Advancing Next Generation Genomic Tools and Big Data Science for Discovery of the WaterVirome

Joan Rose

Team members: Tiong Gim Aw, Yiseul Kim, Samantha Wengert
GLOBAL TRENDS IN THE ERA OF THE ANTHROPOCENE

• Urbanization
• Population Growth
• Regional Growth
• Travel and Tourism
• Global Corporate Growth
• Global Food Market
• Water Recycling, Reuse

• HOW IS WATER QUALITY CHANGING?
• HOW DOES THIS EFFECT HUMAN HEALTH?
• HOW CAN NEW TECHNOLOGY HELP US WITH WATER EXPLORATION & DISCOVERY?
COASTAL SYSTEMS AND FRESH WATER RESOURCES ARE DEGRADING Ecosystems

Recreational

Irrigation

Drinking

Algal blooms

Ecosystems

In waters used for drinking, fishing, recreation
World has 356,000 km of coastline (221,208 miles) US has 19,924 km of coastline Great Lakes 15,043 km of coastline. 44% of the global population (7 billion people) lives within 150 km (93 miles) of the coastline (that is 3 billion people who flush or dispose daily and send fecal pollution into the environment and eventually into waterways). The world's rivers (ten of the longest rivers = 55,734 km or 34,629 miles) are so badly affected by human activity that the water security of 5 billion people are impacted.
THE PROBLEM: THE GREAT ACCELERATION

GLOBAL POPULATION TRENDS 1800S TO 2100


Global Water Consumption 1900 - 2025
(by region, in billion m³ per year)

http://esa.un.org/wpp/unpp/panel_population.htm
Data from FAO (2010).
Figure I. Cultivated areas of the world. Brown regions indicate areas in which at least 30% of the landscape is cultivated. Reproduced from the Millennium Ecosystem Assessment 2005 (http://www.MAweb.org), UNEP.

LOSS OF WETLANDS
KM²
From 1950s to 1990s in the US.

Verhoeven et al. TRENDS in Ecology and Evolution Vol.21 No.2
February 2006
Figure 1.3. Total fertiliser application of N, P and K in industrialised and developing countries plus China and Africa, developing from 1961 to 2002 (kg/ha) (FAOSTAT data, 2005)

Figure 1.4. Imports and exports value of global pesticide sales from 1961 to 2003 (FAOSTAT data, 2005)
Economics of Pollution

Health, Social & Economic impacts

- Morbidity and Mortality
  Health care costs
- Outbreaks/disaster costs
- Productivity loss
- Decreases in educated workforce (girls)
- Quality of life low
- Transboundary blame

Tourism

- Total Tourism Foreign Earnings ($B) US$ 856
- 9.7% of world GDP
- Worlds largest and fastest growing industry
- Tourism contributes to 70% GDP for some of the world’s poorest countries.
- ~$280,000 loss/d/beach closure

Ref: PATA Tourism Forecast / WTO
There are 16,000 publicly owned wastewater treatment plants, 100,000 major pumping stations, 600,000 miles of sanitary sewers, and 200,000 miles of storm sewers in the US.

**Wastewater Grades**

<table>
<thead>
<tr>
<th>Year</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988</td>
<td>C</td>
</tr>
<tr>
<td>1998</td>
<td>D+</td>
</tr>
<tr>
<td>2012</td>
<td>D</td>
</tr>
</tbody>
</table>

**INFRASTRUCTURE IS NOT KEEPING UP**

Discharging billions of gallons of untreated sewage into U.S. surface waters each year. The EPA estimates that the nation must invest $390 billion over the next 20 years to replace existing systems and build new ones to meet increasing demands.

ASCE
The Blue Economy*

• Desalination - $16.6 Billion
• Wastewater Resource Recovery - $37 Billion
• Wastewater Treatment - $70.8 Billion

Overall Market (products/services) - $770B
Growth – As high as 10% Annually!!

* Pittsburgh’s H2Opportunity Report
Water is at the core of the global goals for ONE HEALTH
Fecal contamination of water remains one of the largest threats to the biological safety of water today.
Outbreak of Polio Spreading in Syria
Oct 26th, 2013

In Syria, twenty-two people or more, mostly babies and toddlers, are said to have contracted polio, announced the World Health Organization. Doctors in Syria have also reported a rise in diseases such as typhoid, hepatitis, and the flesh-eating parasite leishmaniasis.

Polio-like illness found in five California children
By Jacque Wilson, CNN
updated 7:11 AM EST, Tue February 25, 2014

Between 1983 and 2005, 270 cases of enterovirus 71 were reported in the United States. Current virus identified as enterovirus 68.

About 83% diseases in Mumbai are water borne:
PTI | Aug 19, 2013, 09.18 PM IST

Hand, foot, mouth disease patients top 130,000 in Japan JIJI, Aug 7, 2013

Wild type Polio virus circulating in sewage in Israel.

Between 1983 and 2005, 270 cases of enterovirus 71 were reported in the United States. Current virus identified as enterovirus 68.
Environmental Virology:

Public Health-Related Viruses
• adenovirus
• coxsackievirus
• Echovirus
• enteroviruses
• Hepatitis A and E
• Norovirus
• poliovirus
• rotavirus

Bacteriophage as indicators
eg. coliphage

Phage Ecology
Detection of viruses – cell culture assay

Solid monolayers of cells – plaque assay

Liquid suspensions – cytopathic effect

Noninfected BGM cells

BGM cells infected with poliovirus
Molecular methods for virus detection & discovery

Molecular methods

Sequence dependent
- PCR (consensus primers)
- Microarray (probe hybridization)

Sequence independent
- Whole genome amplification
- Random PCR
- Subtractive hybridization
- Metagenomic sequencing
Number of prokaryotic cells on earth $\sim 4$ to $6 \times 10^{30}$
Viruses in ocean $\sim 10^{30}$ (Suttle, 2007)

99% of all microbes remain, as yet, uncultured
(Amann et al. 1990)
The Genomic-based tools revolution

**1914**
- The first US Public Health Service Drinking Water Standard adopted a bacteriological standard

**1953**
- Discovery of DNA helix by James Watson & Francis Crick

**1977**
- Sanger sequencing

**1983**
- DNA microarray

**1985**
- Human Genome Project

**1990**
- Human Genome Project - completed

**Mid 1990s**
- Real-time PCR; Pyrosequencing method published

**2003**
- 454 Life Sciences markets a version of pyrosequencing

**2004**
- Illumina next generation sequencing is on the market

**2006**
- Polymerase chain reaction (PCR)

**2010**
- Ion semiconductor sequencing

**2011**
- Pacific Biosciences commercialized single molecule real-time (SMRT) sequencing
THE MICROBIOME

sources

microbiota

Community DNA

Metagenome library & sequence analysis
Viral metagenomics

Next generation sequencing for the water: using metagenomic approach to explore new frontiers in water quality testing

**Virome**: Genomes of all the viruses that inhabit a particular organism (animal, plant or microorganism) or a specific environmental niche.
Why study water/wastewater viromes?

• Better assessment and finer resolution of the waterbiome systems due to the natural host specificity of viruses “What is healthy?”
  • Identify the multitude of new targets in various water environments address characterization and risks

• Support current water quality monitoring practices and ultimately transform how we test for the quality of our water

• Wastewater has been shown to harbour a great numbers of viruses with a wide range of genome sizes
  • Phage-bacteria interaction to improve the efficiency of the wastewater treatment process (e.g., foam and pathogen control)
Objectives of the MSU team

• Multidisciplinary team (Public health microbiologists, virologists, bioinformaticians & environmental engineers)

(1) to use metagenomics with Illumina sequencing to generate a viral biome view of various environmental and food samples:
   - Wastewaters from full scale treatment plants
   - Ballast water system
   - Fresh produce and irrigation water POSTER by Samantha Wengert (Tues)

(2) to use bioinformatics to analyze the metagenomic fingerprints to assess the viral ecology and their diversity

The new science of viral metagenomics will be used to discover potential novel viral indicators or viral genetic markers to better inform on what are healthy waters & food systems.
Methods for viral metagenomics

Experimental methods

Sample

Viral particles concentration & purification

DNA/RNA extraction & amplification

DNA sequencing

Quality control: 16S rRNA gene & 18S rRNA gene PCR

Preservation

- Ultrafiltration,
- precipitation with polyethylene glycol (PEG)
- Passage through 0.45μm and 0.22μm diameter filters
- Nuclease treatment

- Reverse transcription (for RNA viruses)
- Random amplification

- Illumina sequencing

Computational methods

Classification of reads by barcodes

Quality control:
- Filter repeats and sequencing errors
- Screen against contaminating host DNA sequences

Assembly
- Discovery & annotation of viral contigs & genomes

Similarity-dependent analysis
- Query public and custom databases (NCBI viral RefSeq, ACLAME, Phage SEED)

Similarity-independent analysis
- Nucleotide frequencies
- Sequence clustering
Low-cost ultrafiltration system with disposable hollow fiber filters

- Size-exclusion mechanism
- A wide range of viruses can be recovered in a single water sample
# Recovery efficiency of hollow fiber ultrafiltration for bacteriophages seeded into 20 liters of water samples

<table>
<thead>
<tr>
<th>Phage</th>
<th>Groundwater (20 L)</th>
<th>Surface water (20 L)</th>
<th>Tap water (20 L)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No elution</td>
<td>With elution</td>
<td>No elution</td>
</tr>
<tr>
<td>MS2</td>
<td>3</td>
<td>74.9±5.0</td>
<td>109.4±2.4</td>
</tr>
<tr>
<td>PhiX174</td>
<td>3</td>
<td>75.9±2.6</td>
<td>94.4±13.2</td>
</tr>
<tr>
<td>P22</td>
<td>3</td>
<td>77.8±5.1</td>
<td>76.8±12.6</td>
</tr>
</tbody>
</table>
Sampling of viromes in full scale water reclamation plant

Microfiltration effluent (100L)

Microfiltration feed (20L)

Influent of OCWD plant – Activated sludge and trickling filter effluent (20L)
Next-generation sequencing platform

Illumina HiSeq 2500

The Illumina technology immobilizes random DNA fragments on a surface and then performs solid-surface PCR amplification, resulting in clusters of identical DNA fragments. These are then sequenced in a **sequencing-by-synthesis process**. Typically read lengths of 100 bases are possible and **hundreds of millions of bases are typically generated per sequencing run**.
**Bioinformatics approach**

- **Sequencing reads** – filtering, quality control
- **Read mapping** to viral reference genome (Bowtie 2)
- **De novo** assembly (Velvet)
- **Annotation** (BLAST)
- **Taxonomic and functional classification** (MEGAN)

*De novo* sequence assembly is the process whereby we merge together individual sequence reads to form long contiguous sequences sharing the same nucleotide sequence as the original template DNA from which the sequence reads were derived.

### Assembly and sequencing statistics for sewage virome

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total no. of clean read</td>
<td>17,922,804</td>
</tr>
<tr>
<td>Total no. of contigs (&gt; 200 bp)</td>
<td>31,915</td>
</tr>
<tr>
<td>Total assembly length (bp)</td>
<td>11,409,793</td>
</tr>
<tr>
<td>% Reads assembled</td>
<td>36%</td>
</tr>
<tr>
<td>Average contig length (bp)</td>
<td>750.5</td>
</tr>
<tr>
<td>Longest contig length (bp)</td>
<td>4,312</td>
</tr>
</tbody>
</table>
Distribution of sequences for sewage virome based on tBLASTx analysis against NCBI Viral Sequence database

- **No hits, 24739, 77.5%**
- **Assigned viruses, 3571, 11.2%**
  - ssDNA viruses, 1350, 4.5%
  - ssRNA viruses, 42, 0.1%
  - unclassified viruses, 680, 2.1%
  - satellites, 62, 0.2%
  - dsDNA viruses, 1437, 4.5%
- **Not assigned, 3605, 11.3%**

**Human viruses, 76, 3.4%**

- **Animal viruses, 436, 19.4%**
- **Plant viruses, 225, 10.0%**
- **Bacteriophages, 1507, 67.2%**

- **Unclassified phages, 101, 4.5%**
  - **Corticoviridae, 1, 0.0%**
  - **Inoviridae, 10, 0.4%**
  - **Microviridae, 293, 13.1%**

**Caudovirales, 1102, 49.1%**
Diversity of the sewage virome

**Human Viruses identified:**

*Adenoviridae*
- Human adenovirus B
- Human adenovirus C
- Human adenovirus F

*Polyomaviridae*
- JC polyomavirus
- BK polyomavirus
- Human enterovirus B
- Human papillomavirus
Comparison of virome of different wastewater samples using principal coordinate analysis based on Bray-Curtis similarity index.
Diversity of *Tobamovirus* in untreated and treated wastewater

Protein sequence alignment of pepper mild mottle virus detected in microfiltration effluent
Virus host

<table>
<thead>
<tr>
<th>Virus host</th>
<th>Percent</th>
<th>Types of viral homologs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humans</td>
<td>1%</td>
<td><em>Adenoviridae, Picornaviridae</em>, (Rhinoviruses and enteric viruses)</td>
</tr>
<tr>
<td>Animals</td>
<td>15%</td>
<td><em>Circoviridae</em> (Avian and swine circoviruses)</td>
</tr>
<tr>
<td>Plants</td>
<td>6%</td>
<td><em>Geminiviridae, Nanoviridae</em> (Grain, Fruit, and vegetable viruses)</td>
</tr>
<tr>
<td>Bacteria</td>
<td>48%</td>
<td><em>Inoviridae, Microviridae, Myoviridae, Podoviridae, Siphoviridae</em></td>
</tr>
</tbody>
</table>

Bacteriophage shown to move antibiotic resistant genes
An Emerging Waterborne Virus: Found as Part of the sewage viral biome.

**Circoviridae**

- A virus family that comprises two genera, Circovirus that includes porcine circoviruses, pigeon circovirus, and psittacine beak and feather disease virus, and Gyrovirus that includes chicken anemia virus.

- They are the smallest animal viruses, 17 to 24 nm in diameter, contain a single-stranded circular DNA genome composed of about 2500 nucleotides and replicate in the nucleus of cells and are assumed to be dependent on the host cell for many functions required for viral replication and probably, like paroviruses, replicate in cells that are in the S-phase of the cell cycle.

- A cyclovirus - has been isolated from the cerebrospinal fluid of 25 Vietnamese patients with CNS infections of unknown aetiology (2013)

  The same virus has been isolated from the faeces of healthy children and also from pigs and chickens. This suggests a oral faecal route of transmission with a possible animal reservoir.

Viruses in ballast water

- **Ballast water**: Used to control and maintain ship’s stability, trim, draft, and stress.
- Taken on board when a ship is traveling without cargo.
- Discharged *typically containing a variety of biological materials* when loading new cargo on board.
- Abundance of viruses in ballast water based on the estimation of the number of virus-like particles (VLPs)
  - $7.4 \times 10^9$ VLPs per liter of ballast water (Ruiz et al., 2000)
  - $2.6 \times 10^{17}$ VLPs per ship (Drake et al., 2007)
- Diversity and composition of viruses in ballast water have not been studied.

Image from www.globallast.imo.org
Sampling location and source of ballast water
All samples were collected around the Port of Duluth-Superior. The green and red boxes represent the source of port and ballast waters, respectively. BW; ballast water.
Taxonomic (A) and functional (B) profiling of ballast and harbor water viromes

(A) Taxonomic profiling:
- Caudovirales: 57.5%
- Unidentified: 24%
- Other: 15.9%

(B) Functional profiling:
- Phages, prophages, transposable elements: 44.7%
- Unidentified: 48.4%
- Other groups: 6.9%

Ballast water:
- Caudovirales: 65.5%
- Microviridae: 4.8%
- Phycodnaviridae: 1.9%
- Other: 1.9%
- Unclassified: 21.8%

Harbor water:
- Caudovirales: 57.5%
- ssRNA positive-strand viruses: 6.1%
- Microviridae: 5.8%
- Phycodnaviridae: 1.3%
- Circoviridae: 1.1%
- Other: 1.7%
- Unclassified: 24%

(A) Taxonomic profiling:
- DNA metabolism: 3.6%
- Virulence: 1.6%
- Regulation and cell signaling: 1%
- Other: 0.7%

(B) Functional profiling:
- DNA metabolism: 4.6%
- Virulence: 1.8%
- Regulation and cell signaling: 1.1%
- Other: 1%
Global comparison of viromes from different aquatic environments based on BLAST.

Comparison of ballast and harbor water viromes using principal coordinates analysis based on Bray-Curtis similarity index.

Hypersaline and French lakes viromes were from the studies by Rodriguez-Brito et al. (2010) and Roux et al. (2012), respectively.
Technical challenges in viral metagenomics

- Experimental and sampling issues: the number of samples needed, the importance of spatial and temporal aspects

- New and better tools for the recovery and purification viruses from small amounts of starting microbial community biomass, methods for less biased amplification of extracted nucleic acids before sequencing.

- Computational methods and related visualization tools
  - Development of a universal pipeline for viral metagenomic analysis, statistical analysis, data management and storage
  - How do we address “unknown” sequences?
Research Needs

1. Characterization of viruses in the water environ in various types of full scale systems of wastewater and treated waters (including after advanced treatment and in ambient waters receiving discharges);

2. New viral genome data added to the database; improve the bioinformatics approaches for assessing the viral genome;

3. Experimental approaches using Illumina sequencing for viral metagenomics in water (technology transfer workshop)

4. New qPCR assays capable of efficiently targeting viruses of interest from different sources.
Acknowledgements

Funding
- National Science Foundation PIRE, USDA NIFA

Ballast water sampling
- Cordell Manz (Lake Superior Ballast Water Inspector at Wisconsin DNR)
- Chris Scianni & Ballast Water Inspectors (California State Lands Commission)

Wastewater sampling
- East Lansing Wastewater Treatment Plant
- Orange County Water District (OCWD)

Yuma irrigation water and fresh produce sampling
- Channah Rock, Kurt Nolte (The University of Arizona)

MSU Research Technology Support Facility
MSU High Performance Computing Center
Thank You!

Any Questions??