Environmental Metagenomics: Developing guidance for environmental labs interested in NextGen sequencing technology

APHL Annual Meeting

June 7, 2016

National Center for Emerging and Zoonotic Infectious Diseases

Division of Foodborne, Waterborne, and Environmental Diseases

Goals for Environmental Microbiology Testing

- Characterize environmental exposure risks
- Identify disease and outbreak sources
- Understand transmission dynamics
- Identify control and mitigation strategies

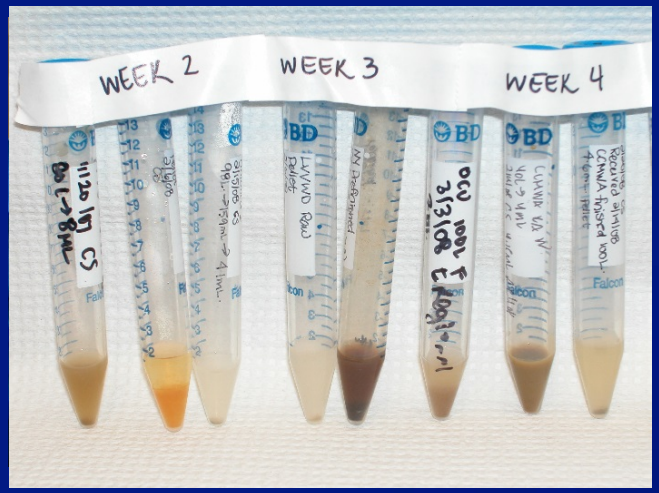
FINDING THE NEEDLE IN THE HAYSTACK: ISSUES & CHALLENGES FOR MOLECULAR TESTING OF ENVIRONMENTAL SAMPLES

Issues & Challenges for Environmental Testing

- **Fecal indicators (e.g., *E. coli*) present in contaminated samples at relatively high levels**
  - Small sample volumes typically sufficient for detection
- **Enteric pathogens present at much lower levels**
  - On population basis, most people shed fecal indicators but relatively few (sick or recently infected) shed pathogens
  - Die-off outside of human/animal host
  - Dilution and transport in the environment
- **Environmental pathogens also often at low levels**
  - Presence and concentration vary according to environmental factors
  - Biofilm-associated or particle-associated

Specialized Methods Needed

- **Sampling**
  - Large-volume samples to improve detection chances
  - Storage and transport
- **Sample preparation**
  - Additional sample concentration
  - Separation of target microbes/nucleic acid from background constituents
- **Amplification/enrichment**
  - Increase signal:noise ratio
- **Quantification**
- **Viability/infectivity**



NextGen Sequencing: Power & Potential for Environmental Testing

- Pathogen detection and identification
- Molecular epidemiology, linking sources to cases
- Understanding evolution of environmentally transmitted pathogens
- Identifying virulence factors
- Detecting antibiotic resistance (AR) and functional genes in samples and microbial isolates
- Understand environmental microbiome dynamics in relation to environmental factors and water system operation & maintenance

Whole Genome Sequencing of Bacterial Isolates from Environmental Samples

genomeA
Journals.ASM.org

Draft Genome Sequence of *Buttiauxella agrestis*, Isolated from Surface Water

Narayanan Jothikumar,[a] Amy Kahler,[a] Nancy Strockbine,[a] Lori Gladney,[a,b] Vincent R. Hill[a]
National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA[a]; IHRC, Inc., Atlanta, Georgia, USA[b]

MI agar is routinely used for quantifying *Escherichia coli* in drinking water. A suspect *E. coli* colony isolated from a water sample was identified as *Buttiauxella agrestis*. The whole genome sequence of *B. agrestis* was determined to understand the genetic basis for its phenotypic resemblance to *E. coli* on MI agar.

Received 8 September 2014 **Accepted** 10 September 2014 **Published** 16 Oc
Citation Jothikumar N, Kahler A, Strockbine N, Gladney L, Hill VR. 2014. Draft ge
e01060-14. doi:10.1128/genomeA.01060-14.

Address correspondence to Narayanan Jothikumar, jin2@cdc.gov.

genomeA
Journals.ASM.org

Draft Genome Sequence of *Raoultella planticola*, Isolated from River Water

Narayanan Jothikumar,[a] Amy Kahler,[a] Nancy Strockbine,[a] Lori Gladney,[a,b] Vincent R. Hill[a]
National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA[a]; IHRC, Inc., Atlanta, Georgia, USA[b]

We isolated *Raoultella planticola* from a river water sample, which was phenotypically indistinguishable from *Escherichia coli* on MI agar. The genome sequence of *R. planticola* was determined to gain information about its metabolic functions contributing to its false positive appearance of *E. coli* on MI agar. We report the first whole genome sequence of *Raoultella planticola*.

Received 8 September 2014 **Accepted** 10 September 2014 **Published** 16 October 2014
Citation Jothikumar N, Kahler A, Strockbine N, Gladney L, Hill VR. 2014. Draft genome sequence of *Raoultella planticola*, isolated from river water. Genome Announc. 2(5):
e01061-14. doi:10.1128/genomeA.01061-14.

Address correspondence to Narayanan Jothikumar, jin2@cdc.gov.

ENVIRONMENTAL TESTING METHOD DEVELOPMENT

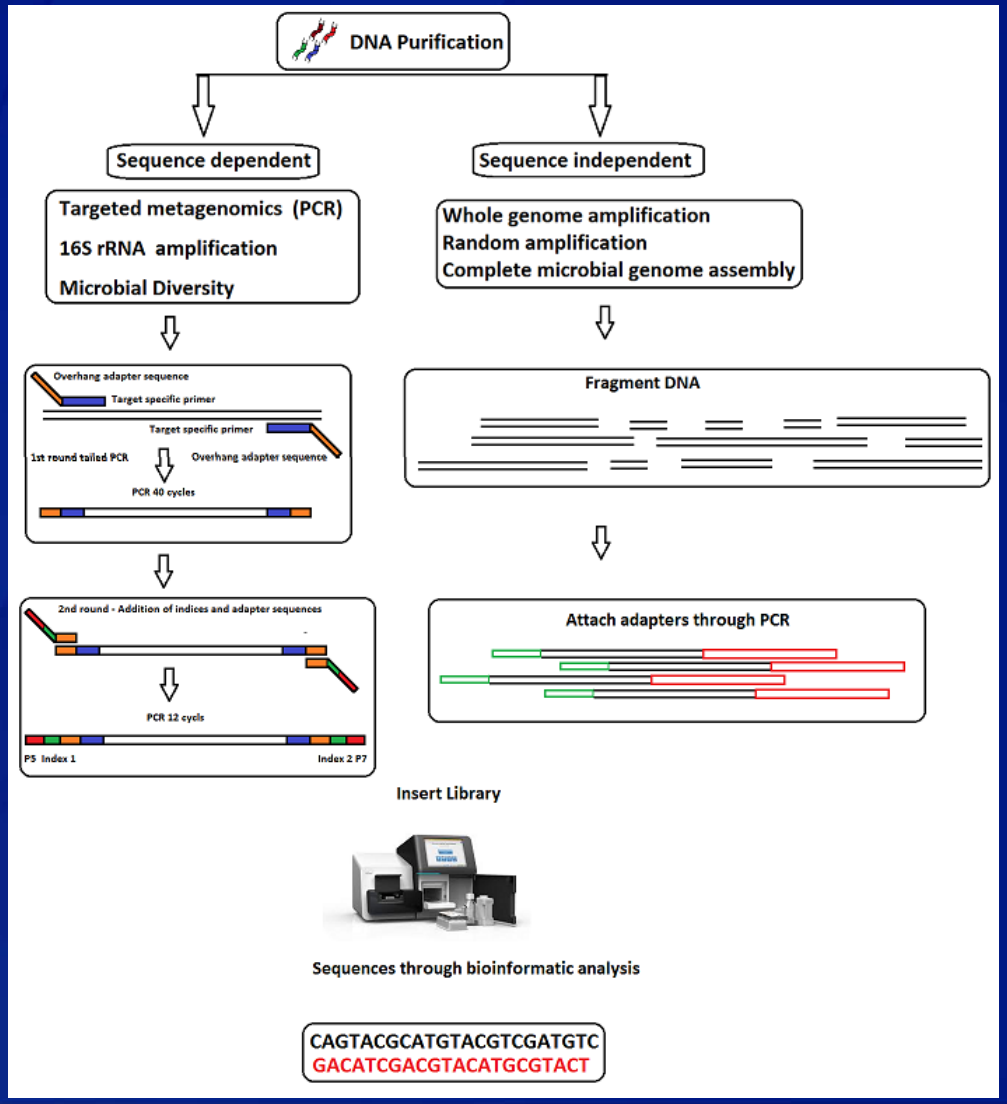
NextGen Sequencing for Environmental Testing

- **Implementing targeted gene sequencing (16S, 18S)**
  - "Microbiome" community analysis
  - Conserved gene sequences enable community characterization, but low resolution often limits differentiation to genus level
  - Bioinformatics methods simpler and more established
- **Shotgun metagenomics**
  - Sequences all nucleic acid in a sample (microbial, human, animal, plant)
  - Enables species identification, virulence, AR, & functional genes
  - Signal:noise ratio a major challenge—necessitates whole genome amplification to increase DNA for sequencing, but amplification introduces bias
  - Clutter mitigation and bioinformatics challenging

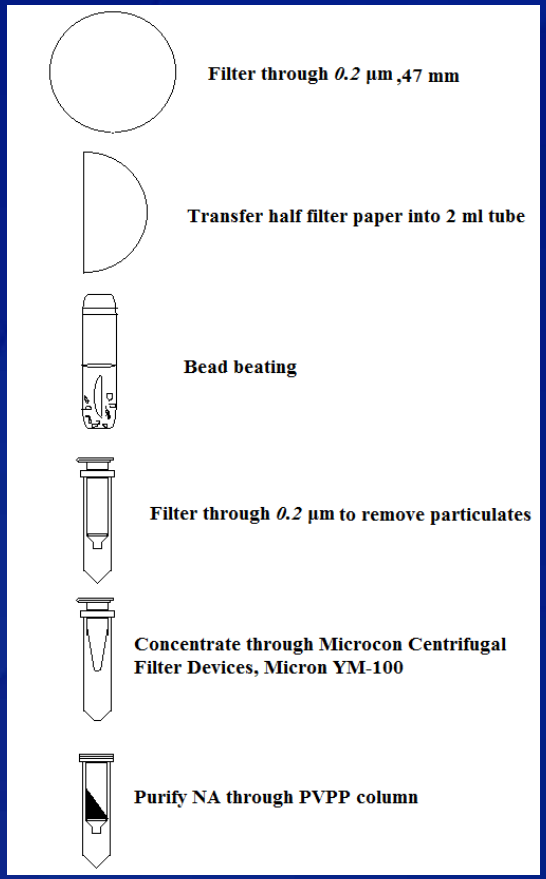
Sample Preparation for NextGen Sequencing of Environmental Samples

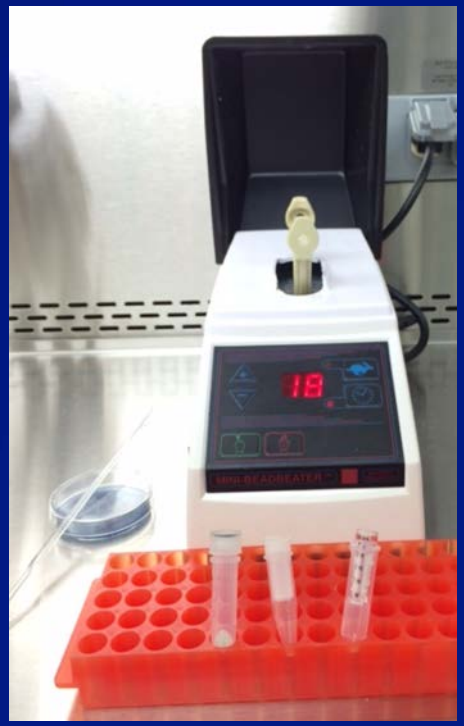
1) Immunomagnetic separation (if available)
2) Microbial lysis (bead beating and enzymatic)
3) Nucleic acid separation and purification
4) DNA quantification and quality assessment
5) Enrichment
6) Fragmentation (for shotgun metagenomics)
7) Library preparation
8) Sequencing
9) Analysis (sequence assembly, QA, comparison to databases [pipelines])

Sequence-dependent vs. Sequence-Independent Workflows



Nucleic acid extraction procedure for 16S NextGen sequencing of drinking water samples

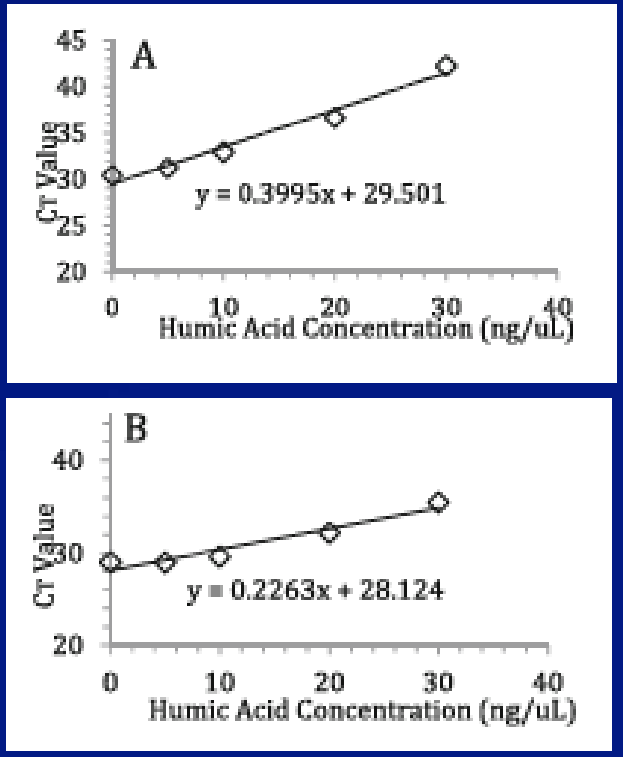




Goal: sample concentration (~500 mL reduced to 100 µL) while optimizing yield and purity

Remove Inhibitors to Maximize DNA Amplification Efficiency

- Organic compounds (e.g., humic and fulvic acids) interfere with DNA polymerase enzymes
- Remove using silica bead/column separation, microconcentrators, PVPP column, resins (e.g., Sephadex)



Streby et al (2015)

Streby A, Mull BJ, Levy K, and Hill VR. (2015) Comparison of real-time PCR methods for the detection of *Naegleria fowleri* in surface water and sediment. *Parasitol Res.*, **114**:1739-1746.

Quantify DNA Before Sequencing

Real-Time PCR prior to 16S/18S Sequencing

			16S metagenomics	
			1st round	1st round
			V1-V3	V4
1	PCUO20116-E1		29.4	23.9
2	PCUO20116-E2		26.6	21.2
3	PCUO20116-E3		35.1	27.7
4	PCUO20116-U1		36.2	28
5	PCUO20116-U2		32.9	24.8
6	PCUO20116-U3		36.7	27.8
7	PCU091615-E1		30.5	23.1
8	PCU091615-E2		23.8	18.2
9	PCU091615-E3		36	30.8
10	PCU091615-U1		25.2	18.6
11	PCU091615-U2		36.7	25.8
12	PCU091615-U3		25.4	19.6
13	OCWO10716-E1		25.9	20.4
14	OCWO10716-E2		26	22
15	OCWO10716-E3		29.1	23.3
16	OCWO10716-U1		30	23.8
17	OCWO10716-U2		25.5	20
18	OCWO10716-U3		26.2	20.8

DNA size characterization using TapeStation

DNA quantity using fluorometer

Enrichment Needed for Environmental Samples

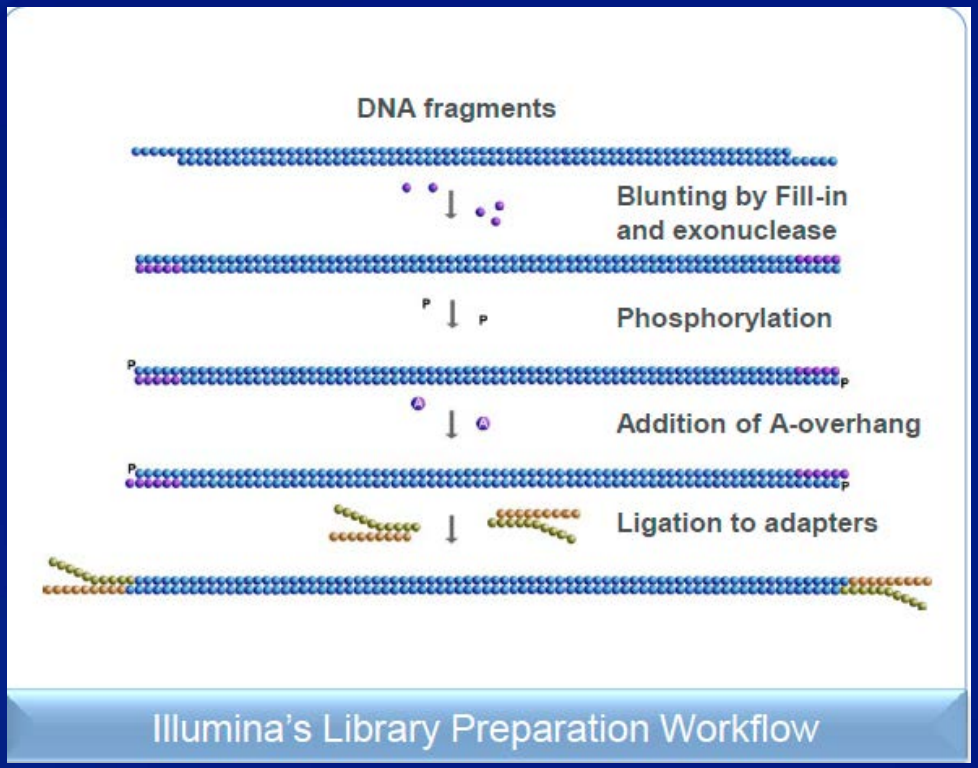
- **Targeted PCR amplification**
  - Advantages: PCR-based amplicon sequencing (e.g. 16S) is inexpensive (less prep work, and more samples can be pooled per run (sequencing short regions)
  - Disadvantages: can only amplify 200 - 2000 bp; less information for microbe ID (often limited to genus level ID)
- **Whole genome amplification**
  - Advantages: 1) various amounts of genomic DNA (0.3–300 ng) can be amplified, 2) representative DNA amplification with minimal risk of locus dropout, 3) produces long stretches of intact DNA, increases downstream sequence quality
  - Disadvantages: 1) may cause preferential amplification of DNA sequences (bias), 2) requires average genomic DNA fragment sizes of approximately 2 kb in order to amplify DNA, 3) relatively high cost

Library Preparation: Step 1 (Fragmentation)

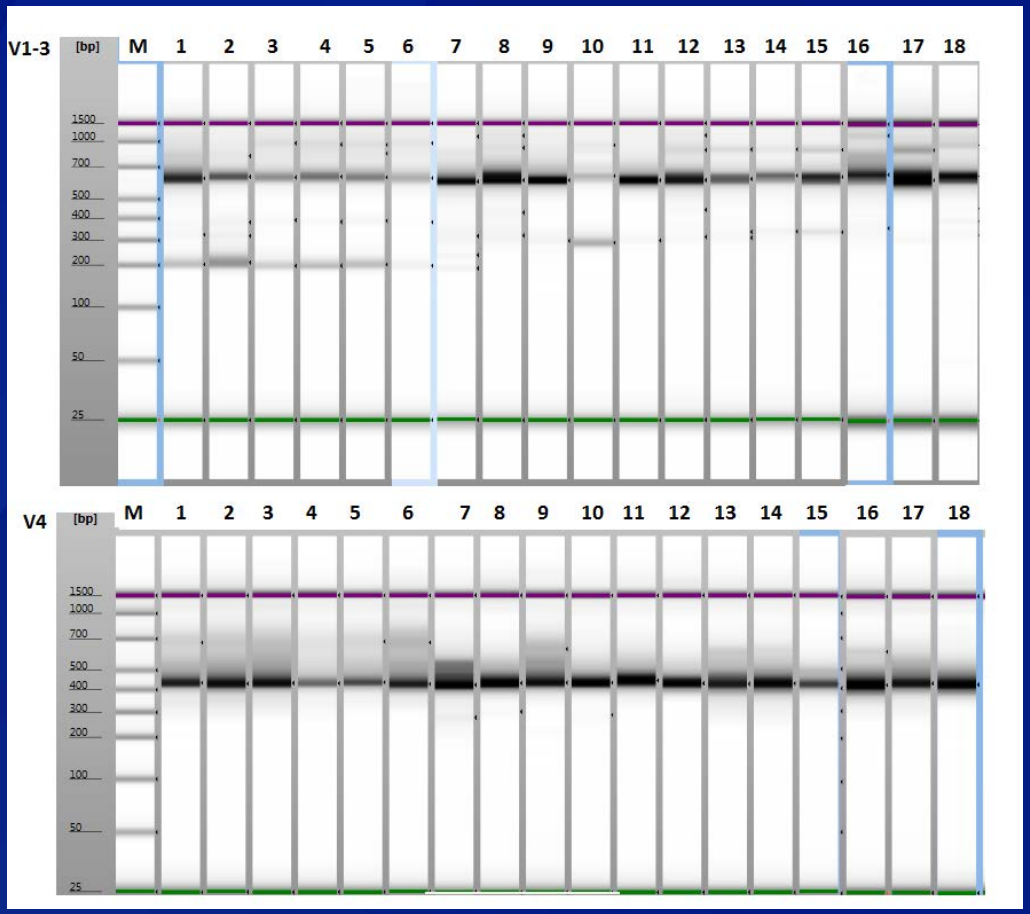
- DNA fragmented to generate double stranded DNA to attach adapters
- Fragmenting done using an ultrasonicator instrument or by enzyme-based fragmenting DNA kit

Library Preparation: Step 2 (Fragment-end repair)

- Adapter sequences are ligated to the ends of the fragmented DNA
- The library is amplified (2nd rnd PCR) in sufficient quantity for DNA sequencing
- Use TapeStation for fragment sizing



TapeStation sizing of 16S amplicons from 18 drinking water samples



Automated Library Prep and NextGen Sequencing

NeoPrep System Library Preparation

Starting with DNA or RNA, the NeoPrep System performs all library preparation steps, including PCR amplification, quantification, and normalization in a single, fully integrated instrument.

MiSeq Benchtop sequencer

Metagenomics, small genome sequencing, amplicon sequencing starting at 10 ng DNA, and HLA typing. Enable up to 15 Gb of output with 25 M sequencing reads and 2x300 bp read lengths.

Application of 16S Amplicon Sequencing

- Epidemiological study of potential health effects from low pressure events in drinking water distribution systems
- Chlorinated and chloraminated systems
- Compared main break areas vs control areas

Chloraminated system (Winter 2016)

Top Species Classification Results

Classification	Number of Reads	% Total Reads
Unclassified at Species level	376,049	75.11 %
Methylobacterium phyllosphaerae	16,320	3.26 %
Mycobacterium neworleansense	12,187	2.43 %
Mycobacterium vanbaalenii	11,196	2.24 %
Mycobacterium anthracenicum	9,892	1.98 %
Bradyrhizobium pachyrhizi	9,586	1.91 %
Methylomonas methanica	5,703	1.14 %
Methylomonas scandinavica	4,092	0.82 %

Total Species-level Taxonomic Categories Identified: 644. This table shows the top 8 of 644 classifications.

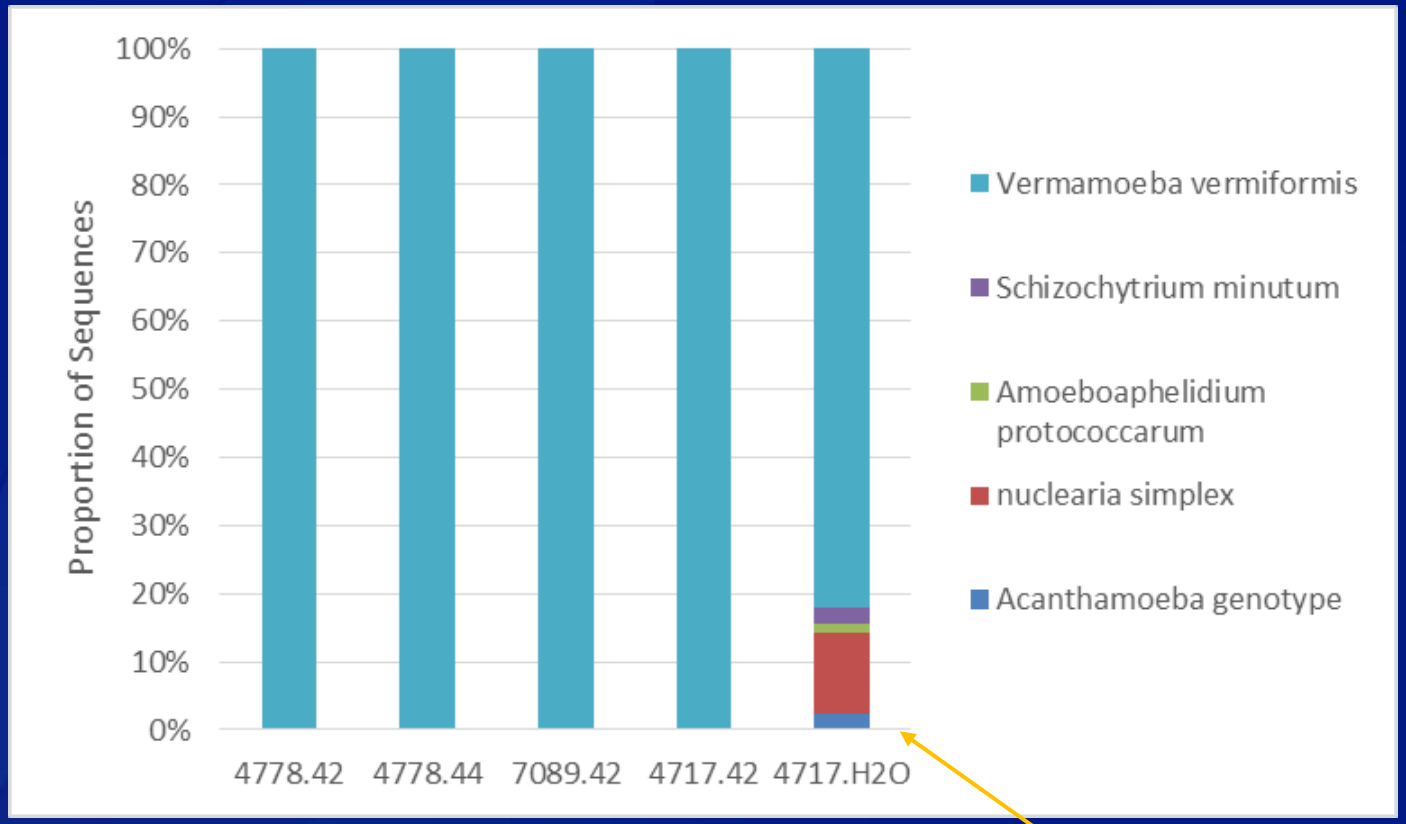
Chlorinated system (Winter 2016)

Top Species Classification Results

Classification	Number of Reads	% Total Reads
Nevskia ramosa	272,580	44.69 %
Unclassified at Species level	130,631	21.42 %
Sphingomonas sanxanigenens	57,091	9.36 %
Sphingomonas asaccharolytica	47,205	7.74 %
Sphingomonas oligophenolica	32,820	5.38 %
Xanthobacter polyaromaticivorans	31,563	5.17 %
Bradyrhizobium canariense	15,345	2.52 %
Sphingomonas elodea	4,531	0.74 %

Total Species-level Taxonomic Categories Identified: 239. This table shows the top 8 of 239 classifications.

18S Metagenomics for Hot Water Heater Samples, Louisiana, 2014



Bioinformatics

- Developing procedures for environmental sampling testing
- Need powerful computing capacity, dedicated specialists
  - Tailor software to filter out unwanted sequences, assemble sequences into contigs, annotate contigs for gene ID, bin sequences to identify organisms (assign contigs to Operational Taxonomic Units (OTUs—order, family, genus, species)
- Databases for metagenomics analysis
  - Illumina (automated 16S analysis not requiring bioinformatics)
  - GenBank (NCBI)
  - Greengenes, Mothur, SILVA: Ribosomal RNA sequence database
  - EUPathDB: Eukaryotic pathogens database
  - Metavir: a web server dedicated to virome analysis

QA/QC Considerations

- **Quality Control**
  - Positive control: Custom or commercial microbial consortium providing expected sequences
  - Minimizing carry-over contamination: instrument cleaning/maintenance (e.g., bleach wash), use alternate sequence adapters between runs
- **Quality Assurance**
  - DNA quality (purity [e.g., absorbance 260/280], fragment size)
  - DNA amount (critical, depends on biomass in the starting sample, appropriate enrichment conditions)
  - Sequence quality (quality score output from instrument software)
    - Low score contributors: background nucleic acid contamination, sample cleanup issues, too little (or too much) DNA, improper DNA fragmentation

What's it cost?

- **Equipment**
  - Extraction ($8K), Library preparation ($26K for ultrasonicator; $30K for automated instrument [optional]), DNA quantity/quality ($15K), sizing ($45K for tapestation)
  - Sequencing ($50K - $100K - >$250-500K)
  - Instrument maintenance costs add up
- **Consumables**
  - Whole genome amplification kits
  - Library preparation kits: very expensive: ~$2500, $50-100/sample
  - Sequencing reagents: prices going down
- **Staffing**
  - Dedicated technicians for sample prep
  - Bioinformaticians for sequence assembly and analysis

Goals and Next Steps

- Establish SOPs for environmental sample preparation and workflow for gene amplicon sequencing and shotgun metagenomics
  - Applying 16S and 18S amplicon sequencing techniques to large-volume drinking water samples
  - Developing shotgun metagenomics SOPs for water samples to detect hepatitis A virus and hepatitis E virus
- Work with partners (APHL, EPA) to transfer NextGen sequencing techniques to public health labs
- Develop workgroup with partners to facilitate communication & collaboration, process of continual improvement

Acknowledgements

CDC Office of Public Health Preparedness and Response

CDC Office of Advance Molecular Detection

Dr. Jothikumar Narayanan, CDC

Questions?

For more information please contact:

Vincent Hill (vhill@cdc.gov)

Jothikumar Narayanan (jin2@cdc.gov)

1600 Clifton Road NE, Atlanta, GA 30333

Telephone: 1-800-CDC-INFO (232-4636)/TTY: 1-888-232-6348

E-mail: cdcinfo@cdc.gov Web: http://www.cdc.gov