Genomic Approaches to Virus Detection and Discovery in Acute Pediatric Illness in Nicaragua

https://escholarship.org/uc/item/5733d92j

Yozwiak, Nathan Lawrence

2011

Peer reviewed|Thesis/dissertation
Genomic Approaches to Virus Detection and Discovery in Acute Pediatric Illness in Nicaragua

by

Nathan Lawrence Yozwiak

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Infectious Disease and Immunity in the Graduate Division of the University of California, Berkeley

Committee in charge:

Professor Eva Harris, Chair
Professor Joseph DeRisi (UCSF)
Professor George Sensabaugh
Professor Jack Colford

Spring 2011
Genomic Approaches to Virus Detection and Discovery in Acute Pediatric Illness in Nicaragua

© 2011

by Nathan Lawrence Yozwiak
Abstract
Genomic Approaches to Virus Detection and Discovery in Acute Pediatric Illness in Nicaragua

by

Nathan Lawrence Yozwiak

Doctor of Philosophy in Infectious Disease and Immunology

University of California, Berkeley

Professor Eva Harris, Chair

Monitoring the global emergence and spread of novel human pathogens is of vital interest to public health efforts aimed at preventing and treating infectious disease. Viral surveillance studies yield important information about the prevalence and seasonality of circulating virus species and the possible existence of novel species, yet many large screening studies that have been conducted in temperate countries have not been repeated in tropical settings. This thesis describes experiments aimed at expanding our understanding of the virus diversity associated with acute pediatric illness in a tropical environment using comprehensive genomic detection strategies.

Enteroviruses (*Picornaviridae* family) are a common cause of human illness worldwide and are associated with diverse clinical syndromes, including asymptomatic infection, respiratory illness, gastroenteritis, and meningitis. In Chapter 2, we report the identification and complete genome sequence of a novel enterovirus (EV109) isolated from a case of acute respiratory illness in a Nicaraguan child. Unbiased deep sequencing of nucleic acid from a nose and throat swab sample enabled rapid recovery of the full-genome sequence. Phylogenetic analysis revealed that EV109 is most closely related to serotypes of enterovirus species C (HEV-C) in all genomic regions except the 5´-untranslated region (5´UTR). Bootstrap analysis indicates that the 5´ UTR of EV109 is likely the product of an inter-species recombination event between ancestral members of the HEV-A and HEV-C groups. Overall, the EV109 coding region shares 67-72% nucleotide sequence identity with its nearest relatives. EV109 isolates were detected in 5/310 (1.6%) of nose and throat swab samples collected from children in a pediatric cohort study of influenza-like illness in Managua, Nicaragua between June 2007 and June 2008. Further experimentation is required to more fully characterize the pathogenic role, disease associations, and global distribution of EV109.
Dengue virus is re-emerging viral disease that infects an estimated 50-100 million people annually worldwide, yet current diagnostic practices cannot detect an etiologic agent in ~40% of dengue-like illness cases. Metagenomic approaches to pathogen detection, such as viral microarrays and deep sequencing, are promising tools to address emerging and non-diagnosable disease challenges, but have not been fully utilized in cases of tropical illness. In Chapter 3, we used the Virochip microarray and deep sequencing to characterize the viral spectrum in human sera from Nicaraguan patients presenting with dengue-like symptoms but testing negative for dengue virus. We utilized a barcoding strategy to simultaneously sequence multiple serum specimens, generating an average of 5.1 million reads per sample. We then implemented a stepwise bioinformatic filtering pipeline to remove the majority of human and low quality sequences to improve the speed and accuracy of subsequent GenBank searches. Virus sequence was detected in 35% (12/34) of previously negative cases using deep sequencing, including six samples with Human Herpesvirus 6 sequence and five samples containing sequence from a putative novel virus related to the Circoviridae family. These results demonstrate the utility of two metagenomic strategies, the Virochip and deep sequencing, as comprehensive platforms to detect known and divergent viruses in the study of tropical febrile illness.
Dedication

For my parents, Lawrence and Donna Yozwiak
# Table of Contents

Dedication.................................................................................................................................i

Table of Contents ..................................................................................................................ii

List of Figures..........................................................................................................................iv

List of Tables ..............................................................................................................................v

Acknowledgements ...............................................................................................................vi

**Chapter 1** ............................................................................................................................ 1

**Introduction** ......................................................................................................................... 1

1.1 Historical perspective of emerging infectious diseases ................................................. 1

1.2 Traditional Viral Diagnostics ..................................................................................... 3

1.3 Viral Microarrays .......................................................................................................... 3

1.4 Deep Sequencing Metagenomics ................................................................................ 4

1.5 Viruses in acute respiratory tract infections ............................................................. 6

1.6 Dengue virus infection ............................................................................................... 7

1.7 Thesis ............................................................................................................................ 8

**Chapter 2** ............................................................................................................................ 10

**Human Enterovirus 109: a novel inter-species recombinant enterovirus discovered in acute pediatric respiratory illness in Nicaragua** ............................................. 10

2.1 Introduction .................................................................................................................. 10

2.2 Experimental Procedures ......................................................................................... 12

2.3 Results ......................................................................................................................... 16
List of Figures

Figure 1.1 Virus detection pipeline. ................................................................. 5

Figure 1.2 Agents associated with cases of acute pediatric influenza-like illness in temperate regions. .......................................................................................... 6

Figure 2.1 Viral recovery workflow .................................................................. 14

Figure 2.2 Deep sequencing results and picornavirus genome coverage ............ 17

Figure 2.3 Relationships between known enteroviruses and EV109 based on full-length genome analysis...................................................................................... 20

Figure 2.4 Sequence analysis and RNA folding predictions reveal conserved picornavirus features in EV109 .................................................................................. 22

Figure 2.5 The EV109 5’UTR region reveals evidence of an ancestral recombination event. ........................................................................................................ 24

Figure 2.6 Amino acid conservation on the enterovirus capsid pentamer subunit. ...................................................................................................................... 25

Figure 2.7 Geographic locations of the five EV109-positive patient households... 26

Figure 3.1 Virochip sensitivity using poliovirus control RNA. ........................... 36

Figure 3.2 Deep sequencing initial reads............................................................. 37

Figure 3.3 Bioinformatic filtering of deep sequencing data. ............................. 38

Figure 3.4 HHV-6B genome coverage across six samples................................. 42

Figure 3.5 Circovirus-like NI sequence coverage and phylogeny ..................... 43
List of Tables

Table 1.1 Recently discovered viruses associated with acute respiratory infections / respiratory distress.................................................................2
Table 1.2 Laboratory confirmed cases from Dengue Pediatric Hospital Study in Managua, Nicaragua.............................................................8
Table 2.1 Deep-sequencing library preparation primers. .................................13
Table 2.2 List of primers used for EV109 sequencing and genome recovery.......18
Table 2.3 Clinical and demographic information from five EV109-positive patients, including patient age, date of illness onset, and symptoms..........26
Table 3.1 Summary of viruses identified on Virochip microarrays. ..................35
Table 3.2 Post-filtering reads per barcode.....................................................39
Table 3.3 Virus sequence hits. .....................................................................41
Table 3.4 Clinical information of patients with detectable virus sequence.........41
Table 4.1 EV109 screening...........................................................................48
Acknowledgements

I would first like to thank Jasmine Hanifi, for her love, support, and confidence.

I am grateful to my graduate advisor, Joe DeRisi, for keeping the excitement up about our projects and their potential. And for showing me how to make sweet slides in Illustrator.

I thank my graduate advisor, Eva Harris, for helping me maintain a broad scientific perspective.

I thank my committee members, George Sensabaugh and Jack Colford for their constructive advice and helpful conversations.

To Peter Skewes-Cox, thanks for keeping the band together.

I would like to thank the DeRisi lab post-doctoral mentors that kept me on course, especially Graham Ruby, Sarah Arron, Amy Kistler, Kael Fischer, Charles Chiu, and Patrick Tang.

I am grateful to all of the other fun members of the DeRisi lab. I have been incredibly fortunate to have worked alongside so many talented minds and entertaining friends.

I am thankful for the helpful clinical advice and direction of Don Ganem.

I am grateful to members of the Harris Lab, especially Aubree Gordon, Karen Clyde, and Molly OhAinle.

I also wish to acknowledge all of our partners in Nicaragua, especially Angel Balmaseda, Roger López, Juan Carlos Mercado, and Christian Cerpas Cruz.

To my high school and college mentors, Paul Nale and Joe Mankowski, for encouraging my nascent interest in biology and the world of scientific research.

To all my friends, thank you for the healthy distractions these past several years.

To my extended family, sisters Andrea and Ali, Grandmother, uncles, aunts, and cousins. Your support is invaluable.

And finally, I would like to thank my parents, Lawrence and Donna Yozwiak, for instilling in me a curiosity about the world.
Chapter 1

Introduction

This thesis describes experiments focused on detecting and discovering known and novel viruses in cases of acute pediatric illness in Nicaragua using viral microarrays and next-generation sequencing technology. Chapter 1 presents background information on traditional viral detection methods and an overview of viral microarray and deep sequencing approaches to pathogen detection and discovery. Chapter 2 describes experiments that reveal the discovery of a novel enterovirus isolated from cases of acute pediatric respiratory illness. The phylogenetic relationship, genomic features, recombination evidence, and clinical cases of Enterovirus 109 are described. Chapter 3 explores the results of experiments seeking to employ deep sequencing and viral microarray technology as a diagnostic tool in a study of unknown tropical febrile illnesses. Chapter 4 offers some concluding perspectives and future directions of genomic approaches to viral diagnostics.

1.1 Historical perspective of emerging infectious diseases

A series of public health triumphs in the first half of the twentieth century fueled a widely held optimism within the microbiology community that infectious diseases were conquerable and could be easily eliminated through human progress. Indeed, the successes of modern sanitation, early antibiotics, DDT to fight malaria, and the smallpox vaccine fueled hubristic confidence within the field; rather than simply control microbial pathogens, a more realistic objective should be complete eradication (69). Overconfidence bred complacency. By the 1960s the belief that infectious diseases had been conquered was widespread among scientific leaders (48) and led to a decline in funding, vigilance, and public health preparedness. Fueling the zeitgeist were misplaced evolutionary assertions that posited a microbial world that was slowly evolving or static and was attached to assumptions that nature was fundamentally benign. Long-term evolution was thought to favor harmless commensalism; in this context, virulence was accidental (138).

The decisive turning point in modern infectious disease awareness occurred in the 1980s during the appearance and subsequent pandemic of the human immunodeficiency virus and AIDS. By 1992, the official alarm bell sounded from the influential Institute of Medicine (IOM) report “Emerging Infections” which warned, “infectious diseases remain the major cause of death worldwide and will not be conquered during our lifetimes” (75). Policy redirection began. The Clinton White House responded with a Presidential Directive proposing funding increases to strengthen the global infectious disease surveillance and response system (1). Nobel laureate Joshua Lederberg proposed the term ‘emerging diseases’ to describe “diseases of infectious origin whose incidence in humans has increased within the past two decades or threatens to increase in the near future” (34). NIH funding doubled between 1998 and 2003.
Accompanying public health anxieties was a growing awareness of the causative factors of emergence and the realization that human actions have altered infectious disease dynamics (75, 80, 102). One commonly theorized cause faulted the impact of globalization and the rapid movement of people, resulting in the exposure of non-immune populations. Additional theorized factors paradoxically blamed medical breakthroughs themselves, such as microbial resistance due to antibiotic misuse, and growing numbers of immunocompromised patients due to life expectancy increases. Commonly, the rise in emerging infectious diseases was thought to be the result of the post-industrial population explosion (138). Extreme growth had led to unplanned urbanization, dense crowding, and environmental destruction, factors that increase host exposure. Social inequalities and escalating poverty resulting from uncontrolled population growth were also understood to play direct roles in infectious disease emergence and spread (47).

Currently, it is accepted that emerging diseases and novel human viruses arise not through purely evolutionary or virological mechanisms, but also due to one or more additional co-factors resulting from modern societal changes. Emerging infectious diseases have increased significantly in the past several decades, with over 300 reported separate events between 1940 and 2004. Many of these events were caused by zoonoses originating in wildlife in hotspots within lower-latitude developing countries (80). Dozens of newly described viruses, many of which are associated with significant clinical pathologies, have been recently discovered, owing both to societal factors favoring their emergence and technological advances confirming their existence (40). At least eleven novel viruses associated with acute respiratory tract infections or severe respiratory distress in humans have been discovered since 1993 (Table 1.1). Increasingly, molecular techniques such as large-scale sequencing, viral microarrays, and polymerase chain reaction (PCR) methods have been utilized for novel virus identification and classification.

Table 1.1 Recently discovered viruses associated with acute respiratory infections / respiratory distress.

<table>
<thead>
<tr>
<th>Virus</th>
<th>Family</th>
<th>Discovery method</th>
<th>Year</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARS coronavirus</td>
<td>Coronaviridae</td>
<td>microarray</td>
<td>2003</td>
<td>Wang et al., 2003; Rota et al., 2003.</td>
</tr>
<tr>
<td>Coronavirus NL-63</td>
<td>Coronaviridae</td>
<td>large-scale sequencing</td>
<td>2004</td>
<td>Van der Hoek, et al., 2004.</td>
</tr>
<tr>
<td>Coronavirus HKU1</td>
<td>Coronaviridae</td>
<td>degenerate PCR</td>
<td>2005</td>
<td>Woo, et al., 2005.</td>
</tr>
</tbody>
</table>

The present challenges posed by emerging infectious diseases require increasingly rapid, cheap, and sensitive diagnostic techniques that can respond to divergent and unencountered viral
pathogens. To better understand the molecular methods used in this dissertation, it is first important to briefly review some traditional methods of viral detection.

1.2 Traditional Viral Diagnostics

Clinical and research laboratories employ viral diagnostics to inform patient antiviral therapy and antibiotic treatment decisions, to study emerging and nosocomial disease epidemiology, and to inform public vaccine policy. Detection of viruses in clinical samples has historically fallen to the cornerstone methods of traditional diagnostic virology, including virus isolation, antigen detection, serology, and electron microscopy (141).

While these methods still comprise the principal repertoire of techniques used to study molecular virology, many possess inherent limitations that constrain their diagnostic effectiveness. Cell culture, for example, can only be used to detect and study viruses with known culture growth conditions, and cannot be used to detect many viruses that are fastidious or unculturable. No isolated or engineered cell line currently exists that can grow all medically relevant human viruses (140). Antigen detection, which includes direct fluorescent antibody (DFA) staining, is widely used to diagnose pathogens, including respiratory viruses such as influenza and respiratory syncytial virus, but these techniques frequently suffer low sensitivity compared to molecular methods (30, 53). Serology is an important tool used to investigate host immune responses to a viral infection by detecting antiviral antibodies and can also be used to detect asymptomatic infections, but the method is hindered by the requisite collection of both acute and convalescent serum samples (141). Electron microscopy requires that a virus be present at high titers in a clinical sample and only provides morphology information (122). The external shape of a virus often cannot inform differences between virus family members.

Utilization of traditional virus diagnostic techniques is being altered as technical developments advance the effectiveness of nucleic acid molecular detection tools. Polymerase chain reaction-based methods achieve sensitive and direct nucleic acid target amplification and identification and can be used to target known agents (103). PCR is simple, broadly applicable, and can target conserved regions to amplify related viruses; however, viruses lack universally conserved regions, such as ribosomal RNA, that can be exploited to amplify all viruses (121). Moreover, PCR is failure-prone due to sequence mismatches and cannot be easily used alone to detect novel or highly divergent viruses for which there exists no a priori sequence information. Additional, more comprehensive molecular approaches have been developed, such as screening cDNA libraries, subtractive hybridization, and representational difference analysis (RDA), but these methods are time and labor-intensive and not cost-effective for large studies (83).

1.3 Viral Microarrays

DNA microarrays have emerged as a method for viral detection that seeks to overcome limitations in other traditional and molecular approaches. The Virochip (DeRisi Lab, UCSF) represents a comprehensive, rapid, unbiased platform for viral detection and discovery (151, 152). The Virochip utilizes DNA microarray technology in which pathogen nucleic acid present in a sample is detected by hybridization to corresponding oligonucleotide (oligo) probes on the microarray. The current array platform is composed of 16,000 70-mer oligos derived from over
400,000 GenBank records (January 2007) providing sequence coverage for over ~20,000 viral taxa. Family level oligos allow for detection of all known viruses within a given family, as well as unknown or unsequenced related viral family members. Genus and species level oligos on the array also permit identification at these taxonomic levels as well.

The Virochip has been used successfully to discover novel viruses and diagnose acute cases of respiratory failure caused by known pathogens in which conventional detection methods failed. In March 2003, the Virochip detected the presence of a previously uncharacterized coronavirus in a cell culture isolate from a SARS patient based on cross-hybridization to conserved astrovirus and coronavirus oligos (152). The array has also identified a novel retrovirus in prostate tumors from patients with a mutation in the RNase L gene (148) and a novel human cardiovirus in a nasal pharyngeal swab from a child with influenza-like illness (28). Additionally, the array has been used to correctly diagnose two cases of life-threatening acute respiratory illness when conventional clinical laboratory testing failed, in one case due to PCR primer point mutations (27) and in the other because the causative pathogen was not part of routine testing (29). More recently, Chiu et al. compared the Virochip’s performance to direct fluorescent antibody (DFA) and PCR in detecting respiratory viruses in 278 nasopharyngeal aspirate samples from 222 children with acute respiratory tract infection and found the array performed better than DFA (19% increase in the detection of seven respiratory viruses) and similar to virus-specific PCR (sensitivity 85-90%, specificity >99%) (30). Array-based screening has also uncovered unexpected viral diversity in known respiratory viruses. In naturally acquired upper respiratory infections, Kistler et al. used the array to identify a highly divergent and uncharacterized clade of unculturable human rhinoviruses and to detect human coronavirus variants less commonly found in respiratory illnesses (85). Thus, of vital interest to the public’s health, the Virochip has a role in viral detection and discovery relevant to monitoring the emergence and spread of existing and novel human pathogens (Figure 1.1).

1.4 Deep Sequencing Metagenomics

High-throughput second-generation sequencing technologies that facilitate ‘deep’ sequencing are undergoing rapid evolution and increased affordability, resulting in a more routine use of genomic approaches to investigate scientific questions (11, 116). These sequencing tools offer improved speed, throughput, and cost per base compared to the traditional Sanger sequencing method (127). Several next generation platforms are currently commercially available: 454 (Roche Applied Science), a pyrosequencing-based method; SOLiD (Applied Biosystems), a sequencing-by-ligation approach; and Illumina (Solexa), which uses reversible-terminator sequencing-by-synthesis chemistry. The experiments described in this work utilized the Illumina (Solexa) Genome Analyzer II (GAII) platform (16). Currently, the Illumina GAII sequencer can generate over 300 million sequence reads per run and greater than 10 gigabases of sequence. Additionally, paired-end sequencing can generate reads 50-100 bases long from opposing ends of DNA molecules.

High throughput second-generation sequencing technologies have spurred the rapid pioneering of metagenomic science (the study of many genes and genomes within complex genetic pools), improved the speed and ease of genome resequencing, and greatly impacted the field of virus discovery (144). In the past two decades, improvements in sequencing technology have led to a rapid growth of available virus sequences and genomes in the viral sequence
Database (93). Recovering the genome of a previously uncharacterized virus has been historically accomplished by screening bacterial or phage libraries, primer-walking, and using PCR with degenerate primers. By contrast, deep sequencing of clinical samples generates hundreds of thousands to millions of sequencing reads per sample, which can then be incorporated into stepwise virus detection pipelines (Figure 1.1). Database searches using basic local alignment search tool (BLAST) variants (6) can be used to detect sequences in samples that correspond to known and novel viruses, including those present in exceedingly low titers or viruses that may be too divergent to be detected with PCR or microarray methods. Recent examples of novel viruses identified in human samples using deep sequencing include a polyomavirus associated with Merkel cell carcinoma (49) and a hemorrhagic fever virus detected in South Africa (22). Genomic approaches to pathogen detection using deep sequencing methods also increasingly recognized as important components of broad emerging disease monitoring and public health preparedness (41, 54, 60, 142).

Figure 1.1

Figure 1.1 Virus detection pipeline.
Staged strategy for virus detection and discovery in clinical samples using molecular diagnostics, including PCR, viral microarray, and deep sequencing. The end stage goal is virus sequence recovery. + denotes positive result and – denotes negative result.
1.5 Viruses in acute respiratory tract infections

Acute respiratory tract infections (ARTI) account for the majority of acute illnesses at all ages worldwide and represent the leading killer of children under five, with the incidence and mortality burden heavily skewed to the developing world (15, 160). The large majority of ARTIs in children (70-90%) are caused by viruses (Figure 1.2), including: respiratory syncytial virus (RSV), influenza virus A and B, parainfluenza virus types 1-4, picornaviruses, adenoviruses, and human metapneumovirus (32, 55, 77, 101, 154, 163). Recent advances in molecular diagnostic assays, such as PCR and RT-PCR, have improved the detection rate of infectious agents (77, 97, 143). Most studies attempting to define the incidence and burden of respiratory viruses in tropical environments, however, have relied primarily on traditional methods such as virus isolation and immunofluorescence, and have not been repeated using more sensitive and specific molecular techniques (132). As a result, the burden and incidence of specific viral respiratory pathogens is not well defined in tropical environments such as Nicaragua. Correctly identifying the agents associated with respiratory infections is critical for guiding policy, treatment, and prevention strategies. Currently, no routinely used single method is able to rapidly and simultaneously detect all known viral respiratory pathogens. Even with the best current viral detection methods, a specific pathogen cannot be identified in 13-50% of ARTIs (Figure 1.2)(32, 55, 77, 101, 154, 163).

Figure 1.2 Agents associated with cases of pediatric influenza-like illness in temperate regions

![Pie chart showing distribution of agents associated with pediatric ILI](image)

**References:** Watts et al., 2003; Monto et al., 2002; Van Gageldonk-Lafeber et al., 2005; Jennings et al., 2004; Zambon et al., 2001.

**Figure 1.2** Agents associated with cases of acute pediatric influenza-like illness in temperate regions.

Approximations for the detection of etiologic agents in ILI cases based on several studies (154, 101, 55, 77, 163). No pathogen can be detected in a sizeable portion of ILI cases in all studies of respiratory tract infections.
Monitoring the emergence and spread of novel human respiratory pathogens is of vital interest to public health. Several new and clinically important respiratory viruses have been identified in the last several years, including human metapneumovirus, human bocavirus (hBoV), and three novel coronaviruses (CoV; Table 1.1) (71, 72, 125, 152, 158). Human metapneumovirus has been implicated as a highly prevalent cause of childhood respiratory infections (52), and SARS-CoV caused a pandemic from November 2002 to July 2003 resulting in over 700 deaths at a 9% mortality rate (161). Most of these clinically relevant viruses were newly discovered using molecular techniques, such as degenerate PCR (CoV HKU1), high-throughput sequencing (CoV NL63 and hBoV), or microarrays (SARS CoV; Table 1.1). Even using the most sensitive and current viral detection methods, a specific agent cannot be identified many ARTI cases (32, 55, 77, 101, 154, 163). Furthermore, many viruses possess known mechanisms of generating genetic variation, among these are error-prone polymerases, high replication rates and frequent recombination (133, 157). Global population expansion, increased urbanization, and global travel add to viral migration/gene flow and cross-species transfer events. As a result, it is likely that some previously undiagnosable RTIs are due to novel or highly divergent viruses. Additionally, differences in climate, agriculture, living conditions, and human-animal interactions may generate a different spectrum of disease-causing pathogens in tropical nations. While many large epidemiologic studies have been conducted in temperate North America and Europe, similar studies have not been repeated in tropical nations, leaving information about the burden and seasonality of viral disease incomplete in those settings. Recently developed unbiased methodologies, such as deep sequencing, pan-viral microarrays, and rapid-multiplex polymerase chain reaction, have expanded knowledge of infectious agents and improved viral detection capabilities, making large-scale broad-spectrum viral surveys and novel viral discovery more feasible than ever before (38, 40, 64, 93, 144).

1.6 Dengue virus infection

Dengue virus (DENV) infection is the most common arthropod-borne viral disease of humans and is a re-emerging infectious disease with a rapidly expanding geographic range. An estimated 3 billion people living in tropical and subtropical regions are at risk for dengue infection, and an estimated 50 million cases and 500,000 hospitalizations occur annually worldwide. The spectrum of dengue illness, including dengue fever and the more severe dengue hemorrhagic fever/dengue shock syndrome (DHF/DSS) is caused by four serotypes of dengue virus (DENV 1-4) of the Flaviviridae family and is spread by the bite of the Aedes mosquito vector (159). Diagnosing dengue relies on detecting viral nucleic acid or antigens in the blood or confirming the presence of anti-DENV IgM and IgG antibodies and therefore traditionally depends on ELISA, RT-PCR, and viral cell culture methods (67, 89).

Dengue diagnostics are of crucial importance due to its global emergence and spread, unique disease epidemiology, and clinical relation to other as-yet unknown tropical febrile pathogens. The global spread of dengue virus and its mosquito vectors over the past several decades, especially in Latin America, has illuminated the risk of widespread dengue appearance in previously dengue-free subtropical regions, including the southern United States, and has necessitated additional methods of viral surveillance (89). Dengue detection is also important to monitor secondary infections, because successive DENV infections with different viral serotypes
is a major risk factor for the more severe DHF/DSS (128, 146). As a result of the severe disease risk associated with secondary heterologous DENV infections, current dengue vaccine initiatives aim to provide cross-protection to all four serotypes. The existence of a related, but unknown cross-reactive DENV serotype could have momentous effects on dengue vaccine efforts. Sentinel discovery and characterization efforts would also be critical to better understand the etiologic agents associated with acute tropical febrile illnesses. Currently available laboratory diagnostic methods cannot detect DENV in 30-50% of pediatric patients presenting with hallmark dengue symptoms (Table 1.2).

Several newly described flaviviruses (T’Ho Virus, Nounane Virus, Lammi Virus, and Nakiwogo Virus) have been recently discovered in tropical mosquitoes and underscore the possible uncharacterized Flaviviridae family diversity (31, 46, 74, 82). The human clinical associations of these newly described viruses have yet to be determined, but their detection in insect vectors further illustrates the importance of virus discovery efforts directed towards emerging undiagnosed tropical illnesses.

### Table 1.2 Laboratory-confirmed cases from the Dengue Hospital Study in Managua, Nicaragua

<table>
<thead>
<tr>
<th>Year</th>
<th>DENV-1</th>
<th>DENV-2</th>
<th>DENV-3</th>
<th>DENV-4</th>
<th>Pos for DENV</th>
<th>Other</th>
<th>Neg for DENV</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aug 2005 - Jul 2006</td>
<td>9</td>
<td>43</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>6</td>
<td>30</td>
<td>101</td>
</tr>
<tr>
<td>Aug 2006 - Jul 2007</td>
<td>2</td>
<td>36</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>4</td>
<td>78</td>
<td>128</td>
</tr>
<tr>
<td>Aug 2007 - Jul 2008</td>
<td>5</td>
<td>61</td>
<td>13</td>
<td>0</td>
<td>6</td>
<td>5</td>
<td>61</td>
<td>151</td>
</tr>
<tr>
<td>Aug 2008 - Jul 2009</td>
<td>18</td>
<td>13</td>
<td>99</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>64</td>
<td>198</td>
</tr>
<tr>
<td>Aug 2009 - Nov 2009</td>
<td>5</td>
<td>6</td>
<td>66</td>
<td>0</td>
<td>2</td>
<td>8</td>
<td>54</td>
<td>141</td>
</tr>
<tr>
<td>TOTALS</td>
<td>39</td>
<td>159</td>
<td>179</td>
<td>0</td>
<td>32</td>
<td>23</td>
<td>287</td>
<td>719</td>
</tr>
</tbody>
</table>

Samples were collected from patients 0-14 years old presenting with suspected dengue (120) at the Hospital Infantil Manuel de Jesús Rivera (HIMJR) in Managua, Nicaragua, and tested with DENV-specific RT-PCR, ELISA, and cell culture (Harris Lab and Centro Nacional de Diagnóstico y Referencia of the Nicaraguan Ministry of Health, unpublished). The number of suspected dengue clinical cases per year that were negative for dengue virus are reported in the red column.

### 1.7 Thesis

This thesis describes experiments aimed at expanding our understanding of the virus diversity associated with acute human illness in a tropical environment. We have discovered and fully sequenced the genome of a novel enterovirus (EV109) that was detected in five children with acute respiratory illness in Managua, Nicaragua. Prior to the work described in this thesis, the existence of EV109 and its sequence was unknown. We used a combined strategy of deep sequencing and polymerase chain reaction methods to detect and fully sequence the virus genome to understand its relationship within the Enterovirus genus of the Picornaviridae family. Genome analysis also revealed evidence of an ancestral recombination event and the conservation of key RNA secondary structures. We developed an EV109-specific PCR assay and
detected the virus in five cases of pediatric respiratory illness in Nicaragua, though we concede the role of disease causality for EV109 remains unknown. Chapter 2 of this thesis describes how we detected and sequenced EV109 and details its phylogenetic relationships, predicted genomic features, and clinical case distribution.

The remainder of the thesis describes experiments aimed at understanding the virus diversity associated with unknown dengue-like illness in Nicaragua. We used deep sequencing to analyze the nucleic acid from acute serum samples collected from 34 children with dengue-like symptoms who tested negative for dengue virus by multiple laboratory procedures. Prior to this work, deep sequencing approaches to viral diagnostics have been limited and have not been performed on human serum from tropical febrile illness cases. In Chapter 3, we present the deep sequencing results demonstrating the presence of nucleic acid from known viruses in serum, including six samples with Human Herpesvirus 6. Additionally, we report the presence of multiple serum samples containing divergent torque teno virus (TTV) sequence and sequence from a novel Circovirus. Our results expand our understanding of the existing and rapidly changing virus spectrum associated with human illness in an urban tropical setting.
Chapter 2

Human Enterovirus 109: a novel inter-species recombinant enterovirus discovered in acute pediatric respiratory illness in Nicaragua

This chapter describes experiments that report the identification and complete genome sequence of a novel enterovirus isolated from a case of acute respiratory illness in a Nicaraguan child. A paper describing some of these experiments has been published in the *Journal of Virology* (162).

Enteroviruses (*Picornaviridae* family) are a common cause of human illness worldwide and are associated with diverse clinical syndromes, including asymptomatic infection, respiratory illness, gastroenteritis, and meningitis. Unbiased deep sequencing of nucleic acid from a nose and throat swab sample enabled rapid recovery of the full-genome sequence. Phylogenetic analysis revealed that EV109 is most closely related to serotypes of enterovirus species C (HEV-C) in all genomic regions except the 5´-untranslated region (5´UTR). Bootstrap analysis indicates that the 5´ UTR of EV109 is likely the product of an inter-species recombination event between ancestral members of the HEV-A and HEV-C groups. Overall, the EV109 coding region shares 67-72% nucleotide sequence identity with its nearest relatives. EV109 isolates were detected in 5/310 (1.6%) of nose and throat swab samples collected from children in a pediatric cohort study of influenza-like illness in Managua, Nicaragua between June 2007 and June 2008. Further experimentation is required to more fully characterize the pathogenic role, disease associations, and global distribution of EV109.

2.1 Introduction

The genus *Enterovirus* in the family *Picornaviridae* is a group of related viruses that are associated with a spectrum of disease, ranging from subclinical infections to acute respiratory and gastrointestinal illness to more severe manifestations, such as aseptic meningitis, encephalitis, and acute flaccid paralysis (76, 115). Enteroviruses are small, non-enveloped viruses that share a common genomic organization. The RNA genome is a ~7.5 kb single-stranded, positive-sense, poly-adenylated molecule, with a single, long open-reading frame flanked by 5´ and 3´ untranslated regions (UTRs). The 5´UTR is ~700 nucleotides in length and contains highly structured secondary elements with internal ribosomal entry site (IRES) function. The ~2200 amino acid (aa) polyprotein is co-translationally processed by viral proteases to yield structural (VP4, VP2, VP3 and VP1) and non-structural (2A, 2B, 2C and 3A, 3B, 3C, and 3D) proteins (115). Current enterovirus classification is based on the high sequence divergence within the VP1 capsid region, which has been shown to correspond with serotype neutralization (108, 109). Human enterovirus types are currently classified into four species, *Human enterovirus A* (HEV-A), HEV-B, HEV-C (including poliovirus), and HEV-D, based on the four
phylogenetic clusters observed when comparing coding region sequence. An enterovirus is designated as a new type within a species if it possesses <75% nucleotide and <85% aa identity with known members across the VP1 sequence (108, 111). Molecular identification methods play a crucial role in rapid, sensitive enterovirus diagnostics and have led to the recent discovery of several novel enteroviruses (110, 112, 135, 136, 145). Most approaches target a limited number of conserved regions in the 5’UTR and VP4-VP2 junction or seek to ascertain serotype information by probing antigenic regions, such as VP1 (20).

Picornavirus RNA-dependent RNA polymerases are highly error-prone and lack proofreading ability, resulting in a misincorporation frequency of one per $10^3$-$10^4$ nucleotides (157). The relative infidelity of these polymerases is believed to enable rapid adaptability under selective pressure. Large impact evolutionary events, such as recombination within and between enterovirus serotypes, also contributes to their evolution and genetic diversity (10, 26, 107, 134) and may lead to changes in disease associations with human enterovirus infections. Human enteroviruses are classified into four species based on coding region sequence phylogeny, and intra-species recombination events between enteroviruses that are closely related in the coding region are well documented (107, 129, 134). All known enterovirus 5’UTR region sequences, however, cluster into two groups, one containing HEV-A and B sequences and the other with HEV-C and D sequences. Recent findings have described enterovirus genomes with a coding region that clusters with one species and a 5’UTR region that clusters with a different species, suggesting possible inter-species recombination events (137, 145). Understanding the recombination-driven evolution of HEV-C viruses is of particular public health concern due to the viruses’ ability to recombine with vaccine poliovirus resulting in circulating, highly neurovirulent vaccine-derived polioviruses (78, 94, 118). It is unclear whether recombination events between poliovirus and HEV-C viruses allow for the rapid acquisition of traits that increase pathogenic and circulation potential.

The enterovirus pathogenicity spectrum is related to tissue tropism and largely determined by cellular receptor usage. Most picornaviruses use receptors from the immunoglobulin superfamily of proteins, such as intracellular adhesion molecule-1 (ICAM-1) or coxsackievirus-adenovirus receptor (CAR; (124)). A distinct subgroup of HEV-C viruses, which includes coxsackievirus A1, A19, A22 and enterovirus 104, has not yet been grown successfully in cell culture, and the receptor molecule for this subgroup is unknown (23). HEV-C viruses are believed to be the ancestral source of poliovirus, which resulted from a capsid mutation that caused a cellular receptor switch from ICAM-1 to CD155 (Poliovirus receptor; PVR; (78)).

In this study, we report the discovery and characterization of a novel human enterovirus type within species HEV-C, for which we propose the designation Human Enterovirus 109 (EV109). Sequence analysis reveals considerable nucleotide divergence in the 5’UTR between EV109 and other HEV-C types, and scanning bootstrap analysis supports the hypothesis that EV109 is the product of an inter-species recombination event with an ancestral member of the HEV-A group. Viral capsid amino acid alignments and homology modeling reveal the predicted three-dimensional arrangement of divergent and conserved residues of EV109 compared with other related enteroviruses. We also report highly similar EV109 isolates within multiple cases of acute pediatric respiratory illness in Managua, Nicaragua.
2.2 Experimental Procedures

Nicaragua Influenza Cohort Study

The Nicaraguan Influenza Cohort Study is a prospective community-based cohort study of viral respiratory illness in ~3,800 children aged 2-13 years in Managua, Nicaragua (58, 88). Study enrollment began June 1, 2007. At enrollment, participants’ families agreed to bring their children to the study health center at the first sign of illness. The study provides medical care free of charge to participants, and data from all medical visits are recorded systematically. Nose and throat swabs are collected from a 25% random sample of patients presenting with influenza-like illness (ILI), as defined by exhibiting fever or history of fever with a cough and/or sore throat with symptom onset within the previous four days. Study participants typically present early in illness, increasing the likelihood that a child will be shedding virus at presentation (92% of participants with ILI presented within three days of symptom onset). At collection, nose and throat swabs are placed immediately into a tube containing 3 ml of viral transport media. Samples are stored at 4°C at the clinical laboratory until they are transported to the Nicaraguan National Virology Laboratory, where they are aliquoted and stored at -80°C.

Specimen nucleic acid extraction, library amplification, and Enterovirus PCR

RNA was extracted from 140 µl of viral transport medium containing the nose/throat swabs using the QIAamp Viral RNA Isolation Kit (Qiagen, Valencia, CA), and cDNA was randomly amplified using the round A/B protocol (153). Specific PCR was performed on cDNA libraries using primers targeting the HEV 5’UTR region (90) and the VP1 region (106). PCR reactions of 25 µl volume were performed using 17 µl water, 2.5 µl 10x Taq buffer, 1 µl MgCl₂ (50mM), 0.5 µl dNTPs (12.5mM), 0.5 µl of each primer (10 µM), and 0.5 µl of Taq polymerase (Invitrogen). Conditions for the 5’UTR reaction were 30 cycles of 94°C for 30 seconds, 50°C for 1 minute, and 72°C for 1 minute, and conditions for the VP1 reaction followed those previously described (106). Amplicons were cloned into plasmid vectors using the TOPO TA Cloning System (Invitrogen) and sequenced on an ABI 3130xl Genetic Analyzer (Applied Biosystems) in both directions using vector primers M13F and M13R and Big Dye terminator sequencing chemistry.

Sequencing library preparation

One paired-end ultra-deep sequencing run was performed using a GAIi Sequencer (Illumina; Figure 2.1). We utilized one lane of the run for the sample 4327 library. Library generation primers (Table 2.1) were modified from “adapter A” and “adapter B” sequences (Illumina). Nose/throat swab RNA extracted from sample 4327 was randomly primed and reverse transcribed using a 14-bp sequence common to the 3’-end of both Illumina adapters attached to a random hexamer (primer 1a). Second-strand synthesis was also primed using pr1a and followed by PCR amplification using the 14-bp common sequence without the hexamer (primer 1b) for 25 cycles. PCR products of ~300 bp were purified on a 4% native polyacrylamide gel (30mM KCl, 1X TBE buffer, 19:1 acrylamide:bis) run at 4°C, then ethanol precipitated and PCR amplified for 17 additional cycles using a 22 bp-long primer consisting of the 3’-end of Illumina adapter A (primer 2) and the full-length 61-bp Illumina adapter B (primer 4) under the following conditions: 2 cycles of 94°C for 30 seconds, 40°C for 30 seconds, and 72°C for 1 minute, followed by 15 cycles of 94°C for 30 seconds, 55°C for 30 seconds, and 72°C
for 1 minute. Amplicons generated with the correct adapter A-adapter B topologies were ~355 bp, and were purified away from adapter A-adapter A amplicons (~394 bp) and adapter B-adapter B amplicons (~316 bp) on a 4% native polyacrylamide gel run at 4°C, then ethanol precipitated. An additional 10 cycles of PCR were then performed on the adapter A-adapter B products using primer 3 and the 24 bases from the 5’-end of adapter B (primer 5).

**Table 2.1**

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>length</th>
<th>Sequence (5’-3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primer 1A</td>
<td>20</td>
<td>CGCTCTTCCGATCTNNNNNN</td>
</tr>
<tr>
<td>Primer 1B</td>
<td>14</td>
<td>CGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Primer 2</td>
<td>22</td>
<td>CTACACGACGCTCTTTCCGATCT</td>
</tr>
<tr>
<td>Primer 3</td>
<td>58</td>
<td>AATGATACGGCGACCACCGAGATCTACACTCTTTTCCCTACACGACGCTCTTTCCGATCT</td>
</tr>
<tr>
<td>Primer 4</td>
<td>61</td>
<td>CAAGCAGAAGACGCGATACGAGATCGGTTCCTCGGCATTCCCTGCTGAACCAGCTCTTTCCGATCT</td>
</tr>
<tr>
<td>Primer 5</td>
<td>24</td>
<td>CAAGCAGAAGACGCGATACGAGAT</td>
</tr>
</tbody>
</table>

**Table 2.1 Deep-sequencing library preparation primers.**
Primers used for deep sequencing library preparation were modified from adapter A and adapter B sequences (Illumina).
The method for viral genome recovery begins with RNA extraction from a nose/throat swab of clinical interest. A sequencing library is generated from the nucleic acid using reverse transcription and random-primed amplification to affix known sequencing adapter molecules necessary for the Illumina platform. The Illumina GAII sequencing is performed using the manufacturer protocol. Sequence analysis includes image analysis and basecalling software, read filtering using quality metrics, and iterative sequence alignments. Reads with similarity to viral sequence of interest are computationally assembled or used for primer design for subsequent specific amplification of remaining viral regions.
Deep sequencing analysis
The image analysis, base-calling, and sequence quality control for the sequencing run was analyzed using the Illumina pipeline (v1.3.2). A total of 10,412,905 sequencing read pairs, each pair consisting of 65 nucleotides, were obtained, for a total of 20,825,810 reads and approximately 1.4 gigabases of sequence (Sequence Read Archive: SRA012708.1).

The first six bases of each read, which correspond to the hexamer binding site of the random primer, were trimmed, leaving 61 nt-long reads. Reads with more than 10% of their bases uncalled (>6 N's) were removed. Reads were next subjected to a complexity filter in which sequences with LZW compression ratio below 0.45 were removed from the dataset (156). The remaining set of higher-complexity reads was aligned to the human genome (UCSC Build 18) using BLAT with the default parameters, and any read with 90% or more of its nucleotides matching identically to the human genome was removed along with its paired-end. The set of remaining reads was then aligned to the human genome using nucleotide BLAST with an E value of $10^{-3}$ and a word size of 20, and any read with 90% or more of its nucleotides matching identically to the human genome was removed along with its paired-end. The remaining quality-filtered, non-human reads were aligned to Haemophilus influenzae (gi: 148826757), Streptococcus pneumoniae (gi: 116515308), and Porphyromonas gingivalis (gi: 34398108) as sequences related to each of these bacteria were collected during quality validation of the sequence library. The alignment to these bacterial genomes employed nucleotide BLAST with an E value of $10^{-3}$ and a word size of 11, and in order to better filter the wide diversity of bacteria, any read with 70% or more of its nucleotides identically matching to one of the bacterial genomes was removed along with its paired-end. Finally, the remaining reads were aligned to NCBI’s NT database in an iterative fashion using nucleotide BLAST, first with an E value of $10^{-5}$ and a word size of 40, followed by an E value of $10^{-3}$ and a word size of 20, and finally with an E value of $10^{-3}$ and a word size of 10. Any query sequence producing alignments to only viral sequences with at least half of its nucleotides matching a known virus was considered viral.

Complete genome sequencing
Amplicons derived from specific EV109 PCR primers (Table 2.1) were gel-purified, cloned, and sequenced as described above. Rapid amplification of cDNA ends (RACE) was performed on RNA from sample 4327 to recover the 5’ and 3’ ends of the genome. The 5’ end was captured using the Invitrogen 5’-RACE kit (ver 2.0) according to the standard protocol. 3’-end recovery employed one round of RT-PCR using the Promega One-Step RT-PCR and AMV reverse transcriptase using primer “EV109 7014F” and an oligo-dT primer attached to a known “primer B” sequence (153) followed by one hemi-nested round of PCR using Taq polymerase (Invitrogen) with primer “EV109 7044F” and primer B. Genome sequence assembly of PCR amplicons, RACE-derived amplicons, and ultra-deep sequencing reads was generated with Geneious (ver 3.6.1) alignment tool using a 20-bp minimum overlap requiring a 95% overlap identity. The completed genome sequence of EV109 isolate 4327 was deposited into GenBank (accession number GQ865517) and reported to the Picornavirus Study Group (August 24, 2009).

EV109 PCR screening
Specific EV109 primers targeting the VP1 region were designed from recovered VP1 sequence (EV109 VP1 123F: 5’-GGAGACTGGAGCAACTAGTAAAG-3’ and EV109 VP1 363R: 5’-GGTGAACATTTCCAATTTCCCTACG-3’). PCR reactions of 25 µl were performed
on cDNA libraries using 17 µl water, 2.5 µl 10x Taq buffer, 1 µl MgCl₂ (50mM), 0.5 µl dNTPs (12.5mM), 0.5 µl of each primer (10µM), and 0.5 µl of Taq polymerase (Invitrogen). Conditions for the 5’UTR reaction were 30 cycles of 94°C for 30 seconds, 50°C for 1 minute, and 72°C for 1 minute. To obtain the full-length VP1 sequences and 5’UTR-VP4 junction from positive samples, PCR was performed using conserved primers flanking these regions, as determined from the full-length 4327 isolate. Amplicons of expected size were gel-purified, cloned, and sequenced as described above.

**Phylogenetic and bootscanning analysis**

Multiple complete and partial genome alignments were constructed using ClustalW (version 2.0.10), and phylogenetic trees were constructed using the neighbor-joining method (100 bootstrap replicates) with Mega software (v4.0). Bootscanning was performed using the Jukes and Cantor method with RDP3 (98) (window size, 400; step size, 20; 500 replicates).

**Comparative protein structure modeling**

EV109 nucleotides 664 to 7281 were translated in the reading frame inferred from all related enteroviruses. The inferred protein sequences homologous to other enterovirus proteins VP1, VP2, VP3, and VP4, were extracted for structure modeling of the viral capsid. These sequences were aligned to the sequence and structure of their coxsackievirus A21 (PDB code: 1Z7S) counterparts using the “align2d” function in MODELLER-9v7 (126). Homology models were built using these alignments and the crystallographic structure of the coxsackievirus A21 capsid proteins using the standard “automodel” routine of MODELLER-9v7.

**2.3 Results**

**2.3.1 Detection and full-genome sequencing of a new HEV-C type from a nose/throat swab of a patient with Influenza-like Illness**

We utilized samples collected in the Nicaraguan Influenza Cohort Study, a prospective community-based cohort study of viral respiratory illness in ~3,800 children aged 2-13 years in Managua, Nicaragua (58, 88). For this analysis, nose and throat swab specimens were examined from patients presenting with influenza-like illness (ILI), as defined by exhibiting fever or history of fever with a cough and/or sore throat with symptom onset in the previous four days. First, total RNA was extracted from the swab samples, reverse-transcribed to cDNA, and tested by RT-PCR for Influenza virus A and B, Parainfluenza virus 1-3, and Respiratory Syncytial Virus. Samples that were negative were then randomly amplified as previously described (153), and tested by PCR for rhinoviruses and enteroviruses using conserved and degenerate picornavirus primers (90, 106). As part of ongoing picornavirus typing studies, 5’UTR and VP1 amplicons of expected size were cloned and sequenced using a capillary electrophoresis sequencer (see Materials and Methods). In one sample, the amplified 5’UTR and VP1 regions identified only 79% and 73% BLASTn identity, respectively, to other known enteroviruses.

To recover additional viral sequence from the divergent sample, ultra-deep sequencing was performed using the GAI sequencer (Illumina, Inc.) using a random-primed cDNA library derived from nose/throat swab RNA from one patient, sample 4327 (Figure 2.1). The resulting 20.8 million 61-nt reads, (10.4 million paired-end reads), were filtered for N-content, LZW
complexity, and aligned to the human genome and three bacterial genomes using BLAT and BLAST, leaving 4.6 million sequencing reads (Figure 2.2A, inset). The remaining reads were aligned to NCBI’s NT database in an iterative fashion (see Materials and Methods). 186 reads were identified as viral sequence and 119 of these reads were tentatively assigned as picornavirus sequence by nucleotide BLAST analysis.

Aligning the 119 picornavirus reads to related enterovirus genomes revealed discontinuous regions of high read coverage distributed across the genome but were not sufficient to fully assemble the complete genome of the new enterovirus (Figure 2.2B). We used read-specific sequence data to design primers to close the remaining genomic gaps (Figure 2.2B, arrowheads; Table 2.2). The 3' end of the genome was recovered using rapid amplification of cDNA ends (3'RACE), and the 5' end was recovered using a hemi-nested 5'RACE strategy. The complete VP1 sequence and genome were submitted to the Picornavirus Study Group (http://www.picornastudygroup.com), compared to other enterovirus sequences, and named as a new proposed enterovirus type, Enterovirus 109 [EV109 (Genbank accession number: GQ865517); Nick Knowles, personal communication].

**Figure 2.2 Deep sequencing results and picornavirus genome coverage**

A. Coverage map of deep sequencing-derived picornavirus reads. Coverage map of deep sequencing-derived picornavirus reads and their adjusted EV109 genome positions before (blue dashed line) and after (orange solid line) full-genome sequencing. Inset: Summary of total 61-nt deep sequencing reads and results from the iterative alignment analysis. B. Schematic diagram of EV109 genome. The black triangles denote primer positions derived from deep sequencing reads used to recover the full-length EV109 genome. The black bars denote sequence regions recovered from all five EV109 isolates. The predicted EV109 genome domains are drawn to scale.
Table 2.2

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>length</th>
<th>start position</th>
<th>end position</th>
<th>Sequence (5'-3')</th>
</tr>
</thead>
<tbody>
<tr>
<td>EV109 5' RACE 268R</td>
<td>19</td>
<td>268</td>
<td>250</td>
<td>ATGGTGCTACTAGGCTTCC</td>
</tr>
<tr>
<td>EV109 5' RACE 274R</td>
<td>20</td>
<td>274</td>
<td>255</td>
<td>CAATCCATGGTGCTACTAGG</td>
</tr>
<tr>
<td>EV109 5' RACE 333R</td>
<td>17</td>
<td>333</td>
<td>317</td>
<td>TATTCCGAGCCTCCTCAG</td>
</tr>
<tr>
<td>EV109 5' UTR 492F</td>
<td>19</td>
<td>492</td>
<td>510</td>
<td>CAACACCGTAGCTGAGTGC</td>
</tr>
<tr>
<td>EV109 VP4 849R</td>
<td>21</td>
<td>849</td>
<td>829</td>
<td>AATCATTACATCTTTAACTGG</td>
</tr>
<tr>
<td>EV109 VP2 942F</td>
<td>20</td>
<td>942</td>
<td>961</td>
<td>TACTCAAGAAGAGCTAACG</td>
</tr>
<tr>
<td>EV109 VP3 1870F</td>
<td>20</td>
<td>1870</td>
<td>1889</td>
<td>GGGTGATGCTCTCTTTGTC</td>
</tr>
<tr>
<td>EV109 VP3 1899R</td>
<td>21</td>
<td>1889</td>
<td>1869</td>
<td>AACACAATGGAGGATACCC</td>
</tr>
<tr>
<td>EV109 VP1 2529F</td>
<td>23</td>
<td>2529</td>
<td>2551</td>
<td>GGAGACTGGAGCAACTGTAAG</td>
</tr>
<tr>
<td>EV109 VP1 2769R</td>
<td>24</td>
<td>2769</td>
<td>2746</td>
<td>GTGAAGACATTTCCCAATTTCTAG</td>
</tr>
<tr>
<td>EV109 VP1 3030R</td>
<td>21</td>
<td>3030</td>
<td>3010</td>
<td>GGAAAAAGCCTCATAAGAG</td>
</tr>
<tr>
<td>EV109 2A deg 3662F</td>
<td>19</td>
<td>3662</td>
<td>3680</td>
<td>TCACGCHGGKGGWGAAG</td>
</tr>
<tr>
<td>EV109 2A deg 3688R</td>
<td>17</td>
<td>3688</td>
<td>3704</td>
<td>GCHTTTGCGWACATYAG</td>
</tr>
<tr>
<td>EV109 2C 4255F</td>
<td>20</td>
<td>4255</td>
<td>4274</td>
<td>ACACGTGTTCAAACATCG</td>
</tr>
<tr>
<td>EV109 2C 4307F</td>
<td>18</td>
<td>4307</td>
<td>4324</td>
<td>CTTGTATGCGCAAGGG</td>
</tr>
<tr>
<td>EV109 2C 4324R</td>
<td>18</td>
<td>4324</td>
<td>4307</td>
<td>CCTCGCTGGCATACAG</td>
</tr>
<tr>
<td>EV109 2C 4513F</td>
<td>17</td>
<td>4513</td>
<td>4529</td>
<td>CCCCTGCACTTGGAG</td>
</tr>
<tr>
<td>EV109 2C 4916R</td>
<td>19</td>
<td>4916</td>
<td>4898</td>
<td>TGACCTTCAATAATTGATG</td>
</tr>
<tr>
<td>EV109 3C deg 5362F</td>
<td>20</td>
<td>5362</td>
<td>5381</td>
<td>GAYTAGAGCTGGAATAAGC</td>
</tr>
<tr>
<td>EV109 3C 5452R</td>
<td>18</td>
<td>5452</td>
<td>5435</td>
<td>CTCACCGTGCTAAATGC</td>
</tr>
<tr>
<td>EV109 3D 5932R</td>
<td>20</td>
<td>5932</td>
<td>5913</td>
<td>CGTTTTGTGCTGTTCTAC</td>
</tr>
<tr>
<td>EV109 3D 6286F</td>
<td>20</td>
<td>6286</td>
<td>6305</td>
<td>ATTCGAAACAGAAAAACAG</td>
</tr>
<tr>
<td>EV109 3' RACE 7014F</td>
<td>20</td>
<td>7014</td>
<td>7033</td>
<td>GTCTTTGAAACCGGTCTTTC</td>
</tr>
<tr>
<td>EV109 3' RACE 7044F</td>
<td>20</td>
<td>7044</td>
<td>7063</td>
<td>ACAATTCCACCAAATGG</td>
</tr>
<tr>
<td>EV109 3D 7090R</td>
<td>25</td>
<td>7090</td>
<td>7066</td>
<td>CATATCTGACATAGGCAAAACAG</td>
</tr>
</tbody>
</table>

Table 2.2 List of primers used for EV109 sequencing and genome recovery.

2.3.2 Genomic description of EV109

The complete genome of EV109 consists of 7,354 nucleotides, excluding the poly(A) tail. The 5’ UTR contains 663 nucleotides and the 3’ UTR consists of 73 nucleotides. EV109 features a single open reading frame from base 664 to 7281 that encodes a 2,206 amino-acid polyprotein. The base composition of the full genome is 27.7% A, 23.8% C, 24.1% G, and 24.3% U. To investigate the relationships among EV109 and other members of the Enterovirus genus, we compared the full genome of EV109 with representative members of different enterovirus species and constructed similarity plots (Figure 2.3A). The scanning pairwise nucleotide identity plots suggested that EV109 was most closely related to serotypes of the HEV-C species in all of the genome regions except the 5’UTR, which was most similar to HEV-A. Overall, EV109 shares only 67-72% nucleotide sequence identity with other HEV-C coding regions, including CAV 19 (71%), CAV 22 (70%), HEV 104 (72%), and Poliovirus 1 (67%). The VP1 capsid subunit region of the EV109 polyprotein is 276 amino acids, and by pairwise similarity, shares 66-71% aa similarity to CAV 22, CAV 19, and HEV 104 (64-65% nucleotide similarity).
Figure 2.3 Relationships between known enteroviruses and EV109 based on full-length genome analysis.

A. Full genome similarity plot depicting scanning pairwise identity using a 100-nt sliding window evaluated at each nucleotide. The EV109 sequence is compared with a close HEV-C relative, Coxsackievirus A19 (CAV19; red line) and more distant HEV-A (HEV 92; blue), and HEV-D (HEV 68; orange) serotypes. The conserved enteroviral domains are depicted to scale. The 5′UTR and VP1 regions are highlighted. B. Phylogenetic tree constructed from complete enterovirus genomes. The EV109 genome sequenced in this study is depicted with a solid black arrow, with Coxsackievirus A19 identified by a red arrow, Enterovirus 92 by a blue arrow, and Enterovirus 68 by an orange arrow. ClustalW and MEGA were used for alignments and tree construction, respectively, using the neighbor-joining method and 500 bootstrap replicates.

A phylogenetic tree constructed from full-length nucleotide sequences using the neighbor-joining method illustrates the four separate enterovirus species clusters and EV109 groups in a sub-branch of HEV-C that also contains CAV 19, 22, 1 and the recently described HEV104 (145) (Fig. 2.3B). As of this writing, this distinct subgroup of enteroviruses has not been successfully grown in cell culture (23).

Several typical picornavirus sequence features are conserved in EV109, including several cis-acting RNA elements, protease catalytic residues, and a nuclear localization sequence (NLS) at the N-terminus of the 3D polymerase gene (Figure 2.4). Specifically, the 5′ and 3′ UTRs maintain conserved RNA secondary structures with known roles in viral translation and replication, including the X and Y domains of the 3′UTR that have been shown in related viruses to play a role in minus-strand synthesis (165). The cis-acting replication element (cre) features a stem loop located in the 2C region with a conserved AAACA motif in the loop (23, 57). The cre hairpin sequence and loop motif are present in EV109 at nucleotide positions 4,354 - 4,412. The EV109 polyprotein sequence also contains canonical precursor viral protease cleavage sites as described in other HEV-C viruses (24) and possesses the predicted catalytic triad of residues in both 2Apro and 3Cpro (14, 63, 117). Also conserved in EV109 is an NLS (PnKTKLnPS) near the N-terminus of the 3Dpol sequence that has been shown to direct the 3CDpro of poliovirus and rhinoviruses to the nucleus, where it cleaves host nuclear factors and inhibits cellular RNA transcription (7, 8, 155).

2.3.3 An inter-species recombinant and truncated 5′UTR

Intra-species recombination events occur frequently in enteroviruses, especially within the non-structural coding region, but inter-species exchanges involving the 5′UTR have also been observed (10, 26, 73, 107, 117, 129, 134, 137). A phylogenetic tree constructed using enterovirus 5′UTR sequences (Figure 2.5A) was inconsistent with the EV109 complete genome tree (Figure 2.3B) because the EV109 5′UTR did not cluster with its counterparts from other members of HEV-C. EV109 fell outside the two major 5′UTR clusters and was closely related to EV104 and HEV92, which further suggested a recombination event.
Figure 2.4 Sequence analysis and RNA folding predictions reveal conserved picornavirus features in EV109.

RNA secondary structure prediction of the EV109 5’UTR using pknotsRG (119) and visualized using PseudoViewer3 (24). Previously described type-1 IRES features maintained in EV109 include a 5’ cloverleaf as stemloop I, a “GNRA” sequence in stemloop IV, and a pyrimidine-rich region in stemloop VI (28). Stemloops II and III are predicted as one combined stemloop structure. Other nucleic acid folding software (mfold and nupack) predicted similar secondary structures (39, 166).

Tapparel et al. (145) recently described EV104 and reported that its 5’UTR had also undergone possible recombination. To evaluate the likelihood of genomic recombination, full-length bootscanning analysis was performed with representative strains of HEV-C, HEV-A, and HEV-D (types CAV 19, HEV92, and HEV68, respectively) using EV109 as the query sequence (Figure 2.5B). Throughout the 5’UTR region, there was high bootstrap support (>75%) for clustering with HEV 92 (HEV-A), while the coding region maintained high bootstrap support for CAV 19 (HEV-C). The bootscanning analysis revealed a phylogenetic conflict between the 5’UTR and the downstream coding sequence, which suggests that EV109 arose from an inter-species recombination event preceding the VP4 start codon between an HEV-A and HEV-C type enterovirus.

Enterovirus 5’UTRs maintain a common structural organization that includes a 5’-end cloverleaf structure and several pseudoknots that function as an internal ribosomal entry site (IRES) required for translation of the viral polyprotein. The IRES of EV109 (Figure 2.4) features several required motifs, such as a GNRA consensus sequence in stemloop IV and a pyrimidine-rich region in stemloop VI, and is followed by a hyper-variable region preceding the VP4 start codon (117). The average 5’UTR length of human enteroviruses is 740nt. Pair-wise alignments to other HEV-A and HEV-C 5’UTR-VP4 junctions revealed a truncated hyper-variable region in the 3’-end of the EV109 5’UTR (Figure 2.5C), resulting in a shorter than average 633nt 5’UTR. In spite of this shortened sequence, RNA folding algorithms (mfold and NuPack) predicted that the EV109 5’UTR maintains the canonical enterovirus cloverleaf and IRES structures, including required sequence motifs (Figure 2.4).

2.3.4 EV109 capsid homology modeling

The non-enveloped picornavirus capsid proteins are subject to the diversifying effects of host immunologic pressure. To gain additional insight into the viral capsid diversity of EV109, we examined the position-specific amino acid conservation of the four structural protein sequences (VP4, VP2, VP3, and VP1) of EV109 compared to other HEV-C relatives. To perform this analysis, the structural protein amino acid sequences were aligned and scored with a position-specific scoring matrix (PSSM) for amino acid position-specific conservation. The PSSM scores were then mapped onto the 3-dimensional viral pentamer crystal structure of coxsackievirus A21 (PDB code: 1Z7S).
Figure 2.5 The EV109 5′UTR region reveals evidence of an ancestral recombination event.
A. Phylogenetic tree constructed from 5′UTR regions of 74 annotated human enteroviruses. A black arrow denotes EV109. Also highlighted are CAV 19 (red arrow), HEV 92 (blue arrow), and HEV 68 (orange arrow), which are used in subsequent bootscanning analysis. ClustalW and MEGA were used for alignments and tree construction, respectively, using the neighbor-joining method and 500 bootstrap replicates. B. Bootscanning analysis of EV109. Bootscanning analysis was performed with other serotypes of HEV-A (HEV 92), HEV-C (CAV 19), and HEV-D (HEV 68) using a word size of 400 and step size of 20. C. Similarity plot and sequence alignment. Sequence alignment is shown depicting the scanning pairwise identity using a 400-nt sliding window evaluated every 20 nucleotides at the 5′UTR-VP4 junction site. The EV109 sequence is compared to representative members of HEV-A and HEV-C.

and revealing positions of amino acid diversity throughout the enterovirus capsid pentamer (Figure 2.6). Non-conserved EV109 residues (as denoted by a negative PSSM score) are located both along protrusions and within the capsid canyons on the external pentamer surface, but are not aggregated within one particular external location (Figure 2.6A and 2.6B). Non-conserved residues are also located on the internal capsid surface and predominantly cluster along the edges of the five tetrameric units that make up the pentamer (Figure 2.6C).

2.3.5 Identification of additional EV109 isolates

EV109-specific PCR primers for the VP1 region were designed and used to screen a total of 310 ILI samples from the Nicaraguan pediatric cohort in a blinded fashion. Four additional EV109-positive samples were detected (1.6% detection rate). Additional PCR primers were then designed and employed to recover the full-length VP1 sequence and a 350bp region spanning the 5′UTR hyper-variable and VP4 junction region from each sample. The five viruses shared >95% nucleotide and >94% amino acid pairwise identity across the full-length VP1 region and >95% nucleotide pair-wise identity in the 5′UTR-VP4 junction (Genbank Accession numbers: GU131224-GU131227). This high relatedness in multiple regions (as denoted by black bars on Figure 2.2B) suggested that all five isolates indeed belong to our newly characterized enterovirus type.

Unmasking the clinical status and demographic information of the five samples revealed symptom onset dates between January 14, 2008 and April 23, 2008 (Table 2.3). In addition to ILI symptoms, 3/5 cases exhibited gastroenteritis symptoms, including abdominal pain or vomiting. One patient was transferred to the National Pediatric Reference Hospital. All five cases originated from separate households within a 2.5 km² area of northwest Managua and involved female patients (Figure 2.7).
Figure 2.6 Amino acid conservation on the enterovirus capsid pentamer subunit.

Amino acid alignment PSSM score of EV109 compared to other HEV-C capsid sequences and mapped on the pentamer crystal structure of Coxsackievirus A21. Residues shaded in blue have a higher PSSM score and are more conserved in EV109, whereas yellow residues have a negative PSSM and are non-conserved (scale bar below panel C). A. External pentamer view. B. Cross-sectional view. C. Internal pentamer view.
Table 2.3 Clinical and demographic information from five EV109-positive patients, including patient age, date of illness onset, and symptoms.

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>5’UTR-VP4 junction accession</th>
<th>Full VP1 accession</th>
<th>Specimen</th>
<th>Age</th>
<th>Sex</th>
<th>Onset</th>
<th>Symptoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>4733.01</td>
<td>GU265819</td>
<td>GU131227</td>
<td>Nose/throat swab</td>
<td>3 yrs</td>
<td>Female</td>
<td>Jan 14, 2008</td>
<td>Fever, sore throat, lymphadenopathy, cough, coryza, abdominal pain</td>
</tr>
<tr>
<td>4323.04</td>
<td>GU265817</td>
<td>GU131225</td>
<td>Nose/throat swab</td>
<td>3 yrs</td>
<td>Female</td>
<td>Jan 27, 2008</td>
<td>Fever, headache, sore throat, lymphadenopathy, cough, coryza, vomiting **</td>
</tr>
<tr>
<td>4327.01</td>
<td>GQ865517*</td>
<td>GQ865517*</td>
<td>Nose/throat swab</td>
<td>3 yrs</td>
<td>Female</td>
<td>Feb 23, 2008</td>
<td>Fever, sore/ red throat, cough, coryza</td>
</tr>
<tr>
<td>1578.01</td>
<td>GU265816</td>
<td>GU131226</td>
<td>Nose/throat swab</td>
<td>11 yrs</td>
<td>Female</td>
<td>Apr 15, 2008</td>
<td>Fever, cough, coryza, vomiting</td>
</tr>
<tr>
<td>2751.01</td>
<td>GU265818</td>
<td>GU131224</td>
<td>Nose/throat swab</td>
<td>11 yrs</td>
<td>Female</td>
<td>Apr 23, 2008</td>
<td>Fever, sore/ red throat, coryza</td>
</tr>
</tbody>
</table>

* Full-length genome  
** Patient was transferred to hospital

Figure 2.7 Geographic locations of the five EV109-positive patient households. Each patient lived within a separate household located in a 2.5 square kilometer area in northwest Managua, Nicaragua. Patient households denoted by black circles and patient ID. The study health center is denoted by a cross symbol. Figure 2.7 was not included in the published Journal of Virology manuscript.
2.4 Discussion

In this study, we report the discovery and full genome sequence of a novel enterovirus isolated from a case of acute pediatric respiratory illness in Nicaragua and we propose the name Enterovirus 109, according to Picornavirus Study Group naming conventions. EV109 was detected as part of respiratory virus surveillance studies of pediatric ILI patients who are participants in a long-term community-based cohort study in Managua (88). EV109 is a member of the HEV-C species and is most closely related to CAV19, CAV22, CAV1 and EV104 (Figure 2.3), a distinct subgroup within HEV-C that has yet to be grown in cell culture (23). The VP1 sequence of EV109 shared <72% amino acid and <66% nucleotide identity to its relatives, fulfilling current classification requirements designating new enterovirus types.

This study demonstrated the utility of deep sequencing as a strategy to quickly sequence the full-length genome of a novel virus too divergent to be easily recovered by other methods. Recovering the genome of a previously uncharacterized virus has been historically accomplished by screening bacterial or phage libraries, primer-walking, and/or PCR using degenerate primers. In the best of cases, recovery of the genome is a straightforward iterative trial and error process. However, several factors, such as the degree of identity to known viruses and the copy number of the genome present in the sample can make this process highly inefficient, time-consuming, and costly. In the case where the presence of the viral target is vanishingly small (< 1:100,000), ultra deep sequencing can provide a set of distributed reads that can then be used to rapidly close the genome, even in the context of a host organism that lacks a reference genome (86). Even when the entire genome cannot be assembled from the primary sequencing reads, as was the case for EV109, this strategy yields an overall cost savings, in terms of supplies, labor, and time; all of which will improve as sequencing output per run continues to grow and cost decreases. The sequence data presented here represents 1/8th of an Illumina GAII sequencing run (1 lane on an 8 lane flowcell) and cost $750. Thus, the cost per recovered viral read (211 total, out of 20 million) is $3.55, which we note is approximately half the cost of a conventional PCR primer (25mer, at $0.28/base). While colony screening, degenerate PCR, and primer walking strategies still have their place in novel genome recovery, the time when these techniques are rendered obsolete by ultra deep sequencing is clearly on the horizon. It is notable that of the complete sequencing data set, only 211 reads out of 20.8 million (0.001%) actually originated from the EV109 genome (Figure 2.2A). Given the non-sterile location of the sample (nose and throat swab) and the unbiased nucleic acid extraction, it may not be surprising that so few reads could be obtained. In this regard, deep sequencing technologies that produce less depth (under 1 million reads), may have missed this species completely. Regardless, using our alignment parameters, more reads identified with the picornavirus family than all other virus families combined. Furthermore, the use of paired-end sequencing greatly diminished the possibility of obtaining spurious hits.

The coding region of EV109 is most closely related to HEV-C species, but the 5’UTR is not closely related to either the HEV-C or HEV-A phylogenetic groups, suggesting ancestral recombination with a divergent species outside the two major groups (Figure 2.5). The observed recombination product between different HEV species is consistent with previous reports that inter-species recombination involving the 5’UTR can occur in enteroviruses, such as results by Smura et al. that describe an enterovirus genome closely related to HEV-A types, except the 5’UTR region, which clusters with HEV-C and D sequences (137). The recombination evidence in EV109 is similar to the recombination event reported by Tapparel et al. in EV104 (145). The
existence of recombinant enteroviruses has practical implications for diagnostic strategies that attempt to type enteroviruses using solely 5′UTR sequence and highlights the importance of obtaining VP1 sequence for definitive identification. The 663nt 5′UTR of EV109 is shorter than average by virtue of a truncated hyper-variable region between the maintained IRES secondary structure (Figure 2.4) and the start codon, indicating that the hyper-variable region is deletion-tolerant. It has been suggested by poliovirus studies (139) that this unstructured variable stretch plays a role as a spacer element preceding the authentic translation start codon of internal initiation and that upstream AUG codons, including two in stemloop VI at base 578 and 584 of EV109, may allow ribosome docking preceding transfer to the downstream AUG at base 664. Beyond its role as a spacer element, we speculate that the hyper-variable region may also be a hotspot for recombination between enterovirus species.

A functional enterovirus with a recombinant 5′UTR would be required to maintain the necessary known interactions with both viral proteins, such as 2B, 2BC, 3A, and 3D, and host proteins, such as polypyrimidine tract-binding protein and poly(rC) binding protein, which are important for viral translation and RNA synthesis (99, 165). To conserve interaction with viral proteins, such a recombination event in the 5′UTR may well have provoked compensatory mutations in the viral coding region. However, the specific sequence motifs necessary for specific viral protein-nucleotide interactions are currently poorly defined, limiting our ability to recognize them. Structural features may also play a functional role, as demonstrated by 3CD-cloverleaf interaction experiments that alter sequence but maintain structure (164). Chimeric polioviruses possessing heterologous 5′UTRs from coxsackievirus or rhinovirus have exhibited growth deficiency in cell culture and attenuated neuropathogenesis in poliomyelitis mouse models, though they share ~70% nucleotide identity across 5′UTRs (62, 79). Others have observed complementary mutations in 3A/3AB and 3C/3CD that alleviate cell type-specific growth defects in enterovirus viruses with incompatible UTRs; although the mutations were confined to the 3A and 3C regions, homogenous mutations were not identified at defined sites (131). Future site-directed mutagenesis studies using EV109 could elucidate the selected coding region mutations needed to maintain UTR-viral protein interactions within the context of a natural recombination event.

A total of five EV109 isolates were detected in nose/throat swabs from five children in separate households between January and April 2008. We screened 310 nose/throat swab samples by PCR for EV109 and obtained a detection rate of 1.6%. In addition to presenting with acute influenza-like illness (fever or history of fever with a cough and/or sore throat), three of the five patients also presented with gastrointestinal symptoms (Table 2). EV109 is genetically similar to other known pathogenic enteroviruses, and though present findings are suggestive, human disease causation has not yet been formally established. Each of the five isolates reported here were negative by RT-PCR testing for a number of additional respiratory viruses, including Influenza A and B, Respiratory Syncytial Virus, and Human Parainfluenza Virus 1, 2, and 3 (data not shown). Deep sequencing did not reveal a high number of reads aligning to other viral families, with the exception of reads with identity to human endogenous retroviruses. Other low-level viral reads observed in the sequencing library are likely false positives derived from spurious alignments. These findings are consistent with a pathogenetic role in the illnesses described here, but further experimentation is required to formally establish such a role. Our findings should make possible development of serologic as well as genomic testing, which would allow (i) documentation of seroconversion during illness and (ii) estimation of seroprevalence in other populations. Additionally, given the wide variety of enterovirus-associated diseases, which
includes ILI, gastroenteritis, encephalitis, aseptic meningitis, and acute flaccid paralysis (76, 115), more expansive screening of both asymptomatic and symptomatic subjects with other clinical syndromes will be required to establish the full spectrum of EV109’s disease associations.

It should be noted that EV109 was successfully detected using conserved 5'UTR primers first derived by Lönnrot et al. in 1999 (90), but subsequent testing in other studies has not detected EV109, or perhaps failed to detect it due to the 5'UTR recombination. Although the global distribution of EV109 is currently unknown, this study describes the first novel picornavirus isolated in Nicaragua. Given the paucity of dedicated molecular diagnostic testing and the large percentage of undiagnosed cases of respiratory illness in developing country settings, our findings underscore the importance of conducting molecular diagnostic and surveillance studies in tropical, developing regions.
2.5 Acknowledgments

We thank the children and families participating in the Nicaragua Influenza Cohort Study and the physicians, nurses, and staff at the Health Center Sócrates Flores Vivas and the Nicaraguan National Virology Laboratory, in particular Angel Balmaseda, Saira Saborio, William Avilés, Roger López, Cristhiam Cerpas, Carolina Flores, Heyri Roa, Moises Navarro, and Patricia Castillo. We would also like to acknowledge the data entry and data managers on the Nicaraguan Influenza Cohort Study. We are grateful to J. Graham Ruby for helpful advice and editing and to Lucille Huang for technical support. This work was supported in part by a grant from the Pacific Southwest Regional Center of Excellence (PSWRCE), NIH grant U54-AI65359, the Pediatric Dengue Vaccine Initiative (#VE-1 to EH), the Howard Hughes Medical Institute, the Doris Duke Foundation, and the Packard Foundation. Joseph DeRisi and Don Ganem are supported by the Howard Hughes Medical Institute.
Chapter 3

Virus Identification in Unknown Tropical Febrile Illness Cases Using Deep Sequencing

This chapter describes experiments that report the detection of known and novel viruses in cases of acute unknown tropical febrile illness using comprehensive metagenomic analysis.

Dengue virus is a re-emerging viral disease that infects an estimated 50-100 million people annually worldwide, yet current diagnostic practices cannot detect an etiologic agent in ~40% of dengue-like illnesses. Metagenomic approaches to pathogen detection, such as viral microarrays and deep sequencing, are promising tools to address emerging and non-diagnosable disease challenges, but have not been fully utilized in cases of tropical illness. In Chapter 3, we used the Virochip microarray and deep sequencing to characterize the viral spectrum in human sera from Nicaraguan patients presenting with dengue-like symptoms but testing negative for dengue virus. We utilized a barcoding strategy to simultaneously sequence multiple serum specimens, generating an average of 5.1 million reads per sample. We then implemented a stepwise bioinformatic filtering pipeline to remove the majority of human and low-quality sequences to improve the speed and accuracy of subsequent GenBank searches. Virus sequence was detected in 35% (12/34) of previously negative cases using deep sequencing, including six samples with Human Herpesvirus 6 sequence and five samples containing sequence from a putative novel virus related to the Circoviridae family. These results demonstrate the utility of two metagenomic strategies, the Virochip and deep sequencing, as comprehensive platforms to detect known and divergent viruses in the study of tropical febrile illness.

3.1 Introduction

Viral infections pose a significant global health burden, especially in the developing world where most infectious disease deaths occur in children and are commonly due to preventable or treatable agents. Effective diagnostic surveillance tools are crucial for reducing disability-adjusted-life-years (DALYs) due to infectious agents and bolstering elimination and treatment programs (96). Previously unrecognized and novel pathogens continually emerge due to globalization, climate change, and environmental encroachment and pose unique and important diagnostic challenges (40, 80).

Traditional viral detection methods, such as serology, virus isolation, and PCR, focus on known agents and can be combined to target many pathogens, but can be time and resource intensive (40). Moreover, novel and highly divergent viruses are not be easily detected by approaches that rely on a priori sequence, antigen, or cell tropism data. PCR-based assays that employ degenerate primers may successfully target conserved regions within related virus
groups, but viruses lack universally conserved genetic regions, such as ribosomal RNA, that can be exploited to amplify all viruses (121).

Viral metagenomic analysis enables more systemic viral detection of both known and novel viral pathogens (17, 38, 64, 144) and is approached through a variety of microarray and sequencing strategies (11, 151). The Virochip is a pan-viral microarray platform that has been previously utilized in the detection and discovery of viruses from both human and animal samples (28, 84, 85, 148, 152). Deep sequencing and shotgun sequencing of human clinical samples has been previously used for viral detection (51, 105, 142, 150), novel virus discovery (22, 50, 61, 114), and divergent virus genome recovery (162). Viral metagenomic approaches have also been employed as a diagnostic supplement to pathogen detection as part of public health monitoring systems (142), but have been limited to shotgun sequencing of viral-enriched libraries and have yet to utilize deep sequencing data. Currently available sequencing platforms can generate millions to hundreds of millions of sequencing reads per run, far exceeding large-scale shotgun sequencing (141). Deep sequencing of clinical samples, in which hundreds of thousands to millions of sequencing reads are generated per sample, can be incorporated into stepwise virus detection pipelines (93). Database searches using Basic Local Alignment Search Tool (BLAST) variants (92) can be used to identify sequences in samples that correspond to known and novel viruses, including those present at exceedingly low titers or deriving from viruses that may be too divergent to be detected with PCR or microarray methods. Deep sequencing represents an unbiased, highly sensitive method for identifying viral nucleic acid in clinical samples, but prior investigations that sought to identify viruses (21, 81) and characterize the human virome (20, 80) in human blood and serum have yet to utilize deep sequencing technologies.

This study describes the use of deep sequencing for the direct viral diagnosis of serum from cases of acute pediatric febrile illness in a tropical urban setting. Patient clinical data and serum samples were collected between 2005 and 2009 as part of an ongoing pediatric dengue virus study at the Hospital Infantil Manuel de Jesús Rivera (HIMJR) in Managua, Nicaragua (120). Deep sequencing was performed on 34 dengue virus-negative serum samples using the Illumina Genome Analyzer II (GAII) platform and generated 184.6 million reads (12.0 billion bases) with a range of 1.5 to 14.9 million reads per samples. Virus nucleic acid was detected in 12/34 samples (35%), with human herpesvirus 6 detected in 6/34 samples (17.6%). Additionally, novel and highly divergent sequence related to the Circoviridae family was detected in 5/34 samples (14.7%).

### 3.2 Experimental Procedures

**Study population**

Acute serum samples were collected from consenting suspected dengue cases at the Hospital Infantil Manuel de Jesús Rivera (HIMJR), the National Pediatric Reference Hospital in Managua, Nicaragua. A suspected dengue case was defined as an acute febrile illness plus two or more of the following signs and symptoms: headache, retro-orbital pain, myalgia, arthralgia, rash, and hemorrhagic manifestations. Suspected dengue cases were tested for DENV at the Centro Nacional de Diagnóstico y Referencia (CNDR) of the Nicaraguan Ministry of Health by: 1) DENV isolation 2) DENV reverse transcriptase-polymerase chain reaction (RT-PCR), 3) IgM capture enzyme-linked immunosorbent assay (ELISA) of paired acute and convalescent sera, and
4) inhibition ELISA of paired acute and convalescent sera. All patients were aged 6 months to 14 years and presented between August 2005 and January 2009. Inclusion criteria for the retrospective Virochip analysis were suspected dengue cases negative by all four DENV assays. Inclusion criteria for the deep sequencing analysis were suspected dengue cases presenting to HIMJR < 4 days after symptom onset with both fever and rash, either on exam or by patient report, and negative by all DENV assays and Virochip. The study protocol was reviewed and approved by the Institutional Review Boards (IRB) of the University of California, Berkeley, and of the Nicaraguan Ministry of Health. Parents or legal guardians of all subjects in both studies provided written informed consent, and subjects 6 years of age and older provided assent.

Virochip analysis and sequence recovery

Three analysis tools were used to analyze Virochip data: E-predict (147), Z-score analysis (29), and cluster analysis (44). An array was deemed positive for a particular virus if the virus was identified by at least 2 of these methods. In some cases, positive results were confirmed by PCR and sequencing of the PCR products.

For dengue virus PCR, primers MAMD (5'-AAC ATG ATG GGR AAR AGR GAR AA-3') and cFD2 (5'-GTG TCC CAG CCG GCG GTG TCA TCA GC-3') were used (130). These primers target a 250-nucleotide region in the NS5 gene conserved in many flavivirus genomes. For hepatitis A virus PCR, primers MDS-47 (5'-CTA AAT TGG GGA CRC AGA TG-3') and MDS-48 (5'-TGA GGA AAA ACC TAA ATG CC-3') were used. These target a 411 nucleotide conserved region in the 5' UTR of many Hepatitis A sequences (corresponding to nucleotides 311-721 of the Hepatitis A virus reference sequence, NC_001489). PCR thermocycling conditions were: 95°C for 2 minutes, 35 cycles of 95°C for 30 seconds, 50°C for 30 seconds, 72°C for 30 seconds, and 72°C for 2 minutes.

For Cvl-NI, we designed primers based on Circovirus-like reads recovered in sample 387. Primers Cvl-N1 1F (5'-GGG TGA GAT TGG TGC TGG TA-3') and Cvl-N1 4R (5'-CCT TTT TGC TCC ACA AGG ACT-3') amplified 203 nucleotides from the putative Rep ORF under the following PCR conditions: 95°C for 5 minutes, 40 cycles of 95°C for 60 seconds, 54°C for 60 seconds, 72°C for 60 seconds, and 72°C for 5 minutes.

Virochip sensitivity

Full-length poliovirus genomic RNA was transcribed from MluI-linearized plasmid prib(+)XpA using T7 polymerase as previously described (68). Poliovirus RNA was mixed with HeLa total RNA in a dilution series ranging from 10⁻² to 10⁻⁴ poliovirus gRNA per HeLa RNA. Randomly-primed dsDNA libraries were prepared, hybridized to the virochip, and analyzed as described above.

Deep sequencing library preparation

Serum nucleic acid was extracted using the QIAamp Viral RNA Isolation Kit (Qiagen, Valencia, CA). Randomly amplified cDNA libraries were generated using a random hexamer, affixed with “adapter A” and “adapter B” sequences (Illumina), and size selected on a 4% native polyacrylamide gel (30mM KCl, 1X TBE buffer, 19:1 acrylamide:bis) as previously described (162). Library generation primers were modified to include a three-nucleotide of four-nucleotide barcode sequence at the 3'-end to allow pooling of multiple samples in each sequencing lane. Pool 1 and pool 2 each included six sample libraries with three-base barcode sequences. Pool 3 included six sample libraries with four-base barcodes, and pool 4 comprised sixteen sample
libraries with three-base barcodes. One paired-end sequencing run was performed using the GAII Sequencer (Illumina) for each of the four sample pools. We utilized one lane of a sequencing flowcell per pool (four lanes total, one lane each from four separate sequencing runs).

Deep sequencing bioinformatics

The initial FASTQ data from each pool’s lane were binned for each barcode. Barcoded reads were identified by matching the barcode positions of each end of the paired reads to the known barcodes, allowing only a single mismatch in either of the paired ends. The barcode-split reads were trimmed of non-template deriving and potentially error-prone sequence: a randomly incorporated nucleotide (N), the barcode (3 or 4 bases), and the sequence corresponding to the random hexamer, leaving 54 or 55 bases per read in our working datasets. The lowest complexity fraction was identified by sequences with LZW-compressed sizes less than 25. Reads were aligned to the human genome (build hg18) first using BLAT with the “–fastMap” flag, and after filtering, the remaining reads were aligned using BLAT without the flag. Paired reads for which at least one of the reads in the pair had at least 80% identity to the database were marked as human and removed from subsequent analyses. After removal of reads identified as human by BLAT, remaining reads were aligned and filtered by mapping to the human transcriptome using nucleotide BLAST (BLASTn, word size 30, E value $10^{-3}$). Remaining reads were next aligned to the human genome using BLASTn (word size 30, E value $10^{-3}$), filtered, and again aligned to the human genome by BLASTn (word size 11, E value 10). After all human filtering, we analyzed the distribution of the complexity of reads and saw a relative enrichment of reads with LZW compressed sizes lower than 30, so those reads were removed from further analysis. In looking for reads with viral homology, we search the non-redundant nucleotide database (nt) using BLASTn (word size 20, E value $10^{-3}$). Reads that did not map to nt were aligned to the non-redundant protein database (nr) using translated BLAST (BLASTx, word size 3, E value 10).

Virus sequence detection

In order to make specific virus positive calls, we implemented a set of rules to minimize false positives while maintaining sensitivity. In order to reduce the number of false positive sequences that may share identity equally with both viral and non-viral genomes, we restricted our analysis to those queries whose best alignments were only to viral sequences. In a number of datasets, we detected human klassevirus 1, a virus identified and studied in our lab (61), as well as Moloney murine leukemia virus (MMLV), whose polymerase was used in the sequence library preparation. We believe these reads represent lab contaminants, and others studies that prepared sequence libraries in the same location have reported similar findings (12) (Arron, et al., J Invest Dermatol, 2011. In press.). To account for these contaminants, positive calls were only made on viruses for which there were more reads supporting its call than there were reads to any known contaminant. Finally, in order to avoid making calls based on potentially spurious alignments, we considered only those viruses for which there were at least 5 reads supporting their presence.

3.3 Results

3.3.1 Virochip analysis of serum from acute pediatric febrile illness

We tested 150 human serum samples using with the Virochip to determine whether a pan-viral microarray could comprehensively detect known viruses in acute pediatric febrile
illness cases (GEO accession series: GSE28142). Samples were collected at HIMJR in Managua, Nicaragua, and selected from cases that fulfilled the WHO clinical definition of dengue fever or dengue hemorrhagic fever/dengue shock syndrome but were negative for dengue virus by dengue specific serology, RT-PCR, and virus isolation. We prepared randomly amplified cDNA libraries from extracted nucleic acid, applied the libraries to the Virochip as previously described (152), and analyzed the microarray results using three analysis methods: E-predict, an automated tool for hybridization pattern interpretation (147), Z-score analysis (29), and visual cluster analysis (44). An array was deemed positive for a particular virus if it was identified by at least two of these methods. In some cases, positive results were confirmed by PCR and sequencing. We detected two samples that were positive for Hepatitis A virus and twelve samples that were positive for torque teno virus (TTV) using the Virochip (Table 3.1).

We included five additional blinded dengue virus positive control samples in our screening (one DENV-1 and four DENV-2 samples). The Virochip successfully detected the appropriate dengue virus type in all five controls, and subsequent sequence recovery yielded dengue virus sequence with high nucleotide similarity (98-100% pairwise identity) to previously described Nicaraguan isolates. We also applied poliovirus RNA to the Virochip as an additional positive control to quantify Virochip sensitivity. We did this by mixing poliovirus genomic RNA with HeLa total RNA and subjecting these mixtures to an identical library preparation and microarray hybridization protocol. Using the E-predict analysis tool, the lowest detectable concentration of poliovirus was 1 viral gRNA per $10^5$ HeLa RNA molecules (approximately 10 polio gRNAs per cell equivalent of HeLa RNA; Figure 3.1).

### Table 3.1

<table>
<thead>
<tr>
<th>Virus identified on Virochip</th>
<th>Number positive microarrays</th>
<th>Confirmatory PCR/sequencing</th>
<th>Accession of top-scoring BLAST hit to recovered sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>DENV-1</td>
<td>1</td>
<td>1 / 1</td>
<td>100% identity to HM631855.1</td>
</tr>
<tr>
<td>DENV-2</td>
<td>4</td>
<td>3 / 4</td>
<td>100% identity to JF357907.1</td>
</tr>
<tr>
<td>Hepatitis A Virus</td>
<td>2</td>
<td>2 / 2</td>
<td>99% identity to K02990.1</td>
</tr>
<tr>
<td>TTV</td>
<td>12</td>
<td>0 / 12</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>131</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.1 Summary of viruses identified on Virochip microarrays.**

All arrays were from the Viro5 platform and submitted to the Gene Expression Omnibus (GEO) database as accession series GSE28142.
Figure 3.1 Virochip sensitivity using poliovirus control RNA.
The Virochip can detect one poliovirus gRNA in a background of $10^5$ HeLa RNA molecules. Poliovirus RNA was mixed with HeLa total RNA and analyzed on the Virochip. Eighty enterovirus Virochip oligos were found to be responsive to the poliovirus RNA and the mean fold above background of the normalized intensity of these oligos is plotted. Background is defined as the normalized intensity for each oligo in the HeLa-only control sample. The top E-predict hit in the $10^{-5}$ to $10^{-2}$ samples was human enterovirus C.

3.3.2 Aggregate deep sequencing analysis

We performed deep sequencing using the Illumina GAIID (Illumina, Inc.) platform on a subset of 34 serum samples collected from cases of acute pediatric febrile illness in Managua, Nicaragua. Each serum sample was selected from the HIMJR study and shared the following clinical criteria: fever or history of fever and rash. Additionally, each patient sample was collected within three days of illness onset and was negative by specific tests for dengue virus and by the Virochip (Table 3.1). Nucleic acid library preparation for deep sequencing was performed as before (162) using random-primed cDNA libraries derived from RNA from serum, with the additional step of adding barcode sequences three to four nucleotides in length at the end of the known adapter sequences to facilitate sample pooling. Six serum samples were simultaneously pooled and sequenced in each of pools 1, 2, and 3, and sixteen samples made up pool 4. One deep sequencing lane was run for each sample pool (one flow cell lane per pool) generating a total of 184.6 million 65-nucleotide long paired-end reads (12.0 billion bases total;
Each of the barcoded samples yielded between 1.5 and 14.9 million reads with an average of 5.1 million reads per sample (Figure 3.2B).

### Figure 3.2 Deep sequencing initial reads

**A. Total barcoded reads**

<table>
<thead>
<tr>
<th>Pool</th>
<th>Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pool 1</td>
<td>39,755,772</td>
</tr>
<tr>
<td>Pool 2</td>
<td>41,510,714</td>
</tr>
<tr>
<td>Pool 3</td>
<td>40,223,892</td>
</tr>
<tr>
<td>Pool 4</td>
<td>53,052,892</td>
</tr>
</tbody>
</table>

**B. Initial reads per barcode**

![Bar chart showing reads per barcode](chart.png)

3.3.3 Bioinformatic filtering

All of the raw reads were first separated into 34 independent read sets by barcode sequence identity and analyzed as individual data sets. We trimmed the raw reads of a randomly-incorporated nucleotide, the sequence barcode, and the sequence corresponding to the random hexamer, and performed all downstream analysis on ~55 nucleotide paired-end reads. We removed sequencing reads of low-quality, low complexity, and high degree of similarity to human sequence from each data set in order to reduce the computing time and to increase the specificity of subsequent virus sequence searches (Figure 3.3). Low quality reads were defined as having more than five indeterminate bases, and low complexity was determined by the Lempel-Ziv-Welch (LZW) compressed size of the sequence string, where the more compressible the sequence, the lower its complexity (156). An average of 2.0% of the total reads from each sample met our criteria for low quality or low complexity and were removed from subsequent analysis. Next, since the vast majority of sequenced nucleic acid from human serum has high identity to human sequences, we removed human reads through BLAT and BLAST alignments to the human genome and transcriptome. In an effort to minimize computing time while maximizing sensitivity, we took an iterative approach to filtering human sequence, filtering the initial datasets with faster, less sensitive BLAT searches, then filtering the smaller, previously-filtered datasets with slower, more sensitive BLAST searches. An average of 28.4% of the initial reads remained per sample following two BLAT searches to the human genome in which read pairs for which at least one read shared at least 80% identity to human sequence were removed. We then performed a BLASTn search to the human transcriptome, followed by a faster, longer word size BLASTn, and a slower, shorter word size BLASTn to the human genome, again.
removing pairs for which at least one read aligned to the database with at least 80% identity over the whole read. We performed an additional, less conservative LZW compression filter (Figure 3.3) to remove the fraction of sequences with complexity too low to be aligned by BLAST to any database. An average of 3.3% of the initial sequencing reads remained after the bioinformatic filtering steps, with an absolute read average of 150,000 remaining per sample (Table 3.2).

Figure 3.3 Bioinformatic filtering of deep sequencing data.

Average percent remaining reads after each of the filtering steps. Low quality and low complexity reads are removed first, followed by iterative BLAT and BLAST comparisons to human sequence. Averages were calculated for all samples (n = 34). Inset: viral search strategies using filtered read sets include BLASTn and BLASTx to GenBank databases, followed by Hidden Markov Model searching and de novo sequence assembly of unidentifiable reads, tools under development.
Table 3.2 Post-filtering reads per barcode.

<table>
<thead>
<tr>
<th>Barcode</th>
<th>Initial reads</th>
<th>Remaining reads</th>
<th>% reads remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>1.98 M</td>
<td>69,161</td>
<td>3.49%</td>
</tr>
<tr>
<td>ATT</td>
<td>8.99 M</td>
<td>455,495</td>
<td>5.07%</td>
</tr>
<tr>
<td>CCA</td>
<td>5.03 M</td>
<td>89,453</td>
<td>1.78%</td>
</tr>
<tr>
<td>CGT</td>
<td>4.59 M</td>
<td>71,341</td>
<td>1.55%</td>
</tr>
<tr>
<td>GAG</td>
<td>13.68 M</td>
<td>245,219</td>
<td>1.79%</td>
</tr>
<tr>
<td>GGA</td>
<td>5.49 M</td>
<td>229,927</td>
<td>4.19%</td>
</tr>
<tr>
<td>CAC</td>
<td>3.22 M</td>
<td>71,523</td>
<td>2.22%</td>
</tr>
<tr>
<td>CCA</td>
<td>2.74 M</td>
<td>55,761</td>
<td>2.04%</td>
</tr>
<tr>
<td>CGT</td>
<td>5.26 M</td>
<td>407,054</td>
<td>7.73%</td>
</tr>
<tr>
<td>CTG</td>
<td>4.16 M</td>
<td>133,603</td>
<td>3.21%</td>
</tr>
<tr>
<td>GAG</td>
<td>11.22 M</td>
<td>276,150</td>
<td>2.46%</td>
</tr>
<tr>
<td>TTA</td>
<td>14.91 M</td>
<td>349,359</td>
<td>2.34%</td>
</tr>
<tr>
<td>AAAA</td>
<td>6.93 M</td>
<td>88,440</td>
<td>1.28%</td>
</tr>
<tr>
<td>AATT</td>
<td>3.63 M</td>
<td>61,121</td>
<td>1.68%</td>
</tr>
<tr>
<td>AAGG</td>
<td>6.75 M</td>
<td>118,089</td>
<td>1.75%</td>
</tr>
<tr>
<td>AACC</td>
<td>2.53 M</td>
<td>31,517</td>
<td>1.25%</td>
</tr>
<tr>
<td>ATAT</td>
<td>10.78 M</td>
<td>109,185</td>
<td>1.01%</td>
</tr>
<tr>
<td>ATTA</td>
<td>9.6 M</td>
<td>126,972</td>
<td>1.32%</td>
</tr>
<tr>
<td>AAA</td>
<td>7.43 M</td>
<td>603,260</td>
<td>8.12%</td>
</tr>
<tr>
<td>ACC</td>
<td>7.53 M</td>
<td>250,078</td>
<td>3.32%</td>
</tr>
<tr>
<td>AGG</td>
<td>3.26 M</td>
<td>65,806</td>
<td>2.02%</td>
</tr>
<tr>
<td>ATT</td>
<td>4.99 M</td>
<td>201,208</td>
<td>4.03%</td>
</tr>
<tr>
<td>CAC</td>
<td>1.91 M</td>
<td>227,606</td>
<td>11.94%</td>
</tr>
<tr>
<td>CCA</td>
<td>3.16 M</td>
<td>42,578</td>
<td>1.35%</td>
</tr>
<tr>
<td>CGT</td>
<td>1.65 M</td>
<td>105,073</td>
<td>6.38%</td>
</tr>
<tr>
<td>CTG</td>
<td>2.29 M</td>
<td>58,416</td>
<td>2.55%</td>
</tr>
<tr>
<td>GAG</td>
<td>3.46 M</td>
<td>56,535</td>
<td>1.64%</td>
</tr>
<tr>
<td>GCT</td>
<td>2.01 M</td>
<td>66,150</td>
<td>3.28%</td>
</tr>
<tr>
<td>GGA</td>
<td>3.7 M</td>
<td>62,202</td>
<td>1.68%</td>
</tr>
<tr>
<td>GTC</td>
<td>1.94 M</td>
<td>84,310</td>
<td>4.35%</td>
</tr>
<tr>
<td>TAT</td>
<td>3.12 M</td>
<td>79,138</td>
<td>2.53%</td>
</tr>
<tr>
<td>TCG</td>
<td>1.47 M</td>
<td>122,978</td>
<td>8.39%</td>
</tr>
<tr>
<td>TGC</td>
<td>2 M</td>
<td>68,351</td>
<td>3.42%</td>
</tr>
<tr>
<td>TTA</td>
<td>3.15 M</td>
<td>73,154</td>
<td>2.32%</td>
</tr>
</tbody>
</table>

Table 3.2 Post-filtering reads per barcode.
Post-filtering reads per barcode depicting the initial absolute reads per barcoded sample, remaining reads, and percentage of remaining reads for each of the 34 sequenced samples. Samples sequenced as part of pool 1 are depicted in blue, pool 2 in green, pool 3 in orange, and pool 4 in red.
3.3.4 Detection of known viruses

The remaining reads not removed by low complexity or human filtering from each of the samples were then compared to sequences in the NCBI non-redundant nucleotide and protein databases using BLASTn and BLASTx respectively. Virus-derived sequences were detected in 12/34 (35%) samples (Table 3.3).

Human Herpesvirus 6 (HHV-6) sequence was detected in 6/34 samples (17.6%). The six HHV-6 positive samples had an average normalized read count of 278 HHV-6 reads per sample (range: 88 – 514), and all reads possessed high sequence identity by BLAST to the HHV-6B reference genome sequence (gi: 9633069). No additional virus reads were detected in 5/6 HHV-6 positive samples, but one HHV-6 positive sample also had reads sharing identity with Torque Teno Virus sequence (Table 3.3).

We then aligned the entire dataset for each barcode containing HHV-6 reads to the HHV-6B genome sequence to investigate the depth and genomic position of the sequence coverage across the HHV-6 genome (Figure 3.4). Though the fraction of each of these barcode datasets deriving from HHV-6 ranges from only 0.002% to 0.02%, there is read coverage across the entire genome and over many genes, making it likely that many viral genes are actively expressed in each of the patient specimens.

We also detected sequence from GB virus C (Hepatitis G Virus) of the Flaviviridae family in one sample and TTV of the Circoviridae family, in 2/34 (5.9%) samples. Though these patients all exhibited symptoms of dengue-like febrile illness and additional symptoms (Table 3.4), neither GB virus C or TTV has yet to be implicated in human disease.

3.3.5 Detection of novel viruses

In addition to detecting known viruses in the serum samples, we also recovered sequence from a novel Circovirus-like virus in 5/34 samples (14.7%). All reads recovered from the divergent virus, which we have denoted Circovirus-like-NI (Cvl-NI), derived from a putative replicase gene related to replicases in the Circoviridae family. We detected a range of 9 to 173 Cvl-NI reads per positive sample by BLASTx (Table 3.3). Due to low sequence coverage, it was not possible to recover the complete Cvl-NI genome sequence from these samples. However, we were able to design primers from the aligned deep sequencing reads and recover a ~220 nt continuous region of the replicase gene (GenBank accession: JF781513) from 3/5 positive nucleic acid libraries by PCR and Sanger sequencing (Figure 3.5). The Cvl-NI sequence is most closely related to Circovirus-like virus RW-E (gi: 254688530), a circular single-stranded DNA virus previously found in reclaimed water samples in Florida (123) and a currently unclassified but possible member of the Circoviridae family. The partial Cvl-NI rep sequence is more closely related to the Circovirus-like viruses than Cycloviruses and Circoviruses (Figure 3.5B). The recovered replicase regions share <70% amino acid identity as determined by BLASTx to its closest relative, indicating that the recovered sequence likely derives from a novel circovirus-like viral species.
Table 3.3 Virus sequence hits.
The relevant virus sequence hits, taxon identification, and normalized virus read counts per sample sets of sequencing reads. BLASTn (n) and BLASTx (x) algorithms were used to search the NCBI nt/nr database.

<table>
<thead>
<tr>
<th>Pool barcode</th>
<th>Patient Code</th>
<th>Relevant viral hit</th>
<th>taxon id</th>
<th>Norm. read count</th>
<th>Number initial reads</th>
<th>BLAST algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-GGA</td>
<td>235</td>
<td>Human Herpesvirus 6 Torque teno virus</td>
<td>10368</td>
<td>116.0</td>
<td>5.5M</td>
<td>n</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>68887</td>
<td>11.0</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>2-CCA</td>
<td>451</td>
<td>Human Herpesvirus 6</td>
<td>10368</td>
<td>88.0</td>
<td>2.7M</td>
<td>n</td>
</tr>
<tr>
<td>3-ATTA</td>
<td>207</td>
<td>Human Herpesvirus 6</td>
<td>10368</td>
<td>390.0</td>
<td>9.6M</td>
<td>n</td>
</tr>
<tr>
<td>4-GAG</td>
<td>432</td>
<td>Human Herpesvirus 6</td>
<td>10368</td>
<td>411.0</td>
<td>3.5M</td>
<td>n</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10368</td>
<td>9.0</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>4-CCA</td>
<td>574</td>
<td>Human Herpesvirus 6</td>
<td>10368</td>
<td>138.0</td>
<td>3.2M</td>
<td>n</td>
</tr>
<tr>
<td>4-AAA</td>
<td>370</td>
<td>Human Herpesvirus 6</td>
<td>10368</td>
<td>514.0</td>
<td>3.2M</td>
<td>n</td>
</tr>
<tr>
<td>4-ATT</td>
<td>118</td>
<td>Circovirus-like genome RW-E</td>
<td>642255</td>
<td>173.0</td>
<td>7.4M</td>
<td>x</td>
</tr>
<tr>
<td>4-CAC</td>
<td>323</td>
<td>Circovirus-like genome RW-E</td>
<td>642255</td>
<td>27.0</td>
<td>5M</td>
<td>x</td>
</tr>
<tr>
<td>4-CGT</td>
<td>363</td>
<td>Circovirus-like genome RW-E</td>
<td>642255</td>
<td>11.0</td>
<td>1.9M</td>
<td>x</td>
</tr>
<tr>
<td>1-ATT</td>
<td>387</td>
<td>Circovirus-like genome RW-E</td>
<td>642255</td>
<td>27.0</td>
<td>9M</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB virus C</td>
<td>54290</td>
<td>171.0</td>
<td></td>
<td>n</td>
</tr>
<tr>
<td>3-AAAA</td>
<td>73</td>
<td>Torque teno midi virus</td>
<td>432261</td>
<td>47.0</td>
<td>6.9M</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Torque teno virus</td>
<td>68887</td>
<td>34.0</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Small anellovirus</td>
<td>393049</td>
<td>29.2</td>
<td></td>
<td>n</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Torque teno mini virus 1</td>
<td>687379</td>
<td>12.2</td>
<td></td>
<td>n</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Torque teno mini virus</td>
<td>432261</td>
<td>8.3</td>
<td></td>
<td>n</td>
</tr>
</tbody>
</table>

Table 3.4 Clinical information of patients with detectable virus sequence.
The clinical variables collected during hospital stay, including age, sex, illness onset date, hospital stay duration, and symptoms in addition to fever and rash.

<table>
<thead>
<tr>
<th>Pool barcode</th>
<th>Patient Code</th>
<th>Relevant viral hit</th>
<th>Patient Age</th>
<th>Sex</th>
<th>Date illness onset</th>
<th>Hospital days</th>
<th>Additional clinical symptoms *</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-GGA</td>
<td>235</td>
<td>Human Herpesvirus 6 Torque teno virus</td>
<td>7 mos</td>
<td>F</td>
<td>19-Aug-2007</td>
<td>2</td>
<td>petechiae</td>
</tr>
<tr>
<td>2-CCA</td>
<td>451</td>
<td>Human Herpesvirus 6 Torque teno virus</td>
<td>1 yr</td>
<td>M</td>
<td>22-Oct-2008</td>
<td>2</td>
<td>petechiae, cough, distal coldness, hoarse, no appetite, tourniquet test positive</td>
</tr>
<tr>
<td>3-ATTA</td>
<td>207</td>
<td>Human Herpesvirus 6</td>
<td>9 mos</td>
<td>M</td>
<td>4-May-2007</td>
<td>2</td>
<td>petechiae, cough</td>
</tr>
<tr>
<td>4-GAG</td>
<td>432</td>
<td>Human Herpesvirus 6</td>
<td>1 yr</td>
<td>M</td>
<td>8-Oct-2008</td>
<td>2</td>
<td>petechiae, cough</td>
</tr>
<tr>
<td>4-TTA</td>
<td>574</td>
<td>Human Herpesvirus 6</td>
<td>9 mos</td>
<td>M</td>
<td>4-Jan-2009</td>
<td>2</td>
<td>facial rash</td>
</tr>
<tr>
<td>4-CCA</td>
<td>370</td>
<td>Human Herpesvirus 6</td>
<td>1 yr</td>
<td>F</td>
<td>14-Jul-2008</td>
<td>3</td>
<td>petechiae, no appetite, tourniquet test positive</td>
</tr>
<tr>
<td>4-AAA</td>
<td>118</td>
<td>Circovirus-like genome RW-E</td>
<td>6 yrs</td>
<td>F</td>
<td>30-Aug-2006</td>
<td>2</td>
<td>petechiae, ascites, cough, tourniquet test positive</td>
</tr>
<tr>
<td>4-ATT</td>
<td>323</td>
<td>Circovirus-like genome RW-E</td>
<td>1 yr</td>
<td>F</td>
<td>11-Dec-2007</td>
<td>4</td>
<td>cough</td>
</tr>
<tr>
<td>4-CAC</td>
<td>363</td>
<td>Circovirus-like genome RW-E</td>
<td>6 yrs</td>
<td>F</td>
<td>11-Jul-2008</td>
<td>4</td>
<td>petechiae, hepatomegaly</td>
</tr>
<tr>
<td>4-CGT</td>
<td>371</td>
<td>Circovirus-like genome RW-E</td>
<td>11 yrs</td>
<td>F</td>
<td>15-Jul-2008</td>
<td>4</td>
<td>arthralgia, cough, hepatomegaly, migraine, myalgia, no appetite, retro-orbital pain, tourniquet test positive</td>
</tr>
<tr>
<td>1-ATT</td>
<td>387</td>
<td>Circovirus-like genome RW-E</td>
<td>12 yrs</td>
<td>M</td>
<td>3-Aug-2008</td>
<td>4</td>
<td>cough, hepatomegaly, migraine, no appetite, splenomegaly, tourniquet test positive</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB virus C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-AAAA</td>
<td>73</td>
<td>Torque teno midi virus</td>
<td>2 yrs</td>
<td>M</td>
<td>22-Nov-2005</td>
<td>1</td>
<td>cough, hepatomegaly, no appetite, sensitive abdomen</td>
</tr>
</tbody>
</table>

* All patients had fever and rash, either on exam and/or by patient report

Table 3.4 Clinical information of patients with detectable virus sequence.
The clinical variables collected during hospital stay, including age, sex, illness onset date, hospital stay duration, and symptoms in addition to fever and rash.
Figure 3.4 HHV-6B genome coverage across six samples.
Histogram of HHV-6B genome coverage generated by aligning reads with minimum 90% identity over the total read length to the genome. The depth of sequence coverage was calculated as the total Kb of aligned sequence per 1 Kb bin over the HHV-6B reference genome. Inset text for each histogram is the pool/barcode information, as well as the total number of reads, the percentage of the dataset the number of aligned reads represents, and the total amount of aligned sequence.
Figure 3.5 Circovirus-like NI sequence coverage and phylogeny.
(A) Depth of coverage from sequences sharing BLASTx homology to the putative replicase (Rep) ORF of Circovirus-like RW-E from sample 387 (Pool 1, ATT) by amino acid (aa) position. A 220 nt (72 aa) amplicon (GenBank accession: JF781513) was recovered from the original sequencing library and sequenced by the Sanger method (denoted by red bar). The divergent circovirus sequence shared 66% aa identity (E value = 5 x 10^{-22}) by BLASTx to the putative Rep ORF of Circovirus-like RW-E (123), as depicted by the circular genome diagram in the inset. A grey arrow depicts the Rep ORF, an unknown
ORF is shown with a black arrow, and numbers indicate the nucleotide positions within the 2,782 nt-long circular genome.

(B) Phylogenetic neighbor-joining tree of amino acid sequences showing the relationship between Circovirus-like NI Rep sequence (red) and 28 other circovirus, circo-like virus, and cyclovirus species. The Rep aa sequences (67-74 amino acids in length) were aligned by MUSCLE. The neighbor-joining tree was generated by Geneious Tree Builder (42).

3.4 Discussion

In this study, we examined the virus diversity in serum samples from Nicaraguan children with unknown acute febrile illness. We performed Virochip microarray analysis on 150 serum nucleic acid samples, followed by complementary deep sequencing on a subset of 34 samples. We employed a barcoding strategy to pool serum-derived nucleic acid libraries to simultaneously sequence multiple samples, producing an average of 5.1 million 55-nt reads per sample. Virus sequence was detected in 12/34 samples, including sequence that shared identity with Human Herpesvirus 6 (HHV-6), Torque Teno Virus (TTV), GB virus C, and Circovirus-like virus. This study demonstrates the utility of deep sequencing as a strategy to detect virus sequence in multiple human serum samples and is the first study to utilize second-generation sequencing to simultaneously investigate many cases of acute unknown tropical illness.

Monitoring the emergence and spread of novel human pathogens in tropical regions is a central public health concern. Metagenomic analysis by both viral microarray and deep sequencing enables more systemic viral detection of both known and novel viral pathogens (17, 28, 38, 64, 85, 105, 144, 150, 152) and can be employed as diagnostic supplements to pathogen detection as part of public health monitoring systems and epidemiologic surveys (51, 142). In this study, the Virochip was utilized as a rapid, first-pass test, and deep sequencing was employed as a more comprehensive follow-up application. As a measure of the relative utility of these two methods, we asked whether deep sequencing could detect viruses not detected on the Virochip and vice versa. Three viruses (HHV-6, circovirus, and GB virus C, Table 3.3) were identified by deep sequencing but were not evident on the corresponding microarrays (Table 3.1). TTV was detected by Virochip in 12 samples. Two of these 12 were also deep sequenced, and TTV-derived reads were identified in one (sample 73) but not the other (sample 506). Using the assumption that a single sequence read suffices to identify a virus in a deep sequencing data set, we conclude that our barcoded sequencing approach, which produces 10^6 – 10^7 sequences per sample (Table 3.2), represents a more sensitive tool for virus detection and discovery in human serum. The Virochips, however, provides a relatively fast and inexpensive method that is perhaps best applied to samples with relatively high virus titers, such as virus isolation supernatant, stool, and respiratory samples. Thus, deep sequencing and the Virochip are complementary virus detection tools.

We applied a stepwise bioinformatic pipeline for deep sequencing analysis to filter low-quality, low-complexity, and human reads from sequencing data sets (Figure 3.3). The vast majority of nucleic acid extracted from human serum without major biochemical filtering or biased selection shares identity with human genome and transcriptome sequences. The iterative computational filtering sought to remove human and poor-quality reads in order to minimize computing time and maximize accuracy of subsequent virus sequence searches. On
average, we removed 96.7% of reads from each sample using seven steps of LZW compression analysis and BLAT and BLAST alignments to human sequences. We report our bioinformatic filtering strategy as an effective approach to reduce the untargeted noise from large clinical sequencing data sets preceding virus searches.

Determining the etiology of human diseases with symptoms that overlap with dengue-like illness is important to understanding the full spectrum of emerging or previously uncharacterized pathogens in tropical populations. HHV-6 sequences were detected in 6/34 serum samples in this study. Primary HHV-6 infection causes undifferentiated febrile illness and exanthem subitum (roseola infantum or sixth disease), an acute illness with high fever and rash that typically resolves in three to seven days (13). Exanthem subitum is a common disease of infants worldwide, and HHV-6 infection most frequently occurs between 6 and 12 months of age (113), with seropositivity estimates of >95% in adult populations in developed countries (36). Each of the six HHV-6 positive patients in this study were between 7-12 months old, presented with fever and rash, and were hospitalized for 2-3 days (Table 3.4). We detected 5-31 kilobases of HHV-6 sequence in each sample with sequence deriving from multiple viral genomic regions (Figure 3.4).

After acute infection, HHV-6 can latently persist in the host quiescently, with no production of infectious virions or with low-levels of viral replication. Latency is believed to endure in several cell types, including monocytes and bone marrow progenitor cells (87, 95), and may undergo chromosomal integration that can be vertically transmitted (33). The confounding effects of chromosomal integration make differentiating between active and latent HHV-6 infections difficult when detecting HHV-6 sequence in serum DNA (3, 25). A previous study detected integrated HHV-6 genomic sequence in ~1% of healthy blood samples (91). Since detection of HHV-6 nucleic acid in serum alone does not prove active viral infection, we cannot definitively confirm that the HHV-6 sequences in these samples were not derived from the vertical transmission of chromosomally integrated virus. However, the clinical, epidemiological, and virus sequence data suggests HHV-6 may be the etiologic agent in these six infant febrile illness cases.

Primary HHV-6 infection is a major cause (~20%) of infant hospitalizations in the United States (66), a clinical burden likely shared throughout the tropical world given similar seroprevalence rates (18). The results of this study illustrate the importance of administering additional HHV-6 tests to cases of suspected dengue-like illness in infants from dengue-endemic regions to differentiate between cases of exanthem subitum, a ubiquitous self-limiting childhood illness, and dengue fever, which carries a greater risk of severe clinical complications and death. In this study, 18% of acute serum samples negative for dengue virus from cases of pediatric dengue-like illness were positive for HHV-6.

This study also reports the discovery of divergent sequence from a novel virus related to the Circoviridae family and unclassified Circovirus-like viruses. Evidence of the circovirus-like virus, which we have denoted Circovirus-like NI (Cvl-NI), was detected in 5/34 samples. Partial Rep sequence of Cvl-NI shared 66% amino acid identity by BLASTx (E value = 5 x 10^{-22}) with the putative Rep ORF of Circovirus-like RW-E and likely represents a novel circular ssDNA virus species. The Circoviridae family is an extraordinarily diverse group of small, single-stranded circular DNA viruses that include: TTV (genus Anellovirus), cyclovirus (genus Cyclovirus), and circovirus (genus Circovirus), and are commonly detected in human stool and blood and environmental samples (70, 92, 123, 149). Some circovirus species, such as beak and feather disease virus and porcine circovirus 2, can infect and cause disease in bird and pig hosts,
respectively, but the pathogenic potential of circoviruses in humans remains unconfirmed. CvLN1 sequence was detected in nucleic acid libraries prepared from acute human serum and was most closely related to Circovirus-like viruses (Figure 3.5B), which were first reported in reclaimed water and marine samples (123). It has not yet been possible to prove that the CvLN1 sequence was not an artifact introduced during sample preparation or that CvLN1 can cause human infection and illness, questions that warrant additional screening and serologic studies. The divergent Circovirus-like NI and the known TTV isolates detected in these serum samples illustrate the need for continued investigation to determine the diversity, distribution, and disease associations of individual circovirus species. The high nucleotide substitution rates of ssDNA viruses (43) contribute to circovirus evolution and the potential emergence of novel human pathogens.

Metagenomic approaches provide an effective high-throughput method to detect uncharacterized circovirus diversity in a tropical setting from many samples simultaneously.

3.5 Acknowledgements

We thank the children and families participating in the Dengue Pediatric Hospital Study and the study staff and physicians at the Hospital Infantil Manuel de Jesús Rivera, in particular Crisanta Rocha, Sheyla Silva, Maria Angeles Pérez, Federico Narvaez, Gamaliel Gutierrez, and Julia Medina. We are also grateful to the study personnel at the Nicaraguan National Virology Laboratory and the Nicaraguan Ministry of Health, especially Angel Balmaseda, Juan Carlos Mercado, Juan Carlos Matute, Saira Saborio, Douglas Elizondo, and Andrea Nuñez. This work was supported in part by a grant from the Pacific Southwest Regional Center of Excellence (PSWRCE), NIH grant U54-AI65359, the Pediatric Dengue Vaccine Initiative (#VE-1 to EH). JD is supported by the Howard Hughes Medical Institute, the Doris Duke Foundation, and the Packard Foundation. PS-C is supported by a PhRMA foundation Pre Doctoral Informatics Fellowship.
Chapter 4
Conclusions and Future Directions

The findings presented in this thesis advance our understanding of known and previously uncharacterized viruses present in samples collected as part of pediatric illness studies in a tropical environment. In particular, our results describe the discovery and sequences of two novel viruses detected in cases of acute respiratory and dengue-like disease in Managua, Nicaragua and advance our understanding of the application of metagenomic approaches to human pathogen detection. Of course, there are additional questions raised by our work. In this chapter, our results are summarized and recommendations are presented for future avenues of investigation.

In Chapter 2, we were able to report the detection of a novel enterovirus, Enterovirus 109 (EV109), in multiple cases of acute respiratory illness in Nicaraguan children. We were also able to recover its complete genome sequence and describe its diversity within the Enterovirus genus. The sequence evidence suggests EV109 underwent an ancestral inter-species recombination event between a member of the Enterovirus C and Enterovirus A species. A more detailed examination using RNA folding software and homology modeling revealed conserved picornavirus hairpin motifs and external capsid amino acid diversity. Specific EV109 screening detected the virus in 5/310 (1.6%) nose and throat swab samples collected between June 2007 and June 2008 from a pediatric cohort study of influenza-like illness in Managua, Nicaragua. The experiments described in Chapter 3 continue the broad goal of defining the viral spectrum in cases of undiagnosed acute tropical illness using molecular diagnostics. In Chapter 3, we report the results of our attempt to comprehensively detect all virus sequence in 150 cases of unknown acute febrile illness in Nicaragua using the Virochip and deep sequencing methods. Using a barcoded sample-prep strategy, we simultaneously evaluated 34 serum samples using deep sequencing, generating an average of 5.1 million reads per sample in which to pursue virus sequence using our bioinformatic filtering pipeline and direct BLAST searches. We detected known and divergent virus sequence in 35% (12/34) of previously negative samples, including five samples that contained sequence from a putative novel virus related to the Circoviridae family, Circovirus-like virus NI (Cvl-NI).

A major outstanding question of this dissertation is whether the novel viruses described in this work, EV109 and Cvl-NI, are responsible for human disease. The presence of a virus associated with an illness is, after all, only a preliminary step towards proving causality. The Henle-Koch postulates long defined the criteria to investigate causal relationships between pathogens and disease, but have been subsequently modified to consider additional evidence, such as detecting neutralizing antibodies and microbial nucleic acid (93). EV109 is a member of the Enterovirus genus and is closely related to known pathogens, including Coxsackievirus A21 and Poliovirus. Though the current findings, including the phylogenetic relatedness of EV109, may suggest a pathogenic role, the question is complicated by the wide spectrum of illnesses associated with members of the Enterovirus genus, including: asymptomatic infection, gastroenteritis, respiratory disease, encephalitis/meningitis, and acute flaccid paralysis. Determining where EV109 lies on the Enterovirus disease spectrum requires extensive global
sample screening of both asymptomatic and symptomatic subjects with varying clinical syndromes, an endeavor that may require several extensive studies. We have begun initial EV109 screening studies to test for viral nucleic acid in clinical samples using EV109-specific PCR directed at the divergent VP1 region and have shared these reagents with partner labs in Nicaragua and South Korea. We have also tested nose/throat swabs from influenza-like illness cases, stool samples from cases of gastroenteritis, and cerebrospinal fluid (CSF) from aseptic meningitis cases collected in California, and CSF from encephalitis cases in India (Table 4.1). Thus far, none of these samples tested positive for EV109. Two EV109 cases have been reported by others, however, including a case of respiratory illness tested by the California Department of Public Health that contained enterovirus sequence sharing ~95% nucleotide identity to EV109 in the VP1 region (Charles Chiu, personal communication). Interestingly, a recent report also detected putative EV109 (and no other enteroviruses) in a rectal swab from one fatal case as part of an acute flaccid paralysis outbreak in the Congo from October to November 2010 (59). The Congo EV109 isolate shared 75-77% nt sequence identity with the VP1 region of the Nicaraguan isolates and 90% nt identity in the 3D region. Though the other positive cases examined thus far in the outbreak are wild-type Poliovirus 1, additional investigation is needed to evaluate the role of EV109 in acute flaccid paralysis. Previous studies on poliovirus evolution has suggested that poliovirus first evolved from other enterovirus C species and may continue to recombine with enteroviruses in the wild while maintaining its ability to cause debilitating illness (78). The grave implications of enterovirus emergence and recombination amidst ongoing polio eradication campaigns warrant strict attention.

Table 4.1 EV109 screening

<table>
<thead>
<tr>
<th>Sample Origin</th>
<th>Specimen</th>
<th>Illness</th>
<th>Number screened</th>
<th>EV109 positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nicaragua</td>
<td>Nose/throat swab</td>
<td>Influenza-like</td>
<td>314</td>
<td>5</td>
</tr>
<tr>
<td>California</td>
<td>Nose/throat swab</td>
<td>Influenza-like</td>
<td>95</td>
<td>0</td>
</tr>
<tr>
<td>California</td>
<td>Stool</td>
<td>Gastroenteritis</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>California</td>
<td>Cerebrospinal fluid</td>
<td>Aseptic meningitis</td>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>India</td>
<td>Cerebrospinal fluid</td>
<td>Encephalitis</td>
<td>96</td>
<td>0</td>
</tr>
<tr>
<td>CA Dept PH</td>
<td>Nose/throat swab</td>
<td>Influenza-like</td>
<td>ND</td>
<td>1</td>
</tr>
<tr>
<td>Congo</td>
<td>Rectal swab</td>
<td>Acute flaccid paralysis</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>Global</td>
<td>Many</td>
<td>Asymptomatic</td>
<td>ND</td>
<td>ND</td>
</tr>
</tbody>
</table>

Table 4.1 EV109 screening.
The number of samples tested for EV109 from multiple geographic origins, specimen types, and human illnesses. Samples denoted as originating in Nicaragua, California, or India were tested in our laboratory, whereas the California Department of Public Health and Congo results were reported by other groups. ND = no data.
Whereas our EV109 investigation was greatly aided by its close relationship to poliovirus, a well-studied model RNA virus, inferring characteristics about Cvl-NI based on its limited sequence homology and incomplete genome is exceedingly more challenging. The Circoviridae family is an extremely diverse group of single-stranded circular DNA viruses with many individual species and genera that are currently taxonomically undefined and incompletely sequenced. Additionally, the tropism and lifecycle of most Circoviruses are presently poorly characterized, with the exception of a few species known to adversely affect the agricultural industry. The family does not yet contain any confirmed human pathogens. Nonetheless, diverse Circoviruses species, as reported here and elsewhere, have been detected by metagenomic analysis of environmental and human samples (19, 92, 123). It remains to be determined if Cvl-NI or other Circoviruses are capable of replicating in human cells and causing acute or chronic illness, but their presence in deep sequencing data sets provokes the obvious question: are these small viruses implicated in disease, or innocuous bystanders? Complete genome recovery is a preliminary step in downstream Cvl-NI analysis, but unfortunately the sequence coverage in our nucleic acid libraries was too low to permit complete computational sequence assembly or allow long-range PCR amplification. Several attempts using long-range and inverse PCR, as well as rolling circle amplification (65) failed to amplify the whole genome from serum nucleic acid samples. Repeating RCA on the original serum samples or re-extracted and ribonuclease-treated material may successfully amplify the genome, if it is in fact circular, though we have not yet attempted these experimental modifications.

Finally, our work also explored the broad utility of genomic approaches to pathogen detection in a resource-limited setting. The past few years have seen the rapid development of second-generation sequencing technologies, which can currently generate many thousands to millions of sequencing reads per experiment (11). Deep sequencing analysis of clinical samples holds tremendous promise as a diagnostic tool by permitting the detection of many different viruses simultaneously, including those present at low levels and of divergent origin. Major remaining barriers to high-throughput sequencing strategies becoming standard diagnostic practice include their high cost, lengthy sample preparation time, and data analysis requirements, limitations that are magnified in resource-poor settings such as Central America. These limitations, however, are gradually being addressed. Industry hardware and technical advancements have steadily decreased the per-base cost of deep sequencing, and the results presented here advance our understanding and expectations of multiplexed sample preparation and bioinformatic data filtering within the framework of a current second-generation sequencing platform. In Chapter 2, we described the application of deep sequencing to aid in quickly recovering the complete genome of a novel picornavirus, and in Chapter 3, we used deep sequencing and the Virochip microarray to characterize the viral spectrum in acute tropical febrile illness. In both studies, we were able to employ genomic methods of virus detection to add to our understanding of viral diversity in Nicaraguan pediatric illness and inform the development and policy of specific virus testing. Our collaboration allowed for rapid outreach and the transfer of PCR-based assays to test for EV109 and other relevant viral pathogens on-site at the National Virology Laboratory in Managua using existing machines and reagents. We plan to continue studies investigating the viral spectrum of unknown dengue-like illness and respiratory illness, building upon these results. Specifically, we plan to test and incorporate additional biochemical and enzymatic sample preparation strategies to increase the virus-to-host sequence ratio, thereby increasing deep sequencing sensitivity and facilitating the pooling of hundreds of samples per experiment, decreasing the cost per sample. We also plan to continue to
examine the epidemiology, seroprevalence, and virological characteristics of novel and divergent viruses discovered in our future studies.

In conclusion, I believe the results presented in this thesis expand our understanding of the virus diversity in pediatric illness in Nicaragua and the application of genomic detection techniques in a tropical setting, findings that are particularly valuable given the pressing need for improved global emerging pathogen surveillance.
References


a Novel Coronavirus, Coronavirus HKU1, from Patients with Pneumonia. J. Virol. 79:884-895.


