DISSERTATION

# DEMYSTIFYING VIRUSES: UNDERSTANDING THE ROLE OF RIVER VIRUSES ON MICROBIAL COMMUNITY STRUCTURE AND BIOGEOCHEMICAL CYCLING THROUGH A MULTI-OMIC LENS

Submitted by

Josué Rodríguez-Ramos

Graduate Degree Program in Ecology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2023

Doctoral Committee:

Advisor: Kelly Wrighton

Edward Hall
Jessica Metcalf
Michael J. Wilkins

Copyright by Josué A. Rodríguez-Ramos 2023

All Rights Reserved

ABSTRACT

DEMYSTIFYING VIRUSES: UNDERSTANDING THE ROLE OF RIVER VIRUSES ON MICROBIAL COMMUNITY STRUCTURE AND BIOGEOCHEMICAL CYCLING THROUGH A MULTI-OMIC LENS

Viruses are the most abundant entity on the planet, with estimates of up to $10^{31}$ viral particles dispersed across the globe in every ecosystem that can sustain life. Today, as the world responds to the COVID-19 pandemic, the word "virus" often evokes a negative response because of their impacts on human health and disease. Yet, most viruses that exist in the world can only infect bacteria and archaea. In fact, it has long been estimated that for every 1 bacterial or archaeal cell, there are 10 viruses that can infect it. While bacteria and archaea are long regarded as essential to overall ecosystem health and functionality, the roles of viruses in natural systems are much less understood and appreciated. Due to a scarcity of genome-resolved multi-omic studies, this lack of understanding is compounded in river ecosystems, which play critical roles modulating global carbon and nitrogen biogeochemistry. *The overarching aims of this dissertation are to harness genome-resolved, multi-omic datasets to 1) decipher the impact that viruses can have on river microbial communities and biogeochemical cycling, and 2) to explain how viral ecology can enhance our understanding of river ecosystem function.*

To define the role that viral and microbial communities have on river function, I first set out to understand what is currently known of river viral ecology. In Chapter 1, I provided a background primer on viruses and their impacts on natural ecosystems. I then zoomed in on viral roles exclusively within rivers and described the current state of river viral ecology. I also

highlighted some of the knowledge gaps addressed specifically by my thesis. My literature review revealed that while there are publicly available metagenomic datasets, there is a drastic underutilization of genome-resolved strategies which are critical for constraining microbial metabolism and viral impacts into informative units. Further, these datasets are largely unused because the data is collected in an un-coordinated manner, leading to the lack of similar sampling methods, and ultimately an inability to make results interoperable. Together, in this chapter I present compelling evidence for the need of genome-resolved, virus-host paired multi-omic analyses that are pivotal to our understanding of river ecosystems and lay the groundwork for the questions I will address throughout my dissertation.

After identifying that there was a gap studies that leverage metagenome assembled genomes (MAGs) and viral metagenome assembled genomes (vMAGs), for Chapter 2 I focused on using a genome-resolved lens to uncover the microbial and viral metabolic underpinnings responsible for the biogeochemical cycling of carbon and nitrogen in the Columbia River system. This chapter used a dataset that was spatially resolved at the centimeter scale for three sediment cores across two transects of the Columbia River and included 33 samples, all of which had metagenomes that were paired to metaproteomes, biogeochemistry, and metabolites. Using this dataset, I created the first river microbial and viral database genome-resolved database called Hyporheic Uncultured MAG and vMAG (HUM-V).

Leveraging metaproteomics paired to HUM-V database, I built a conceptual model outlining microbial and viral contributions to carbon and nitrogen biogeochemistry in these river sediments. With this metabolic reconstruction, I showed an intertwined carbon and nitrogen cycle that can likely contribute to the fluxes of nitrous oxide. Specifically, I demonstrated that well recognized river microbes like those of the phyla *Nitrososphaeraceae* as well as other less

recognized phyla like *Binatia* encode and express genes for denitrification. I also showed that the clade II *nosZ* gene, which is responsible for nitrous oxide production, could possibly act as a nitrous oxide sink without contributing to its production. Linking viral members to microbial hosts demonstrated that viruses may be key modulators of carbon and nitrogen cycling. Specifically, I presented evidence that viruses can infect key nitrifying organisms (i.e., *Nitrospiraceae*) as well as key polymer degrading organisms (i.e., *Actinobacteria*). Highlighting their potential roles, linear regression analyses consistently identified viral organisms as key predictors of ecosystem biogeochemistry. Chapter 2 of my thesis yielded insights that uncovered some of the microbial contributions that were thought to occur but were poorly defined in river sediments (e.g., nitrogen mineralization), and presented a genome-resolved, virus-host paired strategy that I could then use to directly assess how viruses impacted host metabolism and ecosystem function. Ultimately, Chapter 2 highlights the power of genome-resolved database strategies to reduce existing predictive uncertainties in river corridor models.

Having provided a genome-resolved view of metabolic processes in Chapter 2, for Chapter 3 I set out to expand upon our understanding of river viruses by providing insights into their temporal and spatial dynamics. For this, I worked with a finely tuned temporal dataset from an urban stream near Berlin, Germany called the Erpe River. The Erpe River dataset is a metagenomic timeseries where samples were collected every 3 hours for a total of 48 hours across both the surface water (SW) and pore water (PW) compartments. In addition to metagenomes, Fourier-transform ion cyclotron resonance mass spectrometry (FTICR-MS) and biogeochemistry were collected for each sample.

Using this dataset, I created a database consisting of 1,230 vMAGs and 125 MAGs. Only 1% of our vMAGs clustered to known taxonomic representatives, highlighting the

underrepresentation of river viruses in public databases. Due to this underrepresentation, I supplemented my viral taxonomic analyses with over 20,000 vMAGs spanning different publicly available studies that were relevant to rivers and wastewater treatment plants and showed that nearly half of the novel genera identified were cosmopolitan in aquatic ecosystems. I also characterized the spatial and temporal dynamics of the river microbiomes across the surface water (SW) and pore water (PW) compartments. Both the viral and microbial communities were distinct between the SW and PW samples and were both driven by the same chemical drivers. Given that these compartments had distinct communities, I set out to understand how they were changing over time. By employing multiple temporal statistical methods, I show that SW communities are more persistent and more stable relative to the PW communities, likely resulting from the homogeneous selection pressures of the SW, and the heterogeneity within the sediment. In addition to resolving these temporal dynamics, I highlight some specific virus and host genomes that influence biogeochemical cycling. In summary, my third chapter shows how river viral and microbial communities change across spatial and temporal gradients, and highlights how genome-resolved metagenomics enhances our interpretation of microbiome data.

The final chapter of this dissertation (Chapter 4) summarizes the key findings of my thesis and provides future perspectives to inspire research in environmental river viral ecology. This section also showcases several publications that I have worked on throughout my doctoral degree that span multiple ecosystems like mouse guts, human guts, soils, and the development of the computational tool Distilled and Refined Annotation of Metabolism (DRAM). This final chapter also highlights a manuscript that I was involved in that showcases a new scientific framework: Interoperable, Open, Coordinated, and Networked (ICON). I further highlight this framework to address how these ICON strategies are beginning to be implemented in other fields

and propose that in order to move the discipline of river microbial ecology forward, we need to implement ICON frameworks and the standardization and coordination of sampling collection.

In summary, the aims of this dissertation were to summarize what is known in the field of river viral ecology (Chapter 1), to investigate viral roles that viruses play on river organic nitrogen and carbon processing (Chapter 2), to interrogate the temporal and spatial dynamics of viruses within rivers (Chapter 3), and to summarize how this dissertation has added to the understanding of river viral ecology, and what the next big questions for the field should be (Chapter 4). Ultimately, these works shine a spotlight on the viruses found in river ecosystems and shows that they likely play key roles in the regulation of microbial biogeochemical cycles.

ACKNOWLEDGEMENTS

There is a saying from back home in Puerto Rico: “Dime con quién andas, y te diré quien eres”, which translated means “tell me who you are with, and I will tell you who you are”. Throughout the course of my degree, and my entire life, I have been surrounded by people that have, whether through good or bad, shaped me into the person that I am today. This dissertation spans 6 years of arduous science, during which I have lost loved ones, and have dealt with the consequences of moving away from my home, my family, and my friends, and has been accompanied by frustration, self-doubt, and doubt. The reality is that this dissertation would have not been possible without every single one of you and your support.

Specifically, a huge thank you to my advisor, **Dr. Kelly Wrighton**, for being both a mentor, and a friend. I did not think my life and path would move me from Puerto Rico, to Ohio, to Colorado…but your advice, and your belief in me when I did not believe in myself got me through this degree. Whether it was your advice and your consolation, or your genuine care and understanding, or the tough love, you have made all the difference. I remember when I joined your lab and I said that I did not want to touch a wet lab for the entirety of my PhD, and my audacity in saying that with almost zero computational experience aside from hacking my iPod touch when I was 12. Your guidance throughout the science learning process, and the contacts that you have enabled me are the reason why I can do the science that I do today, and I will never forget that. I will cherish and remember our moments for the rest of my life, particularly our multi-hour paper writing days that involved you arguing with Cedric and all of his smells.

I would also like to say a huge thank you to all of the mentors that I have had through my time at the lab. Your patience, and your kindness has gotten me through the grunt of the bad

times. To **Rebecca Daly**, I cannot understate how important and pivotal you were to my success during this degree. You were the turning point when I thought that I could not do it anymore and wanted to give up. Your support and friendship mean the world to me, and I owe you a debt I cannot repay. To **Dr. Mikayla Borton** and **Dr. Bridget McGivern**, it has been a crazy ride since we all left Ohio together, and I will forever be grateful for the friendship, camaraderie, and support that you have provided me throughout my time at the lab. I hope that your amazing science brains continue to do awesome work in the field, and I look forward to your future accomplishments. To my committee members **Dr. Edward Hall**, **Dr. Jessica Metcalf**, and **Dr. Michael Wilkins**, thank you for teaching me to think outside the box and for pushing me and my science during my time here at Colorado State University. To **Dr. Angela Oliverio**, **Dr. Jordan Angle**, and **Dr. Jared Ellenbogen**, I am truly fortunate that my time in the lab co-occurred with all of you. None of this would have been the same without your guidance, the memes, and your constant support. To all of my lab mates, in particular **Dr. Ikaia Leleiwi**, **Kaela Amundson**, and **Amelia Nelson**, thanks for always having great conversations with me and for being a source of happiness during the last few years. To the people that I have met at Colorado State University, I would not have made it through this without all of your support. **María Chavez**, **Caitlin Charlton**, **Jemma Fadum**, **Daniel Rosales**, **Nathan Phipps**, **Siwook Hwang**, the entire **Hydro Punks crew**, and my cohorts, I am lucky to have met you and am eternally grateful for your friendship and your support. A huge thank you as well to the **Colorado State University Counseling Services**, in particular **Jennifer Brandsma**, for their support and counsel.

The scientist that I am today extends far beyond the people that are in the Wrighton and Wilkins laboratories, and I am grateful for all your support. Specifically, thank you to **Dr. Carlos Ríos-Velazquez** and **Moises de Jesus-Cruz** for taking an undergraduate student that had no idea

what he was doing into your lab and project. It jump-started my curiosity, and who I am today. To all of my friends back home in Puerto Rico, I am who I am today because of you all, and although we may not talk often, know that I still value you as much as I did when we last spoke. To **David Rodríguez**, and **Norman Paulino**, thank you for always being available for a conversation, a meme, a game of Dota2, or all of the above.

Last but not at all least, I would like to thank my family for their constant support and love. It was not an easy transition moving away and having you all be so far, but we made it work and I hope I have made you proud. To my mother, **Betzaida Ramos**, thank you for being there for me whenever I called, and for constantly reminding that my mind is my worst enemy, and that I can do anything I set as my goal. To my father, **Alvin Rodríguez**, thank you for reminding me to enjoy the little things, and for the constant motivation to get outside and experience things that are outside of my comfort zone. To my sister, **Ambar Rodríguez**, I am lucky to be your brother, and am very grateful that I can count on you for anything for the rest of my life. To my grandparents, **Elba Bonilla**, **Benjamin Ramos**, **María Charriez**, and **José Rodríguez-Matos**, your never-ending love and support have gotten me through my darkest moments. To my partner, **Cozette Romero**, thank you for always seeing me with as much worth and value as you do. Through your eyes I have learned self-kindness that I did not think I was capable of. To my uncles, my aunts, and my cousins, who if I listed out individually would need several pages, I am eternally grateful to have a loving, caring, and amazing family like you all, and I hope that I continue to make you proud throughout the rest of my career. Los amo.

To every person that is or was a part of my life, I am eternally grateful; and know that if I had to do it again, I would not change a single thing.

# AUTOBIOGRAPHY

*Education*

| | |
|---|---|
| 2017-2023 | Ph.D., Ecology, Colorado State University, Soil and Crop Sciences |
| 2013-2017 | Bachelor of Science, Industrial Microbiology, University of Puerto Rico, Mayagüez |

*Awards and Honors*

| | |
|---|---|
| 2023 | 2023 Front Range Microbiome Symposium Best Contributed Talk Award |
| 2021 | AGU 2021 Biogeosciences DEI Travel Grant |
| 2020 | MLK Jr. Advancing Education Scholarship |
| 2019 | Fall 2019 AGU Meeting award for Outstanding Student Presentation given to top 5% of presenters |
| 2019 | 2019 Front Range Microbiome Symposium Best Poster Award |
| 2019 | 2019-2020 GAUSSI Fellowship at Colorado State University |
| 2019 | 2019 Pre-Doctoral Ford Fellowship Honorable Mention |
| 2019 | 2019 Environmental System Science (ESS) PI Meeting Student Travel Award |
| 2018 | Best poster at the 17th International Symposium for Microbial Ecology, Leipzig, Germany given to top 1% of posters |
| 2016 | Puerto Rico Louis Stokes Alliance for Minority Participation Travel Grant; NSF research grant funded summer internship at Lawrence Berkeley National Laboratory, CA |
| 2016 | Best oral presentation in Microbiology at the 6th Biology Undergraduate Research Symposium at the University of Puerto Rico, Mayaguez Campus |
| 2015-2016 | NSF funded Puerto Rico Louis Stokes Alliance for Minority Participation Scholarship |

*Leadership and Diversity, Equity, and Inclusion (DEI) Positions*

| | |
|---|---|
| 2023 | Ambassador for the National Microbiome Data Collaborative (NMDC) |
| 2022 | Founding member of the ICON Advisory Board |
| 2021-2022 | Graduate student representative for the College of Agricultural Sciences (CAS) Strategic Planning Committee for Equity and Inclusion |
| 2021 | Founding member of the Graduate Students of Color (GSoC) group of the Graduate Student Committee of Colorado State University |
| 2020 | Graduate Students of Color (GSoC) Leadership and founding member |
| 2020 | Leadership / Organizing Member for the ICON-FAIR Science (Integrated, Coordinated, Open, Networked – Findable, Accessible, Interoperable and Reusable) American Geophysical Union Special Collection |

2019 Soil and Crop Sciences Inclusion, Equity and Diversity Committee

*Outreach Activities*

2022 Summer 2022 IMAGINANTES camp at Colorado State University: The importance of water on our daily lives.

2020 2020 Martin Luther King March IMAGINANTES at Colorado State University

2019 Summer 2019 IMAGINANTES camp at Colorado State University. Astro-sciences themed.

2018 Graduate student judge at the 2018 State Science Day in The Ohio State University

2017 Microbial Bioprospecting and Biotechnology Laboratory Outreach at the Puerto Rico EcoExploratorio event "Feria Planeta Digital"

2016 Former President and Co-founder of AEXO: Astrobiology and Exobiology Student Association at the University of Puerto Rico, Mayaguez

*Publications*

Spatial and finely tuned temporal metagenomics of river compartments reveals viral community dynamics in an urban stream, **Rodríguez-Ramos J.**, Oliverio A., Borton M., Mueller B., Schulz H., Flynn R., Daly R., Danczak R., Ellenbogen J., Schopflin L., Shaffer M., Goldman A., Lewandowski J., Stegen J., Wrighton K., *Frontiers (2023).*

Human gut phages harbor sporulation genes, Schwartz D., **Rodríguez-Ramos J.**, Schaffer M., Flynn R., Daly R., Wrighton K., & Lennon J., *mBio (2023)*

Exposing New Taxonomic Variation with Inflammation – A Murine Model-Specific Genome Database for Gut Microbiome Researchers, Leleiwi I., **Rodríguez-Ramos J.**, Shaffer M., Sabag-Daigle A., Kokkinias K., Flynn R., Daly R., Kop L., Solden L., Ahmer B., Borton M., & Wrighton K., *Microbiome (2023)*

Microbial genome-resolved metaproteomic analyses frame intertwined carbon and nitrogen cycles in river hyporheic sediments, **Rodríguez-Ramos J.,** Borton M., McGivern B., Daly R., Graham E., Roux S., Smith G., Solden L., Purvine S., Nelson W., Lipton M., Stegen J., & Wrighton K., *mSystems (2022).*

Integrated, Coordinated, Open, and Networked (ICON) Science to Advance the Geosciences: Introduction and Synthesis of a Special Collection of Commentary Articles, Goldman, A. E., S. Emani, L. Pérez-Angel, **Rodríguez-Ramos J.**, & Stegen J., *Earth and Space Science Open Archive* (2021)

DRAM: Distilled and Refined Annotation of MAGs, Shaffer M., Borton M., McGivern B., Solden L., Zayed A., Bolduc B., **Rodriguez-Ramos J.**, Liu P., Narrowe A., Daly R., Smith G., Vik D., Sullivan M., Roux S., & Wrighton K., *Nucleic Acid Research, 2020*

Towards optimized viral metagenomes from challenging soils, Trubl G., Roux S., Solonenko N., Li Y., Bolduc B., **Rodríguez-Ramos J.**, Eloe-Fadrosh E., Rich V., & Sullivan M.*, PeerJ 2019*

# DEDICATION

Dedicado a mi familia, por su apoyo y amor incondicional. Los llevo siempre en el corazón.

TABLE OF CONTENTS

*ABSTRACT* ..... *ii*

*ACKNOWLEDGEMENTS* ..... *vii*

*AUTOBIOGRAPHY* ..... *x*

*DEDICATION* ..... *xii*

*Chapter 1: River viral ecology rises: An introduction to viruses in river ecosystems* ..... *1*

1.1 A brief overview of viruses and their impacts on microbial community structure and function ..... 1

1.2 The history of river viral ecology ..... 4

1.3 Metagenomics offers new insights into river viral ecology ..... 5

1.4 A multi-omic, genome-resolved approach to address knowledge gaps in viral ecology ..... 8

*Chapter 1 Figures* ..... *10*

*Chapter 1 References* ..... *12*

*Chapter 2: The fellowship of the river biogeochemical cycling: Deciphering viral and microbial underpinnings of carbon and nitrogen cycling in the Columbia River* ..... *16*

2.1 Summary ..... 16

2.2 Introduction ..... 17

2.3 Results and Discussion ..... 19

2.3.1 HUM-V greatly expands the genomic sampling of HZ microbial members ..... 19

2.3.2 HUM-V recruits metaproteomes offering new insights into HZ microbiomes ..... 21

2.3.3 An inventory of processes contributing to microbial carbon dioxide production and consumption ..... 23

2.3.4 Ammonium exchange can support coordinated nitrogen mineralization and nitrification pathways ..... 27

2.3.5 Metabolic contributions from less characterized taxa actively shape nitrogen cycling ..... 30

2.3.6 Viral influence on sediment carbon and nitrogen cycling ..... 31

2.4 Conclusion ..... 34

2.5 Materials and Methods ..... 36

2.5.1 Sample collection, DNA isolation, and chemical characterization ..... 36

2.5.2 Metagenome assembly and binning ..... 39

2.5.3 Metabolic and taxonomic analyses of MAGs ..... 40

2.5.4 Viral Analyses ..... 41

2.5.5 MAG relative abundance calculations and their use in predictions ..... 42

2.5.6 Metaproteome generation and peptide mapping ..... 43

2.5.7 Data availability: ..... 44

2.5.8 Acknowledgements: ..... 45

*Chapter 2 Figures* ..... *47*

*Chapter 2 References* ..... *61*

*Chapter 3: 2018 a space (and time) odyssey: Spatial and temporal metagenomics of river compartments reveals viral community dynamics in an urban impacted stream* ..... *66*

3.1 Summary ..... 66
3.2 Introduction ..... 67
3.3 Methods ..... 69
3.3.1 Sample collection, DNA isolation, and chemical characterization ..... 69
3.3.2 Metagenome data processing and assembly ..... 71
3.3.3 Viral identification, taxonomy, and annotations ..... 71
3.3.4 Bacterial and archaeal metagenomic binning, quality control, annotation, and taxonomy ..... 73
3.3.5 Virus host linkages ..... 73
3.3.6 Genome relative abundance and normalization ..... 74
3.3.7 Temporal and statistical analyses ..... 75
3.3.8 Data availability ..... 77
3.3.9 Funding ..... 77
3.3.10 Acknowledgements ..... 78
3.4 Results ..... 78
3.4.1 Metagenomics uncovers viral novelty and biogeography of River Erpe viruses ..... 78
3.4.2 Viral and microbial river microbiomes are compartment-specific and coordinated with each other ..... 80
3.4.3 Temporally resolved metagenomics unveils compartment-level stability and persistence of viral and microbial communities ..... 81
3.4.4 Genome-resolved virus-host analyses demonstrated viruses could infect highly abundant, phylogenetically diverse microbial genomes ..... 82
3.4.5 Virally encoded auxiliary metabolic genes can potentially alter host metabolic machinery ..... 83
3.4.6 Co-occurrence networks elucidate ecological patterns that inform ecosystem biogeochemistry ..... 86
3.5 Discussion ..... 88
3.5.1 Viral reference databases underrepresent certain habitats, missing cosmopolitan, ecologically relevant lineages ..... 88
3.5.2 Temporally and spatially resolved metagenomics demonstrates that viral and microbial communities are compartment specific, and more stable in surface water than sediment pore water ..... 89
3.5.3 Viruses have the potential to regulate river biogeochemical cycles by predation and metabolic reprograming of microbial hosts ..... 92
*Chapter 3 Figures ..... 96*
*Chapter 3 References ..... 113*
*Chapter 4: Viruses everywhere all at once: Dissertation conclusions, other works, and perspectives on the future of river viral ecology ..... 120*
4.1 Dissertation summary ..... 120
4.2 Other collaborations and works ..... 121
4.2.1 Towards optimized viral metagenomes from challenging soils ..... 121
4.2.2 DRAM for distilling microbial metabolism to automate the curation of microbiome function ..... 122
4.2.3 Human gut phages harbor sporulation genes ..... 122
4.2.4 Exposing New Taxonomic Variation with Inflammation – A Murine Model-Specific Genome Database for Gut Microbiome Researchers ..... 123
4.2.5 Integrated, Coordinated, Open, and Networked (ICON) Science to Advance the Geosciences: Introduction and Synthesis of a Special Collection of Commentary Articles ..... 124
4.3 Future Research Directions ..... 125
4.3.1 A new era of river viral ecology ..... 125
4.3.2 The future of river viral ecology and ICON-FAIR frameworks ..... 129
4.4 Conclusion ..... 130
*Chapter 4 Figures ..... 132*

*Chapter 4 References* ..... *133*

Chapter 1: River viral ecology rises: An introduction to viruses in river ecosystems

### *1.1 A brief overview of viruses and their impacts on microbial community structure and function*

Although viruses are often recognized for their roles in human health and disease, the vast majority of viruses exclusively infect bacteria and archaea, with $10^{31}$ viral-like particles (VLPs) estimated to exist on Earth [1–3]. In other words, there are more viral particles on Earth than there are stars in the universe [4]! Not surprisingly, through their replication processes these viruses have profound implications for ecosystem biogeochemical cycles by altering microbial community metabolism. For this section, I review the impacts of viruses on microbial communities by 3 key processes: 1) predation 2) resource control and 3) auxiliary metabolic genes (AMGs).

Through predation, viruses are key ecological controls on microbial community dynamics (**Figure 1.1a**). Viral replication has two main cycles: the lytic cycle, and the lysogenic cycle. During favorable environmental conditions for the host, a virus replicates via the lytic cycle, where the virus attaches to the host cell, inserts its genome, and then commandeers host replication machinery to produce copies of itself. This results in the utilization of host resources to generate viral genomes and structures, and ultimately leads to the release of multiple viral copies into the environment upon cell lysis. When there are unfavorable environmental conditions, a virus enters the lysogenic cycle, where instead of taking over cellular machinery, the virus integrates its genome into the host genome. During lysogeny, a virus replicates alongside its host whenever the host replicates, and then upon return of favorable environmental conditions, becomes induced and shifts to the lytic cycle. Through these canonical replication cycles, viruses enforce key top-down ecological controls on microbial communities by eliminating bacterial and archaeal organisms. These predation dynamics are highly relevant in natural ecosystems. In fact, in oceans microbial

viruses are estimated to kill between 20-40% of all bacteria daily [5–7]. Further, these predation dynamics can "short-circuit" ecosystem biogeochemical cycles (i.e., carbon and nitrogen cycles) by removal of key microbial metabolic contributors and the restructuring of microbial communities [8,9]. Nonetheless, despite host death and microbial community "pruning" being a key role of viruses in natural systems, it is not the only way viruses impact the environment.

In addition to their role as top-down ecological controls, viruses exert key bottom-up resource controls (**Figure 1.1b**). When a virus lyses a host cell, cellular contents are spilled into the surrounding environment, and these cellular contents are rich in dissolved and particulate organic matter. In ocean systems, viral lysates are hypothesized to stimulate the recycling of bacterial community carbon, ultimately reducing net bacterial production [10], a process termed the viral shunt[11,12]. Due to the viral shunt, viruses alter the direction of carbon transport across ecological trophic levels. In other words, when microbial organic matter gets recycled within microbial communities, it prevents carbon from moving up into higher trophic levels (i.e., plankton, fish). Similarly, viral predation can shift ecological niches for microbial communities through "Kill the Winner" dynamics, where viruses are more likely to infect highly abundant competition specialists (r-strategy organisms) as opposed to defense specialists (k-strategy organisms), making resources available for microorganisms that would otherwise not be able to compete[13]. Conversely, viral lysis in oceans can also lead to the "clumping" of organic matter and microorganisms resulting in colloidal particles or aggregates, which subsequently sink into the deeper oceans thereby sequestering carbon, a process referred to as the viral shuttle[5,14]. Together, the balance between the viral shuttle, shunt, and higher-trophic level predation of microorganisms by eukaryotes can alter the C dynamics of natural systems.

Lastly, viruses are also known to influence microbial community composition by encoding auxiliary metabolic genes (AMGs), or virally encoded genes that upon infection, upregulate or enhance the hosts metabolic capability (**Figure 1.1c**). Upon viral infection, AMGs are transcribed and subsequently translated by the host cell. In one of the best documented AMG examples to date, oceanic cyanobacteria enhanced photosynthesis upon viral infection due to virally encoded photosynthesis system genes[15]. Similarly, cyanophages carry Calvin cycle inhibitor genes that direct carbon flux from the Calvin cycle into the pentose phosphate pathway, leading to the augmented production of NADPH and subsequently, higher levels of nucleotide biosynthesis[16]. Additionally, another study linked viral AMGs to the unlocking of complex carbon degradation potential within soil permafrost communities [17]. This metabolic reprogramming of host genomes, however, is not limited to carbon cycling and nucleic acid production. A study in oceans detected ammonia monooxygenase (*amoC*) genes that could upregulate nitrification [18], as well as denitrification genes like nitrate reductases (*narGHI*) and nitrite oxidoreductases (*nxrAB*) [19]. Additionally, a wide array of widespread, sulfur metabolism AMGs were identified in viruses of multiple environments which could promote anaerobic sulfur respiration (i.e., *dsrA*, *dsrC*) as well as the oxidation of thiosulfate (i.e., *soxC*, *soxD*)[20]. Together, these studies demonstrated that AMGs likely contribute to the overall productivity of microbial communities and introduce a new level of complexity and implications for viruses in natural systems.

In conclusion, given the myriad of ways by which viruses can impact microbial communities, it is not surprising that the field of viral ecology has boomed within the last decade. Given the evidence above, if bacteria and archaea are known as the biogeochemical engines of the earth[21], then viruses should be considered as the oil that keeps that engine running. Nonetheless,

while viral ecology is appreciated in marine [4,7], soil [22,23] and even lake [24,25] ecosystems, the role of viruses in rivers are not well understood. This knowledge gap is addressed by my thesis work.

### *1.2 The history of river viral ecology*

Rivers cover around 773,000km of terrestrial surface across the world, which accounts for less than 1% of Earth's non-glaciated area [26]. Despite their relatively small surface area, rivers act as an important nexus between terrestrial and aquatic environments, transporting or storing nearly 2 petagrams of terrestrial organic carbon per year [27]. Additionally, rivers are estimated to significantly contribute to global greenhouse gas emissions, releasing the equivalent of around 7% of global carbon dioxide emissions (~2350 teragram $CO_2$ $yr^{-1}$), 5% of global methane emissions (30.5 teragram $CH_4$ $yr^{-1}$), and up to 30% of global anthropogenic nitrous oxide emissions (32.3 gigagram $N_2O$ $yr^{-1}$) per year [28–31]. Within these river ecosystems, microorganisms like bacteria and archaea are responsible for a large portion of ecosystem respiration and biogeochemical cycling, catalyzing biochemical conversions along the length of a river [32,33]. Despite this, our knowledge of river microbial ecology is severely lacking, and this is compounded when it comes to viral communities due to the lack of studies that holistically address viral and microbial communities.

Early studies of viruses from rivers provided the first insights into the enumeration of VLPs [34,35], with lower and upper end abundances of 0.07 x $10^7$ VLPs $ml^{-1}$ [36], and 88.8 x $10^7$ VLPs $ml^{-1}$ [37] observed, respectively. Viral abundances within these limits were identified across seasonal, hydrological, and geographic gradients, suggesting environmental factors influence viral community abundances along river transects [38–47]. Further adding complexity, heterogeneity at a fine scale is often observed in river sediments where groundwater (GW) intrusion and surface water (SW) mixing occur in the hyporheic zone (HZ) resulting in centimeter-scale "hotspots" that account for up to 90% of ecosystem respiration [32,48,49]. Reflecting this, a study in 2016 showed

viruses in sediments change at this centimeter scale, and are most abundant in the sediment-surface interface[50].

Viruses exist in a constant evolutionary arms race with their microbial hosts. Bolstering support for constant virus-host interactions in rivers, studies observed that up to 80% of bacterial isolate strains had associated virulent phage that could be isolated with them [51]. Separately, Bettarel et al. showed an average proportion of 7.1% of lysogenic (i.e., integrated within a host genome) to lytic (i.e., separated and free living) viral particles [52]. Similarly, reports estimating viral productivity (i.e. the number of viruses produced per hour) in rivers showed a wide range of measurements from very low values of $4.10 \times 10^8$ $l^{-1}$ $h^{-1}$ virions in sediments [52,53] up to $170.2 \times 10^9$ virions $l^{-1}$ $h^{-1}$ in a eutrophic river [45]. The impact of this predation on river respiration was showcased by Pollard and Ducklow, who demonstrated that viral predation influenced DOC concentrations, expediting the total DOC that becomes respired to $CO_2$, and ultimately causing higher-trophic level loss of organic carbon [45]. Together, these early studies showcased that viruses are likely important for river microbial community structure and function. Nonetheless, the lack of multi-omic datasets hindered the field of river viral ecology from providing a genome-resolved, mechanistic view of the interactions and metabolisms that lead to these cycles, and from understanding how these early findings scaled across (and within) river ecosystems.

### *1.3 Metagenomics offers new insights into river viral ecology*

Recent developments in multi-omics gave rise to a plethora of bioinformatic tools and spurred the "coming of age" of viral ecology [24,54,55]. Metaviromics, or the study of viral communities using metagenomics, enabled researchers to assess the "uncultured majority" of viral communities using a suite of specialized software and methods [56–58]. Genome-resolved strategies now offer new illumination of the presence and distribution of viral genomes, virus-host predation

dynamics, and the impacts viruses have on microbial community assembly and function. Despite the power of these technologies, it should be noted that river genome-resolved metagenomic studies are limited, and remain far more understudied compared to oceans and soils [59].

Since the last river viral ecology review published in 2016 [35], there has been an explosion of river metagenomic sequencing efforts. According to the Joint Genome Institute Genomes Online Database (JGI GOLD), prior to 2016 the repository included 154 publicly available river metagenomes (**Figure 1.1ab**). Since then, that number of metagenomes exceeds 3,000 samples, spanning 30 different countries, and 123 unique studies (**Figure 1.1a-c, Appendix A**). Despite this, there are less than 20 publications resulting from these data, with most focused on public health by the identification of viruses that are potentially harmful to human health, and contribute to the dispersal of antibiotic resistance genes [60–67].

Despite the vast data collection shown in **Figure 1**, prior to my thesis research, only a handful of publications used metagenomics to survey viral diversity in river systems. For the earliest of these studies, read-based approaches were implemented in order to classify the total number of reads that belonged to viruses. In 2008, Leroy et al., used microscopy paired with metagenomes to assess bacteriophage morphotype and diversity in a river sediment, providing preliminary results showing viruses like *Myoviridae*, *Siphoviridae*, and *Podoviridae* are abundant in rivers [68]. In 2014, Satinsky et. al. reported 10 viral organisms that were identified from their multi-station sampling campaign of the Amazon river [69]. Hinting at the importance of viral dispersal in rivers, two of those viruses were identified in two distant stations. Two years later in 2016, Uyaguari-Diaz published a study that optimized extraction methods for metagenomic classification for viruses, and reported a large number of viruses present across different watersheds belonging mostly to *Microviridae*, *Siphoviridae*, and *Podoviridae* [70]. They also

annotated their bulk metagenomes with MG-RAST and reported high abundances of viral replication related genes. A recent publication from 2021 by Aishwarya et. al., also annotated bulk metagenomes and metatranscriptomes to identify the main functions present within a river, and show viruses were both present and abundant [71].

In addition to these studies, a smaller subset of research attempted to go further by using assembled metagenomes, and subsequently viral genomes, to describe what environmental impacts are structuring viral ecological patterns. In 2016. Dann et. al. published their work identifying viruses in a riverine system in Australia by using metagenomics and tBlastx[72]. They showed that human activity possibly impacted viral diversity downstream of a town and showed that river viruses shared similarity to marine viruses. Although they did not have temporal resolution, the authors state that there are likely some persistence dynamics at play given viral taxonomic conservation across upstream and downstream sites. They also highlight that carbohydrate, amino acid, and protein metabolism are potentially important based off of the total number of genes by MG-RAST classifications. Another study by Silva et al. in 2017 showed that viruses within the Amazon river were shown to be ubiquitous and span changes in ecosystem transition zones [73]. These findings were supported by a newer study by Lu et. al. in 2022 who showed viral community similarity across large geographical distances in river surface water [74]. Contrastingly, Reddington et. al. noted distinct viral community differences between rivers in close proximity in 2020 [75].

In addition to helping scientists address viral community structuring, genome-resolved metagenomics has offered a new opportunity to identify AMGs from natural ecosystems. To this end, there have been at least 4 publications that showed the prevalence and function of AMGs in rivers without directly alluding to antibiotic resistance genes or their roles in human health. Studies

within the last 5 years showed viruses encode AMGs related to photosynthesis as well as carbon and nitrogen cycling [73,76,77], and a 2022 study showed that AMGs can be structured by habitat, viral lifestyle, and host [78]. These genes are likely responsible for the upregulation of different metabolisms like glycolysis and organic nitrogen degradation, or for "unlocking" the ability of a microbial host to degrade different carbon types. Unfortunately, these AMG studies do not include paired, genome-resolved microbial analyses, thus these putative AMGs have yet to be confirmed like in marine systems [15], and the full potential of river microbiome analyses are yet to be unlocked.

Summarizing, while initial viral metagenomics research present evidence that there are ecological patterns at play for viral communities, studies linking the bridge between what is observed and why it is observed (i.e., the environmental influences on viral and microbial community metabolism) are scarce. Contradictory results by early studies on foundational ideas like "*Do viruses differ along a river length*?" highlight the need for additional, genome-resolved multi-omic studies that integrate both viral and microbial communities. Further, no study has utilized multi-omic strategies (e.g., metatranscriptomics, metaproteomics and metabolomics) paired to genome-resolved units of both viruses and their hosts at an ecosystem level in an effort to understand the interactions that can contribute to river microbiome metabolisms. Ultimately, this means that the ability to derive ecological relevance and constrain ecological patterns of these river microbiomes is limited.

### *1.4 A multi-omic, genome-resolved approach to address knowledge gaps in viral ecology*

To holistically evaluate the role of viruses, bacteria, and archaea in river ecosystems, I integrated genome-resolved strategies to spatially and temporally resolved multi-omic datasets. This thesis work is some of the first comprehensive viral and microbial multi-omics analyses applied to river systems. As such, given the nascent state of the field, many of the questions

addressed were fundamental in terms of the structure and activity of these viral members. Specifically, the four objectives of my dissertation research were to:

(1) Determine whether viruses were metabolically active in river sediments and how they could impact ecosystem biogeochemistry (Chapter 2).

(2) Identify the microbial members that are impacted by viruses, along with mechanistically understanding their metabolic roles in rivers (Chapter 2, Chapter 3).

(3) Assess the biogeography of river viruses in order to understand the relevance and scalability of our multi-omic analyses (Chapter 2, Chapter 3).

(4) Determine what factors, if any, structure viral membership and diversity across river compartments (Chapter 3).

Together, this dissertation shows that viruses are an important, missing piece of river ecosystem function, and shines the spotlight on viruses as orchestrators of river biogeochemical cycles.

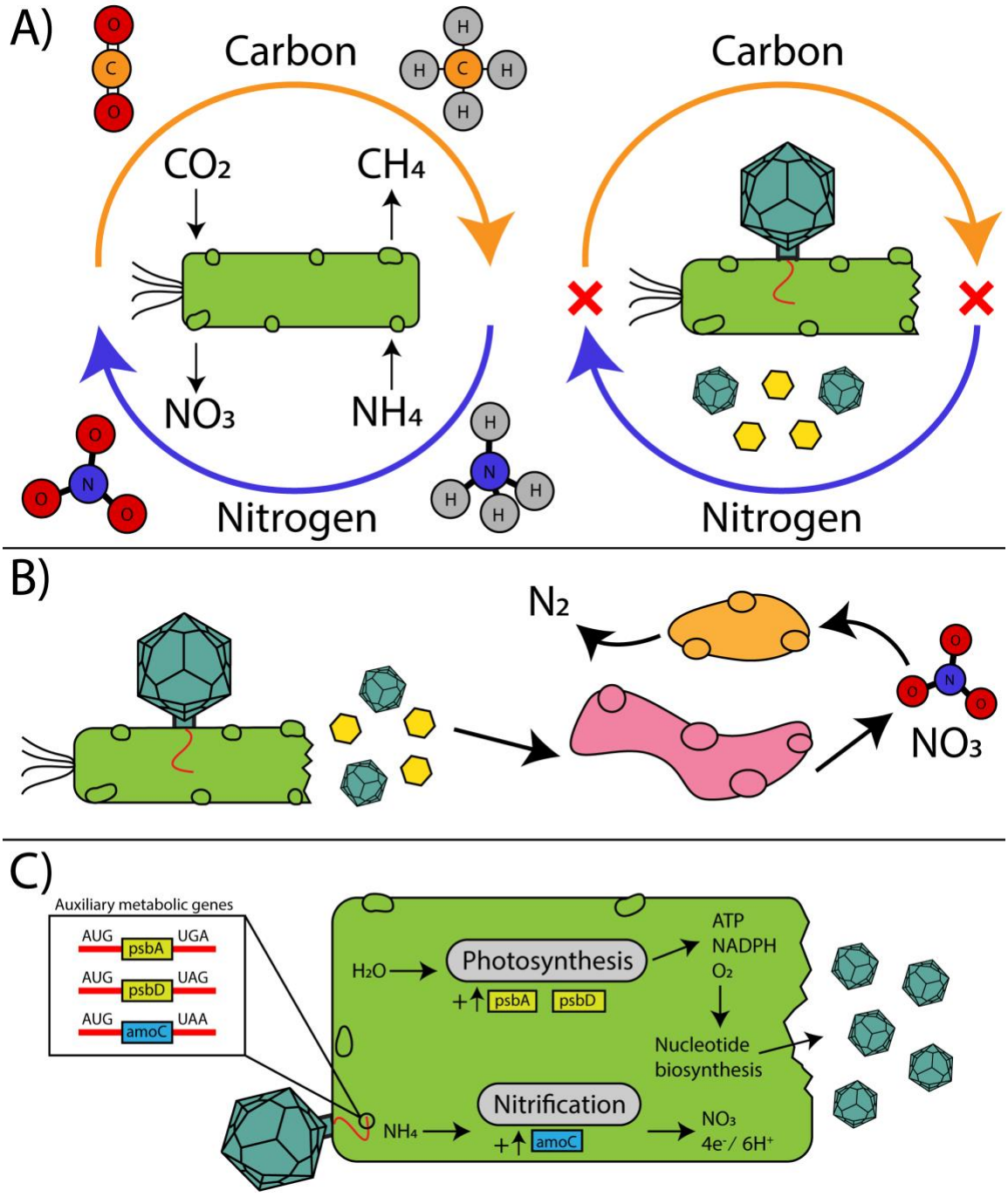



**Figure 1.1** Summary of three major ways by which viruses can control microbial populations and metabolism. A) Viral predation can impact the microbial biogeochemical elemental cycles like carbon and nitrogen cycling. B) Viral host lysis releases organic matter like carbon that can be reutilized within the microbial community. C) Upon infection, a virus encoding for auxiliary metabolic genes can reprogram host metabolisms like photosynthesis and nitrification which can lead to the upregulation of metabolism, replication, and more viral progeny upon lysis.

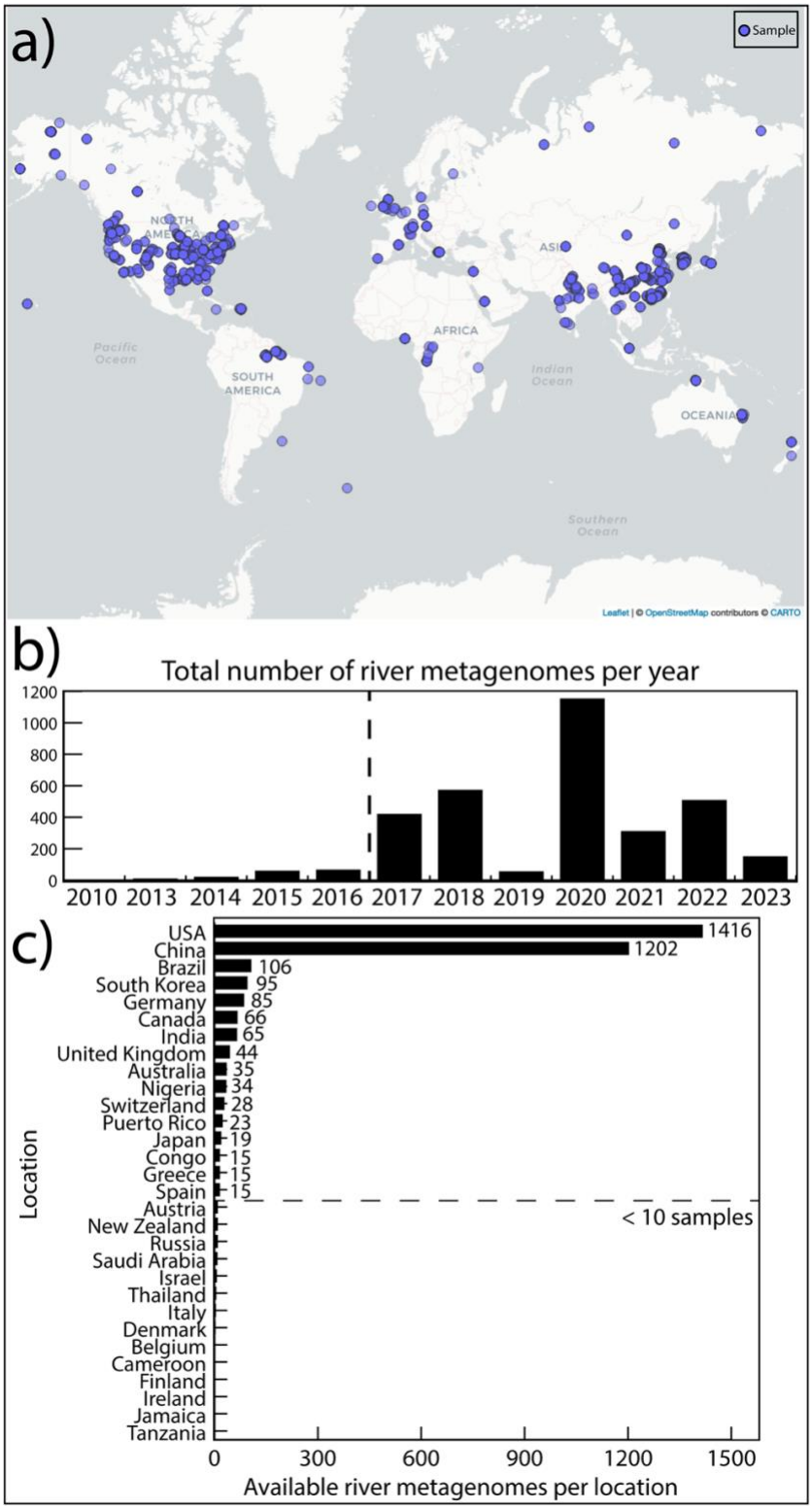



**Figure 1.2** Summary of total river metagenomes available. a) Global map showing all river metagenome samples that are publicly available within the JGI GOLD Repository. b) Bar chart showing the total river metagenomes that were sequenced for each year. c) The geographic location breakdown of metagenomic samples available in JGI GOLD as of April 2023.

Chapter 1 References


1. Mushegian, A. R. Are There 1031 Virus Particles on Earth, or More, or Fewer? *J. Bacteriol.* **202**, (2020).
2. Hendrix, R. W., Smith, M. C. M., Neil Burns, R., Ford, M. E. & Hatfull, G. F. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 2192–2197 (1999).
3. Bar-On, Y. M., Phillips, R. & Milo, R. The biomass distribution on Earth. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 6506–6511 (2018).
4. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the marine microbial realm. *Nat Microbiol* **3**, 754–766 (2018).
5. Weinbauer, M. G. Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* **28**, 127–181 (2004).
6. Weinbauer, M. G. & Rassoulzadegan, F. Are viruses driving microbial diversification and diversity? *Environ. Microbiol.* **6**, 1–11 (2004).
7. Suttle, C. A. Marine viruses--major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812 (2007).
8. Albright, M. B. N. *et al.* Experimental evidence for the impact of soil viruses on carbon cycling during surface plant litter decomposition. *ISME Communications* **2**, 1–8 (2022).
9. Braga, L. P. P. *et al.* Impact of phages on soil bacterial communities and nitrogen availability under different assembly scenarios. *Microbiome* **8**, 52 (2020).
10. Middelboe, M. & Lyck, P. G. Regeneration of dissolved organic matter by viral lysis in marine microbial communities. *Aquat. Microb. Ecol.* **27**, 187–194 (2002).
11. Wilhelm, S. W. & Suttle, C. A. Viruses and Nutrient Cycles in the SeaViruses play critical roles in the structure and function of aquatic food webs. *Bioscience* **49**, 781–788 (1999).
12. Weitz, J. S. & Wilhelm, S. W. Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 Biol. Rep.* **4**, 17 (2012).
13. Breitbart, M. Marine viruses: truth or dare. *Ann. Rev. Mar. Sci.* **4**, 425–448 (2012).
14. Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470 (2016).
15. Sullivan, M. B. *et al.* Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol.* **4**, e234 (2006).
16. Thompson, L. R. *et al.* Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proc. Natl. Acad. Sci. U. S. A.* **108**, E757-64 (2011).
17. Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nat Microbiol* **3**, 870–880 (2018).
18. Ahlgren, N. A., Fuchsman, C. A., Rocap, G. & Fuhrman, J. A. Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. *ISME J.* **13**, 618–631 (2019).
19. Gazitúa, M. C. *et al.* Potential virus-mediated nitrogen cycling in oxygen-depleted oceanic waters. *ISME J.* (2020) doi:10.1038/s41396-020-00825-6.
20. Kieft, K. *et al.* Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. *Nat. Commun.* **12**, 3503 (2021).
21. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034–1039 (2008).

22. Jansson, J. K. & Wu, R. Soil viral diversity, ecology and climate change. *Nat. Rev. Microbiol.* (2022) doi:10.1038/s41579-022-00811-z.
23. Trubl, G., Hyman, P., Roux, S. & Abedon, S. T. Coming-of-Age Characterization of Soil Viruses: A User's Guide to Virus Isolation, Detection within Metagenomes, and Viromics. *Soil Systems* **4**, 23 (2020).
24. Roux, S. *et al.* Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nat. Commun.* **8**, 858 (2017).
25. Berg, M. *et al.* Host population diversity as a driver of viral infection cycle in wild populations of green sulfur bacteria with long standing virus-host interactions. *ISME J.* **15**, 1569–1584 (2021).
26. Allen, G. H. & Pavelsky, T. M. Global extent of rivers and streams. *Science* **361**, 585–588 (2018).
27. Liu, S. *et al.* The importance of hydrology in routing terrestrial carbon to the atmosphere via global streams and rivers. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2106322119 (2022).
28. Friedlingstein, P. *et al.* Global carbon budget 2021. *Earth Syst. Sci. Data* **14**, 1917–2005 (2022).
29. Gómez-Gener, L. *et al.* Global carbon dioxide efflux from rivers enhanced by high nocturnal emissions. *Nat. Geosci.* **14**, 289–294 (2021).
30. Rosentreter, J. A. *et al.* Half of global methane emissions come from highly variable aquatic ecosystem sources. *Nat. Geosci.* **14**, 225–230 (2021).
31. Hu, M., Chen, D. & Dahlgren, R. A. Modeling nitrous oxide emission from rivers: a global assessment. *Glob. Chang. Biol.* **22**, 3566–3582 (2016).
32. Lewandowski, J. *et al.* Is the Hyporheic Zone Relevant beyond the Scientific Community? *Water* **11**, 2230 (2019).
33. Stegen, J. C. *et al.* Groundwater-surface water mixing shifts ecological assembly processes and stimulates organic carbon turnover. *Nat. Commun.* **7**, 11237 (2016).
34. Steenhauer, L. M. Freshwater viruses: from ecosystem dynamics to the cyanobacterial cell. *Griffith University, Australia* (2013).
35. Peduzzi, P. Virus ecology of fluvial systems: a blank spot on the map? *Biol. Rev. Camb. Philos. Soc.* **91**, 937–949 (2016).
36. Olapade, O. A., Brothers, A., Crissman, M., Gao, X. & Leff, L. G. Comparison of planktonic microbial communities among nine North American streams. *Archiv für Hydrobiologie* 221–239 (2006).
37. Ma, L., Sun, R., Mao, G., Yu, H. & Wang, Y. Seasonal and spatial variability of virioplanktonic abundance in Haihe River, China. *Biomed Res. Int.* **2013**, 526362 (2013).
38. Mathias, C. B., Kirschner, A. & Velimirov, B. Seasonal variations of virus abundance and viral control of the bacterial production in a backwater system of the danube river. *Appl. Environ. Microbiol.* **61**, 3734–3740 (1995).
39. Almeida, M. A., Cunha, M. A. & Alcântara, F. Loss of Estuarine Bacteria by Viral Infection and Predation in Microcosm Conditions. *Microb. Ecol.* **42**, 562–571 (2001).
40. Baker, P. W. & Leff, L. G. Seasonal patterns of abundance of viruses and bacteria in a Northeast Ohio (USA) stream. *Archiv für Hydrobiologie* **161**, 225–233 (2004).
41. Auguet, J. C., Montanié, H., Delmas, D., Hartmann, H. J. & Huet, V. Dynamic of virioplankton abundance and its environmental control in the Charente estuary (France). *Microb. Ecol.* **50**, 337–349 (2005).

42. Slováčková, H. & Maršálek, B. Virioplankton and microbial communities in two Czech rivers (Svratka and Morava River). *Aquat. Sci.* **70**, 282–291 (2008).
43. Peduzzi, P. & Luef, B. Viruses, bacteria and suspended particles in a backwater and main channel site of the Danube (Austria). *Aquat. Sci.* **70**, 186–194 (2008).
44. Besemer, K. *et al.* Sources and composition of organic matter for bacterial growth in a large European river floodplain system (Danube, Austria). *Org. Geochem.* **40**, 321–331 (2009).
45. Pollard, P. C. & Ducklow, H. Ultrahigh bacterial production in a eutrophic subtropical Australian river: Does viral lysis short-circuit the microbial loop? *Limnol. Oceanogr.* **56**, 1115–1129 (2011).
46. Peduzzi, P., Agis, M. & Luef, B. Evaluation of confocal laser scanning microscopy for enumeration of virus-like particles in aquatic systems. *Environ. Monit. Assess.* **185**, 5411–5418 (2013).
47. Williamson, K. E., Harris, J. V., Green, J. C., Rahman, F. & Chambers, R. M. Stormwater runoff drives viral community composition changes in inland freshwaters. *Front. Microbiol.* **5**, 105 (2014).
48. Stegen, J. C. *et al.* Influences of organic carbon speciation on hyporheic corridor biogeochemistry and microbial ecology. *Nat. Commun.* **9**, 585 (2018).
49. Naegeli, M. W. & Uehlinger, U. Contribution of the Hyporheic Zone to Ecosystem Metabolism in a Prealpine Gravel-Bed-River. *J. North Am. Benthol. Soc.* **16**, 794–804 (1997).
50. Dann, L. M., Paterson, J. S., Newton, K., Oliver, R. & Mitchell, J. G. Distributions of Virus-Like Particles and Prokaryotes within Microenvironments. *PLoS One* **11**, e0146984 (2016).
51. Lammers, W. T. Stimulation of bacterial cytokinesis by bacteriophage predation. in *Sediment/Water Interactions* 261–265 (Springer Netherlands, 1992).
52. Bettarel, Y., Bouvy, M., Dumont, C. & Sime-Ngando, T. Virus-bacterium interactions in water and sediment of West African inland aquatic systems. *Appl. Environ. Microbiol.* **72**, 5274–5282 (2006).
53. Mei, M. L. & Danovaro, R. Virus production and life strategies in aquatic sediments. *Limnol. Oceanogr.* **49**, 459–470 (2004).
54. Sullivan, M. B., Weitz, J. S. & Wilhelm, S. Viral ecology comes of age. *Environ. Microbiol. Rep.* **9**, 33–35 (2017).
55. Emerson, J. B. Soil Viruses: A New Hope. *mSystems* **4**, (2019).
56. Bekliz, M., Brandani, J., Bourquin, M., Battin, T. J. & Peter, H. Benchmarking protocols for the metagenomic analysis of stream biofilm viromes. *PeerJ* **7**, e8187 (2019).
57. Benler, S. & Koonin, E. V. Fishing for phages in metagenomes: what do we catch, what do we miss? *Curr. Opin. Virol.* **49**, 142–150 (2021).
58. Moon, K. & Cho, J.-C. Metaviromics coupled with phage-host identification to open the viral 'black box.' *J. Microbiol.* **59**, 311–323 (2021).
59. Chu, H., Gao, G.-F., Ma, Y., Fan, K. & Delgado-Baquerizo, M. Soil Microbial Biogeography in a Changing World: Recent Advances and Future Perspectives. *mSystems* **5**, (2020).
60. Colombo, S. *et al.* Viromes As Genetic Reservoir for the Microbial Communities in Aquatic Environments: A Focus on Antimicrobial-Resistance Genes. *Front. Microbiol.* **8**, 1095 (2017).
61. Moon, K. *et al.* Freshwater viral metagenome reveals novel and functional phage-borne antibiotic resistance genes. *Microbiome* **8**, 75 (2020).

62. Chopyk, J. *et al.* Metagenomic analysis of bacterial and viral assemblages from a freshwater creek and irrigated field reveals temporal and spatial dynamics. *Sci. Total Environ.* **706**, 135395 (2020).
63. Kumar, N. *et al.* Abundance and diversity of phages, microbial taxa and antibiotic resistance genes in the sediments of the river Ganges through metagenomic approach. 2020.04.29.067819 (2020) doi:10.1101/2020.04.29.067819.
64. Guerrero-Latorre, L. *et al.* Quito's virome: Metagenomic analysis of viral diversity in urban streams of Ecuador's capital city. *Sci. Total Environ.* **645**, 1334–1343 (2018).
65. Behera, B. K. *et al.* Bacteriophages diversity in India's major river Ganga: a repository to regulate pathogenic bacteria in the aquatic environment. *Environ. Sci. Pollut. Res.* **30**, 34101–34114 (2023).
66. Allsing, N., Kelley, S. T., Fox, A. N. & Sant, K. E. Metagenomic Analysis of Microbial Contamination in the U.S. Portion of the Tijuana River Watershed. *Int. J. Environ. Res. Public Health* **20**, (2022).
67. Bibby, K. *et al.* Metagenomics and the development of viral water quality tools. *npj Clean Water* **2**, 1–13 (2019).
68. Leroy, M., Prigent, M., Dutertre, M., Confalonieri, F. & Dubow, M. Bacteriophage morphotype and genome diversity in Seine River sediment. *Freshw. Biol.* **53**, 1176–1185 (2008).
69. Satinsky, B. M. *et al.* The Amazon continuum dataset: quantitative metagenomic and metatranscriptomic inventories of the Amazon River plume, June 2010. *Microbiome* **2**, 17 (2014).
70. 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).
71. Aishwarya, S. *et al.* Structural, functional, resistome and pathogenicity profiling of the Cooum river. *Microb. Pathog.* **158**, 105048 (2021).
72. Dann, L. M. *et al.* Marine and giant viruses as indicators of a marine microbial community in a riverine system. *Microbiologyopen* **5**, 1071–1084 (2016).
73. Silva, B. S. de O. *et al.* Virioplankton Assemblage Structure in the Lower River and Ocean Continuum of the Amazon. *mSphere* **2**, (2017).
74. Lu, J. *et al.* Metagenomic analysis of viral community in the Yangtze River expands known eukaryotic and prokaryotic virus diversity in freshwater. *Virol. Sin.* **37**, 60–69 (2022).
75. Reddington, K. *et al.* Metagenomic analysis of planktonic riverine microbial consortia using nanopore sequencing reveals insight into river microbe taxonomy and function. *Gigascience* **9**, (2020).
76. Ruiz-Perez, C. A., Tsementzi, D., Hatt, J. K., Sullivan, M. B. & Konstantinidis, K. T. Prevalence of viral photosynthesis genes along a freshwater to saltwater transect in Southeast USA. *Environ. Microbiol. Rep.* **11**, 672–689 (2019).
77. Rajput, V. *et al.* Metagenomic mining of Indian river confluence reveal functional microbial community with lignocelluloytic potential. *3 Biotech* **12**, 132 (2022).
78. Luo, X.-Q. *et al.* Viral community-wide auxiliary metabolic genes differ by lifestyles, habitats, and hosts. *Microbiome* **10**, 190 (2022).

# Chapter 2: The fellowship of the river biogeochemical cycling: Deciphering viral and microbial underpinnings of carbon and nitrogen cycling in the Columbia River[1]

## *2.1 Summary*

Rivers have a significant role in global carbon and nitrogen cycles, serving as a nexus for nutrient transport between terrestrial and marine ecosystems. Although rivers have a small global surface area, they contribute substantially to global greenhouse gas emissions through microbially mediated processes within the river hyporheic zone. Despite this importance, research linking microbial and viral communities to specific biogeochemical reactions is still nascent in these sediment environments. To survey the metabolic potential and gene expression underpinning carbon and nitrogen biogeochemical cycling in river sediments, we collected an integrated dataset of 33 metagenomes, metaproteomes, and paired metabolomes. We reconstructed over 500 microbial metagenome assembled genomes (MAGs), which we dereplicated into 55 unique, near-complete medium and high-quality MAGs spanning 12 bacterial and archaeal phyla. We also reconstructed 2,482 viral genomic contigs, which were dereplicated into 111 viral MAGs (vMAGs) >10kb in size. As a result of integrating gene expression data with geochemical and metabolite data, we created a conceptual model that uncovered new roles for microorganisms in organic matter decomposition, carbon sequestration, nitrogen mineralization, nitrification, and denitrification. Integrated through shared resource pools of ammonium, carbon dioxide, and inorganic nitrogen we show how these metabolic

[1] This chapter was reproduced verbatim from "Rodríguez-Ramos et al. Genome-Resolved Metaproteomics Decodes the Microbial and Viral Contributions to Coupled Carbon and Nitrogen Cycling in River Sediments. mSystems (2022)". The text benefitted from writing and editing from contributing authors and reviewers selected by the publisher. The ordering of the materials in this dissertation are consistent with the content available online but have been renumbered to reflect incorporation into this dissertation.

pathways could ultimately contribute to carbon dioxide and nitrous oxide fluxes from hyporheic sediments. Further, by linking viral MAGs to these active microbial hosts, we provide some of the first insights into viral modulation of river sediment carbon and nitrogen cycling.

### *2.2 Introduction*

The hyporheic zone (HZ) is a transitional space between river compartments, where the mixing of nutrients and organic carbon from river and groundwater stimulate microbial activity [1–3]. Characterized as the permanently saturated interface between the river surface channel and underlying sediments, the HZ is considered a biogeochemical hotspot for microbial biogeochemistry [1–3], ultimately contributing to the majority of river greenhouse gas (GHG) fluxes. For instance, it is estimated rivers contribute up to 85% of inland water carbon dioxide and 30% of nitrous oxide emissions [4–6]. Microorganisms in the HZ also catalyze the transformation of pollutants and natural solutes, all while microbial biomass itself supports benthic food webs [7]. Together, these findings highlight that microbial metabolism in HZ sediments has substantial influence on overall river biogeochemistry and health.

Despite the importance of HZ microorganisms, research linking microbial identity to specific biogeochemical reactions in the carbon and nitrogen cycles is still nascent in these sediments. In conjunction with geochemistry, microbial functional genes or gene products (e.g., *nirS* and *nrfA*) have been quantified to denote microbial contributions to specific biogeochemical pathways (e.g., nitrate reduction) [8]. However, these studies often do not identify the microorganisms catalyzing the process and only focus on a few enzymatic reactions. Thus, a comprehensive assessment of the interconnected microbial metabolisms that fuel carbon and nitrogen cycling in river sediments is underexplored.

More recently, 16S rRNA amplicon sequencing has shed new light on the identity of bacteria and archaeal members in river sediments. These studies revealed that cosmopolitan and dominant members in river sediments belong to six main phyla: *Acidobacteria, Actinobacteriota, Firmicutes, Nitrospirota, Proteobacteria*, and *Thaumarchaeota* [9,10]. Furthermore, in some instances cultivation paired to amplicon sequencing has assigned some of these microorganisms (e.g., *Proteobacteria*) to specific biogeochemical process (e.g., denitrification) [11]. Yet, most functional inferences from taxonomic data alone are unreliable due to the dissociation between microbial taxonomy and metabolic function [12,13]. Thus, many key biogeochemical pathways in rivers (e.g., plant biomass deconstruction, denitrification, nitrogen mineralization) are not holistically interrogated alongside microbial communities [14]. Furthermore, amplicon sequencing fails to sample viral communities. While it is likely viruses are key drivers of HZ microbial mortality and biogeochemical cycling by dynamics of predation and auxiliary metabolic genes, the evidence is even more sparse than for its bacterial and archaeal counterparts [15–18].

Cultivation-independent, community-wide, and genome resolved approaches are key to addressing the knowledge gap of how microbial and viral communities influence river biogeochemical cycling. However, metagenomic studies in river sediments are limited, and have focused primarily on gene content as opposed to reconstructing genomes [19,20]. To our knowledge, only two river sediment studies have generated microbial genomes to link taxonomy to functional processes, and these have focused on the impacts of nitrate oxidizing and comammox microorganisms to nitrification [21,22]. As such, despite these recent advances, the chemical exchange points that interconnect the carbon and nitrogen cycles cannot be discerned from existing HZ microbiome studies.

With the overarching goal of providing enhanced resolution to microbial and viral contributions to carbon and nitrogen cycling in the HZ, we created the first of its kind Hyporheic Uncultured Microbial and Viral (HUM-V) genomic catalog. We then used HUM-V to recruit metaproteomic data collected from 33 laterally and depth distributed HZ sediment samples. We further supported this gene expression data using chemical data from paired metabolomics and geochemical measurements. Our results (i) profiled expressed microbial metabolisms that support organic and inorganic carbon and nitrogen cycling in the HZ, (ii) uncovered roles for viruses that could modulate microbial activity in the HZ, and (iii) created a roadmap of a microbial metabolic circuitry that potentially contributes to greenhouse gas fluxes from rivers. We anticipate that this publicly available community resource will advance future microbial activity-based studies in HZ sediments and is a step towards the development of biologically aware, hydro-biogeochemical predictive models.

## *2.3 Results and Discussion*

### 2.3.1 *HUM-V greatly expands the genomic sampling of HZ microbial members*

We used previously collected samples from HZ sediment cores from the Hanford Reach of the Columbia River in eastern Washington, USA [23], an 80km stretch of cobble-bed river that often experiences rapid discharge fluctuations [24]. From this system, six samples per transect were collected, and each core was subsampled into six 10cm depth increments (0-60 cm) (**Figure 2.1ab**). Of these 36 samples, 33 were subsequently processed for metagenomic sequencing, geochemistry, metaproteomics, and FTICR-MS (**Figure 2.1c**, **Appendix B**, **see methods**). A subset of these (n=17) were also analyzed for NMR metabolites. For our metagenomics data, we obtained 379 Gbp of sequencing across all 33 samples which included i) the original shallow

sequencing of all samples (1.7-4.9 Gbp/sample) [23] and ii) an additional deeper sequencing of 10 samples (15.3-49.2 Gbp/sample), which are reported here for the first time. We reconstructed 655 metagenome assembled genomes (MAGs), of which 102 were denoted as medium or high-quality per current standards [25]. MAGs from both transects were then dereplicated (99% identity, **see methods**) into the 55 unique genomic representatives that constitute the microbial component of HUM-V (**Appendix B** and **Data Availability**). Of the MAGs retained in HUM-V, 36% were obtained from deeply sequenced, assembled, and binned samples, 27% were from co-assemblies performed across samples, and the remaining 37% came from single-assemblies of shallow sequences. The ability to recover additional MAGs relative to our teams prior effort which only used shallow sequencing [23] demonstrates how sequencing depth and integration of co-assembly methods enhanced our ability to sample microbial HZ communities, corroborating findings from other studies with similar methods [26].

Given the few metagenomic studies in HZ sediments, it was not surprising that HUM-V contained the first MAG representatives of highly prevalent microorganisms (**Figure 2.2ab, Appendix B**). Taxonomic assignment of the 55 unique HUM-V MAGs revealed they spanned 2 Archaeal and 9 Bacterial phyla, and that most MAGs (n=35) belonged to a subset of 3 bacterial phyla (*Desulfobacterota*, *Nitrospirota*, and *Proteobacteria*). To our knowledge, the 8 *Desulfobacterota* (Class *Binatia*) and 7 *Proteobacteria* (Orders *Rhizobiales*, *Burkholderiales*, *Steroidobacterales*, *Thiohalobacterales*, and *Woeseiales*) MAGs identified here represent the first HZ MAGs sampled from these commonly reported lineages. For the *Nitrospirota*, a prior study reported 21 MAGs that we dereplicated into 12 unique MAGs (99% ANI) [21], a sampling we further expanded by an additional 20 MAGs. The *Nitrospirota* MAGs sampled here spanned 3 genera that to our knowledge have not been previously sampled from rivers (*Nitrospiraceae 2-*

*02-FULL-62-14*, *40CM-3-62-11*, and *NS7*). Moreover, HUM-V contains one MAG of the *Actinobacteriota* that may represent a new order, as well as 6 new genera from *Acidobacteriota*, *Actinobacteriota*, *CSP1-3*, *Desulfobacterota, Proteobacteria*, and *Thermoplasmatota* (**Fig 2.2ab**). Further highlighting the genomic novelty of this ecosystem, HUM-V contains MAGs from entirely uncultivated members of different phyla (9 MAGs from *CSP1-3* and *Eisenbacteria*) and classes (10 MAGs from *Binatia* and *MOR-1*). Ultimately, HUM-V is a public MAG resource that can be leveraged to enable taxonomic analyses and metabolic reconstruction of microbial metabolisms in HZ sediments.

*2.3.2 HUM-V recruits metaproteomes offering new insights into HZ microbiomes*

Leveraging paired metaproteomes collected with the metagenomes allowed us to assign gene expression to each MAG in HUM-V (**Appendix B**). These MAGs recruited 13,102 total peptides to 1,313 proteins. Because our genome analyses revealed that there were closely related strains (**Appendix B**), we analyzed the proteomic data using two approaches. First, we considered the 'unique' peptides that were only assigned to proteins from a single MAG. These represented 67% of genes expressed in our proteome. Next, we considered proteins that recruited 'non-unique but conserved' peptides which we defined as those assigned to proteins that (i) have identical functional annotation and (ii) are from more than one MAG within the same genus. These proteins are shown in gray on **Figure 2.2b**, and although they accounted for a smaller fraction of the genes expressed (14%), this prevented us excluding data due to strain overlap in our database.

In microbiome studies, dominance is often used as a proxy for microbial activity. Here, we evaluated this assumption using our paired metagenome and metaproteome data. When comparing the MAG relative abundances to protein expression patterns, we observed that the

most abundant MAGs were not necessarily those that were most actively expressing proteins at the time of sampling. The most abundant MAGs included members of the *Binatia, Nitrospiraceae NS7*, and *Nitrososphaeraceae TA-21* (formerly *Thaumarchaeota*) (**Figure 2.2b**). However, only the dominant *Nitrososphaeraceae* MAGs had high recruitment of the uniquely assigned proteome. On the other hand, some low abundance members (e.g., *Actinobacteriota*) accounted for a sizeable fraction (30%) of the uniquely assigned proteome.

Leveraging these metagenomic and metaproteomic datasets, we first examined metabolic traits that were conserved across nearly all HUM-V MAGs. Notably, all but one (CSP1_3_1) of the MAGs recovered from this site encoded the genomic capacity for aerobic respiration. We defined this capability by the recovery of genes indicating a complete electron transport chain and some form of terminal oxidase within each MAG. Consistent with this genomic data, resazurin reduction assays indicated the bulk of sediments were oxygenated and could likely support aerobic microbial respiration (**Figure 2.3a**) [27]. However, while proteomic evidence for aerobic respiration (cytochrome c oxidase *aa3*) was detected in nearly 40% of samples, it could only be confidently assigned to the *Nitrososphaeraceae.* This is likely due to the highly conserved nature of this gene, as well as the limitations of detecting membrane, heme-containing cytochromes with metaproteomic data [28]. As such, we consider it likely this metabolism was more active than was captured in the metaproteomic data.

We performed ordination analyses of our MAG-resolved metaproteomic recruitment and revealed the recruited gene expression in each sample did not cluster significantly by sediment depth or transect position (**Figure 2.4ab**). These MAG-resolved results agree with those previously published [23] using an unbinned metaproteome approach. That is, neither study observed any structuring at the transect level based on any microbial data type (i.e.,

metagenomics or metaproteomics). Consistent with this, over 90% of measured gene expression was shared across both transects (**Figure 2.4c**). Contrastingly, significant differences at the transect level were observed for metabolite concentrations and non-biotic data like molecular weight and carbon types, as previously reported [23]. As such, when considering explanations for this lack of microbiological spatial structuring, it is possible that 1) the microbial gene expression heterogeneity in these samples occurred over a finer spatial resolution (pore or biofilm scale, <10 cm) or larger (>60 cm) than those sampled here and thus were not captured in our analyses, and/or 2) that while the chemical data shows changes across transects, these changes do not differentially shape the metabolic processes of the microorganisms as the same substrate types are still mainly present in both regions.

### *2.3.3 An inventory of processes contributing to microbial carbon dioxide production and consumption*

To uncover the microbial food web contributing to organic carbon decomposition in these HZ sediments, we reconstructed a carbon degradation network using coordinated genome potential, expression, and carbon metabolite data. Based on linkages to specific substrate classes, MAGs were assigned to the following trophic levels in carbon decomposition: (i) plant polymers (ii) smaller organic compounds (e.g., sugars, alcohols, and fatty acids), and (iii) single carbon compounds (carbon monoxide, carbon dioxide, methane) (**Figure 2.5**).

It is well recognized that heterotrophic oxidation of organic carbon derived in HZ sediments largely contributes to river respiration [2]. Despite generally low organic concentrations in our sediments (<10 mg/g), FTICR-MS analysis showed that lignin-like compounds were the most abundant biochemical class detected in all samples regardless of transect or depth suggesting that plant litter was a likely source of organic carbon (**Figure 2.6a**). In support of this,

38% of the HUM-V MAGs encoded genes for degradation of phenolic/aromatic monomers, while 11% could degrade the larger, more recalcitrant polyphenolic polymers. In fact, our analyses revealed that seven unique MAGs constituting a new genus within the uncultivated *Binatia* encoded novel pathways for the decomposition of aromatic compounds from plant biomass (phenylpropionic acid, phenylacetic acid, salicylic acid), and xenobiotics (phthalic acid) (**Appendix B**).

Gene expression of carbohydrate-active enzymes (CAZymes) also supported the degradation of plant polymers like starch and cellulose. We detected the expression of putative extracellular glucoamylase (GH15) and endo-glucanase (GH5) from an *Actinobacteriota* (Microm_1) and *Nitrososphaeraceae* (Nitroso_2) MAG, respectively. Additionally, using our unbinned assembled fractions we detected expression of 3 GH33 CAZymes which could further contribute to the oxidation of organic carbon, and were likely assigned to unbinned *Rokubacteria* and an unknown *Actinobacteria*. The integration of our chemical and biological data revealed that heterotrophic metabolism in these sediments could in part be maintained by inputs of plant biomass. In support of carbon depolymerization, sugars like glucose and sucrose were detected by nuclear magnetic resonance (NMR) (**Figure 2.6b**).

We next sought to identify microorganisms that could utilize these sugars and found that members expressed transporters for fructose (Rhizo-Anders_1), glucose (Microm_1), and general sugar uptake (Actino_1, Nitroso_2, Nitroso_3). In support of further decomposition, we detected organic acids (acetate, butyrate, lactate, pyruvate, propionate) and alcohols (ethanol, methanol, isopropanol) by NMR (**Figure 2.6b**). Similarly, proteomics data supported interconversions of these smaller carbon molecules, with the *Myxococcota* (Anaerom_1) expressing genes for aerobic acetate respiration and the archaeal *Woeseia* (Woese_1) respiring

methanol. In summary, the chemical scaffolding and overlayed gene expression patterns support an active heterotrophic metabolic network in these HZ, likely driven by plant biomass decomposition.

In addition to heterotrophy, our proteomics data revealed autotrophy was also active in these sediments. Dehydrogenase genes for the aerobic oxidation of carbon monoxide (CO) were among the most prevalent across these sediments. This metabolism was expressed by phylogenetically distinct lineages, including members of uncultivated lineages *Binatia* (Binatia_2) and *CSP1-3* (CSP1_3_1), as well as members of *Actinobacteriota* (Actino_1, Microm_1), *Methylomirabilota* (Roku_AR37_2), and *Proteobacteria* (Burk_1, Thioh_1). The wide range of bacteria and archaea that encoded CO dehydrogenase genes, combined with gene expression data, suggests carbon monoxide oxidation may be an important metabolism for persistence in HZ sediments.

Given these sediments have relatively low total carbon concentrations (**Figure 2.3b**), we consider it possible that carbon monoxide may act as a supplemental microbial energy and/or carbon source. Based on genomic content, we cautiously infer members of *Actinobacteriota* (Microm_1), *Binatia*, and *CSP1-3* may be capable of carboxydotrophy (i.e., using carbon monoxide as sole energy and carbon source), while the *Actinobacteriota* (Actino_1) is a likely carboxydovore (i.e., oxidize carbon monoxide, while requiring organic carbon). While this metabolism is poorly resolved environmentally, recent efforts have shown it is induced by organic carbon starvation to mediate aerobic respiration, thereby enhancing survival in oligotrophic conditions [29]. Here we add river sediments to list of oxygenated environments (e.g., ocean and soils) where this metabolism may act as a sink or regulate the emission of this indirect greenhouse gas (GHG) [30,31].

Since proteomics indicated heterotrophy and carbon monoxide oxidation could generate carbon dioxide, we next tracked microorganisms in HUM-V that could fix this compound, sequestering its release. Analyses revealed four pathways for carbon fixation were encoded by 75% of HUM-V MAGs including (i) Calvin-Benson-Bassham cycle, (ii) reductive TCA cycle, (iii) 3-HydroxyPropionate /4-HydroxyButyrate cycle, and (iv) 3-Hydroxypropionate bi-cycle. The two nitrifying lineages were inferred chemolithoautotrophs, with *Nitrososphaeraceae* encoding HydroxyPropionate/4-HydroxyButyrate (3HP/4HB) and the *Nitrospiraceae* encoding the reductive tricarboxylic acid (TCA) cycle. Additionally, phylogenetically diverse lineages, *Acidobacteriota*, *Binatia, CSP1_3, Proteobacteria*, and *Woeseiaceae* encoded redundant fixation pathways. Although expression was not detected for these metabolisms in our MAGs or unbinned data, we hypothesize these are likely relevant given the distribution of this metabolism across the microbial community.

Our genomic and proteomic data revealed the prevalence and activity of single carbon metabolism in these sediments. Carbon monoxide and dioxide are likely the primary substrates, as HUM-V only had minimal evidence for methanol oxidation (*Woeseia*), no methanotrophs, and no methanogens. Along these lines, the unbinned metaproteomics approach also did not detect evidence for methanotrophy or methanogenesis in these samples. Together our findings hint at the importance of carbon monoxide and carbon dioxide in sustaining microbial metabolism in these oxygenated, but low, or fluctuating, carbon environments. Further work is needed to understand physiochemical factors controlling carbon monoxide oxidation and carbon dioxide fixation activity, and the balance between production (via heterotrophy and carbon monoxide oxidation) and consumption (fixation) on overall river sediment carbon dioxide emissions.

*2.3.4 Ammonium exchange can support coordinated nitrogen mineralization and nitrification pathways*

The ratio of total carbon (C) (**Figure 2.3b**) and total nitrogen (N) (**Figure 2.3c**) (e.g., C/N) is a geochemical proxy used to denote the possible microbial metabolisms that can be supported in a habitat [32,33]. Our HZ sediments had C/N ratios with a mean of 6.5 ± 1.1 (maximum 8.4) (**Figure 2.3d**). Geochemical theory posits that sediments with low C/N ratios (<15) support organic mineralization that yields sufficient ammonium such that heterotrophic bacteria are not N-limited and nitrifying bacteria are able to compete successfully for ammonium enabling nitrification [33–35]. Based on our sediment C/N ratios, we hypothesized organic nitrogen mineralization and nitrification co-occurred in these sediments. Here we profiled the microbial substrates (organic nitrogen metabolites, ammonium) and expressed pathways (mineralization and nitrification) to provide biological validation of this established geochemical theory.

To examine the microbial contributions to organic nitrogen mineralization, we examined metaproteomic data for peptidases, genes that mineralize organic nitrogen into amino acids and free ammonium. In support of active microbial N mineralization, FTICR-MS revealed that protein-like and amino sugar like organic nitrogen compounds were correlated to high microbial activity [23], while here we show hydrophobic, polar, and hydrophilic amino acids were prevalent in the $H^1$-NMR characterized metabolites (**Figure 2.6ab**). The expression of peptidases *in situ*, combined with our genomic resolution of their hosts, provided a new opportunity to interrogate the mechanisms underpinning nitrogen mineralization. We first noted which microorganisms expressed extracellular peptidases (inferred from [36–38]), as these enzymes could shape the external organic nitrogen pools in the sediment. We categorized these expressed peptidases as either releasing free amino acids (end terminus cleaving families, e.g., M28) or releasing

peptides (endocleaving families, e.g., S08A, M43B, M36, MO4) (**Figure 2.7a**). Members of the *Actinobacteriota*, *Binatia, Methylomirabilota*, and *Thermoproteota* were found to express extracellular peptidases and as such, are likely candidates that contribute to sediment N mineralization.

We then profiled the expressed amino acid transporters in our MAGs, i.e., genes for the cellular uptake of these smaller organic nitrogen compounds (e.g., branched chain amino acids, glutamate, amines, and peptides were examined) (**Figure 2.7b**). At the time of sampling, some members expressed peptidases for cleavage of organic N to liberating smaller peptides, yet we could not detect expressed genes for the transport of these produced compounds. Other taxa like *Actinobacteriota* and *Binatia* expressed genes for external peptidases and for transporting the organic N products into the cell. Alternatively, we entertain the possibility that members of the *CSP1-3, Proteobacteria*, and *Thermoplasmatota* as these members expressed genes for assimilating peptidase products but did not contribute to the cost of their production.

While it is tempting to speculate that at the time of sampling some members of the community may be operating as producers, and thus extending this concept to organic nitrogen processing as others have noted for carbon decomposition [39–41], we note the metabolic potential of these microorganisms is more diverse than their expressed patterns, as is shown in **Figure 2.7**. Thus, we could have missed the expression of these genes as they may be below the detection in the proteome data or they may be missed entirely because our MAGs are draft genomes (inferred completion mean is 82%, maximum 99%). As such the absence of data here should be interpreted cautiously, and future research using model cultivated strains analogous to the work by Pollak et al for polysaccharides as a public good [39] would be necessary to classify producers and cheaters, and the conditions under which they operate, in organic nitrogen processing.

Finally, we examined the proteomes for evidence that nitrification co-occurred with organic nitrogen mineralization. Supporting this possibility, the substrate ammonium ($NH_4^+$) was detected in all 33 sediment samples (**Fig 2.3e, Figure 2.6b**). We did not detect genomic evidence or expression for comammox or anammox metabolisms in HUM-V MAGs and did not identify these metabolisms in our metaproteomes mapped to the unbinned, assembled data. This suggests aerobic nitrification by different organisms drives nitrification in our metabolic network. Proteomics confirmed aerobic ammonium oxidation to nitrite was performed by archaeal *Nitrososphaeraceae*. In fact, ammonia monooxygenase (*amo*) subunits were within the top 5% most highly expressed functional proteins in this dataset. The next step in nitrification, nitrite oxidation to nitrate, was inferred from nitrite oxidoreductase (*nxr*) which was expressed by members of *Nitrospiraceae*. Demonstrating that new lineages first discovered in HUM-V could shape in situ biogeochemistry, we confirmed that 5 MAGs from two new species of *Nitrospiraceae* expressed nitrification genes (Nitro_40CM-3_1, Nitro_NS7_3, Nitro_NS7_4, Nitro_NS7_5, and Nitro_NS7_14).

The proteome supported archaeal-bacterial nitrifying mutualism outlined here appear well adapted to the low nutrient conditions present in many HZ sediments, warranting future research on the universal variables that constrain nitrification rates (i.e., ammonium availability, dissolved oxygen, pH) and their role in driving nitrogen fluxes from these systems [42]. In conclusion, our microbial data supports the idea that nitrification is concomitant with mineralization in these samples, providing biological evidence to substantiate inferences made from the C/N ratio of these sediments.

*2.3.5 Metabolic contributions from less characterized taxa actively shape nitrogen cycling*

Our proteomics suggests that aerobic nitrification could complement allochthonous nitrate from groundwater discharges, contributing to measured nitrate concentrations in excess of 20 mg/L [2,43]. Based on this, we next sought to summarize the metabolic potential and expressed gene content related to the use of nitrate or oxidized nitrogen species in our metagenome dataset.

HUM-V MAGs with the genomic capacity for dentification were phylogenetically diverse (**Appendix B**). Nitrate reductases (*narG* or *napX)* were encoded in an additional 11 MAGs from the *Actinobacteriota*, *Binatia, Myxococcota*, and *Proteobacteria*, of which 2 had the subsequent capacity to reduce nitrite to nitric oxide gas. Notably the conversion of nitric oxide to nitrous oxide, a potent greenhouse gas was encoded by two *Gammaproteobacteria* (Steroid-FEN-1191_1, Steroid_1) and a member of the *Myxococcota* (Anaerom_1). We also detected that the dissimilatory nitrite reduction to ammonium (DNRA) was encoded by multiple members of the *Binatia* and *Nitrospiraceae* (**Appendix B**).

Based on the gene expression data, three members were assumed to be actively contributing to nitrogen reduction at the time these samples were collected. The only *narG* detected in our proteome for our MAGs was uniquely assigned to the *Binatia*, while proteins for nitrite reduction were assigned to the denitrifying *Burkholderia* and nitrifying *Nitrososphaeraceae* (**Appendix B**). We did detect expression of *narG* and *nirK* within our unbinned data, however they were taxonomically inferred to be from *Nitrospiraceae* and *Nitrososphaeraceae* represented in HUM-V. We did not detect the expression of DNRA in our HUM-V MAGs or the unbinned fractions, but due to the high number of genomes encoding this functionality, we hypothesize it could play important roles in ammonia generation by microbes in these sediments. Additionally, while nitrous oxide production genes were not expressed in either the binned or unbinned fractions, we did detect *nos* gene expression for reducing nitrous oxide to

nitrogen gas by the non-denitrifying *Desulfobacterota_D* (Desulf_UBA2774_1, formerly *Dadabacteria* [44]. Our phylogenetic analysis (**see Data Availability**) revealed this sequence was a "Clade II" *nosZ* sequence, an atypical variant adapted for lower, or atmospheric nitrous oxide concentrations that is often encoded in non-denitrifying lineages [45].

Our metaproteome data adds to emerging interest on the use of untargeted approaches to identify the nitrogen transforming genes that are expressed in hyporheic sediments. To our knowledge we provide the first gene expression evidence for two lineages lacking cultured representatives, the *Binatia* and a member of the *Desulfobacterota*, in hyporheic zone denitrification. We also provide expression data supporting the notion clade II *nosZ* gene expression could act as a nitrous oxide sink (without contributing to its production) [46,47]. Importantly, the activity of this enzyme would have been missed using traditional *nosZ* primers, denoting the value of our untargeted, expression-based approach [48]. In summary, our gene expression data pinpoints new metabolic contributions from less characterized taxa that actively shape nitrogen cycling in the hyporheic zone.

*2.3.6 Viral influence on sediment carbon and nitrogen cycling*

We reconstructed 2,482 vMAGs that dereplicated into 111 dereplicated viral populations (>10kb) in HUM-V (**Figure 2.8a, Appendix B**), making this one of only a handful of genome-resolved studies that include viruses derived from rivers [16,49,50]. To our knowledge, this is the first study to provide a coordinated analyses of microbial and viral community MAGs, and given their sparse sampling, only 5 of the 111 HUM-V vMAGs had taxonomic assignments established from standard reference databases. To better understand if the remaining vMAGs had been previously detected in other virally sampled freshwater systems, we compared the protein content of vMAGs in our system to an additional 1,861 vMAGs we reconstructed *de novo* or

obtained from public metagenomes from North and South America (**Figure 2.8b, Appendix B**). Of the 106 non-taxonomically assigned viruses, 15% (n=17) clustered with these freshwater derived viral genomic sequences indicating possible cosmopolitan viruses, and another 23% (n=26) clustered only with vMAGs recovered in this data set, indicating multiple samplings of the same virus across different sites and depths. The remaining 57% (n=63) of the vMAGs were singletons, meaning they were only sampled from these sediments once. Combined, these results hint at the possible biogeographically diverse, as well as endemic viral lineages warranting further exploration in river sediments.

We then assessed peptide recruitment to the viral portion of HUM-V (**Figure 2.8a, Appendix B**). For viruses and microbes alike, the most abundant vMAGs did not have the highest gene expression. While viral gene expression was not structured by edaphic or spatial factors (**Figure 2.9ab**), it was strongly coordinated to the microbial abundance patterns (**Figure 2.9c**). Like our microbial MAG peptide recruitment, 66% of the vMAGs uniquely recruited peptides. This exceeded prior viral metaproteome recruitment from other environmental systems (e.g., wastewater, saliva, rumen) with ranges from 0.4-15% [51–53]. From this we infer a relatively large portion of the viral community was active at the time of sampling.

The proteomic recruitment of viruses sampled in HUM-V hinted at the possibility that viruses could structure the microbial biogeochemistry through predation. *In silico* analysis assigned a putative host to 29% of the 111 vMAGs. Viruses were linked to 18 microbial MAGs that belong to bacterial members in *Acidobacteriota*, *Actinobacteriota*, *CSP1-3*, *Eisenbacteria, Methylomirabilota*, *Myxococcota*, *Nitrospirota*, and *Proteobacteria* (**Figure 2.8c**, **Figure 2.9de**). Analysis of the metaproteomes for these putative phage-impacted MAGs revealed these hosts expressed genes for carbon monoxide oxidation (*Actinobacteriota*), carbon fixation (*CSP1-3*),

nitrogen mineralization (*Acidobacteriota, and Methylomirabilota*), methanol respiration (*Myxococcota*), nitrification (*Nitrospirota*), and ammonia oxidation (*Proteobacteria*) (**Figure 2.8d**). Thus, viral predation in HZs could impact carbon and nitrogen biogeochemistry and may explain some of the strain and functional redundancy we observe in the microbial communities of these sediments.

We next inventoried HUM-V vMAGs for auxiliary metabolic genes (AMGs) with the potential to augment biogeochemistry. We detected 14 auxiliary metabolic genes (AMGs) which we confirmed were not bacterial in origin and had viral-like genes on both flanks [54]. These putative AMGs had the potential to augment carbon (CAZymes), sulfur (sulfate adenylyltransferase), and nitrogen (amidase to cleave ammonium) metabolism (**Figure 2.9f**). One of our vMAGs that was putatively linked to a *Steroidobacteraceae* (Steroid_1) encoded a polysaccharide lyase gene (PL1). This viral PL1 could enable its host to cleave the backbone of pectin, generating pectin oligosaccharides that could be used via two host-encoded glycoside hydrolases (GH4 and GH2), ultimately freeing galactose for energy metabolism (**Figure 2.9gh**). While theoretical, we include this as an example to illustrate how virally encoded genes could expand the substrate ranges for their hosts and alter biogeochemical cycling in river sediments.

In support of their importance to modulating microbial activity and sediment biogeochemistry, we noted that using the vMAG abundance patterns in addition to MAG abundance patterns improved our predictions of river sediment carbon and nitrogen concentrations (**Figure 2.10**). In summary, these results indicate that viral predation and AMGs may contribute to river sediment biogeochemistry, either through top down or bottom-up controls on the microbial community.

### *2.4 Conclusion*

Despite the importance of the HZ and its relative accessibility in terms of sampling locations, HZ microbial and viral communities are surprisingly under sampled in a genomic context. Previous studies pertaining to this ecosystem are not genome resolved and used 16S rRNA amplicons or unbinned metagenomes [8–11,19,20,22], thus limiting the predictive and explanative power of the study and often times not discriminating between metabolically active and inactive organisms. Further, the few studies which are genome-resolved [21,22] often focus on specific lineages or single processes and not the entire microbial community, missing the complex interplays between the carbon and nitrogen cycles.

Here we created HUM-V, a MAG resolved database to expand on prior non-genome resolved analyses done in a previous publication by our team [23]. Our MAG-resolved proteomic data further provided some of the first activity indicators for members of hyporheic microbiomes. As an example, we focus on new insights gleaned from the 7 recovered *Binatia* MAGs (one genome included a complete 16S rRNA gene), which are uncultured, and have only been described in terms of metabolic potential by one previous publication [55]. Proteomics demonstrated these bacteria (i) aerobically oxidized carbon monoxide, (ii) mineralized organic nitrogen, and (iii) denitrified, contributing to carbon and nitrogen cycling in these sediments (**Fig 2.11a**). Using 16S rRNA recovered from these MAGs, we show closely related strains are biogeographically widespread (**Figure 2.11b**), and thus it is possible the gene expression findings we illuminate here can be more widespread across other habitats. These results illustrate the power of HUM-V in uncovering new roles for members of uncultivated, previously enigmatic microbial lineages in hyporheic zone biogeochemistry.

Empowered by our process-based metaproteomic analyses (**Figure 2.2-Figure 2.11**), we present a conceptual model outlining microbial and viral contributions to carbon and nitrogen biogeochemistry in these sediments (**Figure 2.12**). Together, we demonstrated how these pathways could result in the formation and depletion of nutrients in shared resource pools. From gene expression data we suggest a network of metabolisms that affected organic and inorganic carbon cycling, and is intertwined with nitrogen mineralization, nitrification, and denitrification pathways.

In conclusion, while river carbon and nitrogen budgets are often quantified by direct measurements of inputs, and the concentration of inorganic and organic compounds exported from rivers, our findings put forth an integrated framework that advances the resolution of microbial roles in hyporheic carbon and nitrogen transformations. It yields insights that could inform research strategies to reduce existing predictive uncertainties in river corridor models and resolves some of the microbial contributions that were thought to occur but were poorly defined in river sediments (e.g., nitrogen mineralization). We also highlight previously enigmatic processes that could directly impact river GHG fluxes in unappreciated ways (e.g., carbon dioxide fixation, carbon monoxide oxidation, type II nitrous oxide reduction). Ultimately, we show that a MAG-resolved database allows us to track the consumption and production of carbon dioxide, ammonium, and inorganic nitrogen and helps us explain how these transformations could contribute to overall GHG fluxes.

## *2.5 Materials and Methods*

### *2.5.1 Sample collection, DNA isolation, and chemical characterization*

Samples were collected from the hyporheic zone of the Columbia River (46°22'15.80″N, 119°16'31.52″W) in March 2015 as previously described [23]. Briefly, sediment profiles (0-60cm) were collected along two transects separated by approximately 170 meters (**Figure 2.1**). At each transect, three sediment cores up to 60 cm in depth were collected at 5-meter intervals perpendicular to the river flow. Liquid $N_2$-frozen sediment profiles were collected as detailed previously [56], with a pointed stainless-steel tube (152 cm length, 3.3 cm outside diameter, 2.4 cm inside diameter) driven into the riverbed and liquid $N_2$ poured down the tube for ~15 min. Once sediments were frozen the tube and attached material were removed from the riverbed with a chain hoist suspended beneath a tripod. Profiles were placed over an aluminum foil lined cooler containing dry ice. The material was then wrapped in the foil and transported on dry ice to storage at −80 °C. In the lab, each core was then sectioned into 10 cm segments from 0-60-centimeter depths for downstream analyses, except for core "N2" which had the first 3 depths (0-30cm) subsampled together, and for core "N1" 50-60 cm which was damaged and did not pass quality control. For processing, samples were transferred to an anaerobic glove bag with 95% $N_2$ and 5% $H_2$ (Coy Laboratory Products, Grass Lake, MI) and thawed on clean 2mm stainless steel sieves prior to processing. Approximately 5g was transferred into a 40-mL borosilicate glass vial and stored at -80ºC for chemical analyses. An additional sample was taken for elemental analysis (OC, N) and remaining material was divided into 20g samples collected into 40-mL borosilicate glass vials and stored at -80ºC.

DNA isolation was carried out as previously described [23]. To release biomass from sediment particles, thawed samples were suspended in 20 mL of chilled PBS/0.1% Na-

pyrophosphate solution and vortexed for 1 min. The suspended fraction was decanted to a fresh tube and centrifuged for 15′ at 7000 ×g at 10 °C. DNA was extracted from the resulting pellets using the MoBio PowerSoil kit in plate format (MoBio Laboratories, Inc., Carlsbad, CA) following manufacturer's instructions, with the addition of a 2-hour proteinase-K incubation at 55 °C prior to bead-beating to facilitate cell lysis. DNA was then sent to the Joint Genome Institute (JGI, n=33) for sequencing. The additional deep sequencing described here was performed at the Genomics Shared Resource facility at The Ohio State University (OSU, n=10) using a Nextera XT library System. Libraries at both facilities were sequenced using an Illumina HiSeq 2500 platform. **Appendix B** details all sequencing information, including accession numbers for metagenomes.

Chemical analyses included geochemical and metabolite data, where geochemistry and Fourier-transform ion cyclotron resonance mass spectrometry (FTICR-MS) methods were performed as previously described [23]. Regarding FTICR-MS, sediments were extracted with three solvents with different polarities - water ($H_2O$), methanol ($CH_3OH$, 'MeOH') and chloroform ($CHCl_3$)—to sequentially extract a large diversity of organic compounds from samples, according to previous publications [57,58]. Water extractions were performed first, followed by MeOH and then $CHCl_3$. Ultra-high resolution mass spectrometry of the three different extracts from each sample was carried out using a 12 Tesla Bruker SolariX FTICR-MS located at the Environmental Molecular Sciences Laboratory (EMSL) in Richland, WA, USA.

The total nitrogen, sulfur, and carbon content was determined using an Elementar vario El cube (Elementar Co. Germany). $NH_4^+$ was extracted with KCl and measured with Hach Kit (Hach, Loveland, Co). Aerobic metabolism was inferred by the resazurin reduction assay, based on method previously described [27]. FTICR-MS compounds are reported as relative abundance

values based on counts of C, H, and O. The relative abundance of a biochemical class is defined as (# of formula in class per sample / total # of formula per sample). This was done for the following H:C and O:C ranges: lipids ($0 < O:C \leq 0.3$, $1.5 \leq H:C \leq 2.5$), unsaturated hydrocarbons ($0 \leq O:C \leq 0.125$, $0.8 \leq H:C < 2.5$), proteins ($0.3 < O:C \leq 0.55$, $1.5 \leq H:C \leq 2.3$), amino sugars ($0.55 < O:C \leq 0.7$, $1.5 \leq H:C \leq 2.2$), lignin ($0.125 < O:C \leq 0.65$, $0.8 \leq H:C < 1.5$), tannins ($0.65 < O:C \leq 1.1$, $0.8 \leq H:C < 1.5$), and condensed hydrocarbons ($0 \leq 200$ $O:C \leq 0.95$, $0.2 \leq H:C < 0.8$) [57].

Additional metabolite data was obtained through $^{1}H$ Nuclear Magnetic Resonance (NMR) spectroscopy on water extracted sediments. Thawed sediment samples were mixed with 200, 300, or 600 µL of MilliQ water depending on the sediment mass (**Appendix B**) and centrifuged to remove the sediment [58–60]. Supernatant (180 µL) was then diluted by 10% (vol/vol) with 5 mM 2,2-dimethyl-2-silapentane-5-sulfonate-$d_6$ as an internal standard. All NMR spectra were collected using a Varian Direct Drive 600-MHz NM R spectrometer equipped with a 5-mm triple resonance salt-tolerant cold probe. Chemical shifts were referenced to the 1H or 13C methyl signal in DSS-d6 at 0 ppm. The 1D $^{1}H$ NMR spectra of all samples were processed, assigned, and analyzed using Chenomx NMR Suite 8.3 with quantification based on spectral intensities relative to the internal standard as described previously [61,62]. For the NMR, while obtaining concentrations is possible, many compounds were below the limit of quantitation (2µM), but above the limit of detection (1µM). To still derive meaning from this data, we reported NMR-identified compounds as present (detected) or absent (not >1µM), with "relative abundance" reported as the # formula in class per sample/ total # formula per sample. All geochemical and metabolite data can be found in **Appendix B**.

*2.5.2 Metagenome assembly and binning*

Raw reads were trimmed for length and quality using Sickle v1.33 (https://github.com/najoshi/sickle) and then subsequently assembled using IDBA-UD 1.1.0 [63] with an initial kmer of 40. Two of our samples did not assemble with IDBA-UD 1.1.0 at this kmer and were assembled with metaSPAdes 3.13.0 [64] using default parameters (**Appendix B**). To further increase genomic recovery, for the ten samples that had shallow and deep sequencing, metagenomic reads were co-assembled using IDBA-UD 1.1.0 with an initial kmer of 40 (**Appendix B**). All assemblies, including co-assemblies, were then individually binned using Metabat2 v2.12.1 [65] with default parameters to obtain microbial MAGs.

For each bin, MAG completion was estimated based on the presence of core gene sets using Amphora2 (e-value=$1e^{-3}$, which used RaxML (v8.2.9), HMMER (v3.3), and EMBOSS (v6.6.0.0)) and CheckM v1.1.2 [66,67]. To ensure only quality MAGs were utilized for metabolic analyses, we discarded all MAGs that had completion <70% and contamination >10%. This equates to "high quality" (HQ) bins and a more stringent "medium quality" (MQ) bin cutoff than those used by the genome consortium standard [25]. These 102 MAGs (6 HQ, 96 MQ) that were then dereplicated using dRep [68] with default parameters to result in a final set of 55 MAGs (>99% ANI which represents strain-level MAG distinctions) (**Appendix B**). To further assess bin quality, we used the Distilled and Refined Annotation of MAGs (DRAM) v1.0 tool [54] to identify ribosomal ribonucleic acids (rRNAs) and transfer ribonucleic acids (tRNAs) to ensure they were taxonomically consistent with overall taxonomic assignments. To further curate our bins, we used bowtie2 [69] to calculate the coverage of each scaffold within a MAG and manually checked scaffolds that had 10% higher coverage than the mean, confirming consistency in taxonomic

assignment and annotations of the scaffold in question with the overall bin. MAG quality and taxonomic information reported in **Appendix B**.

*2.5.3 Metabolic and taxonomic analyses of MAGs*

Medium and high-quality MAGs were taxonomically classified using the Genome Taxonomy Database (GTDB) Toolkit v1.3.0 using reference data r95 on September 2020 [70]. Novel taxonomy was identified as the first taxonomic level with no designation using GTDB taxonomy. MAG scaffolds were annotated using the DRAM v1.0 tool which uses PFAM (v33.1), KEGG (v89.1), dbCAN (v9), MEROPS (v120), and VOGDB database for annotations [54]. Phylogenetic analyses were performed on genes annotated as respiratory nitrate reductase (*nar*) and nitrite oxidoreductase (*nxr*) to resolve novel *Binatia* role in nitrogen cycling. Specifically, sequences from [71] were downloaded and combined with *nar* and *nxr* amino acid sequences from dereplicated bins, aligned using MUSCLE (v3.8.31) and run through an in-house script for generating of phylogenetic trees (https://github.com/WrightonLabCSU/columbia_river). Phylogenetic trees are provided in Zenodo: https://doi.org/10.5281/zenodo.6339808. For polyphenol and carbon polymer degradation, we used predicted secretion and functional annotations to characterize these metabolisms. To determine if the predicted genes encoded a secreted protein, pSortb [36] and SignalP [38] were used to predict location; if those methods did not detect a signal peptide, the amino acid sequence was queried to SecretomeP (with a SecP score > 0.5 [37] used as a threshold to report non-canonical secretion signals). Metabolic characterization for each MAG discussed in this manuscript are available in **Appendix B.**

*2.5.4 Viral Analyses*

Metagenomic assemblies (n=43) were screened for DNA viral sequences using VirSorter v1.0.3 with the ViromeDB database option [72], retaining viral contigs ranked 1, 2, 4 or 5 where category 1-2 indicate high confidence predicted lytic viruses and 4-5 indicate high-confidence prophage sequences from VirSorter output [72]. Viral sequences were filtered based on size to retain those greater than or equal to 10kb based on current standards [73]. Viral scaffolds were then clustered into vMAGs at 95% ANI across 85% of the shortest contig using ClusterGenomes 5.1 (https://github.com/simroux/ClusterGenome) [73]. After clustering, vMAGs were manually confirmed to be viral by looking at DRAM-v annotations and assessing the total viral-like genes relative to non-viral genes per scaffold. Using DRAM-v, vMAGs that were assigned the J-flag (which indicate vMAGs containing more than 18% of non-viral genes), were deemed suspicious, manually confirmed to contain no viral hallmark genes, and subsequently discarded [54]. All vMAG information can be found on **Appendix B**.

To determine taxonomic affiliation, vMAGs were clustered to viruses belonging to viral reference taxonomy databases NCBI Bacterial and Archaeal Viral RefSeq V85 with the International Committee on Taxonomy of Viruses (ICTV) and NCBI Taxonomy using the network-based protein classification software vContact2 v0.9.8 using default methods [74,75]. To determine geographic distribution of viruses in freshwater ecosystems, we included viruses mined from publicly available freshwater metagenomes in vContact2 analyses: 1) East River, CO (PRJNA579838); 2) A previous Columbia River, WA study (PRJNA375338); 3) Prairie Potholes, ND (PRJNA365086); 4) the Amazon River (PRJNA237344). The viral sequences that were identified from these systems and the genes used for vContact2 are deposited on Zenodo with doi 10.5281/zenodo.6310084 with more information of downloaded datasets found in **Appendix B**.

Viral contigs were annotated with DRAM-v [54]. Genes that were identified by DRAM-v as being high-confidence possible auxiliary metabolic genes (auxiliary scores 1-3) [54] were subjected to protein modeling using Protein Homology / AnalogY Recognition Engine (PHYRE2) [76]. Auxiliary scores were assigned by DRAM [54], based on the following ranking system: A gene is given an auxiliary score of 1 if there is at least one hallmark gene on both the left and right flanks, indicating the gene is likely viral. An auxiliary score of 2 is assigned when the gene has a viral hallmark gene on one flank and a viral-like gene on the other flank. An auxiliary score of 3 is assigned to genes that have a viral-like gene on both flanks. To identify likely vMAG hosts, oligonucleotide frequencies between virus (n=111) and non-dereplicated hosts (n=102) were analyzed using VirHostMatcher using a threshold of d2* measurements of <0.25 [77]. The lowest d2* value for each viral contig <0.25 was used. All vMAG annotations are reported in **Appendix B**.

*2.5.5 MAG relative abundance calculations and their use in predictions*

To estimate the relative abundance of each MAG and vMAG, the metagenomic reads for each sample were rarified to 3Gbp and mapped to 55 unique MAGs via Bowtie2 [69] [60,61]. For MAGs, a minimum scaffold coverage of 75% and depth of 3x coverage was required for read recruitment at 7 mismatches. For vMAGs, reads were mapped using Bowtie2 [69] at a maximum mismatch of 15, a minimum contig coverage of 75% and a minimum depth coverage of 2x. Relative abundances for each MAG and vMAG were calculated as their coverage proportion from the sum of the whole coverage of all bins for each set of metagenomic reads. MAG relative abundances per sample for MAGs and vMAGs are reported in **Appendix B**. Correlations and sparse Partial Least Squares Regression (sPLS) predictions (PLS R package [78]) used mapping data pertaining to only the 10 deeply sequenced metagenomes rarified to 4.8Gbp (**Appendix B**).

*2.5.6 Metaproteome generation and peptide mapping*

Metaproteomic mapping results for MAGs and vMAGs can be found on **Appendix B**. Sediment samples were prepared for metaproteome analysis as previously reported in Graham et al. 2018 [23] and the protocol outlined by Nicora et al [79]. As previously described [61,80], tandem mass spectrometry (MS/MS) spectra from all liquid chromatography tandem mass spectrometry (LC-MS/MS) datasets were converted to ASCII text (.dta format) using MSConvert (http://proteowizard.sourceforge.net/tools/msconvert.html) and the data files were then interrogated via target-decoy approach [81] using MSGF+ [82]. For protein identification, spectra were searched against two files that included (i) 55 dereplicated MAG and (ii) 111 clustered vMAGs amino acid sequences. Peptide recruitment for each MAG amino acid sequence per sample is reported in **Appendix B**. Hits were divided into 3 categories: (1) unique- peptide hits to a single protein (2) non-unique specialized- peptide hits to multiple amino acid sequences that all had same annotation and MAG taxonomy (3) non-unique peptide hits to multiple amino acid sequences with different annotation or from MAGs with different taxonomy. This designation was necessary as several hits could not be resolved to the MAG level due to functional conservation across closely related MAGs (strains) in the HUM-V database. Data in **Figure 2.2** showcases (1) and (2) categories. Microbial metaproteomes were converted to normalized spectral abundance frequency (NSAF) values and subsequently divided into unique, non-unique specialized, and non-unique categories, and only reported as expressed when detected in >3 samples (unless otherwise noted and subsequently manually inspected to be good hits). Viral metaproteomes were analyzed using peptide counts only from unique hits due to low recruitment.

To understand changes in the metaproteome recruited to MAGs over spatial gradients non-metric multidimensional scaling (NMDS) ordinations were performed on the subset of 29 samples that recruited sufficient peptides. Analyses were run with multiple transformations and with presence/absence and relative abundance data. Relevant code for **Figure 2.4ab**, **Figure 2.9ab** and input data used can be found on GitHub (https://github.com/WrightonLabCSU/columbia_river).

In addition to our binned approach, we also analyzed all proteins that recruited peptides across all assembled contigs (i.e., both binned and unbinned). Proteins that recruited peptides uniquely and in >3 samples were then annotated with DRAM. To determine if the predicted carbon cycling genes encoded a secreted protein, pSortb [36] and SignalP [38] were used to predict location; if those methods did not detect a signal peptide, the amino acid sequence was queried to SecretomeP (with a SecP score > 0.5 [37] used as a threshold to report non-canonical secretion signals). The genes that were relevant to the manuscript and their annotations are included in **Appendix B**, along with the mapping results for all unique genes. The annotations for all unique hits, without QC, are available on Zenodo: https://doi.org/10.5281/zenodo.6607647.

*2.5.7 Data availability:*

The datasets supporting the conclusions of this article are publicly available. Sequencing data information can all be found in **Appendix B** and are available in NCBI under BioProject PRJNA576070. The reads sequenced at JGI are also available on JGI/M ER under Gold ID Gs0114663 alongside their respective JGI Assembly Pipeline. MAG accession numbers and quality information can all be found in **Appendix B** and are deposited under Biosamples SAMN18867633-SAMN18867734. The accession numbers and quality for 111 vMAGs can be found in **Appendix B**, and they are deposited in NCBI under the BioProject ID PRJNA576070.

The raw annotations for each MAG are deposited on Zenodo with the following DOI: https://doi.org/10.5281/zenodo.5128772, and the interactive heatmap for them can be found here: https://zenodo.org/record/5124964. Additionally, the dataset of freshwater viruses used to cluster to the HUM-V vMAGs is provided on Zenodo with DOI https://doi.org/10.5281/zenodo.6310084. Metaproteomics data are deposited in the MassIVE database under accession MSV000087330. Metabolomics data are publicly available and deposited in Zenodo doi https://doi.org/10.5281/zenodo.5076253. Phylogenetic trees are provided in Zenodo here: https://doi.org/10.5281/zenodo.6339808. GTDB-Tk phylogenetic analyses output are provided in Zenodo here: https://doi.org/10.5281/zenodo.6502149. The unbinned metaproteomic mapping annotation data is hosted on Zenodo here: https://doi.org/10.5281/zenodo.6607647. All scripts along with the input files used in this manuscript are available at https://github.com/WrightonLabCSU/columbia_river.

*2.5.8 Acknowledgements:*

This work was supported by the Subsurface Biogeochemical Research (SBR) program (DE-SC0018170) and the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Environmental System Science (ESS) program through subcontract from the River Corridor Scientific Focus Area project at Pacific Northwest National Laboratory. These analyses were also supported by technology developed in the Wrighton laboratory supported by the National Sciences Foundation Division of Biological Infrastructure (1759874) and Department of Energy Office of Biological & Environmental Research (DE-SC0021350).

J.R.R. was partially supported by the National Science Foundation (NRT-DESE) [1450032], a Trans-Disciplinary Graduate Training Program in Biosensing and Computational

Biology at Colorado State University. B.B.M. and K.C.W. were supported by NSF early career award to Wrighton (1912915). The NMR data, FTICR-MS data and MS-proteomics data in this work was collected using instrumentation in the Environmental Molecular Science Laboratory (grid.436923.9), a DOE Office of Science User Facility sponsored by the Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory. Pacific Northwest National Lab is operated by Battelle for the DOE under Contract DE-AC05-76RL01830.

The work (proposal: 10.46936/jejc.proj.2014.48473/60005497) conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science in the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. A portion of the metagenomic sequencing for this research was performed at the Genomics Shared Resource Core at The Ohio State University Comprehensive Cancer Center supported by P30 CA016058.

The authors would like to thank Tyson Claffey and Richard Wolfe for Colorado State University server management; Sandy Shew for management of computing resources retained from The Ohio State University Unity cluster; Dr. Pearlly Yan at the Genomics Shared Resource Core at The Ohio State University Comprehensive Cancer Center for management of metagenomic sequencing; and Dr. J John for the continuous support. The authors declare they have no competing interests. J.R.R. and M.A.B. contributed equally to this work.

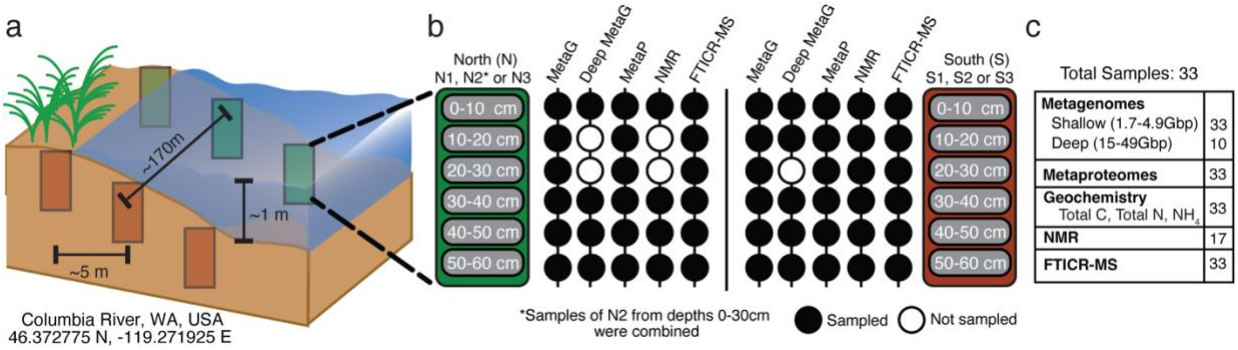



**Figure 2.1**. Overview of hyporheic zone sampling and the microbial genomes included in the HUM-V database a) Samples were collected from two transects, each with 3 sediment cores, with each core sectioned into 10-centimeter segments from 0-60 centimeters in depth and paired with metaproteomics and geochemistry. b) Schematic of the data types available for each of the depth samples within a core. Black-filled circles indicate depth samples for a particular data type, open circles denote missing analyses (due to limited sample availability), and black-filled circles with green outlines indicate new sequencing that was performed as part of this project and not available in Graham et al. c) Summarized catalog of the total samples for each analysis. d) The total recovered genomes (# of MAGs), taxonomic string, inferred genome completion (Comp., %), and contamination (Cont., %) for the dereplicated microbial genomes retained in HUM-V. Asterisks on names indicate uncultivated lineages.

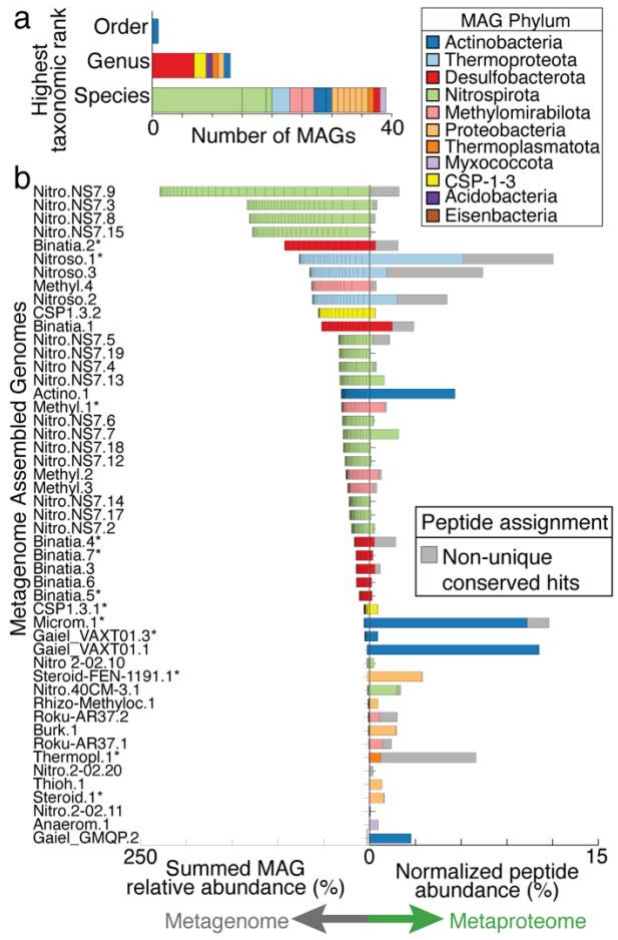



**Figure 2.2** Bacterial and archaeal members of HUM-V database coupled to metaproteomics reveals active members in the hyporheic zone microbiome. a) Stacked bar graph indicates the taxonomic novelty of the de-replicated MAGs colored by phylum and stacked according to the first empty position within the taxonomic string provided by the Genome Taxonomy Database GTDB-Tk. Each color represents a MAG phylum according to the MAG Phylum legend, a coloring maintained across this manuscript. b) Butterfly plot reports the summed genomic relative abundance across all samples (left side) and the normalized mean proteomic relative abundance (right side) for dereplicated MAGs (55 total, 49 shown), with bars colored by phylum. MAGs that contain a partial or complete 16S rRNA sequence are denoted with and asterisk (*). Non-unique specialized peptide assignment is defined in the methods and is shown with grey bars.

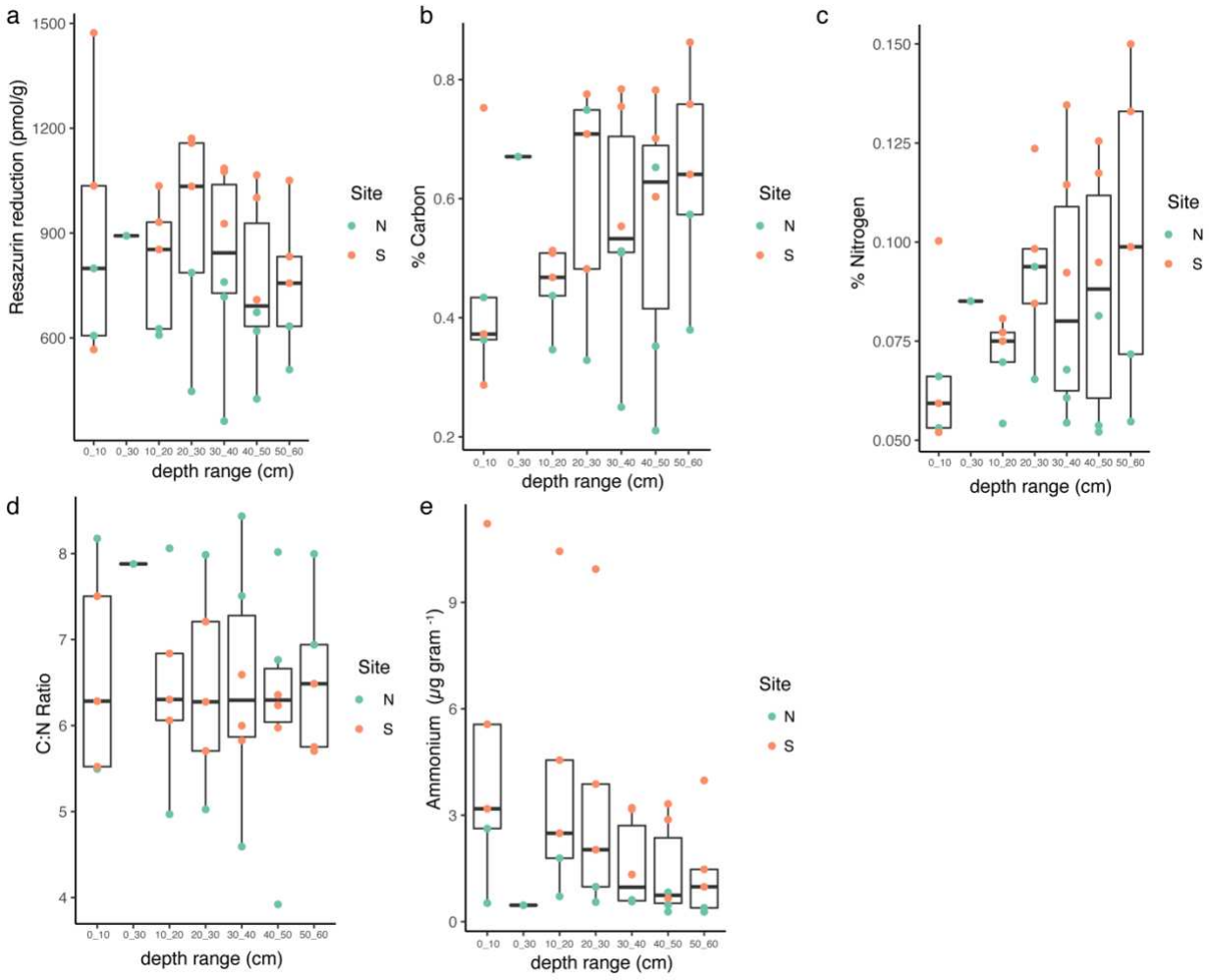



**Figure 2.3** Subset of geochemical measurements used to contextualize the multi-omics data, including the resazurin reduction assay (a), percent carbon (b), percent nitrogen (c), C:N ratio (d), and ammonium (e). Box and whisker plots ($n = 33$ samples) with points colored by sample transect and each box representing the sample depths spanned (0 to 60 cm). No significant differences were observed by site and depth, except for ammonium ($P = 0.026$), which was significantly different by bulk depths (i.e., surface = 0 to 30 cm, and deep = 30 to 60 cm).

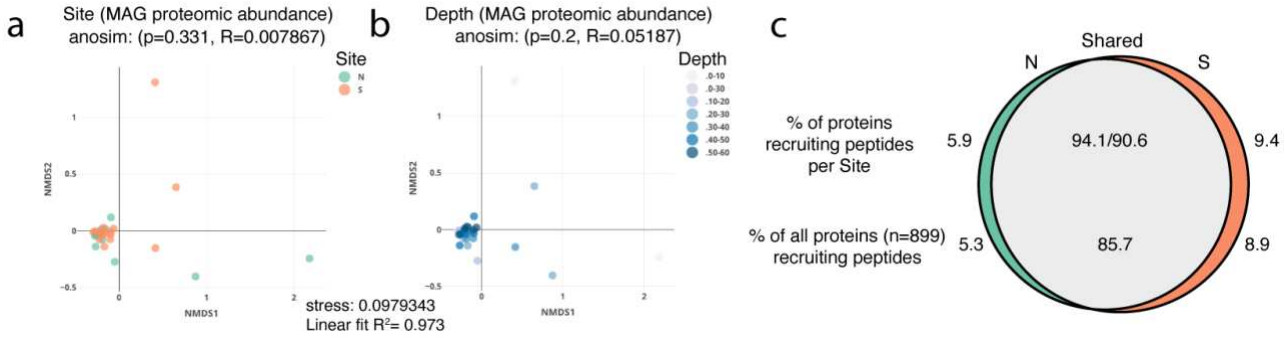



**Fig 2.4** NMDS and Venn diagram of differences in transect (N/S) or depth (0 to 60 cm) for MAG proteome. (a) NMDS ($n = 29$) of north (green) and south (orange) transects of recruited proteome peptides for MAGs. Analysis of similarity (ANOSIM) is reported. (b) NMDS of sediment core depth (0 to 60 cm) of recruited proteome peptides for MAGs. ANOSIM is reported. (c) Euler diagram showing the number of total proteins recruiting peptides in each transect, where the overlap represents the proportion of these proteins recruiting peptides in both transects. Only proteins recruiting two or more total peptides were included ($n = 898$) to allow for recruitment of at least one peptide for a given protein to both transects or both depths and to reduce false positives.

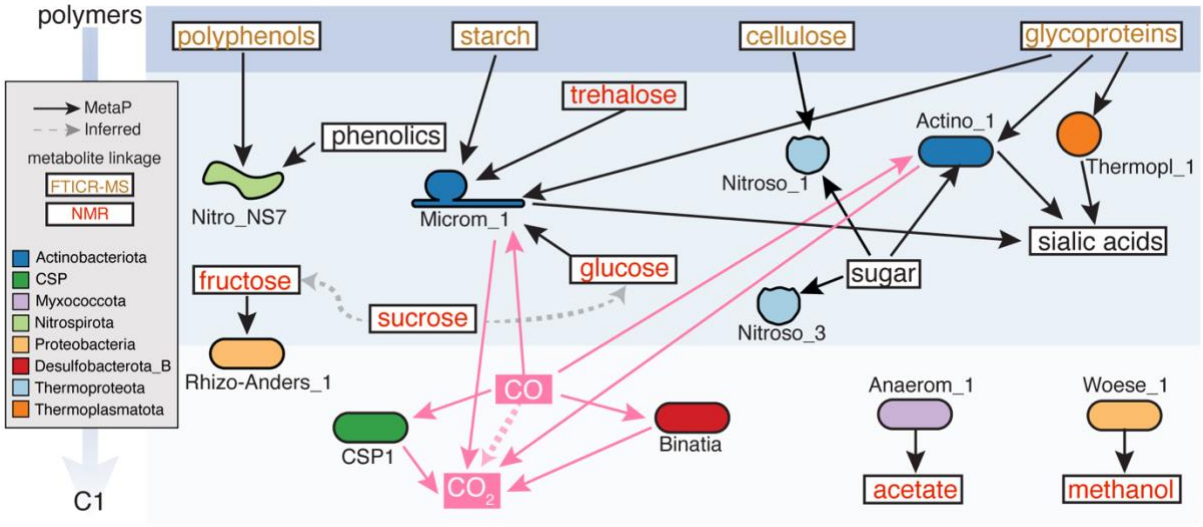



**Figure 2.5** Metaproteomics and metabolomics reveal microbial metabolic handoffs that support carbon cycling in river sediments. Detected metabolites are given in boxes, with NMR-detected compounds listed in red, polymers from FTICR-MS in orange, undetected metabolites in black. These polymers were inferred from FTICR-MS assigned biochemical classes and the specificity of CAZymes detected in metaproteome, where starch and cellulose were within the "polysaccharide-like" class and glycoproteins were in the "amino sugar-like" class. MAG-resolved metaproteome information is indicated by solid arrows, with MAG shape colored by phylum. Red arrows indicate processes leading to $CO_2$ production, while black arrows indicate other microbial carbon transforming genes expressed in the proteome. Shaded bold arrows indicate chemical connections, where (1) grey indicates a metabolite was detected along with putative downstream products (e.g., sucrose conversion to glucose) but metaproteomic lacked evidence for the transformation or (2) red indicates a metabolite not measured but metaproteomic evidence supported transformation (e.g., CO conversion to $CO_2$).

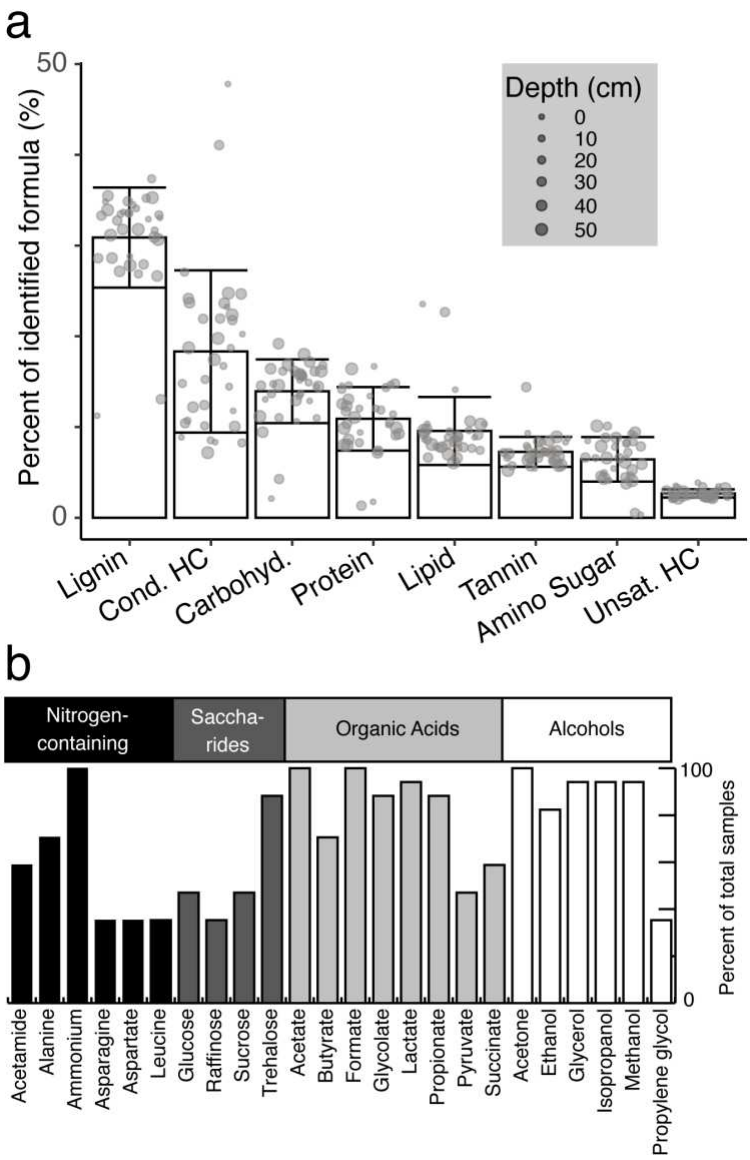



**Figure 2.6** Percentage of identified formula classes by FTICR-MS and prevalence of NMR metabolites across samples. (a) The percentage of identified formula classes identified in FTICR-MS across samples is not statistically structured by depth or site. Bar plots show the average percentage of identified formulas for each class calculated as the number of formulas identified per class divided by the total number of formulas identified per sample, with error bars representing 1 standard deviation (n = 33). Classes were determined using criteria reported previously (E. B. Graham, A. R. Crump, D. W. Kennedy, E. Arntzen, et al., Sci Total Environ 642:742–753, 2018, https://doi.org/10.1016/j.scitotenv.2018.05.256). Individual data points are plotted for each sample, with point size increasing with depth. Peaks were classified as described in Materials and Methods. Raw, full data are provided in Table S1. (b) Bar graphs showing specific NMR metabolites and ammonium with the percentage of total samples in which they were found. Shading denotes the different categories of detected compounds. For detailed information on how these were collected, see Materials and Methods. Only metabolites detected in more than four samples are shown.

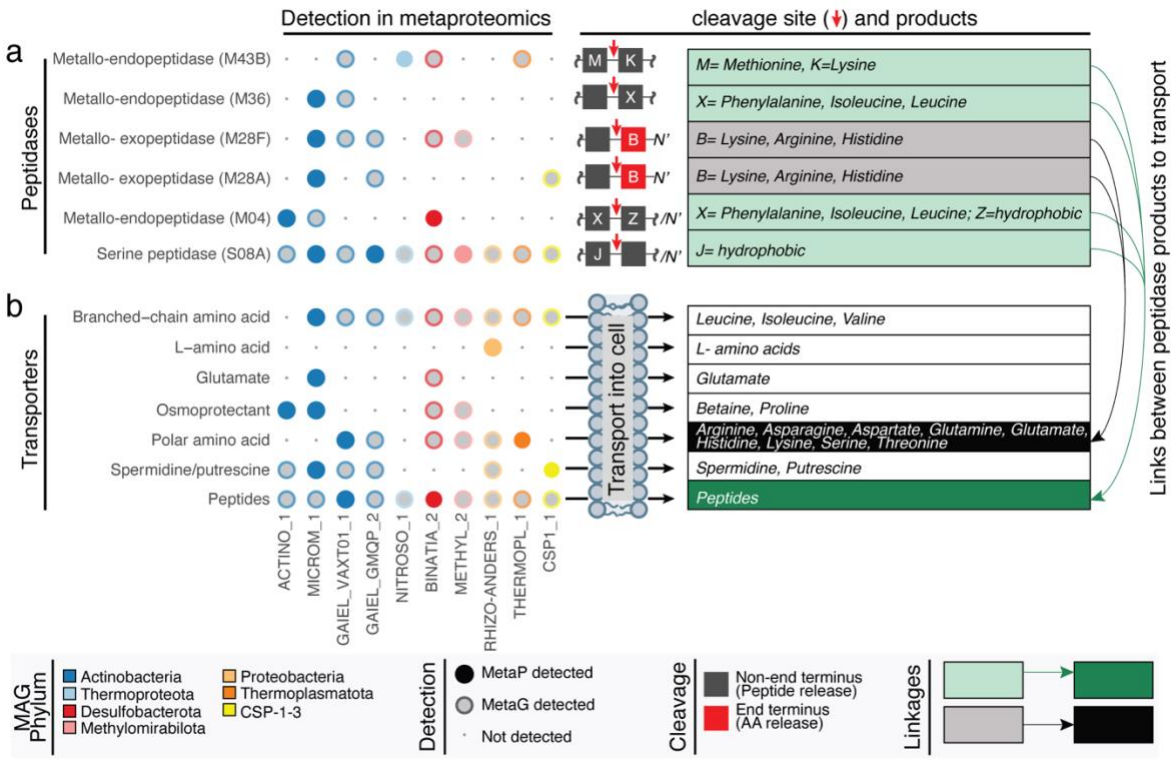



**Figure 2.7** Organic nitrogen mineralization and cellular transport are active microbial processes in river sediments. Bubble plots indicate the expressed genes that were uniquely assigned to specific genome including (a) extracellular peptidases and (b) cellular transporters for organic nitrogen. Unique peptides detected in at least 3 samples are reported as bubbles and colored by phylum. Table on the right shows putative amino acids cleaved or transported by respective peptidases or transporters, shades of color (green or grey) denote peptides that are cleaved into amino acids that could be transported, providing linkages between extracellular organic nitrogen transformation and transport of nitrogen into the cell. White boxes indicate an organic nitrogen transporter that recruited peptides but could not be linked to outputs of specific peptidases.

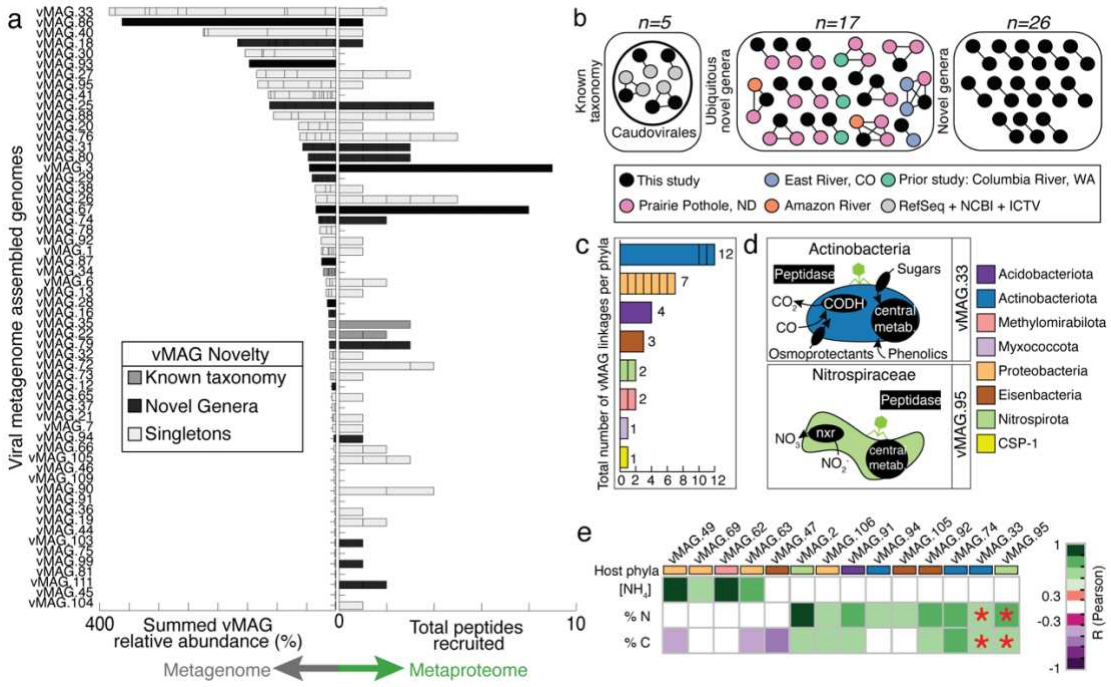



**Figure 2.8** Viruses in HUM-V are active, taxonomically novel, and can play key roles in microbial host metabolism and river geochemistry a) Butterfly plot showing summed genomic relative abundance (left side) and total peptides recruited for each vMAG population (total 111, 58 shown). Bars are colored by clustering of vMAGs from this study with (i) viruses of known taxonomy in RefSeq, ICTV and NCBI Taxonomy (dark grey), (ii) novel genera, both only from this study and ubiquitous (black), and no clustering from any database (light grey, singletons). b) Similarity network of the few vMAGs from our study (black) that clustered to viruses belonging to the default RefSeq, ICTV and NCBI Taxonomy databases (gray), as well as clustering of our vMAGs to other freshwater, publicly available dataset we mined (Pink, Purple, Orange, and Turquoise). The remaining clusters of viruses that were novel (e.g., did not cluster with prior viral genomes) are shown, with the full network file including singletons shown in Table S5. c) Stacked bar chart of the total number of vMAGs (n=32) that have putative host linkages. Each bar represents a phylum and lines within bars indicate the linkages for specific genomes within each phylum. For example, there are three genomes within the Actinobacteriota phylum that collectively have 12 viral linkages and of the three genomes that have linkages, one host has 10 viruses linked, while the other two hosts have 1 virus linked. d) Genome cartoons of microbial metabolisms for two representative genomes that could be predated by vMAGs, with the genes shown in black text boxes denoting processes detected in proteomics. These two microorganisms were selected as examples because they were active members in shaping carbon and nitrogen metabolism in these river sediments but could be impacted by viral predation; other virus-host relationships are reported in Supplementary Figure 9. e) Heatmap reports correlations between a subset of vMAGs with rectangle colors denoting the putative phyla for the respective host. Correlations between these vMAGs and ecosystem geochemistry ($NH_4$ µg/gram, %N, %C) are reported with significant correlation coefficients denoted by purple-green shading according to the legend. Red asterisks (*) indicate the vMAG relative abundance predicted a key environmental variable by sparse partial least squares (sPLS) regression. Note two of these predicted vMAGs are shown in (d).

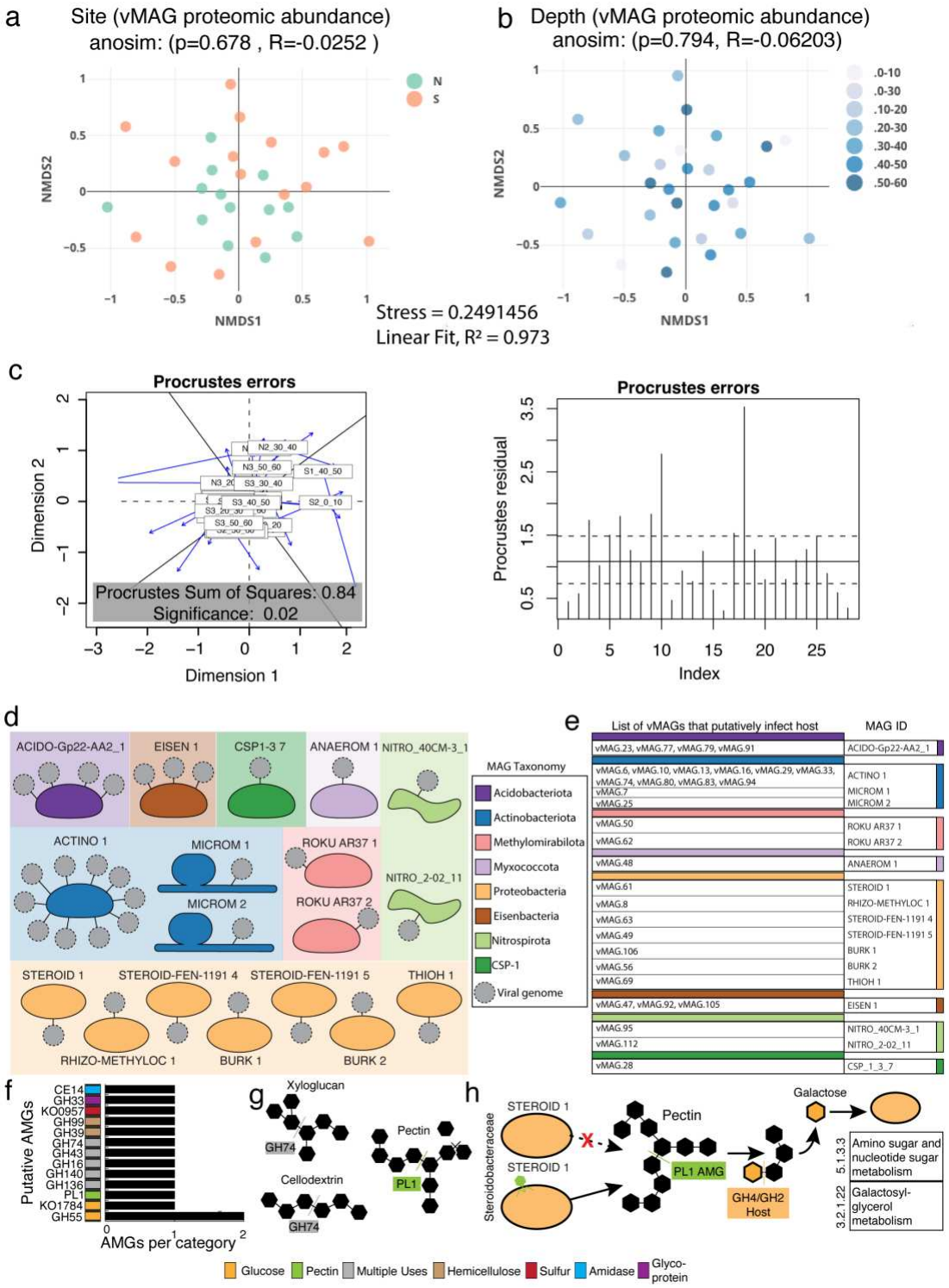



**Figure 2.9** vMAG recruited peptides combined with virus-host linkages indicate possible impacts on river sediment carbon and nitrogen cycling. (a and b) NMDS (n = 29) of north (green) and south (orange) transects (a) or sediment core depth (0 to 60 cm) (b) of vMAG recruited peptides show no significant clustering, with stress and ANOSIM reported. (c) Procrustes ordination of MAG and vMAG NMDS ordinations with relative abundance data across the 28 shallow sequenced samples that recruited reads to our MAG and vMAG communities. (d) Gray

dotted circles represent the number of significant vMAGs that putatively infect each MAG, and MAG cartoons are colored based on phylum. (e) Assignment of vMAGs to their putatively infected host MAGs. Colors match host MAG phylum assignment. (f) Different AMGs identified in vMAGs with putative hosts, with the predicted AMG substrate type denoted by colored boxes on the y axis. (g) Inferred cleavage activities for glycoside hydrolase (GH74) and a polysaccharide lyase (PL1) AMG are shown. (h) A conceptual model for how viral AMGs could impact the metabolism of their hosts is shown. One putative bacterial host, Steroidobacteraceae, has a linkage to a vMAG (vMAG.63) that encodes a PL1. Integration of this vMAG into the host MAG would provide the capacity to cleave the pectin backbone (PL1), a functionality not encoded in the host genome. This would create a new resource for the host which does encode genes for the oligo cleavage but not for backbone cleavage of pectin (via GH4 and GH2). This cleavage would result in galactose monomers that can further support host cell-encoded metabolism.

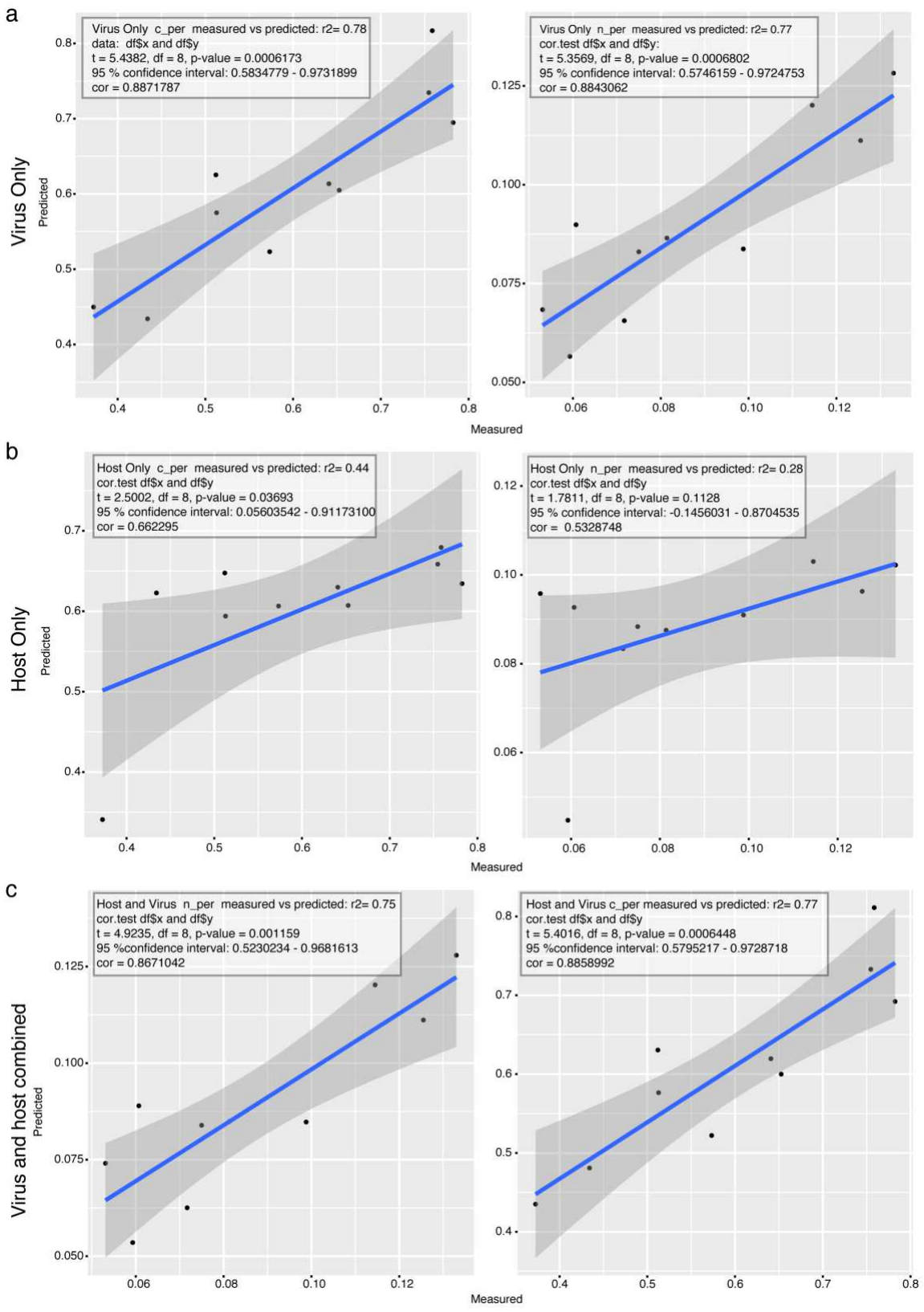

a
Virus Only c_per measured vs predicted: r2= 0.78
data: df$x and df$y
t = 5.4382, df = 8, p-value = 0.0006173
95 % confidence interval: 0.5834779 - 0.9731899
cor = 0.8871787
Virus Only n_per measured vs predicted: r2= 0.77
cor.test df$x and df$y:
t = 5.3569, df = 8, p-value = 0.0006802
95 % confidence interval: 0.5746159 - 0.9724753
cor = 0.8843062
Virus Only
Predicted
Measured
b
Host Only c_per measured vs predicted: r2= 0.44
cor.test df$x and df$y
t = 2.5002, df = 8, p-value = 0.03693
95 % confidence interval: 0.05603542 - 0.91173100
cor = 0.662295
Host Only n_per measured vs predicted: r2= 0.28
cor.test df$x and df$y
t = 1.7811, df = 8, p-value = 0.1128
95 % confidence interval: -0.1456031 - 0.8704535
cor = 0.5328748
Host Only
Predicted
Measured
c
Host and Virus n_per measured vs predicted: r2= 0.75
cor.test df$x and df$y
t = 4.9235, df = 8, p-value = 0.001159
95 %confidence interval: 0.5230234 - 0.9681613
cor = 0.8671042
Host and Virus c_per measured vs predicted: r2= 0.77
cor.test df$x and df$y
t = 5.4016, df = 8, p-value = 0.0006448
95 %confidence interval: 0.5795217 - 0.9728718
cor = 0.8858992
Virus and host combined
Predicted
Measured

**Figure 2.10** sPLS regressions show vMAG relative abundance-enhanced carbon and nitrogen predictions relative to those of MAG relative abundance alone. The predicted versus measured values for percent carbon (c_per) and percent nitrogen (n_per) with vMAGs only (a), MAGs only (b), and vMAGs and MAGs combined (c) are shown. Values corresponding to t, degrees of freedom, P values, confidence intervals, and correlation are also shown for each plot.

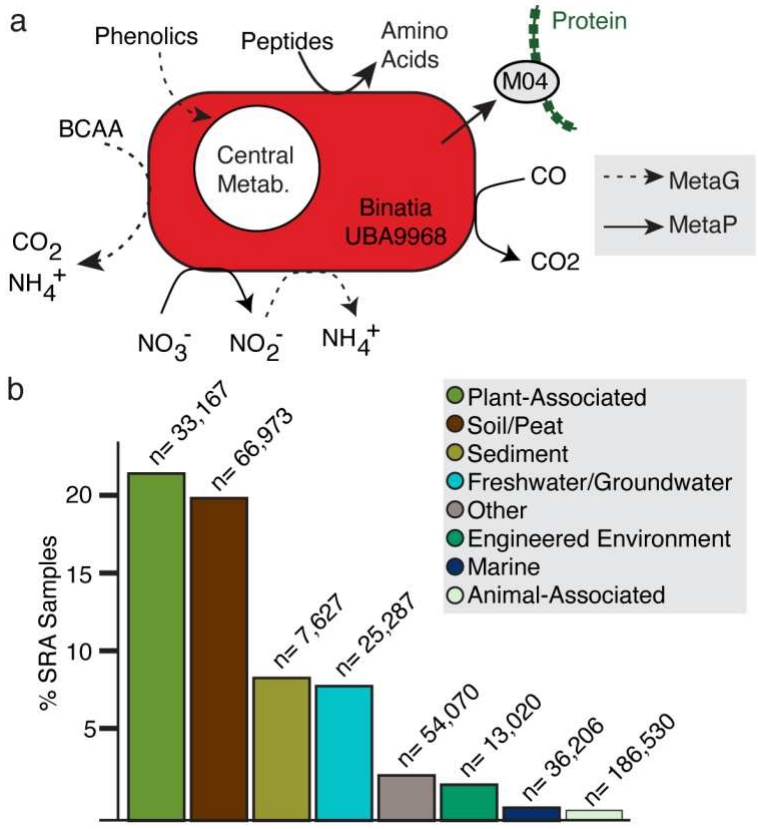



**Figure 2.11** Uncultured Binatia are widely dispersed across ecosystems and express nitrogen and carbon cycling genes in situ. (a) Metabolic MAG cartoon for the major functions encoded by the Binatia MAGs. Dotted arrows indicate functions encoded in the metagenome, and solid arrows correspond to metaproteome-detected enzymes. BCAA, branched-chain amino acids. (b) Using the 16S rRNA gene (from Binatia_7), we inventoried the distribution of closely related species to our HUM-V MAGs (>97% similarity) in the Sequence Read Archive (SRA) samples, uncovering the ecological distribution of these microorganisms from soils, as well as a wide variety of terrestrial, terrestrial-aquatic, and marine samples, indicating that the expressed metabolisms in panel a are likely applicable to a wide range of ecosystems.

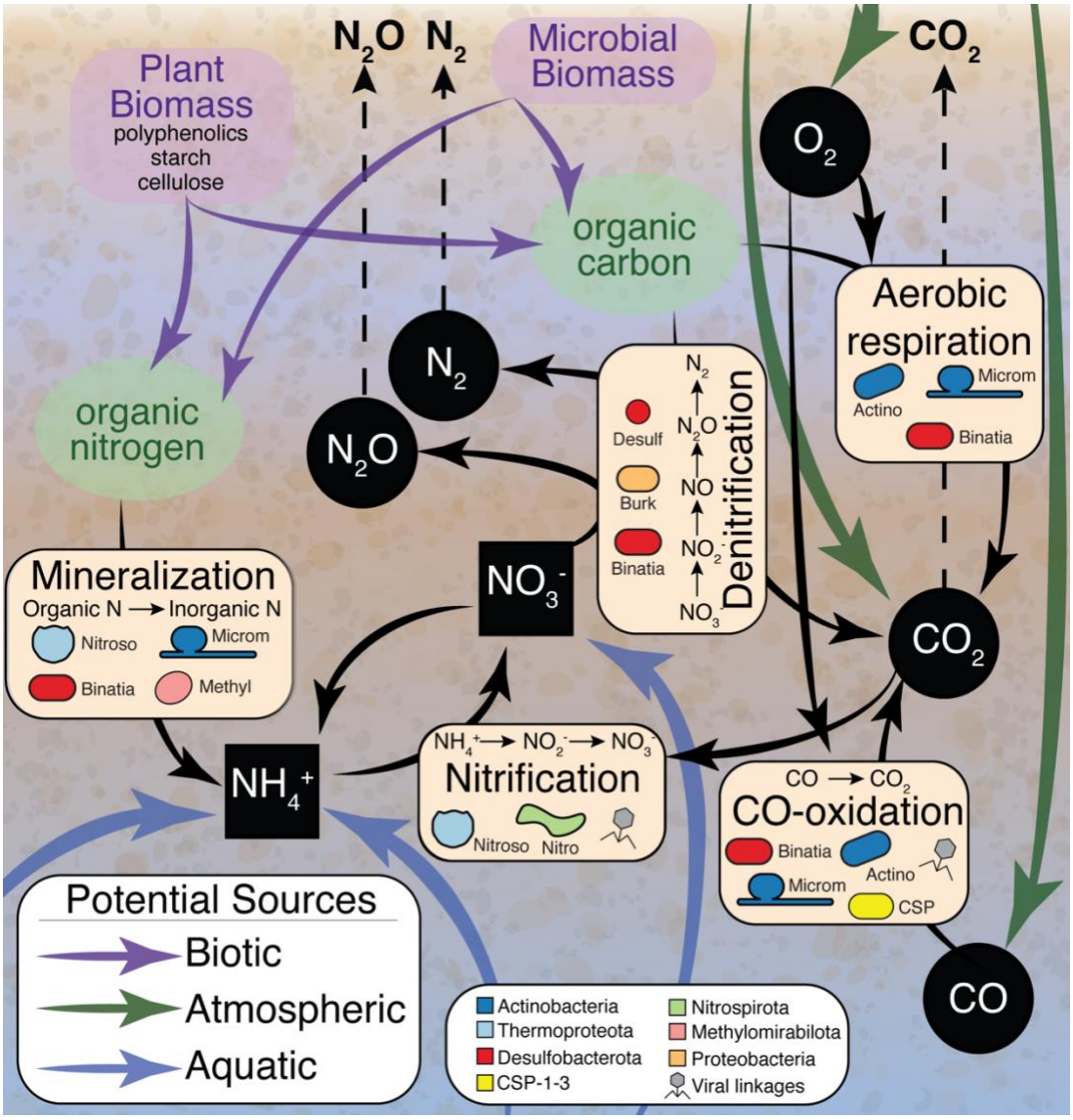



**Figure 2.12** Conceptual model uncovering microbes and processes contributing to carbon and nitrogen cycling in river sediments. Integration of multi-omic data uncovered the microbial and viral effects on carbon and nitrogen cycling in river sediments. Black arrows signify microbial transformations uncovered in our metaproteomic data. Specific processes (e.g., mineralization, nitrification, CO-oxidation, denitrification, and aerobic respiration) are highlighted in beige boxes, with microorganisms inferred to carry out the specific process denoted by overlaid cell shapes colored by phylum. Prior to this research, little was known about the specific enzymes and organisms responsible for river organic nitrogen mineralization and CO-oxidation, thus this research adds new content to microbial roles in carbon and nitrogen transformations in these systems. Possible (biotic, atmospheric, and aquatic) carbon and nitrogen sources are shown by purple, green, and blue arrows respectively. Inorganic carbon and nitrogen sources are shown by black squares (aqueous) and black circles (gaseous) with white text and dashed arrows indicating possible gasses that could be released to the atmosphere. Processes that could be impacted by viruses are marked with grey viral symbols.

Chapter 2 References


1. Boulton, A. J., Findlay, S., Marmonier, P., Stanley, E. H. & Valett, H. M. THE FUNCTIONAL SIGNIFICANCE OF THE HYPORHEIC ZONE IN STREAMS AND RIVERS. *Annu. Rev. Ecol. Syst.* **29**, 59–81 (1998).
2. Stegen, J. C. *et al.* Influences of organic carbon speciation on hyporheic corridor biogeochemistry and microbial ecology. *Nat. Commun.* **9**, 585 (2018).
3. Newcomer, M. E. *et al.* Influence of hydrological perturbations and riverbed sediment characteristics on hyporheic zone respiration of CO2 and N2. *Journal of Geophysical Research: Biogeosciences* **123**, 902–922 (2018).
4. Hu, M., Chen, D. & Dahlgren, R. A. Modeling nitrous oxide emission from rivers: a global assessment. *Glob. Chang. Biol.* **22**, 3566–3582 (2016).
5. Raymond, P. A. *et al.* Global carbon dioxide emissions from inland waters. *Nature* **503**, 355–359 (2013).
6. Gómez-Gener, L. *et al.* Global carbon dioxide efflux from rivers enhanced by high nocturnal emissions. *Nat. Geosci.* **14**, 289–294 (2021).
7. Lewandowski, J. *et al.* Is the Hyporheic Zone Relevant beyond the Scientific Community? *Water* **11**, 2230 (2019).
8. Raes, E. J. *et al.* Can We Use Functional Genetics to Predict the Fate of Nitrogen in Estuaries? *Front. Microbiol.* **11**, 1261 (2020).
9. Hou, Z. *et al.* Geochemical and Microbial Community Attributes in Relation to Hyporheic Zone Geological Facies. *Sci. Rep.* **7**, 12006 (2017).
10. Nelson, A. R. *et al.* Heterogeneity in Hyporheic Flow, Pore Water Chemistry, and Microbial Community Composition in an Alpine Streambed. *J. Geophys. Res. Biogeosci.* **124**, 3465–3478 (2019).
11. Sackett, J. D. *et al.* Microbial Community Structure and Metabolic Potential of the Hyporheic Zone of a Large Mid-Stream Channel Bar. *Geomicrobiol. J.* **36**, 765–776 (2019).
12. Sun, S., Jones, R. B. & Fodor, A. A. Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories. *Microbiome* **8**, 46 (2020).
13. Djemiel, C. *et al.* Inferring microbiota functions from taxonomic genes: a review. *Gigascience* **11**, (2022).
14. Thornton, P. E., Lamarque, J.-F., Rosenbloom, N. A. & Mahowald, N. M. Influence of carbon-nitrogen cycle coupling on land model response to CO2fertilization and climate variability. *Global Biogeochem. Cycles* **21**, (2007).
15. Weitz, J. S. & Wilhelm, S. W. Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 Biol. Rep.* **4**, 17 (2012).
16. Peduzzi, P. Virus ecology of fluvial systems: a blank spot on the map? *Biol. Rev. Camb. Philos. Soc.* **91**, 937–949 (2016).
17. Pollard, P. C. & Ducklow, H. Ultrahigh bacterial production in a eutrophic subtropical Australian river: Does viral lysis short-circuit the microbial loop? *Limnol. Oceanogr.* **56**, 1115–1129 (2011).
18. Ma, L., Sun, R., Mao, G., Yu, H. & Wang, Y. Seasonal and spatial variability of virioplanktonic abundance in Haihe River, China. *Biomed Res. Int.* **2013**, 526362 (2013).
19. Samson, R. *et al.* Metagenomic insights to understand transient influence of Yamuna River on taxonomic and functional aspects of bacterial and archaeal communities of River Ganges. *Sci. Total Environ.* **674**, 288–299 (2019).

20. Huber, D. H. *et al.* Metagenome Sequences of Sediment from a Recovering Industrialized Appalachian River in West Virginia. *Genome Announc.* **6**, (2018).
21. Liu, S. *et al.* Comammox Nitrospira within the Yangtze River continuum: community, biogeography, and ecological drivers. *ISME J.* (2020) doi:10.1038/s41396-020-0701-8.
22. Black, E. M. & Just, C. L. The Genomic Potentials of NOB and Comammox Nitrospira in River Sediment Are Impacted by Native Freshwater Mussels. *Front. Microbiol.* **9**, 2061 (2018).
23. Graham, E. B. *et al.* Multi'omics comparison reveals metabolome biochemistry, not microbiome composition or gene expression, corresponds to elevated biogeochemical function in the hyporheic zone. *Science of the total environment* **642**, 742–753 (2018).
24. Arntzen, E. V., Geist, D. R. & Dresel, P. E. Effects of fluctuating river flow on groundwater/surface water mixing in the hyporheic zone of a regulated, large cobble bed river. *River Res. Appl.* **22**, 937–946 (2006).
25. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
26. Vosloo, S. *et al.* Evaluating de Novo Assembly and Binning Strategies for Time Series Drinking Water Metagenomes. *Microbiol Spectr* **9**, e0143421 (2021).
27. Haggerty, R., Martí, E., Argerich, A., von Schiller, D. & Grimm, N. B. Resazurin as a "smart" tracer for quantifying metabolically active transient storage in stream ecosystems. *J. Geophys. Res.* **114**, (2009).
28. Yang, F. *et al.* Characterization of Purified c-Type Heme-Containing Peptides and Identification of c-Type Heme-Attachment Sites in Shewanella o neidenis Cytochromes Using Mass Spectrometry. *Journal of proteome research* **4**, 846–854 (2005).
29. Cordero, P. R. F. *et al.* Atmospheric carbon monoxide oxidation is a widespread mechanism supporting microbial survival. *ISME J.* **13**, 2868–2881 (2019).
30. Zafiriou, O. C., Andrews, S. S. & Wang, W. Concordant estimates of oceanic carbon monoxide source and sink processes in the Pacific yield a balanced global "blue-water" CO budget. *Global Biogeochem. Cycles* **17**, (2003).
31. Haszpra, L., Ferenczi, Z. & Barcza, Z. Estimation of greenhouse gas emission factors based on observed covariance of CO2, CH4, N2O and CO mole fractions. *Environmental Sciences Europe* **31**, 1–12 (2019).
32. Xia, X. *et al.* The cycle of nitrogen in river systems: sources, transformation, and flux. *Environ. Sci. Process. Impacts* **20**, 863–891 (2018).
33. Strauss, E. A. & Lamberti, G. A. Regulation of nitrification in aquatic sediments by organic carbon. *Limnol. Oceanogr.* **45**, 1854–1859 (2000).
34. Brust, G. E. Chapter 9 - Management Strategies for Organic Vegetable Fertility. in *Safety and Practice for Organic Food* (eds. Biswas, D. & Micallef, S. A.) 193–212 (Academic Press, 2019).
35. Verhagen, F. J. & Laanbroek, H. J. Competition for Ammonium between Nitrifying and Heterotrophic Bacteria in Dual Energy-Limited Chemostats. *Appl. Environ. Microbiol.* **57**, 3255–3263 (1991).
36. Yu, N. Y. *et al.* PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **26**, 1608–1615 (2010).

37. Bendtsen, J. D., Kiemer, L., Fausbøll, A. & Brunak, S. Non-classical protein secretion in bacteria. *BMC Microbiol.* **5**, 58 (2005).
38. Armenteros, J. J. A. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature biotechnology* **37**, 420–423 (2019).
39. Pollak, S. *et al.* Public good exploitation in natural bacterioplankton communities. *Sci Adv* **7**, (2021).
40. Allison, S. D. Cheaters, diffusion and nutrients constrain decomposition by microbial enzymes in spatially structured environments. *Ecol. Lett.* **8**, 626–635 (2005).
41. Smith, P. & Schuster, M. Public goods and cheating in microbes. *Curr. Biol.* **29**, R442–R447 (2019).
42. Strauss, E. A. *et al.* Nitrification in the Upper Mississippi River: patterns, controls, and contribution to the NO3- budget. *Journal of the North American Benthological Society* **23**, 1–14 (2004).
43. Stegen, J. C. *et al.* Groundwater-surface water mixing shifts ecological assembly processes and stimulates organic carbon turnover. *Nat. Commun.* **7**, 11237 (2016).
44. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 13219 (2016).
45. Jones, C. M. *et al.* Recently identified microbial guild mediates soil N2O sink capacity. *Nat. Clim. Chang.* **4**, 801–805 (2014).
46. Conthe, M. *et al.* Life on N2O: deciphering the ecophysiology of N2O respiring bacterial communities in a continuous culture. *ISME J.* **12**, 1142–1153 (2018).
47. Hallin, S., Philippot, L., Löffler, F. E., Sanford, R. A. & Jones, C. M. Genomics and Ecology of Novel N2O-Reducing Microorganisms. *Trends Microbiol.* **26**, 43–55 (2018).
48. Orellana, L. H. *et al.* Detecting nitrous oxide reductase (NosZ) genes in soil metagenomes: method development and implications for the nitrogen cycle. *MBio* **5**, e01193-14 (2014).
49. Moon, K. *et al.* Freshwater viral metagenome reveals novel and functional phage-borne antibiotic resistance genes. *Microbiome* **8**, 75 (2020).
50. Wolf, Y. I. *et al.* Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nat Microbiol* (2020) doi:10.1038/s41564-020-0755-4.
51. Püttker, S. *et al.* Metaproteomics of activated sludge from a wastewater treatment plant--A pilot study. *Proteomics* **15**, 3596–3601 (2015).
52. Rudney, J. D., Xie, H., Rhodus, N. L., Ondrey, F. G. & Griffin, T. J. A metaproteomic analysis of the human salivary microbiota by three-dimensional peptide fractionation and tandem mass spectrometry. *Molecular oral microbiology* **25**, 38–49 (2010).
53. Solden, L. M. *et al.* Interspecies cross-feeding orchestrates carbon degradation in the rumen ecosystem. *Nature Microbiology* **3**, 1274–1284 (2018).
54. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).
55. Murphy, C. L. *et al.* Genomic Analysis of the Yet-Uncultured Binatota Reveals Broad Methylotrophic, Alkane-Degradation, and Pigment Production Capacities. *Mbio* **12**, (2021).
56. Moser, D. P. *et al.* Biogeochemical processes and microbial characteristics across groundwater-surface water boundaries of the Hanford Reach of the Columbia River. *Environ. Sci. Technol.* **37**, 5127–5134 (2003).
57. Tfaily, M. M. *et al.* Advanced solvent based methods for molecular characterization of soil organic matter by high-resolution mass spectrometry. *Anal. Chem.* **87**, 5206–5215 (2015).

58. Tfaily, M. M. *et al.* Sequential extraction protocol for organic matter from soils and sediments using high resolution mass spectrometry. *Anal. Chim. Acta* **972**, 54–61 (2017).
59. RoyChowdhury, T. *et al.* Temporal dynamics of CO2 and CH4 loss potentials in response to rapid hydrological shifts in tidal freshwater wetland soils. *Ecol. Eng.* **114**, 104–114 (2018).
60. Daly, R. A. *et al.* Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales. *Nature Microbiology* **1**, 16146 (2016).
61. Borton, M. A. *et al.* Coupled laboratory and field investigations resolve microbial interactions that underpin persistence in hydraulically fractured shales. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E6585–E6594 (2018).
62. Weljie, A. M., Newton, J., Mercier, P., Carlson, E. & Slupsky, C. M. Targeted profiling: quantitative analysis of 1H NMR metabolomics data. *Anal. Chem.* **78**, 4430–4442 (2006).
63. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
64. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
65. Kang, D. *et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies*. https://peerj.com/preprints/27522/ (2019) doi:10.7287/peerj.preprints.27522v1.
66. Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* **28**, 1033–1034 (2012).
67. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
68. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
69. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
70. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2018).
71. Castelle, C. J. *et al.* Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nat. Commun.* **4**, 1–10 (2013).
72. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
73. Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4306.
74. Merchant, N. *et al.* The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLoS Biol.* **14**, e1002342 (2016).
75. Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
76. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).

77. Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A. & Sun, F. Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic acids research* **45**, 39–53 (2017).
78. Lê Cao, K.-A., Rossouw, D., Robert-Granié, C. & Besse, P. A sparse PLS for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology* **7**, (2008).
79. Nicora, C. D. *et al.* The MPLEx protocol for multi-omic analyses of soil samples. *Journal of visualized experiments: JoVE* (2018).
80. McGivern, B. B. *et al.* Decrypting bacterial polyphenol metabolism in an anoxic wetland soil. *Nat. Commun.* **12**, 1–16 (2021).
81. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for mass spectrometry-based proteomics. in *Proteome bioinformatics* 55–71 (Springer, 2010).
82. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).

# Chapter 3: 2018 a space (and time) odyssey: Spatial and temporal metagenomics of river compartments reveals viral community dynamics in an urban impacted stream [2]

## *3.1 Summary*

Although river ecosystems comprise less than 1% of Earth's total non-glaciated area, they are critical modulators of microbially and virally orchestrated global biogeochemical cycles. However, most studies either use data that is not spatially resolved or is collected at timepoints that do not reflect the short life cycles of microorganisms. As a result, the relevance of microbiome interactions and the impacts they have over time on biogeochemical cycles are poorly understood. To assess how viral and microbial communities change over time, we sampled surface water and pore water compartments of the wastewater-impacted River Erpe in Germany every 3 hours over a 48-hour period resulting in 32 metagenomes paired to geochemical and metabolite measurements. We reconstructed 6,500 viral and 1,033 microbial genomes and found distinct communities associated with each river compartment. We show that 17% of our vMAGs clustered to viruses from other ecosystems like wastewater treatment plants and rivers. Our results also indicated that 70% of the viral community was persistent in surface waters, whereas only 13% were persistent in the pore waters taken from the hyporheic zone. Finally, we predicted linkages between 73 viral genomes and 38 microbial genomes. These putatively linked hosts included members of the *Competibacteraceae,* which we suggest are potential contributors to carbon and nitrogen cycling. Together, these findings demonstrate that microbial and viral communities in

[2] This chapter was reproduced verbatim from "Rodríguez-Ramos et al. Spatial and temporal metagenomics of river compartments reveals viral community dynamics in an urban impacted stream. Frontiers in Microbiology (2023)". The text benefitted from writing and editing from contributing authors and is still pending one reviewer for final publication. Reviewer 1 comments were minor. The ordering of the materials in this dissertation are consistent with the content available online but have been renumbered to reflect incorporation into this dissertation.

surface waters of this urban river can exist as stable communities along a flowing river; and raise important considerations for ecosystem models attempting to constrain dynamics of river biogeochemical cycles.

### *3.2 Introduction*

Rivers are crucial modulators of global biogeochemical cycles and provide a dynamic, moving passageway between terrestrial and aquatic ecosystems [1]. Corresponding to ~7% of global carbon dioxide ($CO_2$) and ~5% of global methane ($CH_4$) emissions per year, rivers contribute up to 2,508 Tg $yr^{-1}$of $CO_2$, and ~30.5 Tg $yr^{-1}$ of $CH_4$ [2–5]. Within rivers, microbial communities are key orchestrators of carbon and nitrogen transformations, where they contribute between 40-90% of total river respiration [6–8]. Despite a general understanding of the importance of microbial metabolism, knowledge of river viral communities and their impacts on microbial communities remains scarce.

Viruses are the most abundant organism on the planet, with estimates of up to $10^{31}$ viral particles worldwide [9–12]. These viral predators are mostly studied in marine ecosystems, where viruses can lyse 20-40% of bacteria daily [13–17] and play key roles reprogramming their bacterial hosts with ecosystem-wide consequences [18–20]. Although research has mostly focused on marine ecosystems, recent efforts have been made to expand our knowledge of natural viral communities in freshwater aquatic environments like lakes [21,22] and estuaries [23,24]. Early studies in these systems have shown viral like particle (VLP) abundances and viral productivity (i.e., the number of viruses produced per hour) in rivers can be equivalent, or higher, than those in marine systems [25–28]. Additionally, early river studies found that up to 80% of bacterial isolate strains from sediments

had virulent phage that could be isolated [29]. Together, these foundational works highlight the importance of viral predation in regulating microbial dynamics in river ecosystems.

There are two key reasons why it remains difficult to link viral communities to river ecosystem function. First, river microbiome studies are rarely genome-resolved, both from a bacterial and viral perspective. While there is still much to explore, most information on aquatic virus dynamics pertains to oceanic studies [30], and rivers are described as one of the most underexplored aquatic ecosystem with metagenomics, second only to glacier microbiomes [31]. Although the taxonomic composition of microbial communities in rivers has been well-described by 16S rRNA gene amplicon surveys [32,33], it remains unclear how microbial membership relates to relevant ecosystem processes. Likewise, our ability to link the viral community to their respective microbial hosts, and subsequently to ecosystem biogeochemistry, remains hindered by a lack of genome-resolved studies. Second, river studies are often temporally constrained. Although significant changes in river chemistry and hydrology are observed at seasonal periods [34], they are also known to change at sub-daily scales [35,36], particularly in human-impacted rivers affected by wastewater treatment plant effluent and reservoirs [37–39]. Additionally, microbial communities double, evolve, and shift metabolically at an hourly basis [40–42]. Nonetheless, river microbiome time-series are often resolved at seasonal scales [43,44], making our understanding of viral and microbial community dynamics across relevant short-scale gradients poorly understood.

To address these knowledge gaps, we collected a finely resolved metagenomic time-series at the River Erpe near Berlin, Germany, a lowland river receiving treated wastewater. Our sampling campaign included biogeochemical measurements every 3 hours for 48 hours across both surface water (SW) and pore water (PW) compartments that were paired to metagenomics and metabolomics **(Figure 3.1A-D)**. This study design provided a metagenomically resolved dataset

which enabled us to interrogate how viral and microbial communities are structured across river compartments, and how this metabolic potential could modulate biogeochemical processes. Additionally, the unparalleled temporal resolution of our dataset allowed us to analyze both the persistence of viral and microbial communities across compartments, as well as the individual genome stability throughout the 48 hours of sampling. Finally, by using genome-resolved metagenomics, we show that viruses can be linked to hosts in river ecosystems, and that these linkages reveal putative interactions that may be central to ecosystem biogeochemistry.

## *3.3 Methods*

### *3.3.1 Sample collection, DNA isolation, and chemical characterization*

The River Erpe is highly influenced by diurnally fluctuating effluent volumes of the Münchehofe wastewater treatment plant and consists of up to 80% treated wastewater [45]. Our sampling site is in a side channel with a mean discharge of 25 l/s [45,46] (**Figure 3.1A**). For sample collection, a sampling station was set up ~1m from the shoreline of the River Erpe side channel "Rechter Randgraben" (52.476416, 13.625710), 1.6km from the wastewater treatment plant outlet leading the same water as in the main channel as previously described [45], and in accordance to the Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS) protocol [47]. Briefly, for surface water (SW), 60ml at a time of SW were collected manually with a syringe and tubing fixed in the water column and then passed through a 0.20µm filter until clogged. A cap was then put on the filter, filled with 3ml RNAlater, and refrigerated until extraction. For pore water (PW), 60ml of PW from 25cm sediment depth were collected with a stainless-steel rod in the middle of the channel. The rods were covered with a filter mesh sock over the screened area at the tip, pushed into the sediment, and equipped with a Teflon suction line.

Samples were then taken by manually pulling 60ml of PW with syringes attached to the suction line and filtering them through a 0.20µm filter until clogged. The filter was then capped, filled with 3ml RNAlater, and refrigerated until extraction. Each of these processes were repeated every 3 hours over a period of 48hrs in September of 2018, resulting in 15 SW and 17 PW metagenomes. 2 SW samples failed due to lack of biomass. For DNA isolation, filters were cut into ~5mm$^2$ pieces and added to the bead bashing tubes of Quick-DNA Soil Microbe Microprep Kit (Zymo). The nucleic acids were then extracted according to the manufacturer protocol and sequenced at the Genomics Shared Resource Anschutz Medical Campus, Colorado. Accession numbers, total metagenomic reads, and sample sizes can be found on **Appendix C** and the original data repository [48].

Chemical characterization was performed as previously described [45]. Water samples were filtered with 0.2µm polyethersulfone Sterivex for Fourier transform ion cyclotron resonance mass spectrometer (FTICR-MS) analysis or regenerated cellulose for all other analytes, then acidified to a pH of 2 with 2M HCl and stored at -18$^o$C until analysis. Samples were analyzed at the Leibniz Institute of Freshwater Ecology and Inland Fisheries for nitrate and sulfate (ion chromatography, Metrohm 930 Compact IC Flex), ammonium and soluble reactive phosphorous (SRP) (segmented flow analyzer Skalar SAN, Skalar Analytical B.V., Netherlands), and manganese and iron (inductively coupled plasma optical emission spectrometry (ICP-OES), (ICP iCAP 6000 series, Thermo Fisher Scientific Inc.). Dissolved organic carbon (DOC) concentrations were analyzed via infrared gas analyzer (NDIR) after combustion (TOC/TN Analyzer, Shimadzu). Dissolved organic matter (DOM) data is part of the WHONDRS dataset [48] and was analyzed using a 12T Bruker SolariX FTICR-MS (Bruker, SolariX, Billerica, MA, USA) at the Environmental Molecular Sciences Laboratory in Richland, WA. Once peaks were picked using the Bruker data analysis

software and formulas were assigned using Formularity [49], DOM was classified into seven compound classes based upon hydrogen to carbon ratio (H:C), and oxygen to carbon (O:C) ratios [50]. FTICR-MS analysis does not allow for a quantitative approach, therefore compound class data was analyzed qualitatively, and DOM composition was evaluated using the number of molecular formulas in every compound class as described in the original publication [45]. The biogeochemical measurements for this study can all be found on **Appendix C**.

*3.3.2 Metagenome data processing and assembly*

Each set of metagenomic reads were trimmed using Sickle v1.33 with default settings [51], and assessed using FastQC (v0.11.2) [52]. Trimmed reads were then assembled with either 1) metaSPAdes BBCMS pipeline (v3.13.0) [53], 2) Megahit (v1.2.9) [54], or 3) IDBA UD (v.1.1.0) [55]. For metaSPAdes pipeline, reads were merged into a single .fa file using fq2fa [56]. Then, bbcms was run with flags "mincount = 2", and "highcountfraction = 0.6", followed by metaSPAdes using kmers 33, 55, 77, 99, 127, and flag "–meta". For Megahit, reads were assembled with flags "k-min = 31", "k-max = 121", "k-step = 10", and "m = 0.4". For IDBA_UD, samples were rarefied to 25% of reads using BBMAP's reformat.sh [57] with flags "samplerate = 0.25" and "sampleseed = 1234". These 25% of subset reads were then merged into a single .fa file using fq2fa [56] and then assembled with default parameters. Assembly statistics for each sample can be found in **Appendix C**.

*3.3.3 Viral identification, taxonomy, and annotations*

Viral metagenome assembled genomes (vMAGs) were identified from each set of assemblies using Virsorter2 and CheckV using the established protocols.io methods [58,59]. Resulting genomes were then screened based on VirSorter2 and checkV output for viral and host gene counts, VirSorter2 viral scores, and hallmark gene counts [59]. Viruses were then annotated with DRAM-v

using the "--use_uniref" flag, and further manually curated according to the established protocol [59,60]. The resulting subset of 6,500 viral genomes were clustered at 95% ANI across 85% of shortest contig per MIUViG standards [61] resulting in 1,230 viral populations.

Viral taxonomic identification of viral populations was performed using protein clustering methods with vContact2 using default methods [62]. We supplemented the standard RefSeq v211 database containing 4,533 vMAGs with viral genomes from an additional 303 river and wastewater treatment plant metagenomes that were publicly available from 1) JGI IMG/VR (6,254 vMAGs ≥10kb), 2) two previously unpublished anaerobic digestor metagenomic datasets that were mined in-house (14,436 vMAGs ≥10kb) (https://doi.org/10.5281/zenodo.7709817), 3) a previously published wastewater treatment plant sludge database (7,443 vMAGs ≥10kb) [63], 4) a previously available reference database that included freshwater ecosystem viruses (2,032 vMAGs ≥10kb) [8], and 5) the 43 TARA Oceans Virome datasets (5,476 vMAGs ≥10kb) [64]. This resulted in an additional 35,641 reference vMAGs in our network. Proteins file for all vMAGs used in the network as well as accession numbers are available on Zenodo (https://doi.org/10.5281/zenodo.7709817). Results from vContact2 can be found in **Appendix C.**

Viral population genome representatives were annotated using DRAM-v [60]. To identify putative auxiliary metabolic genes (AMGs), auxiliary scores were assigned by DRAM-v to each annotated gene based on the following previously described ranking system: A gene is given an auxiliary score of 1 if there is at least one hallmark gene on both the left and right flanks, indicating the gene is likely viral. An auxiliary score of 2 is assigned when the gene has a viral hallmark gene on one flank and a viral-like gene on the other flank. An auxiliary score of 3 is assigned to genes that have a viral-like gene on both flanks [8,60]. Genes identified by DRAM-v as being high-confidence possible AMGs (auxiliary scores 1-3) were subjected to protein modeling using Protein

Homology / AnalogY Recognition Engine (PHYRE2) [65], and manually verified. All files for vMAG quality and annotations can be found in **Appendix C.**

*3.3.4 Bacterial and archaeal metagenomic binning, quality control, annotation, and taxonomy*

Bacterial and archaeal genomes were binned from each set of assemblies with MetaBAT v2.12.1 [66] as previously described [8]. Briefly, reads were mapped to each respective assembly to get coverage information using BBmap [57], and then MetaBAT was run with default settings on each assembly after filtering for scaffolds ≥2,500bp. Quality for each MAG was then assessed using CheckM (v1.1.2) [67]. To ensure that only quality MAGs were utilized for analyses, we discarded all MAGs that were not medium quality (MQ) to high quality (HQ) according to MIMAG standards [68], resulting in 1,033 MAGs. These MAGs were dereplicated using dRep [69] at 95% identity, resulting in 125 MAGs. These 125 MQHQ MAGs were annotated using the DRAM pipeline [60] as previously described [8]. For taxonomic analyses, MAGs were classified using the Genome Taxonomy Database (GTDB) Toolkit v1.5.0 on November 2021 using the r202 database [70]. Genome quality, annotations, and taxonomy are reported in **Appendix C**.

*3.3.5 Virus host linkages*

To identify virus-host linkages, we used 1) CRASS (Direct Repeat / Spacer based) v1.0.1 [71], 2) VirHostMatcher (alignment-free oligonucleotide frequency based) v.1.0.0 [72], and 3) PHIST (all-versus-all exact matches based) v.1.0.0 [73]. CRASS protocol and scripts used are described in detail on GitHub (see **Data availability**). VirHostMatcher was run with default settings, and the best possible hit for each virus was considered only if it had a d2* dissimilarity score of < 0.2. PHIST was run with flag "-k = 25", and a PHIST hit was considered only if it had a significant adjusted p-value of < 0.05. To be classified as a virus-host linkage, a virus-host pair had to be

predicted by the significant consensus of both VirHostMatcher and PHIST or a virus-host pair had to have a CRASS linkage. With this consensus method, CRASS links, which were always considered good hits, agreed across 60% of predictions at the Genus level, 80% of predictions at the Order level, and 87% at the Class level, suggesting high accuracy of consensus-only, non-CRASS linked virus-host pairs. All virus-host predictions are in **Appendix C**.

*3.3.6 Genome relative abundance and normalization*

To estimate the relative abundance of each vMAG and MAG, metagenomic reads for each sample were mapped to a database of vMAGs or MAGs with Bowtie2 [74] at an identity of 95%, with minimum contig coverage of 75% and minimum depth coverage of 3x. To normalize abundances for known temporal omics data biases [75], we performed a library size normalization of abundance tables using TMM [76]. Given that PW and SW organism abundances were drastically different in magnitude, and that abundance zeroes across compartments are likely real zeroes, vMAGs and MAGs were considered to be present if detectable in at least 10% of samples in either compartment. Organisms detected in > 10% PW samples were labeled "pore", organisms detected in > 10% SW samples were labeled "surface", organisms > 10% PW and SW samples were labeled "both", and organisms that were in < 10% SW and PW samples were removed. Based on these groups, the TMM abundances file was split into two different files, one for PW samples (n = 17) including "pore" and "both" organisms, and one for SW samples (n = 15) including "surface" and "both" organisms. Abundances for vMAGs and MAGs can be found in **Appendix C** and **Appendix C**, and specific commands can be found on GitHub.

*3.3.7 Temporal and statistical analyses*

Temporal analyses were all performed in R with the TMM normalized abundances described above. To determine which environmental parameters were significantly driving differences across our compartments, we performed multiple regressions using envfit in the vegan R package [77] across multiple types of ordinations. Principal Coordinate Analysis (PCA) for biogeochemistry were done with vegan in R. Dissimilarities in community composition were calculated with the Bray-Curtis metric in vegan [77] for all vMAGs and MAGs that were present in >3 samples per each compartment. Nonmetric multidimensional scaling (NMDS) was then used with k = 2 dimensions for visualization. An analysis of similarity (ANOSIM) was performed using the base R stats package in order to determine community similarity between river compartments. PERMANOVA analyses were done in R using the adonis function from vegan. The NMDS ordinations of the vMAGs and MAGs were compared using the PROCRUSTES function in vegan. To visualize the relative contribution of each biogeochemical variable, we calculated the envfit vector using function ordiArrowMul and plotted them using ggplot. Shannon's H' were done using TMM normalized values with vegan in R. Species accumulation curves were done using the vegan function specaccum in R. All R code and files are available on GitHub.

To determine the relative stability of surface and pore water communities, we first calculated the differences in Bray-Curtis dissimilarity for each sample and its prior timepoint and then ran an unpaired t test to compare the mean differences across compartments with the vegan package in R. For assigning the persistence of the different genomes, we used previously established metrics to assess persistent (present in ≥ 75% of samples), intermittent (present > 25% <75% of samples), or ephemeral (present in ≤ 25% of samples) categories [78]. For establishing the abundance stability, we assessed the total number of samples in which each individual persistent

genome fluctuated by ± 25% of the median relative abundance value across all samples. Then, using the established cutoffs by Fuhrman and Chow et al [78]., we categorized our genomes as stable (shifting in ≤ 25% of samples), intermediately stable (shifting in > 25% < 75% of samples) and unstable (shifting in ≥ 75% of samples). Fishers exact test for count data was used for assessing the significance of difference in stability metrics using fisher.test from R base stats package. The enrichment analyses for AMGs were performed using a hypergeometric test between the total AMGs in our dataset and the individual groups of AMGs present in either compartment. The code used is available on GitHub. All temporal analyses and results are in **Appendix C**.

To reduce the complexity of our microbial data so we could link viral and microbial communities more concretely to ecosystem biogeochemical cycling, we applied a Weighted Gene Correlation Network Analysis (WGCNA) to identify which groups of organisms co-occurred using TMM normalized values in R with package WGCNA [79,80]. A signed hybrid network was performed with a combined dataset of MAGs and vMAGs on a per-compartment basis. For SW, we used a minimum power threshold of 14 and a minimum module size of 20. For PW, we used a minimum power threshold of 8 with and a minimum module size of 20. For both networks, a reassign threshold of 0, and a merge cut height of 0.3 were used.

To link the modules to ecosystem biogeochemistry, we performed sparse partial least square regressions (sPLS) on the groups of organisms in each module. sPLS were done using TMM normalized values of co-occurring communities that resulted from WGCNA above in R with package PLS [81]. Subnetwork membership was related to the overall genome significance for nitrate as described in the WGCNA tutorials document (see GitHub code) using R and the WGCNA package [79]. Full code for WGCNA and SPLS are available on GitHub along with

detailed instructions and input files. Visualizations for figures 6a, 7b, and 7c were made using RawGraphs [82].

*3.3.8 Data availability*

The datasets supporting the conclusions of this article are publicly available and collected as part of the Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS) collective sequencing project and are all publicly available on ESS-Dive [48]. The 125 MAGs are deposited under BioProject PRJNA946291 and in Zenodo (https://doi.org/10.5281/zenodo.7709817). The raw annotations for each genome are deposited on Zenodo (https://doi.org/10.5281/zenodo.7709817). The 1230 vMAGs have been deposited NCBI under the BioProject ID SAMN34000891 and in Zenodo (https://doi.org/10.5281/zenodo.7709817). Additionally, the dataset of freshwater and wastewater viruses we used to cluster to the HUM-V viruses is also hosted on Zenodo (https://doi.org/10.5281/zenodo.7709817). All scripts, commands, and input data used for this manuscript are available at https://github.com/jrr-microbio/erpe_river.

*3.3.9 Funding*

JRR, AO, MAB, RMF, RAD, JE, LS, and MS were fully or partially supported by awards to KCW, including those from DOE Office of Science, Office of Biological and Environmental Research (BER), grant nos. DE-SC0021350 and DE-SC0023084, as well as the National Science Foundation, grant no. 2149506. A portion of this work was performed by MAB, RD, AEG, and JCS at Pacific Northwest National Laboratory (PNNL) and funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, and Environmental System Science (ESS) Program. This contribution originates from the River Corridor Scientific

Focus Area (SFA) project at Pacific Northwest National Laboratory (PNNL). PNNL is operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract No. DE-AC05-76RL01830. A subcontract to KCW from the River Corridor SFA also supported a portion of this work. Metagenomic sequencing was performed by the University of Colorado Anschutz's Genomics Shared Resource supported by the Cancer Center Support Grant (P30CA046934).

### *3.3.10 Acknowledgements*

Data used in this manuscript were collected as a part of the WHONDRS 2018 sampling campaign and we thank those that participated in the design and implementation of that effort. Samples were sequenced and processed as a part of the Genome Resolved Open Watersheds effort to sequence world rivers. We also thank Tyson Claffey and Richard Wolfe for Colorado State University server management.

## *3.4 Results*

### *3.4.1 Metagenomics uncovers viral novelty and biogeography of River Erpe viruses*

We sampled 17 pore water (PW) and 15 surface water (SW) metagenomes collected over a 48-hour period using a Eulerian sampling scheme (i.e., at a fixed location) and collected 565.5Gbp of paired metagenomic sequencing (10-48Gbp/sample, 17Gbp avg.) (**Figure 3.1, Appendix C**). Assembly of these samples revealed 6,861 viral metagenome assembled genomes (vMAGs), of which 6,500 vMAGs were ≥10kb in length and were subsequently clustered into 1,230 species-level vMAGs (**Appendix C**). The average vMAG genome fragment was 24,164bp (180,216bp max) in the PW, and 19,553bp (153,177bp max) in the SW (**Appendix C**). Viral MAG richness was consistently 8 times higher ($p < 0.01$) in the SW (845.0 ± 124.4) compared to the PW

(108.3 ± 49.7) and likely drove differences ($p < 0.01$) in Shannon's diversity (H') recorded for the SW (SW = 6.05 ± 0.17, PW = 3.67 ± 0.49) (**Figure 3.2**). In addition to our vMAGs, we identified 1033 metagenome assembled genomes (MAGs) that were dereplicated at 95% identity into 125 medium and high-quality genome representatives. Similarly, MAG richness was higher ($p < 0.01$) in the SW (SW = 62.6 ± 7.2, PW = 21.8 ± 9.0), and showed significantly different patterns ($p < 0.01$) in terms of Shannon's (H') (SW = 2.9 ± 0.17, PW = 2.6 ± 0.3) (**Figure 3.1**). Together, these results highlight how metagenomic sequencing can be leveraged to successfully reconstruct representative viral and microbial communities across river compartments.

Viruses from freshwater systems are not well sampled in the databases commonly used for taxonomic assignment in viral studies [83]. To determine the extent of novel viral diversity recovered, we mined additional set of 21,022 vMAGs from a variety of freshwater, wastewater, and marine samples and added this to the original vContact2 database (**Appendix C**, see Materials and Methods). We then performed protein clustering of our unique 1,230 viruses with this modified aquatic database, revealing 3,030 viral clusters (VCs). This network was composed of 19,623 nodes with 679,402 edges, which was simplified to only show protein clusters that contained at least 1 vMAG from this study (**Figure 3.3A-C**).

Of our 1,230 vMAGs, 1% clustered to known taxonomic representatives of the Caudovirales Order (8 *Podoviridae*, 7 *Siphoviridae*, 3 *Myoviridae*). Of the remaining vMAGs, 37% clustered only to Erpe viruses, constituting 189 novel genera. An additional 41% did not cluster to any vMAG in our database and were "singletons" or "outliers". Interestingly, 17% of our total vMAGs and nearly half of our novel genera were cosmopolitan in aquatic ecosystems, meaning that while these vMAGs failed to cluster with taxonomically known strains, they did cluster with vMAGs recovered from other ecosystems (**Figure 3.3B**). Specifically, our

cosmopolitan novel genera clustered with vMAGs from wastewater treatment plant sludge or effluent (n=168), other rivers surface or sediment samples (n=65), and marine samples of the TARA oceans dataset (n=25) (**Figure 3.3C**).

*3.4.2 Viral and microbial river microbiomes are compartment-specific and coordinated with each other*

The collected biogeochemistry was significantly structured across compartments and explained a large portion of the total variation in our samples ($R^2 = 0.79$, $p < 0.01$) (**Figure 3.4A**). The surface water compartment was driven mostly by 1) the accumulation of alternative terminal electron acceptors (i.e., nitrate ($NO_3^-$), and sulfate ($SO_4^+$)), 2) the availability of nitrogen compounds (i.e., total nitrogen, avg. N), and 3) a more negative overall nominal oxidative state of carbon (NOSC) and a higher H:C ratio. Conversely, the pore water was characterized by 1) accumulation of $NH_4$, 2) the availability of soluble reactive phosphorous (SRP), and 3) the overall concentration of carbon (avg. C), its aromaticity index (AI), and the quantity of double bond equivalents per molecule (DBE). In summary our redox data indicated more oxidative conditions in the SW while the FTICR-MS data showed that SW carbon was likely more labile, accessible, and thermodynamically favorable.

To determine how viral and microbial communities were structured across these biogeochemical gradients, we recruited the time-series metagenomic reads to our viral database of 1,230 dereplicated vMAGs and 125 MAGs and then performed non-metric multidimensional scaling (NMDS) ordinations (**Figure 3.4B-C**). Like the geochemical PCA plots, PERMANOVA analyses showed that river compartment explained 67% ($p < 0.01$) and 59% ($p < 0.01$) of the variation in viral and microbial communities, respectively. The drivers of both viral and microbial communities were nearly identical in both magnitude and direction. Similarly, a PROCRUSTES

analyses showed that vMAG and MAG ordinations are highly coordinated with each other (sum of squares = 0.027, corr. = 0.99, $p < 0.01$) emphasizing the dependencies between these communities (**Fig 3.5**). Further highlighting these compartmental distinctions, the abundances of 85% of vMAGs (n = 1051) and 67% of MAGs (n = 87) were indicators of only one compartment (**Appendix C**). Interestingly, across both viral and microbial ordinations as well as our PCA, time only explained an additional 4-5% of the total variation, albeit significantly ($p = 0.03$, $p = 0.02$, and $p < 0.01$, respectively), likely due to long travel times and hydrological separation (**Appendix C**). Together, these results show viral and microbial communities are strongly structured by compartment, and that these in turn are structured due to compartment-dictated aqueous chemistry and the availability of suitable metabolic substrates.

### *3.4.3 Temporally resolved metagenomics unveils compartment-level stability and persistence of viral and microbial communities*

SW metagenomic temporal samples for both vMAGs and MAGs were on average 2-fold more similar than PW by Bray-Curtis dissimilarities (BC) (vMAG $t = 6.3$; MAG $t = 6.2$, $p < 0.01$) (**Figure 3.6AB**). We next evaluated whether the individual temporal persistence of the viral and microbial genomes shared similar patterns to the BC across compartments, and categorized members using persistence metrics that were previously established [84]. Briefly, if a viral genome was in more than 75% of the samples it was designated as persistent, between 25-75% of samples it was intermittent, and in less than 25% it was ephemeral. Of the 1,035 vMAGs detected in the SW compartment, 70% were categorized as "persistent", with the remainder being 25% intermittent and 5% ephemeral. Contrastingly, of the 374 vMAGs detected in the PW, only 11% were categorized as persistent, with the remainder being 26% intermittent and 63% ephemeral (**Figure 3.6CD).** Similarly, the bacterial and archaeal MAGs shared comparable persistence

patterns across the compartments (**Figure 3.6EF**). Combined, these results showed that SW communities were less temporally dynamic in terms of BC and had more persistently sampled genomes than the PW.

We then assessed whether the relative abundance of persistent genomes was also temporally stable. Based on Fuhrman and Chow [78], we tallied the number of samples in which persistent vMAG and MAG relative abundances exceeded ± 25% of their respective median (**Figure 3.6GH, Appendix C**). Our results showed that both the relative abundance of vMAGs and MAGs in the SW fluctuate less over time than the PW as shown by Fishers exact t test ($p < 0.01$). Our persistence and temporal stability results supplement the observation that surface water communities in this urban stream change less over the 48-hour period than pore water communities which are more dynamic. Together, our results show that vMAGs and MAGs can be structured into temporally resolved groups at the genome level with varying stability.

### *3.4.4 Genome-resolved virus-host analyses demonstrated viruses could infect highly abundant, phylogenetically diverse microbial genomes*

We were able to predict hosts for 73 vMAGs, matching 30% of our total microbial genomes to a viral partner (**Figure 3.7)**. A majority (62%) of vMAGs with host associations were from the SW compartment, with 22% of host-associated vMAGs found in the PW, and around 10% found across both compartments. MAGs that had viruses linked to them were highly abundant, with 54% of our linked vMAGs infecting hosts of the top 25% most abundant MAGs. At the phylum level, 11 of the 20 identified phyla had evidence for a viral host. Notably, all the phyla that could not be assigned a viral link had 2 or less MAG representatives, with the exception of *Desulfobacterota* which had 6 MAGs. Additionally, of the 51 *Patescibacteria* MAGs we recovered in this study, we uncovered 12 possible viral genome links, which to our knowledge is one of the few reports of

possible infective agents for members of this phylum [84,85], and is the only one thus far reported in rivers. Ultimately, nearly a third of the genera from our MAG database as defined by GTDB were successfully linked to a vMAG, providing further evidence that viral predation is likely pervasive across these river microbial communities.

To decipher the potential impacts that viral predation could have on ecosystem biogeochemical cycling, we metabolically characterized the 38 viral-linked MAGs from our genome-resolved database and saw a wide array of metabolisms spanning ecosystem redox gradients (**Figure 3.7**). Across both compartments, viruses were inferred to impact hosts that could modulate both aerobic and microaerophilic metabolism (carbon respiration), as well as anaerobic metabolisms (nitrate reduction, fumarate reduction, fermentation, and nitrogen fixation). For example, vMAGs were predicted to infect hosts with metabolisms such as methanogenesis (e.g., *Methanothrix*), and sulfur metabolisms (e.g., *Sulfurimonas*), which were encoded more predominantly by MAGs in the PW. Interestingly, 20% of the vMAGs that infected hosts were cosmopolitan with representatives identified in other freshwater and wastewater systems. Together, these results suggest that viruses can infect hosts that can play key roles in ecosystem biogeochemical cycling, and insinuate that these influences are likely widespread across a wide biogeographical range.

### *3.4.5 Virally encoded auxiliary metabolic genes can potentially alter host metabolic machinery*

In addition to the impact on microbial communities via predation, viruses can also mediate biogeochemical cycles through enhancing host metabolism with Auxiliary Metabolic Genes (AMGs). We mined our 1,230 vMAGs for putative AMGs and found 165 unique viral AMG candidates after quality filtering (**Figure 3.8A**). We failed to see a statistical enrichment for the number of AMGs in either compartment (Fisher's exact $p = 0.77$), indicating their shared

importance. The functionalities of these AMGs at the gene annotation level (e.g., KO number) were mostly conserved across compartments, with only 27% of unique gene IDs present in both compartments. However, at the functional module level (e.g., amino acid metabolism) 69% of metabolisms were present across both our ecological gradients (**Figure 3.8B**). Conserved DRAM categories across compartments pertained to electron acceptor utilization (e.g., oxygen, nitrate), carbon utilization (e.g., CAZyme inferred substrates), and other reactions (heavy metal usage, nitrification). We note that genes necessary for viral replication like nucleotide biosynthesis, ribosomal proteins, host mimicry, glycan biosynthesis, cofactor and vitamin metabolism, and molecular transporter were conserved between compartments.

Nonetheless, there were some AMGs that did show compartment specificity. Within the surface water, we detected AMGs involved in organic nitrogen mineralization and transcriptional regulation (e.g., peptidases (M50)), sugar metabolism (e.g., fructose, mannose), and motility (e.g., flagellar assembly). These unique AMGs could be associated with a lifestyle supported by more favorable carbon, more aquatic environments that would favor mobility, and differences in protein content associated with the SW environment. On the other hand, in the pore water we detected AMGs that encoded for plant hemicellulose degradation and cobalamin biosynthesis, adaptations that could sustain metabolism in a litter impacted, anoxic habitat. Ultimately, these results suggest that like their microbial hosts, viral AMGs can potentially have some degree of functional tuning or filtering by environmental conditions.

We next considered AMGs that either expanded the host metabolism or that were complementary to the host metabolism (i.e., Class I AMGs) [20]. Of the 12 *Patescibacteria* MAGs that had possible viral genome links, MAG representative CSBR16-119 had two possible vMAG linkages. A comparison of the metabolic capabilities of the host and viral genomes indicated

multiple shared genes (**Figure 3.8C, Appendix C**). For example, a peptidase-like protein (M50) that is inferred transcriptional regulator [86] was present in both the *Patescibacteria* MAG and its infecting vMAG and shared 77% nucleotide and 99% amino acid similarity across the length of the open reading frame (**Appendix C**). The microbial host genome also had a single copy of ribosome L28 encoded, and two viral genomes putatively infecting it that contained 1 AMG each for the ribosomal protein L28. Both L28 AMGs shared 93% nucleotide identity (90%, and 100% query coverage) (**Appendix C**). It is possible that these viral proteins show strong homology to replace the cellular versions in the host [87], or can function to benefit the host, potentially leading to growth rate enhancement [88].

A second putatively infected persistent genome was *Proteobacteria* UBA2383 (a novel unclassified *Competibacteraceae*) which had broad metabolic capabilities (**Figure 3.8D**). This MAG was inferred to be a facultative aerobe encoding genes for aerobic respiration and for denitrification. The genome encoded genes supporting a heterotrophic lifestyle including CAZymes necessary for the degradation of complex carbon substrates like chitin, starch, and polyphenol, and complete glycolytic and TCA pathways for oxidation of carbon. This MAG also encoded the ability to fix nitrogen. The two vMAGs that were associated with this genome encoded genes to upregulate host metabolism, e.g., GTP cyclohydrolase, which generates important co-factors for bacterial metabolic processes like oxidative stress (**Appendix C**) [89,90]. Additional AMGs encoded by infecting viruses could potentially enhance nucleotide biosynthesis (dCTP deaminase, dUTP pyrophosphatase, thymidylate synthase) as well as other viral functions like host mimicry genes (i.e., 7-cyano-7-deazaguanine synthase, 6-carboxy-5,6,7,8-tetrahydropterin synthase) to avoid the CRISPR defense mechanisms encoded within the host *Proteobacteria*. Together, these results show that viruses are not only important due to viral predation in river

ecosystems, but that they can also potentially play critical roles in reprogramming host metabolism.

### *3.4.6 Co-occurrence networks elucidate ecological patterns that inform ecosystem biogeochemistry*

To link viral and microbial communities more concretely to ecosystem biogeochemical cycling, we applied a Weighted Gene Correlation Network Analysis (WGCNA) to identify which groups of organisms co-occurred over the 48-hour sampling time. Highlighting the clear distinctions in SW and PW compartments, WGCNA analyses could not be reasonably performed simultaneously on a dataset containing both SW and PW communities (scale free topology model fit max = 0.32 at power = 20). As such, using only microbial and viral genomic abundances from either SW or PW separately, we identified 15 and 4 co-occurring modules in the SW and PW, respectively (**Figure 3.9**). The largest module in both networks (turquoise module) contained 254 genomes in the SW and 71 in the PW. In the SW compartment, the overall modules had an average richness of 66 vMAGs and 5 MAGs, while in the PW they had an average richness of 46 vMAGs and 10 MAGs.

Overall, both surface and pore water communities had modules of co-occurring genomes that were significantly related ($R^2 > 3$, $p < 0.05$) by sparse partial least square regressions (sPLS) to the collected biogeochemical measurements (**Figure 3.10**). Only total Fe concentrations were related to modules in both the SW (brown, salmon modules) and PW (red module). SW modules were uniquely related to variables pertinent to nitrogen (nitrate, average total nitrogen), carbon (average total carbon, aromaticity index, hydrogen:carbon), as well as physical (temperature, water stage) and geochemical (magnesium, calcium, manganese, ammonium, sulfate) features in these samples. Of the 8 modules that were significantly related to ecosystem hydrobiogeochemical

features, viruses had significant variable importance in projection scores (VIP > 1) in 7 of them, and 70% of the most significantly related genomes across all regressions were viral (**Figure 3.10B**).

Of the 73 vMAGs and 38 MAGs that were computationally linked (**Figure 3.7, Figure 3.8**), nearly a quarter of those vMAGs and a third of MAGs were grouped into the same co-occurring modules. Interestingly, the SW brown module was related to the total nitrate concentrations in our dataset and contained a co-occurring virus-host link (**Figure 3.11A**). The host genome was the *Competibacteraceae* genome in **Figure 3.8D** and its putatively infecting a virus, which together could play roles in modulating the nitrogen cycling through both fixation and denitrification. This virus and microbial host pair had significant negative correlations to nitrate concentrations and were the second and fourth most significantly related genomes to nitrate within the brown module. The virus bacterial ratio (VBR) for these two organisms was nearly 1:1 and significantly correlated, which is expected of kill the winner dynamics [84], and ultimately highlighs the possible dependency of an infecting vMAG and its host (**Figure 3.11B**). In support of this relationship, the viral genome coverages were on average 10x more than the putative host MAG coverage, suggesting a possible lytic infection lifestyle. Further underlining the importance of these related genomes, both were designated as persistent (i.e., present in >75% of all collected timepoints) and were the 1$^{st}$ (vMAG) and 9$^{th}$ (MAG) most abundant genomes detected in the surface waters. Together, these results indicate that viral and microbial predation dynamics are strongly related to, and could potentially influence, ecosystem biogeochemistry which suggests viral content may be useful units for modeling river ecosystem biogeochemistry.

## *3.5 Discussion*

### *3.5.1 Viral reference databases underrepresent certain habitats, missing cosmopolitan, ecologically relevant lineages*

Within the 48-hour time-series metagenomes, we identified 1,230 dereplicated vMAGs that spanned surface water (SW) and pore water (PW) compartments. The large majority of vMAGs from our dataset were of unknown taxonomy (99%), a finding previously reported in another genome resolved viral focused publication from river sediments [8]. Together these two studies surmise the current state of genome-resolved viral analyses from river systems, highlighting how underexplored river ecosystems are from a viral perspective [31]. Given that these recent viral river genomic studies have yet to be incorporated into viral reference datasets, as well as the influence of wastewater treatment on this urban stream, we tested whether adding more viral representatives (n = 21,022) from relevant ecosystems (i.e., wastewater treatment plant effluent, freshwater viruses, and ocean viruses) to our analyses could expand the relevance of the viruses recovered here (**Figure 3.12**). Adding these additional genomes reduced the total number of River Erpe vMAGs that were categorized as singletons or outliers, resulting in the addition of 49 novel genera, and giving biogeographical context to 164 novel viral genera (**Figure 3.3A**).

The biogeography of viruses and how they are structured across spatial gradients has been studied previously in oceanic ecosystems [91,92], developing the idea of "global" or "core" viruses [93–96]. Further, it has been shown that factors like the presence of AMGs can enable a virus to potentially survive across a variety of different ecosystems [17,18]. Nearly a quarter of our Erpe viruses formed genus-level clusters with viruses from wastewater and freshwater systems, and of those, 11% encoded a putative AMG with functions for metabolisms such as carbon utilization, organic nitrogen transformations, and housekeeping functions (i.e., transporters and flagellar

assembly). While the protein clustering of River Erpe vMAGs to wastewater viruses was not entirely surprising given the sampling location was downstream from the wastewater outlet [45], we note that we also clustered a similar proportion of viruses to other viral genomes from river systems. Notably, this similar clustering proportion for River Erpe viruses was not observed with the TARA ocean viruses (**Figure 3.3C**, **Figure 3.9**). These results hint at possible ecosystem filtering that may affect the biogeographical patterns of freshwater viruses. Our results also underscore the importance of customized, ecosystem relevant databases in environmental viromics for extending the ecological relevance of these ecosystem modulators, and further understanding the major drivers for river microbiomes.

### *3.5.2 Temporally and spatially resolved metagenomics demonstrates that viral and microbial communities are compartment specific, and more stable in surface water than sediment pore water*

Rivers are characterized by containing distinct ecological compartments that are in relatively close proximity to each other (e.g., surface waters and pore waters) [97]. Furthermore, river microbiomes have been reported to have temporal dynamics that can change at hourly or sub-daily scales [34–36]. Nonetheless, most river microbiome studies focus on a single compartment, and/or are examined at seasonal or yearly scales thus lacking relevant temporal resolution. To address this temporal knowledge gap, we sampled both the SW and PW compartments over a 48-hour period, with sampling every three hours. Within our identified vMAGs and MAGs, we saw very clear structuring of communities at this compartment scale and observed differences in their persistence and stability (**Figure 3.4ABC, Figure 3.6**). Like the sampled biogeochemistry, microbial and viral communities were heavily structured by river compartment (67% and 59%, respectively). These data support previous studies within urban rivers (using 16S rRNA gene sequencing) that have

shown that chemistry like phosphorous, nitrate, and metals covaries with microbial communities across compartments [98].

Based on the inferred redox and our metagenomic data, we infer that the SW compartment microorganisms are preferentially utilizing oxygen, likely contributing to the accumulation of alternative terminal electron acceptors within this compartment (i.e., nitrate ($NO_3^-$), and sulfate ($SO_4^{2-}$)). Likewise, genomes with the capacity for anoxic metabolisms like methanogenesis, as well as sulfur reduction, were only detected in the porewater. In addition to redox features, our FTICR-MS data was also shown to be an important driver of both microbial and viral communities (**Figure 3.4**). Most of our FTICR-MS metrics indicated more labile and accessible carbon in the SW (lower H:C ratio, lower C:N ratio, lower DBE, lower AI) [99–102], however the NOSC values showed the opposite, with a more negative value being associated with the SW [103]. Explaining these unexpected differences, it has been recently shown that variables other than NOSC may carry higher importance for predicting carbon lability [104], particularly in environments with oxic conditions [105]. Ultimately these findings highlight the need for larger, cross river studies to decode the unifying factors, like redox and carbon quality, that control the structure and function of river microbiomes accounting for variables like river order, geography, time, and compartment.

Sampling with a Eulerian method allowed us to detect microbiomes passing through the same space over time in the SW and PW samples. Due to the flow rate of SW, and the potential that PW communities may be more biofilm impacted, we might have expected to see greater microbial and viral changes in the surface compartment than the sediments over the sampled time period. On the contrary, both vMAGs and MAGs were more persistent and had more stable abundance patterns over time in the SW of the River Erpe (**Figure 3.6AB**). One possible explanation could be methodological due to the PW being sampled less completely due to genomic

extraction bias caused by fine grain sediments, less sampling volume, or strain level complexity. However, our species area curves did not signify an obvious difference in sampling exhaustion between these compartments (**Figure 3.2**), leaving open the possibility that this finding may be biological. A possible biological explanation is that the strong influence of the wastewater treatment plant, where inputs were relatively uniform and continuous over time [45], could contribute to the increased temporal stability we observed. It is also possible that the mixing in the PW hyporheic zone was more frequent that the flow rate within this channel. In support of the former, we did observe strong clustering between our viral genomes and wastewater treatment viral genomes (**Figure 3.3**). Importantly, our study validates other research indicating that surface water microbiomes are not unstable, or intractable [106], and could thus be important for the poorly resolved indices of river health and biogeochemistry that currently exist.

Ultimately, given the limited number of genome-resolved metagenomic studies in rivers, our results demonstrate that increasing the number of temporally resolved, multi-omic datasets could lead to the development of theoretical frameworks for understanding river viral and microbial community dynamics in natural and urban impacted rivers. Further, our results suggest that studies performed at these finely resolved scales can be informative of viral and microbial metabolisms, which are known to be critical for overall river respiration [107,108]. Finally, our results support the idea that understanding these dynamics is useful for enacting efficient sampling campaigns that can maximize efforts based on relevant biological and biogeochemical factors (i.e., adjusting frequency and depth of metagenomic sampling based on empirical knowledge of temporal microbiome dynamics, where perhaps SW communities require less efforts than PW communities).

*3.5.3 Viruses have the potential to regulate river biogeochemical cycles by predation and metabolic reprograming of microbial hosts*

Although river viral ecology is only recently becoming appreciated, early works have suggested that viruses likely play key roles in the structuring of river microbial communities [26,109]. By using a combination of computational methods, we were able to link 73 vMAGs to 38 MAGs spanning a wide range of taxonomic identities (**Figure 3.7**). One of the PW vMAGs in our dataset was identified to putatively infect an archaeal genome of the genera *Methanothrix*, a known canonical methanogen that accounts for the majority of *mcrA* transcripts in other freshwater environments [110]. In addition to virally impacted microbial methane metabolisms, we show 11 of the 29 microbial families that were linked to a virus had the metabolic potential for denitrification, which could have ramifications for nitrous oxide emissions [111] (**Figure 3.7**). Together, our genome-resolved database of microbial metabolisms, and their putatively infecting viruses gives insight into the underpinnings of river metabolisms, including climate-critical metabolisms (i.e., nitrous oxide and methane production), and show that river microbiome studies can provide critical perspectives for understanding the impact that viruses can have in river ecosystems.

In addition to predation, viral auxiliary metabolic genes are recognized across aquatic systems to play key roles in host metabolic reprogramming and can encompass a wide range of processes from photosynthesis to the oxidation of sulfur [8,18,19]. We add to the existing literature and show AMGs in urban river systems may also impact redox important reactions involving nitrogen (peptidases), carbon (CAZymes), and sulfur (thiosulfate reduction) (**Figure 3.8**). Additionally, one of the vMAGs that was predicted to infect a *Patescibacteria* genome encoded a ribosomal protein that was similar to the gene encoded in the host genome (**Figure 3.8C**). Candidate phyla radiation (CPR) organisms like *Patescibacteria* are well known to be present in

wastewater treatment plants [112], an idea supported by the fact 90% of the *Patescibacteria* genomes in our datasets originated from SW metagenomes. Interestingly, CPR organisms are also known to contain non-redundant, highly streamlined small genomes [112,113]. As such, our results hint at the possibility that *Patescibacteria* viruses could help maintain those small genome sizes by encoding for necessary host genes, a concept with ecological precedence that has been previously demonstrated for the virus-host dependency of cyanobacterial photosynthesis in oceans [18].

While there are extremely limited studies from rivers examining viral AMGs, a prior study showed that river viruses contain AMGs that are expressed and potentially influence river biogeochemical cycles [8]. Other works looking at vMAGs from freshwater lakes and estuaries have shown that some viruses seem to exhibit endemism for certain environments, meaning their distribution is limited to a small geographic area [114]. This points to an interesting idea that perhaps AMGs are also tuned to the specific ecological functions of the sampled habitat, and as such that we could expect some degree of endemism. Interestingly, although individual AMG genes in the River Erpe were unique across compartments (27%), their functional categories were highly similar (69%) (**Figure 3.8B**). Our finding is different to a study recently reported form an estuary, where significant partitioning of AMG functions was reported between habitat types (water particle and sediment) [115]. It is possible that due to the constant mixing of surface and HZ water in rivers, stratification at the genomic potential may be less notable, and expression information may be necessary to capture habitat specific differences. Lastly, this study highlights how moving forward annotation resolution and expanding reference database(s) are important factors to consider when extrapolating AMG inferences across datasets [20,60].

By combining our computational virus-host linking and our AMG analyses with weighted correlation network analysis (WGCNA), we were able to link a specific virus-host pair to their

potential impacts on nitrate utilization (**Figures 3.10 and Figure 3.11**). A virus linked *Competibacteraceae* genome was detected to co-occur within the brown module with its infecting virus, and this module was highly negatively related to nitrate concentrations. This bacterial genome had the metabolic capability to fix nitrogen as well as denitrify (**Figure 3.8D**), and the latter metabolism could explain the negative relationship between this bacterium (and its crispr-linked virus) to SW nitrate concentration. In addition, we also provide AMG evidence that this virus could further alter host cell nitrogen use through modulating nitrogen fixation. Interactions like these are not unheard of in natural systems [20], and ultimately suggest that viruses in river systems are possibly top-down (by predation) and bottom-up (by resource-control) regulators of ecosystem biogeochemical cycles.

In conclusion, how microbial and viral communities interact and change across spatial, and finely tuned temporal gradients is poorly understood today. To begin to provide insights, here we characterized both spatially and temporally the microbial and viral communities present in the wastewater treatment plant impacted River Erpe. We showed that within this human-impacted river, compartments were distinct in terms of chemical dynamics and microbiome composition. Interestingly, the viral communities recovered from the River Erpe overlapped more with viral communities from other wastewater treated samples than oceanic systems. Leveraging our genome-resolved databases, we were able to designate virus-host linkages for 30% of our dereplicated MAGs and show that viral predation could potentially impact several key metabolic processes from dominant microbial members that could alter carbon and nitrogen resource pools. In addition to predation, vMAGs also encoded for several auxiliary metabolic genes that may alter inorganic nitrogen availability, and possibly complement host ribosomal genes. We also identified groups of vMAGs and MAGs that were differentially stable across river compartments, and

showed that on average, microbiomes in the SW are more stable than those in PW. Finally, using WGCNA networks, we identify ecologically co-occurring communities through time and show these genomes were highly related to ecosystem biogeochemistry, potentially resulting from microbial metabolisms and viral controls. Together, our results highlight the power of temporally resolved metagenomics in understanding river dynamics. Further, we provide methods and analyses that can be implemented across temporal datasets to reveal meaningful ecological patterns. Finally, this research also provides a strong scaffolding foundation for future temporally resolved river studies that leverage tools like metatranscriptomics, metaproteomics, and biogeochemical rates in order to bridge the gap between the potential impacts of river microbiomes on biogeochemistry, and their direct, quantifiable effects.

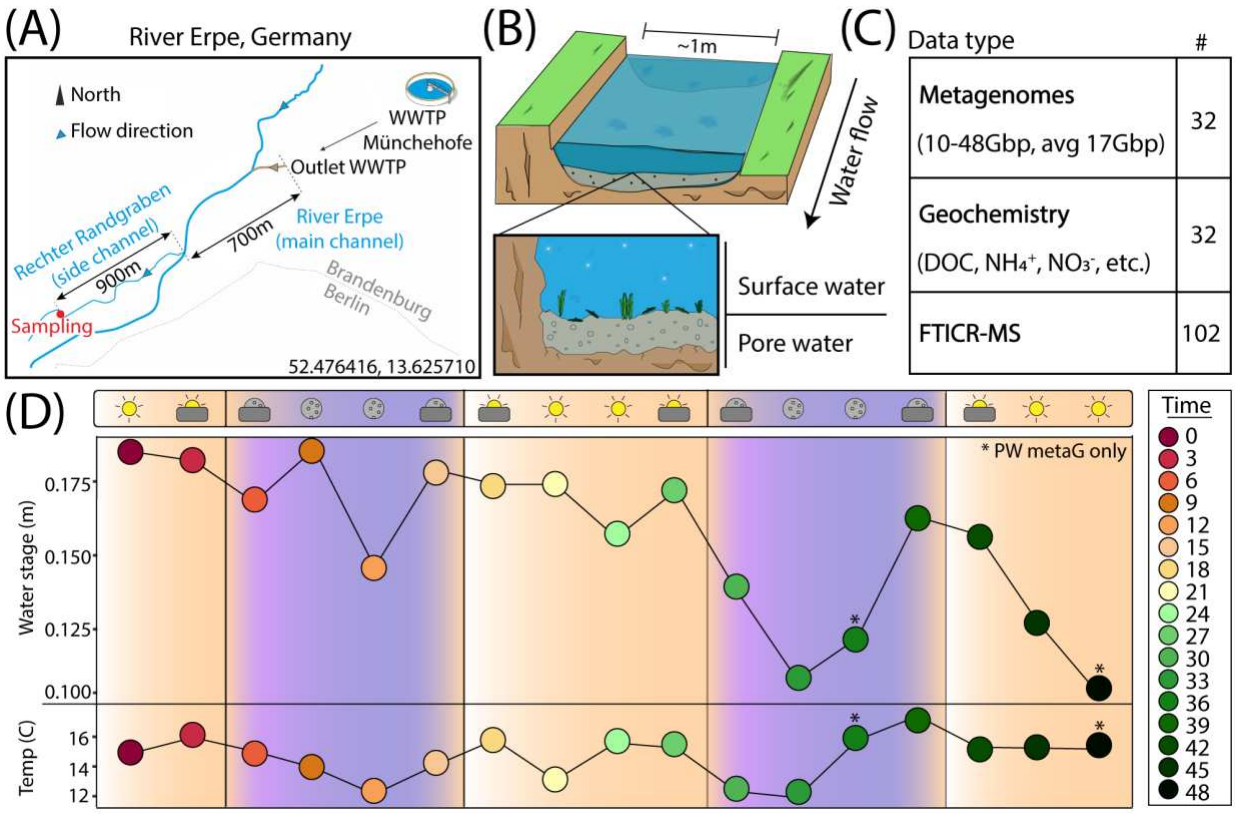



***Figure 3.1: Experimental design enables a genome- and time-resolved view of microbial communities at a finely scaled resolution.*** **A)** River Erpe sampling site that is located near Berlin, Germany. **B)** Conceptual schematic of the surface and pore water compartments that were sampled as part of this research. **C)** Table of data types that were collected as part of this sampling effort. **D)** Sampling schematic over 48-hour period with two ecological variables (water stage, and temperature) shown across the timepoints collected. The colors and icons highlight the hour of the day when samples were collected.

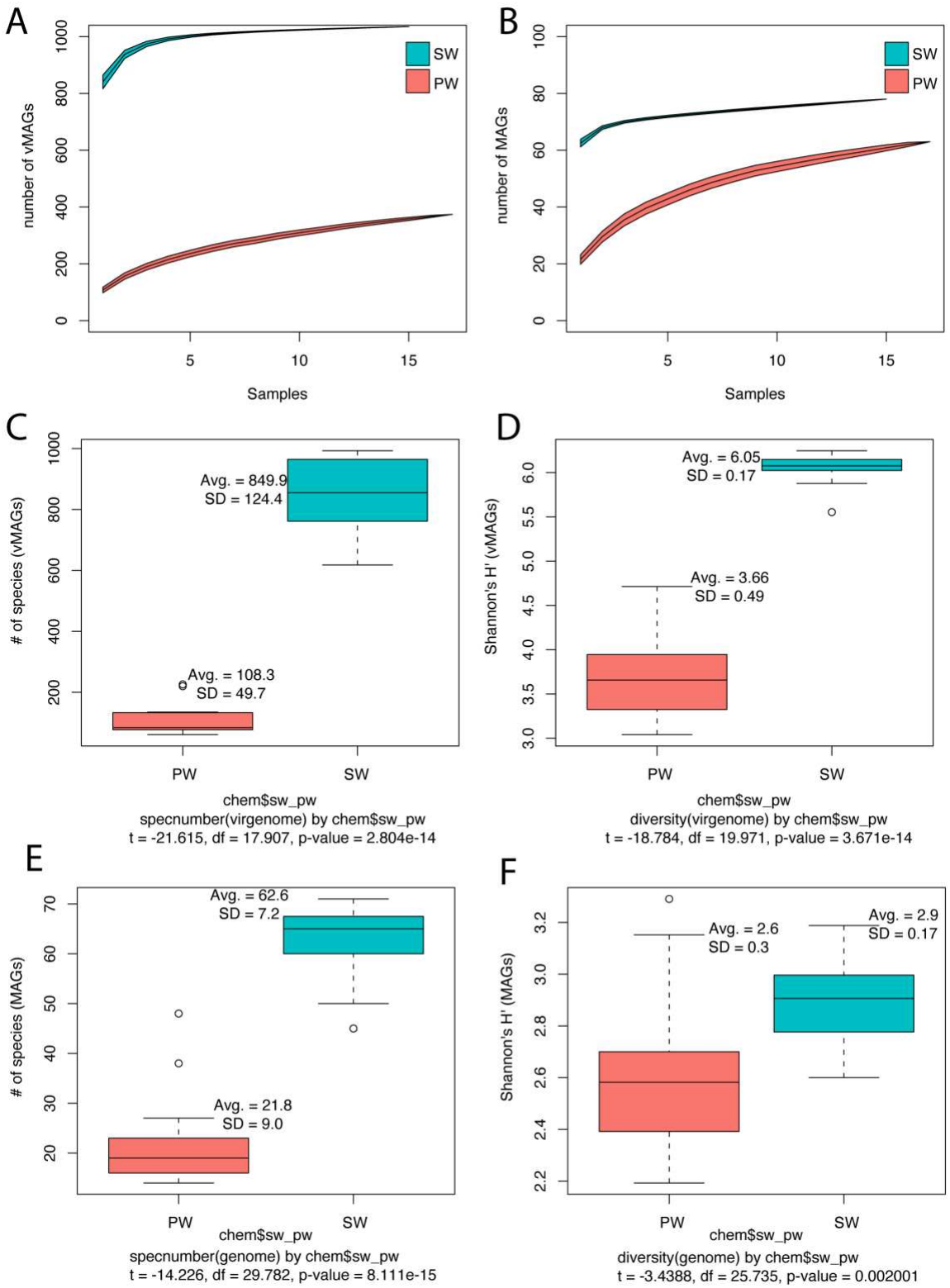



***Figure 3.2****:* A) Viral accumulation curve for identified viral genomes across both the surface (SW) and pore (PW) water compartments. B) Species accumulation curve for identified microbial genomes across SW and PW compartments. C) Total richness for vMAGs across SW and PW

compartments. D) Total richness for MAGs across SW and PW compartments. E) Shannon's H for vMAGs across SW and PW compartments. F) Shannon's H for MAGs across SW and PW compartments.

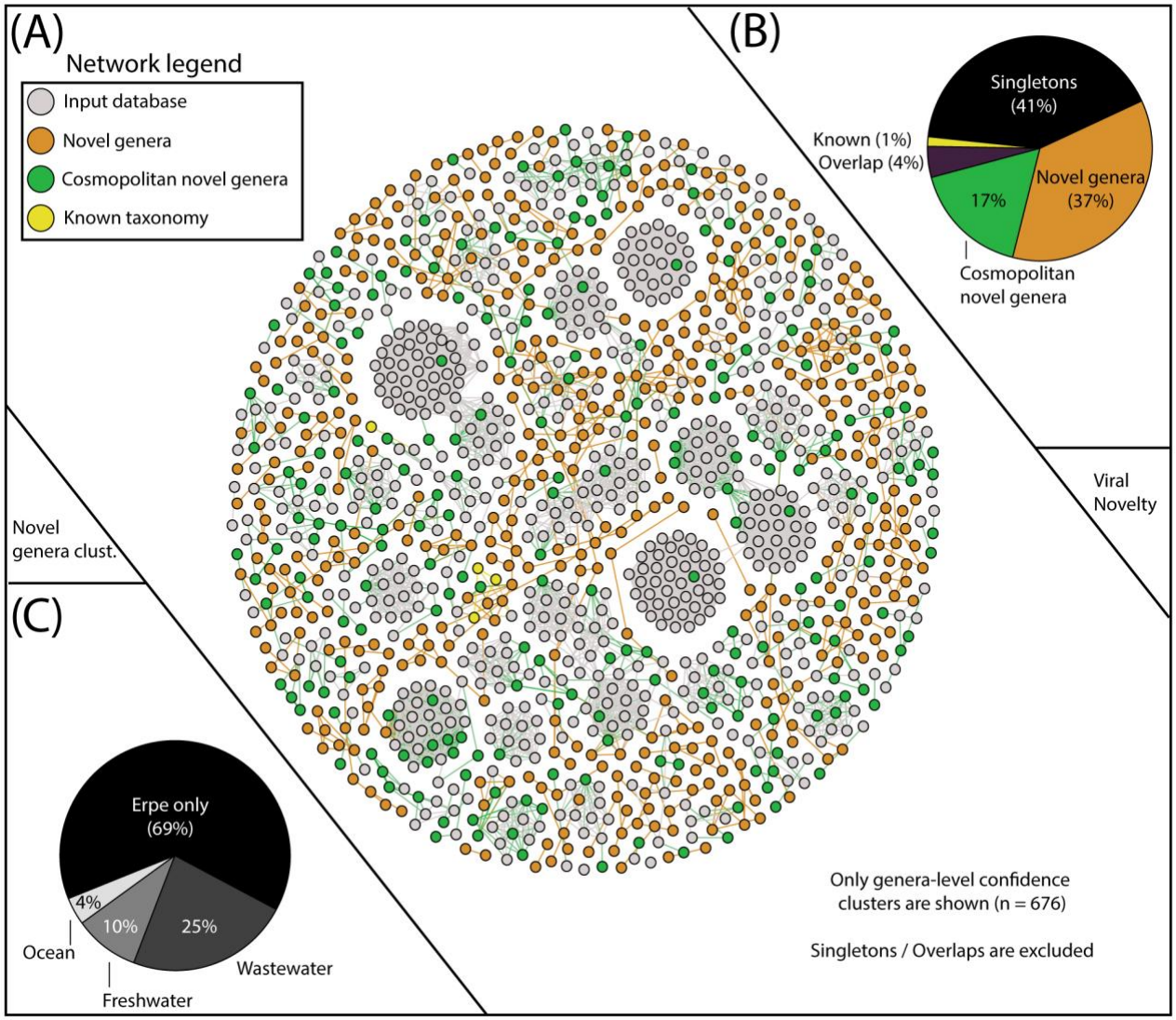



***Figure 3.3: vContact2 reveals Erpe vMAG database constitutes mostly novel genera, and a portion of these are cosmopolitan.*** **A)** vContact2 protein cluster (PC) similarity network where nodes represent vMAGs and edges show similarity across edges. Only high-confidence genera-level clusters are shown (n=676) with node color representing whether the vMAG pertains to our input databases (gray) or other categories assigned to vMAGs recovered here: orange shows novel genera (clustering only with Erpe genomes), green shows cosmopolitan novel genera (clustering with viruses from additional input database not from RefSeq), and yellow represents vMAGs with known taxonomy (clustering with known RefSeq vMAGs). Singletons (genomes that do not cluster with any other genomes) are excluded from the visualization (n=518). **B)** Pie chart shows the distribution of the different categories from the vContact2 network of vMAGs recovered. "Overlap" refers to a category where vContact2 assigns a vMAG to more than one cluster but cannot confidently place in either. **C)** Pie chart shows the proportion of vMAGs from novel genera in this study that were clustering with vMAGs from different environmental input databases.

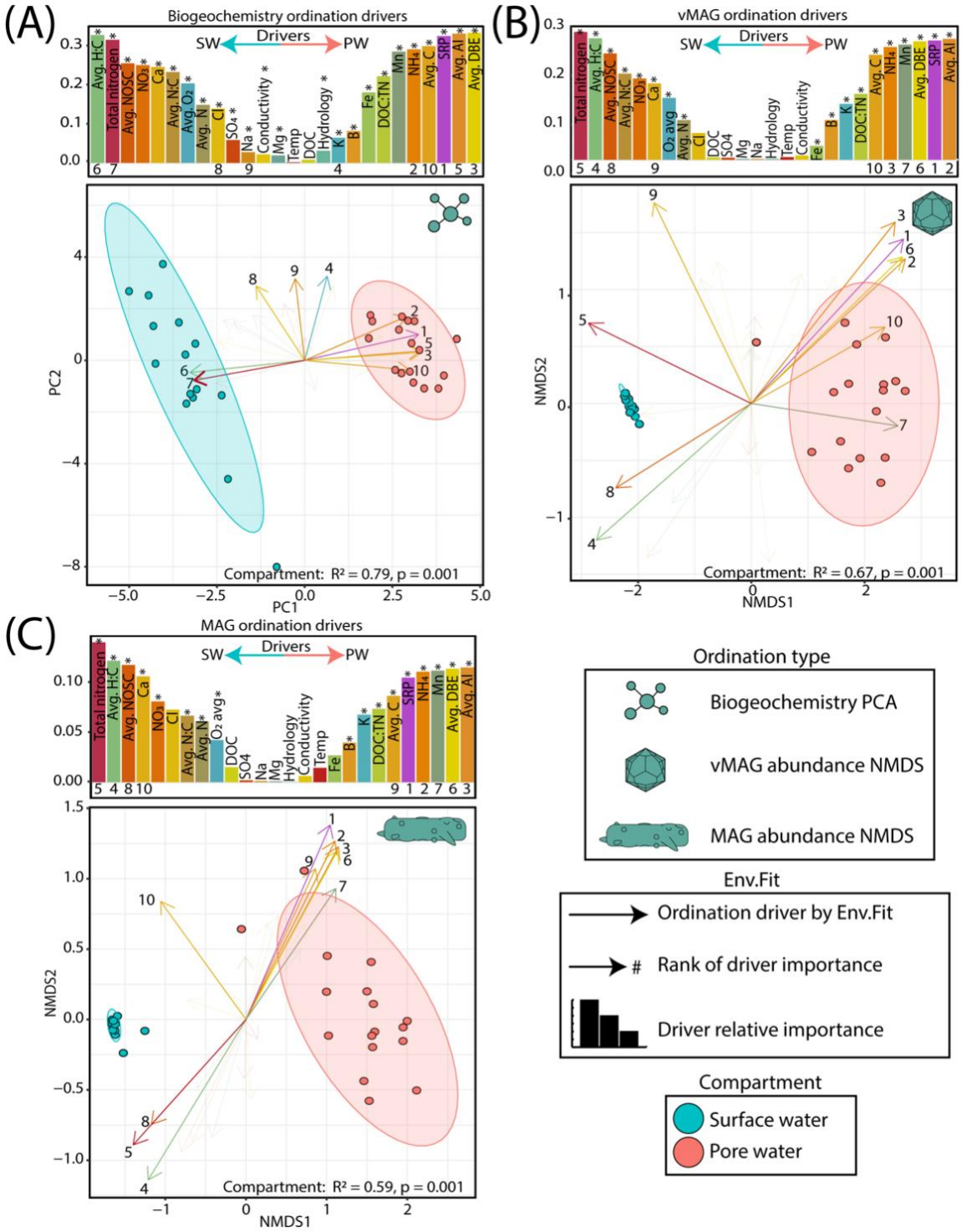



***Figure 3.4: Surface and pore water compartments have distinct viral communities and distributions are driven by biogeochemistry*** **A)** PCA plot of biogeochemical measurements where loadings and bars show the biogeochemical drivers per compartment. The size of bars represents the distance between the end of a loading arrow and the center of the plot. Within each bar plot, the drivers are labeled, and asterisks denote significant drivers by env.fit. The top 10 most significant drivers are numbered below each bar and are shown with solid, numbered arrows within

the ordination below. **B)** NMDS ordination of river pore water and surface water vMAG abundances with bars and arrows showing the same as in (A). **C)** NMDS ordination of river pore water and surface water MAG abundances with bars and arrows showing the same as in (A). Non-compound abbreviations are: nominal oxidative state of carbon (NOSC), calcium (Ca), chlorine (Cl), sodium (Na), magnesium (Mg), dissolved organic carbon (DOC), soluble reactive phosphorous (SRP), aromaticity index (AI), and double bond equivalents (DBE). Note: NOSC values are plotted as the absolute value per value per sample (i.e., a higher SW NOSC driver value translates to a more negative NOSC measurement).

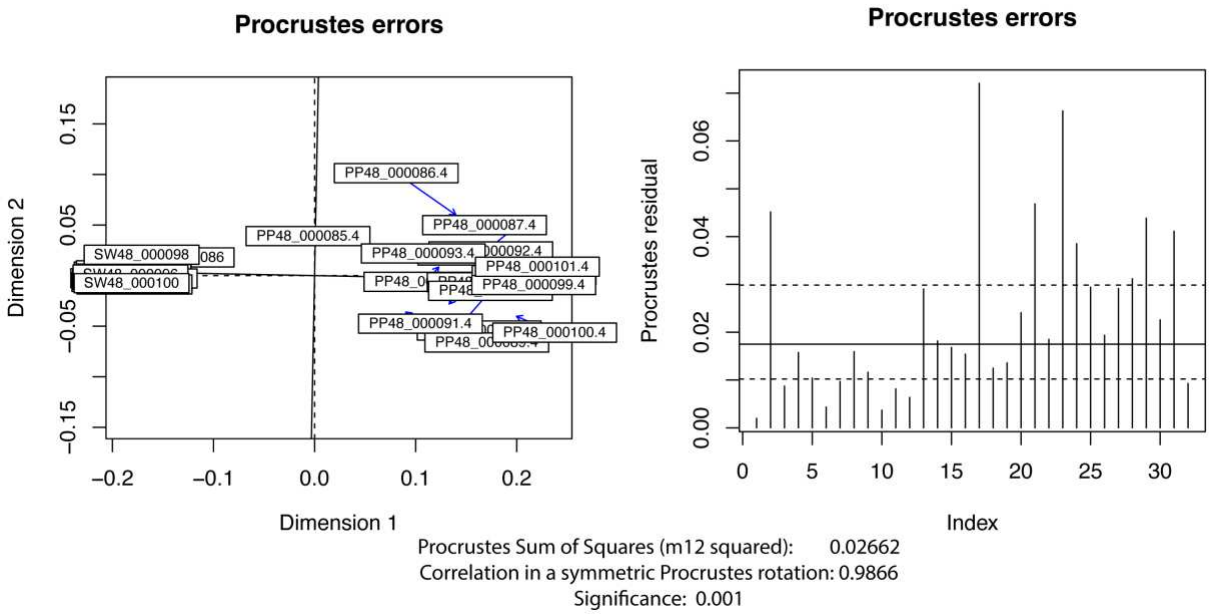



***Figure 3.5****:* Procrustes analysis of the vMAG and MAG non-metric multidimensional scaling (NMDS) ordinations. Figures show the PROCRUSTES results indicating high correlation between the two communities.

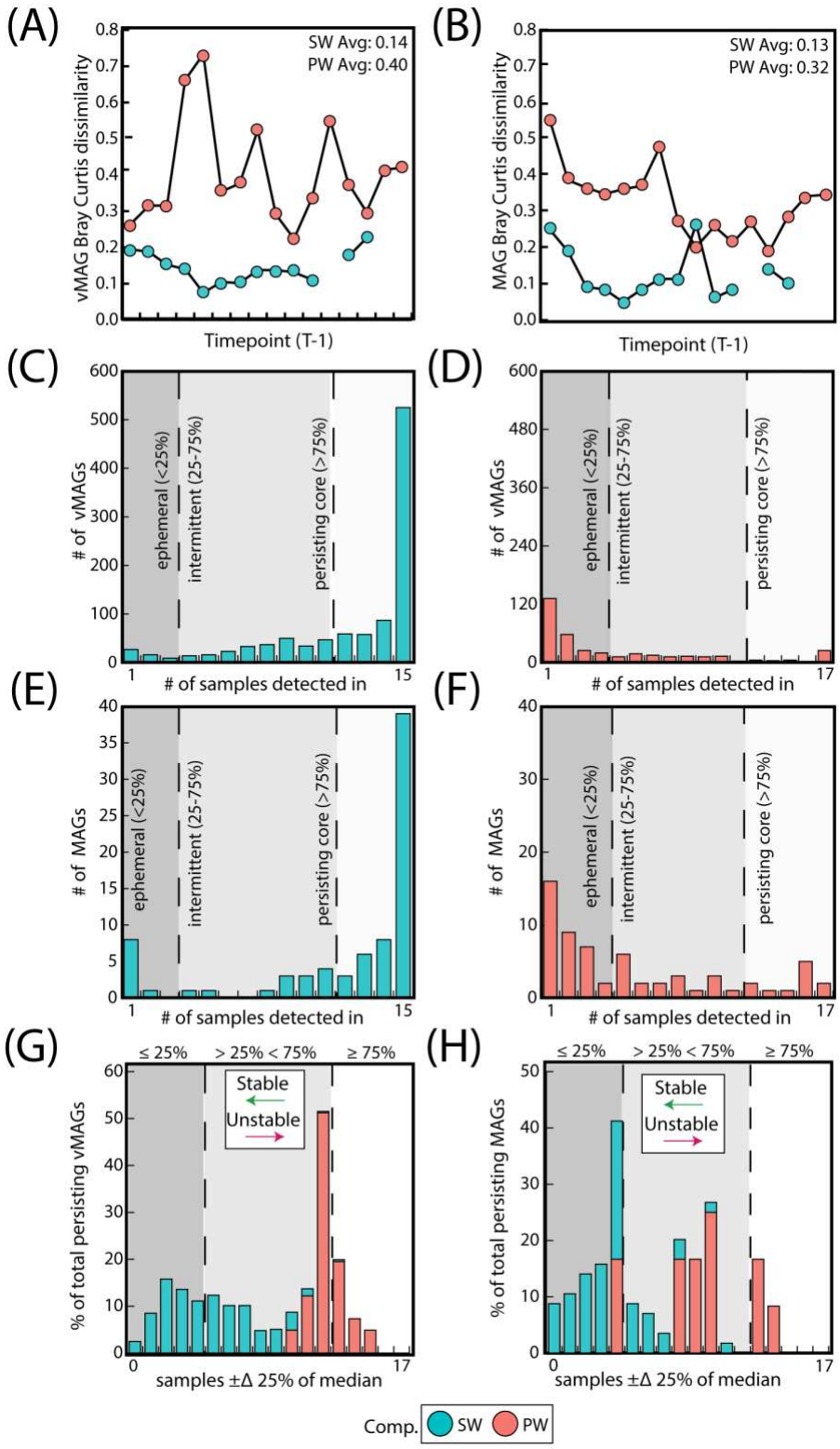



***Figure 3.6: Surface water communities are more stable and persistent than pore water communities.*** **A)** Difference in Bray-Curtis dissimilarities between each sample and its prior timepoint calculated for vMAGs and **B)** MAGs per compartment. **C)** Bar plots show the number of persistent, intermittent, and ephemeral vMAGs in the SW and **D)** the PW. **E)** Bar plots show the

number of persistent, intermittent, and ephemeral MAGs in the SW and **F)** the PW. **G)** Bar plot where the x-axis shows the number of samples where each vMAG that fluctuates above or below 25% of their median values and the y-axis shows the normalized total percentage of persistent genomes per each compartment that are fluctuating. **H)** Identical bar plots to those in **G** but for MAGs.

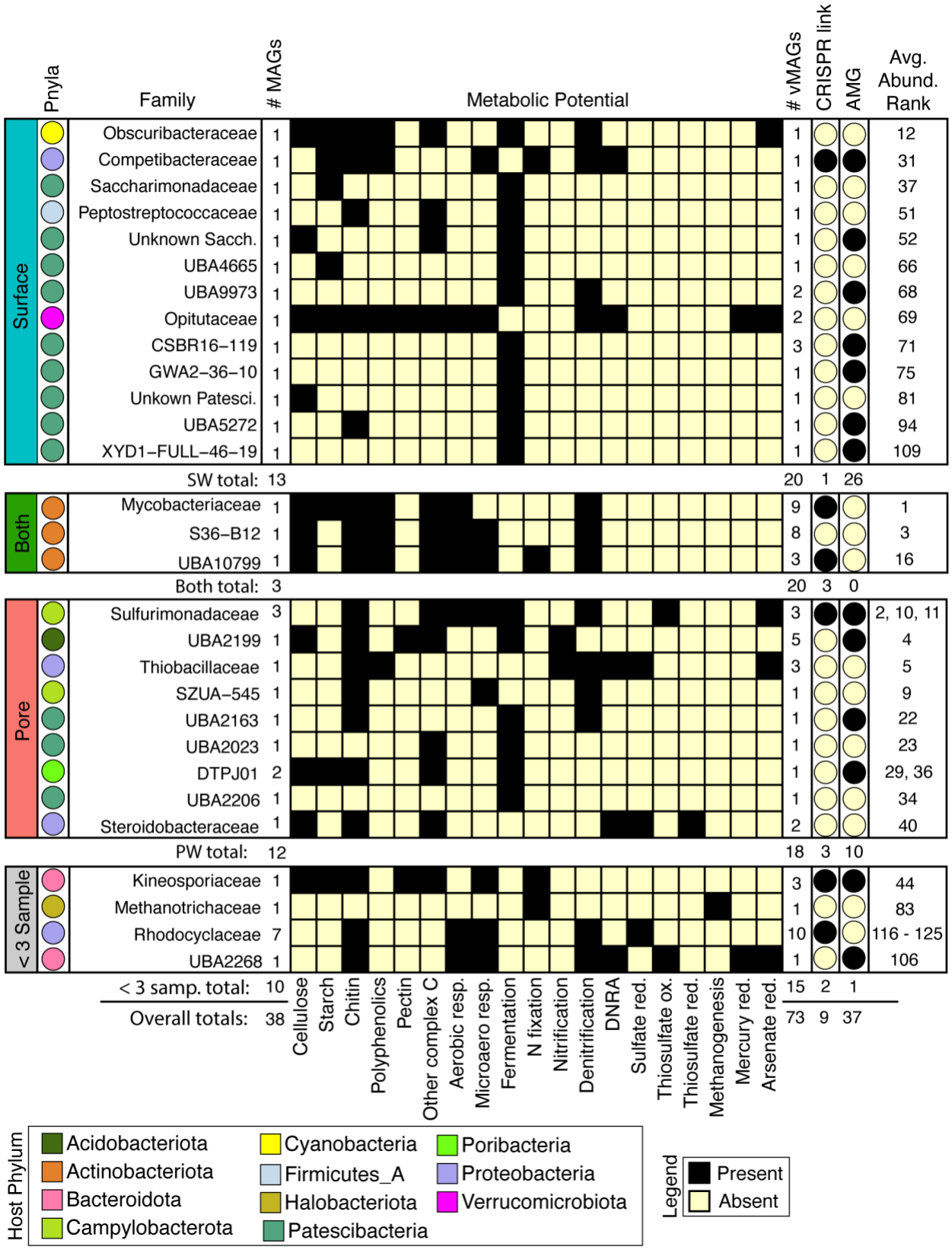



***Figure 3.7: Viruses infect abundant microorganisms in rivers which can influence aerobic and anaerobic C, N, and S cycling by predation or auxiliary metabolic genes.*** MAG families that had

a linkage to a virus are shown and split into their compartment-level distributions. From left to right: Colors of each circle on the leftmost side represent the Phyla, and for each family the total number of MAGs are shown. The presence absence heatmap describes the metabolisms of each family. Following the heatmap are the number of vMAGs that are linked in each family, whether the virus-host link is predicted by CRISPR or consensus method, and if at least 1 infecting vMAG with an AMG is reported. Numbers below each bounding box show totals of above criteria. The overall average rank of each MAG within a family is shown in the rightmost column.

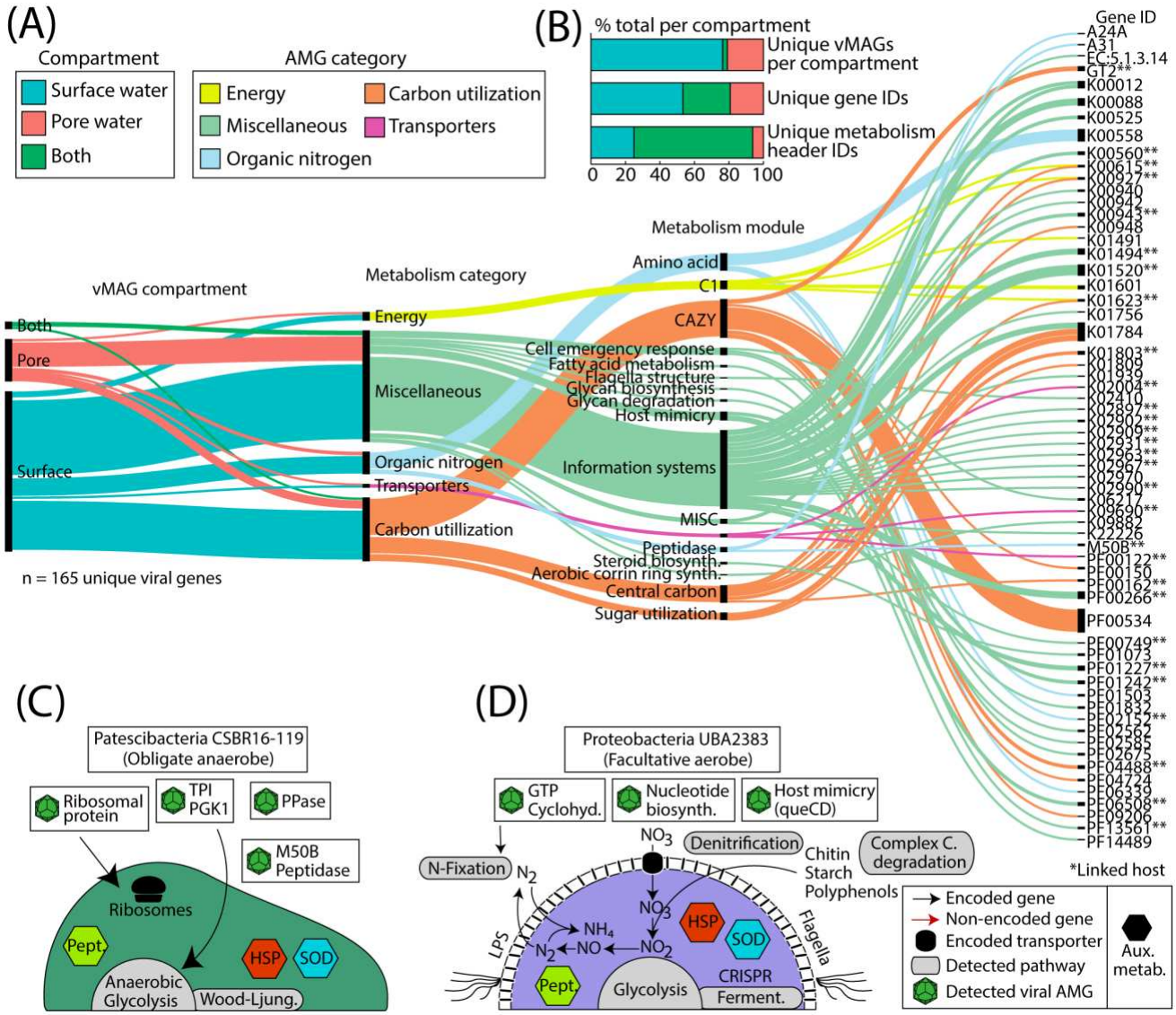



***Figure 3.8: Distribution of viral Auxiliary Metabolic Genes (AMGs) and their function reveals key viral interactions* that *can enhance host metabolism in river ecosystems.* A)** Alluvial plot shows the subset of AMGs (77%, n=165) that had a metabolic function annotated by DRAM-v and were 1) not at the end of a contig and 2) did not contain a transposon like element. In the first vertical line, colors show the compartments that each vMAG with an AMG was detected in. The second vertical line shows the different DRAM-v metabolic categories for each AMG. The next vertical line shows the specific metabolic module name as categorized by DRAM. The final line contains each of the Gene IDs for the detected AMGs. Genes that can have multiple functions (n = 13) are duplicated and treated as individual genes within each category. **B)** Stacked bar charts show the proportion of total AMGs encoded in vMAGs from different compartments at the scaffold, gene ID, and metabolism header ID level as shown in **(A)**. **C-D)** Genome cartoons of two computationally linked bacterial hosts and their respective metabolisms. Detected viral AMGs are shown as viral icons above each genome cartoon. Pept. = peptidases, HSP = heat shock proteins, SOD = superoxide dismutase, queCD = 7-cyano-7-deazaguanine synthase, 6-carboxy-5,6,7,8-tetrahydropterin synthase.

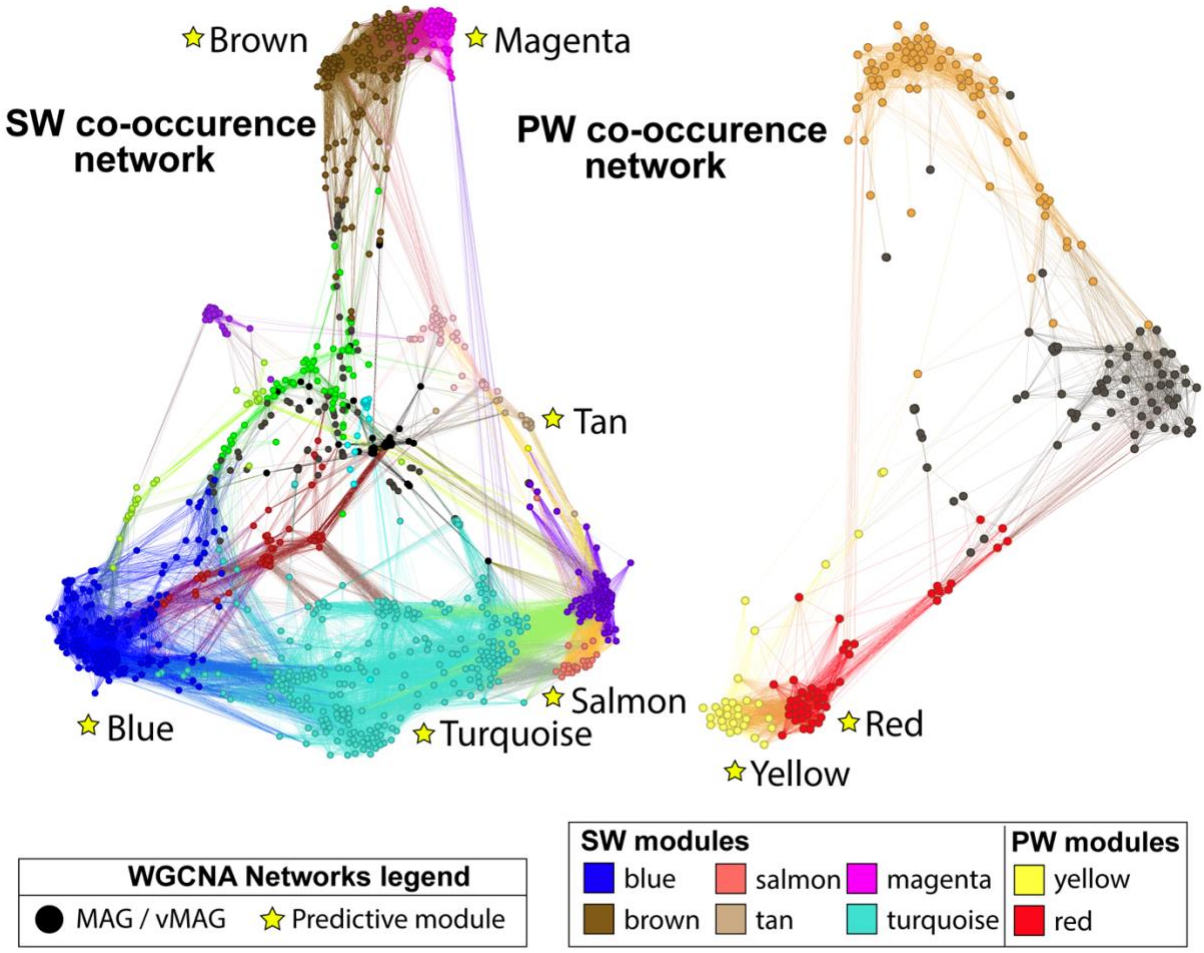



**Figure 3.9***:* WGCNA networks of the surface water and pore water microbial and viral communities. Each circle represents a node (i.e., individual vMAG/MAG), and each line represent an edge which denotes protein cluster similarity. Modules with organisms that are predictive of an environmental variable are denoted by a yellow star.

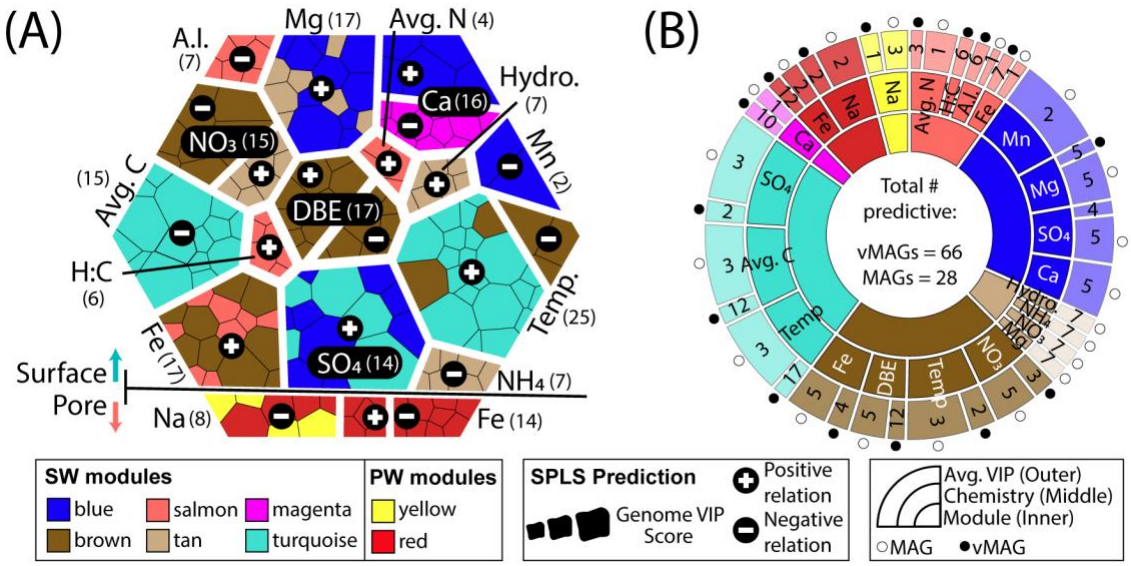



**Figure 3.10: WGCNA co-occurrence networks reveal ecologically similar groups that are related to overall ecosystem biogeochemistry**. **A)** Voronoi diagram shows VIP values of predictions for each predictive genome using a hierarchy structure. Each amorphous square within a group represents a single MAG or vMAG. At the first level (i.e., splitting of the large hexagon into upper and lower groups), SW (top) and PW (bottom) predictions are shown. At the second level (i.e., grouping of individual chemical variables predicted across each compartment), individual chemical variables are shown, per each compartment, and how many vMAGs/MAGs were predictive are denoted by numbers next to each variable name. At the third level (i.e., individual amorphous square or genomes), shapes are sized by the VIP score (>1) of genomes that predict that variable and are colored by their respective WGCNA module. **B)** Sunburst diagram shows the predictive WGCNA modules in the innermost level, followed by what chemical values each module predicts in the middle level. The outer level shows the average variable importance in projection (VIP) score for each genome type: vMAG (black circles) and MAGs (white circles) for that chemical prediction.

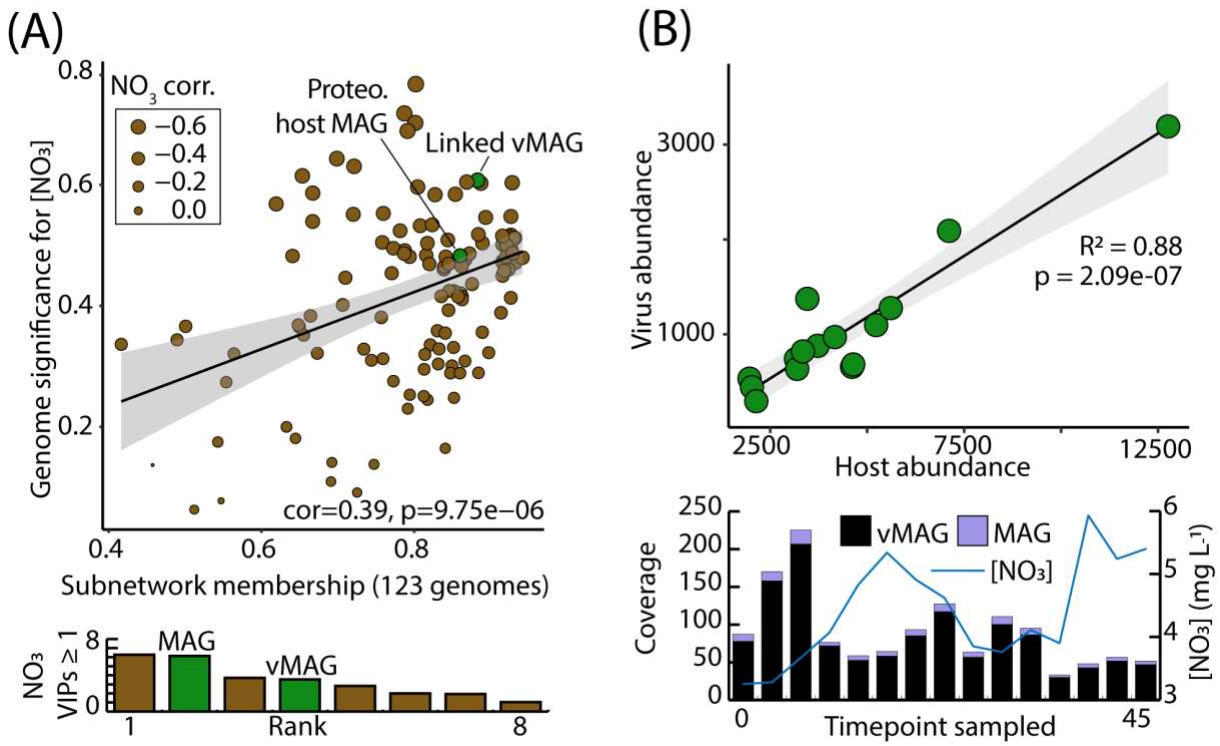



**Figure 3.11: Computationally linked vMAG and MAG pair that share co-occurrence patterns demonstrate high significance for nitrate, and display kill-the-winner dynamics.** A) Scatterplot depicts the genomic significance for nitrate of each of the genomes in the brown module in relation to the membership of those genomes within the WGCNA network modules. Below, bar charts show the VIP score (≥1) of the different organisms in the brown module. **B)** A Virus bacteria ratio (VBR) plot of a viral genome within the brown module that was predicted to infect a Proteobacteria genome. Below it, bar plots show the total coverage across all samples for both the vMAG and the MAG, and a line graph shows the measured nitrate concentrations that these genomes predict.

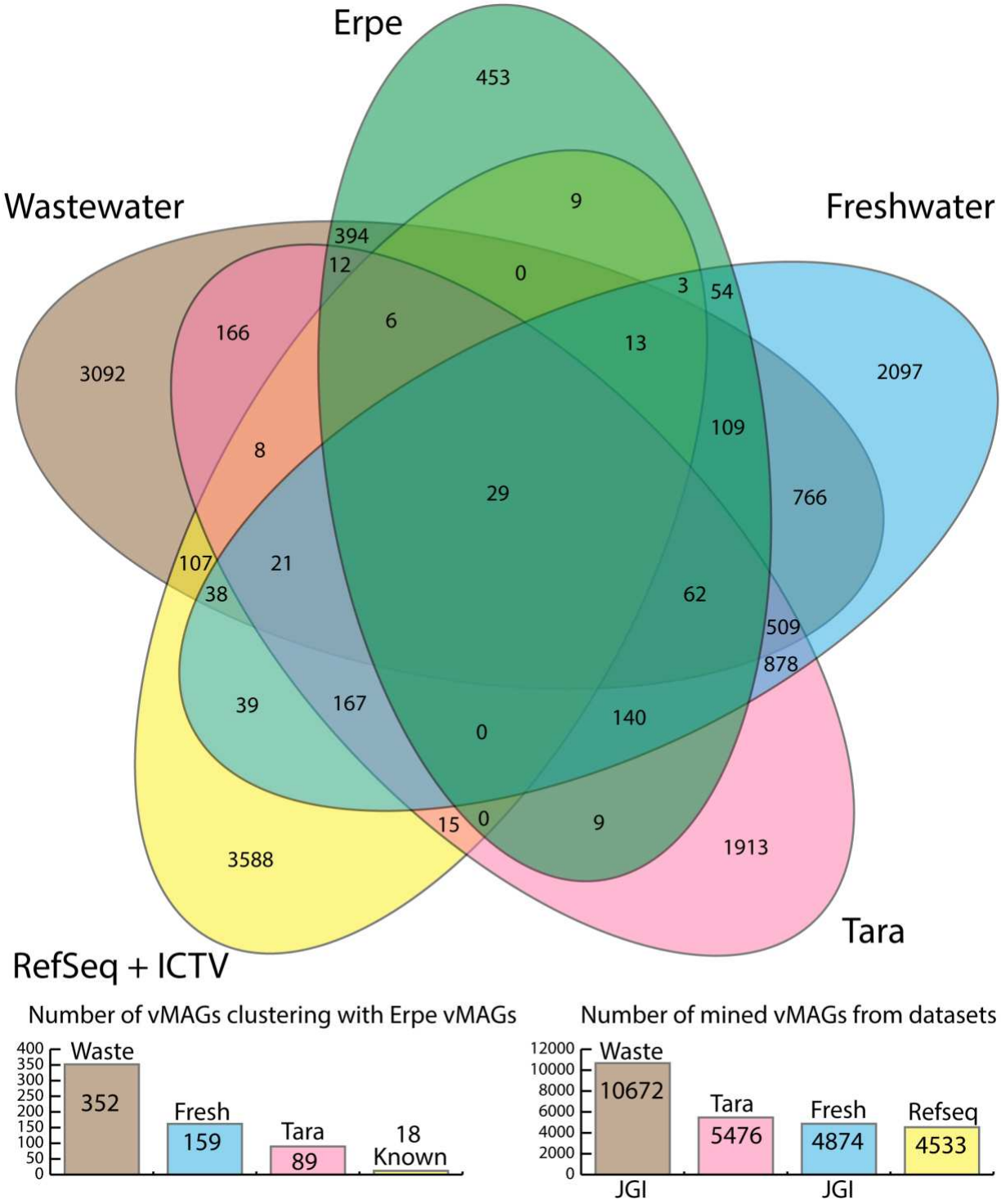



**Figure 3.12***:* The total proportion of clustered vMAGs that were added to the vContact2 protein clustering network. Venn diagram shows the number of vMAGs in each "group" across the different environments. Bar graphs below show the total number of vMAGs that clustered with Erpe vMAGs across ecosystems (left), and the total number of vMAGs that were used as input (right). Note, due to vContact2 quality control parameters, not all vMAGs used in protein

clustering stage are retained, hence the discrepancy between total number of genomes mined from public datasets and total values in the diagram.

*Chapter 3 References*


1. Allen, G. H. & Pavelsky, T. M. Global extent of rivers and streams. *Science* **361**, 585–588 (2018).
2. Villa, J. A. *et al.* Methane and nitrous oxide porewater concentrations and surface fluxes of a regulated river. *Sci. Total Environ.* **715**, 136920 (2020).
3. Rosentreter, J. A. *et al.* Half of global methane emissions come from highly variable aquatic ecosystem sources. *Nat. Geosci.* **14**, 225–230 (2021).
4. Friedlingstein, P. *et al.* Global carbon budget 2021. *Earth Syst. Sci. Data* **14**, 1917–2005 (2022).
5. Liu, S. *et al.* The importance of hydrology in routing terrestrial carbon to the atmosphere via global streams and rivers. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2106322119 (2022).
6. Naegeli, M. W. & Uehlinger, U. Contribution of the Hyporheic Zone to Ecosystem Metabolism in a Prealpine Gravel-Bed-River. *J. North Am. Benthol. Soc.* **16**, 794–804 (1997).
7. Battin, T. J., Kaplan, L. A., Newbold, J. D. & Hendricks, S. P. A mixing model analysis of stream solute dynamics and the contribution of a hyporheic zone to ecosystem function. *Freshw. Biol.* **48**, 995–1014 (2003).
8. Rodríguez-Ramos, J. A. *et al.* Genome-Resolved Metaproteomics Decodes the Microbial and Viral Contributions to Coupled Carbon and Nitrogen Cycling in River Sediments. *mSystems* **7**, e0051622 (2022).
9. Hendrix, R. W., Smith, M. C. M., Neil Burns, R., Ford, M. E. & Hatfull, G. F. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 2192–2197 (1999).
10. Munn, C. B. Viruses as pathogens of marine organisms—from bacteria to whales. *J. Mar. Biol. Assoc. U. K.* **86**, 453–467 (2006).
11. Bar-On, Y. M., Phillips, R. & Milo, R. The biomass distribution on Earth. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 6506–6511 (2018).
12. Mushegian, A. R. Are There 1031 Virus Particles on Earth, or More, or Fewer? *J. Bacteriol.* **202**, (2020).
13. Weinbauer, M. G. & Rassoulzadegan, F. Are viruses driving microbial diversification and diversity? *Environ. Microbiol.* **6**, 1–11 (2004).
14. Weinbauer, M. G. Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* **28**, 127–181 (2004).
15. Suttle, C. A. Marine viruses--major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812 (2007).
16. Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470 (2016).
17. Chow, C.-E. T. & Suttle, C. A. Biogeography of Viruses in the Sea. *Annu Rev Virol* **2**, 41–66 (2015).
18. Sullivan, M. B. *et al.* Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol.* **4**, e234 (2006).
19. Anantharaman, K. *et al.* Sulfur oxidation genes in diverse deep-sea viruses. *Science* **344**, 757–760 (2014).
20. Hurwitz, B. L. & U'Ren, J. M. Viral metabolic reprogramming in marine ecosystems. *Curr. Opin. Microbiol.* **31**, 161–168 (2016).

21. Roux, S. *et al.* Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nat. Commun.* **8**, 858 (2017).
22. Berg, M. *et al.* Host population diversity as a driver of viral infection cycle in wild populations of green sulfur bacteria with long standing virus-host interactions. *ISME J.* **15**, 1569–1584 (2021).
23. Hewson, I., O'Neil, J. M., Fuhrman, J. A. & Dennison, W. C. Virus-like particle distribution and abundance in sediments and overlying waters along eutrophication gradients in two subtropical estuaries. *Limnol. Oceanogr.* **46**, 1734–1746 (2001).
24. Cissoko, M. *et al.* Effects of freshwater and seawater mixing on virio- and bacterioplankton in a tropical estuary. *Freshw. Biol.* **53**, 1154–1162 (2008).
25. Peduzzi, P. & Luef, B. Viruses, bacteria and suspended particles in a backwater and main channel site of the Danube (Austria). *Aquat. Sci.* **70**, 186–194 (2008).
26. Peduzzi, P. Virus ecology of fluvial systems: a blank spot on the map? *Biol. Rev. Camb. Philos. Soc.* **91**, 937–949 (2016).
27. Corinaldesi, C., Dell'Anno, A., Magagnini, M. & Danovaro, R. Viral decay and viral production rates in continental-shelf and deep-sea sediments of the Mediterranean Sea. *FEMS Microbiol. Ecol.* **72**, 208–218 (2010).
28. Rowe, J. M. *et al.* Viral and bacterial abundance and production in the Western Pacific Ocean and the relation to other oceanic realms. *FEMS Microbiol. Ecol.* **79**, 359–370 (2012).
29. Lammers, W. T. Stimulation of bacterial cytokinesis by bacteriophage predation. in *Sediment/Water Interactions* 261–265 (Springer Netherlands, 1992).
30. Vincent, F. & Vardi, A. Viral infection in the ocean-A journey across scales. *PLoS Biol.* **21**, e3001966 (2023).
31. Chu, H., Gao, G.-F., Ma, Y., Fan, K. & Delgado-Baquerizo, M. Soil Microbial Biogeography in a Changing World: Recent Advances and Future Perspectives. *mSystems* **5**, (2020).
32. Hou, Z. *et al.* Geochemical and Microbial Community Attributes in Relation to Hyporheic Zone Geological Facies. *Sci. Rep.* **7**, 12006 (2017).
33. Nelson, A. R. *et al.* Heterogeneity in hyporheic flow, pore water chemistry, and microbial community composition in an alpine streambed. *J. Geophys. Res. Biogeosci.* **124**, 3465–3478 (2019).
34. Tomalski, P., Tomaszewski, E., Wrzesiński, D. & Sobkowiak, L. Relationships of Hydrological Seasons in Rivers and Groundwaters in Selected Catchments in Poland. *Water* **13**, 250 (2021).
35. Lundquist, J. D. & Cayan, D. R. Seasonal and Spatial Patterns in Diurnal Cycles in Streamflow in the Western United States. *J. Hydrometeorol.* **3**, 591–603 (2002).
36. Alonso, C., Román, A., Bejarano, M. D., Garcia de Jalon, D. & Carolli, M. A graphical approach to characterize sub-daily flow regimes and evaluate its alterations due to hydropeaking. *Sci. Total Environ.* **574**, 532–543 (2017).
37. Lu, Q. *et al.* Effect of wastewater treatment plant discharge on the bacterial community in a receiving river. *Ecotoxicol. Environ. Saf.* **239**, 113641 (2022).
38. Wang, J. *et al.* Response of bacterial communities to variation in water quality and physicochemical conditions in a river-reservoir system. *Global Ecology and Conservation* **27**, e01541 (2021).

39. Luo, X. *et al.* Bacterial community structure upstream and downstream of cascade dams along the Lancang River in southwestern China. *Environ. Sci. Pollut. Res. Int.* **27**, 42933–42947 (2020).
40. Gibson, B., Wilson, D. J., Feil, E. & Eyre-Walker, A. The distribution of bacterial doubling times in the wild. *Proc. Biol. Sci.* **285**, (2018).
41. Erbilgin, O. *et al.* Dynamic substrate preferences predict metabolic properties of a simple microbial consortium. *BMC Bioinformatics* **18**, 57 (2017).
42. Wang, J. *et al.* Natural variation in preparation for nutrient depletion reveals a cost-benefit tradeoff. *PLoS Biol.* **13**, e1002041 (2015).
43. Kaevska, M., Videnska, P., Sedlar, K. & Slana, I. Seasonal changes in microbial community composition in river water studied using 454-pyrosequencing. *Springerplus* **5**, 409 (2016).
44. Malki, K. *et al.* Spatial and Temporal Dynamics of Prokaryotic and Viral Community Assemblages in a Lotic System (Manatee Springs, Florida). *Appl. Environ. Microbiol.* **87**, e0064621 (2021).
45. Mueller, B. M., Schulz, H., Danczak, R. E., Putschew, A. & Lewandowski, J. Simultaneous attenuation of trace organics and change in organic matter composition in the hyporheic zone of urban streams. *Sci. Rep.* **11**, 4179 (2021).
46. Lewandowski, J., Putschew, A., Schwesig, D., Neumann, C. & Radke, M. Fate of organic micropollutants in the hyporheic zone of a eutrophic lowland stream: results of a preliminary field study. *Sci. Total Environ.* **409**, 1824–1835 (2011).
47. Stegen, J. C. & Goldman, A. E. WHONDRS: a Community Resource for Studying Dynamic River Corridors. *mSystems* **3**, (2018).
48. Wells, J. R. *et al.* WHONDRS 48 Hour Diel Cycling Study at the Erpe River, Germany. (2019) doi:10.15485/1577260.
49. Tolić, N. *et al.* Formularity: Software for Automated Formula Assignment of Natural and Other Organic Matter from Ultrahigh-Resolution Mass Spectra. *Anal. Chem.* **89**, 12659–12665 (2017).
50. Kim, S., Kramer, R. W. & Hatcher, P. G. Graphical method for analysis of ultrahigh-resolution broadband mass spectra of natural organic matter, the van Krevelen diagram. *Anal. Chem.* **75**, 5336–5344 (2003).
51. Joshi NA, F. J. N. *Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files*. (Github, 2011).
52. Andrews, S. *FastQC: A quality control analysis tool for high throughput sequencing data*. (Github).
53. Metagenome Assembly Workflow (v1.0.1) — NMDC Workflows 0.2a documentation. https://nmdc-workflow-documentation.readthedocs.io/en/latest/chapters/3_MetaGAssemly_index.html.
54. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
55. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
56. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* **11**, e0163962 (2016).

57. Bushnell, B. *BBMap: A Fast, Accurate, Splice-Aware Aligner*. https://www.osti.gov/servlets/purl/1241166 (2014).
58. Guo, J. *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
59. Guo, J., Vik, D., Pratama, A. A., Roux, S. & Sullivan, M. Viral sequence identification SOP with VirSorter2. *protocols.io* https://www.protocols.io/view/viral-sequence-identification-sop-with-virsorter2-5qpvoyqebg4o/v3 (2021).
60. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).
61. Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4306.
62. Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
63. Shi, L.-D. *et al.* A mixed blessing of viruses in wastewater treatment plants. *Water Res.* **215**, 118237 (2022).
64. Brum, J. R. *et al.* Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).
65. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
66. Kang, D. *et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies*. https://peerj.com/preprints/27522/ (2019) doi:10.7287/peerj.preprints.27522v1.
67. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
68. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
69. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
70. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2018).
71. Skennerton, C. T., Imelfort, M. & Tyson, G. W. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Res.* **41**, e105 (2013).
72. Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A. & Sun, F. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.* **45**, 39–53 (2017).
73. Zielezinski, A., Deorowicz, S. & Gudyś, A. PHIST: fast and accurate prediction of prokaryotic hosts from metagenomic viral sequences. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab837.
74. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
75. Coenen, A. R., Hu, S. K., Luo, E., Muratore, D. & Weitz, J. S. A Primer for Microbiome Time-Series Analysis. *Front. Genet.* **11**, 310 (2020).

76. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
77. Oksanen, J. *et al.* vegan: Community Ecology Package. R package version 2.4-3. *Vienna: R Foundation for Statistical Computing. [Google Scholar]* (2016).
78. Chow, C.-E. T. & Fuhrman, J. A. Seasonality and monthly dynamics of marine myovirus communities. *Environ. Microbiol.* **14**, 2171–2183 (2012).
79. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
80. R Core Team. A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna.* (2018).
81. Chung, D., Chun, H. & Keles, S. Spls: sparse partial least squares (SPLS) regression and classification. *R package, version* **2**, 1–1 (2012).
82. Mauri, M., Elli, T., Caviglia, G., Uboldi, G. & Azzi, M. RAWGraphs: A Visualisation Platform to Create Open Outputs. in *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter* 1–5 (Association for Computing Machinery, 2017).
83. Elbehery, A. H. A. & Deng, L. Insights into the global freshwater virome. *Front. Microbiol.* **13**, 953500 (2022).
84. Trubl, G. *et al.* Active virus-host interactions at sub-freezing temperatures in Arctic peat soil. *Microbiome* **9**, 208 (2021).
85. Holmfeldt, K. *et al.* The Fennoscandian Shield deep terrestrial virosphere suggests slow motion "boom and burst" cycles. *Commun Biol* **4**, 307 (2021).
86. Rawlings, N. D. *et al.* The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res.* **46**, D624–D632 (2018).
87. Mizuno, C. M. *et al.* Numerous cultivated and uncultivated viruses encode ribosomal proteins. *Nat. Commun.* **10**, 752 (2019).
88. Brahim Belhaouari, D. *et al.* Metabolic arsenal of giant viruses: Host hijack or self-use? *Elife* **11**, (2022).
89. He, A. & Rosazza, J. P. N. GTP cyclohydrolase I: purification, characterization, and effects of inhibition on nitric oxide synthase in nocardia species. *Appl. Environ. Microbiol.* **69**, 7507–7513 (2003).
90. Holden, J. K. *et al.* Structural and biological studies on bacterial nitric oxide synthase inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 18127–18131 (2013).
91. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).
92. Jian, H. *et al.* Diversity and distribution of viruses inhabiting the deepest ocean on Earth. *ISME J.* (2021) doi:10.1038/s41396-021-00994-y.
93. Heyerhoff, B., Engelen, B. & Bunse, C. Auxiliary Metabolic Gene Functions in Pelagic and Benthic Viruses of the Baltic Sea. *Front. Microbiol.* **13**, 863620 (2022).
94. Ignacio-Espinoza, J. C. & Sullivan, M. B. Phylogenomics of T4 cyanophages: lateral gene transfer in the "core" and origins of host genes. *Environ. Microbiol.* **14**, 2113–2126 (2012).
95. 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).

96. Hurwitz, B. L., Brum, J. R. & Sullivan, M. B. Depth-stratified functional and taxonomic niche specialization in the “core” and “flexible” Pacific Ocean Virome. *ISME J.* **9**, 472–484 (2015).
97. Stegen, J. C. *et al.* Influences of organic carbon speciation on hyporheic corridor biogeochemistry and microbial ecology. *Nat. Commun.* **9**, 585 (2018).
98. Wang, L. *et al.* Shift in the microbial community composition of surface water and sediment along an urban river. *Sci. Total Environ.* **627**, 600–612 (2018).
99. Seo, J.-S., Keum, Y.-S. & Li, Q. X. Bacterial degradation of aromatic compounds. *Int. J. Environ. Res. Public Health* **6**, 278–309 (2009).
100. Mentges, A., Feenders, C., Seibt, M., Blasius, B. & Dittmar, T. Functional Molecular Diversity of Marine Dissolved Organic Matter Is Reduced during Degradation. *Frontiers in Marine Science* **4**, (2017).
101. Bae, E. *et al.* Study of double bond equivalents and the numbers of carbon and oxygen atom distribution of dissolved organic matter with negative-mode FT-ICR MS. *Anal. Chem.* **83**, 4193–4199 (2011).
102. Ghosh, S. & Leff, L. G. Impacts of labile organic carbon concentration on organic and inorganic nitrogen utilization by a stream biofilm bacterial community. *Appl. Environ. Microbiol.* **79**, 7130–7141 (2013).
103. Boye, K. *et al.* Thermodynamically controlled preservation of organic carbon in floodplains. *Nat. Geosci.* **10**, 415–419 (2017).
104. Garayburu-Caruso, V. A. *et al.* Carbon Limitation Leads to Thermodynamic Regulation of Aerobic Metabolism. *Environ. Sci. Technol. Lett.* **7**, 517–524 (2020).
105. Pracht, L. E., Tfaily, M. M., Ardissono, R. J. & Neumann, R. B. Molecular characterization of organic matter mobilized from Bangladeshi aquifer sediment: tracking carbon compositional change during microbial utilization. *Biogeosciences* **15**, 1733–1747 (2018).
106. Graham, E. B. *et al.* Deterministic influences exceed dispersal effects on hydrologically-connected microbiomes. *Environ. Microbiol.* **19**, 1552–1567 (2017).
107. Boulton, A. J., Findlay, S., Marmonier, P., Stanley, E. H. & Valett, H. M. THE FUNCTIONAL SIGNIFICANCE OF THE HYPORHEIC ZONE IN STREAMS AND RIVERS. *Annu. Rev. Ecol. Syst.* **29**, 59–81 (1998).
108. Newcomer, M. E. *et al.* Influence of hydrological perturbations and riverbed sediment characteristics on hyporheic zone respiration of $CO_2$and $N_2$. *J. Geophys. Res. Biogeosci.* **123**, 902–922 (2018).
109. Peduzzi, P. & Luef, B. Viruses. in *Encyclopedia of Inland Waters* (ed. Likens, G. E.) 279–294 (Academic Press, 2009).
110. Angle, J. C. *et al.* Methanogenesis in oxygenated soils is a substantial fraction of wetland methane emissions. *Nat. Commun.* **8**, 1567 (2017).
111. Hu, M., Chen, D. & Dahlgren, R. A. Modeling nitrous oxide emission from rivers: a global assessment. *Glob. Chang. Biol.* **22**, 3566–3582 (2016).
112. Wang, Y. *et al.* Genome-centric metagenomics reveals the host-driven dynamics and ecological role of CPR bacteria in an activated sludge system. *Microbiome* **11**, 56 (2023).
113. Tian, R. *et al.* Small and mighty: adaptation of superphylum Patescibacteria to groundwater environment drives their genome simplicity. *Microbiome* **8**, 51 (2020).
114. Ruiz-Perez, C. A., Tsementzi, D., Hatt, J. K., Sullivan, M. B. & Konstantinidis, K. T. Prevalence of viral photosynthesis genes along a freshwater to saltwater transect in Southeast USA. *Environ. Microbiol. Rep.* **11**, 672–689 (2019).

115. Luo, X.-Q. *et al.* Viral community-wide auxiliary metabolic genes differ by lifestyles, habitats, and hosts. *Microbiome* **10**, 190 (2022).

# Chapter 4: Viruses everywhere all at once: Dissertation conclusions, other works, and perspectives on the future of river viral ecology

## *4.1 Dissertation summary*

Collectively, my dissertation used genome-resolved multi-omic methods to identify viral communities in river ecosystems and examine their roles on microbial community function. I examine the state of the field and provide a detailed summary of the river viral studies published to date (**Chapter 1**). Included in my dissertation work was a database called Hyporheic Uncultured Microbial and Viral (HUM-V), the first of its kind, genome-resolved microbiome database in rivers. It encompasses a collection of bacteria, archaea, and viruses within the Columbia River in Washington, across centimeter depth spatial scales within the river hyporheic zone (**Chapter 2**). I leveraged this database to provide a detailed mechanistic understanding of how organic nitrogen (N) processing within hyporheic sediments contributes to an intertwined carbon (C) and N cycle within rivers. We then expanded on this second chapter by exploring several key ecological gradients of viral and microbial community changes. For this, we leveraged a metagenomic timeseries dataset that was resolved at the compartment level (surface water and pore water) of an urban river located in Germany called the Erpe River; and we identified the prevailing viral and microbial dynamics (**Chapter 3**). Interestingly, surface water and pore water river compartments harbored distinct microbiomes, and the surface water communities were more temporally persistent (i.e., present across the most samples) as well as genomically stable (i.e., changing less in abundance at the genome level across the most samples). Finally, we explored the biogeography of river viruses and showed viral communities are likely subject to environmental filtering, similar to what is observed for their microbial counterparts.

### *4.2 Other collaborations and works*

In addition to the works presented within this dissertation, I have collaborated on several other studies where I contributed my knowledge on viruses and their ecosystem impacts. These collaborative opportunities have given me perspectives and ideas for my own research and resulted in 5 publications. I summarize below, each project indicated with their respective published titles. The author list and full information for these are available in the **Autobiography** section of this dissertation.

#### *4.2.1 Towards optimized viral metagenomes from challenging soils*

Before my time in the Wrighton laboratory, I did rotations while I was at The Ohio State University. One of those rotations was with Dr. Matthew Sullivan, where I worked with Dr. Gary Trubl on optimizing the protocol for the extraction of viral DNA from environmental samples. Within the last 10 years, viral ecology has taken off into unprecedented heights due to the development of multi-omics methods. These multi-omics methods depend upon the extraction of all the DNA that is present in an environmental sample, and then differentiation between the DNA that is viral and non-viral. Given that the largest bacterium is 37,500 times larger than the smallest virus, as such, the small size of viruses and their relatively small genomes presents a key methodological challenge when extracting environmental DNA. In addition, technical challenges related to the separation, isolation, and extraction of DNA viruses from soils also result in difficulties for identifying viral communities in soil ecosystems using standard metagenomic extraction protocols. For this publication [1], we experimentally optimized environmental sample concentration and DNA extraction steps for generating quantitatively-amplified viral metagenomes in order to increase single stranded and double stranded virus yield from

environmental soils samples. These methods have served as the groundwork for different standardization protocols [2] and for understanding viral populations of peat environments [3].

*4.2.2 DRAM for distilling microbial metabolism to automate the curation of microbiome function*

Once in the Wrighton laboratory, as part of my dissertation I analyzed thousands of both viral and microbial genomes across a variety of different ecosystems and metabolisms. During the early years of my PhD, it became very apparent that we would have a scaling issue with regards to annotations and our ability to extract meaningful information from these datasets. To that end, I worked closely with lab members to develop and test the viral portion of an annotation tool called Distilled and Refined Annotation of Metabolism (DRAM) [4]. DRAM profiles metagenomes, as well as both viral and microbial genomes, for metabolisms that are critical to ecosystem function. To accomplish this, DRAM/DRAM-V leverages several annotation databases such as PFAM, KEGG, UniProt, CAZY, MEROPS, and VOGdb. After annotation, it then provides the end user with a distilled and summarized output report of the results. Notably, DRAM-V, the viral counterpart for this tool, also classifies and reports quality of viral auxiliary metabolic genes (AMGs) and has been widely applied to characterize AMGs in several studies [5,6]. Ultimately, DRAM and DRAM-V provide more accessible microbiome datasets, compared to traditional tools that either only leverage a handful of databases, or do not provide simplified outputs.

*4.2.3 Human gut phages harbor sporulation genes*

Building upon the DRAM tool, we implemented an additional resource where users could manually curate and utilize their own “annotation modules” for providing large-scale annotation using the DRAM environment. For this tool, users can provide a customized list of genes or gene

IDs, allowing DRAM to generate more personalized results of their datasets, thereby leveraging the expertise of the user. For this manuscript, we developed a module in DRAM that contained a collection of known sporulation genes, and then used this module to identify sporulation genes across a dataset of human gut metagenomes [7]. We then used our curated viral database to link vMAGs to putative hosts using in-silico computational methods. Our results showed that viruses encode for AMGs that can potentially regulate host sporulation within the human gut. This has important implications in environments that experience abrupt changes in environmental conditions (i.e., gut responses to inflammation upon infection, soil environments that experience wet up events), and links viruses to the survival and cellular emergency response of microbes. The sporulation module, along with other sporulation gene hidden markov models (HMMs) that were developed here, are publicly available in order to serve as a resource for other researchers to identify these genes across their datasets. Ultimately, this publication was the first piece of scientific literature that demonstrated human gut viruses as having a large amount of sporulation AMGs, and links viruses to host gut microbiome stability and its responses to perturbation.

### *4.2.4 Exposing New Taxonomic Variation with Inflammation – A Murine Model-Specific Genome Database for Gut Microbiome Researchers*

While most of my dissertation focused on the river microbiomes, I wanted to extend my skillset outside of this particular ecosystem. For this, I collaborated with Dr. Ikaia Leleiwi as a second author to identify the viral communities of the CBA/J mouse gut. The murine CBA/J mouse model has been widely used in immunology and enteric pathogen research. Notably, the murine CBA/J mouse model has illuminated interactions between *Salmonella* and the gut microbiome due to the fact that pathogen proliferation does not require disruptive pretreatment of the native microbiota, nor does it become systemic. As a result, the CBA/J mouse model is an

analog to gastroenteritis disease progression in humans. Despite the value to broad research communities, microbiota in CBA/J mice are not represented in current murine microbiome genome catalogs. To address this critical gap, we present the first viral and microbial catalog of the CBA/J murine gut microbiome[1]. MAG reconstruction from metagenomic sequencing enabled us to profile the functional potential of the murine gut microbiome during acute *Salmonella* inflammation, contrasting community membership and gene content with uninfected mice. Further, our phage analyses revealed the possibility that phage infection could alter microbial community responses to *Salmonella* infection by means of sporulation induction by auxiliary metabolic genes or through control of microbial populations. Ultimately, this CBA/J microbiome database provides the first genomic sampling of relevant, uncultivated microorganisms within the gut from this widely used laboratory model. Further, it demonstrates that rare and novel viruses sampled across this inflammation gradient advance the utility of this microbiome resource to benefit the broad research needs of the CBA/J scientific community, and those using murine models for understanding the impact of inflammation.

### *4.2.5 Integrated, Coordinated, Open, and Networked (ICON) Science to Advance the Geosciences: Introduction and Synthesis of a Special Collection of Commentary Articles*

Throughout my PhD, I leveraged multiple datasets that were collected as part of sampling campaigns whose data was collected for more than just microbiome analyses purposes. As such, often times datasets were not optimized for microbiome analyses. Moreover, the public availability of these datasets in the era of multi-omics studies can often times be an issue due to public data repositories and what institution hosts them. The accessibility and public availability of data in the field of microbiome sciences has been a topic of fierce discussion within the last few years, given the recent boom in metagenomic studies. Efforts like the Findable Accessible

Interoperable and Reproducible (FAIR) principles provide guidelines for enhancing and maximizing the availability of datasets across diverse studies by making them more open[8]. However, FAIR principles relate to the availability of data, and not necessarily involving the integration of disciplines, the coordination of methods, and the development of mutually beneficial networks[9]. In an effort to bridge this gap, I worked with a group of collaborators from different institutions to develop the Integrated, Coordinated, Open, and Networked (ICON) principles [9]. We then crowdsourced a special collection of manuscripts through the American Geophysical Union (AGU), where 181 researchers, representing 19 of the 25 AGU sections discussed what ICON science looked like within each respective discipline and described their vision of ICON implementation. These publications revealed that scientists agree the ICON framework is a necessary concept that should be implemented, and which will provide the transparency and accessibility to science it desperately needs. As part of the leadership group of this publication, we then created the ICON Advisory board whose responsibility is to advise and promote the use of ICON principles across disciplines and projects. Ultimately, the ICON framework will provide a foundation for future researchers to build their datasets upon that will enable cross-disciplinary, large-scale analyses that is equitable for all.

### *4.3 Future Research Directions*

#### *4.3.1 A new era of river viral ecology*

In 2018, Hall et al. presented a framework to understand how microbiomes influence the systems they inhabit [10]. According to that framework, the first step to fully understanding an environment is to map out its ecosystem processes (i.e., qualitative changes or fluxes within chemical pools), its critical sub-processes (i.e., the reactions that contribute to a change in pool or

flux), and the distribution of the microbial metabolic pathways that drive those sub-processes (e.g., denitrification, nitrification, heterotrophic respiration). While river ecosystem ecology is well appreciated [11], to fully understand and model these climate-critical ecosystems we need to bridge the gap between ecosystem ecology and microbial ecology.

My dissertation repeatedly mentions how river systems act as key intermediates between terrestrial and aquatic ecosystems, and harbor large pools of both N and C [12–14]. However, prior to my PhD study, the field lacked critical knowledge of the sub-processes contributing to C and N fluxes (e.g., nitrification and denitrification) and the microbial pathways that involved in those sub-processes (e.g., chemoautotrophic nitrification, chemolithoautotrophic ammonia oxidation). Specifically, the field of river microbial ecology was missing a genome-resolved link between microbial organisms and the relevant metabolisms that govern biogeochemical cycles. Furthermore, there were previously only a handful of studies which addressed viruses using a host-linked, ecosystem-wide approach. As such, I built upon existing research and utilized multi-omics to elucidate the microbial and viral sub-processes that contribute to biogeochemical cycling and provided the first insights into their spatial and temporal distributions. While this is a key first step in understanding river ecosystems, there remain numerable unanswered questions in the field of river viral ecology. For this section, I offer insights to advance the field, specifically focusing on strategies to constrain the impacts of viral and microbial sub-processes on ecosystem function.

In addition to their ecological roles, viruses themselves are elemental reservoirs in their respective ecosystems. Many virus structures include a head comprised of a capsid that encloses genetic material, and some viruses include a tail that is used for host recognition and attachment. Both of these structures contribute to the total pools of C, N, and phosphorous (P) in the ecosystem. In 2014, Jover et al. analyzed the elemental composition of ocean viruses, and showed they can

constitute >5% of the total dissolved organic P pool and up to 7% of the total N pool, depending on the area [15]. Their estimations relate the total number of atoms in a virus to the radius of the capsid, its thickness, the volume of base pairs in it, the average molecular formula of a base pair, and the filling fraction of genetic material within the capsid. While useful in marine environments, a comparable calculation is not available in rivers as specific morphological characteristics of virus like particles and their distribution are unknown or too variable. River viruses can be as abundant as those in oceanic systems [16], and scientists have reported capsid sizes that range from 85-200nm [17], which is close to the median capsid diameters used in Jover et al. (50-70nm). As such, understanding these viral properties could help quantify the relevance of viruses to river biogeochemical fluxes, especially given that abundance estimates vary based on stream order and trophic state (**Figure 4.1a**).

To further constrain viral impacts, we also need a clear understanding of the environmental, physical, and chemical variables that structure viral communities. Suttle and Chow identified a variety of different viral traits in ocean environments that could facilitate the global virus dispersion. These traits include being primarily lysogenic, having a high burst size, a small vision size, and being able to infect multiple highly abundant hosts (i.e., being generalists). Additionally, recent metagenomic analyses of the TARA Oceans pole-to-pole oceanic dataset revealed geographic zones which may help constrain predictive modeling efforts, and have identified "hotspots" for understanding viral impacts [18,19]. Like oceans, rivers are hydrologically dynamic ecosystems. As such, it is possible that hydrology structures river viral communities. Conversely, similar C types across longitudinal and hydrological gradients may support similar organisms. Viral biogeography, however, remains understudied. Therefore, leveraging the thousands of available metagenomic datasets to make a global, genome-resolved database would enable a

comprehensive assessment of the river virome and address the present knowledge gap of which environmental factors dictate virus distribution. (**Figure 4.1b**).

While microbial studies have highlighted the contributions of microbial communities to greenhouse gas (GHG) emissions [13,20], viruses are rarely directly addressed, and their impacts on GHG flux are not quantified [21,22]. Recent advances in soil ecosystems demonstrate how stable isotope probing (SIP) allows for targeted metagenomics of relevant viral and microbial communities alongside an assessment of $CO_2$ emissions in incubation experiments [6]. SIP has also been used to measure the $CH_4$-derived C shuttled between viruses and their respective hosts [23]. By using these methods, we could directly quantify the influence of viruses on ecosystem GHG pools and fluxes (**Figure 4.1c**). These influences, however, likely differ along the distinct ecological compartments of river ecosystems: the surface water and the sediment (as well as its associated pore water). As such, beyond global biogeography, the spatial distribution of viruses and their hosts across these redox-distinct zones is essential to constraining microbial and viral processes. Additionally, microbial communities seem to exhibit phenotypic plasticity across compartments [5]. Therefore, understanding how viruses differentially express their genes (e.g., AMGs, replication proteins) across compartments can yield key constraints of viral metabolism (**Figure 4.1d**).

One of the most significant findings in ocean viral ecology was the idea that viruses can alter the direction of C flow across ecosystem trophic levels both via the release of organic matter upon host lysis (i.e., viral shuttle) and the export of C to the ocean floor and away from higher trophic levels (i.e., the viral shunt) [24,25]. Similarly, preliminary studies have shown that viruses can influence C dynamics [26] in river ecosystems as well. Viruses release approximately $0.61 \pm 0.5$ petagrams of C $year^{-1}$ as a result of predation which is equivalent to global ocean estimates [22]. However, we are missing information on the ranges of C release across different

hydrobiogeochemical conditions and river compartments. As such, future efforts should focus on determining viral impacts on riverine C cycling and how the underlying mechanisms differ across ecological compartments (**Figure 4.1e**).

The fate of microbial biomass following viral infection (e.g., C and N) deserves special attention specifically directed at the impact river viruses have on microbial priming within the hyporheic zone (**Figure 4.1f**). As surface and ground water mix, thermodynamically favorable compounds "prime" microbial metabolism, leading to the degradation of less energetically favorable compounds [27]. The cellular debris and contents that result from host cell lysis likely contribute to the availability and heterogeneity of C compound types in natural systems. As such, understanding the role that viral predation plays in increasing the availability of thermodynamically favorable C will likely yield key insights into metabolic constraints that are necessary for ecosystem models. Together, by leveraging existing tools and strategies that have been utilized across diverse ecosystems, we can begin to constrain viral impacts and provide quantifiable, direct evidence for the roles they have in climate-critical river ecosystems.

*4.3.2 The future of river viral ecology and ICON-FAIR frameworks*

In addition to the knowledge gaps identified above, the field of river ecology is currently limited by the lack of data interoperability. While there are over 3,000 metagenomes in public repositories as mentioned in the dissertation introduction, the majority are not usable for cross-system comparisons because they are collected with different methods, research aims, and internal standards. Recently, the Integrated, Coordinated, Open, and Networked (ICON) framework was proposed to facilitate the advancement of the geosciences [9]. The goal of ICON is to provide standards that ensure the life cycle of data extends beyond a single manuscript or study. As such, future collected river samples must *integrate* across physical, chemical, biological, spatial, and

temporal scales to enable wholistic sampling campaigns that can be widely utilized beyond a single project. Data collection must also be *coordinated* with consistent protocols and methods to reduce methodological biases that obfuscate real biological patterns (i.e., metagenomic sequencing depth, under sampling of river sediments versus surface water), and to enable transferability of these datasets across disciplines. Furthermore, we must ensure that the collected data is accessible in an *open* manner and is readily available to scientists to enable cross-river global analyses. To this end, organizations like the National Microbiome Data Collaborative (NMDC) strive to make multi-omic data publicly available and usable for all, regardless of computational or financial resources [28]. Finally, it is important to collect these data in a *networked* manner that involves original stakeholders at every step (e.g., indigenous populations, affected communities) in order to ensure mutual benefit of the scientific outcomes. Rivers have long been regarded as life-sources for civilizations, and knowledge on their care and biology extends far beyond that which is captured by modern science. Highlighting the usefulness of ICON, worldwide river sampling campaigns, like the Worldwide Hydrobiogeochemical Observation Network for Dynamic River Systems (WHONDRS,) have collected multi-omic datasets across rivers, leading to multi-disciplinary publications [29–33].

### *4.4 Conclusion*

While the field of river viral ecology is still in its infancy, I present here convincing preliminary evidence of the critical role that viruses play on river ecosystem processes and beyond. While their roles are globally recognized across a wide range of environments from oceans [34] to soils [35], we build upon the existing literature and provide one of the first genome-resolved, metagenomic insights into the role of viruses in rivers. Specifically, we show that

viruses can infect key microbial processes that can alter river respiration [5], and can contribute to the metabolic reprogramming of their host processes including enhanced N cycling and even sporulation [7]. We also show that these viruses are widespread and likely play key and shared roles across ecosystems. Utilizing the tools and methods that I described (and helped develop) during the course of my PhD [1,4,7], these viral communities can greatly aid our ability to understand ecosystem function amidst a changing global climate. Specifically, I believe that with properly curated bacterial and archaeal databases, viruses are a promising diagnostic resource that can make harnessing metagenomic datasets more accessible in the near future. While microbial genome databases are often queried using 16S rRNA genes for the presence and functionality of an organism, querying them instead with viral genomes could also provide insights into additional ecological dynamics like predation. Given that most viruses have narrow host ranges, a curated database of virus-host linkages followed by a subsequent viral genome query would reveal specific viral interactions that influence biogeochemical processes, as well as the taxonomic resolution of the host. It is my hope that viral ecology will allow users who lack access to complex, cost-prohibitive machinery to benefit from modern technology towards the goal of answering questions that are relevant to their homes, their cultures, and their ways of life. As such, I am excited to continue to be a part of this field and look forward to the amazing advances in technology that will help us leap into the full expanse of bridging viral ecology.

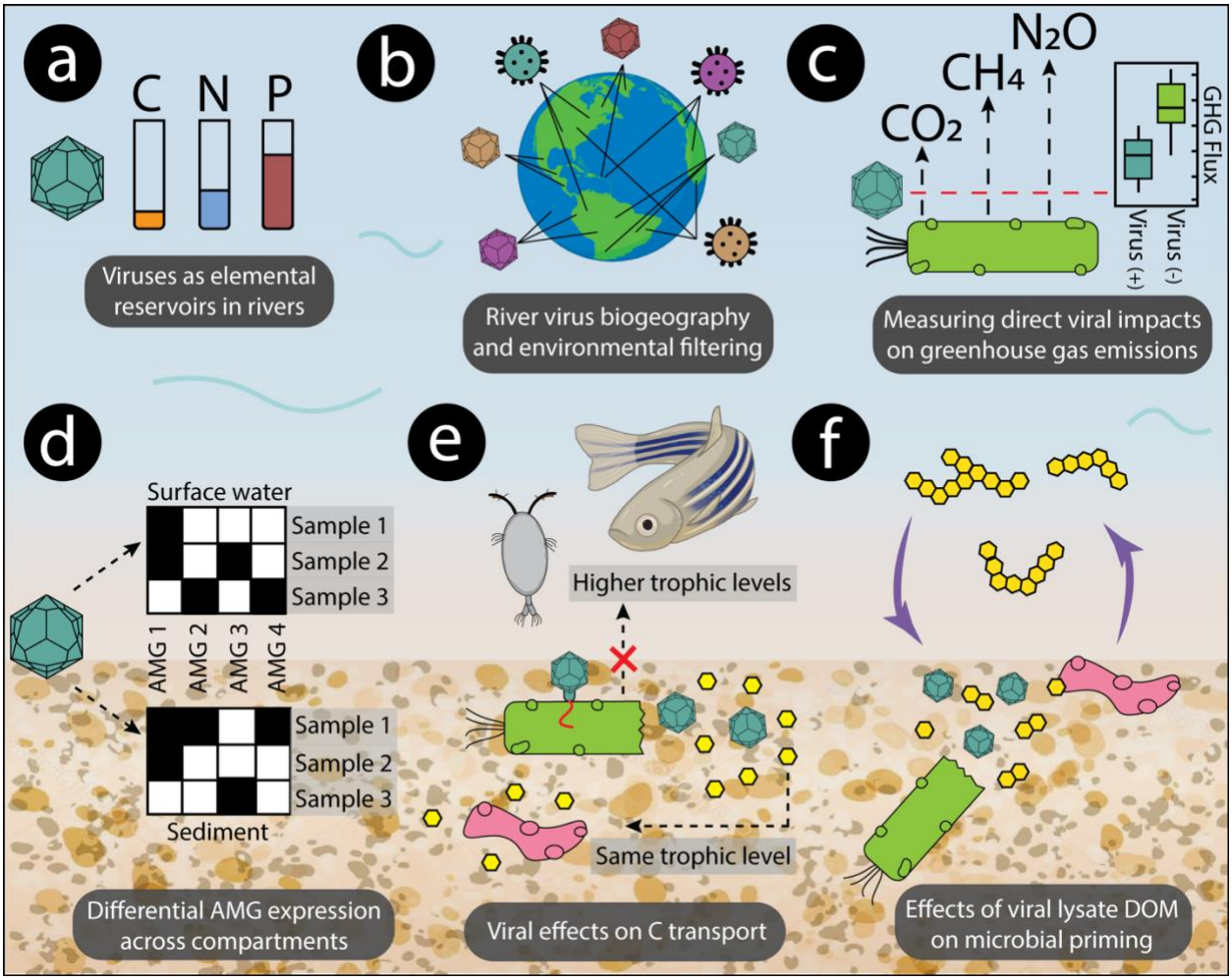



**FIG 4.1:** Conceptual representation of possible future directions in river viral ecology. **a)** Addressing viruses and their roles as elemental reservoirs of carbon (N), nitrogen (N), and phosphorous (P) in rivers **b)** Identifying the global biogeography of river viruses and identifying ecological patterns to their structure **c)** Quantifying the direct impacts of viruses on microbial metabolisms that lead to greenhouse gas (GHG) emissions like carbon dioxide ($CO_2$), methane ($CH_4$), and nitrous oxide ($N_2O$) **d)** Addressing the different expression patterns between surface water and sediments of viral auxiliary metabolic genes (AMGs). **e)** Deciphering the roles of viruses to carbon transport across trophic levels. Yellow hexagons represent carbon monomers. **f)** Understanding how viral lysate impacts the priming of microbial activity during hyporheic zone exchange, and quantifying how much of the lysate becomes biologically unavailable.

Chapter 4 References


1. Trubl, G. *et al.* Towards optimized viral metagenomes for double-stranded and single-stranded DNA viruses from challenging soils. *PeerJ* **7**, e7265 (2019).
2. Trubl, G., Hyman, P., Roux, S. & Abedon, S. T. Coming-of-Age Characterization of Soil Viruses: A User's Guide to Virus Isolation, Detection within Metagenomes, and Viromics. *Soil Systems* **4**, 23 (2020).
3. Ter Horst, A. M. *et al.* Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. *Microbiome* **9**, 233 (2021).
4. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).
5. Rodríguez-Ramos, J. A. *et al.* Genome-Resolved Metaproteomics Decodes the Microbial and Viral Contributions to Coupled Carbon and Nitrogen Cycling in River Sediments. *mSystems* **7**, e0051622 (2022).
6. Trubl, G. *et al.* Active virus-host interactions at sub-freezing temperatures in Arctic peat soil. *Microbiome* **9**, 208 (2021).
7. Schwartz, D. A. *et al.* Human-gut phages harbor sporulation genes. *bioRxiv* 2023.01.19.524802 (2023) doi:10.1101/2023.01.19.524802.
8. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
9. Goldman, A. E., Emani, S. R., Pérez-Angel, L. C., Rodríguez-Ramos, J. A. & Stegen, J. C. Integrated, coordinated, open, and networked (ICON) science to advance the geosciences: Introduction and synthesis of a special collection of commentary articles. *Earth and Space Science Open Archive* (2021) doi:10.1002/essoar.10508554.1.
10. Hall, E. K. *et al.* Understanding how microbiomes influence the systems they inhabit. *Nat Microbiol* **3**, 977–982 (2018).
11. Likens, G. E. *River Ecosystem Ecology: A Global Perspective*. (Elsevier Science, 2016).
12. Liu, S. *et al.* The importance of hydrology in routing terrestrial carbon to the atmosphere via global streams and rivers. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2106322119 (2022).
13. Rosentreter, J. A. *et al.* Half of global methane emissions come from highly variable aquatic ecosystem sources. *Nat. Geosci.* **14**, 225–230 (2021).
14. Hu, M., Chen, D. & Dahlgren, R. A. Modeling nitrous oxide emission from rivers: a global assessment. *Glob. Chang. Biol.* **22**, 3566–3582 (2016).
15. Jover, L. F., Effler, T. C., Buchan, A., Wilhelm, S. W. & Weitz, J. S. The elemental composition of virus particles: implications for marine biogeochemical cycles. *Nat. Rev. Microbiol.* **12**, 519–528 (2014).
16. Ma, L., Sun, R., Mao, G., Yu, H. & Wang, Y. Seasonal and spatial variability of virioplanktonic abundance in Haihe River, China. *Biomed Res. Int.* **2013**, 526362 (2013).
17. Leroy, M., Prigent, M., Dutertre, M., Confalonieri, F. & Dubow, M. Bacteriophage morphotype and genome diversity in Seine River sediment. *Freshw. Biol.* **53**, 1176–1185 (2008).
18. Gregory, A. C. *et al.* Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* (2019) doi:10.1016/j.cell.2019.03.040.
19. Endo, H. *et al.* Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nat Ecol Evol* **4**, 1639–1649 (2020).

20. Villa, J. A. *et al.* Methane and nitrous oxide porewater concentrations and surface fluxes of a regulated river. *Sci. Total Environ.* **715**, 136920 (2020).
21. Zhang, R., Weinbauer, M. G. & Peduzzi, P. Aquatic Viruses and Climate Change. *Curr. Issues Mol. Biol.* **41**, 357–380 (2021).
22. Peduzzi, P. Virus ecology of fluvial systems: a blank spot on the map? *Biol. Rev. Camb. Philos. Soc.* **91**, 937–949 (2016).
23. Lee, S. *et al.* Methane-derived carbon flows into host-virus networks at different trophic levels in soil. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
24. Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470 (2016).
25. Wilhelm, S. W. & Suttle, C. A. Viruses and Nutrient Cycles in the SeaViruses play critical roles in the structure and function of aquatic food webs. *Bioscience* **49**, 781–788 (1999).
26. Pollard, P. C. & Ducklow, H. Ultrahigh bacterial production in a eutrophic subtropical Australian river: Does viral lysis short-circuit the microbial loop? *Limnol. Oceanogr.* **56**, 1115–1129 (2011).
27. Stegen, J. C. *et al.* Influences of organic carbon speciation on hyporheic corridor biogeochemistry and microbial ecology. *Nat. Commun.* **9**, 585 (2018).
28. Eloe-Fadrosh, E. A. *et al.* The National Microbiome Data Collaborative Data Portal: an integrated multi-omics microbiome data resource. *Nucleic Acids Res.* **50**, D828–D836 (2022).
29. Mueller, B. M., Schulz, H., Danczak, R. E., Putschew, A. & Lewandowski, J. Simultaneous attenuation of trace organics and change in organic matter composition in the hyporheic zone of urban streams. *Sci. Rep.* **11**, 4179 (2021).
30. Hartmann, J., Lauerwald, R. & Moosdorf, N. A Brief Overview of the GLObal RIver Chemistry Database, GLORICH. *Procedia Earth and Planetary Science* **10**, 23–27 (2014).
31. Chu, R. K. *et al.* WHONDRS 48 Hour Diel Cycling Study at the East Fork Poplar Creek in Tennessee, USA. (2019) doi:10.15485/1577278.
32. Wells, J. R. *et al.* WHONDRS 48 Hour Diel Cycling Study at the Erpe River, Germany. (2019) doi:10.15485/1577260.
33. Danczak, R. E. *et al.* WHONDRS 48 Hour Diel Cycling Study at the Altamaha River in Georgia, USA. (2019) doi:10.15485/1577263.
34. Suttle, C. A. Marine viruses--major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812 (2007).
35. Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nat Microbiol* **3**, 870–880 (2018).