Developing a Quantitative Metagenomic Approach to Explore Viral Community Dynamics Through Wastewater Treatment

by

Kathryn L. Langenfeld

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Environmental Engineering)
in the University of Michigan
2022

Doctoral Committee:

Assistant Professor Melissa Duhaime, Co-Chair
Associate Professor Krista Rule Wigginton, Co-Chair
Professor Lutgarde Raskin
Assistant Professor Luis Zaman

Kathryn L. Langenfeld

klangenf@umich.edu

ORCID iD: 0000-0002-2741-2254

Dedication

To my two favorite beings, Ian and Ivy.

Acknowledgements

This dissertation was accomplished with the aid of several funding sources, including the NSF PIRE Halting Environmental Antimicrobial Resistance Dissemination (Project No. 1545756), USDA Antibiotic Resistance Project (Project No. 2016-68003-24601), University of Michigan Integrated Training in Microbial Systems (ITiMS) mini-grant, and Rackham research grants. My training was supported by multiple fellowships that encouraged me to pursue interdisciplinary research projects and pursue the research direction of interest to me. These fellowships included the NSF Graduate Research Fellowship, the University of Michigan ITiMS Fellowship, the University of Michigan Jack A. Borchardt Student Support Fellowship, and the University of Michigan Rackham Conference Travel Grant.

I am extremely grateful to the research advisors that supported and mentored me through undergraduate and graduate school. Thank you for the guidance, perseverance, support, and investment that my co-chairs, Dr. Krista Wigginton and Dr. Melissa Duhaime, gave me over the last five and a half years and for allowing me to explore the intersection of environmental engineering and ecology and evolutionary biology. I am so grateful to have had both of their mentorship throughout graduate school and their flexibility and willingness to adapt when challenges or conflicts arose. Thank you to the members of my dissertation committee, Dr. Lutgarde Raskin and Dr. Luis Zaman, for helping develop my doctoral research. Thank you to Dr. Zaman for teaching my favorite graduate school course and introducing me to evolutionary biology. Thank you to Dr. Raskin for asking difficult questions related to experimental design and data analysis that improved my study designs and conclusions. Finally, thank you to the Dr.

Craig Just, Dr. Hunter Schroer, Dr. Ellen Black, and James Shannon for introducing me to research, environmental microbiology, and the possibility of pursuing an academic career path.

This research benefited from the support, feedback, and dedication of several individuals. Thank you to Dr. Bridget Hegarty for asking and answering the crucial question of "What is a viral contig?" and being a constant source of feedback with ideas and manuscripts. Thank you to Ariel Roy, Kaitlyn Chin, Michael Mata, and Ernesto Martinez for helping me with lab work, Santiago Vidaurri for helping develop code and analyze data, and Dr. Emily Crossette with research ideas. Thank you to the Wigginton and Duhaime research groups for listening to presentations, reviewing manuscripts, and engaging in thoughtful discussions on research topics and directions. Thank you to the PIRE HEARD collaborators, especially Mariah Gnegy for traveling to Hong Kong with me to sample wastewater and Dr. Tong Zhang and Dr. Andong Li for hosting us.

While at the University of Michigan, I benefitted from the ITiMS and Environmental Biotechnology communities. Through these programs, I interacted with fantastic peers that helped expand my research ideas and methodology while also benefiting my professional development. Within the ITiMS community, I am appreciative of Dr. Betsy Foxman, Freida Bolstein, Morgan Lindback, and James Tan. I am so appreciative of the colleagues and friendships that have offered advice, support, coffee runs, and fun throughout graduate school. Thank you to Dr. Nicole Rockey and David Barron, Kate Dowdell and Ted Bloch-Rubin, Dr. Yinyin Ye, Elizabeth and Hunter Wilson, Ellen and John Thompson, Dr. Maggie Reuter, Maddy and Tim Fairley-Wax, Rachel Cable, AJ Wing, Eric Bastian, Dr. Alex Szczuka, Mira Chaplin, Enrique Rodriguez, Dr. Brett Wagner, and Dr. Sarah Potgieter. Finally, I am forever grateful for my family and friends that supported my ambitions and helped me enjoy and laugh even during the tough parts of graduate school. Thank you to my parents, Ann and Tom Langenfeld, my sister, Hannah, my brother-in-law and sister-in-law, Aaron and Jenny Nessler, my parents-in-law, Julie and Randy Nessler, and nephew and niece, Callum and Azelie Nessler. Thank you to some fantastic friends, Harper Beasley, Hailey Klabo, Rebecca Rethwisch, and Victoria Dinger. Lastly, thank you to my ultimate support team, Ian Nessler and Ivy, for being with me during every good or bad moment in graduate school.

Table of Contents

List of Tables

List of Figures

List of Appendices

Abstract

Municipal wastewater treatment removes carbon and nutrients from sewage by harnessing a dense microbial community in a biological treatment process. The dynamics of the viral community structure and function through wastewater treatment is not well understood. Viruses are expected to play critical roles in biological wastewater treatment because they are highly abundant, exhibit complex host interactions ranging from predatory to symbiotic, and accelerate host evolution. The lack of rigorous methods for isolating viral communities from environmental samples and quantitative methods for measuring and interpreting viral metagenomes has hindered our understanding of the roles of viruses in the environment, in general, and biological wastewater treatment, in particular. The overall goal of this dissertation research was to develop and apply metagenomic and *in silico* approaches to explore viral community dynamics through biological wastewater treatment and probe the roles that viruses play on the dissemination and emergence of antibiotic resistance.

This dissertation developed rigorous methodologies for studying environmental viromes. To address the issue of virus enrichment from environmental samples, an ultrafiltration approach was compared with an iron chloride flocculation method. Next, to measure the absolute abundances of target viruses in wastewater samples before and after treatment, a rigorous quantitative viral metagenomic method was developed. Specifically, dsDNA and ssDNA standards were added to viral DNA extracts to relate relative and absolute abundances. A bioinformatic pipeline, QuantMeta, was developed to calculate concentrations of targets (e.g., contigs or sequences from databases) and assess target-specific detection thresholds and detect and correct non-specific mapping and assembly errors. QuantMeta was applied to quantitative viromes from wastewater samples and improved quantification confidence and accuracy. QuantMeta is not specific to wastewater viromes and is applicable to whole metagenomes and other environments.

These methods were applied to three samples of wastewater influent and secondary effluent collected in December 2020 from a municipal wastewater treatment plant. The wastewater viromes were highly purified for viruses with 75.5-78% of contigs classified as viral. Mean total virus concentrations in influent and secondary effluent were 10.3 and 10.6 $\log_{10}$ gc/mL, respectively, approximately two-orders of magnitude higher than previous concentrations made with viral particle counting-based methods. 12.9% of influent viral populations persisted and replicated through biological treatment to be 10.3 $\log_{10}$ gc/mL more abundant in secondary effluent. Viruses rarely carried antibiotic resistance genes, with only 59 viral populations identified.

Finally, compounding effects of phage-host coevolution and antibiotic stress on antibiotic resistance emergence and expression in chemostat environments, such as biological treatment and the gut, was assessed using *in silico* evolution experiments. An Avida environment was developed that simulated an antibiotic with an evolvable trait to confer antibiotic resistance. Experiments demonstrated that phage-host coevolution accelerated the emergence of antibiotic resistance and the presence of phages and antibiotics occasionally resulted in decreased susceptibility to antibiotics. The results indicate that phages alter outcomes of antibiotic resistance evolution.

Overall, this dissertation provides critical tools for quantitative studies of viromes. Their application provides insight on viral community dynamics through wastewater treatment, including their overall abundances, diversity, and potential role in the spread of antimicrobial resistance. The tools developed here can be applied in future studies of viral and microbial communities in metagenomes to directly compare between samples.

Chapter 1 Introduction

Viruses are ubiquitous, abundant, and diverse, yet ecologists have just scratched the surface to uncover their impacts on microbial community composition, structure, and functions[1-4]. Most studies on wastewater viruses focused on common enteric viruses[5-9]; to date, much remains unknown about the other virus populations in wastewater. Wastewater treatment implements a biological treatment process using microorganisms to remove carbon and nutrients from wastewater before it is released into the environment[10]. Biological reactors contain a dense and active microbial community that supports high concentrations of phages (i.e., viruses that infect bacteria)[11]. Phages exhibit complex host interactions and accelerate host evolution to potentially impact the structure and function of microbial communities in biological reactors[12-16]. For example, recent research has linked certain phages to bacteria causing foaming and bulking issues in activated sludge[12, 16].

In addition to affecting the performance of the biological reactors, phage-host interactions may impact public health through the dissemination of antibiotic resistance genes (ARGs)[17-19]. Wastewater contains subinhibitory concentrations of antibiotics that may spur bacteria to obtain or evolve resistance mechanisms[20]. Phages may aid their hosts by horizontally transferring ARGs and accelerating evolution of novel ARGs.

This dissertation research advances quantitative metagenomics methods in order to explore viral community dynamics in wastewater. The role of viruses in ARG dissemination and evolution in wastewater are probed with quantitative viral metagenomes and an *in silico* approach to simulate ARG evolution.

1.1 Challenges of existing viral metagenomic methods

Compared to bacteria, viruses are studied less often in engineered systems. This is largely due to challenges associated with studying viruses in metagenomes. There have been 12 studies evaluating viral communities in wastewater using metagenomics[1-4, 12, 16, 18, 21-25]. These and other studies have taken one of two approaches to study viruses in metagenomes: data-mining whole metagenomes or sequencing isolated viruses (i.e., viromes). Data-mining whole metagenomes for viruses reduces the sequencing effort devoted to viral nucleic acids and relies on bioinformatic methods to distinguish DNA of viral origin. Viruses are the most abundant entity in an environmental sample, but their average genome size is several orders of magnitude smaller than the genomes of prokaryotes and eukaryotes[26, 27]. Consequently, viral nucleic acids comprise a small fraction of the total nucleic acids in a sample. Viruses are difficult to distinguish in whole metagenomes because viruses have high mutation rates, incorporate fragments of host genomes, and viral databases are sparse[28]. Alternatively, viral nucleic acids may be purified from environmental samples to apply sequencing effort to viruses, but it relies on establishing rigorous methods to concentrate viruses while removing non-viral nucleic acids. Chapter 2 compares two methods to concentrate and purify viral communities from environmental samples to establish best practices for preparing viral DNA extracts for sequencing.

Another challenge of metagenomics is that the data is inherently relative making direct comparisons between samples difficult. Several quantitative metagenomic methods were developed for whole metagenomes including spiking-in synthetic DNA[29, 30] or genomic DNA[31-34] at known concentrations or normalizing by total cell estimates based on flow cytometry or 16S rRNA or housekeeping gene copies[35-38]. These quantitative metagenomic methods were not previously validated for viromes and did not establish requirements for confident detection and accurate quantification. In Chapter 3, quantitative viromics was performed by adding synthetic DNA standards to wastewater viral communities. The quantification accuracy was improved by developing a method to determine target-specific detection and quantitative limitations by assessing the variability in reads mapping across standard sequences.

1.2 Viruses impact the structure and function of microbial communities

Viruses are up to two orders of magnitude more abundant than bacteria in the environment[39-41] resulting in lysis of approximately a third of the world's bacteria each day[42]. Viruses exhibit complex host interactions that can have important implications on the structure and function of microbial communities. For example, viruses can hijack host metabolism to express auxiliary metabolic genes[42]. They can also cause horizontal gene transfer between host organisms via transduction[42]. Transduction is typically an inefficient route for gene transfer but is known to play an important role in the evolution of the clinically relevant Methicillin-resistant *Staphylococcus aureus* (MRSA)[43, 44]. MRSA is resistant to several antibiotics and most often causes skin infections, but can also cause sepsis, pneumonia, and surgical site infections. *S. aureus* developed an autotransduction and lateral transduction routes to efficiently transduce genes, including ARGs[45, 46]. *S. aureus* typically carries several prophages and events triggering prophages to enter the lytic cycle are important precursors to transduction.

The role of transduction in ARG dissemination within wastewater treatment remains poorly understood. Recent metagenomic studies have found ARGs incorporated on viral contigs[17-19, 47, 48]. However, there are conflicting arguments for the potential importance of transduction in the environment. There is evidence that viruses rarely carry ARGs and most ARGs in viromes are packaged in putative vesicles[21, 24, 49]. Others argue that while transduction may be a rare occurrence compared to genomic ARGs such as plasmids, viral-associated ARGs likely have a larger temporal and spatial scale[17]. Quantitative metagenomics allows the direct comparison of ARG abundances through wastewater treatment and if viruses carrying ARGs persist through wastewater treatment. The quantitative metagenomic methods developed in this research facilitated quantitative measurements of viruses carrying ARGs through wastewater treatment in Chapter 4.

In addition to their role in horizontal gene transfer, viruses impact the evolution of bacteria by acting as parasites through lytic infection cycles. This predator-host relationship exerts an evolutionary pressure on bacteria to evolve mechanisms to resist phage infection. The parasite-host interactions follow the Red Queen hypothesis to accelerate evolution as phages and bacteria race to evolve new mechanisms of infection and resistance[15]. The Red Queen hypothesis proposes that competing organisms must constantly evolve, adapt, and proliferate to survive in environments with an evolving opposition. When subinhibitory concentrations of antibiotics are present, phage-host coevolution may have a compounding effect on bacteria as multiple stressors pressure bacteria to evolve resistance to phage infection and antibiotics. Antagonistic coevolution was previously shown to drive evolutionary changes in bacteria and phage as demonstrated in controlled laboratory communities and soil microbiomes[13-15] as well as *in silico* using Avida digital organisms and parasites[50]. To date, *in silico* approaches have not been applied to probe compounding effects of multiple environmental stressors on organism evolution.

Classic studies of evolution rely on environmental observation or laboratory scale experiments of living organisms. Limitations on time and the ability to measure events complicate study design and limit the scope of evolution experiments. Naturally occurring organisms are complex and it is impossible to track every change occurring during an *in vitro* or *in situ* experiment. *In silico* approaches with digital organisms in Avida resolve these shortcomings. The Avida digital environment provides a perfectly controlled setting to efficiently observe evolution for thousands of generations with organisms that inherit and mutate traits[51]. The evolution of genes, or “tasks”, in digital organisms provides instances of evolution where everything occurring in the environment is known and measurable (Figure 1.1). Parasites, acting like phages, may be injected into Avida simulations to infect organisms performing a task encoded on their genome and overtake the hosts’ CPU cycles effectively killing or limiting hosts’ replication, analogous to phages lysing or sickening hosts. This approach to evolutionary studies was implemented in Chapter 5 to study ARG evolution in the presence of phage-host coevolution. An Avida environment was created with an environmental stressor, such as an antibiotic, to test the influence of antagonistic coevolution on the rate antibiotic resistance evolves. The experiments were used to determine if coevolution and environmental stressors have a compounding effect on the emergence of resistance.

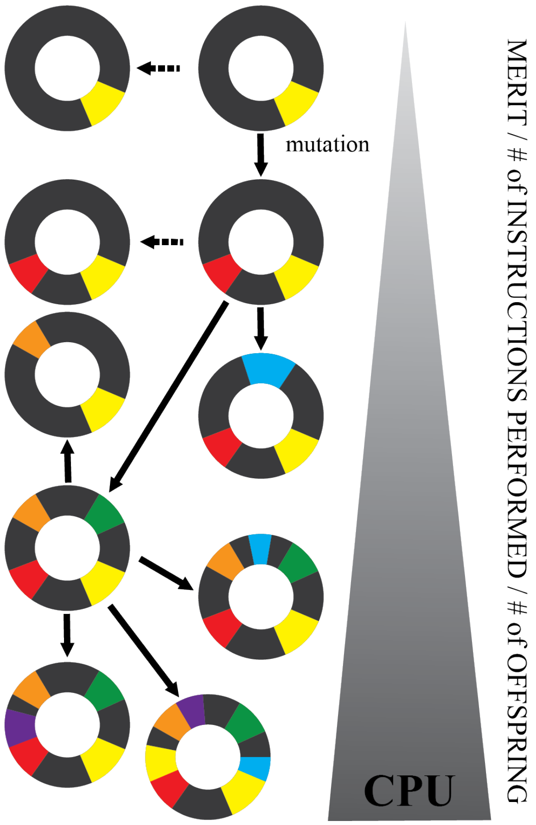

Figure 1.1: Digital organisms gain CPU or merit when they complete tasks (organisms are represented by circles and tasks are represented by colored sections on circles). The more CPU a digital organism has, the more instructions it is able to execute and the more offspring it produces. Digital organisms may mutate during reproduction creating more or less fit offspring.

1.3 Dissertation summary

Scientific conclusions are as reliable as the methods used to make observations. Therefore, Chapters 2 and 3 are devoted to developing a rigorous quantitative viral metagenomic method beginning with sample collection and sequencing preparation to bioinformatic analyses and establishing quantitative limitations. The method was applied in Chapter 4 to wastewater viral communities from influent and secondary effluent to explore the dynamics of viruses through biological treatment and identify antibiotic resistance genes integrated on viral genomes. The evolutionary impacts of phage-host interactions were explored in Chapter 5 by conducting experiments with Avida simulations. Lastly, the primary findings and implications of this dissertation and future research directions are summarized in Chapter 6.

1.4 References

1. Cantalupo, P.G. et al. Raw sewage harbors diverse viral populations. *mBio* **2** (2011).
2. Gulino, K. et al. Initial Mapping of the New York City Wastewater Virome. *mSystems* **5**, e00876-00819 (2020).
3. Tamaki, H. et al. Metagenomic analysis of DNA viruses in a wastewater treatment plant in tropical climate. *Environ Microbiol* **14**, 441-452 (2012).
4. Wang, Y., Jiang, X., Liu, L., Li, B. & Zhang, T. High-Resolution Temporal and Spatial Patterns of Virome in Wastewater Treatment Systems. *Environ Sci Technol* **52**, 10337-10346 (2018).
5. Bhatt, A., Arora, P. & Prajapati, S.K. Occurrence, fates and potential treatment approaches for removal of viruses from wastewater: A review with emphasis on SARS-CoV-2. *J Environ Chem Eng* **8**, 104429 (2020).
6. Corpuz, M.V.A. et al. Viruses in wastewater: occurrence, abundance and detection methods. *Sci Total Environ* **745**, 140910 (2020).
7. Haramoto, E. et al. A review on recent progress in the detection methods and prevalence of human enteric viruses in water. *Water Res* **135**, 168-186 (2018).
8. Ibrahim, Y. et al. Detection and removal of waterborne enteric viruses from wastewater: A comprehensive review. *Journal of Environmental Chemical Engineering* **9** (2021).
9. Saawarn, B. & Hait, S. Occurrence, fate and removal of SARS-CoV-2 in wastewater: Current knowledge and future perspectives. *J Environ Chem Eng* **9**, 104870 (2021).
10. Grady, C.L., Daigger, G.T., Love, N.G. & Filipe, C.D. Biological Wastewater Treatment. (CRC press, 2011).
11. Du, B. et al. Responses of bacterial and bacteriophage communities to long-term exposure to antimicrobial agents in wastewater treatment systems. *J Hazard Mater* **414**, 125486 (2021).
12. Chen, Y., Wang, Y., Paez-Espino, D., Polz, M.F. & Zhang, T. Prokaryotic viruses impact functional microorganisms in nutrient removal and carbon cycle in wastewater treatment plants. *Nat Commun* **12**, 5398 (2021).
13. Gómez, P. & Buckling, A. Bacteria-Phage Antagonistic Coevolution in Soil. *Science* **332**, 106-109 (2011).
14. Koskella, B. & Brockhurst, M.A. Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol Rev* **38**, 916-931 (2014).
15. Paterson, S. et al. Antagonistic coevolution accelerates molecular evolution. *Nature* **464**, 275-278 (2010).
16. Yang, Q., Zhao, H. & Du, B. Bacteria and bacteriophage communities in bulking and non-bulking activated sludge in full-scale municipal wastewater treatment systems. *Biochemical Engineering Journal* **119**, 101-111 (2017).
17. Debroas, D. & Siguret, C. Viruses as key reservoirs of antibiotic resistance genes in the environment. *ISME J* **13**, 2856-2867 (2019).
18. Li, X. et al. Metagenomic and viromic data mining reveals viral threats in biologically treated domestic wastewater. *Environmental Science and Ecotechnology* **7** (2021).
19. Moon, K. et al. Freshwater viral metagenome reveals novel and functional phage-borne antibiotic resistance genes. *Microbiome* **8**, 75 (2020).

20. Lindberg, R. et al. Behavior of Fluoroquinolones and Trimethoprim during Mechanical, Chemical, and Active Sludge Treatment of Sewage Water and Digestion of Sludge. *Environmental Science & Technology* **40**, 1042-1048 (2006).
21. Maestre-Carballa, L. et al. Insights into the antibiotic resistance dissemination in a wastewater effluent microbiome: bacteria, viruses and vesicles matter. *Environ Microbiol* **21**, 4582-4596 (2019).
22. Osunmakinde, C., Selvarajan, R., Mamba, B. & Msagati, T. Viral Communities Distribution and Diversity in a Wastewater Treatment Plants Using High-throughput Sequencing Analysis. *Polish Journal of Environmental Studies* **30**, 3189-3201 (2021).
23. Petrovich, M.L. et al. Viral composition and context in metagenomes from biofilm and suspended growth municipal wastewater treatment plants. *Microb Biotechnol* (2019).
24. Petrovich, M.L. et al. Microbial and Viral Communities and Their Antibiotic Resistance Genes Throughout a Hospital Wastewater Treatment System. *Front Microbiol* **11**, 153 (2020).
25. Wang, H. et al. Variations among Viruses in Influent Water and Effluent Water at a Wastewater Plant over One Year as Assessed by Quantitative PCR and Metagenomics. *Appl Environ Microbiol* **86** (2020).
26. Madigan, M.T., Martinko, J.M., Bender, K.S., Buckley, D.H. & Stahl, D.A. Brock Biology of Microorganisms, Edn. 14th. (Pearson Education, Inc., Glenview, IL; 2015).
27. Noble, R. & Fuhrman, J. Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. *Aquatic Microbial Ecology* **14**, 113-118 (1998).
28. Rose, R., Constantinides, B., Tapinos, A., Robertson, D.L. & Prosperi, M. Challenges in the analysis of viral metagenomes. *Virus Evol* **2**, vew022 (2016).
29. Hardwick, S.A. et al. Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. *Nat Commun* **9**, 3096 (2018).
30. Reis, A.L.M. et al. A universal and independent synthetic DNA ladder for the quantitative measurement of genomic features. *Nat Commun* **11**, 3609 (2020).
31. Crossette, E. et al. Enhancing metagenomic quantification of genes in environmental samples with internal standards. *mBio* **in press** (2021).
32. Lin, Y., Gifford, S., Ducklow, H., Schefield, O. & Cassar, N. Towards Quantitative Microbiome Community Profiling Using Internal Standards. *Applied and Environmental Microbiology* **85**, e02634-02618 (2019).
33. Satinsky, B.M., Gifford, S.M., Crump, B.C. & Moran, M.A. Use of internal standards for quantitative metatranscriptome and metagenome analysis. *Methods Enzymol* **531**, 237-250 (2013).
34. Stammler, F. et al. Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome* **4**, 28 (2016).
35. Jian, C., Luukkonen, P., Yki-Jarvinen, H., Salonen, A. & Korpela, K. Quantitative PCR provides a simple and accessible method for quantitative microbiota profiling. *PLoS One* **15**, e0227285 (2020).
36. Ju, F. et al. Antibiotic resistance genes and human bacterial pathogens: Co-occurrence, removal, and enrichment in municipal sewage sludge digesters. *Water Res* **91**, 1-10 (2016).
37. Majeed, H.J. et al. Evaluation of Metagenomic-Enabled Antibiotic Resistance Surveillance at a Conventional Wastewater Treatment Plant. *Front Microbiol* **12**, 657954 (2021).

38. Vandeputte, D. et al. Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507-511 (2017).
39. Cael, B.B., Carlson, M.C.G., Follett, C.L. & Follows, M.J. Marine Virus-Like Particles and Microbes: A Linear Interpretation. *Front Microbiol* **9**, 358 (2018).
40. Parsons, R.J., Breitbart, M., Lomas, M.W. & Carlson, C.A. Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea. *ISME J* **6**, 273-284 (2012).
41. Wigington, C.H. et al. Re-examination of the relationship between marine virus and microbial cell abundances. *Nat Microbiol* **1**, 15024 (2016).
42. Breitbart, M. Marine viruses: truth or dare. *Ann Rev Mar Sci* **4**, 425-448 (2012).
43. Gogarten, M.B., Gogarten, J.P. & Olendzenski, L. Horizontal Gene Transfer: Genomes in Flux. (Humana Press, New York, NY; 2009).
44. Haaber, J., Penades, J.R. & Ingmer, H. Transfer of Antibiotic Resistance in Staphylococcus aureus. *Trends Microbiol* **25**, 893-905 (2017).
45. Chiang, Y.N., Penades, J.R. & Chen, J. Genetic transduction by phages and chromosomal islands: The new and noncanonical. *PLoS Pathog* **15**, e1007878 (2019).
46. Haaber, J. et al. Bacterial viruses enable their host to acquire antibiotic resistance genes from neighbouring cells. *Nat Commun* **7**, 13333 (2016).
47. Colombo, S., Arioli, S., Guglielmetti, S., Lunelli, F. & Mora, D. Virome-associated antibiotic-resistance genes in an experimental aquaculture facility. *FEMS Microbiol Ecol* **92** (2016).
48. Subirats, J., Sanchez-Melsio, A., Borrego, C.M., Balcazar, J.L. & Simonet, P. Metagenomic analysis reveals that bacteriophages are reservoirs of antibiotic resistance genes. *Int J Antimicrob Agents* **48**, 163-167 (2016).
49. Enault, F. et al. Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J* **11**, 237-247 (2017).
50. Zaman, L. et al. Coevolution Drives the Emergence of Complex Traits and Promotes Evolvability. *PLoS Biology* **12** (2014).
51. Ofria, C. & Wilke, C. Avida: A Software Platform for Research in Computational Evolutionary Biology. *Artificial Life* **10**, 191-229 (2004).

Chapter 2 Comparing Ultrafiltration and Iron Chloride Flocculation to Prepare Aquatic Viromes from Various Matrices

2.1 Publication Information

Langenfeld, K., Chin, K., Roy, A., Wigginton, K., & Duhaime, M. B. (2021). Comparison of ultrafiltration and iron chloride flocculation in the preparation of aquatic viromes from contrasting sample types. *PeerJ*, 9.

This chapter was adapted from its published version for this dissertation. The purpose of this chapter is to identify a robust method to concentrate and purify viral communities for metagenomics.

2.2 Abstract

Viral metagenomes (viromes) are a valuable untargeted tool for studying viral diversity and the central roles viruses play in host disease, ecology, and evolution. Establishing effective methods to concentrate and purify viral genomes prior to sequencing is essential for high quality viromes. Using virus spike-and-recovery experiments, we stepwise compared two common approaches for virus concentration, ultrafiltration and iron chloride flocculation, across diverse matrices: wastewater influent, wastewater secondary effluent, river water, and seawater. Viral DNA was purified by removing cellular DNA via chloroform cell lysis, filtration, and enzymatic degradation of extra-viral DNA. We found that viral genomes were concentrated 1-2 orders of magnitude more with ultrafiltration than iron chloride flocculation for all matrices and resulted in higher quality DNA suitable for amplification-free and long-read sequencing. Given its widespread use and utility as an inexpensive field method for virome sampling, we nonetheless sought to optimize iron flocculation. We found viruses were best concentrated in seawater with 5-fold higher iron concentrations than the standard used, inhibition of DNase activity reduced purification effectiveness, and five-fold more iron was needed to flocculate viruses from freshwater than seawater—critical knowledge for those seeking to apply this broadly used method to freshwater virome samples. Overall, our results demonstrated that ultrafiltration and purification performed better than iron chloride flocculation and purification in the tested matrices. Given that the method performance depended on the solids content and salinity of the samples, we suggest spike-and-recovery experiments be applied when concentrating and purifying sample types that diverge from those tested here.

2.3 Introduction

Viruses are important members of natural and engineered aquatic ecosystems that can outnumber other microbes by up to two orders of magnitude[1-3], and influence their host's ecology and evolution through metabolic reprogramming and mortality[4, 5]. To better understand the fate and role of viruses in aquatic systems, whole community sequencing ('metagenomics') is used for the untargeted exploration of viruses in a community context[6]. Metagenomics has led to the unprecedented discovery of viral diversity and function[7-10], but its ability to deliver an unbiased representation of viral communities is nonetheless hindered by methodological challenges and biases[11].

The central challenge is that viral DNA comprises a small fraction of total community DNA due to viral genomes being one to four orders of magnitude smaller than eukaryotic and prokaryotic genomes[12]. As a result, in the preparation of virus-enriched metagenomes, 'viromes', it can be difficult to recover sufficient viral genomic material to generate sequencing libraries without biases introduced by amplification steps[10, 13]. Some studies avoid this obstacle by studying viruses in metagenomes prepared from whole community DNA, rather than viromes. But this approach may not be appropriate for all research questions, as it will capture temperate and actively replicating viruses, rather than predominantly free viruses. Further, when viral genomes are sequenced amidst the overwhelming cellular DNA, the sequencing effort dedicated to viruses is drastically limited. Purifying samples to increase the ratio of viral DNA to cellular DNA results in a more comprehensive representation of the viral community. This enhancement will increase the likelihood of sequencing low abundance and rare viruses and increase the sensitivity of viral detection studies. Overall, several of these challenges can be mitigated during sample preparation by efficiently concentrating and purifying viral genomes to focus maximal sequencing effort on viruses.

A number of methods to concentrate and purify viruses have been developed and are broadly used. Viruses are concentrated in water by exploiting their unique physical and structural properties, such as size (20 to 300 nm) and surface charge (commonly negative), with varying degrees of effectiveness. Methods that rely on charge include flocculation or precipitation (iron chloride, skimmed milk, lanthanum chloride, aluminum sulfate, aluminum chloride, and polyethylene glycol)[14-19] and virus adsorption-elution[15, 20-22]. Ultrafiltration takes advantage of particle sizes to concentrate viruses in dead-end, tangential, or axial flow configurations[14, 18, 20, 22-29]. Alternate virus concentration methods that do not rely on size or surface charge include ultracentrifugation and lyophilization[14, 15, 18]. Most of these concentration methods were developed for PCR-based detection where targets are selectively amplified and thus purification is not necessary. When purification is conducted on virome samples, non-viral biological material is removed from the aqueous samples via a number of approaches. For instance, submicron filtration can be used to separate cells from viruses, chloroform can be used to solubilize lipids in the cell membrane and cause cell lysis[30-32], non-encapsidated extra-viral DNA can be enzymatically degraded[33], and density gradients can be used to separate phages from cells based on their buoyant densities[25, 34, 35].

Despite decades of research on virus concentration and purification methods, a knowledge gap remains regarding the preparation of virome samples. Of the existing studies that have evaluated aquatic virome sample preparation[15, 25, 36], few have applied comparable methods or assessed performance at each of the concentration and purification steps, making cross-study comparisons difficult. While a multitude of studies have demonstrated the ability to produce aquatic viral metagenomes using various protocols[15, 25, 36, 37], none have evaluated viral recovery or removal of cellular DNA. Typically, these method assessments have focused on a single aquatic matrix. The performance, and thus suitability, of different concentration and purification methods across a variety of water sample types is limited.

In this study, we evaluated the impact of sample matrix on the performance of two commonly applied concentration methods, ultrafiltration and iron chloride flocculation. Ultrafiltration and iron chloride flocculation were selected because they have been widely used for marine and freshwater virome studies[7, 25, 38-42]. Using virus spike-and-recovery experiments, a step-wise assessment of each method was performed for four contrasting sample types that varied in their solids content and salinity: wastewater influent (i.e., raw sewage), wastewater secondary effluent (i.e., post carbon removal, pre-disinfection), river water, and seawater. Our findings will inform the sample preparation of future virome studies, especially as the quantitative rigor of metagenomics is further pursued.

2.4 Materials and Methods

2.4.1 Sample Collection

Grab samples of secondary effluent and raw influent were collected from automatic samplers at the Ann Arbor wastewater treatment plant (Ann Arbor, MI) in November 2018 through November 2019 (Table A.1). Grab samples of river water were collected from a boat ramp upstream of the Ann Arbor wastewater discharge site along the Huron River in Ann Arbor at the surface from May through July 2019 (Table A.1). Raw influent, secondary effluent, and river water were collected in autoclaved bottles and carboys. Samples were transported to the laboratory on ice within 1 h of collection and began processing immediately upon arrival in the lab. Seawater was collected from the Shedd Aquarium (Chicago, IL) on February 27, 2020 and immediately transported to the lab in Ann Arbor, MI on ice. Samples were stored at 4˚C until processing for a maximum of one week (Table A.2).

2.4.2 Sample characterization

Sample volumes were determined in the lab by weighing the sample. Immediately prior to processing, the pH was measured with a Mettler Toledo pH meter calibrated with a 4, 7, and 10 pH standards prior to measurement (Table A.1). The total suspended solids (TSS) and volatile suspended solids (VSS) were determined for each sample using standard methods with 20 mL of influent or river water in duplicate and 40 mL of secondary effluent stored at -20˚C until analysis (Table A.1)[43]. The seawater was tested for changes in pH during the week of storage (Table A.2).

2.4.3 Phages for spike and recovery

Phage spike-and-recovery experiments evaluated the performance of concentration and purification methods. Several different phages were spiked into the matrices to estimate the total viral recovery: *Enterobacteria* phage T3 (GenBank accession no. NC_003298, ATCC® BAA-1025-B1™), *Enterobacteria* phage T4 (GenBank accession no. NC_000866), *Enterobacteria* phage PhiX174 (GenBank accession no. NC_001422), *Pseudoalteromonas* phage HS2 (GenBank accession no. KF302036), *Pseudoalteromonas* phage HM1 (GenBank accession no. KF302034.1), and *Sulfitobacter* phage ICBM5, a ssDNA *Microviridae* phage provided by the Moraru Phage Lab (Institute for Chemistry and Biology of the Marine Environment, Oldenburg, Germany) (Table 2.1). Phages originating from freshwater (T3, T4, and PhiX174) and seawater (HS2, HM1, and ICBM5) were spiked into freshwater and seawater, respectively (Table 2.2), to avoid compromising the integrity of protein capsids due to salinity differences, as was observed in preliminary experiments (Section A.3). For secondary effluent and seawater samples, multiple phages were spiked into the same samples. T3, T4, HS2, and HM1 have tails protruding from the protein capsid[44-46] and PhiX174 and ICBM5 have small icosahedral capsids[6]. None of the phages were enveloped.

Table 2.1: Three freshwater phages and three marine phages were used in spike-and-recovery experiments. Characteristics of phages spiked into water samples to access viral recovery during concentration and purification processes.

Phage	**Genome Length (bp)**	**Genome Composition**	**Length (nm)**
Enterobacteria phage T3	38,208	dsDNA	70 (head and tail)
Enterobacteria phage T4	168,903	dsDNA	203 (head and tail)
Enterobacteria phage PhiX174	5,386	ssDNA	26
Pseudoalteromonas phage HS2	38,208	dsDNA	210 (head and tail)
Pseudoalteromonas phage HM1	129,401	dsDNA	~200 (head and tail)
Sulfitobacter phage ICBM5	5,581	ssDNA	29

Table 2.2: Phage spike additions for concentration and purification method evaluation. See Table 2.1 for characteristics.

Matrix	**Phage Spiked**

	Iron Chloride Flocculation and Purification	**Ultrafiltration and Purification**
Influent	T3	T3
Secondary Effluent	T3	T3, T4, PhiX174
River Water	T3	T3
Seawater	HS2, HM1, ICBM5	HS2, HM1, ICBM5

All phages were cultured prior to spike-and-recovery studies. T3 and T4 host (*E. coli* ATCC® 11303) and PhiX174 host (*E. coli* ATCC® 13609) were grown overnight in 25 mL of host media (Table A.3) at 37˚C and 180 rpm from glycerol stocks. HS2 host (*Pseudoalteromonas* sp. 13-15), HM1 host (*Pseudoalteromonas* sp. H71), and ICBM5 host (*Sulfitobacter* sp. SH24-1b) were suspended in 25 mL of host media (Table A.3) from culture plates and allowed to grow overnight at 25˚C and 180 rpm. The plaque overlay method generated plates of completely lysed bacterial lawns for each phage. Briefly, 100 µL of 10^6 pfu mL^{-1} T3, T4, or PhiX174 were combined with 100 µL of respective host in 3.5 mL of soft nutrient agar and poured over a hard nutrient agar plate (Table A.3), then incubated at 37˚C overnight. For HS2, HM1, and ICBM5, the same plaque overlay method was used with 300 µL of respective host and incubation at 25˚C overnight. The top layer of soft agar was gently collected and 5 mL of respective buffer (Table A.3) was poured on each completely lysed plate. The plates were gently mixed and incubated on the benchtop for 20 minutes. The buffer was combined with the soft agar, gently shaken, and incubated at 4˚C for 2 hours. The mixture was vortexed for 30 seconds and treated with 0.5 mL chloroform, vigorously mixed for 2 minutes, and centrifuged (Table A.4). The supernatant was collected and aerated to remove trace chloroform. Stocks were 0.45-µm (T3, T4, HM1, HS2, ICBM5) or 0.22-µm (PhiX174) polyethersulfone (PES) filtered (CellTreat Scientific Products, Cat. No. 229771 and 229747, respectively) and stored at 4˚C.

2.4.4 Optimizing iron chloride flocculation

To optimize iron chloride flocculation, modified jar tests were performed in triplicate for each sample matrix. Samples were pre-filtered through a 100-µm Long-Life filter bag for water made of polyester felt (McMaster-Carr, Cat. No. 6835K58) then 0.45-µm Express PLUS PES filters (MilliporeSigma^{TM}, Cat. No. HPWP09050 or HPWP14250) and aliquoted in 500 mL increments into 6 autoclaved glass bottles with stir bars. Approximately 5 x 10^5 gene copies (gc) μL^{-1} of T3 were spiked into each jar for freshwater and approximately 10^7 gc μL^{-1} of HS2 were spiked into seawater jars. Samples were collected immediately following the spike addition for recovery analysis. A sterile 10 g Fe L^{-1} iron chloride solution was made immediately prior to use. For consistency with the previously established method, half of the iron dose was added to each jar followed by a minute of turbulent mixing on a stir plate and then repeated[47]. Iron concentrations of 0, 0.1, 1, 5, 10, and 25 mg Fe L^{-1} were tested (Table A.5). Following the addition of iron, samples were left for an hour at room temperature, then flocs were captured on 0.45-µm Express PLUS filters. Samples of the filtrate, or material passing through the filter, were collected for recovery analysis (Figure 2.1). The floc-containing filters were carefully placed in sterile centrifuge tubes along with freshly-made 1x oxalic acid resuspension buffer, as described previously[48] (Table A.5). All samples were placed on a shaker table in the dark at 4˚C and 180 rpm overnight to dissolve the flocs. The concentrate (i.e., the resuspension buffer with dissolved flocs), was separated from the filters as described previously[47] and samples were collected for recovery analysis (Figure 2.1).

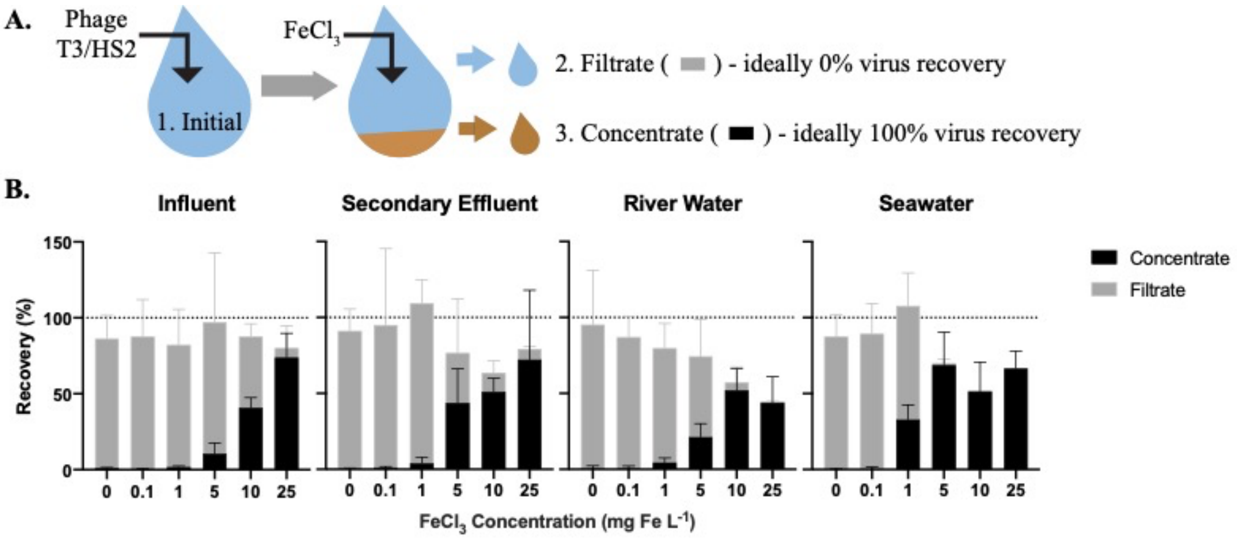

Figure 2.1: The efficiency of virus flocculation was tested with jar tests at several iron chloride concentrations. (A) T3 for freshwater matrices and HS2 for seawater was spiked into jars and an initial sample was collected to determine the initial T3 or HS2 genome concentration (1). Iron chloride was added to the samples, which underwent turbulent mixing, then flocs were captured on filters. A sample of the water flowing through the filter (the 'filtrate' (2)) was collected to determine the number of T3 or HS2 genomes not flocculated. The filter was placed in resuspension buffer to dissolve flocs. A concentrate sample (3) was then collected to determine the number T3 or HS2 captured in flocs and successfully resuspended. (B) The recoveries of T3 and HS2 genomes in the sample concentrate (black) and filtrate (gray). Bars are stacked. Error bars represent the standard deviations around the geometric mean of experimental triplicates. A perfect recovery of T3 or HS2 genomes would result in the filtrate and concentrate bars stacking to the 100% recovery dashed line.

2.4.5 Iron chloride flocculation and purification method (Figure 2.2A)

Step 1: Iron Chloride Flocculation. The protocol was adapted from John et al. (2011) to concentrate viruses. 500-mL influent, secondary effluent, and river water samples and 20-L seawater were pre-filtered through 0.45-µm Express PLUS filters. Rather than 0.22-µm pore size filters, 0.45-µm pore size filters were used in this step to recover giant phages[49, 50]. Half of the iron required for the best concentration, as determined in the jar tests (above), was added to the sample. Samples were rapidly mixed for a minute with a magnetic stir bar for freshwater matrices or shaken in carboys for seawater. This process was repeated with the remaining half of the required iron. Samples were left at room temperature for 1 hour to flocculate. The flocculated viruses were captured on 0.45-µm Express PLUS PES filters. PES matrix filters were used for this step because of their superior flow rates, as compared to the limited flow rates achieved with

PC isopore filters, an observation consistent with prior knowledge[51]. Further, preliminary experiments comparing 0.22-µm PES filters versus 1-µm polycarbonate filters demonstrated the PES filter had superior flow rates (2.9-fold greater than the PC) and equivalent recoveries of nucleic acids, as measured by a Nanodrop spectrophotometer (data not included). The filters were carefully folded into sterile centrifuge tubes. 1 mL of 1x oxalic acid resuspension buffer per mg Fe used to flocculate the sample was added to the filter containing tubes. The viruses were resuspended from the filters by shaking at 4˚C and 180 rpm overnight. The solution was transferred to a clean centrifuge tube as described previously[47]. Only 500-mL was used for freshwater samples because the concentration by volume is independent of initial sample volume (i.e., mass of iron added controls the volume of the resuspension buffer added) and a final volume of 12.5 mL is sufficient for subsequent steps.

Step 2: Chloroform and filtration. To isolate the viruses in the sample, cells were lysed with chloroform. 1 mL chloroform was added to the sample and vortexed for approximately 2 minutes. Chloroform settled out of suspension by sitting undisturbed on the benchtop for 15 minutes. The bulk of the chloroform was pipetted off the bottom of the sample and disposed. The trace remainder of chloroform was aerated from the sample in a fume hood for approximately 10 minutes. The cell debris was filtered from the sample with 0.45-µm Express PLUS filters.

Step 3: DNase treatment. DNase treatment was performed with a previously established method with the reaction time reduced to one hour[52]. Briefly, lyophilized DNase 1, grade II from bovine pancreas (Roche, Cat. No. 10104159001) was resuspended in a storage buffer (10mM Tris-Cl pH 7.5, 2 mM $CaCl_2$ in 50% glycerol) to a concentration of 40,000 U mL^{-1} and stored at -20˚C. Immediately prior to DNase treatment of sample, DNase 1 in storage buffer was diluted 1:40 in a 10x reaction buffer (100 mM Tris-HCl pH 7.6, 25 mM $MgCl_2$, 5 mM $CaCl_2$) and gently mixed. 100 U mL^{-1} DNase 1 in the reaction buffer was added to the sample and reacted for 1 hour on the bench top. The DNase reaction was inhibited with the addition of 100 mM EDTA and 100 mM EGTA. Chloroform and DNase treatments were performed to purify the viral nucleic acids as tested in a purification optimization experiment provided in the Section A.4.

2.4.6 Ultrafiltration and purification method (Figure 2.2B)

Step 1: Tangential ultrafiltration. 10 L influent or 20 L of other matrices were pre-filtered through 100-µm polyester filter bag then 0.45-µm ExpressPLUS filters to remove large particles. Tangential ultrafiltration was performed with hollow-fiber ultrafilters, specifically Dialyzer Rexeed single use dialysis filters with a surface area of 2.5 m^2 and approximate molecular weight cut-off of 30 kDa (Asahi Kosei Medical Co., Ltd, Cat. No. 6292966), as described previously[24]. Briefly, the sample and filter were configured such that the sample passed through the filter tangential to the membrane surface. The sample volumes progressively decreased as water passed through the membrane pores and particles were retained. New filters were used for each sample. The sample flowed through the filter in the direction labeled blood until the sample volume was minimized and air began to enter the tubing. The minimized volume was approximately 350 mL. At the minimal volume, the flow direction was reversed, and the concentrated sample was collected. The exact volume after tangential ultrafiltration was determined by weighing the sample and assuming a density of 1 g mL^{-1}.

Step 2: Chloroform and filtration. The same chloroform and filtration method as performed during the iron chloride flocculation and purification method was implemented in this method.

Step 3: Dead-end ultrafiltration. The sample was concentrated an additional 20-fold with dead-end ultrafiltration. Dead-end ultrafiltration was performed with 100 kDa MWCO and 1 cm^2

surface area Amicon™ Ultra Centrifugal filter units (MilliporeSigma™, Cat. No. UFC510096). Four new filters were used for each sample and processed in parallel. The sample was centrifuged at 3,000*xg* and 4˚C while incrementally refilling the filters until 4 mL of sample was reduced to 200 µL per filter. Additional pre-washing, incubation, or sonication steps were not included in the protocol because results from preliminary experiments indicated that pre-washing the filter with water, incubating concentrate with BSA, and sonicating the filter cartridge prior to collecting concentrate did not improve viral genome recoveries (Section A.5; Figure A.2). The concentrate was collected by inverting the filter into a clean collection tube and centrifuging at 1,000*xg* for 1 minute. The contents from individual filters were combined. No additional treatment of the ultrafilters was performed prior or following dead-end ultrafiltration in accordance with the results from a dead-end ultrafiltration optimization experiment summarized in Section A.5.

Step 4: DNase treatment. The same DNase treatment as performed during the iron chloride flocculation and purification method was implemented for this method.

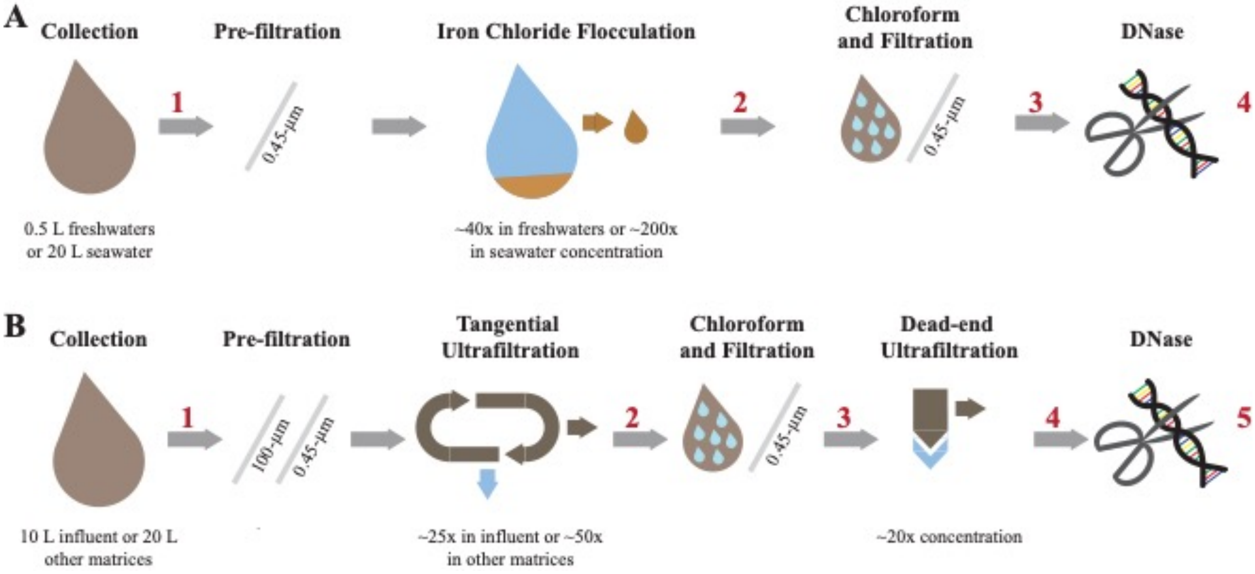
A
Collection
Pre-filtration
Iron Chloride Flocculation
Chloroform and Filtration
DNase
1
0.45-µm
2
0.45-µm
3
4
0.5 L freshwaters or 20 L seawater
~40x in freshwaters or ~200x in seawater concentration
B
Collection
Pre-filtration
Tangential Ultrafiltration
Chloroform and Filtration
Dead-end Ultrafiltration
DNase
1
100-µm
0.45-µm
2
0.45-µm
3
4
5
10 L influent or 20 L other matrices
~25x in influent or ~50x in other matrices
~20x concentration

Figure 2.2: Overview of each step involved in concentrating and purifying viruses with iron chloride flocculation and ultrafiltration. Concentration and purification process for the iron chloride flocculation and purification

method (A) and the ultrafiltration and purification method (B). The red numbers indicate where aliquots were collected to measure viral genome and 16S rRNA gene copy concentrations.

2.4.7 Phage concentration and 16S rRNA removal analysis

Approximately 10^4 gc μL^{-1} per phage were spiked into samples prior to pre-filtering to monitor the recovery after each step of iron chloride flocculation and purification and ultrafiltration and purification (Table 2.2). We used removal of the ~500 bp long 16S rRNA V3 region amplicon to approximate the removal of non-viral DNA in the sample. 16S rRNA is commonly used to estimate total bacteria counts in samples because it is a conserved region of bacterial genomes[53]. A before sample was collected prior to the phage addition to examine background concentrations of spiked phages in the samples. An "initial" sample was collected immediately after spike additions to determine the exact concentration of each phage representing total phage recovery. For the iron chloride flocculation and purification method, additional samples were collected after 0.45-µm filtering and iron chloride flocculation, chloroform and 0.45-µm filtering, and DNase treatment (Figure 2.2A). Samples were collected after 0.45-µm filtering and tangential ultrafiltration, chloroform and 0.45-µm filtering, dead-end ultrafiltration, and DNase treatment for the ultrafiltration and purification method (Figure 2.2B). The initial 16S rRNA concentration was determined with the "initial" sample for iron chloride flocculation and purification and ultrafiltration and purification experiments. Four key parameters were calculated to assess concentration and purification performance: virus concentration factor, virus recovery, 16S rRNA concentration factor, and virus to 16S rRNA enrichment. The virus concentration factor is the concentration of viral genomes after a step divided by the initial concentration (Equation 2.1). The virus recovery builds from the virus concentration factor by accounting for the change in volume occurring throughout the concentration and purification processes to identify losses of virus (Equation 2.2). The same virus concentration factor calculation was applied to 16S rRNA gene copies to calculate the 16S rRNA concentration factor (Equation 2.3). Lastly, the virus to 16S rRNA enrichment is the ratio of the virus concentration factor and the 16S rRNA concentration factor to determine if viral genomes were selectively concentrated throughout the processes (Equation 2.4). Gene copy concentrations were determined with ddPCR probe assays for iron chloride flocculation and purification and ultrafiltration and purification methods or qPCR SYBR green assays for optimizing iron chloride flocculation defined in *T3 and HS2 qPCR assays* and *Phage and 16S rRNA ddPCR assays* sections.

$$Virus\ Concentration\ Factor = \frac{[Phage]_{Step}\ (gc\ \mu L^{-1})}{[Phage]_{Initial}\ (gc\ \mu L^{-1})} \tag{2.1}$$

$$Recovery\ (\%) = Virus\ Concentration\ Factor \cdot \frac{Volume_{Step}(mL)}{Volume_{Initial}(mL)} \cdot 100\% \tag{2.2}$$

$$16S\ rRNA\ Concentration\ Factor = \frac{[16S\ rRNA]_{Step}\ (gc\ \mu L^{-1})}{[16S\ rRNA]_{Initial}\ (gc\ \mu L^{-1})} \tag{2.3}$$

$$Virus\ to\ 16S\ rRNA\ Enrichment = \frac{Virus\ Concentration\ Factor}{16S\ rRNA\ Concentration\ Factor} \tag{2.4}$$

2.4.8 DNA extraction

DNA extraction was performed with QIAamp UltraSens Virus Kit (QIAGEN, Cat. No. 53706). The manufacturer’s protocol was followed with minor changes. The first six steps were modified to combine 140 µL of sample with 5.6 µL carrier RNA and vortexed briefly. All DNA extractions occurred within 3 hours of sample generation.

2.4.9 T3 and HS2 qPCR assays

Primers (5’ to 3’) specific to T3 were selected (351 bp; forward, CCA ACG AGG GTA AAG TGA TAG; reverse, CGA CGA TAG CGA ATA GGA TAA G). Primers specific to HS2 were selected (300 bp; forward, GGT TGA TGA AAA GTC ACT; reverse, CGG GGC AGA TCT AAA TGA). The 10 µL reaction contained 5 µL 2x Biotium Fast-Plus EvaGreen master mix, 0.5 μM T3 or HS2 primers, 0.625 mg mL^{-1} bovine serum albumin, and 1 μL of DNA template. Standard curves were prepared in triplicate between 100 and 10^6 gene copies μL^{-1} with gBlocks dsDNA fragments of the amplicon sequence (IDT, Coralville, IA) (Table A.7). Ten replicates of the previously determined limit of quantification (T3 = 100 gene copies μL^{-1}, HS2 = 30 gene copies μL^{-1}) were measured on each plate, and two ddH_2O negative controls and two positive controls of DNA extracts from virus stocks were included on each plate. All positive controls were positive and all negative controls were negative for T3 or HS2. qPCR was performed with the realplex^2 Mastercycler epgradient S automated real-time PCR system (Eppendorf®, New York City, NY) with standard reaction conditions (Table A.8). All efficiencies were greater than 80% and R^2 values were greater than 0.98. Inhibition was assessed by comparing measurements for a random sampling of undiluted and 1:10 diluted freshwater DNA extracts. Wastewater samples were not found to have inhibition, but river water samples had inhibition at high iron concentrations and T3 qPCR was performed with 1:10 diluted samples and 1:100 diluted samples for 10 and 25 mg Fe L^{-1} concentrate samples. All HS2 qPCR measurements for seawater samples were diluted 1:10. Each sample was measured in duplicate and the geometric mean was reported.

2.4.10 Phage and 16S rRNA ddPCR assays

Singlet ddPCR reactions were performed with the QX200 AutoDG Droplet Digital PCR System (Bio-Rad Laboratories, Inc., Hercules, CA) with at least two ddH_2O negative controls and one positive control DNA extract from virus stocks per 96-well plate. Samples were multiplexed with two targets per reaction. Specific primer, probe, and annealing temperatures are provided in Table A.10. 22 μL reactions contained 11 μL of 2x ddPCR™ Supermix for Probes (No dUTP) (Bio-Rad Laboratories, Inc., Cat. No. 1863023), 0.4 μM of all probes and primers, and 3 μL of DNA template. Droplets were generated to a 20 μL reaction volume using the automated droplet generation oil for probes (Bio-Rad Laboratories, Inc., Cat. No. 1864110) and the plate was sealed. PCR was performed on the C1000 Touch^{TM} Thermal Cycler (Bio-Rad Laboratories, Inc., Hercules, CA) within 15 minutes of droplet generation with reaction conditions provided in Table A.11. Plates were run on the droplet reader within 1 hour of PCR completion with the exception of one plate that was stored at 4°C for 60 hours due to an error with the droplet reader. Thresholds were set for each ddPCR reaction to determine the absolute abundance of 16S rRNA amplicons and phage amplicons using a previously defined method from Lievens et al. that categorizes droplets as positive, negative, or rain based on kernel density estimates for each reaction[54]. Reactions were rerun if there were more than two fluorescence populations, more than 2.5% of droplets were classified as rain, or less than 30% compartmentalization. Background signals were present in all 16S rRNA ddH$_2$O negative controls (n=22, geometric mean = 263 gc μL^{-1}, 99% CI: 233-305 gc μL^{-1}), as observed in previous studies[55-57]. The 16S rRNA negative controls were significantly less than sample 16S rRNA measurements (*p*-value < 0.000001) with 16S rRNA concentrations greater than the upper limit of the 99% confidence interval deemed acceptable (i.e., limit of quantification = 305 gc μL^{-1})[58]. The mean background signal concentration of each 16S rRNA ddPCR run was subtracted from 16S rRNA sample measurements for the respective run to correct for 16S rRNA background signal. Alternatively, negative controls for the virus assays rarely resulted in target detection (T3: n = 3, max = 6.2 gc μL^{-1}; T4: n = 2, max = 14.8 gc μL^{-1}; PhiX174: n = 0; HS2: n =1, max = 2.8 gc μL^{-1}; HM1: n = 1, max = 13.3 gc μL^{-1}; ICBM5: n = 1, max = 2.7 gc μL^{-1}). Given that viruses were spiked into samples at 10^4 gc μL prior to concentrating for spike-and-recovery experiments, the rare virus detections in negative controls was deemed negligible.

2.4.11 DNA concentration and quality assessment

After the complete ultrafiltration and purification method and iron chloride flocculation and purification method, DNA concentration and fragmentation were assessed. The dsDNA concentration was measured with Qubit^{TM} dsDNA HS Assay (Invitrogen^{TM}, Cat. No. Q32851) with 1 or 2 µL of DNA template added to each 200 µL assay. The ssDNA concentration was determined by taking the difference between the measurement from Qubit^{TM} ssDNA Assay (Invitrogen^{TM}, Cat. No. Q10212) and the dsDNA measurement. The ssDNA assay was performed with 1 µL of DNA template added to each 200 µL assay. DNA fragmentation for each matrix and method (triplicates pooled, 9 total samples) was assessed by Agilent TapeStation for DNA lengths up to 60,000 bp (Agilent, Cat. No. 5067-5365) according to manufacturer protocols. TapeStation processing was carried out in the Advanced Genomics Core at the University of Michigan.

2.4.12 Statistical analysis

All statistical analysis and graphs were completed in Prism (version 8.4.3, GraphPad Software, LLC). Reported means are geometric with their respective 95% confidence intervals (CI) included (Table A.12 and Table A.13). Single phage spike recovery experiments were assessed with one-way ANOVA with Tukey's multiple comparison tests to generate *p*-values (Table A.14 and Table A.15). Multiple phage spike recovery experiments were assessed with two-way ANOVA with Tukey's multiple comparison test to generate *p*-values (Table A.14 and Table A.15). The outcomes from the two methods were compared with paired two-tailed t-tests with Holm-Sidak method to correct for multiple comparisons on each tested matrix for the final dsDNA and ssDNA concentrations, virus concentration factors, and 16S rRNA concentration factors (Table 2.3 and Table A.16). Significance for all comparisons was any *p*-value less than 0.05 for all tests.

2.4.13 Data availability

Jar test qPCR data and stepwise ultrafiltration and purification and iron chloride flocculation and purification ddPCR data is available in csv format in the "Viral Concentration and Purification Methods" Github repository (github.com/klangenf/Viral-Concentration-and-Purification-Methods).

2.5 Results

Two methods to concentrate and purify viruses were evaluated in four distinct matrices using virus spike-and-recovery tests. First, iron chloride concentrations were optimized for flocculation in each matrix using jar tests. Then, a two-step ultrafiltration method was compared to iron chloride flocculation stepwise through concentration and purification. The methods were evaluated based on viral concentration factors, viral recoveries, 16S rRNA concentration factors, virus to 16S rRNA enrichments, and final DNA quantity and quality. Finally, multiple DNA viruses were spiked into effluent and seawater samples to determine the extent to which the performance of each method was virus-specific.

2.5.1 Optimization of iron chloride flocculation

The highest virus recoveries in the concentrate of influent, secondary effluent, and river water were achieved with 25 mg Fe L^{-1}, where T3 genomes were recovered at 74%, 72%, and 44%, respectively (Figure 2.1B), signaling successful flocculation and capture of the viruses. The lowest viral recoveries in the filtrate of influent, effluent, and river water were achieved with 25 mg Fe L^{-1} where T3 genomes were recovered at 6.2%, 7.1%, and 1.4%, respectively, indicating successful flocculation. The sum of the concentrate and filtrate recoveries, which should theoretically equal 100% (Figure 2.1A), were 80%, 79%, and 46% of the spiked viral genomes for influent, effluent, and river water, respectively. For influent and secondary effluent, the viral genomes recovered in the filtrate decreased with increasing iron concentrations up to 25 mg Fe L^{-1}. We postulated that the iron chloride flocculation performance would continue to improve with increased iron chloride concentrations. However, this would require more oxalic acid resuspension buffer to dissolve the formed flocs and the increased volume would be counterproductive to concentrating the viruses. Furthermore, solubility limits of the oxalic acid were reached when preparation of a 2x more concentrated resuspension buffer was attempted[48]. Based on these limitations, we concluded that an iron chloride concentration of 25 mg Fe L^{-1} was the best option for recovering viral DNA in the freshwater samples.

For seawater, best recoveries in the concentrate were observed with Fe concentrations of 5, 10, and 25 mg L^{-1}, where 69%, 52%, and 67% of the spiked HS2 genomes were recovered, respectively (Figure 2.1B). In the filtrate, low recoveries of HS2 genomes were observed with Fe concentrations of 5, 10, and 25 mg L^{-1}, where 1.5%, 0.3%, and 0.1% of the spiked viruses were recovered, respectively. The filtrate and concentrate recoveries summed to 70%, 52%, and 67% with 5, 10, and 25 mg Fe L^{-1}, respectively. Given that iron chloride flocculation performs similarly at 5, 10, and 25 mg Fe L^{-1} and 5 mg Fe L^{-1} requires the smallest volume of resuspension buffer, 5 mg Fe L^{-1} was chosen for recovering viral DNA from seawater. Notably, at 1 mg Fe L^{-1}, the current standard in seawater flocculation[16], HS2 genome recoveries were 33% in the concentrate and 75% in the filtrate, indicative of poorer flocculation than with the higher Fe concentrations tested here.

The sum of the filtrate and filter concentrate recoveries was often less than 100%, regardless of matrix. Control experiments confirmed that T3 and HS2 genomes did not degrade over the length of the iron chloride flocculation process in any matrix (Table A.6). Thus, we suspect that the loss of viral genomes was due to inefficiencies in dissolving the flocs from the filters, which reduced recoveries in the concentrate.

2.5.2 Evaluation of virus concentrating and recovery

Overall, the viruses were concentrated more following ultrafiltration and purification than with iron chloride flocculation and purification (Table 2.3). Specifically, the iron chloride flocculation and purification approach resulted in T3 and HS2 concentration factors of 7.6, 6.7, 7.6, and 25 for influent, effluent, river water, and seawater, respectively. The ultrafiltration and purification approach resulted in T3 and HS2 genome concentration factors of 220, 440, 410, and 150-fold for influent, effluent, and river water, and seawater, respectively. These virus concentration factors for the entire concentration and purification approaches were the result of two effects, the volume reduction ($volume_{final}$/$volume_{initial}$) and the virus recovery through all of the processes. During iron chloride flocculation, the volume of the freshwater samples was reduced from 500-mL to 12.5-mL and seawater was reduced from 20-L to 100-mL. The amount that the volume is reduced with iron chloride flocculation depends on the mass of iron added to the sample, so increasing the initial volume will not increase the relative reduction in sample volume. Alternatively, for the ultrafiltration and purification approach, the volumes were reduced approximately 500-fold for influent and 1,000-fold for effluent, river water, and seawater. The T3 and HS2 genome recoveries after iron chloride flocculation and purification were 25%, 21%, 25%, and 15% in influent, effluent, river water, and seawater, respectively. The final recovery of

T3 and HS2 genomes after ultrafiltration and purification was 47%, 42%, 43%, and 18% in influent, effluent, river water, and seawater, respectively.

Virus concentration factors increased following the iron chloride flocculation and ultrafiltration concentration steps, as expected. The magnitude of the virus concentration factor correlated with the reduction in volume for iron chloride flocculation, which can account for the high virus concentration factors in seawater compared to freshwater samples (Figure 2.3A). Virus concentration factors increased twice during the ultrafiltration and purification process due to two ultrafiltration steps (Figure 2.4A). In the freshwater matrices, the higher virus concentration factors in effluent and river water compared to influent correlates to a greater reduction in volume. Conversely, influent and seawater virus concentration factors were similar despite a greater volume reduction for seawater due to poorer recovery of HS2 in seawater, as compared to T3 in freshwater.

Despite demonstrated concentration of viruses, all of the concentration methods resulted in statistically significant viral losses (Figure 2.3B and Figure 2.4B). The ranges of recoveries were 63-80% following iron chloride flocculation, 57-82% following tangential ultrafiltration, and 35-86% following dead-end ultrafiltration. Recovery significantly decreased in effluent and seawater following iron chloride flocculation (Table A.14). Recovery significantly decreased in effluent and seawater after both tangential ultrafiltration and dead-end ultrafiltration steps and in river water after only tangential ultrafiltration (Table A.15).

The purification steps aimed to remove non-viral DNA, not concentrate viruses, so changes in the virus concentration factor through chloroform and DNase treatments depended solely on virus recovery. DNase treatment caused significant viral genome losses following iron chloride flocculation, but not following ultrafiltration concentration. Specifically, the range of recoveries following DNase treatments were 25-48% and 80-120% following iron chloride flocculation and ultrafiltration, respectively. DNase treatment during iron chloride flocculation and purification resulted in statistically significant decreases in viral genome recoveries in the influent, effluent, and seawater samples (Table A.16). DNase treatment during ultrafiltration and purification did not result in a statistically significant reduction in viral genome recoveries in any of the sample types.

Based on the larger volume reductions and higher virus recoveries, we concluded that ultrafiltration and purification outperforms iron chloride flocculation and purification, particularly in freshwater matrices.

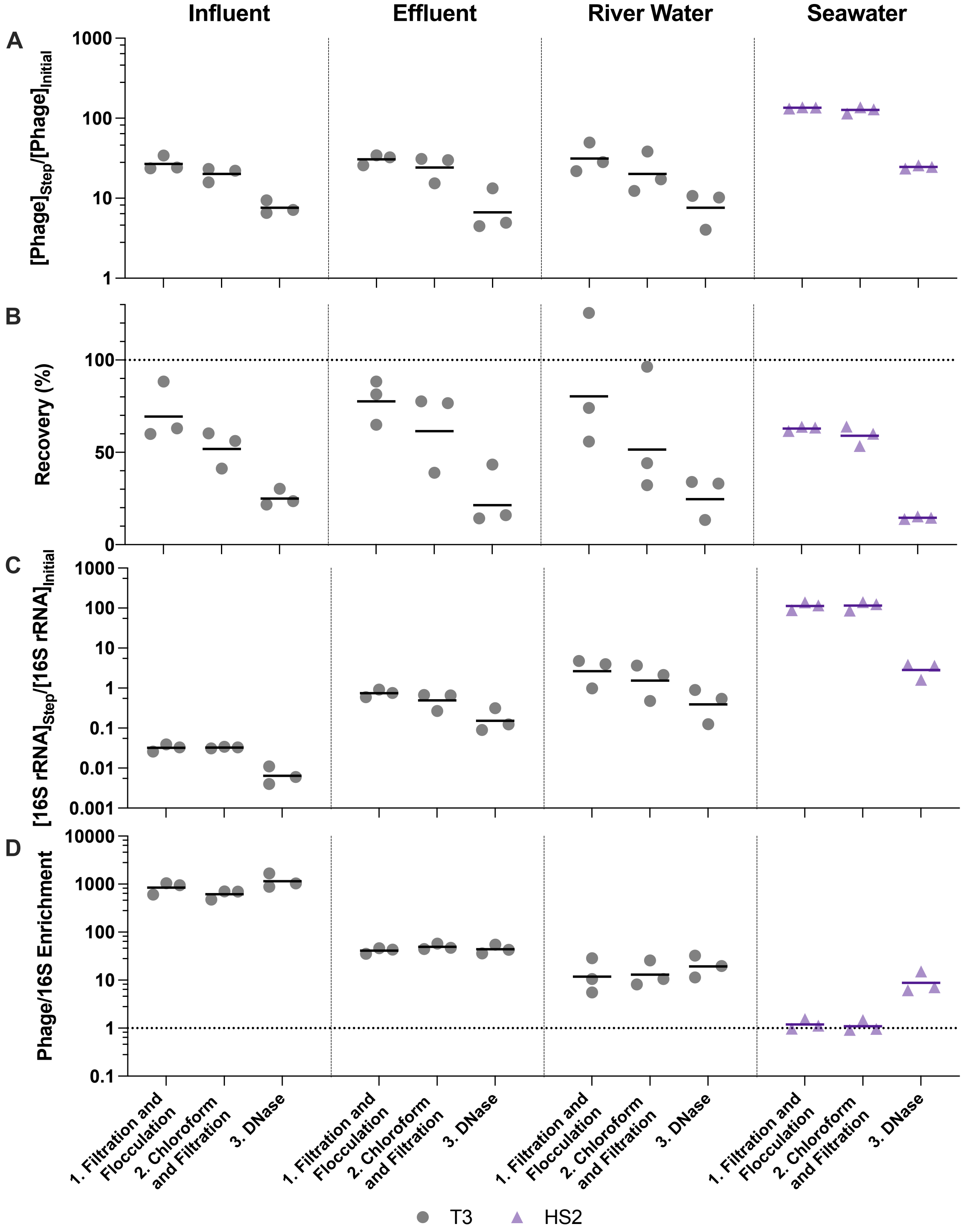

Figure 2.3: Stepwise evaluation of iron chloride flocculation and purification with 40-fold concentration of freshwater matrices and 200-fold concentration of seawater. 25 mg Fe L^{-1} was used for freshwater matrices and 5 mg Fe L^{-1} was used for seawater during iron chloride flocculation. (A) T3 or HS2 con- centration factors at each

step on a log-scale. Values greater than 1 indicated an increase in viral genomes. (B) T3 and HS2 recovery for each step of iron chloride flocculation demonstrated step-wise viral loss. Perfect recovery is indicated by the dotted line at 100%. (C) 16S rRNA concentration factor on a log-scale indicated non-viral DNA removal. Concentration factor values less than 1 indicated reduction of 16S rRNA gene copies. (D) T3 or HS2 to 16S rRNA enrichment with each step on a log-scale, as calculated by the virus concentration factor divided by the 16S rRNA concentration factor. Enrichments greater than 1 demonstrated viral genomes were concentrated more than 16S rRNA gene copies during the process. Individual measurements are points with a bar indicating the geometric mean of experimental replicates.

Table 2.3: Virus and 16S rRNA concentration factors after DNase treatment for both methods in each matrix. The geometric mean and geometric standard deviation from the triplicate data is provided. The methods were compared with individual t-tests corrected for multiple comparisons with the Holm-Sidak method were performed for each matrix and phage spike.

Matrix	**Phage Spike**	**Virus Concentration Factor**			**16S rRNA Concentration Factor**		
		Ultrafiltration	**Flocculation**	***p*-values**	**Ultrafiltration**	**Flocculation**	***p*-values**
Influent	T3	220 (120, 420)	7.6 (4.7, 12)	8.4E-3 (**)	0.062 (0.016, 0.23)	6.4E-3 (1.8E-3, 0.023)	0.15 (ns)
Secondary Effluent	T3	440 (360, 540)	6.7 (1.5, 30)	6.8E-4 (***)	6.3 (3.0, 13)	0.15 (0.030, 0.76)	0.19 (ns)
	T4	200 (71, 550)	NA	NA			
	PhiX174	45 (20, 100)	NA	NA			
River Water	T3	410 (280, 610)	7.6 (1.9, 30)	2.4E-3 (**)	0.94 (0.48, 1.8)	0.39 (0.031, 5.0)	0.33 (ns)
Seawater	HS2	150 (50, 450)	25 (22, 28)	0.055 (ns)	1.6 (0.50, 5.1)	2.8 (0.85, 9.3)	0.33 (ns)
	HM1	250 (75, 850)	48 (40, 58)	0.055 (ns)			
	ICBM5	110 (55, 220)	3.7 (2.3, 5.9)	9.3E-3 (**)			

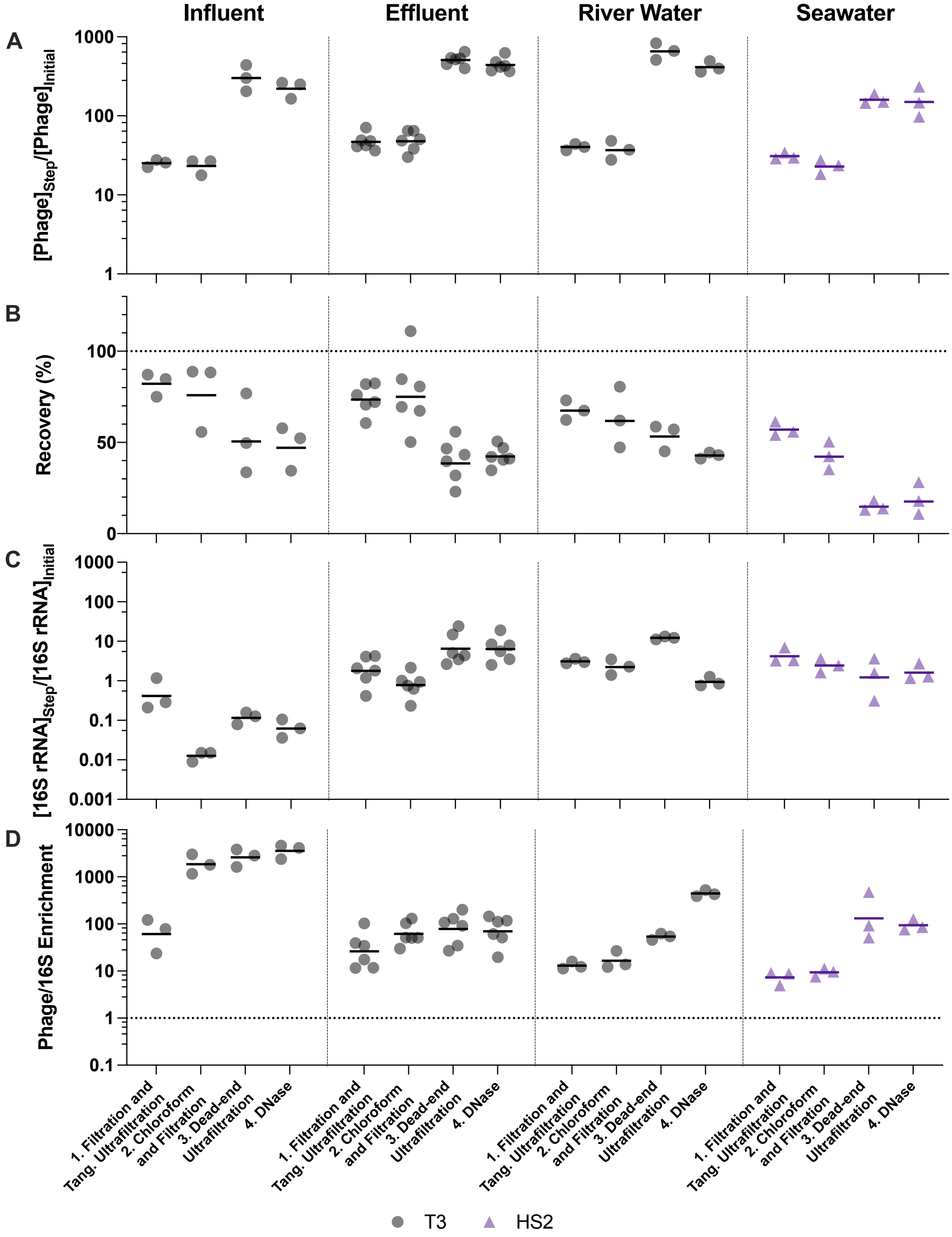

Figure 2.4: Stepwise evaluation of ultrafiltration and purification with approximately 500-fold concentration by volume of influent and 1,000-fold concentration by volume of all other matrices. (A) T3 or HS2 concentration factors on a log-scale at each step. Values greater than 1 indicated an increase in viral genomes. (B) T3 and HS2

recovery with each step of ultrafiltration showed losses of viruses throughout the process. Ideal recovery is indicated by the dotted line at 100%. (C) 16S rRNA concentration factor on a log-scale indicated non-viral DNA removal. Concentration factor values less than 1 indicated reduction of 16S rRNA gene copies. (D) T3 or HS2 to 16S rRNA enrichment with each step on a log-scale, as calculated by the virus concentration factor divided by the 16S rRNA concentration factor. Enrichments greater than 1 demonstrate viral genomes were concentrated more than 16S rRNA gene copies during the process. Individual measurements are points with a bar indicating the geometric mean of experimental replicates.

2.5.3 Non-viral DNA removal performance

Filtration, chloroform, and DNase treatments are purification steps that aim to enrich the viral genomes relative to other organisms' genomes in order to focus sequencing effort on viral DNA. Overall, greater than 98% and 97% of 16S rRNA were removed from all matrices, following ultrafiltration and purification and iron chloride flocculation and purification, respectively (Figure A.3). Ultimately, selectively concentrating viruses compared to non-viral DNA is most important. In every case, viruses were concentrated by a greater factor than 16S rRNA (Table 2.3). This relationship was evaluated using the virus to 16S rRNA enrichment factor, whereby an enrichment factor greater than one indicated that viruses were concentrated more than 16S rRNA gene copies. Virus to 16S rRNA enrichment factors were 1000, 44, 19, and 8.8 after iron flocculation and purification and 3600, 70, 440, and 94 after ultrafiltration and purification for influent, effluent, river water, and seawater, respectively (Table A.16). Virus to 16S rRNA enrichment factors for both methods were statistically significantly greater than one, except after iron chloride flocculation and purification with river water (Table A.16). Of the different sample types, the influent samples resulted in the lowest 16S rRNA concentration factors and highest virus to 16S rRNA enrichment factors with both methods, likely due to the high initial concentrations of 16S rRNA in influent relative to the other sample matrices (Figure A.3).

The purification steps are designed to enrich viral genomes. However, viral genomes in many sample types were not enriched by chloroform and DNase purification steps (Figure 2.3D and Figure 2.4D). During ultrafiltration and purification, enrichment of T3 relative to 16S rRNA was observed only in the influent samples following chloroform treatment and in the river water sample after DNase treatment. During iron chloride flocculation and purification, the only sample in which enrichment of T3 relative to 16S rRNA was observed was the seawater sample. Interestingly, the seawater samples that were concentrated with iron chloride flocculation and purification exhibited HS2 to 16S rRNA enrichment factors greater than one *only* following DNase treatment. This suggests that viral DNA was purified in this matrix by DNase or a combination of chloroform and DNase. With both concentration approaches, most of the 16S rRNA in freshwater was removed after the first step of pre-filtering (0.45-μm) and concentrating. This was likely because the prefiltration step removes a large fraction of the cells.

2.5.4 Final DNA concentrations and fragmentation

Sequencing a sample requires a minimum quantity of DNA. Specifically, Illumina sequencing typically requires ~200 ng of DNA and Oxford Nanopore flow cells require ~1 μg of high molecular weight DNA. DNA yields ranged from 160 ng to 520 ng and 430 ng to 16 μg for iron chloride flocculation and ultrafiltration approaches, respectively (Figure A.4). These ranges are sufficient for generating amplification-free dsDNA and ssDNA virome libraries for Illumina sequencing. Following ultrafiltration and purification, secondary effluent yields were sufficient for Oxford Nanopore sequencing, but other matrices would require multiplexing with one or two additional samples per flow cell to be sufficient. An additional 10-fold virus DNA concentration would be necessary for amplification-free long read sequencing after applying the iron flocculation approach.

To evaluate whether final DNA extracts contained high molecular weight DNA, fragmentation was assessed with gel electrophoresis. Following the iron chloride flocculation method, DNA from freshwater samples resulted in faint streaks with darker regions below the 10 kb rung of the high range ladder and seawater samples had no visible DNA (Figure A.5). No T3 or HS2 genome bands were visible, suggesting the viral DNA had been sheared. All of the ultrafiltration and purification samples had clear high molecular weight DNA streaks between 10 and 50 kb with visible bands at the T3 or HS2 genome size (38 kb) indicating the genomes were not fragmented (Figure A.5). Together, the DNA concentration and fragmentation results suggest that the ultrafiltration and purification method may be better suited for Nanopore sequencing than the iron chloride flocculation and purification method.

2.5.5 Phage-specific recovery through concentration and purification

We expanded the number of phages tested under select experimental conditions to better understand the degree to which T3 and HS2 results were representative of other DNA viruses. As ultrafiltration and purification resulted in the greater viral concentration in freshwater, we applied this method to determine the concentration of two other phage types, dsDNA phage T4 and ssDNA phage PhiX174, in a freshwater secondary effluent sample (Table 2.2). As both the ultrafiltration and iron chloride flocculation methods provided equivalently effective viral concentration, we tested both methods with dsDNA phage HM1 and ssDNA phage ICBM5, in seawater.

Viral genome recoveries varied amongst the phages with both concentration methods. The iron chloride flocculation method applied to seawater resulted in genome recoveries of 15, 28, and 2.2% for HS2, HM1, and ICBM5, respectively. HM1 recovery was significantly greater than HS2 (*p*-value = 0.011) and ICBM5 (*p*-value = 1.9E-5) and HS2 recovery was significantly greater than ICBM5 (*p*-value = 0.021) (Figure 2.5). HM1 and ICBM5 genome losses occurred at the same steps as the HS2 genome losses, namely following iron chloride flocculation and

DNase treatment. The ultrafiltration and purification method applied to the secondary effluent resulted in genome recoveries of 42, 17, and 3.8% for T3, T4, and PhiX174, respectively, whereas T3 recovery was significantly greater than both T4 (*p*-value = 0.0058) and PhiX174 (*p*-value = 0.012). The ultrafiltration and purification method applied to the seawater resulted in genome recoveries of 18, 30, and 13% for HS2, HM1, and ICBM5, whereas HM1 recovery was significantly greater than HS2 (*p*-value = 0.083) and ICBM5 (*p*-value = 0.0098). The ssDNA phages, PhiX174 and ICBM5, had the lowest genome recovery of the spiked viruses after ultrafiltration and purification. As seen in the T3- and HS2-only experiments, the ultrafiltration steps resulted in the largest viral genome losses across the entire ultrafiltration and purification method. Following iron chloride flocculation, significant reductions in recovery after DNase treatment were common regardless of the virus or matrix (Table A.14). Conversely, only ICBM5 recovery was significantly reduced by the DNase treatment following ultrafiltration (*p*-value = 0.0060).

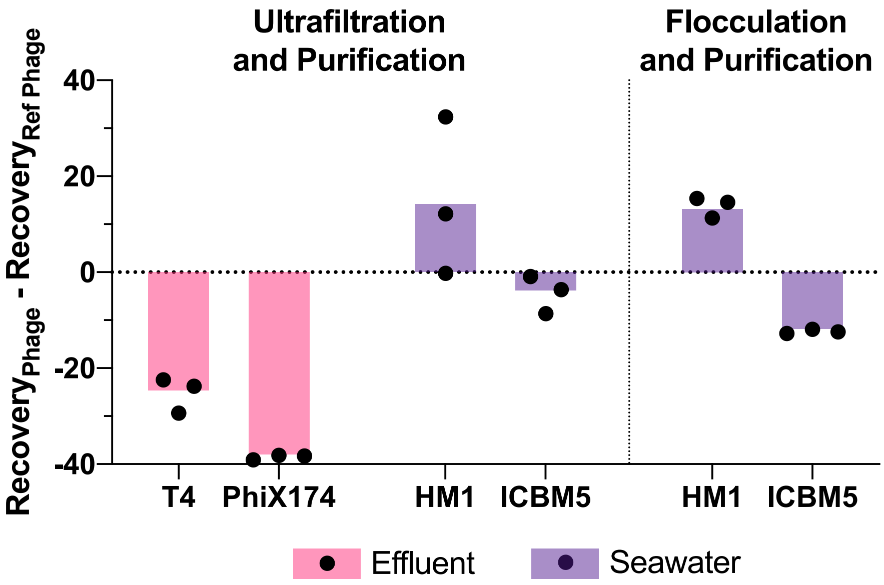

Figure 2.5: Variability in virus recovery depending on the phage was observed. T3 in effluent and HS2 in seawater recoveries were used to evaluate the final genome recoveries of different phage types. The individual points are the difference between a final recovery individual measurement and the geometric mean final recovery of the reference

phage in experimental replicates. In effluent, T4 and PhiX174 were com- pared to T3. In seawater, HM1 and ICBM5 were compared to HS2. The bars represent the mean of the individual points.

2.6 Discussion

We compared two approaches for concentrating and purifying viruses, iron chloride flocculation and ultrafiltration, and applied them to four water matrices for the preparation of high-quality viral DNA extracts for metagenomics. Both concentration methods are widely used for virome studies[7, 38, 40-42, 59-66], but there has not been a systematic study evaluating stepwise performance and viral recovery through the concentration and purification processes across multiple matrices. As our end goal was the preparation of amplification-free sequencing libraries suitable for both short and long read sequencing, the ideal preparation method would generate sufficient mass of high molecular weight DNA for these purposes. As such, we evaluated iron chloride flocculation and ultrafiltration in terms of the resulting DNA concentration, purity, fragmentation, and the relative number of viral to cellular genome copies in the four different sample matrices.

2.6.1 Five-fold more iron is needed to flocculate viruses from freshwater than seawater

Iron chloride flocculates viruses because it neutralizes the electrostatic repulsion layer of negatively charged particles. The behavior of this process is influenced by pH, the abundance of particles that vary in surface charge, and electrolyte strength, such as salinity[67]. Due to the dependence of flocculation efficiency on the relationship between sample matrix properties and flocculant concentration, we both expected and observed iron chloride flocculation performance to be strongly matrix-specific. The best iron chloride concentration for the flocculation and removal of viruses from freshwater matrices was 25 mg Fe L^{-1}. In contrast, seawater required less iron chloride, attaining equally high removal at concentrations at 5, 10, and 25 mg Fe L^{-1}.

The concentrations required for maximal removal are higher than previously reported for both freshwater[68, 69] and seawater[16] when a similar filtration-based approach was applied to capture flocs. This may be due to the fact that these previous studies only tested concentrations up to 10 and 1 mg Fe L^{-1} for freshwater and seawater, respectively. Supporting the finding that less Fe was needed for flocculation in seawater, a recent estuary study demonstrated that increased iron flocculation occurred with increasing salinity[67]. As more viral ecologists apply the commonly used iron chloride flocculation method first developed in seawater [16] to freshwater samples, it is important to note that the best viral genome recovery in the freshwater samples was achieved at 25 mg Fe L^{-1}, 25-fold higher than the standard concentration used for ocean samples.

Iron chloride flocculation performance varied amongst the freshwater matrices (virus recoveries of 44-74%). The pH values of the different matrices were similar (range: 6.89-8.05, Table A.1) and the T3 and HS2 have negative surface charges in this pH range[70, 71]. Consequently, the different flocculant concentrations necessary for optimized virus recoveries in the freshwater matrices were likely due to other water characteristics. For instance, viral genome recoveries in the freshwater jar tests decreased with increasing total solids content in the samples (Figure 2.2, Table A.1). Previous studies identified that flocculation performance decreased as solids content increased and additional iron was required for sufficient viral flocculation[68, 69, 72].

Our application of the jar test method, an elementary technique used to optimize flocculant concentrations in environmental engineering applications, emphasized the matrix-specific nature of iron chloride flocculation performance. We recommend these performance tests before applying iron chloride flocculation to recover viruses from novel matrices, especially when salinity and solids content are expected to differ from those tested here.

2.6.2 Ultrafiltration and purification was best at concentrating viruses and preserving their genome integrity

Preparation of PCR-free sequencing libraries requires high masses of DNA (e.g., 200 ng for Illumina and 1 µg for Nanopore). Due to the low abundance of viral DNA relative to total community DNA in aquatic samples, the recovery of high viral DNA masses requires a concomitant significant reduction of water volumes. We concentrated seawater, river water, and secondary effluent approximately 1000-fold and influent approximately 500-fold during ultrafiltration and purification to achieve sufficient DNA for sequencing. During the iron chloride flocculation process, only 40-fold concentration by volume of influent, effluent, and river water and 200-fold volume concentration for seawater was achieved. The iron chloride flocculation method is limited in its ability to concentrate viruses. Increasing the sample volume requires the addition of more iron chloride, which subsequently increases the required volume of resuspension buffer. Therefore, unlike with ultrafiltration, increasing sample volume does not ultimately increase the virus concentration factor.

Due to the ability of ultrafiltration to reduce sample volume more than iron chloride flocculation, we anticipated virus concentration factors to be larger after ultrafiltration and purification. In freshwater matrices, the ultrafiltration method concentrated T3 genomes 29, 66, and 54-fold more than iron chloride flocculation in influent, effluent, and river water, respectively. In seawater, the concentration factors obtained with the two methods were not different (6-fold more with ultrafiltration than with iron chloride flocculation). The limited ability for iron chloride flocculation to concentrate viruses in freshwater samples makes it a poorer choice for viral metagenomics applications, as the resultant DNA concentrations are too low to generate sequencing libraries without applying additional steps. Two steps commonly applied to iron chloride flocculated samples include an additional concentration step, such as dead-end ultrafiltration[7, 40, 64, 73] and either PCR or enzymatic amplification of the genomic material[74-78]. Our results indicate that, given its ability to generate sufficient viral DNA for amplification-free sequencing, ultrafiltration and purification is the more suitable method for quantitative virome studies, where biases in sequence representation must be minimized.

Long-read data can capture complete viral genomes with single reads[39], overcoming the challenges of assembling viral genomes from short-read data. Long-read sequencing, however, requires a large mass of high integrity, high molecular weight DNA. Our results suggest that the iron chloride flocculation concentration and purification method caused more DNA fragmentation than the ultrafiltration concentration and purification method. Two ocean virome studies have successfully applied long-read sequencing to obtain reads that were several kilobases long, with median lengths of 30 kb[39] and 4 kb[73]. Beaulaurier et al. (2020) successfully applied microfiltration followed by tangential ultrafiltration to generate high masses of high molecular weight DNA for amplification-free Nanopore libraries. Whereas, Warwick et al. (2019) used iron chloride flocculation followed by dead-end ultrafiltration. This study used 100 ng of input DNA and required PCR-adaptor ligation amplification, which reduced their read length potential to less than 8 kb. Our findings are consistent with these limited studies, namely that ultrafiltration concentration approaches produce greater viral concentration factors and higher quality DNA than iron chloride flocculation and are thus more suitable for long-read sequencing applications.

2.6.3 Viral DNA is enriched more by ultrafiltration than iron chloride flocculation, while 16S rRNA is removed equally well during purification

Minimizing non-viral DNA contamination focuses sequencing effort on viral DNA and facilitates the capture of low abundance viruses. Non-viral DNA contamination was evaluated by the 16S rRNA gene concentration factor and the virus to 16S rRNA enrichment. 18S rRNA gene concentrations were not assessed in this study, as previous studies have demonstrated greater contamination of prokaryotic DNA in viromes as compared to eukaryotic DNA[25, 34, 42]. Both methods performed similarly at removing 16S rRNA gene copies, but the greater ability to concentrate viruses with the ultrafiltration steps resulted in a greater enrichment of viral genomes relative to 16S rRNA gene copies. The similar final 16S rRNA removals, 97-99% (Figure A.3), indicated we had reached a threshold for 16S rRNA removal. The remaining 16S rRNA gene copies may be encapsidated in gene transfer agents[79] or otherwise protected from removal by the chloroform and DNase purification processes.

2.6.4 Iron chloride flocculation disrupted inhibition of DNase reactions causing viral genome loss

DNase treatment was found to impact virus recovery after iron chloride flocculation, but not after ultrafiltration. This suggests a mechanism whereby the iron chloride flocculation and purification method increased viral genome susceptibility to DNase enzymes. DNase activity requires calcium and magnesium ions; the DNase activity is therefore inhibited prior to genome extraction by adding EDTA and EGTA to chelate calcium and magnesium ions. The high level of Fe^{3+} in these samples may have reduced the effectiveness of the EDTA at quenching the DNase activity prior to DNA extraction, thus leading to viral DNA degradation. Alternative viral purification methods to DNase or other approaches to DNase inhibition may improve viral retention during purification following iron chloride flocculation.

2.6.5 Phage-specific recovery in both ultrafiltration and iron chloride flocculation is biased against ssDNA viruses

Significant phage-specific genome recoveries were observed with both methods. Regardless of the concentration method, genomes of the ssDNA phages, PhiX174 and ICBM5, were recovered less than those of the dsDNA phage genomes at the end of the concentration and purification processes. For all matrices concentrated with ultrafiltration, no single step in the method resulted in statistically significantly higher losses of ICBM5 or PhiX174, as compared to losses of the dsDNA viruses. However, by the end of the ultrafiltration method, the ssDNA virus recoveries were statistically significantly lower than those of dsDNA viruses. In addition to having ssDNA genomes, PhiX174 and ICBM5 are smaller in diameter and have shorter genomes than the other viruses used in this study (Table 2.1). Previous studies have shown that smaller phages have poorer particle[18] and genome[34] recoveries than larger dsDNA phages. Our observed differences in final recoveries demonstrated a source of bias that could impact downstream viral community representations in virome data. In the case of processing seawater, we observed greater variance in final virus recoveries with iron chloride flocculation and purification than with ultrafiltration and purification. Previously, Hurwitz et al. (2013) compared the concentration of seawater viruses with tangential ultrafiltration and iron chloride flocculation, and then evaluated the viromes produced by each protocol. Viromes prepared from iron chloride flocculation had more viral reads relative to non-viral and captured more rare viral reads than observed using tangential ultrafiltration. It was concluded that iron chloride flocculation introduced fewer virus-specific biases than tangential ultrafiltration, in contrast with our results. Given the different downstream purification methods applied (cesium chloride density gradient [25] versus chloroform/filtration and DNase), direct comparisons of our results were not possible.

To resolve this uncertainty, additional work with an expanded set of dsDNA and ssDNA viruses will be needed to conclusively evaluate the relative impact of iron chloride flocculation versus ultrafiltration on the diversity of recovered viruses.

2.6.6 Limitations of phage spike-and-recovery experiments, selected matrices, and focus on DNA viruses

The spiked phages did not represent all known virus diversity. We selected the six tested phages to span a range of characteristics common to aquatic viruses, such as genome length, particle size, icosahedral capsid shapes and tails (Table 2.1). Although eukaryotic viruses were absent, phages are commonly accepted as surrogates for eukaryotic viruses in spike-and-recovery experiments[19, 27, 80]. Enveloped, double jelly roll capsid, and filamentous phages were absent from our spiked viruses, which has implications for the generalizability of our findings. Although enveloped viruses are considered a minor fraction of the known aquatic viromes[18, 81, 82], inovirus filamentous phages were recently deemed more widespread and pervasive in the environment than previously thought[83]. Enveloped, double jelly roll capsid, and filamentous viruses are known to lose infectivity when exposed to chloroform[32, 84, 85], but this does not absolutely render their genomes susceptible to DNase enzymes that follow. Enveloped herpes virus genome recovery was not impacted by chloroform treatment[30, 86], whereas the recoveries of enveloped coronavirus and mimivirus genomes, and those of enveloped reverse-transcribing viruses broadly, have been found to decrease following chloroform treatment[86, 87]. Although chloroform introduces biases, it facilitates the removal of 16S rRNA gene copies in subsequent DNase steps (Section A.4). Further investigation of the impact of chloroform treatment on viromes is needed to weigh the benefits (less cellular DNA, more low abundance viruses, lower limits of virus detection and quantification) versus drawbacks (possible biases in viral community representation) of this purification step.

All samples tested here fell within a narrow pH range (6.9-8.2), and some waters of interest may be outside of this range (e.g., acid mine drainage, alkaline lakes, treated drinking water). The pH of the water matrix impacts the overall charge of viruses and this may affect virus recoveries in the tested methods. Virus isoelectric points are typically less than seven[71, 88], although some have values as high as 8.4[71]. Viruses with isoelectric points greater than the pH range of our samples were not used in spike-and-recovery experiments. Since pH will change viral surface charges and the charge of the dominant iron species, applying iron chloride flocculation to matrices that have pH values deviating greatly from circumneutral may affect virus recoveries. An expanded set of viruses and water samples should be tested in the future to assess the impact of pH on virus recoveries.

This work focuses on methods that recover and purify DNA viruses for preparing viromes from seawater and freshwater samples. RNA viruses are also important members of viral communities, estimated to represent half of ocean viruses[89]. Recent work has also expanded our knowledge of RNA phages in seawater[90], though less work has been done on the prevalence and diversity of freshwater RNA viruses. Methods designed for effectively concentrating and purifying DNA viral genomes may not be as effective for RNA viruses. The necessary reverse transcription step to form cDNA and the need to deplete host RNA, for example, add unique challenges in preparing unbiased RNA viral communities from sequencing[91]. The unknown relative abundance of RNA viruses in water samples further complicates RNA viral library preparation. Future work is therefore necessary to assess which methods work best for capturing

RNA viral metagenomes, as well as the impact of initial relative abundance on the recovery of a given virus type through concentration and purification processes.

2.7 Conclusions

We demonstrated the importance of assessing viral genome recovery and non-viral DNA removal prior to sequencing. Differences in aquatic matrices alter concentration and purification performances. Assuming a method performs adequately across matrices is inadvisable, as evidenced by the 5-fold differences in best iron chloride concentrations for the freshwater and seawater samples. The ultrafiltration and purification method resulted in higher virus concentration factors and higher concentrations of high molecular weight DNA than iron chloride flocculation and purification for all tested matrices. We demonstrated that our ultrafiltration and purification protocol was superior to iron chloride flocculation and purification for influent, effluent, river water, and seawater samples. Given the demonstrated impact of solids content and salinity on the performance of these concentration and purification methods, we encourage future virome studies with matrices not tested here to assess virus concentration factors, recovery, and non-viral DNA removal with spike-and-recovery tests prior to sample preparation.

2.8 References

1. Cael, B.B., Carlson, M.C.G., Follett, C.L. & Follows, M.J. Marine Virus-Like Particles and Microbes: A Linear Interpretation. *Front Microbiol* **9**, 358 (2018).
2. Parsons, R.J., Breitbart, M., Lomas, M.W. & Carlson, C.A. Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea. *ISME J* **6**, 273-284 (2012).
3. Wigington, C.H. et al. Re-examination of the relationship between marine virus and microbial cell abundances. *Nat Microbiol* **1**, 15024 (2016).
4. Abedon, S.T. Bacteriophage ecology: population growth, evolution, and impact of bacterial viruses, Edn. 15th. (Cambridge Univeristy Press, 2008).
5. Breitbart, M. Marine viruses: truth or dare. *Ann Rev Mar Sci* **4**, 425-448 (2012).

6. Dion, M.B., Oechslin, F. & Moineau, S. Phage diversity, genomics and phylogeny. *Nat Rev Microbiol* **18**, 125-138 (2020).
7. Brum, J. et al. Patterns and ecological drivers of ocean viral communities. *Science* **348** (2015).
8. Dutilh, B.E. et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun* **5**, 4498 (2014).
9. Koonin, E.V. & Yutin, N. The crAss-like Phage Group: How Metagenomics Reshaped the Human Virome. *Trends Microbiol* **28**, 349-359 (2020).
10. Roux, S. et al. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* **4**, e2777 (2016).
11. Duhaime, M.B. & Sullivan, M.B. Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology* **434**, 181-186 (2012).
12. Mahmoudabadi, G. & Phillips, R. A comprehensive and quantitative exploration of thousands of viral genomes. *Elife* **7** (2018).
13. Brinkman, N.E., Villegas, E.N., Garland, J.L. & Keely, S.P. Reducing inherent biases introduced during DNA viral metagenome analyses of municipal wastewater. *PLoS One* **13**, e0195350 (2018).
14. Calgua, B. et al. New methods for the concentration of viruses from urban sewage using quantitative PCR. *J Virol Methods* **187**, 215-221 (2013).
15. Hjelmso, M.H. et al. Evaluation of Methods for the Concentration and Extraction of Viruses from Sewage in the Context of Metagenomic Sequencing. *PLoS One* **12**, e0170199 (2017).
16. John, S.G. et al. A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ Microbiol Rep* **3**, 195-202 (2011).
17. Randazzo, W. et al. Interlaboratory Comparative Study to Detect Potentially Infectious Human Enteric Viruses in Influent and Effluent Waters. *Food Environ Virol* **11**, 350-363 (2019).
18. Ye, Y., Ellenberg, R.M., Graham, K.E. & Wigginton, K.R. Survivability, Partitioning, and Recovery of Enveloped Viruses in Untreated Municipal Wastewater. *Environ Sci Technol* **50**, 5077-5085 (2016).
19. Zhang, Y., Riley, L.K., Lin, M., Purdy, G.A. & Hu, Z. Development of a virus concentration method using lanthanum-based chemical flocculation coupled with modified membrane filtration procedures. *J Virol Methods* **190**, 41-48 (2013).
20. Kunze, A., Pei, L., Elsasser, D., Niessner, R. & Seidel, M. High performance concentration method for viruses in drinking water. *J Virol Methods* **222**, 132-137 (2015).
21. Millen, H.T. et al. Glass wool filters for concentrating waterborne viruses and agricultural zoonotic pathogens. *J Vis Exp*, e3930 (2012).
22. Shi, H., Xagoraraki, I., Parent, K.N., Bruening, M.L. & Tarabara, V.V. Elution Is a Critical Step for Recovering Human Adenovirus 40 from Tap Water and Surface Water by Cross-Flow Ultrafiltration. *Appl Environ Microbiol* **82**, 4982-4993 (2016).
23. Gallardo, V.J., Morris, B.J. & Rhodes, E.R. The use of hollow fiber dialysis filters operated in axial flow mode for recovery of microorganisms in large volume water samples with high loadings of particulate matter. *J Microbiol Methods* **160**, 143-153 (2019).

24. Hill, V.R. et al. Development of a rapid method for simultaneous recovery of diverse microbes in drinking water by ultrafiltration with sodium polyphosphate and surfactants. *Appl Environ Microbiol* **71**, 6878-6884 (2005).
25. Hurwitz, B.L., Deng, L., Poulos, B.T. & Sullivan, M.B. Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ Microbiol* **15**, 1428-1440 (2013).
26. Pei, L. et al. Combination of crossflow ultrafiltration, monolithic affinity filtration, and quantitative reverse transcriptase PCR for rapid concentration and quantification of model viruses in water. *Environ Sci Technol* **46**, 10073-10080 (2012).
27. Rhodes, E.R., Huff, E.M., Hamilton, D.W. & Jones, J.L. The evaluation of hollow-fiber ultrafiltration and celite concentration of enteroviruses, adenoviruses and bacteriophage from different water matrices. *J Virol Methods* **228**, 31-38 (2016).
28. Smith, C.M. & Hill, V.R. Dead-end hollow-fiber ultrafiltration for recovery of diverse microbes from water. *Appl Environ Microbiol* **75**, 5284-5289 (2009).
29. Thurber, R.V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nat Protoc* **4**, 470-483 (2009).
30. Breitbart, M. & Rohwer, F. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. *BioTechniques* **39**, 729-736 (2005).
31. Hannigan, G., Duhaime, M.B., Ruffin, M., Kaumpouras, C. & Schloss, P. Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome. *mBio* **9** (2018).
32. Kauffman, K.M. et al. A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature* **554**, 118-122 (2018).
33. Maruyama, A., Oda, M. & Higashihara, T. Abundance of Virus-Sized Non-DNase-Digestible DNA (Coated DNA) in Eutrophic Seawater. *Applied and Environmental Microbiology* **59**, 712-717 (1993).
34. Kleiner, M., Hooper, L.V. & Duerkop, B.A. Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics* **16**, 7 (2015).
35. Trubl, G. et al. Towards optimized viral metagenomes for double-stranded and single-stranded DNA viruses from challenging soils. *PeerJ* **7**, e7265 (2019).
36. Uyaguari-Diaz, M.I. et al. A comprehensive method for amplicon-based and metagenomic characterization of viruses, bacteria, and eukaryotes in freshwater samples. *Microbiome* **4**, 20 (2016).
37. Wang, H. et al. Variations among Viruses in Influent Water and Effluent Water at a Wastewater Plant over One Year as Assessed by Quantitative PCR and Metagenomics. *Appl Environ Microbiol* **86** (2020).
38. Aguirre de Carcer, D., Lopez-Bueno, A., Pearce, D. & Alcami, A. Biodiversity and distribution of polar freshwater DNA viruses. *Science Advances* **1** (2015).
39. Beaulaurier, J. et al. Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. *Genome Research* (2020).
40. Gregory, A.C. et al. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109-1123 e1114 (2019).
41. Moon, K. et al. Freshwater viral metagenome reveals novel and functional phage-borne antibiotic resistance genes. *Microbiome* **8**, 75 (2020).
42. Moon, K., Kim, S., Kang, I. & Cho, J.C. Viral metagenomes of Lake Soyang, the largest freshwater lake in South Korea. *Sci Data* **7**, 349 (2020).

43. E.W. Rice, R.B.B., A.D. Eaton Standard Methods for the Examination of Water and Wastewater, Edn. 23. (American Public Health Association, American Water Works Association, Water Environment Federation, Washington, D.C.; 2017).
44. Duhaime, M.B. et al. Comparative Omics and Trait Analyses of Marine Pseudoalteromonas Phages Advance the Phage OTU Concept. *Front Microbiol* **8**, 1241 (2017).
45. Leiman, P.G., Chipman, P.R., Kostyuchenko, V.A., Mesyanzhinov, V.V. & Rossmann, M.G. Three-dimensional rearrangement of proteins in the tail of bacteriophage T4 on infection of its host. *Cell* **118**, 419-429 (2004).
46. Matsuo-Kato, H., Fujisawa, H. & Minagawa, T. Structure and Assembly of Bacteriophage T3 Tails. *Virology* **109**, 157-`164 (1981).
47. John, S., Poulos, B. & Schirmer, C. (protocols.io, 2015).
48. John, S.G., Poulos, B. & Schirmer, C. (protocols.io; 2015).
49. Schultz, F. et al. Giant viruses with an expanded complement of translation system components. *Science* **356**, 82-85 (2017).
50. Uchiyama, J. et al. Intragenus generalized transduction in Staphylococcus spp. by a novel giant phage. *ISME J* **8**, 1949-1952 (2014).
51. Ho, C. & Zydney, A. Effect of membrane morphology on the initial rate of protein fouling duirng microfiltration. *Journal of Membrane Science* **155**, 261-275 (1999).
52. Thornton, J. (protocols.io, 2015).
53. Nadkarni, M.A., Martin, F.E., Jacques, N.A. & Hunter, N. Determination of bacterial load by real-time PCR using a broad-range (universal) probe and primers set. *Microbiology* **148**, 257-266 (2002).
54. Lievens, A., Jacchia, S., Kagkli, D., Savini, C. & Querci, M. Measuring Digital PCR Quality: Performance Parameters and Their Optimization. *PLoS One* **11**, e0153317 (2016).
55. Dickson, R.P. et al. The Lung Microbiota of Healthy Mice Are Highly Variable, Cluster by Environment, and Reflect Variation in Baseline Lung Innate Immunity. *Am J Respir Crit Care Med* **198**, 497-508 (2018).
56. Rehbinder, E.M. et al. Is amniotic fluid of women with uncomplicated term pregnancies free of bacteria? *Am J Obstet Gynecol* **219**, 289 e281-289 e212 (2018).
57. Sze, M.A., Abbasi, M., Hogg, J.C. & Sin, D.D. A comparison between droplet digital and quantitative PCR in the analysis of bacterial 16S load in lung tissue samples from control and COPD GOLD 2. *PLoS One* **9**, e110351 (2014).
58. Huggett, J.F. et al. The digital MIQE guidelines: Minimum Information for Publication of Quantitative Digital PCR Experiments. *Clin Chem* **59**, 892-902 (2013).
59. Beaulaurier, J. et al. Assembly-free single-molecule nanopore sequencing recovers complete virus genomes from natural microbial communities. *bioRxiv* (2019).
60. Breitbart, M. et al. Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* **185**, 6220-6223 (2003).
61. Breitbart, M. et al. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* **99**, 14250-14255 (2002).
62. Mohiuddin, M. & Schellhorn, H.E. Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Front Microbiol* **6**, 960 (2015).

63. Petrovich, M.L. et al. Microbial and Viral Communities and Their Antibiotic Resistance Genes Throughout a Hospital Wastewater Treatment System. *Front Microbiol* **11**, 153 (2020).

64. Roux, S. et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689-693 (2016).

65. Roux, S. et al. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* **7**, e33641 (2012).

66. Skvortsov, T. et al. Metagenomic Characterisation of the Viral Community of Lough Neagh, the Largest Freshwater Lake in Ireland. *PLoS One* **11**, e0150361 (2016).

67. Jilbert, T. et al. Impacts of flocculation on the distribution and diagenesis of iron in boreal estuarine sediments. *Biogeosciences* **15**, 1243-1271 (2018).

68. Chang, S., Stevenson, R., Bryant, A., Woodward, R. & Kabler, P. Removal of coxsackie and bacterial viruses in water by flocculation. *American Journal of Public Health* **48**, 159-169 (1958).

69. Manwaring, J., Chaudhuri, M. & Engelbrecht, R. Removal of viruses by coagulation and flocculation. *Journal American Water Works Association* **63**, 298-300 (1971).

70. Ghanem, N. et al. Marine Phages As Tracers: Effects of Size, Morphology, and Physico-Chemical Surface Properties on Transport in a Porous Medium. *Environ Sci Technol* **50**, 12816-12824 (2016).

71. Michen, B. & Graule, T. Isoelectric points of viruses. *J Appl Microbiol* **109**, 388-397 (2010).

72. Leiknes, T. The effect of coupling coagulation and flocculation with membrane filtration in water treatment: A review. *Journal of Environmental Sciences* **21**, 8-12 (2009).

73. Warwick-Dugdale, J. et al. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ* **7**, e6800 (2019).

74. Kim, K.H. & Bae, J.W. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl Environ Microbiol* **77**, 7663-7668 (2011).

75. Kim, K.H. et al. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Appl Environ Microbiol* **74**, 5975-5985 (2008).

76. Laver, T.W. et al. Pitfalls of haplotype phasing from amplicon-based long-read sequencing. *Sci Rep* **6**, 21746 (2016).

77. Woyke, T. et al. Assembling the marine metagenome, one cell at a time. *PLoS One* **4**, e5299 (2009).

78. Zhang, K. et al. Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol* **24**, 680-686 (2006).

79. Lang, A.S., Zhaxybayeva, O. & Beatty, J.T. Gene transfer agents: phage-like elements of genetic exchange. *Nat Rev Microbiol* **10**, 472-482 (2012).

80. Pecson, B.M., Ackermann, M. & Kohn, T. Framework for using quantitative PCR as a nonculture based method to estimate virus infectivity. *Environ Sci Technol* **45**, 2257-2263 (2011).

81. Casanova, L.M. & Weaver, S.R. Inactivation of an Enveloped Surrogate Virus in Human Sewage. *Environmental Science & Technology Letters* **2**, 76-78 (2015).

82. Gundy, P.M., Gerba, C.P. & Pepper, I.L. Survival of Coronaviruses in Water and Wastewater. *Food and Environmental Virology* **1**, 10-14 (2008).

83. Roux, S. et al. Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nat Microbiol* **4**, 1895-1906 (2019).

84. Petrenko, V., Smith, G., Gong, X. & Quinn, T. A library of organic landscapes on filamentous phage. *Protein Engineering* **9**, 797-801 (1996).

85. Taniguchi, H., Sato, K., Ogawa, M., Udou, T. & Mizuguchi, Y. Isolation and characterization of a filamentous phage, Vf33, specific for Vibrio parahaemolyticus. *Microbiology and immunology* **28**, 327-337 (1984).

86. Conceicao-Neto, N. et al. Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci Rep* **5**, 16532 (2015).

87. Weynberg, K.D., Wood-Charlson, E.M., Suttle, C.A. & van Oppen, M.J. Generating viral metagenomes from the coral holobiont. *Front Microbiol* **5**, 206 (2014).

88. Mayer, B.K., Yang, Y., Gerrity, D.W. & Abbaszadegan, M. The Impact of Capsid Proteins on Virus Removal and Inactivation during Water Treatment Processes. *Microbiology Insights* **8s2** (2015).

89. Steward, G.F. et al. Are we missing half of the viruses in the ocean? *ISME J* **7**, 672-679 (2013).

90. Wolf, Y.I. et al. Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nat Microbiol* **5**, 1262-1270 (2020).

91. Hong, P.Y., Mantilla-Calderon, D. & Wang, C. Metagenomics as a tool to monitor reclaimed-water quality. *Applied and Environmental Microbiology Minireview* **86** (2020).

Chapter 3 Evaluating Limitations of Quantitative Metagenomics with Synthetic dsDNA and ssDNA Standards

3.1 Abstract

Metagenomic data is inherently relative which limits the potential scope for studying microbial communities. Here, we present QuantMeta, a quantitative metagenomic tool, that calculates concentrations of targets and accounts for detection and quantification limitations. We developed a set of synthetic ssDNA standards to compliment a previously developed set of dsDNA standards for application to dsDNA and ssDNA viromes. The methods were applied to concentrated and purified wastewater influent and effluent virus communities. We probed target-specific limitations of detection and quantification to establish detection thresholds and quantification requirements. We developed a method to detect non-specific mapping and assembly errors by establishing read depth variability thresholds based on the observed read depth variability across standard sequences. When the method was applied to wastewater viromes, target concentrations were comparable to those measured with ddPCR. Furthermore, we determined that read-based and contig-based quantification resulted in statistically similar concentrations. While this method was tested on wastewater viromes, the results are applicable to other types of metagenomes and environments.

3.2 Key Terminology

Coverage: Fraction of the reference gene or genome with at least one read mapped onto it, i.e., what fraction of the reference sequence is covered by mapped reads.

Distribution: Number of reads mapped to each basepair of a target (e.g., contig or reference gene or genome), i.e., how the mapped reads are distributed on a target, also “read depth”.

3.3 Introduction

Metagenomics has allowed unprecedented insight into the diversity of microbial communities. However, metagenomic data is inherently relative, complicating comparisons between samples where it is valuable to understand variations in population concentrations in response to changing conditions. Virome studies are usually limited to relative abundance data[1-6]. Previous virome studies have uncovered the diversity, host ranges, and infection dynamics of viruses in many environments that impact the structure and function of microbiomes and biogeochemical cycling[7, 8]. However, relative abundance data limits direct sample-to-sample comparisons as absolute abundances of viruses are unknown in each sample. When absolute abundances are necessary, quantitative PCR is traditionally employed, but limited to specific targets. Several methods have been introduced to make metagenomes quantitative.

Previous methods to normalize metagenomic data for quantitative comparisons between samples include spiking-in synthetic DNA[9, 10] or genomic DNA at known concentrations[11-14] or normalizing by total cell estimates based on flow cytometry or 16S rRNA or housekeeping gene copies[15-18]. These methods have not been validated for viral metagenomes (i.e., viromes) nor has an in-depth analysis of requirements for confident detection and accurate quantification in quantitative metagenomes been conducted. Despite the progress made with previously developed quantitative metagenomic methods, each method has limitations. Normalizing target genes or genomes within a sample by total cell counts is highly dependent on the accuracy of the method used to measure cell counts. Normalizing targets by 16S rRNA or housekeeping genes assumes a fixed ratio of 16S rRNA or housekeeping genes to the number of cells. However, this is not always the case and conclusions are not always reliable[15, 19] with the established standard of 1.8 16S rRNA gene copies per bacterium at odds with the current average 16S rRNA gene copy number per bacteria of 5.29 from the Ribosomal Database Project[19]. Previously, results of normalizing by 16S rRNA gene copies were compared to flow cytometry cell counts and yielded different results[20] demonstrating compounding errors when relying on an additional method to normalize metagenomes. Another drawback of these approaches is that they are not feasible for viral communities because viruses do not have conserved genes to measure with quantitative PCR and are too small to quantify accurately with flow cytometry.

Alternatively, DNA standards have been added to samples at known concentrations to measure absolute abundances. With this approach, target gene or genome counts in the metagenome are quantified using the number of reads mapping to the standards and the known concentration of the standards[9-13]. The standard nucleic acids may be from a foreign genome[11, 13], a mock microbial community[12], or synthetic DNA mimicking a microbial community[9, 10]. A drawback of using foreign genomes or mock communities is that they likely share sequences with the sample's microbial community leading to non-specific mapping to spike-in genomes where the origin of reads mapping to those shared sequences is ambiguous. The use of synthetic DNA standards can reduce non-specific mapping because the synthetic DNA sequences are designed to be unique from known organisms[9, 10]. Furthermore, synthetic DNA standards can be designed to better capture the range of a sample microbiome's characteristics, including genome lengths, GC contents, and concentrations. To date, quantification with standards has been applied to directly compare population abundances in samples with varying environmental conditions to test if specific populations' abundance correlate with environmental variables[21, 22].

In any quantitative method, well-defined limits of detection and limits of quantification are necessary to meet quality targets for bias, imprecision, and total error[23, 24]. Identifying rational limitations for quantitative metagenomics with spike-in standards is needed[25]. Defining the limitations of quantitative metagenomics is complicated by the thousands of targets, and because each measured read comprises only a small fraction of a target gene or genome. Previously, as little as one read was considered sufficient for detection and quantification[9, 11]. Considering only read counts in the detection limit, however, results in target length bias as a single read covers a larger proportion of a short target than longer targets, therefore, requiring higher concentrations to detect short targets. Detection thresholds for quantitative metagenomics based on read distribution and coverage across targets remediate target length bias[26, 27]. [27] proposed minimum requirements for viral detection based on read coverage, distribution, and count based on read mapping with Kallisto to the RefSeq viral database. Read distribution and coverage may be summarized by measuring mapping entropy (i.e., measure of randomness). Entropy is commonly used to summarize community diversity and richness for the Shannon's Diversity Index. We propose removing read count requirements to prevent target length biases and defining detection thresholds with minimum mapping entropy detection thresholds in quantitative metagenomes.

Here, we applied synthetic DNA standards[9] to municipal wastewater viromes to determine absolute virus abundances. The spike-in standards provided an opportunity to develop metrics and criteria for evaluating whether a target is detectable and quantifiable in a metagenomic dataset. We created QuantMeta, a bioinformatic tool to assess target-specific detection thresholds and improve quantification accuracy in quantitative metagenomes. Ultimately, we compare the absolute abundances of two groups of viruses determined with our approach to values in other wastewaters quantified with qPCR. The guiding principles defining limitations of quantitative metagenomics can be applied broadly to whole metagenomes.

3.4 Importance

Establishing quality thresholds for quantitative metagenomics will improve confidence and accuracy of absolute abundance measurements. By defining limitations of quantitative metagenomics, future studies will be able to inform experimental designs to determine how much sequencing depth is required and when supplementing sequencing results with alternative methods, such as qPCR, that have more sensitive detection and quantification thresholds, is necessary.

3.5 Methods

3.5.1 Sample collection and processing

Two grab samples of secondary effluent (20 L) or raw influent (10 L) were collected in autoclaved carboys from automatic samplers at the Ann Arbor wastewater treatment plant (Ann Arbor, MI) per day from December 19 to 24, 2020. Samples were transported to the lab on ice within 2 h of collection and immediately began concentration and purification of the viral community using the ultrafiltration and purification method previously described[28] (Figure 3.2). Briefly, secondary effluent and raw influent were pre-filtered through 100-µm sized pores (Long-Life polyester felt filter bag, McMaster-Carr, Cat. No. 6835K58) then 0.45-µm sized pores (Express PLUS PES filters, MilliporeSigma™, Cat. No. HPWP14250) and concentrated with tangential ultrafiltration (approximately 50-fold and 25-fold, respectively, using 30 kDa MWCO dialysis filters, Asahi Kosei Medical Co., Ltd, Cat. No. 6292966). The concentrated samples were treated with 1 mL chloroform to lyse any remaining cells, filtered through 0.45-µm filters, further concentrated with dead-end ultrafiltration (approximately 20-fold, using 100 kDa MWCO Amicon™ Ultra Centrifugal filter units, MilliporeSigma™, Cat. No. UFC510096), and then the extra-viral nucleic acids were degraded with 100 U/mL DNase (Roche, Cat. No. 10104159001) for 1 h on the bench top. The DNase enzymatic reaction was ended by adding 100 mM EDTA and 100 mM EGTA. DNA was then extracted immediately with QIAamp UltraSens Virus Kit (QIAGEN, Cat. No. 53706). The manufacturer's instructions were followed except for the first six steps, in which 140 µL of sample was combined with 5.6 µL carrier RNA and vortexed briefly. DNA extracts were stored at -20˚C. On December 26, 2020, a 20 L deionized water sample was processed with the same viral concentration and purification process to account for contamination during sample processing.

3.5.2 Sample characterization

Each raw influent and secondary effluent sample was analyzed for pH, turbidity, solids content, fold concentration, and viral recovery (Table B.1). Sample volumes were determined by weighing samples and assuming a density of 1 g/mL. To more accurately determine virus recovery through concentration and purification, a replicate grab sample was always collected and processed in parallel with the grab samples intended for sequencing. The sample-specific viral recovery was determined by adding *Enterobacteria* phage T3 (GenBank accession no. NC_003298, ATCC® BAA-1025-B1™) to a concentration of 10^5 gene copies/µL in grab samples. The concentration of the phage T3 remaining after the concentration and purification steps was determined with a previously developed ddPCR probe assay[28] (Table B.3).

3.5.3 Sequencing standard spike additions and ssDNA standard development

A set of 91 dsDNA and ssDNA standards were spiked into each DNA extract prior to library preparations. Sequins metagenome mix A dsDNA standards with lengths varying 981 to 9,120 bp and GC content varying 20 to 71% (Figure 3.2)[9] were spiked into each sample. The dsDNA concentration of the sequin metagenome mix A was measured with the Qubit^{TM} dsDNA HS Assay (ThermoFisher Scientific, Cat. No. Q32851) with 2 µL of DNA template added to the 200 µL assay. The set of standards was expanded with five ssDNA standards. A similar approach to the Sequin metagenome mix was employed by inverting ssDNA viral genomes in the NCBI viral genome database (downloaded on 3/19/2019) and dividing the genomes into 1-kb long fragments. The fragments were mapped to the NCBI nr database (downloaded on 8/15/2019) with bowtie2 (v2.3.5). Inverted genome fragments with no alignments were selected as potential ssDNA standard candidates. A random selection of five sequences from the potential candidates were selected with GC contents of 31.7, 40, 45, 50, and 60% and made into Megamer® ssDNA fragments without the complementary strand (IDT, Coralville, IA, Table B.2). The ssDNA fragments were resuspended in molecular biology grade ddH_2O (Fisher Scientific, Cat. No. BP28191) to an approximate concentration of 7.5 ng/µL, aliquoted into 10 µL increments, and stored at -20˚C. Standards underwent a maximum of one freeze-thaw cycle. Immediately prior to use, the concentration of the ssDNA aliquots were measured with the Qubit^{TM} ssDNA Assay (ThermoFisher Scientific, Cat. No. Q10212). A mix of the ssDNA standards was then prepared to match the range of concentrations in the sequins metagenome mix A, namely 10^7, 10^4, 10^6, 10^8, 10^5 gc/µL of the 31.7, 40, 45, 50, and 60% GC content ssDNA standards, respectively. The sequins dsDNA metagenome mix A and the ssDNA standard mix were spiked into each sample at 10 gc/ng DNA extract for the standards at the lowest abundance in the mixes based on Qubit measurements. We predicted that the standards spiked at 10 gc/ng DNA extract would be near the detection threshold at our sequencing depth (approximately 200 million reads per sample). The spike-in concentrations were confirmed with ddPCR assays designed for one dsDNA standard and one ssDNA standard (see section below for details, Table B.4).

3.5.4 Foreign marine phage HM1 genome spike additions

To further examine quantification accuracy of low abundance viruses at our sequencing depth, we spiked in marine phage HM1 (Genbank accession no. KF302034.1) that is foreign to our wastewater samples. The dsDNA HM1 genome (129,401 bps) has an average GC content of 35.7%. HM1 DNA was extracted using the QIAamp UltraSens Virus Kit with the modified protocol described above. Then, 15 µL of DNA extract with 1 µL of loading dye was run on a 0.3% agarose gel with 1 µL of 10,000x SYBR™ Gold nucleic acid gel stain (Invitrogen™, Cat. No. S11494) per 10 mL of gel at 3 V cm^{-1} for 90 minutes. GeneRuler High Range DNA ladder (Thermo Scientific™, Cat. No. FERSM1351) was run according to the manufacturer instructions. HM1 genomes were extracted from the gel with the QIAEX II Gel Extraction Kit (QIAGEN, Cat. No. 20021). The concentration of the purified HM1 genomes was measured with the Qubit dsDNA HS Assay. HM1 was spiked into all samples at approximately 50 gc/ng DNA and subsequently checked with ddPCR (see section below for details, Table B.4).

3.5.5 Spike-in ddPCR assays

Spike-in standards and HM1 concentrations were checked with singlet ddPCR reactions performed with the QX200 AutoDG Droplet Digital PCR System (Bio-Rad Laboratories, Inc., Hercules, CA). For each plate, at least two ddH2O negative controls and two positive controls from spike-in stocks were included. Specific primers and probes were developed for T3, HM1, dsDNA standard S1106_MG_020_A, and the ssDNA standard with 45% GC content (Table

B.3). 22µL reactions were prepared with 11 µL of 2x ddPCR^{TM} Supermix for Probes (No dUTP) (Bio-Rad Laboratories, Inc., Cat. No. 1863023), 0.4 µM of all probes and primers, and 3 µL of template. Droplets were generated using the automated droplet generation oil for probes (Bio-Rad Laboratories, Inc., Cat. No. 1864110) to a 20 µL volume, then PCR was performed on the C1000 Touch^{TM} Thermal Cycler (Bio-Rad Laboratories, Inc., Hercules, CA) immediately after droplet generation. The ssDNA assays consisted of 40 cycles of denaturation for 30 seconds at 95˚C, annealing for 1 minute at 56˚C, and extension for 2 minutes at 72˚C, then enzyme deactivation for 5 minutes at 4˚C and 5 minutes at 95˚C, and a final hold at 4˚C. The same PCR reaction for dsDNA assays was performed with an initial denaturation step at 95˚C for 10 minutes. Plates were run on the droplet reader within 1 hour of PCR completion. For each ddPCR reaction, thresholds were set using a previously defined method that uses kernel density to categorize droplets as positive, negative, or rain[29]. Reactions with more than 2 populations, more than 2.5% of droplets classified as rain, or less than 30% compartmentalization were rerun. Inhibition was checked by running 10 and 100-fold dilutions on an influent and effluent sample and was not found to significantly alter the concentration; therefore, 10-fold dilutions were used for all dsDNA and ssDNA standard reactions and 1- or 2-fold dilutions were used for HM1 reactions. ddH$_2$O negative controls infrequently resulted in positive droplets, and the corresponding concentrations were significantly lower than samples. Specifically, for T3 dsDNA extracts, dsDNA standards, ssDNA standards, and HM1 dsDNA extracts, the ddH$_2$O negative controls resulted in 20.8 gc/µL template (1/6 reactions), 17.3 gc/µL template (1/4 reactions), 18.5-36.1 gc/µL template (2/4 reactions), and 55.6 gc/µL template (1/4 reactions), respectively.

3.5.6 Illumina NovaSeq sequencing

Libraries were prepared with the Accel-NGS® 1S Plus DNA Library Kit (Swift Biosciences, Cat. No. 10024) using 50 ng DNA. Samples were sequenced on the Illumina NovaSeq 600 with five samples sequenced per paired-end 500 cycle SP flow cell yield 251-bp long reads. Library preparations and sequencing were conducted by the Advanced Genomics Core at the University of Michigan. Quality control was performed by trimming Illumina adaptors and an additional 15 bp from the rightmost and leftmost of each read to remove adaptors from the Accel-NGS® 1S Plus DNA Library Kit, and reads were decontaminated of PhiX174 with BBDuk (BBTools, v37.64) (Table B.4).

3.5.7 Oxford Nanopore GridION sequencing

We combined short and long read sequencing technologies because it improves viral assemblies since viral genomes commonly have repetitive regions, high mutation rates, and contain host genome fragments (Figure 3.2)[30-32]. DNA extracts from each experimental replicate without standards or phage HM1 were cleaned up prior to library preparations with the Zymo Genomic DNA Clean and Concentrate-10 kit (Cat. No. D4011, Zymo Research Corporation). Long read libraries were prepared with the Ligation Sequencing Kit (Cat. No. SQK-LSK109, Oxford Nanopore Technologies) and barcoded with the Native Barcoding Expansion 1-12 (Cat. No. EXP-NBD104, Oxford Nanopore Technologies). The six samples were sequenced on two flow cells (R9.4.1, Cat. No. FLO-MIN106, Oxford Nanopore Technologies) (Table B.4). Basecalling was performed using Guppy (v4.2.3) and called reads were classified as either pass or fail depending on their mean quality score (≥ 7). Library preparations, sequencing, and basecalling were conducted by the Advanced Genomics Core at the University of Michigan.

3.5.8 Assemblies: long read only and hybrid co-assemblies

Two assemblies were run with the reads from each sample. Hybrid co-assemblies were performed with long read and short reads from each sample separately with metaSPAdes (v3.15.2) using kmer sizes of 21, 33, 55, 77, 89, and 127. Long read only assemblies were performed with Flye (v2.8.3) for each sample separately followed by several polishing steps including four rounds of Racon (v1.4.10), one round of medaka (v1.3.2), and short-read error correction with pilon (v1.24)[32]. Following both assemblies, the contigs for each sample were pooled and contigs less than 1,000-bp were removed. The remaining contigs were assessed for likelihood of viral or proviral origin. Five viral detection methods were run on the contigs. VirSorter (v1.0.5)[33] with the virome flag, VirSorter2 (v2.2.2)[34], VIBRANT (v1.2.1)[35] with the virome flag, VirFinder (v1.1)[36], and CheckV (v0.7.0)[37] end-to-end were run to assess likelihood of each contig being viral or proviral. Potential viral or proviral contigs were sorted into high and low confidence categories. High confidence contigs were any contigs where at least one viral detection method indicated the contigs was highly likely to be viral or proviral. The high confidence cut-offs were a VirSorter score less than 2 or 4, VirSorter2 score greater than 0.9, VirFinder score greater than 0.9, VIBRANT score less than 3 or included in the "prophage" list, and included in the CheckV "prophage" list. Low confidence contigs were any contigs where at least one viral detection method indicated some confidence in the contigs to be viral or proviral. The low confidence cut-offs were a VirSorter score of 3 or greater than 4, VirSorter score between 0.5 and 0.9, VirFinder score between 0.7 and 0.9, and a VIBRANT score of 4. The low confidence viral contigs were assessed with CheckV run end-to-end and any contig containing one or more viral genes or no host genes were assigned as viral and included with the high confidence viral contigs. Of the contigs greater than 1000 bp, 75.5-78.0% of contigs were classified as viral. The viral contigs were dereplicated by clustering contigs with clustergenomes.pl (v5.1)[38] with 95% ANI similarity and sharing at least 70% coverage, then retaining the longest contig in each cluster. After clustering, 16.5-23.3% of contigs were removed from the viral contig pool. Remaining viral contigs are referred to as viral populations.

3.5.9 Mapping

Read mapping with short reads was performed using bowtie2 (v 2.4.2) using the default mapping parameters with deinterleaved fastq formatted files of QC reads. Bowtie2 indexes were built with default parameters for the NCBI viral database as whole genomes and broken into separate genes (downloaded 9/6/2021), VirSorter curated database (downloaded 5/5/2021), dsDNA and ssDNA standard sequences, mutated standard sequences, HM1 sequence, RefSeq crAssphage database (downloaded 10/4/2021), and RefSeq polyomavirus database (downloaded 10/4/2021). JC and BK polyomaviruses were measured based on the concentration of reads or contigs mapping to NC_001699.1 and NC_001538.1, respectively. Primer blast to the NCBI nr database with the crAssphage CPQ056 primer set[39] was performed to determine the list of NCBI accessions captured by the primer set (performed on 10/26/21). Viral populations from each sample were indexed using the “large-index” parameter. The bowtie2 sam file output was converted to a tabular format with the number of reads mapping to each basepair of a target using idxstats (samtools, v1.11). Minimap2 (v2.17) was used to map contigs onto standard sequences and genes or genomes from databases using default parameters. Assembly quality of contigs derived from standards were assessed by performing Blastn (v2.9.0) with a custom database made from the standard sequences using default settings.

3.5.10 Detection threshold parameters and test dataset

To summarize read distribution and coverage, we measured the relative entropy of each target in a metagenome. Entropy is commonly applied in ecology to summarize community diversity with the Shannon's diversity index. In the context of quantitative metagenomics, the collection of basepairs mapping to a target gene or genome is analogous to a community with each base pair representing an individual and each position along the target representing a population (Figure 3.1).

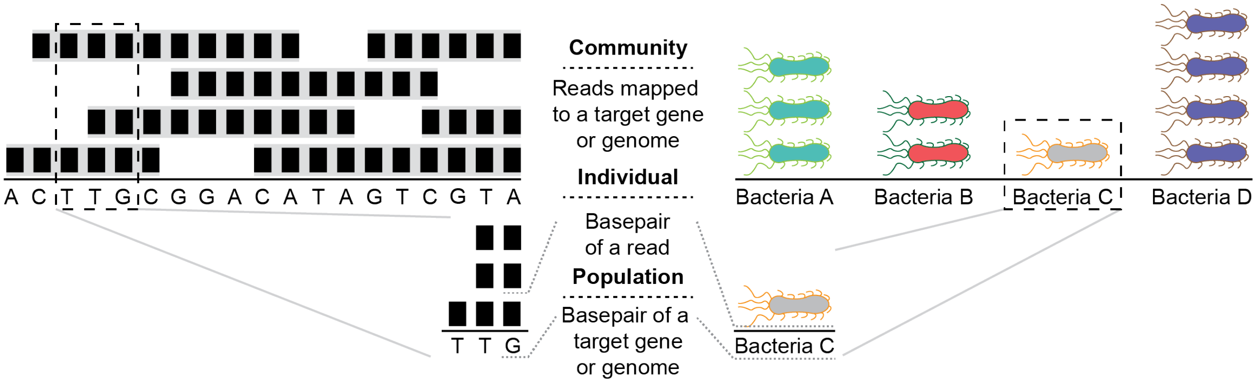
A C T T G C G G A C A T A G T C G T A
T T G
Community
Reads mapped to a target gene or genome
Individual
Basepair of a read
Population
Basepair of a target gene or genome
Bacteria A
Bacteria B
Bacteria C
Bacteria D
Bacteria C

Figure 3.1: The collection of reads mapped onto a target gene or genome is analogous to a community composed of individuals categorized into populations.

Entropy for reads mapping to a target are defined by equations (3.1-3.3).

$$I = \sum_{i=1}^{L} \frac{b_i}{B} \ln\left(\frac{b_i}{B}\right) \qquad (3.1)$$

$$I_{max} = \ln(L) \qquad (3.2)$$

$$R = \frac{I}{I_{max}} \qquad (3.3)$$

"I" refers to the distribution and coverage index for a target and "I_{max}" is the maximum value of "I". In this context, the minimum possible value of "I" is 0 equating to a single basepair aligning to a target. "b_i" is the read depth for basepair i on a target, "B" is the total number of basepairs mapping to a target, and "L" is the length of a target. Relative entropy, "R", summarizes the evenness of read distribution with 1 referring to complete coverage and perfectly even distribution of reads across a genome.

Binary logistic regression models were developed to test the ability of entropy to summarize coverage and read distribution. We applied an expected distribution to actual distribution greater than 0.3 and a minimum 10% coverage; these thresholds were previously applied by FastViromeExplorer[27]. We did not include a minimum number of reads mapping because including a minimum number of reads introduces a target length bias by increasing the detection threshold for short targets, such as genes or small viral genomes, whereas coverage requirements eclipse minimum read counts for longer targets (Figure B.1). Regressions were created with the results of mapping reads to spike-in standards supplemented with downsampling viromes and creating a set of "failure" standards. Downsampling was performed by randomly sampling 1% and 20% of reads with seqtk (v1.3) from each virome to expand the existing spike-in standards dataset to include lower standard concentrations. The set of "failure" standards were created by modeling single nucleotide polymorphisms (SNPs) for each standard with a Mutation Simulator[40]. 5 sequential rounds of mutations were performed with 0.1 rate of SNPs and 0.02 rate of insertions, deletions, duplications, inversions and translocations with lengths of 5-bp. Reads without downsampling from each virome were mapped to mutated standard sequences from each mutation round. All mutated sequences with a coverage less than 10% or expected read distribution based on a Poisson distribution to actual read distribution less than 0.3 were retained as "failure" standards to capture the variability in mapping to targets that are not confidently detected. Results from the logistic regressions were used to establish minimum relative entropy with respect to target length as a threshold for detection. Model performance was tested by measuring the relative entropy of reads aligning to individual basepairs across gene or genome targets from the NCBI viral database and VirSorter curated database[33].

3.5.11 Quantifying targets with metagenomes

Synthetic DNA standards were spiked into samples to create a linear regression model relating relative abundance to absolute concentration using an approach similar to (Hardwick et al.) (Figure 3.2). The concentration of viral genes and genomes were determined by relating the average read depth per mass of library insert size to the absolute gene copy concentration per ng of DNA extract, then converting the DNA extract concentration to the concentration in the original wastewater sample as defined by equations (3.4-3.6).

$$predicted_x \left[\frac{gc}{\mu L}\right] = \frac{\sum_i^n (read\ depth_i)}{n} [gc] \cdot \frac{c_{DNA}\left[\frac{ng}{\mu L}\right]}{M_{library}[ng]} \tag{3.4}$$

$$Gene_x \left[\frac{gc}{\mu L}\right] = slope \cdot \left(predicted_x \left[\frac{gc}{\mu L}\right]\right) + intercept \tag{3.5}$$

$$Gene_x \left[\frac{gc}{mL}\right] = Gene_x \left[\frac{gc}{\mu L}\right] \cdot \frac{1}{R \cdot CF} \tag{3.6}$$

The average read depth of target is converted to gene copies per ng of DNA in the library insert (i.e., $predicted_x$) in eq. (3.4) where read $depth_i$ is the number of read mapping to the i-th basepair along a target with n basepairs and $M_{library}$ is the mass of the library insert. $Predicted_x$ is converted to $Gene_x$ (i.e., the concentration of a target in the DNA extract) in eq. (3.5) with a linear regression model based on the observed average read depths of spike-standards to their known concentration in the DNA extract. $Gene_x$ is then converted to its absolute abundance with eq. (3.6), in which C_{DNA} is the DNA concentration of the DNA extract, R is the viral recovery through concentrating and purifying viruses from wastewater, and CF is the fold concentration by volume of wastewater through concentration and purification.

3.5.12 Statistical analysis

All statistical analysis was performed in R (v4.0.3). Linear and logistic regression analyses and student's t-tests were performed with the R stats package (v4.0.3). Paired and unpaired t-tests were performed with 0.95 confidence levels with *p*-values less than 0.05

considered significant. ROC curves and optimal cutpoints for logistic regressions were performed with the cutpointr package (v1.1.1). Graphs were created with ggplot2 (v3.3.5).

3.5.13 Data and programming availability

All data will be made available on JGI and relevant code will be compiled as QuantMeta and available on Github upon submission of this work to bioRxiv.

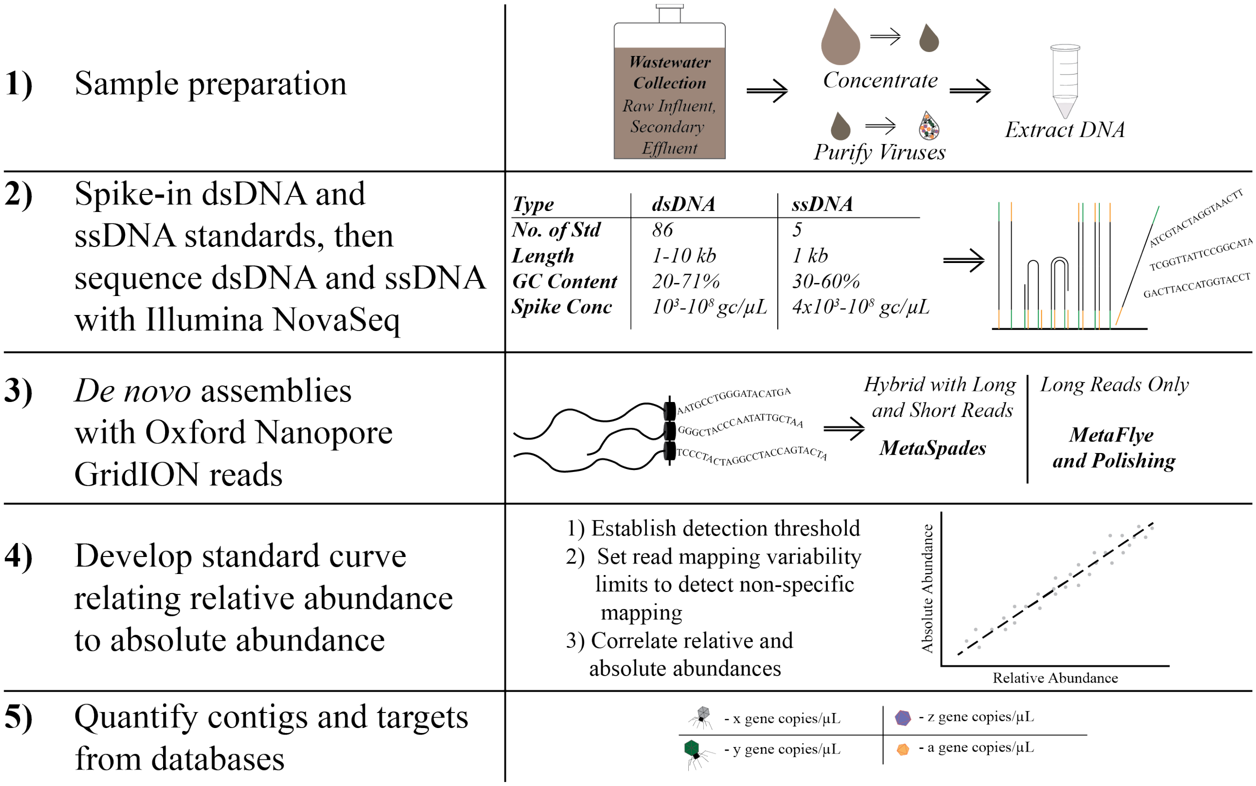
1) Sample preparation
Wastewater Collection
Raw Influent, Secondary Effluent
Concentrate
Purify Viruses
Extract DNA
2) Spike-in dsDNA and ssDNA standards, then sequence dsDNA and ssDNA with Illumina NovaSeq
Type | dsDNA | ssDNA
No. of Std | 86 | 5
Length | 1-10 kb | 1 kb
GC Content | 20-71% | 30-60%
Spike Conc | 10^3-10^8 gc/µL | 4x10^3-10^8 gc/µL
ATCGTACTAGGTAACTT
TCGGTTATTCCGGCATA
GACTTACCATGGTACCT
3) *De novo* assemblies with Oxford Nanopore GridION reads
AATGCCTGGGATACATGA
GGGCTACCCAATATTGCTAA
TCCCTACTAGGCCTACCAGTACTA
Hybrid with Long and Short Reads
MetaSpades
Long Reads Only
MetaFlye and Polishing
4) Develop standard curve relating relative abundance to absolute abundance
1) Establish detection threshold
2) Set read mapping variability limits to detect non-specific mapping
3) Correlate relative and absolute abundances
Absolute Abundance
Relative Abundance
5) Quantify contigs and targets from databases
- x gene copies/µL
- y gene copies/µL
- z gene copies/µL
- a gene copies/µL

Figure 3.2: Overview of the quantitative metagenome workflow. Briefly, (1) a wastewater sample is collected, then concentrated and purified to isolate the viral community and dsDNA and ssDNA is extracted. (2) The DNA extract is spiked with a set of dsDNA and ssDNA standards and sequenced with Illumina NovaSeq. (3) The same DNA extracts with spike-in standards are sequenced with Oxford Nanopore GridION and the long and short reads are de novo assembled. (4) A relationship between known concentrations and observed relative abundance with the spike-in standards is developed to (5) determine the concentration of unknown targets.

3.6 Results and Discussion

3.6.1 Entropy-based detection threshold improves standard curve

To confidently detect targets in our metagenomes, we developed detection thresholds that account for read coverage and distribution across a target. Standard curves relating ddPCR measurements to predicted concentrations of standards and entropy-based detection thresholds were defined using the reads mapped to our spike-in DNA standards. We developed entropy-based detection thresholds (R_{detect}) based on target coverage, read distribution, and length (L) (equation 3.7, Section B.2).

$$R_{detect} = 0.894 + 0.732(L) \qquad (3.7)$$

An observed entropy measurement less than the length-specific entropy-based detection threshold is discarded because the target is not confidently detected by failing coverage and read distribution requirements.

To test the application of this entropy cut-off, spike-in standards were added at concentrations expected to span the range of method detection thresholds and downsampled 20% and 1% to further test detection limits (Figure 3.3). At low concentrations, large variability between the ddPCR measured and predicted concentrations demonstrated low abundance standards were not confidently detected and should be excluded from analysis. Of 910 standards across all samples, 890 exceeded entropy cut-offs. When standards not meeting length-specific R_{detect} were removed, the detection limit was determined to be approximately 500 gc/µL of DNA extract based on the lowest standard concentrations remaining. As expected, the entropy-based detection threshold improved the relationship between the expected and measured standard concentrations, which should theoretically be 1 if the measured concentrations were perfectly captured. Without the entropy-based detection threshold applied, the regression slope was 0.895 ($R^2 = 0.94$) and with the entropy-based detection threshold applied, the regression slope was improved to 0.993 ($R^2 = 0.94$). Our detection limit is 5-fold lower when adjusted for sequencing depth than a previous quantitative metagenomic study that relied on a foreign genome spike-in and applied a detection threshold of a single read[11]. This is likely due to differences in methods to determine detection limits.

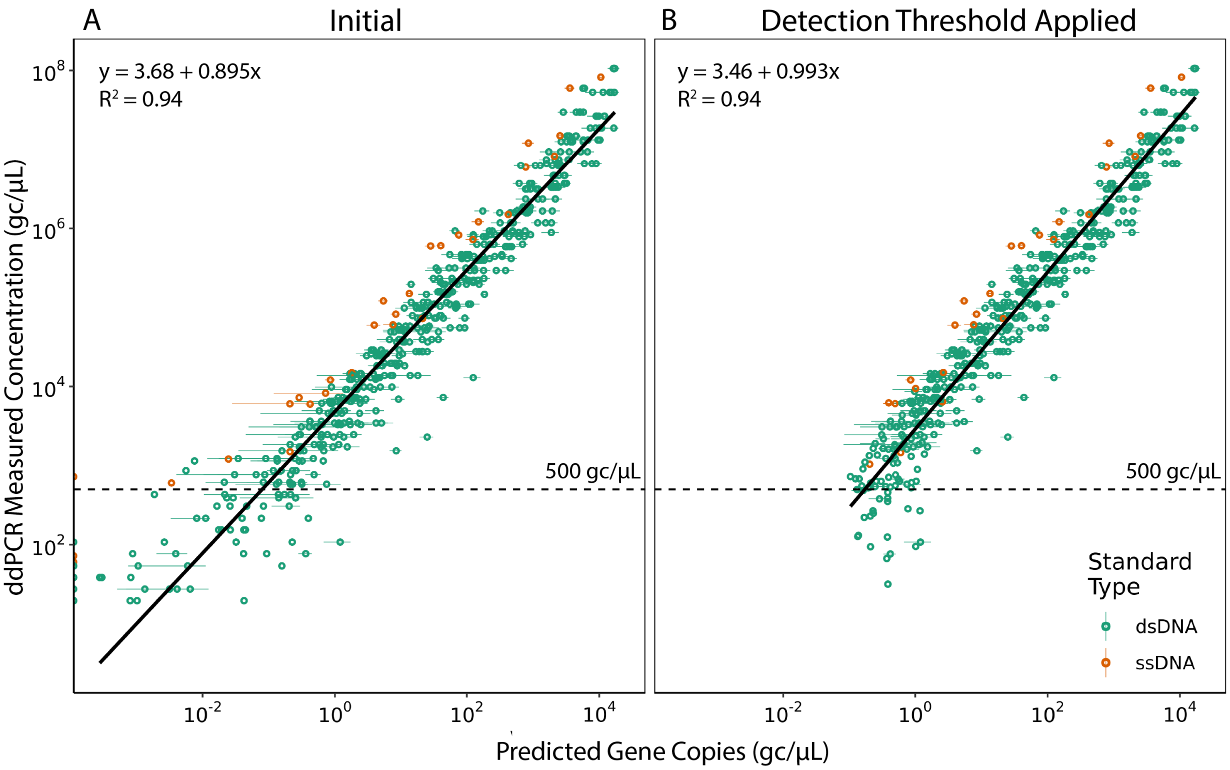

Figure 3.3: Predicted gene copies per µL of DNA extract in viromes using equation 4 compared to the known spike-in concentration based on ddPCR measurements for all of the standards across all samples with no, 20%, and 1% downsampling (replicates are reported as single values with 95% confidence intervals). dsDNA standards are in green and ssDNA standards are orange. (A) shows the initial results without the detection threshold applied and (B) shows the results with the detection threshold applied and standards failing to meet the detection threshold removed.

3.6.2 Incorporating assessment of mapping variability improves the accuracy of quantification

Whereas target detection relies on sufficient evidence that a target is present in a metagenome, target quantification relies on accurately transforming read mapping results to target concentrations. Unknown targets can be quantified using two mapping approaches: database-dependent read mapping, where reads are mapped to known genes or genomes, and *de novo* read mapping, where reads are mapped to contigs assembled from the same sample. We mapped short reads from our wastewater viromes onto viral databases and the contigs identified as the spiked-in standards. In either case, non-specific mapping could lead to the inaccurate derivation of target concentrations. Non-specific mapping may occur when reads map to an identical or similar mobile element on a reference sequence that is different from its true source or reads map to a conserved element present on several reference sequences[41]. In both cases, non-specific mapping can change the observed read depth. For the case where reads are mapped to assembled contigs, assembly errors may also cause quantification errors.

We designed an approach for identifying and accounting for nonspecific mapping and assembly errors prior to quantification to improve quantification accuracy. Locating non-specific mapping or assembly errors is complicated by intrinsic variability in the number of reads mapping across a target sequence[42, 43]. Browne et al. (2020) observed a quadratic relationship between local GC content and local read depth along a target sequence. We distinguished intrinsic read depth variability from non-specific mapping or assembly errors using read mapping results to standard sequences because standards do not share DNA sequence homology with any known organism and, therefore, are assumed not to have non-specific mapping[9]. We observed that this assumption held because there were no outliers with high RMSE based on reads to standard sequences (Figure 3.4A). Using the observed read depth variability of the spiked in standards, we related local GC content to the observed local read depth to establish the intrinsic variability in read depth across a target (Section B.3, Table B.6). Targets with non-specific mapping or assembly errors were identified by setting maximum allowable root mean square errors (RMSE) based on observed read depth variability RMSE with the standards (Figure 3.4A, see Section B.3). Targets with read depth variability RMSE above the threshold then underwent an iterative correction process (Figure B.4) involving the following steps: (1) identification of outlier regions along the target (Figure 3.4A), (2) establishing acceptable read depth ranges for outlier regions (Figure 3.4B), (3) setting outlier regions to the maximum or minimum of the acceptable read depth range (Figure 3.4B), and (4) iterating until the RMSE of the whole target was less than the threshold or 20 iterations were completed. At the end of the iteration process, if targets had more than 20% of 49-bp sliding windows altered, targets were identified as not quantifiable.

To assess our ability to identify assembly errors, we evaluated the RMSE of reads mapping to the subset of contigs originating from the standards (Figure 3.4C and Figure 3.4D). We assessed results with all standard derived contigs and quality controlled standard derived contigs, whereby redundant contigs of fragmented standard assemblies or contigs with less than 80% alignment to standards were removed. Of 910 standards spiked in across all samples, 837 had *de novo* contigs and 811 had quality controlled contigs which is less than the fraction of standards that were above detection thresholds (890/910). 103 standards' contigs had too high of read depth variability RMSE caused by assembly errors resulting in unpredictable read mapping distribution that impacts quantification (Table B.8). In the quality controlled contigs, only 45 had high read depth variability RMSE. 27 standards' contigs in both sets of contigs had high read depth variability RMSE but were not altered during quality control. These standards' contigs had additions to the ends of the standard sequences on the contigs that impacted the read depth distribution.

Previously, incomplete and over-extended contigs were found to be unreliable estimates of transcript abundances[44]. 15 standards' contigs with high read depth variability RMSE were correctable in each set of standards' contigs. Correcting standards' contigs significantly improved quantification accuracy by reducing differences to ddPCR measurements with and without contig quality control (*p*-value = 0.0063 and 0.0065, respectively) while corrected concentrations were not dependent on whether contigs were quality controlled (*p*-value = 0.351). These results demonstrate that quantification detection and correction improved quantification accuracy by detecting assembly errors that impacted a target's average read depth.

To examine the prevalence of non-specific mapping, we mapped reads to viral databases. Of 95,351 genes and genomes across all samples exceeding the detection threshold, 1,290 targets were flagged as having non-specific mapping with 702 identified as not quantifiable (Figure 3.4E). Not quantifiable targets were likely false positives. The percent of reads mapped as false positives was 3.8% (n = 525/90,012 targets), 14.5% (n = 58/2,426 targets), and 56.0% (n = 119/2,913 targets) to the NCBI viral gene database, NCBI viral genome database, and VirSorter curated genome database, respectively. These observations are consistent with previous read mapping false positives to gene-centric databases of 11.9-23.6%[45]. These percentages likely vary depending on target length, abundance of false negative targets, and database size. Spike-in standards are a valuable tool to study read mapping variability in metagenomes to identify where nonspecific mapping or assembly errors occur.

We evaluated if contig-based standard concentrations differed from ddPCR measurements and read-based concentrations (Figure 3.5). Contig-based standard concentrations after correction are statistically similar to ddPCR measurements where the set of all standards' contigs regression slope was 0.971 (R^2 = 0.92) and quality-controlled standards' contigs regression slope was 0.998 (R^2 = 0.93). Read-based concentrations were not significantly different from all or quality controlled contig-based concentrations after correction (*p*-value = 0.182 and 0.727, respectively). Previously, read-based transcript quantification matched full-length contig transcript quantification however assembly issues increased quantification error[44].

Our results indicate that read-based and contig-based quantification in metagenomes are equally valid methods to quantify targets in metagenomes and are not significantly different from ddPCR measurements. We found that ssDNA and dsDNA standards resulted in significantly different residuals from the relative to absolute abundance regression (Figure B.5).

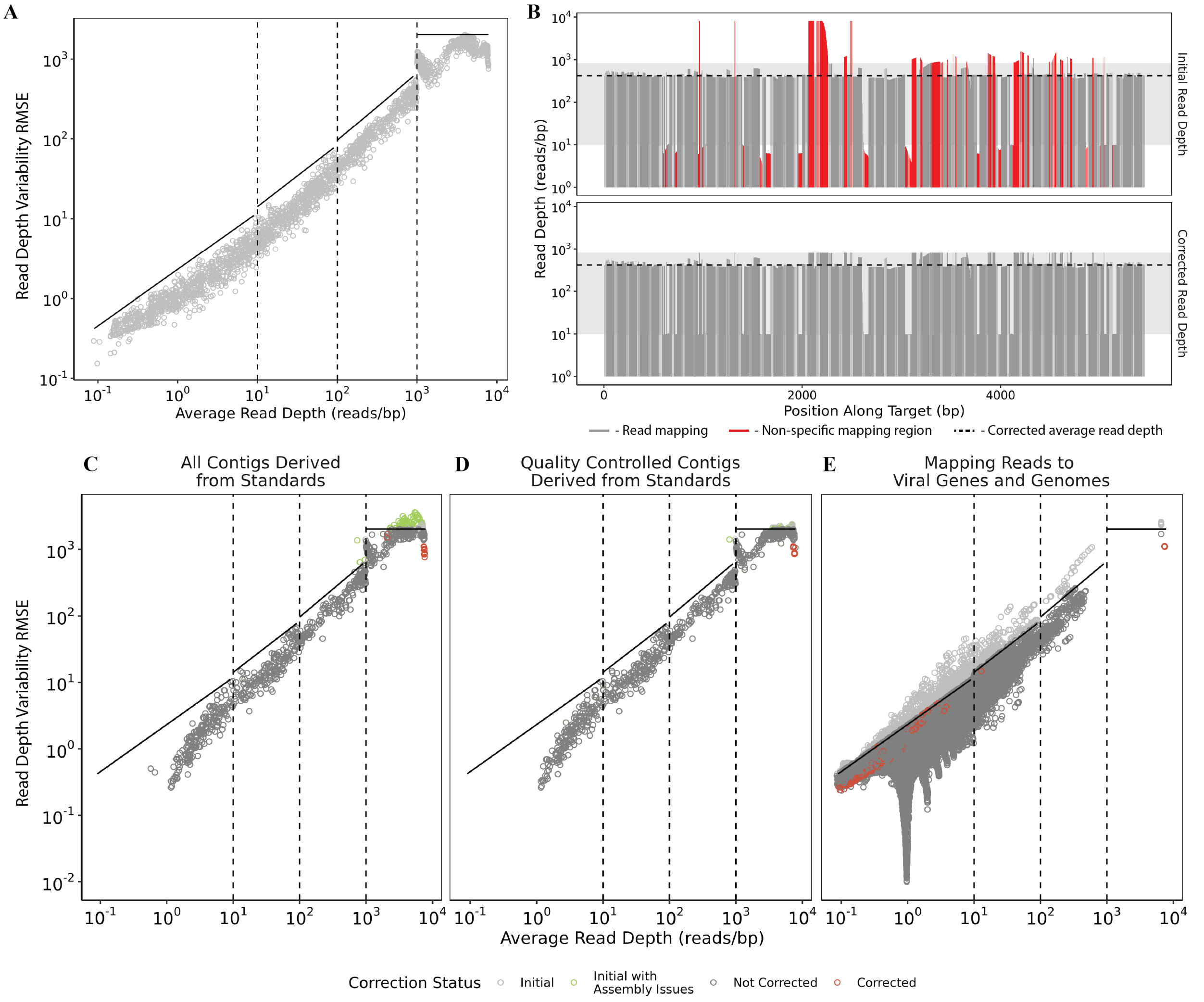

Figure 3.4: (A) Observed RMSE of the reads mapped to known standard sequences was measured with respect to the total average read depths. Dashed, vertical lines indicate the average read depth bins used for the read depth variability regression (Table B.5). Thresholds for maximum acceptable RMSE (black lines) were set e0.25 above the highest observed RMSE with respect to the average read depths (Table B.6). (B) Targets with too high of RMSE with respect to its average read depth were flagged for nonspecific mapping or, for contigs, assembly errors. An example of a contig with nonspecific mapping or assembly errors is shown. Regions before correction in red indicate areas outside of 1.5 standard deviations from the mean read depth where correction is required. The corrected read depth is shown with the black, dashed line. The corrected read depths are shown in the bottom graph. In this particular case, 21.2% of sliding windows required correction and, therefore, the quantification accuracy is questionable and the target requires further inspection before accepting the results. (C) and (D) show detection and correction applied to all contigs derived from spike-in standards and quality controlled standards' contigs, respectively. Grey

points represent initial RMSE with points falling above the black RMSE threshold lines requiring correction. Points in green are initial RMSE that were flagged as having assembly errors (fragmentation and low alignment). Corrected RMSE for all targets with less than 20% of sliding windows requiring correction are shown in red. (E) Nonspecific mapping detection and correction was applied to reads mapping to genes and genomes in the NCBI viral database and VirSorter curated database. Points in grey falling above the black RMSE threshold lines indicate targets corrected for nonspecific mapping. If less than 20% of sliding windows for the respective targets were flagged as requiring correction, their corrected RMSE is represented by the red points.

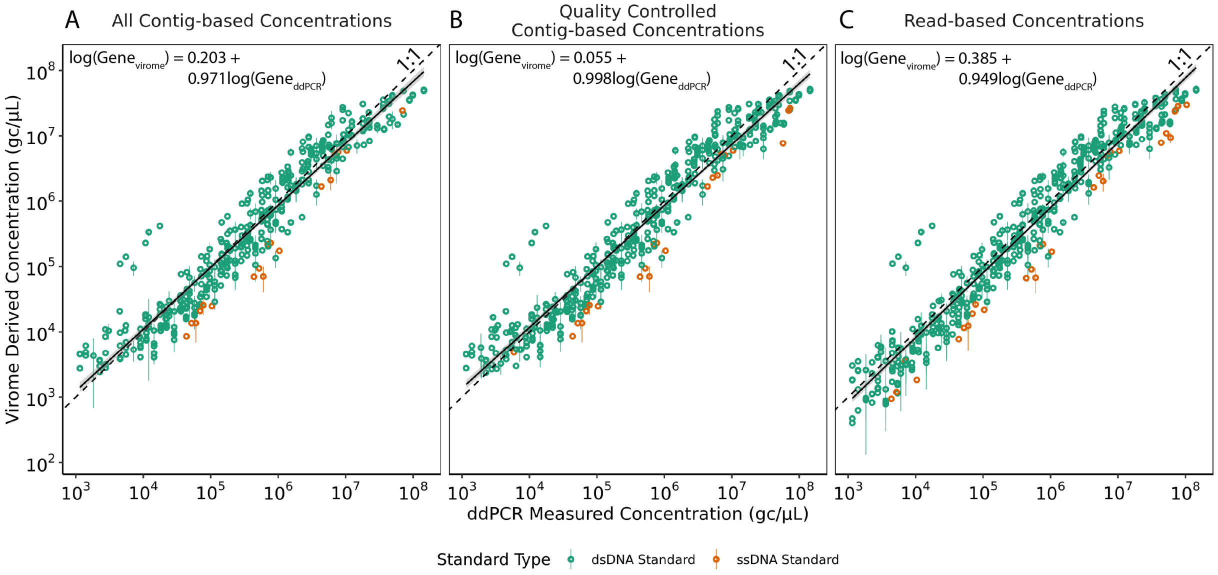

Figure 3.5: ddPCR measured concentrations of spike-in standards are compared to the virome derived concentrations in gc/μL of DNA extract. 95% confidence intervals are reported with error bars for the three technical replicates from the 12/21/21 influent and 12/22/21 effluent samples. Green and orange circles represent dsDNA and ssDNA standards, respectively. The dashed line indicates the ideal 1:1 relationship if the virome derived concentrations completely agree with the known concentrations. The solid black line and reported regressions are the observed relationship between virome derived and ddPCR measured concentrations of standards with the 95% confidence intervals in grey. (A) Results from all contig-based concentrations after correction (R^2 = 0.92), (B) results from quality controlled contig-based concentrations after correction (R^2 = 0.93), and (C) results from read-based concentrations (R^2 = 0.94).

3.6.3 Viral abundances through wastewater treatment

To test our quantitative approach against a standard molecular quantification method, we quantified the concentration of the spiked-in marine phage HM1 in influent and effluent viromes using our quantitative metagenomics approach and also with ddPCR. We tested both quantification with reads mapped to HM1-derived contigs (contig-based) as well as reads mapped to its known genome (read-based). The mean HM1 concentrations in the DNA extracts based on ddPCR were 3.03 $\log_{10}$ gc/μL (Table B.9). Only three of ten samples had a HM1-derived contig present whereas read-based HM1was above detection thresholds in all samples and quantifiable in eight of ten samples demonstrating read-based measurements have lower detection and quantification thresholds than contig-based measurements. When quantifiable the, read-based HM1 quantification resulted in concentrations that were similar to those measured with ddPCR. Specifically, the read-based values were only 1% lower than those measured with ddPCR in influent samples and 8.4% higher than those values measured in effluent. This aligns with previous work that observed qPCR measurements to be 22% greater than spike-in quantitative metagenomic measurements[11]. These results highlight the higher detection limit of the quantitative metagenomics but suggest that when they are above the detection threshold, the quantities measured are similar to those measured with other molecular quantification techniques.

The quantitative metagenomics approach was applied to quantify the total abundance of DNA virus populations in the wastewater samples. The measured concentrations were higher in effluent than influent (*p*-value = 0.12) with mean concentrations of 10.9 $\log_{10}$ gc/mL in effluent (s.d. = 0.16 $\log_{10}$ gc/mL) and 10.6 $\log_{10}$ gc/mL in influent (s.d. = 0.14 $\log_{10}$ gc/mL) (Figure B.6). Previous studies have used epifluorescent microscopy and flow cytometry to measure virus-like particle (VLP) concentrations in influent and effluent and reported concentration ranges of 8.00-8.85 $\log_{10}$ VLP/mL and 8.00-8.60 $\log_{10}$ VLP/mL in influent and effluent, respectively[3, 46, 47]. The 1-2 orders of magnitude higher total viral concentrations measured with quantitative metagenomics compared to previous measurements may be due to epifluorescent microscopy and flow cytometry not capturing all viruses, including ssDNA[48]. Epifluorescent microscopy differs from dsDNA and ssDNA qPCR measurements by one and two orders of magnitude, respectively[48]. Likewise, flow cytometry significantly underestimates ssDNA phage concentrations and dsDNA quantification is highly staining temperature dependent where staining at 30˚C underestimates dsDNA phage concentrations by an order of magnitude[48]. Despite values that are 1-2 orders of magnitude higher than those reported previously, we suspect the metagenomics approach also underestimated total DNA virus concentrations due to limitations in identifying all viral contigs and sequencing depth constraints missing rarer viral populations.

CrAssphage and JC and BK polyomaviruses were detected and quantified in influent and effluent samples as they are ubiquitous in municipal wastewater (Table 3.1). Polyomavirus concentrations in effluent samples traversed detection thresholds in our quantitative metagenomes and were therefore only quantifiable in some samples. CrAssphage was quantifiable in every sample. When polyomaviruses or crAssphage were quantifiable with both the read-based and contig-based approaches, the results were not statistically different between the two approaches for the low abundance polyomaviruses (*p*-value = 0.40 and 0.098 for JC and BK polyomaviruses, respectively), but the contig-based quantification was significantly higher for crAssphage (*p*-value = 0.0048). The crAssphage genomes were not able to undergo quantification correction due to their long lengths, we expect that the difference between contig and read-based quantification will be reduced after the 49-bp sliding window code is corrected.

Mean crAssphage measurements in influent and effluent were 7.71 $\log_{10}$ gc/mL (3/3 samples) and 6.68 $\log_{10}$ gc/mL (3/3 samples), respectively. This is consistent with previous qPCR measurements of crAssphage in influent and effluent that ranged from 1.84-9.03 $\log_{10}$ gc/mL and 2.75-6.00 $\log_{10}$ gc/mL, respectively[49-55]. As our values were on the higher end of those concentrations reported previously with qPCR, we tested for potential biases caused by qPCR primer specificity by limiting crAssphage virome derived quantification to NCBI accession numbers captured with the CPQ056 primer set[39]. Limiting quantification to the crAssphage CPQ056 primer set resulted in lower levels of total crAssphage concentrations (*p*-value = 0.016).

Influent and effluent samples contained mean JC polyomavirus concentrations of 4.57 $\log_{10}$ gc/mL (3/3 samples) and 3.20 $\log_{10}$ gc/mL (1/3 samples), respectively. Similarly, mean BK polyomavirus concentrations in influent and effluent were 4.53 $\log_{10}$ gc/mL (3/3 samples) and 2.93 $\log_{10}$ gc/mL (1/3 samples), respectively. This is similar to previous measurements of JC and BK polyomaviruses in influent and effluent that ranged from 0.95-5.30 $\log_{10}$ gc/mL and 1.43-2.70 $\log_{10}$ gc/mL[51, 53-56].

Table 3.1: Concentrations of crAssphages (total crAssphages and CPQ056 primer specific crAssphages) and JC and BK polyomaviruses were measured in each sample by mapping contigs and reads to genomes from RefSeq. Concentrations are reported as log10 gc/mL of wastewater. (n.d. = not detected)

Sample	**Concentration ($\log_{10}$ gc/mL wastewater)**			
	crAssphage/CPQ056		**JC/BK Polyomavirus**	
	Contig-based	**Read-based**	**Contig-based**	**Read-based**
12/19/20 Influent	7.80/7.13	7.33/6.62	4.73/4.92	4.76/4.80
12/21/20 Influent	7.57/6.99	7.17/6.47	4.39/4.54	4.51/4.50
12/23/20 Influent	7.72/6.98	7.23/6.51	n.d./4.49	4.32/4.29
12/20/20 Effluent	6.71/6.64	5.82/5.08	n.d./n.d.	n.d./n.d.
12/22/20 Effluent	6.82/6.70	5.88/5.17	n.d./n.d.	n.d./2.93
12/24/20 Effluent	6.41/6.31	5.56/4.85	n.d./n.d.	3.20/n.d.

3.7 Conclusions

We developed a quantitative metagenomic method, QuantMeta, to confidently and accurately quantify targets in metagenomes using information from synthetic DNA standards. The method was applied to three influent and three effluent wastewater viral communities. We found a detection limit of approximately 500 gc/µL with a sequencing depth of approximately 200 million reads and applying our method to detect and correct non-specific mapping and assembly errors improved quantification accuracy. Furthermore, we demonstrated that read-based and contig-based quantification resulted in statistically similar concentrations and aligned with ddPCR measured concentrations of standards and marine phage HM1. Our measurements of DNA virus populations in influent and effluent yielded higher concentrations than previously reported concentrations measuring viral-like particle counts. Despite only testing this method on viromes, the approach is applicable to other types of metagenomes although future work will evaluate its performance on whole metagenomes. The current method may be constrained to Illumina NovaSeq SP flowcells with 251-bp paired-end reads. The method should be evaluated with other sequencing technologies and read lengths as sequencing technology may alter read depth variability regressions and RMSE limits. Quantitative metagenomics merges the benefits of metagenomics and quantitative PCR methods to provide unparalleled insight into the composition and abundance of microbiomes. Future work will apply this method to evaluate dynamics of viral communities through wastewater treatment and may be applied to other environments and types of metagenomes.

3.8 References

1. Gregory, A.C. et al. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109-1123 e1114 (2019).
2. Petrovich, M.L. et al. Microbial and Viral Communities and Their Antibiotic Resistance Genes Throughout a Hospital Wastewater Treatment System. *Front Microbiol* **11**, 153 (2020).
3. Rosario, K., Nilsson, C., Lim, Y.W., Ruan, Y. & Breitbart, M. Metagenomic analysis of viruses in reclaimed water. *Environ Microbiol* **11**, 2806-2820 (2009).
4. Roux, S. et al. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* **7**, e33641 (2012).

5. Subirats, J., Sanchez-Melsio, A., Borrego, C.M., Balcazar, J.L. & Simonet, P. Metagenomic analysis reveals that bacteriophages are reservoirs of antibiotic resistance genes. *Int J Antimicrob Agents* **48**, 163-167 (2016).
6. Tamaki, H. et al. Metagenomic analysis of DNA viruses in a wastewater treatment plant in tropical climate. *Environ Microbiol* **14**, 441-452 (2012).
7. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N.A. Phage puppet masters of the marine microbial realm. *Nat Microbiol* **3**, 754-766 (2018).
8. Trubl, G. et al. Towards optimized viral metagenomes for double-stranded and single-stranded DNA viruses from challenging soils. *PeerJ* **7**, e7265 (2019).
9. Hardwick, S.A. et al. Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. *Nat Commun* **9**, 3096 (2018).
10. Reis, A.L.M. et al. A universal and independent synthetic DNA ladder for the quantitative measurement of genomic features. *Nat Commun* **11**, 3609 (2020).
11. Crossette, E. et al. Enhancing metagenomic quantification of genes in environmental samples with internal standards. *mBio* **in press** (2021).
12. Lin, Y., Gifford, S., Ducklow, H., Schefield, O. & Cassar, N. Towards Quantitative Microbiome Community Profiling Using Internal Standards. *Applied and Environmental Microbiology* **85**, e02634-02618 (2019).
13. Satinsky, B.M., Gifford, S.M., Crump, B.C. & Moran, M.A. Use of internal standards for quantitative metatranscriptome and metagenome analysis. *Methods Enzymol* **531**, 237-250 (2013).
14. Stammler, F. et al. Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome* **4**, 28 (2016).
15. Jian, C., Luukkonen, P., Yki-Jarvinen, H., Salonen, A. & Korpela, K. Quantitative PCR provides a simple and accessible method for quantitative microbiota profiling. *PLoS One* **15**, e0227285 (2020).
16. Ju, F. et al. Antibiotic resistance genes and human bacterial pathogens: Co-occurrence, removal, and enrichment in municipal sewage sludge digesters. *Water Res* **91**, 1-10 (2016).
17. Majeed, H.J. et al. Evaluation of Metagenomic-Enabled Antibiotic Resistance Surveillance at a Conventional Wastewater Treatment Plant. *Front Microbiol* **12**, 657954 (2021).
18. Vandeputte, D. et al. Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507-511 (2017).
19. Starke, R., Pylro, V.S. & Morais, D.K. 16S rRNA Gene Copy Number Normalization Does Not Provide More Reliable Conclusions in Metataxonomic Surveys. *Microb Ecol* **81**, 535-539 (2021).
20. Galazzo, G. et al. How to Count Our Microbes? The Effect of Different Quantitative Microbiome Profiling Approaches. *Front Cell Infect Microbiol* **10**, 403 (2020).
21. Kong, Z. et al. Bacterial ecosystem functioning in organic matter biodegradation of different composting at the thermophilic phase. *Bioresour Technol* **317**, 123990 (2020).
22. Santini, N.S. et al. Natural and Regenerated Saltmarshes Exhibit Similar Soil and Belowground Organic Carbon Stocks, Root Production and Soil Respiration. *Ecosystems* **22**, 1803-1822 (2019).
23. Armbruster, D. & Pry, T. Limit of Blank, Limit of Detection and Limit of Quantitation. *Clinical Biochemistry Review* **29** (2008).

24. Forootan, A. et al. Methods to determine limit of detection and limit of quantification in quantitative real-time PCR (qPCR). *Biomol Detect Quantif* **12**, 1-6 (2017).
25. Shen, J., McFarland, A.G., Young, V.B., Hayden, M.K. & Hartmann, E.M. Toward Accurate and Robust Environmental Surveillance Using Metagenomics. *Front Genet* **12**, 600111 (2021).
26. Castro, J.C. et al. imGLAD: accurate detection and quantification of target organisms in metagenomes. *PeerJ* **6**, e5882 (2018).
27. Tithi, S.S., Aylward, F.O., Jensen, R.V. & Zhang, L. FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data. *PeerJ* **6**, e4227 (2018).
28. Langenfeld, K., Chin, K., Roy, A., Wigginton, K. & Duhaime, M.B. Comparison of ultrafiltration and iron chloride flocculation in the preparation of aquatic viromes from contrasting sample types. *PeerJ* **9**, e11111 (2021).
29. Lievens, A., Jacchia, S., Kagkli, D., Savini, C. & Querci, M. Measuring Digital PCR Quality: Performance Parameters and Their Optimization. *PLoS One* **11**, e0153317 (2016).
30. Brown, B.L., Watson, M., Minot, S.S., Rivera, M.C. & Franklin, R.B. MinION nanopore sequencing of environmental metagenomes: a synthetic approach. *Gigascience* **6**, 1-10 (2017).
31. Giordano, F. et al. De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Sci Rep* **7**, 3935 (2017).
32. Warwick-Dugdale, J. et al. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ* **7**, e6800 (2019).
33. Roux, S., Enault, F., Hurwitz, B.L. & Sullivan, M.B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
34. Guo, J. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
35. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).
36. Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A. & Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).
37. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* **39**, 578-585 (2021).
38. Roux, S. (GitHub, 2015).
39. Stachler, E. et al. Quantitative CrAssphage PCR Assays for Human Fecal Pollution Measurement. *Environ Sci Technol* **51**, 9146-9154 (2017).
40. Kuhl, M.A., Stich, B. & Ries, D.C. Mutation-Simulator: fine-grained simulation of random mutations in any genome. *Bioinformatics* **37**, 568-569 (2021).
41. Zhou, Z., Luhmann, N., Alikhan, N.-F., Quince, C. & Achtman, M. Accurate Reconstruction of Microbial Strains Using Representative Reference Genomes. *bioRxiv* (2017).
42. Browne, P.D. et al. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. *Gigascience* **9** (2020).

43. Ekblom, R., Smeds, L. & Ellegren, H. Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria. *BMC Genomics* **15** (2014).
44. Hsieh, P.H., Oyang, Y.J. & Chen, C.Y. Effect of de novo transcriptome assembly on transcript quantification. *Sci Rep* **9**, 8304 (2019).
45. Commichaux, S. et al. A critical assessment of gene catalogs for metagenomic analysis. *Bioinformatics* **37**, 2848-2857 (2021).
46. Ma, L. et al. Rapid quantification of bacteria and viruses in influent, settled water, activated sludge and effluent from a wastewater treatment plant using flow cytometry. *Water Sci Technol* **68**, 1763-1769 (2013).
47. Wu, Q. & Liu, W.T. Determination of virus abundance, diversity and distribution in a municipal wastewater treatment plant. *Water Res* **43**, 1101-1109 (2009).
48. Kaletta, J. et al. A rigorous assessment and comparison of enumeration methods for environmental viruses. *Sci Rep* **10**, 18625 (2020).
49. Ahmed, W. et al. Evaluation of the novel crAssphage marker for sewage pollution tracking in storm drain outfalls in Tampa, Florida. *Water Res* **131**, 142-150 (2018).
50. Ahmed, W., Payyappat, S., Cassidy, M., Besley, C. & Power, K. Novel crAssphage marker genes ascertain sewage pollution in a recreational lake receiving urban stormwater runoff. *Water Res* **145**, 769-778 (2018).
51. Crank, K. et al. CrAssphage abundance and correlation with molecular viral markers in Italian wastewater. *Water Res* **184**, 116161 (2020).
52. Garcia-Aljaro, C., Balleste, E., Muniesa, M. & Jofre, J. Determination of crAssphage in water samples and applicability for tracking human faecal pollution. *Microb Biotechnol* **10**, 1775-1780 (2017).
53. Malla, B., Ghaju Shrestha, R., Tandukar, S., Sherchand, J.B. & Haramoto, E. Performance Evaluation of Human-Specific Viral Markers and Application of Pepper Mild Mottle Virus and CrAssphage to Environmental Water Samples as Fecal Pollution Markers in the Kathmandu Valley, Nepal. *Food Environ Virol* **11**, 274-287 (2019).
54. Tandukar, S., Sherchan, S.P. & Haramoto, E. Applicability of crAssphage, pepper mild mottle virus, and tobacco mosaic virus as indicators of reduction of enteric viruses during wastewater treatment. *Sci Rep* **10**, 3616 (2020).
55. Wu, Z., Greaves, J., Arp, L., Stone, D. & Bibby, K. Comparative fate of CrAssphage with culturable and molecular fecal pollution indicators during activated sludge wastewater treatment. *Environ Int* **136**, 105452 (2020).
56. McQuaig, S.M., Scott, T.M., Lukasik, J.O., Paul, J.H. & Harwood, V.J. Quantification of human polyomaviruses JC Virus and BK Virus by TaqMan quantitative PCR and comparison to other water quality indicators in water and fecal samples. *Appl Environ Microbiol* **75**, 3379-3388 (2009).

Chapter 4 Viral Community Dynamics Through Wastewater Treatment Using Quantitative Metagenomics

4.1 Abstract

We conducted an in-depth exploration of the viral community at a wastewater treatment plant in Michigan, U.S.A. by implementing a rigorous quantitative viromic method. By evaluating the community composition, functional potential, and abundance of viruses in influent and secondary effluent wastewater, we demonstrated how the viral community alters after biological treatment. We found that our influent and secondary effluent communities were significantly different from each other and highly diverse compared to previous wastewater viromes. Many viral populations were removed during biological treatment, but a few highly abundant viral populations became more abundant in secondary effluent, which we attribute to their ability to replicate during biological treatment. Additionally, we identified viral-associated ARGs in influent and secondary effluent. However, viral-associated ARGs were rare occurrences and influent viruses containing ARGs did not persist through biological treatment.

4.2 Introduction

Wastewater treatment removes carbon and nutrients from raw sewage to lower oxygen demand of effluent thereby protecting receiving surface waters from eutrophication. However, wastewater also contains highly abundant and diverse microbial communities that include viruses. Pathogenic viruses have been well characterized in influent and secondary effluent wastewater[1-5]. Few studies have explored viruses in wastewater beyond pathogens which may have indirect impacts on the outcomes of wastewater treatment and thus the ecology of receiving water environments.

Viral communities in wastewater are primarily composed of phages (i.e., viruses that infect bacteria)[6-9], which have been reported as five to seven-fold more abundant than bacteria in wastewater[10]. Phage-host interactions impact the structure and function of wastewater microbial communities due to their high abundance and complex host interactions. Additionally, viruses may hijack hosts' metabolisms and express their own auxiliary metabolic genes and horizontally transfer genes via transduction[11, 12]. Previous metagenomic studies have demonstrated that phages impact bulking and foaming issues within activated sludge during biological treatment[13, 14]. While a number of studies are now taking a whole viral community perspective through the use of viral metagenomics (i.e., viromics)[6-9, 13, 15-17] or sorting viral sequences from whole metagenomes[14, 18-20], most of the immense diversity of viruses and their abundances in wastewater communities remains undescribed and unknown[6, 8, 9, 18]. Further, these studies were not quantitative, not as highly purified for viruses, and not sequenced with as much depth as our viromes, which limits their potential to deeply probe and accurately describe viral community composition, structure, and functional potential.

To address these challenges and expand upon existing characterizations of wastewater viromes, here we implemented a new quantitative viromics (Chapter 3) approach that allowed for direct comparisons of influent and secondary effluent dsDNA and ssDNA viral communities. Using rigorous viral purification methods, high sequencing depth, and supplementing the typical short read sequencing with a long read sequencing technology to improve assemblies, we were able to quantitatively evaluate the impacts of wastewater treatment on viral community structure and metabolic gene content.

4.3 Methods

4.3.1 Sample collection and preparation

Influent and secondary effluent wastewater samples were collected and prepared for sequencing as described previously (Chapter 3). Briefly, three influent and secondary effluent grab samples were collected from automatic samples from the Ann Arbor wastewater treatment plant (Ann Arbor, MI) between December 19 and 24, 2020. Samples underwent an ultrafiltration and purification method[21] to concentrate and purify viruses. DNA extraction was performed with QIAamp UltraSens Virus Kit (QIAGEN, Cat. No. 53706) according to manufacturer's instructions except for the first six steps, in which 140 µL of sample was combined with 5.6 µL carrier RNA and vortexed briefly. DNA extracts were stored at -20°C. A 20 L deionized "blank" water sample was processed with the same viral concentration and purification process to account for contamination during sample processing.

4.3.2 Quantitative Illumina NovaSeq sequencing

A set of sequins metagenome mix A dsDNA standards[22] and 5 ssDNA standards (Chapter 3) were spiked into each DNA extract prior to library preparations as described previously (Chapter 3). The ssDNA standards were prepared to match the range of concentrations in the sequins metagenome mix A with 10^7, 10^4, 10^6, 10^8, 10^5 gc/µL of the 31.7, 40, 45, 50, and 60% GC content ssDNA standards, respectively. Mean spike-in concentrations were $6.16\text{x}10^5$ gc/µL of DNA extract and $6.81\text{x}10^5$ gc/µL of DNA extract for sequin S1106_MG_020_A and 45% GC ssDNA standard, respectively, as measured with ddPCR assays (Chapter 3). Illumina NovaSeq libraries were prepared with the Accel-NGS® 1S Plus DNA Library Kit (Swift Biosciences, Cat. No. 10024) using 50 ng DNA. The blank sample failed library preparations and quality thresholds for sequencing due to inadequate DNA concentrations, so contamination was determined to be negligent. Samples were sequenced on the Illumina NovaSeq 600 on two 500 cycle SP flow cells yielding 251-bp long paired-end reads. Library preparations and sequencing were performed by the Advanced Genomics Core at the University of Michigan. Reads were quality controlled by trimming Illumina adaptors plus an additional 15 bp from the rightmost and leftmost of each read to remove Accel-NGS® 1S Plus DNA Library Kit adaptors, and decontaminated of PhiX174 with BBDuk (BBTools, v37.64).

4.3.3 De novo assemblies utilizing Oxford Nanopore GridION sequencing

Illumina NovaSeq sequencing was supplemented with long read sequencing to improve viral assemblies because viral genomes typically contain repetitive regions, high mutation rates, and host genome fragments[23-25]. DNA extracts without standards were cleaned with the Zymo Genomic DNA Clean and Concentrate-10 kit (Cat. No. D4011, Zymo Research Corporation), then libraries were prepared with the Ligation Sequencing Kit (Cat. No. SQK-LSK109, Oxford Nanopore Technologies) and barcoded with the Native Barcoding Expansion 1-12 (Cat. No. EXP-NBD104, Oxford Nanopore Technologies). Samples were sequenced on two flow cells (R9.4.1, Cat. No. FLO-MIN106, Oxford Nanopore Technologies). Basecalling was performed with Guppy (v4.2.3) and called reads were classified as pass or fail depending on their mean quality score (≥ 7). Library preparations, sequencing, and basecalling were conducted by the Advanced Genomics Core at the University of Michigan. Each sample underwent two *de novo* assemblies: hybrid co-assembly and long read only assembly. Hybrid co-assemblies with long and short reads for each sample were performed with metaSPAdes (v3.15.2) using kmer sizes of 21, 33, 55, 77, 89, and 127. Long read assemblies for each sample were performed with Flye (v2.8.3) followed by polishing with four rounds of Racon (v1.4.10), one round of medaka (v1.3.2), and pilon (v1.24)[25]. Following assemblies, contigs greater than 1,000-bp from each sample were pooled.

4.3.4 Creating viral populations from contigs

Viral populations were determined based on the likelihood of contigs being of viral origin. Contigs originating from viruses were identified by running VirSorter (v1.0.5)[26] with the virome flag, VirSorter2 (v2.2.2)[27], VIBRANT (v1.2.1)[28] with the virome flag, VirFinder (v1.1)[29], and CheckV (v0.7.0)[30] end-to-end. High confidence viral contigs were any contig with a VirSorter score 1, 2, or 4, VirSorter2 score greater than 0.9, VirFinder score greater than 0.9, VIBRANT "complete", "high", or "medium" quality or included in the "prophage" list, and included in the CheckV "prophage" list. Low confidence viral contigs were any contig with a VirSorter score of 3, 5, or 6, VirSorter2 score between 0.5 and 0.9, VirFinder score between 0.7 and 0.9, and VIBRANT "low quality". Low confidence viral contigs were further analyzed by running with CheckV end-to-end. Any low confidence viral contig containing one or more viral genes or no host genes were determined to be of viral origin. Viral contigs were dereplicated by retaining the longest contigs in clusters sharing 95% ANI and at least 70% coverage assessed with clustergenomes.pl (v5.1)[31]. 16.5-23.3% of viral contigs were removed by dereplication. The remaining dereplicated viral contigs represent individual viral populations. Each viral population was classified as dsDNA or ssDNA based on the VirSorter2 assignment. Non-viral contigs content was assessed with Blastn (v2.9.0) using default parameters to the nt database (downloaded 6/15/2019). Blast alignments were considered significant if more than 50% of the contig aligned to the database target and shared more than 80% identity.

4.3.5 Quantification of viral populations

Viral populations' concentrations were determined with QuantMeta (Chapter 3) based on ddPCR measured concentrations of standards. Detection thresholds and regressions from (Chapter 3) were used for quantification. Quantification correction was not performed because creating 49-bp sliding windows for all viral contigs became a computational bottleneck.

4.3.6 Alpha diversity analysis

Shannon's diversity index was calculated for each sample based on the concentrations and RPKM of viral populations using vegan (v2.5-7). The three technical replicates of 12/21/20 influent and 12/22/20 secondary effluent were distilled for Shannon's diversity analysis by calculating mean concentrations or RPKMs of each viral population. Richness was calculated as the number of viral populations in each sample. All samples were summarized in a PCoA plot based on concentrations using Bray Curtis dissimilarity with capscale from the R vegan package (v2.5-7), then visualized with ggplot2 (v3.3.5).

4.3.7 Functional potential assessment

The metabolic potential of viral populations was assessed with Distilled and Refined Annotation of Metabolism (DRAM, v1.2.4)[32] to genes in the KEGG and Pfam databases (installed 9/29/21). Open reading frames were annotated, then distilled with default parameters for each separate sample. Viral-associated metabolism genes were summarized by summing concentrations of genes within each KEGG ontology level 2 group. Viral-associated ARGs were identified by running DeepARG (v1.0.2)[33] on the viral populations from each sample separately. For technical replicates of the 12/21/20 influent and 12/22/20 secondary effluent samples, mean concentrations were calculated for each gene with not detected values removed.

4.3.8 Taxonomic assignment

Taxonomic assignment of phages was assigned with vConTACT2 (v0.9.15)[34] with the Viral RefSeq database (v.85, January 2018) using open reading frames created by the DRAM annotate function with all of the viral populations from all of the samples combined. vConTACT2 created a network based on the portion of genes shared between different clusters that was amended by removing RefSeq reference genomes without direct connections to any wastewater viral populations. The resulting network contained 10,794 nodes and 115,972 edges and was visualized with Cytoscape (v3.9.0). Nodes were colored based on if a population was specific to influent, effluent, both influent and effluent, or if it was a RefSeq reference genome.

4.3.9 Statistical analysis

Statistical analysis was performed in R (v4.0.3). Unpaired t-tests were performed with the R stats package (v4.0.3) using 0.95 confidence levels with *p*-values less than 0.05 considered significant. All graphs were created with ggplot2 (v3.3.5) and heatmaps were created with ComplexHeatmap (v2.6.2).

4.3.10 Data availability

All sequencing data will be available at JGI upon submission of this work to bioRxiv.

4.4 Results and Discussion

4.4.1 High quality quantitative viromes indicate viruses in wastewater more abundant than previously reported

The influent and secondary effluent viral metagenomes proved to be highly purified, with 75.5-78.0% of the contigs identified as viral. The remaining contigs were predominately unclassified with 19.9% of the contigs lacking a significant match to any sequence in the NCBI nt database (downloaded 6/15/2019) and 2.2% and 0.6% classified as bacteria and plasmids, respectively. The high degree of viral sequence representation in the viromes of this study contrasts with other wastewater viromes, where an estimated 1-18% of reads were classified as viral[8, 17] and 80-85% of reads classified as bacterial[8]. Furthermore, an analysis of 1,445 viromes found that contigs originating from viruses ranged 1-60% of contigs per sample[35]. We attribute the high ratio of viral to non-viral contigs in our sequenced dataset to the extensive effort invested in the development of robust concentration and purification steps prior to sequencing[21].

Viruses in influent and secondary effluent communities were highly abundant and diverse, as compared to previous measurements of total viral particles[36-38] and Shannon's diversity index[8, 20] in wastewater. Mean total virus concentrations were 10.6 and 10.9 $\log_{10}$ gene copies (gc)/mL of wastewater liquid fractions in influent and secondary effluent, respectively (Chapter 3). These values are two orders of magnitude higher than previous studies that quantified virus-like particle (VLP) measurements using epifluorescent microscopy and flow cytometry and reported between 8.0-8.9 $\log_{10}$ VLP/mL[36-38].

Given that libraries were prepared to intentionally capture ssDNA and dsDNA viruses[39], we were able to compare the absolute abundances of these viral types. dsDNA phages were significantly more abundant than ssDNA in influent and secondary effluent (*p*-values = 0.033 and 0.035, respectively) accounting for 92.3% and 92.8% of the viral populations in influent and secondary effluent, respectively. We suspect that the disparity between dsDNA and ssDNA viruses is two-fold. First, given that the majority of *in silico* viral contig sorting methods available for our contig curation step reference and are trained on databases that are substantially biased towards dsDNA viruses (ssDNA phages represent approximately 11% of phages in the ICTV database[40]; and include four families with 124 genomes in the RefSeq viral database[41]), ssDNA viruses are likely not as readily recognized, thus not assigned as viral and missing from downstream analyses. Second, the chloroform purification step employed to reduce the cellular DNA contamination of the wastewater virome may have reduced recovery of ssDNA viruses. Of the ssDNA viruses recognized in the ICTV database, 64% are classified as *Inoviridae*, which are known to be sensitive to chloroform[42]. Loss of Inoviruses due to chloroform treatment may explain why previous studies found *Inoviridae* to be highly abundant in influent[6], whereas no *Inoviridae* were identified in our viromes. However, this discrepancy may also be explained by the use of MDA in previous influent virome studies, which is known to be heavily biased towards the amplification of ssDNA viruses[43]. While the use of a chloroform treatment greatly improves viral recoveries when concentrating samples with dead-end ultrafiltration and removes non-viral DNA[21], given that chloroform is known to reduce the recovery of some viruses by removing lipid envelopes or impacting the structure of filamentous viruses[44-46], studies aimed at capturing chloroform-sensitive viruses should employ alternate methods to reduce cellular DNA contamination.

4.4.2 Wastewater harbors highly diverse viral communities that differ between influent and effluent

Despite originating from the same wastewater influent source community, the influent and secondary effluent viral communities are distinct in both alpha and beta diversity (Figure 4.1). Influent richness was higher than secondary effluent with mean viral population counts of 203,000 and 76,400 in influent and secondary effluent, respectively (*p*-value = 0.00698). Previous wastewater viromes have captured less richness, ranging from 50-10,000 populations[8, 15, 20]. The Shannon's diversity index, which captures both richness and evenness, was also higher in influent, with mean values of 11.0 and 9.50 for influent and secondary effluent, respectively (*p*-value = $1.44\text{x}10^{-4}$). Shannon's diversity indexes calculated with relative abundances (e.g., RPKM; as is the approach taken by nearly all prior virome studies) were slightly lower than those calculated with absolute abundances, with mean values of 10.8 and 9.29 for influent and secondary effluent, respectively. Shannon's diversity indices of viral communities from suspended growth municipal wastewater treatment plants in Wisconsin and Singapore were roughly half of the values calculated in our study, ranging from 3-4 and 4.5-5.5 in influent and secondary effluent, respectively[8, 20]. However, because sequencing depth can influence richness measures, the low viral richness of previous studies, which ranged from 50-600 viral populations[8, 20], was likely due to the five to ten times lesser sequencing effort that led to lower Shannon's diversity estimates. Viruses from viromes and whole metagenomes collected within activated sludge reactors in Hong Kong resulted in Shannon's diversity index values from 5.22-7.89[13, 14], which is likely similar to secondary effluent diversity because activated sludge and secondary effluent were previously reported to have a greater than 49% similarity[8]. The observed higher diversity and richness compared to previous studies is likely because our sequencing effort was substantially higher and more of it dedicated to viral genome content given the exceptionally low contamination of non-viral DNA. Additionally, recent advances in viral prediction have improved our ability to sort contigs of viral origin[26-30], thus leading to higher richness and Shannon's diversity measures in our samples.

Influent and effluent viral community structures were significantly different (PERMANOVA; Figure 4.1C). When Bray-Curtis dissimilarity measures were used to evaluate differences in sample to sample viral community structure, the influent and secondary effluent communities separated along the first principle component of a principal component analysis (PCoA), which explained 48% of the variability in the dataset (Figure 4.1C). To evaluate the nature of these community-level differences, we next evaluated how specific populations, their relatedness, and their functional gene repertoire differed between the influent and effluent viral communities.

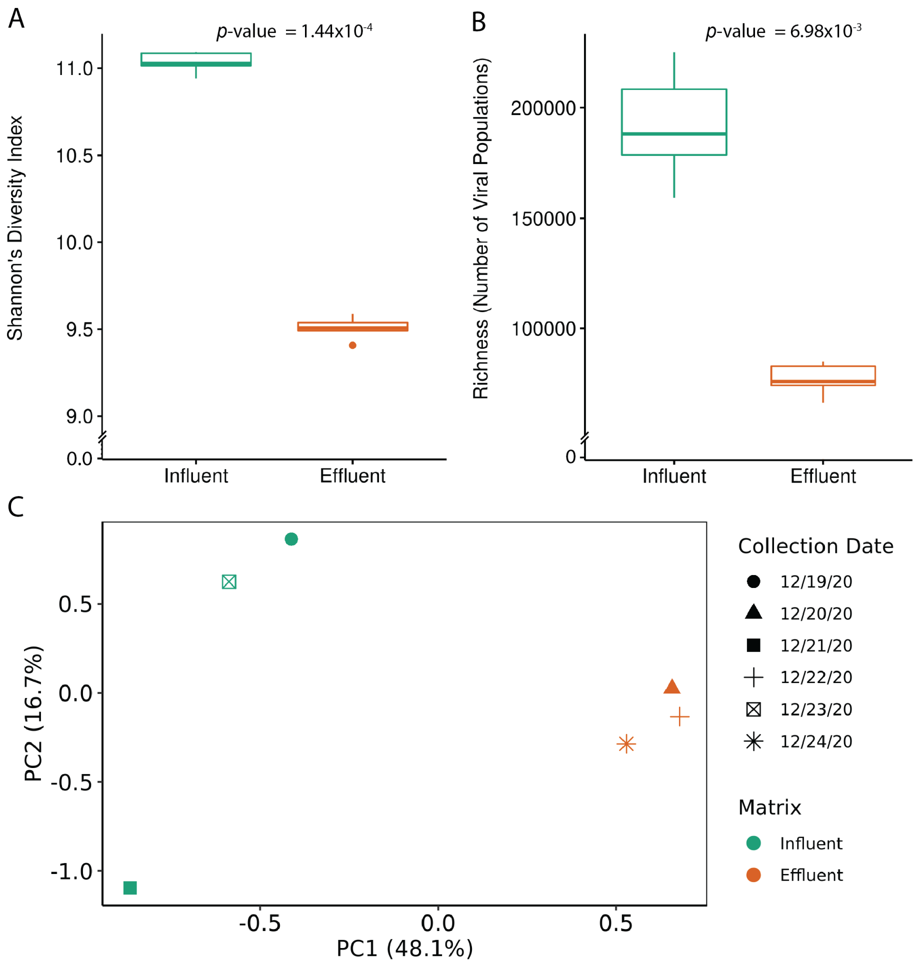

Figure 4.1: (A) Shannon's diversity index boxplot of influent and secondary effluent samples. (B) Richness boxplot of influent and secondary effluent samples. (C) Ordination of PCoA representing the Bray Curtis dissimilarity between wastewater viral community structures. Color indicates matrix (green for influent, orange for effluent) and shape corresponds to collection date.

4.4.3 Most viral populations are novel and removed during biological treatment

Most wastewater virus populations were novel and not annotated by vConTACT2 (Figure 4.2A), a tool that integrates distance-based hierarchical clustering to determine virus taxonomy. However, owing to its underlying reference dataset, taxonomic assignment with vConTACT2 is limited to viruses that infect bacteria and archaea, therefore any human or plant viruses present in influent or secondary effluent samples were not assigned taxonomy with this method. The viral populations assigned a taxonomy by vConTACT2 represented only 0.79% and 0.33% of total viral concentrations in influent and secondary effluent, respectively (Figure 4.2A-B). These results demonstrated how the sole reliance on databases of known viral genomes to identify viruses based on sequence similarity will miss a substantial portion, in some cases *nearly all*, of the viral populations that may be present. Previous virome studies from a variety of environments have similarly noted that a significant portion of viruses found in metagenomes are novel[47-49]. These trends serve as a caution to expand viral detection methods beyond genome sequence alignment based (e.g., blast) approaches and they support the need for continued viral biodiversity discovery in novel environments to broaden the global scope of known viral diversity.

As is documented in a number of studies of viral biogeography[47, 50] the vast majority of viral populations were not-ubiquitous (Figure 4.2B), with 84.3-95.6% of viral populations observed in only one sample. Of the populations found in multiple samples, the degree of ubiquity was not equivalent in influent and effluent samples. Largely due to the overall reduction in viral load from influent to effluent, a greater proportion, 37%, of the effluent viral populations were also present in the influent, whereas only 12.9% of influent populations were present in effluent. This could be explained by a selective loss of influent viral populations that may be physically sensitive to the treatment process, in combination with kill-the-winner dynamics[51], whereby the influent viruses, as well as endemic effluent viruses, whose hosts also persist through the treatment process, are able to continue replicating and maintain high abundance in the effluent. This observation of few viral populations conserved in influent and secondary effluent is at odds with a previous study[8] where 82% of viral populations were conserved throughout wastewater treatment. This discrepancy is likely due to advances both in sequencing that allow for greater depth and in algorithms used to identify viruses from metagenomic datasets, both of which allowed us to capture many low abundance viral populations only present in influent samples (Figure 4.2D) that may have been missed with shallower sequencing. The viruses conserved between influent and secondary effluent were highly abundant accounting for 52.3% and 83.6% of total virus concentrations in influent and secondary effluent, respectively (Figure 4.2D). The absolute abundance of conserved viruses in influent and secondary effluent were 10.3 $\log_{10}$ gc/mL more abundant in secondary effluent indeed suggested that highly abundant viruses are able to replicate during biological treatment. Previous work has also observed viruses to persist from influent into secondary effluent, as well as secondary effluent endemicity[8, 20].

When shared gene-content was considered to distinguish the evolutionary relatedness of the viral populations, the influent-specific viral populations clustered with those found in secondary effluent (Figure 4.2C), indicating the viruses that persist through wastewater treatment were not genomically distinct from those lost from the original influent. This suggests that factors, such as host availability or robustness to conditions through wastewater treatment, may determine whether viruses persist[52]. The most abundant viruses in influent continued to dominate in secondary effluent viral communities indicating virus concentration may be a predictor of persistence. Future work examining host ranges of viral populations will determine if these highly abundant and persistent viral populations are able to replicate throughout wastewater treatment to maintain high concentrations due to their ability to infect multiple hosts.

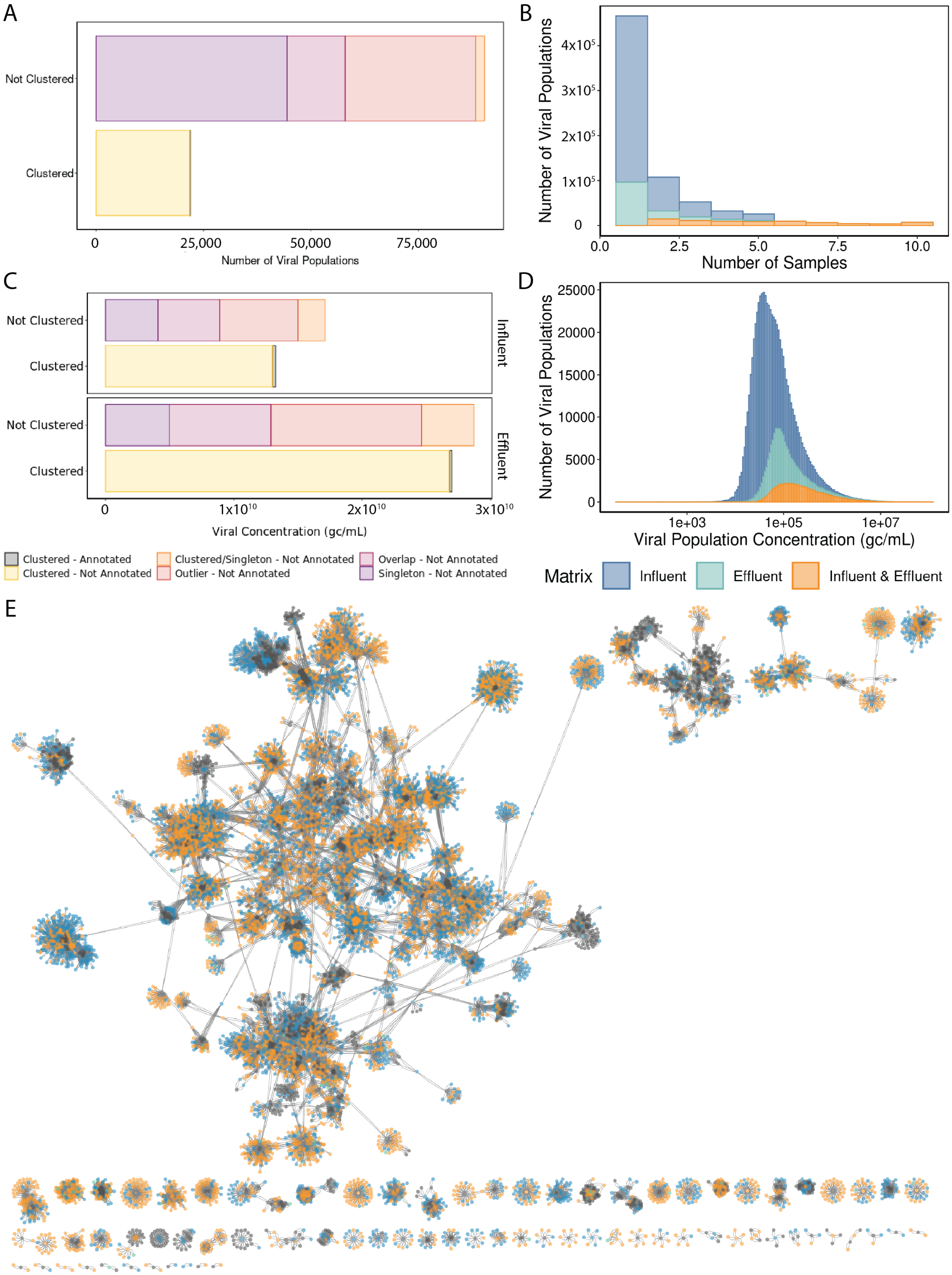
A
Not Clustered
Clustered
0
25,000
50,000
75,000
Number of Viral Populations
B
Number of Viral Populations
$4x10^5$
$3x10^5$
$2x10^5$
$1x10^5$
0
0.0
2.5
5.0
7.5
10.0
Number of Samples
C
Not Clustered
Clustered
Influent
Not Clustered
Clustered
Effluent
0
$1x10^{10}$
$2x10^{10}$
$3x10^{10}$
Viral Concentration (gc/mL)
D
Number of Viral Populations
25000
20000
15000
10000
5000
0
1e+03
1e+05
1e+07
Viral Population Concentration (gc/mL)
Clustered - Annotated
Clustered - Not Annotated
Clustered/Singleton - Not Annotated
Outlier - Not Annotated
Overlap - Not Annotated
Singleton - Not Annotated
Matrix
Influent
Effluent
Influent & Effluent
E

Figure 4.2: (A) vContact clustering and annotation results reported as number of viral populations and (B) concentration in wastewater (gc/mL). (C) Number of samples each viral population was found in and (D) the mean concentration (gc/mL) of each viral population for all samples the population was quantified in. Colors indicate if the population was found in influent only (blue), secondary effluent only (green), or both matrices (orange). (E) Cytoscape network analysis of viral populations with all samples. Node color indicates if the viral population is present in influent (blue), secondary effluent (green), or both steps of wastewater treatment (orange). Grey nodes indicate closely related viruses with reference genomes present in vConTACT2's database, but not specifically found in our wastewater samples.

4.4.4 Wastewater treatment does not alter the distribution of virus-encoded metabolic genes

Influent and secondary effluent viral populations contained similar metabolic genes, despite containing distinct viral populations (Figure 4.3). The most abundant genes were associated with nucleotide metabolism, with genes involved in some pathways of xenobiotic biodegradation and metabolism and cofactor and vitamin metabolism also highly abundant (Figure 4.3). Other studies have similarly observed nucleotide metabolism to be highly prevalent in wastewater virus communities[8, 16, 18], though this trend is not universally observed in other environments[8, 48, 53, 54]. The high abundance of nucleotide metabolism genes in wastewater viral populations may be due to the higher biomass and microbial activity in wastewater, as compared to well-studied systems like the open ocean, that could lead to higher rates of viral replication that might select for viral genes supporting the formation of nucleotide precursor molecules required for DNA replication[55].

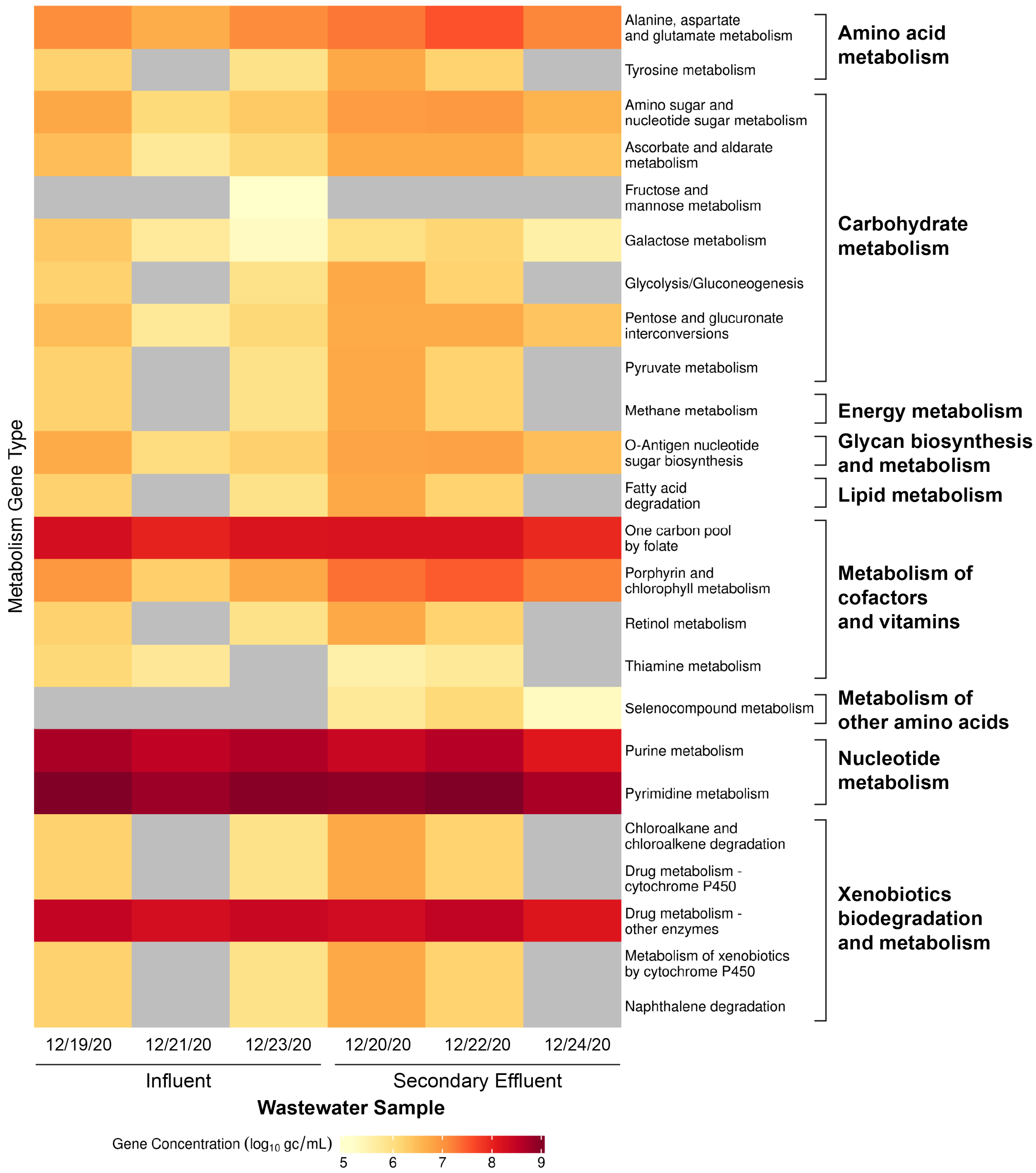

Figure 4.3: KEGG ontology heatmap for metabolism genes identified in the viral populations organized into metabolism categories for each influent and secondary effluent. Gene concentrations ($\log_{10}$ gc/mL) are represented with a color gradient where light yellow and dark red indicate low and high concentrations, respectively. Not detected is indicated in grey.

4.4.5 Viral-associated antibiotic resistance genes are rare and do not persist through wastewater treatment

Motivated by previous estimates that viruses contain 0.001-0.1% of ARGs in an environmental sample[56], we sought to apply our quantitative viromics method to evaluate the prevalence of virus-encoded ARGs on a whole community-level. Of the high confidence viral contigs that contained bonafide viral genes, the majority of antibiotic resistance genes (ARGs) identified by DeepARG[33] were dihydrofolate reductase inhibitors, comprising 7.00 and 7.61 $\log_{10}$ gc/mL of viral-associated ARGs in influent and secondary effluent, respectively. Because dihydrofolate reductase genes are known to be involved in conferring resistance to dihydrofolate reductase inhibitor (e.g., trimethoprim) in bacteria by reducing how much dihydrofolate is needed for DNA replication[57], dihydrofolate reductase genes are identified by DeepARG as ARGs. However, viral-associated dihydrofolate reductase genes have been long identified as auxiliary metabolic genes involved in nucleotide metabolism[58-60], rather than for conferring antibiotic resistance. Therefore, we did not consider viral-associated dihydrofolate reductase genes to confer antibiotic resistance and caution against future virome analyses interpreting the presence of this enzyme as support for viruses as genomic vectors of ARGs. Of what were deemed bonafide virus-encoded ARGs, those of the influent samples were found to be more diverse than those in the secondary effluent, a phenomenon previously observed in other metagenomes [61]. This phenomenon is unique to ARGs and not observed for metabolic genes. Further, the distribution of ARG functions in the influent and secondary effluent viral-associated resistomes suggest similar forces may shape their distributions, as influent and effluent ARG profiles distinctly cluster (Figure 4.4).

Concentrations of bonafide viral-encoded ARGs determined with our quantitative viromes were consistent with concentrations previously reported via other methods. Total virus-encoded ARG concentrations were 6.26 $\log_{10}$ gc/mL and 5.89 $\log_{10}$ gc/mL in influent and secondary effluent, respectively, and did not significantly differ (*p*-value = 0.131). Reported ARG concentrations measured in whole wastewater microbial communities with high throughput qPCR have ranged from 7.5 to 10 $\log_{10}$ gc/mL for the sum of all PCR targets[62]. The most abundant class of ARGs in our wastewater analysis confers resistance to vancomycin. VanS and vanU concentrations ranged from 5.18-6.18 and 4.44-4.93 $\log_{10}$ gc/mL, respectively, for samples with vanS and vanU above detection. While vanS and vanU have not previously been measured in wastewater, vanA and vanB concentrations have been reported to range 0-4 $\log_{10}$ gc/mL[63, 64]. Further, vancomycin resistant bacteria have been detected in wastewater samples at concentrations ranging 1.33-3.76 CFU/mL[65, 66]. Recent studies found β-lactamase, vancomycin, tetracycline resistance genes on contigs with phage structural genes and phage integrases[67, 68]. Likewise, we identified contigs with phage genes and ARGs present.

Despite the undeniable detection of ARGs encoded on viral contigs, the question remains "are viruses a meaningful reservoir of ARGs?". Previous studies have argued that viruses rarely carry ARGs[7, 19, 69]. Of these studies that investigated wastewater viromes[7, 19], they generated less than 10 Gb of sequences per sample, approximately 6-10 times less sequencing depth than our viromes. Even with this greater sequencing depth, we found that only 59 contigs (0.008% of all viral populations) contained ARGs, strongly supporting the claim that virus-encoded ARGs in wastewater are rare occurrences. Even though not many ARGs are detected on contigs, we still do not know the role or prevalence of vesicle-packaged ARGs that were previously found to be more prevalent than virus-encoded ARGs[19].

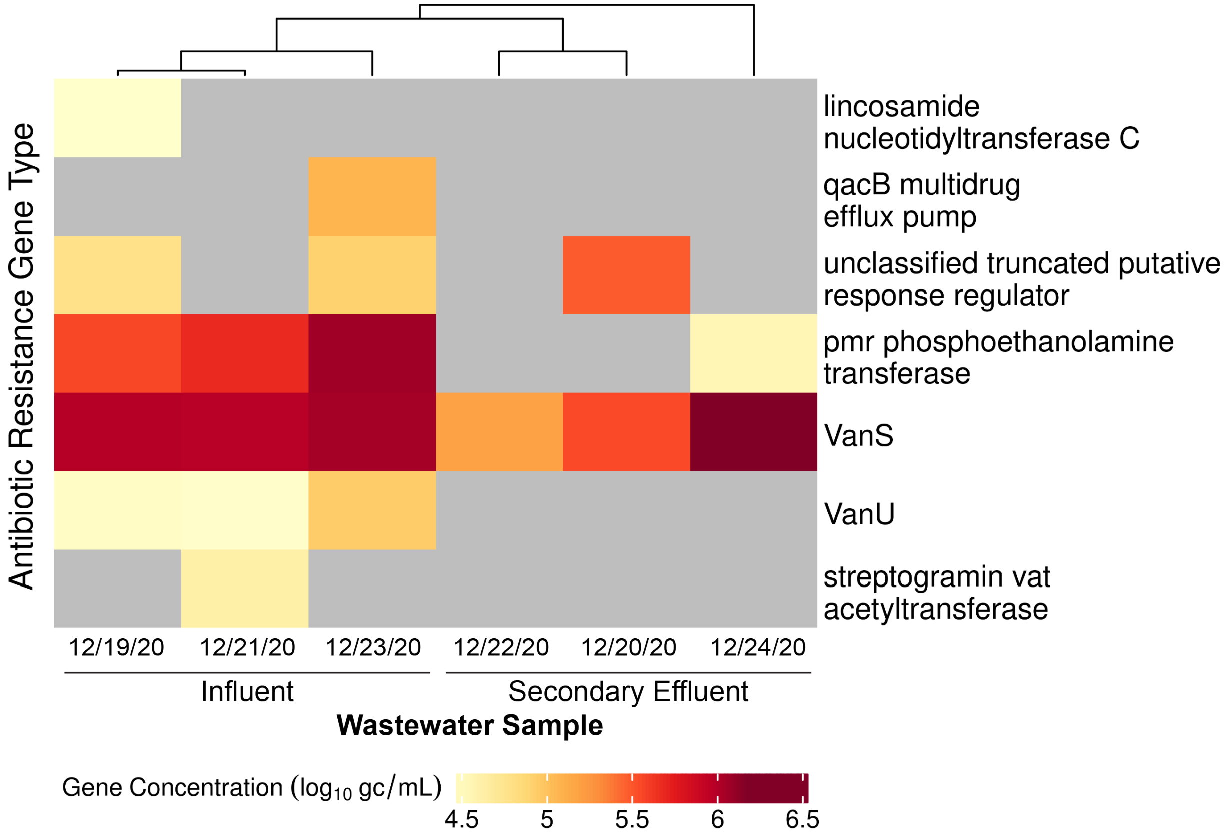

Figure 4.4: Virus-associated ARG heatmap for genes identified by DeepARG in the viral populations organized by resistance gene type. Gene concentrations ($\log_{10}$ gc/mL) are represented with a color gradient where light yellow and dark red indicate low and high concentrations, respectively. Not detected is indicated in grey. A dendrogram arranges samples based on similarity of ARG concentrations.

4.5 Conclusions

Our quantitative metagenomic analysis of influent and secondary effluent wastewater viruses demonstrated that these viral communities are highly diverse, novel, and abundant. Further, most influent virus populations do not persist through wastewater treatment. Rather, select virus populations are likely able to replicate during wastewater treatment making them able to persist at greater abundances in secondary effluent. We demonstrated that metabolic functional potential does not alter significantly from influent to secondary effluent, but that, while viral-associated ARGs are rare across the total viral community, the distribution of ARG functions does change from influent to secondary effluent. Of the small number of ARGs identified on viral contigs, vancomycin resistance genes were among the most abundant and were identified in both influent and secondary effluent.

4.6 References

1. Bhatt, A., Arora, P. & Prajapati, S.K. Occurrence, fates and potential treatment approaches for removal of viruses from wastewater: A review with emphasis on SARS-CoV-2. *J Environ Chem Eng* **8**, 104429 (2020).
2. Corpuz, M.V.A. et al. Viruses in wastewater: occurrence, abundance and detection methods. *Sci Total Environ* **745**, 140910 (2020).
3. Haramoto, E. et al. A review on recent progress in the detection methods and prevalence of human enteric viruses in water. *Water Res* **135**, 168-186 (2018).
4. Ibrahim, Y. et al. Detection and removal of waterborne enteric viruses from wastewater: A comprehensive review. *Journal of Environmental Chemical Engineering* **9** (2021).
5. Saawarn, B. & Hait, S. Occurrence, fate and removal of SARS-CoV-2 in wastewater: Current knowledge and future perspectives. *J Environ Chem Eng* **9**, 104870 (2021).
6. Cantalupo, P.G. et al. Raw sewage harbors diverse viral populations. *mBio* **2** (2011).
7. Petrovich, M.L. et al. Microbial and Viral Communities and Their Antibiotic Resistance Genes Throughout a Hospital Wastewater Treatment System. *Front Microbiol* **11**, 153 (2020).
8. Tamaki, H. et al. Metagenomic analysis of DNA viruses in a wastewater treatment plant in tropical climate. *Environ Microbiol* **14**, 441-452 (2012).
9. Wang, Y., Jiang, X., Liu, L., Li, B. & Zhang, T. High-Resolution Temporal and Spatial Patterns of Virome in Wastewater Treatment Systems. *Environ Sci Technol* **52**, 10337-10346 (2018).
10. Du, B. et al. Responses of bacterial and bacteriophage communities to long-term exposure to antimicrobial agents in wastewater treatment systems. *J Hazard Mater* **414**, 125486 (2021).
11. Breitbart, M. Marine viruses: truth or dare. *Ann Rev Mar Sci* **4**, 425-448 (2012).
12. Paterson, S. et al. Antagonistic coevolution accelerates molecular evolution. *Nature* **464**, 275-278 (2010).
13. Chen, Y., Wang, Y., Paez-Espino, D., Polz, M.F. & Zhang, T. Prokaryotic viruses impact functional microorganisms in nutrient removal and carbon cycle in wastewater treatment plants. *Nat Commun* **12**, 5398 (2021).
14. Yang, Q., Zhao, H. & Du, B. Bacteria and bacteriophage communities in bulking and non-bulking activated sludge in full-scale municipal wastewater treatment systems. *Biochemical Engineering Journal* **119**, 101-111 (2017).
15. Adriaenssens, E.M. et al. Viromic Analysis of Wastewater Input to a River Catchment Reveals a Diverse Assemblage of RNA Viruses. *mSystems* **3** (2018).
16. Osunmakinde, C., Selvarajan, R., Mamba, B. & Msagati, T. Viral Communities Distribution and Diversity in a Wastewater Treatment Plants Using High-throughput Sequencing Analysis. *Polish Journal of Environmental Studies* **30**, 3189-3201 (2021).

17. Wang, H. et al. Variations among Viruses in Influent Water and Effluent Water at a Wastewater Plant over One Year as Assessed by Quantitative PCR and Metagenomics. *Appl Environ Microbiol* **86** (2020).
18. Gulino, K. et al. Initial Mapping of the New York City Wastewater Virome. *mSystems* **5**, e00876-00819 (2020).
19. Maestre-Carballa, L. et al. Insights into the antibiotic resistance dissemination in a wastewater effluent microbiome: bacteria, viruses and vesicles matter. *Environ Microbiol* **21**, 4582-4596 (2019).
20. Petrovich, M.L. et al. Viral composition and context in metagenomes from biofilm and suspended growth municipal wastewater treatment plants. *Microb Biotechnol* (2019).
21. Langenfeld, K., Chin, K., Roy, A., Wigginton, K. & Duhaime, M.B. Comparison of ultrafiltration and iron chloride flocculation in the preparation of aquatic viromes from contrasting sample types. *PeerJ* **9**, e11111 (2021).
22. Hardwick, S.A. et al. Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. *Nat Commun* **9**, 3096 (2018).
23. Brown, B.L., Watson, M., Minot, S.S., Rivera, M.C. & Franklin, R.B. MinION nanopore sequencing of environmental metagenomes: a synthetic approach. *Gigascience* **6**, 1-10 (2017).
24. Giordano, F. et al. De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Sci Rep* **7**, 3935 (2017).
25. Warwick-Dugdale, J. et al. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ* **7**, e6800 (2019).
26. Roux, S., Enault, F., Hurwitz, B.L. & Sullivan, M.B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
27. Guo, J. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
28. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).
29. Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A. & Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).
30. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* **39**, 578-585 (2021).
31. Roux, S. (GitHub, 2015).
32. Shaffer, M. et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res* **48**, 8883-8900 (2020).
33. Arango-Argoty, G. et al. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* **6** (2018).
34. Bolduc, B. et al. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ* **5**, e3243 (2017).
35. Zolfo, M. et al. Detecting contamination in viromes using ViromeQC. *Nat Biotechnol* **37**, 1403-1412 (2019).

36. Ma, L. et al. Rapid quantification of bacteria and viruses in influent, settled water, activated sludge and effluent from a wastewater treatment plant using flow cytometry. *Water Sci Technol* **68**, 1763-1769 (2013).
37. Rosario, K., Nilsson, C., Lim, Y.W., Ruan, Y. & Breitbart, M. Metagenomic analysis of viruses in reclaimed water. *Environ Microbiol* **11**, 2806-2820 (2009).
38. Wu, Q. & Liu, W.T. Determination of virus abundance, diversity and distribution in a municipal wastewater treatment plant. *Water Res* **43**, 1101-1109 (2009).
39. Roux, S. et al. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* **4**, e2777 (2016).
40. Szekely, A.J. & Breitbart, M. Single-stranded DNA phages: from early molecular biology tools to recent revolutions in environmental microbiology. *FEMS Microbiol Lett* **363** (2016).
41. Koonin, E.V. & Dolja, V.V. in Encyclopedia of Virology, Edn. 4 (Elsevier, Ltd. , Oxford, United Kingdom; 2021).
42. Roux, S. et al. Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nat Microbiol* **4**, 1895-1906 (2019).
43. Kim, K.H. & Bae, J.W. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl Environ Microbiol* **77**, 7663-7668 (2011).
44. Kauffman, K.M. et al. A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature* **554**, 118-122 (2018).
45. Petrenko, V., Smith, G., Gong, X. & Quinn, T. A library of organic landscapes on filamentous phage. *Protein Engineering* **9**, 797-801 (1996).
46. Taniguchi, H., Sato, K., Ogawa, M., Udou, T. & Mizuguchi, Y. Isolation and characterization of a filamentous phage, Vf33, specific for Vibrio parahaemolyticus. *Microbiology and immunology* **28**, 327-337 (1984).
47. Gregory, A.C. et al. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109-1123 e1114 (2019).
48. Hegarty, B. et al. A Snapshot of the Global Drinking Water Virome: Diversity and Metabolic Potential Vary with Residual Disinfectant Use. *bioRxiv* (2021).
49. Trubl, G. et al. Towards optimized viral metagenomes for double-stranded and single-stranded DNA viruses from challenging soils. *PeerJ* **7**, e7265 (2019).
50. Roux, S. et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689-693 (2016).
51. Winter, C., Bouvier, T., Weinbauer, M.G. & Thingstad, T.F. Trade-offs between competition and defense specialists among unicellular planktonic organisms: the "killing the winner" hypothesis revisited. *Microbiol Mol Biol Rev* **74**, 42-57 (2010).
52. Liu, R. et al. Bacteriophage ecology in biological wastewater treatment systems. *Appl Microbiol Biotechnol* **105**, 5299-5307 (2021).
53. Coutinho, F.H. et al. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nat Commun* **8**, 15955 (2017).
54. Jian, H. et al. Diversity and distribution of viruses inhabiting the deepest ocean on Earth. *ISME J* **15**, 3094-3110 (2021).
55. Hatfull, G.F. & Hendrix, R.W. Bacteriophages and their genomes. *Curr Opin Virol* **1**, 298-303 (2011).

56. Debroas, D. & Siguret, C. Viruses as key reservoirs of antibiotic resistance genes in the environment. *ISME J* **13**, 2856-2867 (2019).
57. Levings, R.S., Lightfoot, D., Elbourne, L.D., Djordjevic, S.P. & Hall, R.M. New integron-associated gene cassette encoding a trimethoprim-resistant DfrB-type dihydrofolate reductase. *Antimicrob Agents Chemother* **50**, 2863-2865 (2006).
58. Asare, P.T. et al. Putative type 1 thymidylate synthase and dihydrofolate reductase as signature genes of a novel Bastille-like group of phages in the subfamily Spounavirinae. *BMC Genomics* **16**, 582 (2015).
59. Kozloff, L.M., Verses, C., Lute, M. & Crosby, L.K. Bacteriophage Tail Components: Dihydrofolates reductase in T4D bacteriophage. *Journal of Virology* **5**, 740-753 (1970).
60. Mathews, C.K. Evidence That Bacteriophage-induced Dihydrofolate Reductase Is a Viral Gene Product. *Journal of Biological Chemistry* **242**, 4083-4086 (1967).
61. Martin, C. et al. Nanopore-based metagenomics analysis reveals prevalence of mobile antibiotic and heavy metal resistome in wastewater. *Ecotoxicology* **30**, 1572-1585 (2021).
62. An, X.L. et al. Tracking antibiotic resistome during wastewater treatment using high throughput quantitative PCR. *Environ Int* **117**, 146-153 (2018).
63. Furukawa, T., Hashimoto, R. & Mekata, T. Quantification of vancomycin-resistant enterococci and corresponding resistance genes in a sewage treatment plant. *J Environ Sci Health A Tox Hazard Subst Environ Eng* **50**, 989-995 (2015).
64. Le, T.H., Ng, C., Tran, N.H., Chen, H. & Gin, K.Y. Removal of antibiotic residues, antibiotic resistant bacteria and antibiotic resistance genes in municipal wastewater by membrane bioreactor systems. *Water Res* **145**, 498-508 (2018).
65. Blanch, A.R. et al. Comparison of enterococcal populations related to urban and hospital wastewater in various climatic and geographic European regions. *Journal of Applied Microbiology* **94**, 994-1002 (2003).
66. Caplin, J.L., Hanlon, G.W. & Taylor, H.D. Presence of vancomycin and ampicillin-resistant Enterococcus faecium of epidemic clonal complex-17 in wastewaters from the south coast of England. *Environ Microbiol* **10**, 885-892 (2008).
67. Li, X. et al. Metagenomic and viromic data mining reveals viral threats in biologically treated domestic wastewater. *Environmental Science and Ecotechnology* **7** (2021).
68. Moon, K. et al. Freshwater viral metagenome reveals novel and functional phage-borne antibiotic resistance genes. *Microbiome* **8**, 75 (2020).
69. Enault, F. et al. Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J* **11**, 237-247 (2017).

Chapter 5 Parasite-host Coevolutionary Impacts on the Emergence of Antibiotic Resistance

5.1 Abstract

Antibiotic resistance is a public health threat and phage therapy is a potential alternative or supplement to antibiotics to treat infections. However, rapid bacteria evolution to phage infection challenges the efficacy of phage therapy. Phage-host coevolution is known to accelerate evolution and its impact on the emergence and expression of antibiotic resistance is not well understood. Therefore, we conducted evolution experiments with Avida to test if phage-host coevolution and antibiotics exert compounding effects on antibiotic resistance. We found that phage-host coevolution accelerated the emergence of antibiotic resistance. In some experiments, pleiotropy of evolving phage infection defenses resulted in decreased susceptibility to antibiotics. The results demonstrate that phage-host coevolution should be considered when determining the efficacy of supplementing antibiotics with phage therapy to treat infections.

5.2 Introduction

Infections with antibiotic resistant bacteria are a pressing public health threat that are predicted to cause as many as 444 million deaths by 2050[1]. Phage therapy is a potential alternative or adjuvant to antibiotics[2]. In clinical applications, phage therapy treats infections by targeting a pathogen with a lytic phage, or a virus that infects bacteria, that takes over the cellular machinery to create tens-to-thousands of new infectious viral particles and lyses the pathogen. Phage therapy-like applications are also being considered for wastewater treatment, surface disinfection, aquaculture, and the food industry to control bacteria populations that have deleterious effects[3, 4]. For example, wastewater treatment relies on biological treatment that can suffer from certain deleterious bacteria that cause foaming and dewaterability issues which may be resolved with phage therapy to remove the deleterious bacteria[3]. However, a challenge phage therapy faces is the bacteria's potential to rapidly evolve defense mechanisms countering phage infection[5]. Because evolution has the potential to undermine design goals, outcomes of phage-host coevolution are an important consideration in the efficacy of phage therapy applications. We implemented Avida experiments to study antibiotic resistance evolution in the presence of phage-host coevolution.

The parasitic relationship between phages and bacteria accelerates evolution of phage and bacteria as each races to evolve new mechanisms of infection and defense[6]. In well-mixed chemostat-like environments, such as wastewater biological treatment or the gut, bacteria are exposed to antibiotics while growing in a dense microbial community with a high abundance of phage. The dynamics between phage and bacteria may impact the emergence and prevalence of antibiotic resistance in such environments[5, 7]. Bacteria-phage coevolution often exhibits Red Queen interactions[6], where competing organisms constantly evolve, adapt, and proliferate to survive in environments with an evolving opposition. Antagonistic coevolution is predicted to drive evolutionary changes in bacteria and phage as demonstrated in controlled laboratory communities and soil microbiomes[6, 8, 9]. This phenomenon was also observed *in silico* using digital organisms and parasites in Avida, software that generates self-replicating and evolving computer programs (i.e., digital organisms), in a user-defined environment[10].

Avida provides a controlled setting to efficiently observe evolution for thousands of generations with organisms that inherit and mutate traits[11]. Avida digital organisms are self-replicating and evolving computer programs[11], similar to bacteria in laboratory experiments. The organisms are capable of replicating their genome to create a new digital organism. Mutation generates diversity in the population and users define the probability of mutations occurring in offspring. The evolution of genes, or logic tasks, in digital organisms provides instances of evolution where everything occurring in the environment is known and measurable. Organisms compete for CPU cycles where more CPU cycles allow organisms to replicate quicker. Parasites, acting like phages, may be injected into Avida experiments to infect organisms performing a task encoded on a parasite's genome resulting in the parasite overtaking the organism's CPU cycles and effectively slowing down the organism's replication. Parasites may use all or some of a host's CPU cycles therefore killing or sickening the host.

We hypothesized that in the presence of antibiotics, phage-host coevolution accelerates the evolution of antibiotic resistance. To address the hypothesis, we created an Avida environment that simulated sub-inhibitory concentrations of an antibiotic with a specific logic task selected to confer antibiotic resistance. The emergence and fraction of digital organisms expressing resistance were compared to runs without antibiotics or phage present. The results demonstrated a compounding effect of coevolution and antibiotics on the emergence and prevalence of resistance. The presence of parasites accelerated the evolution of antibiotic resistance and, occasionally, organisms evolving in an environment with antibiotics and parasites resulted in a decreased susceptibility to antibiotics. Potential outcomes of phage-host coevolution should be considered when assessing the efficacy of phage therapy applications.

5.3 Methods

5.3.1 Avida configuration

Experiments were performed with Avida (v2.14.0) using conditions similar to Zaman et al. (2014). The environment mimicked a well-mixed chemostat containing a beneficial resource with an initial abundance of 125 units, a steady inflow of 125 units, and 20% removal with each update. Organisms consumed the beneficial resource by performing any logic task (Not, Not-And, And, Or-Not, Or, And-Not, Not-Or, Exclusive-Or, Equals). All beneficial resource reactions awarded organisms the same merit benefits with enzyme type rewards that mimic Michaelis-Menton kinetics. Organisms were required to perform a task prior to reproducing. Parasites were able to infect organisms if organisms performed a function in common and infecting parasites acquired 80% of their host's CPU cycles. The environment held a maximum of 14,400 organisms and all experiments were run for 500,000 updates where each organism performs an average of 30 instructions per update. Experiments began with a single organism containing the Not-And task, then 400 parasites containing the Not-And task were added at 2,000 updates. Organisms and parasites acquired new tasks through mutations. Host organisms had a point mutation rate, insertion rate, and deletion rate of 0.000703, 0.00003906, and 0.00003906, respectively. Parasites had a point mutation rate, insertion rate, and deletion rate of 0.005625, 0.000625, and 0.000625, respectively.

5.3.2 Avida environment with a pseudo-antibiotic and potential evolution of resistance

A pseudo-antibiotic, referred to here simply as antibiotic, was introduced to the Avida environment by creating a deleterious resource that decreased an organisms' available CPU cycles after performing a logic task. When organisms encountered antibiotics, performing a task reduced organisms' merit by a specified multiplicity factor where values near one exert a smaller impact than values near zero. A specific logic task was selected to confer resistance, referred to as resistance task. An antibiotic reaction was required to be completed prior to a beneficial resource reaction. The antibiotic reaction multiplicity factor, antibiotic abundance, and resistance task were user-specific inputs that impacted results of experiments. Therefore, experiments were performed with multiplicity factor, antibiotic abundance, and resistance task varied to prevent user-specified inputs limiting the generalizability of conclusions. Multiplicity factors were randomly selected between 0.25 and 0.5 for all tasks except the task selected to confer resistance which had a multiplicity factor of 0.99 to represent the fitness cost of carrying and expressing antibiotic resistance in bacteria. If the resistance task was performed, then other more detrimental antibiotic reactions did not occur. Benefits of evolving new tasks were limited to evading parasite infection or obtaining antibiotic resistance.

5.3.3 Evolution experiments

Control experiments without antibiotics, without parasites, or without both antibiotics and parasites were performed to compare the compounding effects of antibiotics and parasites on the evolution of resistance. Evolution experiments were conducted with six replicates for each set of conditions except for the control without antibiotics or parasites, which was performed in duplicate (Table 5.1). For simulations containing antibiotics, the resistance task was specified as Not-And, And, or Not with the multiplicity factor and antibiotic abundance randomly selected from 0.25-0.50 and 10-15%, respectively. Not-And resistance task experiments served as a control where evolution of resistance did not occur because organisms began each experiment containing the Not-And task. Specific conditions of each simulation are provided in Table C.1.

Table 5.1: Evolution experiments were performed with six replicates for each subset of conditions except for the control with no parasites or antibiotics that had two replicates. The multiplicity factor and antibiotic abundance were randomly varied from 0.25-0.5 and 10-15%, respectively. The specified resistance task was either Not-And,

And, or Not where Not-And served as a no resistance evolution control because the initial organism could perform the Not-And task.

Condition	**Number of Parasites** *(injected at 2,000 updates)*	**Resistance Task**	**Multiplicity Factor**	**Antibiotic Initial and Inflow Abundance** (*percent of food inflow*)
Experiment	400	Not-And	0.25-0.50	10-15%
	400	And	0.25-0.50	10-15%
	400	Not	0.25-0.50	10-15%
Control: No Parasites	0	Not-And	0.25-0.50	10-15%
	0	And	0.25-0.50	10-15%
	0	Not	0.25-0.50	10-15%
Control: No Antibiotics	400	-	-	-
Control: No Parasites or Antibiotics	0	-	-	-

5.3.4 Statistical analysis

Percent resistance was calculated as the number of organisms performing the resistance task was divided by the number of organisms in the environment per update. The update when significant resistance expression occurred was defined as the first update when at least 10% resistance was observed. The results of the experiments were assessed by comparing the arithmetic mean and 95% confidence intervals for each condition and resistance task per update (Figure 5.1). Statistical analysis was performed with R (v4.0.3). Mean and maximum percent resistance across all updates for different conditions were compared with unpaired t-tests using the R stats package (v4.0.3) with 0.95 confidence levels where *p*-values less than 0.05 were considered significant. All graphs were created with ggplot2 (v3.3.5).

5.3.5 Data availability

Configuration and data files will be publicly available on GitHub upon submission of the work to bioRxiv.

5.4 Results

5.4.1 Introduction of environmental stressors reduced the frequency of endemic Not-And resistance expression

Initial organisms and parasites contained the Not-And task whereas organisms had to evolve Not and And tasks. When organisms did not face environmental stressors such as antibiotics or parasites, the organisms did not deviate from Not-And task expression with mean Not-And expression of 97.5% over 500,000 updates and mean And and Not expression of 2.97% and 0.73%, respectively (Figure 5.1). The frequency that Not-And tasks were performed decreased regardless of the task conferring resistance when antibiotics and parasites were present (*p*-values = 0.074 and 0.011, respectively). Parasites and antibiotics had a compounding effect where mean Not-And resistance task expression was less relative to only parasites or antibiotics present (p-values = 0.0021 and $7.73\text{x}10^{-5}$, respectively). When any stressor was added to an environment, organisms evolved new tasks even if the most advantageous task (i.e., Not-And resistance task with antibiotics present) was already endemic in the population.

5.4.2 Parasite-host coevolution accelerated emergence of resistance

Significant antibiotic resistance expression, or the earliest update with at least 10% of organisms expressing resistance, was accelerated by the addition of parasites to experiments. When parasites were present in addition to antibiotics, significant resistance expression occurred on average 54,300 and 134,500 updates earlier for Not and And tasks, respectively (*p*-values = 0.0325 and 0.044, respectively). The introduction of parasites rapidly altered which tasks were performed as organisms evaded infection (Figure 5.1). After parasite introduction, the performance of Not-And tasks decreased rapidly from a mean of 67.2% expression at 2,000 updates to a mean of 50.2% expression at 2,600 updates across all updates containing parasites (*p*-value = 0.00033). When parasites and antibiotics were present, significant resistance expression was not statistically significantly different than when only parasites were present for And and Not resistance tasks (*p*-value = 0.263 and 0.159, respectively). Therefore, antibiotic presence did not alter when resistance emerged with parasites present indicating that parasites were the primary drivers of accelerated resistance emergence in populations.

5.4.3 Magnitude of resistance expression varied by resistance task and was altered by presence of antibiotics and parasites

The task selected to confer resistance impacted the amount that parasites or antibiotics altered resistance expression (Figure 5.1). When resistance tasks were evolved by organisms, mean And resistance task expression increased by 6.4% and 38.1% with parasite or antibiotic addition, respectively, relative to controls without antibiotics or parasites (*p*-value = $6.38\text{x}10^{-4}$). Similarly, mean Not resistance task expression increased by 5.4% and 14.3% with parasite or antibiotic addition, respectively (*p*-value = 0.0214). When And tasks conferred antibiotic resistance, antibiotic presence correlated with significantly greater mean resistance expression whether parasites were present or not (*p*-values = 0.020 and 0.014, respectively). This was not observed when Not tasks conferred antibiotic resistance. However, the presence of parasites significantly increased the maximum observed resistance expression when the Not task conferred resistance whether antibiotics were present or not (p-values = 0.031 and 0.0045, respectively). Therefore, organisms evolving to resist antibiotics and evade parasite infection exhibit pleiotropy that may result in decreased susceptibility to antibiotics.

5.4.4 Multiple stressors decreased parasite extinction occurrence

Parasite extinction always occurred when parasites were the only stressor introduced to an environment (n = 6). Organisms successfully evaded infection to the extent that parasite extinction occurred after an average of 32,950 updates (range = 5,500-163,300 updates). When organisms were challenged by antibiotics and parasites simultaneously, parasite extinction occurred in two-thirds of experiments (n = 12/18). When parasite extinction occurred, organisms faced with antibiotics and parasites did not evade parasitic infection as quickly with parasite extinction occurring after an average of 212,225 updates (range = 700-442,900), a statistically significantly longer time (*p*-value = 0.00267). Additionally, we observed that parasite extinction coincided with decreases in resistance expression. When parasite concentrations significantly decreased, resistance expression also decreased and did not recover to levels observed before parasite abundance decreased (Figure C.1). Lastly, the presence of antibiotics and parasites decreased the average generation present after 500,000 updates. The mean generation after 500,000 updates was 6,110 when parasites and antibiotics were present as opposed to 8,830 when one or neither stressor was present (*p*-value = 1.82×10^{-4}). The presence of multiple stressors made it more difficult for organisms to outcompete parasites which delayed or prevented parasite extinction during simulations.

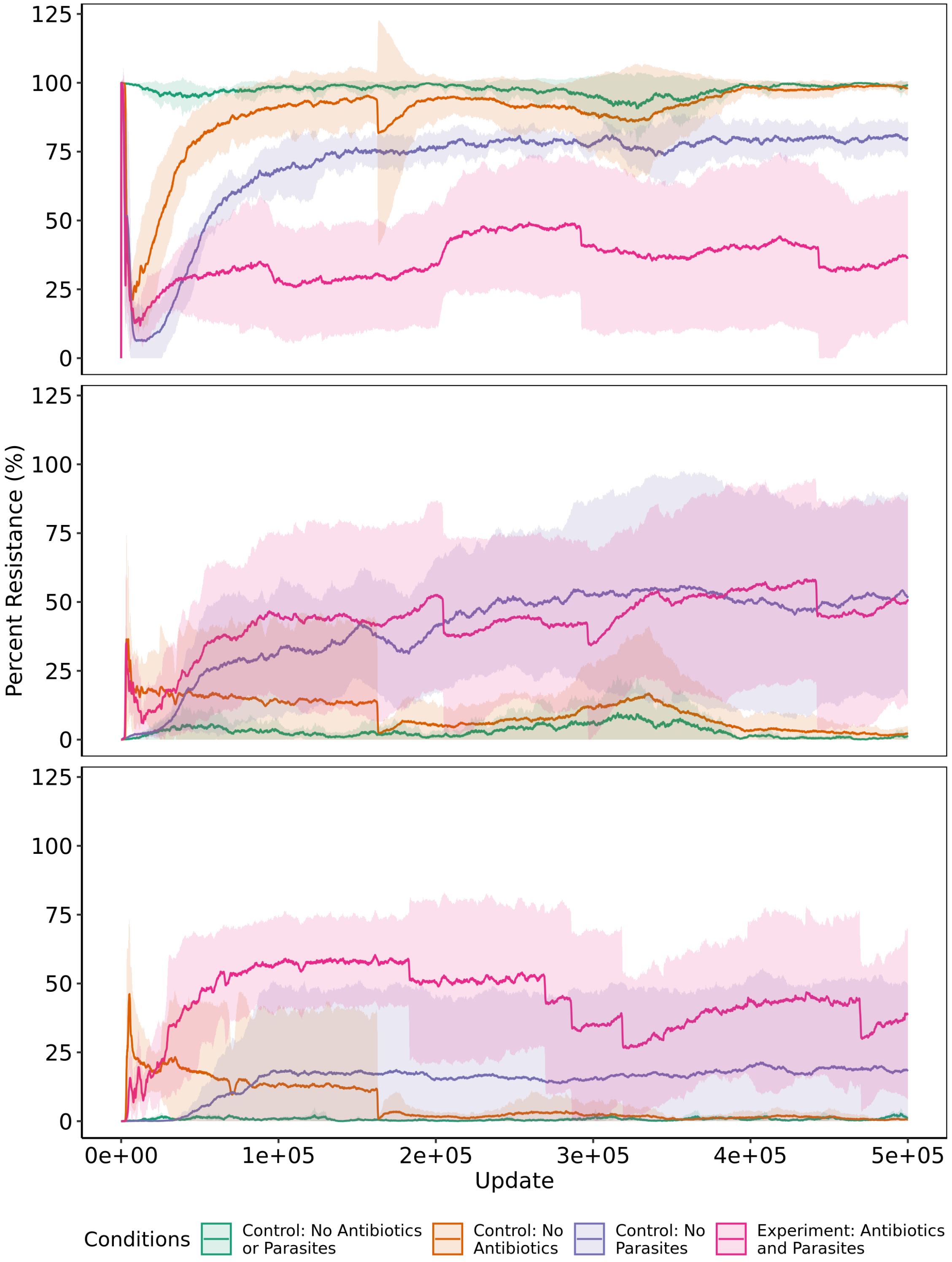

Figure 5.1: Organisms began with Not-And task and evolved And and Not tasks. Each panel represents a different task selected to confer resistance: Not-And (top), And (middle), and Not (bottom). The mean percent resistance with 95% confidence intervals is shown per update where percent resistance is the number of resistance tasks performed per organism. Colors separate the simulation conditions where Control: No Antibiotics or Parasites, Control: No

Antibiotics, Control: No Parasites, Experiment: Antibiotics and Parasites are represented by green, orange, purple, and magenta. All replicate experiments (with all updates including those following parasite extinction) are included.

5.5 Discussion

We observed that parasite introduction to an environment with endemic antibiotic resistance resulted in the organisms becoming more susceptible to antibiotics by decreasing Not-And resistance task expression. Given the rise in antibiotic resistant infections, recent research has focused on using phage therapy to replace or complement antibiotics[2, 12-14]. Phage therapy is challenged by bacteria rapidly evolving defense mechanisms against phage infection. Previous research proposed methods to overcome this challenge such as developing phage cocktails[12], evolving phages prior to treating *in vivo* infections[15], and administering phages and antibiotics together[16]. When phage therapy and antibiotics were administered together, a pleiotropic effect was observed, where increased defenses against phage infection also increased susceptibility to antibiotics[7, 17-19]. The experiments with Not-And tasks conferring resistance also displayed this pleiotropic effect, where the presence of parasites and antibiotics increased susceptibility to the antibiotics.

However, for simulations where antibiotic resistance was evolved, parasites and antibiotics had a compounding effect that accelerated antibiotic resistance emergence and altered antibiotic resistance expression. Previous work has established that parasite-host interactions accelerate evolution by initiating arms race or Red Queen dynamics[6, 8-10]. Parasite-host coevolution allowed organisms to traverse valleys in the adaptive landscape causing rapid diversification of organisms[20, 21]. Occasionally, pleiotropy of evolving defenses to phage infection corresponded with an increased resistance to antibiotics, as was the case for the Not task (Figure 5.1). Pleiotropy was recently observed to occur with an *E. coli* phage that has an infection pathway involving two proteins that confer antibiotic resistance[17]. Some phage resistant

E. coli mutants resulted in increased tetracycline resistance through mutations to the structural barrier molecule lipopolysaccharide[17]. However, this was not always the case and the authors did not observe increased resistance to colistin. Likewise, increased antibiotic resistance expression for evolved tasks, such as Not or And, was not always observed.

In these experiments, parasite infection strategies and antibiotic resistance were linked such that the same tasks conferring resistance to antibiotics may also be used as an infection mechanism for parasites. While this system is true for some phage-host pairs[17, 19, 22], previous work has argued that these interactions are rare[23]. Future experiments will separate parasite infection tasks and antibiotic resistance tasks to evaluate pleiotropic effects in a system that is more prevalent in nature. Additionally, these experiments were always performed with antibiotics preceding parasite introduction. A recent study found that introducing phages prior to antibiotics increases the susceptibility of bacteria to antibiotics[24]. The impacts of altering the order antibiotics and parasites are introduced into an environment will be explored in future work. Furthermore, these simulations do not take into account the community context of natural systems that may have important ecological and evolutionary effects[25]. For example, phage-host interactions in nature range from predatory to symbiotic. Previous research has suggested that piggyback-the-winner dynamics occur in ecosystems with high microbial densities[26, 27] where symbiotic phage-host relationships are prevalent with lysogeny and superinfection exclusion decreasing viral particle concentrations. To elucidate the community context of phage-host coevolution, controlled medium-scale experiments should be performed to determine the generalizability of our conclusions[25].

Here, we demonstrated that phage-host interactions accelerated evolution such that antibiotic resistance emerged sooner than if organisms were only exposed to antibiotics.

Additionally, we found a pleiotropic effect between evolving phage infection defenses and antibiotic resistance that occasionally resulted in increased antibiotic resistance expression. These results indicate that phage therapy applications should consider the evolutionary impacts of phage-host coevolution when determining the efficacy of treatment. Additional research is needed to predict when phage-host coevolution will result in increased antibiotic resistance.

5.6 References

1. Aslam, B. et al. Antibiotic resistance: a rundown of a global crisis. *Infection and Drug Resistance* **11**, 1645-1658 (2018).
2. Gorski, A. et al. Phage Therapy: What Have We Learned? *Viruses* **10** (2018).
3. Withey, S., Cartmell, E., Avery, L.M. & Stephenson, T. Bacteriophages--potential for application in wastewater treatment processes. *Sci Total Environ* **339**, 1-18 (2005).
4. Stachler, E., Kull, A. & Julian, T.R. Bacteriophage Treatment before Chemical Disinfection Can Enhance Removal of Plastic-Surface-Associated Pseudomonas aeruginosa. *Appl Environ Microbiol* **87**, e0098021 (2021).
5. Torres-Barcelo, C. Phage Therapy Faces Evolutionary Challenges. *Viruses* **10** (2018).
6. Paterson, S. et al. Antagonistic coevolution accelerates molecular evolution. *Nature* **464**, 275-278 (2010).
7. Gurney, J., Brown, S.P., Kaltz, O. & Hochberg, M.E. Steering Phages to Combat Bacterial Pathogens. *Trends Microbiol* **28**, 85-94 (2020).
8. Gómez, P. & Buckling, A. Bacteria-Phage Antagonistic Coevolution in Soil. *Science* **332**, 106-109 (2011).
9. Koskella, B. & Brockhurst, M.A. Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol Rev* **38**, 916-931 (2014).
10. Zaman, L. et al. Coevolution Drives the Emergence of Complex Traits and Promotes Evolvability. *PLoS Biology* **12** (2014).
11. Ofria, C. & Wilke, C. Avida: A Software Platform for Research in Computational Evolutionary Biology. *Artificial Life* **10**, 191-229 (2004).
12. Chan, B.K., Abedon, S.T. & Loc-Carrillo, C. Phage cocktails and the future of phage therapy. *Future Microbiology* **8**, 769-783 (2013).
13. Loc-Carrillo, C. & Abedon, S.T. Pros and cons of phage therapy. *Bacteriophage* **1**, 111-114 (2011).
14. Gordillo Altamirano, F. & Barr, J.J. Phage Therapy in the Postantibiotic Era. *Clinical Microbiology Reviews* **32**, e00066-00018 (2019).
15. Abdelsattar, A. et al. How to Train Your Phage: The Recent Efforts in Phage Training. *Biologics* **1**, 70-88 (2021).
16. Segall, A.M., Roach, D.R. & Strathdee, S.A. Stronger together? Perspectives on phage-antibiotic synergy in clinical applications of phage therapy. *Curr Opin Microbiol* **51**, 46-50 (2019).

17. Burmeister, A.R. et al. Pleiotropy complicates a trade-off between phage resistance and antibiotic resistance. *Proc Natl Acad Sci U S A* **117**, 11207-11216 (2020).

18. Oechslin, F. et al. Synergistic Interaction Between Phage Therapy and Antibiotics Clears Pseudomonas Aeruginosa Infection in Endocarditis and Reduces Virulence. *J Infect Dis* **215**, 703-712 (2017).

19. Chan, B.K., Brown, K., Kortright, K., Mao, S. & Turner, T. Extending the lifetime of antibiotics: how can phage therapy help? *Future Microbiology* **11**, 1105-1107 (2016).

20. Williams, H. Phage-induced diversification improves host evolvability. *BMC Evolutionary Biology* **13** (2013).

21. Weitz, J.S., Hartman, H. & Levin, S.A. Coevolutionary arms races between bacteria and bacteriophage. *Proc Natl Acad Sci U S A* **102**, 9535-9540 (2005).

22. Tkhilaishvili, T., Winkler, T., Muller, M., Perka, C. & Trampuz, A. Bacteriophages as Adjuvant to Antibiotics for the Treatment of Periprosthetic Joint Infection Caused by Multidrug-Resistant Pseudomonas aeruginosa. *Antimicrob Agents Chemother* **64** (2019).

23. Allen, R.C., Pfrunder-Cardozo, K.R., Meinel, D., Egli, A. & Hall, A.R. Associations among Antibiotic and Phage Resistance Phenotypes in Natural and Clinical Escherichia coli Isolates. *mBio* **8** (2017).

24. Kumaran, D. et al. Does Treatment Order Matter? Investigating the Ability of Bacteriophage to Augment Antibiotic Activity against Staphylococcus aureus Biofilms. *Front Microbiol* **9**, 127 (2018).

25. Blazanin, M. & Turner, P.E. Community context matters for bacteria-phage ecology and evolution. *ISME J* **15**, 3119-3128 (2021).

26. Knowles, B. et al. Lytic to temperate switching of viral communities. *Nature* **531**, 466-470 (2016).

27. Silveira, C.B. & Rohwer, F.L. Piggyback-the-Winner in host-associated microbial communities. *NPJ Biofilms Microbiomes* **2**, 16010 (2016).

Chapter 6 Significance and Future Research Directions

Viruses are highly abundant and diverse with the ability to impact the structure, function, and evolutionary trajectory of microbial communities. Recent advances in metagenomics have increased our knowledge of viral communities in the environment. Yet, viruses are studied less often than prokaryotes partly due to methodological challenges. In this dissertation, we advanced viral isolation and metagenomic methods to resolve challenges of studying viral communities in environmental samples and applied the methods to explore viral community dynamics through wastewater treatment. Furthermore, we applied *in silico* approaches to study the compounding effects of phage-host coevolution and antibiotics on the emergence and expression of antibiotic resistance in well-mixed chemostats, such as biological reactors or the gut.

There are two methods to study viral communities with metagenomics: sorting viruses from whole metagenomes or enriching viruses from environmental samples before sequencing (i.e., viromes). Viromes have several advantages over sorting viruses from whole metagenomes because viruses are difficult to assemble and identify in whole metagenomes and capture more rare viral populations by focusing sequencing effort on viruses. However, virome quality is dependent on the methods used to concentrate and purify viruses from environmental matrices. Despite widespread use of these methods, there had not been a rigorous study of concentration and purification methods to enrich viruses in different water matrices prior to sequencing. We compared two common approaches for concentrating viruses, namely iron chloride flocculation and ultrafiltration on different water samples. We also assessed the efficacy of virus purification steps, including filtration, chloroform, and DNase reaction. Ultrafiltration performed better at recovering viruses and removing cellular DNA, particularly for freshwater matrices. The combination of purification methods removed most of the 16S rRNA gene copies from samples. The results from this study will inform future research performing viromics in a range of water matrices to reduce biases created during viral enrichment. In this dissertation, the ultrafiltration and purification method was used to generate wastewater DNA virus samples for metagenomic sequencing.

Metagenomics is inherently relative, complicating sample-to-sample comparisons. Therefore, metagenomics can resolve the composition of microbiomes, but not the concentrations of populations within microbiomes. Quantitative PCR (e.g., qPCR or ddPCR) techniques are commonly used methods, but the resulting measurements are constrained to specific targets. Quantitative metagenomics is a promising approach that resolves the composition and quantities of microbial communities and facilitates direct quantitative comparisons between samples. Therefore, we implemented a quantitative metagenomic method by adding synthetic dsDNA and ssDNA standards in known concentrations to correlate relative to absolute abundance. To our knowledge, this was the first application of ssDNA standards and quantitative metagenomics to viromes.

Furthermore, we addressed the poorly defined detection and quantification limitations of quantitative metagenomics. To do this, we performed an in-depth analysis of the limitations of quantitative metagenomics to improve confidence and accuracy of quantification. We assessed target-specific detection thresholds using the entropy of reads mapping to targets. Quantification was improved by detecting and correcting non-specific mapping and assembly errors by establishing limitations of read depth variability. Read depth variability limitations were developed with respect to local GC content using the observed read depth variability along standard sequences. The method was tested with three influent and secondary effluent viromes where one of each matrix type was sequenced with three technical replicates. The method improved quantification accuracy by reducing the difference between ddPCR and quantitative metagenomic measurements. QuantMeta, a bioinformatic tool, was created to make the quantitative metagenomic pipeline publicly accessible. The methods developed here are applicable to whole metagenomes in addition to viromes and to environments other than wastewater to resolve questions related to how functional potential and microbial composition shifts through reactors or changing environmental conditions. Future research should explore amending this method for RNA viruses and evaluate if the regressions and cut-offs hold with other sequencing technologies.

We evaluated the dynamics of viruses through wastewater treatment by comparing influent and secondary effluent viral communities using the quantitative metagenomic method developed in Chapter 3. As a result of our rigorous enrichment method, the viromes were highly purified with 75.5-78% of contigs classified as originating from viruses. Viral populations were 92.5% dsDNA viruses, likely a result of biases in databases used to train viral sorting tools and chloroform purification removing filamentous ssDNA *Inoviridae*. Influent viral communities were significantly more diverse with higher richness than secondary effluent (*p*-values = 1.44×10^{-4} and 0.00698, respectively). The metabolic functional potential of influent and secondary effluent were not significantly different with high concentrations of nucleotide metabolism genes.

In our wastewater viromes, we evaluated how viruses impact antibiotic resistance dissemination and emergence. Antibiotic resistance is a pressing public health threat. Viruses impact antibiotic resistant bacteria by disseminating ARGs via transduction and accelerating the evolution of antibiotic resistance. In our viromes, only 59 viral contigs were carrying ARGs; this indicates that viral-encoded ARGs are rare occurrences in wastewater. In the future, the developed methods can be implemented in studies to explore specific routes of ARG transduction that may occur during biological treatment and quantify ARG transduction frequency.

Transduction is a horizontal gene transfer mechanism that may occur when ARGs are incorporated into viral genomes. We found ARGs incorporated on viral genomes in our viromes, however, it rarely occurred with only 59 viral populations encoding ARGs. Future work should explore if these observations are true at other wastewater treatment plants, in other types of wastewater, and for other biological treatment technologies. Furthermore, we did not examine if secondary effluent viruses carrying ARGs were present in final effluent following disinfection. Future studies should explore the temporal and spatial potential of viruses carrying ARGs in secondary effluent. For example, benchtop experiments should explore if DNA damaging antibiotics may increase transduction events by inducing the SOS response system causing prophages to enter the lytic cycle and transduce genes.

To further examine potential impacts of viruses on antibiotic resistance, we performed *in silico* evolution experiments using Avida. We developed an Avida environment with a detrimental resource that mimics an antibiotic and created the ability for digital organisms to evolve resistance. The experiments demonstrated that there are compounding effects of phage-host coevolution and antibiotics on the outcomes of antibiotic resistance evolution. In the presence of phages, antibiotic resistance evolved more rapidly. Occasionally, when phages and antibiotics were present, pleiotropic results of organisms evolving phage infection defenses resulted in a decreased susceptibility to antibiotics. Phage therapy is gaining traction as a potential supplement or replacement for antibiotics. Phage therapy in clinical applications uses lytic phages to target pathogens and relieve infections. These results demonstrate that potential outcomes of phage-host coevolution should be considered when determining the efficacy of phage therapy. Future work with Avida should decouple the antibiotic resistance tasks and phage infection tasks to evaluate the generalizability of phage-host coevolution of antibiotic resistance evolution. Other applications for phage therapy are being explored to improve wastewater treatment or disinfect surfaces. Additional research is required to assess the efficacy of adding lytic phage to activated sludge to resolve foaming, bulking, or dewaterability issues. With the methods developed in this dissertation, direct (e.g., removal of problematic bacteria populations) and indirect (e.g., evolution of phage infection defenses and pleiotropic results, altering microbial community structure) outcomes of adding lytic phages to microbial communities.

This dissertation advanced methods to study microbial communities by developing a quantitative metagenomic method that can be applied to viromes and whole metagenomes from a variety of environments. Future applications of this work to viral communities could elucidate the impact of viruses on the structure and function of microbial communities such as during wastewater treatment, in phage therapy applications, or in response to changing environmental conditions.

Appendix A Supplementary Information for Comparing Ultrafiltration and Iron Chloride Flocculation to Prepare Aquatic Viromes from Various Matrices

A.1 Sample Characteristics

*Table A.1: pH, total suspended solids (TSS), and volatile suspended solids (VSS) content of all the samples gathered for experiments. *The method for TSS and VSS measurements does not work on high salinity samples, seawater TSS and VSS were unable to be determined.*

Matrix	Collection Date	Experiment	pH	Solids Content (mg/L)	
				TSS	VSS
Influent	1/3/19	Jar Test	NA	374	261
	1/8/19	Jar Test	NA	267	180
	1/21/19	Jar Test	NA	259	185
	2/15/19	Viral Purification Optimization	NA	179	109
	5/6/19	Ultrafiltration and Purification	7.57	287	189
	5/9/19	Ultrafiltration and Purification	7.57	164	85
	5/23/19	Ultrafiltration and Purification	7.42	182	92
	6/25/19	Iron Chloride Flocculation and Purification	7.43	188	122
	6/27/19	Iron Chloride Flocculation and Purification	7.53	189	108
	7/1/19	Iron Chloride Flocculation and Purification	7.42	164	98
Secondary Effluent	10/30/18	Jar Test	NA	NA	NA
	11/6/18	Jar Test	NA	NA	NA
	12/4/18	Jar Test	NA	30	16
	12/11/18	DE UF Optimization	NA	NA	NA
	3/14/19	Viral Purification Optimization	NA	61	23
	5/16/19	Ultrafiltration and Purification	6.95	60	19
	5/20/19	Ultrafiltration and Purification	6.89	72	14
	5/28/19	Ultrafiltration and Purification	6.98	63	17

	6/25/19	Iron Chloride Flocculation and Purification	7.08	55	17
	6/27/19	Iron Chloride Flocculation and Purification	7.15	70	10
	7/1/19	Iron Chloride Flocculation and Purification	7.13	58	16
	10/31/19	Ultrafiltration and Purification (3 phage)	7.14	45	15
	11/7/19	Ultrafiltration and Purification (3 phage)	7.07	35	15
	11/14/19	Ultrafiltration and Purification (3 phage)	7.07	37	12
River Water	5/20/19	Ultrafiltration and Purification	7.92	75	35
	5/23/19	Ultrafiltration and Purification	8.03	57	20
	5/28/19	Ultrafiltration and Purification	7.94	56	14
	5/30/19	Jar Test	7.9	60	20
	6/11/19	Jar Test	7.93	73	33
	6/12/19	Jar Test	7.87	64	20
	6/25/19	Iron Chloride Flocculation and Purification	7.96	47	10
	6/27/19	Iron Chloride Flocculation and Purification	8.1	43	10
	7/1/19	Iron Chloride Flocculation and Purification	8.19	52	20
Shedd Aquarium Seawater	2/27/20	Jar Test, Ultrafiltration and Purification, Iron Chloride Flocculation and Purification	7.98	*	*

Table A.2: pH of the Shedd Aquarium seawater stored over a week of experiments and measured immediately prior to the start of each experiment.

Date	Experiment	pH
2/27/20	Immediately After Collection	NA
2/27/20	Jar Test (Immediately after arrival to lab)	7.98
2/28/20	Jar Test	8 (filtered)
2/29/20	Jar Test	8.05 (filtered)
3/2/20	Ultrafiltration and Purification	8
3/3/20	Ultrafiltration and Purification	7.91
3/4/20	Ultrafiltration and Purification	7.88
3/5/20	Iron Chloride Flocculation and Purification Iron Chloride Flocculation and Purification	7.95

A.2 Phage Propagation

Table A.3: Respective media recipes for host media, hard nutrient agar plates, soft nutrient agar, and buffer used to culture each phage.

Phage	T3, T4, PhiX174	HS2, HM1	ICBM5
Host Media	8.0 g Nutrient Broth (Fisher Scientific, catalog no. BD23400), 5.0 g NaCl, 1 L H_2O	2.5 g Peptone, 0.5 g Yeast Extract, 100 mL Widdel 10x Salt Solution*, 900 mL H_2O, pH 7.6	37.4 g Marine Broth 2216 (Fisher Scientific, catalog no. DF0791174), 1 mL Balch Vitamin Solutionˆ, 1 L H_2O
Hard Nutrient Agar	8.0 g Nutrient Broth, 5.0g NaCl, 15.0g Agar, 1 L H_2O	1.0 g Peptone, 0.2 g Yeast Extract, 12 g Bacto Agar, 100 mL Widdel 10x Salt Solution*, 900 mL H_2O, pH 7.6	37.4 g Marine Broth 2216, 18 g Bacto agar, 1 mL Balch Vitamin Solutionˆ, 1 L H_2O
Soft Nutrient Agar	8.0 g Nutrient Broth, 5.0 g NaCl, 7.0 g Agar, 1 L H_2O	5 g Peptone, 1 g Yeast Extract, 6 g Bacto Agar, 100 mL Widdel 10x Salt Solution*, 900 mL H_2O, pH 7.6	29.9 g Marine Broth 2216, 4.8 g Agar, 1 mL Balch Vitamin Solutionˆ, 1 L H_2O
Buffer	0.6 g NaH_2PO_4, 0.58 g NaCl, 0.1 g NaOH, 1 L H_2O	5.85 g NaCl, 20.0 g $MgSO_4{\cdot}7H_2O$, 7.88 g Tris-HCl, 1 L H_2O, pH 7.6	

*Widdel 10x Salt Solution: 50 g NaCl, 7.5 g $MgCl_2{\cdot}6H_2O$, 0.28 g $CaCl_2{\cdot}2H_2O$, 0.63 g NH_4Cl, 0.5 g KH_2PO_4, 1.25 g KCl, 250 mL H_2O

ˆBalch Vitamin Solution: 25 mg para-Aminobenzoic acid, 10 mg Folic acid, 10 mg Biotin, 25 mg Nicotinic acid, 25 mg Ca pantothenate, 25 mg Riboflavin, 25 mg Thiamine hydrochloride, 50 mg Pyridoxine hydrochloride, 5 mg Cyanocobalamine, 25 mg Lipoic acid

Table A.4: Phage culturing centrifuge conditions for each phage following chloroform addition.

Phage	Centrifuge Force (*xg*)	Time (minutes)
T3	5000	15
T4	5000	15
PhiX174	1000	25
HS2	3000	10
HM1	3000	10
ICBM5	3000	10

A.3 Iron Chloride Flocculation Jar Tests

*Table A.5: Tested range of iron chloride concentrations for jar tests and specifications of resuspension buffer added to dissolve the iron flocs. Resuspension buffer was diluted into sterile water (** sterile water was added instead of resuspension buffer).*

Iron chloride concentration (mg Fe L^{-1})	Resuspension buffer dilution	Volume of resuspension buffer added (mL)
0	NA**	5
0.1	1:100	5
1	1:10	5
5	1:2	5
10	1:1	5
25	1:1	12.5

Table A.6: Recovery of benchtop controls at the time filtrate and concentrate samples were collected during flocculation jar tests. Benchtop controls are 0.5-L samples of matrix set aside and spiked with T3 or HS2 at the same time as the samples that underwent flocculation. The benchtop control remained at the same temperature as the flocculation samples throughout the experiment to assess degradation of spike viruses in each matrix over the duration of the flocculation and resuspension. The geometric mean and 95% confidence intervals are reported for all benchtop controls.

Matrix	Recovery (%)	
	Filtrate	Concentrate
Influent	110.3 (59.31, 205.0)	106.0 (54.87, 204.9)
Effluent	109.2 (73.67, 161.8)	89.50 (83.37, 96.08)
River Water	81.66 (47.46, 140.5)	132.8 (102.1, 172.7)
Seawater	89.89 (54.90, 147.2)	102.8 (54.09, 195.4)

A.4 Optimizing Purification Methods

Three methods to purify viral nucleic acids were tested on pre-concentrated influent and secondary effluent samples. Each purification method tested has a different mechanism for purifying the viral nucleic acids. Two chloroform treatments, chloroform and DNase treatments, and filtering followed by DNase treatment were evaluated by the ratio of phage T3 gene copies to 16S rRNA gene copies with a higher ratio indicating a better purification performance. For the influent samples, the highest phage T3 to 16S rRNA ratios are observed in the chloroform and DNase treated samples. The phage T3 to 16S rRNA ratio is not significantly less than the two chloroform treated samples (*p*-value = 0.28) in the secondary effluent as seen in Figure A.1. The chloroform and DNase treatment method was selected for purifying viruses in all of the sample matrices for the iron chloride flocculation and purification and ultrafiltration and purification methods.

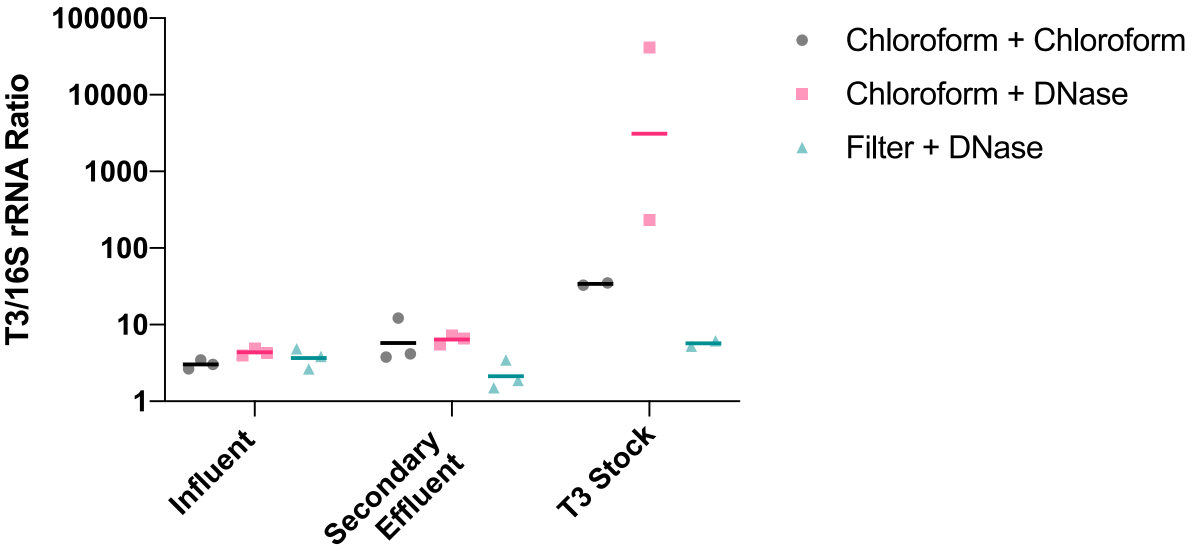

Figure A.1: T3 to 16S rRNA concentration factors after purification for pre-concentrated influent and secondary effluent. The geometric mean surrounded by individual points is plotted. A 2-way ANOVA test was performed to investigate variability in the data. No statistically significant differences between the three treatments in the influent were observed (all p-values > 0.25). No statistical difference between two chloroform treatments and chloroform and DNase treatments in secondary effluent was observed (p-value = 0.30). A statistical difference between two chloroform treatments and filtering and DNase treatments in secondary effluent was observed (p-value = 0.04). Chloroform and DNase treatments and filtering and DNase treatments in secondary effluent had no statistically significant difference (p-value = 0.45).

Methods. Three different methods to purify viruses were tested: two chloroform treatments, a chloroform and DNase treatment, and filtering and DNase treatment. The methods were tested in triplicate on a concentrated influent and concentrated secondary effluent. 10-L of influent and 20-L of secondary effluent were collected the day before the experiment and concentrated approximately 30-fold and 65-fold, respectively, by tangential ultrafiltration with the same method described above then 0.45-µm Express PLUS filtered. Samples were stored at 4˚C overnight. The day of the experiment, the wastewater samples were concentrated an additional 20-fold by dead-end ultrafiltration using 100 kDa MWCO and 1 cm^2 surface area Amicon™ filter units. Initially, 500 µL of sample was added to each filter, then centrifuged at 3,000*xg* and 4˚C until approximately 200 µL remained. The process continued with more sample added to each filter until a total of 4 mL of sample was added and a 200 µL final volume remained on each filter. The concentrate was collected by inverting the filter into a clean collection tube and centrifuging at 1,000*xg* for 1 minute. Approximately 10^6 T3 gene copies μL^{-1} were added to the influent and secondary effluent after concentrating. A sample was collected for recovery analysis after T3 addition and stored at 4˚C until DNA extraction. Each replicate had an initial volume of 400 µL.

Chloroform treatments were performed by adding 100 µL of chloroform and vortexing for approximately 2 minutes then settled for 10 minutes and centrifuged briefly. A majority of the chloroform was removed by pipetting the chloroform off of the bottom of the samples, then evaporating the remainder from the sample by aerating the sample in a fume hood. The chloroform treatment was completed by filtering samples through 0.45-µm PES and 13-mm diameter syringe filters (CellTreat Scientific Supplies, Cat. No. 229748). For the two chloroform treatments samples, the entire chloroform treatment protocol was repeated. Filtering for the filtering and DNase treatment samples was performed with the same syringe filters as the final step of the chloroform treatment. DNase treatment was performed as described for iron chloride flocculation and purification methods. Samples were collected for recovery analysis after purification and stored at 4˚C until DNA extraction. Previously described T3 and 16S rRNA probe ddPCR assays were used in the recovery analysis.

A.5 Optimizing Dead-end Ultrafiltration

MilliQ pre-wash, 1% BSA incubation, and sonication have been previously reported to improve viral recoveries after dead-end ultrafiltration [1]. We tested these three methods alone and in combination with our samples spiked with T3 and MS2 to confirm previous findings. We found no significant difference in the geometric mean of any of the treatment recoveries or the no treatment control (one-way ANOVA: *p*-value = 0.23). We did not perform any additional treatments before or after dead-end ultrafiltration for the remainder of the experiments.

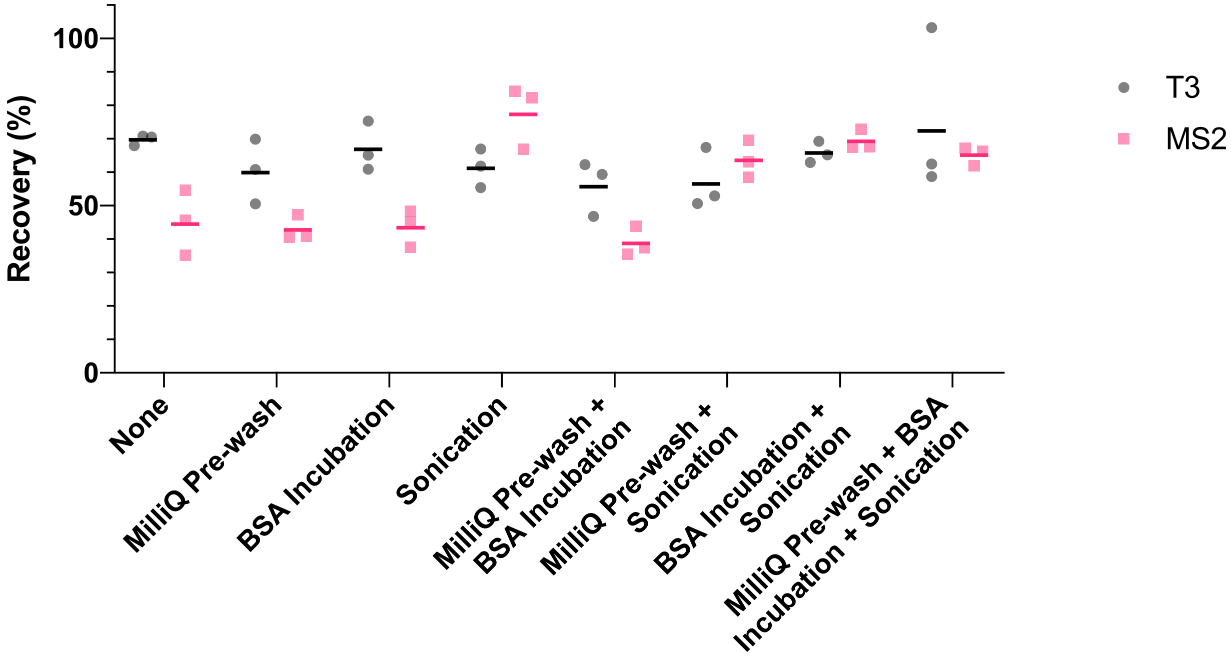

Figure A.2: MilliQ pre-washing, 1% BSA incubation, and sonication with TE buffer and combinations of each were tested to optimize dead-end ultrafiltration of secondary effluent. The secondary effluent was previously tangentially ultrafiltered (60-fold concentration) and 0.45-µm filtered. T3 and MS2 were spiked in to evaluate the recovery after dead-end ultrafiltration (5-fold concentration) for each method. The experiment was performed in triplicate. The individual measurements and geometric mean for each virus and dead-end ultrafiltration condition are shown above. A one-way ANOVA test concluded there is no statistical difference between the means of the different dead-end ultrafiltration methods (p-value = 0.23).

Methods. Three different methods to optimize dead-end ultrafiltration were tested: milliQ pre-wash, bovine serum albumin (BSA) treatment pre-ultrafiltration, and sonication post-ultrafiltration. These treatments were all tested in triplicate separately and in combination against a no treatment control. The experiment was conducted with secondary effluent collected the day of the experiment and concentrated approximately 70-fold by tangential ultrafiltration then 0.45-µm Express PLUS filtered. Tangential ultrafiltration was performed as described for the ultrafiltration and purification method. Approximately 10^5 gene copies μL^{-1} T3 and 10^8 gene copies μL^{-1} *Escherichia coli* phage MS2 (ATCC® 15597-B1™) were added to the concentrated effluent. A sample was collected after the spike addition for recovery analysis and stored at 4˚C until DNA extraction. The milliQ pre-wash was conducted by adding 500 µL of sterilized milliQ to each filter then centrifuging at 3,000*xg* and 4˚C until all the milliQ passed through the filter. The milliQ was 0.02-µm filtered with Anotop™ 25 Plus sterile syringe filters (GC Healthcare Whatman™, Cat. No. 0992626) to sterilize. The BSA treatment was performed with a sterile 1% BSA solution made by diluting 50 mg mL^{-1} UltraPure BSA (Invitrogen™, Cat. No. AM2616) in phosphate buffer (5 mM NaH_2PO_4 and 10 mM NaCl, pH 7.5). 500 µL of the 1% BSA solution was added to each filter and incubated at room temperature for 1 h, then removed from the filter by pipetting. Dead-end ultrafiltration was performed with 100 kDa MWCO and 1 cm^2 surface area Amicon™ Ultra Centrifugal filter units by adding 500 µL of concentrated effluent to each filter and centrifuging at 3,000*xg* and 4˚C until 100 µL of sample remains on the filters. Filters were sonicated by adding 50 µL of sterile 1x TE buffer and sonicating for 3 minutes at 50 W and 42 kHz. All filters were inverted into clean collection tubes and centrifuged at 1000*xg* and 4˚C for 1 minute. After treatments, samples were collected for recovery analysis and stored at 4˚C until DNA extraction. T3 qPCR and MS2 RT-qPCR assays were used in the recovery analysis (Section A.7).

A.6 Phage HS2 Survivability in Wastewater Influent

After two hours of incubation at room temperature and DNase treatment, 2.1% and 2.9% of marine phage HS2 was recovered in two 0.45-µm filtered influent samples. This indicates that differences in seawater and wastewater matrices impact the survivability of HS2. Based on this result, marine phages (HS2, HM1, ICBM5) were spiked into seawater and freshwater phages (T3, T4, PhiX174) were spiked into freshwater matrices for all experiments.

Methods. Two 25 mL of influent was 0.45-µm PES filtered and spiked with 10^6 gc μL^{-1} of HS2. The samples incubated on the benchtop for 2 hours, then were treated with DNase, as described previously, with a 2 hour incubation time. DNase treatment was performed to remove non-encapsidated HS2 genomes and constrain qPCR measurements to HS2 viral particles. DNA extractions were performed before, immediately after HS2 addition, and after incubation and DNase treatment, as described previously and HS2 recovery was determined based on measurements from the HS2 qPCR assay.

A.7 PCR Reaction Conditions and Amplicons

Table A.7: qPCR amplicons developed into qPCR standards as gBlocks® Gene Fragments (IDT, Coralville, IA) for each phage used in experiments.

Target Phage	dsDNA Fragments of qPCR Amplicons
T3	CCA ACG AGG GTA AAG TGA TAG GCT TTA GTG TGC TTC TTG AGA CTG GTC GTT TAG TAG ACG CCA ACA ACA TCT CTC GCG CAT TGA

	TGG ACG AGT TCA CAT CCA ACG TTA AAG CCC ACG GTG AAG ACT TCT ACA ATG GTT GGG CCT GTC AGG TCA ACT ACA TGG AAG CGA CCC CGG ACG GCT CCC TGC GAC ACC CTA GCT TCG AGA AGT TCC GAG GAA CTG AGG ACA ACC CTC AAG AGA AAA TGT AAC CAA CTC ACT GGC TCA CCT TCA CCT TCA CGG GTG GGC CTT TCT TCG TTC CGG GCA TTA ACC CTC ACT AAC AGG AGA CAC ACA CCA TGT GGC TTA TCC TAT TCG CTA TCG TCG
HS2	GGT TGA TGA AAA GTC ACT AGG CTG TAA ATC GCA TTC TGT AAA TAA ATC GGC ATT GTT AAG CAA TAC GCC AAT GAC TAA AGA TTC CTG CTC TAA AAT ATC CTT GTT CAT AGT TTA AAT TCC TTC ACT GCT GGT CTT GAT TGT TGT GGT TTG CCG TTA AAG CCA TTT GAT GCA GCT TTC ATT TTA GCT GAC AAG TCT GGG TAT TTA TCC CTA AGC TTT GCA AGG CTG AGA ATA TTA ACA CTC CAA AAG CTA TCA GCA TTA GCC CAT GAG AAA ACT TTC CAA CAC TCA TTT AGA TCT GCC CCG
MS2	CCG CTA CCT TGC CCT AAA CGA AGA TCG AAA GTT TCG ATC AAA ACA CGT GGC CGG CAG GTG GTT GGA GTT GCA GTT CGG TTG GTT ACC ACT AAT GAG TGA TAT CCA GGG TGC ATA TGA GAT GCT TAC GAA GGT TCA CCT TCA AGA GTT TCT TCC TAT GAG AGC CGT ACG TCA GGT CGG TAC TAA CAT CAA GTT AGA TGG CCG TCT GTC GTA TCC AGC TGC AAA CTT CCA GAC AAC GTG CAA CAT ATC GCG ACG TAT CGT GAT ATG GTT TTA CAT AAA CGA TGC ACG TTT GGC ATG GTT GTC GTC

Table A.8: qPCR thermocycler conditions for T3 and HS2 with Biotium Fast-Plus EvaGreen mastermix.

Step	**T3**			**HS2**		
	Duration	**Temperature (˚C)**	**Cycles**	**Duration**	**Temperature (˚C)**	**Cycles**
Initial Denaturation	2 minutes	95		2 minutes	95	
Denaturing	5 seconds	95	35	5 seconds	95	35
Annealing	5 seconds	60		5 seconds	47	
Extension	25 seconds	72		25 seconds	72	
Final Denaturation	5 minutes	95	1	5 minutes	95	1
Melting Curve Initial	15 seconds	55		15 seconds	55	
Melting Curve End	15 seconds	95		15 seconds	95	
Melt Curve Duration	20 minutes			20 minutes		

Final Hold	infinite	4		infinite	4	

Phage MS2 RT-qPCR assay. Primer (5’ to 3’) specific to phage MS2 were selected (303 bp; forward, CCG CTA CCT TGC CCT AAA C; reverse, GAC GAC AAC CAT GCC AAA C)[2]. The 20 µL reaction contained 10 µL 2x GoTaq^{TM} one-step RT-qPCR master mix (Promega, Cat. No. PRA6020), 0.4 µL 50x reverse transcriptase mix, 0.3 µM primers, and 2 µL of DNA template. Standard curves were prepared in triplicate between 100 and 10^6 gene copies µL^{-1} with gBlocks dsDNA fragments of the amplicon sequence (IDT, Coralville, IA) shown in Table A.7. qPCR was performed on a realplex^2 Mastercycler epgradient S automated real-time PCR system (Eppendorf®, New York City, New York) with reaction conditions are provided in Table A.8. All efficiencies were greater than 80% and R^2 values were greater than 0.95. Each sample was measured in duplicate and the geometric mean was taken.

Table A.9: One step RT-qPCR reaction conditions for MS2 with GoTaqTM one-step RT-qPCR mastermix.

Step	MS2		
	Duration	Temperature (˚C)	Cycles
Reverse Transcription	15 minutes	40	1
Initial Denaturation	10 minutes	95	
Denaturing	15 seconds	95	45
Annealing	30 seconds	60	
Extension	45 seconds	72	
Melting Curve Initial	45 seconds	68	1
Melting Curve End	5 seconds	95	
Melt Curve Duration	5 minutes		
Final Hold	infinite	4	

Table A.10: ddPCR assays for all phage targets (Lim et al.) and 16S rRNA (Nadkarni et al.) with maximum two targets per reaction (IDT, Coralville, IA).

Targets	Primers/Probes	Amplicon Length (bp)	Annealing Temperature (˚C)
16S rRNA [2]	Forward: 5’- TC CTA CGG GAG GCA GCA GT-3’	466 bp	56

	Reverse: 5’- GG ACT ACC AGG GTA TCT AAT CCT GTT-3’ Probe: 5’-/FAM/-CG TAT TAC CGC GGC TGC TGG CAC-/BHQ_1/-3’		
T3	Forward: 5’- CCA ACG AGG GTA AAG TGA TAG-3’ Reverse: 5’- CGA CGA TAG CGA ATA GGA TAA G-3’ Probe: 5’-/HEX/-CC AAC AAC ATC TCT CGC GCA TT-/BHQ_2/-3’	351 bp	56
T4[3]	Forward: 5’-CCA CAA CTA ACC GAG GAA GTA A-3’ Reverse: 5’-TGC GAT ATG CTA TGG GTC TTG-3’ Probe: 5’-/FAM/-TGC TCC ATC AGA GGA AGA ATG CGA-/BHQ_1/-3’	107 bp	56
PhiX174	Forward: 5’-GGG ATA CCC TCG CTT TCC TG-3’ Reverse: 5’-CAA AGA CGA GCG CCT TTA CG-3’ Probe: 5’-/HEX/-TAC GTG CGG AAG GAG TGA TGT AAT G-/BHQ_2/-3’	353 bp	56
HS2	Forward: 5’-GGT TGA TGA AAA GTC ACT-3’ Reverse: 5’-CGG GGC AGA TCT AAA TGA-3’ Probe: 5’-/HEX/-TTT AGT CAT TGG CGT ATT GCT TAA C-/BHQ_2/-3’	300 bp	57
HM1	Forward: 5’-CGT CTG CAG TAG ATT GGG CA-3’ Reverse: 5’-AGA TGG GGT GTT GGA GGA AAG-3’ Probe: 5’-/FAM/-CAA GAA CAG GAC TTG CCA GAA GTG T-/BHQ_1/-3’	216 bp	56
ICMB5	Forward: 5’-ATC CGA TCC GCC GAA GTA AC-3’ Reverse: 5’-AAA CGC TCC GTT CTT CTC GT-3’ Probe: 5’-/HEX/-AAG GTG TAA CCG CTG GTC GGC ATA A-/BHQ_2/-3’	275 bp	56

*Table A.11: ddPCR reaction conditions for targets listed in Table A.10 with ddPCRTM Supermix for Probes (no dUTP). The initial denaturation step was only used on dsDNA targets (ssDNA targets were PhiX174 and ICBM5). With ssDNA phage targets, the initial denaturing began to degrade the ssDNA targets and produced two positive fluorescence levels. *Annealing temperatures for each target are provided in Table A.10.*

Step	Temperature (˚C)	Duration (minutes)	Cycles

Initial Denaturation (dsDNA targets only)	95	10	1
Denaturation	95	0.5	40
Annealing	*	1	
Extension	72	2	
Final Annealing	4	5	1
Final Denaturation	95	5	1
Hold	4	Infinite	1

A.8 Iron Chloride Flocculation Compared to Ultrafiltration

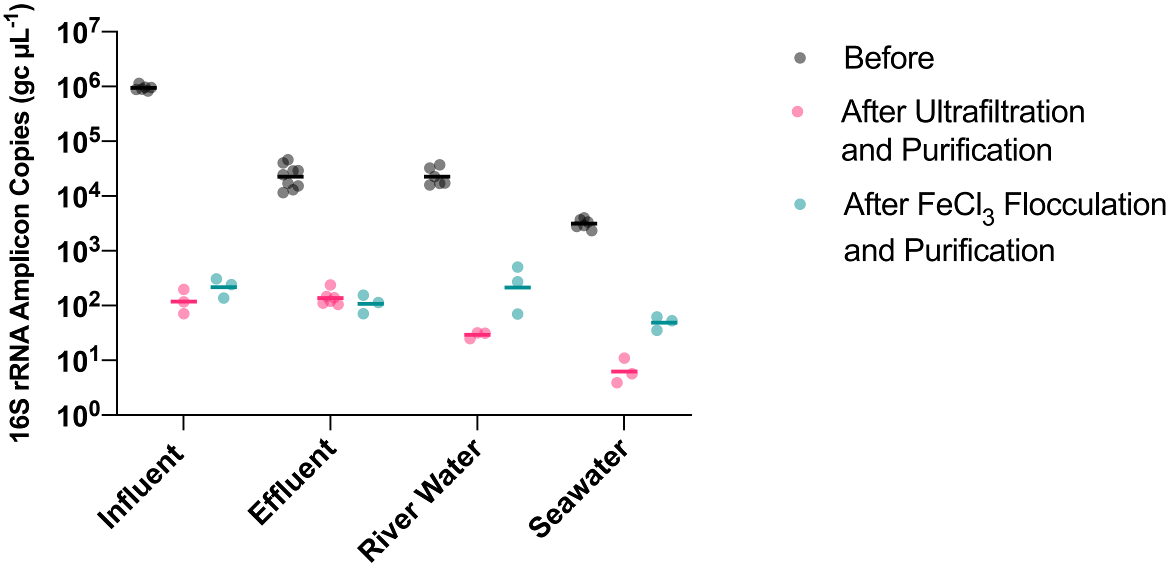

Figure A.3: Absolute 16S rRNA amplicon concentrations before and after ultrafiltration or iron chloride flocculation for each matrix. The geometric mean and individual points are plotted for each matrix and treatment.

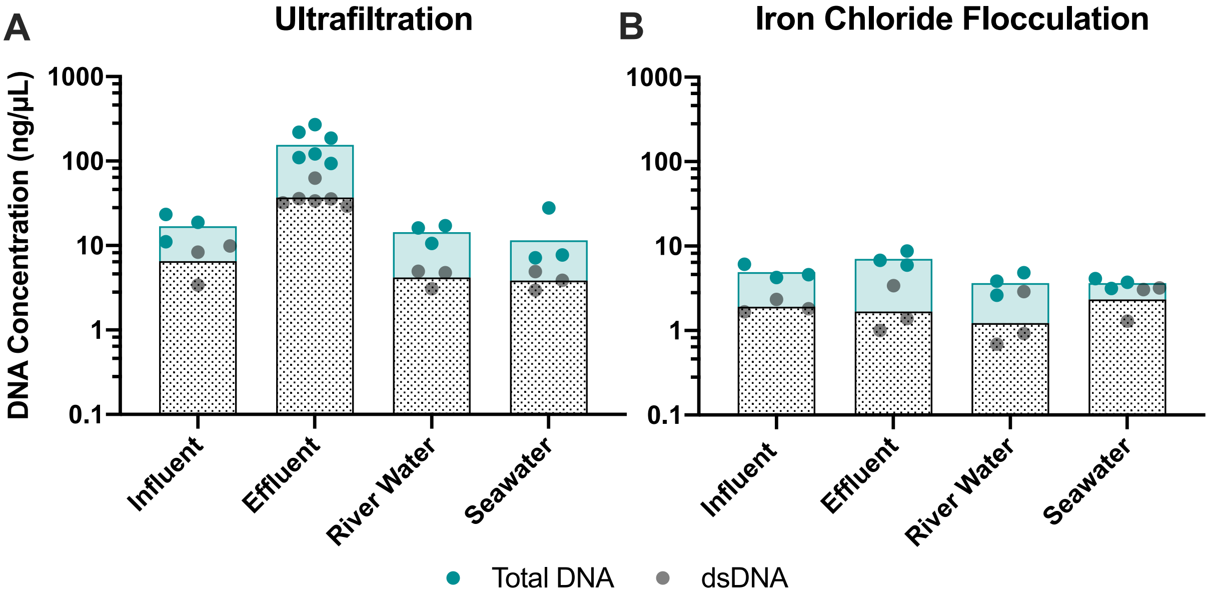

Figure A.4: dsDNA and ssDNA concentrations were measured in each sample after concentrating and purifying. DNA concentrations after ultrafiltration and purification (A) and iron chloride flocculation and purification (B). Geometric means of ssDNA concentrations with individual experimental replicates are stacked on top of geometric means of dsDNA concentrations with individual experimental replicates.

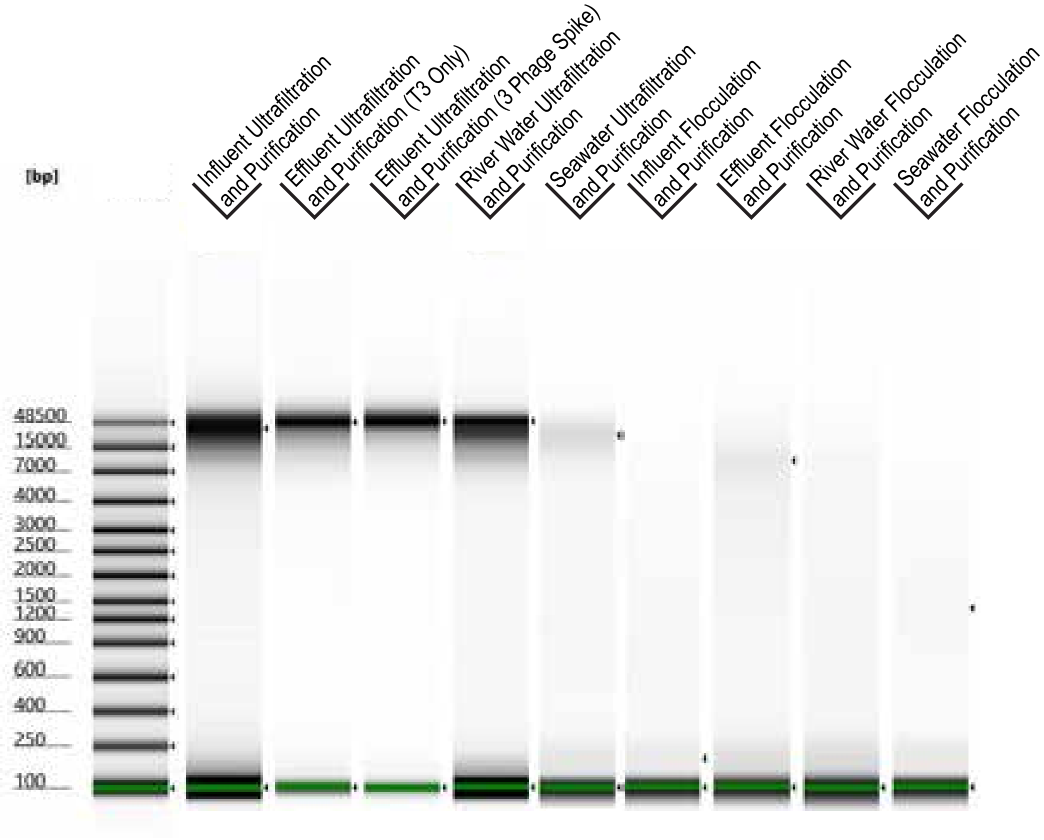

Figure A.5: DNA fragmentation was compared between ultrafiltration and iron chloride flocculation for each matrix with Agilent TapeStation. DNA fragmentation for each matrix and method (triplicates pooled, 9 total samples) was assessed by Agilent TapeStation for DNA lengths up to 60,000 bp by the Advanced Genomics Core at the University of Michigan. The clearly defined bands in the ultrafiltration samples are relics of the T3 spikes in freshwater matrices and HS2 spike in seawater that notably do not appear in the iron chloride flocculation samples. This observation indicates potential genome shearing during the iron chloride flocculation and purification process that may not occur in the ultrafiltration and purification process.

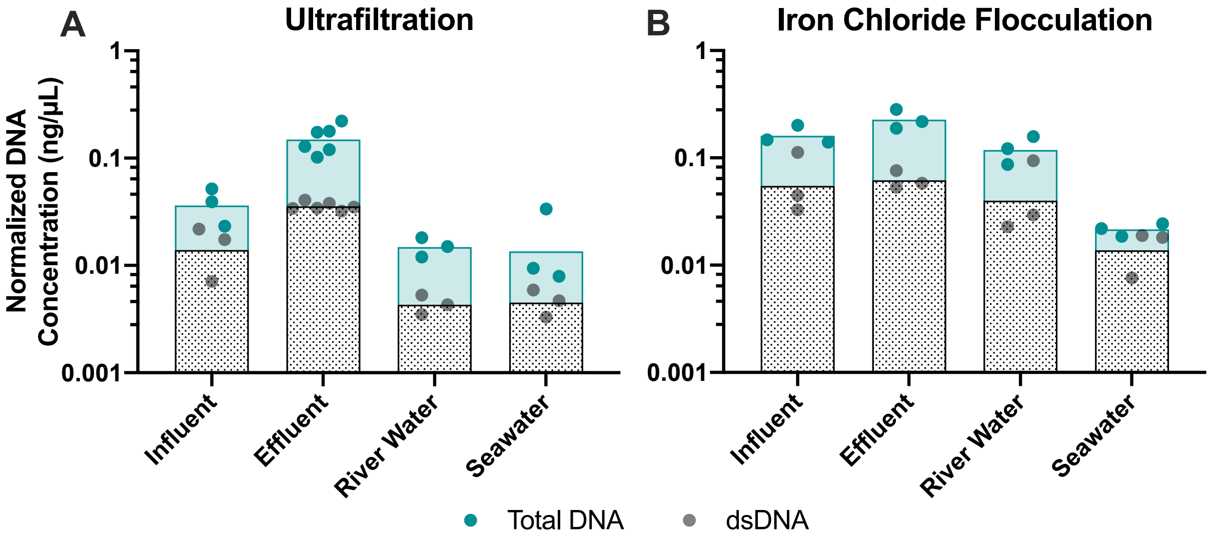

Figure A.6: dsDNA and ssDNA concentrations were normalized by the volumetric concentration factor in each sample. Normalized DNA concentrations are reported after ultrafiltration and purification (A) and iron chloride flocculation and purification (B). Geometric means of normalized ssDNA concentrations with individual experimental replicates are stacked on top of geometric means of normalized dsDNA concentrations with individual experimental replicates.

*Table A.12: Iron chloride flocculation and purification geometric means and 95% confidence intervals for the spike-and-recovery experiments presented in Figure 2.3. *Phage identity does not factor into the [16S rRNA]Step/[16S rRNA]Initial calculation, therefore all ratios are the same for a specific matrix.*

			Geometric Mean (95% CI)			
Matrix	**Phage**	**Step**	**[Phage]$_{Step}$/ [Phage]$_{Initial}$**	**Recovery (%)**	**[16S rRNA]$_{Step}$/ [16S rRNA]$_{Initial}$**	**Phage Enrichment**
Influent	T3	After 0.45-µm filter and iron chloride flocculation	26.89 (16.13-44.85)	69.36 (41.08-117.1)	0.032 (0.019-0.053)	842.9 (411.6-1726)
		After 0.45-µm filter and chloroform	20.12 (11.93-33.94)	51.88 (31.41-85.69)	0.033 (0.029-0.037)	615.7 (354.3-1070)
		After DNase	7.61 (4.75-12.20)	24.94 (16.21-38.35)	0.0064 (0.0018-0.0227)	1043 (678.5-1603)
Effluent	T3	After 0.45-µm filter and iron chloride flocculation	30.61 (20.88-44.89)	77.60 (52.28-115.2)	0.743 (0.429-1.29)	41.18 (29.04-58.40)
		After 0.45-µm filter and chloroform	24.23 (9.10-64.49)	61.44 (23.03-163.9)	0.491 (0.133-1.82)	49.35 (35.32-68.96)

		After DNase	6.66 (1.49-29.81)	21.42 (4.66-98.46)	0.152 (0.030-0.757)	43.89 (25.83-74.59)
River Water	T3	After 0.45-µm filter and iron chloride flocculation	31.35 (11.02-89.21)	80.36 (28.96-223)	2.64 (0.313-22.31)	11.87 (1.51-93.35)
		After 0.45-µm filter and chloroform	20.14 (4.73-85.84)	51.56 (12.70-209.3)	1.55 (0.113-21.07)	13.04 (2.89-58.71)
		After DNase	7.60 (1.94-29.79)	24.69 (6.62-92.07)	0.393 (0.031-4.98)	19.37 (5.29-70.88)
Seawater	HS2	After 0.45-µm filter and iron chloride flocculation	135.0 (128.4-142.0)	62.91 (59.97-65.99)	112.9 (64.72-197.0)	0.836 (0.472-1.48)
		After 0.45-µm filter and chloroform	126.7 (100.8-159.1)	59.02 (47.15-73.87)	115.1 (62.00-213.8)	0.910 (0.470-1.76)
		After DNase	24.67 (21.96-27.71)	14.58 (12.96-16.39)	2.82 (0.854-9.31)	8.75 (2.52-29.23)
	HM1	After 0.45-µm filter and iron chloride flocculation	138.2 (95.10-200.8)	64.4 (44.47-93.28)	NA*	1.22 (0.534-2.81)
		After 0.45-µm filter and chloroform	180.1 (175.8-184.5)	83.93 (81.67-86.26)	NA*	1.57 (0.860-2.85)
		After DNase	47.83 (39.52-57.89)	28.27 (23.35-34.21)	NA*	16.97 (4.47-64.40)
	ICBM5	After 0.45-µm filter and iron chloride flocculation	122.4 (105.7-141.6)	57.01 (49.12-66.18)	NA*	1.08 (0.633-1.85)
		After 0.45-µm filter and chloroform	87.45 (47.94-159.5)	40.75 (22.31-74.45)	NA*	0.760 (0.336-1.72)
		After DNase	3.69 (2.30-5.94)	2.18 (1.35-3.52)	NA*	1.31 (0.540-3.18)

*Table A.13: Ultrafiltration and purification geometric means and 95% confidence intervals for the spike-and-recovery experiments presented in Figure 2.4. *Phage identity does not factor into the [16S rRNA]Step/[16S rRNA]Initial calculation, therefore all concentration factors are the same for a specific matrix.*

Matrix	Phage	Step	Geometric Mean (95% CI)

			$[\text{Phage}]_{\text{Step}}/[\text{Phage}]_{\text{Initial}}$	**Recovery (%)**	**$[\text{16S rRNA}]_{\text{Step}}/[\text{16S rRNA}]_{\text{Initial}}$**	**Phage Enrichment**
Influent	T3	After 0.45-µm filter and tangential ultrafiltration	25.19 (19.44-32.63)	82.16 (67.41-100.1)	0.413 (0.042-4.04)	60.96 (7.34-506.7)
		After 0.45-µm filter and chloroform	23.28 (12.90-42.02)	75.90 (39.05-147.6)	0.013 (0.0061-0.026)	1850 (564.8-6059)
		After dead-end ultrafiltration	301.1 (116.3-779.8)	50.49 (18.1-140.8)	0.116 (0.049-0.277)	2591 (883.4-7598)
		After DNase	220.8 (116.5-418.5)	47.02 (23.80-92.89)	0.062 (0.016-0.234)	3567 (1476-8619)
Effluent	T3	After 0.45-µm filter and tangential ultrafiltration	46.89 (36.81-59.73)	73.55 (65.23-82.94)	1.78 (0.719-4.43)	26.27 (10.83-63.69)
		After 0.45-µm filter and chloroform	47.83 (34.93-65.49)	75.01 (56.79-99.06)	0.779 (0.362-1.68)	61.38 (34.85-108.1)
		After dead-end ultrafiltration	508.9 (428.0-605.2)	38.57 (27.76-53.58)	6.47 (2.59-16.17)	78.61 (34.61-178.6)
		After DNase	441.1 (358.8-542.2)	42.34 (36.96-48.51)	6.31 (2.99-13.32)	69.88 (32.18-151.7)
	T4	After 0.45-µm filter and tangential ultrafiltration	24.38 (14.86-40.01)	35.71 (25.54-49.91)	NA*	25.31 (6.04-106.1)
		After 0.45-µm filter and chloroform	23.72 (10.50-53.55)	34.73 (18.68-64.55)	NA*	49.68 (20.06-123.1)
		After dead-end ultrafiltration	204.7 (120.8-346.9)	13.78 (6.03-31.52)	NA*	59.56 (53.98-65.71)
		After DNase	197.8 (71.15-549.9)	16.89 (9.55-29.86)	NA*	53.60 (49.71-57.80)
	PhiX174	After 0.45-µm filter and tangential ultrafiltration	39.99 (17.25-92.68)	58.56 (35.04-97.88)	NA*	41.52 (14.19-121.5)

		After 0.45-µm filter and chloroform	32.43 (16.41-64.08)	47.49 (26.05-86.61)	NA*	67.95 (25.84-178.7)
		After dead-end ultrafiltration	230.9 (114.4-466.0)	15.55 (4.50-53.74)	NA*	67.19 (38.12-118.4)
		After DNase	44.58 (19.80-100.4)	3.81 (2.74-5.29)	NA*	12.08 (9.38-15.57)
River Water	T3	After 0.45-µm filter and tangential ultrafiltration	40.33 (32.04-50.76)	67.49 (55.38-82.24)	3.09 (2.21-4.33)	13.03 (8.31-20.44)
		After 0.45-µm filter and chloroform	36.91 (18.59-73.28)	61.83 (31.96-119.6)	2.24 (0.720-6.95)	16.51 (5.82-46.88)
		After dead-end ultrafiltration	654.9 (359.3-1194)	53.30 (37.23-76.31)	12.19 (9.76-15.21)	53.73 (36.77-78.50)
		After DNase	413.7 (279.2-612.9)	42.87 (38.94-47.19)	0.936 (0.479-1.83)	442.7 (305.8-640.8)
Seawater	HS2	After 0.45-µm filter and tangential ultrafiltration	30.91 (24.60-38.85)	57.02 (48.49-67.04)	4.21 (1.43-12.44)	7.34 (3.11-17.28)
		After 0.45-µm filter and chloroform	22.90 (13.68-38.33)	42.24 (27.10-65.84)	2.44 (0.894-6.65)	9.39 (5.74-15.39)
		After dead-end ultrafiltration	160.3 (112.5-228.5)	14.83 (9.67-22.73)	1.22 (0.055-27.06)	131.3 (7.48-2303)
		After DNase	149.9 (50.30-446.8)	17.59 (5.34-57.93)	1.60 (0.503-5.11)	93.59 (48.29-181.4)
	HM1	After 0.45-µm filter and tangential ultrafiltration	35.90 (30.05-42.89)	66.23 (56.58-77.51)	NA*	8.52 (3.43-21.19)
		After 0.45-µm filter and chloroform	23.81 (20.37-27.83)	43.92 (41.40-46.60)	NA*	9.77 (3.98-24.00)
		After dead-end ultrafiltration	215.7 (155.9-298.4)	19.94 (12.89-30.83)	NA*	176.5 (10.88-2864)
		After DNase	252.0 (74.96-847.4)	29.55 (7.95-109.9)	NA*	157.2 (79.49-310.8)

	ICBM5	After 0.45-µm filter and tangential ultrafiltration	25.66 (20.23-32.55)	47.34 (41.11-54.50)	NA*	6.09 (2.35-15.82)
		After 0.45-µm filter and chloroform	20.63 (12.93-32.89)	38.05 (23.96-60.43)	NA*	8.46 (4.23-16.94)
		After dead-end ultrafiltration	347.4 (150.9-799.6)	32.11 (14.24-72.39)	NA*	284.3 (11.59-6974)
		After DNase	108.9 (54.85-216.1)	12.76 (5.74-28.40)	NA*	67.90 (29.18-158.0)

Table A.14: Iron chloride flocculation and purification step-by-step geometric means of recovery and phage to 16S rRNA enrichment for the spike-and-recovery experiments presented in Figure 2.3. A significant stepwise reduction in recovery or change in enrichment was assessed with p-values from ANOVA analysis with Tukey's correction for multiple comparisons with p-values less than 0.05 bolded.

Matrix	**Phage**	**Step**	**Recovery (%)**		**Phage Enrichment**	
			Stepwise Geometric Mean	**Stepwise *p*-value**	**Stepwise Geometric Mean**	**Stepwise *p*-value**
Influent	T3	After 0.45-µm filter and iron chloride flocculation	69	**0.022**	840	**7.6E-04**
		After 0.45-µm filter and chloroform	75	0.18	0.73	0.33
		After DNase	36	**0.032**	1.7	**0.045**
Effluent	T3	After 0.45-µm filter and iron chloride flocculation	78	0.35	41	**3.0E-04**
		After 0.45-µm filter and chloroform	79	0.68	1.2	0.46
		After DNase	35	**0.046**	0.89	0.78
River Water	T3	After 0.45-µm filter and iron chloride flocculation	80	0.89	12	0.33

		After 0.45-µm filter and chloroform	64	0.58	1.1	1.0
		After DNase	48	0.49	1.5	0.84
Seawater	HS2	After 0.45-µm filter and iron chloride flocculation	63	**<1E-06**	0.84	1.0
		After 0.45-µm filter and chloroform	94	0.73	1.1	1.0
		After DNase	25	**<1E-06**	9.6	**0.027**
	HM1	After 0.45-µm filter and iron chloride flocculation	64	**<1E-06**	1.2	1.0
		After 0.45-µm filter and chloroform	130	**1.2E-04**	1.3	1.0
		After DNase	34	**<1E-06**	11	**1.6E-05**
	ICBM5	After 0.45-µm filter and iron chloride flocculation	57	**<1E-06**	1.1	1.0
		After 0.45-µm filter and chloroform	71	**1.3E-03**	0.70	1.0
		After DNase	5.3	**<1E-06**	1.7	1.0

Table A.15: Ultrafiltration and purification step-by-step geometric means of recovery and phage to 16S rRNA enrichment for the spike-and-recovery experiments presented in Figure 2.4. A significant stepwise reduction in recovery or change in enrichment was assessed with p-values from ANOVA analysis with Tukey's correction for multiple comparisons with p-values less than 0.05 bolded.

			Recovery (%)		**Phage Enrichment**	
Matrix	**Phage**	**Step**	**Stepwise Geometric Mean**	**Stepwise *p*-value**	**Stepwise Geometric Mean**	**Stepwise *p*-value**
Influent	T3	After 0.45-µm filter and tangential ultrafiltration	82	0.58	61	1.0

		After 0.45-µm filter and chloroform	92	0.99	30	0.10
		After dead-end ultrafiltration	67	0.30	1.4	0.80
		After DNase	93	0.99	1.4	0.63
Effluent	T3	After 0.45-µm filter and tangential ultrafiltration	74	**2.8E-04**	26	0.37
		After 0.45-µm filter and chloroform	102	0.98	2.3	0.42
		After dead-end ultrafiltration	51	**<1E-06**	1.3	0.56
		After DNase	107	0.99	0.89	0.95
	T4	After 0.45-µm filter and tangential ultrafiltration	36	**<1E-06**	25	0.85
		After 0.45-µm filter and chloroform	97	1.0	2.0	0.91
		After dead-end ultrafiltration	40	0.076	1.2	1.0
		After DNase	123	1.0	0.90	1.0
	PhiX174	After 0.45-µm filter and tangential ultrafiltration	59	**5.9E-05**	42	0.50
		After 0.45-µm filter and chloroform	81	0.65	1.6	0.85
		After dead-end ultrafiltration	33	**2.2E-03**	0.99	1.0
		After DNase	25	0.48	0.18	0.25
River Water	T3	After 0.45-µm filter and tangential ultrafiltration	67	**6.3E-03**	13	0.99
		After 0.45-µm filter and chloroform	92	0.97	1.3	1.0

		After dead-end ultrafiltration	86	0.65	3.3	0.61
		After DNase	80	0.56	8.2	**<1E-06**
Seawater	HS2	After 0.45-µm filter and tangential ultrafiltration	57	**<1E-06**	7.3	1.0
		After 0.45-µm filter and chloroform	74	0.076	1.3	1.0
		After dead-end ultrafiltration	35	**1.2E-04**	14	0.72
		After DNase	119	0.94	0.71	0.95
	HM1	After 0.45-µm filter and tangential ultrafiltration	66	**5.0E-06**	8.5	1.0
		After 0.45-µm filter and chloroform	66	**1.9E-03**	1.1	1.0
		After dead-end ultrafiltration	45	**9.1E-04**	18	0.48
		After DNase	148	0.17	0.89	0.96
	ICBM5	After 0.45-µm filter and tangential ultrafiltration	47	**<1E-06**	6.1	1.0
		After 0.45-µm filter and chloroform	80	0.47	1.4	1.0
		After dead-end ultrafiltration	84	0.86	34	**0.027**
		After DNase	40	**6.0E-03**	0.24	0.066

Table A.16: After DNase treatment phage gene copy recoveries and phage to 16S rRNA enrichments for ultrafiltration and iron chloride flocculation for each matrix. The geometric mean and geometric standard deviation from the triplicate data is provided. Individual t-tests were performed for each matrix and phage spike.

Matrix	**Phage Spike**	**Virus Recovery (%)**			**Virus to 16S rRNA Enrichment**		
		Ultrafiltration	**Flocculation**	***p*-values**	**Ultrafiltration**	**Flocculation**	***p*-values**
Influent	T3	47 (24, 93)	25 (16, 38)	0.066 (ns)	3600 (1500, 8600)	1000 (680, 44)	0.049 (*)
Secondary Effluent	T3	42 (37, 49)	21 (4.7, 99)	0.21 (ns)	70 (32, 150)	44 (26, 75)	0.97 (ns)
	T4	17 (9.6, 30)	NA	NA	54 (50, 58)	NA	NA
	PhiX174	3.8 (2.7, 5.3)	NA	NA	12 (9.4, 16)	NA	NA

River Water	T3	43 (39, 47)	25 (6.6, 92)	0.11 (ns)	440 (310, 640)	19 (5.3, 71)	8.0E-3 (**)
Seawater	HS2	18 (5.3, 58)	15 (13, 16)	0.47 (ns)	94 (48, 180)	8.8 (2.6, 29)	0.034 (*)
	HM1	30 (7.9, 110)	28 (23, 34)	0.68 (ns)	160 (79, 310)	17 (4.5, 64)	0.034 (*)
	ICBM5	13 (5.7, 28)	2.2 (1.4, 3.5)	0.042 (*)	68 (29, 160)	1.3 (0.54, 3.2)	0.041 (*)

A.9 References

1. Deng, L. et al. Contrasting life strategies of viruses that infect photo- and heterotrophic bacteria, as revealed by viral tagging. *mBio* **3** (2012).
2. Nadkarni, M.A., Martin, F.E., Jacques, N.A. & Hunter, N. Determination of bacterial load by real-time PCR using a broad-range (universal) probe and primers set. *Microbiology* **148**, 257-266 (2002).
3. Lim, S.W., Lance, S.T., Stedman, K.M. & Abate, A.R. PCR-activated cell sorting as a general, cultivation-free method for high-throughput identification and enrichment of virus hosts. *Journal of Virological Methods* **242**, 14-21 (2017).

Appendix B Supplementary Information for Evaluating Limitations of Quantitative Metagenomics with Synthetic dsDNA and ssDNA Standards

B.1 Sample characteristics and sequencing results

Table B.1: Sample Characteristics. pH was measured with a Mettler Toledo pH meter calibrated immediately prior to measurement with 4, 7, and 10 pH standards. Turbidity was measured with a Hach 2100N Laboratory Turbidimeter. The total suspended solids (TSS) and volatile suspended solids (VSS) were measured in each sample using standard methods with 80 mL of sample stored at -20˚C until analysis [1].

Matrix	Collection Date	Total Fold Concentration	pH	Turbidity (NTU)	Solids Content		Phage T3 Recovery (%)
					TSS (g/L)	VSS (g/L)	
Raw Influent	12/19/20	554	7.47	31.5	0.103	0.057	32
	12/21/20	534	7.44	42.0	0.127	0.074	45
	12/23/20	554	7.55	54.5	0.121	0.082	33
Secondary Effluent	12/20/20	973	7.04	2.26	0.029	0.011	33
	12/22/20	1099	7.19	14.8	0.056	0.034	22
	12/24/20	1031	7.20	9.65	0.045	0.025	33
Deionized Water Blank	12/26/20	969	7.64	0.039	NA	NA	NA

Table B.2: ssDNA standard sequences.

>NC_000936.1_S3_length_1000_GC_0.317
GAATGCACGTCTATTACTTTGACACTTTAGGAAAGAACTACGATTTTCACTCCTTTACACAATAACCT GAAACGTACTGAAATTATGAAAAATAAGTGTTGGTATAACATTAGGTTGTAAATACCGACCAATAGT TGGTAATCTTCTATTTATGATAAGACAATTACTAAGTAACTGTTAAAATTTATTCGATAATCGTATAG TCAAACTACTACTGCAACAAGTCTAATTTAGTAAAAATTTATGTAAAATATTACACGAACTCGCATTT AATGGTTTCACATTATGTTATACACATACCAGAGGAGGTTCACGTCCTTTCTTGAAGAAAAAACTACA CATAGATGTAATAGATTACTTGATACCTGTTAATCCATATTACTTATTCTGATTATTAAAGTCAAATGT TCTTCGTTGTAGATTTGCACACGAAAACACCTTACTTGGTTTGATACTCCTACGTATATGTCTGTGAAA CTTTTACAATTGTCCACCACTACGAGATACACATTCTCATTTTGTTTTTTTTCTAACAGTACATATATT CTGTGGTGAATAACATGAATGTTTATTATACTAACCTAAGTACGTACTTGATCGTAAACATCTATCTC ACTTCCATATATCTACCTTTGTTCGAGGTAAGGATCGACTCATATTATTCTTTGGATTAGGCGAACATC GAAAACTTTATAACCACATAACCCTTTATTAAGGTTCCTATTAACTTGTTATTTTATATTTAAGTAAGT ACAATGACAACAAAATAACGAGACTGTTCCTTCGTAATTGTATGTTAGGTCGTTTACCTTGTCATAAT GACAGTTTTCCATCACTAAAGGGTAATACAGGAGAATAAAGGAGAGGTCGTCGTGTTAATATAAGAC

TCCGTTTACCAGACAGTCATATACAAGACATAGGATGTTGTGTAAGTGTTCATTGTAGTTTTATACTAGACCATAGTCATTTACTACTTAATCATAGAGGTGGTCATTAACGT
>NC_027637.1_S2_length_1000_GC_0.4
CTTAGAAGTCCTTATTGTGGACTCTTTATTAGTCGATAAACCATAAATTCATTTCGAACAAGCGGCACTCGAAAGAACTCCTATATGGAACGTTGTATTCGTCGTCTCAGTTTCCAATGCCTTGGTTATCTTTGTTCTGGAAATGGGGTCTAAGCCCTTTCCCATAGTCTTCATATTTCTCCATATGGTAATGTTAGGAACTCTATAAACGCTAGTTGTTTGACCTCTAATTTGTTTTATCGTACTATTACTCCTGCTGGCCTTCTAAGCGGCATGAGTGGTGTGTACATTATCGCCTCTAATCTTACTAATCTAAAATCTTTTTCTAGTCCTCGTAGTCTTTAGGTAAGCGCATTGGCGAATTATTTTTCGACTTTTAGTGCGAATTGATTTCTTGGTATGAATCCATTTACGAGACCCCTTATTCATTCTCACTATTTTGTTAACATTAGTATTCTTAGTCGGAGTACGTATGTGGAATTTCTGGTGAGTGCTCTTGCAAATTCTGCTGTTGGTTGTGTAGCCGGTGTTGGCTATCATTATCCCAATTCTTGCGGTCGTGGCCGGAAATTTACCAATTTTTGTTGGAACGGTCTCCATCCAACTGCCTTTGGAACTTACCCACTCATTGTTCCGTATCTTATAGAGGACACATCTCAACTATTTTTTTGCTGTCTGTGTGGGGAAGGCTAGGCCAGAGTTGGAAACCTTCCCCATGGAATTTTGATGTTAGCTTATGGTGGTGCGTAGCCAAAGGATTGTACCAAAAAAGACACGTTAAACGCTCGTACTGCTAATATTCAGAACGCGGGAAATACACGTCTGTACCTTGCAGCAAAGTATTTTTTTGGAGGTTTTTTACATTTTTTTTGTTTTTGTCTTCCTTGTCCATCAAGTTTTTCGCTTGATGGACAGTCAGTCTGGTATATGTAGTTCACCCATATACCAGACTGATTATTTTCGAGTATAAAAGAGGAGGAAGCCGCCTCTTTATGCTAGT
>NC_039057.1_S1_length_1000_GC_0.45
CTAAGTATGTCCCACTATGTGTATAAATTCATCATCAATTCCACTATTGCAATGACCCACTTAATTATTACCAAATGGCGTCGATACCTATATCTTCCCGAATGTCTGGGTAATCCGCAAAGTCTATGGTACCCGCGTCTTCCACCACTTTACCAGTGGCACAGATAGCTTCTCTACATGCTATACCGATGGTCAGAATTATTCTAACCATAGTATCCATATGTATGGGGGTTAGAACCGGTCTAACGCGCGTCCACCCGTCCGGAGTACTAATTAGTATTCAAGGACGCACAATTCAGAACACTACAAGTAGAACGAACGCGAAGATACGAAGGACGACTAGGAAACGTCTAACCGTGATGATGACCTTGATAACGTGGCGTCCTATACTACTTAGGAGAAGAGATAGCGCGTCAAAGATTACTCTCAACCTTACGAAAATGAAGAGCAGAGAAACGGTGAAGACCGCGAAGACAATTAGTCTTAAGACAGTTCCACAAATGACTTCTCAAGAGATGGGTCAGACTGCTTTTATTACTCTTCTATATGATATTACGGGAGAGACTTGGACCGACCTTCTTCCGAAAGGGGGTCAGACCAGAGCTATAGTTACCCGAATTCGGAGAACATATGGGGTAAGAATGGGTCAAGCCATTGCCACTCCGTCCTTAGTTGTTGAAGCCAGAAGAGCCGTCATCGCAAGTCCGATGATGAGGCCGGCCTTTAGGATGGTGACAGTGATTAGGATGATGACCATGACGTGTCAACCTATTGCGGTCTATAAAGGCACCGTTCCGAGTTGGATACGGGTCTAAGGGCTGAAGTATAGGAATACAGTCATGACTATGACCAAGATGACACCCACTATGAACCTGAGAGTACAAAGTCCTAAGGATAGTCTATGGGGCATGGATACAGCGAACATAAGATTGAGGAGGACCCTTTGAGTAGTTCAAAATAATGTCCGAGGCGTAGCAGACCATGTATCTTAAGAGACTAT
>NC_025708.1_S1_length_1000_GC_0.5
TGGGGCAGTGTATAGACGAAGCCGGGACTGCCCGGCTGTGCATGGAGCGTGCGTGTATAAGACTCAAAAGTCAGAAGTAAAGTAAAGTAAGAAAATACAGATAAGGTCGAGTCTGGACCTTATGAAATCCGAAACCATGGTCTCGATGGCGGCGGCTCCGTGATCGTCCTCTGCGTAATTAGCGTCGTCCGGGTCATCGGAATCCTCGTCGTCGGGATCCTCCATATCGAATGCGACCTATGTTTGTTTACCGTGCTGCTGCGGGTGCTAGTAAGGGTCCTGGTATGGCTGGCTTCTTTGCACCAAGGATGCCTGGTTCTTCGTCTGGCGGAAGGTCCGCTATGTCCGACGGTTTGAACGCTTAACCTCCAACCGAATTAGCTCTTGAATTTAAATTCCTGCAGGTCGCATTACACAGGTAAAGTTGTTGATAACGCCCCTCACTTTAGCTAGGACTGCCGCCACAGTTGACGGAGTTGCGGTAGTCAGTTCCACTGCCCCTTCTTGTTGCGCTGCCGGCACCAATGTTTTAGTGTAGTCAGGTGGAATCTCCAATGCAGGACAAGCGGCCCGTTAGGCCGCGTCCTCGTTGGAGCCTGGTGTACTCCTAGTAGGATGATGTCCTATGTGTTTGCTTGCCTTGTGTCAAGTTACGGCTTCTGCAGTAACTATTGAGGAGGCCTCAATTACTCGATGTCTGGCAACGAAAGGCTTTAGAGCTCTTATGCTTAGCGAAATTCTAGAATTTCCTGTAGCAACATGTGTTCGGCTGATGGCCTGAACGTCCGTTGGTTCGATTGGGGCCACTGCAACTTAGGTTAAGGCGATGTTACGGTTAAGTTTACCTGCAGTTGAAGTTCATGGGTCAATTGCAGGAAACGTGGCCGTGTCCACCTTGCCAAAGATTAGAATGGCTATTAAGCAAGGTAAATTACCGTTAGTCACAGTTAAGGCCCTTGTCAAAGGCTATGTAAACGGCTTGCGCAAAGCAACCGATTTA
>NC_010429.1_S4_length_1000_GC_0.6

```
CGCGTATCCACCGGCGGGTACGGCTTTCCTAGTAGCACTGCCGGTGCGGCTTGGGCCACATTAGCTTT
ATCCGCTGTTCATGCCCCGGCTCCCTCCTCTAAAACGAGACTTACTTCGAGAACTGCCGACCATCACG
GTGGAGCACTCACTGCTCCGGTTATGGCTTTCGTTCGTCTTGCGCGGGCTACTACGGGCCATCCATGA
CCGCCTCGCGGACGTCCTCGTACGAGTGTTCGCGCCGTTACTAGTAACCGTAATCAAGAAGGGAGGT
GACCATTCCTACTTCCGCGTGCACTGACAACGCCTTTCTTACTAGTGACGCGCTTAGTCCTACTACATC
GACGTCTTCTTCAGCCCTCCGCTACGGTAGCGAACGCGTTTCGGGCTCGGCTACGGCGACGGCGGCG
ACTGGCGTTTAGGCTACGTCGACGGCGGTGGCGGCTTCGGCCTAACTAGGTTGACGAGCTGGGGCTC
GGAGAACTGGCAATCCTGAAAGCACAGTAGCCCACTTCGTGGTTCAGGGCCTAGCCACCTGTACGGC
GGGTGGAATCAGGTGAGTCCGACGCGCGGTAGAGGCGGCAACAACAACGGGACCGACCGCTCGAAG
AGACGCCGTGTGCGCCACGCGGTAACGTAGTTCTCCGGACGCATGAGGTAGCGTAGCGTGAAGAGTG
GGCATGTTTGTCCGTAGCGCCGTGTCCTTGGAGGACGCTACAAGGCCAAGGAAGCACATAACGTCAG
TTACGCGGTAACCTACGCCCGGACTGGCGTTTCCTAGCCACTAGGTGATGTGCTCCCGCCGTACGACG
TTAGTAGCGGTGGACGTTTCGAGAGGAAAAGCGGGCCGAGTAGCCCGACTGGTAGCAACAGCCTTAG
GAAGAACGTCCAAGGTAGACCTGGCGCTTGGAGCGGAAACGCGTGACGGGCAGTGGCCCATTCGT
GGCCGGTAGACCGCTGGAGTGGACGGGACGACCTGTAAGCAAGAGGAAGAAGAACGTG
```

Table B.3: ddPCR primers and probes.

Target *[2]	**Primers/Probes**	**Amplicon Length (bp)**
*Enterobacteria*phage T3*	Forward: 5'- CCA ACG AGG GTA AAG TGA TAG-3'	351
	Reverse: 5'- CGA CGA TAG CGA ATA GGA TAA G-3'	
	Probe: 5'-/HEX/-CC AAC AAC ATC TCT CGC GCA TT-/BHQ_2/-3'	
Sequin Metagenome Mix A Standard S1106_MG_020_A	Forward: 5'-CGA CCA CCC AAA CAG GTA CA-3'	170
	Reverse: 5'-CCA CGC AAC TTT TTA CGG CA-3'	
	Probe: 5'-/FAM/-CG ACC ATG GTG GAC GTA TAG GCA AT-/ZEN/3IBFQ/-3'	
ssDNA Standard NC_039057.1_S1 (45% GC content)	Forward: 5'-CGC GAA GAT ACG AAG GAC GA-3'	150
	Reverse: 5'-GCG GTC TTC ACC GTT TCT CT-3'	
	Probe: 5'-/FAM/-TG ACC TTG ATA ACG TGG CGT CCT AT-/ZEN/3IBFQ/-3'	
Marine Phage HM1*	Forward: 5'-CGT CTG CAG TAG ATT GGG CA-3'	216
	Reverse: 5'-AGA TGG GGT GTT GGA GGA AAG-3'	
	Probe: 5'-/FAM/-CAA GAA CAG GAC TTG CCA GAA GTG T-/BHQ_1/-3'	

Table B.4: Spike-in concentrations and Illumina NovaSeq and Oxford Nanopore GridION sequencing summaries.

Matrix	**Collection Date**	**Total DNA Conc**	**Technical Replicate**	**Illumina NovaSeq Spike-in ddPCR Measured Concentrations (gc/µL DNA Extract)**	**Number of Illumina NovaSeq**	**ONT GridION Sequencing Statistics**

		(ng/µL DNA Extract)		dsDNA Standard S1106_MG _020_A	ssDNA Standard NC_ 039057.1 _S1	Marine Phage HM1	251 bp Paired-end Reads after Quality Control	Passed Bases (Gb)	Mean Read Length (kb)
Influent	12/19/20	26.0	1	3.54E+05	5.20E+05	9.17E+02	197,484,150	4.68	11.1
	12/21/20	25.7	1	4.18E+05	5.74E+05	8.62E+02	113,211,076	4.79	10.0
			2	4.06E+05	5.88E+05	8.76E+02	133,609,632		
			3	5.57E+05	6.40E+05	1.11E+03	143,723,666		
	12/23/20	21.8	1	2.90E+05	4.36E+05	7.23E+02	206,205,456	3.02	10.2
Effluent	12/20/20	58.3	1	1.13E+06	1.04E+06	1.34E+03	190,828,706	4.81	17.8
	12/22/20	54.3	1	6.60E+05	7.30E+05	1.30E+03	189,889,444	4.60	16.8
			2	8.57E+05	8.44E+05	1.28E+03	181,632,578		
			3	7.97E+05	7.44E+05	1.07E+03	218,629,348		
	12/24/20	48.5	1	6.90E+05	6.95E+05	1.60E+03	168,440,530	5.33	12.3

B.2 Establishing Detection Thresholds

To evaluate the accuracy of relative entropy, “R”, to summarize coverage and read distribution, a binary logistic regression model was created with results from mapping reads to the spike in standard sequences and the fail test set across all samples including results of 20% and 1% downsampling each virome (equation B.1). Passing or failing detection was assigned using the cut-offs proposed by FastViromeExplorer[3]. If mapping to a standard passes detection, the coverage is greater than or equal to 10% and the observed to expected read distribution ratio is greater than or equal to 0.3. We converted the number of reads requirement, as used for the FastViromeExplorer pipeline, to the number of basepairs from reads because read lengths were longer at approximately 235 bp with the NovaSeq SP flow cell than the 150 bp long reads used to validate the FastViromeExplorer pipeline. We tested basepair count cut-offs of 1,000, 400, and 0 bp. We observed that employing a basepair count cut-off only impacted short target lengths (Figure B.1). Therefore, we did not include a minimum number of basepairs mapped to our target in our logistic regression model. The final logistic regression model has a ß$_1$of 164.97 (p-value =1.39 x 10^{-9}) and ß$_0$ of -120.45 (p-value = 8.64 x 10^{-10}).

$$P(detection = 1) = \frac{1}{1+\exp(-(\beta_1 \cdot R+\beta_0))} \qquad \text{(B.1)}$$

The logistic regression models were tested with results of mapping each virome to the NCBI DNA viral database divided into genomes and their respective genes and the VirSorter curated database. Ideal logistic regression performance has an area of one under an ROC curve. With all of the mapping results from each database combined, the area under the ROC curve is 0.998 indicating the model has a high sensitivity and specificity. The detection threshold was calculated by bootstrapping to maximize the sum of sensitivity and specificity at the optimal "R_{detect}" threshold. We observed that the optimal "R_{detect}" threshold differed between databases composed of genomes and genes. The optimal "R_{detect}" for the NCBI viral gene database, NCBI viral genome database, and VirSorter curated database was 0.704, 0.744, and 0.743, respectively. We hypothesized that this observation was due to a difference in target lengths between databases composed of genomes compared to genes. Targets were binned based on their lengths and the optimal "R_{detect}" was calculated for each bin (Table B.5). There is a significant relationship between length and optimal "R_{detect}" (Figure B.2, R^2=0.92). Therefore, the detection threshold varies with length of the target. However, the mean length of the final bin is 350 kb and a maximum length of 2.5 Mbp, so it is unclear if the relationship extends to longer targets. Based on this observed relationship, a target is confidently detected in a metagenome if its respective "R" is above its length dependent "R_{detect}" value.

Table B.5: Detection threshold with length raw data (for Figure B.2).

Range of Target Lengths (bp)	Area Under the ROC Curve	Optimal $R_{G,detect}$	Probability of Detection	Number of Targets
0-500	1.000	0.557	0.0000%	39,045

500-1,000	1.000	0.644	0.0001%	41,465
1,000-2,000	0.999	0.693	0.1979%	37,887
2,000-3,000	0.999	0.704	1.22%	15,117
3,000-4,000	1.000	0.710	3.31%	5,397
4,000-5,000	1.000	0.723	27.13%	2,047
5,000-6,000	1.000	0.735	68.00%	1,150
6,000-7,000	1.000	0.729	44.13%	1,142
7,000-8,000	1.000	0.732	55.16%	1,094
8,000-9,000	1.000	0.753	97.97%	1,006
9,000-10,000	1.000	0.748	95.28%	948
10,000-12,500	1.000	0.738	79.68%	4,039
12,500-15,000	0.999	0.740	83.12%	3,126
15,000-17,500	0.997	0.747	94.34%	2,755
17,500-20,000	0.999	0.736	71.88%	2,472
20,000-22,500	0.999	0.743	87.90%	2,441
22,500-25,000	1.000	0.738	74.74%	2,017
25,000-27,500	0.999	0.741	83.42%	2,112
27,500-30,000	0.999	0.762	99.39%	2,489
30,000-35,000	1.000	0.758	99.24%	5,603
35,000-40,000	0.997	0.735	77.27%	7,549
40,000-50,000	1.000	0.764	99.56%	11,743
50,000-60,000	1.000	0.770	99.86%	4,959
60,000-75,000	0.999	0.762	99.37%	3,830
75,000-100,000	1.000	0.772	99.94%	2,497
100,000-200,000	1.000	0.774	99.94%	6,097
>200,000	1.000	0.808	99.9998%	1,526

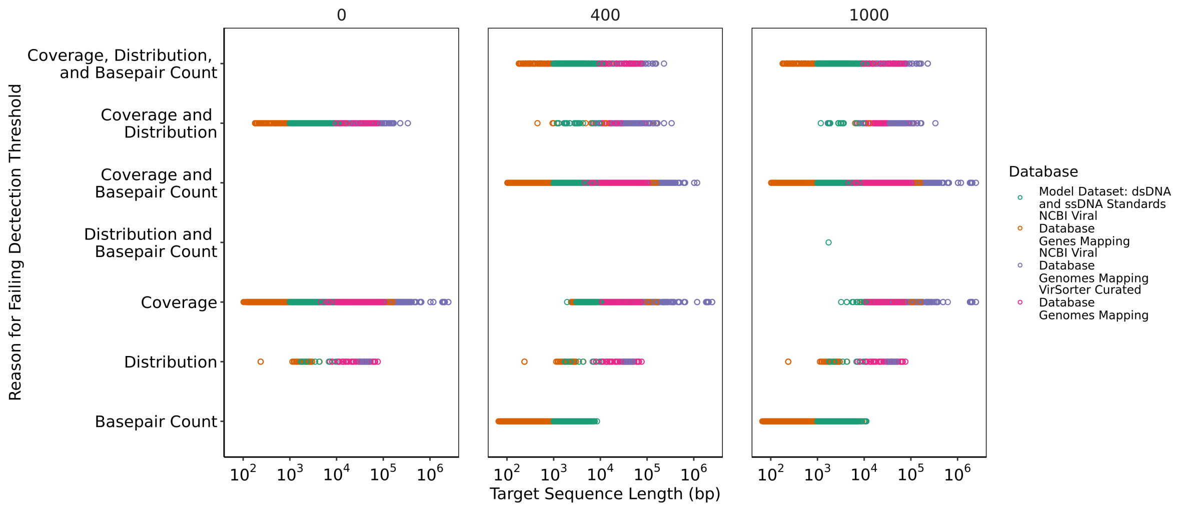

Figure B.1: Reasons for failing to meet detection threshold for different minimum basepair requirements (0, 400, and 1,000 bp per target) with respect to the target sequence length. Basepair count is redundant for long targets and only serves to increase the detection threshold of short targets.

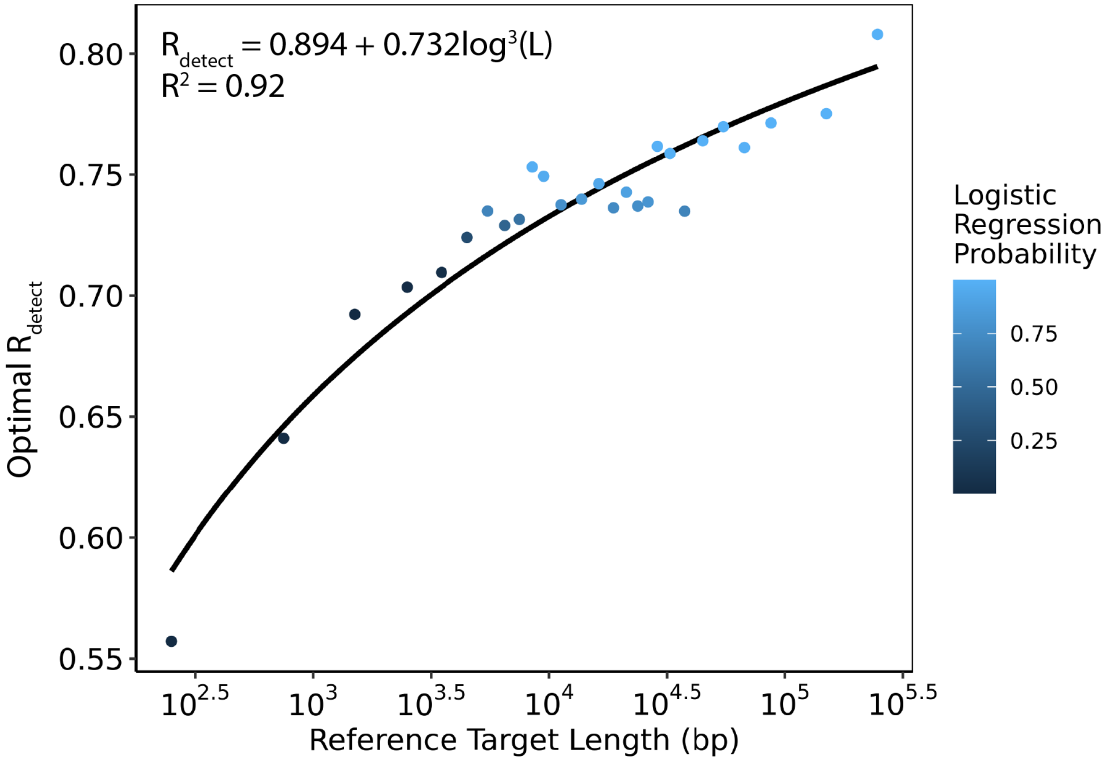

Figure B.2: The optimal R_{detect} threshold varies with target length with a significant relationship (R^2=0.92). Mapping results from all databases were pooled and divided into subsets based on sequence lengths, then the optimal R_{detect} was calculated for each length subset (Table B.5).

B.3 Read Distribution Patterns Across Detected Targets

To improve quantification, mapping errors need to be detected and corrected. To do so, we evaluated the variability in read distribution across spike-in standards and created regressions to predict read depth variability without non-specific mapping. The number of reads mapping to each basepair along standard sequences varied, as expected. The spike-in standards were designed to be unique from known microbial genomes and, therefore, we assumed non-specific mapping did not occur to the spike-in standards. We determined how the observed read depth varied across 49-bp long windows shifted by 1-bp along each spike-in standard related to the respective local GC content (i.e., GC content of the 49-bp window). Separate regressions were developed for subsets of average read depths (e.g., 0-10, 10-100, 100-1,000, greater than 1,000 reads per bp) (Table B.6). Each regression included a quadratic polynomial of local GC content and the logarithmic transformed average read depth across the entire target. Each regression was amended to prevent overfitting and underfitting while striving for a normal distribution of standardized residuals (Figure B.3). A quadratic polynomial of the logarithmic transformed average read depth was included to improve fit for the lower average read depth ranges (e.g., 0-10 and 10-100 reads/bp). For the lowest average read depth range (0-10 reads/bp), the mapping coverage was highly variable so a quadratic polynomial of R_G was included to summarize coverage and read distribution. The regressions were developed with all dsDNA and ssDNA spike-in standards from all samples. However, DNA type was a statistically significant factor in each regression (*p*-value < 0.001) based on ANOVA of DNA type added as a factor to each regression. Creating separate regressions for dsDNA and ssDNA targets or including a DNA type factor into existing regressions is infeasible because DNA type is occasionally unknown for targets in databases such as the NCBI viral collection and challenging to predict in viral genomes originating from viromes or metagenomes[4-7]. Therefore, we chose to use the regressions developed with the complete pool of standards without a DNA type factor for predicting read depth distribution in unknown targets to test for non-specific mapping.

Table B.6: Regressions summarizing the variability in read depth along 49-bp long windows shifted by 1-bp for each spike-in standard with respect to the local GC content (i.e., average GC content of each 49-bp window). Separate regressions were developed for four ranges of average read depth with polynomial terms and RG incorporated to prevent underfitting or overfitting. The normality of standardized residuals for each regression is plotted in Figure B.3.

Average Read Depth Range	**Read Depth Variability Regression**	**R^2**
≥ 1,000 reads/bp	$Read\ Depth_x = -22981 + 519(GC_x)^2 - 1732(GC_x) + 3426 \cdot \ln\ (Avg\ Read\ Depth)$	0.76
100 - 1,000 reads/bp	$\ln(Read\ Depth_x) = -0.090 + 0.370(GC_x)^2 - 0.183(GC_x) + 1.003 \cdot \ln(Avg\ Read\ Depth)$	0.83
10 - 100 reads/bp	$\ln(Read\ Depth_x + 1) = -0.667 - 0.141(GC_x)^2 + 0.221(GC_x) - 0.047 \cdot \ln(Avg\ Read\ Depth)^2 + 1.323 \cdot \ln(Avg\ Read\ Depth)$	0.71
0 - 10 reads/bp	$\ln(Read\ Depth_x + 1) = -2.982 + 0.373(GC_x)^2 - 0.337(GC_x) + 0.146 \cdot \ln(Avg\ Read\ Depth)^2 + 0.349 \cdot \ln(Avg\ Read\ Depth) - 2.551(R_G)^2 + 6.319(R_G)$	0.58

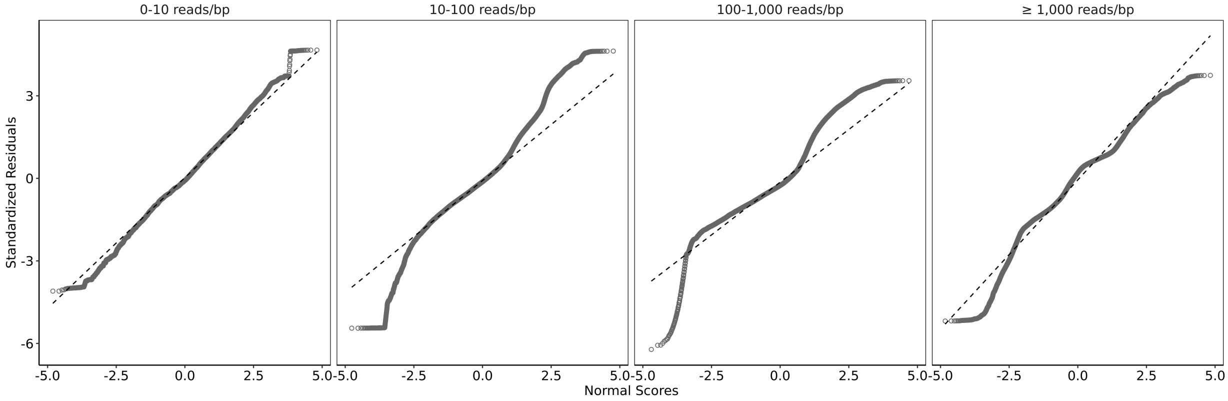

Figure B.3: Each regression was amended to prevent overfitting and underfitting while striving for a normal distribution of standardized residuals. Q-Q plots were examined to assess the normal scores and standardized residuals generally follow a 1:1 relationship.

Separate acceptable read depth variability thresholds were established for each read depth variability regression (Table B.6). The root mean square error (RMSE) for reads mapping to each standard summarized the difference between predicted and observed read depth long each target. RMSE was calculated for read mapping to each spike-in standard from all samples including without downsampling and 20% and 1% downsampling results. The RMSE of read mapping to spike-in standards linearly increased for average read depths between zero and 1,000 reads/bp and remained relatively constant for average read depths greater than 1,000 reads/bp (Figure 3.4A). RMSE was expected to increase with average read depth because the magnitude of read depth variability is relative to the average read depth. The acceptable read depth variability thresholds are the highest RMSE of standards with respect to their average read depths translated up by $e^{0.25}$ RMSE (Table B.7). Targets with a RMSE greater than the acceptable read depth variability threshold likely have non-specific mapping and require correction before calculating the targets' absolute abundance.

Table B.7: Acceptable read depth variability thresholds based on the root mean square error (RMSE) comparing the observed and predicted read depth variability for spike-in standards. Separate thresholds were created for each read depth variability regression. The thresholds are the linear trend lines through the standards with the highest

RMSE translated up by 0.25 (Figure 3.4A). Targets with a RMSE greater than the RMSE$_{max}$ at its average read depth are considered to have non-specific mapping.

Average Read Depth Range	**Acceptable Read Depth Variability Threshold**
≥ 1,000 reads/bp	$RMSE_{max} = 2026$
100 - 1,000 reads/bp	$\ln(RMSE_{max}) = 0.696 + 0.841(\ln(Avg\ Read\ Depth))$
10 - 100 reads/bp	$\ln(RMSE_{max}) = 0.880 + 0.770(\ln(Avg\ Read\ Depth))$
0 - 10 reads/bp	$\ln(RMSE_{max}) = 0.840 + 0.709(\ln(Avg\ Read\ Depth))$

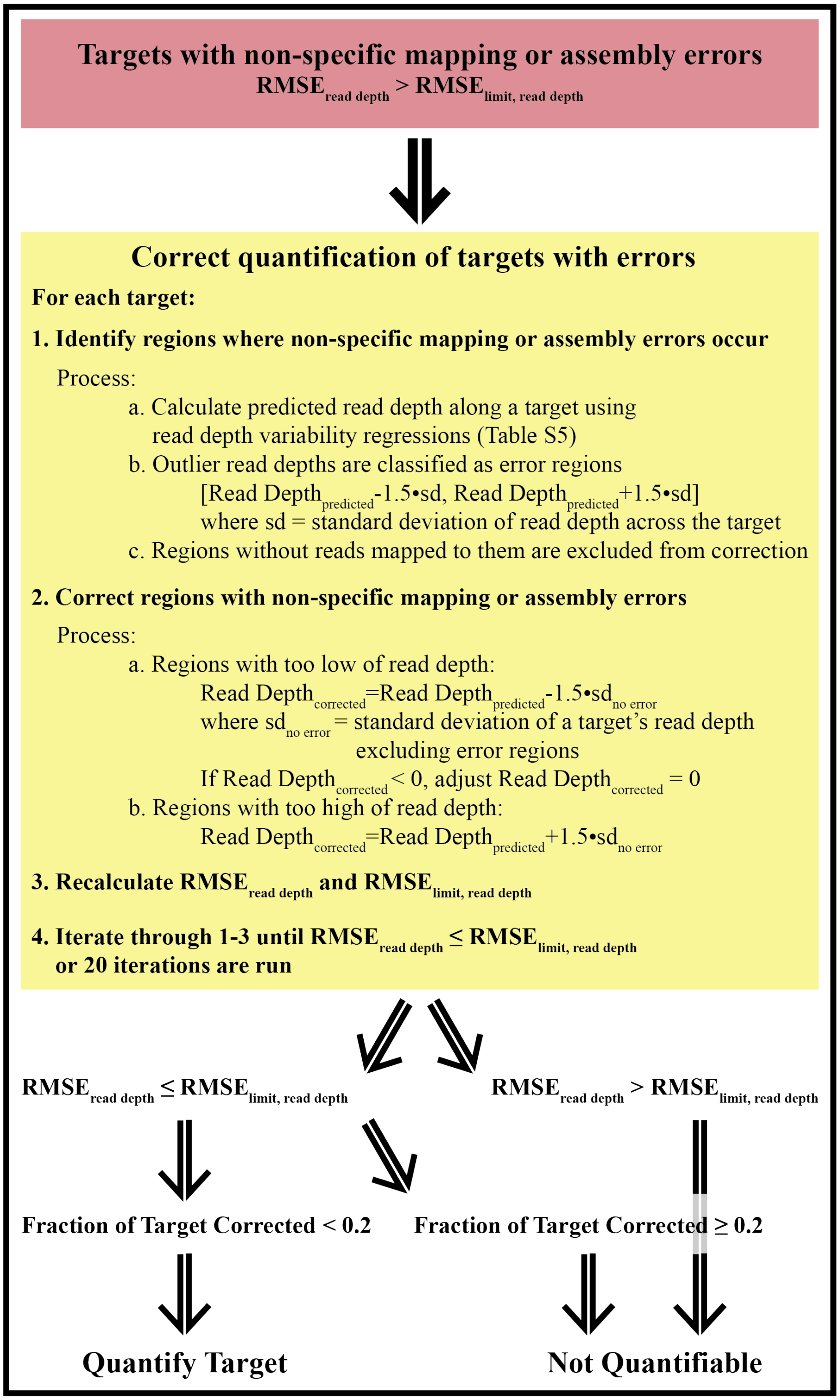
Targets with non-specific mapping or assembly errors
$\mathrm{RMSE}_{\text{read depth}} > \mathrm{RMSE}_{\text{limit, read depth}}$
Correct quantification of targets with errors
For each target:
1. Identify regions where non-specific mapping or assembly errors occur
Process:
a. Calculate predicted read depth along a target using read depth variability regressions (Table S5)
b. Outlier read depths are classified as error regions
$[\text{Read Depth}_{\text{predicted}} - 1.5 \cdot \text{sd}, \text{Read Depth}_{\text{predicted}} + 1.5 \cdot \text{sd}]$
where sd = standard deviation of read depth across the target
c. Regions without reads mapped to them are excluded from correction
2. Correct regions with non-specific mapping or assembly errors
Process:
a. Regions with too low of read depth:
$\text{Read Depth}_{\text{corrected}} = \text{Read Depth}_{\text{predicted}} - 1.5 \cdot \text{sd}_{\text{no error}}$
where $\text{sd}_{\text{no error}}$ = standard deviation of a target's read depth excluding error regions
If $\text{Read Depth}_{\text{corrected}} < 0$, adjust $\text{Read Depth}_{\text{corrected}} = 0$
b. Regions with too high of read depth:
$\text{Read Depth}_{\text{corrected}} = \text{Read Depth}_{\text{predicted}} + 1.5 \cdot \text{sd}_{\text{no error}}$
3. Recalculate $\mathrm{RMSE}_{\text{read depth}}$ and $\mathrm{RMSE}_{\text{limit, read depth}}$
4. Iterate through 1-3 until $\mathrm{RMSE}_{\text{read depth}} \leq \mathrm{RMSE}_{\text{limit, read depth}}$ or 20 iterations are run
$\mathrm{RMSE}_{\text{read depth}} \leq \mathrm{RMSE}_{\text{limit, read depth}}$
$\mathrm{RMSE}_{\text{read depth}} > \mathrm{RMSE}_{\text{limit, read depth}}$
Fraction of Target Corrected < 0.2
Fraction of Target Corrected ≥ 0.2
Quantify Target
Not Quantifiable

Figure B.4: Overview of the method to detect and correct quantification of targets with non-specific mapping or assembly errors.

Table B.8: Reads were mapped to standard derived contigs with and without quality control. Quality control removed low alignments and redundant fragment contigs from the pool of all standard derived contigs (see methods for additional details). Quality control reduced the number of standards with high read depth variability RMSE and

that were not quantifiable. However, without quality control, inaccurate standard quantities due to assembly issues were detected.

Standard Derived Contig Statistic	All Standard Derived Contigs	Quality Controlled Standard Derived Contigs
Number of Contigs	1,262	1,019
Number of Standards Represented	837	811
Number of Standards with High Read Depth Variability RMSE	103	45 (subset of the 103)
Number of Standards Altered by Quality Control	167	141
Number of Standards with High Read Depth Variability and Altered by Quality Control	76	18
Number of Standards with High Read Depth Variability and Not Altered by Quality Control	27	27 (same 27 standards)
Number of Standards without High Read Depth Variability and Altered by Quality Control	91	123
Number of Standards that were correctable	15	15
Number of Standards that were not quantifiable	88	30

B.4 Read-based and Contig-based Concentrations

Table B.9: Measurements of marine phage HM1 spiked in at a low concentration (~1000 gc/µL) into each DNA extract prior to Illumina NovaSeq sequencing. Contig-based virome derived measurements are not detected in 7/10 samples. (n.d. = not detected, n.q. = not quantifiable)

Sample	Marine Phage HM1 Concentration ($\log_{10}$ gc/µL DNA extract)		
	ddPCR	Contig-based	Read-based
12/19/20 Influent	2.96	3.51	3.02

12/21/20 Influent (Replicate 1)	2.94	n.d.	n.q. (2.65)
12/21/20 Influent (Replicate 2)	2.94	3.63	2.91
12/21/20 Influent (Replicate 3)	3.05	n.d.	2.94
12/23/20 Influent	2.86	n.d.	n.q. (2.92)
12/20/20 Effluent	3.13	n.d.	3.50
12/22/20 Effluent (Replicate 1)	3.11	3.98	3.28
12/22/20 Effluent (Replicate 2)	3.11	n.d.	3.34
12/22/20 Effluent (Replicate 3)	3.03	n.d.	3.44
12/24/20 Effluent	3.20	n.d.	3.33

B.5 Differences Between dsDNA and ssDNA Standard Regressions

To determine if there are differences between dsDNA and ssDNA in our viromes, the outcomes of the ssDNA standards were compared to the dsDNA standards. Two linear regressions were created to relate the known spike-in concentration to the predicted concentration with one regression including if a standard was dsDNA or ssDNA for all of the samples combined (n=890). The models were compared with ANOVA where the additional consideration of ssDNA or dsDNA standard type significantly impacted the linear regressions (*p*-value $< 2.2\text{x}10^{-16}$). The average residual of each standard from the linear regression relating the known spike-in concentrations to the predicted concentrations were calculated (Figure B.5). Differences between residuals for dsDNA and ssDNA standards is significant as determined with a two-tailed t-test (*p*-value $< 2.2\text{x}10^{-16}$). The differences in dsDNA and ssDNA standards may be due sequencing biases that preferentially sequence ssDNA slightly more than dsDNA.

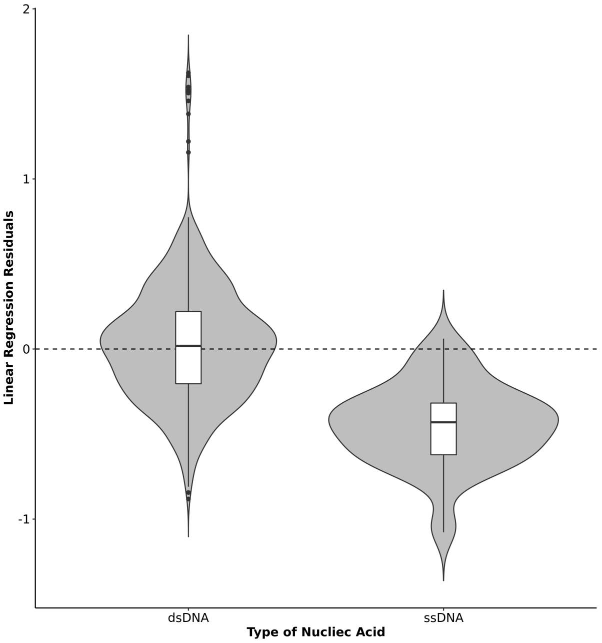

Figure B.5: A linear regression combining all standards across all samples was performed to relate the known concentrations of dsDNA and ssDNA standards to the predicted concentrations of standards. The regression residual for each standard per sample was calculated and differences between dsDNA and ssDNA standards are plotted in the violin plot.

B.6 Virome Quantification Results

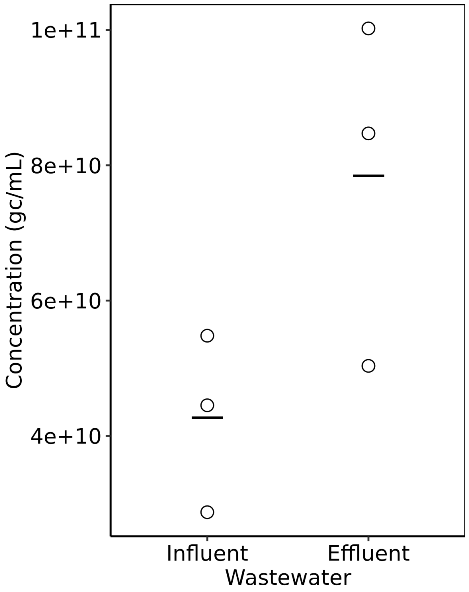

Figure B.6: Concentrations of viral populations in influent and effluent samples in gc/mL of wastewater. Dots represent viral population concentrations in individual samples with means indicated by black bars. Effluent has a higher abundance of viruses than influent (p-value = 0.12).

B.7 References

1. E.W. Rice, R.B.B., A.D. Eaton Standard Methods for the Examination of Water and Wastewater, Edn. 23. (American Public Health Association, American Water Works Association, Water Environment Federation, Washington, D.C.; 2017).
2. Langenfeld, K., Chin, K., Roy, A., Wigginton, K. & Duhaime, M.B. Comparison of ultrafiltration and iron chloride flocculation in the preparation of aquatic viromes from contrasting sample types. *PeerJ* **9**, e11111 (2021).
3. Tithi, S.S., Aylward, F.O., Jensen, R.V. & Zhang, L. FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data. *PeerJ* **6**, e4227 (2018).
4. Guo, J. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
5. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).
6. Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A. & Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).
7. Roux, S., Enault, F., Hurwitz, B.L. & Sullivan, M.B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).

Appendix C Supplementary Information for Parasite-host Coevolutionary Impact on the Emergence of Antibiotic Resistance

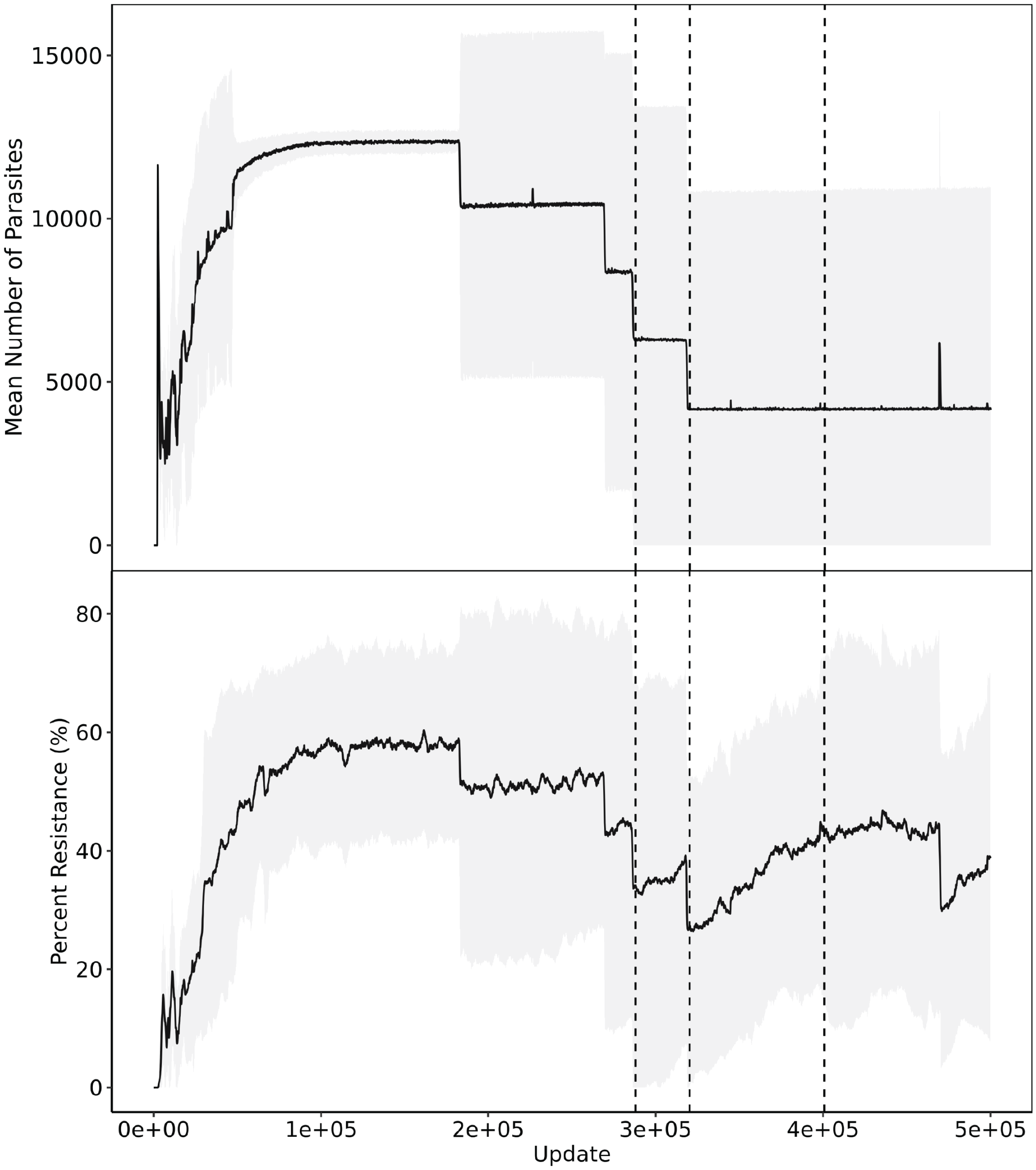

Figure C.1: Mean number of parasites (upper) and mean percent resistance (lower) for each update in the experiments with antibiotics and parasites where the Not task conferred resistance. Mean values for the six replicates with the shaded region representing the 95% confidence intervals. Dashed lines indicate updates when parasite extinction occurred in a replicate. Decreases in parasite abundance coincided with decreases in resistance expression.

Table C.1: Avida simulation conditions. The antibiotic inflow inflow rate and multiplicity factor were randomly selected. Not-And, And, and Not tasks were performed with 6 replicates for the no parasite control and experiment simulations.

Type	Number of Parasites Injected at 2,000 Updates	Replicate No.	Antibiotic Abundance/ Positive Resource Abundance	Antibiotic Initial and Inflow Abundance	Positive Resource Initial and Inflow Abundance	Multiplicity Factor	Resistance Task
Control: No Antibiotics	400	1	0	0	125	NA	NA
		2	0	0	125	NA	NA
		3	0	0	125	NA	NA
		4	0	0	125	NA	NA
		5	0	0	125	NA	NA
		6	0	0	125	NA	NA
Control: No Parasite	0	1	0.129	16.124	125	0.369	Not-And
		2	0.112	14.050	125	0.274	And
		3	0.138	17.311	125	0.286	Not
		4	0.145	18.159	125	0.257	Not-And
		5	0.101	12.614	125	0.271	And
		6	0.115	14.357	125	0.399	Not
		7	0.107	13.351	125	0.344	Not-And
		8	0.115	14.340	125	0.388	And
		9	0.116	14.556	125	0.321	Not
		10	0.110	13.798	125	0.323	Not-And
		11	0.145	18.169	125	0.351	And
		12	0.108	13.454	125	0.440	Not
		13	0.139	17.437	125	0.270	Not-And
		14	0.138	17.252	125	0.256	And
		15	0.113	14.111	125	0.410	Not
		16	0.138	17.286	125	0.269	Not-And
		17	0.143	17.824	125	0.491	And
		18	0.101	12.664	125	0.469	Not
Experiment	400	1	0.130	16.288	125	0.394	Not-And
		2	0.126	15.720	125	0.365	And

		3	0.147	18.375	125	0.484	Not
		4	0.148	18.524	125	0.353	Not-And
		5	0.110	13.759	125	0.352	And
		6	0.124	15.546	125	0.319	Not
		7	0.117	14.607	125	0.417	Not-And
		8	0.131	16.422	125	0.333	And
		9	0.124	15.486	125	0.351	Not
		10	0.117	14.649	125	0.354	Not-And
		11	0.115	14.415	125	0.438	And
		12	0.147	18.377	125	0.297	Not
		13	0.127	15.898	125	0.270	Not-And
		14	0.128	15.972	125	0.449	And
		15	0.122	15.192	125	0.385	Not
		16	0.115	14.344	125	0.271	Not-And
		17	0.101	12.600	125	0.480	And
		18	0.147	18.385	125	0.395	Not
Control: No Antibiotics or Parasites	0	1	0	0	125	NA	NA
		2	0	0	125	NA	NA