possibility of dietary exposure. However, the potential piscine origin of this virus would not necessarily preclude its role as an etiologic agent of human disease.

The micro-mass sequencing approach, like any other experimental methodology capable of detecting novel viruses (such as culture or degenerate PCR), cannot of course by itself determine whether the newly detected agent is pathogenic. However, this strategy can generate novel, testable hypotheses such as "Are these novel viruses involved in the

etiology of human diarrhea?" and "What is the true host of these viruses?" that could not be asked in the absence of the knowledge that these viruses existed.

Unassigned reads. 556 out of the 2013 (28%) unique high quality sequences were binned as unassigned by the BLAST criteria. Of these, 23 were identified as containing repetitive elements or low-complexity sequence by RepeatMasker (123,124) thus explaining the lack of meaningful BLAST alignments. The origin of the remaining 533 sequences that were unassigned is uncertain, but they could be derived from unannotated host genome, novel or unsequenced microbes, or dietary sources which have not been sequenced. However, it is also possible that some of these sequences could represent viruses that have no appreciable similarity to sequences of currently known viruses. Extracting more telling information from these sequences is a challenging problem that will require the development of new computational measures capable of detecting more distant evolutionary relationships than is possible with existing methods. In addition, as more genome sequences from diverse organisms and other genomic/metagenomic projects become available, sequence similarity based methods may identify a greater fraction of these currently unassigned sequences.

Diagnostic Applications and Implications. Our data suggest that micro-mass sequencing might be of great diagnostic utility for a number

of reasons. First, viruses escaping detection in conventional assays were detected by micro-mass sequencing. In theory, the sensitivity of this strategy is limited only by the depth of sequencing. As demonstrated here, even shallow sequencing performed better than conventional diagnostics in some instances. In addition, the unbiased nature of the method enabled detection of viruses not conventionally tested for. Moreover, co-infections were detected in multiple samples. Furthermore, for multi-segmented viruses such as rotaviruses, reassortment of segments between species is a major mechanism of viral evolution that can lead to the emergence of more virulent strains (125). Complete genome sequencing of all segments simultaneously would yield completely unambiguous identification of the viral genotype. In contrast to typical PCR or antibody based assays that target a single segment or protein, micro mass sequencing detected all 11 genomic RNA segments of rotavirus. In terms of technical practicality, samples were only minimally manipulated relative to traditional metagenomic sequencing (78,107,109,112), thereby avoiding the time, labor, and use of specialized equipment required to concentrate the specimens, rendering this methodology potentially amenable to use in diagnostic laboratories. As sequencing costs diminish and efficiencies improve, mass sequencing could become a powerful diagnostic tool.

In summary, we have shown that micro-mass sequencing can define the diversity of viral communities found in fecal samples from diarrhea patients. Both known viruses and novel viruses were detected by sequencing only a few hundred colonies from each sample library. These studies will serve as the springboard for further interrogations of the roles of these diverse viruses in the gastrointestinal tract. Finally, our detection of multiple novel viruses in this initial, limited exploration of a dozen samples suggests that broader sampling of patient specimens is likely to be highly fruitful in terms of identification of additional novel viruses.

MATERIALS AND METHODS

Clinical Archived Stool Specimens.

Melbourne Cohort. Stool samples were collected from children under the age of 5 who were admitted to the Royal Children's Hospital, Melbourne, Victoria, Australia with acute diarrhea between 1978 and 1999.

Seattle Cohort. Stool samples were collected between 2003-2005 at the Emergency Department of the Children's Hospital and Regional Medical Center in Seattle, Washington, USA as part of a prospective study attempting to discern the cause of unexplained pediatric diarrhea.

Diagnostic testing of stool specimens for known microbial diarrheagenic agents.

Melbourne Cohort. Specimens were tested by routine enzyme immunoassays (EIA) and culture assays for rotaviruses, adenoviruses, and common bacterial and parasitic pathogens as previously described (85). RT-PCR assays were used to screen specimens for the presence of caliciviruses and astroviruses (85,126) .

Seattle Cohort. Specimens were tested for the presence of a number of bacterial species (*Campylobacter jejuni*, *Escherichia coli* O157:H7 and non-O157:H7 Shiga toxin-producing *E. coli*, *Salmonella*, *Shigella*, and *Yersinia)* following standard culture assays, *Clostridium difficile* toxin by a cytotoxicity assay, parasites by microscopy and antigen testing (84). Additionally, samples were tested by EIA for rotaviruses, adenoviruses, noroviruses 1 & 2, and astroviruses (Meridian Biosciences, DAKO). This study was approved by the institutional review boards of the CHRMC and of Washington University.

Library construction and mass sequencing.

Chips of frozen archived fecal specimens (~30-150mg) were resuspended in 6 volumes of PBS. A subset of the archived specimens had been

previously diluted and were further diluted 1:1in PBS. The stool suspensions were centrifuged (9,700 x g, 10 minutes) and supernatants were harvested and then passed through 0.45µm filters. RNA was extracted from 100µL of the filtrates using RNA-Bee (Tel Test, Inc., Friendswood, Texas) according to manufacturers instructions. Approximately, 100-300 nanograms of RNA from each sample was randomly amplified following the Round AB protocol as previously described (41). The amplified nucleic acid was cloned into pCR4 using the TOPO cloning kit (Invitrogen, Carlsbad, CA), and transformed into Top10 bacteria. Positive colonies were subcloned into 384 well plates, DNA was purified using magnetic bead isolation, and followed by sequencing using standard Big Dye terminator (v3.1) sequencing chemistry and the universal primer M13 reverse. Reactions were ethanol precipitated and resuspended in 25uL of water prior to loading onto the ABI 3730xl sequencer.

Analysis of sequence reads.

Sequence traces were subjected to quality assessment and base-calling using Phred (127,128). Lucy (129) was used to trim vector and low quality sequences. Default parameters were used except that high quality sequences identified by Lucy were allowed to be as short as 75 nucleotides. To define the set of reads with unique sequence content in each library, sequences that passed the quality filter were clustered using

BLASTClust from the 2.2.15 version of NCBI BLAST to eliminate redundancy. Sequences were clustered based on 98% identity over 98% sequence length, and the longest sequence from each cluster was aligned to the NCBI nr database using the tBLASTx algorithm (130). An E-value cutoff of 1e-5 was applied. Sequences were phylotyped as human, bacterial, phage, viral, or other based on the identity of the best BLAST hit. Sequences without any hits having an E-value of 1e-5 or better were placed in the "Unassigned" category. All eukaryotic viral sequences were further classified into viral families in similar fashion.

Trimmed, high quality sequences that were not found by RepeatMasker to contain repetitive or low-complexity sequence have been deposited in Genbank (Accession numbers ET065304 through ET067293).

Phylogenetic analysis.

ClustalX (1.83) was used to perform multiple sequence alignments of the protein sequences associated with select sequence reads. Available nucleotide or protein sequences from known viruses were obtained from Genbank for inclusion in the phylogenetic trees. Selected sequences from Genbank included those with the greatest similarity to the sequence read in question based on the BLAST alignments as well as representative sequences from all major taxa within the relevant virus family. The protein alignments created by ClustalX were input into PAUP (118), and maximum

parsimony analysis was performed using the default settings with 1,000 replicates.

Astrovirus trees: Human astrovirus 1 (NC_001943); Human astrovirus 2 (L13745); Human astrovirus 3 (AAD17224); Human astrovirus 4 (DQ070852); Human astrovirus 5 (DQ028633); Human astrovirus 6 (CAA86616); Human astrovirus 7 (AAK31913); Human astrovirus 8 (AF260508); Turkey astrovirus 1 (Y15936); Turkey astrovirus 2 (NC_005790); Turkey astrovirus 3 (AY769616); Chicken astrovirus (NC_003790); Ovine astrovirus (NC_002469); and Mink astrovirus (NC_004579).

Nodavirus tree: Striped Jack Nervous Necrosis virus (Q9QAZ8); Macrobrachium rosenbergii nodavirus (Q6XNL5); Black Beetle virus (YP_053043.1); Flockhouse virus (NP_689444.1); Epinephelus tauvina nervous necrosis virus (NC_004136.1); Nodamura virus (NC_002691.1); Boolarra virus (NC_004145.1); Pariacoto virus (NC_003692.1); and Redspotted grouper nervous necrosis virus (NC_008041.1).

Picornavirus trees: Human coxsackievirus A1 (AAQ02675.1), Human coxsackievirus A18 (AAQ04836.1), Human coxsackievirus A19 (AAQ02681.1), Human coxsackievirus A21 (AAQ04838.1), Human coxsackievirus A24 (ABD97876.1), Human poliovirus 1 (CAD23059.1), Human coxsackievirus A2 (AAR38840.1), Human coxsackievirus A4 (AAR38842.1), Human coxsackievirus A5 (AAR38843.1), Human coxsackievirus A16 (AAV70120.1), Human enterovirus 89 (AAW30683.1), Human enterovirus 91 (AAW30700.1), Human enterovirus 90 (BAD95475.1), Human enterovirus 71 (CAL36654.1), Echovirus 1 strain Farouk (AAC63944.2), Human coxsackievirus B2 (AAD19874.1), Human enterovirus 86 (AAX47040.1), Human coxsackievirus B5 (AAF21971.1), Human echovirus 29 (AAQ73089.1), Human enterovirus 68 (AAR98503.1), Human enterovirus 70 (BAA18891.1), Bovine enterovirus (NP_045756.1), Porcine enterovirus A (NP_653145.1), Porcine enterovirus B (NP_758520.1), Simian enterovirus A (NP_653149.1), Human rhinovirus A (ABF51203.1), Human rhinovirus B (NP_041009.1).

Picobirnavirus trees: Human picobirnavirus strain 1-CHN-97 (AF246939.1), Human picobirnavirus strain 4-GA-91 (AF246940.1), Human picobirnavirus strain Hy005102 (NC_007027.1), Human picobirnavirus strain 2-GA-91

(AF245701.1), Human picobirnavirus strain 1-GA-91 (AF246612.1), Porcine picobirnavirus 2 (EU104360.1).

Anellovirus trees: TGP96 Torque teno virus (AB041962), Pt-TTV8-II Torque teno virus (AB041963), CBD231 TTV-like mini virus (AB026930), Mf-TTV9 Torque teno virus (AB041959), Mf-TTV3 Torque teno virus (AB041958), KC009/G4 Torque teno virus (AB038621), TA278/G1 Torque teno virus (AB008394), Pt-TTV6 Torque teno virus (AB041957), TUS01/G3 Torque teno virus (AB017613), PMV/G2 Torque teno virus (AF261761), JT33F/G5 Torque teno virus (AB064606), MD1-073 Torque teno midi virus (AB290918), MD2-013 Torque teno midi virus (AB290919), Tbc-TTV14 Torque teno virus (AB057358), Sd-TTV31 Torque teno virus (AB076001), Fc-TTV4 Torque teno virus (AB076003), Cf-TTV10 Torque teno virus (AB076002), So-TTV2 Torque teno virus (AB041960), At-TTV3 Torque teno virus (AB041961).

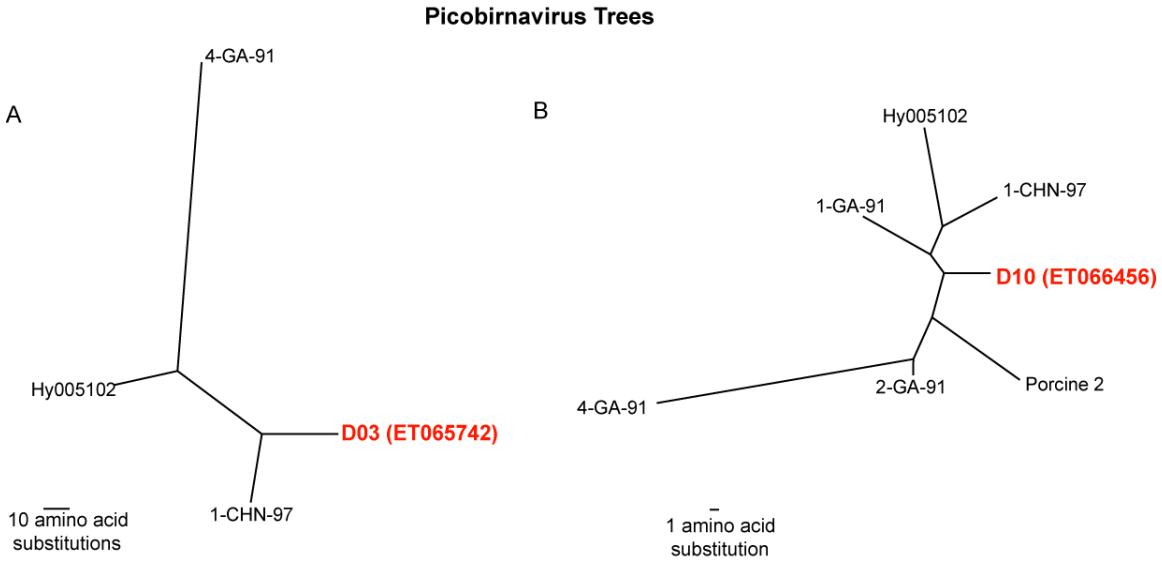
Calicivirus trees: Camberwell (AAD33960.1), MD-2004 (ABG49508.1), Carlow(ABD73935.1), Snow Mountain virus (AAN08111.1), Mc37 (AAS47823.1), Hawaii(AAB97767.2), Norwalk(AAB50465.1), Southampton (AAA92983.1), Chiba(BAB18266.1), Hesse(AAC64602.1), BoJena-DEU-98 (CAA09480.1), Murine (AAO63098.2), SU17(BAC11827.1), Dumfries (AAM95184.2), SU25-JPN(BAC11830.1), SU1-JPN(BAC11815.1), Desert Shield (AAA16284.1), Melksham (CAA57461.1), Toronto-24 (AAA18929.1), Sw918 (BAB83515.1), OH-QW101 (AAX32876.1).

Endonuclease-like sequences for D03 tree (mimvirus-like sequences): *Bacteroides caccae* (ZP_01959575.1), *Acanthamoeba* mimivirus (YP_142599.1), *Eubacterium dolichum* (ZP_02077753.1), *Staphylococcus* phage K (YP_024462.1), *Lactobacillus* phage LP65 (YP_164778.1), *Lactococcus* phage bIL170 (NP_047162.1), *Lactococcus* phage r1t (NP_695069.1), *Burkholderia vietnamiensis* G4 (YP_001119011.1), *Streptococcus pyogenes* (NP_607538.1), *Tetrahymena thermophila* (XP_001029162.1), *Bacteroides vulgatus* (YP_001300673.1)

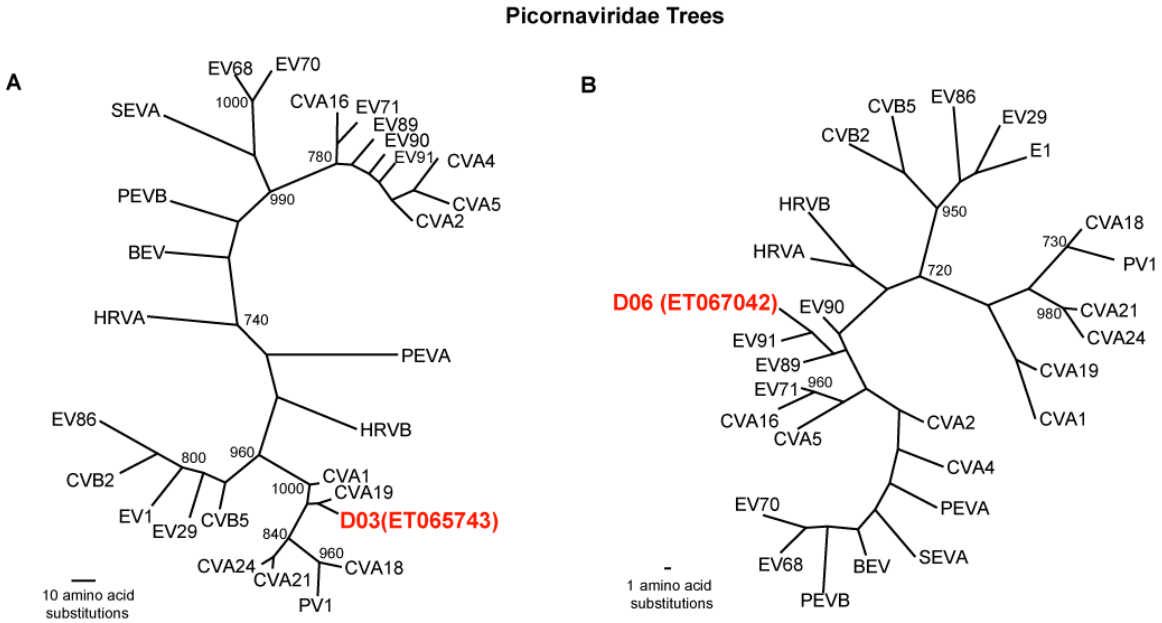
ACKNOWLEDGEMENTS

We would like to thank Henry Huang for helpful advice regarding the phylogenetic analysis. This work was funded in part by the Pilot Sequencing Program sponsored by the Center for Genome Sciences at Washington University (DW), an NHMRC RD Wright Research Fellowship (CK), and by a USDA Grant NRI 2002-35212-12335 (PT).

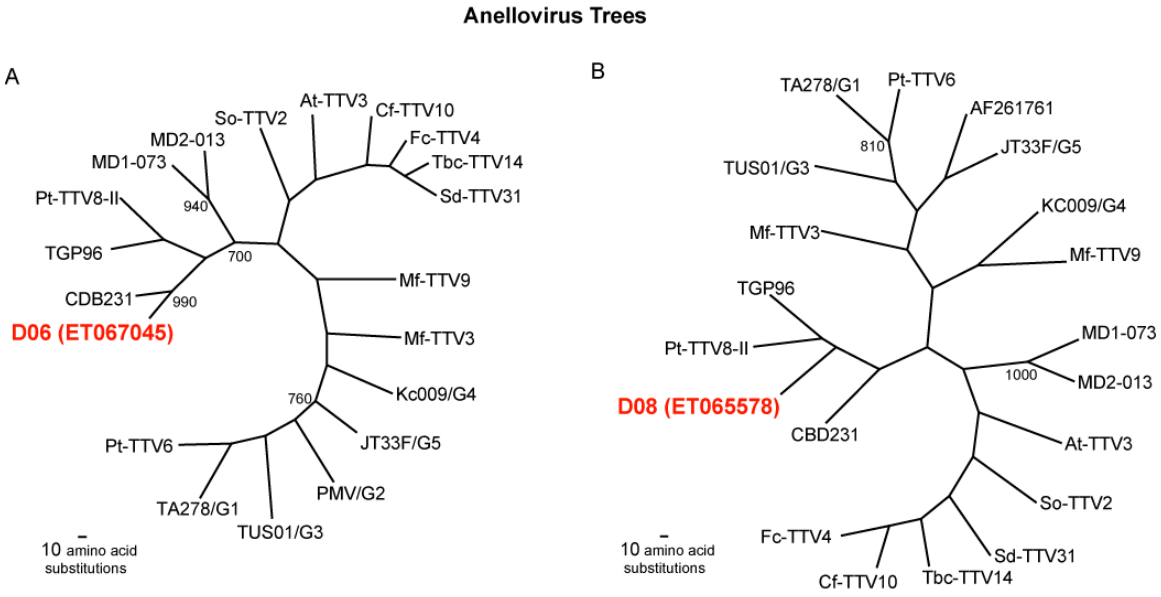
Supplemental Figures:



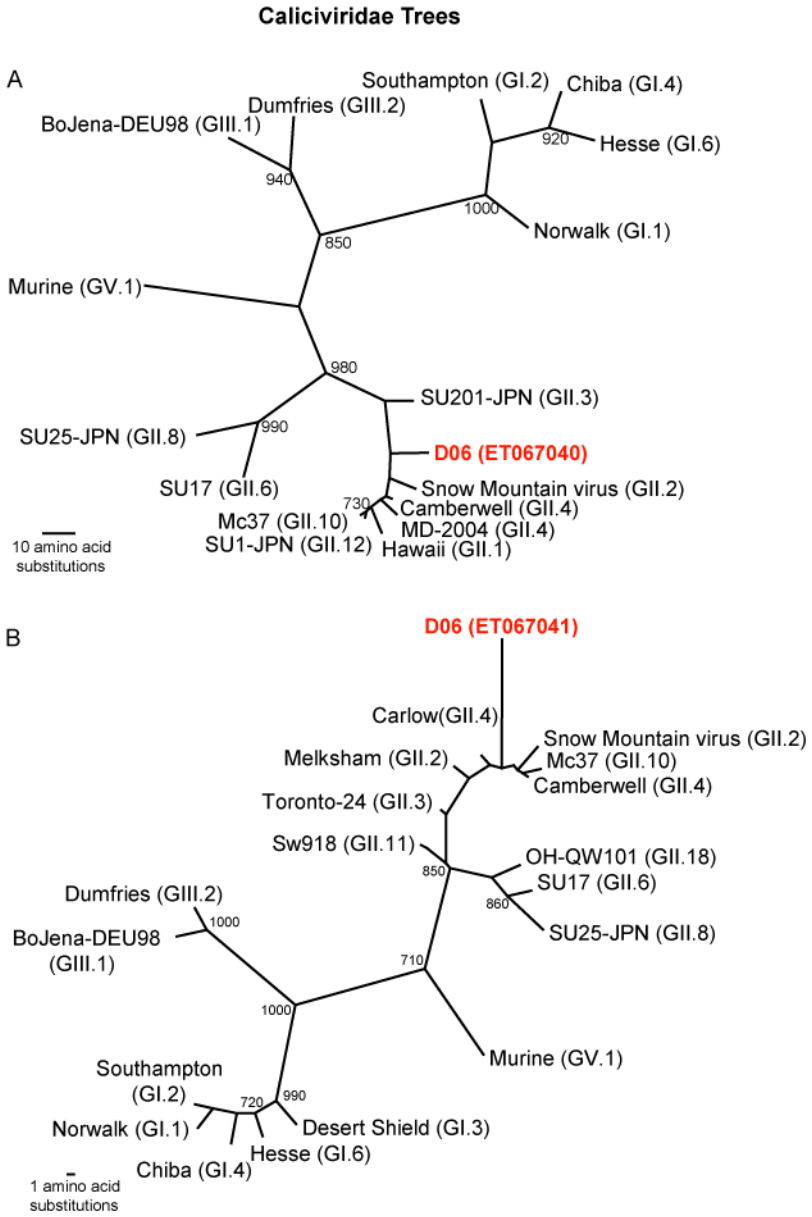
Supplemental Figure 2.S1: Phylogenetic analysis of picobirnavirus-like sequence reads. Phylogenetic trees were generated by comparing the translated amino acid sequence of individual sequence reads to members of the unclassified taxa picobirnavirus. The trees were created using the maximum parsimony method with 1,000 replicates. Bootstrap values over 700 are shown.



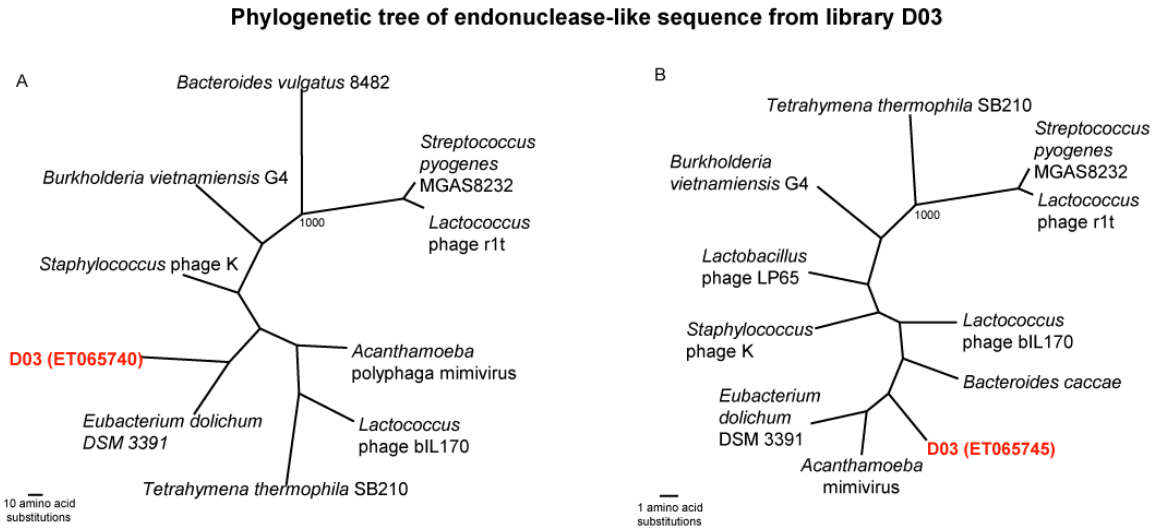
Supplemental Figure 2.S2: Phylogenetic analysis of *Picornaviridae*-like sequence reads. Phylogenetic trees were generated by comparing the translated amino acid sequence of individual sequence reads to members of the Picornaviridae family. The trees were created using the maximum parsimony method with 1,000 replicates. Bootstrap values over 700 are shown. CVA=Coxsakievirus A, CVB=Coxsackievirus B, BEV=Bovine Enterovirus, EV=Enterovirus, HRVA=Human Rhinovirus A, HRVB=Human Rhinovirus B, PEV=Porcine Enterovirus, PV=Poliovirus, SEVA=Simian Enterovirus A.



Supplemental Figure 2.S3: Phylogenetic analysis of anellovirus-like sequence reads. Phylogenetic trees were generated by comparing the translated amino acid sequence of individual sequence reads to Anelloviruses. The trees were created using the maximum parsimony method with 1,000 replicates. Bootstrap values over 700 are shown.



Supplemental Figure 2.S4: Phylogenetic analysis of *Caliciviridae*-like sequence reads. Phylogenetic trees were generated by comparing the translated amino acid sequence of individual sequence reads to the A) NS4 (3A-like) protein or B) NS7 (RNAP) protein of Caliciviruses. The trees were created using the maximum parsimony method with 1,000 replicates. Bootstrap values over 700 are shown.



Supplemental Figure 2.S5: Phylogenetic analysis of endonuclease-like sequence reads. Phylogenetic trees were generated by comparing the translated amino acid sequence of two individual sequence reads to endonuclease sequences derived from mimivirus, phage, and bacterial species representing some of the top scoring BLAST hits. The trees were created using the maximum parsimony method with 1,000 replicates. Bootstrap values over 700 are shown.

Chapter 3:
Complete genome sequence of a highly divergent astrovirus isolated from a child with acute diarrhea

This work is published in *Virology Journal* (2008), 5:117

Contributors: Stacy R. Finkbeiner[1], Carl D. Kirkwood[2], David Wang[1]

[1] Departments of Molecular Microbiology and Pathology & Immunology, Washington University School of Medicine, St. Louis, MO USA
[2] Enteric Virus Research Group, Murdoch Childrens Research Institute, Royal Children's Hospital, Victoria, Australia.

ABSTRACT

Astroviruses infect a variety of mammals and birds and are causative agents of diarrhea in humans and other animal hosts. We have previously described the identification of several sequence fragments with limited sequence identity to known astroviruses in a stool specimen obtained from a child with acute diarrhea, suggesting that a novel virus was present. In this study, the complete genome of this novel virus isolate was sequenced and analyzed. The overall genome organization of this virus paralleled that of known astroviruses, with 3 open reading frames identified. Phylogenetic analysis of the ORFs indicated that this virus is highly divergent from all previously described animal and human astroviruses. Molecular features that are highly conserved in human serotypes 1-8, such as a 3'NTR stem-loop structure and conserved nucleotide motifs present in the 5'NTR and ORF1b/2 junction, were either absent or only partially conserved in this novel virus. Based on the analyses described herein, we propose that this newly discovered virus represents a novel species in the family Astroviridae. It has tentatively been named Astrovirus MLB1.

BACKGROUND

Astroviruses are non-enveloped, single stranded, positive sense RNA viruses. Their genomes range from approximately 6 to 8 kb in length, are

polyadenylated, and have both 5' and 3' non-translated regions (NTR) (131). Their genomes have three open reading frames (ORFs) organized from 5' to 3' as follows: ORF 1a, which encodes a serine protease; ORF1b, which encodes the RNA dependent polymerase; and ORF 2, which encodes the structural proteins. A frameshift must occur during the translation of ORF1a in order for ORF1b to be translated. ORF 2 is translated from a sub-genomic RNA and produces a polyprotein which is cleaved by cellular proteases (131).

The family *Astroviridae* includes 8 closely related human serotypes as well as additional members that infect cattle, sheep, cats, dogs, deer, chickens, turkeys, and ducks (102). Although some of the animal astroviruses are known to cause hepatitis or nephritis (104), astroviruses typically cause diarrhea in their hosts. Human astrovirus infections most frequently cause watery diarrhea lasting 2-4 days, and less commonly vomiting, headache, fever, abdominal pains, and anorexia in children under the age of 2, the elderly, and immunocompromised individuals (104). The known human astroviruses account for up to ~10% of sporadic cases of non-bacterial diarrhea in children (84,85,105,132,133).

Diarrhea is the third leading infectious cause of death worldwide and is responsible for approximately 2 million deaths each year as well as (80) an estimated 1.4 billion non-fatal episodes (81,82). In children, rotaviruses, caliciviruses, adenoviruses and astroviruses are responsible for

the greatest proportion of cases (84-88). Most epidemiological studies fail to identify an etiologic agent in ~40% of diarrhea cases (89-93). Recently, we conducted viral metagenomic analysis of diarrhea samples using a mass sequencing approach with the explicit goal of identifying novel viruses that may be candidate causes of diarrhea. One of the stool samples we analyzed was collected in 1999 at the Royal Children's Hospital in Melbourne, Australia from a 3-yr old boy with acute diarrhea. Seven sequence reads were identified in this sample that shared ≤ 67% amino acid identity to known astrovirus proteins, suggesting that a novel astrovirus was present in the sample (134). In this paper, we report the full sequencing and characterization of the genome of this astrovirus, referred to hereafter as astrovirus MLB1 (AstV-MLB1).

RESULTS/DISCUSSION

Genome sequencing and analysis. In the previous metagenomic study (134), we identified seven sequence reads with limited identity to known astroviruses that could be assembled into two small contigs in a clinical stool sample. The contigs had 42-44%, and 59-61% amino acid identity to human astrovirus serine proteases and RNA-polymerases, respectively. In this study, the complete genome of the astrovirus present in the original stool specimen was sequenced to an average of >3X coverage (GenBank: FJ222451). The virus has been tentatively named

Astrovirus MLB1 (AstV-MLB1). Analysis of the genome showed that AstV-MLB1 has the same genomic organization as other astroviruses. Like other astroviruses, the AstV-MLB1 genome was predicted to encode three open reading frames (ORF1a, ORF1b, and ORF2) and contained both 5' and 3' non-translated regions (NTR), as well as a poly-A tail. The complete genome length of AstV-MLB1 was 6,171bp, excluding the poly-A tail, slightly shorter when compared to other astrovirus genomes which range in size between ~6,400 and 7,300bp (131). A comparison of AstV-MLB1 genomic elements with those of fully sequenced astroviruses is shown in Table 3.1.

The ORF 1a of astroviruses encodes a non-structural polyprotein which contains a serine-like protease motif. Pfam analysis revealed a region of ORF1a that has homology to a peptidase domain. In addition, alignment of AstV-MLB1 with other astroviruses revealed that AstV-MLB1 contains the amino acids of the catalytic triad (His, Asp, Ser) which are conserved in the 3C-like protease motif found in other viruses (data not shown) (135). The residues RTQ which have been suggested to be involved in substrate binding are conserved among the human astroviruses, but vary in other viruses which have the 3C-like motif (135). In AstV-MLB1, the predicted substrate binding residues (ATR) are identical to those found in *Ovine astrovirus* and not those of the human astroviruses (data not shown).

Table 3.1: Genome Comparison of MLB1 to other astroviruses						
Virus	**Genome (bp)**	**5' UTR (bp)**	**ORF1a**	**ORF1b**	**ORF2**	**3' UTR**
Chicken AstV-1	6,927	15	3,017	1,533	2,052	305
Turkey AstV-1	7,003	11	3,300	1,539	2,016	130
Turkey AstV-2	7,325	21	3,378	1,584	2,175	196
Mink AstV	6,610	26	2,648	1,620	2,328	108
Ovine AstV	6,440	45	2,580	1,572	2,289	59
Human AstV-1	6,813	85	2,763	1,560	2,361	80
Human AstV-2	6,828	82	2,763	1,560	2,392	82
Human AstV-4	6,723	84	2,763	1,548	2,316	81
Human AstV-5	6,762	83	2,763	1,548	2,352	86
Human AstV-8	6,759	80	2,766	1,557	2,349	85
AstV-MLB1	**6,171**	**14**	**2,364**	**1,536**	**2,271**	**58**

A second feature of astrovirus ORF1a is the presence of a bipartite nuclear localization signal (NLS) found in human, chicken, and ovine astroviruses, but not turkey astroviruses (136). A bipartite NLS is characterized as having two regions of basic amino acids separated by a 10 aa spacer. The protein alignment of ORF1a revealed that AstV-MLB1 has a sequence motif similar to the putative NLS of human astroviruses. This region of the genome has also been predicted to potentially encode for a viral genome-linked protein (VPg) (137). The high sequence similarity observed between AstV-MLB1 and other astroviruses in the motifs identified as essential for a putative VPg suggests that AstV-MLB1 may also encode a VPg (data not shown). While no experimental data exists supporting the prediction of the presence of a Vpg being encoded in any of the astrovirus genomes, we should note that we did encounter difficulty

in obtaining the 5' end of the MLB1 genome until treatment of the RNA with proteinase K prior to RNA extraction was added to the experimental protocol.

Finally, the 2,364nt sequence of AstV-MLB1 ORF1a is shorter than ORF1a sequences of other astroviruses, which range between ~2,500-3,300nt (Table 3.1). The shorter length of AstV-MLB1 ORF1a relative to the human astroviruses is largely attributable to two deletions totaling 57 amino acids located within a highly conserved motif near the carboxyl terminus of human astroviruses 1-8. This deletion falls within a 144 aa region that has been mapped as being an immunoreactive epitope in human astroviruses (138) and is located in the non-structural protein p38 (135). Recently, p38 has been reported to lead to apoptosis of the host cell which results in efficient virus replication (139) and particle release (140). However, it is unclear how the genome deletion identified in AstV-MLB1 might influence these activities.

Astrovirus ORF1b is classically generated by a -1 ribosomal frameshift induced by the presence of a heptameric `slippery sequence' (AAAAAAAC). (102). A conserved slippery sequence was identified near the end of ORF1a of Ast-MLB1 and FSFinder was used to determine if the downstream sequence was capable of forming a stem-loop structure, as found in other astoviruses (141). The predicted start position of ORF1b was then determined by selecting the first amino acid in frame with the

slippery sequence. The 1b open reading frame of astroviruses encodes an RNA-dependent RNA polymerase (RNAP). Pfam analysis revealed that AstV-MLB1 ORF1b contains the RNA-dependent RNA polymerase domain found in other positive strand RNA viruses, suggesting this ORF does in fact encode for an RNAP.

Astrovirus ORF2 encodes a large structural polyprotein that is cleaved by cellular proteases to generate the viral capsid proteins. Following the convention of human astroviruses (142,143) by choosing a start codon for ORF2 located two nucleotides upstream of the ORF 1b stop codon resulted in a predicted protein length of 756aa. Pfam analysis of the predicted protein encoded by ORF2 identifies an astrovirus capsid motif, thereby congruent with the paradigm of astrovirus genome organization in which ORF2 encodes the structural capsid proteins.

The AstV-MLB1 ORF2 protein sequence was divided into four subregions for more detailed analysis as described (144). Pair-wise comparisons of each region were conducted between the AstV-MLB1 sequence and the sequences of all astroviruses for which sequences were available. Consistent with previous reports, region I appeared to be the most conserved of the four regions and in each of the regions, AstV-MLB1 shared the most similarity to known human astroviruses. However, even in region I, AstV-MLB1 only exhibited 33-35% identity to known human astroviruses. In the less conserved regions II-IV, AstV-MLB1 shared only 5-

27% amino acid identity to the known human astroviruses. By contrast, the range of identities between human astrovirus serotypes 1-8 were, 43-75%, 16-66% and 28-77% for regions II, III and IV, respectively. Overall, ASTV-MLB1 maintained higher conservation in region I of ORF2 than in other regions, consistent with paradigms established by analysis of other astroviruses.

Non-coding features.

Multiple independent 5' RACE experiments were performed to determine the precise 5' end of the genome. Based on these experiments, the AstV-MLB1 5' NTR was determined to be 14nt long. This is similar in length to the ~10-20nt 5'NTRs of avian astroviruses (131), but much shorter than the 80-85 nt long 5'NTRs of the 8 human astrovirus serotypes (Table 3.1). Notably, the human astroviruses share a 20nt consensus sequence at the terminal 5' nucleotides of the genome which is not conserved in other astroviruses (data not shown). AstV-MLB1 contained 13 out of the 20 consensus nucleotides, including the most 5'CCAA motif within the this region (145) (Figure 3.1A). These data support the notion that the sequence we generated does contain the very 5' terminus of the genome.

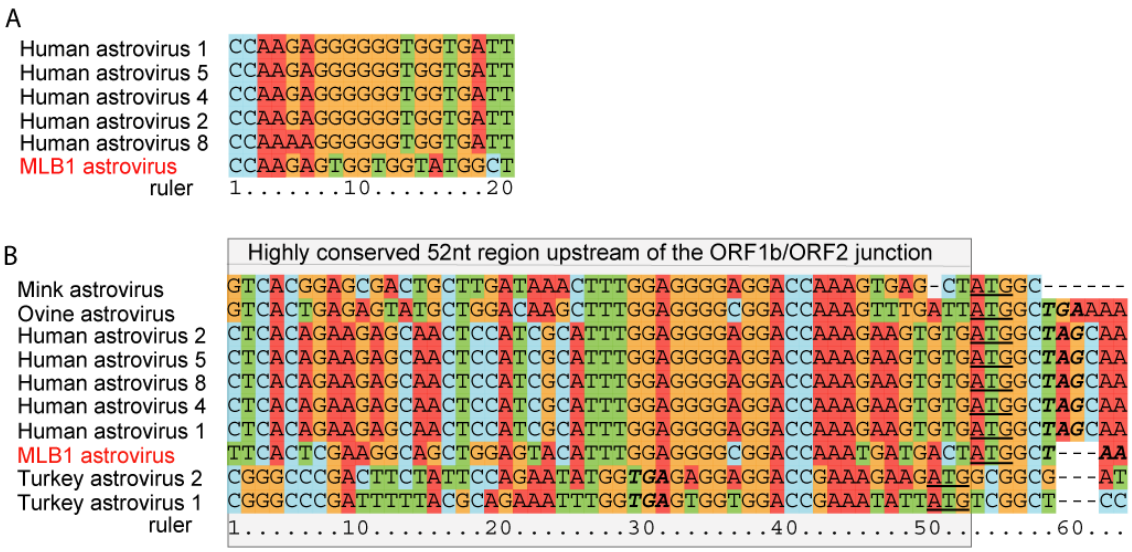


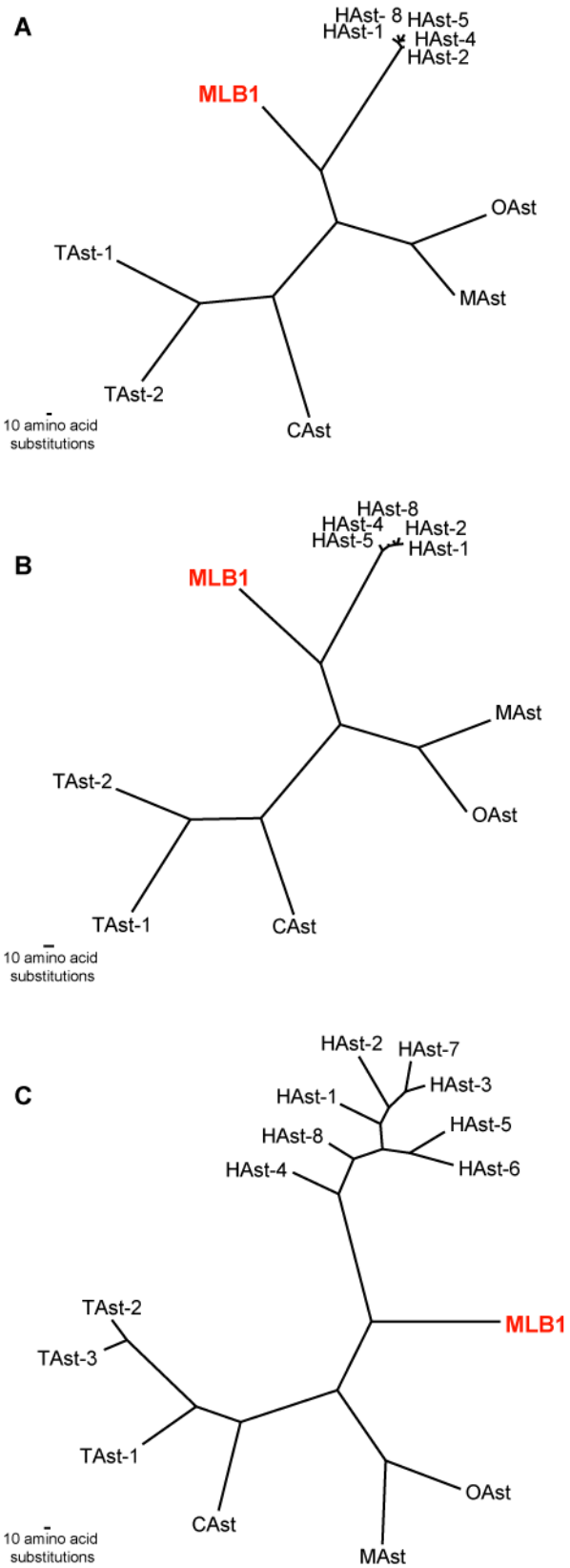
Figure 3.1: Multiple sequence alignments of putative astrovirus regulatory regions.
A.) Alignment of the 20 nucleotides at the very 5′ end of the Astrovirus MLB1 genome with those of fully sequenced astroviruses. MLB1 only shares 13 of the 20 conserved nucleotides present in human strains 1-8. B.) Alignment of the 52nt highly conserved nucleotide motif (shown in box) present immediately upstream of the ORF1b/ORF2 junction of Astrovirus MLB1 and other astroviruses. (Note: there is no overlap in the Turkey Astroviruses). MLB1 lacks the high degree of sequence identity seen between the human astroviruses. The start codon of ORF2 is shown underlined and the stop codon of ORF1b is shown italicized in bold for each virus.

Human astroviruses contain a 120nt region at the junction between ORF1b and ORF2 that is ~95-97% conserved between serotypes (146). The most highly conserved core 52nt region of this sequence is 99-100% identical among the human astrovirus serotypes. The exact role of this sequence is not known, but it is hypothesized to be a regulatory element of the sub-genomic RNA that encodes for ORF2. Alignment between AstV-MLB1 and other human astroviruses of the highly conserved 52nt at the ORF1b/ORF2 junction revealed that AstV-MLB1 possessed only 61.5% identity in this region (Figure 3.1B). By contrast, the known animal astroviruses share only 44-59.6% identity in this 52nt region with human

astroviruses as determined by pair-wise comparisons. Interestingly, AstV-MLB1 shares 71.2% identity in this region to *Ovine Astrovirus*.

All of the previously described astroviruses, with the exception of turkey astrovirus 2, have a conserved RNA secondary structure referred to as the stem-loop II-like motif (s2m) found at the 3′ end of the genome in the 3′ NTR (42). This motif is also present in some coronaviruses and equine rhinovirus serotype 2. Mutations within this motif are generally accompanied by compensatory mutations that restore base pairing (42). The conservation of such a sequence motif across multiple viral families suggests that it may play a broad role in the biology of positive stranded RNA viruses (42). The exact function of this stem loop is not known, but it is hypothesized to interact with viral and cellular proteins needed for RNA replication. Nucleotide alignment of the 150 nucleotides at the 3′ terminus of the AstV-MLB1 genome and other viruses known to contain the stem-loop motif suggested that AstV-MLB1 does not have this conserved nucleotide motif (data not shown). Furthermore, it also has the shortest 3′NTR reported to date for an astrovirus. (Table 3.1) (131).

Phylogenetic analysis. Multiple sequence alignments of the three astrovirus open reading frames were performed and bootstrapped



maximum parsimony trees were generated (Figure 3.2). The trees confirmed initial assessments that AstV-MLB1 is a novel astrovirus(134). The trees for ORFs 1a and 1b (Figure 3.2A, B) both indicated that AstV-MLB1 is most closely related to the human astroviruses, although it is highly divergent from them. AstV-MLB1 ORF1a only has 9-28% amino acid identity to other astrovirus ORF1a proteins and the pairwise sequence alignments of ORF1b revealed 35-54% amino acid identity between ORF1b proteins of AstV-MLB1 and other astroviruses (Table 3.2).

Figure 3.2: Phylogenetic analysis of AstV-MLB1 open reading frames.
Phylogenetic trees are based on amino acid sequences and were generated using the maximum parsimony method with 1,000 bootstrap replicates. Significant bootstrap values are shown. (A) ORF1a; (B) ORF1b; (C) ORF2. HAstV = Human astrovirus; CAstV = Chicken astrovirus; MAstV = Mink astrovirus; TAstV = Turkey astrovirus; OAstV = Ovine astrovirus.

Table 3.2: Comparison of astrovirus proteins to predicted AstV-MLB1 proteins.															
Gene	**Est. Size (aa)**	**% Amino Acid Identity to:**													
		HAstV-1	HAstV-2	HAstV-3	HAstV-4	HAstV-5	HAstV-6	HAstV-7	HAstV-8	TAstV-1	TAstV-2	TAstV-3	ChAstV-1	OAstV	MAstV
ORF 1a	787	28	28	NA	29	29	NA	NA	29	9	9	NA	10	22	24
ORF 1b	511	54	54	NA	54	54	NA	NA	54	36	35	NA	36	47	44
ORF 2	756	24	24	24	23	23	24	24	24	15	16	16	11	18	19

The maximum parsimony tree for ORF2 (Figure 3.2C) shows that there is greater divergence among all of the sequences for ORF2, as is to be expected of the capsid region. However it is still evident that AstV-MLB1 is quite divergent from any of the known human astroviruses. Based on the predicted 756aa protein of ORF2, AstV-MLB1 has only 11-24% amino acid identity to other astrovirus capsid precursor proteins (Table 3.2).

Origin of virus. At this point, the origin of AstV-MLB1 is unclear. AstV-MLB1 may be a bona fide human virus capable of infecting and replicating within the human gastrointestinal tract that had evaded detection until now. Alternately, it may be a passenger virus present simply as a result of dietary ingestion, as has been described previously for plant viruses detected in human stool (78). Of course, viruses derived from dietary intake that appear to cause human disease, such as Aichi virus, have been described previously (147,148). Another possibility is that this virus may represent zoonotic transmission from some other animal species

that is the true host for Astrovirus MLB1. Traditionally it has been thought that astroviruses have a strict species tropism. However, recent evidence has emerged that suggests that interspecies transmission does occur. For example, chicken astrovirus antibodies have been detected in turkeys (149) and an astrovirus was isolated from humans whose capsid sequence most closely resembled that of feline astrovirus(131). Because of the uncertainty as to the identity of the true host species and the host range for this virus, we have tentatively named this novel virus Astrovirus MLB1 (AstV-MLB1). Efforts to define whether AstV-MLB1 is a novel human pathogen are underway.

CONCLUSION

Complete sequencing and genome analysis of Astrovirus MLB1 revealed that the virus has three open reading frames sharing the same organization as other astroviruses. Phylogenetic analysis of the open reading frames clearly demonstrated that AstV-MLB1 is highly divergent from any of the known astroviruses. Furthermore, AstV-MLB1 lacks the conservation seen between human astroviruses 1-8 in the non-translated regions of the genome such as the 5' and 3' NTR and the ORF1b/2 junction. The aggregate analysis of the non-coding features and ORFs as well as the phylogentic analysis clearly indicates that AstV-MLB1 is highly divergent from all previously described astroviruses.

The divergence of AstV-MLB1 from known astroviruses in the non-translated regions of the genome is particularly interesting because these regions are nucleotide motifs that are thought to play regulatory roles in viral replication. This suggests that AstV-MLB1 may behave very differently from the known astroviruses and that additional studies on the regulation of AstV-MLB1 transcription and replication may broaden our understanding of astrovirus paradigms.

Astroviruses are associated with diarrhea predominantly in young children and immunocompromised individuals. The discovery of AstV-MLB1 in a liver transplant patient fits well with the known clinical parameters of astrovirus infection. We previously reported that the only other virus detected in this stool was a TT virus (134), which is thought to be non-pathogenic (150). It is therefore tempting to speculate that AstV-MLB1 is the pathogenic agent that caused this case of diarrhea. However, whether AstV-MLB1 is a bona fide human virus capable of causing diarrhea will have to be established by further experimentation and epidemiological surveys.

Materials and Methods

Specimen. A stool sample was collected from a 3 year old boy admitted to the Royal Children's Hospital with acute diarrhea in 1999. The child had

previously undergone a liver transplant one year prior to this episode of diarrhea, however the immunological status was unknown.

RNA extraction. RNA was isolated from the primary stool filtrate using RNA-Bee (Tel-Test, Inc.) according to manufacturer's instructions. In some cases, the stool filtrate was treated with 2.5 mg\ml proteinase K (Sigma) for 30 min prior to RNA extraction.

Genome amplification and sequencing. The astrovirus sequence reads previously detected in the primary stool filtrate (134) (GenBank accessions: ET065575, ET065576, ET065577, ET065579, ET065580, ET065581, ET065582) were assembled into two contigs, and the nucleic acid between the contigs was obtained by RT-PCR. For reverse transcription reactions, cDNA was generated with MonsterScript RT at 65ºC and amplified with Taq (Invitrogen). Subsequent 5' and 3' RACE reactions were done to obtain the entire genome. To generate high quality sequence coverage, 7 pairs of specific primers that spanned the complete genome in overlapping ~1kb fragments were used in RT-PCR reactions and then cloned and sequenced using standard Sanger sequencing chemistry. All amplicons were cloned into pCR4.0 (Invitrogen). These 7 primer pairs were used to confirm the sequence of the viral genome from both the primary stool sample and the passage 2

tissue culture sample. The complete genome sequence of AstV-MLB1 has been deposited in (GenBank: FJ222451).

ORF prediction and annotation. Open reading frames 1a and 2 were predicted for AstV-MLB1 using the NCBI ORF Finder program. ORF1b was predicted based on the frameshift paradigm that occurs in other astroviruses by identifying a heptameric slippery sequence (151). Conserved motifs were identified using Pfam (152).

Pair-wise alignments. Bioedit was used to determine the percent identity between sequences as determined by pair-wise alignments.

Phylogenetic analysis. ClustalX (1.83) was used to carry out multiple sequence alignments of the protein sequences associated with all three of the open reading frames of representative astrovirus types. Maximum parsimony trees were generated using PAUP with 1,000 bootstrap replicates (118).

Available nucleotide or protein sequences of the following astroviruses were obtained: Human Astrovirus 1 (GenBank: NC_001943); Human Astrovirus 2 (GenBank: L13745); Human Astrovirus 3 (GenBank: AAD17224); Human Astrovirus 4 (GenBank: DQ070852); Human Astrovirus 5 (GenBank: DQ028633); Human Astrovirus 6 (EMBL: CAA86616); Human Astrovirus 7

(Gen Bank: AAK31913); Human Astrovirus 8 (GenBank: AF260508); Turkey Astrovirus 1 (GenBank: Y15936); Turkey Astrovirus 2 (GenBank: NC_005790); Turkey Astrovirus 3 (GenBank: AY769616); Chicken Astrovirus (GenBank: NC_003790); Ovine Astrovirus (GenBank: NC_002469); and Mink Astrovirus (GenBank: NC_004579).

Acknowledgements

This work was funded in part by an NHMRC RD Wright Research Fellowship (ID 334364, CK), and by the Food Safety Research Response Network, a Coordinated Agricultural Project, funded through the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service, grant number ##2005-35212-15287.

Chapter 4:
Detection of newly described Astrovirus MLB1 in pediatric stools

This work is published in Emerging Infectious Diseaseas (2009), 15(3): 441-444

Contributors: Stacy R. Finkbeiner[1], Binh-Minh Le[1], Lori R. Holtz[1], Gregory A. Storch[1], David Wang[1]

[1] Washington University School of Medicine, St. Louis, MO USA

ABSTRACT

The prevalence of the recently identified astrovirus MLB1 in a pediatric diarrhea cohort in St. Louis, USA was defined by RT-PCR. Of 254 stool specimens collected in 2008, 4 were positive for astrovirus MLB1. These results demonstrate that astrovirus MLB1 is currently circulating in North America.

BACKGROUND

Astroviruses infect a variety of hosts including humans, turkeys, chicken, cattle, sheep, dogs, cats, deer, ducks and bats (102,103). There are 8 known human serotypes which are genetically very closely related. Astroviruses typically cause diarrhea in their hosts and in humans symptoms normally last 2-4 days (104). Children under the age of 2, elderly people, or otherwise immunocompromised individuals are most commonly affected (104). Epidemiological studies suggest Human astroviruses 1-8 are responsible for up to ~10% of cases of acute, non-bacterial diarrhea in children (84,85,105,132,133).

Recently, a highly divergent astrovirus, referred to as astrovirus MLB1 (AstV-MLB1), was identified in the stool of a three year old boy in Australia (134). The entire genome of this novel virus was subsequently sequenced and characterized (153) . To date, there have been no published reports describing the presence of AstV-MLB1 outside of the

index case. In this study, the prevalence of this novel virus was determined by RT-PCR screening of pediatric stool samples collected at the St. Louis Children's Hospital in St. Louis, MO, USA.

RESULTS/DISCUSSION

Pediatric stool specimens sent to the clinical microbiology lab for bacterial culture at the Saint Louis Children's Hospital were analyzed for the presence of AstV-MLB1. This study was approved by the Human Research Protection Office of Washington University in St. Louis. Samples were collected January-May of 2008. Stools were diluted in PBS at a 1:6 ratio (w/v) and total nucleic acid (TNA) was extracted from 200µL of each stool suspension using the MagNAPure LC Automated Nucleic Acid Extraction System (Roche).

Previously described astrovirus primers Mon269 and Mon270 (154) have frequently been used for the detection of human astrovirus serotypes 1-8 in clinical stool specimens. However, the extensive divergence of AstV-MLB1 to the known human astroviruses rendered these primers unable to amplify AstV-MLB1 (data not shown). Since it is possible that AstV-MLB1 represents a new grouping of astroviruses that could include multiple subtypes, we designed primers to conserved regions of the AstV-MLB1 genome in order to maximize the likelihood of detection of

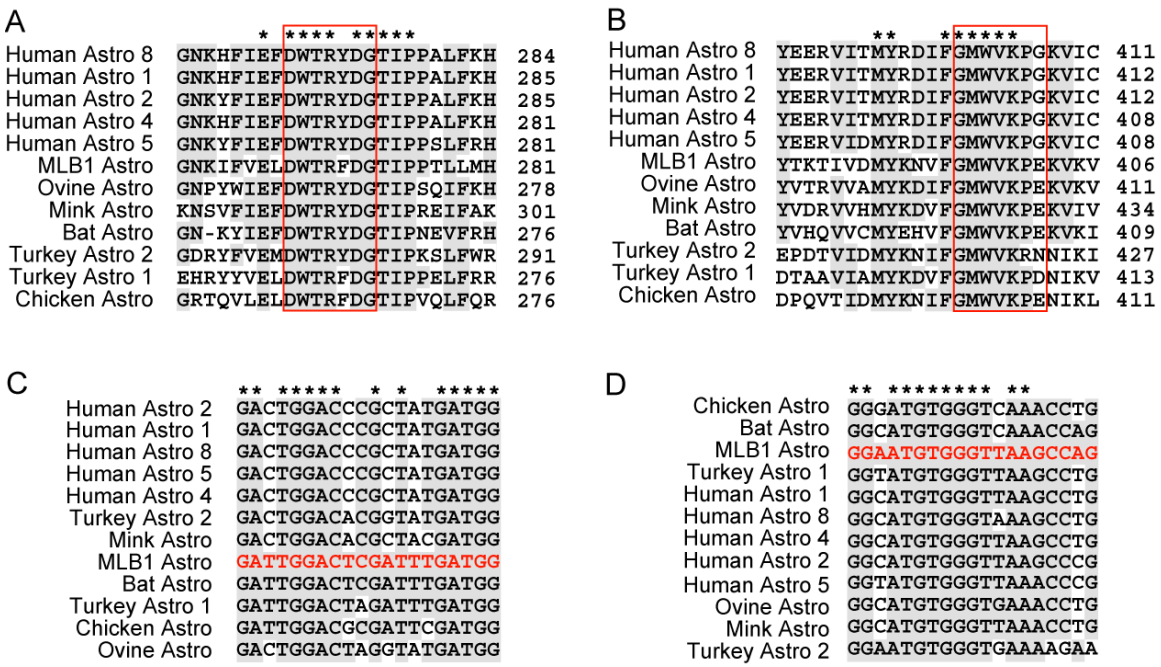


Figure 4.1: Astrovirus ORF1b alignments for design of pan-astrovirus primers
Astrovirus RNA polymerase sequences (ORF1b) were aligned at the amino acid level in order to define the conserved regions used for the design of primers SF0073 (A) and SF0076 (B). The numbers to the right of the sequences indicate the position of the last amino acid within each ORF1b sequence. The amino acids outlined in the red boxes represent the specific regions that were reverse translated into the corresponding nucleic acid sequences used for the design of SF0073 (C) and SF0076 (D). The red sequences shown in the nucleotide alignments are the actual primer sequences for SF0073 (C) and SF0076 (D).

any AstV-MLB1 variant viruses, or even other novel astroviruses. Conserved regions were identified by performing multiple sequence alignments of AstV-MLB1 amino acid sequences to all fully sequenced astrovirus genomes (Figure 4.1A, 4.1B). The corresponding nucleotide sequences for these regions were then aligned in order to define the most highly conserved regions (Figure 4.1C, 4.1D). Two regions within ORF 1b were identified that yielded primers SF0073 (5'GATTGGACTCGATTTGATGG) and SF0076 (5'CTGGCTTAACCCACATTCC) that are predicted to generate a

~409 bp product. Control experiments validated that this primer pair could detect AstV-MLB1 as well as *Human astrovirus 1* (Figure 4.2). Given that some of the canonical human astroviruses are identical in the primer binding sites, this data suggests that at least some of the canonical human astroviruses can be detected by the primer pair SF0073/SF0076. In theory, under appropriate experimental conditions, these primers may also be able to detect all other known human and animal astroviruses, although that remains to be experimentally tested. These primers were used with the QIAGEN One-Step RT-PCR kit using the following cycling conditions: 30 min RT step, 94ºC hold for 10 min, followed by 40 cycles of 94ºC for 30s, 52ºC for 30s, and 72ºC for 50s.

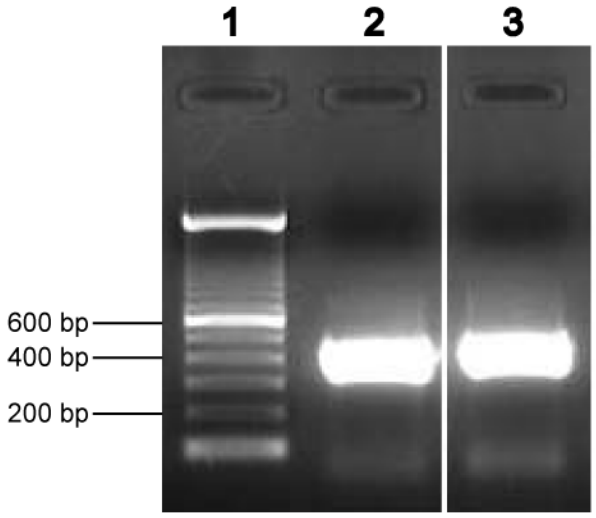


Figure 4.2: Validation of screening primers SF0073 and SF0076
Primers SF0073 and SF0076 were tested on stool filtrate made from the original AstV-MLB1 positive stool (Lane 2) as well as a *Human astrovirus 1* positive stool specimen (Lane 3) using the QIAGEN One-Step RT-PCR kit as described in the text. The products were visualized on a 1.2% agarose gel. The expected size of the RT-PCR product generated with these primers is ~400bp. Lane 1 shows the Invitrogen 100bp DNA ladder for a size comparison.

Samples that tested positive with primers SF0073 and SF0076 were then tested in a second round of screening with two different primer sets in parallel to determine if the samples contained canonical human

astrovirus serotypes 1-8 or AstV-MLB1. The previously reported Mon269 (5'CAACTCAGGAAACAGGGTGT) and Mon270 (5'TCAGATGCATTGTCATTGGT) primers, which generate a 449 bp amplicon, were used to detect canonical human astroviruses (154). Another set of primers, SF0053 (5'CTGTAGCTCGTGTTAGTCTTAACA) and SF0061 (5'GTTCATTGGCACCATCAGAAC), was designed to exclusively detect AstV-MLB1 and produce a 402bp PCR product. These primers target a region of the capsid gene. The second round of screening with both sets of primer pairs was performed as described above with the exception that an annealing temperature of 56ºC was used.

Of 254 stool specimens screened, 9 (3.5%) tested positive in the initial round of screening using the newly designed pan-astrovirus primers, SF0073 and SF0076. Secondary screening demonstrated that 5 (2% of all samples) were canonical human astroviruses. This is likely to be an underestimate of the astrovirus serotype 1-8 prevalence in the cohort since the initial screening primers were biased towards the detection of AstV-MLB1. The remaining 4 (1.6% of all samples) were positive for AstV-MLB1 using primers SF0053 and SF0061. For each of the 4 samples positive for AstV-MLB1, two additional fragments were generated by RT-PCR for

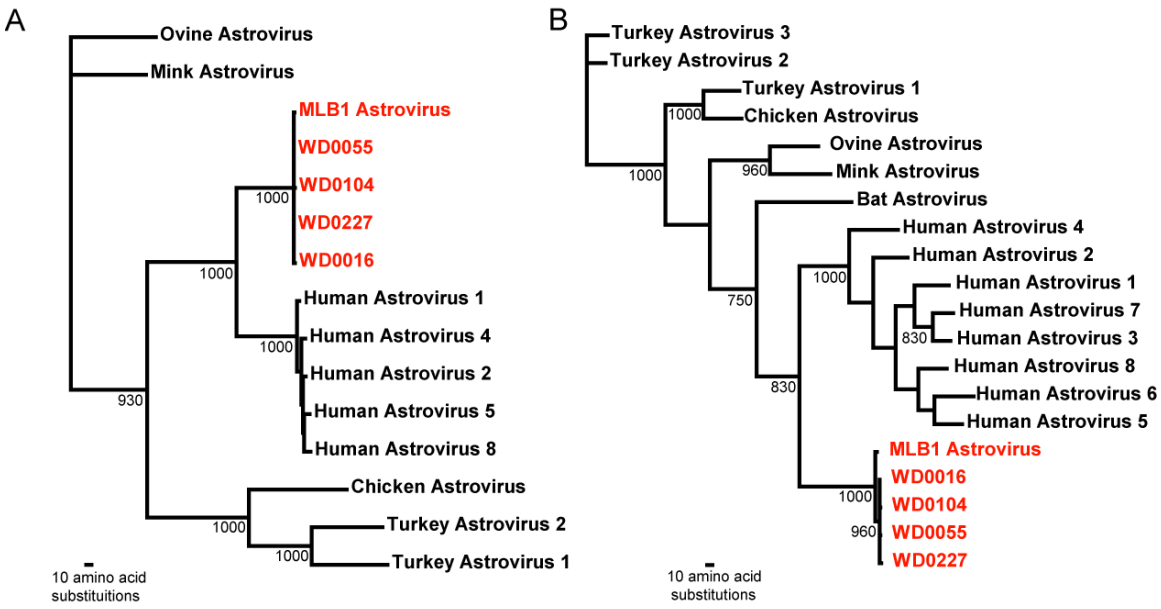


Figure 4.3: Phylogenetic Analysis of AstV-MLB1 isolates
A region of the serine protease (A) and the capsid (B) of each virus detected by the AstV-MLB1 specific primers was amplified and sequenced. Multiple sequence alignments were then generated with these sequences and the corresponding regions of known astroviruses using ClustalX. PAUP was used to generate phylogenetic trees and bootstrap values (>700) from 1,000 replicates are shown.

phylogenetic analysis. A 1228 bp fragment of ORF1a, which encodes the serine protease, and a 920 bp fragment of ORF2, which encodes the capsid proteins, were amplified using AstV-MLB1 specific primers from each of the 4 samples designated WD0016, WD0055, WD0104, WD0227. The primers used for the ORF1a fragment are SF0080 (5'-AAGGATAGTGCTGGTAAAGTAGTTCAGA-3') and SF0094 (5'-CAAGAGCCTTATCAACAACGTA-3') and the primers used for the ORF2 fragment are SF0064 (5'-GTAAGCATGGTTCTTGTGGAC-3') and SF0098 (5'-TGCATACATTTATGCTGGAAGA-3'). The ORF1a fragments (Genbank #'s: FJ227120-FJ227123) from these samples all shared ~92%

Table 4.1: Similarity of fully sequenced WD0016 genome to MLB1			
	ORF1a (Serine Protease)	ORF1b (RNA Polymerase)	ORF2 (Capsid)
Nucleotide identity between WD0016 and MLB1	92.60%	93.90%	91.90%

nucleotide identity to the reference astrovirus MLB1 sequence (Genbank #: FJ222451) and 99% amino acid identity, indicating that most mutations were synonymous. The ORF2 fragments (Genbank #'s: FJ227124-FJ227127) shared ~91-92% nucleotide identity and 95-96% amino acid identity to the reference astrovirus MLB1 sequence. The 4 positive St. Louis samples shared ~99% nucleotide identity to each other. The ORF1a and ORF2 sequences were aligned to other astroviruses for which full genome sequences were available using ClustalX (1.83), and then maximum parsimony trees were generated using PAUP with 1,000 bootstrap replicates (118) (Figure 4.3). The entire genome of one of the isolates, WD0016 (Genbank #: FJ402983), was sequenced and was determined to have 92.6% identity overall to that of AstV-MLB1 based on a pairwise nucleotide alignment. Table 4.1 shows the identity of WD0016 to the original MLB1 isolate for each open reading frame of the genome.

Clinical and demographic information of patients with AstV-MLB1 positive stools is shown in Table 4.2. The patients with AstV-MLB1 positive stools were between the ages of ~4mo-4yrs. All patients had symptoms of

Table 4.2: Clinical and demographic information of patients with AstV-MLB1 positive stools				
	WD0016	**WD0055**	**WD0104**	**WD0227**
Age (months)	15	17	4	43
Gender	Female	Female	Male	Male
Diarrhea	No*	Yes	Yes	Yes
Other Symptoms	abdominal pain	vomiting, fever	fever, seizures, respiratory distress	fever
Hospitalization	Yes	No	Yes	Yes
Bacterial Cultures†	Negative	Negative	Negative	Positive for *E. coli* 0157:H7
* Patient had diarrhea two days prior to stool collection, but not at time of collection				
† Tests were conducted for *E. coli*, *Campylobacter*, *Shigella*, *Salmonella*, and *Yersinia*				

diarrhea at the time of stool collection, except for patient WD0016 who reported having diarrhea 2 days prior to the collection of the stool specimen. All specimens were tested for the presence of *E. coli, Campylobacter, Yersinia, Shigella, and Salmonella* by standard bacterial culture. WD0227 tested positive for *E. coli* 0157:H7, while the other samples were negative for all bacterial cultures. A pan-viral microarray, the

ViroChip (20), was used to examine whether there were other viruses present in the stool of three (WD0055, WD0104, and WD0227) of the four AstV-MLB1 positive samples for which there was enough material left for analysis. WD0055 and WD0104 were negative by array, but WD0227 was positive for rotavirus as determined by the ViroChip.

CONCLUSIONS

The newly identified AstV-MLB1 virus was discovered in a stool specimen collected in Melbourne, Australia in 1999. In this study, we describe the detection of AstV-MLB1 in a cohort from St. Louis, USA collected in 2008. This observation provides the first evidence that AstV-MLB1 is present outside of Australia, and suggests that AstV-MLB1 is likely to be globally widespread. In addition, these data demonstrate that AstV-MLB1 is currently circulating in the human population. The observed sequence divergence of ~8% at the nucleotide level between the reference AstV-MLB1 genome and the viruses detected in this study suggests that there may be significant sequence heterogeneity within the AstV-MLB1 group of viruses. It is possible that multiple serotypes or subtypes of AstV-MLB1 exist, as is the case with the canonical human astroviruses. More extensive screening of stool samples with PCR primers targeted toward detection of AstV-MLB1 such as those described in this paper may provide insight into the true diversity and prevalence of AstV-MLB1-like viruses. Finally, a critical direction for future investigation is determining whether AstV-MLB1, like the canonical astrovirus serotypes 1-8, is a causal agent of human diarrhea, and if so, what is the disease burden associated with this virus. To begin addressing this question, further epidemiologic studies, including both case-control prevalence studies

and seroprevalence assays, and efforts to fulfill Koch's postulates should be pursued.

ACKNOWLEDGEMENTS

This work was supported in part by National Institutes of Health grant U54 AI057160 to the Midwest Regional Center of Excellence for Biodefense and Emerging Infectious Diseases Research (MRCE).

Chapter 5:
Identification of a novel astrovirus (Astrovirus VA1) associated with an outbreak of acute gastroenteritis

This work has been submitted to *PLoS Pathogens* for review

Contributors: Stacy R. Finkbeiner[1†], Yan Li[2†], Susan Ruone[2], Christina Conrardy[2], Nicole Gregoricus[2], Denise Toney[3], Herbert W. Virgin[1], Larry J. Anderson[2], Jan Vinjé[2*], David Wang[1*], Suxiang Tong[2*]

[1]Departments of Molecular Microbiology and Pathology and Immunology, Washington University School of Medicine, St. Louis, Missouri 63110, USA
[2]Division of Viral Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30333 USA
[3]Commonwealth of Virginia Division of Consolidated Laboratory Services Richmond, VA 23219 USA

ABSTRACT

Viral gastroenteritis is one of the most common human illnesses worldwide and is a major cause of morbidity and mortality in infants and young children. Despite the availability of improved molecular diagnostics to detect known viral agents, the etiology of a large proportion of diarrheal cases is unknown. The advent of metagenomic sequencing approaches has revolutionized the process of pathogen discovery. In this study, parallel efforts applying both random Sanger sequencing and high throughput pyrosequencing were used to analyze fecal specimens obtained from an outbreak of acute gastroenteritis in a child care center. The specimens tested negative for known bacterial, parasitic and viral enteric pathogens by conventional assays. Sequences consistent with astroviruses were identified by both techniques, 72 sequence reads by the low through-put Sanger sequencing and 1339 sequencing reads by high through-put pyrosequencing. Assembly of these reads into contigs and subsequent RT-PCR and RACE experiments yielded a complete genome of 6,586 nucleotides. Phylogenetic analysis of the three predicted open reading frames from this newly identified astrovirus, tentatively named Astrovirus VA1 (AstV-VA1), demonstrated that AstV-VA1 was highly divergent from all previously described human and animal astroviruses and most closely related to mink and ovine astroviruses. Using AstV-VA1 specific RT-PCR assays, 3 of 5 fecal specimens from symptomatic

individuals in this outbreak tested positive, raising the possibility that AstV-VA1 may be a causal agent of acute gastroenteritis.

INTRODUCTION

Astroviruses are known to infect a variety of avian and mammalian species (102,103) and typically cause diarrhea (104). They are thought to be host specific with little evidence for cross-species transmission (131). In humans, 8 serotypes of astroviruses have been described (131). Clinical symptoms usually last 2-4 days and consist of watery diarrhea and, less commonly, vomiting, headache, fever, abdominal pains, and anorexia (104).

Astroviruses consist of a family of small, single-stranded, positive-sense RNA viruses. Their genomes are organized into three open reading frames denoted ORFs 1a, 1b, and 2, which encode a serine protease, RNA-dependent RNA polymerase (RdRP), and a capsid precursor protein, respectively (131). At both the 5' and 3' ends, non-translated regions (NTR) flank the 6.1-7.3kb sized genomes (131,153). Two characteristic features of astroviruses are the dependency on a ribosomal frameshift for the translation of ORF1b and the generation of a sub-genomic RNA from which ORF2 is translated (131).

Human astroviruses have been associated with up to ~10% of sporadic cases of viral diarrhea in children (84,85,105,132,133) and with

0.5-15% of outbreaks (94,95,155). Significantly, in some reports the etiologies of 12-41% of the outbreaks remain undetermined even after extensive testing, suggesting that there is a diagnostic gap (94,95). Similarly, on average, approximately 40% of the cases of sporadic diarrhea are unexplained (89-93). In previous efforts to identify novel candidate pathogens present in sporadic cases of diarrhea in humans, we used high throughput sequencing to identify a novel astrovirus (Astrovirus MLB1) (134,153) and a novel picornavirus (Cosavirus E1) (156). The role of these viruses in causing diarrhea remains unclear. In this paper, we applied mass sequencing to analyze specimens obtained from an unexplained outbreak of gastroenteritis. We report the identification and complete genome sequencing of yet another novel astrovirus, referred to as Astrovirus VA1 (AstV-VA1), associated with a gastroenteritis outbreak at a child care center.

RESULTS

Genome sequencing and analysis

Five fecal specimens (labeled A, B, C, D and E) were collected from a gastroenteritis outbreak at a child care center in Virginia (Table 5.1). Following high throughput pyrosequencing of RNA and DNA extracted from samples A, B, C and D (average of 12,730 reads per sample), we found 313 unique high quality sequence reads in sample B and 1,017 unique high quality reads in sample C most closely related to astroviruses.

Table 5.1: Epidemiologic data of the 5 specimens from a child care center outbreak of acute gastroenteritis.

Sample ID	Sex	Age	Onset Date	Sample Date	Symptoms
A	M	2 years	8/19/08	8/19/08	Diarrhea, vomiting
B	F	36 years	8/26/08	8/28/08	Diarrhea, vomiting
C	M	6 months	8/25/08	8/25/08	Diarrhea
D	M	19 months	8/5/08	8/26/08	Diarrhea
E	Unknown	20 months	8/5/08	8/27/08	Diarrhea

A 6,376 nucleotide (nt) contig was assembled from the astrovirus-like sequences detected in sample B and 4 contigs totaling 6,026 nucleotides were assembled from sample C. The translated contigs had only limited sequence similarity (37-71% aa identity) to proteins from mink and ovine astroviruses, suggesting the presence of a potentially novel astrovirus in these samples. Because the nucleotide sequences obtained in samples B and C were nearly identical, the five original contigs were assembled to generate a larger contig of 6,581 nucleotides in length.

Independently, four of the five fecal samples (stool samples A, B, C and E) were analyzed by Sanger sequencing. 3 out of 96 clones from sample B and 69 out of 152 clones from sample C contained sequence signatures that were most closely related to previously known astroviruses by Blastn similarity searches. Sequencing of 100 clones each from samples A and E yielded no clones with detectable similarity to astroviruses. The

sequences of the 69 clones from sample C were assembled into 4 contigs. Primers were then designed to generate a series of eight overlapping RT-PCR amplicons with an average size of ~900 bp that yielded a genomic sequence of 6,537 nt. In order to define the 5' end of the genome, three independent 5'RACE reactions were performed and a total of 23 clones from these reactions were sequenced. All clones extended the genome by 49 nt and yielded the identical 5'end sequence, suggesting that the genome was complete with a total length of 6,586 nt, excluding the poly-A tail. Comparison of the genome sequences generated by the two sequencing methods yielded nearly identical sequences, with the exception of 5 missing nucleotides at the 5' end of the contig generated by pyrosequencing and 3 nucleotide substitution differences. These were resolved by direct PCR sequencing to generate the final, corrected sequence. This virus has been provisionally named Astrovirus VA1 (AstV-VA1).

The genome of AstV-VA1 had three predicted open reading frames as well as non-translated regions (NTRs) at both the 5' and 3' ends of the genome. Several conserved protein motifs were identified including a serine protease in ORF1a, an RNA dependent RNA polymerase in ORF1b, and capsid protein in ORF 2. ORFs 1a and 2 were predicted by the NCBI ORF Finder program; however the full coding region for ORF1b was not predicted by the program because translation of ORF1b is dependent on

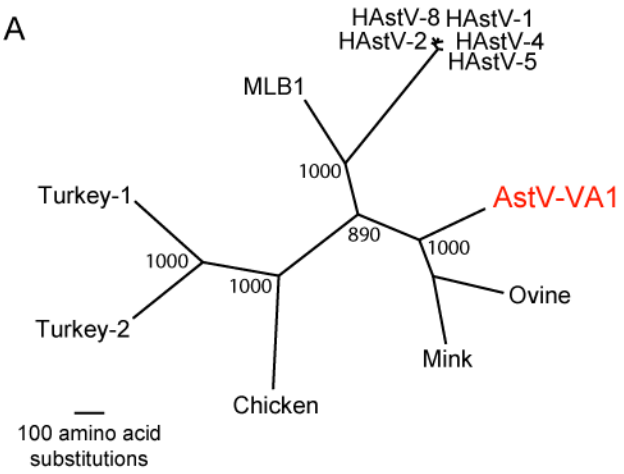
Table 5.2: Genome Comparison of MLB1 to other astroviruses						
Virus	**Genome (bp)**	**5' UTR (bp)**	**ORF1a**	**ORF1b**	**ORF2**	**3' UTR**
Chicken AstV-1	6,927	15	3,017	1,533	2,052	305
Turkey AstV-1	7,003	11	3,300	1,539	2,016	130
Turkey AstV-2	7,325	21	3,378	1,584	2,175	196
Mink AstV	6,610	26	2,648	1,620	2,328	108
Ovine AstV	6,440	45	2,580	1,572	2,289	59
Human AstV-1	6,813	85	2,763	1,560	2,361	80
Human AstV-2	6,828	82	2,763	1,560	2,392	82
Human AstV-4	6,723	84	2,763	1,548	2,316	81
Human AstV-5	6,762	83	2,763	1,548	2,352	86
Human AstV-8	6,759	80	2,766	1,557	2,349	85
AstV-MLB1	6,171	14	2,364	1,536	2,271	58
AstV-VA1	**6,586**	**38**	**2,661**	**1,575**	**2,277**	**98**

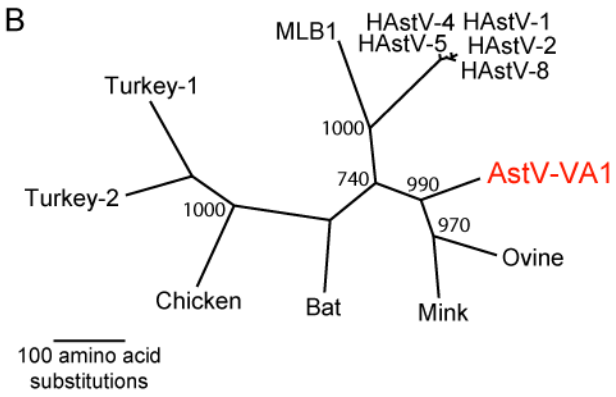
a -1 ribosomal frameshift that occurs during translation (151). This frameshift is thought to be mediated by the presence of a heptameric 'slippery sequence' (AAAAAAAC) near the end of ORF1a (151), which was also conserved in the AstV-VA1 sequence, suggesting that this new virus follows the same paradigm. The sequence AUUUGGAGNGGNGGACCNAAN$_{5-8}$*AUG*NC located upstream of ORF2, which has been proposed as the promoter for subgenomic RNA synthesis in all previously known astroviruses (131), is also present in AstV-VA1 with only 2 nt differences. The predicted size for each of the open reading frames is 2,661 nt, 1,575 nt, and 2,277 nt for ORFs 1a, 1b, and 2, respectively. These sizes are similar to the ORF sizes of mink and ovine astroviruses (Table 5.2).

The 5′ non-translated region (NTR) of AstV-VA1 is 38 nt in length, which is between the lengths of the 5′ NTRs of mink astrovirus (26 nt) and ovine astrovirus (45 nt). The 3′ NTR is 98 nt in length, which again is intermediate between the length of the NTRs of ovine astrovirus (59 nt) and mink astrovirus (108 nt). The 3′ NTR of nearly all astroviruses contains a highly conserved RNA secondary structure called the stem-loop II-like motif (s2m), which has also been identified in several coronaviruses and in equine rhinovirus 2 (42,157). An alignment of the 150 nt just upstream of the poly-A tail of AstV-VA1 along with the 3′ terminal sequences of other astroviruses known to contain the s2m motif indicated that AstV-VA1 contains the highly conserved ~33 nucleotide core of the s2m motif, with 100% identity to other astroviruses in this region. The exact role of this motif is not understood; however its presence in multiple viral families suggests it may play an important role in the replication of these viruses.

Phylogenetic analysis

Multiple sequence alignments were independently carried out for each of the three predicted ORFs. Maximum parsimony trees confirmed that AstV-VA1 was highly divergent from but most closely related to mink and ovine astrovirus in all three ORFs (Figure 5.1). Furthermore, the greatest sequence identity between AstV-VA1 and mink and ovine astroviruses is in ORF1b with 61% amino acid identity to mink astrovirus and 62% to ovine





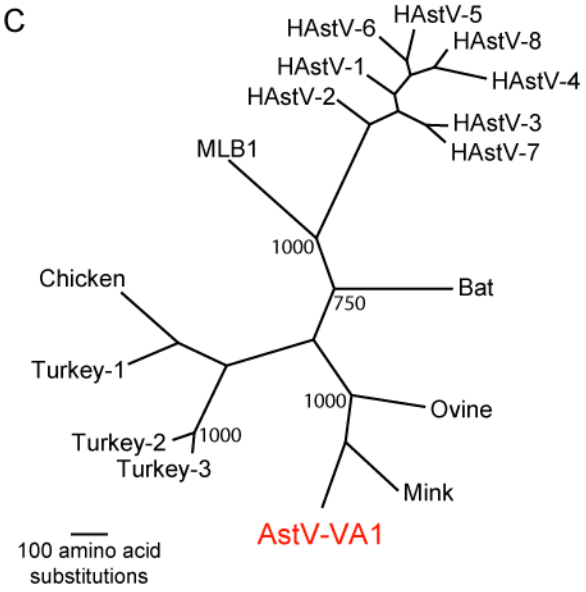


Figure 5.1: Phylogenetic analysis of AstV-VA1 open reading frames. Phylogenetic trees were generated in PAUP using the maximum parsimony method with 1,000 bootstrap replicates. Significant bootstrap values are shown. (A) ORF1a-serine protease; (B) ORF1b-polymerase; (C) ORF2-capsid. HAstV =Human astrovirus

astrovirus. The ORF1a (serine protease) coding region was more divergent with 39% and 40% amino acid identity with ovine astrovirus and mink astrovirus, respectively. In ORF2, AstV-VA1 shared 41% amino acid identity to mink astrovirus and 42% to ovine astrovirus.

RT-PCR screening for AstV-VA1

High throughput pyrosequencing yielded many AstV-VA1 sequences in samples B and C, but none were detected in

samples A or D. Sample E was not analyzed by pyrosequencing due to technical problems with the sample preparation. Similarly, Sanger sequencing detected AstV-VA1 positive reads in samples B and C, but not in samples A and E (sample D was not initially tested). To determine whether low levels of AstV-VA1 might be present in samples A, D and E, real time RT-PCR and semi-nested RT-PCR assays were developed targeting regions in ORF1b and ORF2, respectively. Using these assays, sample D tested positive and sequencing of the 250 bp amplicon confirmed the presence of AstV-VA1.

DISCUSSION

Despite the availability of improved molecular diagnostic methods for the detection of gastroenteritis viruses in humans such as norovirus, rotavirus, astrovirus, adenovirus, and sapovirus, the etiology of 12-41% of the outbreaks of gastroenteritis remain unexplained (94,95), In this study, we identified a novel astrovirus (AstV-VA1) in fecal samples from an outbreak of acute gastroenteritis in a child care center by two sequence independent genome amplification and sequencing methods, high throughput pyrosequencing and low throughput Sanger sequencing. Both methods identified and thus confirmed the presence of a novel astrovirus.

Complete genome sequencing and phylogenetic analysis demonstrated that AstV-VA1 was highly divergent from all previously described astroviruses including the 8 human astrovirus serotypes and recently described astrovirus MLB1 (AstV-MLB1). AstV-VA1 appears to have diverged from a common ancestor of the mink and ovine astroviruses following their separation from the branch containing human astroviruses 1-8 and astrovirus MLB1. The discovery of AstV-VA1 following the recent identification of AstV-MLB1 clearly demonstrates that a much greater diversity of astroviruses exists in humans than is commonly recognized.

The detection of AstV-VA1 in three out of five samples of this gastroenteritis outbreak suggests a potential association between AstV-VA1 and symptomatic infection. The fact that AstV-VA1 was only detected in sample D by targeted PCR assays and not by either of the mass sequencing methods may be due to the late timing of sample acquisition relative to the onset of symptoms (Table 1). Further studies defining the frequency of detection of AstV-VA1 in additional samples from individuals with and without acute gastroenteritis are needed to define the role of AstV-VA1 in human diarrhea. It is likely that the application of sequence independent amplification and sequencing methods to other outbreaks of gastroenteritis of unknown etiology will

identify other novel viruses and expand our ability to determine the cause of diarrheal disease.

MATERIALS AND METHODS

Outbreak

On Monday, 8/18/2008, the Eastern Shore Health District in Virginia was notified of cases of gastrointestinal illness among 26 teachers and children in a child care center over a period of 2 to 3 weeks. Symptoms included vomiting, and/or diarrhea. Control measures were put in place immediately at the center including exclusion of symptomatic children, mandated testing of all symptomatic staff, testing of symptomatic children, environmental disinfection of surfaces and ultimately, temporary closing of the facility. Five fecal specimens (A- E) (Table 5.1) that tested negative for enteric parasites, enteric bacteria by standard microscopy and culture, and negative for enteric viruses including rotavirus (RotaClone EIA), norovirus, sapovirus, human astrovirus, and adenovirus gp F by (RT)-PCR (92,158,159), were available for further testing.

Genome amplification and Sequencing.

The fecal specimens were further analyzed independently in two laboratories. At Washington University, the specimens were diluted in PBS at a 1:6 ratio (w/v) and total nucleic acid was extracted from 200µL of each fecal suspension using the MagNAPure LC Automated Nucleic Acid

Extraction System (Roche). Total nucleic acid was randomly amplified using the Round AB protocol as previously described with the exception that each sample was independently amplified with a different modified primer B containing a unique 6-nucleotide barcode at the 5' end of the primer (41). Amplification products from multiple samples were pooled, adaptor-ligated, and sequenced using the Roche GS-FLX Titanium platform (Roche) at the Washington University Genome Sequencing Center.

Sequences from each sample were identified by the unique barcodes introduced during the Rd B amplification. Primer and barcode sequences were then trimmed off prior to analysis of the sequences. Sequences were clustered using CD-HIT (160) to reduce redundancy with the requirement that they had to be 98% identical over 98% of their lengths. The longest sequence from each cluster was selected for inclusion in the pool of unique sequences to be analyzed. Unique sequences were filtered for repetitive sequences and compared with the human genome using BLASTn with an e-value cutoff of $1e^{-10}$. Sequences without significant similarity to the human genome were then compared to the GenBank nucleic acid nt database using BLASTn (cutoff: $1e^{-10}$) and tBLASTx (cutoff: $1e^{-5}$), and remaining sequences without significant hits to sequences in the database were then compared to the NCBI All Viral Genome database (ftp://ftp.ncbi.nih.gov/refseq/release/viral/) using

tBLASTx (cutoff: $1e^{-5}$) (130). Overlapping sequences with significant sequence identity were assembled into contigs using Newbler (454 Life Sciences) or CAP3 (161).

At CDC, 10% fecal suspensions were first clarified by centrifugation at

6,000×g for 10 minutes and the supernatant was then filtered through a 0.22-um filter (Ultrafree MC; Millipore, Bedford, MA). Total nucleic acid (TNA) was extracted from 200 µl of the cleared supernatant fluid with the QIAamp MinElute Virus Spin kit (QIAGEN, Valencia, CA) according to the manufacturer's instructions. After elution from the column in 50 µl of RNase-free water, TNA was randomly amplified using the Round AB protocol as previously described (41). The 300-800bp amplicons were then cloned using the TOPO TA cloning kit (Invitrogen, Carlsbad, CA) and plasmids were sequenced with a BigDye Terminators v3.1 ready reaction cycle sequencing kit on an ABI Prism 3130 automated sequencer (Applied Biosystems, Foster City, CA). Sequence analysis and generation of contigs were performed using Sequencher software (Ann Arbor, MI, USA). Sequence identification was performed through NCBI nucleotide–nucleotide BLASTn similarity searches (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi). In addition, a set of eight overlapping RT-PCR products with an average size of 900 bp which cover the entire genome including the 3' end poly A tail were generated by

primer pairs designed from clone sequences as described above, using the SuperScript III First-Strand Synthesis System for RT-PCR and AccuPrime High Fidelity Taq DNA polymerase (Invitrogen, Carlsbad, CA, USA). Both strands of each amplicon were sequenced with a BigDye Terminators v3.1 ready reaction cycle sequencing kit as described above. The 5' end genome sequence was amplified and determined using the 5' / 3' RACE Kit (Roche, Mannheim, Germany) following the manufacturer's instructions. The complete genome sequence of AstV-VA1 has been deposited in Genbank (number will be added once acquired).

ORF Prediction and annotation.

Open reading frames (ORFs) 1a and 2 were predicted by the NCBI ORF finder (http://www.ncbi.nlm.nih.gov/projects/gorf/). The end of ORF1b was predicted by the program, however the start of ORF1b was predicted based on the location of the heptameric slippery sequence found in other astroviruses (151). Protein motifs were identified by conserved domain searches using BlastX and Pfam (152,162-164).

Pair-wise alignments.

Bioedit was used to determine the percent identity between sequences as determined by pair-wise alignments.

Phylogenetic analysis.

ClustalX (1.83) was used to carry out multiple sequence alignments of the protein sequences associated with all three of the open reading frames of astroviruses for which sequences were available. Maximum parsimony trees were generated using PAUP with 1,000 bootstrap replicates (118). Available nucleotide or protein sequences of the following astroviruses were obtained: Human Astrovirus 1 (GenBank: NC_001943); Human Astrovirus 2 (GenBank: L13745); Human Astrovirus 3 (GenBank: AAD17224); Human Astrovirus 4 (GenBank: DQ070852); Human Astrovirus 5 (GenBank: DQ028633); Human Astrovirus 6 (EMBL: CAA86616); Human Astrovirus 7 (Gen Bank: AAK31913); Human Astrovirus 8 (GenBank: AF260508); Turkey Astrovirus 1 (GenBank: Y15936); Turkey Astrovirus 2 (GenBank: NC_005790); Turkey Astrovirus 3 (GenBank: AY769616); Chicken Astrovirus (GenBank: NC_003790); Ovine Astrovirus (GenBank: NC_002469); Mink Astrovirus (GenBank: NC_004579), Astrovirus MLB1 (GenBank: NC_011400), and Bat Astrovirus (GenBank: EU847155).

Real Time assay:

The real-time RT-PCR assay was performed using the SuperScript™ III One-Step RT-PCR kit (Invitrogen Corp., Carlsbad, CA) and the Mx4000® system (Stratagene, La Jolla, CA). Each 50ul reaction mixture contained 900 pmol

of forward primer (5' TAT CCA TAG TTG TGG ATA TTT GTC CA 3'), 1000 pmol of reverse primer (5' TGT CTT AGG GGA GAC TTG CAA A 3') and 100 pmol of probe (5' TT CC CCCT GTC CTG GAT TGT CAC TTC 3'), 1x buffer, 6.0 mM $MgSO_4$ (final concentration), 20 units of RNase inhibitor, a 5 µl aliquot of RNA extracts, and 1 unit of SuperScript III RT/Platinum Taq Mix. Water was added to achieve a final volume of 50 µl. The RT-PCR reaction mixture was incubated at 60ºC for 1 minute for denaturing, 50ºC for 30 minutes (for RT), 94ºC for 2 minutes (for hot start), then 40 cycles at 94ºC for 15 seconds; 55ºC for 30 seconds; 72ºC for 30 seconds and a final extension at 72ºC for 7 minutes. Fluorescence measurements were taken and the threshold cycle (CT) value for each sample was calculated by determining the point at which fluorescence exceeded a threshold limit set at the mean plus 10 standard deviations above the baseline.

Semi-nested RT-PCR assay:

The first round RT-PCR in the semi-nested assay was performed according to the protocol described previously (165) using forward primer (5' AGG GGT CGC TGG GAG TTT G 3') and reverse primer (5' GTC TAT TGT TTT GGG CGT CTG C 3'). The 2nd round PCR in the semi-nested assay PCR assay in 50ul reaction mixture contained 1x buffer (Platinum Taq kit; Invitrogen), 2 mM MgCl2, 200uM (each) of deoxynucleoside triphosphates, 50 pmol (each) of forward primer (5' AGG GGT CGC TGG GAG TTT G 3') and

reverse primer (5' CGG GGG TGG TGC GAC AT 3') 1 U Platinum Taq, one 2-ul aliquot from the first reaction, and water to achieve a final volume of 50 ul. The mixture was first heated to 94°C for 2 min. The cycling conditions were 40 cycles with the same conditions as for the first amplification: 94°C for 15 s, primer annealing at 55°C for 30 s, and 72°C for 30 s. A final extension was carried out at 72°C for 7 min. The final seminested PCR products were visualized by UV light after electrophoresis on a 2% agarose gel containing 0.5 ug/ml ethidium bromide in 0.5 x Tris-borate buffer. Amplicons from the final round of PCR were purified using the QIAquick PCR purification kit (Qiagen, Inc., Valencia, CA). Both strands of the amplicons were sequenced with a BigDye Terminators v3.1 ready reaction cycle sequencing kit as described above.

ACKNOWLEDGMENTS

This work was supported in part by National Institutes of Health grant 2 U54 AI057160-06 to the Midwest Regional Center of Excellence for Biodefense and Emerging Infectious Diseases Research. DW holds an Investigators in the Pathogenesis of Infectious Disease Award from the Burroughs Wellcome Fund.

Chapter 6: Conclusions

In the developed world, diarrhea is not commonly thought of as a major health concern. However, diarrhea is a yet unsolved problem given that it is still the third leading cause of death due to an infectious disease. Since ~40% of all cases are of unknown etiology, there is a major gap in our understanding of the causes of diarrhea. Not fully understanding the problem hinders the development of comprehensive tactics for prevention and treatment of diarrhea. My doctoral thesis work was aimed at addressing the existing gap in knowledge and has served to make important strides towards potentially closing this gap.

Currently, there is building excitement about the Human Microbiome Project which is a National Institutes of Health multi-institute iniative to use mass sequencing to determine the microbial populations associated with humans at various sites of the body. However, this is a new found excitement. Few people were talking about metagenomic analysis of human-associated microbial communities at the time in which this thesis work began. While some groups had examined the spectrum of viruses present in stools from healthy adults, and that is in part what the Human Microbiome Project proposes to do further, no one had previously examined what spectrum of viruses can be found in diarrhea samples. My initial analysis of 12 diarrheal stools was therefore the first snapshot of the diarrhea virome. This analysis revealed that the spectrum of viruses associated with diarrhea is highly varied in terms of the families of viruses

found in a given sample, the numbers of different viruses in a given sample, and the relative abundance of viruses in the samples. This suggests that the complexity of the human diarrhea virome is quite complex and highlights the difficulty of assessing the role of an individual virus in causing diarrhea when there is a complex mixture of viruses present in a sample.

Perhaps the most exciting finding that came out of the metagenomic analysis of the first 12 diarrhea samples was that there were many novel viruses present in the samples. The identification of multiple novel viruses in just a small random sampling of diarrhea specimens provides confidence that there are still many unknown viruses to be discovered and provides further encouragement to continue searching for them.

One of the novel viruses that was identified belongs to the family *Nodaviridae*. The identification of this virus perfectly exemplifies the possibility for paradigm shifting discoveries. Nodaviruses are known to infect fish and insects, but do not naturally infect humans. The identification of a novel nodavirus in a human diarrhea specimen presents the possibility that the virus infects humans. There are other explanations though. For example, it is possible that the virus was derived from some dietary source, perhaps an infected fish, and was simply passed through the gastrointestinal tract. However, if it can be demonstrated that this

novel nodavirus does infect humans, it would be an exciting finding since this virus would be the first nodavirus known to infect humans. Discoveries of viruses like this one points out the need to have an unbiased approach to search for novel viruses and furthermore for thinking about what the potential causes of diarrhea might be. Of course, much more work needs to be done in order to discern between the possible explanations for the nodavirus' presence in the diarrhea sample.

The most promising discovery of the initial metagenomic analysis was that of a novel astrovirus, AstV-MLB1, in an Australian diarrhea specimen. Phylogenetic analysis revealed that AstV-MLB1 is highly distinct from any of the known astroviruses, including the known human astroviruses, confirming that AstV-MLB1 is in fact a highly divergent, novel astrovirus. Given that astroviruses are known to infect humans and that they typically cause diarrhea in their hosts it is easy to believe that this astrovirus could also cause diarrhea in humans, however that has yet to be proven. The detection of AstV-MLB1 in additional diarrhea specimens from Saint Louis is the first report examining the prevalence of AstV-MLB1 and the first detection of AstV-MLB1 outside of Australia. Detection of AstV-MLB1 on two continents suggests that the virus may be globally widespread, suggesting that if it is shown to be a human pathogen it may affect a wide distribution of people around the world.

The identification of MLB1 was followed by the identification of yet another highly divergent, novel astrovirus. This astrovirus, AstV-VA1, was found in 3/5 samples from a gastroenteritis outbreak. The identification of AstV-VA1 in association with a diarrhea outbreak presents a very strong possibility that this astrovirus is in fact associated with diarrhea.

While astroviruses are known to cause diarrhea, they have generally been associated with at most 10% of both sporadic and outbreak cases of acute diarrhea. The identification of two novel astroviruses in human diarrhea samples hints that astroviruses may have a larger role in causing human diarrhea than previously believed. There are 8 known human astroviruses which are all very closely genetically related. Following that example, it is possible that each of these viruses represents a whole new cluster of astroviruses. One could furthermore hypothesize that these novel astroviruses may eventually be shown to cause diarrhea and allowing us to begin to close the gap in our understanding of the causes of diarrhea.

Significant technological developments within recent years have greatly enhanced our ability to detect both known and novel viruses as evidenced by the work presented in this dissertation as well as work done by others. Continuing evolution of these technologies will certainly lead to further increases in this arena. Critically, while the identification of novel viruses is becoming increasingly more facile, the discovery is merely the introductory chapter into each virus' story. For example, with each

new virus identified, there are a host of questions that arise regarding each virus' tropism, epidemiology, and potential link to disease. To answer these questions will require many years of further investigation, using myriad tools of biology, virology and medicine.

The scientific questions that need to be addressed for each virus are initially all the same. The major questions for both AstV-MLB1 and AstV-VA1 are whether they are actually human pathogens and if so, whether they do in fact cause diarrhea. These questions are difficult to answer short of doing human infection trials. The founding principles for demonstrating microbial pathogenesis are based on satisfying Kochs postulates. Today we know that Kochs postulates cannot always be satisfied for every pathogen. The reasons for this can vary from ethical considerations of carrying out human experiments to the fact that not all pathogens can be cultured, which is a requirement for satisfying the postulates.

Culturing of these novel viruses may present a major challenge for future research investigations. Since traditional virus discovery methods relied heavily on the ability to culture viruses, one might argue that these viruses are unlikely to grow well under standard culturing conditions or else they would have already been discovered. There are many important humans pathogens that have proven difficult to culture such as human noroviruses and hepatitis C virus. For these viruses, significant effort has been exerted since their discoveries in trying to get them to grow in the

laboratory. It might be the case that significant creativity as well as a lot of trial and error will be required to establish culturing conditions for these novel astroviruses. This is a very important direction for future research on these viruses because having a culturable virus makes exploring other scientific questions easier and also facilities production of vaccines and is useful for other translational applications.

Kochs first postulate is that an organism must be found in all disease cases and in no healthy cases in order for it to be considered a pathogen. Again, we know that these absolutes are not always true, especially when considering a polymicrobial disease like diarrhea. There are many factors such as the host immune system, co-infecting microorganisms, and environmental factors that are speculated to be involved in viral pathogenesis, but the extent of how these factors affect different viruses is not well understood. However, if AstV-MLB1 and AstV-VA1 can be statistically associated with diarrhea cases rather than healthy cases, then that will provide compelling evidence to say that these viruses cause diarrhea. Therefore, screening large cohorts of case-controlled diarrhea specimens is also an essential next step for the progression of research on these viruses.

Finally, some of the best evidence to demonstrate that these viruses infect humans is seroconversion upon exposure to the virus. It is assumed that the generation of antibodies implies that the virus actually infected

the host as opposed to having been passed through the gastrointestinal tract. Again since there are ethical concerns for doing human experiments with novel viruses about which very little is known, the best proxy for directly monitoring seroconversion is to examining existing immunity to these viruses in the general population. Seroprevalence studies can give us some information about how common these viruses are in the human population and potentially at what age people are frequently exposed to them.

Historically, it has been postulated that the identification of novel viruses in a given specimen or syndrome is the rate-limiting step in addressing the question of disease causality. With the advent of new technologies, a shift has occurred such that in many instances, the rate-limiting step is no longer discovery but understanding the biological relevance and impact of newly discovered viruses. The application of unbiased, culture independent discovery methods has already yielded many fruit in the form of new viruses, which are likely to be just the tip of the proverbial iceberg. Undoubtedly many additional new viruses will be uncovered in the upcoming years, thereby providing new substrates for investigation and understanding of virology, virus host interactions, virus-environment interactions, and of course disease pathogenesis.

REFERENCES

1. Nichol ST, Spiropoulou CF, Morzunov S, Rollin PE, Ksiazek TG, et al. (1993) Genetic identification of a hantavirus associated with an outbreak of acute respiratory illness. Science 262: 914-917.
2. Woo PC, Lau SK, Chu CM, Chan KH, Tsoi HW, et al. (2005) Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. J Virol 79: 884-895.
3. Rose TM, Schultz ER, Henikoff JG, Pietrokovski S, McCallum CM, et al. (1998) Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. Nucleic Acids Res 26: 1628-1635.
4. Jabado OJ, Palacios G, Kapoor V, Hui J, Renwick N, et al. (2006) Greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments. Nucleic Acids Res 34: 6605-6611.
5. Lisitsyn N, Lisitsyn N, Wigler M (1993) Cloning the differences between two complex genomes. Science 259: 946-951.
6. Chang Y, Cesarman E, Pessin MS, Lee F, Culpepper J, et al. (1994) Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. Science 266: 1865-1869.
7. Karst SM, Wobus CE, Lay M, Davidson J, Virgin HWt (2003) STAT1-dependent innate immunity to a Norwalk-like virus. Science 299: 1575-1578.
8. Choo QL, Kuo G, Weiner AJ, Overby LR, Bradley DW, et al. (1989) Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. Science 244: 359-362.
9. Delgado S, Erickson BR, Agudo R, Blair PJ, Vallejo E, et al. (2008) Chapare virus, a newly discovered arenavirus isolated from a fatal hemorrhagic fever case in Bolivia. PLoS Pathog 4: e1000047.
10. Chua KB, Crameri G, Hyatt A, Yu M, Tompang MR, et al. (2007) A previously unknown reovirus of bat origin is associated with an acute respiratory disease in humans. Proc Natl Acad Sci U S A 104: 11424-11429.
11. van den Hoogen BG, de Jong JC, Groen J, Kuiken T, de Groot R, et al. (2001) A newly discovered human pneumovirus isolated from young children with respiratory tract disease. Nat Med 7: 719-724.
12. Welsh J, McClelland M (1990) Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res 18: 7213-7218.
13. Ralph D, McClelland M, Welsh J (1993) RNA fingerprinting using arbitrarily primed PCR identifies differentially regulated RNAs in mink

lung (Mv1Lu) cells growth arrested by transforming growth factor beta 1. Proc Natl Acad Sci U S A 90: 10710-10714.
14. Lindner J, Modrow S (2008) Human bocavirus--a novel parvovirus to infect humans. Intervirology 51: 116-122.
15. Reyes GR, Kim JP (1991) Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. Mol Cell Probes 5: 473-481.
16. Allander T, Emerson SU, Engle RE, Purcell RH, Bukh J (2001) A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. Proc Natl Acad Sci U S A 98: 11609-11614.
17. Jones MS, Kapoor A, Lukashov VV, Simmonds P, Hecht F, et al. (2005) New DNA viruses identified in patients with acute viral infection syndrome. J Virol 79: 8230-8236.
18. Jones MS, 2nd, Harrach B, Ganac RD, Gozum MM, Dela Cruz WP, et al. (2007) New adenovirus species found in a patient presenting with gastroenteritis. J Virol 81: 5978-5984.
19. Jones MS, Lukashov VV, Ganac RD, Schnurr DP (2007) Discovery of a novel human picornavirus in a stool sample from a pediatric patient presenting with fever of unknown origin. J Clin Microbiol 45: 2144-2150.
20. Chiu CY, Greninger AL, Kanada K, Kwok T, Fischer KF, et al. (2008) Identification of cardioviruses related to Theiler's murine encephalomyelitis virus in human infections. Proc Natl Acad Sci U S A.
21. Abed Y, Boivin G (2008) New saffold cardioviruses in 3 children, Canada. Emerg Infect Dis 14: 834-836.
22. Drexler JF, Luna LK, Stocker A, Almeida PS, Ribeiro TC, et al. (2008) Circulation of 3 lineages of a novel Saffold cardiovirus in humans. Emerg Infect Dis 14: 1398-1405.
23. van der Hoek L, Pyrc K, Jebbink MF, Vermeulen-Oost W, Berkhout RJ, et al. (2004) Identification of a new human coronavirus. Nat Med 10: 368-373.
24. van der Hoek L, Sure K, Ihorst G, Stang A, Pyrc K, et al. (2005) Croup is associated with the novel coronavirus NL63. PLoS Med 2: e240.
25. de Souza Luna LK, Baumgarte S, Grywna K, Panning M, Drexler JF, et al. (2008) Identification of a contemporary human parechovirus type 1 by VIDISCA and characterisation of its full genome. Virol J 5: 26.
26. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, et al. (1991) Light-directed, spatially addressable parallel chemical synthesis. Science 251: 767-773.
27. Maskos U, Southern EM (1992) Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and

hybridization properties of oligonucleotides synthesised in situ. Nucleic Acids Res 20: 1679-1684.
28. Schena M (1996) Genome analysis with gene expression microarrays. Bioessays 18: 427-431.
29. Carpten J, Nupponen N, Isaacs S, Sood R, Robbins C, et al. (2002) Germline mutations in the ribonuclease L gene in families showing linkage with HPC1. Nat Genet 30: 181-184.
30. Urisman A, Fischer KF, Chiu CY, Kistler AL, Beck S, et al. (2005) E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns. Genome Biol 6: R78.
31. Kistler A, Avila PC, Rouskin S, Wang D, Ward T, et al. (2007) Pan-viral screening of respiratory tract infections in adults with and without asthma reveals unexpected human coronavirus and human rhinovirus diversity. J Infect Dis 196: 817-825.
32. Chiu CY, Urisman A, Greenhow TL, Rouskin S, Yagi S, et al. (2008) Utility of DNA microarrays for detection of viruses in acute respiratory tract infections in children. J Pediatr 153: 76-83.
33. Chiu CY, Alizadeh AA, Rouskin S, Merker JD, Yeh E, et al. (2007) Diagnosis of a critical respiratory illness caused by human metapneumovirus by use of a pan-virus microarray. J Clin Microbiol 45: 2340-2343.
34. Korimbocus J, Scaramozzino N, Lacroix B, Crance JM, Garin D, et al. (2005) DNA probe array for the simultaneous identification of herpesviruses, enteroviruses, and flaviviruses. J Clin Microbiol 43: 3779-3787.
35. Lin B, Wang Z, Vora GJ, Thornton JA, Schnur JM, et al. (2006) Broad-spectrum respiratory tract pathogen identification using resequencing DNA microarrays. Genome Res 16: 527-535.
36. Wong CW, Heng CL, Wan Yee L, Soh SW, Kartasasmita CB, et al. (2007) Optimization and clinical validation of a pathogen detection microarray. Genome Biol 8: R93.
37. Palacios G, Quan PL, Jabado OJ, Conlan S, Hirschberg DL, et al. (2007) Panmicrobial oligonucleotide array for diagnosis of infectious diseases. Emerg Infect Dis 13: 73-81.
38. Watson M, Dukes J, Abu-Median AB, King DP, Britton P (2007) DetectiV: visualization, normalization and significance testing for pathogen-detection microarray data. Genome Biol 8: R190.
39. Rehrauer H, Schonmann S, Eberl L, Schlapbach R (2008) PhyloDetect: a likelihood-based strategy for detecting microorganisms with diagnostic microarrays. Bioinformatics 24: i83-89.
40. Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, et al. (2003) A novel coronavirus associated with severe acute respiratory syndrome. N Engl J Med 348: 1953-1966.

41. Wang D, Urisman A, Liu YT, Springer M, Ksiazek TG, et al. (2003) Viral discovery and sequence recovery using DNA microarrays. PLoS Biol 1: E2.
42. Monceyron C, Grinde B, Jonassen TO (1997) Molecular characterisation of the 3'-end of the astrovirus genome. Arch Virol 142: 699-706.
43. Peiris JS, Lai ST, Poon LL, Guan Y, Yam LY, et al. (2003) Coronavirus as a possible cause of severe acute respiratory syndrome. Lancet 361: 1319-1325.
44. Drosten C, Gunther S, Preiser W, van der Werf S, Brodt HR, et al. (2003) Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N Engl J Med 348: 1967-1976.
45. Urisman A, Molinaro RJ, Fischer N, Plummer SJ, Casey G, et al. (2006) Identification of a novel Gammaretrovirus in prostate tumors of patients homozygous for R462Q RNASEL variant. PLoS Pathog 2: e25.
46. Casey G, Neville PJ, Plummer SJ, Xiang Y, Krumroy LM, et al. (2002) RNASEL Arg462Gln variant is implicated in up to 13% of prostate cancer cases. Nat Genet 32: 581-583.
47. Rokman A, Ikonen T, Seppala EH, Nupponen N, Autio V, et al. (2002) Germline alterations of the RNASEL gene, a candidate HPC1 gene at 1q25, in patients and families with prostate cancer. Am J Hum Genet 70: 1299-1304.
48. Rennert H, Bercovich D, Hubert A, Abeliovich D, Rozovsky U, et al. (2002) A novel founder mutation in the RNASEL gene, 471delAAAG, is associated with prostate cancer in Ashkenazi Jews. Am J Hum Genet 71: 981-984.
49. Carter BS, Bova GS, Beaty TH, Steinberg GD, Childs B, et al. (1993) Hereditary prostate cancer: epidemiologic and clinical features. J Urol 150: 797-802.
50. Dong B, Kim S, Hong S, Das Gupta J, Malathi K, et al. (2007) An infectious retrovirus susceptible to an IFN antiviral pathway from human prostate tumors. Proc Natl Acad Sci U S A 104: 1655-1660.
51. Mihindukulasuriya KA, Wu G, St Leger J, Nordhausen RW, Wang D (2008) Identification of a novel coronavirus from a beluga whale by using a panviral microarray. J Virol 82: 5084-5088.
52. Kistler AL, Gancz A, Clubb S, Skewes-Cox P, Fischer K, et al. (2008) Recovery of divergent avian bornaviruses from cases of proventricular dilatation disease: identification of a candidate etiologic agent. Virol J 5: 88.
53. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74: 5463-5467.
54. Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P (1996) Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem 242: 84-89.

55. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24: 133-141.
56. Xu Y, Stange-Thomann N, Weber G, Bo R, Dodge S, et al. (2003) Pathogen discovery from human tissue by sequence-based computational subtraction. Genomics 81: 329-335.
57. Allander T, Tammi MT, Eriksson M, Bjerkner A, Tiveljung-Lindell A, et al. (2005) Cloning of a human parvovirus by molecular screening of respiratory tract samples. Proc Natl Acad Sci U S A 102: 12891-12896.
58. Allander T, Andreasson K, Gupta S, Bjerkner A, Bogdanovic G, et al. (2007) Identification of a third human polyomavirus. J Virol 81: 4130-4136.
59. Gaynor AM, Nissen MD, Whiley DM, Mackay IM, Lambert SB, et al. (2007) Identification of a novel polyomavirus from patients with acute respiratory tract infections. PLoS Pathog 3: e64.
60. Bialasiewicz S, Whiley DM, Lambert SB, Jacob K, Bletchly C, et al. (2008) Presence of the newly discovered human polyomaviruses KI and WU in Australian patients with acute respiratory tract infection. J Clin Virol 41: 63-68.
61. Abed Y, Wang D, Boivin G (2007) WU polyomavirus in children, Canada. Emerg Infect Dis 13: 1939-1941.
62. Han TH, Chung JY, Koo JW, Kim SW, Hwang ES (2007) WU polyomavirus in children with acute lower respiratory tract infections, South Korea. Emerg Infect Dis 13: 1766-1768.
63. Lin F, Zheng M, Li H, Zheng C, Li X, et al. (2008) WU polyomavirus in children with acute lower respiratory tract infections, China. J Clin Virol 42: 94-102.
64. Neske F, Blessing K, Ullrich F, Prottel A, Wolfgang Kreth H, et al. (2008) WU polyomavirus infection in children, Germany. Emerg Infect Dis 14: 680-681.
65. Payungporn S, Chieochansin T, Thongmee C, Samransamruajkit R, Theamboolers A, et al. (2008) Prevalence and molecular characterization of WU/KI polyomaviruses isolated from pediatric patients with respiratory disease in Thailand. Virus Res.
66. Le BM, Demertzis LM, Wu G, Tibbets RJ, Buller R, et al. (2007) Clinical and epidemiologic characterization of WU polyomavirus infection, St. Louis, Missouri. Emerg Infect Dis 13: 1936-1938.
67. Feng H, Shuda M, Chang Y, Moore PS (2008) Clonal integration of a polyomavirus in human Merkel cell carcinoma. Science 319: 1096-1100.
68. Lemos B, Nghiem P (2007) Merkel cell carcinoma: more deaths but still no pathway to blame. J Invest Dermatol 127: 2100-2103.
69. Goessling W, McKee PH, Mayer RJ (2002) Merkel cell carcinoma. J Clin Oncol 20: 588-598.

70. Becker JC, Houben R, Ugurel S, Trefzer U, Pfohler C, et al. (2008) MC Polyomavirus Is Frequently Present in Merkel Cell Carcinoma of European Patients. J Invest Dermatol.
71. Foulongne V, Kluger N, Dereure O, Brieu N, Guillot B, et al. (2008) Merkel cell polyomavirus and Merkel cell carcinoma, France. Emerg Infect Dis 14: 1491-1493.
72. Garneski KM, Warcola AH, Feng Q, Kiviat NB, Leonard JH, et al. (2008) Merkel Cell Polyomavirus Is More Frequently Present in North American than Australian Merkel Cell Carcinoma Tumors. J Invest Dermatol.
73. Kassem A, Schopflin A, Diaz C, Weyers W, Stickeler E, et al. (2008) Frequent detection of Merkel cell polyomavirus in human Merkel cell carcinomas and identification of a unique deletion in the VP1 gene. Cancer Res 68: 5009-5013.
74. Poulin DL, DeCaprio JA (2006) Is there a role for SV40 in human cancer? J Clin Oncol 24: 4356-4365.
75. Palacios G, Druce J, Du L, Tran T, Birch C, et al. (2008) A new arenavirus in a cluster of fatal transplant-associated diseases. N Engl J Med 358: 991-998.
76. Allander T, de Lamballerie X, Simmonds P (2008) A new arenavirus in transplantation. N Engl J Med 358: 2638; author reply 2638-2639.
77. Kapoor A, Victoria J, Simmonds P, Wang C, Shafer RW, et al. (2008) A highly divergent picornavirus in a marine mammal. J Virol 82: 311-320.
78. Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, et al. (2006) RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol 4: e3.
79. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, et al. (2003) Metagenomic analyses of an uncultured viral community from human feces. J Bacteriol 185: 6220-6223.
80. (2004) World Health Report. World Health Organization.
81. O'Ryan M, Prado V, Pickering LK (2005) A millennium update on pediatric diarrheal illness in the developing world. Semin Pediatr Infect Dis 16: 125-136.
82. Kosek M, Bern C, Guerrant RL (2003) The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000. Bulletin of the World Health Organization. pp. 197-204.
83. Cheng AC, McDonald JR, Thielman NM (2005) Infectious diarrhea in developed and developing countries. J Clin Gastroenterol 39: 757-773.
84. Klein EJ, Boster DR, Stapp JR, Wells JG, Qin X, et al. (2006) Diarrhea Etiology in a Children's Hospital Emergency Department: A Prospective Cohort Study. Clin Infect Dis 43: 807-813.

85. Kirkwood CD, Clark R, Bogdanovic-Sakran N, Bishop RF (2005) A 5-year study of the prevalence and genetic diversity of human caliciviruses associated with sporadic cases of acute gastroenteritis in young children admitted to hospital in Melbourne, Australia (1998-2002). J Med Virol 77: 96-101.
86. Nataro JP, Mai V, Johnson J, Blackwelder WC, Heimer R, et al. (2006) Diarrheagenic Escherichia coli infection in Baltimore, Maryland, and New Haven, Connecticut. Clin Infect Dis 43: 402-407.
87. Clark B, McKendrick M (2004) A review of viral gastroenteritis. Curr Opin Infect Dis 17: 461-469.
88. Wilhelmi I, Roman E, Sanchez-Fauquier A (2003) Viruses causing gastroenteritis. Clin Microbiol Infect 9: 247-262.
89. Wigand R, Baumeister H, Maass G, Kuhn J, Hammer H (1983) Isolation and Identification of Enteric Adenoviruses. Journal of Medical Virology 11: 233-240.
90. Kurtz JB, Lee TW, Craig JW, Reed SE (1979) Astrovirus infection in volunteers. J Med Virol 3: 221-230.
91. Thornhill T, Kalica A, Wyatt R, Kapikan A, Chanock R (1975) Pattern of Shedding of the Norwalk Particle in Stools during Experimentally Induced Gastroenteritis in Volunteers as Determined by Immune Electron Microscopy. The Journal of Infectious Diseases 132: 28-34.
92. Davidson G, Townley R, Bishop RF, Holmes I, Ruck B (1975) Importance of a new virus in acute sporadic enteritis in children. The Lancet: 242-246.
93. Kapikan A (1993) Viral Gastroenteritis. The Journal of the American Medical Association 269: 627-630.
94. Lyman WH, Walsh JF, Kotch JB, Weber DJ, Gunn E, et al. (2009) Prospective study of etiologic agents of acute gastroenteritis outbreaks in child care centers. J Pediatr 154: 253-257.
95. Svraka S, Duizer E, Vennema H, de Bruin E, van der Veer B, et al. (2007) Etiological role of viruses in outbreaks of acute gastroenteritis in The Netherlands from 1994 through 2005. J Clin Microbiol 45: 1389-1394.
96. Frenzen PD (2003) Mortality due to gastroenteritis of unknown etiology in the United States. J Infect Dis 187: 441-452.
97. Caul EO, Paver WK, Clarke SK (1975) Letter: Coronavirus particles in faeces from patients with gastroenteritis. Lancet 1: 1192.
98. Paver WK, Caul EO, Clarke SK (1975) Letter: Parvovirus-like particles in human faeces. Lancet 1: 691.
99. Mathan M, Mathan VI, Swaminathan SP, Yesudoss S (1975) Pleomorphic virus-like particles in human faeces. Lancet 1: 1068-1069.
100. Madeley CR, Cosgrove BP (1975) Letter: 28 nm particles in faeces in infantile gastroenteritis. Lancet 2: 451-452.

101. Appleton H, Higgins PG (1975) Letter: Viruses and gastroenteritis in infants. Lancet 1: 1297.
102. Koci MD, Schultz-Cherry S (2002) Avian astroviruses. Avian Pathol 31: 213-227.
103. Chu DK, Poon LL, Guan Y, Peiris JS (2008) Novel astroviruses in insectivorous bats. J Virol 82: 9107-9114.
104. Moser LA, Schultz-Cherry S (2005) Pathogenesis of astrovirus infection. Viral Immunol 18: 4-10.
105. Glass RI, Noel J, Mitchell D, Herrmann JE, Blacklow NR, et al. (1996) The changing epidemiology of astrovirus-associated gastroenteritis: a review. Arch Virol Suppl 12: 287-300.
106. Matsui SM, Greenberg HB (2001) Astroviruses. In: Knipe DM, Howley PM, editors. Fields Virology. 4 ed. Philadelphia: Lippincott WIIIliams & Wilkins. pp. 875-894.
107. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, et al. (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A 99: 14250-14255.
108. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428: 37-43.
109. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66-74.
110. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554-557.
111. Hallam SJ, Putnam N, Preston CM, Detter JC, Rokhsar D, et al. (2004) Reverse methanogenesis: testing the hypothesis with environmental genomics. Science 305: 1457-1462.
112. Culley AI, Lang AS, Suttle CA (2006) Metagenomic analysis of coastal RNA virus communities. Science 312: 1795-1798.
113. Dennehy PH (2005) Acute diarrheal disease in children: epidemiology, prevention, and treatment. Infect Dis Clin North Am 19: 585-602.
114. Denno DM, Klein EJ, Young VB, Fox JG, Wang D, et al. (2007) Explaining unexplained diarrhea and associating risks and infections. Anim Health Res Rev 8: 69-80.
115. Bon F, Fascia P, Dauvergne M, Tenenbaum D, Planson H, et al. (1999) Prevalence of group A rotavirus, human calicivirus, astrovirus, and adenovirus type 40 and 41 infections among children with acute gastroenteritis in Dijon, France. J Clin Microbiol 37: 3055-3058.
116. Chikhi-Brachet R, Bon F, Toubiana L, Pothier P, Nicolas JC, et al. (2002) Virus diversity in a winter epidemic of acute diarrhea in France. J Clin Microbiol 40: 4266-4272.

117. Berns KP, CR (2007) Parvoviridae. In: Knipe DH, PM, editor. Fields Virology. 5th ed: Lippincott Williams & Wilkins. pp. 2437-2477.
118. Swofford DL (1998) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Sunderland, Massachusettes: Sinauer Associates.
119. Xing L, Tikoo SK (2004) Viral RNAs detected in virions of porcine adenovirus type 3. Virology 321: 372-382.
120. Mannhalter C, Koizar D, Mitterbauer G (2000) Evaluation of RNA isolation methods and reference genes for RT-PCR analyses of rare target RNA. Clin Chem Lab Med 38: 171-177.
121. Gallimore CI, Appleton H, Lewis D, Green J, Brown DW (1995) Detection and characterisation of bisegmented double-stranded RNA viruses (picobirnaviruses) in human faecal specimens. J Med Virol 45: 135-140.
122. Friesen P (2007) Insect Viruses. In: Knipe DH, PM, editor. Fields Virology. 5th ed: Lippincott Williams & Wilkins. pp. 725-727.
123. Smit A, Hubley, R & Green, P. (1996-2004) RepeatMasker Open-3.0.
124. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110: 462-467.
125. Iturriza-Gomara M, Isherwood B, Desselberger U, Gray J (2001) Reassortment in vivo: driving force for diversity of human rotavirus strains isolated in the United Kingdom between 1995 and 1999. J Virol 75: 3696-3705.
126. Mustafa H, Palombo EA, Bishop RF (1998) Improved sensitivity of astrovirus-specific RT-PCR following culture of stool samples in CaCo-2 cells. J Clin Virol 11: 103-107.
127. Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8: 175-185.
128. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8: 186-194.
129. Chou HH, Holmes MH (2001) DNA sequence quality trimming and vector removal. Bioinformatics 17: 1093-1104.
130. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.
131. Mendez E, Arias CF (2007) Astroviruses. In: Knipe DM, Howley PM, editors. Fields Virology. 5th ed. Philadelphia: Lippincott WIIIliams & Wilkins. pp. 981-1000.
132. Soares CC, Maciel de Albuquerque MC, Maranhao AG, Rocha LN, Ramirez ML, 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. J Med Virol 80: 113-117.

133. Caracciolo S, Minini C, Colombrita D, Foresti I, Avolio M, et al. (2007) Detection of sporadic cases of Norovirus infection in hospitalized children in Italy. New Microbiol 30: 49-52.
134. Finkbeiner SR, Allred AF, Tarr PI, Klein EJ, Kirkwood CD, et al. (2008) Metagenomic analysis of human diarrhea: viral detection and discovery. PLoS Pathog 4: e1000011.
135. Kiang D, Matsui SM (2002) Proteolytic processing of a human astrovirus nonstructural protein. J Gen Virol 83: 25-34.
136. Jonassen CM, Jonassen TT, Sveen TM, Grinde B (2003) Complete genomic sequences of astroviruses from sheep and turkey: comparison with related viruses. Virus Res 91: 195-201.
137. Al-Mutairy B, Walter JE, Pothen A, Mitchell DK (2005) Genome Prediction of Putative Genome-Linked Viral Protein (VPg) of Astroviruses. Virus Genes 31: 21-30.
138. Matsui SM, Kim JP, Greenberg HB, Young LM, Smith LS, et al. (1993) Cloning and characterization of human astrovirus immunoreactive epitopes. J Virol 67: 1712-1715.
139. Guix S, Bosch A, Ribes E, Dora Martinez L, Pinto RM (2004) Apoptosis in astrovirus-infected CaCo-2 cells. Virology 319: 249-261.
140. Mendez E, Salas-Ocampo E, Arias CF (2004) Caspases mediate processing of the capsid precursor and cell release of human astroviruses. J Virol 78: 8601-8608.
141. Moon S, Byun Y, Kim HJ, Jeong S, Han K (2004) Predicting genes expressed via -1 and +1 frameshifts. Nucleic Acids Res 32: 4884-4892.
142. Monroe SS, Jiang B, Stine SE, Koopmans M, Glass RI (1993) Subgenomic RNA sequence of human astrovirus supports classification of Astroviridae as a new family of RNA viruses. J Virol 67: 3611-3614.
143. Willcocks MM, Carter MJ (1993) Identification and sequence determination of the capsid protein gene of human astrovirus serotype 1. FEMS Microbiol Lett 114: 1-7.
144. Wang QH, Kakizawa J, Wen LY, Shimizu M, Nishio O, et al. (2001) Genetic analysis of the capsid region of astroviruses. J Med Virol 64: 245-255.
145. Mendez-Toss M, Romero-Guido P, Munguia ME, Mendez E, Arias CF (2000) Molecular analysis of a serotype 8 human astrovirus genome. J Gen Virol 81: 2891-2897.
146. Walter JE, Briggs J, Guerrero ML, Matson DO, Pickering LK, et al. (2001) Molecular characterization of a novel recombinant strain of human astrovirus associated with gastroenteritis in children. Arch Virol 146: 2357-2367.
147. Yamashita T, Sakae K, Ishihara Y, Isomura S, Utagawa E (1993) Prevalence of newly isolated, cytopathic small round virus (Aichi strain) in Japan. J Clin Microbiol 31: 2938-2943.

148. Yamashita T, Kobayashi S, Sakae K, Nakata S, Chiba S, et al. (1991) Isolation of cytopathic small round viruses with BS-C-1 cells from patients with gastroenteritis. J Infect Dis 164: 954-957.
149. Baxendale W, Mebatsion T (2004) The isolation and characterisation of astroviruses from chickens. Avian Pathol 33: 364-370.
150. Bendinelli M, Pistello M, Maggi F, Fornai C, Freer G, et al. (2001) Molecular properties, biology, and clinical implications of TT virus, a recently identified widespread infectious agent of humans. Clin Microbiol Rev 14: 98-113.
151. Jiang B, Monroe SS, Koonin EV, Stine SE, Glass RI (1993) RNA sequence of astrovirus: distinctive genomic organization and a putative retrovirus-like ribosomal frameshifting signal that directs the viral replicase synthesis. Proc Natl Acad Sci U S A 90: 10539-10543.
152. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, et al. (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34: D247-251.
153. Finkbeiner SR, Kirkwood CD, Wang D (2008) Complete genome sequence of a highly divergent astrovirus isolated from a child with acute diarrhea. Virol J 5: 117.
154. Noel JS, Lee TW, Kurtz JB, Glass RI, Monroe SS (1995) Typing of human astroviruses from clinical isolates by enzyme immunoassay and nucleotide sequencing. J Clin Microbiol 33: 797-801.
155. Akihara S, Phan TG, Nguyen TA, Hansman G, Okitsu S, et al. (2005) Existence of multiple outbreaks of viral gastroenteritis among infants in a day care center in Japan. Arch Virol 150: 2061-2075.
156. Holtz LR, Finkbeiner SR, Kirkwood CD, Wang D (2008) Identification of a novel picornavirus related to cosaviruses in a child with acute diarrhea. Virol J 5: 159.
157. Jonassen CM, Jonassen TO, Grinde B (1998) A common RNA motif in the 3' end of the genomes of astroviruses, avian infectious bronchitis virus and an equine rhinovirus. J Gen Virol 79 (Pt 4): 715-718.
158. Oka T, Katayama K, Hansman GS, Kageyama T, Ogawa S, et al. (2006) Detection of human sapovirus by real-time reverse transcription-polymerase chain reaction. J Med Virol 78: 1347-1353.
159. Trujillo AA, McCaustland KA, Zheng DP, Hadley LA, Vaughn G, et al. (2006) Use of TaqMan real-time reverse transcription-PCR for rapid detection, quantification, and typing of norovirus. J Clin Microbiol 44: 1405-1412.
160. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658-1659.
161. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9: 868-877.

162. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, et al. (2005) CDD: a Conserved Domain Database for protein classification. Nucleic Acids Res 33: D192-196.
163. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, et al. (2007) CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res 35: D237-240.
164. Marchler-Bauer A, Bryant SH (2004) CD-Search: protein domain annotations on the fly. Nucleic Acids Res 32: W327-331.
165. Tong S, Chern SW, Li Y, Pallansch MA, Anderson LJ (2008) Sensitive and broadly reactive reverse transcription-PCR assays to detect novel paramyxoviruses. J Clin Microbiol 46: 2652-2658.