Enhancing Virus Surveillance through Metagenomics:

Water Quality and Public Health Applications

by

Karyna Rosario-Cora

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
College of Marine Science
University of South Florida

Major Professor: Mya Breitbart, Ph.D.
    John Cannon, Ph.D.
    John Paul, Ph.D.
    Kathleen Scott, Ph.D
    Ted Van Vleet, Ph.D.

Date of Approval:
October 28, 2010

Keywords: Enteric Viruses, Bioindicators, Wastewater, Reclaimed Water, Fecal Pollution, Single-stranded DNA Viruses

Copyright © 2010, Karyna Rosario-Cora
DEDICATION

Este logro está dedicado a mi familia. A mi esposo, Luke A. LeMond, por entender el sacrificio que implica completar un doctorado y por su apoyo incondicional durante todos estos años. A mis padres, Myrna L. Cora y Rafael Rosario, por hacerme entender lo importante que es ir tras tus metas sin importar que tan incansables sean, aunque eso conlleva alejarme físicamente de la familia. A mis hermanas, Shakyra Rosario Cora y Yahayra Rosario Cora, por siempre creer en mí y por su amistad porque más que hermanas somos amigas. Gracias a todos por siempre estar ahí para escucharme, apoyarme en todo y dejarme volar…
ACKNOWLEDGEMENTS

This dissertation would not be possible without the guidance of my major professor, Dr. Mya Breitbart. There are no words that can express my gratitude to Dr. Breitbart, an advisor like no other. Her commitment to students, good science, flexibility, and sense of humor have made my Ph.D. journey a fun and unforgettable experience. I would also like to thank my committee members, Dr. Kathleen Scott, Dr. John Cannon, Dr. John Paul and Dr. Ted Van Vleet for all their advice and support throughout my graduate work.

I am really grateful for all the help and friendship that my labmates (Camille Daniels, Terry Ng, Erin Symonds, Dawn Goldsmith, Kim Pause, Darren Dunlap, and Bhakti Dwivedi) offered throughout the years. Their support and advice contributed to the completion and quality of this work.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>CHAPTER 1: INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Background</td>
<td>2</td>
</tr>
<tr>
<td>Virus Surveillance</td>
<td>2</td>
</tr>
<tr>
<td>Limitations of Current Methods Used for Virus Surveillance</td>
<td>5</td>
</tr>
<tr>
<td>Metagenomics for Discovery and Community Analysis</td>
<td>6</td>
</tr>
<tr>
<td>Overall Research Objectives</td>
<td>7</td>
</tr>
<tr>
<td>Research Overview</td>
<td>8</td>
</tr>
<tr>
<td>References</td>
<td>12</td>
</tr>
<tr>
<td>CHAPTER 2: RESEARCH IMPACTS AND CONCLUSIONS</td>
<td>20</td>
</tr>
<tr>
<td>Research Impacts</td>
<td>21</td>
</tr>
<tr>
<td>Viral Metagenomics and Clinical Surveillance</td>
<td>21</td>
</tr>
<tr>
<td>Viral Metagenomics and Water Quality</td>
<td>24</td>
</tr>
<tr>
<td>References</td>
<td>29</td>
</tr>
<tr>
<td>APPENDIX A: METAGENOMIC SEQUENCING FOR VIRUS IDENTIFICATION</td>
<td>31</td>
</tr>
<tr>
<td>IN A PUBLIC HEALTH SETTING</td>
<td></td>
</tr>
<tr>
<td>APPENDIX B: METAGENOMIC ANALYSIS OF VIRUSES IN RECLAIMED WATER</td>
<td>32</td>
</tr>
<tr>
<td>APPENDIX C: PEPPER MILD MOTTLE VIRUS AS AN INDICATOR OF FECAL POLLUTION</td>
<td>33</td>
</tr>
<tr>
<td>APPENDIX D: DIVERSE CIRCOVIRUS-LIKE GENOME ARCHITECTURES</td>
<td>34</td>
</tr>
<tr>
<td>REVEALED BY ENVIRONMENTAL METAGENOMICS</td>
<td></td>
</tr>
<tr>
<td>APPENDIX E: METHOD DEVELOPMENT FOR THE CHARACTERIZATION OF SINGLE-STRANDED DNA VIRUSES</td>
<td>35</td>
</tr>
<tr>
<td>APPENDIX F: AUTHOR CONTRIBUTIONS AND COPYRIGHT CLEARANCES</td>
<td>36</td>
</tr>
</tbody>
</table>
ABSTRACT

Monitoring viruses circulating in the human population and the environment is critical for protecting public and ecosystem health. The goal of this dissertation was to incorporate a viral metagenomic approach into virus surveillance efforts (both clinical and water quality control programs) to enhance traditional virus detection methods.

Clinical surveillance programs are designed to identify and monitor etiological agents that cause disease. However, the ability to identify viruses may be compromised when novel or unsuspected viruses are causing infection since traditional virus detection methods target specific known pathogens. Here we describe the successful application of viral metagenomics in a clinical setting using samples from symptomatic patients collected through the Enterovirus Surveillance (EVS) program in the Netherlands (Appendix A). Despite extensive PCR-based testing, the viruses in a small percentage of these samples (n = 7) remained unidentified for more than 10 years after collection. Viral metagenomics allowed the identification of viruses in all seven samples within a week using minimal sequencing, thus rapidly filling the diagnostic gap. The unexplained samples contained BK polyomavirus, Herpes simplex virus, Newcastle disease virus and the recently discovered Saffold viruses (SAFV) which dominated the unexplained samples (n = 4). This study demonstrated that metagenomic analyses can be added as a routine tool to investigate unidentified viruses in clinical samples in a public-health setting. In addition, metagenomic data gathered for SAFV was used to complete four
genotype 3 SAFV (SAFV-3) genomes through primer walking, doubling the number of SAFV-3 full genomic sequences in public databases.

In addition to monitoring viruses in symptomatic patients, it is also important to monitor viruses in wastewater (raw and treated) to protect the environment from biological contamination and prevent further spread of pathogens. To gain a comprehensive understanding of viruses that endure wastewater treatment, viral metagenomics was used to survey the total DNA and RNA viral community in reclaimed water (the reusable end-product of wastewater treatment) (Appendix B). Phages (viruses that infect bacteria) dominated the DNA viral community while eukaryotic viruses similar to known plant and insect viruses dominated RNA metagenomic libraries suggesting that highly stable viruses may be disseminated through this alternative water supply. A plant virus, the *Pepper mild mottle virus* (PMMoV), was identified as a potential indicator of wastewater contamination based on metagenomic data and quantitative PCR assays (Appendix C). The metagenomic analysis also revealed a wealth of novel single-stranded DNA (ssDNA) viruses in reclaimed water. Further investigation of sequences with low-level similarities to known ssDNA viruses led to the completion of ten novel ssDNA genomes from reclaimed water and marine environments (Appendix D). Unique genome architectures and phylogenetic analysis suggest that these ssDNA viruses belong to new viral genera and/or families. To further explore the ecology of the novel ssDNA viruses, a strategy was developed to take metagenomic analysis to the next level by combining expression analysis and immunotechnology (Appendix E). This dissertation made a significant contribution to current microbiological data regarding
wastewater by uncovering viruses that endure the wastewater treatment and identifying a new viral bioindicator.
CHAPTER 1:

INTRODUCTION
Background

Virus Surveillance

The detection of pathogenic viruses circulating in the human population is a fundamental component of public health monitoring schemes to minimize risks associated with pathogen transmission. In clinical settings it is critical to monitor viral pathogens for early detection and intervention (1, 2). For this purpose, pathogens are usually monitored by collecting samples from patients expressing disease symptoms. However, the emergence of epizootic diseases has necessitated the ecological surveillance of viral pathogens in order to identify conditions that may lead to human infection (e.g. (3-5)). Ecological surveillance involves the detection and identification of epizootic viruses in animal hosts and environmental reservoirs, including vectors responsible for transmitting the viruses (e.g. (6-8)). Since pathogenic viruses can be shed in high numbers in human feces (9-12), wastewater is considered an important source of pathogens to the environment (13, 14). Therefore it is important to monitor viruses in wastewater (raw and treated) to protect environments exposed to wastewater discharges from biological contamination and prevent further spread of pathogens. This type of surveillance is accomplished through water quality control programs. Although surveillance efforts in both clinical and water quality control programs aim to limit pathogen dissemination, their overall objective is different and thus the approaches used for monitoring viruses differ.

Clinical surveillance programs provide vital information for improving disease management strategies such as vaccine implementation and vector control (e.g. (15-17)). Therefore it is of utmost importance to identify and monitor etiological agents that cause
disease. The detection and identification of viruses in clinical samples relies on a range of traditional and modern techniques including cell culture, electron microscopy, serology and molecular assays designed to screen for specific viral species or closely related viruses (18). The combination of these molecular techniques is often successful but fails to produce conclusive results when novel viruses or divergent variants of a known viral family are involved. This may result in a high percentage (up to 50%) of cases with unidentified etiological agents (19-22). Moreover, new viruses are frequently discovered, creating the need to constantly update PCR assays (23-26). Microarrays have been proposed as a more sensitive tool to monitor viruses with high mutability and have been shown to provide significant advantages over isolation in culture, immunoassays, and PCR-based methods (27-29). Notably, microarrays can be used to simultaneously test for multiple species of viruses, offering an opportunity for massively parallel virus surveillance (30). However, microarrays often fail to detect etiological agents in clinical samples from patients suspected to have viral infections (28). New approaches are needed to identify divergent and novel viruses that standard clinical surveillance methods fail to detect.

In contrast to clinical scenarios, water quality control programs that monitor the microbiological quality of wastewater are not designed to detect specific pathogens. The main goal of water quality programs is to limit the dissemination of fecal-associated pathogens through wastewater discharges into the environment, consequently minimizing public exposure to these pathogens. Due to the wide variety of possible pathogens and the large number of samples that must be tested on a frequent basis, current quality control methods do not test the presence of pathogens directly (31). Instead, standard detection
methods focus on detecting indicator organisms that serve as surrogates for pathogenic organisms (32). It has been shown that bacterial indicators currently used for microbiological monitoring of wastewater, such as fecal coliforms and Enterococci, are easily inactivated compared to viruses and often do not correlate with the occurrence of viral pathogens in wastewater (33-39). These findings have led several scientists to propose a suite of viral bioindicators, including coliphage, human adenoviruses and polyomaviruses, as a more sensitive tool to detect viral pathogens (37, 39-41).

Due to the historical dependence on indicator organisms to monitor water quality, the virological content of wastewater is still largely unknown. Most microbiological and risk assessment studies regarding wastewater have examined either indicator organisms or specific human pathogens (42). However, wastewater may also be a reservoir for non-human pathogens that are present in human waste. For example, it has been shown that plant viruses dominate the RNA viral community in human feces (43). Other studies have identified animal rotavirus strains of unknown origin co-circulating with human strains in sewage and treated wastewater (44, 45). Furthermore studies investigating the viral community in stool from South Asian children have identified an abundance of novel picornaviruses related to animal and insect pathogens (46, 47). Therefore, in addition to human pathogens, the diverse viral flora in human feces may contain plant, insect, and animal viruses. Since wastewater is ultimately discharged into the environment, it is relevant for ecosystem health to evaluate if these non-human pathogens are also present in wastewater. If current treatments fail to remove plant and insect pathogens from wastewater this may have implications for the use of treated effluent (i.e. reclaimed water) for agricultural irrigation. In order to protect public and environmental health,
there is a need to assess the total viral community in wastewater instead of focusing exclusively on human viruses.

**Limitations of Current Methods Used for Virus Surveillance**

The methods conventionally used for virus surveillance are limited as these methods heavily depend on *a priori* knowledge of the viruses that are being targeted. In clinical settings if novel or unsuspected viruses are causing infection, then current methods that target a specific virus or group of viruses will fail to identify the etiological agent. Although water quality control programs aim to detect bioindicators rather than specific pathogens, the scarce knowledge regarding the viral content in wastewater demonstrates the need to survey the viral community in wastewater. Information from these surveys can later be used to identify new and improved viral bioindicators, test the efficacy of different wastewater treatment techniques, and evaluate potential impacts on ecosystem health. However, current methods in water quality programs do not allow for total viral community analyses since no single monitoring assay can target all viruses. In view of the limitations of methods that target specific viruses, new approaches are required for the identification of novel or “unsuspected” viruses in surveillance schemes. One promising approach for viral identification and community analysis is the use of virus particle purification and metagenomic sequencing (viral metagenomics).
Metagenomics for Virus Discovery and Viral Community Analysis

Traditional approaches used to identify viruses in surveillance programs are problematic because they require prior knowledge of the viruses that are being targeted. For example, culture-based methods depend on the availability and selection of appropriate susceptible host cells to propagate viruses. Molecular approaches, such as PCR assays, require knowledge regarding genomic information for primer design. Moreover, viruses do not have ubiquitously conserved genetic elements such as ribosomal DNA that can be used to amplify and identify all viruses (48). Thus, there is no universal PCR assay that can target all the viruses in a sample. Although microarrays can be used for massive parallel detection of multiple viral species, the assay still depends on oligoprobes based on known viruses and, thus, may fail to detect divergent viral species. Furthermore each microarray will only detect viruses that can bind to the oligoprobes included in the assay (30). These limitations make traditional methods insufficient for virus surveillance strategies that need to characterize novel viruses in clinical samples or investigate the total viral community found in wastewater.

Metagenomic (whole community) analyses offer an opportunity to circumvent limitations found in current assays used for virus surveillance and directly describe the composition and structure of uncultured viral communities. In contrast to specific assays that are designed to recover a single virus or a group of closely related viruses, viral metagenomics allows the identification of viruses in a sample without a priori knowledge of the viral types present (49, 50). It is important to distinguish viral metagenomics, where viruses are purified before shotgun sequencing and yield of viral sequences is high (51), from a direct metagenomics approach, where total homogenates are sequenced and
viruses only account for a small proportion of the sequences (52). Viral metagenomics has been used to describe viruses in mammalian feces (43, 47, 53-57), cell cultures (46, 58), respiratory tract aspirates (25, 59), blood (58, 60), and animal tissues (61, 62) as well as to characterize viral communities present in different environments (63-67). Applying a viral metagenomic approach into virus surveillance efforts will allow the description of unidentified viruses in clinical samples and characterization of the entire viral community in wastewater, as opposed to using specific assays for a limited number of viruses. This novel approach will enhance and complement current methods used for virus surveillance.

**Overall Research Objectives**

The overarching goal of this research project was to apply viral metagenomics in virus surveillance efforts including clinical and water quality control programs. The project objectives included:

1) Characterization of unidentified viruses in samples collected through an established clinical surveillance program

2) Description of the complete DNA and RNA viral community in treated wastewater

3) Identification of a potential new viral bioindicator for water quality assessments

4) Characterization of newly described single-stranded DNA (ssDNA) viruses identified in wastewater
Research Overview

Viral metagenomics was used to enhance current methods for virus surveillance in clinical and water quality control programs. The information gathered during the project was used to make recommendations for clinical surveillance programs and add to the current microbiological data regarding viruses in wastewater. The objectives of the project were accomplished through the following studies:

Appendix A: Metagenomic Sequencing for Virus Identification in a Public Health Setting

- This study describes systematic analysis of 1834 clinical specimens cultured from symptomatic patients as part of the Enterovirus Surveillance program in the Netherlands through a combination of PCR-based assays and viral metagenomics. During the investigated 13-year period (1994-2007), a total of seven samples exhibited reproducible cytopathogenic effects in cell culture and tested negative for standard PCR assays and, thus, remained unexplained. In order to fill the diagnostic gap, metagenomic sequencing was applied to virus particles purified from the unexplained cell culture samples. Viral metagenomics resulted in the rapid identification of viruses in all the samples with minimal sequencing. This study demonstrated that viral metagenomics is a powerful tool that can be integrated into public health monitoring efforts to investigate unexplained infections that standard PCR assays fail to detect. In addition, four
genomes of the recently discovered Saffold virus were completed and reported as part of this work.

**Appendix B: Metagenomic Analysis of Viruses in Reclaimed Water**

- The use of reclaimed water (i.e., the reusable end product of wastewater treatment) is an important component of sustainable water resource management. The primary goal of this study was to characterize the DNA and RNA viral community found in reclaimed water in an effort to detect viruses that endure the wastewater treatment process. The data suggested that reclaimed water may play a role in the dissemination of highly stable plant viruses and other novel viruses that have not been previously described. Stable viruses found in reclaimed water may share the same resistance to wastewater treatment processes as some pathogens of concern to public health. Therefore reclaimed water represents an untapped source for discovering new bioindicators that can serve as surrogates for pathogen detection in water supplies.

**Appendix C: Pepper Mild Mottle Virus as an Indicator of Fecal Pollution**

- Accurate indicators of fecal pollution are needed to minimize public health risks associated with wastewater contamination in recreational waters. A plant virus, *Pepper mild mottle virus* (PMMoV), was identified as a potential novel indicator of wastewater contamination metagenomic sequencing of reclaimed water. Quantitative PCR assays showed that PMMoV is widespread and abundant in wastewater throughout the United
States, including both raw sewage and treated effluent. Further testing revealed the presence of PMMoV in seawater samples collected near point sources of secondary-treated wastewater where it co-occurred with several human pathogens and other proposed indicators of fecal pollution. The data collected during this study demonstrated that PMMoV is a promising indicator of fecal pollution in coastal environments.

**Appendix D:** Diverse Circovirus-like Genome Architectures Revealed by Environmental Metagenomics

- One of the biggest advantages of metagenomics is the potential to uncover new viruses directly from environmental communities without having prior knowledge of their existence. The reclaimed water DNA viral metagenome contained a wealth of sequences with low levels of similarities to single-stranded DNA (ssDNA) viruses. This study focused on further investigating these sequences through the complete sequencing of five novel genomes. Data-mining of environmental datasets resulted in the completion of an additional five genomes with similar characteristics from metagenomic libraries of three different marine environments (Chesapeake Bay, British Columbia coastal waters, and Sargasso Sea). The ten novel genomes shared similarities with ssDNA circoviruses; however, only half exhibited genomic features consistent with known circoviruses. Unique genome architectures and phylogenetic analysis suggest that these viruses belong to new viral genera and/or families. This
study revealed an unprecedented diversity of ssDNA viruses with unique features in the environment.

Appendix E: Method Development for the Characterization of Novel Single-stranded DNA Viruses

- Metagenomic analyses have led to the discovery of a diversity of unknown ssDNA viruses in the environment. Since the novel ssDNA viruses have been identified directly from environmental sequence datasets and the hosts are unknown we are extremely limited in our ability to characterize these viruses. This study aimed to take metagenomic analysis to the next level by combining expression analysis and immunotechnology to describe novel ssDNA viruses in reclaimed water. Recombinant expression of structural proteins would allow us to isolate unknown viruses from environmental samples by virtue of the antigenic properties of these proteins. For this purpose, the potential structural genes of novel viruses were identified from assembled genomes based on conserved genome organizations. However, the expression of the environmental ssDNA viruses proved to be difficult in the selected bacterial expression system and future studies should evaluate eukaryotic expression systems. This research will provide a method for isolation and characterization of unknown ssDNA viruses in the environment based on structural genes found in metagenomic libraries.
References


42. **Westrell, T., C. Schonning, T. A. Stenstrom, and N. J. Ashbolt.** 2004. QMRA (quantitative microbial risk assessment) and HACCP (hazard analysis and critical control points) for management of pathogens in wastewater and sewage sludge treatment and reuse. Water Science and Technology **50**:23-30.


metagenomic survey of microbes in honey bee colony collapse disorder. Science 318:283-287.


CHAPTER 2:
RESEARCH IMPACTS AND CONCLUSIONS
Research Impacts

Virus surveillance programs are an integral component of public health monitoring efforts designed to minimize public exposure to pathogens. However, monitoring efforts are limited in their ability to detect pathogens since current methods specifically test for a limited number of pathogens, thus relying on prior knowledge of the targeted viruses. The goal of this project was to incorporate a virus metagenomic approach into virus surveillance schemes, including clinical and water quality control programs, to circumvent methodological limitations traditionally associated with virus detection. Metagenomics enhanced viral surveillance in a clinical setting by enabling the characterization of infectious viral agents in unidentified clinical specimens. In addition, sequencing of the viral metagenome from reclaimed water (the reusable end-product of wastewater treatment) described viruses that endure wastewater treatment, identifying new viral bioindicators that can be used to improve water quality monitoring and uncovering viruses that are not commonly thought to be associated with wastewater.

Viral Metagenomics and Clinical Surveillance

The application of viral metagenomics in a clinical setting was tested using samples collected through the Enterovirus Surveillance (EVS) program in the Netherlands. In its current form, the EVS program monitors the circulation of enteroviruses to ensure the eradication of poliovirus from the Netherlands (1, 2). Although routine virological laboratories across the country perform a range of tests for enteroviruses, all cell culture samples exhibiting cytopathogenic effects (CPE) that cannot be identified are submitted to the Center for Infectious Disease Control, National
Institute for Public Health and the Environment (RIVM) in the Netherlands for virus identification and typing by PCR. Between 1994 and 2007, a total of 1742 unidentified cell culture isolates exhibiting consistent CPE were submitted to RIVM. PCR assays for enteroviruses and parechoviruses performed at RIVM successfully identified viruses in approximately 98% of these isolates. The remaining 2% of the samples were subjected to extensive PCR testing for a wide range of viruses, including noroviruses, rotaviruses A, B and C, adenoviruses, astroviruses, sapoviruses, vesiviruses, reoviruses, a generic PCR that detects both enteroviruses and rhinoviruses, hepatitis A and E viruses, influenza A and B viruses, Aichi virus, coronaviruses 229E, NL63 and OC43, human respiratory syncytial viruses A and B and human metapneumovirus. In addition, samples that remained negative had to be retrospectively analyzed every time PCR assays were updated to include a broader diversity of viruses. The process of successive PCR testing for each of these viruses was extremely time-consuming and labor-intensive, with diminishing returns as some of these virus groups were not detected in any of the samples.

The EVS program needs to identify all viral agents associated with clinical specimens collected throughout the Netherlands to ensure that infections are not associated with poliovirus and track enteric viruses circulating in the population. However, extensive PCR testing failed to identify viruses in seven cell culture samples that exhibited consistent CPE. Some of the samples from the EVS program remained unidentified for more than 10 years after collection, despite frequent updates to improve the PCR assays over time. As presented in this dissertation, viral metagenomics allowed the identification of viruses in all seven samples within a week using minimal sequencing
(< 50 clones per sample). The unexplained samples contained BK polyomavirus, Herpes simplex virus, Newcastle disease virus and the recently discovered Saffold viruses (SAFV) which dominated the unexplained samples \( (n = 4) \). Although all of these viruses had been previously described, they were not included in assays performed at RIVM because they were not suspected to cause an infection in the samples collected. This assumption precluded the identification of viruses in all the samples, keeping the EVS program from meeting its goal. Metagenomic analysis does not require prior knowledge of the viruses in a sample for identification and, thus, this approach was successful in identifying infectious agents in all the samples.

Although viral metagenomics was originally developed to describe total viral communities in environmental samples (3), this dissertation proved the effectiveness of viral metagenomics as a strategy for clinical surveillance. Viral metagenomics was a more efficient and cost-effective alternative than individual PCR assays for different viruses, other than enteroviruses and parechoviruses, in the EVS program. In addition, the discovery of genotype 3 SAFV (SAFV-3) in the EVS samples contributed to genomic data regarding these newly discovered RNA viruses. Prior to the completion of the SAFV-3 found in the EVS samples there were only three SAFV-3 genomes in public databases. The metagenomic data gathered for SAFV was used to complete four SAFV-3 genomes through primer walking, thus doubling the number of SAFV-3 full genomic sequences in the database. This study demonstrated that metagenomic analyses can be added as a routine tool to investigate unidentified viruses in cell cultures from clinical samples in a public-health setting. Therefore, we recommend the use of viral
metagenomics on cell culture samples that test negative for established PCR assays in virus surveillance programs.

**Viral Metagenomics and Water Quality**

Viral metagenomics was used to describe the total DNA and RNA viral community in reclaimed water. Reclaimed water is an important component of water reuse programs that aim to reduce the discharge of wastewater effluent into surface waters and conserve water by supplying water for activities that do not require drinking water quality standards. However, this means that the public will be exposed to treated wastewater effluent through the use of reclaimed water for non-potable public water supply, agricultural irrigation, environmental enhancement and industrial uses. As water reuse applications increase and reclaimed water distribution expands, there are some concerns that need to be addressed to ensure protection of public health and the health of the environment, including the potential for pathogen dissemination through this alternative water supply. Since the virological content of wastewater is still largely unknown, viral metagenomics was used to survey the total viral community in an effort to gain a comprehensive understanding of viruses that endure wastewater treatment. The information gathered during this study was used to identify potential new bioindicators of fecal pollution and bring attention to non-human viruses that have been overlooked when considering water quality.

The DNA viral community in reclaimed water was dominated by phages (viruses that infect bacteria). After comparing this viral community with a potable water viral metagenome, reclaimed water had a distinct phage community based on phage family
distributions and host representation within each family. Therefore, although phages dominate both conventional and alternative water supplies, the types of phages in each water supply differ. From a water quality standpoint, it is useful to evaluate which types of phages endure wastewater treatment processes but are not present in potable water in order to identify potential viral bioindicators. Finding strong bioindicators that correlate with the presence of human viruses is not an easy task as different viruses exhibit varying levels of resistance to wastewater treatment (4). Natural phage populations found in wastewater offer a range of resistance to disinfection (chlorination) that may represent most of the viruses that can be found in sewage (5). Therefore, phage populations in reclaimed water offer an untapped source of potential bioindicators.

Alternatively, eukaryotic viruses, such as plant pathogens, may also be explored as potential viral bioindicators. Since plant pathogens found in human waste are suspected to be dietary in origin (6, 7), these viruses may be more abundant in the healthy human population than viruses that cause human disease. Data from this dissertation suggest that plant viruses may be good indicators of human fecal pollution. The plant pathogen *Pepper mild mottle virus* (PMMoV) was identified as a potential bioindicator due to its abundance in the reclaimed water RNA libraries and previous findings indicating that this virus dominates the RNA viral community in human feces (6). Quantitative PCR assays showed that PMMoV is widespread and abundant (> $10^4$ copies/ml) in wastewater (both raw and treated) across the United States. In addition, PMMoV was detected in seawater samples collected near point sources of secondary treated wastewater, where it co-occurred with several other pathogens and indicators of fecal pollution. PMMoV was not found in non-polluted seawater samples and could be
detected in surface seawater for approximately 1 week after its initial introduction, indicating that the presence of PMMoV in the marine environment reflects a recent contamination event. These findings suggest PMMoV is a promising indicator of fecal pollution in marine environments. However, the abundance of PMMoV in both raw sewage and treated wastewater demonstrated that this viral indicator cannot be used to distinguish between these two sources of wastewater. Instead, PMMoV serves as a conservative viral tracer of fecal pollution that can be used to represent microconstituents and pathogens that may not be removed effectively through wastewater treatment processes. The extremely high concentrations of PMMoV detected in human sewage (up to $10^7$ copies/ml) compared to the concentration of any human pathogen reported to date ($< 10^5$ copies/ml) (8-11) suggest that this virus would be a good indicator of human fecal pollution. As a plant pathogen, PMMoV is different from other proposed viral indicators of fecal contamination since its presence in sewage is dietary in origin and is not dependent on active human infection.

In addition to potential bioindicators, the reclaimed water metagenome uncovered a wealth of novel eukaryotic viruses related to viruses that are not commonly associated with wastewater. Notably, DNA metagenomic libraries revealed the presence of viruses similar to single-stranded DNA (ssDNA) animal viruses from the Circoviridae family, which are known avian and porcine pathogens. Further investigation of sequences related to circoviruses resulted in the completion of five novel circovirus-like genomes. The unprecedented abundance of these viral sequences in reclaimed water prompted a search for related viruses in other environmental samples. Data-mining of environmental datasets resulted in the completion of an additional five circovirus-like genomes from
metagenomic libraries of three different marine environments (Chesapeake Bay, British Columbia coastal waters, and Sargasso Sea). The ten novel genomes shared similarities with ssDNA circoviruses; however, only half exhibited genomic features consistent with known circoviruses. Some of the genomes exhibited a mixture of genomic features associated with different families of ssDNA viruses (i.e. circoviruses, geminiviruses and parvoviruses). The abundance of circovirus-like sequences in environmental metagenomic studies and the presence of unique genome sequences and architectures suggest that there is a complex and largely unexplored community of ssDNA viruses in the environment.

The identification of circovirus-like sequences in reclaimed water led to the discovery of novel circovirus genomes in wastewater and the marine environment for the first time. However, we are extremely limited in our ability to further characterize these novel viruses since they were identified directly from metagenomic data and the hosts are unknown. Isolation of novel ssDNA viral particles from the environment will allow further biological and physicochemical characterization of these unknown viruses. This task may be possible through recombinant expression of viral structural proteins and development of immunoassays, such as immunoprecipitation, to isolate native viruses from environmental samples by virtue of the antigenic properties of these proteins. However, in this dissertation, efforts to express and purify recombinant circovirus proteins in a bacterial system were unsuccessful. Refinement of recombinant expression strategies, such as using a eukaryotic expression system, will allow expression of divergent structural proteins and future isolation of unknown viruses from the environment. The combination of metagenomic sequencing, protein expression, and
immunotechnology is a promising strategy that will help overcome the biggest current limitation in metagenomic analyses, namely, the connection between new genes, viral particles, and biological properties.

Other than data on specific human pathogens, information regarding viruses in wastewater is extremely sparse. Since pathogen transport through wastewater is an important concern for public and environmental health, there was a need to survey the total viral community in wastewater to identify viral types that endure wastewater treatment. The metagenomic analysis of viruses in reclaimed water performed in this dissertation significantly contributed to current microbiological data regarding treated wastewater. The genetic information gathered during this study can be used to design molecular assays to detect viral types of interest and assess their abundance in wastewater and ecosystems exposed to wastewater discharge. The diversity of both phages and eukaryotic viruses suspected to be dietary in origin, such as PMMoV, may be examined to find new and improved viral bioindicators of fecal pollution. Since reclaimed water contained a wealth of novel single-stranded DNA and RNA viruses related to plant, animal and insect viruses, this alternative water supply may play a role in the dissemination of highly stable viruses. Future research needs to evaluate the host range, infectivity and ecological impacts of novel viruses identified in reclaimed water to ensure the appropriate use of this important alternative water supply.
References


APPENDIX A:

METAGENOMIC SEQUENCING FOR VIRUS IDENTIFICATION

IN A PUBLIC HEALTH SETTING
INTRODUCTION

Diagnostic assays for pathogen detection are a critical component of public-health monitoring efforts. The detection and identification of viruses in clinical samples rely on a range of traditional and modern techniques (Leland & Ginocchio, 2007). Traditionally, electron microscopy and cell culture have been used to identify the presence of viral agents. These traditional techniques have now been combined or replaced by molecular or serological assays designed to screen for specific known viruses. Subsequently, if no viral pathogens are found, more investigative methods, such as generic PCR assays that target viral groups rather than species-specific assays, are used (Svraka et al., 2007, 2009a, b). The combination of these molecular techniques is often successful, but fails to produce conclusive results when novel viruses or divergent variants of a known virus family are involved. Moreover, viruses do not have ubiquitously conserved genetic elements such as tDNA that can be used to amplify and identify all viruses and, thus, there is no universal PCR assay that can target all viruses in a sample (Roehver & Edwards, 2002). As novel viruses are discovered frequently, there is a need to constantly update PCR assays (Abed & Boivin, 2008; Allander et al., 2005; van der Hoek et al., 2004). In view of the limitations of methods that target specific viruses, new approaches are required for the identification of novel or ‘unsuspected’ viruses in public-health monitoring schemes. One promising approach for rapid virus identification is the use of virus purification and metagenomic sequencing (viral metagenomics).

In contrast to specific assays that are designed to recover a single virus or a group of closely related viruses, viral metagenomics can identify all viruses in a sample without a priori knowledge of the virus types present (Delwart, 2007; Edwards & Roehver, 2005). It is important to distinguish...
viral metagenomics, where viruses are purified before sequencing and yield of viral sequences is high (Thurber et al., 2009), from a direct metagenomics approach, where total homogenates are sequenced and viruses account for only a small proportion of the sequences (Cox-Foster et al., 2007). Viral metagenomics has been used to characterize virus communities present in the environment (Angly et al., 2006; Breitbart et al., 2004; Dinsdale et al., 2008; Dijkeng et al., 2009) and to describe viruses in mammalian faeces (Blinova et al., 2010; Breitbart et al., 2003, 2008; Cann et al., 2005; Li et al., 2010; Victoria et al., 2009; Zhang et al., 2006), cell culture (Jones et al., 2007; Kapoor et al., 2008), respiratory-tract aspirates (Allander et al., 2005; Willner et al., 2009) and blood (Breitbart & Rohwer, 2005; Jones et al., 2005), and animal tissues (Ng et al., 2009a, b).

The present study describes the use of a combination of molecular approaches, including specific and generic PCRs for various viruses as well as metagenomic sequencing, to identify viruses from clinical specimens in a public-health context. Virological data from 1834 cultured clinical specimens collected between 1994 and 2007 through the Enterovirus Surveillance (EVS) programme in the Netherlands are summarized. Metagenomic sequencing led to the identification of viruses in all samples that exhibited consistent cytopathogenic effects (CPE) and remained unexplained after PCR testing.

**RESULTS**

Characterization of isolates from enterovirus surveillance at RIVM

From 1994 to 2007, a total of 1834 cell-culture isolates were submitted to the RIVM to identify viruses in cell cultures with CPE. Upon passaging at RIVM, 92 cell-culture isolates did not show reproducible CPE on cell lines and thus were not investigated further (Table 1). Of the 1742 samples with reproducible CPE, 1513 were confirmed as being enterovirus-positive and 189 as parechovirus-positive.

<table>
<thead>
<tr>
<th>PCR assay</th>
<th>Total no. samples (1994–2007)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enterovirus</td>
<td>1513 (82.5)</td>
</tr>
<tr>
<td>Parechovirus</td>
<td>189 (10.3)</td>
</tr>
<tr>
<td>Rhinovirus</td>
<td>9 (0.5)</td>
</tr>
<tr>
<td>Adenovirus</td>
<td>14 (0.8)</td>
</tr>
<tr>
<td>Reovirus</td>
<td>9 (0.5)</td>
</tr>
<tr>
<td>Astrovirus</td>
<td>1 (0.1)</td>
</tr>
<tr>
<td>No CPE</td>
<td>92 (5.0)</td>
</tr>
<tr>
<td>No virus identified (consistent CPE)</td>
<td>7 (0.4)</td>
</tr>
<tr>
<td>Total no. samples</td>
<td>1834 (100)</td>
</tr>
</tbody>
</table>

Extended PCR testing resulted in the identification of other viral agents, such as adenoviruses, reoviruses, rhinoviruses and astroviruses, in 53 cell-culture samples (Table 1). The seven remaining cell-culture samples exhibiting consistent CPE tested negative for all viral PCR assays and were processed for viral metagenomics (Table 2).

**Viral metagenomics of cell-culture samples with unidentified viruses**

Viral metagenomics was performed on seven cell-culture samples where PCR assays failed to identify potential viral pathogens (Table 2). This approach allowed the identification of viral agents in all samples with minimal sequencing (<50 clones per sample). Viral sequences from four of the cell cultures were related to the recently discovered Saffold viruses (SAFV). The other unexplained cell-culture samples contained a paramyxovirus related to Newcastle disease virus (one isolate, to be described elsewhere), BK polyomavirus (one isolate) and herpes simplex virus (<100% identical at the nucleotide level to known virus sequences and were not examined further. As SAFV-like sequences were identified in four isolates, the complete genomes of these viruses were sequenced. The genomes were identified as SAFV_NL1999-590, NL2007-2686, NL2007-2690 and NL2005-1053 (GenBank accession nos HM181996-HM181999, respectively). These isolates represent samples collected between 1999 and 2005 (Table 2).

**Phylogenetic analysis of identified SAFV**

Available SAFV complete or near-complete genome sequences were used for phylogenetic analysis (Fig. 1). All of the SAFV isolates from EVS samples clustered with SAFV genotype 3 (SAFV-3) genomes from the Netherlands and Germany. A similarity-plot analysis (using SimPlot software; see Methods) was performed to evaluate sequence similarities among the different SAFV genomes (Fig. 2). SAFV-3_NL2007 (GenBank accession no. FM207487) was selected as the query sequence to evaluate how the new genomes compared with this genome, which was discovered in the Netherlands from a stool sample collected in 2007 (Zoll et al., 2009). The SAFV_NL2007-2686, NL2007-2690 and NL2005-1053 isolates were almost identical to each other, sharing 98–99% nucleotide identity. The SAFV_NL1999-590 isolate showed the lowest sequence similarity to SAFV-3_NL2007 and the other SAFV-like genomes detected in this study (88% overall nucleotide identity to all SAFV isolates from the Netherlands). The region encoding the structural protein of the SAFV_NL1999-590 genome showed the most divergence from other SAFV-3 sequences, including those from Germany (Fig. 2). NL1999-590 shared between 80 and 87% nucleotide identity with the query sequence, whilst the other SAFV-3 genomes, including those from

---

**Table 1.** Total number of isolates obtained through the EVS programme from 1994 to 2007 and results from the PCR-based virological analysis (van der Sanden et al., 2008, 2009)
Germany, ranged between 90 and 95% nucleotide identity. The SimPlot analysis did not reveal any potential recombination among SAFV-3 and other genotypes (data not shown).

**Frequency of SAFV infections in cell-culture samples with unidentified viruses from EVS and phylogenetic analysis of identified SAFV isolates**

A PCR assay targeting the VP1 region of SAFV was used to screen eight additional EVS samples with CPE of unidentified origin, collected between 2008 and 2009. SAFV was detected in one of the samples collected in 2008, bringing the total number of SAFV-positive cell-culture isolates to five in the period from 1994 to 2009. Overall, SAFV accounted for 33% of the unidentified cell-culture isolates, but <0.3% of the total EVS samples.

To further describe the five SAFV-like viruses identified in samples from the EVS programme, complete VP1 sequences were acquired and used for classification. Phylogenetic analysis using a neighbour-joining tree was performed on approximately 825 nt of the VP1 protein gene (Fig. 3). All of the detected SAFV clustered with SAFV-3 sequences from the Netherlands, California, and Germany (Fig. 3) (Chiu et al., 2008; Drexler et al., 2008; Zoll et al., 2009).

**DISCUSSION**

RIVM is a poliovirus reference laboratory, to which laboratories across the Netherlands submit cell-culture isolates for confirmation and typing of enteroviruses through PCR assays and serological testing. This programme started in 1994, after a poliovirus type 3 outbreak in 1992–1993, to monitor the eradication of poliovirus from the Netherlands (Cornyn-van Spaelendonck et al., 1996; Oostvogel et al., 1994). For samples collected between 1994 and 2007, RIVM confirmed the presence of a range of known viruses in all but seven samples. Enteroviruses (including human rhinoviruses) accounted for 82.5% of the samples, and parechoviruses were the second most common group of viruses detected, accounting for 10.3% of the samples collected through EVS (van der Sanden et al., 2008, 2009). Extended PCR testing allowed the identification of other viruses in a low proportion of samples (<1% for each virus). Finally, viral metagenomics enabled the identification of additional DNA and RNA viruses in all of the remaining unexplained samples. Of note, none of these were enteroviruses, indicating that the existing enterovirus PCR assays effectively capture the diversity of enteroviruses circulating in the Netherlands.

As expected, two genera from the family Picornaviridae, Enterovirus and Parechovirus, comprised the bulk of viruses detected during the 13 year EVS period investigated in this study. The newly described SAFV dominated the unexplained cases (approx. 33% of the cell cultures with
unidentified viruses from 1994 to 2009), and also belong to the family *Picornaviridae*, but to a third genus, named *Cardiovirus*.

Since its initial discovery in 2007 from a stool sample collected in 1981 in California, USA (Jones et al., 2007), SAFV has been detected in stool and respiratory samples from patients in Europe, North and South America and Asia (Abed & Boivin, 2008; Blinkova et al., 2009; Chiu et al., 2008; Drexler et al., 2008; Liang et al., 2008; Ren et al., 2009; Xu et al., 2009; Zoll et al., 2009). Currently, there are eight SAFV genotypes based on variants of the VPI protein (SAFV-1–8), which is a determinant of viral tropism and antibody-neutralization sensitivity (Blinkova et al., 2009). To date, most of the SAFV diversity has been identified from samples collected in Pakistan, where genotypes 2–6 and 8 have been identified (Blinkova et al., 2009). In contrast, SAFV-3 is the only genotype that has been detected in samples collected in the Netherlands, both previously (Zoll et al., 2009) and in our study.

One of the SAFV-3 genomes, SAFV_NL1999-590, is divergent in the structural protein-encoding region of the genome from other SAFV-3 genomes, but conservation of amino acids in this region suggests that SAFV_NL1999-590 belongs to the same genotype (Fig. 4). It is possible that the divergence observed in SAFV_NL1999-590 is due to mutational changes over time, indicating prolonged circulation of these viruses, as has been described for other enteroviruses (van der Sanden et al., 2008, 2009). As there are limited data regarding SAFV-3 isolates, it is difficult to calculate their substitution rates. Alternatively, SAFV_NL1999-590 may be recombinant relative to genotype 3 strains from the USA, as the VPI phylogenetic analysis indicated that SAFV_NL1999-590 clusters with SAFV-3 strains from California instead of strains from the Netherlands and Germany. Within-genotype recombination has been reported for SAFV-2 (Drexler et al., 2010), but BootScan analysis failed to show evidence for recombination between SAFV_NL1999-590, SAFV-3 sequences from the Netherlands and Germany, and SAFV-1 and -2 sequences from California, Canada and Brazil (data not shown). For definitive conclusions, more complete-genome sequences of SAFV strains identified in the USA are needed.

The identification of different SAFV-3 isolates in 0.25% of the tested samples is in the same range as other viruses...
from the family Picornaviridae, such as coxsackie A viruses (S. van der Sanden, unpublished data), indicating endemic circulation in the Netherlands. Based on a virus-neutralization assay, it has been shown that SAFV-3 infection is indeed widespread in humans in at least three continents (Europe, Africa and Asia; Zoll et al., 2009). Consistent with the hypothesis of endemic circulation, infection seems to occur in early life, as 92–98% of a group of 150 individuals between 4 and 50 years old from the Netherlands had neutralizing antibodies against SAFV-3 (Zoll et al., 2009).

Seroprevalence data also indicated that SAFV-3 NL2007 or a related virus has been present in the Netherlands for at least 10 years (Zoll et al., 2009), confirmed by the identification of SAFV_NL1999-590 (Fig. 2). In the USA, it has been shown that 91% of adults carry antibodies against SAFV-2, with 80% of studied individuals generating neutralizing antibodies against this genotype (Chiu et al., 2010).

Although SAFV infection is highly prevalent, the pathogenicity of this virus remains unknown. SAFV has been detected in stool and respiratory-secretion samples from children with gastroenteritis (Chiu et al., 2008; Drexler et al., 2008; Ren et al., 2009; Xu et al., 2009), respiratory illness (Abiri & Boirin, 2008; Chiu et al., 2008) and non-polio acute flaccid paralysis (Blíňková et al., 2009), as well as in asymptomatic individuals (Blíňková et al., 2009; Chiu et al., 2008; Xu et al., 2009; Zoll et al., 2009). In several cases, SAFV has been detected in addition to other viruses known to cause gastroenteritis, thus hindering interpretation regarding SAFV infection and observed pathology (Chiu et al., 2008; Drexler et al., 2008; Ren et al., 2009). SAFV-2 has recently been linked to infection in a child with diarrhoea and vomiting, based on an acute seroconversion event supporting an association between this virus and diarrhoeal disease (Chiu et al., 2010). Although SAFV-3 was the only viral agent identified in five of 15 of the unexplained cell-culture samples in this study, it cannot be concluded that these viruses were causing illness or clinical symptoms. It is well-known that enteroviruses may manifest illness in a minority of infected persons, but this does not mean that infections are trivial. Polio, for instance, develops in <1% of infected persons, but is a debilitating disease that results in lifelong disabilities. Therefore, further research is needed to unravel the possible role of SAFV in human illness.

Sequence-independent methods are becoming more important to identify emerging viruses in public-health
Appendix A (Continued)

Fig. 3. Neighbour-joining phylogenetic tree of SAFV VP1 sequences. All SAFV sequences used for phylogenetic analysis are identified by their genotype (SAFV-#), the name used in their original publication and the country where the sequences were identified. Sequences reported in this study are marked with an asterisk. The tree was constructed in MEGA4 using the maximum composite likelihood method to calculate evolutionary distances. Bar, 0.05 base substitutions per site. One thousand bootstrap resamplings were performed to assess statistical support.

Fig. 4. Alignment of putative SAFV-3 VP1 CD (top) and VP2 EF (bottom) loop amino acid sequences. aa 716–785 and 261–330 are shown for the VP1 CD and VP2 EF loops, respectively. The alignment shows conservation of amino acids in these structural loop sequences, except for one amino acid in the VP1 CD loop II sequence (highlighted as bold white text on a black background) in SAFV-3 NL2007.
monitoring efforts (Abed & Boivin, 2008). Viral metagenomics is a relatively new technique that has been used increasingly to identify viruses in clinical specimens (Allander et al., 2005; Breitbart & Rohwer, 2005; Breitbart et al., 2003, 2008; Jones et al., 2005, 2007; Kapoor et al., 2008; Victoria et al., 2009; Willner et al., 2009; Zhang et al., 2006). Purifying virus particles before sequencing minimizes the amount of sequencing that needs to be performed for virus identification. If sufficient budget exists, combining the virus-purification procedure presented here with next-generation sequencing technologies will enable rapid sequencing of the complete genomes of viruses of interest.

In this study, viruses were identified in a relatively small amount of time (less than a week) with limited sequencing (<50 clones per sample), demonstrating that metagenomic analysis can be added as a routine test to investigate unidentified viruses in cell cultures from clinical samples in a public-health setting. This method proved to be a more efficient and cost-effective alternative than individual PCR assays for different viruses. Therefore, we recommend the use of viral metagenomics on cell-culture samples that test negative for established PCR assays in virus-surveillance programmes. However, we emphasize that embedding these techniques with follow-up studies is essential for establishing relevance of the findings for public health.

METHODS

Clinical specimens. All clinical specimens reported in this study were collected through the EVS programme in the Netherlands. The goal of this programme is to monitor poliovirus circulation and ensure its eradication from the Netherlands. Although 90% of enterovirus infections are asymptomatic, these infections can result in a broad spectrum of clinical symptoms (Zaoutis & Klein, 1998). Therefore, the EVS collects samples from patients with a variety of clinical manifestations such as meningitis, hepatitis, respiratory illness and gastrointestinal disorders. As virus-shedding of both symptomatic and asymptomatic patients occurs through faeces, most of the samples collected through the EVS programme are stool specimens, which provide information about the general circulation of enteroviruses. The surveillance in its current form was started after a poliovirus type 3 outbreak in 1992-1993 (Coomen-van Spaendonck et al., 1996; Oostwegel et al., 1991). In order to detect polioviruses, all samples were initially tested for a broad range of enteroviruses in routine virological laboratories (Nix et al., 2008; Oberte et al., 1999), which are linked to academic hospitals or to regional public-health centres throughout the entire country. For enterovirus surveillance, clinical samples were cultured on combinations of cell lines used routinely in the different laboratories, including human rhabdomyosarcoma (RD), terry monkey kidney (SK), vero monkey kidney (LLC-MK2), Vero-Hep2, human epithelial carcinoma (HeLa), human embryonic lung fibroblast (Gaith) and various human fibroblast cell lines, and confirmed as enteroviruses by local laboratories. Cell-culture samples that exhibited CPE but tested negative for enteroviruses or were not typed genetically were submitted to RIVM for virus identification and typing by PCR. Between 1991 and 2007, a total of 1834 cell-culture isolates collected throughout the Netherlands were submitted to RIVM for analysis.

Systematic analysis of cell-culture isolates from EVS samples at RIVM. Cell-culture isolates submitted to RIVM were passaged once, and subsequently tested using PCR assays for enteroviruses (Nix et al., 2006; Oberte et al., 1999), followed by parechoviruses (van der Sanden et al., 2008). Samples that remained negative for enteroviruses and parechoviruses were screened using a broad range of PCR assays for noroviruses, rotaviruses A, B and C, adenoviruses, astroviruses, sapoviruses, picornaviruses, reoviruses, a generic PCR that detects both enteroviruses and rhinoviruses, hepatitis A and E viruses, influenza A and B viruses, Aichi virus, coronaviruses 29E, NL63 and OC43, human respiratory syncytial viruses A and B and human metapneumovirus, as described previously (Syrka et al., 2007, 2009a, b; van der Sanden et al., 2008; van Gageldonk-Leiber et al., 2005). Samples that tested negative in all assays but showed consistent CPE were processed for metagenomic sequencing (Tables 1 and 2). Information regarding sample description for these unexplained cell cultures and associated patient clinical symptoms is given in Table 2.

Virus-particle purification and metagenomic sequencing. Virus particles were purified from selected cell cultures prior to metagenomic sequencing. For this purpose, cell-culture samples were vortexed vigorously for 2 min. A 1 ml aliquot was collected and mixed with 1 ml SM buffer [0.1 M NaCL, 50 mM Tris/HCl (pH 7.4), 10 mM MgSO4], vortexed once again and filtered through a 0.22 μm Sterivex filter (Millipore). The filtrate (i.e. virus fraction) was treated with 20% (v/v) chloroform, followed by treatment with DNase I (2.5 U μl-1) and RNase A (0.25 U μl-1) to remove free nucleic acids. If no viruses were detected after this purification procedure, the purification was repeated without chloroform and treated with only DNase I (2.5 U μl-1) to recover enveloped viruses. The methods used for virus purification for each of the samples are summarized in Table 2. Prior to extracting total viral nucleic acids, EDTA (20 mM final concentration) was added to a 200 μl aliquot of purified virus particles to chemically inactivate nucleases. For analysis of RNA viruses, total RNA was extracted from virus particles by using a Qiagen RNeasy Mini kit, followed by random primer RNA amplification with a Transcriptor Whole Transcriptome Amplification (WTA) kit (Strata-Galizic) according to the manufacturer's protocol using 50 and 65 PCR cycles. For analysis of DNA viruses, total viral nucleic acids were extracted using a Qiagen MinElute Virus Spin kit (Qiagen), followed by random amplification using a GenomiPhi V2 DNA Amplification kit (GE Healthcare) according to the manufacturer's protocol. GenomiPhi products were amplified further using a Whole Genome Amplification (WGA) kit (Strata-Galizic)

As it was unknown whether the selected cell-culture samples contained DNA or RNA viruses, WTA and WGA products were analysed on a 2% agarose gel, and positive reactions (defined as a smear between 200 and 400 bp) were purified by using an UltraClean PCR Clean-Up kit (MO BIO Laboratories Inc.). The purified products were cloned into competent Escherichia coli cells using the TOPO TA system (Invitrogen) and sequenced. To identify the different viruses in the cell-culture samples, approximately 68 clones were selected from each sample for sequencing. The sequenced fragments were compared against the GenBank non-redundant (nr) nucleotide database through BlastX (E-value <0.001) (Altschul et al., 1997).

Genome completion of SAFV. WTA fragments sequenced from four samples had significant similarities to SAFV. In order to complete the genome of the SAFV-like viruses, WTA fragments were assembled against the SAFV genome with the highest similarities to the sequenced WTA fragments (i.e. SAFV-3_NL2007) using the GenBank accession no. FM207487. The assembly was performed using the Bambus software (version 4.1; GeneCoes), using the assembly algorithm for draft data with optimization for gap alignment, 70% minimum match percentage and a minimum overlap of 10 bp. As WTA fragments did not overlap into a complete genome, the assembly was used to design primers for genome walking based on the consensus sequence between WTA fragments and SAFV-3_NL2007.
Table 3. Primers used for genome walking of SAFV genomes

<table>
<thead>
<tr>
<th>Genome/Primer ID*</th>
<th>Sequence</th>
<th>$T_m$ (°C)†</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>TGAACCTTCTGCTGTATAGCC</td>
<td>RACE 5'</td>
</tr>
<tr>
<td>400R</td>
<td>GCCGCTGTTACGCCTGCTGCT</td>
<td>RACE 5'</td>
</tr>
<tr>
<td>HM181996</td>
<td>AAAGACCGCTGTATGAGAT</td>
<td>RACE 5'</td>
</tr>
<tr>
<td>HM181997, HM181998, HM181999</td>
<td>CACCTATTAGAGAGCACGGA</td>
<td>RACE 5'</td>
</tr>
<tr>
<td>650R</td>
<td>TTACCAAGGGGCCCCCTG</td>
<td>49</td>
</tr>
<tr>
<td>610R</td>
<td>AAAAGGCGGGTTGAGAT</td>
<td>49</td>
</tr>
<tr>
<td>640F</td>
<td>ATTCAGCCGAGGGGCTTGG</td>
<td>49</td>
</tr>
<tr>
<td>140R</td>
<td>GTTTGCTGGCGGCTGTTG</td>
<td>49</td>
</tr>
<tr>
<td>1418F</td>
<td>TGGACACCCTGTTGAGGAT</td>
<td>49</td>
</tr>
<tr>
<td>240R</td>
<td>TTTTCACCAATGAGGAGCCT</td>
<td>49</td>
</tr>
<tr>
<td>7200F</td>
<td>GCGCAGACAGAGACTAGG</td>
<td>49</td>
</tr>
<tr>
<td>7400R</td>
<td>GGTTAACAGAGGAGGCTG</td>
<td>49</td>
</tr>
<tr>
<td>7124F</td>
<td>ACTGTGACCAGAGCGAGTC</td>
<td>49</td>
</tr>
<tr>
<td>8125R</td>
<td>TTTTTTTTTTTTTTGTCATTCCTCAA</td>
<td>49</td>
</tr>
<tr>
<td>HM181997, HM181998, HM181999</td>
<td>TTTCAAAAGGGGCCCCCTG</td>
<td>49</td>
</tr>
<tr>
<td>580R</td>
<td>CACCTATTAGAGAGCACGGA</td>
<td>49</td>
</tr>
<tr>
<td>520F</td>
<td>GGCAAGAACCGCGCTGCTG</td>
<td>60</td>
</tr>
<tr>
<td>1095R</td>
<td>CAAAGGCGGGTTGAGGAT</td>
<td>49</td>
</tr>
<tr>
<td>10B7F</td>
<td>CAGGTGCTGTTGCTGCTCT</td>
<td>60</td>
</tr>
<tr>
<td>1350R</td>
<td>AAAAAAGGCGGGTCTGCT</td>
<td>49</td>
</tr>
<tr>
<td>1418F</td>
<td>TGACACCTGCTGCTGAGG</td>
<td>49</td>
</tr>
<tr>
<td>2400R</td>
<td>TTTTCACCAATGAGGAGCCT</td>
<td>49</td>
</tr>
<tr>
<td>2300F</td>
<td>GTCCAGGAAACTGAGGAC</td>
<td>49</td>
</tr>
<tr>
<td>2700R</td>
<td>TCCCAAAATGAGGCTAG</td>
<td>49</td>
</tr>
<tr>
<td>2600F</td>
<td>TGAGGAGGAGGAGGACCA</td>
<td>49</td>
</tr>
<tr>
<td>3750R</td>
<td>TTGTCTGCTTCTGCTGTTG</td>
<td>49</td>
</tr>
<tr>
<td>3600R</td>
<td>TAGGTGCTGCTGCTGAGG</td>
<td>49</td>
</tr>
<tr>
<td>3700R</td>
<td>CAAAGCAGAACCGACAGGA</td>
<td>49</td>
</tr>
<tr>
<td>5050R</td>
<td>GCGCAGCAAGAGACAGTTG</td>
<td>49</td>
</tr>
<tr>
<td>5175F</td>
<td>TTTCAAAATGAGGCTGAG</td>
<td>49</td>
</tr>
<tr>
<td>6700R</td>
<td>GAATTCAGGAGGAGGACCA</td>
<td>49</td>
</tr>
<tr>
<td>7100R</td>
<td>GCTGTGCAGAGCACTGCT</td>
<td>49</td>
</tr>
<tr>
<td>7500F</td>
<td>AGCTTACAGCAATAGGAGA</td>
<td>49</td>
</tr>
<tr>
<td>8125R</td>
<td>TTTTTTTTTTTTTTTTTTTTTTTCTC</td>
<td>49</td>
</tr>
</tbody>
</table>

*Primer number indicates approximate nucleotide position in reference genome SAFV-3 (GenBank accession no. F2071477).
†Annoing temperature. All PCRs were performed with a touchdown program to decrease the annoing temperature incrementally by 0.2 °C each cycle. RACE 5' indicates gene-specific primers used for the 5' RACE system (Invitrogen) and nested reactions according to the manufacturer's instructions.
cDNA was produced from RNA extracted from purified virus particles by using a synthesis SuperScript III reverse transcription kit (Invitrogen) with random primers. Each PCR was performed in a 50 μl reaction containing 5 μl cDNA, 1x RED-Taq PCR buffer (Sigma-Aldrich), 1 μM each primer (Table 3), 0.2 μM dNTPs and 1 U RED-Taq DNA polymerase (Sigma-Aldrich). Each PCR proceeded at 94°C for 5 min, followed by 50 cycles of 94°C for 1 min, annealing temperature (Table 3) for 45 s, and 72°C for 2 min and a final extension at 72°C for 10 min. The 5'-UTR region was verified with the 5'-RACE system (Invitrogen) using gene-specific primers (Table 3) according to the manufacturer’s suggested protocol for high-capacity cDNA synthesis. The resulting PCR products were subcloned into pCRII-TOPO and sequenced to determine nucleotide sequence agreement with the published sequence.

Phylogenetic analysis of identified SAFV isolates. The SAFV genome sequences were used for phylogenetic analysis (Fig. 1). All alignments were performed using the CLUSTAL W algorithm (Thompson et al., 1994) and MUSCLE as the similarity matrix in BioEdit version 7.0.9.0 (Hall, 1999). A neighbour-joining phylogenetic tree was constructed in MEGA4 (Tamura et al., 2007), using the maximum composite likelihood method to calculate evolutionary distances. All gaps and missing data were eliminated from the dataset and 1000 bootstrap resamplings were performed to assess statistical support.

A similarity-plot analysis showing the relationship among the detected SAFV sequences and previously reported genomes was performed using the SimPlot software, version 3.5.1 (Lole et al., 1999). The following parameters were used to calculate the similarity plot: 200 bp window size, 20 bp step size, and an empirical transition to transversion ratio. Sequences that originated from the same region and were 98–99% identical at the nucleotide level were grouped for the SimPlot analysis.

SAFV screening of additional unexplained samples from EVS. The original cell-culture isolates examined for this study were collected through the EVS programme between 1994 and 2007. To further assess the prevalence of SAFV in unexplained cell-culture samples, eight additional unidentified cell-culture isolates collected between 2008 and 2009 were tested by using a PCR assay targeting the VP1 region of SAFV (Blinkova et al., 2009). The VP1 region is the most diverse protein of picornaviruses and has been used to classify different SAFV genotypes (Blinkova et al., 2009). For the PCR assay, cDNA was obtained from randomly primed reverse transcription of total RNA from cell cultures using a SuperScript III reverse transcription kit (Invitrogen).

The RT-PCR was prepared by adding 2.5 μl reverse transcription mixture with random primers to 22.5 μl PCR mixture containing 12.5 μl HotStarTaq mastermix (Qiagen), 1.25 μl forward and reverse primers (10 pmol μl$^{-1}$) and 7.5 μl H$_2$O. For the VP1 region, VP1F1 (5'-ACWCCTGGTTCGDDGAGG-GG-3') and VP1R1 (5'-TCGCCCATATCAYACGRGAA-3') primers were used in the first round of nested PCR, while VP1F2 (5'-GACTTAYCCTGCTTGAAGCC-3') and VP1R2 (5'-ACCTGTTACTCRTGAACTGGTATA-3') were used in the second round (Blinkova et al., 2009). The PCR proceeded as follows: denaturation at 95°C for 15 min, four cycles of amplification at 95°C for 1 min, 55°C for 1 min and 68°C for 1 min 30 s, followed by 35 cycles of amplification at 95°C for 30 s, 52°C for 30 s and 68°C for 1 min 30 s, then a final extension at 68°C for 10 min. A 2.5 μl sample from the first round PCR was used as template for the nested PCR. The same PCR mixture and conditions were used for both rounds of PCR (Blinkova et al., 2009).

SEQUENCES FROM THE SAFV VP1 PCR PRODUCTS, ALONG WITH PUBLICLY AVAILABLE SEQUENCES THAT WERE PREVIOUSLY CLASSIFIED INTO SAFV GENOTYPES (3), WERE USED FOR PHYLOGENETIC ANALYSIS. A NEIGHBOUR-JOINING TREE WAS CONSTRUCTED IN MEGA4, USING THE SAME PARAMETERS THAT WERE USED TO CONSTRUCT THE PHYLOGENETIC TREE OF FULL-LENGTH SAFV GENOMES (FIG. 3).

ACKNOWLEDGMENTS

The authors declare that they have no conflicting interests in relation to this work. The authors thank all of the virologists participating in the Weekly Sentinel System of the Dutch Working Group for Clinical Virology for collecting and providing cell-culture isolates over the complete period of the study, and Gokhan Usub and Edin Jusko for technical assistance. M. B. was funded through a grant from the Alfred P. Sloan Foundation (BR-4772). K.R. was funded through the NSF Bridge to the Doctorate Fellowship and the Elise and William Knight Jr Endowed Fellowship.

REFERENCES


APPENDIX B:

METAGENOMIC ANALYSIS OF VIRUSES IN RECLAIMED WATER
Metagenomic analysis of viruses in reclaimed water

Karyna Rosario,¹ Christina Nilsson,² Yan Wei Lim,² Yijun Ruan³ and Mya Breitbart⁴*¹
¹College of Marine Science, University of South Florida, St. Petersburg, Florida, USA.
²Genome Institute of Singapore, Singapore.

Summary

Reclaimed water use is an important component of sustainable water resource management. However, there are concerns regarding pathogen transport through this alternative water supply. This study characterized the viral community found in reclaimed water and compared it with viruses in potable water. Reclaimed water contained 1000-fold more virus-like particles than potable water, having approximately 10¹⁴ VLPs per millilitre. Metagenomic analyses revealed that most of the viruses in both reclaimed and potable water were novel. Bacteriophages dominated the DNA viral community in both reclaimed and potable water, but reclaimed water had a distinct phage community based on phage family distributions and host representation within each family. Eukaryotic viruses similar to plant pathogens and invertebrate picornaviruses dominated RNA metagenomic libraries. Established human pathogens were not detected in reclaimed water viral metagenomes, which contained a wealth of novel single-stranded DNA and RNA viruses related to plant, animal and insect viruses. Therefore, reclaimed water may play a role in the dissemination of highly stable viruses. Information regarding viruses present in reclaimed water but not in potable water can be used to identify new biocides for water quality. Future studies will need to investigate the infectivity and host range of these viruses to evaluate the impacts of reclaimed water use on human and ecosystem health.

Introduction

Increasing urbanization on a global scale places enormous pressure on finite freshwater supplies. The use of alternative water supplies is therefore an important component of sustainable water resource management practices across the world (Levine and Asano, 2004). Reclaimed water (i.e. the reusable end-product of wastewater treatment) is an important alternative water supply since it reduces the discharge of wastewater effluent into surface waters and contributes to water conservation by supplying water for activities that do not require drinking water quality standards. For more than 20 years Florida has been on the forefront of water reuse efforts in the USA (Young and York, 1996). Local freshwater supplies are insufficient for supporting rapid population growth in Florida, which has lead regulatory agencies to increase emphasis on beneficial water reuse, such as water reclamation (Overman and Pirozzoli, 1996). Reclaimed water is currently used in Florida for non-potable public water supply, agricultural irrigation, environmental enhancement, industrial uses and groundwater recharge (Florida Department of Environmental Protection, 2007).

Reclaimed water has successfully been used as an alternative water resource for decades (Young and York, 1996; Levine and Asano, 2004). Nevertheless, as water reuse applications increase and reclaimed water distribution expands, there are some concerns that need to be addressed to ensure protection of public health and the health of the environment. One of the biggest issues regarding reclaimed water use is pathogen transport. Since the microbiological content of this water supply is still largely unknown, it is difficult to assess which pathogens can potentially be spread through this alternative water supply. Viruses are a group of particular concern because they include highly stable pathogens that can be resistant to standard wastewater treatment processes. Although reclaimed water meets water quality standards, for practical reasons, current quality control methods do not test the presence of pathogens directly (Salgot et al., 2001). The spread of viral pathogens through reclaimed water is a real possibility as several studies have detected enteric viruses in treated wastewater, including noroviruses, astroviruses, sapoviruses, rotaviruses, noroviruses, adenoviruses, hepatitis A viruses and enteroviruses (Morace et al., 2002; Sedmak et al., 2005; Botli-Mas et al., 2006; Haramoto et al., 2006; 2008; Arraj et al., 2008; Katayama et al., 2008; Meleg et al., 2008). In spite of these findings, most of the microbiological research in treated wastewater has been directed towards biocidal organisms, such as faecal coliforms, to indirectly reflect the presence of enteric bacteria and viruses. It has been shown that bacterial indicators, such as faecal coliforms, do not correlate with the occurrence of viral pathogens in wastewater (Harwood et al., 2005;
Appendix B (Continued)

Haramoto et al., 2006; Carducci et al., 2000). These findings have led several scientists to propose a suite of bioremediation organisms, including bacteria and coliphages, as well as viral indicators, such as human adenoviruses and polyomaviruses as a more sensitive tool to detect viral pathogens (Harewood et al., 2000; Bobillier-Mas et al., 2005; McCuaig et al., 2000; Carducci et al., 2008).

Concerns regarding reclaimed water use mainly focus on human pathogens, since most of the water is derived from domestic (human) wastewater. However, reclaimed water may also be a reservoir for non-human pathogens that are present in human waste. For example, it has been shown that plant viruses dominate the RNA virus community in human faeces (Zhang et al., 2006). Other studies have identified animal rotavirus strains of unknown origin co-culturing with human strains in sewage and treated wastewater (Viteria et al., 2003; Malleg et al., 2008). Recently, a metagenomic study of viruses in stool from South Asian children identified an abundance of novel polyomaviruses related to the Enterovirus genus (i.e., cosaviruses) and four new viral species related to the Deltaviridae, Adenoviridae, Circoviridae families and the Dicistroviridae genus (Kapoor et al., 2009; Victoria et al., 2009). Therefore, the diverse viral flora in human faeces may contain plant, insect and animal viruses, in addition to human pathogens.

It is critical to have a comprehensive understanding of viruses in reclaimed water as this alternative resource becomes more widely used. Information regarding viruses will help regulatory agencies to make informed decisions about reclaimed water use to minimize negative impacts upon human and environmental health. The main objective of this study was to examine the abundance and diversity of viruses found in reclaimed and potable water samples from southwest Florida, USA, through direct epifluorescent microscopy and metagenomic sequencing of purified viral particles from these water sources. Bacteriophages (phages) were abundant in both potable and reclaimed water; however, differences in phage community composition between these water supplies can be exploited to identify potential biodegradable water quality. Reclaimed water also contains a wealth of novel viruses related to plant, animal and insect pathogens, suggesting that highly stable viruses can spread through the use of this alternative water supply.

Results and discussion

Abundance of virus-like particles

The concentration of virus-like particles (VLPs) in various traditional and alternative water supplies, including well, potable and reclaimed water, was determined through SYBR Gold staining and epifluorescent microscopy. Potable and well water had VLP concentrations on the order of $10^4$ and $10^5$ VLPs ml$^{-1}$, respectively (Fig. 1A). The VLP concentrations found in potable water are similar to those observed in other studies, which have found an abundance of bacteria (on the order of $10^5$ cells ml$^{-1}$) and VLPs (on the order of $10^5$–$10^6$ VLPs ml$^{-1}$) using direct counts (Finta-Kanto et al., 2004; Berney et al., 2008). Reclaimed water samples contained approximately 1000-

![Fig. 1. A. Epifluorescent microscopy counts of virus-like particles (VLPs) found in conventional water supplies (well and potable water), raw sewage and reclaimed water (RW). B. Transmission electron micrographs of virus-like particles found in reclaimed water (the bar in each panel is 100 nm). Viral counts include well water samples collected from private wells (n = 5); potable water samples collected at a plant nursery (n = 3); raw sewage and RW effluent samples collected at a wastewater treatment plant (n = 3 for each); and RW samples at the point-of-use obtained from different public spriiters and a plant nursery (n = 3). Error bars represent one standard deviation.](image-url)
fold more VLPs (on the order of $10^3$ VLPs ml$^{-1}$) than the conventional water supplies studied. In order to compare the abundance of VLPs in raw sewage, after treatment, and at the downstream point-of-use, several samples originating from a single wastewater treatment plant were collected. The average VLP concentrations for treated reclaimed water effluent and at the point-of-use were similar to the VLP concentrations in raw sewage (Fig. 1A). Examination of purified reclaimed water VLPs by transmission electron microscopy (TEM) showed an abundance of viruses resembling known phages and plant pathogens (Fig. 1B).

Although average VLP concentrations were similar in raw sewage and reclaimed water (~$10^4$ VLPs ml$^{-1}$), this should not be considered an indication of treatment efficiency. Virus-like particles counted with epifluorescent microscopy are not necessarily infectious virus particles. The infectivity of the virus particles will rely heavily upon the type of wastewater treatment. This study focused on reclaimed water from a treatment plant using secondary treatment (activated-sludge) with chlorine disinfection, which is typical of wastewater reclamation facilities in Florida (Florida Department of Environmental Protection, 2007). In addition, high VLP concentrations in reclaimed water samples may reflect the abundance of phages in treated effluent. A study investigating bacteriophage populations in an activated sludge system found that there was a net production of phages in the system, with total phage concentrations in the supernatant of an activated-sludge reactor and the unchlorinated effluent measuring as high or higher than those in sewage entering the reactor (Ewert and Paynter, 1980). The abundance of phages in reclaimed water was also supported by TEM of viral concentrates, where an abundance of phage-like particles was observed compared with other VLPs (examples of VLPs shown in Fig. 1B) and by the dominance of phage-like sequences in the reclaimed water DNA viral metagenomes (see below).

Overview of metagenomic analyses

This study utilized a metagenomic sequencing approach to examine the viruses present in reclaimed water. The advantage of this method is that it studies the entire viral community, without selection based on host or sequence similarity to known viruses. DNA viral metagenomes were obtained from a potable water sample (‘Potable’), and reclaimed water samples at the point-of-discharge (‘Effluent’) and the point-of-use (‘Nursery’ and ‘Park’) (Table 1). Corresponding RNA viral metagenomes were sequenced from reclaimed water samples at the point-of-discharge (‘Effluent’) and the point-of-use (‘Nursery’) (Table 1). It was not possible to obtain enough RNA from the potable water sample for pyrosequencing and, thus, a comparison between reclaimed and potable water RNA viral communities could not be achieved.

All DNA and RNA reclaimed water libraries were analysed individually. However, there were no notable differences in the distribution of phages and viral types between the Effluent and the point-of-use, for both DNA and RNA libraries, suggesting that the viral community composition does not change significantly in the distribution system. Therefore, all reclaimed water DNA libraries sequence distributions, including Effluent, Park and Nursery, were averaged together and shown as ‘Reclaimed DNA’ in the results. The Effluent and Nursery RNA library distributions were averaged together as well, and are referred to as ‘Reclaimed RNA’.

For metagenomic sequence analysis, more than 230 000 raw reads for each library were used for contig assembly in the SeqMan program (DNASTAR) (Table 1). The average raw sequence length for Effluent, Nursery and Potable libraries was 245 nt while the average length for the Park library was 136 nt. The majority (68–97%) of the sequences in each metagenome assembled into contigs larger than 200 nt (Table 1). Contigs larger than 200 nt were utilized for BLASTX searches against the

Table 1. List of samples used to construct metagenomic libraries and overview of the total number of sequences and contigs for each library.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Date of collection</th>
<th>Volume collected</th>
<th>Description</th>
<th>Library</th>
<th>No. of raw reads</th>
<th>No. of contigs</th>
<th>Raw reads in contigs (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Park</td>
<td>9/7/2006</td>
<td>501</td>
<td>Reclaimed water collected from a public sprinkler</td>
<td>DNA</td>
<td>307 069</td>
<td>8 327</td>
<td>67.60</td>
</tr>
<tr>
<td>Effluent</td>
<td>5/10/2007</td>
<td>100</td>
<td>Reclaimed water collected at a treatment plant</td>
<td>DNA</td>
<td>260 304</td>
<td>26 729</td>
<td>78.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RNA</td>
<td>232 529</td>
<td>5 352</td>
<td>96.70</td>
</tr>
<tr>
<td>Nursery</td>
<td>5/10/2007</td>
<td>100</td>
<td>Reclaimed water collected at a plant nursery</td>
<td>DNA</td>
<td>291 542</td>
<td>27 372</td>
<td>76.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RNA</td>
<td>297 014</td>
<td>13 987</td>
<td>92.00</td>
</tr>
<tr>
<td>Potable</td>
<td>5/10/2007</td>
<td>100</td>
<td>Potable water collected at a plant nursery</td>
<td>DNA</td>
<td>231 715</td>
<td>4 923</td>
<td>95.60</td>
</tr>
</tbody>
</table>

All samples were collected in Manatee County, Bradenton, FL. Samples collected at the ‘point-of-use’ (Park and Nursery) receive reclaimed water from the wastewater treatment plant where point-of-discharge samples (i.e. Effluent) were collected. The number of contigs includes all contigs analysed by BLASTX (i.e. contigs larger than 200 bases).

© 2009 Society for Applied Microbiology and Blackwell Publishing Ltd, Environmental Microbiology, 11, 2806–2820
GenBank non-redundant (nr) protein database since longer read length increases the chances of identifying homologies in the database (Wommack et al., 2008).

**Novel sequences**

Summaries of the BLASTX similarities for the reclaimed and potable water viral metagenomes are shown in Fig. 2. Over 50% of the viral metagenomic sequences (both DNA and RNA) identified in reclaimed water metagenomes had no significant similarity to proteins in GenBank, suggesting the novelty of viruses in this alternative water source (Fig. 2A). This is similar to previous studies, in which the majority of the sequences in environmental viral metagenomes had no similarities to known genes in the database, indicating the high proportion of unknown viruses in the environment (Breitbart et al., 2002; Breitbart et al., 2004; Angly et al., 2006; Culley et al., 2006; Bench et al., 2007). The Potable library had a higher proportion of sequences (56%) with significant similarities to sequences in GenBank compared with reclaimed water libraries, but a large fraction of the viral community still could not be identified.

**Mobile genetic elements**

All BLASTX results with an E-value of $<0.001$ were analysed using the MEGAN software to identify the different taxa present in reclaimed and potable water viral communities. Since viral particles were purified extensively prior to isolating DNA and RNA, contigs with their best matches to proteins from cellular organisms (i.e., bacteria, archaea and eukaryotes) were further compared against the ACLAME database to identify proteins related to mobile genetic elements (i.e., plasmids and phages). For all metagenomes, more than 60% of the sequences that were classified as bacteria were re-classified as plasmids or phages after performing a BLASTX analysis against the ACLAME database. Although sequences similar to archaea were not as abundant as the bacterial sequences, 15–56% of eukaryotic and 40–70% of archaeal sequences were also re-classified in the same manner. The discrepancy between the results obtained from searching the GenBank versus ACLAME databases may be due to an abundance of unidentified prophage-like sequences within microbial genomes in GenBank (Fouts, 2006). Similarities to mobile genetic elements are common in previously sequenced DNA virus metagenomes from other sources (Breitbart et al., 2002; 2003; 2004; Bench et al., 2007; Kim et al., 2006).

Hits to mobile genetic elements dominated all the viral metagenomes (Fig. 2B). For reclaimed water DNA libraries, 51% of the known sequences were similar to viral proteins, whereas the majority of the sequences in the reclaimed water RNA libraries and potable water DNA library were similar to proteins found in plasmids. More than 99% of plasmid-like protein sequences were identified after re-analysing sequences with hits to cellular organisms through the ACLAME database. Although most of the sequences identified were similar to hypothetical proteins, numerous integrases, transposases, recombinases and replication-associated proteins, among others, were identified. Since viral particles were purified by CsCl gradients and DNase treatment to remove contaminating cells and free DNA, plasmids should have been eliminated before viral DNA and RNA isolation. However, phages and plasmids share a number of characteristics as both contain machinery for gene transfer and replication. Moreover, there may be genetic exchange between plasmids, phages and other mobile genetic elements within a bacterial host (Böttner et al., 2002; Mark Osborn and Böttner, 2002), and this exchange may lead to gene organization and protein similarities between plasmids and phages (Haazen et al., 2007). In addition, some phages (e.g. pK02, PY4, P1, N15, LE1, (2O and dB8-1) replicate in their hosts as low-copy-number plasmids instead of integrating into the host genomes (Ikeda and Tomizawa, 1968; Inal and Karunakaran, 1996; Eggers et al., 2000; Girons et al., 2000; Raven et al., 2000; Brian et al., 2001; Casjens et al., 2004). It is possible that there is an abundance of previously undescribed phages containing plasmid-like proteins in potable water as 52% of the identified sequences had similarities to plasmid proteins (Fig. 2B). The abundance of plasmid-like sequences in reclaimed water RNA libraries may reflect the abundance of novel RNA viruses with plasmid-like properties such as the endornaviruses. Currently, four species of dsRNA viruses with plasmid-like properties found in some rice and bean species have been classified by the International Committee on Taxonomy of Viruses as members of the Endornavirus genus (Gibbs et al., 2005). In addition, double-stranded RNA viruses with plasmid-like properties have been found in plants, algae, fungi, protozoa and insects and, thus, endorna-like viruses may be widely distributed among eukaryotes (Horiuchi and Fukushima, 2004; Fukushima et al., 2006; Ooasik et al., 2006).

**Comparison of DNA and RNA viral sequences**

The fraction of metagenomic sequences with similarities to known viral proteins suggests there are fundamental differences between the DNA and RNA viral communities in reclaimed water (Fig. 2C). The potable and reclaimed water DNA viral communities were dominated by phages (more than 98% of the contigs), with only one contig similar to an archaeal virus protein. In contrast, the RNA viral community in reclaimed water was dominated by eukaryotic viruses. Similar results have been found in viral...
Fig. 2. Overview of the average distribution of contigs larger than 200 nt for potable and reclaimed water metagenomic libraries based on BLASTX analysis (E-value < 0.001).
A. Percentage of contigs that had homologues in the GenBank protein database (assigned) versus contigs for which no homologues were found (no-hit).
B. Distribution of "assigned" contigs according to their top BLASTX homologues in GenBank and ACLAME databases.
C. Host distribution of top viral homologues in GenBank and ACLAME. For reclaimed water DNA libraries, eukaryotic hosts included animals, insects, algae, and protozoa. For RNA libraries, all viruses identified as prokaryotic viruses had hits to DNA phage. The environmental category refers to contigs with similarities to marine viruses believed to infect prokaryotes.
metagenomic studies of human faeces where the DNA community was dominated by phages (Breitbart et al., 2003), whereas the RNA community was dominated by eukaryotic viruses (Zhang et al., 2006; Finkbeiner et al., 2008; Kapoor et al., 2008). The RNA community has also been observed in marine viral metagenomes (Breitbart et al., 2002; Angly et al., 2006; Culley et al., 2006; Bench et al., 2007). This trend may represent an important fundamental difference between these viral types that persists across systems; however, it may simply reflect the underrepresentation of RNA phage genomes in the database. To date, there are only 11 complete RNA phage genomes in GenBank. The small number of RNA phages in the database compared with DNA phages may lead to an underestimation of RNA phages identified through metagenomic surveys of natural populations in different environments.

**Phages**

Since the DNA viral community was dominated by phages, DNA metagenomic sequences were analysed against the Phage Sequence Databank, containing the complete genomes of 512 phages and prophages. Although phages were abundant in DNA viral metagenomes from both potable water and reclaimed water, there were distinct dominant phage types identified in each of these water sources based on the distribution of top phage homologues identified in the Phage Sequence Databank. The phages in potable water were dominated by prophages while reclaimed water libraries had a similar proportion of hits to phages in the Siphoviridae family and prophages (Fig. 3A). The abundance of prophages belonging to the Siphoviridae family in reclaimed water is consistent with a viral metagenomic study performed in human faeces (Breitbart et al., 2003) and suggests that members of the Siphoviridae are potential indicators of faecal pollution. In addition, the prevalence of Siphoviridae in reclaimed water suggests that members of this viral family are resistant to chlorination. This is consistent with previous studies, which demonstrated that some phages from the Siphoviridae family are more resistant to chlorination than host-specific phages (i.e., MS2) and some members of the Microviridae (i.e., phiX174) and Myoviridae (i.e., M13) families (Duran et al., 2003). Together, these findings suggest that members of the Siphoviridae family could be further explored as potential bioindicators as they represent an abundant group of viruses in human sewage that are fairly resistant to wastewater treatment (chlorination).

For all sequences that had significant similarities to the Phage Sequence Databank, the host for the top phage hit was examined. Differences in host representation within each phage family indicate that reclaimed water has a distinct phage community when compared with potable water (Fig. 3). Although prophages were abundant in both potable and reclaimed water libraries, the host distribution was quite different. The Potable library was dominated by hits to prophages identified in different strains of *Escherichia coli* (-12% of total phage hits), whereas prophages in reclaimed water libraries had a more even distribution of hosts including, but not limited to, *Xylella* spp., *Ralstonia* spp., *Streptococcus* spp., and *Pseudomonas* spp. (less than 4% of the total phage hits for each) (Fig. 3B). Differences in host distributions were also noted within the *Myoviridae* (Fig. 3D) and *Podoviridae* (Fig. 3E) families.

Total somatic coliphages (viruses that infect *E. coli*) and male-specific (F-) RNA coliphages have been used as viral indicators of faecal contamination (Griffin et al., 2000; United States Environmental Protection Agency, 2001; Cole et al., 2003; Harwood et al., 2005). However, the results presented here suggest that coliphages are not appropriate for this purpose. DNA coliphages were found in potable and reclaimed water, indicating that their presence does not reflect water quality. In addition, the metagenomic data gathered during this study demonstrate that coliphages were not the most abundant phages in reclaimed water (Fig. 3). Phages infecting *Salmonella* spp. were more abundant in reclaimed water (~7%) than coliphages (~5%) and their abundance in potable water was not as high (~4%) as coliphages (~12%). The most abundant host in reclaimed water libraries was *Burkholderia* spp. (~9.5%), while phages infecting this host were less abundant in potable water (~2.5%). Therefore, phages that infect hosts other than *E. coli* should be explored as potential bioindicators of water quality.

**Eukaryotic DNA viruses**

The role of reclaimed water in the dissemination of human, animal, and plant viral pathogens is currently unknown. Therefore, one of the goals of this study was to examine the prevalence of different eukaryotic viral groups in reclaimed water. The types of eukaryotic viruses found in reclaimed water metagenomic libraries, based on their best BLASTX homologies in GenBank, are summarized in Tables 2 and 3. No eukaryotic viruses were detected among the sequences from potable water (Fig. 2C).

Eukaryotic viral sequences in the DNA libraries were dominated by viruses containing single-stranded DNA (ssDNA) circular genomes, including plant pathogens from the *Geminiviridae* and *Nanoviridae* families (Tables 2 and 3) and animal pathogens from the *Circoviridae* family (Table 3) (Rosario et al., 2009). It is important to note that the amino acid identities to known viral proteins were less than 60%, suggesting that these are novel viruses with weak similarities to known ssDNA viruses. The

Fig. 3. Distribution of phage families (A), identified through TBLASTX analysis of DNA contigs against the Phage Sequence Database and host representation within prophage (B), Siphoviridae (C), Myoviridae (D) and Podoviridae (E).
Identification of ssDNA viruses in reclaimed water suggests that they may be resistant to chlorination. This is consistent with the known resistance of small ssDNA viruses to wastewater treatment (Nwachukwu and Gerba, 2004) and suggests the presence of these viruses in treated effluent should be further explored.

Circular ssDNA viruses are likely overrepresented in the libraries due to the multiple displacement amplification (MDA) step utilized before pyrosequencing. MDA has been shown to selectively amplify circular single-stranded genomes by 2 or 3 orders of magnitude compared with other DNA types in a mixed community (Kim et al., 2008). Despite this enrichment, this study demonstrates that circular ssDNA viruses can be disseminated through reclaimed water. To date, most of the research regarding circular ssDNA viruses has focused on pathogens infecting agriculturally important crops and animals (e.g. Todd, 2000; Seal et al., 2006). However, a recent metagenomic study used MDA to identify novel circular ssDNA viral genomes in a rice paddy soil sample (Kim et al., 2008), suggesting they are more abundant in the environment than previously recognized.

© 2009 Society for Applied Microbiology and Blackwell Publishing Ltd, Environmental Microbiology, 11, 2806–2820
Table 3. Eukaryotic viruses, other than plant viruses, identified in reclaimed water metagenomic libraries.

<table>
<thead>
<tr>
<th>Virus host</th>
<th>Virus homologue</th>
<th>Amino acid identity range (%)</th>
<th>Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA libraries</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>Rhinovirus sp.</td>
<td>24–35</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Enterovirus sp.</td>
<td>23–33</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Parechovirus</td>
<td>26–35</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Aichi virus</td>
<td>34</td>
<td>Nursery</td>
</tr>
<tr>
<td>Animal</td>
<td>Porcine enterovirus sp.</td>
<td>25</td>
<td>Effluent</td>
</tr>
<tr>
<td></td>
<td>Bovine enterovirus</td>
<td>25</td>
<td>Effluent</td>
</tr>
<tr>
<td></td>
<td>Simian enterovirus sp.</td>
<td>26–30</td>
<td>Nursery</td>
</tr>
<tr>
<td></td>
<td>Senecio valley virus</td>
<td>32</td>
<td>Effluent</td>
</tr>
<tr>
<td></td>
<td>Lyssavirus</td>
<td>25</td>
<td>Effluent</td>
</tr>
<tr>
<td></td>
<td>European brown hare syndrome virus</td>
<td>25</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Hepatitis C virus subtype 1a</td>
<td>29</td>
<td>Effluent*</td>
</tr>
<tr>
<td>Invertebrates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tauro syndrome virus</td>
<td>25–57</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Cricket parvovirus</td>
<td>26–56</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Drosophila C virus</td>
<td>27–35</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Plautia stali intestine virus</td>
<td>27–51</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Rhopalosiphum padula virus</td>
<td>34</td>
<td>Nursery</td>
</tr>
<tr>
<td></td>
<td>Hymenolepis P virus</td>
<td>25–42</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Triatoma virus</td>
<td>22–42</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Aphid lethal paralysis virus</td>
<td>27–56</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Kashmir bee virus</td>
<td>38–43</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Acute bee paralysis virus</td>
<td>32–43</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Israel acute paralysis virus of bees</td>
<td>30–74</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Solenopsis inimicus virus 1</td>
<td>25–48</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Homaloctogaster coagulata virus-1</td>
<td>24–54</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Honey bee slow paralysis virus</td>
<td>27–31</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Nora virus</td>
<td>26–54</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Bemisia virus</td>
<td>31</td>
<td>Effluent</td>
</tr>
<tr>
<td></td>
<td>neath domesticus virus</td>
<td>31</td>
<td>Effluent</td>
</tr>
<tr>
<td>Other Euk (Algae, diatoms, fungi)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Heterosigma akashiwoi RNA virus</td>
<td>27–51</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Schizochytrium single-stranded RNA virus</td>
<td>31–50</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Rhodosolenia sedigeri RNA virus</td>
<td>24–57</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Sclerotiphora macrospora virus A</td>
<td>36–57</td>
<td>Effluent*, Nursery*</td>
</tr>
<tr>
<td>DNA libraries</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Animal</td>
<td>Bird circovirus sp.</td>
<td>27–65</td>
<td>Effluent*, Nursery*, Park</td>
</tr>
<tr>
<td></td>
<td>Swine circovirus sp.</td>
<td>35–50</td>
<td>Effluent*, Nursery*, Park</td>
</tr>
<tr>
<td></td>
<td>Canary circovirus sp.</td>
<td>51–57</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Crocodile circovirus</td>
<td>31–33</td>
<td>Effluent</td>
</tr>
<tr>
<td></td>
<td>Cercopithecine herpesvirus sp.</td>
<td>31–33</td>
<td>Effluent</td>
</tr>
<tr>
<td></td>
<td>Lymphocystis disease virus 1</td>
<td>30</td>
<td>Effluent</td>
</tr>
<tr>
<td></td>
<td>Rock bream iridovirus</td>
<td>31</td>
<td>Effluent</td>
</tr>
<tr>
<td></td>
<td>Parapoxivirus</td>
<td>41</td>
<td>Effluent</td>
</tr>
<tr>
<td>Invertebrates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Equine herpesvirus sp.</td>
<td>26</td>
<td>Nursery</td>
</tr>
<tr>
<td></td>
<td>Epiphytes poitiviana NPV</td>
<td>60</td>
<td>Effluent</td>
</tr>
<tr>
<td></td>
<td>Courtelina zostadica iridescent virus</td>
<td>42</td>
<td>Effluent</td>
</tr>
<tr>
<td></td>
<td>Invertebrate iridescent virus 6</td>
<td>31–35</td>
<td>Nursery</td>
</tr>
<tr>
<td>Other Euk (Algae, protists)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pamecium buana Chlorella virus sp.</td>
<td>25–62</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Acantamoeba tibetica Chlorella virus 1</td>
<td>35–51</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Ostreococcus virus OsV5</td>
<td>32–45</td>
<td>Effluent, Nursery</td>
</tr>
<tr>
<td></td>
<td>Eutylorus hueyi virus 88</td>
<td>39</td>
<td>Nursery</td>
</tr>
<tr>
<td></td>
<td>Acanthamoeba polyphaga minima</td>
<td>30–41</td>
<td>Effluent, Nursery</td>
</tr>
</tbody>
</table>

The identity range refers to the amino acid level identity between contigs and their homologous in the nr protein database. Sequences similar to DNA viruses that were found in RNA libraries and vice versa.

Eukaryotic RNA viruses

All the viral-like sequences in the RNA viral metagenomes were similar to positive-sense RNA eukaryotic viruses (Tables 2 and 3). Sequences similar to proteins from the proposed Picornavirales order (Le Gall et al., 2008) dominated the libraries, suggesting an unprecedented abundance of novel picorna-like viruses in reclaimed water. Some sequences were similar to picornaviruses that infect vertebrates, including members of the Enterovirus, Parechovirus, Rhinovirus and Kobuvirus genera, which have been associated with diseases in humans and

© 2009 Society for Applied Microbiology and Blackwell Publishing Ltd, Environmental Microbiology, 11, 2806-2820
animals (e.g. Yamashita et al., 1995; 2001; Rotbart, 2002; Yamada et al. 2004; Benshop et al., 2008). However, this
does not necessarily mean that there are viruses in
reclaimed water that cause human disease, since amino
acid identities to known human pathogens were very low
(<36%). Furthermore, the majority of the picorna-like
viruses were similar to viruses that infect invertebrates.
Numerous sequences were similar to marine picorna-like
viruses, Marine JF-A and JP-B, believed to infect protists
based on phylogeny and genome characteristics (Culley
et al., 2007). These sequences were classified in the envi-
nronmental category (Fig. 2C) because the Marine JF-A
and JP-B genomes were assembled from marine metagen-
omic libraries and a definitive host has not been
identified (Culley et al., 2006).
Reclaimed water RNA libraries also contained
sequences similar to insect and arthropod picornaviruses
from the family Dicistroviridae and an insect virus (i.e.
Nora virus) that may belong to a new picorna-like family
(Habayeby et al., 2006). A recent study of stool from
South Asian children also identified an abundance of
sequences related to picornaviruses of invertebrates,
including members of the Dicistroviridae and Nodaviridae
families (Victoria et al., 2009). However, these types of
viruses have not been reported in human faeces from
the USA population and, thus, the source of novel
picorna-like viruses found in reclaimed water remains to
be determined. Future studies need to investigate the
relationship between picorna-like viruses in reclaimed
water and viruses from known hosts to determine if the
reclaimed water viruses belong to known picornavirus
families and if virus hosts can be inferred based on
phylogenetic analysis (Culley et al., 2003; Culley and
Steward, 2007).
Both DNA and RNA reclaimed water metagenomes
contained a diverse group of sequences related to plant
viruses (Table 2). Most of the viruses were novel as they
shared less than 60% amino acid identities to known
viral proteins. However, all the viruses belonging to the
Tobamovirus genus and the Melon necrotic spot virus
(MNSV) from the Tombusviridae family had high amino
acid identity (>80%) to known plant pathogens. Viruses
from these groups are known to have strict host special-
isations to organic solvents and non-ionic detergents
(Fauquet et al., 2005). In the early 1980s, the tombusvi-
rus Tomato bushy stunt virus was used to demonstrate
that humans can act as carriers of plant pathogens by
consuming infected produce and shedding infective viral
particles in their faeces (Tomlinson and Faithfull, 1982). It
was suggested that plant viruses with no known vectors,
such as most tobamoviruses and tombusviruses, may
have certain ‘alimentary resistance’ (i.e. stay intact after
passing through alimentary tract), which enables humans
and other animals to act as carriers of these viruses
(Tomlinson and Faithfull, 1982). Interestingly, all the
tobamoviruses and MNSV detected in the reclaimed
water metagenomes have also been previously detected
in faecal samples from healthy individuals (Zhang et al.,
2005). The tobamovirus Pepper mild mottle virus
(PMMoV) was the most abundant virus in RNA viral
metagenomes from individual faecal samples, and was
still capable of infecting plants after passing through the
human gut (Zhang et al., 2006). PMMoV was subse-
quently detected at concentrations greater than 10^6
copies ml^-1 in both raw sewage and treated effluent samples
collected throughout the USA (E. Symonds, K. Rosario
and M. Breitbart, unpublished). MNSV and most of the
tobamoviruses, including PMMoV, were detected in
reclaimed water samples both at the point-of-discharge
(Effluent library) and at the point-of-use (Nursery library),
indicating that stable viruses may reach irrigation
systems. The tobamovirus Tomato mosaic virus (ToMV)
and MNSV have been detected in irrigation systems in
other countries (Gosalvez et al., 2005; Boben et al.,
2007); however, data regarding the presence of plant
pathogens in irrigation systems in the USA are not avail-
able. These findings indicate that reclaimed water may
serve as a mechanism for the spread of highly stable
plant pathogens that exhibit alimentary resistance. The
infectivity of these plant viruses needs to be examined
to determine if reclaimed water use represents a potential
problem for the agricultural sector.

Conclusions
This study identified the dominant DNA and RNA viral
types in reclaimed water, thus making a significant con-
tribution to current microbiological data regarding treated
wastewater. The DNA viral community in both reclaimed
and potable water was dominated by phages. However,
there were clear differences between both communities
as demonstrated by phage family representation and host
distribution in the different libraries. From a water quality
standpoint, it is useful to evaluate which types of phages
endure wastewater treatment processes in order to iden-
tify potential viral bioindicators. Finding strong bioindica-
tors that correlate with the presence of human viruses is
not an easy task as different viruses exhibit varying levels
of resistance to wastewater treatment (Nwachukwu and
Gerba, 2004). Natural phage populations found in waste-
water offer a range of resistance to disinfection (chlo-
rina) that may represent most of the viruses that can be
found in water (Duran et al., 2003). Therefore, phage
populations in reclaimed water offer an untapped source
of potential bioindicators.

The metagenomes also uncovered a wealth of novel
eukaryotic viruses present in reclaimed water. DNA
metagenomic libraries revealed the presence of viruses similar to ssDNA viruses, including plant and animal pathogens from the Geminiviridae, Nanoviridae and Circoviridae families. The RNA metagenome contained an abundance of plant pathogens known to be resistant to environmental degradation, including members of the Tobamovirus genus and Tombusviridae family. The RNA metagenome also contained an abundance of picorna-like viruses. Some of these picorna-like viruses may be related to human pathogens; however, the majority of these novel viruses are most closely related to insect and plant viruses. None of the established human pathogens (e.g., enteroviruses, hepatitis viruses and caliciviruses) were detected during this study, suggesting that these viruses were not abundant relative to phages and other eukaryotic viruses in the reclaimed water samples. The genetic information gathered during this study can be used to design molecular assays to detect viral types of interest and assess their abundance in wastewater and ecosystems exposed to wastewater discharge. Future research needs to evaluate the host range, infectivity and ecological impacts of novel viruses identified in reclaimed water to ensure the appropriate use of this important alternative water supply.

**Experimental procedures**

**Enumeration and visualization of virus-like particles**

SYBR Gold staining and epifluorescent microscopy (Chen et al., 2001; Shibata et al., 2006; Patel et al., 2007) was used to enumerate VLPs in well, potable and reclaimed water samples as well as in raw sewage. Potable water samples (n = 3) were collected from spigots at the point-of-use (plant nursery) in Manatee County (Bradenton, FL). Well water samples (n = 5) were collected from private shallow wells in Pinellas County (St. Petersburg, FL). Reclaimed water samples were collected at the point-of-discharge (i.e. wastewater treatment plant, n = 3) and at the point-of-use (i.e. public sprinklers, fountains, spigots at a plant nursery, n = 8) in Pinellas County (St. Petersburg, FL) and Manatee County (Bradenton, FL). Raw sewage samples (n = 3) were collected from a raw inflow stream at a wastewater treatment plant. Raw sewage and reclaimed water samples at the point-of-discharge and point-of-use originated from the same wastewater treatment facility, which uses activated sludge followed by chlorination for disinfection. All samples for VLP enumeration were collected in sterile 50 ml conical tubes and processed within 3 h. Samples were fixed with 2% paraformaldehyde and then subsamples were filtered onto a 0.22 μm Anodisc (Whatman, Maidstone, Kent, UK). For raw sewage and reclaimed water samples, 10 μl was diluted into 1 ml of sterile water to prepare the slides, whereas 3–5 ml of well and potable water samples was directly filtered onto the Anodisc. Filters were stained with 1% SYBR Gold (Invitrogen, Carlsbad, CA, USA) for 10 min in the dark, and virus particles were counted digitally (Chen et al., 2001) in at least eight fields of view for each sample.

Transmission electron microscopy was used to examine the morphology of virus particles in reclaimed water. For this purpose, purified viral particles (see below) were dried down overnight onto copper 400-mesh carbon coated formvar grids. Grids were stained with 1% uranyl acetate for 20 s and air-dried before visualization on a Hitachi 7100 transmission electron microscope.

**Viral isolation and extraction of nucleic acids**

Virus samples were purified from four different samples: potable water (‘Potable’), reclaimed water at the point-of-discharge from the wastewater treatment plant (‘Effluent’) and reclaimed water at two points-of-use, a plant nursery (‘Nursery’) and a public park sprinkler (‘Park’) (Table 1). For each sample, viruses were concentrated and purified from 50 to 100 l of water (Table 1) using a combination of tangential flow filtration, density-dependent centrifugation and nucleic acid treatment (Breitbart et al., 2002; 2003; 2004; Zhang et al., 2006; Thorber et al., 2009). Each water sample was first filtered through a 0.2 μm tangential flow filter (TFF) (GE Healthcare, Westborough, MA, USA) to remove bacteria, eukaryotes and large particles. Viruses in the filtrate were concentrated using a 100 kDa TFF until the final sample volume was less than 1 l. All TFF viral concentrates were then filtered through a 0.22 μm Starwex filter (Millipore, Billerica, MA, USA) to remove any bacterial contamination. The viral concentrate from potable water was further concentrated through polyethylene glycol (PEG 8000) precipitation in order to obtain enough nucleic acids for metagenomic sequencing, and treated with 10% chloroform to remove contaminating microbial cells. TFF viral concentrates from reclaimed water and the PEG-precipitated concentrates from potable water were loaded onto a cesium chloride (CsCl) density gradient, ultracentrifuged at 61 000 g for 3 h at 12°C, and the 1.2–1.5 g ml⁻¹ fraction was collected. Examination of the different CsCl fractions by SYBR Gold staining and epifluorescent microscopy revealed a significant number of VLPs that remained in the sample reservoir, therefore the sample fraction was re-loaded onto a second CsCl density gradient and processed with the same procedure. The 1.2–1.5 g ml⁻¹ fractions collected from both gradients were pooled together. This double CsCl gradient procedure allowed the recovery and isolation of the vast majority of VLPs. After CsCl purification, viral fractions were treated with DNase I to further eliminate free DNA. Samples were not treated with RNase due to the potential for RNase to destroy the nucleic acids of some RNA viruses (Griffin et al., 2000; Cole et al., 2003). DNase-treated VLPs were further concentrated using centrifugal concentration filters (Microcon Ultracel YM-30; Millipore, Bedford, MA, USA) before nucleic acid extraction. After final purification and concentration, viral DNA and RNA were simultaneously extracted from viral concentrates using the QIAamp MinElute Virus Spin Kit (Qiagen, Valencia, CA, USA). Extracted nucleic acids were split into two fractions, one for DNA libraries and one for RNA libraries. The RNA fraction was treated with DNase I using the DNA-free Kit (Ambion, Austin, TX, USA) followed by random-primed cDNA synthesis using the SuperScript III First-Strand Synthesis System for RT-PCR (Invitrogen).
Library construction and pyrosequencing

Four DNA and two RNA viral metagenomes were pyrosequenced at the Genome Institute of Singapore (Table 1). For this purpose, DNA and cDNA samples were amplified in triplicate reactions with GenomiPhi V2 DNA Amplification Kit (GE Healthcare, Little Chalfont, Buckinghamshire, UK). All GenomiPhi reactions were purified using standard phenol/ chloroform extraction and ethanol precipitation (Sambrook and Russell, 2001). The amplification products were pooled and 4 μg was processed for pyrosequencing with the GS-FLX sequencer (454 Life Sciences, Roche) according to the manufacturer's protocol. The Park-DAH sample was amplified in duplicate by the GenomiPhi reaction and purified amplification products (4 μg) were processed for shotgun sequencing using a GS20 sequencer (454 Life Sciences, Roche). Sequences have been deposited to the Short Read Archive (SRA) at NCBI (accession numbers: SRA008294).

Bioinformatics

For metagenome sequence analysis, raw reads longer than 100 nt were assembled into contigs using SeqMan (DNASTAR, Madison, WI) with a criteria of ≥ 95% identity over at least 35 nt (Table 1). Contigs larger than 200 nt were then compared against the GenBank non-redundant (nr) protein database using BLASTX (Altschul et al., 1997) (E-value < 0.001). These BLASTX results were analysed using the Metagenome Analysis System (Huson et al., 2007) to identify the different taxa present in reclaimed and potable water viral metagenomes. Contigs with best matches to cellular organisms (i.e. bacteria, archaea or eukaryotes) were further compared against the ACLAME (A Classification of Mobile Genetic Elements) database using the same parameters as in GenBank to identify mobile genetic elements, including phage and plasmids (Lepage et al., 2004; 2006). All contigs from the DNA libraries with significant hits to phages and cellular organisms were also analysed against the Phage Sequence Database (http://scmu.sdsu.edu/philage/) using the BLASTX program (E-value < 0.001) in order to identify the dominant phage families in the DNA metagenomes.

A number of chimeric sequences were identified after manual examination of contigs. This chimera problem was more pronounced in the RNA libraries, making genome assemblies impossible from the RNA libraries. These artifacts were most likely the result of the multiple displacement amplification (MDA) used to obtain enough nuclei acids for pyrosequencing (Lasken and Stockwell, 2007). In addition, MDA is known to have amplification biases selecting for circular ssDNA templates (Kim et al., 2008). Therefore, the relative abundance of different viral sequences, such as ssDNA and RNA viruses, was not used to infer the relative abundance of different DNA and RNA eukaryotic viruses. Instead, this study focused on contigs to evaluate the different types of viruses present in reclaimed water as opposed to investigating the absolute abundance of individual raw reads. Analysis of contigs rather than individual raw reads has several advantages. First, this strategy reduces the number of sequences to be analysed by eliminating redundant sequences (i.e. identical sequences assemble into the same contig). Second, contigs are longer than individual reads, which increases the chances of finding significant matches in the database by increasing sequence length (Wommack et al., 2000). Third, analysis of contigs may help collapse the effects of artifacts caused by MDA, as chimeras are likely to form within the same template (Lasken and Stockwell, 2007).

Acknowledgements

This research was funded by grants to M.B. from the National Science Foundation (MCB-0701854) and Alfred P. Sloan Foundation (BR-4772). K.R. was funded by a Bridge to the Doctorate Fellowship from the National Science Foundation and the Saint Petersburg Progress Endowed Fellowship. The Genome Institute of Singapore (GIS) sequencing team provided the Roche 454 sequencing data. Y.F. was funded by Singapore A*STAR and NHGRI (R01 HG004456-01 and R01HG003521-01). The authors would like to acknowledge Tony Greco for the TEM analysis and Dana Hall and Wah Heng Lee for assistance with the bioinformatics analysis. Reclaimed water sampling was possible thanks to the collaboration of Oshang Kashel, Jan Tracy, David Shulmister and Chris Collins.

References


Florida Department of Environmental Protection (2007) 2006 Reuse Inventory. Tallahassee, FL, USA: Florida Department of Environmental Protection.


Ikeda, H., and Tomizawa, J.I. (1969) Prophage P1 an extra-


Villena, C., El-Senousy, W.M., Abad, F.X., Pinto, R.M., and


APPENDIX C:

PEPPER MILD MOTTLE VIRUS AS AN INDICATOR

OF FECAL POLLUTION
Pepper Mild Mottle Virus as an Indicator of Fecal Pollution

Karyna Rosario, Erin M. Symonds, Christopher Sinigalliano, Jill Stewart, and Mya Breitbart

College of Marine Science, University of South Florida, St. Petersburg, and NOAA National Ocean Service, Center for Coastal Environmental Health and Biomolecular Research, Charleston, South Carolina; and Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina

Received 18 February 2009/Accepted 31 August 2009

Accurate indicators of fecal pollution are needed in order to minimize public health risks associated with wastewater contamination in recreational waters. However, the bacterial indicators currently used for monitoring water quality do not correlate with the presence of pathogens. Here we demonstrate that the plant pathogen Pepper mild mottle virus (PMMoV) is widespread and abundant in wastewater from the United States, suggesting the utility of this virus as an indicator of human fecal pollution. Quantitative PCR was used to determine the abundance of PMMoV in raw sewage, treated wastewater, seawater exposed to wastewater, and fecal samples and oral and intal oral homogenates from a wide variety of animals. PMMoV was present in all wastewater samples at concentrations greater than 1 million copies per milliliter of raw sewage. Despite the ubiquity of PMMoV in human feces, this virus was not detected in the majority of animal fecal samples tested, with the exception of chicken and seagull samples. PMMoV was detected in four out of six seawater samples collected near point sources of secondary treated wastewater off southeastern Florida, where it co-occurred with several other pathogens and indicators of fecal pollution. Since PMMoV was not found in nonpolluted seawater samples and could be detected in surface seawater for approximately 1 week after its initial introduction, the presence of PMMoV in the marine environment reflects a recent contamination event. Together, these data demonstrate that PMMoV is a promising new indicator of fecal pollution in coastal environments.

Existing wastewater treatment practices are not always effective at removing the large number of pathogens (bacteria, protozoa, and viruses) present in human feces (17, 42, 47–49, 51). Therefore, wastewater discharges into the environment can have a negative impact on human health. Recreational waters throughout the United States are monitored for the presence of fecal pollution as a means of limiting public exposure to pathogens in areas impacted by wastewater discharges (44). The presence of pathogenic viruses in aquatic environments is an important parameter to consider in the evaluation of water quality. However, the bacterial indicators currently used to detect fecal contamination, such as fecal coliforms and enterococci, often do not correlate with the presence of feces-associated viruses and other pathogens (5, 10, 26, 33, 37, 51). In response, several researchers have proposed the use of viral indicators as a more effective method for monitoring wastewater contamination and the associated risks to public health (11, 14, 31).

To date, the majority of the proposed viral indicators of fecal pollution are enteric viruses transmitted via the fecal-oral route (4). Enteric viruses present in raw sewage (including members of the families Adenoviridae, Caliciviridae, Picornaviridae, and Reoviridae, and of the genus Anellovirus) have been used in several previous studies to identify fecal pollution in the environment (7, 8, 11, 12, 13, 18, 19, 27, 28, 32–36, 38, 50, 51). Of the enteric viruses that have been used as indicators, only the adenoviruses were ubiquitously found in raw sewage samples collected throughout the United States (41). Picornaviruses and Torque teno virus are abundant in raw sewage from some regions and have also been proposed as indicator viruses (15, 41). However, one potential problem with the use of human viruses as indicators is that their abundance in wastewater depends on the degree of infection and shedding in the human population at any given time.

In addition to viruses infecting humans, other viruses shed in feces may be useful for indicating wastewater pollution. The plant pathogen Pepper mild mottle virus (PMMoV) was the most abundant virus found in a metagenomic survey of RNA viruses from human feces (52). PMMoV is a positive-sense, single-stranded RNA virus that belongs to the Tobamovirus genus and infects hot, bell, and ornamental peppers (Capsicum spp.) (9). The nonenveloped, rod-shaped PMMoV virions are extremely stable (9) and have been demonstrated to retain their infectivity for plants after passage through the human gut (52). PMMoV originates from processed pepper products (e.g., hot sauce and curry) and is excreted in human feces at concentrations of 1 million to 1 billion viruses per g (dry weight) (52). Since the presence of PMMoV in human feces is dietary in origin, this plant pathogen may be more abundant in the healthy human population than viruses that cause human disease.

This study analyzed the presence of PMMoV in raw sewage and treated wastewater samples collected from wastewater treatment facilities throughout the coastal United States. To determine if PMMoV is a human-specific indicator useful for
tracking the source of fecal pollution, fecal samples from numerous animals were tested for this virus. Finally, the presence of PMMoV in marine environments exposed to wastewater was determined and compared to that of other microbial indicators. The results of this work demonstrate that PMMoV is a promising indicator of fecal pollution.

MATERIALS AND METHODS

Detection of PMMoV: Quantitative qPCR (qPCR) targeting the virus- associated protein was used to quantify the abundance of PMMoV. For this purpose, RNA was extracted from purified PMMoV samples (provided by Scott Ackert, U.S. Department of Agriculture, Agricultural Research Service [USDA-ARS]) by using the QiaAmp MinElute Virus spin kit (Qiagen, Valencia, CA). After the extraction of nucleic acid, cDNA was synthesized using the First Strand Synthesis SuperScript III reverse transcription kit (Invitrogen, Carlsbad, CA) with random hexamer primers, and 2 μl of this cDNA was used for qPCR. In order to create a standard curve, PMMoV-cDNA was purified using the UltraClean PCR clean-up kit (Mo Bio Laboratories, Inc.), and the amount of cDNA was quantified using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE). The number of targets in this positive-control PMMoV-cDNA was determined using the measured cDNA concentration (μg) and the PMMoV genome size (i.e., 6,857 nucleotides) according to the following equation:

\[
\text{ng} = \frac{1 \mu g}{10^9 \text{ng}} \times 10^9 \mu g = 6.82 \times 10^7 \text{ng} 
\]

\[
1 \mu l = 1 \text{ target} 
\]

The PMMoV cDNA was then serially diluted to 1 target/μl for use as a standard curve in each qPCR assay.

The 50-μl qPCR reaction contained 2 μl of target cDNA, 1X TaqMan Universal PCR Master Mix No AmpErase UNG (Applied Biosystems, Foster City, CA), 400 μM each primer (PMMV-PF, 5′-GAG TGG TTG GAC CTT GAT CTA 3′; PMMV-RP, 5′-GCT GCT TCC GCA AGC CAA 3′), and 125 nM TaqMan probe (PMMV probe) (FAM-CCT ACC GAA CCA AAT G-GTT C) (52). The qPCR mixture was incubated at 95°C for 10 min, followed by 45 cycles of 95°C for 10 s, 57°C for 1 min, and 72°C for 1 min; finally, the mixture was incubated at 72°C for 10 min. To account for any PCR inhibition, each sample, both undiluted and diluted 1:10, was assayed in duplicate. The qPCR assay was capable of detecting 100 copies of PMMoV per reaction. It should also be noted that all detection limits reported are based on purified target DNA in pure water and not on an environmental water matrix.

Preparation of wastewater samples. Ten-milliliter samples of raw sewage and treated wastewater were gathered from 12 wastewater treatment facilities throughout the coastal United States (Table 1) and were concentrated according to the methods of Symonds et al. (41). One raw sewage sample and one treated- wastewater sample each were gathered from Alabama, California, Connecticut, Louisiana, Maine, Maryland, New Jersey, North Carolina, Oregon, and Washington. Two sets of samples were collected from different wastewater treatment facilities in the state of Florida (one in mainland Florida and one in the Florida Keys). To determine the daily variability in PMMoV abundance, six samples of raw sewage and treated wastewater were collected over a 2-week period from the wastewater treatment plant in the Florida Keys. For each 10 ml sample, a 0.45 μm pore-size polyethersulfone membrane filter cartridge (Millipore, Billerica, MA) was used to remove bacteria and larger particles. Using the Centriprep YM-50 and Microcon Ultrafree YM-30 centrifugal concentration device (Millipore), the filtrate was concentrated to less than 200 μl. Nucleic acid was extracted and cDNA synthesized using the methods mentioned above, and 2 μl of this cDNA was used for qPCR.

Efficiency of isolation and extraction methods. Known quantities of PMMoV were added to a treated wastewater sample in order to determine the effectiveness of the methods employed to isolate viral cDNA. Ten milliliters of treated wastewater was spiked with a final concentration of 2.52 × 10³ PMMoV copies/ml. Another 10 ml aliquot of treated wastewater served as a control throughout the entire process of detection. Viral concentration, nucleic acid isolation, reverse transcription, and viral detection processes. Isolation of viral cDNA and qPCR for PMMoV were performed for both samples as described above.

Analysis of PMMoV and other fecal indicators in seawater exposed to wastewater discharge. Seawater samples influenced by wastewater discharge were collected and analyzed in order to determine the concentrations of PMMoV and other microbial water quality markers. For this purpose, seven seawater samples were collected aboard the NOAA Nancy Foster Ship during a 6-day cruise along the southeast coast of Florida in February 2008. Samples were collected as part of a large multi-institutional research collaboration led by the NOAA Florida Area Coastal Environment (FACE) program, utilizing many different analytes simultaneously in order to better characterize the discharge of these pollutants to the coastal environment (broad data from the FACE study will be published separately). Sample sites consisted of six treated-wastewater outfall sites from the following locations (in order from north to south): South Carolina (28°27′37″N, 80°02′52″W), Boca Raton (26°20′37″N, 80°03′28″W), Broward (26°15′12″N, 80°03′50″W), Hollywood (26°01′50″N, 80°15′19″W), Miami Beach (25°55′18″N, 80°15′06″W), and Miami Central (25°45′32″N, 80°16′42″W). These six outfalls (from north to south) have average daily discharge flow rates of 12, 11, 36, 40, 80, and 105 million gallons per day, respectively (25). Samples were collected from Gulf Stream surface water to serve as an unpolished offshore deep-water site control (26°37′39″N, 79°54′16″W). In addition to PMMoV quantitation, all samples were assayed for 15 other microbial water quality markers (culturable enterococci, total enterococci, 254 Bacteria genes, the human-specific Enterococcus faecalis esp gene, culturable Bacteroides, a human-specific Bacteroides detected by the 16S rDNA-CCD assay, the human-specific Bacteroides spp pH 9 affinity, Aeromonas, Cryptobacterium novyi, Vibrio cholerae, the human adenosine deaminase gene, the Campylobacter jejuni espG gene, the Salmonella sp. espG gene, the coagulase-negative Staphylococcus aureus espF gene, and the Endococcus sp. espG577 gene).
Appendix C (Continued)
FIG. 2. PMMoV concentrations over time during incubation experiments to determine the stability of PMMoV in coastal seawater. Error bars represent 1 standard deviation for triplicate samples.

**DISCUSSION**

The primary goal of this study was to assess the utility of PMMoV as an indicator of fecal pollution in the coastal marine environment. For this purpose, the abundance of PMMoV was determined in raw sewage and treated wastewater, as well as in seawater samples collected in areas exposed to wastewater. Large quantities of PMMoV were identified in all of the raw-sewage and treated-wastewater samples collected throughout the United States. In this study, PMMoV was found at concentrations between $1.50 \times 10^9$ and $2.16 \times 10^9$ copies/mL of raw sewage. These concentrations are higher than the maximum concentrations reported in the literature for human viruses that have been proposed as indicators of fecal pollution. For example, qPCR has shown that raw sewage can contain concentrations as high as $1.16 \times 10^9$ copies of adenovirus, $6.5 \times 10^8$ copies of *Tauraena virus*, $1.8 \times 10^8$ copies of norovirus, and $8.9 \times 10^7$ copies of poliovirus per milliliter (3, 6, 16, 25). The extremely high concentrations of PMMoV detected in human sewage suggest that this virus would be a good indicator of human fecal pollution. One of the advantages of using PMMoV instead of human enteric viruses to indicate fecal pollution is that the presence of PMMoV in wastewater is independent of active human infection. Repeated sampling of a single treatment plant demonstrated only limited variability in the levels of PMMoV in raw sewage over a 2-week period, which consistently averaged 1 million copies per milliliter. However, PMMoV concentrations in treated wastewater at the same treatment plant varied by more than 2 orders of magnitude during the same period, suggesting day-to-day variation in wastewater treatment efficiency. Seasonal variability was not examined in this study; however, since PMMoV is dietary in origin and is not dependent on active human infection, no large seasonal variations are expected. Before PMMoV can be used as a fecal indicator in other parts of the world with different dietary preferences, studies will need to determine the prevalence of PMMoV in sewage from each geographic region, as well as the baseline presence of PMMoV in local recreational waters.

In addition to being abundant in raw sewage, PMMoV was also found in all of the treated-wastewater samples examined. This demonstrates that PMMoV cannot be used to differentiate between contamination with raw sewage versus input of treated wastewater. Instead, PMMoV serves as a conservative viral tracer of fecal pollution that can be used to represent microconstituents and pathogens that may not be removed effectively through wastewater treatment processes. Further work is needed to understand the correlation between the detection of PMMoV and the infectivity of human pathogens throughout the wastewater treatment process.

The high concentration of PMMoV in wastewater suggests that this virus could be useful for detecting wastewater contamination in recreational waters. PMMoV was found in four out of six surface bores from oceanic wastewater outlets on the southeastern coast of Florida. PMMoV was not detected in the two northernmost wastewater outlets, which have lower daily discharge flow rates and tested negative for many of the water quality markers (Table 2). The co-occurrence of PMMoV with many other sewage-associated indicators and pathogens in oceanic wastewater outlets shows that PMMoV is a promising indicator of human fecal pollution in the marine environment. PMMoV may also be used to indicate fecal contamination in freshwater samples; however, the natural abundance of PMMoV and the persistence of positive signals in these environments will need to be determined.

Experiments were also performed to determine the stability of PMMoV in seawater in order to understand if the detection of PMMoV in the environment reflects a recent or past pollution event. Detectable concentrations of PMMoV were present for approximately 1 week after its initial introduction in seawater at temperatures ranging from 31 to 33°C (Fig. 2). An additional experiment conducted in seawater at temperatures ranging from 22 to 27°C showed similar results, with detectable amounts of PMMoV present for approximately 1 week (data not shown). This demonstrates that detection of PMMoV in the marine environment reflects a recent contamination event. Future work needs to determine if a correlation exists between the presence of PMMoV and disease risks in polluted environments.

An ideal indicator of fecal pollution would also be able to distinguish the source of the contamination. Since PMMoV is dietary in origin, we hypothesized that this virus might be specific to human sewage. To determine if PMMoV could

<table>
<thead>
<tr>
<th>Source of sample</th>
<th>Location (type) of sample</th>
<th>No. of samples</th>
<th>$10^8$ PMMoV copies/ mg (range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chicken</td>
<td>Arkansas (I)</td>
<td>1</td>
<td>ND</td>
</tr>
<tr>
<td>Chicken</td>
<td>California (I)</td>
<td>4</td>
<td>ND</td>
</tr>
<tr>
<td>Chicken</td>
<td>Delaware (I)</td>
<td>1</td>
<td>ND</td>
</tr>
<tr>
<td>Chicken</td>
<td>Georgia (I)</td>
<td>5</td>
<td>190–119</td>
</tr>
<tr>
<td>Chicken</td>
<td>Missouri (I)</td>
<td>1</td>
<td>5.34–14.6</td>
</tr>
<tr>
<td>Chicken</td>
<td>North Carolina (I)</td>
<td>3</td>
<td>ND</td>
</tr>
<tr>
<td>Chicken</td>
<td>Idaho (F)</td>
<td>1</td>
<td>190–216</td>
</tr>
<tr>
<td>Seagull</td>
<td>Florida (F)</td>
<td>12</td>
<td>ND</td>
</tr>
<tr>
<td>Seagull</td>
<td>California (F)</td>
<td>5</td>
<td>5.84–9.55</td>
</tr>
</tbody>
</table>

*I, intestinal homogenate; F, feces.

*ND, not detected; the level of PMMoV was below the detection limit of the assay.*
distinguish between sources of fecal pollution, the abundance of PMoV in the feces of several animal species was examined. PMoV was not detected in the feces of the majority of animals tested (e.g., pigs, horses, dogs, and cows); however, PMoV was detected in samples from chickens and seagulls. Therefore, the presence of PMoV in environmental samples does not necessarily represent human fecal contamination. However, PMoV is consistently found in high concentrations in human sewage, in contrast to the less frequent detection of lower concentrations in the bird feces; thus, high concentrations in recreational waters may reflect a human source. Baseline levels of any indicator should be investigated before its implementation in new areas.

In conclusion, this study has shown that PMoV is a promising indicator of fecal pollution in the marine environment. As a plant pathogen, PMoV is different from other proposed viral indicators of fecal contamination in that its presence in sewage is dietary in origin and is not dependent on active human infection. Further research is necessary to understand how the presence of PMoV correlates with the infectivity of enteric pathogens of concern throughout the wastewater treatment process and in aquatic environments exposed to fecal pollution.

ACKNOWLEDGMENTS

This research was funded by grants to M.R. from the U.S. Environmental Protection Agency (R710643-9-01-0) and to the Alfred P. Sloan Foundation (BR-4772). Methods comparison was funded in part by the NOAA Oceans and Human Health Initiative. K.R. was funded by a Bridge to the Doctorate Fellowship from the National Science Foundation and the Saint Petersburg Program Endowed Fellowship. Ship time and the coastal sampling and analysis of treated-wastewater ocean outfalls for other indicators and pathogens were supported and conducted by the NOAA FACE Program.

This work could not have been completed without the cooperation of the wastewater treatment facilities where sampling was performed. Thanks to the following people and organizations for supplying animal feces and/or animal intestines: Canadian Department of Agriculture, Marine-Loe Lake Islands (Agriculture and Agric-Food Canada), Parker Hall (USDA-APHIS, Wildlife Services), Daniel McManus (USDA), Pat and Gary Rohwer, the Rosario family, Suncoast Sea Bird Sanctuary (Indian Shores, FL), John Griffin (Southern Coastal California Coastal Water Research Project, Costa Mesa, CA), and Gabi Vargo (Boyd Hill Nature Preserve, St. Petersburg, FL). Chris Johnston and Laura Webster of the NOAA Center for Coastal Environmental Health and Biocontainment Research are gratefully acknowledged for their analysis of the wastewater samples for M. anulatus and norovirus, respectively.

REFERENCES

Appendix C (Continued)
APPENDIX D:

DIVERSE CIRCOVIRUS-LIKE GENOME ARCHITECTURES REVEALED

BY ENVIRONMENTAL METAGENOMICS
Diverse circovirus-like genome architectures revealed by environmental metagenomics

Karyna Rosario,1 Siobain Duffy2 and Mya Breitbart1

1College of Marine Science, University of South Florida, St. Petersburg, FL 33701, USA
2School of Environmental and Biological Sciences, Rutgers, New Brunswick, New Jersey, USA

Single-stranded DNA (ssDNA) viruses with circular genomes are the smallest viruses known to infect eukaryotes. The present study identified 10 novel genomes similar to ssDNA circoviruses through data-mining of public viral metagenomes. The metagenomic libraries included samples from reclaimed water and three different marine environments (Chesapeake Bay, British Columbia coastal waters and Sargasso Sea). All the genomes have similarities to the replication (Rep) protein of circoviruses; however, only half have genomic features consistent with known circoviruses. Some of the genomes exhibit a mixture of genomic features associated with different families of ssDNA viruses (i.e. circoviruses, geminiviruses and parvoviruses). Unique genome architectures and phylogenetic analysis of the Rep protein suggest that these viruses belong to novel genera and/or families. Investigating the complex community of ssDNA viruses in the environment can lead to the discovery of divergent species and help elucidate evolutionary links between ssDNA viruses.

Single-stranded DNA (ssDNA) viruses that replicate through a rolling circle mechanism are the smallest viruses (≤3 kb) known to infect plants and animals. Circular ssDNA viruses are mechanistically predisposed to recombination (LeFeuvre et al., 2009) and experience high nucleotide substitution rates (Duffy et al., 2008; van der Walt et al., 2008). These characteristics may contribute to the emergence of ssDNA viruses as serious pathogens (Hou & Gilbertson, 1996; LeFeuvre et al., 2009; Ma et al., 2007; Varsani et al., 2008; Zhou et al., 1997). Circular ssDNA eukaryotic viruses include plant pathogens from the families Nanoviridae and Geminiviridae and animal viruses from the genera Circovirus, Gyrovirus and Anellovirus (de Villiers & zur Hausen, 2009; Greenborn, 2004; Gutierrez, 1999; Mankertz et al., 1997; Naviald et al., 2008; Noteborn & Koch, 1995; Okamoto & Mayumi, 2001). Although circular ssDNA viruses include pathogens of agricultural, veterinary and clinical concern (Almeida et al., 2009; Chae, 2005; de Villiers & zur Hausen, 2009; Moffat, 1999; Rishi, 2009; Todd, 2000), little is known about their prevalence or diversity in natural environments.

To date, most of the research regarding eukaryotic circular ssDNA viruses has focused on recombinant expression or specific detection methods, such as PCR, to further characterize known, economically important pathogens or detect related viruses in known hosts (e.g. Almeida et al., 2009; Banda et al., 2007; Johnne et al., 2003, 2006; Kakkoila et al., 2007; Mankertz et al., 2000; Ninomiya et al., 2009; Seal et al., 2006; Zhou et al., 2005). However, the incorporation of multiple displacement amplification (MDA) into metagenomic analyses provides a method to detect ssDNA viruses in environmental viral communities. MDA can be used to enrich for ssDNA circular genomes as it can selectively amplify circular ssDNA genomes by two or three orders of magnitude in a mixed community (Haiblé et al., 2006; Kim et al., 2008). A number of recent viral metagenomic studies utilizing an MDA step have identified novel sequences related to circular ssDNA eukaryotic viruses in different samples, including rice paddy soil (Kim et al., 2008), reclaimed water (Rosario et al., 2009) and marine animal tissues (Ng et al., 2009a,b; Vega Thurber et al., 2008). Therefore, small circular ssDNA viruses may be more widespread in the environment than previously recognized.

This study identified circular ssDNA viral genomes in the environment through data-mining of public viral metagenomes generated with an MDA step. Specifically, ssDNA viruses were found in reclaimed water (RW) (Rosario et al., 2009) and three different marine environments, including the Chesapeake Bay (CB) (MOVE858 shotgun dataset), British Columbia coastal waters (BBC) (Angly et al., 2006) and the Sargasso Sea (SAR) (Angly et al., 2006). These datasets were chosen because viral concentrates from these environments were available for confirmation of metagenomic sequence assemblies through PCR. For metagenomic sequence assemblies, the reclaimed water dataset was retrieved from the NCBI short read archive (accession no. SRA008294.7), the MOVE858 dataset was collected from the community cyberinfrastructure for advanced
marine microbial ecology research and analysis (CAMERA) website (http://camera.calit2.net/index.php), while the BBC (SEED accession no. 4440305.3) and SAR (SEED accession no. 4440322.3) metagenomes were obtained from the SEED platform (http://www.jisoeed.org/DinsdaleSupplementalMaterial; Dinsdale et al., 2008). Sequences from each dataset were assembled using the SeqMan program (DNASTAR) with a criteria of >95% identity over at least 35 nt. To identify eukaryotic circular ssDNA viruses, contigs larger than 1000 nt were compared against GenBank using BLASTX (E-value ≤0.001) and results were summarized using the metagenome analyser (MEGAN) software (Huson et al., 2007).

All of the datasets contained hundreds of sequences that assembled into contigs with similarities to viruses in the genus Circovirus. Therefore, this study focused on circovirus-like sequences. It is important to note that contigs similar to nanoviruses were detected in the SAR, CB and RW datasets and geminivirus-like sequences were present in the CB and RW datasets. No anellovirus- or gyrovirus-like sequences were detected in any of the datasets. Contigs containing complete circular genomes with similarities to circoviruses were verified by PCR (see Supplementary Table S1, available in JGV Online) and are described further below. Any contigs that contained only partial genomes or that could not be confirmed by PCR (possibly due to chimeras) were excluded from the analysis. Since only a small portion of the circovirus-like contigs met these criteria, this study is a conservative estimate of the novelty and diversity of circoviruses in the environment. In addition, there was a high degree of sequence variability between clones consistent with the “quasispecies” diversity observed in ssDNA viruses (Ng et al., 2009a).

Ten circovirus-like genomes, ranging in size from 1739 to 2819 nt, were reconstructed from uncloned water and marine metagenomic sequences (Fig. 1; GenBank accession nos FJ959077-FJ959086). All the genomes contained a major open reading frame (ORF) with similarities to the viral replication protein (Rep) of circoviruses. According to the Pfam database classification (Finn et al., 2008; http://Pfam.sanger.ac.uk/search), all of the putative Reps contained a domain similar to the viral Rep family (PF02407), but some Reps also had similarities to an RNA helicase domain (PF00910) (Table 1). Pfam analysis revealed that some known circoviruses only have the viral Rep domain (e.g., canary, finch, columbids and raven circoviruses) while others have both the viral Rep and the RNA helicase domains (e.g., bovine, porcine, beak and feather disease, and duck circoviruses). Overall, amino acid identities to known Rep proteins were less than 40% for all the environmental circovirus-like genomes, suggesting that these are novel viruses (Table 1).

A phylogenetic analysis was performed on the viral Reps to evaluate the relationship between the novel circovirus-like genomes and known viruses. For this purpose, deduced Rep sequences from environmental circovirus-like genomes were aligned against members of the viral Rep family (PF02407) in the Pfam database. This protein family also includes non-viral replication-associated proteins from a plasmid (Bifidobacterium pseudocatenulatum pN4) and protists (Giardia intestinalis and Entamoeba histolytica) (Gibbs et al., 2006). The CB_B (GenBank accession no. FJ959083) Rep was not included in the alignment as this sequence was too divergent and missed several conserved amino acids identified through the Hidden Markov model logo for the viral Rep family (Schuster-Boeckler et al., 2004). Although the SAR_A (GenBank accession no. FJ959084) Rep had several conserved amino acids present in the viral Rep family, this sequence was also excluded due to its small size (180 aa). Amino acid sequence alignments indicate that the novel circular genomes are more related to circoviruses than to nanoviruses (Fig. 2). None of the genomes clustered with known circovirus, plasmid or protist Rep proteins, suggesting that these viruses belong to novel families and/or genera of ssDNA viruses.

Although all the genomes have similarities to the Rep protein of circoviruses, only half of them contained other genomic features consistent with known circoviruses. Circoviruses are characterized by a small (≤2063 nt), circular ssDNA genome that contains two major ORFs, encoding the Rep and capsid proteins (Cap), in an ambisense organization (Todd et al., 2005). The only genomes with this organization are RW_A, RW_B, RW_C, SAR_A and BBC_A (Type I; Fig. 1). These genomes contain two major ORFs, a putative Rep and an unknown ORF (no significant homologues in GenBank, E-value <0.001) that are divergently organized (Fig. 1). Known circoviruses have an origin of replication located upstream of both major ORFs and this region contains the nonanucleotide motif at the apex of a stem-loop (Mankertz, 2008; Todd, 2000). RW_B, RW_C and SAR_A contain a conserved nonanucleotide motif (TACTTATAC) at the apex of a potential stem-loop. Genomes RW_A and BBC_A contain a similar motif (Table 1). Although RW_C and BBC_A genomes contained a genome organization similar to circoviruses, the intergenic region containing the potential stem-loop is located downstream from major ORFs (Fig. 1).

In contrast with the genomes discussed above, genomes RW_D, RW_E, SAR_B, CB_A and CB_B (Types II and III; Fig. 1) do not have genomic features consistent with circoviruses. The CB_B genome (Type II) contained broad organizational similarities to circoviruses, but the Rep is split into three overlapping ORFs and no potential stem-loops were identified near the conserved nonanucleotide sequence (Fig. 1). The ORF upstream (CB_BRep1) had hits to the viral Rep in the Pfam database, the middle ORF (CB_BRep2) did not have any significant matches, while the partial ORF downstream (CB_BRep3) had hits to the RNA helicase family (E-value=0.0053; Table 1). This suggests that the Rep domains of CB_B are in an order consistent with circovirus Reps. Furthermore CB_BRep3 had significant hits to the Rep protein of porcine
circoviruses (PCV) in GenBank (Table 1). PCV Reps, which contain the viral Rep–RNA helicase architecture, are differentially transcribed to produce two proteins essential for virus replication, Rep and Rep' (Cheung, 2003, 2004; Mankertz, 2008). PCV transcriptional splicing results in the removal of an intron and the expression of the Rep’ C terminus in a different reading frame (Mankertz & Hillenbrand, 2001). It is possible that CB_B has a similar
Table 1. Circovirus-like features identified in the environmental genomes

<table>
<thead>
<tr>
<th>Genome</th>
<th>Nucleotide sequence</th>
<th>Rep family domains identified in Pfam (E-value)</th>
<th>Best hit*</th>
<th>Arginine/lysine-rich region†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Top hit (E-value)</td>
<td>Amino acid</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Identity (%)</td>
<td>Length (aa)</td>
<td></td>
</tr>
<tr>
<td>RW_A</td>
<td>CAGTATTAC</td>
<td>Viral Rep (1.3 × 10⁻⁴)</td>
<td>Canary circovirus (1.3 × 10⁻⁴)</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RNA helicase (7.2 × 10⁻⁴)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RW_B</td>
<td>TAGTATTAC</td>
<td>Viral Rep (4.3 × 10⁻⁶)</td>
<td>Porcine circovirus 1 (1.3 × 10⁻⁵)</td>
<td>33</td>
</tr>
<tr>
<td>RW_C</td>
<td>TAGTATTAC</td>
<td>Viral Rep (1.4 × 10⁻⁶)</td>
<td>Columbidi circovirus (6.5 × 10⁻⁶)</td>
<td>38</td>
</tr>
<tr>
<td>RW_D</td>
<td>AAGTATTAC</td>
<td>Viral Rep (5.2 × 10⁻⁶)</td>
<td>Porcine circovirus 2 (6.8 × 10⁻⁶)</td>
<td>32</td>
</tr>
<tr>
<td>RW_E</td>
<td>AAGTATTAC</td>
<td>Viral Rep (6.5 × 10⁻⁶)</td>
<td>Porcine circovirus 2 (1.3 × 10⁻⁵)</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RNA helicase (7.9 × 10⁻⁵)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CB_A</td>
<td>TAGTATTAC</td>
<td>Viral Rep (5.2 × 10⁻⁶)</td>
<td>Gull circovirus (1.3 × 10⁻⁵)</td>
<td>32</td>
</tr>
<tr>
<td>CB_B</td>
<td>TAGTATTAC</td>
<td>ORF1: Viral Rep (8.6 × 10⁻⁶) ORF3: RNA helicase (5.3 × 10⁻⁶)</td>
<td>Genus parvovirinae symptomless virus (5.5 × 10⁻⁶)</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR_A</td>
<td>TAGTATTAC</td>
<td>Viral Rep (6.2 × 10⁻⁶)</td>
<td>Porcine circovirus 1 (5.2 × 10⁻⁵)</td>
<td>36</td>
</tr>
<tr>
<td>SAR_B</td>
<td>TAGTATTAC</td>
<td>Viral Rep (5.7 × 10⁻⁶)</td>
<td>Gull circovirus (1.1 × 10⁻⁴)</td>
<td>36</td>
</tr>
<tr>
<td>BRC_A</td>
<td>TAGTATTAC</td>
<td>Viral Rep (2.5 × 10⁻⁶)</td>
<td>Starling circovirus (2.3 × 10⁻⁵)</td>
<td>40</td>
</tr>
</tbody>
</table>

*Best hit in GenBank using Rep as the query.
†The arginine/lysine-rich region is present (P) or not present (NP) at the N terminus of the major unknown ORF.

replication mechanism, needing two differentially transcribed proteins.

Similar to known circoviruses, the type III genomes (RW_D, RW_E, SAR_B and CB_A) were also characterized by two major ORFs, a putative Rep and an unknown ORF. However, these ORFs are non-overlapping and organized in the same orientation (Fig. 1). These genomes contained the conserved nonamer (CB_A and SAR_B) or a similar sequence (RW_D and RW_E) (Table 1) located within a potential stem-loop. These genome organizations have not been observed in circoviruses or other eukaryotic circular ssDNA viruses. On the other hand, these genome organizations are characteristic of linear ssDNA (4-6 kb) viruses from the family Parvoviridae. Some parvovirus genomes possess two major ORFs, encoding the Rep and Cap proteins, with the same polarity (Tattersall et al., 2005). In addition, other parvoviruses have the Rep and Cap proteins encoded on complementary strands and the Rep is split into three minor ORFs (Tattersall et al., 2005).

Based on similarities to the Rep of circoviruses but genome organizations that have only been observed in linear ssDNA viruses, the type II and III circular genomes most likely represent members of novel virus families.

Extrapolating from similar genome organizations in known ssDNA viruses, the unknown ORFs in the circovirus-like genomes may encode a Cap. Caps of many eukaryotic ssDNA viruses contain a region rich in basic amino acids at the N terminus (Niagro et al., 1998). The RW_A, RW_C and CB_A genomes had a lysine/arginine-rich region at the N terminus of the major unknown ORF which is characteristic of known circovirus Caps (Table 1). A search in Pfam revealed similarities to circovirus capsid proteins (PF022443) for the RW_A and RW_C genomes; however, the matches were not well supported (E-values=0.09 and 0.004, respectively). The major unknown ORF from RW_B had weak similarities (E-value=0.006) to geminivirus coat proteins (PF00844). This is interesting since the putative Rep from this genome had significant matches to the Rep from circoviruses (PF02447). Although geminiviruses and circoviruses have similar genome organizations (Niagro et al., 1998), they belong to different families and infect different hosts. None of the unknown ORFs in the other circovirus-like genomes had similarities to known Pfam protein families. Although the functions of these unknown ORFs are still undefined, this study places these sequences into circovirus-like genomes, thus assigning a portion of the large percentage of unknown sequences in metagenomic surveys (Angly et al., 2006) to genomes. This process may help identify divergent structural viral genes in the future.

Five of the novel circovirus-like genomes identified in this study originated from reclaimed water, the end product of wastewater treatment. The discovery of circovirus-like genomes in reclaimed water suggests that these viruses can be disseminated through the discharge of treated wastewater (Rosario et al., 2009). Although circovirus-like genomes have also been identified in stool from children (Victoria et al., 2009), the source of these viruses in reclaimed water remains to be determined.
Fig. 2. Condensed maximum-parsimony phylogenetic tree of deduced Rep amino acid sequences showing the relationship between reclaimed water (RW), Chesapeake Bay (CB), Sargasso Sea (SAR) and British Columbia coastal waters (BBC) genomes and members of the viral Rep protein family (PF02407). The tree includes a putative Rep sequence from a circovirus-like genome from soil. Alignments were performed using the CLUSTALW algorithm (Thompson et al., 1994) and clustalw2 as the similarity matrix in BioEdit version 7.0.9.0 (Hall, 1999). All sequences were trimmed to match a lysine residue at position 15 of the porcine circovirus 2 translated rep sequence (NC_005148) and the arginine residues at positions 276–277 of the same sequence. Alignments were performed over at least 200 aa and inspected manually. Phylogenetic trees were constructed in MEGA (Tamura et al., 2007) using the close-neighbour-interchange algorithm (CNI=3) with random addition of sequences (10 replicates). All alignment gaps were treated as missing data. The tree was manually rooted between the nanovirus and circovirus clades. One-thousand bootstrap resampling were performed to assess statistical support (only bootstrap values >70 are shown). Samples isolated for this study are printed in a larger font.

The five marine circovirus-like genomes originated from an estuarine environment (CB), coastal waters (BBC) and the open ocean (SAR). To our knowledge, the only known marine virus with similarities to circoviruses is the Chaetoceros salinigenum nuclear inclusion virus (CnIV; Nagasaki et al., 2005). The inferred Rep protein sequence of CnIV had weak hits to bird circoviruses. However, CnIV is not similar to known circoviruses in terms of size and genome organization (Nagasaki et al., 2005; Park et al., 2009). In contrast with CnIV, the circovirus-like SAR, BBC and CB genomes described in this study have stronger hits to circovirus Rep proteins, have similar genome sizes to known circoviruses and some have similar genome organization. Therefore, the novel genomes found in this
study are the first circovirus-like viruses identified in the marine environment.

The data presented here demonstrate the diversity of circovirus-like genome architectures in the environment. Intriguingly, some of the genomes revealed a mixture of genomic features associated with different families of ssDNA viruses. The abundance of circovirus-like sequences in environmental metagenomic studies and the presence of unique genome sequences and architectures suggest that there is a complex community of ssDNA viruses in the environment. Investigating these communities may lead to the discovery of divergent species that could illuminate the evolutionary links between ssDNA viruses. Future studies need to continue to explore the diversity of circular ssDNA viruses and investigate the ecology of these novel viruses.

Acknowledgements

This research was funded by a grant to M.B. from the National Science Foundation (MCB-0701981). K.R. was funded by a Bridge to the Doctorate Fellowship from the National Science Foundation and the Saint Peter’s Progress Endowed Fellowship. The authors would like to acknowledge Eric Womack and Danielle Winget for providing viral concentrates from the Chesapeake Bay.

References


Varsani, A., Shepherd, D. N., Monjane, A. L., Owor, B. E., Erdmann, J. B., Rybicki, E. P., Peterschmitt, M., Bridgon, R. W., Markham, P. G. & other authors (2008). Recombination, decreased host specificity and increased mobility may have driven the emergence of maize streak virus as an agricultural pathogen. *J Gen Virol* 89, 2063–2074.


### Supplementary Table S1. List of primers used for PCR verification and sequencing of circovirus-like genomes

<table>
<thead>
<tr>
<th>Genome</th>
<th>Primer ID</th>
<th>Sequence (5’→3’)</th>
<th>Annealing temperature (°C)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>RW_A</td>
<td>RWAcirc_541F</td>
<td>GAGGGCAGCTCTTGTGC</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>RWAcirc_1430R</td>
<td>ACCCTCTCTGCTGAGTCC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RWAcirc_346out</td>
<td>ACCACACGTACCCCGCTTT</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>RWAcirc_1607out</td>
<td>CTCAAACGACAAGACCTTACCGAA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RWAcirc_937F</td>
<td>ACGTCCGCTGTCAGTGA</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>RWAcirc_1828R</td>
<td>GATTGCGAAGACGTAGGT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RWAcirc_37F</td>
<td>CACCCTACCTGCTTCTCATATT</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>RWAcirc_729R</td>
<td>TCCGCTTGATTTAGTCTCT</td>
<td></td>
</tr>
<tr>
<td>RW_B</td>
<td>RWBcirc_1002F</td>
<td>AAAACACATTGCTGCTTCCGAG</td>
<td>53. TD (−0.1 °C per cycle)</td>
</tr>
<tr>
<td></td>
<td>RWBcirc_1537R</td>
<td>AGTTCCCTGGGAATTCTTCTT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RWBcirc_1963F</td>
<td>AGCAACCAACGGGAGATATTG</td>
<td>53. TD (−0.1 °C per cycle)</td>
</tr>
<tr>
<td></td>
<td>RWBcirc_2509R</td>
<td>TCCGTCTACGCTGTCTTCCA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RWBcirc_68F</td>
<td>TTCTAAAGCTGCTAGCTACCT</td>
<td>53. TD (−0.1 °C per cycle)</td>
</tr>
<tr>
<td></td>
<td>RWBcirc_502R</td>
<td>TGCGCCTATTTGTGTCCACT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RWBcirc_1470F</td>
<td>GTTTCAGCCTGCACCATAG</td>
<td>53. TD (−0.1 °C per cycle)</td>
</tr>
<tr>
<td></td>
<td>RWBcirc_1781R</td>
<td>TCCTGACATCTACAGGCTAAT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RWBcirc_1537out</td>
<td>AGCAAGAGATCCACGGGACTAAT</td>
<td>60. TD (−0.2 °C per cycle)</td>
</tr>
<tr>
<td></td>
<td>RWBcirc_502R</td>
<td>TGGGCAATTGTGTCTCAACT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RWBcirc_1002F</td>
<td>AAAAAAGTTGCTGCTTAGAG</td>
<td>60. TD (−0.2 °C per cycle)</td>
</tr>
<tr>
<td></td>
<td>RWBcirc_502R</td>
<td>TGCGCCTATTTGTGTCCACT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RWBcirc_68F</td>
<td>TTCTAAAGCTGCTAGCTACCT</td>
<td>60. TD (−0.2 °C per cycle)</td>
</tr>
<tr>
<td></td>
<td>RWBcirc_1781R</td>
<td>TCCTGACATCTACAGGCTAAT</td>
<td></td>
</tr>
<tr>
<td>RW_C</td>
<td>RWCcirc_31F</td>
<td>GTGACATTGGGAGCTGAAT</td>
<td>53. TD (−0.1 °C per cycle)</td>
</tr>
<tr>
<td></td>
<td>RWCcirc_638R</td>
<td>GAACGTAAAGCTGAGGCA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RWCcirc_1124F</td>
<td>TGGCTATGAGAATCTGTTT</td>
<td>53. TD (−0.1 °C per cycle)</td>
</tr>
<tr>
<td></td>
<td>RWCcirc_1779R</td>
<td>TCTCCCTGACCACTATACC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RWCcirc_1124F</td>
<td>TGGCTATGAGAATCTGTTT</td>
<td>53. TD (−0.1 °C per cycle)</td>
</tr>
<tr>
<td></td>
<td>RWCcirc_982R</td>
<td>TGTTTAGCACTGTCTGCTGTA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RWCcirc_31F</td>
<td>GTGACATTGGGAGCTGAAT</td>
<td>53. TD (−0.1 °C per cycle)</td>
</tr>
<tr>
<td></td>
<td>RWCcirc_1779R</td>
<td>TCTCCCTGACCACTATACC</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RW_D</th>
<th>RW_Dcirc_638out</th>
<th>TGTGCGCTAGCTTACGTTTC</th>
<th>Used for sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>RW_Dcirc_1838F</td>
<td>ACATCAATGGGTGCTTAGG</td>
<td>53. TD (-0.1 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>RW_Dcirc_2193R</td>
<td>GGAGGCTGTTGGACGAGGA</td>
<td>53. TD (-0.1 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>RW_Dcirc_1453F</td>
<td>CCAAGGCGCTGGTAGATATAG</td>
<td>53. TD (-0.1 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>RW_Dcirc_2298F</td>
<td>CGCATGATCGTCTAAATCG</td>
<td>53. TD (-0.1 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>RW_Dcirc_315F</td>
<td>GACAAAGGATCCGGAATTTG</td>
<td>53. TD (-0.1 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>RW_Dcirc_1030F</td>
<td>CGACCATCTGTGAAAATGG</td>
<td>52.5. TD (-0.1 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>RW_Dcirc_1086F</td>
<td>CAATTTTACATATAAGCGATGTTG</td>
<td>52.5. TD (-0.1 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>RW_Dcirc_1795R</td>
<td>CCCGAATGTTTTCTTATCATTC</td>
<td>60. TD (-0.2 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>RW_Dcirc_315F</td>
<td>GACCAAGGATCCGGAATTTG</td>
<td>60. TD (-0.2 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>RW_Dcirc_1453out</td>
<td>CTAATATCTACAGGCGCTTGG</td>
<td>60. TD (-0.2 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>RW_E</td>
<td>RW_Ecirc_250F</td>
<td>TGGTTCTGGACAATCCCAAGA</td>
<td>53. TD (-0.1 °C per cycle)</td>
</tr>
<tr>
<td>RW_Ecirc_697R</td>
<td>GGTACCGCCTGAGGCGCATCTA</td>
<td>53. TD (-0.1 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>RW_Ecirc_260F</td>
<td>CCGAATCTGCTAAGTGACGCA</td>
<td>53. TD (-0.1 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>RW_Ecirc_905R</td>
<td>AAAACCTAACAGCCGACCGAG</td>
<td>53. TD (-0.1 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>RW_Ecirc_2019F</td>
<td>ATCCCAATGTTGCACTTCCC</td>
<td>53. TD (-0.1 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>RW_Ecirc_2666F</td>
<td>GGGTGATTGTTGGCGACCTCT</td>
<td>53. TD (-0.1 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>RW_Ecirc_12389out</td>
<td>CACGAAAAATCCAGATTACC</td>
<td>50. TD (-0.1 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>RW_Ecirc_905out</td>
<td>CTGTTGCGGTGATAGTTT</td>
<td>50. TD (-0.1 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>RW_Ecirc_1238F</td>
<td>GGGTACTGATGGGTCTTGG</td>
<td>50. TD (-0.1 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>RW_Ecirc_905R</td>
<td>AAAACCTAACAGCCGACCGAG</td>
<td>50. TD (-0.1 °C per cycle)</td>
<td></td>
</tr>
<tr>
<td>CB_A</td>
<td>CBA1circ_1326F</td>
<td>CCCCCCTACAGTTGGCCCTATACAC</td>
<td>50</td>
</tr>
<tr>
<td>CBA1circ_1844R</td>
<td>GGATGATGATTGAGGACATCCA</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>CBA1circ_868F</td>
<td>TAAGAGAGTGCGCAGAAACAAA</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>CBA1circ_1632R</td>
<td>GATCGAAGGGGTAGGCTGAAG</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>CBA1circ_906F</td>
<td>GGCGATGAGGCTAGGATTCG</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>CBA1circ_1347R</td>
<td>GTGGATAGGGCAACTGTAAGGG</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>CBA1circ_32F</td>
<td>TGGTCAATCAAGGAATTTTGG</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>CBA1circ_617R</td>
<td>CGAATATCCACCTTATAGCC</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>CBA1circ_32out</td>
<td>CACCAATTCTCTGATCTGACA</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>CBA1circ_1844out</td>
<td>TGATATGCTCTATCATCATGCC</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>CBA1circ_125out</td>
<td>CTTTCCGGTTCTACCTACTT</td>
<td>53. TD (-0.1 °C per cycle)</td>
<td></td>
</tr>
</tbody>
</table>

| CBAcirc_1681out | GAATTTGTGCATAGCCACCT  |
| CBAcirc_32F     | TGTCGAGATCGAAGATTCTTG  |
| CBAcirc_1844R   | GCATGATGAGGAGACATCCA  |
| CBBrirc_1909F   | AGAAGATGGCGGAAATAATG   |
| CBBrirc_2466R   | TATAAGGGCTTCCACGCTTT  |
| CBBrirc_281F    | CCGTGGTGCTGATCCTATAAT  |
| CBBrirc_1123R   | GCTTAAATCTCCACGACAGAC |
| CBBrirc_1643F   | TGAATTGTAGTGTCGCCCTT  |
| CBBrirc_2466R   | TATAAGGGCTTCCACGCTTT  |
| CBBrirc_751F    | ATTTGAAGATCTGCTGTAATG  |
| CBBrirc_1365R   | CATACGCTCAAGCGCAACAGT  |
| CBBrirc_1045F   | CCTATGACTGGGATGATGATA  |
| CBBrirc_1661R   | GAGGGACACTCACGCTCAAG  |
| CBBrirc_281out  | ATTATAGGATCAACACGGCG  |
| CBBrirc_2466out | AACAGGAGTGAAACCGCTTATA |
| CBBrirc_gapF1   | ACCGGCAGAGACTGCACTT   |
| CBBrirc_gapR1   | TGAATCTCTGACGGTGAA    |
| CBBrirc_outF1   | ACTTGGGCCGATAGGACAAGA |
| CBBrirc_outR1   | GAATCCAAAGAGACGCTTG   |
| CBBrirc_1365out | ACTTTGTGGCCTTTAGCGTATG |
| CBBrirc_1045F   | CCTATGACTGGGATGATGATA  |
| CBBrirc_2466R   | TATAAGGGCTTCCACGCTTT  |
| CBBrirc_2466out | AACAGGAGTGAAACCGCTTATA |
| CBBrirc_1365R   | CATACGCTCAAGCGCAACAGT  |
| SAR_A           | TCCGTTGGCTGTGACAAATA  |
| SARAirc_707F    | CCTACAGTTGACATGCAAT   |
| SARAirc_700R    | TACCCTGTCGACCATTTCA   |
| SARAirc_1558out | ACCACCTGCAACACGAGG    |
| SARAirc_1340R   | GCCAAGAACATGAAAGAGA   |
| SARAirc_339F    | CTCAACCGGCTCCAAAACAA  |
| SARAirc_1558R   | TGAATTTGTGCACAGGAGTA  |
| SARAirc_8F      | CCCCCTATGTGTGTTTGAC   |
| SARAirc_339out  | TTGTTGGAGCGGTAGGAG    |
| SAR_B           | ACACGGTTTCTGAGGAACACC  |
| SARBirc_142F    | GATGGTGGCAGGTACAGG    |
| SARBirc_1941R   | TGGAGATCTGGGAAATTGTCTTG |

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
<th>Temperature</th>
<th>Cycle Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARBcirc_360F</td>
<td>TGTATC TGCCACCACGAAAA</td>
<td>60</td>
<td>TD (−0.2 °C per cycle)</td>
</tr>
<tr>
<td>SARBcirc_142out</td>
<td>GGTGTGCTTGTGAACGGTGT</td>
<td>60</td>
<td>TD (−0.2 °C per cycle)</td>
</tr>
<tr>
<td>SARBcirc_1089F</td>
<td>GGTCAAGGCTAAACAGGTTCG</td>
<td>60</td>
<td>TD (−0.2 °C per cycle)</td>
</tr>
<tr>
<td>SARBcirc_1534R</td>
<td>CTACACGGATTTGTCCATT</td>
<td>60</td>
<td>TD (−0.2 °C per cycle)</td>
</tr>
<tr>
<td>SARBcirc_360F</td>
<td>TGTAATCGTCCACCAGAAAA</td>
<td>60</td>
<td>TD (−0.2 °C per cycle)</td>
</tr>
<tr>
<td>SARBcirc_1158R</td>
<td>ACACGTGCTTGTCAACTGC</td>
<td>60</td>
<td>TD (−0.2 °C per cycle)</td>
</tr>
<tr>
<td>SARBcirc_1512F</td>
<td>AGTGTGCTGTGCAACTCTC</td>
<td>60</td>
<td>TD (−0.2 °C per cycle)</td>
</tr>
<tr>
<td>SARBcirc_1941R</td>
<td>GAAATTCTGGAATTTGCTTGT</td>
<td>60</td>
<td>TD (−0.2 °C per cycle)</td>
</tr>
<tr>
<td>SARBcirc_142F</td>
<td>ACAACGTTATCGAAGCAACC</td>
<td>60</td>
<td>TD (−0.2 °C per cycle)</td>
</tr>
<tr>
<td>SARBcirc_914R</td>
<td>TGCCAGCGAAGCTGTATTCA</td>
<td>60</td>
<td>TD (−0.2 °C per cycle)</td>
</tr>
<tr>
<td>BBC_A BBCAcirc_62F</td>
<td>GATTAGAAATCGGCTACGG</td>
<td>60</td>
<td>TD (−0.2 °C per cycle)</td>
</tr>
<tr>
<td>BBCAcirc_655F</td>
<td>TCGTGATGGCATTTTTGTT</td>
<td>60</td>
<td>TD (−0.2 °C per cycle)</td>
</tr>
<tr>
<td>BBCAcirc_62out</td>
<td>CGGTAGCGATTTTCTCAATT</td>
<td>53</td>
<td>TD (−0.1 °C per cycle)</td>
</tr>
<tr>
<td>BBCAcirc_618F</td>
<td>AACAAAAATCGACATCAGA</td>
<td>53</td>
<td>TD (−0.1 °C per cycle)</td>
</tr>
<tr>
<td>BBCAcirc_1277F</td>
<td>GAGGGTGTGAGTGTGCAAGGA</td>
<td>53</td>
<td>TD (−0.1 °C per cycle)</td>
</tr>
<tr>
<td>BBCAcirc_1814R</td>
<td>CAGGGTGTTGGAAATCAACT</td>
<td>Used for sequencing</td>
<td></td>
</tr>
<tr>
<td>BBCAcirc_1951R</td>
<td>CACACGTCTTATTCAAGGTA</td>
<td>Used for sequencing</td>
<td></td>
</tr>
</tbody>
</table>

*Some PCRs were performed with a touchdown (TD) to incrementally decrease the annealing temperature.

**PCR method.**

For some genomes, an MDA reaction was performed prior to the PCR using the GenomiPhi V2 DNA Amplification kit (GE Healthcare). Each PCR was performed in a 50 μl reaction containing 6 μl template (i.e. DNA extracted from viral concentrates) or 1 μl GenomiPhi product, 1× REDTag PCR Reaction Buffer (10 mM Tris/HCl pH 8.3, 50 mM KCl, 1.1 mM MgCl₂, 0.01% gelatin; Sigma-Aldrich), 1 μM each primer, 0.2 mM dNTPs, and 1 U REDTag DNA Polymerase (Sigma-Aldrich). Each PCR proceeded at 94 °C for 5 min, followed by 40 cycles of 94 °C for 1 min, annealing temperature (see Supplementary Table S1) for 45 s and 72 °C for 1 min, and a final extension at 72 °C for 8 min. Some PCRs were performed with a touchdown (TD) to incrementally decrease the annealing temperature.

APPENDIX E:

METHOD DEVELOPMENT FOR THE CHARACTERIZATION OF NOVEL SINGLE-STRANDED DNA VIRUSES
Appendix E

Introduction

Metagenomic analyses of environmental samples have led to the discovery of a diversity of novel single-stranded DNA (ssDNA) circoviruses in reclaimed water, marine habitats, lake water, and soil indicating that these viruses are more widespread than previously recognized (1-3). Generic PCR assays have also revealed a wealth of novel circoviruses in sewage and stool from humans and animals in different continents including Africa, South Asia, and North America (4, 5). However, nothing is known about the ecology of these novel ssDNA viruses. Further characterization of the novel ssDNA viruses is extremely difficult since these viruses have been identified directly from environmental metagenomes and their hosts are unknown. Therefore there is a need to develop strategies to bridge genomic data gathered from metagenomic datasets to actual biological properties of novel viruses. We propose that the isolation of novel ssDNA viral particles from the environment will allow further biological and physicochemical characterization of these unknown viruses. Isolation of novel viruses may be possible through recombinant expression of divergent structural proteins and development of immunoassays, such as immunoprecipitation, to isolate native viruses from environmental samples by virtue of the antigenic properties of structural proteins. Here we describe efforts to express a putative structural protein from a novel ssDNA circovirus identified in reclaimed water.

The environmental ssDNA virus chosen for this study has characteristics similar to members of the Circovirus genus within the Circoviridae family. Known circoviruses include pathogens of agricultural and veterinary concern as they can cause fatal diseases in swine and birds (6). However, to date it has not been possible to culture most
Appendix E (Continued)
circoviruses, with the exception of porcine circoviruses, making their study difficult. The lack of current culturing techniques for most circoviruses is a big limitation and, thus, many studies have taken advantage of recombinant expression technologies in hopes of better understanding virus life cycles and developing subunit vaccines for avian and porcine circoviruses (7-15). Furthermore, recombinantly expressed circovirus capsid (Cap) proteins sometimes spontaneously self-assemble into virus-like particles (VLPs) that are morphologically similar to wild type circoviruses and have immunogenic activity (8, 9, 13). These studies have established the immunogenicity of recombinant circovirus structural proteins. Therefore successful expression of the cap gene of environmental circovirus genomes may allow us to design immunological assays to detect and select for these viruses in the environment.

Methods

Plasmid Construction

The coding sequence of an unknown open reading frame (ORF) suspected to encode the capsid protein (designated UCap for unknown capsid) of the RW-A circovirus (Genbank accession no. FJ959077) was used for recombinant expression. Total viral DNA used to produce the reclaimed water metagenome was used to amplify the entire ucap ORF from RW-A using PCR (see Table 1 for primers). The PCR product was ligated into the pETBlue-1 vector (Novagen, Gibbstown, New Jersey). This pETBlue-UCap plasmid was digested with BglII and EcoRI to excise the ucap ORF, which was then ligated directionally into the pGEX-6P-2 vector (GE Healthcare, Piscataway, New Jersey). This vector is designed to express proteins fused to glutathione S-transferase
Appendix E (Continued)

(GST) which is advantageous for affinity purification of recombinant proteins. In addition to full length UCap, several truncated versions of this ORF were expressed. PCR primers containing SfiI sites were designed to obtain these UCap truncated versions from the pETBlue1-UCap vector (see Table 1 for primers). Truncated UCap versions were also ligated directionally into the pGEX-6P-2 vector. All pGEX-UCap constructs were propagated in *Escherichia coli* DH10β cells and plasmids were purified from cells using the QIAprep Spin Miniprep Kit (Qiagen, Valencia, California). The final plasmids were sequenced to confirm the GST-UCap fusion and proper phasing.

*Protein Expression and Purification*

The pGEX-UCap plasmids were expressed in *E. coli* Tuner cells (Novagen, Gibbstown, New Jersey). Briefly, transformed cells carrying the recombinant plasmid were grown to a desired optical density in LB medium containing carbenicillin (50 μg/ml), chloramphenicol (34 μg/ml), and glucose (1%) at 37°C. The expression of recombinant protein in Tuner cells was induced by adding isopropyl-β-thiogalactopyranoside (IPTG) and incubating at 37°C. To optimize the expression and cell culture conditions a range of incubation temperatures (25°C, 30°C, and 37°C), IPTG concentrations (0, 0.25, 0.50, 1, and 2 mM) and incubation times (3 – 5 hrs) were tested.

Cells were centrifuged at 1,500 xg for 5 min after induction to retrieve expressed proteins. The pelleted cells were lysed chemically by resuspending the cells in a mixture of 1X BugBuster Protein Extraction Reagent (Novagen, Gibbstown, New Jersey) (50 μl/ml of culture), Protease Inhibitor Cocktail III- EDTA-free (Calbiochem, Gibbstown, New Jersey) (5 μl/ml resuspended cells), and Lysonase Bioprocessing Reagent (Novagen,
Appendix E (Continued)

Gibbstown, New Jersey) (3 μl/ml resuspended cells) followed by 30 min incubation at room temperature. The lysates were centrifuged at 10,200 xg for 10 min at 4°C to separate soluble and insoluble protein fractions. Recombinant proteins in the soluble fraction were purified using a 50% slurry of Glutathione Sepharose 4B (GE Healthcare, Piscataway, New Jersey) following the manufacturer’s instructions for batch purification. All slurry washes were performed with a phosphate-buffered saline (PBS) solution containing 0.1% Tween. Once the recombinant protein was purified, the GST tag fused to the ucap ORF was removed by cleaving with the PreScission Protease™ (GE Healthcare, Piscataway, New Jersey). Expression and purification of recombinant GST-UCap were verified by SDS-PAGE and western blot using an anti-GST antibody.

Nuclear Localization Signal

A possible nuclear localization signal (NLS) sequence was identified on the UCap of RW-A. To test if this ORF is involved in nuclear localization, human embryonic kidney cells (293T cells) were transfected with a pcDNA3 vector (Invitrogen) containing the putative cap gene. The plasmid was constructed by amplifying the UCap ORF from the pETBlue-UCap vector through PCR using primers containing SfiI restriction sites (Sense 5’GTGGCCATTAAGGCCACCATGGGAAAGTACACAAAGCGA3’; Antisense 5’GTGG CCACCGCGGCTAAACCCTGTGCTTCC3’). The PCR products were digested with Sfi I and ligated directionally into the pcDNA3 vector. This pcDNA3-UCap plasmid allowed the expression of the putative Cap fused to an Avi-tag (biotin-acceptor peptide tag) at the C-terminus. The plasmid was propagated in E. coli DH10β cells and plasmids were purified from cells using a commercial kit. The final plasmid was sequenced to confirm the ucap
Appendix E (Continued)

insert and proper phasing. Human 293T cells were transfected using Lipofectamine 2000 (Invitrogen, Carlsbad, California) and incubated at 37°C for 72 hrs.

In order to evaluate where the expressed UCap-Avi tag localized, the cytosol and nuclei of harvested cells were lysed separately. To separate nuclei from cytosol, cells were centrifuged at 500 xg for 1 min. The pelleted cells were resuspended with PBS containing 0.1% Triton X and the resuspension was then centrifuged at 10,200 xg for 10 min to pellet nuclei. The supernatant (cytosol fraction) was collected and the pelleted nuclei were resuspended with dithiothreitol (50 mM) and incubated at 100°C for 10 min. Expression and localization of the UCap-Avi tag fused protein was verified by western blotting and immunofluorescent assays based on anti-Avi monoclonal antibodies. Human 293T cells expressing the green fluorescent protein (GFP) were used as a control.

Results and Discussion

This study explored a method for the recombinant expression of an unknown gene suspected to encode the structural protein of the novel ssDNA virus RW-A. The RW-A circovirus was identified from a reclaimed water viral metagenome (1, 16). The RW-A genome was confirmed by PCR and the virus was detected in raw sewage and treated effluent samples collected in two different years (2007 and 2009) suggesting that the virus is consistently present in wastewater. RW-A genomic features are consistent with known circoviruses including a small genome (2,162 nt) that encodes two major open reading frames (ORFs) in an ambisense organization and a conserved nonanucleotide motif located at the apex of a potential stem-loop structure (Fig. 1). However, RW-A only shares 36% amino acid identity within the replication-associated (Rep) protein of an
avian circovirus and there are no matches in the database for the second ORF. We hypothesized that the unknown ORF (i.e. ucap) encodes for the Cap protein due to its orientation with respect to the rep, size, and N-terminus amino acid composition.

The goal of this study was to obtain a purified recombinant protein that could be used to develop immunoassays to detect and isolate RW-A from environmental samples in the future. For this purpose, the RW-A UCap ORF was expressed using a bacterial expression system. The E. coli system was chosen because bacterial expression techniques are relatively simple and produce results in a short amount of time. The biggest concerns with this system are related to the possible outcomes of expressing eukaryotic proteins in a prokaryotic cell environment. It is possible that expressed proteins are not properly modified and expression of insoluble proteins or overproduction may result in the precipitation of the foreign protein into inclusion bodies. Nevertheless various studies have successfully expressed recombinant circovirus proteins in E. coli and retained their antigenic properties causing an immune response and reacting with antibodies for wild type viruses (10, 12, 14, 17).

Unfortunately, efforts to express and purify the RW-A UCap protein using a bacterial system were unsuccessful in this study. Although the GST-UCap fused protein was expressed in Tuner cells, low protein yields were obtained and degradation products were more concentrated than full length products. This outcome may be due to an arginine-rich region at the N-terminus of the predicted UCap amino acid sequence. This is characteristic of known circoviruses which contain a region high in basic residues at the N-terminus of the Cap, most notably arginine (6, 18). It has been confirmed that this sequence high in basic amino acids interferes with the cap gene expression in E. coli
resulting in low yields (12, 14) probably because of rare codon usage in *E. coli* (17). A few studies have successfully expressed the Cap protein by removing the arginine-rich region from the amino acid sequence (10, 14). Another recent strategy was to use an engineered *E. coli* strain that contains extra copies of genes that encode rare tRNAs (17). However, the expression of truncated versions of the GST-UCap and the use of an *E. coli* strain (Rosetta-gami™) with extra copies of rare tRNAs were also unsuccessful in this study. Truncated products were expressed, but precipitated into inclusion bodies and could not be purified.

The *E. coli* system was chosen due to its simplicity compared to other expression systems. Since the recombinant expression of RW-A UCap in *E. coli* resulted in typical problems encountered when expressing eukaryotic proteins in a prokaryotic environment, we suggest future studies express this protein in an eukaryotic expression system such as baculovirus. The baculovirus expression system is the most widely used expression system for the preparation of VLPs (19). In this system the gene of interest is inserted into an insect virus (vector) and the desired foreign protein is produced by growing the recombinant virus in cultured insect cells. This system has many advantages for the expression of eukaryotic viral proteins, most notably the expressed proteins are usually properly folded and transported to the proper cellular compartment (i.e. membrane proteins are localized to the insect cell membrane, nuclear proteins to the nucleus, and secreted proteins are secreted into the medium) (19).

The arginine-rich amino acid sequence at the N-terminus of circoviurs Cap proteins is believed to be a nuclear localization signal (NLS) involved in viral DNA translocation across the host nucleus (7, 15, 20). This potential NLS was identified at the
N-terminus of RW-A UCap ORF and, thus, we hypothesized this protein localizes to the nucleus. In order to test this hypothesis, the RW-A UCap ORF was expressed in human 293T cells. Transfected 293T cells successfully expressed the recombinant protein. Western blotting (Fig. 1) and immunofluorescent assays (Fig. 2) suggest that the putative Cap protein does localize to the nucleus. The karyophilic nature of the capsid protein has been shown for both porcine and avian circoviruses (7, 15, 20). The concentration of the RW-A UCap in the nucleus further supports that this unknown ORF encodes a capsid protein.

It is important to obtain recombinant structural proteins in order to develop strategies to study novel ssDNA viruses. In general, the combination of metagenomic sequencing, protein expression, and immunotechnology will allow us to make a connection between genes and viral particles. Purified capsid proteins and/or VLPs can be used to obtain polyclonal antibodies that could be used to isolate wild-type viruses from the environment. Immunomagnetic separation assays can then be developed to concentrate and isolate wild-type viruses corresponding to the sequences of interest. Once wild-type viruses are isolated we can study the physicochemical properties of the particles, such as morphology and virion stability. This information will be beneficial for taxonomic classification since the International Committee on Taxonomy of Viruses requires visualization of the viral particles in order to classify novel viruses (21). In addition, the isolation of viral particles will allow us to correctly annotate structural proteins expanding our knowledge regarding viral capsid proteins. This information will contribute correctly annotated viral sequences to the database, which will in turn allow for the identification of more divergent structural genes in the future. Finally, one of the
biggest challenges in viral metagenomics is finding potential hosts for unknown viral sequences. Immunoassays can also be used to determine potential hosts of unknown viruses. Western blot assays can be developed to screen a panel of sera from potential hosts including humans and animals. These Western blot assays can allow us to determine whether a reaction will occur between the sera and the recombinant protein (a positive reaction suggests that the host has been exposed to the protein in the past). In addition, if immunomagnetic separation assays are successful, it will be possible to perform infectivity studies with novel viruses on different hosts. All these efforts will add valuable information regarding unknown ssDNA viruses, which is critical as these viruses continue to be discovered through environmental metagenomics.
**Table 1.** Primers used to obtain the full length coding region for the RW-A unknown protein (UCap) and truncated versions of the protein

<table>
<thead>
<tr>
<th>Primer Pair (5’ – 3’)</th>
<th>UCap product</th>
</tr>
</thead>
</table>
| Sense: TTTGCATATGGAATGGGAAA  
Antisense: CAATCATAAACCCCTGTGCTTCC | Full length |
| Sense: GTGGCCATTATGGCCCGCAGAAAGAATCCCTTC  
Antisense: GTGGCCACCGCGGCCCTCATAAACCCCTGTGCTTCC | No arginine-rich region at N-terminus |
| Sense: GTGGCCATTATGGCCCGCAGAAAGAATCCCTTC  
Antisense: GTGGCCACCGCGGCCCTCATAAACCCCTGTGCTTCC | N-terminus half of the protein (excluding arginine-rich region) |
| Sense: GTGGCCATTATGGCCCAAGCCTTGTGGATTTC  
Antisense: GTGGCCACCGCGGCCCTCATAAACCCCTGTGCTTCC | C-terminus half of the protein |
Figure 1. Schematic genome organization of known circoviruses (left) and the novel RW-A circovirus (right) showing major open reading frames (ORFs) and a potential stem-loop structure (green feature). RW-A contains two major ORFs, the replication-associated (rep) gene (36% amino acid level identity to known Rep proteins from circoviruses) and an unknown ORF (no significant homologs in the database). This genome organization is consistent with known circovirus genomes including swine (e.g. Porcine circovirus, NC001792) and bird (e.g. Gull circovirus, NC008521) pathogens. The unknown ORF is believed to encode for the capsid protein due to its orientation with respect to the rep ORF, size, and N-terminus amino acid composition.
Figure 2. Western blot autoradiograph of 293T cells expressing the green fluorescent protein (GFP) and the putative capsid protein from the circovirus RW-A. The nuclear and cytosol fractions were separated. GFP (lanes G) is a cytoplasmic protein while the putative capsid (lanes CP) is expected to localize to the nucleus due to a nuclear localization signal at the N-terminus. The GFP signal is concentrated in the cytosol fraction (left) while the putative capsid protein is concentrated in the nuclear fraction (right), confirming that this protein localizes to the nucleus.

Figure 3. Immunofluorescence microscopy of 293T cells expressing the putative capsid protein from circovirus RW-A. The signal of the expressed capsid protein is concentrated in the nucleus of the cells suggesting the capsid proteins localize to the nucleus.
References:


APPENDIX F:

AUTHOR CONTRIBUTIONS AND COPYRIGHT CLEARANCES
Author Contributions

Appendix A: Metagenomic Sequencing for Virus Identification in a Public Health Setting

S. Svraka and K. Rosario designed and carried out experiments, analyzed data, and wrote manuscript

E. Duizer and H. van der Avoort assisted with research

M. Breitbart and M. Koopmans designed experiment and wrote manuscript

Appendix B: Metagenomic Analysis of Viruses in Reclaimed Water

K. Rosario designed and carried out experiments, analyzed data, and wrote manuscript

C. Nilsson and Y. W. Lim performed sequencing

Y. Ruan and M. Breitbart designed experiment and wrote manuscript

Appendix C: Pepper mild mottle virus as an Indicator of Fecal Pollution

K. Rosario and E. Symonds designed and carried out experiments, analyzed data, and wrote manuscript

C. Sinigalliano and J. Stewart carried out research

M. Breitbart designed experiment and wrote manuscript

Appendix D: Diverse Circovirus-like Genome Architectures Revealed by Environmental Metagenomics

K. Rosario designed and carried out experiments, analyzed data, and wrote manuscript

S. Duffy and M. Breitbart analyzed data and wrote paper

Note: All the publications included in this dissertation were included with the approval of each journal and necessary copyright clearances (see below).
Appendices A and D:

Dear Ms Rosario Cora,

Thank you for your email. SGM Copyright Permission is not required for a PhD thesis when the person requesting the permission is an author on the papers concerned. However, you are required to acknowledge the General Virology as the original source.

Paper No. 012955 (84143): Diverse circovirus-like genome architectures revealed by environmental metagenomics, Rosario et al.
Your paper was published by the Journal of General Virology in volume 90, part 10, pages 2418-2424.

Paper No. 024612 (84524): Metagenomic sequencing for virus identification in a public-health setting, Svraka et al.
Your paper will be published by the Journal of General Virology in volume 90, part 11, pages 2846-2856.

Good luck with your PhD.

Yours sincerely

Marianne K. Asbury
Editorial Assistant
Journal of General Virology Editorial Office
Direct Tel. No. +44 (0)118 988 1825
Fax No. +44 (0)118 988 1834
Web: http://www.sgm.ac.uk
On-line journals: http://www.sgmjournals.org

Society for General Microbiology
Marlborough House
Basingstoke Road
Spencers Wood
Reading RG7 1AG, UK

Registered Office as above. Registered as a Charity in England and Wales, No. 264017
A charity registered in Scotland, No SC039250
This is a License Agreement between Karyna Rosario ("You") and John Wiley and Sons ("John Wiley and Sons") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by John Wiley and Sons, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

<table>
<thead>
<tr>
<th>License Number</th>
<th>2518961399468</th>
</tr>
</thead>
<tbody>
<tr>
<td>License date</td>
<td>Sep 30, 2010</td>
</tr>
<tr>
<td>Licensed content publisher</td>
<td>John Wiley and Sons</td>
</tr>
<tr>
<td>Licensed content publication</td>
<td>Environmental Microbiology</td>
</tr>
<tr>
<td>Licensed content title</td>
<td>Metagenomic analysis of viruses in reclaimed water</td>
</tr>
<tr>
<td>Licensed content author</td>
<td>Karyna Rosario, Christina Nilsson, Yan Wei Lim, Yijun Ruan, Mya Breitbart</td>
</tr>
<tr>
<td>Licensed content date</td>
<td>Nov 1, 2009</td>
</tr>
<tr>
<td>Start page</td>
<td>2806</td>
</tr>
<tr>
<td>End page</td>
<td>2820</td>
</tr>
<tr>
<td>Type of use</td>
<td>Dissertation/Thesis</td>
</tr>
<tr>
<td>Requestor type</td>
<td>Author of this Wiley article</td>
</tr>
<tr>
<td>Format</td>
<td>Print and electronic</td>
</tr>
<tr>
<td>Portion</td>
<td>Full article</td>
</tr>
<tr>
<td>Will you be translating?</td>
<td>No</td>
</tr>
<tr>
<td>Order reference number</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0.00 USD</td>
</tr>
</tbody>
</table>
This is a License Agreement between Karyna Rosario ("You") and American Society for Microbiology ("American Society for Microbiology") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by American Society for Microbiology, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

<table>
<thead>
<tr>
<th>License Number</th>
<th>2518970901452</th>
</tr>
</thead>
<tbody>
<tr>
<td>License date</td>
<td>Sep 30, 2010</td>
</tr>
<tr>
<td>Licensed content publisher</td>
<td>American Society for Microbiology</td>
</tr>
<tr>
<td>Licensed content publication</td>
<td>Applied and Environmental Microbiology</td>
</tr>
<tr>
<td>Licensed content title</td>
<td>Pepper Mild Mottle Virus as an Indicator of Fecal Pollution</td>
</tr>
<tr>
<td>Licensed content author</td>
<td>Karyna Rosario, Mya Breitbart</td>
</tr>
<tr>
<td>Licensed content date</td>
<td>Nov 15, 2009</td>
</tr>
<tr>
<td>Volume</td>
<td>75</td>
</tr>
<tr>
<td>Issue</td>
<td>22</td>
</tr>
<tr>
<td>Start page</td>
<td>7261</td>
</tr>
<tr>
<td>End page</td>
<td>7267</td>
</tr>
<tr>
<td>Type of Use</td>
<td>Dissertation/Thesis</td>
</tr>
<tr>
<td>Format</td>
<td>Print and electronic</td>
</tr>
<tr>
<td>Portion</td>
<td>Full article</td>
</tr>
<tr>
<td>Order reference number</td>
<td></td>
</tr>
<tr>
<td>Title of your thesis / dissertation</td>
<td>Enhancing virus surveillance through metagenomics: water quality and public health applications</td>
</tr>
<tr>
<td>Expected completion date</td>
<td>Dec 2010</td>
</tr>
<tr>
<td>Estimated size(pages)</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>0.00 USD</td>
</tr>
</tbody>
</table>