ABSTRACT

Title of dissertation:	DIVERSITY, DYNAMICS, & DISSEMINATION OF MICROBIAL COMMUNITIES IN RECLAIMED & UNTREATED SURFACE WATERS USED FOR AGRICULTURAL IRRIGATION
	Jessica Chopyk Doctor of Philosophy, 2019
Dissertation directed by:	Professor Amy Sapkota School of Public Health

High quality freshwater is a vital resource for sustaining agriculture and feeding a growing global population. Yet, due to increasing declines in groundwater, key food production regions across the world face uncertainty with regard to water availability. Nontraditional irrigation water sources, such as reclaimed water (advanced treated municipal wastewater) and untreated surface water (e.g. creeks, ponds, and brackish rivers), may contribute to sustainable solutions to conserve groundwater supplies. However, the microbial community composition and dynamics within these water sources are typically poorly characterized and comparative analysis of their microbial communities are rare. Using high-throughput, cultivation-independent sequencing methodologies, this dissertation research focused on three aims: 1) exploring the functional and taxonomic features of bacteria in nontraditional irrigation water sources; 2) assessing the bacterial and viral communities of agricultural pond water in relation to seasonality; and 3) describing the dynamics, composition, and potential dissemination of irrigation water microbiota from a freshwater creek to an irrigated field. The first aim was addressed through a broad investigation of bacteria within agricultural ponds, freshwater creeks, brackish rivers, and reclamation facilities. Through metagenomic-based analyses, features of the bacterial community, such as antimicrobial resistance genes (ARGs) and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) arrays, were found to vary by sampling date and specific site. For the second aim, agricultural pond water was sampled over two time periods and found to harbor diverse bacteria and bacteriophage species, the abundance and composition of which were influenced by factors characteristic of the pond's topography and seasonality. For the final aim, samples from a creek used actively for agricultural irrigation, as well as samples of pre- and post-irrigated soil, were analyzed. ARGs and virulence factors were identified in the water and soil samples, with the majority being specific to their respective environment. Moreover, analyses of CRISPR arrays from the creek samples indicated the persistence of certain bacterial lineages, as well as specific interactions between creek bacteriophage and their hosts. Overall, this research improves scientific knowledge of bacterial and viral composition, dynamics, and interactions that can be utilized to assess the suitability and safety of nontraditional irrigation water sources.

DIVERSITY, DYNAMICS, & DISSEMINATION OF MICROBIAL COMMUNITIES IN RECLAIMED AND UNTREATED SURFACE WATERS USED FOR AGRICULTURAL IRRIGATION

by

Jessica Chopyk

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Dr. Amy R. Sapkota, Chair/Advisor
Dr. Mihai Pop
Dr. Shirley Micallef
Dr. Emmanuel Mongodin
Dr. Amir Sapkota

Acknowledgments

First and foremost, I'd like to thank my advisor, Dr. Amy Sapkota. Her passion and patience is an inspiration and has helped to shape every aspect of the research detailed in this dissertation. Amy encouraged me to explore other areas of research outside my discipline and gave me the time and space to do so effectively. I look forward to continue working closely with Amy both as a friend and a collaborator.

I would like to thank my committee members, Dr. Mihai Pop, Dr. Emmanuel Mongodin, Dr. Amir Sapkota, and Dr. Shirley Micallef who each have provided their time and scientific expertise to ensure not only is this research of quality, but that I am of the quality suitable as a future scientific leader. It is impossible to be in a room with this group of scientists and not feel electrified, not just by their research endeavors, but the scientific pursuit of knowledge itself.

To my fellow graduate students, lab members (current and former), and Business Manager extraordinaire, Maurice Rocque. The camaraderie we have shared during these years has been something I truly cherish and I wish all of you the best of luck.

I would also like to extend a special thanks to my family and friends, especially my brother Gob, my human-brother Alex, Aunt Sandy, Uncle George, Sam Battista, Ashley Guy, and Kim Rogan. In each of their own ways they have reminded me that life exists outside grad school. It has been a special privilege to have such a great group of people in my life. To my parents, June and John Chopyk your unending support and love is humbling. I am honored and proud to *still* be your (favorite) and now smartest child.

Finally, I want to thank Daniel Nasko, the research in this dissertation would truly not be possible without him. Not just for his valued scientific guidance, but also his unrelenting role as my biggest critic, #1 fan, rock, stress ball, and partner.

Table of Contents

Acknowledgements ii

List of Tables ix

List of Figures xi

List of Abbreviations xv

1 Introduction 1
1.1 Global Management of Irrigation Water 1
1.1.1 Systems, sources, and stresses 1
1.2 Nontraditional Sources of Irrigation Water 3
1.2.1 Reclaimed water 3
1.2.2 Untreated surface water 5
1.2.2.1 Ponds 5
1.2.2.2 Rivers and creeks 6
1.2.3 Irrigation water quality guidelines and regulations 7
1.3 Bacteria and Viruses in Nontraditional Irrigation Water Sources . . . 9
1.3.1 General role in the biosphere 9
1.3.2 Risks to public health 11
1.3.2.1 Food-borne pathogens and outbreaks 11
1.3.2.2 Phage mediated transduction 13
1.3.3 Methods for identification and detection 14
1.3.3.1 Culture-based 15
1.3.3.2 Sequence-based 16
1.3.4 Taxonomic composition and dynamics 20
1.3.4.1 Waste to reclaimed water 20
1.3.4.2 Lentic freshwater 22
1.3.4.3 Lotic fresh and brackish waters 24
1.4 Outline of Dissertation 25

2 Comparative Metagenomic Analysis of Microbial Taxonomic and Functional Variations in Untreated Surface and Reclaimed Waters Used in Irrigation Applications 30
2.1 Abstract . . . 30
2.2 Introduction . . . 31
2.3 Materials and Methods . . . 34
2.3.1 Study sites . . . 34
2.3.2 Sample collection . . . 35
2.3.3 Sample processing . . . 35
2.3.4 Shotgun sequencing . . . 36
2.3.5 Metagenomic assembly . . . 36
2.3.6 Taxonomic and functional classification . . . 37
2.3.7 Peptide ORF clustering . . . 38
2.3.8 Identification of antibiotic resistance genes . . . 38
2.3.9 Prediction and analysis of CRISPRs . . . 38
2.3.10 Data availability . . . 39
2.4 Results . . . 39
2.4.1 Sampling site characteristics . . . 39
2.4.2 Sequencing effort and assembly . . . 39
2.4.3 Taxonomic composition . . . 40
2.4.4 Peptide ORF clustering . . . 41
2.4.5 Functional analysis of bacterial-assigned ORFs . . . 42
2.4.6 Antibiotic resistance . . . 43
2.4.7 CRISPR array abundance and taxonomy . . . 44
2.4.8 CRISPR spacers within and among sites . . . 45
2.5 Discussion . . . 46
2.5.1 Conclusions . . . 51
2.6 Figures . . . 52
2.7 Tables . . . 61

3 Agricultural Freshwater Pond Supports Diverse and Dynamic Bacterial and Viral Populations 64
3.1 Abstract . . . 64
3.2 Introduction . . . 65
3.3 Materials and Methods . . . 68
3.3.1 Study site and sample collection . . . 68
3.3.2 Sample preparation . . . 69
3.3.3 Viral DNA extraction and shotgun sequencing . . . 70
3.3.4 Microbial DNA extraction, 16S rRNA gene PCR amplification, and sequencing . . . 70
3.3.5 16S rRNA gene data analysis . . . 71
3.3.6 Virome metagenomic analysis . . . 72
3.3.7 Data deposition . . . 74
3.4 Results . . . 74
3.4.1 Water characteristics . . . 74

3.4.2 16S rRNA gene sequencing effort . . . 75
3.4.3 Bacterial community composition and temporal variations . . 75
3.4.4 Relationships between water characteristics and bacterial abundance . . . 77
3.4.5 Bacterial alpha diversity . . . 77
3.4.6 Bacterial beta diversity . . . 78
3.4.7 Core OTUs . . . 78
3.4.8 Shotgun sequencing effort and assembly . . . 80
3.4.9 Viral taxonomic composition and abundance . . . 80
3.4.10 Viral functional composition . . . 81
3.4.11 Viral marker gene: Polymerase A . . . 82
3.5 Discussion . . . 83
3.6 Figures . . . 88
3.7 Tables . . . 98

4 Seasonal Dynamics in Taxonomy and Function within Bacterial and Viral Metagenomic Assemblages Recovered from a Freshwater Agricultural Pond 103
4.1 Abstract . . . 103
4.2 Introduction . . . 104
4.3 Materials and Methods . . . 107
4.3.1 Study site and sample collection . . . 107
4.3.2 Water physicochemical assessment . . . 107
4.3.3 Water sample processing . . . 108
4.3.4 Viral concentration and DNA extraction . . . 108
4.3.5 Microbial DNA extraction . . . 109
4.3.6 16S rRNA sequencing and analysis . . . 110
4.3.7 Shotgun sequencing for microbial metagenomes and viromes . 111
4.3.8 Microbial metagenomic and virome assembly . . . 111
4.3.9 Microbial and viral taxonomic and functional classification . . 111
4.3.10 ARGs prediction and host assignment . . . 112
4.3.11 Viral Pol I prediction and phylogenetic analysis . . . 113
4.3.12 CRISPRs prediction from microbial metagenomes . . . 114
4.3.13 Statistical analysis . . . 114
4.4 Results . . . 114
4.4.1 Sequencing effort and assembly . . . 114
4.4.2 Temporal variations in physicochemical characteristics and bacterial diversity . . . 115
4.4.3 Temporal variations in bacterial phyla . . . 116
4.4.4 Temporal variations in bacterial genera . . . 117
4.4.5 Microbial functional potential . . . 118
4.4.6 Antibiotic resistance and host taxonomy . . . 119
4.4.7 Viral taxonomic composition . . . 120
4.4.8 Viral Pol I phylogeny . . . 121
4.4.9 Phage-host relationships . . . 121
4.5 Discussion . . . 123

4.6 Figures . . . 129
4.7 Tables . . . 138

5 Metagenomic Analysis of a Freshwater Creek and Irrigated Field Reveals Temporal and Spatial Dynamics in Bacterial and Viral Assemblages 141
5.1 Abstract . . . 141
5.2 Introduction . . . 142
5.3 Materials and Methods . . . 146
5.3.1 Site description . . . 146
5.3.2 Water sample collection . . . 147
5.3.3 Soil sample collection . . . 147
5.3.4 Assessment of water characteristics . . . 148
5.3.5 Water sample processing for microbial metagenomes . . . 148
5.3.6 Water sample processing for viromes . . . 149
5.3.7 Soil sample processing for microbial metagenomes . . . 150
5.3.8 Shotgun sequencing for microbial metagenomes and viromes . 150
5.3.9 Metagenomic assembly . . . 150
5.3.10 Taxonomic and functional classification . . . 151
5.3.11 ORF clustering . . . 152
5.3.12 Identification of ARGs and VFs . . . 152
5.3.13 Prediction and analysis of CRISPRs in microbial metagenomes 153
5.3.14 Statistical tests . . . 153
5.3.15 Data availability . . . 153
5.4 Results . . . 154
5.4.1 Creek water physicochemical characteristics . . . 154
5.4.2 Sequencing effort and assembly . . . 154
5.4.3 Bacterial phyla in water and soil microbial metagenomes . . . 155
5.4.4 Bacterial genera of water and soil microbial metagenomes . . . 156
5.4.5 Creek source water characteristics and bacterial genera abundance . . . 156
5.4.6 Functional potential of water and soil microbial metagenomes 157
5.4.7 Relationships between water characteristics and functional potential . . . 158
5.4.8 ORF clustering in water and soil microbial metagenomes . . . 159
5.4.9 ARGs in water and soil microbial metagenomes . . . 159
5.4.10 Putative hosts of antimicrobial resistance genes . . . 160
5.4.11 Virulence factors in water and soil microbial metagenomes . . 161
5.4.12 Putative hosts of virulence factors . . . 162
5.4.13 Viral taxonomy and ARG/VF in source and point-of-use viromes 163
5.4.14 CRISPR arrays in water and soil microbial metagenomes . . . 164
5.4.15 Phage-host relationships . . . 164
5.5 Discussion . . . 165
5.5.1 Composition and dynamics of bacteria in water and soil samples 166
5.5.2 Diversity and abundance of ARGs in water and soil samples . 168
5.5.3 Diversity and abundance of VFs in water and soil samples . . 169

5.5.4 Phage community structure and interactions in creek water . . 171
5.5.5 Limitations and summary 172
5.5.6 Conclusions 173
5.6 Figures 175
5.7 Tables 186

6 Zero-valent Iron Sand Filtration Reduces Concentrations of Virus-like Particles and Modifies Virome Community Composition in Reclaimed Water Used for Agricultural Irrigation 200
6.1 Abstract 200
6.2 Introduction 201
6.3 Materials and Methods 202
6.3.1 Sample collection 202
6.3.2 ZVI-Sand filter and filtration process 203
6.3.3 VLP quantification 205
6.3.4 Virome preparation 205
6.3.5 Virome DNA sequencing 206
6.3.6 Metagenome assembly and analysis 206
6.4 Results 207
6.4.1 VLP abundance 207
6.4.2 Sequencing effort and assembly 207
6.4.3 ORF clusters 208
6.4.4 Taxonomic assignment 208
6.4.5 Functional assignment 209
6.5 Discussion 209
6.5.1 Limitations 212
6.6 Figures 213
6.7 Tables 217

7 Conclusions, Public Health Significance, and Future Work 219

A Mentholation Affects the Cigarette Microbiota by Selecting for Bacteria Resistant to Harsh Environmental Conditions and Selecting Against Potential Bacterial Pathogens 226

B Temporal Variations in Cigarette Tobacco Bacterial Community Composition & Tobacco Specific Nitrosamine Content are Influenced by Brand and Storage Conditions 264

Bibliography 303

List of Tables

2.1 Descriptions of study sites . . . 61
2.2 Sampling site characteristics by site and date . . . 62
2.3 Sequencing effort and assembly characteristics . . . 63

3.1 Agricultural pond water characteristics during sampling period . . . 98
3.2 Difference (%) in relative abundance between 1 μm and 0.2 μm fractions for the dominant bacterial phyla. . . . 99
3.3 Difference (%) in relative abundance of dominant bacterial phyla between sampling dates in 1 μm and 0.2 μm filter fractions. . . . 100
3.4 Difference (%) in relative abundance between 1 μm and 0.2 μm fractions for the dominant bacterial genera. . . . 101
3.5 Difference (%) in relative abundance between 1 μm and 0.2 μm fraction for the dominant bacterial genera. . . . 102

4.1 Sequencing effort and assembly characteristics for water and soil microbial metagenomes . . . 138
4.2 Descriptive sequencing statistics for viromes . . . 139
4.3 Contig taxonomic assignments for microbial metagenomes . . . 140

5.1 Water physicochemical characteristics . . . 187
5.2 Sequencing effort and assembly characteristics for water and soil microbial metagenomes . . . 188
5.3 Sequencing effort and assembly characteristics for viromes . . . 189
5.4 Taxonomic assignments for contigs from each water and soil microbial metagenome . . . 190
5.5 Bacterial genera assignments for contigs with putative ARGs . . . 191
5.6 Virulence factor descriptions for gene IDs . . . 192
5.7 Bacterial genera assignments for contigs with putative VFs . . . 194
5.8 CRISPR array abundance in soil and water microbial metagenomes . 198
5.9 CRISPR spacers shared with water samples collected at the point-of-use 199

6.1 Sequencing effort and assembly characteristics . . . 217
6.2 Contigs assigned taxonomy . . . 218

A.1 Descriptions of cigarette products tested. 263

B.1 Descriptions of cigarette products tested at three different experimental conditions (pocket, room, and refrigerator) over time (day 0, 5, 9 and 14) 302

List of Figures

2.1 Normalized relative abundance of the bacterial taxa present in reclaimed and untreated surface water sites at each sampling date. . . . 52
2.2 Taxonomic heatmap of the bacterial communities present in reclaimed and untreated surface water sites at each sampling date. 53
2.3 Shared and unique peptide ORFs in reclaimed and untreated surface water sites. 54
2.4 Functional profiles of bacterial communities present in reclaimed and untreated surface water sites at each sampling date. 55
2.5 Taxonomic heatmap of the bacteria associated with pathogenesis GO-terms present in reclaimed and untreated surface water sites at each sampling date. 56
2.6 Antibiotic resistance genes (ARGs) predicted in reclaimed and untreated surface water sites at each sampling date. 57
2.7 Bipartite network of the bacterial taxa with predicted antibiotic resistance genes (ARGs) from reclaimed and untreated surface water sites. 58
2.8 CRISPR array abundance and spacer overlap within and among reclaimed and untreated surface water sites. 59
2.9 Taxonomic origin of contigs containing putative CRISPR arrays. . . . 60

3.1 Bacterial community composition and diversity of 1 and 0.2 μm filter fractions over time 88
3.2 Heatmaps of the Pearson's correlation coefficients between the water characteristics and relative abundance of bacterial (A) phyla and (B) genera for the 1 μm and 0.2 μm filter fractions 89
3.3 Alpha diversity for each filter fraction in late season pond water . . . 90
3.4 Beta diversity for each filter fraction in late season pond water. . . . 91
3.5 PCoA plots of beta diversity (by date) measured using Bray-Curtis. . 92
3.6 Core OTUs for 1 and 0.2 μm filter fractions. 93
3.7 Core *Actinobacteria*, *Proteobacteria*, and *Bacteroidetes* OTUs for 1 μm and 0.2 μm filter fractions. 94
3.8 Viral taxonomy and function for each sampling date. 95

3.9 Heatmaps of the Pearson's correlation coefficients between dominant viral families and the relative abundance of bacterial phyla and genera in both the 1 μm and 0.2 μm filter fractions. 96
3.10 Phylogenic tree of viral Pol I peptides for each sampling date. 97

4.1 Temporal dynamics of physicochemical properties and bacterial diversity in agricultural pond water 129
4.2 Temporal dynamics of bacterial composition in agricultural pond water 130
4.3 Bacterial genera abundance and correlations with physicochemical factors in agricultural pond water across sampling dates 131
4.4 Functional composition in agricultural pond water across sampling dates 132
4.5 Antibiotic resistance genes (ARGs) in agricultural pond water across sampling dates 133
4.6 ARG host network 134
4.7 Viral composition in agricultural pond water across sampling dates . 135
4.8 Cladogram of Pol I peptides across sampling dates 136
4.9 Detection of CRISPR spacers and linkage to viral contigs in agricultural pond water 137

5.1 Bacterial composition in creek source water, water at the point-of-use, and soil 175
5.2 Bacterial genera in creek source water, point-of-use water, and soil . . 176
5.3 Heatmap of the Pearson's correlation coefficients between the water characteristics and normalized abundance of bacterial genera in the creek source water 177
5.4 Functional composition in creek source water, point-of-use water, and soil 178
5.5 Heatmap of the Pearson's correlation coefficients between the water characteristics and normalized abundance of functional GO-terms . . 179
5.6 Shared ORFs at each time point in creek water samples 180
5.7 Shared ORFs at each time point in soil samples 181
5.8 Antibiotic resistance genes (ARGs) predicted in creek source water, point-of-use water, and soil 182
5.9 Virulence factors (VFs) predicted in the creek water source, point-of-use, and soil samples 183
5.10 Taxonomic composition of viral communities from creek water sampled at the source and at the point-of-use 184
5.11 CRISPR spacer persistence in the creek source water 185
5.12 Phage-host network from creek water samples collected at the source and the point-of-use 186

6.1 Epifluorescent microscopy counts of virus-like-particles (VLPs) in reclaimed water (RW) and zero-valent iron sand filtered reclaimed water (ZW) 213

6.2 Peptide ORF clustering in paired reclaimed water (RW) and ZVI-sand filtered reclaimed water (ZW) samples from July and August . . 214
6.3 Taxonomic composition of reclaimed water (RW) and zero-valent iron sand filtered reclaimed water (ZW) 215
6.4 Functional composition of reclaimed water (RW) and zero-valent iron sand filtered reclaimed water (ZW) 216

A.1 Rarefaction curves for each product 248
A.2 Heat map showing the relative abundances of the most dominant bacterial genera identified (>1%) in tested cigarette products 249
A.3 Box plots showing alpha diversity variation across samples on non-rarefied data and with data rarefied to the minimum sampling depth 250
A.4 PCoA analysis plots of Bray-Curtis computed distances between cigarette products 251
A.5 PCoA analysis plots of weighted and unweighted Unifrac computed distances between cigarette products) 252
A.6 Overview of relative abundances of bacterial OTUs that were statistically significantly different between custom-mentholated Camel Kings (CKM) and non-mentholated Camel Kings (CK) 253
A.7 Overview of relative abundances of OTUs that were statistically significantly different between mentholated Camel Crush (CCM) and non-mentholated Camel Crush (CC) 254
A.8 *Pseudomonas* phylogenetic tree 255
A.9 *Acinetobacter* phylogenetic tree 256
A.10 *Stenotrophomonas* phylogenetic tree 257
A.11 *Anoxybacillus* phylogenetic tree 258
A.12 *Deinococcus* phylogenetic tree 259
A.13 *Vagococcus* phylogenetic tree 260
A.14 *Thermus* phylogenetic tree 261
A.15 *Proteus* phylogenetic tree 262

B.1 Bacterial community composition of cigarette products over time and differing storage conditions 287
B.2 Comparison of the relative abundance of the most dominant genera . 288
B.3 Network of core bacterial operational taxonomic units (OTUs) in each brand 289
B.4 PCoA analysis plots of Bray-Curtis computed distances between cigarette products 290
B.5 PCoA analysis plots of Bray-Curtis computed distances between cigarette products 291
B.6 PCoA analysis plots of Bray-Curtis computed distances between individual cigarette products 292
B.7 Alpha diversity comparison by brand, condition, and time point . . . 293

B.8 Overview of relative abundances of bacterial OTUs that were statistically significantly different ($p \leq 0.001$) between day 0 and day 14 for refrigerator, room temperature, and pocket conditions for CC and CCM . . . 294
B.9 Overview of relative abundances of bacterial OTUs that were statistically significantly different ($p \leq 0.001$) between day 0 and day 14 for refrigerator, room temperature, and pocket conditions for CK and CKM . . . 295
B.10 Overview of relative abundances of bacterial OTUs that were statistically significantly different ($p \leq 0.001$) between day 0 and day 14 for refrigerator, room temperature, and pocket conditions for NMB . . . 296
B.11 Overview of relative abundances of bacterial OTUs that were statistically significantly different ($p \leq 0.001$) for NMB . . . 297
B.12 Overview of relative abundances of bacterial OTUs that were statistically significantly different ($p \leq 0.001$) between lots for CK . . . 298
B.13 Overview of relative abundances of bacterial OTUs that were statistically significantly different ($p \leq 0.001$) between lots for CKM . . . 299
B.14 Tobacco-specific nitrosamine levels over time at pocket conditions . . . 300
B.15 TSNA levels in mg/g of tobacco over time at refrigerator conditions . . . 301

List of Abbreviations

ANOSIM	Analysis Of Similarities
ARG	Antibiotic Resistance Gene
ASV	Amplicon Sequence Variant
BP	Base Pairs
BLAST	Basic Local Alignment Search Tool
BSA	Bovine Serum Albumin
CAFO	Concentrated Animal Feeding Operation
CARD	Comprehensive Antibiotic Resistance Database
CDC	Center for Disease Control
CFU	Colony Forming Units
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DO	Dissolved Oxygen
DOM	Dissolved Organic Material
EPA	Environmental Protection Agency
FDA	Food and Drug Administration
FSMA	Food Safety Modernization Act
GM	Geometric Mean
GO	Gene Ontology
GRACE	Gravity Recovery and Climate Experiment
MAFFT	Multiple Alignment using Fast Fourier Transform
MWQP	Microbial Water Quality Profile
ORF	Open Reading Frame
OTU	Operational Taxonomic Unit
PBS	Phosphate-Buffered Saline
PCR	Polymerase Chain Reaction
PSR	Produce Safety Rule
QIIME	Quantitative Insights Into Microbial Ecology
RDP	Ribosomal Database Project

rRNA	Ribosomal Ribonucleic Acid
RW	Reclaimed Water
SDS	Sodium Dodecyl Sulfate
STV	Statistical Threshold Value
TPM	Transcripts Per Million
TSNA	Tobacco Specific Nitrosamines
USDA	United States Department of Agriculture
VF	Virulence Factor
VFDB	Virulence Factor Database
VLP	Viral Like Particles
WWTP	Wastewater Treatment Plant
ZVI	Zero Valent Iron
ZW	Zero Valent Iron Filtered Reclaimed Water

Chapter 1: Introduction

1.1 Global Management of Irrigation Water

1.1.1 Systems, sources, and stresses

Irrigated agriculture is responsible for 40% of the total food produced globally, covering over 275 million hectares of land [1]. However, as the world population grows and the climate changes, competition for water resources is projected to increase, particularly in the agricultural sector [2]. Currently, irrigation for agriculture is one of the largest sectors of global water usage, accounting for upwards of 70% of freshwater withdrawals and 90% of consumptive use (water that is not returned to a resource system) [1,3]. Irrigation-related water sources typically include groundwater and surface water, which account for 38% and 60% of the global agricultural area equipped for irrigation, respectively [4]. While the majority of countries rely on surface water for irrigation, there are about 25 countries in which >50% of agricultural areas equipped for irrigation rely on groundwater. In the United States, groundwater is the chief source of irrigation water in California, Nebraska, Texas, Kansas, South Dakota, and Oklahoma [5]. Additionally, according to the 2013 Farm and Ranch Irrigation Survey, groundwater was reported to account for 55% of on-farm irrigation water applied in 2013 [6,7]. However, intensive groundwater withdrawals have contributed to overdrawn aquifers where water usage exceeds rates of natural replenishment, especially in key food-producing regions around the world [8,9].

A study by Richey et al. 2015 computed renewable groundwater stress for the 37 largest global aquifer systems using data from NASA's GRACE (Gravity Recovery and Climate Experiment) mission [10]. The GRACE satellites collected gravity anomalies over 10 years (2003-2013), which were used to measure monthly changes in total terrestrial water. With these data, they found that more than half of global groundwater aquifers are being depleted, with the California Central Valley Aquifer (-8.87 mm/yr), Atlantic and Gulf Coastal Plains Aquifer (-5.9 mm/yr), Arabian Aquifer System (-9.13 mm/yr), North Caucasus Basin (-16.1 mm/yr), Ganges-Brahmaputra Basin (-19.6 mm/yr), North China Aquifer System (-7.5 mm/yr), and Canning Basin (-9.4 mm/yr) being the most troubled.

Groundwater aquifers are also susceptible to saltwater intrusion that may affect its quality for use in irrigation. In costal regions under normal conditions groundwater and saltwater are separated by a transition zone, a mix of fresh and saline water formed by the seaward movement of freshwater [11]. However, increased groundwater pumping can cause saltwater to be drawn into the freshwater zones of costal aquifers. Saltwater has already intruded into coastal aquifers of the United States, Mexico, and Canada [12] and is expected to be further exacerbated by rising sea levels, reduced precipitation, and higher temperature brought on by climate change [13]. Given this critical situation with regard to groundwater abundance and quality, there is an urgent need to explore alternative irrigation water sources such as reclaimed water (advanced treated wastewater) and untreated surface waters.

1.2 Nontraditional Sources of Irrigation Water

1.2.1 Reclaimed water

Globally, only about 1% of the water used in agriculture is considered non-traditional, largely in the form of treated wastewater (also referred to as reclaimed water) or desalinated water [14, 15]. However, in some regions of the world the rate is much higher, especially for arid and semi-arid areas [14]. Israel has the highest national percentage of wastewater reuse, wherein over 80% of treated wastewater effluent is reused, largely for agricultural irrigation (~60%). In Europe, wastewater reuse for agricultural irrigation is growing, especially in Mediterranean countries such as Spain, which reuses as estimated 17-20% of its wastewater [16].

In the United States, approximately 7-8% of wastewater is reused [17]. According to the U.S. Department of the Interior and U.S. Geological Survey in 2015 reclaimed water was reported for use in irrigation in: California, Florida, Arizona, Texas, Utah, Nevada, New Mexico, Colorado, Kansas, and Illinois [5]. Of these California, is one of the leaders in water reuse in the U.S., with an estimated 13% of its 5 million acre-feet of municipal wastewater produced each year being reused [18]. Reuse applications in California include: agricultural irrigation (37%), urban irrigation (23%), groundwater recharge (19%), commercial and industrial (9%), recreational (4%), environmental (4%), geothermal energy (2%), and other (2%). While most of the wastewater reuse projects are located largely in arid and semi-arid regions, climate change and growing urbanization has hastened development of wastewater reuse projects in other areas, such as the Mid Atlantic and Northeast. In fact, the U.S. Environmental Protection Agency (EPA) 2004 Guidelines for Water Reuse reports 26 states have regulations for water recycling and 15 states have guidelines [19].

Reclamation facilities or wastewater treatment plants (WWTP) can vary enormously with regard to their treatment practices. However, in general they employ between two and three water treatment stages. During the first few stages of treatment (preliminary/primary) influent (the untreated wastewater or raw sewage) is passed through a screen to remove large debris and into a grit chamber where solids (e.g. stones, sand etc.) settle to the bottom. The wastewater then enters a sedimentation tank where suspended solids sink to the bottom. After leaving the settling tank, the wastewater enters the secondary treatment stage where it is pumped into an aeration tank and mixed with air and sludge (raw primary biosolids). Here, bacteria degrade and remove dissolved organic matters and inorganic nutrients. This partially treated wastewater then flows into another sedimentation tank. In some cases, tertiary or advanced treatment is carried out to further remove suspended particles, specific pollutants (e.g. nitrogen, phosphorous), and microorganisms. Here, technologies such as chlorination, sand filtration, microfiltration, ultrafiltration, reverse osmosis, chemical coagulation, ozone, UV light, activated carbon absorption, and/or ion exchange can be employed [20]. Once treated, the effluent is then discharged into the environment, either through release into natural waterbodies (e.g. rivers), through groundwater recharge or through reuse of the water in downstream applications including non-potable and potable uses.

1.2.2 Untreated surface water

In addition to reclaimed water, the agricultural use of untreated surface water sources such as ponds, brackish rivers, and creeks proximal to agricultural facilities may also help to attenuate the burden on diminishing groundwater aquifers. The 2013 Farm and Ranch Irrigation Survey reported on-farm surface water accounted for 10% of irrigation water and included both lentic (standing bodies of water) and lotic (flowing bodies of water) sites [6, 7].

1.2.2.1 Ponds

Lentic ecosystems that are often employed as a means of capture and storage of freshwater for localized irrigation are ponds. Ponds are common features across the United States, with estimates between 2.6 and 9 million, outnumbering larger lakes by a ratio of about 100 to 1 [21, 22]. They are generally defined as small (1 m^2 to 50,000 m^2), shallow, standing water bodies that can either permanently or temporarily collect water [23–25]. However, there is no universally accepted definition of a pond, and limnologists argue that pond criteria should also encompass depth (max of 8 m), tidal forces (none), and wave action (none) [21]. Nevertheless, ponds can occur naturally (e.g. floodplains, isolated depressions), but are often human-constructed for a variety of utilitarian and aesthetic purposes [22]. Some anthropogenic uses of ponds include aquaculture, wastewater treatment, waste stabilization, flood alleviation, storage of urban storm water, capture and storage of freshwater for irrigation, and urban heat mitigation [26, 27]. Aside from human associated purposes, ponds are also valued for their important ecological roles. Ponds contribute to Earth's biogeochemical cycling, estimated to sequester carbon amounts comparable to the global oceans [28]. Moreover, they are critical in supporting a rich tapestry of aquatic plant and macroinvertebrate species, even greater than that of other larger water bodies, such as lakes and rivers [29].

1.2.2.2 Rivers and creeks

In the United States, there are over 3.5 million miles of lotic ecosystems, including rivers, creeks and streams [30]. They are generally defined as linear landforms with clearly defined bed and banks that permanently or temporarily carry a concentrated flow of water [31, 31, 32]. These ecosystems serve a variety of utilitarian purposes including, irrigation, electricity generation, recreation, routes for navigation, and waste disposal. However, similar to ponds, the criteria that defines the varied lotic ecosystems is not universal. The U.S. Board on Geographic Names considers all "linear flowing bodies of water" as streams. However, within this category there exists at least 121 other generic terms. While some general observations are used to differentiate lotic water bodies, such as size (creeks are generally <8.25 m) and flow direction (creeks flow into rivers), they are not universally accepted and are often indicative of local or regional characteristics [31, 31, 32]. Nevertheless, lotic sites, generally rivers, begin at a source (e.g. lake, marsh, spring, glacier), then follow a path (course) from a higher altitude to a lower altitude until it ends at a mouth or mouths (generally at an ocean, sea, or lake) [33]. Along the way, they are often fed by tributaries, which can be other rivers, streams, and/or creeks. As a result of this topography, lotic sites can be impacted by connected waterways, as well as their catchment area, which tends to be greater than that of lentic waters and can encompass a variety of different area types (e.g. agricultural, urban, forested) [34]. Furthermore, lotic sources are subjected to point source and non-point source pollution, habitat degradation, and hydrological changes brought on by human-associated flow modifications [35].

1.2.3 Irrigation water quality guidelines and regulations

To ensure food safety, some countries have published guidelines on appropriate microbial water quality criteria for surface and/or reclaimed water to be used for irrigation [36]. In the U.S. the EPA has published water quality guidelines for the use of reclaimed water [19]. For crops that are intended for human consumption (consumed raw) the EPA standards for reclaimed water applied via surface or spray irrigation are as follows: pH = 6-9, no detectable fecal coliform per 100 mL of water (seven-day median value), $\leq$ 10 mg/L biochemical oxygen demand (5-day BOD test), and an average turbidity of $\leq$ 2 NTU based on a 24-hour time period. Using these guidelines each state government establishes their own water quality standards, which vary in scope. For instance, California recommends a seven-day median value of 2.2 colony forming units (CFU) per 100 mL of water, with a maximum value of

240 CFU per 100 mL for reused water intended to be used on food crops [37].

In addition to these standards, President Obama signed the Food Safety Modernization Act (FSMA) into law on January 4, 2011, which established standards for irrigation water used for agricultural application. Since the enactment of FSMA, the Food and Drug Administration (FDA) has been working to develop the final rules that the act requires them to implement. Under the Produce Safety Rule (PSR, 21 CFR 112), farmers are required to test agricultural water for generic *Escherichia coli*, an indicator of fecal contamination, to form a microbial water quality profile (MWQP) [38]. Using a rolling four-year data set, the MWQP, with at least four samples for groundwater sources and at least 20 samples for surface water sources, is used to produce two statistical calculations, a geometric mean (GM) and a statistical threshold value (STV). The GM measures the central tendency or the average amount of *E. coli* in a water source. The STV measures the expected deviations of the *E. coli* levels from the average. For agricultural water that is directly applied to growing produce (besides sprouts) the GM of samples must be 126 or less CFU of generic *E. coli* per 100 mL of water and the STV of samples must be 410 CFU or less of generic *E. coli* in 100 mL of water. If the water does not meet these criteria, the farmers have time (within a year) to employ corrective measures, such as water treatment (e.g. sanitizers, disinfectants) or implementing an appropriate die off interval (e.g. a die-off rate of 0.5 log CFU per day between last irrigation and harvest). Despite these strengthening guidelines, FSMA only focuses on fecal coliforms, which is just one aspect of the varied microbial community present within these potential irrigation sources.

1.3 Bacteria and Viruses in Nontraditional Irrigation Water Sources

By definition, microbes, or microorganisms, are microscopic organisms that may exist in a single-celled form or in a colony of cells and include bacteria, archaea, microeukaryotes, fungi, and viruses. However, for this dissertation I will focus primarily on bacteria and their viruses (bacteriophage).

1.3.1 General role in the biosphere

Bacteria are single celled microorganisms that are thought to be one of the first forms of life on the planet [39]. In fact, *Cyanobacteria* or "blue-green algae" are oxygen-producing bacteria that are believed to be responsible for the initial production of atmospheric oxygen [40]. Today bacteria are found in nearly every biome on Earth, present in areas as extreme as acidic hot springs and deep subsurface environments [41]. Generally, there are $\sim 10^6$ bacteria per mL in lake and ocean water and $\sim 10^9$ per gram in sediment [42]. The ubiquity of these microbes is matched only by their functional potential, with bacteria capable of sourcing energy from sunlight (e.g *Cyanobacteria*), inorganic compounds (e.g *Nitrospirae*), and organic compounds (e.g. *Enterobacteriaceae*). Because of this wide range of metabolic potential, including many variations of heterotrophy and autotrophy, bacteria are critical mediators in biogeochemical/nutrient cycling (e.g. phosphate, carbon, nitrogen) and, as a result, form complex relationships with other organisms [42, 43]. In soil, some bacteria fix nitrogen gas into ammonium, which can then be easily absorbed by terrestrial plants. Moreover, in the oceans diazotrophs fix roughly 140 x

10^{12} g N y^{-1} [42], nearly equivalent to the that of the 120 x 10^{12} g N y^{-1} produced by fertilization manufacturing (Haber-Bosch) [44]. In aquatic systems, bacteria (mainly phytoplankton) are situated at the base of the food web, supporting the growth of consumers and maintaining a healthy ecosystem [42].

Bacteria can also form mutualistic relationships in and on humans and animals. For instance, ruminants (e.g. cows, sheep) depend on a complex community of bacteria to aid in the breakdown of polysaccharides [45]. In humans, while pathogenic bacteria are responsible for illness (discussed below), commensal bacteria play many roles, including immune system maturation, vitamin synthesis, digestion, and the competitive exclusive of foreign bacteria [46].

While not technically alive, viruses are essential biological components in microbial communities. Environmental viruses are ubiquitous and extremely abundant, ranging from 10^7 per mL in natural aquatic habitats to 10^{10} per gram in sediments [47, 48]. This vast profusion of viral particles contains a wide array of viral groups, including those capable of infecting amoeba, plants, fungi, and vertebrates. However, in most environments, bacteriophage (phage), viruses that infect bacteria, dominate [49]. After phage infection, the bacteria's fate is determined by the replication cycle of the phage. Virulent phage replicate only through the lytic cycle, whereas temperate phage replicate with both the lytic and lysogenic cycles. These different phage life cycles may determine the extent of selective pressure that bacteria are under by their phage. In the lytic cycle, phage infect their host, take over the cell's biochemical machinery, and begin rapidly producing progeny until cell lysis. This can result in the diversification and evolution of the bacterial hosts.

Additionally, phage lysis results in the release of the host's internal cellular contents (e.g. organic carbon, nitrogen), which then become a part of the pool of dissolved organic material (DOM). This phenomenon, known as the viral shunt, increases the level of available DOM for other microbes and is suggested to promote bacterial respiration and growth [50–52]. Conversely, in the lysogenic cycle, phage integrate their DNA into the bacterial chromosome and replicate passively with the host until inducing signals (e.g. UV light) drive transition to the lytic cycle. This life cycle can influence the hosts' phenotype through horizontal transfer of genes, such as those for toxins [53], as well as those that promote host fitness and adaptability (e.g. energy metabolism [54], platelet adhesion [55], antibiotic resistance [56]).

1.3.2 Risks to public health

1.3.2.1 Food-borne pathogens and outbreaks

While microorganisms are essential in maintaining the biosphere, they can also be responsible for negatively impacting environmental and public health, especially when augmented by anthropogenic activities. The Centers for Disease Control and Prevention (CDC) reports each year in the United States that approximately 48 million people are sickened, 128,000 are hospitalized, and 3,000 people die from food associated pathogens [57]. Of the eight known pathogens that account for the majority of foodborne illness, hospitalization and death, six are bacterial: *Salmonella*, *Clostridium perfringens*, *Campylobacter*, *Staphylococcus aureus*, *E. coli* O157:H7, and *Listeria monocytogenes* [57]. This raises a major concern when utilizing untreated surface and reclaimed waters for agricultural application as a variety of these pathogens have been identified in water surveillance studies across the United States [58–61]. For example, *Salmonella* has been identified in surface water samples from the Virginia Eastern Shore (*S.* Newport, *S.* Javiana) and Central Florida (*S.* Muenchen, *S.* Rubislaw, *S.* Anatum, *S.* Gaminara, *S.* IV_50:z4,z23), as well as in the Little River watershed (*S.* Muenchen, *S.* Rubislaw) and Suwannee watershed (*S.* Newport, *S.* Enteritidis, *S.* Muenchen, *S.* Javiana, *S.* Thompson) in Georgia [58–61].

Moreover, some of these pathogenic bacteria have demonstrated the ability in field studies to persist on crops for weeks following irrigation [62–64]. Studies investigating spray irrigation with contaminated water found *E. coli* O157:H7 to persist on lettuce anywhere from 15 [63] to 27 [64] and 56 days [62]. *Salmonella enterica* was also found to persist on parsley after spray irrigation for four weeks [65]. As a result, it is not surprising that irrigation water has been identified as an entry of pathogenic bacteria into the food production chain, which has led to major multi-state outbreaks of illness in the U.S.

In the early 2000s, two outbreaks of *Salmonella* Newport on sliced tomatoes, which caused illness in over 500 patients, were traced back to irrigation water from agricultural ponds on the Eastern Shore of Virginia [66]. An investigation into an outbreak (205 cases) from 2006 of *E. coli* O157:H7 in prepackaged spinach traced the source to nearby river water, which acted as a vector between animal fecal runoff and the irrigation wells used on the crop [67]. More recently, the CDC identified tainted canal water from the Yuma growing region of Arizona as the likely source of a 2018 *E. coli* O157:H7 outbreak of romaine lettuce that left 210 people ill and 5 dead [68].

1.3.2.2 Phage mediated transduction

While they are not direct human pathogens, phage are responsible for shaping the diversity and genetic architecture of their hosts, which can ultimately impact environmental and public health. For example, Stx phage, lambdoid bacteriophage encoding the Shiga-like toxin producing genes (*stx1* and *stx2*), confer pathogenicity to *E. coli* O157:H7 through prophage integration [69, 70]. This is also the case for the toxin genes in *Bordetella avium* (Pertussis toxin), *Clostridium botulinum* (Botulinum toxin), *Corynebacterium diphtheria* (Diphtheria toxin), *Pseudomonas aeruginosa* (cytotoxins), *Shigella dysenteriae* (Shiga toxin), *Staphylococcus aureus* (enterotoxins, exfoliative toxins, Toxic shock syndrome toxin), *Streptococcus pyogenes* (erythrogenic, scarlatinal exotoxins), and *Vibrio cholera* (Cholera toxin) [71], which are all phage encoded. In addition to these toxins, phage have been reported to encode genes that alter bacterial colonization, adhesion, invasion, transmission, and antibiotic resistance [71].

Antibiotic resistance occurs when bacteria are able to withstand the presence of antibiotics either through genetic mutation or by acquiring antibiotic resistance genes (ARGs) from other bacteria. For the latter, conjugation, transformation, and transduction are the primary mechanisms of gene acquisition [72]. As a result, phage, the chief vectors of bacterial transduction, are becoming increasingly recognized for their potential role in the dissemination of ARGs [73]. For example, the horizontal transfer of penta-resistance in *Salmonella typhimurium* DT104 is hypothesized to be facilitated by two P22-like phage [74]. Similarly, Bearson et al. 2014, reported that exposure of *Salmonella enterica* serovar Typhimurium to the veterinary antibiotic, carbadox, induces prophage that could then transfer virulence and ARGs to susceptible hosts [75]. In fact, ARGs have been identified *in vitro* on the lysogenic phage of a number of other human pathogens, including *Bacillus anthracis* [76], *Staphylococcus aureus* [77], and *Salmonella enterica* [78]. Despite this, the predominance of phage containing ARGs *in vivo* is still uncertain. Metagenomic and qPCR surveys of viral populations have identified putative phage harboring ARGs in human lungs [79], hospital wastewater [80], activated sludge [81], urban impacted and nonimpacted river water [82], among others. However, this is not always consistent across studies and may be predicated on the analysis methods utilized [83].

1.3.3 Methods for identification and detection

The methods employed to investigate bacteria and viruses, including phage, have evolved drastically over the centuries beginning with Anton van Leeuwenhoek, in the 17th century who was the first to observe microorganisms under a microscope of his own design [84]. Since then, two prevailing methods have been widely employed: culture and sequencing.

1.3.3.1 Culture-based

For traditional culture-based studies, samples are inoculated on a range of media and incubated to test for the presence of specific or general bacteria. Selective media and biochemical tests can also be employed to enhance identification. Microbiological water quality is traditionally screened by culture-based identification and enumeration of fecal coliforms. This is largely conducted by either membrane filtration or multiple-tube fermentation [85]. With the former, water samples are passed through membrane filters and placed on growth media in a Petri plate. Following incubation, bacterial colonies can be counted directly from each plate. In multiple-tube fermentation, samples are incubated in tubes containing nutrient broth and monitored for the development of gas and/or turbidity [85]. While total coliform counts are still widely used to assess water quality, as depicted by their use as criteria indicators in the EPA water quality guidelines and FSMA, there are some limitations to culture-based methodologies. The most notable is the observation that most bacteria cannot be grown in culture. Despite the fact that culturing technologies and methodologies have advanced over the decades, only a fraction (1-2%) of bacteria can be grown in the laboratory today [86]. This makes culturing phage all the more difficult, considering that you must first properly culture its bacterial host.

For phage, plaque assays were developed to obtain abundance and infectivity data. In this method bacterial host cells are grown on an agar plate until a continuous lawn is formed. Following this, a phage-containing sample can then be poured overtop and allowed to incubate. Any holes or plaques that form during the incubation are then counted as one phage, which infected its host, replicated, and lysed the cells. However, this method neither differentiates phage species nor provides data on the community structure.

1.3.3.2 Sequence-based

An alternative to (or in addition to) cultured-based identification is sequence-based methodologies. One of the first popular sequencing technologies was developed by Fred Sanger in the 1970s and relied on chain termination [87]. Although some updates have been made, Sanger sequencing is still used today. Briefly, a mixture is generated that contains the following: the DNA of interest, a primer (a short piece of single-stranded DNA that binds to the DNA of interest), DNA nucleotides, dideoxynucleotides (chain-terminating versions of DNA nucleotides each labeled with a different color dye), and a DNA polymerase. The mixture is heated and cooled to allow the primer to bind to the DNA of interest and the polymerase to synthesize new DNA. The polymerase will add new nucleotides until it incorporates a dideoxynucleotide and the chain is terminated. This process is then repeated until a dideoxynucleotide has been incorporated at every position from the DNA of interest. The resulting fragments are run through a gel matrix and then excited with a laser. The fluorescence intensity is recorded and used to determine the DNA nucleotide at each position. From this methodology, long stretches of DNA, about 700-800 base pairs (bp), could be assessed. However, since the early 2000s, Sanger sequencing has been eclipsed, in most studies, by next-generation sequencing technologies, such as pyrosequencing (e.g. 454), Single Molecule Real-Time (SMRT) sequencing (e.g. PacBio), and sequencing by synthesis (e.g. Illumina), which can generate high-throughput data rapidly and at a lower cost for multiple samples.

Currently, the most widely utilized platform for DNA sequencing is the suite of Illumina sequencers (e.g. Illumina MiSeq, HiSeq). These technologies function by cluster generation and then sequencing by synthesis [88, 89]. First, the DNA fragments of interest are immobilized on a flow cell, which is then subjected to solid-phase amplification to create copies of each single fragment in close proximity. The "clusters" of DNA are then denatured, allowing a primer and polymerase to anneal and begin incorporating fluorescently tagged nucleotides, one per cycle. Base calls are then made directly from the signal intensity at each cycle. From this, Illumina technologies can sequence hundreds of millions of sequence reads per run.

The collection of DNA to be sequenced is termed a "library". DNA libraries are derived from one of two DNA survey strategies: targeted and shotgun sequencing. Targeted sequences can be selected by PCR with specific primers. For bacteria, the most commonly sequenced marker gene is 16S rRNA; a gene that is present in all bacteria and archaea and codes for the small subunit of ribosomal RNA [90]. The 16S rRNA gene can provide taxonomic and phylogenetic data on bacteria in a complex sample without cultivation. This is in large part due to the different amounts of sequence polymorphism that exist throughout the gene, which range from highly conserved to highly variable regions of DNA. The conserved regions serve as targets for PCR primers, while the variable regions are used to distinguish bacterial groups from one another.

Once amplified, the 16S rRNA gene PCR products are sequenced and then analyzed computationally. In general, quality sequences are clustered by similarity (generally 97%) into Operational Taxonomic Units (OTUs). From each cluster a representative sequence is chosen and assigned a known taxa, which is then applied to all of the sequences within that cluster. Alternatively, a finer-scale equivalent of OTUs can be generated known as amplicon sequence variants (ASVs). Essentially, the ASV approach forgoes the 97% clustering in favor of utilizing unique, identical 16S rRNA sequences for downstream community analyses. With this method, ASVs can differ by as little as one base pair and, thus, is thought to improve taxonomic resolution [91]. Despite the tremendous amount of insight 16S rRNA gene studies have provided on bacterial community structure and diversity, there are some limitations. One well-known problem is the discrepancy in 16S copy number, which vary among bacteria. For example, *Photobacterium profundum* has 15 copies of the 16S rRNA gene compared to the average of four copies in other bacteria [92, 93]. This discrepancy limits the accuracy of using 16S to estimate bacterial abundance and diversity. Additionally, biases can be introduced at the PCR stage due to primer-binding efficiencies, which can lead to under/over representation of some species or taxa.

Moreover, viruses are not known to carry the 16S rRNA gene and, thus, it cannot be used to describe their diversity and dynamics. Consequently, several structural and functional genes have been used as markers of phage diversity for specific groups, including structural genes *g20* and *g23* for *Myoviridae* [94, 95] and

DNA Polymerase A (*polA*) for *Podoviridae* [96,97] However, due to their polyphyletic nature, there are no genes universally conserved among all viruses and, as a result, shotgun metagenomics is currently the best method to explore unknown viruses [98, 99].

In contrast to the single gene approach, shotgun sequencing enables researchers to sequence the entire genome of a single organism or multiple organisms of interest. When shotgun sequencing the genome of a microbe of interest, the DNA is extracted from the isolated organism, randomly sheered, and then randomly sequenced by the machine. Shotgun sequencing of a metagenome works on the same principle except that all of the DNA from the mixture of organisms within a sample is sheered and sequenced together. Consequently, the metagenomic data are usually more computationally challenging to analyze than data derived from amplicon surveys. Some analysis methods/tools (e.g. MetaPhlAn2 [100]), sourmash [101]), kracken [102]) can utilize short sequence reads, but most analyses will begin with assembly. There are two broad methods for assembly, reference-based and *de novo*.

In both cases, short sequences (reads) are assembled together to form long contiguous sequences (contigs) that represent the original piece of DNA extracted from a culture, or a consensus sequence from a population. For reference-based assembly, a reference genome of the organism of interest is used as a guide, while in *de novo* no reference is used [103]. After assembly, open reading frames (ORFs) can be predicted on the contigs, largely by tools (e.g. MetaGene [104]) that are able to predict the start and stop positions of genes, as well as ribosomal binding sites. The resulting putative peptide ORFs can then be assigned taxonomic and functional features generally through sequence homology searches (e.g. BLAST [105]) against a database of interest (e.g. UniRef [106], CARD [107]).

Additionally, by aligning reads used in the assembly to the contigs we can estimate the number of reads used to build the contig. This, along with the contig length, can be used to calculate coverage of the contigs and serves as an estimate of the abundance of that contig in the sample, a valuable metric for describing community structure. It is important to note, however, that even though metagenomic studies have been a useful tool for almost a decade, the framework for analyzing these data are still underdeveloped and analyses can vary widely depending on the study details, questions asked, and computation resources available.

1.3.4 Taxonomic composition and dynamics

After sequencing, scientists can use the data to describe, hypothesize about, and catalog microbes in an environment. While not as broadly studied as large freshwater lakes and marine systems, nontraditional sources of irrigation water have been investigated in previous studies to determine the bacterial and viral community composition.

1.3.4.1 Waste to reclaimed water

16S rRNA gene and metagenomic studies on wastewater treatment and reclaimed water have been a growing area of interest. During treatment the bacterial community composition of wastewater undergoes dynamic changes. Wastewater enters a treatment facility as influent where it has been characterized to be dominated by *Proteobacteria*, as well as *Actinobacteria*, *Bacteriodetes*, and *Firmicutes* [108,109]. However, the bacterial community composition of the influent may vary based on local population demographics, industry, and climate. For instance, several studies have identified *Gammaproteobacteria* as the dominant class of *Proteobacteria* in wastewater [108, 109]. However, this is not consistent across all studies, with some identifying *Deltaproteobacteria* [110] or *Betaproteobacteria* [111] as the most abundant [110]. Nevertheless, the dynamics of the influent bacterial populations vary as they progress through the treatment pipeline depending largely on the processes employed by the individual treatment facility. A previous study from our lab utilized 16S rRNA gene sequencing to characterize total bacterial communities present throughout the treatment process [112]. From these data, we observed that the overall bacterial alpha diversity decreased after treatment, but then increased after open-air storage in a pond. Moreover, some potentially pathogenic genera were still present in the treated water, such as *Legionella* spp., *Mycobacterium* spp., and *Streptococcus* spp.

Previous studies on viruses in reclaimed water have focused, in large part, on human enteric viruses and plant RNA viruses. However, there have been a few studies that have reported the abundance and diversity of dsDNA viral communities during wastewater treatment [113] and in reclaimed water [114]. In reclaimed water collected at the point-of-discharge from a WWTP in Florida, there were a staggering 1000 times higher abundance of virus-like particles in reclaimed water than in potable water (10^8-10^{10} VLPs per mL), largely phage from the family *Siphoviri-*

dae [114]. *Siphoviridae* are a generally temperate family of phage also found to be abundant in sediments [115], terrestrial subsurface environments [116], and human feces [117]. Similar to the bacterial population, during the various stages of wastewater treatment the viral community has been reported to change. Initially, influent delivered to a domestic municipal WWTP in Singapore was found to be dominated by dsDNA phage of the order *Caudovirales*: *Myoviridae* (38% of the total assigned reads), *Podoviridae* (29%), and *Siphoviridae* (27%) [113]. However, during activated sludge treatment the abundance of *Podoviridae* increased to 45%, which was still prevalent in the resulting effluent (38%).

1.3.4.2 Lentic freshwater

Freshwater lentic environments are critical sources of drinking, recreation, and irrigation water in the U.S. and, like reclaimed waters, have been previously surveyed by 16S rRNA gene and metagenomic sequencing [118, 119]. However, despite the evidence that they are functionally different from lakes and "hotspots" of macrodiveristy, freshwater ponds have been left principally unexplored outside of extreme environments (e.g. saline/hypersaline [120–122], thermokarst [123] ponds) and aquaculture facilities [124, 125]. As a result, the majority of what we know regarding the bacterial community composition of freshwater bodies stems from studies conducted on other lentic ecosystems, such as large lakes. Previous studies have reported a widespread distribution of *Actinobacteria* (lineage acI and acIV) in freshwater lakes, in some cases composing greater than 50% of the total bacterial community abundance [119, 126]. The *Proteobacteria* phyla in these lakes are also abundant, composed largely of *Betaproteobacteria*. This is in contrast to marine and brackish systems where *Alphaproteobacteria* typically dominates [119]. Furthermore, freshwater microbial communities are influenced by a variety of environmental factors, such as pH, temperature, water retention time, and seasonal forces [127, 128], as well as top-down regulation from predators (e.g. bacteriophage) [129, 130].

Similar to most bacterial analyses, the majority of studies of viruses are limited to marine samples [131, 132]. However, there have been some analyses of viral communities originating from large freshwater lakes in the Artic [133], Antarctic [95], France [134], North America [135–138], and the United Kingdom [139], as well as freshwater reservoirs in Taiwan [140], China [141], and Signapore [142]. While limited in their scope, these studies have provided some of the first data on freshwater phage ecology, demonstrating that, like their hosts, phage diversity is influenced by environmental factors. For instance, viral genotype diversity, has been found to increase in the summer in a freshwater Antarctic lake [95] and subtropic reservoir [140], likely due to increasing host activity prompted by rising temperatures and nutrient load [143, 144]. Moreover, these studies have identified the dominance of traditionally virulent dsDNA phage families of *Myoviridae* and *Podoviridae* in freshwater lakes Ontario, Erie, Matoaka, and Michigan [136, 137, 145] and freshwater reservoirs in Taiwan, China, and Signapore [140–142].

1.3.4.3 Lotic fresh and brackish waters

Flowing bodies of water are also understudied with regard to their microbial ecology, especially compared to marine and lake ecosystems [146]. However, there exists a growing body of literature that has begun to detail the bacterial community composition of major riverine systems in the U.S. and abroad. For instance, the freshwater James River in Virginia, was described to be dominated by *Proteobacteria*, predominantly from the family *Comamonadaceae* and genera *Acidovorax* [147]. In contrast to ponds, lotic ecosystems can also be impacted by upstream actions and connected waterways. As previously shown, discharge of wastewater effluent into Chicago urban and suburban rivers was reported to increase organic and inorganic nutrient content downstream, decreasing overall bacterial diversity in the sediment [148]. An additional study comparing the bacterial communities within the Santa Ana River watershed impacted by various pollutant sources (agricultural, urban, and wastewater) found that all of the samples were dominated by *Proteobacteria* [149]. However, sites impacted by agricultural runoff were reported to have a greater abundance of *Bacteroidetes* and *Cyanobacteria*. Moreover, other environmental factors (largely driven by seasonality, flow rate, and/or land cover) such as nutrient load [150], pH [150], salinity [149, 151], temperature [149, 150], DO [149], and turbidity [149] have demonstrated an effect on the bacterial community structure in lotic sites. A year-long study surveying three rivers in southwestern British Columbia, Canada found that metagenomes clustered by water chemistry, even when collected from unconnected watersheds 130 km apart [152].

Viruses in lotic ecosystems have only been surveyed in a handful of studies, mostly of large riverine watersheds. DNA viral metagenomes from the brackish water of the Jiulong River Estuary in China found *Caudovirales* to be the most dominant viral type, accounting for 84% of the sequences, with *Podoviridae* being the most abundant family (45%), followed by *Siphoviridae* (33%) and *Myoviridae* (17%) [153]. In contrast, DNA viral metagenomes constructed from the freshwater Murray River upstream and downstream of a small rural town in South Australia were dominated by *Myoviridae* at both upstream (30%) and downstream locations (34%), followed by *Siphoviridae* (23% upstream, 32% downstream) [154]. This is consistent with other freshwater river systems (e.g. Bess River in Spain [155], Amazon River in Brazil [156], Ile River in China [141]), in which *Myoviridae* dominated the *Caudovirales* order.

1.4 Outline of Dissertation

Climate variability and growing urbanization have placed immense pressure on the finite supply of water for agricultural irrigation. As a result, the exploration of nontraditional sources of irrigation water (e.g. reclaimed, pond, brackish) and water treatment technologies (e.g. zero valent iron filtration, ozone) has become a global priority. However, irrigation with untreated surface and reclaimed water may pose a risk to environmental and public health. This is because surface water sites are subjected to environmental and anthropogenic variables that can largely be avoided when using properly constructed and protected deep aquifers (e.g. animal fecal contamination and runoff water from adjacent fields). Similarly, irrigation with reclaimed water may disseminate pathogenic microorganisms and/or antibiotic-resistant bacteria originating from wastewater. Therefore, assessing the microbial community composition in these nontraditional irrigation water sources is critical with regard to completing a comprehensive characterization of their biodiversity, and evaluating their suitability for agricultural applications.

The purpose of my dissertation research was to use high-throughput, cultivation-independent sequencing methodologies to produce novel data on bacterial and viral community structure, dynamics, and potential dissemination that can be utilized to improve our understanding of nontraditional sources of irrigation water. My primary research objectives and the chapters within which they are addressed are as follows:

Objective 1: Explore the functional and taxonomic features of bacteria in nontraditional irrigation water sources. (Chapter 2)

Objective 2: Assess the bacterial and viral communities of agricultural pond water in relation to seasonality. (Chapter 3, 4)

Objective 3: Describe the dynamics, composition, and potential dissemination of irrigation water microbiota from a freshwater creek to an irrigated field. (Chapter 5)

My first research objective is addressed in Chapter 2, which presents a manuscript entitled "Comparative metagenomic analysis of microbial taxonomic and functional variations in untreated surface and reclaimed waters used in irrigation applications", currently in review in *Water Research*. This chapter explores microbial metagenomes from multiple nontraditional sources of irrigation water collected over two months. Here, the bacterial composition, functional potential, and antibiotic resistance profiles are explored across diverse irrigation water sources. Additionally, viral populations are addressed by leveraging phage signals (e.g. CRISPR arrays) present within the bacterial genomes.

Chapter 3 addresses part of my second objective in a manuscript entitled "Agricultural freshwater pond supports diverse and dynamic bacterial and viral populations" published in *Frontiers in Microbiology* [157]. This chapter describes the bacterial and viral populations within a freshwater agricultural pond throughout the late season (October-November) using 16S rRNA gene and shotgun metagenomics and assesses variations in these populations with regard to environmental factors such as dissolved oxygen, pH and nitrate levels.

Chapter 4 extends the study detailed in the previous chapter in a manuscript entitled "Seasonal dynamics in taxonomy and function within bacterial and viral metagenomic assemblages recovered from a freshwater agricultural pond". Here, the same freshwater pond is surveyed over a full calendar year using shotgun metagenomics to characterize the viral fraction and a combination of 16S rRNA gene and shotgun sequencing to survey the bacterial fraction. This chapter focuses in detail on the bacterial community, its relationship to the viral community, as well as its variability with regard to environmental factors.

My final objective is addressed in Chapter 5 with a manuscript entitled "Metagenomic analysis of a freshwater creek and irrigated field reveals temporal and spatial dynamics in bacterial and viral assemblages". This chapter focuses mainly on the bacterial communities present in a creek from the Mid Atlantic, United States throughout two growing seasons. Here, the bacterial composition, functional potential, and antibiotic resistance and virulence profiles are explored. Additionally, this chapter provides preliminary data on the bacterial and viral populations present at the point-of-use (drip irrigation spigot), as well as bacterial populations present in the soil before and after an irrigation event.

To conclude the research chapters of this dissertation an additional research note is presented in Chapter 6, entitled "Zero-valent iron sand filtration reduces concentrations of virus-like particles and modifies virome community composition in reclaimed water used for agricultural irrigation" in review at *BMC Research Note*. This chapter provides pilot data on zero-valent iron filtration (ZVI-filtration), a technology poised to aid in the remediation of reclaimed water, through the use of metagenomic sequencing and epifluorescent microscopy on reclaimed water (RW) and ZVI-sand filtered reclaimed water (ZW) sampled three times over 49 days.

Finally, Chapter 7 presents the conclusions, public health significance, and future work emerging from the research described herein.

Included in Appendix A and B are two manuscripts that were used to develop my skills in 16S rRNA gene sequencing and analysis, as well as manuscript development, formatting, and publication. The first is entitled "Mentholation affects the cigarette microbiota by selecting for bacteria resistant to harsh environmental conditions and selecting against potential bacterial pathogens" published in *Microbiome* [158] and the second is entitled "Temporal variations in cigarette tobacco bacterial community composition and tobacco-specific nitrosamine content are influenced by brand and storage conditions" published in *Frontiers in Microbiology* [159].

Chapter 2: Comparative Metagenomic Analysis of Microbial Taxonomic and Functional Variations in Untreated Surface and Reclaimed Waters Used in Irrigation Applications

2.1 Abstract

The use of irrigation water sourced from reclamation facilities and untreated surface water bodies may be a practical solution to attenuate the burden on diminishing groundwater aquifers. However, comprehensive microbial characterizations of these water sources are generally lacking, especially with regard to variations through time and across multiple water types. To address this knowledge gap we used shotgun metagenomic sequencing to characterize the taxonomic and functional variations of microbial communities within two agricultural ponds, two freshwater creeks, two brackish rivers, and three water reclamation facilities located in the Mid-Atlantic, United States. Water samples (n=24) were collected from all sites between October and November 2016, and filtered onto 0.2 μm membrane filters. Filters were then subjected to total DNA extraction and shotgun sequencing on the Illumina HiSeq platform. From these data, we found that *Betaproteobacteria* dominated the majority of freshwater sites, while *Alphaproteobacteria* were abundant at

times in the brackish waters. One of these brackish sites was also host to a greater abundance of the bacterial genera *Gimesia* and *Microcystis*. Furthermore, predicted microbial features (e.g. antibiotic resistance genes (ARGs) and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) arrays) varied based on specific site and sampling date. ARGs were found across samples, with the diversity and abundance highest in those from a reclamation facility and a wastewater-impacted freshwater creek. Additionally, we identified over 600 CRISPR arrays, containing roughly 2,600 unique spacers, suggestive of a diverse and often site-specific phage community. Overall, these results provide a better understanding of the complex microbial community in untreated surface and reclaimed waters, while highlighting possible environmental and human health impacts associated with their use in agriculture.

2.2 Introduction

Steady declines in groundwater supplies, coupled with the estimation that by 2050 water withdrawals for irrigation will increase by roughly 10%, has strengthened the demand for alternative sources of water for agricultural applications [160]. The use of reclaimed (advanced treated municipal wastewater) and untreated surface waters (e.g. brackish rivers, freshwater creeks, and ponds) may provide effective solutions to reduce pressures on groundwater sources; however, the microbial communities of these water sources are typically poorly characterized and few inter-water microbial comparisons have been made.

Currently, reclaimed water is used as an alternative source of non-potable water, especially in arid and semi-arid regions around the world [5]. However, because it is the end-result of wastewater treatment, concerns remain about the levels of harmful microbiological constituents that may persist in the water. Previous studies have identified that, while wastewater treatment does reduce bacterial diversity, potential human pathogens (*Legionella* spp., *Mycobacterium* spp., and *Streptococcus* spp.) may still be present and, in some cases, selectively enriched by the disinfection process [112, 161]. Additionally, antibiotics introduced through municipal influent and agricultural runoff can persist throughout the treatment process, resulting in high selection pressure for antibiotic-resistant bacteria [162–164]. However, reclaimed water characteristics are likely dependent on the quality of the influent and the treatment practices employed by the wastewater treatment facility [109, 165].

While reclaimed water is commonly considered the standard for water reuse, untreated surface water sources may also represent practical alternatives for water management. For instance, ponds are common features across the United States, with an estimated abundance between 2.6 and 9 million [22]. They can occur naturally (e.g. floodplains, isolated depressions), but are often human-constructed for a variety of utilitarian and aesthetic purposes [22]. In agricultural settings, ponds may be constructed as a means of capture and storage of freshwater for localized irrigation. While ponds are not as widely studied as large lakes and marine systems, their unique topography may influence their microbial community composition. For instance, ponds, and other surface water sites, are subjected to environmental and anthropogenic factors that can largely be avoided when using groundwater (e.g.

animal fecal contamination and agricultural/urban runoff) [166]. Ponds also tend to have a higher terrestrial-aquatic interchange compared to larger bodies of water (e.g. lakes) that may drive the abundance of terrestrial microorganisms [157].

In contrast, lotic ecosystems (e.g. rivers, creeks) are marked by a natural flow, usually toward another body of water such as an ocean or lake. Therefore, in addition to traditional environmental factors, lotic systems may be impacted by connected waterways and upstream discharge facilities. For instance, discharge of wastewater effluent into urban and suburban rivers was reported to increase downstream organic and inorganic nutrient content and decrease overall bacterial diversity in sediment [148]. Alterations in bacterial community structure have also been associated with catchment area (e.g. agricultural, forested, urban), likely due to variations in nutrient concentration stemming from runoff [150, 167].

Despite the potential importance of these water sources to regional biodiversity and water management, few studies have sought to characterize not only the taxonomic and functional profile of the microbiota, but also their complex genomic features, such as CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) arrays. Because the CRISPR-Cas system functions by integrating pieces of foreign DNA (i.e. plasmid, phage) into recognizable arrays, predicting CRISPR can provide valuable information on phage infection history, as well as provide a potential means to subtype pathogens [168]. Previous studies exploring CRISPR-Cas from human body sites [169], dairy operations [170], and extreme environments (e.g. microbial mats [171], hot springs [172], Antarctic snow [173]) have given us profound insights into the infection history and defense strategies of their microbial communities. This is of particular importance when assessing the quality of irrigation waters, because, while they are not direct human pathogens, phage are responsible for shaping the diversity and genetic architecture of their host(s) and are often left unexplored in microbial studies [174].

In this present study, we used high-throughput shotgun metagenomic sequencing to: 1) characterize and compare bacterial community composition; 2) predict and compare functional, pathogenic, and antibiotic resistance genes using the Gene Ontology (GO) and Comprehensive Antibiotic Resistance Database (CARD) framework; and 3) identify CRISPR arrays in various untreated surface and reclaimed water sites sampled over the course of five weeks from October to November 2016 in the Mid Atlantic, United States.

2.3 Materials and Methods

2.3.1 Study sites

Water samples (n=24) were collected from nine sites of four different water body types within the Mid-Atlantic, United States: two tidal brackish rivers (MA04, MA08); two freshwater ponds (MA10, MA11); two non-tidal freshwater creeks (MA03, MA07); and three water reclamation facilitates (MA01, MA02, MA06) (Table 2.1).

2.3.2 Sample collection

Sites were sampled on the following dates: 10/10/16, 10/24/16, and 11/14/16, with the exception of the reclamation facilities (MA01, MA02, MA06) where samples were only collected 10/24/16 and 11/14/16. At each site, 1 L of water was collected. For the surface sites (e.g. creek, pond, river), sterile polypropylene sampling containers (Thermo Fisher Scientific, MA, USA) were submerged 15-30 cm below the surface using a long-range grabbing tool. For the reclamation facilities, 1 L of water was collected from a spigot, irrigation line or storage lagoon, depending on the facility and sampling feasibility on that date. Samples were transported on ice to the laboratory and stored at 4 °C. In addition, a ProDSS digital sampling system (YSI, Yellow Springs, OH, United States) was used to measure, in triplicate, the water temperature (°C) and pH. Ambient temperature was also collected for the time and date of sampling via the Nation Weather Services historical data archive.

2.3.3 Sample processing

To remove the cellular fraction, each water sample was vacuum filtered through a 0.2 μm membrane filter (Pall Corporation, MI, USA). Microbial DNA was then extracted from the filters using an enzymatic and mechanical lysis procedure currently used in our lab to extract DNA from various environmental biomes [158,159]. Briefly, the filters were added to lysing matrix tubes along with a cocktail of PBS buffer, lysozyme, lysostaphin, and mutanolysin. After incubating, samples were subjected to a second lysing cocktail (Proteinase K and SDS) followed by another incubation and mechanical lysis via bead beating. The resulting DNA was purified with the QIAmp DNA mini kit (Qiagen, CA, USA) and assessed with the NanoDrop 2000 Spectrophotometer.

2.3.4 Shotgun sequencing

For each sample, DNA was used in a tagmentation reaction, followed by 12 cycles of PCR amplification using Nextera i7 and i5 index primers per the modified Nextera XT protocol. The final libraries were then quantitated by Quant-iT hs-DNA kit. The libraries were pooled based on their concentrations as determined by Quantstudio 5 and loaded onto an Agilent High Sensitivity D1000 ScreenTape System. The samples were run across 8 lanes of an Illumina Hiseq X10 flow (Illumina, San Diego, CA, United States) cell targeting 100 bp paired end reads per sample.

2.3.5 Metagenomic assembly

The resulting paired-end reads were quality trimmed using Trimmomatic ver. 0.36 (sliding window:4:30 min len:60) [175], merged with FLASh ver. 1.2.11 [176], and assembled *de novo* with metaSPAdes ver. 3.10.1 (without read error correction) [177]. Open reading frames (ORFs) were predicted from the assembled contigs using MetaGene [104].

2.3.6 Taxonomic and functional classification

Predicted peptide ORFs were searched against UniRef 100 (retrieved May 2018) using protein-protein BLAST (BLASTp ver. 2.6.0+) (E value < $1e^{-3}$) [105, 178]. Taxonomic classifications were then made to contigs by max cumulative bit score. This was calculated by summing the bit scores of all taxa with a hit to peptide ORFs encoded by the contig. Functional assignments were made by assigning Gene Ontology (GO) terms to peptide ORFs. UniProt sequences are continually assigned GO terms by the Gene Ontology Annotation (GOA) program [179]. Peptide ORFs were assigned all GO terms that were linked to UniRef 100 peptides within 3% of the top hit's bit score.

Coverage was calculated for each contig by recruiting quality-controlled reads to assembled contigs using Bowtie2 ver. 2.3.3 (very sensitive local mode) and then using the "depth" function of Samtools ver. 1.4.1 to compute the per-contig coverage [180] . To normalize abundances across libraries, contig and ORF coverages were divided by the sum of coverage per million, similar to the transcripts per million (TPM) metric used in RNA-Seq [181]. Scripts performing these assignments and normalization are available at https://github.com/dnasko/baby_virome. Taxonomic and functional data were visualized using the R packages ggplot2 ver. 3.1.0 and pheatmap ver 1.0.10 [182, 183]. Significance tests were conducted using a Tukey's HSD Test.

2.3.7 Peptide ORF clustering

To assess the shared and unique functional profiles among the sampling sites, all complete peptide ORFs were clustered at 60% with CD-HIT [184]. Cluster files were then parsed using the clstr2txt.pl script.

2.3.8 Identification of antibiotic resistance genes

Predicted peptide ORFs were searched against the "Comprehensive Antibiotic Resistance Database" (CARD; retrieved July 2018) using protein-protein BLAST (BLASTp ver. 2.6.0+) (E value < $1e^{-3}$) [105, 107]. A queried peptide ORF was regarded as ARG-like if >40% coverage and >80% amino-acid identity to a protein in the CARD database [83, 185]. Using the CARD database as a reference, the putative ARG was assigned a gene name and a drug class. For the genes conferring resistance through mutations (i.e. *KasA*, *gyrA*, *gyrB*, *murA*, *ndh*, *thyA*, *rpsL*, *rpsJ*), a post-processing step (MAFFT alignment with reference sequences available at CARD) was taken to confirm the presence of resistance-conferring mutations [186].

2.3.9 Prediction and analysis of CRISPRs

CRISPR arrays were predicted from assembled contigs using the CRISPR detection and validation tool, CASC available at https://github.com/dnasko/CASC. CASC utilizes a modified version of the CRISPR Recognition Tool (CRT) to call putative CRISPR spacers [187]. CASC then validates these spacers by searching against a database of Cas proteins and CRISPR repeats to remove false positives and outputs FASTA files containing: (1) valid CRISPR spacers; (2) false-positive CRISPR spacers; (3) valid CRISPR repeats; and (4) false-positive CRISPR repeats. Valid CRISPR spacers were clustered with CD-HIT at 97% nucleotide similarity to determine the number of unique and shared spacers within and among the sites [184]

2.3.10 Data availability

Metagenomic reads were submitted to NCBI's Sequence Read Archive under the BioProject accession number PRJNA473136 (SRX498566- SRX4985689).

2.4 Results

2.4.1 Sampling site characteristics

Sampling site characteristics (pH, water temperature, ambient temperature, precipitation) are described in Table 2.2.

2.4.2 Sequencing effort and assembly

All processed samples (n=24) were sequenced on the Illumina HiSeq for a total of 803,403,499 read pairs (Table 2.3), with an average of 33,475,146 per metagenome (± SD 9,122,444). After metagenomic assembly, there were a total of 11,383,447 contigs, with an average of 474,310 contigs (± 150,989 SD) per sample.

2.4.3 Taxonomic composition

On average, 65% of contigs could be confidently assigned a taxonomic representative. Of these, between 78 and 98% (mean: 91%) were assigned as Bacteria, followed by Eukaryota (min: 0.6%, max: 17%, mean: 4%), and Viruses (min: 0.6%, max: 10%, mean: 3%). For the contigs that could be identified, a normalized abundance was calculated to account for sequencing effort and assembly/recruitment proficiency. For those assigned as Eukaryota, the majority of the abundance was classified as *Streptophyta* (min: 6%, max: 51%, mean: 21%), *Arthropoda* (min: 6%, max: 43%, mean: 15%), and *Chordata* (min: 8%, max: 30%, mean: 15%).

For those assigned as Bacteria, the most frequently observed bacterial phyla relative to each sample was *Proteobacteria* (min: 35%, max: 83%, mean: 55%) (Figure 2.1). This was followed by *Actinobacteria* (min: 2%, max: 25%, mean: 13%), *Bacteriodetes* (min: 6%, max: 21%, mean: 13%) or *Firmicutes* (min: 4%, max: 12%, mean: 7%). Within the *Proteobacteria* phyla, the class *Betaproteobacteria* was the most abundant at each of the sites, with the exception of sampling dates 10/24/16 and 11/14/16 within the tidal brackish river, MA04, and 11/14/16 within MA08, in which *Alphaproteobacteria* was the most abundant *Proteobacterial* class (Figure 2.1).

To further classify the taxonomic composition we considered the bacterial assignments at the genus level and compared them within and among sites (Figure 2.2). The 74-90% of contigs that could be assigned at this level were distributed among 2,207 different genera, with 789 genera identified at some abundance in all of the samples. Of these, 44 occurred at a relative abundance $\geq$1% in at least one sample. Within the majority of sites, *Variovorax* had the greatest relative abundance (min: 2%, max: 32%, mean: 11%). However, this was not the case for the freshwater pond, MA10, at all sampling dates and MA01, a reclamation facility, at its first sampling data (10/24/16), in which *Streptomyces* was the most abundant. Similarly, *Pusillimonas* was the most abundant in MA03, a freshwater creek (11/14/16), and MA04, a brackish river (10/10/16), while *Nostoc* was the most abundant in MA04 (10/24/16). Furthermore, when we compared the normalized abundance of the dominant genera among the different sites we found that MA04, a brackish river, had a significantly higher abundance of *Gimesia* and *Microcystis* than all of the other sites except MA01, a reclamation facility. Alternatively, MA01 had a significantly higher abundance of *Prochlorococcus* than all of the other sites except MA04.

2.4.4 Peptide ORF clustering

To assess the shared and unique functional profiles among the sampling sites, all complete peptide ORFs were clustered at 60% (Figure 2.3) [184]. The majority of these peptide ORFs clustered within site, with the two tidal brackish rivers, MA04 and MA08, having the greatest fraction of unique peptide ORFs, 63% and 51%, respectively. This was followed by MA07 (48%), MA11 (41%), MA03 (36%), MA02 (36%), MA01 (34%), MA10 (33%), and MA06 (31%). The remaining fraction of peptide ORFs from each site clustered with one or more different sites and, in some cases, showed a trend. Peptide ORFs from reclaimed sites (MA01, MA02) clustered highly with peptide ORFs from pond sites (MA11, MA10): 51% of MA02's peptide ORFs clustered with peptide ORFs from MA11, 41% of the peptide ORFs from MA06 clustered with peptide ORFs from MA10, and 38% of peptide ORFs from MA01 clustered with peptide ORFs from MA10. In addition to MA10, a large percentage of peptide ORFs (47%) from MA06 clustered with peptide ORFs from the non-tidal freshwater creek, MA03. Additionally, within between 4 and 9% of peptide ORFs clustered with all the other sites, representing a "functional core" among all the sampled irrigation water sites.

2.4.5 Functional analysis of bacterial-assigned ORFs

To characterize the functional profiles of the sampled sites, Gene Ontology (GO) annotations were assigned to peptide ORFs based on BLASTp matches to UniRef100 proteins. On average, 56% (min: 31%, max: 75%) of peptide ORFs were assigned at least one GO-term, with the majority (min: 81%, max: 98%, mean: 93%) coming from contigs assigned as bacteria. Within this fraction of bacterial peptide ORFs the GO-terms assigned at the greatest frequency were those related to the following: transferase activity (GO:0016740) (min: 25%, max: 29%, mean: 27%), hydrolase activity (GO:0016787) (min: 25%, max: 27%, mean: 26%), ATP binding (GO:0005524) (min: 18%, max: 25%, mean: 20%), oxidation-reduction process (GO:0055114) (min: 17%, max: 20%, mean: 18%), and catalytic activity (GO:0003824) (min: 15%, max: 18%, mean: 17%) (Figure 2.4).

Furthermore, we explored the GO-term for pathogenesis (GO:0009405), as well as its child term, toxin activity (GO:0090729), and terms associated with antibiotic GO:0046677) and drug resistance (GO:0042493) [188]. Again, the normalized abundance was calculated and totaled for bacterial contigs containing ORF(s) assigned to one or more of these GO-terms (Figure 2.4). Between 0.7 and 4% (mean: 2%) of the total bacterial contig abundance was attributed to contigs containing peptide ORF(s) assigned to pathogenesis (GO:0009405), with the largest portion at MA03 (11/14/16). Within this fraction we were able to annotate between 80 and 92% of the associated abundance at the genera level and found the majority of pathogenesis containing contigs were assigned as *Pseudomonas* (min: 6%, max: 14%, mean: 9%) and *Pusillimonas* (min: 1%, max: 12%, mean: 5%) (Figure 2.5).

2.4.6 Antibiotic resistance

To further assess antibiotic resistance, we conducted a stringent BLAST analysis of peptide ORFs against CARD. Across all samples, 114 peptide ORFs were identified as 32 unique ARGs conferring resistance to over ten drug classes, including resistance mechanisms associated with target mutations and those associated with dedicated antibiotic resistance gene products [107]. For the former, proteins were confirmed to carry the following mutations: *kasA*, R121K [189]; *gyrA*, S95T [190]; *murA*, C117D [191], *rpsL*, K88R, K43R [192, 193], *rpsJ*, V57M [194], and *EF-Tu* Q124K [195]. Overall, the reclamation water sampled on 11/14/16 from MA06, a reclamation facility, contained the greatest diversity of ARGSs, with 22 unique

ARGSs (identified from 25 ARG-like peptide ORFs) (Figure 2.6). This was followed by the non-tidal freshwater creek, MA03, sampled on 10/24/16, which had 13 unique ARGs (identified from 14 ARG-like peptide ORFs). The other non-tidal freshwater creek, MA07, contained the lowest diversity of ARGs, with just one unique ARG identified throughout the entire sampling period.

We also identified the source genera and phyla of each ARG-like peptide ORF by parsing the contig taxa identified previously (Figure 2.7). All the ARG-like peptide ORFs originated from contigs assigned as Bacteria, with the majority coming from *Actinobacteria* (59 of 114; 52%). While the majority of these were *rpsL* genes, Actinobacterial genera were also associated with 12 other unique ARGs. Furthermore, 31 ARG-like peptide ORFs were classified as *Proteobacteria* (11 alpha, 6 beta, 14 gamma) and encompassed the majority of ARG diversity, with 26 unique ARGs.

2.4.7 CRISPR array abundance and taxonomy

In addition to identifying traditional genes of concern we also sought to determine the phage-host relationships within and among sites using CRISPR arrays. CRISPR arrays were predicted in every library for a total of 612 arrays on 604 contigs. For the contigs that had a predicted CRISPR array, we calculated their normalized abundance (Figure 2.8). Overall, CRISPR containing contigs accounted for between 0.003 and 0.04% of the total contig abundance among all samples. On average, the tidal brackish water site, MA04, had both the greatest number of detected CRISPR arrays (238 across 234 contigs) and the highest normalized abundance of the CRISPR-containing contigs.

To identify the taxonomy of the contigs containing putative CRISPR arrays, we parsed the BLASTp assigned taxa. However, similar to previous studies [196] making taxonomic predictions for CRISPR-containing contigs was difficult. Only 22% (130/604) of CRISPR containing contigs could be assigned a taxa (Figure 2.9), the majority of which were of the phyla *Proteobacteria* (45%) made up of 38% *Gammaproteobacteria*, 29% *Alphaproteobacteria* and 17% *Betaproteobacteria*. This was followed by *Firmicutes* (21%) and *Cyanobacteria* (11%).

2.4.8 CRISPR spacers within and among sites

To determine the number of unique spacers, we clustered all the sampling dates within each site at 97% nucleotide similarity. Overall, MA04, a brackish river site, had the greatest number of unique spacers (1173 spacers) followed by MA10 (398 spacers), MA11 (321 spacers), MA01 (293 spacers), MA06 (269 spacers), MA03 (124 spacers), MA08 (161 spacers), MA07 (25 spacers), and MA02 (21 spacers) (Figure 2.8). These unique spacers were then clustered together to produce the number of shared spacers among the different sites. Overall, the reclamation site, MA02, (8 spacers) and the freshwater pond, MA10, (120 spacers) shared the greatest portion of their spacers with other sites, at 38% and 30%, respectively. This was followed by MA03 at 29% (36 spacers), MA01 (75 spacers) at 26%, MA06 at 18% (49 spacers), MA11 at 3% (nine spacers), and MA04 at 0.3% (four spacers). Both MA07 (a freshwater creek) and MA08 (a brackish river) shared no spacers with other sites.

Additionally, similar to the results from the peptide ORF clustering, we observed that reclaimed sites shared spacers with pond sites. Specifically, MA01 and MA10 shared the greatest number of spacers (69).

2.5 Discussion

Water reuse is an important practice to mitigate our dependence on dwindling groundwater supplies. However, across the farm-to-fork continuum, irrigation water is a known source of microbial contamination of fresh produce and, therefore, must be subjected to scrutiny. In the present study, we utilized metagenomics to assess multiple facets of the microbial community present in a variety of irrigation water sites. Overall, we found that *Proteobacteria*, especially *Betaproteobacteria*, dominated the bacterial community at the phylum level (Figure 2.1). This agrees with previous research of aquatic environments, including fresh surface [119] and reclaimed waters [149]. However, on several sampling dates for the two tidal brackish rivers, MA04 and MA08, there was a greater proportion of *Alphaproteobacteria*, a phylum traditionally abundant in marine systems [197]. This is not surprising as brackish waters have previously shown to house a co-occurrence of bacteria typically associated with freshwater and marine systems [198, 199].

At a lower taxonomic level, genera commonly associated with human disease (e.g. *Streptococcus* and *Enterococcus*) were identified in all samples. However, typical environmental genera such as *Variovorax* were the most abundant (Figure 2.2). *Variovorax* falls within the family *Comamonadaceae* and is phylogenetically closely related to *Acidovorax*, another widespread environmental genera [200, 201]. *Variovorax* has been found throughout a variety of environments, such as: drinking water [201, 202], freshwater riverine water [203], groundwater [204, 205], and soil [206]. As a result, it has been suggested that *Variovorax* can adapt to different environmental constraints, likely due to its ability to degrade a variety of organic compounds, including pollutants [207]. This heterogeneity in metabolic potential may explain its dominance among the variety of sites sampled here and was also reflected in the bacterial functional profile, in which a high abundance of peptide ORFs were annotated with functions associated with metabolism (e.g. hydrolase and oxidoreductase activity) (Figure 2.4).

While the dominating bacteria phyla were similar across sites there were some trends in specific genera that differentiated sites, especially MA04. MA04 is one of the brackish water sites and the largest lotic body studied in this analysis with a width of roughly 78 meters. Here, we found a significantly higher abundance of the dominant genera *Gimesia* and *Microcystis* compared to the majority of other sites. Species of these genera can tolerate high salinity and, therefore, are likely able to persist in the brackish environment of MA04 [208, 209]. This may be of concern considering that species of *Microcystis* are capable of producing hepatotoxins that have demonstrated some ability to bioaccumulate in produce [210] - a potential source of toxin exposure not often considered when assessing irrigation water quality.

To further investigate the potential public health impacts of these water sources we also investigated the presence of ARGs. Previous studies have identified a diverse range of ARGs in reclaimed waters [211], natural surface water sites [212], and even pristine environments (e.g. ancient permafrost [213, 214]. These studies reflect both the natural production of antibiotics by bacteria and those introduced to the environment by way of human pollution. In this study, we identified ARGs in metagenomes from the majority of sites. One of the genes that was identified in multiple peptide ORFs from a wide range of samples was the antibiotic resistance mutant of *rpsL* (Figure 2.6). Mutants of *rpsL* are resistant to aminoglycosides, including streptomycin, an antibiotic produced by the soil *Actinobacteria Streptomyces griseus* and used widely in clinical and agricultural applications [192, 193]. Mutations in *rpsL* confer resistance by disrupting interactions between the ribosomal protein and the antibiotic. These types of target mutations are frequently found in environmental bacteria and are thought to be the result of spontaneous pleotropic mutations [215, 216]. This is consistent with the potential taxonomic origin of the majority of *rpsL* mutants, *Ferrimicrobium*, an iron-oxidizing *Actinobacteria* found in mine waters [217], geothermal soils [218], and a waterlogged bog [219].

While pathogen associated bacterial genera (e.g. *Listeria*, *Vibrio*, *Escherichia*) were identified to be potential hosts of ARGs, we also saw a diversity of other hosts, including traditional environmental genera (Figure 2.7). This is in agreement with a previous study that determined the taxonomic origin of ARGs in activated and anaerobically digested sludge [220]. The authors identified ARGs from microbiota belonging to indigenous environmental bacteria and attributed this to horizontal gene transfer between pathogen and environmental bacteria within the wastewater treatment plant.

We also identified sites of high ARG diversity, including the reclamation facility, MA06, and the non-tidal freshwater creek, MA03 (Figure 2.6). In MA06 most of the ARGs confer resistance to antibiotics commonly used in clinical and agricultural applications including aminoglycosides, sulfonamides, rifamycins, macrolides, cephalosporins, fluoroquinolones, and tetracyclines that function through a suite of resistance mechanisms, including antibiotic inactivation and efflux, as well as antibiotic target alteration, protection, and replacement [107, 221, 222]. This broad range of resistance is not surprising as wastewater treatment plants are considered "hotspots" of antimicrobial resistance due to traces of antibiotics in the wastewater driving selection pressure [220, 223]. One of the most abundant ARGs detected in the reclaimed water was *ErmF*, which confers resistance to macrolide-lincosamide-streptogramin (MLS) antibiotics. These antibiotics are used frequently to treat Gram-positive infections and have been found to withstand wastewater treatment, persisting at high levels in the effluent [224, 225].

For the freshwater creek sample MA03 we also saw a high ARG diversity. This is likely due to discharge from a wastewater treatment facility located upstream of the sampling location. Previous studies have identified that wastewater effluent discharge into natural lotic systems may increase the presence of antibiotics, antibiotic-resistant bacteria, and ARGs downstream [226,227]. A study of the Grote Beerze River in the Netherlands found increased amounts of antibiotics and ARGs up to 20 km downstream of an effluent discharge point compared to upstream samples [227]. However, it is important to note in our study that ARG diversity was not consistently high throughout the entire sampling period. Due to the limited number of replicates, it is difficult to determine whether this was the result of changing abiotic and biotic factors throughout time or variability in sampling. Nevertheless, the presence of these genes in irrigation water sources, at any point, raises potential concerns for their use in agricultural applications and should be investigated further to determine their dissemination and persistence on food crops.

In addition to pathogenic genes of concern, we leveraged the CRISPR-Cas system to facilitate an investigation into the bacteria-phage infection history at each of the sites. Overall, we identified CRISPR arrays at all time points and sampling locations, suggesting a widespread use of this defense system in bacteria from surface and reclaimed waters. The tidal brackish river site, MA04, had the greatest abundance of CRISPR arrays (Figure 2.8). This may reflect the unique bacterial composition of MA04, which also had the greatest proportion of unique peptide ORFs (Figure 2.3) in addition to an increased abundance of the dominant genus *Microcystis* (Figure 2.2). Previously, CRISPR arrays have been identified in strains of the cyanobacterium *Microcystis aeruginosa* isolated from a shallow eutrophic reservoir [228]. In addition the authors found that the *Microcystis* spacers were rarely shared among the strains, which is similar to the results observed in this study.

Within the CRISPR arrays the majority of spacers were unique to each site, suggesting specific interactions between phage and hosts. This is likely due to the native bacteria having adapted to the local phage populations at each site [229]. As a result, it has been hypothesized that spacers can be used as a molecular fingerprint to subtype bacteria and potentially track pathogen outbreaks [168]. However, there were samples across sites that demonstrated shared spacers, with the most apparent between the reclaimed water site, MA01, and the freshwater pond site, MA10. This suggests a similarity in community composition that may be indicative of similar bacterial lineages and/or environmental exposures. For instance, the reclaimed water sites were stored at least temporarily in irrigation ponds/lagoons, which have similar topographical features to ponds. This may also explain the high percentage of peptide ORF clustering observed between the two sites (Figure 2.3). However, this connection needs to be explored in greater detail.

2.5.1 Conclusions

Crucial for the use of surface and reclaimed water sources is knowledge of their risks to environmental health and food safety. While limited in the number of biological replicates and time points, this study provides valuable data on bacterial community and functional composition, phage infection history, and the presence of pathogenic genes and ARGs in untreated surface and reclaimed waters used in irrigation activities. These data can be used to inform future studies and support the implementation of adaptive on-farm technologies (e.g. drip irrigation) that can reduce the spread of pathogenic microbial components.

2.6 Figures

Figure 2.1: Normalized relative abundance of the bacterial taxa present in reclaimed and untreated surface water sites at each sampling date. (A) Stacked bar chart depicting the community structure at the phylum level. (B) Stacked bar chart of the *Proteobacteria* phyla split into classes. Sites are grouped by water type and arranged in the order of sampling date. Normalized abundance measured as contig coverage divided by the sum contig coverage per million and presented as relative.

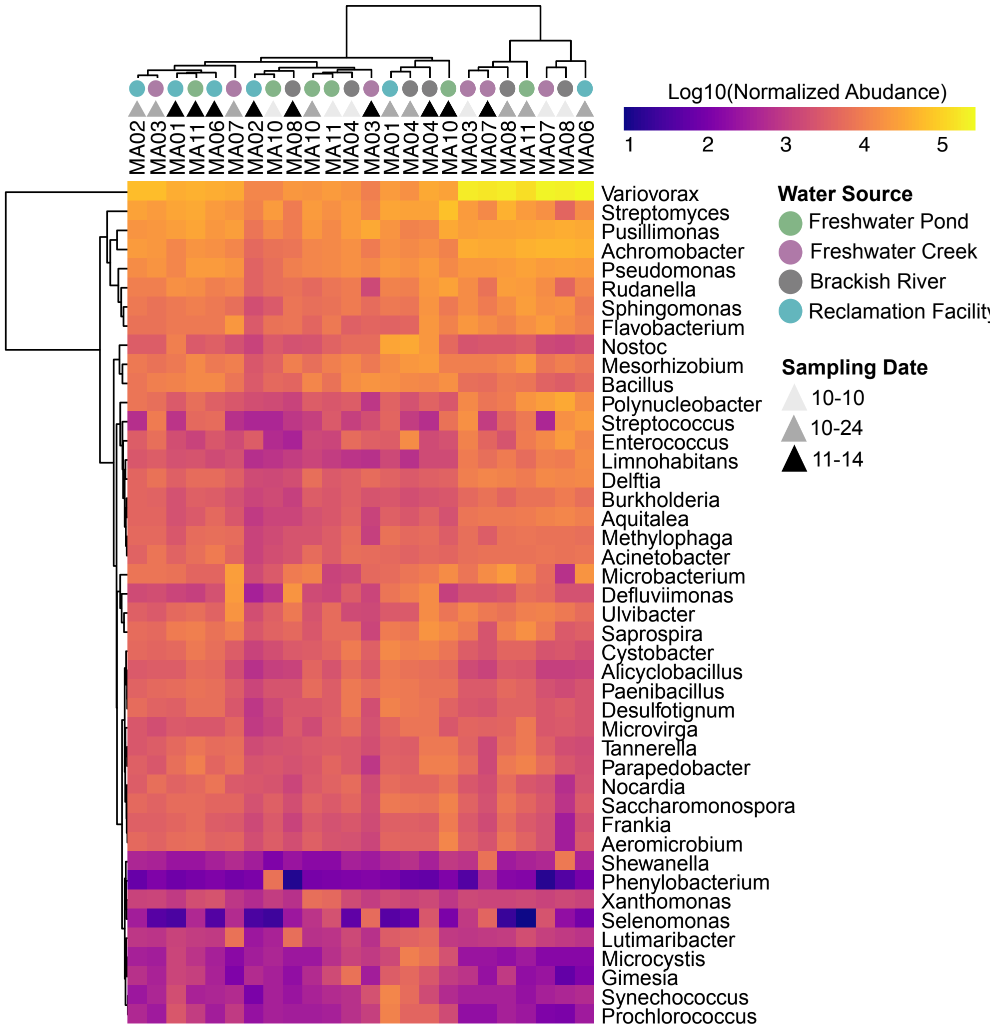

Figure 2.2: Taxonomic heatmap of the bacterial communities present in reclaimed and untreated surface water sites at each sampling date. Heatmap based on the log-transformed normalized abundance of the most dominant genera (>1% in at least one sample). Normalized abundance measured as contig coverage divided by the sum contig coverage per million.

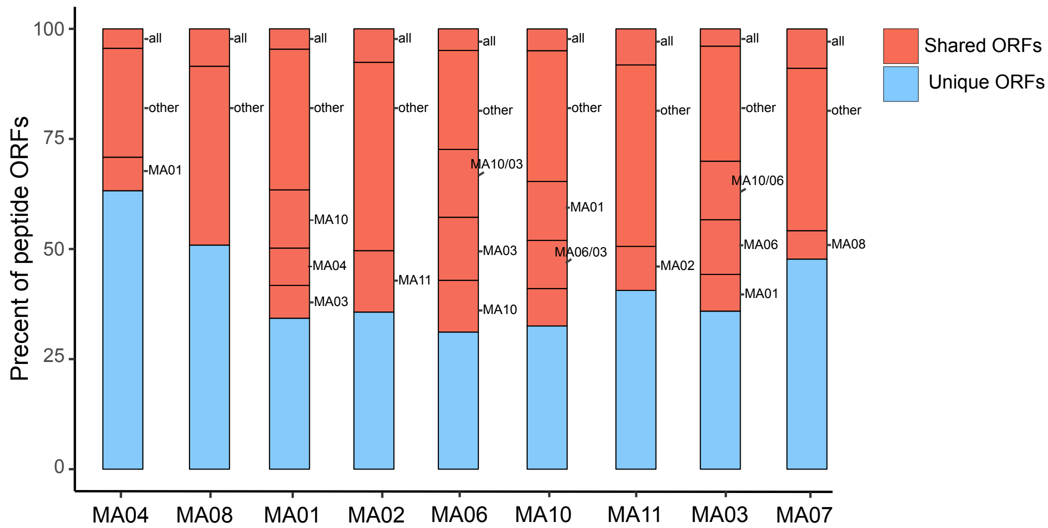

Figure 2.3: Shared and unique peptide ORFs in reclaimed and untreated surface water sites. Stacked bars depict the percentage of peptide ORFs from each site contained within 60% similarity peptide clusters either with themselves (unique; colored blue) or with other sites (shared; colored red). The fractions of shared peptide ORFs in each site are labeled with the site or combination of sites they clustered with at >5%. The "all" label represents, in each site, the fraction of shared peptide ORFs that clustered with at least one peptide ORF from all other sites.

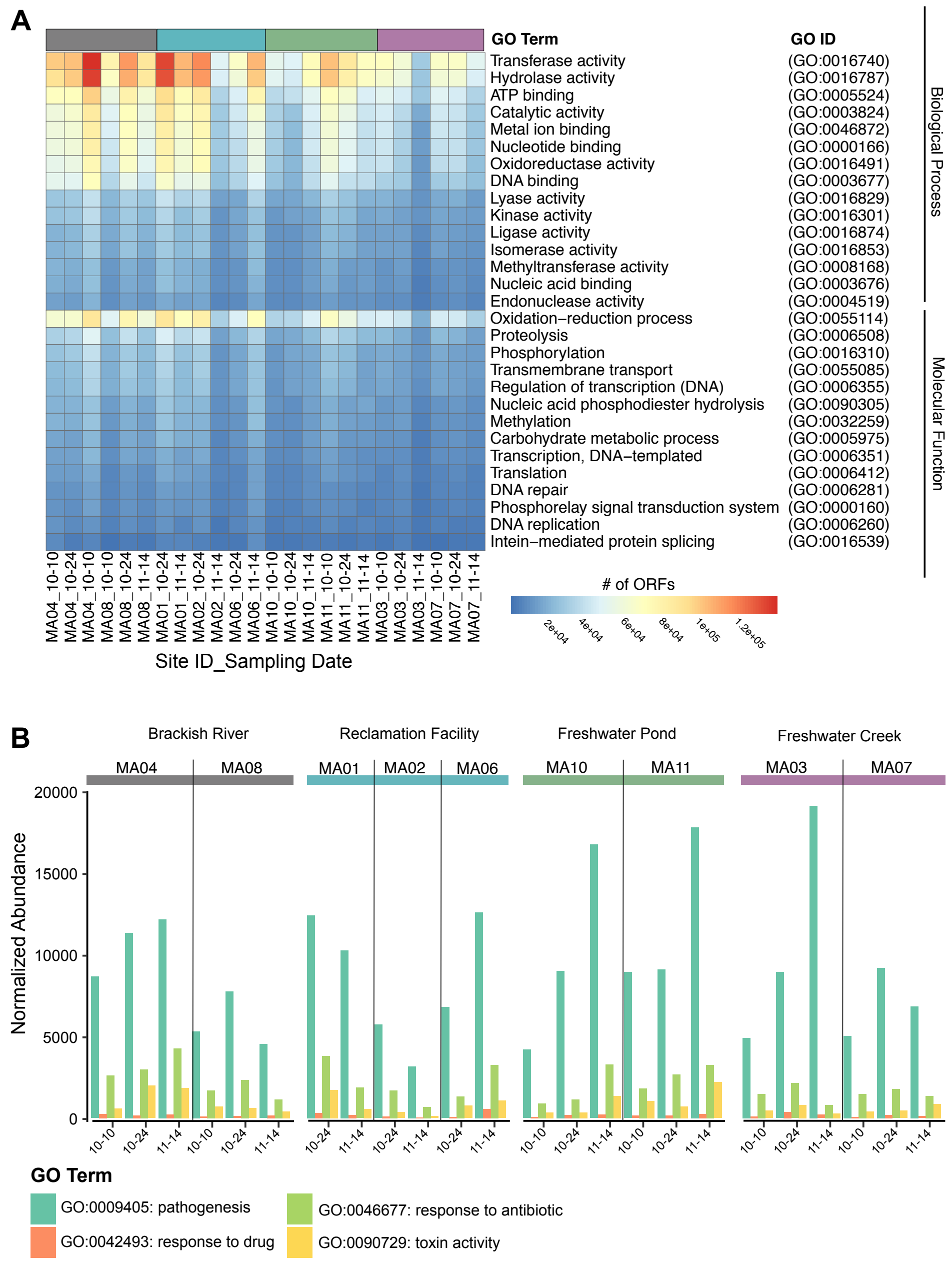

Figure 2.4: Functional profiles of bacterial communities present in reclaimed and untreated surface water sites at each sampling date. (A) Heatmap of the number of peptide ORFs assigned to the top ten GO terms for the biological and molecular categories in each sample. The corresponding GO IDs are presented in parentheses. One peptide ORF may be matched to multiple GO terms. (B) Normalized abundance of contigs containing peptide ORFs assigned at each site to the following GO terms: pathogenesis (GO:0009405), toxin activity (GO:0090729), response to antibiotic GO:0046677), response to drug (GO:0042493). Sites are grouped by water type and arranged in the order of sampling date.

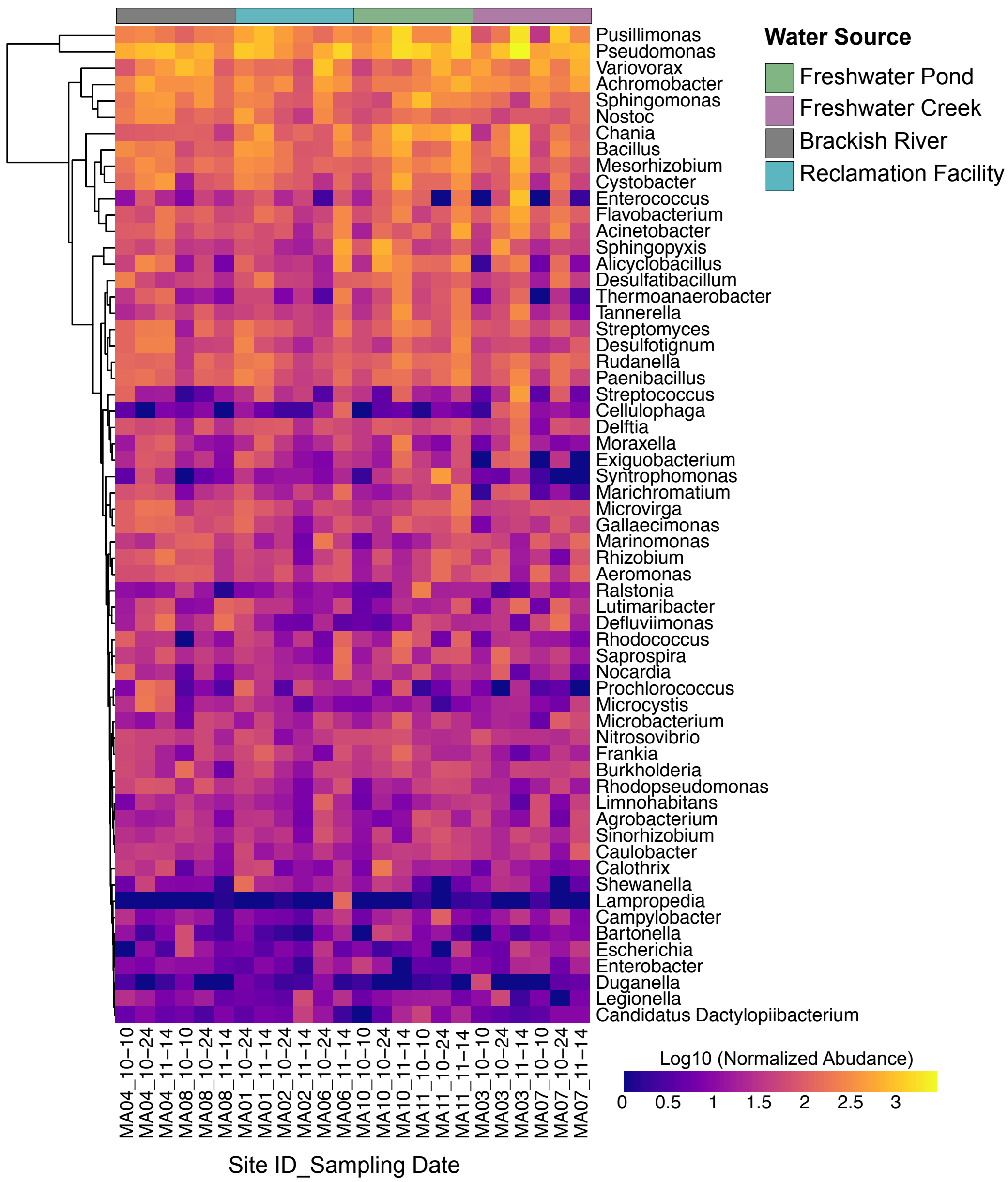

Figure 2.5: Taxonomic heatmap of the bacteria associated with pathogenesis GO-terms present in reclaimed and untreated surface water sites at each sampling date. Heatmap based on the log + 1 transformed normalized abundance of the most dominant contigs containing peptide ORF(s) assigned to pathogenesis (GO:0009405) (>1% in at least one sample). Sites are grouped by water type and arranged in the order of sampling date. Normalized abundance measured as contig coverage divided by the sum contig coverage per million.

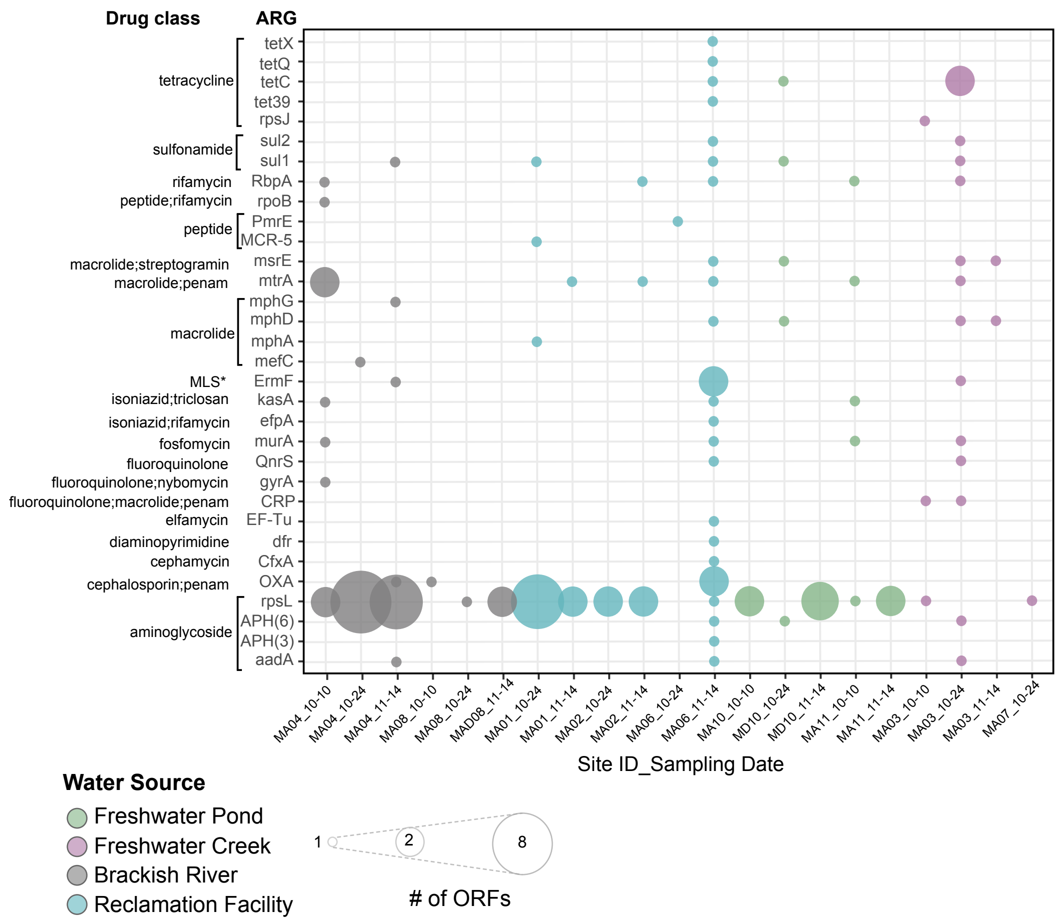

Figure 2.6: Antibiotic resistance genes (ARGs) predicted in reclaimed and untreated surface water sites at each sampling date. Dotplot showing the ARG-like peptide ORFs present at each water site, with the size of each dot equivalent to the number of peptide ORFs with homology to each ARG listed on the y-axis, and the color representative of the water type. ARG drug class designations consistent with the ontology from the Comprehensive Antibiotic Resistance Database. *MLS: macrolide, lincosamide, streptogramin antibiotic.

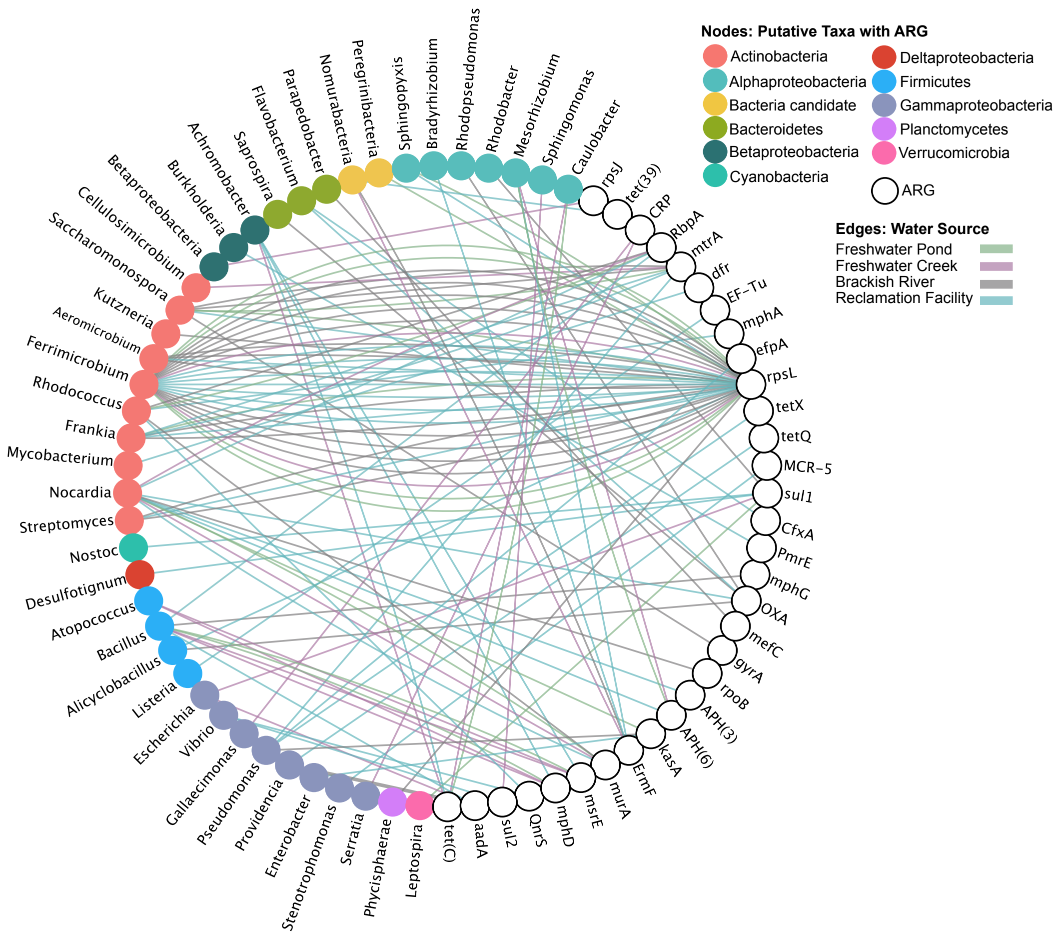

Figure 2.7: Bipartite network of the bacterial taxa with predicted antibiotic resistance genes (ARGs) from reclaimed and untreated surface water sites. Uncolored circles represent ARGs connected by an edge to its putative bacterial host, with each edge colored by the water type of the sample it was identified in. Bacterial host defined as the taxa assigned to the contig the ARG-like peptide ORF originated from and colored accordingly.

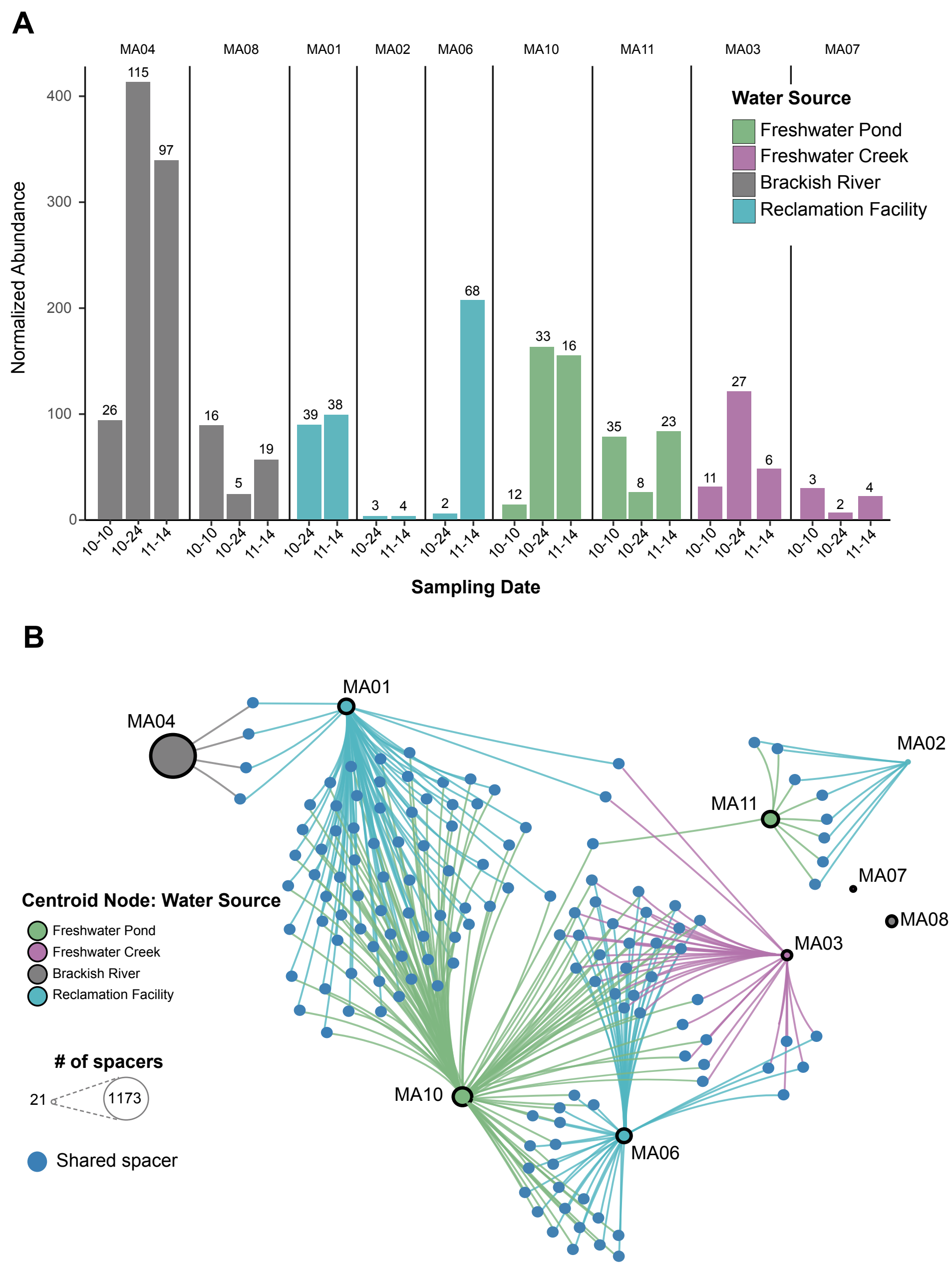

Figure 2.8: CRISPR array abundance and spacer overlap within and among reclaimed and untreated surface water sites. (A) Bar plot of the normalized abundance of contigs containing CRISPR arrays. Bars are colored by water type and labeled with the number of contigs containing CRISPR arrays. Normalized abundance measured as contig coverage divided by the sum contig coverage per million. (B) Network of shared spacers (97% identity) among nontraditional water sites. Donut-shaped centroid nodes represent each of the nine sampling sites, with the size equivalent to the number of spacers within that site, and the color representative of the water type. Nodes connecting the centroids represent shared spacers between and among sites.

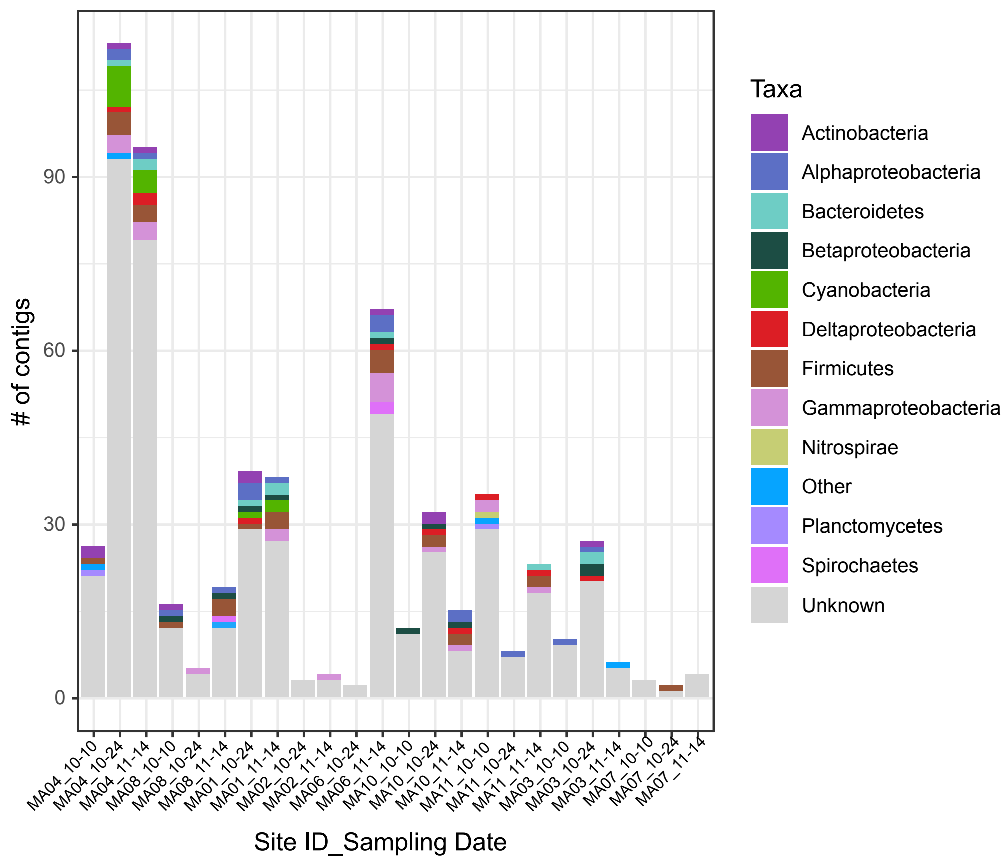

Figure 2.9: Taxonomic origin of contigs containing putative CRISPR arrays. Stacked bar chart depicting the community structure at the phylum level. Sites are grouped by water type and arranged in the order of sampling date.

2.7 Tables

Table 2.1: Descriptions of study sites

Water Type	Site ID	Catchment Area	Description
Pond	MA10	Agricultural	Freshwater pond with a maximum depth of ~3.35 m and surface area of ~0.26 ha
Pond	MA11	Agricultural	Freshwater pond with a maximum depth of ~3 m and surface area of ~0.40 ha
River	MA04	Marshland/ Forested	Tidal brackish river flowing into the Choptank River. width of ca. 76 m., depth of ~0.3-0.6 m.
River	MA08	Marsh grasses ~25-50m wide then pine woods	Tidal brackish river flowing into the Chesapeake Bay. width of ~15 m., depth of ~2-3 m. Within 1-1.5 miles downstream broiler concentrated animal feeding operations (CAFOs).
Creek	MA03	Wooded Agronomic cropland	Non-tidal freshwater creek tributary of the Nanticoke River. width of ~3 m., depth of ~1 m. Within 1 mile downstream wastewater treatment discharge facility.
Creek	MA07	Flood plain grasses and woodland (hardwood)	Non-tidal freshwater creek tributary of the Nanticoke River width of ~10 m., depth of ~1 m. Within 2.5 miles downstream from CAFO poultry houses.
Reclaimed	MA01	Wooded pines grass lanes	Influent is treated through activated sludge processing (Sequential Batch Reactor), filtration, UV light, chlorination, and then stored in an open-air lagoon before land application.
Reclaimed	MA02	Agronomic cropland (Corn and Soybeans)	Influent is treated through activated sludge processing (Sequential Batch Reactor), filtration, UV light, chlorination, and then stored in an open-air lagoon before land application.
Reclaimed	MA06	Native grass	Influent is treated through grinding, activated sludge processing secondary clarification, and then stored in an open-air lagoon. Chlorinated prior to land application.

Table 2.2: Sampling site characteristics by site and date.

Site ID	Sampling Date	Ambient Temp. (°C)	Precipitation (in.)[a]	Water Temp. (°C)	pH
MA01	10/24/16	17.8	0	16.3	8
	11/14/16	12.2	0.05	12.67	7.13
MA02	10/24/16	19	0	17.82	9
	11/14/16	11.1	0	15.98	7.3
MA06	10/24/16	13.9	0	18.97	7.14
	11/14/16	10	0	11.2	7.32
MA03	10/10/16	14.4	0	15.46	6.62
	10/24/16	20	0	14.66	8.04
	11/14/16	11.7	0	10.06	7.65
MA07	10/10/16	14.4	0	14.31	6.48
	10/24/16	21.1	0	15.17	8.05
	11/14/16	12.8	0	9.54	7.55
MA04	10/10/16	12.8	0.25	10.49	7.19
	10/24/16	17.8	0.01	16.48	7.26
	11/14/16	19.4	0	13.23	7.41
MA08	10/10/16	9.4	0	14.14	7.94
	10/24/16	16.1	0	14.49	7.3
	11/14/16	5.6	0.12	6.28	5.9
MA10	10/10/16	10	0	18	7.59
	10/24/16	17.2	0	19.83	7.7
	11/14/16	10.6	0	10.9	7.56
MA11	10/10/16	16	0	18.33	7.5
	10/24/16	15	0	18.4	7.71
	11/14/16	12	0	11.2	7.86

[a]Precipitation 24 hr prior to sampling

Table 2.3: Sequencing effort and assembly characteristics.

Site ID	Sampling Date	no. Read Pairs	no. Contigs	Mean Contig Size	Median Contig Size	Max Contig Size	% GC
MA01	10/24/16	31,266,431	672,177	528	333	96,619	60
	11/14/16	35,432,350	655,732	505	319	366,989	45
MA02	10/24/16	54,186,842	713,283	403	299	302,568	49
	11/14/16	39,765,597	521,318	350	261	346,809	44
MA06	10/24/16	28,245,595	306,996	407	300	169,877	52
	11/14/16	43,345,116	600,037	587	330	363,489	49
MA03	10/10/16	35,372,707	342,129	382	298	65,404	51
	10/24/16	30,597,692	378,757	527	321	272,084	49
	11/14/16	28,800,485	313,464	601	323	564,267	48
MA07	10/10/16	27,006,964	289,868	376	301	19,150	51
	10/24/16	24,222,993	404,914	515	329	177,909	47
	11/14/16	28,830,706	278,159	426	306	79,768	53
MA04	10/10/16	127,815,317	569,628	474	326	136,266	52
	10/24/16	30,063,212	494,474	492	328	710,391	57
	11/14/16	30,310,282	576,622	561	336	448,625	53
MA08	10/10/16	35,208,468	312,378	466	307	130,837	53
	10/24/16	30,241,203	475,763	490	326	109,586	51
	11/14/16	45,151,333	799,931	430	298	140,426	46
MA10	10/10/16	26,751,522	480,211	338	273	152,051	45
	10/24/16	28,003,231	425,677	546	322	496,485	49
	11/14/16	27,979,092	376,297	620	345	245,098	45
MA11	10/10/16	60,316,235	656,403	482	305	554,432	51
	10/24/16	23,901,357	363,433	482	330	230,703	51
	11/14/16	30,588,769	375,796	631	335	614,325	46

Chapter 3: Agricultural Freshwater Pond Supports Diverse and Dynamic Bacterial and Viral Populations

Jessica Chopyk, Sarah Allard, Daniel J Nasko, Anthony Bui, Emmanuel F Mongodin, and Amy Rebecca Sapkota. Agricultural freshwater pond supports diverse and dynamic bacterial and viral populations. *Frontiers in Microbiology*, 9:792, 2018.

3.1 Abstract

Agricultural ponds have a great potential as a means of capture and storage of water for irrigation. However, pond topography (small size, shallow depth) leaves them susceptible to environmental, agricultural, and anthropogenic exposures that may influence microbial dynamics. Therefore, the aim of this project was to characterize the bacterial and viral communities of pond water in the Mid-Atlantic United States with a focus on the late season (October to December), where decreasing temperature and nutrient levels can affect the composition of microbial communities. Ten liters of freshwater from an agricultural pond were sampled monthly, and filtered sequentially through 1 and 0.2 μm filter membranes. Total DNA was then extracted from each filter, and the bacterial communities were characterized using 16S rRNA gene sequencing. The remaining filtrate was chemically concentrated

for viruses, DNA-extracted, and shotgun sequenced. Bacterial community profiling showed significant fluctuations over the sampling period, corresponding to changes in the condition of the pond freshwater (e.g., pH, nutrient load). In addition, there were significant differences in the alpha-diversity and core bacterial operational taxonomic units (OTUs) between water fractions filtered through different pore sizes. The viral fraction was dominated by tailed bacteriophage of the order *Caudovirales*, largely those of the *Siphoviridae* family. Moreover, while present, genes involved in virulence/antimicrobial resistance were not enriched within the viral fraction during the study period. Instead, the viral functional profile was dominated by phage associated proteins, as well as those related to nucleotide production. Overall, these data suggest that agricultural pond water harbors a diverse core of bacterial and bacteriophage species whose abundance and composition are influenced by environmental variables characteristic of pond topography and the late season.

3.2 Introduction

Growing urbanization and climate variability have placed immense pressure on the finite supply of groundwater available for agricultural irrigation. As a result, the exploration of alternative irrigation water sources, including recycled water and pond water, has become a global priority [230–232]. While there is no universal standard, ponds are generally defined as small (~1 m^2 to 50,000 m^2), shallow, standing water bodies that can either permanently or temporarily collect freshwater [23–25]. These small water bodies are known to house a rich tapestry of aquatic plant and macroinvertebrate species, even greater than that of other larger water bodies (e.g., lakes and rivers) [23]. Moreover, the high terrestrial and aquatic interchange of ponds may enable both small free-living (>1 μm) and large/particle-associated bacteria to proliferate [233, 234]. Therefore, assessing the microbial diversity and interactions of these complex water bodies is a critical first step with regard to completing a comprehensive characterization of pond biodiversity, and evaluating the suitability of pond freshwater for agricultural applications, such as the irrigation of food crops. A growing body of literature has defined several bacterial phyla that are abundant in freshwater and markedly different from that of marine systems [118, 119]. Previous studies have reported a widespread distribution of *Actinobacteria* (lineage acI and acIV) among various freshwater sites, in some cases composing greater than 50% of the total bacterial community abundance [119, 126]. The *Proteobacteria* phylum in freshwater is also abundant, composed largely of *Betaproteobacteria*. This is in contrast to marine systems where *Alphaproteobacteria* typically dominates [119]. Furthermore, aquatic microbial communities are influenced by a variety of seasonal factors, such as pH, temperature, and water retention time [127, 128], as well as top-down regulation from predators (e.g., bacteriophage) [129, 130]. However, the majority of studies on the bacterial community composition of freshwater habitats come from large lakes and rivers, and very few have included an analysis of viral populations.

Bacteriophage, viruses that infect bacteria, are the most abundant biological entities in aquatic systems and play an important role in microbial community composition and ecology [235, 236]. For instance, phage lysis results in the release of the host's internal cellular contents (e.g., organic carbon, nitrogen), which then becomes a part of the pool of dissolved organic material (DOM). This phenomenon, known as the viral shunt, increases the level of available DOM for other microbes and is suggested to promote bacterial respiration and growth [51, 52]. Several studies have surveyed viral diversity through the use of widely, although not universally, distributed marker genes, such as *polA*. Family A DNA polymerase, *polA*, which encodes the Pol I protein, is the principal polymerase for phage genome replication, and is suggested to be predictive of viral lifestyle based off a single amino acid substitution [97, 237, 238]. A phenylalanine (wildtype) or tyrosine at amino acid position 762 (relative to *Escherichia coli*) is predictive of virulent phage (i.e., lytic replication), while a leucine substitution at this site seems to be predictive of a temperate lifestyle (i.e., lysogenic replication) [238]. Other studies have surveyed viral communities and diversity through shotgun sampling of genomic DNA from viral concentrates [97].

Similar to most bacterial analyses, the majority of studies of viral metagenomes (viromes) have been created from marine samples, which have provided astounding insights into how phage affect the ecology and biology of their hosts [51, 132]. In addition, there have been several viromes created from large freshwater lakes in the Artic [133], the Antarctic [95], France [134], North America [135–138], and Ireland [139]. While limited in their scope, these studies have provided some of the first data on freshwater phage ecology, demonstrating that, like their hosts, phage diversity is influenced by environmental factors. However, only a few studies have evaluated freshwater viromes from small lakes/ponds and fewer have looked at freshwater viromes in conjunction with a temporal analysis of fine scale host diversity [145, 239].

Therefore, we aimed to assess the bacterial and viral components of a temperate agricultural pond in the Mid-Atlantic, United States during the late growing season (October to December), a time when declining temperature and nutrient levels may impact the structure and function of the microbial assemblages. Specifically, we used 16S rRNA gene and shotgun metagenomic sequencing to: (i) survey the bacterial consortium utilizing different filter pore sizes (1 and 0.2 μm); (ii) characterize the diversity and abundance of the bacteriophage within the viral community; and (iii) compare the phylogeny of pond viromes across time using the phylogenetically relevant, and biologically meaningful, Pol I protein.

3.3 Materials and Methods

3.3.1 Study site and sample collection

Ten-liter water samples were collected in October 2016, November 2016, and December 2016 from a temperate freshwater agricultural pond in central Maryland, United States (maximum depth of ~3.35 meters and a surface area of ~0.26 ha). A Honda WX10TA (32 GPM) water pump was used to collect water 15 to 30 cm below the surface into a sterile polypropylene carboy. Samples were kept in the dark at 4 °C and processed within 24 h of collection. In addition, a ProDSS digital sampling system (YSI, Yellow Springs, OH, United States) was used to measure, in triplicate: the water temperature (°C), conductivity (SPC uS/cm), pH, dissolved oxygen (%), oxidation/reduction potential (mV), turbidity (FNU), nitrate (mg/L), and chloride (mg/L).

3.3.2 Sample preparation

Viral and microbial fractions were separated through peristaltic filtration followed by an iron-based flocculation and resuspension of viral particles. Two 142 mm polycarbonate in-line filter holders (Geotech, CO, United States), one equipped with a 142-mm diameter Whatman 1 μm polycarbonate filter (Sigma-Aldrich, MO, United States) and one with a 142-mm diameter 1 μm polycarbonate filter followed by the 0.2 μm membrane filter using a Watson Marlow 323 Series Peristaltic Pump (Watson-Marlow, Falmouth, Cornwall, United Kingdom). No prefiltration steps were conducted prior to the sample processing described above. After filtration, both filters (1 and 0.2 μm) were dissected into four equal quadrants and stored at $-80\,^{\circ}$C until DNA extraction. The iron chloride procedure [240] was then used on the resulting filtrate to concentrate viral particles. Briefly, 1 mL $FeCl_3$ solution (4.83 g $FeCl_3$ into 100 ml H_2O) was added to the filtered pond water and incubated in the dark for 1 h. Flocculated viral particles were then filtered onto 142-mm 1 μm polycarbonate filters (Sigma-Aldrich, MO, United States) and stored at $4\,^{\circ}$C in the dark until resuspension. For viral resuspension, filters were rocked overnight at $4\,^{\circ}$C in 10 mL of 0.1M EDTA-0.2M $MgCl_2$-0.2 M Ascorbate Buffer, described in detail elsewhere [240]. To ensure total removal of free DNA contamination, resuspended viral particles were subjected to a DNase I (Sigma-Aldrich, MO, United States) treatment for 2 h and again passed through a 33-mm diameter sterile syringe filter with a 0.2 μm pore size (Millipore Corporation, MA, United States).

3.3.3 Viral DNA extraction and shotgun sequencing

For the virome analysis, DNA was extracted from 500 μl of the treated viral concentrate using the AllPrep DNA/RNA Mini Kit (Qiagen, CA, United States) per the manufacturer's instructions and quantified with a HS DNA Qubit fluorescent concentration assay. For each sample, DNA was used in the tagmentation reaction, followed by 13 cycles of PCR amplification using Nextera i7 and i5 index primers and 2 μl Kapa master mix per the modified Nextera XT protocol. The final libraries were then quantitated by KAPA SYBR FAST qPCR kit and sequenced on the Illumina HiSeq 4000 (Illumina, San Diego, CA, United States).

3.3.4 Microbial DNA extraction, 16S rRNA gene PCR amplification, and sequencing

For the 16S rRNA gene analysis, DNA was extracted from each of the four filter quadrants from the 1 and 0.2 μm filters utilizing an enzymatic and mechanical lysis procedure described in detail elsewhere [158]. The V3-V4 hypervariable region of the 16S rRNA gene was PCR-amplified and sequenced on the Illumina HiSeq (Illumina, San Diego, CA, United States) utilizing a dual-indexing strategy for multiplexed sequencing developed at the Institute for Genome Sciences [241], described in detail elsewhere [159]

3.3.5 16S rRNA gene data analysis

16S rRNA gene reads were initially screened for low quality bases and short read lengths [241], paired reads were merged using PANDAseq [242], de-multiplexed, trimmed of artificial barcodes and primers, and assessed for chimeras using UCHIME in de novo mode, as implemented in Quantitative Insights Into Microbial Ecology (QIIME; version 1.9.1-20150604) [243]. Quality-controlled reads were clustered at 97% *de novo* into operational taxonomic units (OTUs) with the SILVA 16S database [244] in QIIME [243]. All sequences taxonomically assigned to chloroplasts were removed from further downstream analysis. When appropriate, data was normalized using metagenomeSeq's cumulative sum scaling to account for uneven sampling depth [245].

To visualize the relative abundance of bacterial phyla, stacked bar charts were generated using ggplot2. In addition, bacterial taxa were summarized at the genera level in QIIME (level = 6) and those with a maximum relative abundance greater than 1% in at least one sample were used to build a heatmap via R (ver. 3.3.2) and vegan heatplus [246]

Significance tests were conducted using a Tukey's HSD Test between filter size fractions and among sampling dates. Additionally, Pearson correlation coefficients were calculated to identify associations between the water characteristics and the relative abundance of the bacterial phyla/genera.

Alpha diversity was calculated using the R packages: Bioconductor [247], metagenomeSeq [245], vegan [248], phyloseq [249], fossil [250], biomformat [249,251], and ggplot2 [183] on data rarefied to an even sampling depth (55,307 sequences) and tested for significance using a Tukey's HSD Test.

Beta diversity was determined through principle coordinates analysis (PCoA) plots of Bray-Curtis, Weighted and Unweighted UniFrac distances calculated using the R package phyloseq and tested for significance with ANOSIM (9,999 permutations) [249, 252–254].

Core bacterial OTUs for each filter fraction, and the sample as a whole, were defined as OTUs in 100% of samples determined with QIIME's compute core microbiome.py script [244]. Core OTUs were then visualized using ggplot2, and Krona [183, 255].

3.3.6 Virome metagenomic analysis

The paired-end reads were quality trimmed using Trimmomatic (ver. 0.36) [175], merged with FLASh (ver. 1.2.11) [176], and assembled *de novo* with metaSPAdes (ver. 3.10.1) without read error correction [177]. Taxonomic classifications were assigned to contigs by searching predicted peptide open reading frames (ORFs) against the peptide SEED and Phage SEED databases (retrieved 11/17/2017) [256]. Briefly, peptide ORFs were predicted from virome contigs using MetaGene [104] and were searched against the SEED and Phage SEED databases using protein-protein BLAST (BLASTp ver. 2.6.0+) (E value $\leq 1e^{-3}$) [105]. Taxonomy was assigned to contigs using ORF BLASTp hits to SEED sequences with NCBI taxonomy IDs. A contig is assigned the NCBI taxonomy ID with the maximum sum bit score across all ORF BLASTp hits in the contig.

Functional classifications were assigned to peptide ORFs by searching predicted peptide ORFs against the same peptide SEED database. Peptide ORFs were searched against the SEED databases using BLASTp (E value $1e^{-3}$). Peptide ORFs were assigned to a SEED subsystem with the maximum sum bit score among all of the ORF's hits. Only functions associated with viral hits were considered for this analysis.

Counts for taxonomy and functions identified in each virome are based on normalized contig and peptide ORF abundances, respectively. The abundance for each contig was estimated by recruiting all quality controlled reads to the assembled contigs using Bowtie2 (ver. 2.3.3) in very sensitive local mode [180]. Coverage for each contig was calculated using Samtools depth [257] and a custom parser available at https://github.com/dnasko/baby_virome. ORF abundances are calculated by computing the coverage of the contig within the ORF's start and stop coordinates. To compare abundance measurements between viromes they are normalized to account for sequencing effort and assembly/recruitment proficiency. Briefly, each contig/ORF's coverage is divided by the giga base pairs (Gbp) of reads able to recruit back to contigs/ORFs. Taxonomic and functional data were visualized using ggplot2 [183]. Additionally, Pearson correlation coefficients were calculated to identify the associations between the water characteristics and the abundance of the predicted viral taxa.

For the phylogenetic marker gene, predicted peptide ORFs were queried against Pol I UniRef90 [106] clusters using protein-protein BLAST (BLASTp) [105] with an

E value cutoff $\leq 1e^{-5}$. Significant hits were filtered based on length ($\geq$ 100 amino acids) and then confirmed to be Pol I via NCBI's Conserved Domain BLAST online tool [258]. The sequences were then aligned with MAFFT using the FFT-NS-i-1000 algorithm [186]. To obtain biologically meaningful data on the Pol I-containing phage, a region of interest (I547 to N923 in *E. coli* IAI39) containing the Phe762 position relative to *E. coli* IAI39 was selected and used to create an unrooted maximum likelihood tree with 100 bootstrap replicates using Geneious 6.0.5 [259] with PhyML [260]. Those with a Phe762 or Tyr762 were defined as generally virulent phage, while those with a Leu762 were defined as generally temperate [238]. Abundances for each Pol I peptide were calculated as described above.

3.3.7 Data deposition

16S rRNA gene sequences were deposited in NCBI Sequence Read Archive under the accession numbers SRX3387709-SRX3387732. Viral metagenomic reads were also deposited in NCBI's Sequence Read Archive under the accession numbers SRS2698857, SRS2698856, and SRS2698858.

3.4 Results

3.4.1 Water characteristics

Water properties are described in Table 3.1. Overall, ambient temperature during sampling, water temperature, nitrate and chloride levels, and turbidity decreased during the study period. Conversely, pH, precipitation levels, conductivity

(SPC uS/cm), oxidation/reduction potential (mV), and dissolved oxygen (%) were highest in December.

3.4.2 16S rRNA gene sequencing effort

In total, 24 samples were PCR-amplified for the 16S rRNA gene and sequenced: four quadrants each from the 1 and 0.2 μm filters from each sampling date (October, November, and December). After sequence quality filtering, 7,489 OTUs (97% identity) were identified from a total of 2.5 million sequences across all samples, with an average number of ~103,000 ($\pm$ ~36,000 SD) sequences per sample and an average of ~2,100 ($\pm$ ~500) OTUs.

3.4.3 Bacterial community composition and temporal variations

The most abundant pond water phyla were *Actinobacteria*, *Proteobacteria*, and *Bacteroidetes* in all samples, however, their average relative abundance fluctuated over the time course and between filter pore sizes (Figure 3.1 and Table 3.2). For instance, *Actinobacteria* was significantly ($p \leq 0.05$) higher at all time points in the 0.2 μm fraction, whereas *Chloroflexi*, *Firmicutes*, *Cyanobacteria*, and *Proteobacteria* were significantly higher in the 1 μm fraction at all time points (Figure 3.1 and Table 3.1). From October to December in both fractions the relative abundance of *Bacteroidetes* increased significantly ($p \leq 0.05$), whereas *TM7* and *Cyanobacteria* decreased 3.3).

At the genera level, summarized via QIIME, there were 31 taxa that were greater than 1% in at least one sample Figure 3.1 . Of these, eight had a significantly ($p \leq 0.05$) higher relative abundance at all time points in the 0.2 μm fraction than the 1 μm fraction (*ACK.M1*, *Limnohabitans*, *Microbacteriaceae*, *Sediminibacterium*, *Polynucleobacter*, *Actinomycetales*, *Cytophagaceae*, *ZB2*), while six were significantly ($p \leq 0.05$) greater in the 1 μm fraction compared to the 0.2 μm fraction (*Synechococcus*, *Chitinophagaceae*, *Dolichospermum*, *Rhizobiales*, *SC.I.84*, *Gemmataceae*) (Figure 3.1 and Table 3.4).

Moreover, within each fraction there were significant ($p \leq 0.05$) changes in the relative abundance of the bacterial taxa over the sampling period (Figure 3.1 and Table 3.5). For instance, in the 1 and 0.2 μm fractions, *ACK.M1*, *Fluviicola*, *Sphingomonadales*, *Dolichospermum*, *Flavobacterium*, *Bacteroidetes*, *SC.I.84*, and *Betaproteobacteria* increased significantly ($p \leq 0.05$) from October to December, while *C111*, *Synechococcus*, *Chitinophagaceae*, *Rhodoluna*, *Actinomycetales*, *Rhodobacter*, *Rhizobiales*, *Mycobacterium*, *Rhizobiales*, *SC3*, and *Gemmataceae* decreased significantly ($p \leq 0.05$). In addition to the taxa above, in the 1 μm fraction there was also a significant ($p \leq 0.05$) increase in *Comamonadaceae* and *Polynucleobacter* and a significant decrease in *Sphingobacteriales* and *OD1* between the October and December sampling dates. For the 0.2 μm fraction, there was also a significant ($p \leq 0.05$) increase in the relative abundance of *Sediminibacterium*, *Cytophagaceae*, *ZB2*, and *ABY1* between the October and December sampling dates.

3.4.4 Relationships between water characteristics and bacterial abundance

Despite the small sample size ($n = 3$), several bacterial phyla showed significant correlations ($p \leq 0.05$) with the measured water characteristics Figure 3.2. In both filter fractions, the relative abundance of *Actinobacteria* was negatively correlated with the level of dissolved oxygen. In just the 1 μm fraction, the relative abundance of *Verrucomicrobia* and *Proteobacteria* were positively correlated with dissolved oxygen and pH, respectively, while *TM7* was positively correlated with both water temperature and turbidity. In addition, in just the 0.2 μm fraction, the relative abundance of *Firmicutes* showed positive correlations with pH and oxidation/reduction. Conversely, the relative abundance of *Bacteroidetes*, *TM7*, and *OD1* were negatively related to, chloride, pH, and nitrate, respectively ($p \leq 0.05$). At the genera level in both filter fractions, *Synechococcus* was positively correlated with nitrate (Figure 3.2).

3.4.5 Bacterial alpha diversity

Alpha diversity was computed for Observed OTUs and Shannon diversity and tested for significance between filter size and over time within each fraction (Figure 3.3). Despite the 0.2 μm fraction containing significantly higher levels of some of the dominant genera, the 1 μm fraction had a significantly higher Shannon index and observed OTU value in the November and December ($p \leq 0.05$) samples (Figure

3.3). When testing within each fraction over time, the Shannon index values in the 1 μm fraction were significantly ($p \leq 0.05$) higher in October than November. In the 0.2 μm fraction the Shannon index values were significantly higher in October than both November and December (Figure 3.3). However, the Observed OTU values in the 1 μm fraction were significantly higher in December than both October and November, while in the 0.2 μm fraction December was only significantly higher than November.

3.4.6 Bacterial beta diversity

Beta diversity comparisons using PCoA plots of Bray-Curtis distances computed for all samples revealed significant clustering by date (R = 0.94, p = 0.001) along axis 1, which accounted for nearly 45% of the variation (Figure 3.4). Samples along axis 2 (17% of the variation) appeared to cluster by filter size (1 μm vs. 0.2 μm). This trend was also observed both utilizing weighted (R = 0.7458, p = 0.001) and unweighted (R = 0.8524, p = 0.001) UniFrac (Figure 3.4) distances. Additionally, within each date, samples clustered by filter pore size: December (R = 1, p = 0.022), November (R = 1, p = 0.022), and October (R = 1, p = 0.029) (Figure 3.5).

3.4.7 Core OTUs

There were 277 core OTUs present in the farm pond water over the 3 months across both fractions (Figure 3.6). These were largely *Actinobacteria* (127 OTUs), followed by: *Proteobacteria* (82 OTUs), *Bacteroidetes* (43 OTUs), *TM7* (9 OTUs), *OD1* (4 OTUs), *Verrucomicrobia* (2 OTUs), *Firmicutes* (2 OTUs), *Chloroflexi* (2 OTUs), *Cyanobacteria* (1 OTU), *WS5* (1 OTU), *SR1* (1 OTU), *GN02* (1 OTU), *Fusobacteria* (1 OTU), and *Chlamydiae* (1 OTU). For the 0.2 μm fraction, there was a unique core of 152 OTUs, largely *Actinobacteria* (90 OTUs), followed by *Proteobacteria* (33 OTUs), and *Bacteroidetes* (23 OTUs), *Verrucomicrobia* (2 OTUs), TM7 (1 OTU), GN02 (1 OTU), *Chlamydiae* (1 OTU), and *Spirochaetes* (1 OTU). The unique core for the 1 μm fraction was more diverse in bacterial phyla compared to the 0.2 μm. It consisted largely of *Proteobacteria* (78 OTUs), followed by *Actinobacteria* (16 OTUs), *Bacteroidetes* (16 OTUs), *Cyanobacteria* (6 OTUs), *Firmicutes* (6 OTUs), *Chloroflexi* (5 OTUs), *TM7* (5 OTUs), *Planctomycetes* (4 OTUs), *Acidobacteria* (4 OTUs), *Verrucomicrobia* (3 OTUs), *Gemmatimonadetes* (2 OTUs), *OD1* (1 OTU), and *WS5* (1 OTU).

When looking at the core bacterial taxa at a lower taxonomic level, in the 0.2 μm fraction the *Actinobacteria* OTUs were dominated by *Actinomycetales* ACK-M1 (70%), followed by *Rhodoluna* (4%) under the family *Microbacteriaceae* (Figure 3.7). The *Proteobacteria* were largely *Limnohabitans* (24%) and *Polynucleobacter* (30%) and the *Bacteroidetes* were *Sediminibacterium* (22%) and a large unclassified family of *Cytophagaceae* (35%) (Figure 3.7). Within the unique core of the 1 μm fraction, the only prominent genus of the *Proteobacteria* was *Rhobacter* at 5%, followed by *Novosphingobium* at 3% and 17 other genera (each at 1%) (Figure 3.7). At a higher taxonomic level, a family of largely unclassified *Rhizobiales* was also abundant (23%). The *Actinobacteria* were dominated by *Acidimicrobiales C111* (25%), *Solirubrobacterales* (13%), and a large proportion of mostly unclassified *Actinomyc-*

etales (44%). Finally, in the *Bacteroidetes* phylum of the 1 μm fraction there was mostly *Fluviicola* at 13%, as well as *Leadbetterella*, *Runella*, and *Sediminibacterium*, which were each at 6%.

3.4.8 Shotgun sequencing effort and assembly

Each sample from the viral fraction was sequenced on the Illumina HiSeq for a total of 89,645,509 read pairs (35,522,822 from October; 32,988,451 from November; 21,134,236 from December). We assembled the reads from all sampling dates to construct a total of 872,200 contigs (272,687 from October; 409,758 from November; 189,755 from December). The mean contig length was 593 nucleotides in October (range from 55 to 366,802 nucleotides), 641 in November (range from 55 to 286,691 nucleotides), and 655 in December (range from 55 to 331,733 nucleotides). The GC content was similar among the three sampling dates: October (Mean 46.51%, Median 45.42%); November (Mean 46.18%, Median 45.59%); and December (Mean 46.12%, Median 45.21%).

3.4.9 Viral taxonomic composition and abundance

Similar to other studies, a large portion of the assembled contigs within the viral fraction had no known homologs (42% October, 51% November, and 42% December could be assigned taxonomy) [261]. For those that did have a hit we calculated the normalized abundance. From this we found that 47% (October), 27% (November), and 53% (December) of the assigned taxonomic composition were homologous to viruses; the vast majority of which were from the dsDNA bacteriophage *Caudovirales* (99%, 98%, and 99%) (Figure 3.8). Within the *Caudovirales*, *Siphoviridae* dominated at all time points followed by *Myoviridae* and then *Podoviridae*. Other viral categories included those homologous to ssDNA viruses *Inoviridae*, dsDNA viruses *Tectiviridae*, *Ligamenvirales*, *Bicaudaviridae*, and Figure 3.8). For the dominant viral families (*Podoviridae*, *Siphoviridae*, and *Myoviridae*), there were no water characteristic that significantly correlated with their abundances; however, there were several bacterial taxa whose relative abundance correlated with the abundance of the dominant viral families (Figure 3.9). For instance, at the phylum level the relative abundance of OD1 in the 1 μm fraction negatively correlated with the abundance of *Siphoviridae*, while in the 0.2 μm the relative abundance of *TM7* negatively correlated with the abundance of *Podoviridae* (Figure 3.9).

3.4.10 Viral functional composition

Peptide ORFs from the virus-assigned contigs were functionally annotated using the SEED Subsystems [256]. Again, to compare these viral categories across the three months, we calculated the normalized abundance for each of the peptide ORFs assigned to the SEED functional categories (Figure 3.8). While peptide ORFs homologous to virulent genes were present (e.g., Multidrug Resistance Efflux Pumps, Zinc Resistance, Copper Homeostasis) they were not abundant within the time period. The majority (93% October, 80% November, and 92% December) were Phage Elements, defined by the SEED subsystem hierarchy [256] as either "Phages, Prophages, Transposable elements" , which were largely phage structural genes (e.g., capsid, scaffolding) or "Phages, Prophages, Transposable elements, and Plasmids", which were genes related to phage replication and packaging (e.g., terminase, integrase, helicase, primase). This was followed by "Nucleosides and Nucleotides" (4% October, 6% November, and 5% December), which were largely genes involved in ribonucleotide reduction.

3.4.11 Viral marker gene: Polymerase A

A total of 842 confirmed Pol I peptides were extracted from our assembled pond water viromes (271 October, 320 November, and 251 December). From these, only 228 spanned the region of interest and were included in the phylogenic analysis: 68 in October, 83 in November, and 77 in December. The phylogenetic analysis (Figure 3.10) showed that the Pol I peptides grouped largely by their 762 position, whereas the sampling dates were distributed among the different clades. Again, to compare the Pol Is we calculated the normalized abundance for each. While the majority of Pol I peptides had the Leu substitution at the 762 site (184), followed by Tyr (27) and then the wildtype Phe (17), the clade at the highest abundance was those with the wildtype Phe mutation. Because this clade was so abundant, we used BLASTp to assess the top hit, which were uncultured bacteriophage from the Dry Tortugas surface water [238] and the seawater collected from the deep chlorophyll maximum of the Mediterranean Sea [262] (E value $< 1e^{-300}$).

3.5 Discussion

While ponds represent a potential source of irrigation water, provided adequate filtering and monitoring technologies are employed, it is important to keep in mind their role in the ecosphere. Their small size and shallow depth enables a complex community of aquatic plants and macroinvertebrate species, as well as an interacting community of microorganisms. However, changing water levels and anthropogenic factors associated with irrigation systems may interrupt the delicate balance of microbial life, which can ultimately impact higher trophic levels. Therefore, it is critical to carefully manage our use, or contamination, of these systems in search of irrigable water.

Here, the studied pond was dominated by *Actinobacteria*, *Proteobacteria*, and *Bacteroidetes*, common phyla found in most freshwater lakes [119]. However, as the late season progressed there appeared to be changes in the bacterial community composition that correlated with specific environmental factors. For instance, *Actinobacteria* fluctuated throughout the sampling period, and was negatively correlated with the percentage of dissolved oxygen (Figure 3.1, 3.2). A previous study of the Luhuitou fringing coral reef also reported a negative correlation between dissolved oxygen and the abundance of several *Actinobacteria* species that may be due to their anaerobic capabilities [263]. This relationship may help to explain the fluctuations in freshwater *Actinobacteria* abundances and diversity described in other seasons [264, 265]. At a lower taxonomic level, the relative abundance of *Synechococcus* decreased significantly throughout the late season (Figure 3.1). This is not surprising as the growth rate of *Synechococcus* is known to decrease with temperature and nitrogen levels, which occurred during the sampling period [266, 266].

While the changes in relative abundance described above were apparent in both filter fractions (1 and 2 μm), there were differences in alpha diversity, beta diversity, and core bacterial composition between the two fractions (Figures 3.3, 3.4, 3.6). For instance, the 0.2 μm core microbiota was dominated by *Actinobacteria*, largely *Actinomycetales ACK-M1*, a known representative in most freshwater habitats [267] and free-living *Proteobacteria*, *Limnohabitans*, and *Polynucleobacter*, [268–270]. Whereas, the core of the 1 μm filter fraction was dominated by a diverse set of *Proteobacteria*, largely of the order *Rhizobiales* and *Rhodobacteriales*, previously found to dominate the particle-associated fraction in a marine pelagic trench [271].

Furthermore, the bacterial richness was significantly higher in the 1 μm fraction in December and November (Figure 3.3). This suggests that the diversity of the large/attached bacteria recovered from the 1 μm filter were greater in the later months (November and December) compared to the smaller/free-living bacterial communities of the 0.2 μm filter. This increased diversity in the larger fraction agrees with studies investigating the differences between the particle-associated and free-living bacteria in other aquatic systems, such as the Baltic and Mediterranean Seas [265, 272]. While the alpha diversity trends were similar between the two filter fractions, the increasing richness was more prominent in the 1 μm filter fraction. There can be several factors that may influence this discrepancy in temporal diversity, such as changing protozoan grazing rates and/or viral lysis impacting the microbial food web [270,273]. Additionally, it can be suggested that the attached/large bacteria of the 1 μm fraction may have been better equipped to compete in an environment of decreased water temperature, nitrate, and chloride levels (Figure 3.1). In fact, previous studies have reported particle-associated bacteria have larger genomes compared to streamlined free-living bacteria [274]. This reservoir of genetic material may allow for a more adaptive lifestyle in the face of changing conditions, competing bacterial populations and predators.

The viral fraction was dominated by sequences homologous to the tailed, double stranded DNA bacteriophage of the order *Caudovirales* (Figure 3.8). Within the *Caudovirales* order, however, the family of generally temperate *Siphoviruses* was the most abundant at each time point. This is in contrast to other freshwater systems, where *Myoviruses* and *Podoviruses* (Lakes Ontario, Erie, Matoaka, and Michigan [136, 137, 145] were found to dominate. Conversely, *Siphoviruses* were reported to be abundant in sediments [115] and terrestrial subsurface environments [116]. *Siphovirus* abundance within the pond water sampled here may reflect the pond's unique topography compared to larger lakes. Ponds and small lakes have a higher terrestrial-aquatic interchange than larger freshwater systems due to their close contact with the adjacent terrestrial environment [233,275]. This large littoral zone may promote the influx of terrestrial/sediment *Siphoviruses*. Additionally, because the dominant lifecycle of cultured *Siphoviruses* [276] is lysogenic, their abundance may be indicative of environmental stress, (e.g., changes in nutrients, pH or temperature) activating the lytic-lysogenic switch and thus increasing their presence in the free phage fraction [277]. This agrees with the phylogenetic analysis of the informative viral marker gene family A DNA polymerase, polA (Pol I protein). In this case, the Pol I proteins were temporally persistent and dominated by the Leu762 mutation, suggested to be indicative of temperate phage (Figure 3.10) [238].

This abundance of temperate phage is important to note as they can alter the genetic architecture of their hosts, which can in turn influence the surrounding microbial community and environment. For instance, temperate phage can transduce bacterial DNA and potentially alter host biology, encoding virulence and antibiotic resistance genes [71, 278]. In this case, while present, genes involved in virulence/antimicrobial resistance were not enriched within the viral fraction during the study period (Figure 3.8). However, the samples did contain a high abundance of genes involved in Nucleosides and Nucleotides production, largely ribonucleotide reduction genes. This is not surprising as ribonucleotide reductases (RNRs) have been observed frequently within aquatic viral metagenomic libraries and namely in *Myoviruses* [279, 280, 280].

From these data we were able to characterize the bacterial and viral taxonomic and functional components within an agricultural pond over time. In several cases, we observed the abundance of the dominant viral families correlated with the relative abundance of bacterial taxa. However, more work is necessary to track and model these interacting species. A more thorough analysis connecting phage with their hosts, such as through the use of the Clustered regularly interspaced short palindromic repeats (CRISPR) system, is warranted to provide a more nuanced assessment of their relationship in pond water and potential impact in the microbial food web (e.g., viral shunt). Additionally, other parameters like phosphate, dissolved organic carbon, chlorophyll concentration, and protozoan grazing rates might also exert some influence on the temporal dynamics observed here and should be included in future studies of freshwater ponds. Overall, this analysis serves as a baseline of the diversity and dynamics of three distinct microbial fractions in agricultural pond water.

3.6 Figures

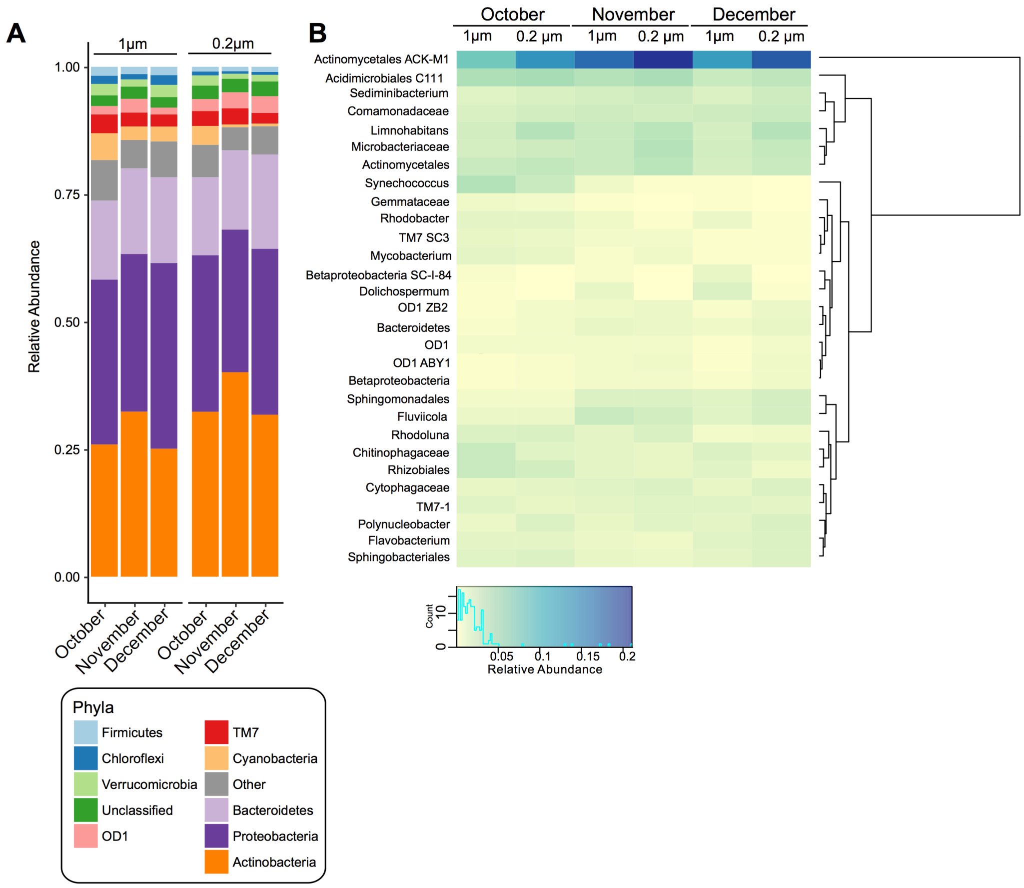

Figure 3.1: Bacterial community comporsition and diversity of 1 and 0.2 μm filter fractions over time. (A) Stacked bar chart of the relative abundance of the bacterial community composition at the phylum level within pond samples from each month (October, November, and December) and each filter fraction. (B) Heatmap based on the relative abundance of the bacterial community composition at the genus level. Displayed are genera or lowest available taxonomic assignments representing more than 1% in at least one of the pond samples. Pooled samples are clustered using Manhattan distance. Color key depicts the spectrum of relative abundance with a histogram of the counts of individual values.

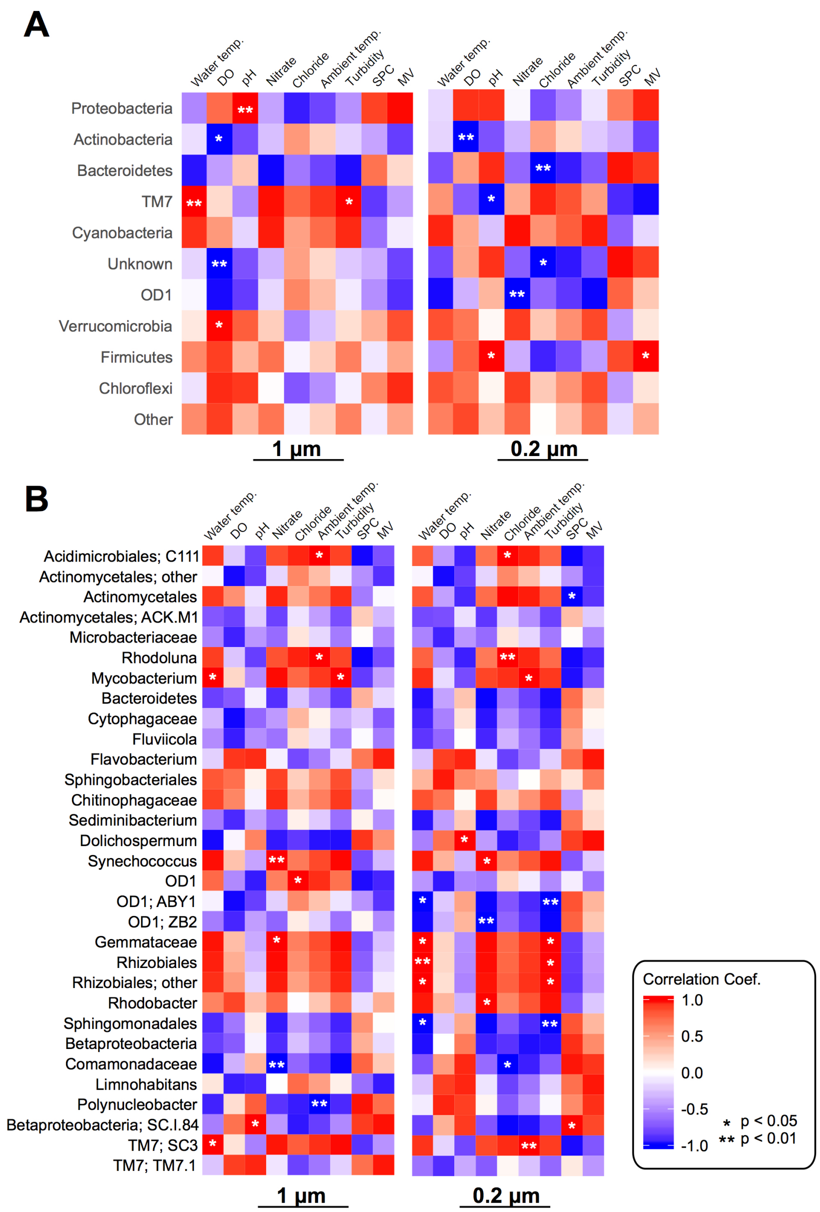

Figure 3.2: Heatmaps of the Pearson's correlation coefficients between the water characteristics and relative abundance of bacterial (A) phyla and (B) genera for the 1 μm and 0.2 μm filter fractions. Color gradients reflect the different values of Pearson's correlation coefficients. MV: Oxidation/reduction (mV), SPC: Conductivity (SPC uS/cm), DO: Dissolved Oxygen (%).

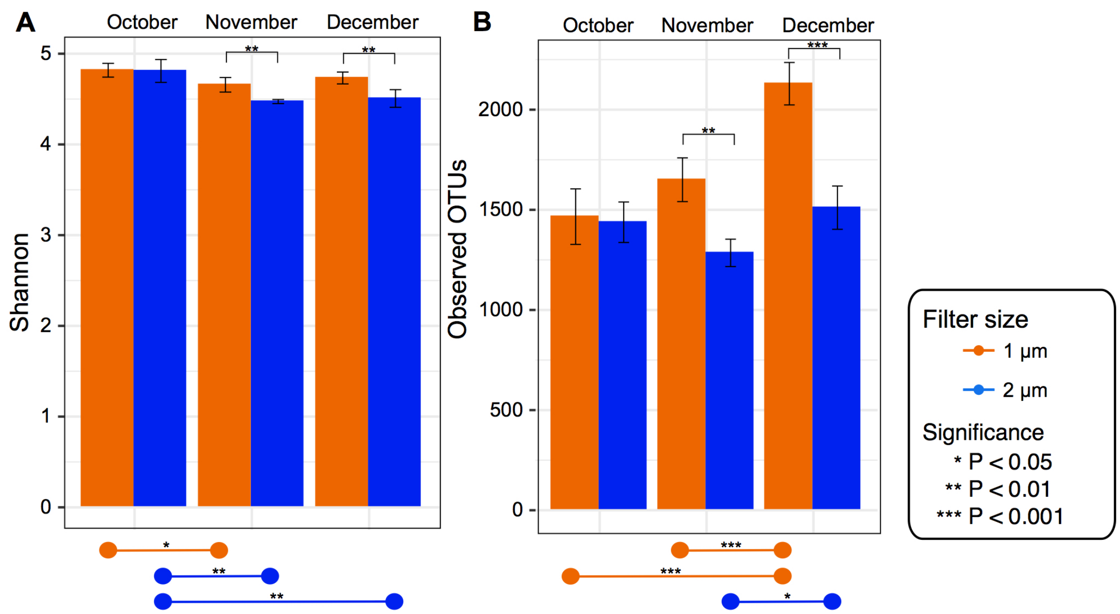

Figure 3.3: Alpha diversity for each filter fraction in late season pond water. Bar charts of alpha diversity measured using (A) Shannon index and (B) Observed OTUs. Color denotes filter pore size, either 1 μm (orange) or 0.2 μm (blue). Pairwise significance between filter size denoted by brackets within graph for each date. Significance within each filter among time points denoted by lines at bottom of the figure. Error bars denote standard deviation.

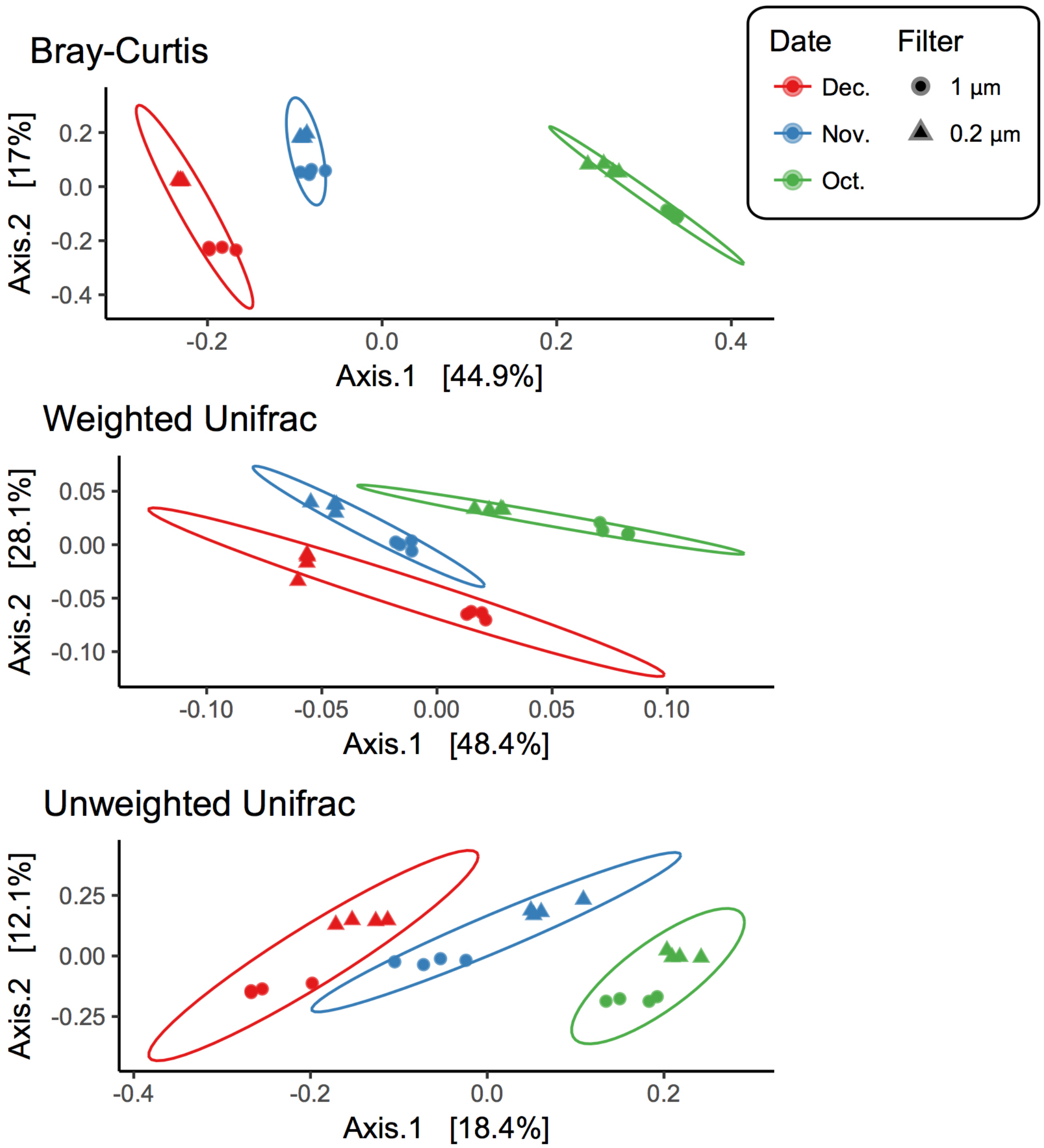

Figure 3.4: Beta diversity for each filter fraction in late season pond water. PCoA plots of beta diversity measured using (A) Bray-Curtis, (B) Weighted UniFrac, and (C) Unweighted UniFrac. Shape denotes filter pore size, either 1 μm (circle) or 0.2 μm (triangle), and color denotes month when water was sampled, October (green), November (blue), and December (red). Ellipses are drawn at 95% confidence intervals for each month.

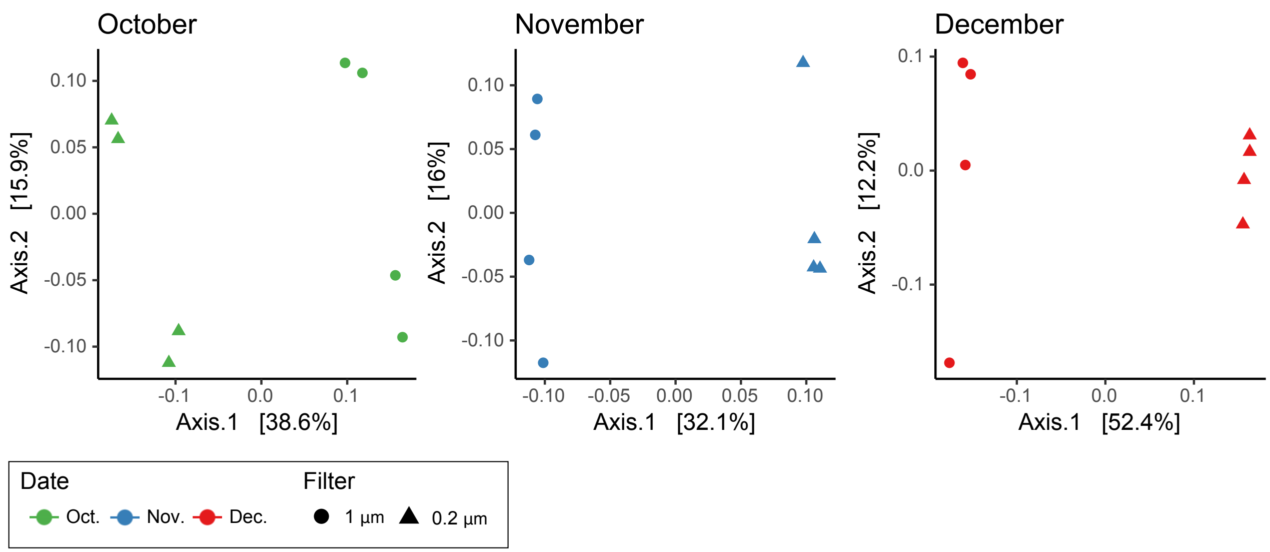

Figure 3.5: PCoA plots of beta diversity (by date) measured using Bray-Curtis. Shape denotes filter pore size, either 1 μm (circle) or 0.2 μm (triangle), and color denotes the month that water was sampled, October (green), November (blue), and December (red).

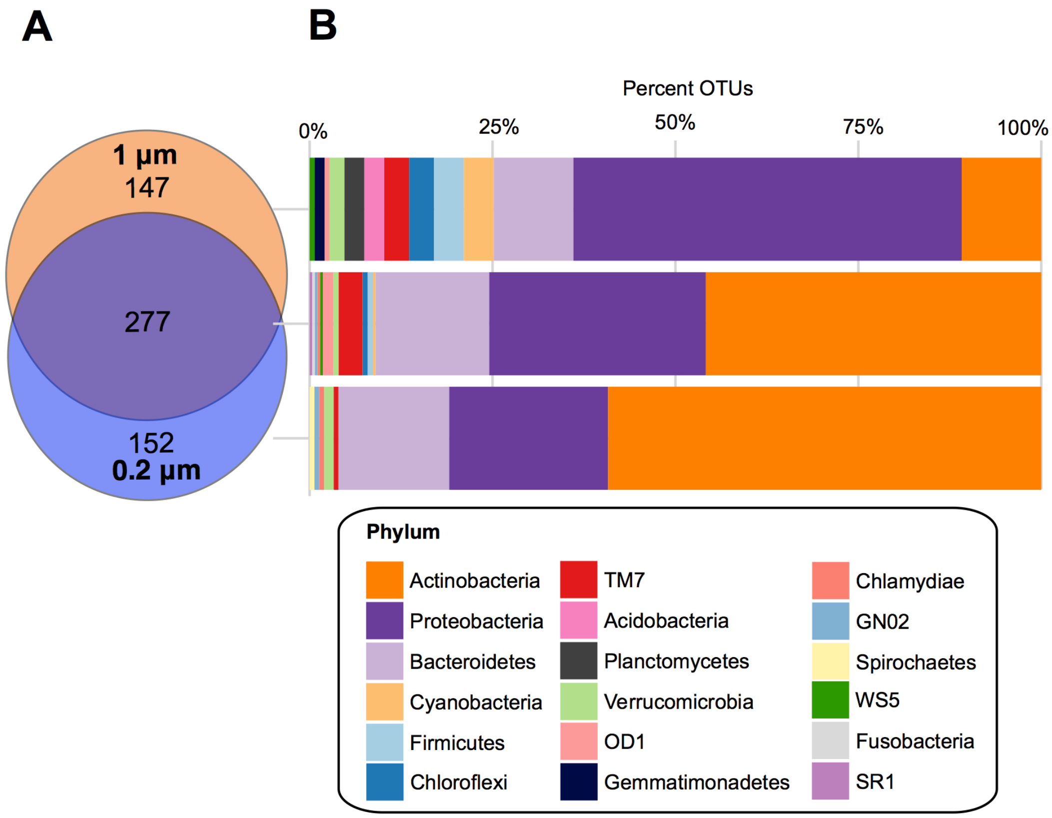

Figure 3.6: Core OTUs for 1 and 0.2 μm filter fractions. (A) Venn diagram depicting the unique and shared OTUs that occurred in 100% of samples in the 1 μm (orange) and 0.2 μm (blue) filter fractions. (B) Stacked bar charts of the percentage of OTUs assigned each phylum within the shared and unique cores.

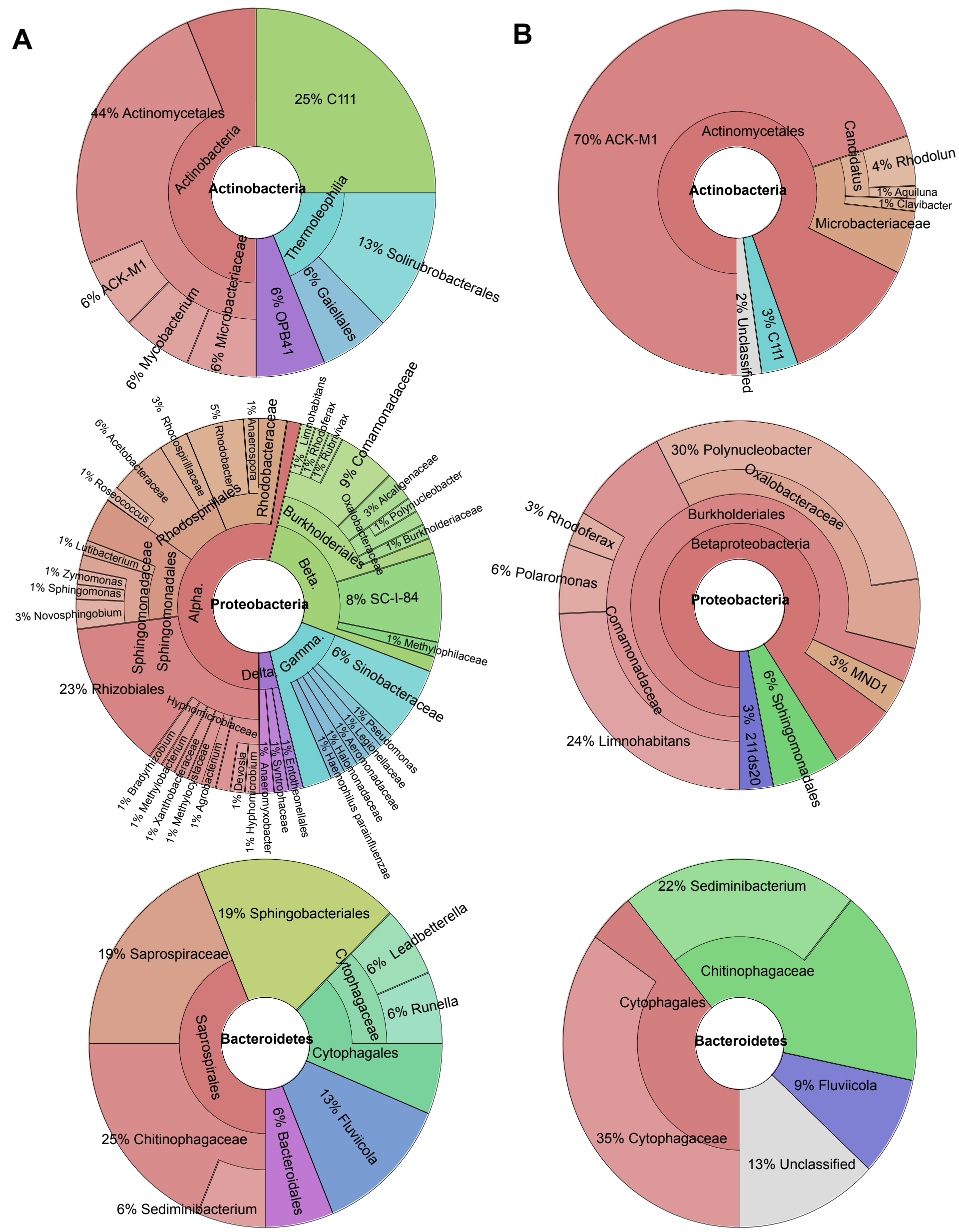

Figure 3.7: Core *Actinobacteria*, *Proteobacteria*, and *Bacteroidetes* OTUs for 1 μm and 0.2 μm filter fractions during the entire sampling period. Krona plots depicting the core OTUs in the (A) 1 μm and (B) 0.2 μm filter fractions for those assigned to the *Actinobacteria*, *Proteobacteria*, and *Bacteroidetes* phyla.

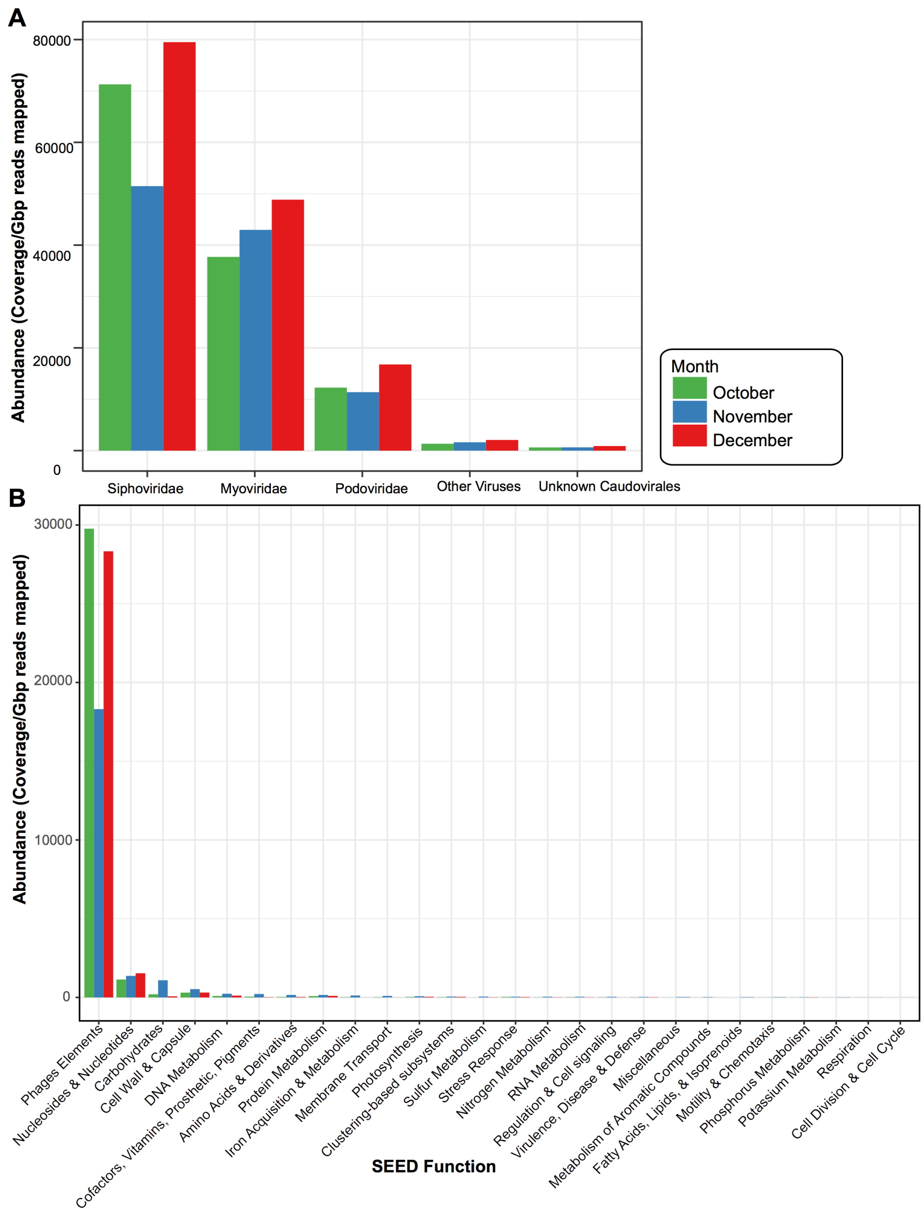

Figure 3.8: Viral taxonomy and function for each sampling date. Bar plots comparing the normalized abundance of viruses with homology to known (A) viral taxa and (B) functional SEED categories. Bar color denotes month when water was sampled, October (green), November (blue), and December (red). Phage elements refer to the SEED functional category "Phages, Prophages, Transposable elements" and Phage, Plasmids, etc. denote the SEED functional category "Phages, Prophages, Transposable elements, Plasmids". Abundance determined by calculating coverage/Gbp reads mapped.

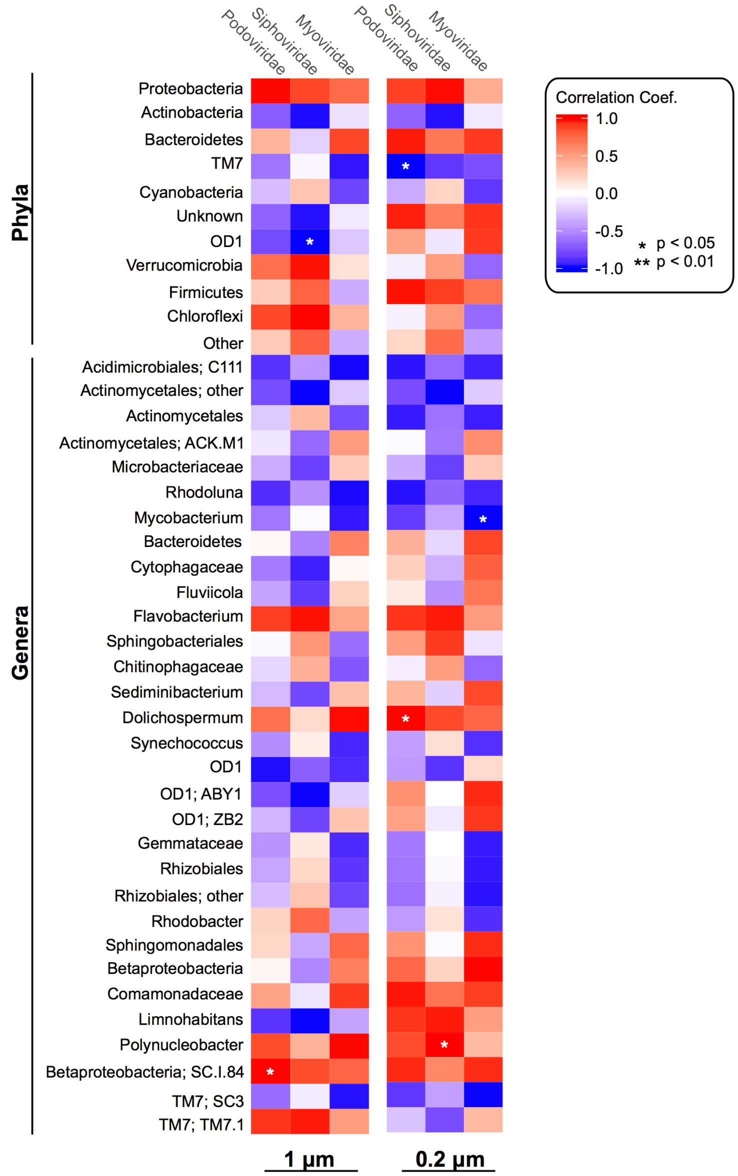

Figure 3.9: Heatmaps of the Pearson's correlation coefficients between dominant viral families and the relative abundance of bacterial phyla and genera in both the 1 μm and 0.2 μm filter fractions. Color gradients reflect the different values of Pearson's correlation coefficients.

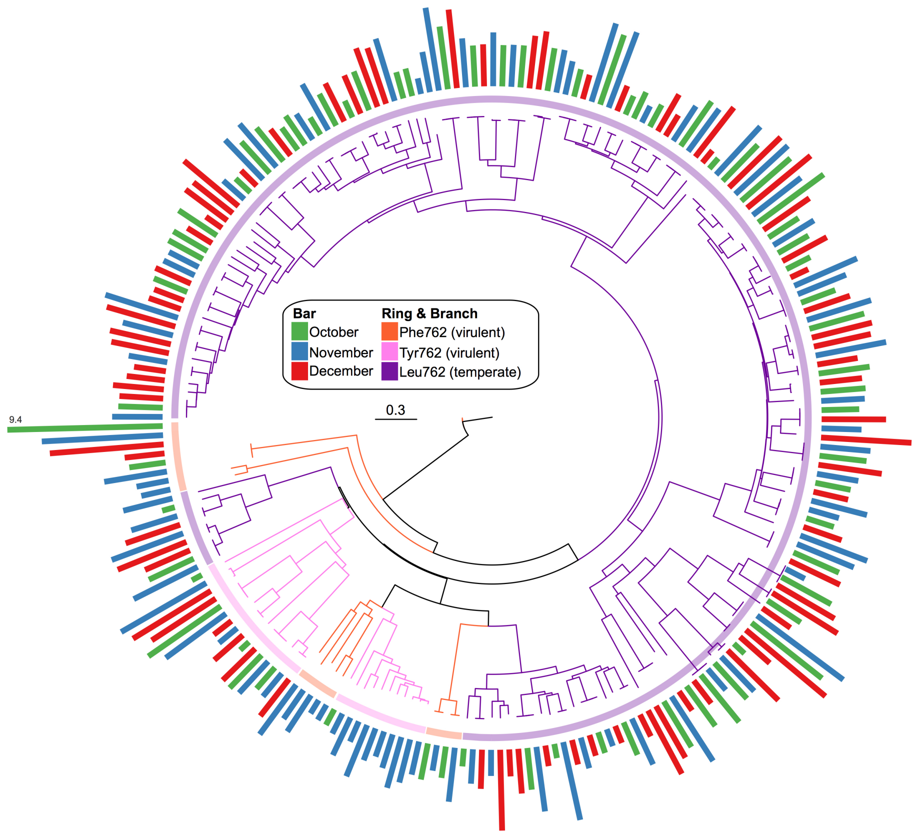

Figure 3.10: Phylogenic tree of viral Pol I peptides for each sampling date. Unrooted maximum likelihood tree with 100 bootstrap replicates of representative Pol I peptide sequences. Branches and inner ring colored by 762 position residues, Phe762 (orange), Tyr762 (pink), and Leu762 (purple). Radial bar chart represents the log2-normalized abundance of each peptide (coverage/Gbp reads mapped) and is colored by sampling date, October (green), November (blue), and December (red).

3.7 Tables

Table 3.1: Agricultural pond water characteristics during sampling period.

Water property	**October**	**November**	**December**
Ambient Temp. (°C)	17.2	12.2	3.9
Water Temp. (°C)	19.8	10.9	7.4
pH	7.7	7.56	8.08
Dissolved Oxygen (%)	116.4	96.4	117.7
Nitrate (mg/L)	0.63	0.26	0.19
Chloride (mg/L)	13.8	13.3	7.9
Turbidity (FNU)	30.2	9.6	3.4
Precipitation (in.)[a]	0	0	0.2
Conductivity (SPC uS/cm)	158.9	160.8	167.1
Oxidation/reduction (mV)	189.7	159.8	243.9

[a]Precipitation 24 hr prior to sampling

Table 3.2: Difference (%) in relative abundance between 1 μm and 0.2 μm fractions for the dominant bacterial phyla.

Phyla	**October**	**November**	**December**
Proteobacteria	1.65*	2.93**	3.89**
Actinobacteria	-6.42**	-7.72**	-6.67**
Bacteroidetes	0.19	1.24**	-1.67*
TM7	0.75**	-0.44**	0.26
Cyanobacteria	1.55**	2.13**	2.38**
Unknown	-0.53	-0.24	-0.82**
OD1	-0.7	-0.46	-1.91**
Verrucomicrobia	0.23	0.42**	1.09**
Firmicutes	0.83**	0.55**	0.62**
Chloroflexi	0.83**	0.56**	1.35**
Other	1.6*	1.02	1.49**

$^*p \leq 0.05$, $^{**}p \leq 0.01$

Table 3.3: Difference (%) in relative abundance of dominant bacterial phyla between sampling dates in 1 μm and 0.2 μm filter fractions

	1 μm			2 μm		
Phyla	**Oct-Nov**	**Nov-Dec**	**Oct-Dec**	**Oct-Nov**	**Nov-Dec**	**Oct-Dec**
Proteobacteria	1.42	-5.5**	-4.08**	2.7	-4.55**	-1.85
Actinobacteria	-6.44**	0.073**	0.83	-7.74**	8.33**	0.58
Bacteroidetes	-1.28*	-0.05	-1.33*	-0.22	-2.96**	-3.19**
TM7	0.95**	0.37*	1.32**	-0.24	1.07**	0.83**
Cyanobacteria	2.58**	-0.23	2.35**	2.79**	-0.36	2.8**
Unknown	-0.31	0.34	0.02	-0.03	-0.24	-0.27
OD1	-1.03*	1.33**	0.29	-0.79	-0.13	-0.92
Verrucomicrobia	0.85**	-0.98**	-0.13	1.03**	-0.31	0.72**
Firmicutes	0.32	-0.19	0.13	0.04	-0.13	-0.09
Chloroflexi	0.56**	-0.88**	-0.32*	0.29**	-0.09	0.2*
Other	2.39**	-1.47**	0.92	1.81*	-1	0.8

$^{*}p \leq 0.05$, $^{**}p \leq 0.01$

Table 3.4: Difference (%) in relative abundance between 1 μm and 0.2 μm fractions for the dominant bacterial genera.

Phyla	**October**	**November**	**December**
Actinomycetales; ACK.M1	-5.74**	-3.96**	-5.31**
Acidimicrobiales; C111	-0.39	-0.76**	-0.58**
Limnohabitans	-1.34**	-0.74**	-1.75**
Microbacteriaceae	-0.43**	-1.07**	-0.64**
Synechococcus	1.35**	0.74**	0.35**
Comamonadaceae	-0.33	0.17	-0.23
Fluviicola	0.01	0.22	-0.62**
Chitinophagaceae	1.24**	0.24*	0.40**
Sediminibacterium	-0.44**	-0.35**	-0.80**
Polynucleobacter	-0.99**	-0.42**	-0.69**
Sphingomonadales	0.00002	-0.09	-0.66**
Actinomycetales; other	-0.69*	-0.55**	-0.73**
Rhodoluna	-0.29	-0.59**	-0.28**
Sphingobacteriales	0.3*	0.21	-0.42**
Dolichospermum	0.2**	1.16**	1.67**
Flavobacterium	-0.02	0.11	-0.23**
Cytophagaceae	-0.36**	-0.18*	-0.57**
Actinomycetales	0.28	-0.27**	0.03
TM7; TM7.1	0.41*	-0.17	0.28
Rhodobacter	0.03	0.47**	0.95**
OD1; ZB2	-0.4**	-0.27*	-0.76*
Rhizobiales	0.3**	0.15*	0.55**
Mycobacterium	0.28*	-0.28*	0.02
Rhizobiales; other	0.06	0.06	0.23**
Bacteroidetes	-0.25	0.02	-0.28**
TM7; SC3	0.21**	-0.18**	0.05
Betaproteobacteria; SC.I.84	0.63**	0.42**	1.24**
OD1; ABY1	-0.16	-0.17	-0.61**
Betaproteobacteria	-0.2*	-0.09	-0.39**
Gemmataceae	0.33**	0.13**	0.2**
OD1	0.04	-0.03	-0.38**

*$p \leq 0.05$, **$p \leq 0.01$

Table 3.5: Difference (%) in relative abundance between 1 μm and 0.2 μm fraction for the dominant bacterial genera.

	1 μm			2 μm		
Phyla	**Oct-Nov**	**Nov-Dec**	**Oct-Dec**	**Oct-Nov**	**Nov-Dec**	**Oct-Dec**
Actinomycetales; ACK.M1	-9.17**	4.21**	-4.96**	-7.39**	2.86**	4.53**
Acidimicrobiales; C111	0.54**	0.99**	1.53**	0.17	1.17**	1.34**
Limnohabitans	-0.46**	0.73**	0.27	0.14	-0.28	0.13
Microbacteriaceae	-0.77**	0.52**	-0.24	-1.4**	0.95**	-0.45
Synechococcus	3.34**	0.61**	3.94**	2.73**	0.22*	2.95**
Comamonadaceae	-0.52**	-0.08	-0.6**	-0.02	-0.48	-0.5
Fluviicola	-1.7**	1.23**	-0.48**	-1.5**	0.39*	-1.11**
Chitinophagaceae	1.48**	-0.3**	1.18**	0.48**	-0.14	0.34
Sediminibacterium	-0.68**	0.43*	-0.25	-0.59**	-0.02	-0.61**
Polynucleobacter	-0.19**	-0.3**	-0.49**	0.38**	-0.58**	-0.19
Sphingomonadales	-1.19**	0.2**	-0.99**	-1.28**	-0.38**	-1.65**
Actinomycetales; other	-0.51**	0.65**	0.14	-0.38	0.48	0.1
Rhodoluna	0.41*	0.85**	1.26**	0.11	1.16**	1.28**
Sphingobacteriales	0.41**	-0.14	0.27*	0.92**	-0.77**	0.15
Dolichospermum	-0.93**	-0.67**	-1.6**	0.04	-0.16**	-0.12
Flavobacterium	0.39*	-0.69**	-0.3*	0.53**	-1.03**	-0.5**
Cytophagaceae	-0.46**	0.43**	-0.04	-0.29**	0.03	-0.25**
Actinomycetales	0.59**	-0.09	0.51**	0.04	0.22	0.26
TM7; TM7.1	0.08	–0.14	-0.07	-0.51**	0.3	-0.2
Rhodobacter	0.66**	-0.37**	0.28**	1.1**	0.1	1.2**
OD1; ZB2	-0.62**	0.40**	-0.22	-0.49	-0.09	-0.58*
Rhizobiales	0.89**	-0.08	0.81**	0.74**	0.32**	1.06**
Mycobacterium	0.93**	0.34**	1.26**	0.44**	0.64**	1.08**
Rhizobiales; other	0.51**	0.02	0.52**	0.51**	0.19*	0.7**
Bacteroidetes	-0.79**	0.26	-0.53**	-0.52**	-0.03	-0.55**
TM7; SC3	0.76**	0.36**	1.12**	0.37**	0.59**	0.95**
Betaproteobacteria; SC.I.84	0.19*	-0.89**	-0.7**	-0.01	-0.07*	-0.08**
OD1; ABY1	-0.28	0.35	0.07	-0.29	-0.09	-0.38
Betaproteobacteria	-0.33**	0.11	-0.23**	-0.23**	-0.19*	-0.42**
Gemmataceae	0.67**	0.1*	0.76**	0.47**	0.16**	0.63**
OD1	0.02	0.39*	0.41*	-0.04	0.03	-0.01

$^{*}p \leq 0.05$, $^{**}p \leq 0.01$

Chapter 4: Seasonal Dynamics in Taxonomy and Function within Bacterial and Viral Metagenomic Assemblages Recovered from a Freshwater Agricultural Pond

4.1 Abstract

Ponds are important freshwater habitats that support both human and environmental activities. However, relative to their larger counterparts (e.g. rivers, lakes) ponds are understudied, especially with regard to their microbial communities. Our study aimed to fill this knowledge gap by using culture-independent, high-throughput sequencing to assess the dynamics, taxonomy, functionality, and interaction history of bacterial and viral communities in a freshwater agricultural pond. Water samples (n=14) were collected from a Mid-Atlantic pond between June 2017 and May 2018 and filtered sequentially through 1 and 0.2 μm filter membranes. Total DNA was then extracted from each filter, pooled, and subjected to 16S rRNA gene and shotgun sequencing on the Illumina HiSeq platform. Additionally, on eight occasions water filtrates were processed for viral metagenomes (viromes) using chemical concentration and then shotgun sequenced. Ubiquitous freshwater phylum *Protebacteria* (e.g. *Variovorax*) and *Actinobacteria* (e.g. *Streptomyces*)

were abundant at all sampling dates throughout the year. However, environmental characteristics appeared to drive the structure of the community. For instance, the abundance of *Cyanobacteria* (e.g. *Nostoc*) increased with rising water temperatures, while a storm event appeared to trigger an increase in overall bacterial diversity, as well as the abundance of *Bacteroidetes*. This event was also associated with an increase in diversity of antibiotic resistance genes. The viral fractions were dominated by dsDNA of the order *Caudovirales*, namely *Siphoviridae* and *Myovirdae*. Moreover, phylogenetic analysis of the viral *polA* marker-gene revealed a diversity of putative lysogenic phage. Overall, this study provides foundational data on the temporal variability of bacterial and viral communities in an agricultural pond, a site underrepresented in freshwater studies.

4.2 Introduction

Ponds are small (1 m^2 to ~50,000 m^2), shallow, standing water bodies that are found ubiquitously among Earth's terrestrial biomes, with an estimated 2.6 to 9 million within the U.S. alone [22, 281]. Globally, ponds occupy a greater total area than lakes and are considered to be functionally and ecologically distinct, playing a major role in global cycling and supporting a high level of macro- and micro-species diversity [22, 29, 157, 233, 281, 282]. Along with those that are formed by natural processes, there are many ponds that are human constructed for a variety of recreational, industrial, agricultural, and aesthetic purposes [22, 283]. For instance, in areas where municipal and ground water sources are limited or unavailable, ponds are built to capture and store water for irrigation [23, 284]. Despite the importance of ponds to both environmental and human activities, the majority of research on freshwater resources is focused on large water systems (e.g. lakes). As a result, outside of extreme environments (e.g. saline/hypersaline [120–122], thermokarst [123]) and aquaculture facilities [124, 125], ponds remain understudied [285], especially with regard to their microbial communities.

Microbial communities are vital to the health and maintenance of aquatic ecosystems [42]. However, in some cases, they can cause severe environmental and public health problems. Ponds, in particular, are uniquely susceptible to microbial disruptions due to their small size and shallow depth [286]. Nonpoint source nutrient pollution, coupled with warm temperatures, and long water residence times can result in a high abundance of algal and cyanobacterial concentrations, in some cases leading to blooms that deplete oxygen levels and produce toxins [287–290]. Storm events can also trigger the influx of fecal pathogens that can contaminate irrigation supplies and subsequently crops [291–293]. For instance, a 2002 multistate outbreak of *Salmonella* Newport on tomatoes was traced back to contaminated pond water used for irrigation [66]. In addition to pathogens, runoff can introduce pollutants originating from land use practices (e.g. antibiotics, pesticides) [294]. Because of the long water retention times of ponds, these pollutants may then diffuse and accumulate, potential leading to changes in bacterial community dynamics, including increased selection pressures for antibiotic-resistant populations [72]. However, the persistence of these disruptions and foreign bacterial agents depends on complex factors such as sedimentation, temperature, UV light, and predation [295].

Despite the value in surveying the microbial composition in ponds, the limited collection of previous studies have been largely restricted to PCR or culture-based methodologies and often comprise just a static "snap shot" of the microbial community. Thus, we are restricted in our understanding of microbial functionality, dynamics, and response under multiple conditions. Shotgun metagenomics, however, makes it possible to observe and analyze a broad sampling of microbial diversity without cultivation, providing new insights into their genomic complexity and functional potential [296]. In addition, because shotgun metagenomics does not rely on a universally distributed marker gene, such as 16S rRNA, it can also be used to explore the viral community; a component of the microbial world often left unexplored [97].

Bacteriophage (phage), viruses that infect bacteria, are the most abundant biological entities in aquatic systems and are critical in shaping the evolution, diversity, abundance, and genetic composition of bacteria [143]. Temperate phage (forming prophage) can influence their host's phenotype through the horizontal transfer of genes, such as those for antibiotic resistance/toxins and those that promote host fitness and adaptability [53, 56]. However, phage composition, diversity, and host-interactions are often linked to fluctuating environmental characteristics [297]. Therefore, assessing phage ecology and relationships with their host(s) is critical with regard to completing a comprehensive characterization of pond biodiversity.

In the present study, we periodically sampled surface water from a freshwater agricultural pond located in the Mid Atlantic, United States. From these samples, we employed culture-independent high-throughput sequencing to characterize the dynamics, taxonomy, functionality, and interaction history of their bacterial and viral communities over time.

4.3 Materials and Methods

4.3.1 Study site and sample collection

Water samples (total n=14) were collected on the following dates: 6/12/17, 7/17/17, 8/8/17, 8/21/17, 9/11/17, 9/25/17, 10/30/17, 11/13/17, 12/18/17, 1/22/18, 2/12/18, 3/12/18, 4/9/18, and 5/7/18 from a temperate freshwater agricultural pond located in a rural area of central Maryland, United States (maximum depth of ~3.35 meters and a surface area of ~0.26 ha). At each date, a utility transfer pump (0.08 W; Everbilt, Atlanta, GA) powered by a EU1000i generator (American Honda Motor Co., Ltd., Alpharetta, GA) and connected to a sampling cartridge via vinyl braided tubing (1.9 cm inner diameter, Sioux Chief, Peculiar, MO) was submerged 15-30 cm below the surface and used to pump roughly 10 L of water into a sterile polypropylene carboy. Samples were kept in the dark at 4 °C and processed within 24 hr of collection.

4.3.2 Water physicochemical assessment

In addition, at each time point a ProDSS digital sampling system (YSI, Yellow Springs, OH, United States) was used to measure the following physicochemical properties of the pond water: temperature (°C), pH, dissolved oxygen (% DO), conductivity (SPC uS/cm), oxidation-reduction potential (ORP, mv), turbidity (FNU), nitrate (mg/L), and chloride (mg/L). Using the Nation Weather Services historical data archive, ambient temperature was recorded for the time and date at each sampling event.

4.3.3 Water sample processing

Microbial DNA was isolated as described in detail previously [157]. Briefly, for each sample 10 L of water was filtered sequentially through a Whatman 1 μm polycarbonate filter (Sigma-Aldrich, MO, United States) and a 142-mm diameter 0.2 μm membrane filter (Pall Gelman Sciences, MI, United States) attached via sterile 1.6 mm PVC tubing with a Watson Marlow 323 Series Peristaltic Pump (Watson-Marlow, Falmouth, Cornwall, United Kingdom). Following filtration, filters (1 and 0.2 μm) containing the cellular fraction were dissected into four equal quadrants and stored at $-80\,^{\circ}$C until DNA extraction.

4.3.4 Viral concentration and DNA extraction

On 6/12/17, 7/17/17, 8/8/17, 8/21/17, 9/11/17, 9/25/17, 10/30/17, and 5/7/18, the iron chloride procedure was used on the pond water after 1 μm and 0.2 μm filtration. A 1 mL solution of $FeCl_3$ (4.83 g $FeCl_3$ into 100 ml H_2O) was added to the filtered pond water and incubated in the dark for 1 hr. The samples were then filtered onto 142-mm 1 μm polycarbonate filters (Sigma-Aldrich, MO, United States) to capture flocculated viral particles. Filters were stored at $4\,^{\circ}$C in the dark until resuspension. For resuspension, filters were rocked overnight at

4 °C in 10 mL of 0.1M EDTA-0.2M $MgCl_2$-0.2 M Ascorbate Buffer, described in detail elsewhere [240]. Resuspended viral particles were then subjected to a DNase I (Sigma-Aldrich, MO, United States) treatment for 1 hr and passed through a 33-mm diameter sterile syringe filter with a 0.2 μm pore size (Millipore Corporation, MA, United States). DNA was extracted from 500 μl of the viral concentrate using the AllPrep PowerViral DNA/RNA Kit (Qiagen, CA, United States) per the manufacturer's instructions. Prior to sequencing, viral DNA was tested for the presence of bacterial contamination via 16S rRNA PCR.

4.3.5 Microbial DNA extraction

Microbial DNA was extracted from the filters using an enzymatic and mechanical lysis procedure [158, 159]. Each filter quadrant was placed in a lysing matrix tube with a cocktail of PBS buffer, lysozyme, lysostaphin, and mutanolysin. After incubation at 37 °C for 30 min, a second lysing cocktail (Proteinase K and SDS) was added followed by another incubation at 55 °C for 45 min and mechanical lysis via bead beating with a FastPrep Instrument FP-24 (MP Biomedicals, CA) (6.0 m/s for 40s). The resulting DNA was purified with the QIAmp DNA mini kit (Qiagen, CA, USA) and assessed with the NanoDrop 2000 Spectrophotometer. To create a composite sample, microbial DNA extracts from all four quadrants of both filter types were pooled for each date.

4.3.6 16S rRNA sequencing and analysis

From each of the pooled microbial DNA extractions (n=14), the V3-V4 hypervariable region of the 16S rRNA gene was PCR-amplified and sequenced on the Illumina HiSeq (Illumina, San Diego, CA, United States) utilizing a dual-indexing strategy for multiplexed sequencing developed at the Institute for Genome Sciences [158, 159, 241].

The resulting 16S rRNA reads were screened for low quality bases and short read lengths, merged with PANDAseq, de-multiplexed, and trimmed of artificial barcodes and primers [241–243]. Using VSEARCH, reads were then checked for chimeras with the UCHIME algorithm and the ChimeraSlayer RDPGold_Trainset reference training dataset [298]. Chimera-free reads were then clustered (*de novo* into Operational Taxonomic Units (OTUs) using VSEARCH with a minimum confidence threshold of 0.97. Following OTU clustering, taxonomic assignments were performed using the Quantitative Insights Into Microbial Ecology software package (QIIME; release v. 1.9.1) using the Greengenes database [244]. The resulting OTU table, OTU reference sequences and phylogenetic tree files were imported to the R Statistical computing software (v. 3.4.3) using the Phyloseq R package (v. 1.22.3) [249].

Alpha diversity assessed via Observed OTUs was calculated using the R packages: Bioconductor [247], metagenomeSeq [245], vegan [248], phyloseq [249], fossil [250], biomformat [249], and ggplot2 [183] on unrarefied and data rarefied to an even sampling depth (13,956 sequences).

4.3.7 Shotgun sequencing for microbial metagenomes and viromes

For both the microbial (n=14) and viral (n=8) samples, DNA extracts were shotgun sequenced. Briefly, for each sample DNA was used in a tagmentation reaction, followed by 12 cycles of PCR amplication using Nextera i7 and i5 index primers per the modified Nextera XT protocol. The final libraries were then quantitated by Quant-iT hsDNA kit. The libraries were pooled based on their concentrations as determined by Quantstudio 5 and loaded onto an Agilent High Sensitivity D1000 ScreenTape System. Samples were then sequenced on an Illumina Hiseq X10 flow (Illumina, San Diego, CA, United States) cell targeting 100 bp paired end reads per sample.

4.3.8 Microbial metagenomic and virome assembly

The resulting paired-end reads from both microbial and viral libraries were quality trimmed using Trimmomatic ver. 0.36 (sliding window:4:30 min len:60) [175], merged with FLASh ver. 1.2.11 [176], and assembled *de novo* with MEGAHIT [299]. Open reading frames (ORFs) were predicted from the assembled contigs from each library using MetaGene [104].

4.3.9 Microbial and viral taxonomic and functional classification

For the microbial metagenomes, predicted peptide ORFs were searched against UniRef 100 (retrieved May 2018) using protein-protein BLAST (BLASTp ver. 2.6.0+) (E value $\leq$ 1e^{-3}) [105, 106]. Taxonomic classifications were then made to contigs by max cumulative bit score. This was calculated by summing the bit scores of all taxa with a hit to peptide ORFs encoded by the contig. Peptide ORFs were searched against the SEED databases using BLASTp (E value $\leq 1e^{-3}$). Peptide ORFs were assigned to a SEED subsystem with the maximum sum bit score among all of the ORF's hits. Taxonomic and functional classification of viromes were conducted as described in Chopyk et al., 2017 [157].

For both viromes and microbial metagenomes, coverage was calculated for each contig by recruiting quality-controlled reads to assembled contigs using Bowtie2 ver. 2.3.3 (very sensitive local mode) and then using the "depth" function of Samtools ver. 1.4.1 to compute the per-contig coverage [180]. To normalize abundances across libraries, contig and ORF coverages were divided by the sum of coverage per million, similar to the transcripts per million (TPM) metric used in RNA-Seq [181, 257]. Scripts performing these assignments and normalization are available at https://github.com/dnasko/baby_virome. All taxonomic and functional data were visualized using the R packages ggplot2 ver. 3.1.0 and pheatmap ver 1.0.10 [182,183].

4.3.10 ARGs prediction and host assignment

Peptide ORFs from both viromes and microbial metagenomes were searched against the "Comprehensive Antibiotic Resistance Database" (CARD; retrieved July 2018) using protein-protein BLAST (BLASTp ver. 2.6.0+) (E value $\leq 1e^{-3}$) [105, 107]. Hits to CARD proteins were considered valid only under the conservative criteria described in previous studies (>40% coverage and >80% amino-acid identity) [83, 185]. In addition, for the ARGs conferring resistance through target mutations, a post-processing step (MAFFT alignment with reference sequences available at CARD) was taken to confirm the presence of resistance-conferring mutations [186]. Taxonomic assignments were parsed for contigs containing ARG-like peptide ORFs. Networks were visualized by Cytoscape software [300].

4.3.11 Viral Pol I prediction and phylogenetic analysis

To extract the biological informative marker gene, polA, from the viromes, predicted peptide ORFs were queried against Pol I UniRef90 clusters using protein-protein BLAST (BLASTp ver. 2.6.0+) (E value $\leq$ 1e^{-5}) [105, 106]. Along with informative references, each sequence was annotate with its Phe762 position relative to *E. coli* IAI39, a position previously reported to be indicative of phage lifecycle: a Phe762 or Tyr762 defined as generally virulent, while a Leu762 defined as generally temperate [238,301]. The extracted sequences, which spanned the 762 position, were then aligned with MAFFT using the FFT-NS-i 1000 algorithm and trimmed to a region of interest (I547 to N923 in *E. coli* IAI39) with Geneious 11.1.5 [157,186,259]. The alignment was then used to create an unrooted maximum likelihood tree with PhyML [260]. A cladogram was produced, annotated and colored with FigTree version 1.4.2 [302].

4.3.12 CRISPRs prediction from microbial metagenomes

For each of the microbial metagenomes, CRISPR arrays were predicted from the assemblies via the CRISPR detection and validation tool, CASC [303]. Bonafide CRISPR spacers were clustered at 97% nucleotide similarity with CD-HIT-EST to determine the number of unique spacers at each time point [184]. Additionally, CRISPR spacers from each library were queried against the eight viromes via BLASTn (E value $\leq$ 1e^{-1}), word size seven.

4.3.13 Statistical analysis

Significance tests were conducted using a Tukey's HSD Test between meteorological seasons, defined by the American Meteorological Society [304]. Additionally, to identify associations between the water physicochemical characteristics and the normalized abundance of the bacterial genera, as well as between the abundance of bacterial genera and viral families, Pearson's correlation coefficients were calculated in RStudio version 1.0.153.

4.4 Results

4.4.1 Sequencing effort and assembly

All samples (n=22) were sequenced on the Illumina HiSeq, 14 microbial and eight viral. In total, there were 907,056,944 read pairs for the microbial metagenomes with an average of 64,789,782 read pairs per metagenome (± 7,936,115 Standard

Deviation, SD) Table 4.1. For the viral metagenomes, there were 489,222,408 read pairs with an average of 61,152,801 read pairs per metagenome (± 9,064,079 SD) Table 4.2. After assembly, there were a total of 9,979,705 contigs, with an average of 712,836 contigs per sample (±142,125 SD) for the microbial metagenomes and a total of 1,913,254 contigs, with an average of 239,157 contigs per sample (±45,658 SD) for the viromes.

4.4.2 Temporal variations in physicochemical characteristics and bacterial diversity

physicochemical variables for each sampling date are shown in Figure 4.1. Water temperature ranged from 29 °C (7/17/17) to 4 °C (1/22/18). By meteorological season, winter (12/18/17, 1/22/18, 2/12/18) had an average water temperature of 6 °C. This was significantly ($p \leq 0.05$) lower than autumn (9/11/17, 9/25/17, 10/30/17, 11/13/17) and summer (6/12/17, 7/17/17, 8/8/17, 8/21/17), which had an average water temperature of 18 °C and 27 °C, respectively. In addition, the water temperature in summer was significantly higher than spring (3/12/18, 4/9/18, 5/7/18). The only other environmental factor that was significantly different by meteorological season was ORP, which was significantly higher in spring compared to autumn. Precipitation 24-hr prior to sampling occurred only on 8/8/17, 10/30/17, and 2/12/18.

Furthermore, we examined the bacterial diversity at each time point by way of amplification and sequencing of the 16S rRNA gene. Overall, the diversity, surveyed by rarefied and unrarefied Observed OTUs, was generally steady throughout the year, with no significant differences found between rarefied, unrarefied diversity and meteorological season. However, by physicochemical parameter we did find some differences, specifically with the spike in diversity on 2/12/18 (Figure 4.1). This date corresponded to an increase in precipitation 24 hr prior to sampling, as well as turbidity. In fact, precipitation and turbidity were both significantly ($p \leq 0.05$) positively correlated with the abundance of rarefied and unrarefied Observed OTUs.

4.4.3 Temporal variations in bacterial phyla

For the microbial metagenomes collected throughout the year, on average 78% (± 4% SD) of contigs could be assigned a taxonomic representative (Table 4.3). Of these, the majority was homologous to Bacteria 93% (± 2% SD), followed by Eukaryota 3% (± 1% SD), and Viruses 3% (± 0.5% SD).

For each of the contigs, a normalized abundance was calculated to account for assembly proficiency and sequencing depth and parsed for those assigned as Bacteria. Of these, the most frequently observed bacterial phylum was *Proteobacteria*, which accounted for 43% (± 5%) of the total bacterial assigned abundance (Figure 4.2). The next most abundant phyla were *Actinobacteria* at 28% (± 8%), *Bacteroidetes* at 12% (± 4% SD), and *Firmicutes* at 7% (± 1%). The largest phylum, *Proteobacteria*, was composed chiefly of the class *Betaroteobacteria* 50% (± 6% SD) and *Alphaproteobacteria* 23% (± 5% SD), with the largest spike in *Alphaproteobacteria* occurring on 2/12/18. By meteorological season, winter had a significantly (p

≤ 0.05) higher abundance of *Bacteroidetes* than all other seasons, while summer had a significantly ($p \leq 0.05$) higher abundance of *Cyanobacteria* compared to all other seasons. Summer and autumn both had a high abundance of *Firmicutes*, with both significantly ($p \leq 0.05$) higher than spring and winter.

In addition to differences by meteorological season, the normalized abundance of some of these top phyla correlated with physicochemical parameters surveyed in the water: *Actinobacteria* (R= 0.65, $p \leq 0.01$) correlated with conductivity, *Bacteroidetes* correlated with precipitation (R= 0.63, $p \leq 0.05$), conductivity (R=-0.67, $p \leq 0.01$), and turbidity (R=0.74, $p \leq 0.01$), *Cyanobacteria* correlated with water temperature (R= 0.83, $p \leq 0.001$), *Firmicutes* correlated with water temperature (R=0.80, $p \leq 0.001$) and ORP (R=-0.64 , $p \leq 0.01$), *Planctomycetes* correlated with water temperature (R=0.74 $p \leq 0.01$) and conductivity (R=0.58, $p \leq 0.05$) and *Chloroflexi* correlated with precipitation (R=-0.68 , $p \leq 0.01$) and turbidity (R=-0.63 , $p \leq 0.05$).

4.4.4 Temporal variations in bacterial genera

Within the bacterial assignments classified at the genera level, *Streptomyces* (11% ±3% SD), *Variovorax* (7% ±2% SD), *Pusillimonas* (4% ±1% SD), and *Pseudomonas* (3% ±0.5% SD) were the most abundant (Figure 4.3). By metrological season, winter had a significantly ($p \leq 0.05$) higher abundance of *Ulvibacter*, *Rudanella*, and *Flavobacterium* compared to all seasons. Spring had a significantly ($p \leq 0.05$) higher abundance of *Polynucleobacter* compared to all seasons and a significantly higher abundance of *Nitrosovibrio* compared to autumn. Summer had a significantly ($p \leq 0.05$) higher abundance of *Nostoc* compared to all seasons. Autumn had a significantly ($p \leq 0.05$) higher abundance of *Ferrimicrobium* compared to winter.

Similar to the analysis of the bacterial phyla, we calculated Pearson's correlations between the normalized abundance of the dominant bacterial genera and the physicochemical parameters of the pond water (Figure 4.3). In total, precipitation and turbidity correlated with the greatest number of genera, followed by conductivity and water temperature.

4.4.5 Microbial functional potential

On average, 40% (± 3%) of peptide ORFs from the microbial metagenomes could be assigned at a SEED functional category (Figure 4.4). Of these, "Carbohydrate Metabolism" was the most abundant representing on average 16 % (± 1%) of the total assigned functional abundance followed by "Amino Acids and Derivatives" at 12% (± 0.3%), "Protein Metabolism" at 9% (± 0.4%), and either "Cofactors, Vitamins, Prosthetic Groups, Pigments" at 7% (±0.2 %) or "DNA Metabolism" at 6% (± 0.5%). By meteorological season, the only SEED functional category that was significantly ($p \leq 0.05$) different between seasons was "Motility and Chemotaxis", which was significantly higher in winter compared to autumn and summer. Similar to the bacterial abundance, precipitation was significantly correlated with the abundance of a diversity of functional SEED systems including: “Potassium metabolism” (R=0.78, $p \leq 0.001$), “Regulation and Cell signaling” (R=0.76, $p \leq 0.01$), “Iron acquisition and metabo physicochemicallism” (R=0.74, $p \leq 0.01$), “Virulence Disease and Defense” (R=0.72, $p \leq 0.01$), “Miscellaneous” (R=0.69, $p \leq 0.01$), “Phages Prophages Transposable elements etc. ” (R=-0.69, $p \leq 0.05$), “Carbohydrates” (R=0.68, $p \leq 0.05$), “Membrane Transport” (R=0.67, $p \leq 0.01$), “Sulfur Metabolism” (R=0.61, $p \leq 0.05$), and “Nitrogen Metabolism” (R=0.58, $p \leq 0.05$). Likewise, turbidity was also correlated with “Iron acquisition and metabolism” (R=0.76, $p \leq 0.01$), “Membrane Transport” (R=0.69, $p \leq 0.01$), “Regulation and Cell signaling” (R=0.66, $p \leq 0.01$), and “Potassium metabolism” (R=0.56, $p \leq 0.05$), and “Virulence” (R=0.55, $p \leq 0.05$)

Other physicochemical factors that had significant correlations with the normalized abundance of the SEED function systems included the following: water temperature with “Photosynthesis” (R=0.63, $p \leq 0.05$), conductivity with “Iron acquisition and metabolism” (R=-0.53, $p \leq 0.05$), pH with “Dormancy and Sporulation” (R=0.65, $p \leq 0.01$), and ORP with “Virulence” (R=0.67, $p \leq 0.01$).

4.4.6 Antibiotic resistance and host taxonomy

To assess antibiotic resistance in the microbial and viral metagenomes, we conducted a BLAST analysis of peptide ORFs against CARD. No peptide ORFs within the viral metagenomes had significantly homology to ARGs within CARD. However, in the microbial metagenomes, 184 peptide ORFs were identified as 21 unique ARGs conferring resistance to over 15 drug classes (Figure 4.5). For the ARGs whose resistance is associated with target mutations, they were confirmed to carry the following mutations: *rpsL*, K88R [192]; *gyrA*, S95T [190]; *murA* C117D [191]; *rpoB* H526T [305]; *EF-Tu* Q124K [195]; *ndh* V300G, V246A [306]. A normalized abundance was also calculated for each ARG-like peptide ORF. From this, the greatest abundance of ARG-like peptide ORFs was attributed to the sampled collected on 10/30/17, followed by 9/25/17 and 9/11/17. However, the greatest diversity of ARGs was identified on 2/12/18, with 11 unique ARGs.

For each ARG-like peptide ORF, the source genera and phyla were parsed (Figure 4.6). All the ARG-like peptide ORFs originated from contigs assigned as Bacteria. Of these, 71% of were contigs assigned to the phylum *Actinobacteria* (9 unique ARGs), largely of the genus *Ferrimicrobium* (30 *rpsL*), *Saccharomonospora* (1 *RbpA*, 4 *gyrA*, 12 *mtrA*, 4 *murA*, 2 *rpsL*), and *Aeromicrobium* (5 *EF-Tu*, 1 *rpoB*, 13 *rpsL*). The next largest phylum assigned to contigs with an ARG-like peptide ORF was *Proteobacteria*, which accounted for 21% of the contigs, but had a wide diversity of ARGs (14 unique ARGs). Within this phylum, *Sphingopyxis* (1 *ESP-1*, 1 *PEDO-2*, 12 *rpsL*) and *Pseudomonas* (3 *rpsL*, 1 *CpxR*, 1 *mtrA*) were assigned to the most contigs.

4.4.7 Viral taxonomic composition

For the viromes, on average 47% of contigs (± 1%) contigs could be assigned a taxa, which is in agreement with results described in other viral metagenomic studies [157,261]. For those that could be assigned, a normalized abundance was calculated.

The vast majority of viral abundance was assigned to the tailed bacteriophage of the order *Caudovirales* (Figure 4.7). Of these, the majority were similar to members of the *Siphoviridae* (49% ± 4%) family, followed by the *Myoviridae* (34% ± 5%) and *Podoviridae* families (14% ± 2%). The remaining proportion were either viral contigs that could not be assigned a family (2% ± 0.1%) or were other viral families (1% ± 0.2%). The other viral families included viruses infecting other bacteria and archaea, ssDNA bacteriophage *Microviridae* and *Inoviridae*, plant viruses from the family *Tymoviridae*, and animal/arthropod viruses from the family *Poxviridae*.

4.4.8 Viral Pol I phylogeny

The Pol I peptide was used as a marker gene to analyze the taxonomic affiliations and potential life cycle of phage in pond water (Figure 4.8). In total, we identified 3,749 Pol I peptides that spanned the 762 amino acid position. Of these, 55% had leucine at the 762 position (Leu762), 29% had the wildtype phenylalanine (Phe762), and 16% had a tyrosine (762Tyr). A subsection of these peptides (Phe762: 230, Leu762: 512, Tyr762: 164) could then be used to build a phylogenetic tree. Similar to previous studies, Pol I peptides claded largely by their 762 positions [157, 301].

4.4.9 Phage-host relationships

To determine the interaction between phage and potential host(s) we calculated any correlation between the abundance of the dominant viral families and the dominant bacterial genera and phyla. At the phylum level, the normalized abundance of *Siphoviridae* correlated with the normalized abundance of *Proteobacteria* (R=0.74, $p \leq 0.05$) and *Podooviridae* normalized abundance correlated with the normalized abundance of *Firmicutes* (R=-0.71, $p \leq 0.01$). At the genera level, *Siphoviridae* correlated with *Sphingomonas* (R=0.85, $p \leq 0.05$), *Sphingopyxis* (R=0.82, $p \leq 0.01$), *Nitrosovibrio* (R=0.80, $p \leq 0.05$), *Achromobacter* (R=0.77, $p \leq 0.05$), *Pusillimonas* (R=0.77, $p \leq 0.05$), and *Polynucleobacter* (R=0.75, $p \leq 0.05$). *Podoviridae* correlated with *Polynucleobacter* (R=0.91, $p \leq 0.01$), *Nitrosovibrio* (R=0.73, $p \leq 0.05$), and *Ferrimicrobium* (R=-0.71, $p \leq 0.05$). *Myoviridae* correlated only with *Microvirga* (R=0.80, $p \leq 0.05$)

Additionally, we predicted CRISPR arrays from the microbial metagenomes. Because CRISPRs arrays contain short segments of cleaved viral DNA (termed spacers) they can be used to provide a record of past infections [307,308]. In total, there were 319 CRISPR arrays detected in the metagenomes, with 1,041 unique spacers (Figure 4.9). To assess the relationship between microbial species and viromes, the unique spacers within each time point were queried against contigs from the eight viromes (Figure 4.9). Overall, 26% of spacers had a significant hit to 1,161 contigs from the eight viromes (10/30/17: 165, 5/7/18: 147, 6/12/18: 136, 7/17/18: 81, 8/21/18: 202, 8/8/18: 210, 9/11/18: 105, and 9/25/18: 115). These contigs were largely unknown, with only 24% assigned a taxonomy (31%: *Siphoviridae*, 35%: *Myoviridae*, and 5%: *Podoviridae*). For the microbial metagenomes, 6/12/17 and 5/7/18 had the greatest portion of spacers that had hits to any of the eight viromes, 42% and 27%, respectively (Figure 4.9). For both of these dates, the greatest number of hits was to the viromes sampled on the same date.

4.5 Discussion

Freshwater is a finite natural resource essential to life on Earth. It is critical in supporting urban, agricultural, and industrial activities, as well as providing a home for a rich diversity of macro- and micro- organisms [21, 22, 29, 157, 282]. Yet, anthropogenic activities, climate change, and a growing global population threaten its quality and availability worldwide [309, 310]. Here, we focused our attention on one freshwater resource that has been historically disregarded in favor of studies on larger aquatic systems, ponds.

In this study, the pond freshwater was dominated by *Proteobacteria*, largely that of *Betaproteobacteria*, a class found ubiquitously in freshwater [119]. However, there were seasonal changes in the abundance of the bacterial phyla that corresponded to environmental conditions. For instance, during the summer months, the abundance of *Cyanobacteria*, as well as the abundance of genes designated for photosynthesis, increased with increasing ambient water temperature (Figure 4.2 and 4.4). This is not surprising as water temperature has been found in multiple prior studies to be a predictor of the abundance of *Cyanobacteria* [311, 312]. Moreover, these results agree with those reported in an earlier study from our lab, where we found, through 16S rRNA sequencing, the abundance of *Cyanobacteria*, *Synechococcus*, decreased significantly with declining temperature [157]. In this study, *Cyanobacteria* peaked in the summer season (specifically on 7/17/17), but continued at high abundance into autumn, where mild temperatures likely sustained their growth. During these peak seasons, the genus *Nostoc* was the most abundant within the *Cyanobacteria* phylum (Figure 4.3).

The *Nostoc* genus includes a highly diverse range of nitrogen-fixing species, commonly found in aquatic environments either free-living, engaged in cooperative growth on plants and fungi, or in gelatinous colonies on rocks and stones [313]. While *Nostoc* blooms in freshwater ponds and lakes are often just considered a nuisance, they may also carry concern for use recreationally or for agricultural irrigation [314, 315]. *Nostoc* spp. are becoming increasingly recognized for their role in the production of cyanotoxins, as well as other bioactive compounds that can cause serious health problems in humans and animals [314, 315]. In fact, *Nostoc* is reported by the EPA as one of the eight most common *Cyanobacteria* for the production of microcystin [316]. In humans, microcystin exposure is associated with both acute health effects (e.g. abdominal pain, headache, diarrhea, pneumonia, etc.), as well as chronic conditions (e.g. primary liver cancer, colon and rectum carcinomas) [317, 318]. While we do not know from the data presented in this study if the *Nostoc* spp. are toxin-producing, their persistence in the summer months is cause for future investigation to protect environmental and public health.

In addition to fluctuations driven by seasonal trends, we saw a large shift in bacterial composition that correlated with a sizable precipitation event on 2/12/18. Likely, this event triggered an influx of upland runoff into the pond, resulting in an increase in bacterial diversity, as well as an increase in the abundance of *Bacteroidetes* (e.g. *Rudanella*, *Flavobacterium*) and *Proteobacteria* (e.g. *Alphaprotebac-*

teria) (Figure 4.1, 4.2). *Bacteroidetes* are often limited in freshwater environments, likely due to their dependency on organic matter [119, 319]. However, previous studies have found *Bacteroidetes* increased in abundance within freshwater creeks following storm events [320, 321]. In these studies, the authors suggested that the increase in *Bacteroidetes* may be a concern, as they are often indicative of human fecal and sewage material contamination [322, 323]. In fact, they have been suggested as better alternatives to traditional fecal indicators such as *E. coli* or fecal coliforms [322–324]. Along with potential pathogens and a diversity of terrestrial microorganisms, runoff can also introduce upland pollutants, such as antibiotics.

While antibiotics and ARGs are both naturally occurring, nonpoint and point source pollution of human and animal-derived wastes may select for an abundance that is atypical and may ultimately have repercussions for environmental and public health [213, 325]. Freshwater environments have become established as important reservoirs for the potential maintenance and dissemination of ARGs, especially small lakes and pond [326]. These lentic bodies tend to have longer water retention times compared to lotic environments, which can result in the accumulation of antibiotics and selection for resistant bacteria [327, 328]. In this study, we identified ARGs at all the sampling dates conferring resistance through a wide range of mechanisms across clinical, veterinary, and agricultural antibiotics. This varied resistome may be attributed to the selective forces driven by the pond topography, environmental contributions, and the commensal bacterial community composition. Unlike other surface freshwater sites, the pond surveyed here was dominated by *Streptomyces* of the phylum *Actinobacteria*. *Actinobacteria*, particularly *Streptomyces*, produce many clinically-significant antibiotics [213,329,330]. As a result, they can contain a wide array of ARGs for self-protection, as well as those inherited horizontally from other *Actinobacteria* [331,332]. Thus, it was not surprising to the see the majority or ARG putative hosts were *Actinobacteria* (Figure 4.6).

As for the environmental contributions, the largest spike in ARG diversity was on 2/12/18, which corresponded to a large precipitation event. Here, we saw the emergence of seven unique ARGs (*JOHN-1*, *ESP-1*, *CRP*, *PEDO-2*, *CPS-1*, *CpxR*, and *bacA*) conferring resistance to a broad range of clinically relevant antibiotics, including three beta-lactamases against carbapenem. The majority of these ARGs, unlike in the other months, were identified on contigs assigned as *Gammaproteobacteria*. This is consistent with the idea that these ARGs were introduced by an influx of upland runoff, as *Gammaproteobacteria* are not common in freshwater and are thought to be transient members introduced from the surrounding environment [119].

While we did not observe any ARGs in the viral fraction, we did identify an abundance of *Siphoviridae*, a family of largely temperate dsDNA bacteriophage previously found to dominate in this agricultural pond [157]. This is in line with the results reported in the Pol I peptide phylogenetic analysis (Figure 4.8), where we saw a high diversity of Pol I peptides with leucine at the 762 amino acid position. In a previous metanalysis of phage genomes, Leu762 mutation occurred primarily in temperate phage [157,238]. This was suggested, from previous biochemical analyses, to be due to the mutation producing a slower, but more accurate polymerase that would be advantageous to a lysogenic lifestyle [333]. Lysogeny is not uncommon to freshwater environments, especially in freshwater sediments [334, 335].

For phage, lyosgeny is suggested to be advantageous when conditions are poor, such as during times of nutrient-starvation [336]. Whereas, the lytic lifestyle is suggested to dominate when the bacterial community is the most productive (e.g. summer) [337]. While we do not have viral data that spans the coldest months of the year, the abundance of *Siphoviridae* did decrease and the abundance of *Myoviridae*, a traditionally virulent phage family, did increase during the warmer months (7/17/17-9/11/17) surveyed. However, the dominance of phage lifestyle strategy may be more complex then previously thought, as not all studies find lysogeny to be prevalent only in times of low bacterial productivity [338]. For instance, the "piggy-back-the-winner" model was born by observations that show lysogeny is more prevalent at higher host cell densities [339, 340]. In this study, the prevalence of lysogenic phage may also be due to the composition of the host taxa, as the dominant bacterial phylum of the pond, *Actinobacteria*, and *Proteobacteria*, have been previously reported in some environments to be ideal hosts for temperate phage [341].

The phage-host relationship is often left unexplored in microbial ecology studies, largely because it is difficult to link a phage with its host(s) due to heterogeneity of phage host range and culture limitations [342, 343]. Here, we utilized the phage regions (termed spacers) within the microbial CRISPR-Cas system to investigate bacteria-phage interactions in this freshwater system [307, 308]. CRISPR arrays were detected in all the microbial metagenomes, suggesting both the widespread use of this defense system and previous infections with sympatric phage species (Figure 4.9). For the majority of dates, spacers had more hits to the viromes collected on the same date or neighboring months compared to dates further away in time. This agrees with the notion that CRISPR spacers are dynamically added into arrays, matching coexisting phage species [344,345]. In addition to their significance as a freshwater resource for human industrial and agricultural activities, ponds are also a "hot spot" of biodiversity that significantly contributes to global ecosystem health [27]. Here, we provide one of the largest datasets on pond water microbial ecology to date. We expect these data will serve to not only improve understanding of the factors that may contribute to the disruption of pond biodiversity, but also further our knowledge regarding the potential microbial risks of using pond water for agricultural irrigation.

4.6 Figures

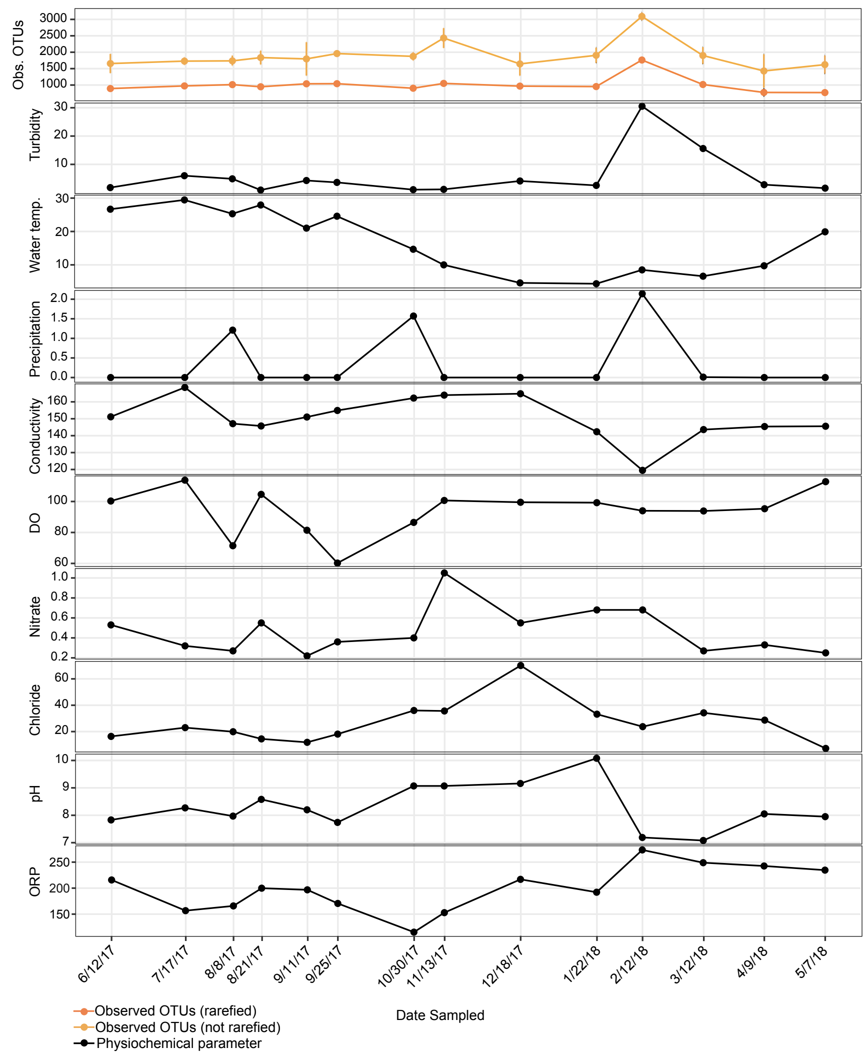

Figure 4.1: Temporal dynamics of physicochemical properties and bacterial diversity in agricultural pond water. Line graph displaying the alpha diversity (Observed OTUs, rarefied; dark orange, raw; light orange) and physicochemical properties (black) through time. The following physicochemical properties were surveyed: temperature (°C), pH, dissolved oxygen (% DO), conductivity (SPC uS/cm), oxidation-reduction potential (mv), turbidity (FNU), nitrate (mg/L), and chloride (mg/L). Sampling dates ordered temporally.

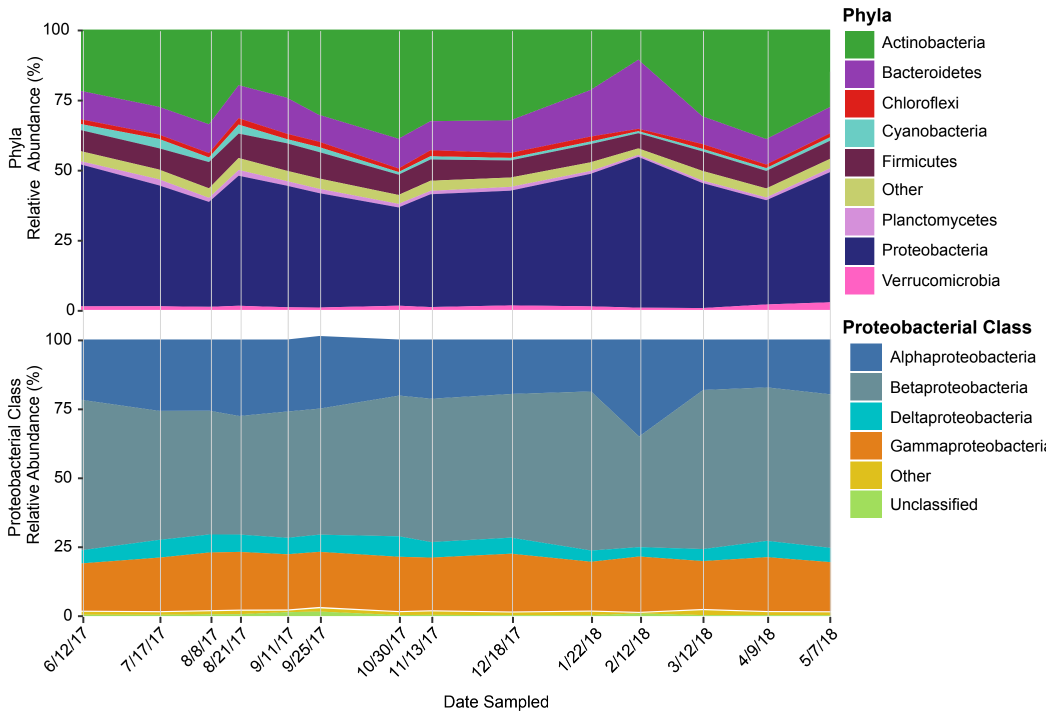

Figure 4.2: Temporal dynamics of bacterial composition in agricultural pond water. Stacked area chart depicting the normalized relative abundance of the bacterial communities at the (A) phylum level, as well as with the (B) *Proteobacteria* phyla split into classes. Sampling dates ordered temporally.

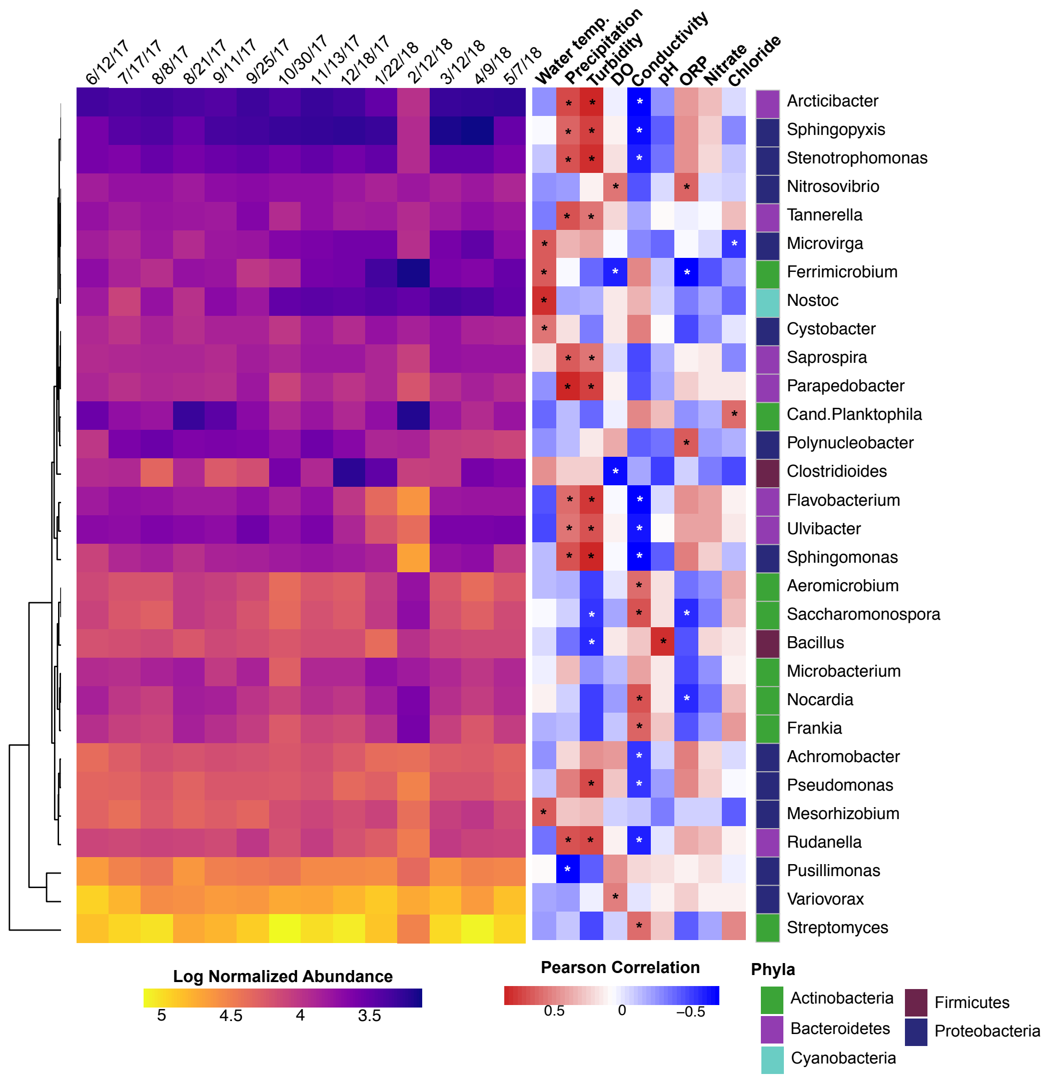

Figure 4.3: Bacterial genera abundance and correlations with physicochemical factors in agricultural pond water across sampling dates. Heatmaps based on the log-transformed normalized abundance of the most dominant genera (>1% in at least one sample) and the Pearson's correlation coefficients between the water physicochemical factors and the bacterial genera normalized abundance listed on the Y-axis. Genera annotated with colors representative of their phylum (*Proteobacteria*: dark blue, *Actinobacteria*: green, *Firmicutes*: burgundy, *Bacteriodetes*: purple, *Cyanobacteria*: light blue). Hierarchical clustering of samples was performed using the complete clustering method with Euclidean distances on the bacterial abundances. Asterix denote significant correlations ($p \leq 0.05$).

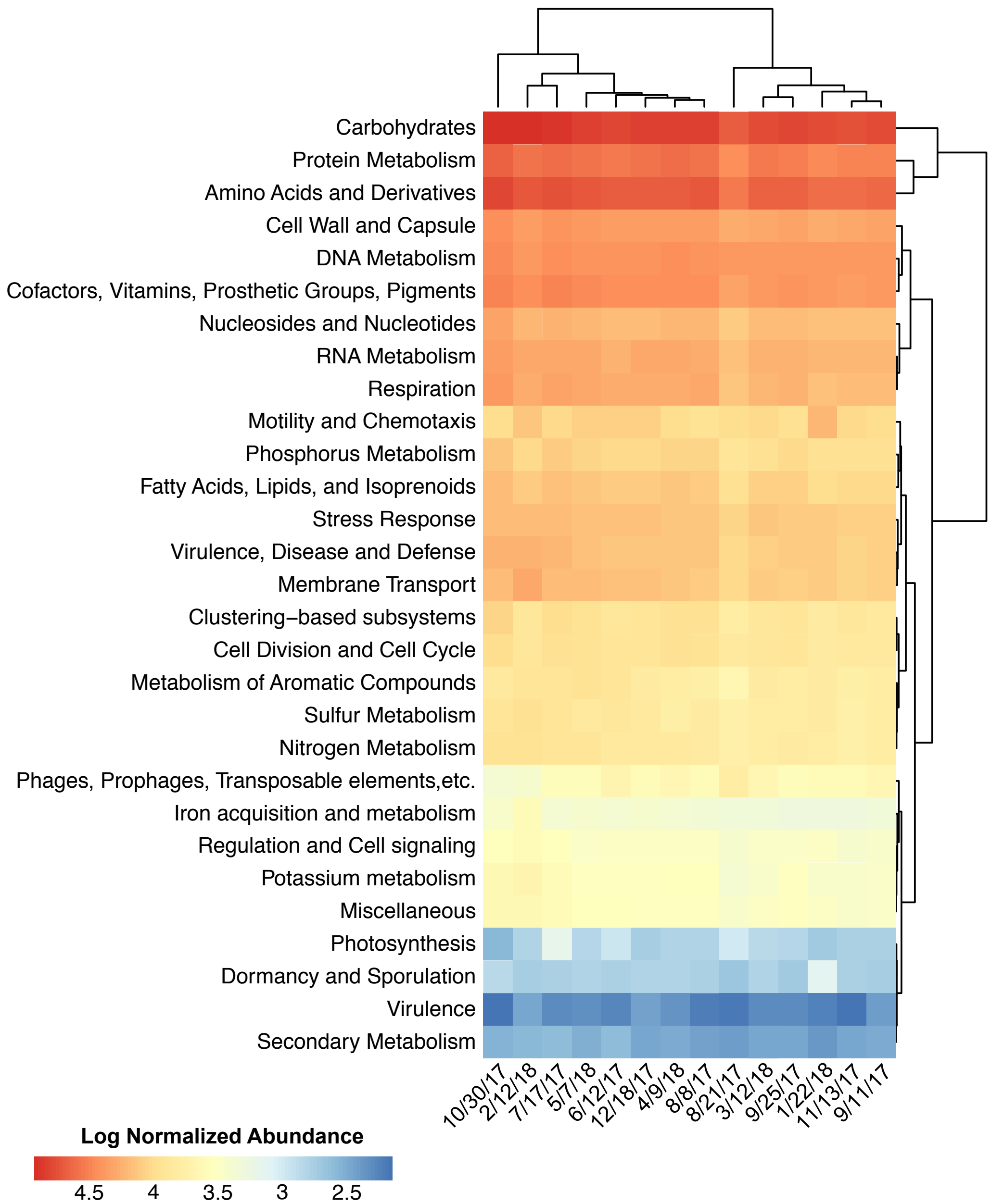

Figure 4.4: Functional composition in agricultural pond water across sampling dates. Heatmap of the normalized abundance assigned to the SEED systems at each sampling date for the microbial metagenomes. Hierarchical clustering of samples was performed using the complete clustering method with Euclidean distances.

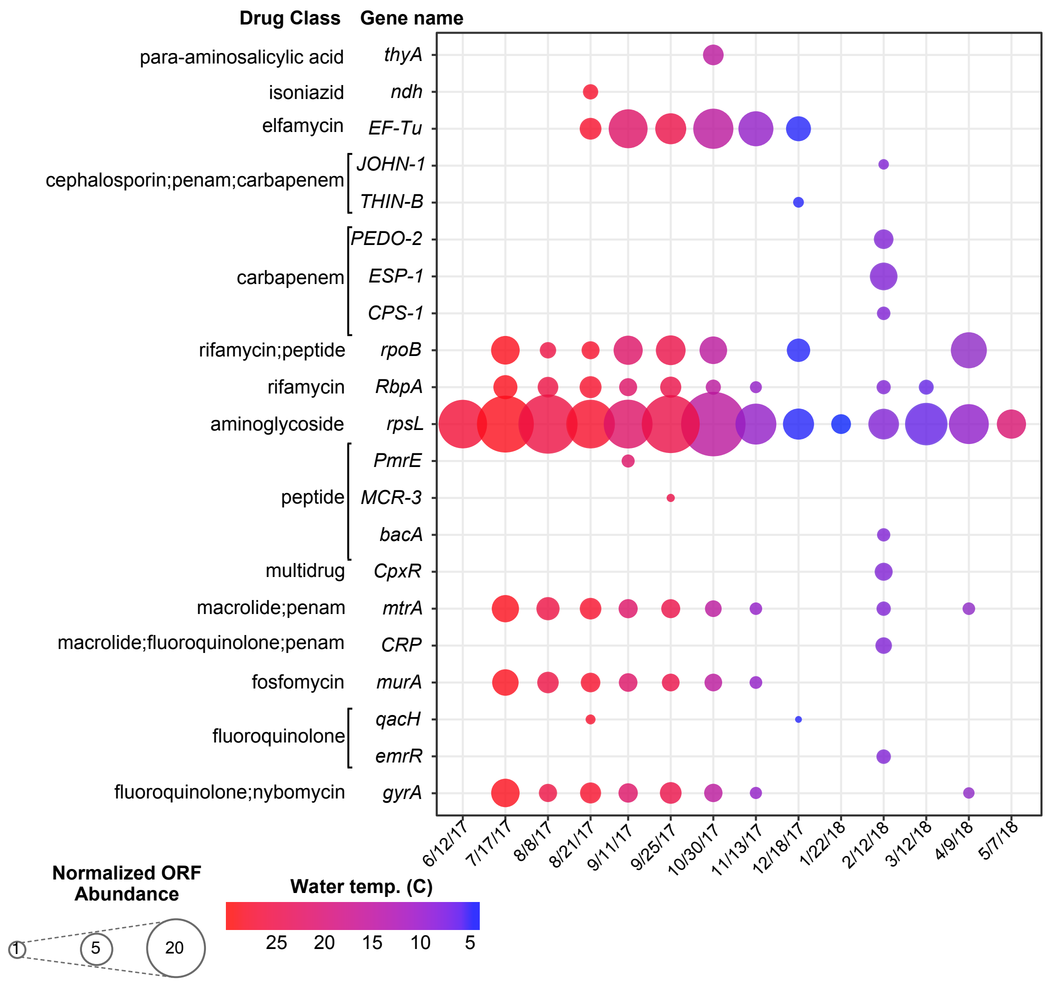

Figure 4.5: Antibiotic resistance genes (ARGs) in agricultural pond water across sampling dates. Dotplot of the ARG-like peptide ORFs predicted from the microbial metagenomes at each sampling date. The size of each dot is equivalent to the normalized peptide ORF abundance with homology to each ARG listed on the y-axis, and the color representative of the temperature of the water at the time of sampling.

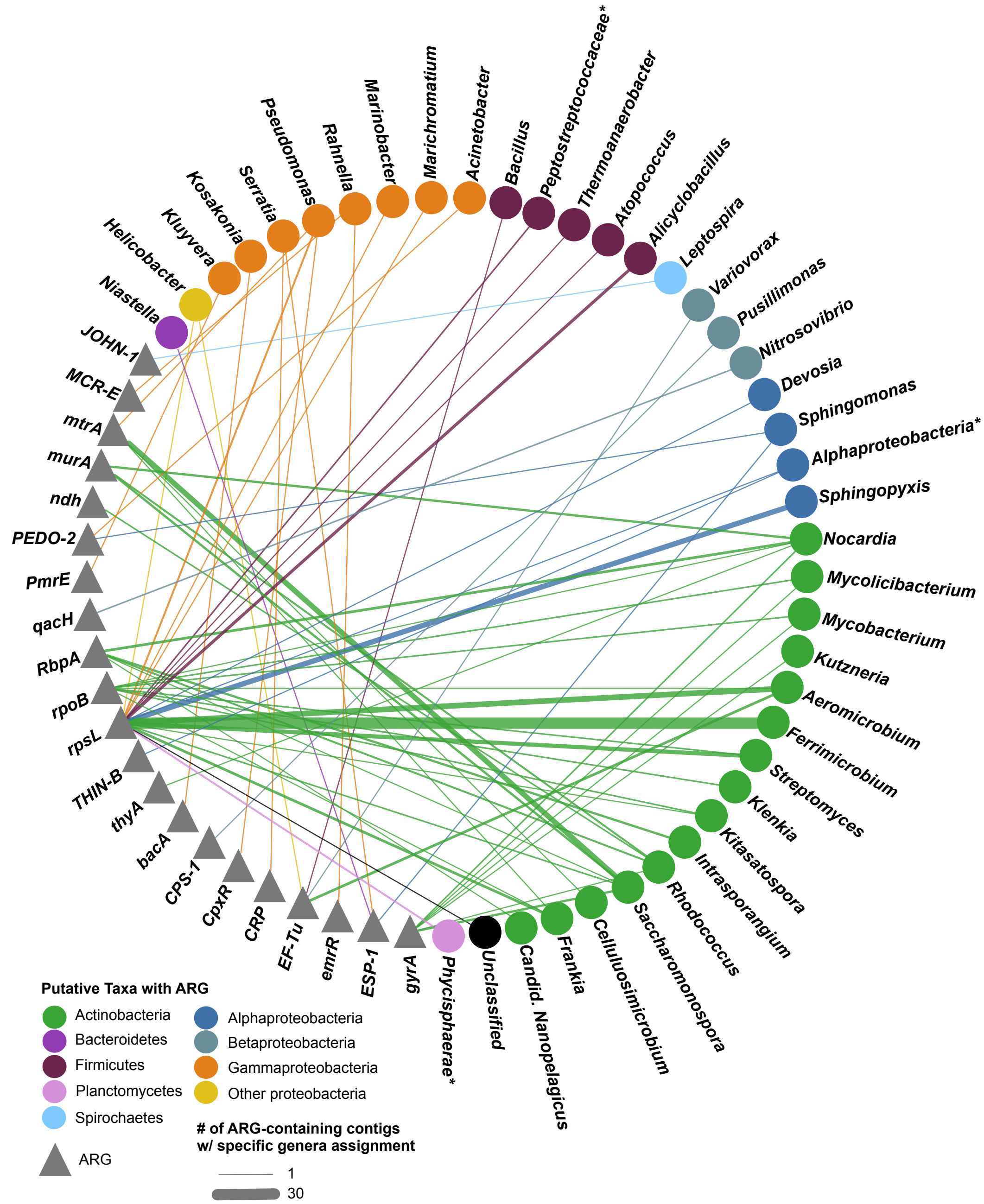

Figure 4.6: ARG host network. Bipartite network of the bacterial taxa with predicted antibiotic resistance genes (ARGs). Grey triangles represent ARGs connected by an edge to its putative bacterial host, with each edge colored by the host phyla. Bacterial host defined as the taxa assigned to the contig the ARG-like peptide ORF originated from and colored accordingly. Asterix represent taxa that could not be assigned at the genera level.

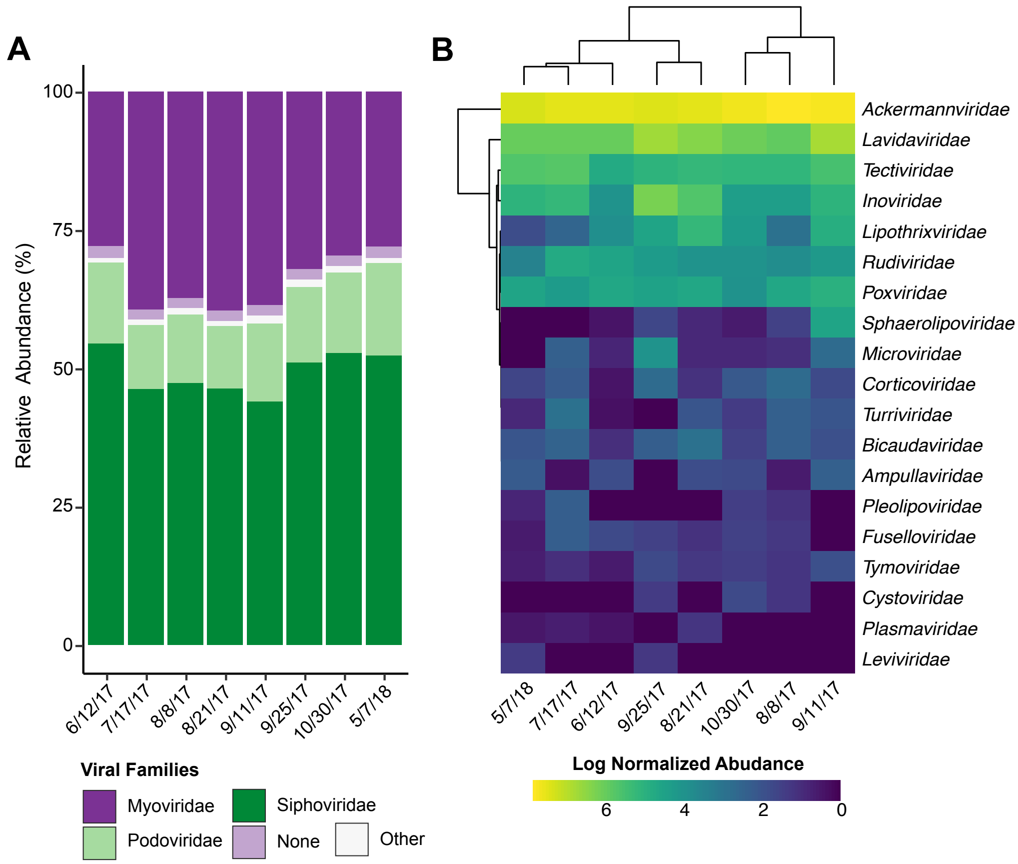

Figure 4.7: Viral composition in agricultural pond water across sampling dates. (A) Stacked bar charts depicting the normalized relative abundance of the viral communities at the family level; (B) Heatmap based on the (log+1)-transformed normalized abundance of “other” viral families. Hierarchical clustering of samples was performed using the complete clustering method with Euclidean distances.

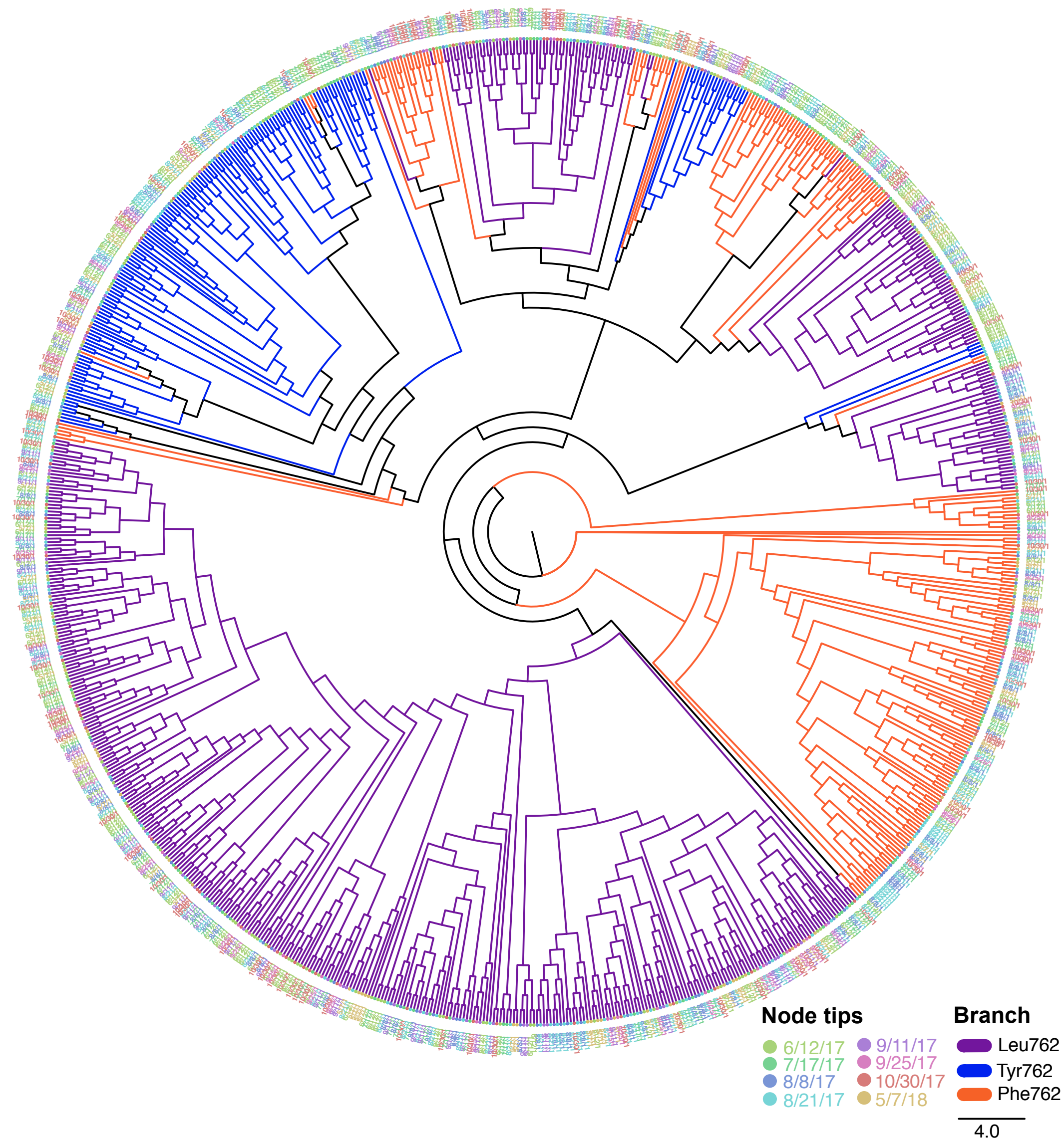

Figure 4.8: Cladogram of Pol I peptides across sampling dates. Unrooted maximum likelihood tree of representative Pol I peptide sequences predicted from each sampling date. Branches colored by 762 position residues, Phe762 (orange), Tyr762 (blue), and Leu762 (purple). Node tip labels indicate the date samples were collected and are colored accordingly.

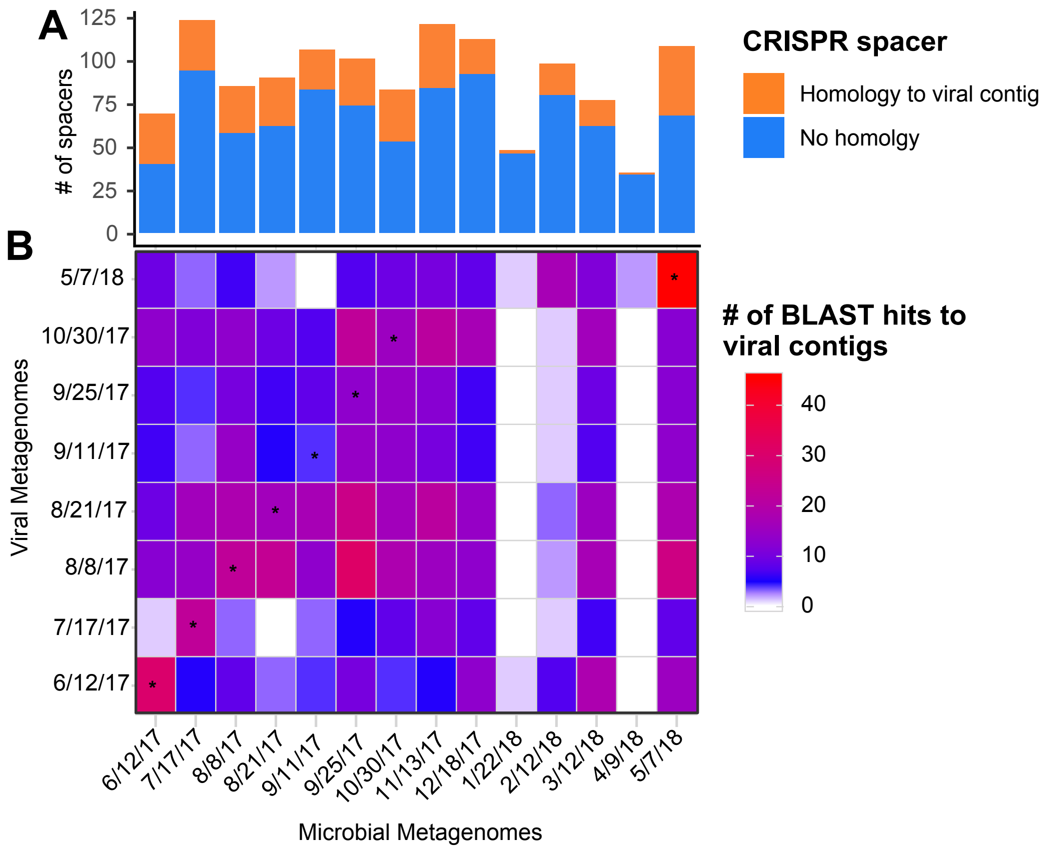

Figure 4.9: Detection of CRISPR spacers and linkage to viral contigs in agricultural pond water. (A) Stacked bar chart of the number of unique CRISPR spacers predicted from the microbial metagenomes at each sampling date. Bars are colored by whether the spacers had a significant BLAST hit to any contigs within the eight viromes. Spacer counts are reported after clustering at 97% with CD-HIT-EST; (B) Heatmap depicting the number of hits spacers had to contigs from each virome by date. Asterix denotes viromes and microbial metagenomes that were sampled on the same date.

4.7 Tables

Table 4.1: Descriptive sequencing statistics for microbial metagenomes.

Date	no. Read Pairs	no. Contigs	Mean Contig Size	Median Contig Size	Max Contig Size	% GC
6/12/17	56253234	710332	850	487	277707	49
7/17/17	63779336	825414	807	477	117814	51
8/8/17	64585947	785337	763	461	205041	49
8/21/17	69809556	805980	843	486	322391	49
9/11/17	64832833	815793	794	464	164272	49
9/25/17	63781249	840001	777	466	227816	50
10/30/17	59929487	665594	866	490	356886	48
11/13/17	71006872	821949	831	476	500323	48
12/18/17	69728632	759516	820	470	266589	48
1/22/18	85726602	317644	681	441	133727	48
2/12/18	56732600	762015	576	404	155163	49
3/12/18	54167570	541845	799	453	456402	47
4/9/18	65786989	651522	797	450	442516	46
5/7/18	60936037	676763	868	473	443357	48

Table 4.2: Descriptive sequencing statistics for viromes.

Date	**no. Read Pairs**	**no. Contigs**	**Mean Contig Size**	**Median Contig Size**	**Max Contig Size**	**% GC**
6/12/17	64110963	276729	865	470	180347	44.67
7/17/17	62502798	273781	667	438	169626	46.72
8/8/17	81168407	258993	729	452	226706	45.62
8/21/17	56676868	254571	873	472	304756	44.08
9/11/17	58609455	194360	704	451	163936	45.27
9/25/17	56219160	153900	670	442	133962	46.72
10/30/17	50692882	280625	715	454	89836	46.39
5/7/18	59241875	220295	786	445	191403	45.58

Table 4.3: Contig taxonomic assignments for microbial metagenomes.

Date	no. Contigs Assigned	Bacteria	Archaea	Eukaryota	Virus
6/12/17	595070	558560	4446	10428	17018
7/17/17	682721	643284	5246	15071	13909
8/8/17	613967	575392	5086	13920	14510
8/21/17	638184	592345	6260	15298	18359
9/11/17	621912	580804	5118	14605	16113
9/25/17	647567	603594	5287	17045	16056
10/30/17	551826	520367	4081	11014	12117
11/13/17	601780	537650	4785	34534	18709
12/18/17	596552	550304	4303	19664	16942
1/22/18	259134	240201	1702	7537	7786
2/12/18	619155	588965	2679	15543	8704
3/12/18	402761	371034	2688	12175	13094
4/9/18	449222	409371	2896	20077	12875
5/7/18	499438	455695	3087	23363	13349

Chapter 5: Metagenomic Analysis of a Freshwater Creek and Irrigated Field Reveals Temporal and Spatial Dynamics in Bacterial and Viral Assemblages

5.1 Abstract

Lotic surface water sites (e.g. creeks) are important resources for localized agricultural irrigation. However, there is a concern about the presence of microbial contaminants within untreated surface water that may be transferred onto irrigated soil and crops. To evaluate this issue water samples were collected between January 2017 and August 2018 from a freshwater creek used to irrigate kale and radish plants on a small farm in the Mid-Atlantic, United States. Furthermore, on one sampling date a field survey was conducted in which additional water (creek source and point-of-use) and soil samples were collected to assess the viral and bacterial communities pre- and post- irrigation. All samples were processed for DNA extracts and shotgun sequenced on the Illumina HiSeq platform. The resulting metagenomic libraries were assembled *de novo* and taxonomic and functional features were assigned at the contig and peptide level. From these data we observed that *Betaproteobacteria* (e.g. *Variovorax*) dominated the water, both at the source and point-of-use, and *Al-*

phaproteobacteria (e.g. *Streptomyces*) dominated both pre- and post-irrigated soil. Additionally, in the creek source water there were variations in the abundance of the dominant bacterial genera and functional annotations associated with seasonal characteristics (e.g. water temperature). Antibiotic resistance genes and virulence factors were also identified in the creek water and soil, with the majority specific to their respective habitat. Moreover, an analysis of clustered regularly interspaced short palindromic repeat (CRISPR) arrays showed the persistence of certain spacers through time in the creek water, as well as specific interactions between creek bacteriophage and their hosts. Overall, these findings provide a more holistic picture of bacterial and viral composition, dynamics, and interactions within a freshwater creek that can be utilized to assess its suitability and safety for irrigation.

5.2 Introduction

Lotic ecosystems (e.g. rivers, streams, creeks), which span an estimated 3.5 million miles in the United States, are reliable sources of agricultural irrigation water, especially in the Eastern states [30]. However, concern remains about the microbial communities that inhabit these water sources. In the farm-to-fork continuum, surface water used for irrigation is a known contributor to microbial contamination of fresh produce, responsible for multiple foodborne illness outbreaks [66, 346]. The National Rivers and Streams Assessment of 2008-2009 found that 23% of the evaluated river and stream miles in the U.S. exceeded the EPA's recreational water quality standard for enterococci [347]. Despite this, irrigation water quality is assessed primarily through culture and PCR-based tests for fecal indicator bacteria, frequently with emphasis on *Escherichia coli* [38], which provides limited data and lacks perspective on the complete profile of resident microbial communities. Therefore, there is a compelling need for more comprehensive characterization of these complex microbial communities and the relationship between communities in surface water sources and irrigated soil in order to ensure the safe use of lotic water sources in agriculture.

While there exist some general criteria that help differentiate lotic water bodies, such as size, depth and flow direction, they are not universally accepted and are often indicative of local or regional characteristics [31]. For instance, creeks are often considered smaller, shorter and shallower than rivers, with no branches or tributaries. Nevertheless, microorganisms can be introduced into these water systems via connected waterways, overland runoff and groundwater flow, aerial deposition, and upstream discharge (e.g. sanitary sewer flows, and wastewater treatment plants) [293,348,349]. Once present, these microorganisms can disperse downstream or associate to the stream bank/bed where they are subjected to biotic and abiotic factors that will ultimately control their composition and persistence [350].

Data on the microbial communities of small lotic water bodies (e.g. creeks) are limited, but previous reports on riverine waterways have given us insight into the major factors influencing the composition of planktonic bacterial taxa, such as geography [351], nutrient concentration [150, 152, 352], number of daylight hours [152], pH [353], water temperature [291, 352], water residence time [353, 354], and storm events [352]. The presence and persistence of pathogenic bacteria have also been associated with some of these factors. For instance, rain events have been associated with an increase in pathogen isolation (e.g. *Escherichia coli* O157:H7) from surface water bodies, likely due to the increased influx of runoff water and/or resuspension of streambed bacteria [290–292]. This incursion of pathogens into surface waters that are used for agricultural irrigation raises concern with regard to their potential transmission to crops and ultimately consumers, as several foodborne pathogens are capable of long-term persistence on fruit and vegetable crops following irrigation [62–64].

In addition to foodborne pathogens, antibiotic resistance determinants have been identified in lotic waterways [355, 356], although their persistence on soil and food crops is less clear [351, 351, 357]. While antimicrobial compounds are ancient and can occur naturally, their large-scale manufacture and use, both in clinical and agricultural applications, has led to an increased selection of antimicrobial-resistant bacteria worldwide [213, 325]. Like pathogens, antimicrobials can enter lotic ecosystems through upstream anthropogenic inputs and/or runoff from adjacent fields and pasture [358]. Here, antimicrobial residues can drive selection pressures on native bacterial communities, restricting susceptible populations while favoring resistant taxa or the spread and acquisition of resistance traits [359]. When resistant bacteria contaminate lotic ecosystems, their antimicrobial resistance genes (ARGs) can be disseminated amongst aquatic bacterial communities via conjugation, transformation, and transduction mechanisms [72]. Consequently, bacteriophage (phage), the chief vectors of bacterial transduction, are becoming increasingly recognized for their potential role in the dissemination of ARGs, as well as other genes that shape bacterial community composition (e.g. virulence factors, auxiliary metabolic genes) [73, 360].

Despite this, phage in lotic ecosystems have only been surveyed in a handful of studies, largely through studies of viral metagenomes (viromes) created from samples of large riverine watersheds [141, 153–156]. Moreover, very few studies have sought to integrate phage with their bacterial hosts. Metagenomic analyses of CRISPR (clustered regularly interspaced short palindromic repeats) arrays can provide unique insights into the bacteria-phage interactions in environmental samples. CRISPR spacers, pieces of foreign phage DNA integrated into the host genome, represent a record of phage infection through time [307, 308]. As a result, they can be used as a molecular fingerprint to identify unique host strains, link phage with potential hosts, and reveal information about the heterogeneity of phage present in an environment; data which are critical in determining the potential impact of phage within and between environments.

In this present study, we utilized culture-independent high-throughput shotgun metagenomic sequencing to obtain a more comprehensive picture of the dynamics and potential dissemination of creek microbiota. Briefly, we generated and characterized microbial metagenomes from water samples collected from an agriculturally utilized freshwater creek throughout two growing seasons and from soil pre-irrigation, immediately post-irrigation, and 24 hours post-irrigation. From these microbial metagenomes, we focused specifically on the bacterial component, assessing their taxonomy, functional potential, and resistance/virulence profile. We also performed a preliminary analysis that enabled us to compare the bacterial and viral community composition in the creek water at the source and point-of-use.

5.3 Materials and Methods

5.3.1 Site description

The study site was selected from a group of sampling sites that were included in a 2-year water sampling effort by CONSERVE: A Center of Excellence at the Nexus of Sustainable Water Reuse, Food and Health (www.conservewaterforfood.org). The non-tidal freshwater creek that runs through the site is a tributary of the Patuxent River, is surrounded by forested and peri-urban land, and is characterized by *E. coli* levels that consistently exceed the Food Safety Modernization Act Produce Safety Rule standards for agricultural water [38, 361]. In addition, the creek consistently tests positive for *Salmonella enterica* and *Listeria monocytogenes* [361].

To test whether microbial contaminants present in the creek could be transferred to irrigated soil, a field was established by the farmer ~260 meters from the creek and planted with kale and radish plants, as described in detail in Allard et al. [361]. Briefly, kale and radish seeds were started in a greenhouse and transplanted to the field in rows when plants grew to seedling size. Drip irrigation tape (10 mm wall thickness, 1.0 GPM flow with 10 cm spacing, Nolt's Produce Supplies, Leola, PA #CH-101041) was laid when plants were at the seedling stage, and drip irrigation was carried out intermittently throughout the season using unfiltered water pumped from the creek.

5.3.2 Water sample collection

Creek water samples were collected on the following dates: 1/23/17, 2/27/17, 7/17/17, 8/8/17, 9/11/17, 10/2/17, 7/16/18, and 8/6/18. On each date, a sterile 1 L polypropylene sampling container (Thermo Fisher Scientific, MA, USA) was submerged 15-30 cm below the creek surface (adjacent to the hose intake) using a long-range grabbing tool. On 10/2/17, an additional water sample (5 L) was collected from the creek for virome processing using a utility transfer pump (0.08 W; Everbilt, Atlanta, GA) powered by a EU1000i generator (American Honda Motor Co., Ltd., Alpharetta, GA) and connected to the cartridge via vinyl braided tubing (1.9 cm inner diameter, Sioux Chief, Peculiar, MO).

In addition, on 10/2/17 water samples were collected at the point-of-use. Roughly 6 L of water was collected into sterile 20 L Nalgene polypropylene carboys from a junction in the irrigation hose next to the field. Of this sample, 5 L was allocated for virome processing and 1 L for microbial metagenomic processing. Samples were transported on ice to the laboratory and stored at 4 °C.

5.3.3 Soil sample collection

Starting on 10/2/17, soil samples were collected immediately before irrigation, immediately post-irrigation, and 24 h post irrigation. Prior to 10/2/17, irrigation was halted for a full week to allow for microbial die-off. For the initial irrigation event on 10/2/17, water was turned on and run for ~2 h. Composite soil samples were collected using sterile scoops from within the kale rows, directly underneath

2 parallel lines of drip tape and places into sterile Whirl-Pak bags. Soil samples were collected from the same ~15 cm sampling zone at each soil-sampling event: pre-irrigation, immediately post-irrigation, and 24 h post irrigation. Samples were transported on ice to the laboratory and stored at 4 °C until processing.

5.3.4 Assessment of water characteristics

At each time point, a ProDSS digital sampling system (YSI, Yellow Springs, OH, United States) was used to measure, in triplicate, the water temperature (°C), pH, dissolved oxygen (% DO), conductivity (SPC μS/cm), oxidation-reduction potential (mv), turbidity (FNU), nitrate (mg/L), and chloride (mg/L). Ambient temperature and precipitation levels (24 h prior to sampling) were also collected for each sampling event via the Nation Weather Services historical data archive.

5.3.5 Water sample processing for microbial metagenomes

To isolate the cellular fraction for microbial metagenomes, 1 L samples were vacuum filtered through a 0.2 μm membrane filter (Pall Corporation, MI, USA). Microbial DNA was then extracted from the filters using an enzymatic and mechanical lysis procedure [158, 159]. Briefly, the filters were added to lysing matrix tubes along with a cocktail of PBS buffer, lysozyme, lysostaphin, and mutanolysin. After incubating (30 min, 37 °C), samples were subjected to a second lysing cocktail (Proteinase K and SDS) followed by another incubation (55 min, 45 °C) and mechanical lysis via bead beating. The resulting DNA was purified with the QIAmp DNA mini kit (Qiagen, CA, USA) and assessed with a NanoDrop 2000 Spectrophotometer.

5.3.6 Water sample processing for viromes

Viral DNA extracts were generated as previously described [157]. Briefly, to remove the cellular fraction, 5 L of creek water sampled at the source (10/2/17) and at the point-of-use were filtered sequentially through a Whatman 1 μm polycarbonate filter (Sigma-Aldrich, MO, United States) and a 142-mm diameter 0.2 μm membrane filter (Pall Gelman Sciences, MI, United States) attached via sterile 1.6 mm PVC tubing with a Watson Marlow 323 Series Peristaltic Pump (Watson-Marlow, Falmouth, Cornwall, United Kingdom).

The resulting filtrate was chemically concentrated for viruses using an iron chloride procedure [240], in which 0.5 mL of $FeCl_3$ solution (4.83 g $FeCl_3$ into 100 mL H_2O) was added and incubated in the dark for 1 h. Flocculated viral particles were then filtered onto 142-mm 1 μm polycarbonate filters (Sigma-Aldrich, MO, United States) and stored at 4 °C in the dark until resuspension. For viral resuspension, filters were rocked overnight at 4 °C in 5 mL of 0.1M EDTA-0.2M $MgCl_2$-0.2 M Ascorbate Buffer, described in detail elsewhere [240]. Additionally, to ensure total removal of free DNA contamination, resuspended viral particles were subjected to a DNase I (Sigma-Aldrich, MO, United States) treatment for 2 h and passed through a 33-mm diameter sterile syringe filter with a 0.2 μm pore size (Millipore Corporation, MA, United States). DNA was then extracted from the viral concentrate using the AllPrep DNA/RNA Mini Kit (Qiagen, CA, United States) per the manufacturer's instructions and quantified with an HS DNA Qubit fluorescent concentration assay.

5.3.7 Soil sample processing for microbial metagenomes

Soil samples were hand-homogenized in Whirl-Pak bags, and 0.2 g per sample was transferred into Lysing Matrix B tubes (MP Biomedicals). Samples were stored at $-80\,^{\circ}\mathrm{C}$ until DNA extraction utilizing the same methodology as performed for the water samples.

5.3.8 Shotgun sequencing for microbial metagenomes and viromes

For each sample, DNA was used in a tagmentation reaction followed by 12 cycles of PCR amplification using Nextera i7 and i5 index primers per the modified Nextera XT protocol. The final libraries were then quantitated by Quant-iT hs DNA kit. The libraries were pooled based on their concentrations as determined by Quantstudio 5 and loaded onto an Agilent High Sensitivity D1000 ScreenTape System. Samples were sequenced on an Illumina Hiseq X10 flow cell (Illumina, San Diego, CA, United States) generating 100 bp paired end reads.

5.3.9 Metagenomic assembly

The resulting paired-end reads from both microbial metagenomes and viromes were quality trimmed using Trimmomatic ver. 0.36 (sliding window:4:30 min len:60) [175], merged with FLASh ver. 1.2.11 [176], and assembled *de novo* with MEGAHIT [299]. Open reading frames (ORFs) were predicted from the assembled contigs using

MetaGene as described in Noguchi et al. [104].

5.3.10 Taxonomic and functional classification

For the microbial metagenomes, peptide sequences encoded by the predicted ORFs were searched against UniRef 100 (retrieved May 2018) using protein-protein BLAST (BLASTp ver. 2.6.0+) (E value $\leq$ 1e^{-3}) [105]. Taxonomic classifications were then made to contigs by max cumulative bit score. This was calculated by summing the bit scores of all taxa with a hit to peptides encoded by the contig. Peptide sequences were assigned all Gene Ontology GO terms that were linked to UniRef 100 peptides within 3% of the top hit's bit score [179]. Taxonomic classification of viromes were conducted as described in Chopyk et al. [157].

For all microbial metagenomes and viromes, coverage was calculated for each contig by recruiting quality-controlled reads to assembled contigs using Bowtie2 ver. 2.3.3 (very sensitive local mode) and then using the "depth" function of Samtools ver. 1.4.1 to compute the per-contig coverage [180]. To normalize abundances across libraries, contig and ORF coverages were divided by the sum of coverage per million, similar to the transcripts per million (TPM) metric used in RNA-Seq [181,257]. Scripts performing these assignments and normalization are available at https://github.com/dnasko/baby_virome. Taxonomic and functional data were visualized using the R packages ggplot2 ver. 3.1.0 and pheatmap ver 1.0.10 [182,183].

5.3.11 ORF clustering

To assess the shared and unique nucleotide composition among the water and soil microbial metagenomes, all complete nucleotide ORFs (i.e. ORFs with both a start and a stop codon) were clustered at 97% with CD-HIT-EST and subsequently parsed with the clstr2txt.pl script [184].

5.3.12 Identification of ARGs and VFs

For both the microbial metagenomes and viromes peptide sequences encoded by the predicted ORFs were searched against the "Comprehensive Antibiotic Resistance Database" (CARD; retrieved July 2018) and the core dataset of virulence factors (VFs) (genes associated with experimentally verified VFs) from the "Virulence Factor Database" (VFDB; retrieved December 2018) [362] using protein-protein BLAST (BLASTp ver. 2.6.0+) (E value $\leq$ 1e^{-3}) [105, 107]. A queried translated ORF was regarded as ARG or VF-like if >40% coverage and >80% amino-acid identity to a protein in the CARD or VF database, a conservative threshold defined previously [83, 185]. In addition, for the ARGs conferring resistance through mutations (i.e. *KasA*, *gyrA*, *gyrB*, *murA*, *ndh*, *ThyA*, *rpsL*, *rpsJ*), a post-processing step (MAFFT alignment with reference sequences available at CARD) was taken to confirm the presence of resistance-conferring mutations [186]. ARG and VF data were visualized using the R package ggplot2 ver. 3.1.0 [183].

5.3.13 Prediction and analysis of CRISPRs in microbial metagenomes

CRISPR arrays were predicted from assembled contigs using the CRISPR detection and validation tool, CASC [303]. Valid CRISPR spacers were clustered with CD-HIT-EST at 97% nucleotide similarity to determine the number of unique and shared spacers both through the course of the experiment and between soil and water samples [184]. Spacers were also used as query sequences to search against the viromes from the creek source and the point-of-use using BLASTn (E value $\leq$ 1e^{-1}, word size 7). Only alignment of length 28-40 bp (inclusively), with bit score $\leq$ 42, and $\leq$ 3 mis-matches were considered valid and used in the network analysis. Only viral contigs that were assigned a taxon were shown in the network analysis, visualized by Cytoscape software [300].

5.3.14 Statistical tests

Pearson correlation coefficients were calculated to identify associations between the water characteristics and the normalized abundance of the bacterial genera and functional assignments.

5.3.15 Data availability

Metagenomic reads were submitted to NCBI's Sequence Read Archive under the accession numbers SRS4362885-SRS4362895 for the water samples and SRS4378913-SRS4378915 for the soil samples.

5.4 Results

5.4.1 Creek water physicochemical characteristics

Water physicochemical factors measured in both the creek source water and at the point-of-use are described in Table 5.1. During the course of the study, the creek source water ranged in temperature from 7 °C to 24 °C, with an average temperature of 17 °C. The only dates with precipitation 24 hours prior to sampling were on 1/23/17 and 8/8/17. Additionally, the physicochemical factors measured at the point-of-use were similar to those measured at the creek source water on the same sampling date (10/2/17), with only slight increases in ambient temperature, dissolved oxygen, conductivity, ORP, turbidity, and nitrate levels.

5.4.2 Sequencing effort and assembly

In total, 12 samples were processed for microbial metagenomes, eight from 1 L samples collected from the freshwater creek periodically over two years, one from a 1 L sample collected from the point-of-use on 10/2/17, and three from soil samples collected beginning on 10/2/17 (one pre-irrigation, one immediately post-irrigation, and one 24 h post-irrigation). Additionally, two samples were processed for viromes: one from a 5 L sample collected from the creek source water and one from a 5 L sample collected from the point-of-use, both on 10/2/17. All samples were shotgun sequenced for a total of 989,086,930 read pairs at an average of 70,649,066 read pairs per metagenome (± 13,168,798) (Table 5.2, Table 5.3). Following assembly, there were a total of 13,607,046 contigs, with an average of 971,932 contigs per sample (± 338,945).

5.4.3 Bacterial phyla in water and soil microbial metagenomes

In both the soil and water microbial metagenomes between 71 and 88% of contigs could be assigned a taxonomic representative (Table 5.4), with the majority (>95%) homologous to Bacteria. Of the contigs that could be assigned a taxon, normalized abundance was calculated to account for sequencing effort and assembly proficiency (TPM-like normalization). For the creek source water *Proteobacteria* dominated at all time points in creek source water, accounting for 62% (± 3%) of the total normalized abundance assigned to Bacteria. The next most abundant phyla in the creek source water were *Actinobacteria* (15% ± 3%), *Bacteriodetes* (13% ± 2%), and *Firmicutes* (5% ± 2%). The distribution of these dominant phyla was similar at the point-of-use, especially compared to the creek water collected at the source on the same date (Figure 5.1).

In soil, *Proteobacteria* also dominated (pre = 44%, post = 46%, 24 h = 47%) (Figure 5.1). However, the *Proteobacteria* phylum was composed largely of the class *Alphaproteobacteria* (pre = 44%, post = 43%, 24 h. = 45%) instead of *Betaproteobacteria*, which dominated the creek source water (71% ± 4) (Figure 5.1). In soil, the next largest phyla were *Actinobacteria*, (pre = 25%, post = 20%, 24 h = 23%) followed by *Firmicutes* (pre = 11%, post = 11%, 24 h = 10%), and *Acidobacteria* (pre = 5%, post = 7%, 24 h = 5%), all of which exhibited a much higher normalized abundance in soil than in water.

5.4.4 Bacterial genera of water and soil microbial metagenomes

For the contigs assigned to bacterial taxa we also report the normalized abundance of the dominant genera (>1% of the total bacterial genera in a least one sample) detected in water and soil (Figure 5.2). *Variovorax* was the most abundant genus in the creek source water accounting for an average of 21% (± 4%) of the total normalized abundance of the bacteria assigned contigs. Following *Variovorax*, *Streptomyces* (5% ± 1%), *Pusillimonas* (5% ± 0.4%), *Achromobacter* (4% ± 0.4%), and *Microbacterium* (3% ± 2%) were the most abundant. In agreement with the phyla distributions described above, the abundance of these bacterial genera was similar between the creek source and the point-of-use.

In soil, the distribution of dominant genera was as follows: *Streptomyces* (pre = 9%, post = 7%, 24 h = 9%), *Mesorhizobium* (pre = 5%, post = 5%, 24 h = 6%), *Sphingomonas* (pre = 4%, post = 3%, 24 h = 4%), *Bacillus* (pre = 3%, post = 3%, 24 h. = 3%), and *Achromobacter* (pre = 2%, post = 3%, 24 h = 3%).

5.4.5 Creek source water characteristics and bacterial genera abundance

To investigate the dynamics of the bacterial genera over time in the creek source water, correlation analysis was performed between the dominant bacterial genera and the measured water characteristics described in Table 5.1 (Figure 5.3).

Water temperature had the greatest significant impact on the normalized abundance of the most genera. For instance, *Mesorhizobium*, *Rhodopseudomonas*, *Microvirga*, *Ferrimicrobium*, *Sphingopyxis*, *Syntrophomonas*, *Saprospira*, *Desulfotignum*, *Sphingomonas*, and *Alicyclobacillus* were all positively correlated (Pearson's correlation, $p \leq 0.05$) with water temperature, whereas the abundance of *Flavobacterium* and *Clostridioides* was negatively correlated with water temperature.

Other water characteristics that were significantly ($p \leq 0.05$) correlated with the normalized abundance of the dominant bacterial genera were as follows: dissolved oxygen was negatively correlated with *Sphingopyxis*, *Pseudomonas*, *Methylophaga*, *Desulfotignum*; conductivity (SPC uS/cm) was negatively correlated with *Polynucleobacte*r; oxygen reduction potential was positively correlated with *Ulvibacter*, *Flavobacterium*, *Clostridioides* and negatively correlated with *Syntrophomonas*, *Ferrimicrobium*, *Mesorhizobium*, and *Microvirga*; turbidity was positively correlated with *Clostridium* and *Sphingomonas*; and precipitation positively correlated with *Clostridium*.

5.4.6 Functional potential of water and soil microbial metagenomes

To characterize the functional profiles of the microbial metagenomes, Gene Ontology (GO) annotations were assigned to peptide sequences encoded by the predicted ORFs based on BLASTp matches to UniRef100 proteins (Figure 5.4). On average, 70% (± 9%) of peptide sequences in the source creek water and 74% in the point-of-use water were assigned at least one GO-term. Similarly, for the soil an average of 72% (± 3%) of peptide sequences were assigned at least one GO-term. For both sample types, the GO-term "transferase activity" (GO:0016740) was assigned to the greatest percentage of the normalized abundance assigned to the predicted peptides, accounting for 27% (± 1%) in the creek source water, 28% in the point-of-use, and 28% in the soil (pre = 28%, post = 28%, 24 h. = 29%) (Figure 5.4). Other functions exhibiting high normalized abundance in both water and soil metagenomes included "metal ion binding", "catalytic activity", "oxidation-reduction process" and "oxidoreductase activity".

5.4.7 Relationships between water characteristics and functional potential

Correlation analysis was performed between the normalized abundance of functional profiles (GO-terms) and the measured water characteristics described in Table 5.1 (Figure 5.5). Again, water temperature was the most significant factor and was positively correlated ($p \leq 0.05$) with the normalized abundance of following GO-terms: "catalytic activity" (GO: 0003824), "lyase activity" (GO: 0016829), "proteolysis" (GO: 0006508), "metabolic process" (GO: 0008152), "oxidation reduction process" (GO: 0055114), "hydrolase activity" (GO: 0016787), and "oxidoreductase activity" (GO: 0016491).

5.4.8 ORF clustering in water and soil microbial metagenomes

To explore the persistence and overlap of microbiota within the samples collected at the creek, as well as with the point-of-use water and soil samples, all ORFs were clustered at 97% similarity (Figure 5.6). Between 19 and 70% of ORFs from water clustered with at least one other sample. These were largely other water samples from the same general season. For example, 20% of ORFs from 8/8/17 clustered with ORFs from the following month, 9/11/17, and 11% from the preceding month, 7/17/17. In contrast, only 2 and 3% of ORFs from 8/8/17 clustered with the ORFs from the winter months, 1/23/17 and 2/27/17, respectively. For the point-of-use, 45% of ORFs formed clusters with ORFs from the creek water source collected on the same date, 10/2/17.

Few ORF clusters contained ORFs from both soil and water. In total ~0.002% of ORFs from the pre-irrigated soil, ~0.01% of ORFs from the post-irrigated soil, and ~0.02% of ORFs from the 24 h post-irrigated soil clustered with ORFs from any of the water samples. Not surprisingly, ORFs from the soil clustered at a greater proportion with each other (Figure 5.7). For instance, 10% of the ORFs in the pre-irrigation soil formed clusters with ORFs from the post-irrigation soil and 9% with ORFs from the 24 h post-irrigation soil.

5.4.9 ARGs in water and soil microbial metagenomes

In total, 138 peptide sequences encoded by the predicted ORFs from both the water and soil microbial metagenomes were classified as 23 unique ARGs (Figure

5.8). Of these, eight ARGs were identified only in creek source water, conferring resistance to cephalosporin, penam, fluoroquinolone, phenicol, elfamycin, peptide, tetracycline, and aminoglycoside antibiotics. Ten ARGs were identified only in the soil, conferring resistance to aminoglycosides, glycopeptide, fosfomycin, rifamycin, isoniazid, macrolides streptogramin and tetracycline antibiotics. Five ARGs were identified in at least one creek water source and one soil sample, conferring resistance to aminoglycoside, peptide, rifamycin, diaminopyrimidine, macrolide, and penam antibiotics. Finally, three ARGs were detected in the point-of-use, one of which was identified also in the creek source water and two in both the creek source water and soil samples.

The greatest diversity of ARGs in the creek source water was identified in the sample retrieved on 8/6/18, which had five unique ARGs. However, the majority of the ARG diversity was identified from soil, with 11 unique ARGs predicted in the pre-irrigation soil, eight unique ARGs in post-irrigated soil, and seven unique ARGs in the 24 h post-irrigated soil. Additionally, for the ARGs that confer resistance through target mutation, the following mutations were identified: *murA* C117D [191], *rpsL*, K88R [192], *rpsJ* V57M [363], *rpoB* H526T [305], and *EF-Tu* Q124K [195] Y161N [364].

5.4.10 Putative hosts of antimicrobial resistance genes

All of the ARGs that could be confidently assigned a taxonomic representative (134/138) were assigned to Bacteria (Figure 5.8, Table 5.5). For each ARG, we calculated its normalized abundance. In the creek source water, 49% of the normalized abundance assigned to ARGs was attributed to *Acidobacteria* (*Saccharomonospora*, *Streptomyces*, *Ferrimicrobium*, and *Nocardia*) and 43% to *Proteobacteria* (*Variovorax*, *Stenotrophomonas*, *Achromobacter*, *Burkholderia*, *Nitrosovibrio*, *Caulobacter*, *Chania*, *Pseudomonas*, and *Agrobacterium*). In the point-of-use 59% of the normalized abundance assigned to ARGs originated from *Acidobacteria* (*Saccharomonospora*, *Streptomyces*) and 41% from *Proteobacteria* (*Sphingopyxis* and *Limnohabitans*).

For the soil samples, while the abundance and diversity of ARGs was much greater than that in water, most (96%) of the normalized abundance assigned to ARGs was attributed to one phylum, *Acidobacteria* (*Streptomyces*, *Nocardia*, *Frankia*, *Rhodococcus*, *Ferrimicrobium*, *Actinoplanes*, *Aeromicrobium*, *Cellulosimicrobium*, *Mycobacterium*, *Nocardia*, and *Williamsia*) (Figure 5.8, Table 5.5).

5.4.11 Virulence factors in water and soil microbial metagenomes

In addition to ARGs, we identified several putative virulence factors (VFs) in the soil and water microbial metagenomes. In total, 629 peptide sequences encoded by the predicted ORFs from both the water and soil were classified as 67 unique VFs (Figure 5.9, Table 5.6, 5.7). Of these, 37 VFs were predicted only in creek source water and included proteins involved in regulation, stress response, mobility, the formation of the lipopolysaccharide (LPS) and/or capsule, hemolysis, adherence, and type II, III and VI secretion systems. Additionally, 13 VFs were identified only in the soil and included proteins involved in binding at the cell surface, mobility, cell wall formation, iron uptake, metabolic adaption, and type VII secretion system. There were also 16 VFs found in at least one creek source water and one soil sample, which included proteins involved in the formation of the LPS and/or capsule, regulation, mobility, adherence, and type VI secretion system. Finally, 12 VFs were identified in the point-of-use, one of which, *algU*, was identified only in the point-of-use.

Overall, the creek freshwater in August had the highest diversity of VFs, with the 8/8/17 sampling date characterized by 22 unique VFs and the 8/6/18 sampling date characterized by 30 unique VFs. In the soil, there was also a high diversity of VFs, with 19 unique VFs predicted in the pre-irrigation soil, 19 unique VFs predicted in the post-irrigated soil, and 17 unique VFs in the 24 h post-irrigated soil.

5.4.12 Putative hosts of virulence factors

Similar to the ARGs, all VFs that could be confidently assigned a taxonomic representative were assigned to Bacteria (Figure 5.9, Table 5.7). However, in this case, the host phyla were largely *Proteobacteria* as opposed to *Actinobacteria*, which hosted the ARGs. In the creek source water 98% of the normalized abundance of VFs was attributed to *Proteobacteria* (*Polynucleobacter*, *Pseudomonas*, *Lutimaribacter*, *Aeromonas*, *Desulfotignum*, *Delftia*, *Nitrosovibrio*, *Variovorax*, and 33 other genera). Likewise, in the point-of-use water, 98% of the normalized abundance assigned to

VFs originated from *Proteobacteria* (*Aeromonas*, *Polynucleobacter*, *Pseudomonas*, *Lutimaribacter* and 12 other genera). Finally, in the soil 76% of the normalized abundance assigned to VFs was also attributed to the *Proteobacteria* phylum (*Lutimaribacter*, *Sphingomonas*, *Sphingopyxis*, *Pseudomonas*, *Rhodopseudomonas*, *Microvirga*, and 21 other genera) (Figure 5.9, Table 5.7).

5.4.13 Viral taxonomy and ARG/VF in source and point-of-use viromes

For the creek water source and the point-of-use water viromes, the majority of assigned contigs were homologous to Viruses (source = 61%, point-of-use = 58%), followed by Bacteria (source = 10%, point-of-use = 15%). In the same manner as was conducted for the microbial metagenomes, a normalized abundance was calculated for these viromes (Figure 5.10). At both the source and point-of-use, the majority of the normalized abundance of viral contigs was assigned to the ds-DNA bacteriophage of the order *Caudovirales* (source = 98%, point-of-use = 98%). Within the *Caudovirales*, *Siphoviridae* (source = 42%, point-of-use = 42%) dominated at all time points followed by *Podoviridae* (source = 31%, point-of-use = 27%), *Myoviridae* (source = 23%, point-of-use = 27%), and the newly introduced *Ackermannviridae* (source = 1%, point-of-use = 1%) [365]. No VFs or ARGs were identified in either virome.

5.4.14 CRISPR arrays in water and soil microbial metagenomes

To determine the prevalence of the CRISPR-Cas defense system, as well as track specific microbial strains among sample types and through time we predicted CRISPR arrays in the soil and water microbial metagenomes (Table 5.8). A total of 370 arrays were identified with 1,645 CRISPR spacers, of which 1,301 were unique. No spacers were shared between the soil and water samples. However, for the soil samples, the pre-irrigated soil shared five spacers with the soil 24 h post irrigation.

Moreover, it appeared that some spacers persisted throughout the study within the creek water source and, in some cases, appeared over a year apart (Figure 5.11). Of the 1,114 unique spacers found in the creek source water, 124 were found in at least two time points. For example, three spacers with shared homology were predicted in the creek source water from 2/27/17, 9/11/17 and 8/6/18 (Figure 5.11).

The point-of-use sample also shared spacers with creek source water (Table 5.9). Of the 333 spacers predicted in the point-of-use water, 42% (140 spacers) were also found in the creek source water on the same sampling date (10/2/17). Additionally, the point-of-use shared spacers with creek source water for the following dates: 7/17/17 (34 spacers), 7/16/18 (22 spacers), and 9/11/17 (10 spacers).

5.4.15 Phage-host relationships

To make connections between the viromes and microbial metagenomes the CRISPR spacers identified in the microbial metagenomes were utilized to link phage with putative hosts (Figure 5.12). Here, we observed that 61 spacers from 38 different contigs from the water metagenomes matched 18 different viral contigs from the creek water source virome (7 *Siphoviridae*, 4 *Podoviridae*, 5 *Myoviridae*, and 2 unclassified) and 39 from the point-of-use virome (19 *Siphoviridae*, 8 *Podoviridae*, 11 *Myoviridae*, and 1 unclassified). For the total 57 viral contigs that were matched to microbial metagenomes, five hit more than one microbial contig.

Of the 38 contigs from the water microbial metagenomes, 16 had specific one-to-one relationships, in other words, one contig from any of the water microbial metagenomes matched one viral contig either from the creek water source virome or the point-of-use virome. Similarly, there were many one-to-two relationships, in which 16 contigs from any of the water microbial metagenomes each matched to two viral contigs from the viromes. For these, the two matched viral contigs were usually split between the two sites, one from the creek source water virome and one from the point-of-use virome and were likely the same viral population present at both sites. Finally, there were some one-to-many relationships, in which six contigs from any of the water microbial metagenomes each matched at least three viral contigs from the creek source water virome and/or from the point-of-use virome.

5.5 Discussion

Comprehensive surveillance of microbial community composition and functional potential in lotic agricultural water is essential, not only for maintaining food safety, but also for ensuring the health of the entire waterway. Despite this, microbial communities of lotic surface waters are far less studied than those of marine and lake ecosystems, especially small lotic systems like creeks [146]. Here, we present a temporal survey of microbial communities from water samples collected from a freshwater creek used actively for agricultural irrigation of a small produce farm. We also report preliminary data on the dynamics and potential dissemination of these communities from water source to soil.

5.5.1 Composition and dynamics of bacteria in water and soil samples

Similar to other freshwater lotic environments, such as the James River in Virginia [147] and the Santa Ana River in California [149], the freshwater creek sampled here was dominated by *Proteobacteria*, specifically *Betaproteobacteria* at all time points (Figure 5.1). However, there appeared to be strong seasonal variability that influenced the taxonomic and functional composition of the community. Seasonal changes in the abundance and functional profile of bacteria, likely brought on by increases in water temperature, organic matter and nutrient availability, are common in aquatic systems [350]. In this study, we found that out of the physicochemical factors tested, water temperature appeared to have the most marked effect on the abundance of functional genes related to production and metabolism, as well as the abundance of several dominant bacterial genera (Figure 5.3, 5.5). In addition, ORFs were shared between samples that were collected at the same time points and seasons (Figure 5.7), suggesting a persistence and reemergence of specific bacterial lineages over time. This seasonal synchrony in microbial diversity has been previously reported in river water, as well as epilithon (algae on river rocks) and sediment in lotic systems [366, 367].

In contrast to the creek water, the agricultural soil was dominated by *Alphaproteobacteria*, largely *Streptomyces*, *Mesorhizobium*, and *Sphingomonas*. These genera have been previously identified at high abundance in soil associated with asparagus and sugar beets [368, 369]. Moreover, at all irrigation stages the community appeared rather stable, especially after 24 h (Figure 5.1). This confirms previous observations from a study on agricultural soils from Illinois, in which the microbial community composition was largely stable year-round, especially when compared to aquatic systems [370]. The authors attributed this to the large average genome sizes of soil bacteria, which may enable them to withstand varying environmental conditions through gene modulation rather than changes in abundance [370]. Additionally, because soil is composed of heterogeneous microenvironments shaped by contrasting physicochemical and biological properties its microbial diversity is nearly unparalleled [371]. This competitive environment may negatively influence the ability of aquatic microbes to establish a niche. However, it is important to note that due to the limited number of soil samples collected in this study we only captured a snapshot of the diversity of the soil microbiota and its response to irrigation.

5.5.2 Diversity and abundance of ARGs in water and soil samples

Further variations between water and soil samples, as well as creek water across time and location were highlighted in the abundance and diversity of VFs and ARGs (Figure 5.8, 5.9). ARGs were detected over time in the creek source water and across the different sample types. However, soil had the greatest pool of ARG diversity. This agrees with a previous study that surveyed 71 environmental and host-associated shotgun metagenomic libraries and found that soil, regardless of anthropogenic impact, had the largest diversity of ARGs [372]. Soil is increasingly recognized as a natural reservoir of antimicrobial resistance, not only due to its close association with agricultural and livestock antibiotics, but also due to the presence of *Streptomyces* spp. [373–375]. In this study, the majority of soil ARGs were predicted from contigs assigned to *Streptomyces* spp. and are potentially capable of conferring resistance to a broad range of antibiotics including: aminoglycosides, rifamycins, macrolides, and glycopeptides. *Streptomyces* are responsible for the production of a large number of clinically significant antibiotics and are suggested to carry resistance determinants even in the absence of anthropogenic antimicrobial contamination [213, 329, 330]. Given this, the soil environment may provide a reservoir of ARGs that could potentially flow via horizontal gene transfer to pathogen populations. A previous study found that resistance genes encoded by soil *Proteobacteria* had a perfect nucleotide match to resistance genes sequenced from clinical isolates of several human pathogens including species of *Escherichia*, *Enterobacter*, and *Salmonella* [374]. More work is needed to determine the potential drivers of bacterial transduction and conjugation between pathogens that can persist in the soil and the reservoir of soil ARGs.

5.5.3 Diversity and abundance of VFs in water and soil samples

In addition to ARGs we observed the presence of VFs. Although the identification of VF genes does not indicate pathogenicity, it can be a useful guide to highlight samples that should be explored in greater detail (e.g. transcriptomic, proteomic, and/or culture-based analyses). We found that the majority of VFs from both water and soil samples were largely opportunistic factors essential for survival (e.g. motility, cell wall formation) and, as a result, are not necessarily pathogen-specific [376]. In the soil there was an abundance of genes associated with the type VII secretion system, largely those involved in the ESX-3 and ESX-5 systems. *Mycobacteria*, including the pathogens *M. tuberculosis*, *M. leprae*, *M. marinum*, *M. ulcerans*, and *M. avium* all use type VII protein secretion systems for both housekeeping functions (e.g. metal homeostasis, cell wall stability) and the secretion of virulence factors [377–379]. In contrast, the creek water had an abundance of genes involved in type II and III secretion systems. While type II and III secretion systems are instrumental to bacterial survival and found widespread in non-virulent gram-negative bacteria, a previous study reported that the type III effectors, such as *aopN*, are directly involved in virulence and are likely to be pathogen-specific [376, 380, 381]. *AopN* in pathogenic *Aeromonas* spp. is responsible for controlling the secretion of translocator proteins and suppressing immunity inside host cells [382]. This gene along with other *Aeromonas* spp. associated VFs were identified in the creek water on 8/6/18. *Aeromonas* spp. are becoming increasingly recognized as potential enteric pathogens and their presence in irrigation water may be of concern [383], especially since they often do not correlate with fecal indicators [384–386].

Moreover, there was one VF, *algU* from a *Pseudomonas spp.*, that was identified just in the point-of-use water sample and at no other time within the creek water source or soil [387]. *AlgU* is a key regulator involved in alginate biosynthesis in *Pseudomonas aeruginosa* [387]. In response to certain conditions, alginate is secreted and contributes to the formation of the extracellular matrix, which provides enhanced adhesion to solid surfaces and protection from external stresses [388]. This provides preliminary evidence of the capability of biofilm formation in the water distribution system that transports the creek water to the field. While the water collected at the source and at the point-of-use were similar with regard to the functional and taxonomic composition of the microbiota, the presence of biofilms may have important public health implications. Biofilms have been reported to form in irrigation and drinking water systems and have been suggested to act as a reservoir for pathogenic microorganisms [389]. Once associated in a biofilm, these pathogens, along with other microbiota, may be released periodically into the water supply [389]. In fact, a previous study from our lab recovered, through culture based analysis, *Salmonella enterica* only at the point-of-use and not in the creek source water [361]. As a result, upstream water quality sampling may not fully capture the risk at the point-of-use.

5.5.4 Phage community structure and interactions in creek water

In addition to assessing the bacterial components from a variety of microbial metagenomes, we were also able to characterize viral communities from a subset of water samples. Viral communities are a component of lotic freshwater microbial ecology that only a handful of studies have investigated, with the majority of research limited to large riverine systems. Here, we found that the viral communities of the water sampled at the creek and at the point-of-use were dominated by dsDNA bacteriophage belonging to the order *Caudovirales*. This has been observed in previous research on large riverine watersheds (e.g. Bess River in Spain [155], Amazon River in Brazil [156], Ile River in China [141], Murray River in Australia [154]). However, these studies identified *Myoviridae* as the most abundant family of *Caudovirales* in their samples, while we found *Siphoviridae* to be the most abundant family across all samples. *Siphoviridae* have been previously found to prevail in terrestrial environments, as well as in some freshwater ponds and lakes [139,157,390]. This difference may be attributed to the presence of *Siphoviridae* in both free-flowting freshwater and suspended sediment from the creek bank/bed. In a sediment sample of the Seine River, the majority of viral sequences were homologous to temperate phage infecting *Proteobacteria* [334]. The majority of cultured representatives of *Siphoviridae* are temperate and are thus capable of horizontal gene transfer through lysogenic integration [276]. While no ARGs or VFs were identified within the viromes, we did find evidence for the persistent infection of bacteria by phage in the creek water.

By leveraging the CRISPR-Cas system we were able to go beyond simple taxonomic classification and begin to survey the history of phage infection within the creek. Interestingly, we found that 11% of unique spacers identified in the creek water source persisted, as they were observed months and years apart (Figure 5.11). Likely these spacers have been inherited over multiple generations to ensure protection against certain phage species. This provides further evidence that, despite being a flowing water body, some microbial populations are maintained in the creek over multiple years.

Additionally, 4% of the total unique CRISPR spacers matched phage present at the creek source water and/or the point-of-use (Figure 5.12). Although the virome data was limited, this CRISPR analysis showcases the varying degrees of phage infections through time. Notably, we identified the presence of some generalist phage, capable of infecting more than one bacterial species [391]. This suggests the potential for cross-infection and horizontal gene transfer across different bacterial classes or even phyla; a broad-host range previously reported in phage isolated from Lake Michigan [392]. However, it is important to note that the spacers matching contigs from these two viromes only represents a small fraction of the total spacers that were identified over time, owing to the dynamic viral populations that are likely present in the creek water.

5.5.5 Limitations and summary

Our research provides an integrated picture of the bacterial and viral communities in creek freshwater and their potential impact on soil health and microbial community structure. However, this study is not without its limitations, mostly related to the small number of metagenomes created at the point-of-use and from the soil pre- and post- irrigation. Replicate samples for the point-of-use and each of the soil stages would be necessary to build a more confident picture of their variability [393]. Despite this, we anticipate that these data will serve as a foundation for future studies regarding the impact of creek freshwater from source to soil, as well as inform research on best practices for water management and monitoring.

5.5.6 Conclusions

One route by which pathogenic microorganisms enter the food production chain is through irrigation water. As a result, water sources utilized for agricultural application must be carefully and comprehensively characterized from source to soil. In this study, we utilized culture independent shotgun metagenomics to analyze multiple features of the bacterial communities in an agricultural freshwater creek. Overall, we found that seasonality was strongly associated with certain bacterial genera, functional potential, and VF diversity in the creek freshwater. By leveraging the CRISPR-Cas system within the creek source water microbial metagenomes we also demonstrated the persistence of specific bacterial and phage lineages in this lotic environment. Moreover, some of these CRISPR spacers matched co-existing phage present in the creek water at the source and point-of-use. From a small subset of samples we also found creek water at the source and at the point-of-use shared a large percentage of ORFs and similar taxonomic and functional composition, both viral and bacterial. However, there were some potential differences between these two sites, such as the presence of biofilm forming genes at the point-of-use. Moreover, soil, regardless of irrigation status, was dominated by *Streptomyces* and, as a result, had a high diversity of ARGs. In conclusion, these findings provide a more complete picture of the microbiota within a freshwater creek that can be utilized to form more inclusive surveys on their potential environmental and public health risks.

5.6 Figures

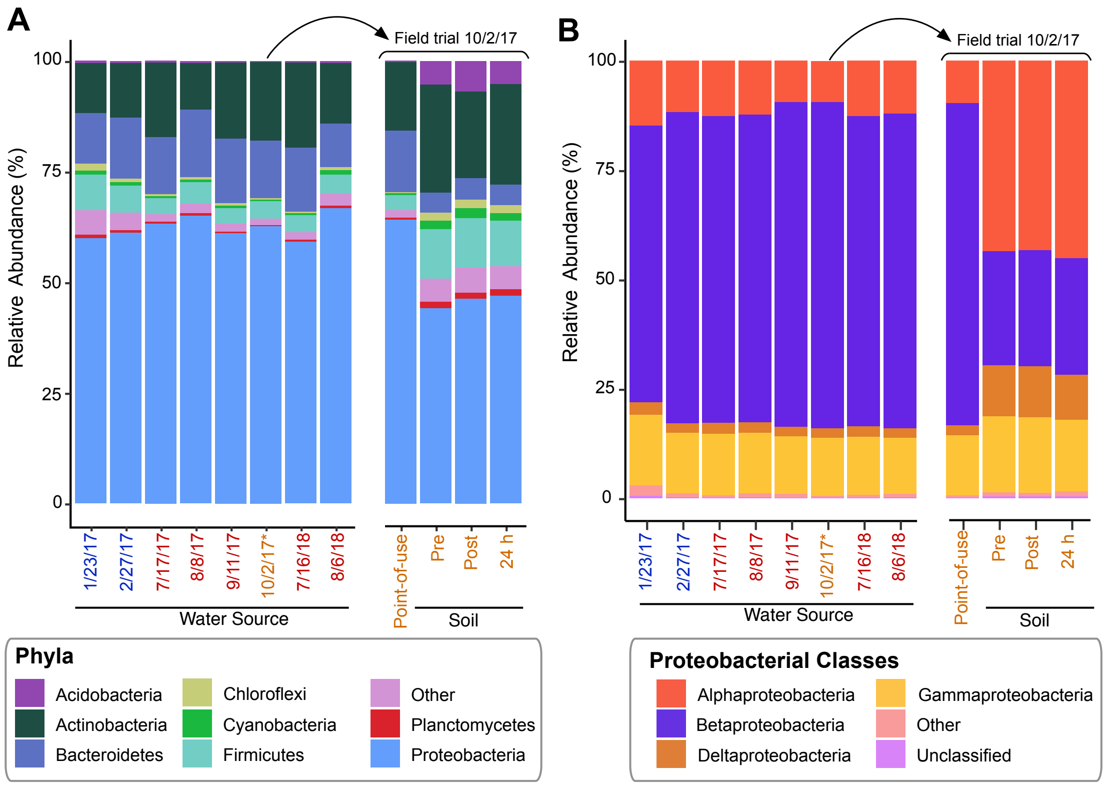

Figure 5.1: Bacterial composition in creek source water, water at the point-of-use, and soil. (A) Stacked bar chart depicting the relative abundance within the bacterial communities at the phylum level. (B) Stacked bar chart depicting the relative abundance within the *Proteobacteria* phyla split into classes. Samples are organized temporally and separated by sample type. Sample labels are colored by season collected (winter: blue, summer: red, autumn: brown). *denotes the date the point-of-use and soil samples were also collected.

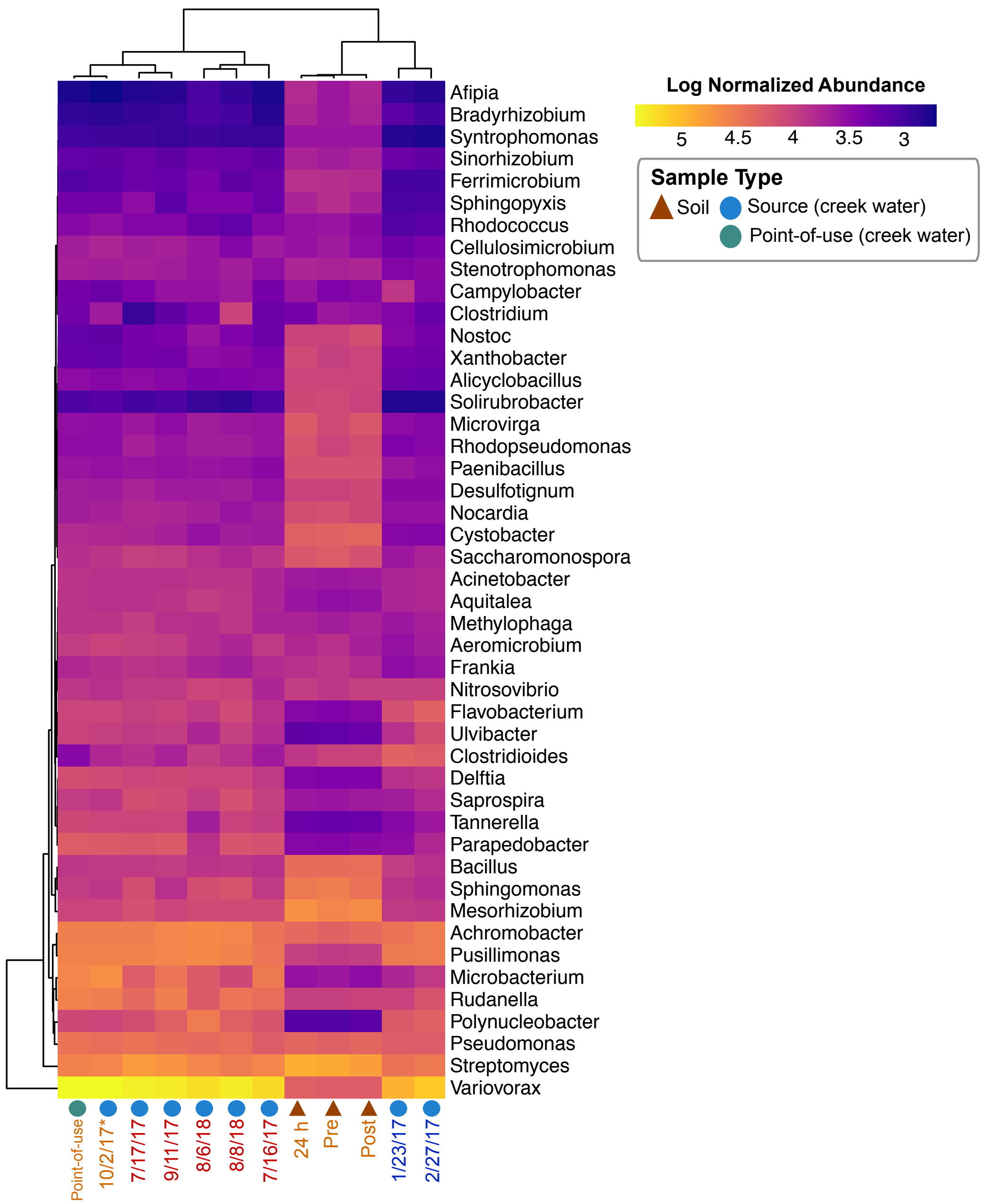

Figure 5.2: Bacterial genera in creek source water, point-of-use water, and soil. Heatmap based on the log-transformed normalized abundance (TPM-like normalization) of the most dominant genera (>1% in at least one sample). Hierarchical clustering of samples was performed using the complete clustering method with Euclidean distances. Sample labels are colored by season collected (winter: blue, summer: red, autumn: brown) and annotated with a symbol corresponding to the sample type (source water: blue circle, point-of-use teal circle, soil: brown triangle). *denotes the date the point-of-use water and soil samples were also collected.

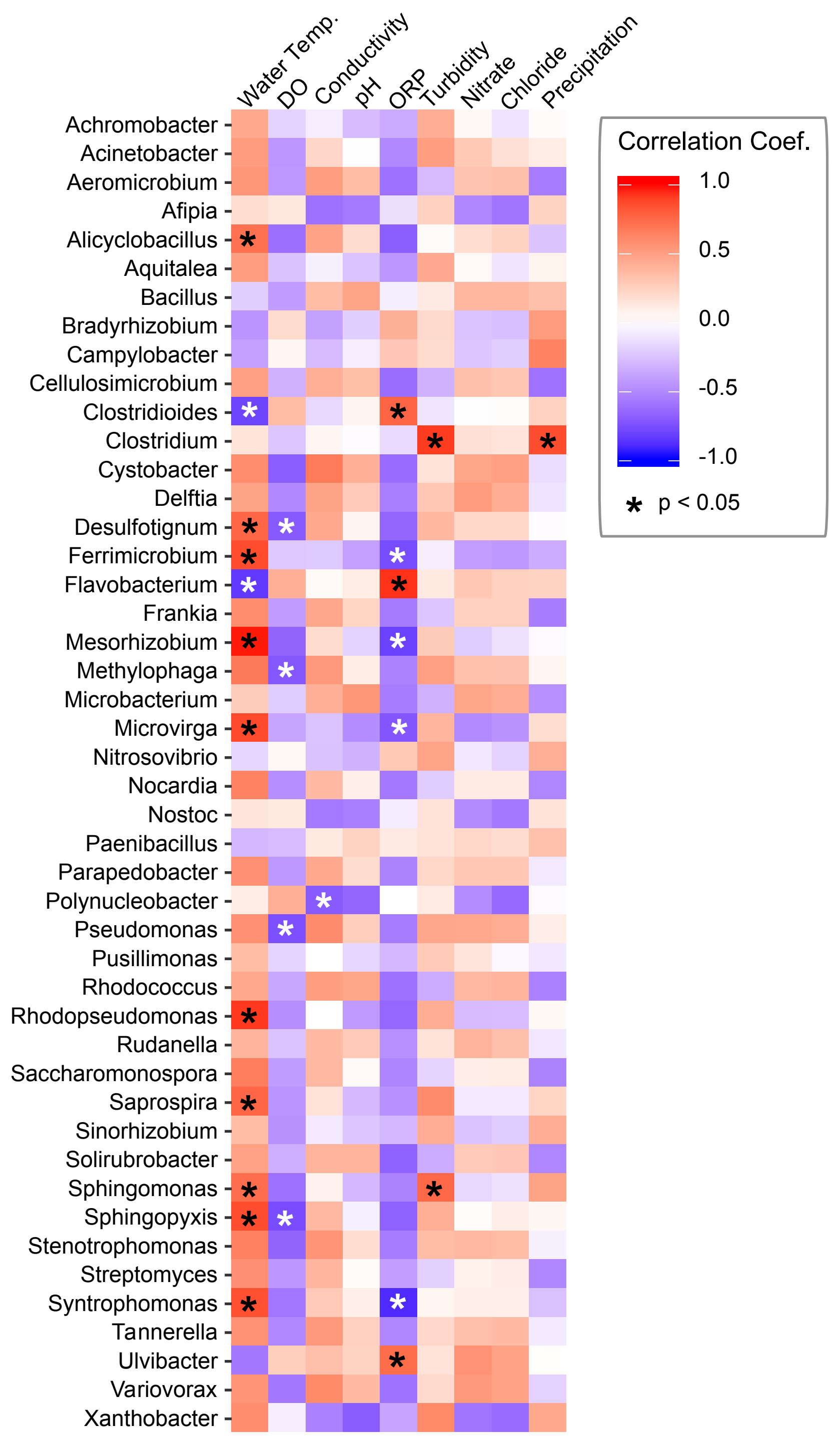

Figure 5.3: Heatmap of the Pearson's correlation coefficients between the water characteristics and normalized abundance of bacterial genera in the creek source water. Color gradients reflect the different values of Pearson's correlation coefficients. ORP: Oxidation/reduction (mV), DO: Dissolved Oxygen (%).

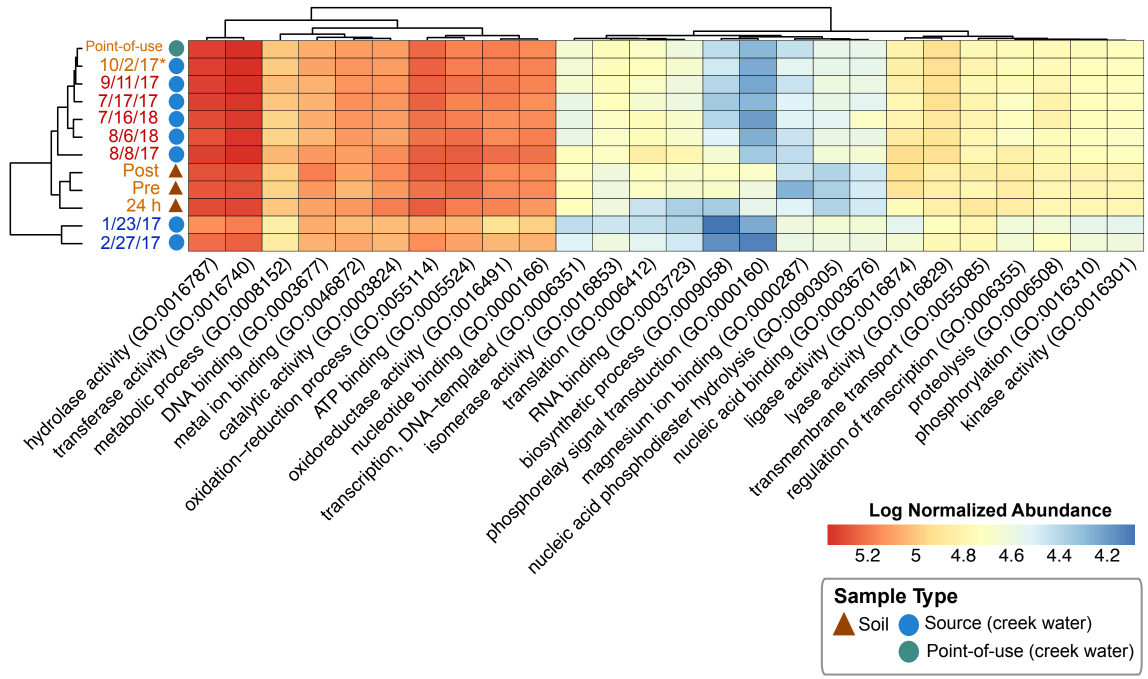

Figure 5.4: Functional composition in creek source water, point-of-use water, and soil. Heatmaps of the normalized abundance assigned to peptide sequences annotated with GO terms for biological and molecular categories. The corresponding GO IDs are presented in parentheses. Note that sequences may be assigned multiple GO terms. Hierarchical clustering of samples was performed using the complete clustering method with Euclidean distances. Sample labels are colored by season (winter: blue, summer: red, autumn: brown) collected. *denotes the date the point-of-use water and soil samples were also collected.

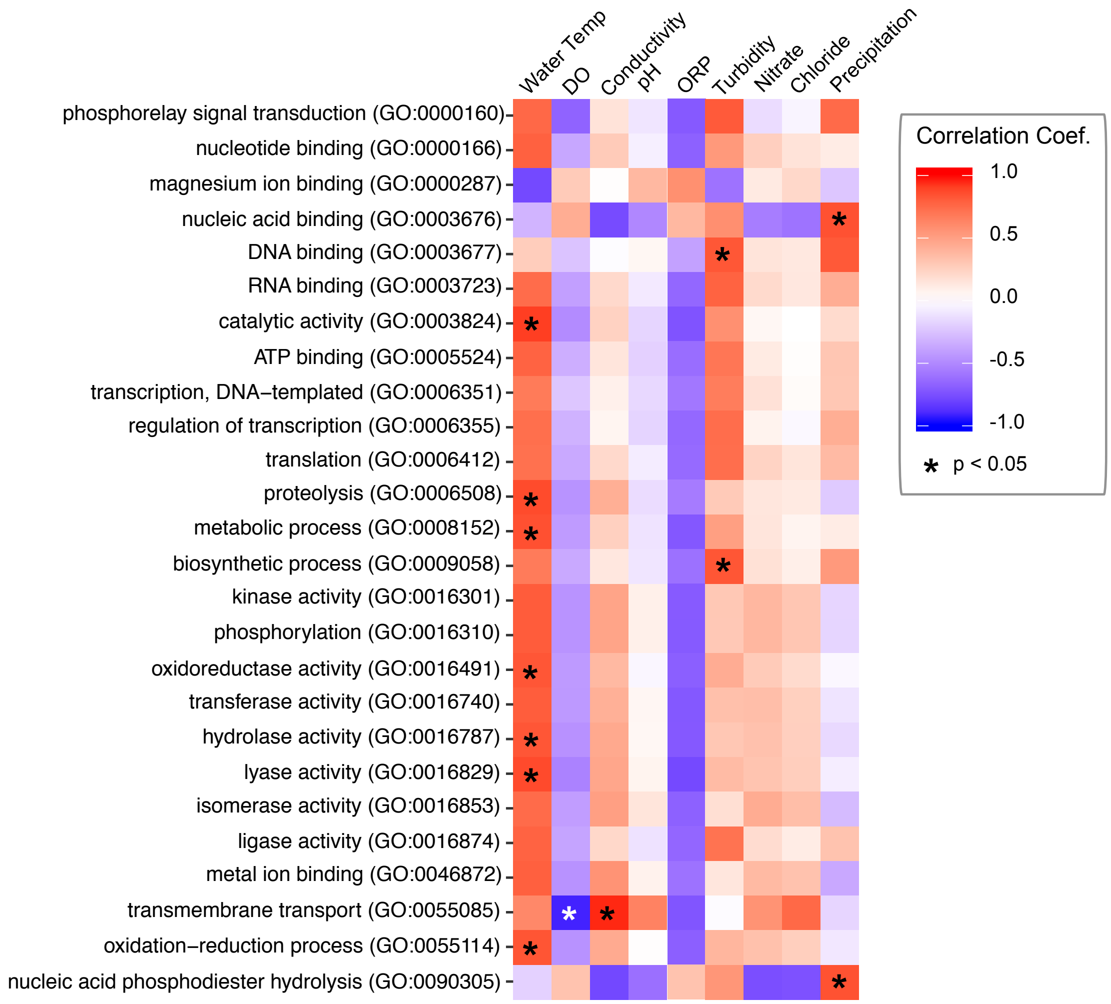

Figure 5.5: Heatmap of the Pearson's correlation coefficients between the water characteristics and normalized abundance of functional GO-terms. Color gradients reflect the different values of Pearson's correlation coefficients. ORP: Oxidation/reduction (mV), DO: Dissolved Oxygen (%).

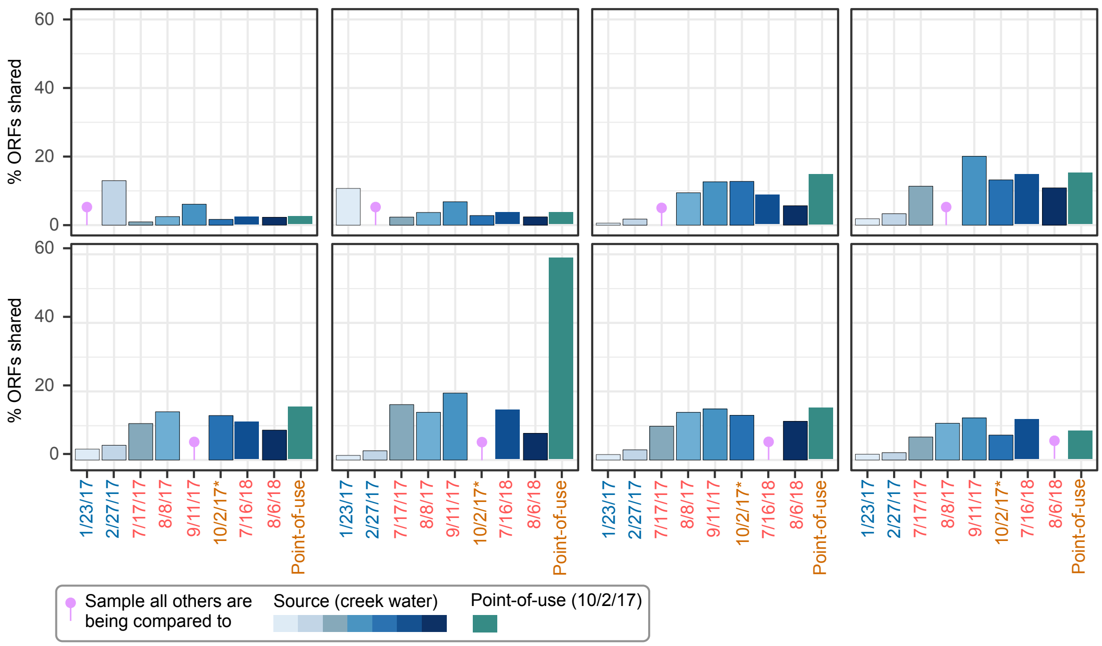

Figure 5.6: Shared ORFs at each time point in creek water samples. Bar plots representing the percentage of ORFs shared between water samples collected at each date and the reference water sample collected on 9/11/17, denoted by the purple pin. Creek water samples are depicted in blue shades and the point-of-use water sample (collected on 10/2/2017) is depicted in teal. Sample labels are colored by season (winter: blue, summer: red, autumn: brown) collected. *denotes the date the point-of-use water and soil samples were also collected.

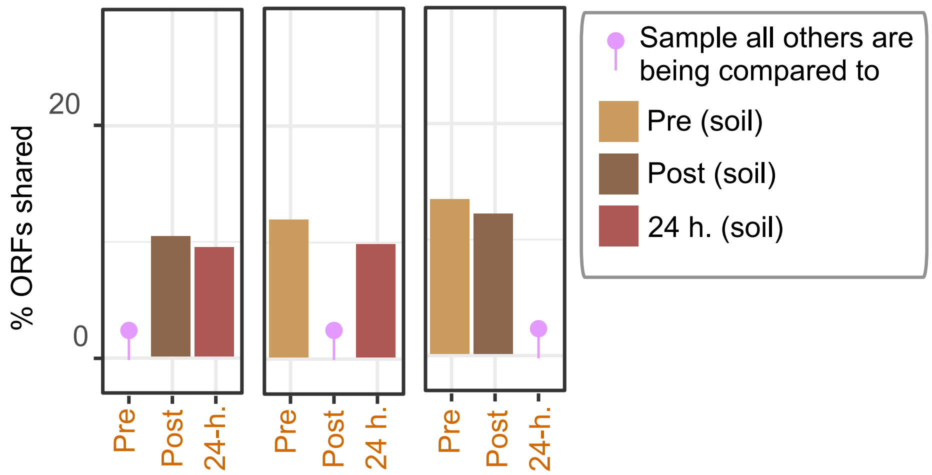

Figure 5.7: Shared ORFs at each time point in soil samples. Bar plots representing the percentage of ORFs shared between soil samples at each stage of irrigation, denoted by the purple pin.

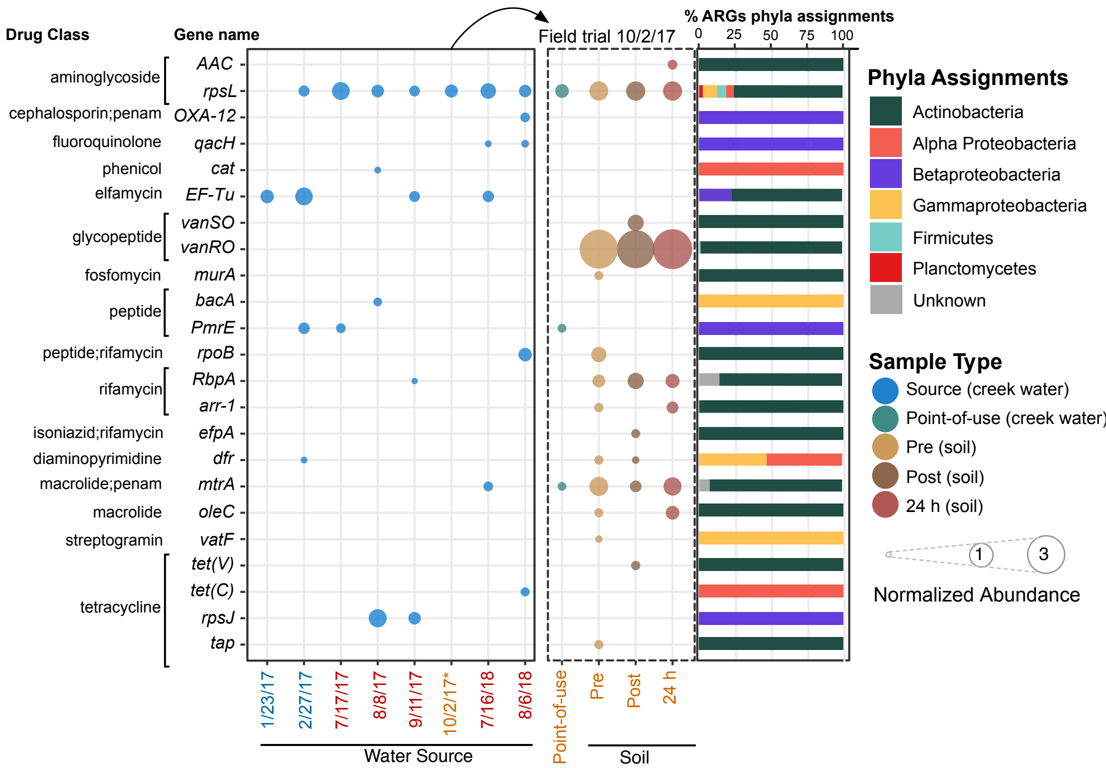

Figure 5.8: Antibiotic resistance genes (ARGs) predicted in creek source water, point-of-use water, and soil. Dotplot of the ARGs present in creek water collected at the source and at the point-of-use, as well as soil samples pre-irrigation, immediately post-irrigation, and 24 h post irrigation. The size of each dot is equivalent to the normalized abundance with homology to each ARG listed on the y-axis, and the color representative of the sample type. Samples are organized temporally and separated by sample type. Bar plot to the right of the dotplot shows the proportion of the normalized abundance for each ARG assigned to a phylum. Sample labels are colored by season (winter: blue, summer: red, autumn: brown) collected. *denotes the date the point-of-use water and soil samples were also collected.

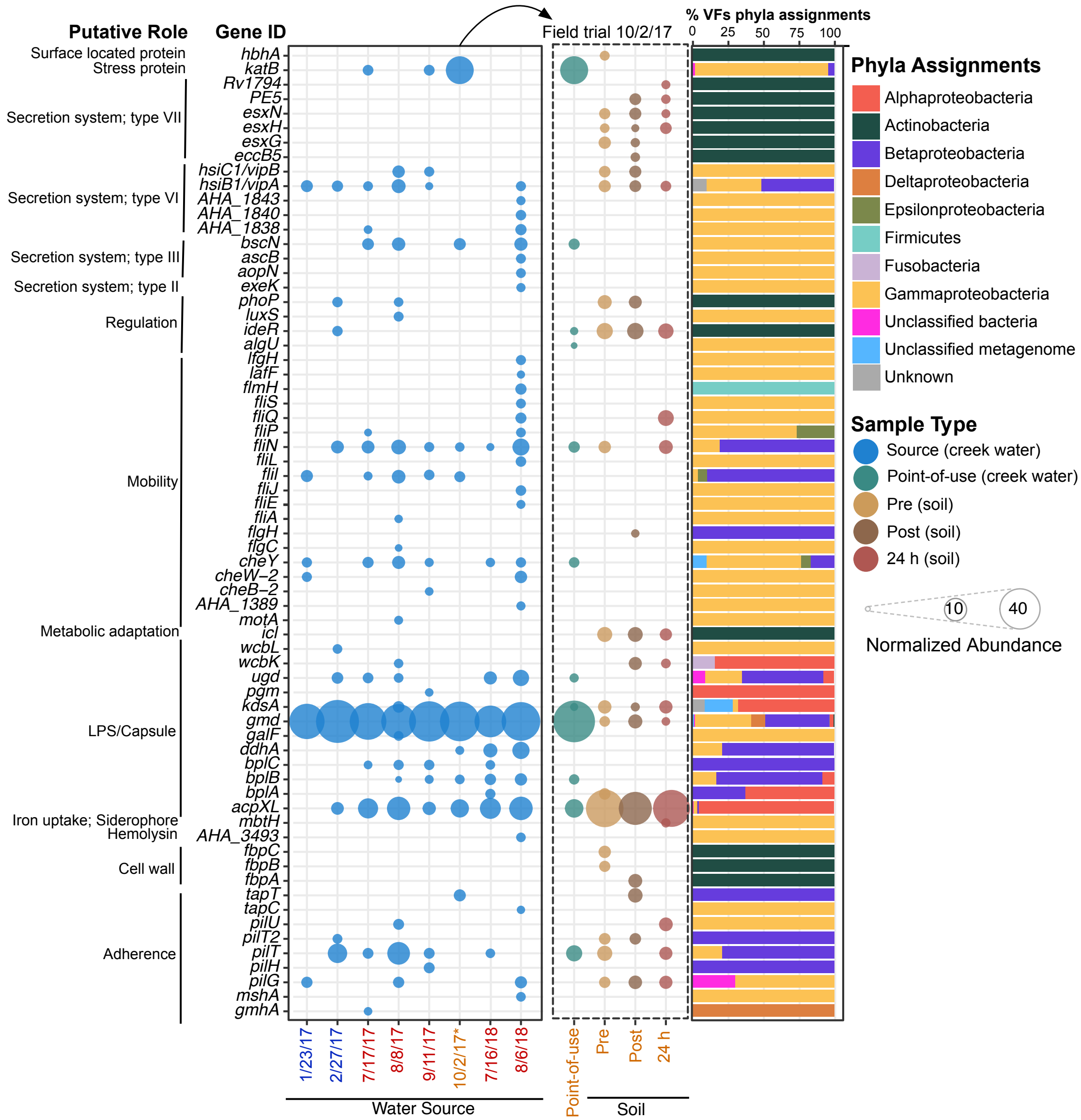

Figure 5.9: Virulence factors (VFs) predicted in the creek water source, point-of-use, and soil samples. Dotplot of the VFs present in creek water collected at the source and at the point-of-use, as well as soil samples collected pre-irrigation, immediately post-irrigation, and 24 h post irrigation. The size of each dot is equivalent to the normalized abundance with homology to each VF listed on the y-axis, and the color is representative of the sample type. Samples are organized temporally and separated by sample type. Bar plot to the right of the dotplot shows the proportion of the normalized abundance for each VF assigned to a phylum. Sample labels are colored by season (winter: blue, summer: red, autumn: brown) collected. *denotes the date the point-of-use water and soil samples were also collected.

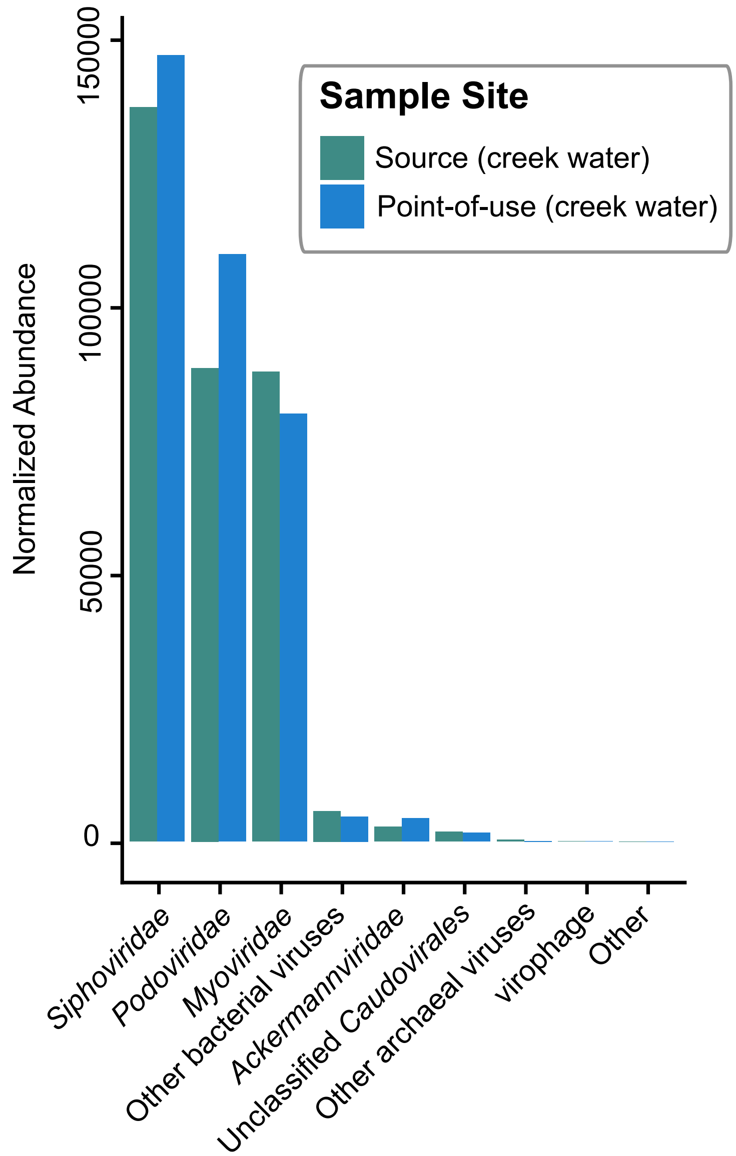

Figure 5.10: Taxonomic composition of viral communities from creek water sampled at the source and at the point-of-use. Grouped bar charts of the relative abundances of the taxonomic assignments for source and point-of-use viromes. Creek water samples are depicted in blue and the point-of-use water sample (on 10/2/2017) is depicted in teal.

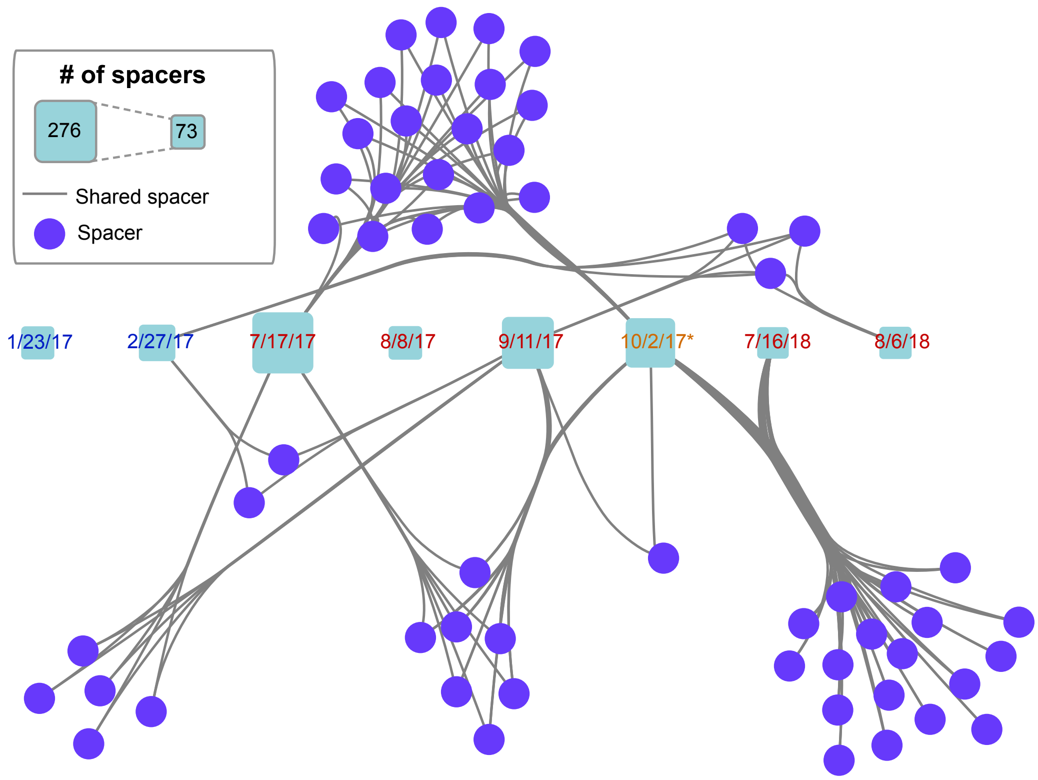

Figure 5.11: CRISPR spacer persistence in the creek source water. Network of shared spacers (97% identity) among samples collected at the creek source. Square-shaped centroid nodes represent each of the eight sampling dates, with the size equivalent to the number of spacers within that site. Nodes connecting the centroids represent shared spacers between and among sites.

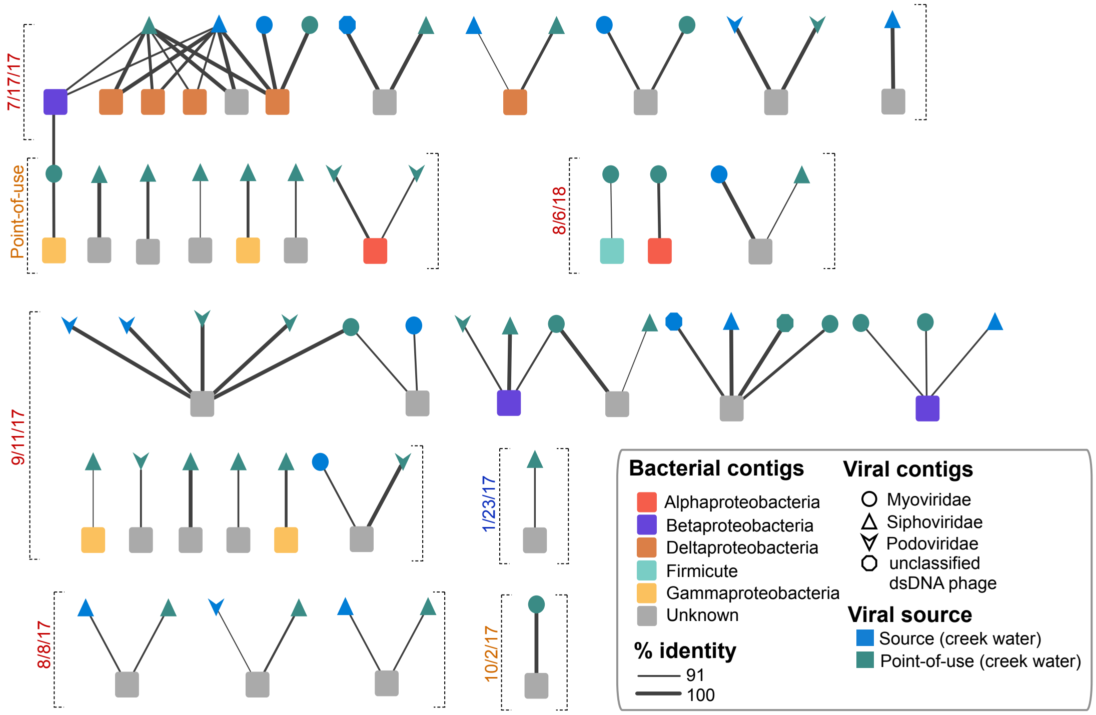

Figure 5.12: Phage-host network from creek water samples collected at the source and the point-of-use. Edges link bacterial contigs (square-shaped nodes), colored by taxonomic representative, and viral contigs by a BLAST matched from at least one spacer. Shape of viral contigs denotes the assigned family. Edge width corresponds to the % identify between a CRISPR spacer present in a contig from the microbial metagenome and a viral contig from either the source (blue) or point-of-use (teal) viromes.

5.7 Tables

Table 5.1: Water physicochemical characteristics.

	Creek Source Water								Point-of-use
Parameter	**1/23/17**	**2/27/17**	**7/17/17**	**8/8/17**	**9/11/17**	**10/2/17**	**7/16/18**	**8/6/18**	**10/2/17**
Ambient Temp. (°C)	7.77	10.5	28.89	21.67	20.1	20.13	31.1	30.5	21.61
Water Temp. (°C)	8.33	6.63	23.55	21.03	15.73	14.8	22.73	22.15	17.83
DO (%)	92.2	101.17	81.83	88.9	96.6	86.97	95.6	93.17	107.03
Conductivity (SPC uS/cm)	305.5	301.57	579.67	322.37	266.37	614.33	205.07	163.05	618
ORP (mv)	163.17	231.63	105.63	114.47	136.57	90.13	87.97	79.73	121.13
Turbidity (FNU)	3.7	3.8	6.47	27	3.2	3.4	0.2	5.47	3.77
Nitrate (mg/L)	0.55	0.7	0.69	0.59	0.59	1.23	0.23	0.38	1.87
Chloride (mg/L)	51.37	57.14	81.15	45.31	30.39	146.98	0.09	0.64	79.79
pH	8	7.64	7.89	7.49	7.6	8.71	7.47	7.36	7.98
Precipitation[a]	0.66	0	0	1.21	0	0	0	0	0

[a]Precipitation 24 h prior to sampling

Table 5.2: Sequencing effort and assembly characteristics for water and soil microbial metagenomes.

Sample Description	Date	no. Read Pairs	no. Contigs	Mean Contig Size	Median Contig Size	Max Contig Size	% GC
Water, Source	1/23/17	58,787,323	793,611	487	389	202,534	47
	2/27/17	60,025,856	918,110	530	405	118,474	48
	7/17/17	65,939,596	793,638	727	455	169,756	52
	8/8/17	65,391,811	978,548	585	419	97,872	51
	9/11/17	65,825,464	1,100,483	661	453	100,674	50
	10/2/17	56,919,057	610,662	728	461	286,247	51
	7/16/18	81,430,175	910,012	629	428	174,984	50
	8/6/18	104,274,389	1,194,393	522	404	61,869	50
Water, Point-of-use	10/2/17	79,086,736	813,975	708	450	181,875	51
Soil, Pre Irrigation	10/2/17	74,570,879	1,616,089	467	399	50,087	63
Soil, Post Irrigation	10/2/17	66,728,971	1,455,733	460	395	29,271	61
Soil, 24 h Post Irrigation	10/3/17	57,579,290	1,216,982	453	392	30,478	62

Table 5.3: Sequencing effort and assembly characteristics for viromes.

Sample Description	**Date**	**no. Read Pairs**	**no. Contigs**	**Mean Contig Size**	**Median Contig Size**	**Max Contig Size**	**% GC**
Water Source	10/2/17	67,415,843	271,124	658	435	126,482	54
Water Point-of-use	10/2/17	85,111,540	933,686	637	425	255,328	52

Table 5.4: Taxonomic assignments for contigs from each water and soil microbial metagenome.

Sample Description	Date	no. Contigs Assigned	Bacteria	Archaea	Eukaryota	Virus
Water Source	1/23/17	560927	507389	5597	15300	23604
	2/27/17	698209	651365	5356	13929	19585
	7/17/17	695217	674121	3845	7391	4674
	8/8/17	844211	814306	5532	7075	10555
	9/11/17	957570	922886	5503	6565	15685
	10/2/17	539273	523808	2246	3419	5678
	7/16/18	685689	632411	5219	37443	5680
	8/6/18	996672	954565	7309	9684	16778
Water, Point-of-use	10/2/17	707669	684943	3567	4348	8367
Soil, Pre Irrigation	10/2/17	1341986	1306210	21505	9146	499
Soil, Post Irrigation	10/2/17	1269513	1230868	24640	8965	429
Soil, 24 h Post Irrigation	10/2/17	1068741	1038865	18511	7155	355

Table 5.5: Bacterial genera assignments for contigs with putative ARGs

Gene Name	Sample Type	(no. contigs) Host Genera
AAC	Soil:	(1) *Actinoplanes*
arr-1	Soil:	(1) *Streptomyces*, (1) *Rhodococcus*
bacA	Water (Source):	(1) *Chania*
cat	Water (Source):	(1) *Agrobacterium*
dfr	Water (Source):	(1) *Pseudomonas*
	Soil:	(1) *Pseudomonas*, (1) *Tropicimonas*
EF-Tu	Water (Source):	(1) *Achromobacter*, (3) *Saccharomonospora*
efpA	Soil:	(1) *Mycobacterium*
mtrA	Water (Source):	(1) *Saccharomonospora*
	Water (POU[a]):	(1) *Saccharomonospora*, (1) *Cellulosimicrobium*, (1) Unknown
	Soil:	(9) *Saccharomonospora*
murA	Soil:	(1) *Williamsia*
oleC	Soil:	(3) *Streptomyces*
OXA-12	Water (Source):	(1) *Achromobacter*
PmrE	Water (Source):	(2) *Burkholderia*
	Water (POU):	(1) *Limnohabitans*
qacH	Water (Source):	(2) *Nitrosovibrio*
RbpA	Water (Source):	(1) *Nocardia*
	Soil:	(6) *Nocardia*, (1) *Rhodococcus*, (1) Unknown
rpoB	Water (Source):	(1) *Streptomyces*
	Soil:	(1) *Frankia*
rpsJ	Water (Source):	(2) *Variovorax*
rpsL	Water (Source):	(5) *Ferrimicrobium*, (1) *Stenotrophomonas* (4) *Streptomyces*, (1) *Thermoanaerobacter*, (1) unclassified *Planctomycetes*
	Water (POU):	(1) *Sphingopyxis*, (1) *Streptomyces*
	Soil:	(2) *Ferrimicrobium*, (1) *Frankia*, (1) *Lutimaribacter*, (6) *Streptomyces*, (1) *Tepidibacillus*
tap	Soil:	(1) *Rhodococcus*
tet(C)	Water (Source):	(1) *Caulobacter*
tet(V)	Soil:	(1) *Aeromicrobium*
vanRO	Soil	(48) *Streptomyces*, (1) *Alicyclobacillus*, (1) Unknown
vanSO	Soil:	(1) *Streptomyces*
vatF	Soil:	(1) *Aeromonas*

[a]POU denotes point-of-use

Table 5.6: Virulence factor descriptions for gene IDs.

Gene Name	VFDB Gene Description
AHA_1389	CobQ/CobB/MinD/ParA family protein
AHA_1838	Type VI secretion system protein
AHA_1840	Type VI secretion system protein DotU
AHA_1843	Type VI secretion system protein
AHA_3493	hemolysin III
algU	alginate biosynthesis protein AlgZ/FimS
aopN	secretion control of translocators and immune suppressor
ascB	chaperone protein
bplA	probable oxidoreductase
bplB	probable acetyltransferase
bplC	lipopolysaccharide biosynthesis protein
bscN	ATP synthase in type III secretion system
cheB-2	chemotaxis-specific methylesterase
cheW-2	chemotaxis protein
cheY	chemotaxis protein
ddhA	glucose-1-phosphate cytidylyltransferase
eccB5	ESX-5 type VII secretion system protein
esxG	Type VII secretion system protein
esxH	Type VII secretion system protein ESXH
esxN	ESX-5 type VII secretion system EsxA (ESAT-6) homolog
exeK	general secretion pathway protein K
fbpA	Secreted antigen mycolyl transferase 85A
fbpB	Secreted antigen 85 complex B
fbpC	Secreted antigen 85-C antigen 85 complex C)
flgC	flagellar basal body rod protein
flgH	flagellar basal body L-ring protein
fliA	flagellar biosynthesis sigma factor
fliE	flagellar hook-basal body complex protein
fliI	flagellum-specific ATP synthase
fliJ	flagellar export protein
fliL	flagellar basal body-associated protein
fliL	flagellar basal body-associated protein
fliN	flagellar motor switch protein
fliP	flagellar biosynthesis protein
fliQ	flagellar biosynthetic protein
fliS	flagellar protein
flmH	short chain dehydrogenase/reductase family oxidoreductase
galF	UTP-glucose-1-phosphate uridylyltransferase subunit
gmd	GDP-mannose 4,6-dehydratase
gmhA	phosphoheptose isomerase
hbhA	iron-regulated heparin binding hemagglutinin
icl	Isocitrate lyase
ideR	Iron-dependent repressor and activator
katB	catalase-peroxidase

kdsA	2-dehydro-3-deoxyphosphooctonate aldolase
lafF	lateral flagella
lfgH	lateral flagellar L-ring protein
luxS	S-ribosylhomocysteinase
mbtH	MbtH-like protein from the pyoverdine cluster
motA	flagellar motor protein
mshA	mannose-sensitive hemagglutinin pili minor prepilin protein
PE5	PE family protein
pgm	phosphoglucomutase
phoP	possible two component system response transcriptional positive regulator
pilG	pilus assembly protein
pilH	pilin-like protein may involving in pseudopilus formation
pilT	twitching motility protein
pilT2	twitching motility protein
pilU	twitching motility protein
Rv1794	ESX-5 locus protein
tapC	type IV fimbrial assembly protein
tapT	twitching ATPase
ugd	UDP-glucose 6-dehydrogenase
vipA	type VI secretion system tubule-forming protein
vipB	type VI secretion system tubule-forming protein
wcbK	GDP sugar epimerase/dehydratase
wcbL	sugar kinase

Table 5.7: Bacterial genera assignments for contigs with putative VFs.

Gene Name	**Sample Type**	**(no. contigs) Host Genera**
acpXL	Water (Source):	(1) *Cystobacter*, (1) *Loktanella*, (23) *Lutimaribacter*, (2) *Mesorhizobium*, (2) *Microvirga*, (2) *Rhodopseudomonas*, (1) *Sinorhizobium*, (11) *Sphingomonas*, (6) *Sphingopyxis*, (1) unclassified *Alphaproteobacteria*
	Water (POU[a]):	(1) *Defluviimonas*, (1) *Loktanella*, (6) *Lutimaribacter*, (1) *Sphingopyxis*
	Soil:	(1) *Acinetobacter*, (2) *Aeromonas*, (20) *Lutimaribacter*, (3) *Mesorhizobium*, (13) *Microvirga*, (1) *Pacificibacter*, (9) *Rhodopseudomonas*, (6) *Sinorhizobium*, (16) *Sphingomonas*, (11) *Sphingopyxis*, (1) *Variovorax*, (1) Unknown, (6) unclassified *Alphaproteobacteria*, (1) unclassified Bacteria
AHA 1389	Water (Source):	(1) *Aeromonas*
AHA 1838	Water (Source):	(1) *Chania*, (1) *Aeromonas*
AHA 1840	Water (Source):	(1) *Chania*
AHA 1843	Water (Source):	(1) *Aeromonas*
AHA 3493	Water (Source):	(1) *Aeromonas*
algU	Water (POU):	(1) *Pseudomonas*
aopN	Water (Source):	(1) *Aeromonas*
ascB	Water (Source):	(1) *Aeromonas*
bplA	Water (Source):	(1) *Variovorax*
	Soil:	(2) *Rhodopseudomonas*
bplB	Water (POU):	(1) *Variovorax*
bplC	Water (Source):	(2) *Variovorax*, (1) *Aquitalea*, (1) *Pusillimonas*
bscN	Water (POU):	(1) *Aeromonas*
	Water (Source):	(4) *Aeromonas*, (1) *Succinivibrio*
cheB-2	Water (Source):	(1) *Vibrio*
cheW-2	Water (Source):	(1) *Pseudomonas*, (1) *Aeromonas*, (1) *Gallaecimonas*
cheY	Water (Source):	(1) *Aliiarcobacter*, (1) *Chania*, (1) *Methylophaga*, (3) *Rahnella*, (1) *Succinivibrio*, (1) Unclassified Bacteria, (2) unclassified *Betaproteobacteria*
	Water (POU):	(2) *Chania*, (1) *Pseudomonas*
ddhA	Water (Source):	(5) *Achromobacter*, (2) *Aeromonas*, (1) *Aquitalea*, (1) *Pusillimonas*
eccB5	Soil:	(1) *Mycolicibacterium*
esxG	Soil:	(3) *Mycolicibacterium*

esxH	Soil:	(2) *Mycobacterium*, (1) *Mycolicibacterium*
esxN	Soil:	(3) *Mycolicibacterium*, (1) *Mycobacterium*
exeK	Water (Source):	(1) *Aeromonas*
fbpA	Soil:	(1) *Mycolicibacterium*
fbpB	Soil:	(1) *Mycobacterium*
fbpC	Soil:	(2) *Mycolicibacterium*
flgC	Water (Source):	(1) *Aeromonas*
flgH	Soil:	(1) *Paraburkholderia*
fliA	Water (Source):	(1) *Rahnella*
fliE	Water (Source):	(1) *Aeromonas*
fliI	Water (Source):	(2) *Nitrosovibrio*, (1) *Delftia*, (1) *Pseudomonas*, (1) *Massilia*, (1) *Campylobacter*
fliJ	Water (Source):	(1) *Aeromonas*
fliL	Water (Source):	(1) *Aeromonas*
fliN	Water (Source):	(1) *Massilia*, (3) *Nitrosovibrio*, (3) *Pseudomonas*, (1) *Stenotrophomonas*, (1) *Succinivibrio*, (3) unclassified *Betaproteobacteria*
	Water (POU):	(1) *Nitrosovibrio*, (1) *Pseudomonas*, (1) unclassified *Betaproteobacteria*
	Soil:	(4) *Nitrosovibrio*
fliP	Water (Source):	(1) *Campylobacter*, (1) *Aeromonas*
fliQ	Water (Source):	(1) *Aeromonas*
	Soil:	(1) *Pseudomonas*
fliS	Water (Source):	(1) *Rahnella*
flmH	Water (Source):	(2) *Caldicellulosiruptor*
galF	Water (Source):	(1) *Kosakonia*
gmd	Water (Source):	(4) *Achromobacter*, (5) *Defluviimonas*, (16) *Delftia*, (16) *Desulfotignum*, (1) *Kitasatospora*, (1) *Marichromatium*, (2) *Methylophaga*, (11) *Nitrosovibrio*, (1) *Parapedobacter*, (89) *Polynucleobacter*, (53) *Pseudomonas*, (1) *Sphingopyxis*, (5) *Stenotrophomonas*, (2) unclassified Bacteria, (4) *Variovorax*, (1) *Xanthobacter*, (1) unclassified *Alphaproteobacteria*, (2) unclassified Bacteria, (3) unclassified *Betaproteobacteria*
	Water (POU):	(1) *Celeribacter*, (3) *Delftia*, (5) *Desulfotignum*, (1) *Limnohabitans*, (14) *Polynucleobacter*, (9) *Pseudomonas*, (1) *Rudanella*, (1) unclassified Bacteria
	Soil:	(1) *Delftia*, (1) *Pseudomonas*, (1) Marinobacter
gmhA	Water (Source):	(1) *Desulfovibrio*
hbhA	Soil:	(1) *Nocardia*

icl	Soil:	(2) *Nocardia*, (2) *Pseudarthrobacter*
ideR	Water (Source):	(1) *Kitasatospora*
	Water (POU):	(1) *Nocardia*
	Soil:	(4) *Nocardia*, (1) *Saccharomonospora*, (3) *Streptomyces*, (1) *Cellulosimicrobium*
katB	Water (POU):	(1) *Pseudomonas*, (1) unclassified *Betaproteobacteria*
	Water (Source):	(1) *Curvibacter*, (1) unclassified Bacteria, (1) *Allochromatium*, (1) *Pseudomonas*
kdsA	Water (Source):	(1) unclassified metagenome
	Water (POU):	(1) Mannheimia
	Soil:	(5) *Microvirga*, (1) Unknown, (1) *Mesorhizobium*
lafF	Water (Source):	(1) *Aeromonas*
lfgH	Water (Source):	(1) *Pseudomonas*
luxS	Water (Source):	(1) *Chania*
mbtH	Soil:	(1) *Pseudomonas*
motA	Water (Source):	(1) *Chania*
mshA	Water (Source):	(1) *Aeromonas*
PE5	Soil:	(2) *Mycobacteroides*, (1) *Mycobacterium*
pgm	Water (Source):	(1) *Sphingomonas*
phoP	Water (Source):	(2) *Nocardia*
	Soil:	(4) *Nocardia*
pilG	Water (Source):	(1) *Pseudomonas*
	Soil:	(3) *Pseudomonas*, (3) unclassified Bacteria
pilH	Water (Source):	(1) *Variovorax*
pilT	Water (Source):	(2) *Acinetobacter*, (1) *Methylophaga*, (9) unclassified *Betaproteobacteria*
	Water (POU):	(3) unclassified *Betaproteobacteria*
	Soil:	(1) *Acinetobacter*, (2) unclassified *Betaproteobacteria*
pilT2	Water (Source):	(1) unclassified *Betaproteobacteria*
	Soil:	(1) unclassified *Betaproteobacteria*, (1) *Variovorax*
pilU	Water (Source):	(1) *Pseudomonas*
	Soil:	(1) *Pseudomonas*
Rv1794	Soil:	(1) *Mycobacterium*
tapC	Water (Source):	(1) *Aeromonas*
tapT	Water (Source):	(1) unclassified *Betaproteobacteria*
	Soil:	(1) unclassified *Betaproteobacteria*
ugd	Water (Source):	(1) *Achromobacter*, (3) *Burkholderia*, (2) *Gallaecimonas*, (1) *Sphingopyxis*, (1) unclassified Bacteria
	Water (POU):	(1) *Limnohabitans*
vipA	Water (Source):	(4) *Variovorax*, (1) *Chania*, (1) *Pseudomonas*

	Soil:	(2) *Pseudomonas*, (1) *Chania*, (1) *Variovorax*, (1) Unknown
vipB	Water (Source):	(2) *Pseudomonas*
	Soil:	(2) *Pseudomonas*
wcbK	Water (Source):	(1) *Fusobacterium*
	Soil:	(2) unclassified *Rhodospirillales*
wcbL	Water (Source):	(1) *Allochromatium*

[a]POU denotes point-of-use

Table 5.8: CRISPR array abundance in soil and water microbial metagenomes.

Sample Description	Date	no. Contigs Assigned	Contigs Abundance	no. Spacers	no. Unique Spacers
Water Source	1/23/17	33	54.873	103	84
	2/27/17	33	22.8259	122	107
	7/17/17	49	56.8421	282	276
	8/8/17	30	21.8518	97	87
	9/11/17	51	40.2514	221	213
	10/2/17	31	176.426	207	196
	7/16/18	20	12.6793	79	73
	8/6/18	29	14.0952	90	78
Water, Point-of-use	10/2/17	64	171.13	349	333
Soil, Pre Irrigation	10/2/17	10	6.59726	34	30
Soil, Post Irrigation	10/2/17	12	3.23286	36	30
Soil, 24 h Post Irrigation	10/2/17	8	6.36801	25	21

Table 5.9: CRISPR spacers shared with water samples collected at the point-of-use.

Cluster Representatives	**no. Spacers**
7/17/17	2
7/17/17 and 10/2/17	23
7/16/18 and 10/2/17	18
9/11/17 and 10/2/17	1
7/17/17, 9/11/17, and 10/2/17	9
10/2/17	89
7/16/18	4

Chapter 6: Zero-valent Iron Sand Filtration Reduces Concentrations of Virus-like Particles and Modifies Virome Community Composition in Reclaimed Water Used for Agricultural Irrigation

6.1 Abstract

Objective: Zero-valent iron (ZVI) sand filtration can remove a broad range of contaminants, including some types of pathogenic bacteria, from contaminated water. However, its efficacy at removing complex viral populations, such as those found in reclaimed water used for agricultural irrigation, has not been fully evaluated. Therefore, the objective of this study was to use metagenomic sequencing and epifluorescent microscopy to enumerate and characterize the viral populations found in reclaimed water and ZVI-sand filtered reclaimed water sampled three times over 49 days from a larger ongoing study. **Results**: ZVI-sand filtered reclaimed water samples had significantly less virus-like particles than reclaimed water samples at all collection dates, with the reclaimed water averaging between 10^9-10^8 and the ZVI-sand filtered reclaimed water averaging between 10^6-10^7 virus-like particles per mL. In addition, for both sample types viral metagenomes (viromes) were dominated by

bacteriophage of the order *Caudovirales*, largely *Siphoviridae*, and genes related to DNA metabolism. However, the proportion of sequences homologous to bacteria, as well as the abundance of genes possibly originating from a bacterial host, was higher in ZW viromes. Overall, ZVI-sand filtered reclaimed water had a lower total concentration of virus-like particles and a different virome community composition compared to unfiltered reclaimed water.

6.2 Introduction

The intense use of groundwater resources for agricultural irrigation and other activities continues to overstress aquifers and has led to substantial groundwater depletions globally [8, 9]. Consequently, demand has grown for the development of technologies, such as zero-valent iron (ZVI) sand filtration, to treat alternative irrigation water sources (e.g. reclaimed water, return flows) and allow for their use. ZVI-sand filters, which were initially designed to remove chlorinated compounds from groundwater supplies, are composed of mixtures of sand and zero-valent iron [394]. Currently, they are being developed and utilized to remove or inactivate a broad range of contaminants, including microorganisms, from multiple water sources [395–401]. Specifically, ZVI has reduced *Escherichia coli* populations in water, likely due to the physical disruption of the cell membranes caused by reactive oxygen species produced by the interaction between ZVI (Fe^0) particles and water molecules during filtration [398, 399]. For viruses, ZVI has been shown to reduce titers of Aichi virus, Murine norovirus, Tulane virus, and bacteriophage MS2 and ΦX174

[400, 402]. However, ZVI has not been evaluated on its ability to remove complex viral populations, such as those found in reclaimed water.

In reclaimed water, virus-like particles (VLPs) are estimated to be 1000-fold higher than in potable water, with the majority of these viruses showing homology to bacteriophage [114]. Bacteriophage are among the most abundant biological entities on earth and are critical components in food web-dynamics and nutrient cycling [143]. Moreover, phage can influence its bacterial host's phenotype through the horizontal transfer of genes, such as those coding for antibiotic resistance determinants and toxins [53, 54, 56, 403]. This is potentially of concern for wastewater treatment plants, which are reported to be reservoirs for antibiotic resistance genes [404]. Therefore, this study aimed to characterize and quantify the DNA viruses in reclaimed water and ZVI-sand filtered reclaimed water collected during a larger greenhouse study that assessed the overall effectiveness of ZVI-sand filtration in treating reclaimed water used to irrigate lettuce.

6.3 Materials and Methods

6.3.1 Sample collection

Samples were collected as part of a larger greenhouse study that evaluated the impacts of ZVI-sand filtration on levels of multiple antimicrobials, *E. coli* and total bacterial communities in conventionally-treated reclaimed water used to irrigate lettuce [405]. Briefly, in the summer of 2016, 240 L of chlorinated effluent was collected every two weeks from a tertiary wastewater treatment plant in the Mid-

Atlantic United States, which processes between 1,136 and 1,420 m^3 of domestic wastewater daily from a rural area. Incoming raw wastewater is treated through grinding, (for removal of large debris) activated sludge processing, and secondary clarification, and is then stored in an open-air lagoon where it is chlorinated prior to land application.

After collection, the reclaimed water was delivered to the University of Maryland Research Greenhouse Complex, where it was stored in multiple 189 L storage barrels (Algreen Products Inc., Ontario, Canada) prior to ZVI-sand filtration.

6.3.2 ZVI-Sand filter and filtration process

A commercially available biosand filter (HydrAid BioSand Water Filter, NativeEnergy, Burlington, VT, USA) was modified for use in this experiment, and has been previously described in detail [405].

Briefly, the filter vessel is made of opaque plastic and has a total volume of ~55.5 L. Fine filtration sand, provided with the filter [406, 407], and ZVI particles (Peerless Metal Powders and Abrasives Company, Detroit, MI) were sieved (resulting in a particle size range of 400 μm to 625 μm) and mixed in equal proportions, generating a 50:50 volume/volume mixture. The ZVI-sand filter was then established in two steps: 1) the empty plastic vessel was filled with 20 L ultrapure water; and 2) the ZVI-sand mixture was added to the vessel, displacing the water. During the filtration events (described below), gravel filled diffuser plate (NativeEnergy, Burlington VT, USA) was then used to pour reclaimed water into the filter, preventing preferential flow. The porosity of the filter was approximately 0.52 [408], the average volumetric flow rate was ~5.6L/min, the filtration rate was 118 L/min/m^2 and the approximate ZVI contact time was 2.58 minutes [405].

To mimic the applied use of sand filters in agricultural settings like the Mid-Atlantic, United States, where irrigation water is not needed every day due to periodic rain events, reclaimed water was filtered through our ZVI-sand filter every five days during the larger greenhouse study. During each filtration event, a 20 L composite of the stored reclaimed water was generated from the storage barrels and then gravity filtered through the ZVI and filter to accommodate the irrigation needs of the greenhouse study. To maintain the ZVI-sand filter between filtration events and mimic a real-life agricultural scenario, the filter was kept submerged in reclaimed water, and right before filtration, the five-day old water within the filter was completely flushed out by pouring 20 L reclaimed water through the filter and discarding it. A new 20 L composite reclaimed water sample was then obtained from the storage barrels and poured completely through the filter. From the total ~20 L ZVI filtrate, a 1 L subsample of filtrate (ZW sample) was collected for analysis along with 1 L of unfiltered reclaimed water (RW samples). ZW and RW samples were collected once a month for the present study on 6/21/2016, 7/30/2016, and 8/9/2016, and then subjected to direct viral counts and DNA extraction for sequencing (detailed below).

6.3.3 VLP quantification

Viral enumeration was performed using the filter mount method adapted from existing protocols [409]. Briefly, aliquots (1 μL for RW and 100 μL for ZW) of formalin fixed samples were suspended in sterile deionized water (total volume of 1000 μL), vacuum filtered onto a 13 mm 0.02-μm Anodisc filter (Whatman, USA), and stained with SYBR Gold (Thermo Fisher Scientific, USA). Triplicate slides for each sample were made and counted within 24 hours with an Olympus BX61 microscope (20 random fields, 1000X magnification). VLPs were quantified with iVision software and a paired t-test with Bonferroni correction was performed to test for differences in VLP counts between the RW and ZW samples at each date.

6.3.4 Virome preparation

Each sample (1L) was vacuum filtered through a 0.2 μm membrane filter (Pall Corporation, Port Washington, NY) to remove the cellular fraction and collected in sterile receiving flasks. Viral particles were then concentrated using a chemical flocculation method where 100 μL of a 4.83 gL^{-1} $FeCl_3$ solution was added to each filtrate and incubated for one hour in the dark and then filtered onto a 25-mm 0.2 μm membrane filter [240]. For resuspension filters were rocked overnight at 4 °C in 1 mL of 0.1M EDTA-0.2M $MgCl_2$-0.2M ascorbate buffer. The resulting released viral particles were then subjected to a two-hour treatment with DNase I (Sigma-Aldrich, MO, USA) and filtered through a 33-mm diameter sterile syringe filter with a 0.2 μm pore size (Millipore Corp., MA, USA). DNA was then extracted from 500 μL of the treated viral concentrates via the AllPrep DNA/RNA Mini Kit (Qiagen, CA, USA).

6.3.5 Virome DNA sequencing

For shotgun DNA sequencing, per the modified Nextera XT protocol, each of the viral DNA extracts were used in a tagmentation reaction, followed by 13 cycles of PCR amplification with the Nextera i7 and i5 index primers and 2X Kapa master mix. The resulting DNA libraries were then sequenced on the Illumina HiSeq 4000 platform (Illumina, San Diego, CA, USA).

6.3.6 Metagenome assembly and analysis

Viromes were assembled as described in detail previously [157]. Briefly, paired-end reads were trimmed, merged, and *de novo* assembled using Trimmomatic ver. 0.36 (slidingwindow:4:30 minlen:60) [175], FLASh ver.1.2.11 [176], and metaSPAdes ver. 3.10.1 (without read error correction) [177], respectively. MetaGene was then used to predict open reading frames (ORFs) from each assembly [104]. Contigs were queried against the peptide SEED and Phage SEED databases (retrieved 11/17/2017) using protein-protein BLAST (BLASTp ver. 2.6.0+) (E value $\leq$ 1e^{-3}) to assigned taxonomic and functional classifications [105, 256].

Coverage was calculated for each contig by: (i) recruiting quality-controlled reads to assembled contigs using Bowtie2 ver. 2.3.3 (very sensitive local mode), (ii) processing the BAM file for artificial duplicates using Picard, and then (iii) using the "depth" function of Samtools ver. 1.4.1 to compute the per-contig coverage [180,257]. To normalize abundances across libraries, contig and ORF coverages were divided by the sum of coverage per million, similar to the transcripts per million (TPM) metric used in RNA-Seq [181]. Scripts performing these assignments and normalization are available at github.com/dnasko/baby_virome. Taxonomic and functional data were visualized using ggplot2 and heatplus [183, 246].

6.4 Results

6.4.1 VLP abundance

VLP counts from RW and ZW samples were compared at each sampling date (June, July, and August). At each date the VLPs were significantly ($p \leq 0.05$) less abundant in the ZW samples compared to the RW samples (Figure 6.1). RW samples contained an average of 1.6×10^9, 6.7×10^8, and 7.0×10^8 VLPs mL^{-1} in June, July, and August, respectively. The ZW samples contained an average of 8.6×10^6, 2.8×10^7, and 4.2×10^7 VLPs mL^{-1} in June, July, and August, respectively.

6.4.2 Sequencing effort and assembly

Viral DNA was extracted from the six samples; however, it was not possible to obtain enough DNA from the June ZW sample for shotgun sequencing. The remaining samples were sequenced on the Illumina HiSeq for a total of 136,267,357 read pairs (Table 6.1), with an average of 27,253,471 read pairs per virome (±3,234,104 SD). Metagenomic assembly produced a total of 825,658 contigs, with an average of

165,132 contigs (±30,305 SD) and 278,196 ORFs (±63,500 SD) per virome.

6.4.3 ORF clusters

To assess the percentage of functional similarity between RW and ZW viromes ORF peptides originating from the same sampling dates (July and August) were clustered using CD-HIT (60% peptide similarity) [184]. In July, 42% of the RW peptide ORFs clustered with 68% of the ZW peptide ORFs. For August the percentage of shared function increased for the reclaimed sample; 60% of the RW peptide ORFs clustered with 61% of the ZW peptide ORFs (Figure 6.2).

6.4.4 Taxonomic assignment

Similar to other virome studies [114], between 32-38% of contigs could be assigned a taxonomic representative (Table 6.2). For the contigs that could be identified, a normalized abundance was calculated. Both the RW and ZW viromes were dominated by sequences homologous to viral phyla (51-67%), followed by bacterial (11-29%) and unknown (17-23%). However, the proportion of bacteria-assigned contigs was greater in the ZW viromes (~29%) than the RW (11-17%) (Figure 6.3).

The most abundant viral taxonomic classifications for each virome (~98% of all viral classified taxa) belonged to the dsDNA phage of the order *Caudovirales* (Figure 6.3), largely *Siphoviridae* (51-54%), followed by *Myoviridae* (28-31%), and *Podoviridae* (13-16%). The remaining ~ 2% of viral sequences were assigned as unclassified phage and viruses infecting archaea, amoeba, plants, or vertebrates.

6.4.5 Functional assignment

Peptide ORFs from all contigs were functionally annotated using SEED Subsystems [256]. Of those assigned, a normalized abundance was calcuated (Figure 6.4). The majority of functional assignments were classified as DNA metabolism (20-30%), followed by phage elements (11-17%), and protein metabolism (8-10%). Annotated SEED Subsystem assignments were parsed for those assigned as resistant to antibiotics and toxic compounds, which were only between 1-2% of the assignments. Among the antibiotic and toxic compound annotations, genes for Beta-lactamase were dominant in both of the August viromes. Additionally, the ZW viromes had a greater normalized abundance than all of the RW viromes for: cobalt-zinc-cadmium resistance, copper homeostasis, multidrug resistance efflux pumps, fluoroquinolone resistance, methicillin resistance in staphylococci, zinc resistance, mercuric reductase, and mercury resistance operon (Figure 6.4).

6.5 Discussion

Reclaimed water is an important emerging resource that can help alleviate stress on surface and groundwater systems and is already being implemented in a number of potable and non-potable applications (e.g. agricultural irrigation) [231, 410]. However, concerns remain about the levels of microbial and chemical constituents that may persist in reclaimed water and whether treatment technologies can be used for further remediation. In this study, we found that reclaimed water harbors 10^8-10^9 VLPs mL^{-1}, similar to the abundances published in previous studies on VLPs in reclaimed water [113,114]. After ZVI-sand filtration the number of VLPs was significantly lower at all sampling dates, ranging between 10^6 and 10^7 VLPs mL^{-1}, a roughly 1-2 log reduction. Previous studies have suggested that virus removal from water during ZVI-sand filtration is likely attributed to adsorption and inactivation via iron oxides within the iron [397,411]. You et al., 2005 posited that over time, as water flows through a ZVI-sand filter, new iron oxides are formed continuously, generating additional adsorption sites that could extend the life of the filter [397].

Our findings are similar to recent results on the reduction efficiency of sand filtration alone for viruses ΦX174, MS2, and AiV (<1-2 log) and lower than previous studies on ZVI-sand filtration, which reported that ΦX174 and MS-2 were reduced by 4-6 logs [397, 400]. However, these studies focus largely on the removal of a few specific viral species and, even so, have found that removal efficiencies vary among species [400]. Here, we used epifluorescence microscopy to look at the entire viral population. This includes hundreds to thousands of different species, with a range of capsid sizes and isoelectric points, which may help explain the smaller removal efficiency [412]. Moreover, while the log-reduction is lower than expected for the overall population, the total VLP concentration post ZVI filtration is still comparable to those described in well and potable water [114, 413].

In terms of viral taxonomic composition, both RW and ZW viromes were dominated by *Siphoviridae*, which are known to be abundant in human waste and reclaimed water [114, 117]. These viruses present a unique risk, as the majority of cultured representatives are capable of lysogeny and, thus, may facilitate horizontal gene transfer among bacteria [276]. Additionally, in both sample types the functional profiles were largely composed of genes related to DNA metabolism (Figure 6.4). A previous study that characterized DNA viruses from a wastewater treatment plant also found these genes to be highly abundant and attributed this to the elevated metabolic activity within treatment plants [113].

While the viral composition at the taxa level was similar between the two samples types, there were some differences that may be attributed to the filtration process. For instance, between 39-32% of peptide ORFs from the ZW virome did not cluster with any peptide ORFs from the RW virome collected on the same date (Figure 6.2). This may be due to the changing microbial ecosystem within the biosand, which may give rise to new functional potential as it becomes established [414]. In addition, ZW viromes had a greater relative abundance of contigs homologous to bacteria. In virome studies, sequences with significant homology to bacteria are sometimes unknown prophage embedded in a bacterial genome present in the database, or phage carrying host genes [114, 415]. During use, ZVI produces reactive oxygen species, which can promote prophage induction [416, 417]. It could be suggested that ZVI-sand filtration may stimulate the induction of integrated prophage present within the bacteria passed through the filter. However, it is important to note that, while the abundance of some genes is higher in the ZW viromes, the overall number of gene copies is likely still higher in the RW sample due to the increased number of total VLPs. Therefore, additional work is necessary to determine whether the bacterial assigned contigs are indeed prophage and whether this may have an impact on the dissemination of bacterial genes in water reuse systems.

6.5.1 Limitations

Our study was limited in the number of samples (n=6). Therefore, a rigorous statistical analysis could not be performed. In addition, we could not obtain enough DNA from the June ZW water sample for shotgun sequencing and, therefore, a comparison between RW and ZW water for June was not possible. Finally, because we were not able to include a second sand-only filter control, due to the set-up of the larger greenhouse based study, within which the present study was designed, we were unable to tease out the individual effects of ZVI versus sand in terms of virus removal.

6.6 Figures

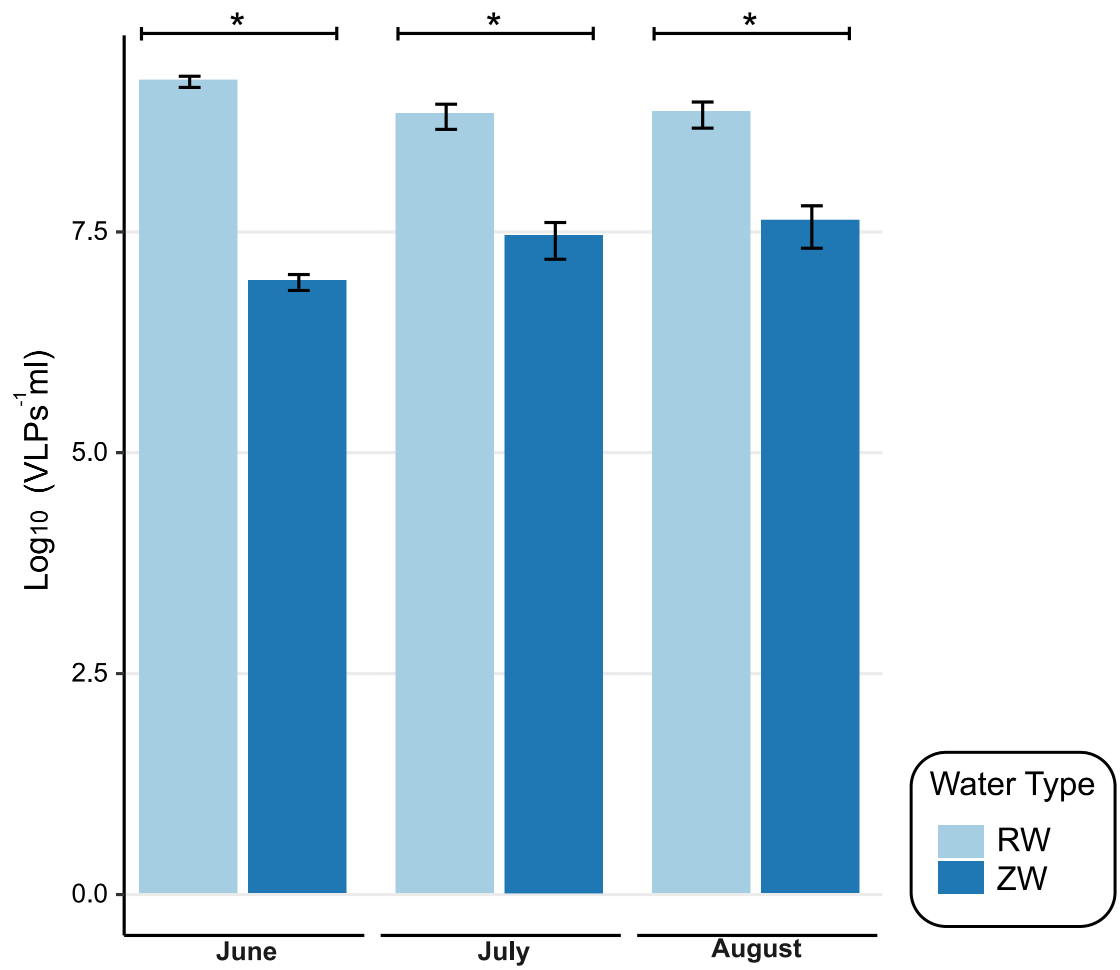

Figure 6.1: Epifluorescent microscopy counts of virus-like-particles (VLPs) in reclaimed water (RW) and zero-valent iron sand filtered reclaimed water (ZW). Samples were collected monthly from June through August. Data presented as mean ± SD, denoted by error bars. Significance determined relative to unfiltered reclaimed water at corresponding sampling dates (*$p \leq 0.05$).

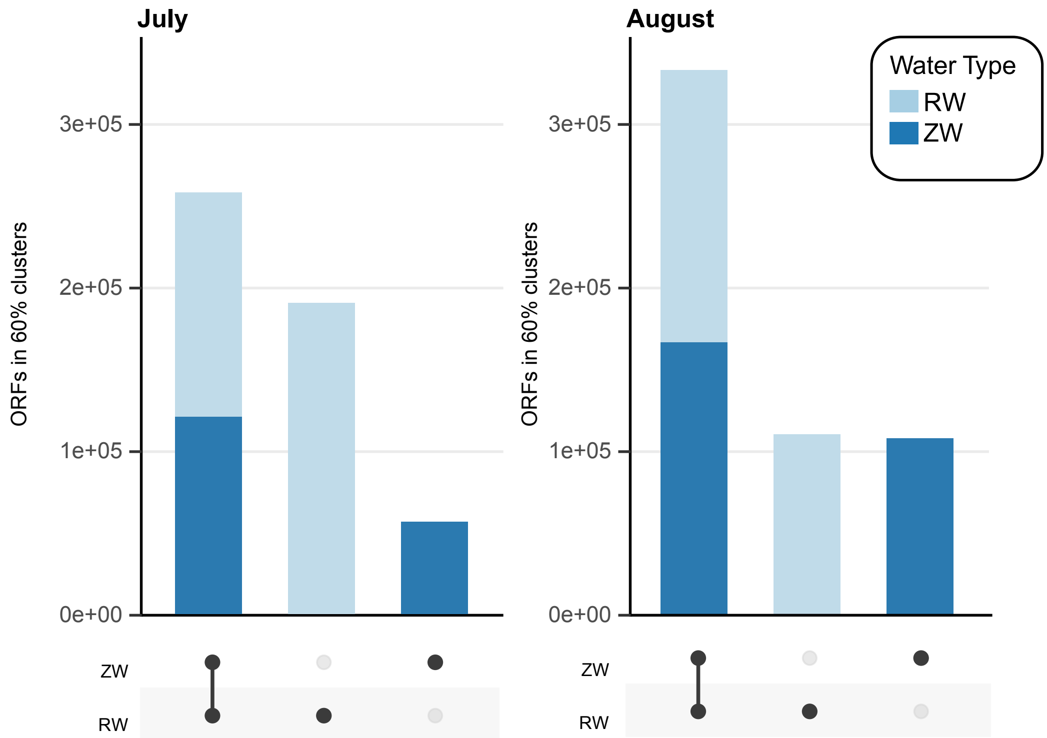

Figure 6.2: Peptide ORF clustering reveals unique and shared functional content in paired reclaimed water (RW) and ZVI-sand filtered reclaimed water (ZW) samples from July and August. Bars denote the number of peptide ORFs from each sample type contained within 60% similarity peptide clusters. Single bars depict the unique peptide ORFs that clustered within water type, while stacked bars depict the peptide ORFs that clustered between water types (e.g. shared peptide ORFs).

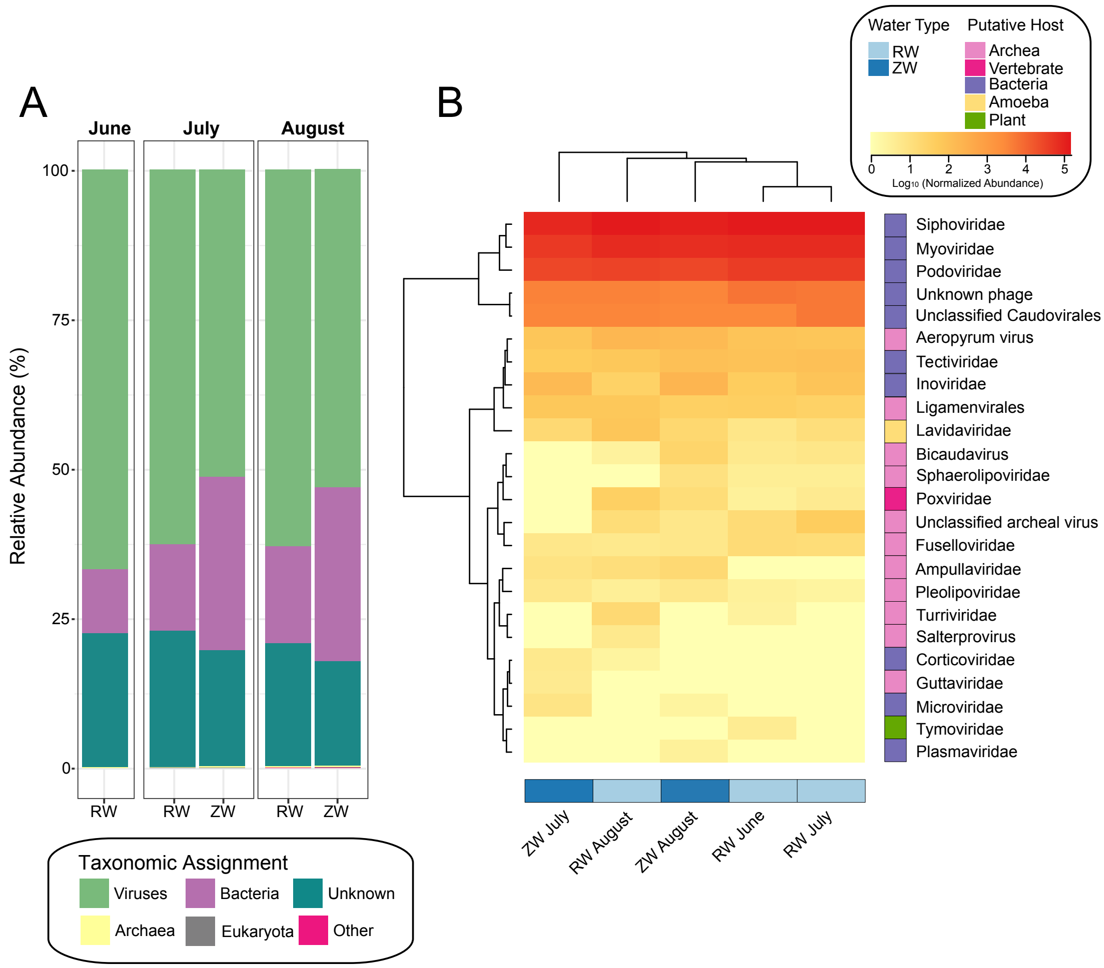

Figure 6.3: Taxonomic composition of reclaimed water (RW) and zero-valent iron sand filtered reclaimed water (ZW). (A) Stacked bar charts of the relative abundances of the taxonomic assignments for RW and ZW viromes. (B) Heatmap showing the abundances (logx + 1 transformed) of the viral taxa in the RW and ZW viromes. The heatmap has samples as columns (colored by water type) and viral taxa as rows (colored by putative host). Normalized abundance measured as contig coverage divided by the sum contig coverage per million.

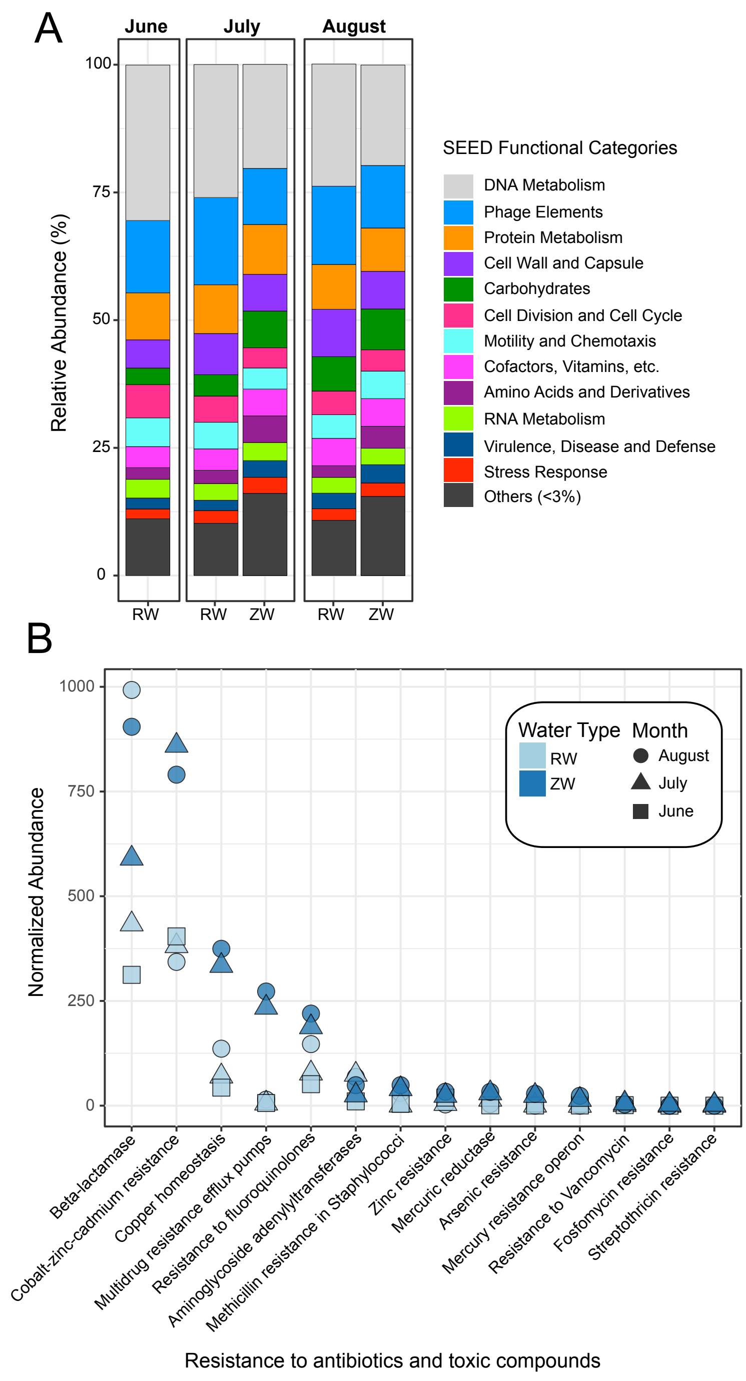

Figure 6.4: Functional composition of reclaimed water (RW) and zero-valent iron sand filtered reclaimed water (ZW). (A) Relative abundances of the SEED subsystems assignments for RW and ZW viromes. (B) Abundance of antibiotic resistance genes in RW and ZW viromes. Normalized abundance measured as ORF coverage divided by sum ORF coverage per million. Shapes denote month samples were collected (June, square; July, triangle; August, circle) and color denotes water type (RW, light blue; ZW, dark blue).

6.7 Tables

Table 6.1: Sequencing effort and assembly characteristics.

Water Type	**Sampling Date**	**no. Read Pairs**	**no. Contigs**	**Mean Contig Size**	**no. peptide ORFs**	**% GC**
RW	6/21/16	23,726,492	194,953	712	337,289	51.1
	7/30/16	27,521,057	181,705	750	327,041	51.3
	8/9/16	24,228,281	162,620	688	276,213	49.9
ZW	7/30/16	29,933,826	115,436	591	177,082	51.4
	8/9/16	30,857,701	170,944	630	273,354	52.4

Table 6.2: Contigs assigned taxonomy.

Water Type	**Sampling Date**	**% Assigned**	**Bacteria**	**Virus**	**Unknown**	**Archaea**	**Eukaroyota**
RW	6/21/16	32.8	8,698	39,377	15,656	193	26
	7/30/16	32.6	9,221	35,074	14,804	165	19
	8/9/16	32.8	10,029	30,711	12,308	181	69
ZW	7/30/16	35.8	15,669	17,519	7,936	137	61
	8/9/16	38.2	24,713	28,839	11,410	244	91

Chapter 7: Conclusions, Public Health Significance, and Future Work

As we continue to face a global freshwater crisis there is an urgent need to explore alternative irrigation water sources. However, critical for the use of these water sources is knowledge of their threats to both environmental and public health. Currently, in the farm-to-fork continuum, irrigation water accounts for one of the chief sources of microbial contamination on fresh produce [346]. In 2018 alone, two major outbreaks across the U.S. of pathogenic *E. coli* on romaine lettuce were preliminarily linked to irrigation water [68]. Although the Food Safety Modernization Act (FSMA) has made strong strides in shifting the focus of food safety from response to prevention, culture-based water quality monitoring–which serves as the basis of the FSMA Produce Safety Rule– is still limited, especially with regard to assessing the overall capability and diversity of microbes in an environment [38]. Next generation sequencing (NGS), however, has proven to be a powerful tool to characterize biodiversity within a variety of ecosystems [13, 418]. Despite their limitations, NGS technologies enable us to cast a wide net and search for potential public health risks that would otherwise be missed by culture alone. My dissertation research utilized culture independent amplicon and shotgun metagenomic sequencing to generate foundational data on the microbial communities in reclaimed and untreated surface waters.

As a primarily exploratory endeavor, this work proceeded from a broad survey of a variety of nontraditional irrigation water sources to a more specific assessment of bacteria and bacteriophage dynamics, composition, and interactions in agriculturally utilized irrigation waters. In total, over a billion reads of novel sequence data was generated on sites rarely studied or studied in isolation including: (i) water from agricultural ponds, freshwater creeks, brackish rivers, and reclamation facilities; (ii) agricultural pond water collected throughout the late season (October-November): (ii) agricultural pond water collected throughout a full calendar year; (iv) water collected temporally from a freshwater creek and soil from an irrigation field survey; and (iiv) reclaimed water before and after ZVI-sand filtration. These samples were used to provide a comprehensive survey of various features of the bacterial populations such as taxonomy, functional potential, antibiotic resistance and virulence gene presence, persistence, and relationship with abiotic and biotic factors. In addition, innovative viral enrichment methodologies were employed to examine the viral communities (Chapter 3, 4, 5), a critical and vastly understudied component in microbiome research. Together these data present a more holistic picture of the bacterial and viral community composition and dynamics within nontraditional irrigation waters. Nevertheless, it has also brought an onslaught of additional questions and avenues of future work.

From this work, it is apparent that untreated lotic and lentic surface waters are rich in microbial life, largely that of environmental bacteria such as *Variovorax*, *Streptomyces*, and *Pusillimonas*, and dsDNA bacteriophage of the order *Caudovi-*

rales. However, they are also home to microorganisms that may have serious public health implications if spread onto irrigated crops. Completely removing or strictly limiting microbial growth in natural surface waters is not feasible and, frankly, a detriment to environmental health. However, this dissertation provides preliminary data that may be used to inform decisions and criteria for water management and safe use.

For instance, when assessing the genetic components of the microbiota in surface and reclaimed waters, our data supported the known limitations of our current culture-based water quality assessment methodologies. The 2012 EPA Guidelines for Water Reuse suggests total or fecal coliforms (fecal indicator bacteria) as the only microorganisms to monitor for water reuse and infer the presence of pathogens [17]. However, bacterial taxa hosting potential virulence factors and antibiotic resistance genes were found in microorganisms that naturally occur in aquatic and soil environments (Chapter 2, 4, 5) and would not necessarily correlate with fecal indicator bacteria. For instance, *Aeromonas* in surface waters were found in several studies not to correlate with indicator bacteria and *Aeromonas*, *Legionella*, *Mycobacterium*, and *Pseudomonas* have all been previously isolated from reclaimed water in the absence of coliforms [384–386]. In this dissertation, *Aeromonas* virulence genes were identified in a freshwater creek and *Aeromonas*-hosting ARGs were detected in the irrigated soil (Chapter 5). While *Aeromonas* spp. are not among the most infamous food-borne pathogens, their presence in irrigation water may be of concern. Over the last few decades, species of *Aeromonas*, which naturally occur in soil and water environments, are becoming increasingly recognized as enteric pathogens acquiring a number of virulence determinants linked with human infection (e.g. gastroenteritis, hemolytic-uremic syndrome, septicemia) [383]. In fact, *Aeromonas hydrophila* was implicated in a massive foodborne outbreak in China where 200 college students were sickened, likely due to the ingestion of salad ingredients washed with contaminated tank water [418]. As a result, a concerted effort should be made by researchers and policy makers to produce more comprehensive assays/models and guidelines for microbial quality monitoring.

Additional evidence that may suggest the need for updates to current culture-based water quality assessment and sampling methodologies was described in Chapter 5, where preliminary evidence showed the potential presence of biofilm-forming bacteria at the point of use (drip irrigation spigot). Although these data came from one sample (and should be explored in greater detail in future studies), biofilm formation in irrigation and drinking water systems is well documented [346]. Because bacteria can persist in biofilms and then detach into flowing waters, their concentrations may be sporadic and vary between the intake and the point-of-use. This may impact the accuracy of agricultural water quality monitoring strategies, which, for the most part, do not specify a required location for water sample collection [346]. Moreover, biofilms, where a vast number of cells live close together, may be a perfect environment for the horizontal transfer of genes [419, 420]. In fact, gene transfer by phage, plasmid conjugation, and DNA transformation has been reported previously within biofilms [419, 420]. This could potentially induce a bottleneck where resistance and virulence genes are spread among bacterial populations. Therefore, more work should be focused on the potential change in microbial community composition during transfer in irrigation systems. In fact, the guidelines for drinking water quality authored by the World Health Organization suggest for effective surveillance and implementation of remediation strategies water samples should be collected both at the source and the point-of-use [421].

The research described in this dissertation can also be used to make more informed decisions regarding irrigation water source management. For instance, seasonal characteristics (Chapter 3, 4, 5), upstream input systems (Chapter 2), sampling date (Chapter 2, 3, 4, 5), and storm events (Chapter 4) were factors found to contribute to the composition and dynamics of microbiota in surface and reclaimed waters. Many of these may be more easily managed on small farm ponds compared to larger aquatic systems, such as rivers. Lotic sites can be impacted by connected waterways and input sources, as well as their catchment area (e.g. agricultural, urban, forested), which tend to be greater than lentic waters and may traverse multiple varied landscapes [34]. For instance, in Chapter 2 a freshwater creek was heavily impacted from an upstream wastewater discharge, resulting in a diversity of ARGs. Farm ponds can be built with conservation buffers, areas of permanent vegetation (e.g. native grasses, shrubs, and trees) designed to intercept pollutants before they reach surface water. This may help mitigate the effects (i.e. increased bacterial diversity, ARGs) observed in the farm pond following a storm event described in Chapter 4. In preliminary studies, conservation buffers have been reported to reduce nutrients, pesticides, pathogens, and sediments by upwards of 50% [422]. However, the full scope of their impact on pond microbial communities and the influx/persistence of ARGs have not been elucidated and represent an avenue for future research.

Furthermore, because ponds are often employed as reference models for larger aquatic systems they present an ideal candidate to test field-portable systems that employ advanced remediation technologies (e.g. UV, ozone, and ZVI-filtration) [423]. Increasing global temperature is poised to change the landscape of climates globally. Throughout the U.S., normally water-rich regions, such as the Mid Atlantic and Southeast, are expected to succumb to intense long-term drought conditions and increased hurricane frequency that may compromise surface water quality and availability [424]. Further remediation technologies may be necessary (and are currently necessary in some locations) to ensure their safe use. ZVI sand filtration is an emerging technology presented in pilot data at the conclusion of the research chapters of this dissertation (Chapter 6). It has been found to remove viral-like-particles from reclaimed water and in previous studies to remove *Escherichia coli* populations, as well as titers of Aichi virus, Murine norovirus, Tulane virus, and bacteriophage MS2 and ΦX174 from water [397, 398, 400]. However, in Chapter 6 the effectiveness of ZVI sand filtration for viral removal was found to decrease over time. As a result, continued studies are needed to examine bacterial and viral community dynamics at fine resolutions over both short and long timeframes to ensure ZVI efficacy, with the ultimate goal of utilizing them for the treatment of reclaimed water.

While many questions are still left unanswered and new questions have formed, this research has provided foundational evidence to aid in our understanding of bacteria and viruses in untreated surface and reclaimed waters. With continued improvements in scientific research and technologies it is conceivable that we will further unravel the complexity of these microbial systems and ensure the safe use of nontraditional water for agricultural applications.

Appendix A: Mentholation Affects the Cigarette Microbiota by Selecting for Bacteria Resistant to Harsh Environmental Conditions and Selecting Against Potential Bacterial Pathogens

Jessica Chopyk, Suhana Chattopadhyay, Prachi Kulkarni, Emma Claye, Kelsey R Babik, Molly C Reid, Eoghan M Smyth, Lauren E Hittle, Joseph N Paulson, Raul Cruz-Cano, Mihai Pop, Stephanie S. Buehler, Pamela I. Clark, Amy R. Sapkota, and Emmanuel F. Mongodin. Mentholation affects the cigarette microbiota by selecting for bacteria resistant to harsh environmental conditions and selecting against potential bacterial pathogens. *Microbiome*, 5(1):22, 2017.

Abstract

There is a paucity of data regarding the microbial constituents of tobacco products and their impacts on public health. Moreover, there has been no comparative characterization performed on the bacterial microbiota associated with the addition of menthol, an additive that has been used by tobacco manufacturers for nearly a century. To address this knowledge gap, we conducted bacterial community

profiling on tobacco from user- and custom-mentholated/non-mentholated cigarette pairs, as well as a commercially-mentholated product. Total genomic DNA was extracted using a multi-step enzymatic and mechanical lysis protocol followed by PCR amplification of the V3-V4 hypervariable regions of the 16S rRNA gene from five cigarette products (18 cigarettes per product for a total of 90 samples): Camel Crush, user-mentholated Camel Crush, Camel Kings, custom-mentholated Camel Kings, and Newport Menthols. Sequencing was performed on the Illumina MiSeq platform and sequences were processed using the Quantitative Insights Into Microbial Ecology (QIIME) software package. In all products, *Pseudomonas* was the most abundant genera and included *Pseudomonas oryzihabitans* and *Pseudomonas putida*, regardless of mentholation status. However, further comparative analysis of the five products revealed significant differences in the bacterial compositions across products. Bacterial community richness was higher among non-mentholated products compared to those that were mentholated, particularly those that were custom-mentholated. In addition, mentholation appeared to be correlated with a reduction in potential human bacterial pathogens and an increase in bacterial species resistant to harsh environmental conditions. Taken together, these data provide preliminary evidence that the mentholation of commercially available cigarettes can impact the bacterial community of these products.

Background

In 2014, an estimated 264 billion cigarettes were sold in the USA, about one-quarter of which were mentholated products [425, 426]. Menthol, a cyclic terpene alcohol, is known to activate cold receptors and provide a "cooling" sensation [427, 428]. In the 1920s, cigarette companies began using this additive to reduce the harshness of cigarette products and to appeal to a wider spectrum of consumers [429, 430]. Although non-menthol cigarettes do contain low levels of menthol, levels in cigarette products labeled as mentholated are 50 to 5000 times higher [431]. For commercially produced menthol cigarettes, menthol, which is usually plant-derived or produced synthetically, is added directly to the tobacco or to other parts of the cigarette (e.g., filter, filter paper) [432]. In addition, several brands of cigarettes (e.g., Camel Crush) have capsules embedded in the filter, which can be "crushed" by the user to release a menthol-containing solution. Today, young adults, minority groups, adult women, and members of low-income households are the primary consumers of menthol cigarettes [426, 433, 434].

Previous studies have provided evidence that menthol smokers are characterized by decreased nicotine metabolism, enhanced systemic nicotine exposure [434], increased serum cotinine levels [435], and increased levels of carboxyhemoglobin [435, 436]. The presence of menthol in some cigarette products has also been shown to increase levels of volatile organic compounds in mainstream smoke [437] and inhibit the detoxification of carcinogens in liver microsome studies [438]. Although results are mixed [439, 440], it appears that menthol cigarettes may be more addictive and may convey a greater risk of cancer and other tobacco-related diseases compared to non-mentholated cigarettes [441, 442]. However, there are relatively few studies that have evaluated other physiological and toxicological health effects associated with exposure to menthol cigarettes, including the impact of the bacteria associated with these products on smokers' oral health.

The antibacterial nature of menthol has been shown to inhibit human and plant pathogenic microorganisms; however, its reaction with the bacterial constituents of the cigarette microenvironment has yet to be explored [443]. The history of microorganisms in tobacco has been documented by several groups [444], with researchers as early as the late 1890s beginning to characterize the microbiology of tobacco before and during fermentation. Fast-forwarding to the 1950s and 1960s, major tobacco companies and researchers began to produce reports describing total numbers of cultivable bacteria in tobacco products [444,444–446]. More recently, several groups have used traditional, culture-dependent methods to identify and characterize specific bacterial and fungal species present in tobacco products including *Actinomycetes* spp. [447], *Pantoea* spp. [448], *Kurthia* spp. [449], *Bacillus* spp. [449], and *Mycobacterium avium* (an important respiratory pathogen) [450].

One study, in particular, recovered viable *M. avium* from cigarette tobacco, tobacco paper and the cigarette filters before cigarettes were smoked and subsequently recovered viable *M. avium* from the cigarette filters after the cigarettes were smoked [450]. These data provide evidence that *M. avium* can survive exposures to high temperatures and gases generated during the cigarette combustion process and potentially be inhaled in mainstream smoke [450]. Other studies have shown that the mainstream smoke of combustible tobacco products also contains other microbial constituents, including lipopolysaccharides, peptidoglycan fragments and fungal components [448]. The same study also showed that cigarettes kept at 94% relative humidity for over 8 days were characterized by additional bacterial and fungal growth within the cigarette tobacco, further demonstrating that microorganisms present in the tobacco are viable and metabolically active [448]. Moreover, in a study by Pauly et al. [444], bacteria growing on single tobacco flakes from multiple cigarette brands were characterized, and the authors hypothesized that these tobacco-associated microorganisms could represent a health risk to the smoker as they are carried to the lungs on the surface of tobacco particulate matter generated during smoking. The impact of these microbial exposures on tobacco users' health is still unclear, as very few epidemiologic studies have focused on the public health impacts associated with the microbiological components of tobacco products. However, bacteria in cigarettes have been previously associated with acute eosinophilic pneumonitis in military personnel deployed in operation Iraqi Freedom, emphasizing the critical role that these microorganisms might play in acute and chronic conditions among tobacco users [449].

Culture-based methods that are used to assess the microbiology of cigarettes, as well as the impacts of menthol on bacterial populations, are limited due to the small percentage of bacterial species that can be cultured in the laboratory. Previous work by our group aimed to address this knowledge gap by applying a 16S rRNA gene-based taxonomic microarray approach to evaluate total bacterial diversity of commercially available cigarettes [451]. In all tested products, 15 different classes of bacteria and a broad array of potentially pathogenic microorganisms were identified, including *Acinetobacter* spp., *Bacillus* spp., *Clostridium* spp., *Klebsiella* spp., *Pseudomonas aeruginosa* spp., and *Serratia* spp. [451]. This initial study also provided some preliminary evidence that the bacterial microbiota of menthol vs. non-menthol cigarettes may vary. However, due to the relatively small number of bacterial taxa represented on the microarray used in the previous study, our view of the bacterial diversity within the tested products was limited.

Therefore, in this study, we applied high-throughput next generation sequencing, which provides a much broader view of total bacterial diversity to characterize five cigarette products: Camel Crush, user-mentholated Camel Crush, Camel Kings, custom-mentholated Camel Kings, and Newport Menthols. In addition to comparing mentholated and non-mentholated cigarette pairs we aimed to identify potential bacterial pathogens that users may be exposed to when they smoke these products, and expand our understanding of the scope of bacterial diversity present in mentholated and non-mentholated cigarette tobacco.

Methods

Sample collection

In the Spring of 2014, menthol and non-menthol cigarettes were either purchased from selected tobacco stores in College Park, Maryland or provided by our collaborators at The Battelle Public Health Center for Tobacco Research (Columbus, OH) (Table A.1). The following products were purchased from selected tobacco stores in College Park, MD: (1) Camel Crush, regular, fresh (CC) (Camel Crush; R.J. Reynolds Tobacco Co., Winston-Salem, NC, USA), where the capsule within the filter was subsequently not crushed during the study; (2) Camel Crush, regular, fresh (CCM), where the capsule was subsequently crushed during the study to release a menthol-containing solution into the cigarette filter (user mentholated) (CCM); and (3) a commercially mentholated brand, Newport Menthol Box (NMB) (Lorillard Tobacco Co., Greensboro, NC, USA). The following products were provide by Battelle: 4) Camel full flavor, hard pack, king (CK) (Camel Kings; R.J. Reynolds Tobacco Co., Winston-Salem, NC, USA); and 5) Camel Kings that were custom-mentholated by Battelle (CKM) using a vapor deposition technique described in detail in MacGregor et al. [452]. The custom-mentholated Camel Kings were prepared concurrently in three separate chambers [452]. The Camel Kings that were not mentholated went through the same motions and preparations and were handled in the same exact way as those that were mentholated. The only difference was that the non-mentholated Camel Kings were not exposed to the mentholation chamber. All custom-mentholated and non-mentholated Camel Kings were shipped from Battelle on the day that custom-mentholation was completed via overnight carrier without refrigeration and all cigarettes were subsequently stored at room temperature until processing. We included two pairs of mentholated and non-mentholated products (custom-mentholated Camel Kings versus non-mentholated Camel Kings; and "non-crushed" Camel Crush cigarettes versus "crushed" Camel Crush cigarettes, as described above) so that we could specifically evaluate the influence of the addition of menthol into two different products on the bacterial community composition of those products. Three lots of each cigarette product were tested in replicates of 6 for a total of 90 samples (18 cigarettes per brand) tested during the study.

DNA extraction

DNA extraction was performed on cigarettes from freshly opened packages, with the exception of the custom-mentholated and non-mentholated Camel Kings (CK and CKM), which were opened at Battelle, processed and shipped as described above. Our total DNA extraction protocol was adapted from procedures previously published [453, 454]. Briefly, each cigarette was dissected under sterile conditions, and 0.2 g of tobacco was weighed out and aseptically placed in Lysing Matrix B tubes (MP Biomedicals, Solon, OH). Enzymatic lysis was initiated by adding the following to the tubes containing cigarette tobacco and lysing matrix: 1 ml of ice cold 1 molecular grade PBS buffer (Gibco by Life Technologies, NY), 5 μl lysozyme from chicken egg white (10 mg/ml, Sigma-Aldrich, MO), 5 μl lysostaphin from Staphylococcus staphylolyticus (5 mg/ml, Sigma-Aldrich, MO) and 15 μl of mutanolysin from *Streptomyces globisporus* ATCC 21553 (1 mg/ml, Sigma-Aldrich, MO). Tubes were then incubated at 37 °C for 30 min, after which a second enzymatic cocktail was added to each tube, composed of 10 μl Proteinase K (20 mg/ml, Invitrogen by Life Technologies, NY) and 50 μl of SDS (10% w/v, BioRad). Following incubation at 55 °C for 45 min, the samples were then further lysed mechanically using a FastPrep Instrument FP-24 (MP Biomedicals, CA) at 6.0 m/s for 40s. The resulting lysate was centrifuged for 3 min at 10,000 rcf and DNA was purified using the QIAmp DSP

DNA mini kit 50, v2 (Qiagen, CA), according to the manufacturer's protocol. Six replicate DNA extractions were completed on each sample and negative extraction controls were included to ensure that no exogenous DNA contaminated the samples during extraction. DNA quality control/quality assurance was performed using spectrophotometric measurements on a NanoDrop (Thermo Scientific, City, State), as well as gel electrophoresis.

16S rRNA gene PCR amplification and sequencing

The V3-V4 hypervariable region of the 16S rRNA gene was PCR-amplified and sequenced on Illumina MiSeq (Illumina, San Diego, CA) using a dual-indexing strategy for multiplexed sequencing developed at the Institute for Genome Sciences and described in detail previously [241].

Briefly, PCR reactions were set-up in 96-well microtiter plates using the 319 F (ACTCCTACGGGAGGCAGCAG) and 806R (GGACTACHVGGGTWTCTAAT) universal primers, each of which also included a linker sequence required for Illumina MiSeq 300 bp paired-ends sequencing, and a 12 bp heterogeneity-spacer index sequence aimed at minimizing biases associated with low-diversity amplicons sequencing [241, 243]. This sample multiplexing approach ensured that 500 samples could be multiplexed in a single Illumina MiSeq run. PCR amplifications were performed using Phusion High- Fidelity DNA polymerase (Thermo Fisher, USA) and 2 ng of template DNA in a total reaction volume of 25 μl . Because of the presence of PCR inhibitors in the DNA solution, an additional 0.375 μl of bovine serum albumin

(BSA) (20 mg/ml, Sigma) was added to the PCR reactions. Reactions were run in a DNA Engine Tetrad 2 thermo cycler (Bio-Rad, USA) using the following cycling parameters: 30 s at 98 °C, followed by 30 cycles of 10 s at 98 °C, 15 s at 66 °C, and 15 s at 72 °C, with a final step of 10 min at 72 °C. Negative controls without DNA template were performed for each primer pair. The presence of amplicons was confirmed using gel electrophoresis, after which the SequalPrep Normalization Plate kit (Invitrogen Inc., CA, USA) was used for clean-up and normalization (25 ng of 16S PCR amplicons from each sample were included), before pooling and 16S rRNA sequencing using the Illumina MiSeq (Illumina, San Diego, CA) according to the manufacturer's protocol.

Sequence quality filtering and analysis of 16S rRNA gene sequences

16S rRNA reads were initially screened for low quality bases and short read lengths [241]. Paired-end read pairs were then assembled using PANDAseq [242] and the resulting consensus sequences were de-multiplexed (i.e., assigned to their original sample), trimmed of artificial barcodes and primers, and assessed for chimeras using UCHIME in *de novo* mode implemented in Quantitative Insights Into Microbial Ecology (QIIME ; release v. 1.9) [243]. Quality trimmed sequences were then clustered *de novo* into Operational Taxonomic Units (OTUs) with the SILVA 16S database [244] in QIIME [243], with a minimum confidence threshold of 0.97 for the taxonomic assignments. All sequences taxonomically assigned to chloroplasts were removed from further downstream analysis. Data were normalized to account for uneven sampling depth with metagenomeSeq's cumulative sum scaling [244], a novel normalization method that has been shown to be less biased than the standard approach (total sum normalization).

Taxonomic assignments of the most abundant genera, contributing >1% of the total abundance in at least one sample, were obtained through QIIME [243] and visualized with RStudio Version 0.99.473 and vegan [248], gplots [455], RColorBrewer [456] and heatplus [246] R packages. Prior to normalization, alpha diversity was estimated with the Chao1 estimator [457], and the Shannon Index [300] through the R packages: Bioconductor [247], metagenomeSeq [245], vegan [248] phyloseq [249] and fossil [250]. To account for uneven sampling depth, the data were also rarefied to the minimum sampling depth of 631 sequences. Alpha diversity data was tested for significance using a Tukey test. Beta diversity was determined through principle coordinates analysis (PCoA) plots of Bray-Curtis distance performed through QIIME and tested for significance with ANOSIM (9,999 permutations) [253].

Determination of statistically significant differences ($p \leq 0.05$) in OTU bacterial relative abundance between mentholated cigarette products and their non-mentholated counterpart (mentholated Camel Crush vs. non-mentholated Camel Crush and custom-mentholated Camel Kings vs. non-mentholated Camel Kings) was performed using DESeq2 [458] through QIIME [243], which utilizes Benjamini-Hochberg multiple-inference correction. DESeq was used due to its high power in computing smaller sample sizes (<20 samples per group) [459]. The significant OTUs ($p \leq 0.05$) were visualized with RStudio Version 0.99.473 and R packages ggplot2 [183], vegan [248], and phyloseq [249]. In addition, species-level assignments were performed for OTUs of interest: reference sequences matching assigned genera of each OTU were extracted from the Ribosomal Database Project (RDP; http://rdp.cme.msu.edu/), aligned with the sequences from the OTU(s) of interest via MAFFT [186], and the V3-V4 region extracted. An unrooted maximum likelihood tree with 10 bootstrap replicates was generated with PhyML [460] for each of the alignments. Trees were visualized with FigTree [302] and branches colored based on species.

Results

Sequencing data and taxonomic assignments

All 90 cigarette samples were successfully PCR amplified and sequenced, thus validating our DNA extraction and purification protocol. A total of 2046 different bacterial OTUs (97% identity) were identified from a total of 909,053 sequences across all samples, and the average number of sequences per sample was 10,100 (± 5004 SD Figure A.1).

The average relative abundance of the most dominant genera (>1.0% in at least one sample) showed that, across all brands, bacteria from the genus *Pseudomonas* dominated, followed by unclassified members of the *Enterobacteriaceae* family, and members of the *Pantoea* and *Bacillus* genera (Figure A.2). Members from the *Pseudomonas* genus were comprised of 15 unique OTUs, with 7 *Pseudomonas* OTUs shared between all mentholation states (OTU#s 1532, 10, 134, 1868, 1886, 8, and

3). Some of these shared *Pseudomonas* OTUs were assigned via RDP classification using SILVA to *Pseudomonas oryzihabitans* (OTUs #1868 and 6) and *Pseudomonas putida* (OTU #3) species. Certain OTUs were also unique to the different menthol products, including OTU #1250 and OTU #1137 for NMB and OTU #77 for CCM. Species level taxonomic information was assigned only to OTU #1137, *Pseudomonas fulva* species. In addition, heatmap hierarchical clustering of the samples revealed that the bacterial community profiles were more similar between the non-menthol cigarette products CK and CC, compared to the commercially mentholated (NMB) and custom-mentholated (CKM) products (Figure A.2).

Alpha and beta diversity metrics by product and menthol state

Because sequence coverage can have an impact on measuring alpha diversity, a quantification of intra-sample diversity, we employed Chao1 and Shannon indices on both non-rarefied and rarefied data (Figure A.3). Tobacco-associated microbiota from the custom-mentholated Camel King (CKM) exhibited significantly lower Chao1 diversity ($p \leq 0.05$) compared to its non-mentholated counterpart (CK), regardless of rarefaction (Figure A.3).

To quantify inter-sample diversity (beta diversity), principal coordinate analyses using the Bray Curtis distance, a measure widely used to measure the compositional dissimilarity between two different sites in ecology and microbiome studies, were performed. Separation of the tobacco-associated bacterial profiles was evident along the first principal component (PC1), which explained 8.59% of the total variability between communities, and the second principal component (PC2), which explained 5.95% of the total variability between communities by brand (ANOSIM R=0.35, p=0.0001) and mentholation status (ANOSIM R=0.43, p=0.0001) (Figure A.4). This was especially evident for the commercially-mentholated, custom-mentholated and non-mentholated products. Unweighted and weighted UniFrac distances [461] were also used to measure beta diversity between the brands (Figure A.5), (ANOSIM R value=0.25, p=0.0001) and (ANOSIM R=0.16, p=0.0001), respectively.

Taxonomic analysis by product and menthol status

There were 173 OTUs at statistically significantly ($p \leq 0.05$) different relative abundances between custom-mentholated Camel Kings and non-mentholated Camel Kings (Figure A.6). Out of these, 167 OTUs were at lower relative abundance in the custom-mentholated Camel Kings, of which 116 were Gram-negative (Figure A.6), with species level assignments including *Achromobacter* sp. HJ-31-2 (OTU #16), *Azospirillum irakense* (OTU #167), *Acinetobacter calcoaceticus* (OTU #40), *Pseudomonas putida* (OTU #3), *Stenotrophomonas maltophilia* (OTU #15), *Pseudomonas aeruginosa* (OTU #420), *Erwinia chrysanthemi* (OTU #446), *Proteus mirabilis* (OTU #450), *Acinetobacter baumannii* (OTU #29), *Agrobacterium tumefaciens* (OTU #1998), and *Pseudomonas oryzihabitans* (OTU #1868). The remaining 51 OTUs at lower relative abundance in custom-mentholated Camel Kings were Gram-positive (Figure A.6), with species level assignments including *Paeni-*

bacillus amylolyticus (OTU #37), *Paenibacillus montaniterrae* (OTU #91), *Paenibacillus* sp. icri4 (OTU #51), *Streptomyces* sp. KP17 (OTU #52), *Bacillus pumilus* (OTU #1937, 5, 1948), *Bacillus cereus* (OTU #176), *Bacillus novalis* (OTU #530 and 1442), *Bacillus clausii* (OTU #9), and *Bacillus licheniformis* (OTU #41). In addition, six OTUs were at higher relative abundance in the custom-mentholated Camel Kings and were composed of two Gram-negative bacteria, *Schlegelella* sp. (OTU #87) and *Silanimonas* sp. (OTU #207), and four Gram-positive bacteria, *Anoxybacillus* sp. (OTU #31), *Vagococcus* sp. (OTU #54), *Deinococcus* sp. (OTU #272), and *Thermus* sp. (OTU #266).

There were 60 OTUs at statistically significantly ($p \leq 0.05$) different relative abundances between mentholated Camel Crush and non-mentholated Camel Crush (Figure A.7). Twenty-two OTUs were at lower relative abundance in the mentholated Camel Crush and, of these, 10 were Gram-negative OTUs including *Aeromonas* sp. (OTU #285), *Cedecea* sp. (OTU #783), unknown *Sphingomonadales* (OTU #333), *Stenotrophomonas* sp. (OTU #1682), *Paracoccus* sp. (OTU #289), unknown *Enterobacteriaceae* (OTU #1969, 2017), *Sphingobacterium* sp. (OTU #124), and *Pantoea* sp. (OTU #398 and 1448). The remaining 12 OTUs at lower relative abundance in mentholated Camel Crush were Gram-positive and included *Bacillus* sp. (OTU #30), *Facklamia* sp. (OTU #104), *Jeotgalicoccus* sp. (OTU #73), *Staphylococcus* sp. (OTU #143), *Saccharopolyspora* sp. (OTU #293), unknown *Streptomycetaceae* (OTU #1729), *Nocardioides* sp. (OTU #86), *Paenibacillus* sp. (OTU #128 and 340), unknown *Bacillaceae* (OTU #296), unknown *Bogoriellaceae* (OTU #193), and unknown *Bacillales* (OTU #667). Additionally, 38 OTUs were at higher relative abundance in the mentholated Camel Crush samples and consisted of 26 Gram-negative OTUs, with species assignments for *Azospirillum irakense* (OTU #167) and *Pectobacterium carotovorum* (OTU #48). The remaining 12 were Gram-positive and included *Sporosarcina* sp. (OTU #228), *Lysinibacillus* sp. (OTU #93), *Solibacillus* sp. (OTU #90), *Anoxybacillus* sp. (OTU #31), *Corynebacterium* sp. (OTU #21 and 551), *Aerococcus* sp. (OTU #14), unknown *Bacillales* (OTU #1182), *Brevibacterium* sp. (OTU #153), *Deinococcus* sp. (OTU #272), *Lactobacillus plantarum* (OTU #359), and *Bifidobacterium* sp. (OTU #535).

OTUs of interest were selected to confirm or predict species-level assignments via phylogenic analyses (Figures A.8-A.15). OTUs included *Pseudomonas putida* (OTU #3), *Pseudomonas oryzihabitans* (OTU #8, 1868), *Pseudomonas* sp. (OTU #10, 77, 132, 134, 163, 251, 608, 972, 1250, 1532, 1872, 1886), *Pseudomonas aeruginosa* (OTU #420), *Pseudomonas fulva* (OTU #1137), *Acinetobacter* sp. (OTU #12, 182, 247, 870, 1900), *Acinetobacter baumannii* (OTU #29), *Acinetobacter calcoaceticus* (OTU #40, 496), *Proteus mirabilis* (OTU #450),*Anoxybacillus* sp. (OTU #31), *Vagococcus* sp. (OTU #54), *Deinococcus* sp. (OTU #272), *Thermus* sp. (OTU #266), *Stenotrophomonas* sp. (OTU #1682, 1899, 1913), and *Stenotrophomonas maltophilia* (OTU #15).

Distinct phylogenetic clustering could be seen among the OTUs and representative species (Figures A.8-A.15). *Pseudomonas oryzihabitans* OTU #8 and 1868 and *Pseudomonas aeruginosa* OTU #420 claded with strains of their assigned species. *Pseudomonas* sp. OTU #10 and 1886 grouped closely with strains of *Pseudomonas putida*, while OTUs #251, 1250, and 134 claded with strains of *Pseu-*

domonas stutzeri (Figure A.8). Although close to several strains of *Pseudomonas putida*, OTU #3 did not group within the large clades of this species (Figure A.8).

Acinetobacter baumannii (OTU #29) and *Acinetobacter calcoaceticus* (OTU #496) also clustered with strains of their assigned species (Figure A.9). Additionally, *Acinetobacter* sp. OTU 12 grouped with *Acinetobacter baumannii* (Figure A.9). *Stenotrophomonas* sp. OTUs #1913 and 1682 appeared close to one another within a clade of *Stenotrophomonas maltophilia*. *Stenotrophomonas maltophilia* OTU #15 claded further away from OTUs #1913 and 1682, but was also with strains of *Stenotrophomonas maltophilia* (Figure A.10). *Stenotrophomonas* sp. OTU #1899 appeared most phylogenetically related to strains of *Stenotrophomonas chelatiphaga* (Figure A.10). *Anoxybacillus* sp. (OTU #31) claded closely to strains of *Anoxybacillus flavithermus*, *Deinococcus* sp. (OTU #272) appeared closely related to strains of *Deinococcus geothermalis*, and *Thermus* sp. (OTU #266) claded closely to *Thermus scotoductus* (Figure A.11-A.15).

Discussion

It has been well established that smokers and those exposed to secondhand smoke are more susceptible to bacterial infections than are non-smokers [57]. Therefore, characterizing this exposure and, more specifically, the bacterial components of cigarette tobacco and their additives, is an important step in uncovering the relationship between tobacco products and user-health. This study aimed to provide comprehensive data concerning bacterial communities present in mentholated and non-mentholated cigarettes by utilizing next-generation sequencing technologies that, to date, have been underutilized in the field of tobacco regulatory science.

The most abundant genus detected in all cigarette products tested, regardless of mentholation status, was *Pseudomonas* (Figure A.2). This was not unexpected as species of *Pseudomonas* are ubiquitous in aquatic and terrestrial environments and have been hypothesized to be a part of the core pulmonary bacterial microbiome [462]. *Pseudomonas* spp. have also been implicated as the dominant genus in cases of chronic obstructive [463], cystic fibrosis [464], and subjects with decreased lung function [463], making their high prevalence and high abundance within cigarette tobacco a potential human health concern. *Pseudomonas putida* due to its metabolical versatility has a distinct association with tobacco and human disease [465]. For instance, several strains of *Pseudomonas putida* (e.g. S16, J5, SKD, and ZB-16A) have the ability to degrade nicotine [466–470], while others have emerged as significant human pathogens causing urinary tract infections [471, 472] and nosocomial pneumonia [471, 473], particularly in ill or immunocompromised patients. In addition, it has been suggested that the clinical isolate strain, HB3267, acquired antibiotic and biocide resistance genes from opportunistic human pathogens, including *Acinetobacter baumannii* [474], which was one of the species found at higher relative abundance in Camel Kings compared to its mentholated counterpart (Figure A.6). *Acinetobacter baumannii* is a Gram-negative opportunistic pathogen of particular global concern due to its increasing rates of antibiotic resistance [475–477] and connection to nosocomial pneumonia and ventilator-associated pneumonia in patients with underlying lung disease [475, 478, 479].

Additional common and rare bacterial species' some of which are known to cause respiratory illnesses' were found at higher relative abundances in the non-mentholated Camel Kings compared to the custom-mentholated Camel Kings, including *Pseudomonas oryzihabitans* and *Pseudomonas aeruginosa*. *Pseudomonas aeruginosa* is noteworthy as a member of the tobacco microenvironment not only due to its association with the occurrence and exacerbation of COPD but also due to its response to cigarette smoke [480–482]. A study preformed on murine models showed that exposure to cigarette smoke followed by infection with *Pseudomonas aeruginosa* resulted in delayed clearance of infection and increased morbidity compared to controls [482].

Despite the overall decrease in bacterial diversity and potential human pathogens that we observed in custom-mentholated compared to non-mentholated Camel Kings, we detected statistically significant ($p \leq 0.05$) increases in the relative abundance of four Gram-positive bacterial species (*Thermus* sp., *Deinococcus* sp., *Vagococcus* sp., and *Anoxybacillus* sp.) and two Gram-negative species (*Silanimonas* sp. and *Schlegelella* sp.) in the mentholated product. Interestingly, *Anoxybacillus* and *Deinococcus* include species that are able to withstand extreme environmental conditions (e.g., elevated pH, industrial processes, UV treatment, radiation) [41,483–485], possibly due to the production of protective carotenoids found in strains of both genera [486–488]. Furthermore, species of *Thermus* [489], *Silanimonas* [490] and *Schlegelella* [480] are known to be thermophilic, hyperthermophilic and/or alkaliphilic. For example, strains of *Thermus scotoductus* have been isolated from a hot water pipeline [491], a South African gold mine [492], and a sulfide-rich neutral hot spring [493].

These data suggest that menthol may be effective against Gram-negative bacteria in cigarette products and select and/or introduce resilient bacterial species that can tolerate the antibacterial activity of menthol. Menthol, although known to be active against both Gram-positive and Gram-negative bacteria [443], has shown, in some instances, to be more effective against Gram-negative bacteria, especially compared to other essential oils [494]. Nevertheless, this overall trend was not observed in our comparisons between the user-mentholated Camel Crush and the non-mentholated Camel Crush. This finding may be due to the degree and rate of menthol exposure in the user-mentholated Camel Crush products. Because these cigarettes are user-mentholated (by crushing a capsule within the cigarette filter and releasing a menthol-containing solution immediately before use), the tobacco is generally exposed to the antibacterial effects of menthol only for a brief period of time before consumption, if at all. For these products, we only evaluated a single time point, just following menthol release; as a result the menthol may not have had the opportunity to migrate fully to the tobacco.

Our study had other limitations as well. We detected more than 2000 OTUs, but as with all DNA-based 16S rRNA gene-sequencing studies, future studies are required to confirm whether these bacteria are active and capable of potentially colonizing a user exposed to these microorganisms. It is also important to note that, chemically, the only difference between the tobacco content of the custom-mentholated Camel King and non-mentholated Camel King cigarettes was the addition of L-menthol. However, the mentholation process used to produce the custom-mentholated Camel King could not be performed under DNA-free conditions, and the introduction of low levels of contaminating foreign bacterial DNA, although unlikely, could be a possibility. Furthermore, commercially available cigarettes may differ from each other in more ways than menthol content, such as tobacco blend [432]. However, the presence of increasing *Anoxybacillus* and *Deinococcus* OTUs in both custom and user-mentholated products suggests a relationship with menthol that should be further tested. Finally, we evaluated the bacterial communities of cigarette products stored under one environmental condition. Characterization of products stored under varying temperature and relative humidity conditions would enable us to better predict the impact of typical daily storage conditions (e.g., pocket conditions) on the dynamics of the bacterial communities in mentholated and non-mentholated cigarettes. Such experiments are currently ongoing in our laboratory. Even with these caveats, our study provides new knowledge regarding the bacterial constituents of commercially mentholated and non-mentholated tobacco products and the potential importance of these bacterial communities to human health.

From pre-harvest to puff, cigarette-associated bacteria are a culmination of ecosystems and commercial manipulations that result in a complex and diverse bacterial community, which may contribute to the acquisition and exchange of pathogenic and antibiotic resistance genes and/or species selection. Our data suggest that tobacco flavor additives, such as menthol, can affect the bacterial community composition of tobacco products and may lead to the selection or introduction of more resilient species. The bacteria and bacterial components present in non-mentholated and mentholated cigarettes may be introduced into the lung and oral cavity during the smoking process, carried by the filter-end of the cigarette butt and/or the tobacco particulate matter within mainstream smoke [444,444,450,495]. These bacterial communities could play a direct role in the development of infectious and/or chronic illnesses among users or exacerbate existing negative health effects associated with smoking.

Conclusions

This study comprehensively characterizes the complex bacterial communities residing in mentholated and non-mentholated cigarette products, which include bacterial pathogens of importance to public health. Most importantly, our study also shows that mentholation of cigarette products, a process used to reduce the harshness of cigarette products and appeal to a wider spectrum of consumers, significantly impacts the bacterial community of these products. Mentholation appeared to be correlated with a reduction in potential human bacterial pathogens and an increase in bacterial species resistant to harsh environmental conditions. These findings have critical implications regarding exposure to potentially infectious pathogens among cigarette smokers, and can be used to inform future tobacco control policies focused on the microbiology of tobacco, an understudied focus area in tobacco regulatory science.

Figures

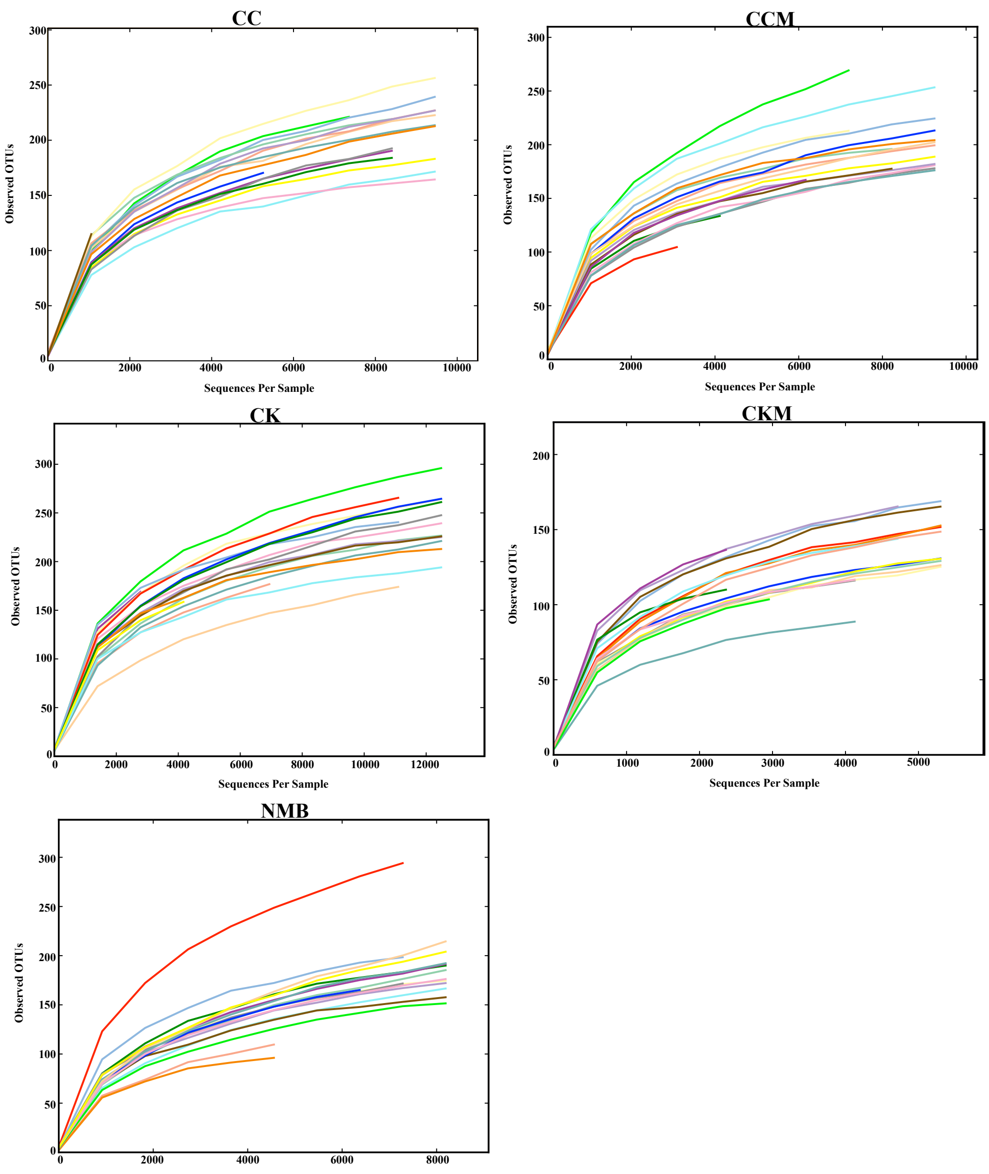

Figure A.1: Rarefaction curves for each product.

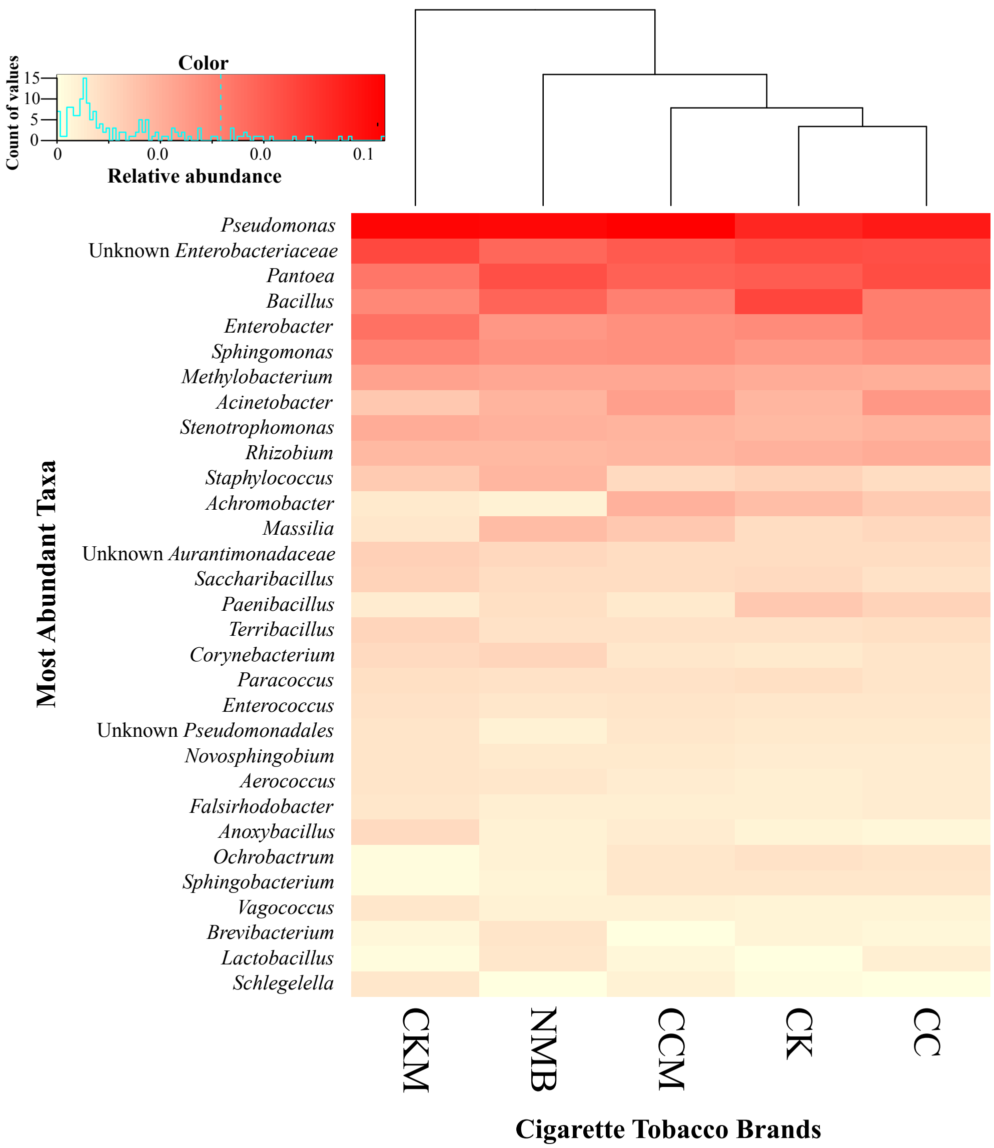

Figure A.2: Heat map showing the relative abundances of the most dominant bacterial genera identified (>1%) in tested cigarette products. Samples pooled by product type: Camel Crush (CC), mentholated Camel Crush (CCM), Camel Kings (CK), custom-mentholated Camel Kings (CKM), and Newport Menthol Box (NMB). Hierarchical clustering of the pooled samples is represented by the dendrogram at the top and inside the color key shows a histogram of the count of the individual values.

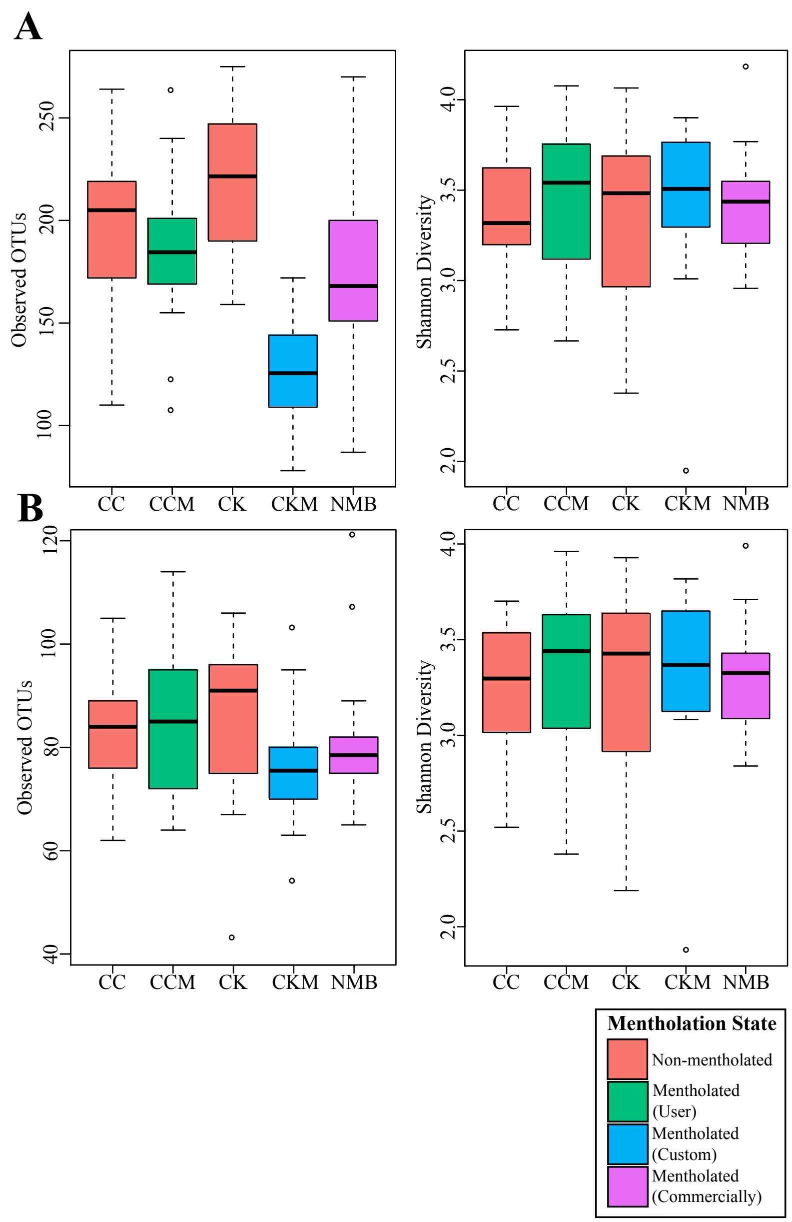

Figure A.3: Box plots showing alpha diversity (Chao1 richness estimator and Shannon Index) variation across samples on non-rarefied data (A) and with data rarefied to the minimum sampling depth (B). Bars are colored by mentholation status: red bars–non-mentholated; green bars–user mentholated; blue bars–custom-mentholated; purple bars–commercially-mentholated.

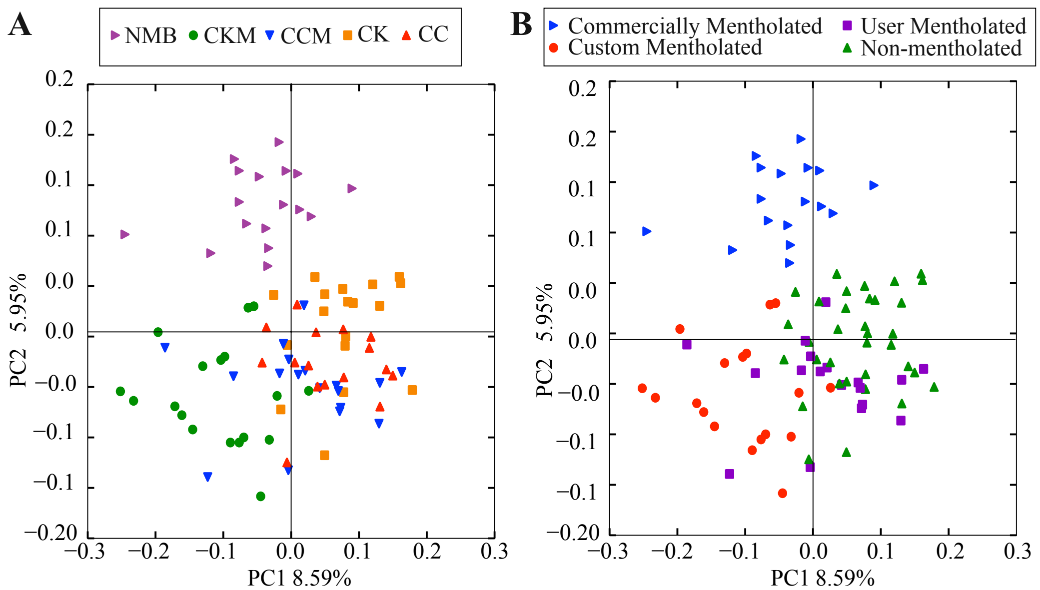

Figure A.4: PCoA analysis plots of Bray-Curtis computed distances between cigarette products. (A) Points colored by brand: purple–Newport Menthol (NMB); green–mentholated Camel King (CKM); blue–mentholated Camel Crush (CCM); orange-Camel Kings (CK); red-Camel Crush (CC) (ANOSIM R value=0.35, p value=0.0001); (B) Points colored by mentholation status: green–non-mentholated; purple-user mentholated; blue–commercially-mentholated; red–custom mentholated. (ANOSIM R=0.43, p=0.0001).

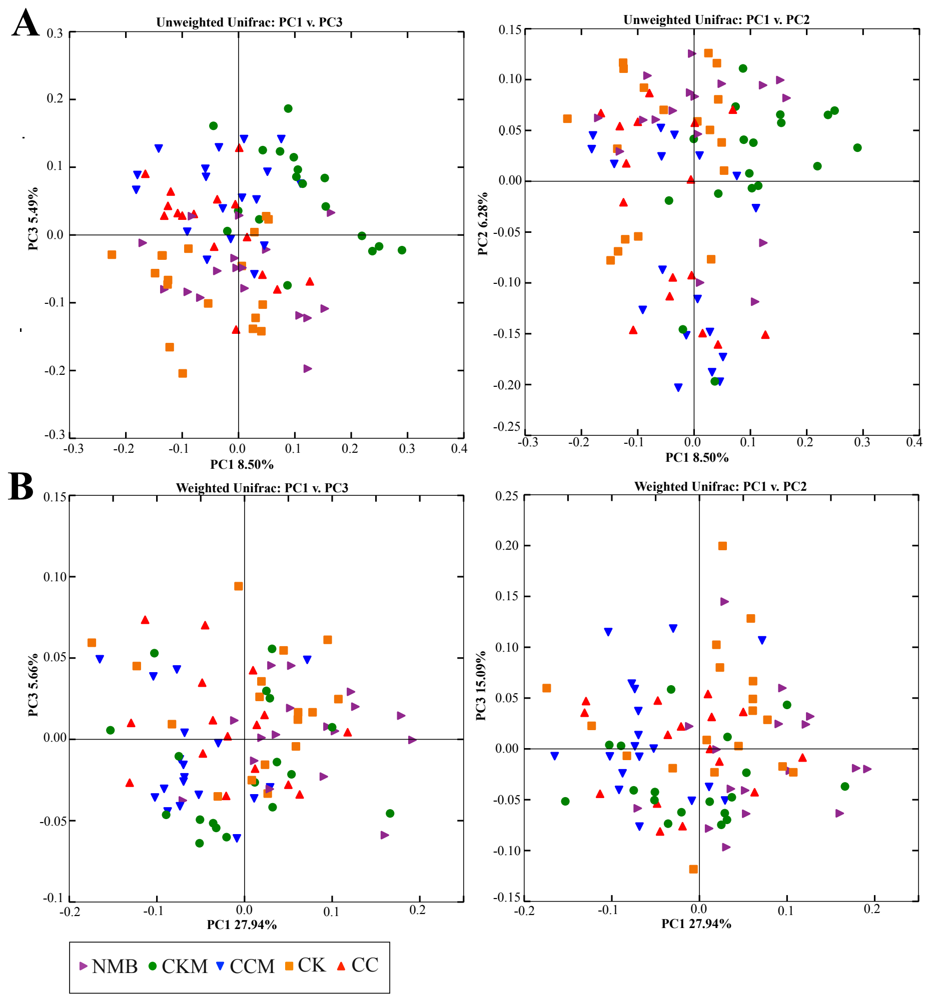

Figure A.5: PCoA analysis plots of weighted and unweighted Unifrac computed distances between cigarette products.

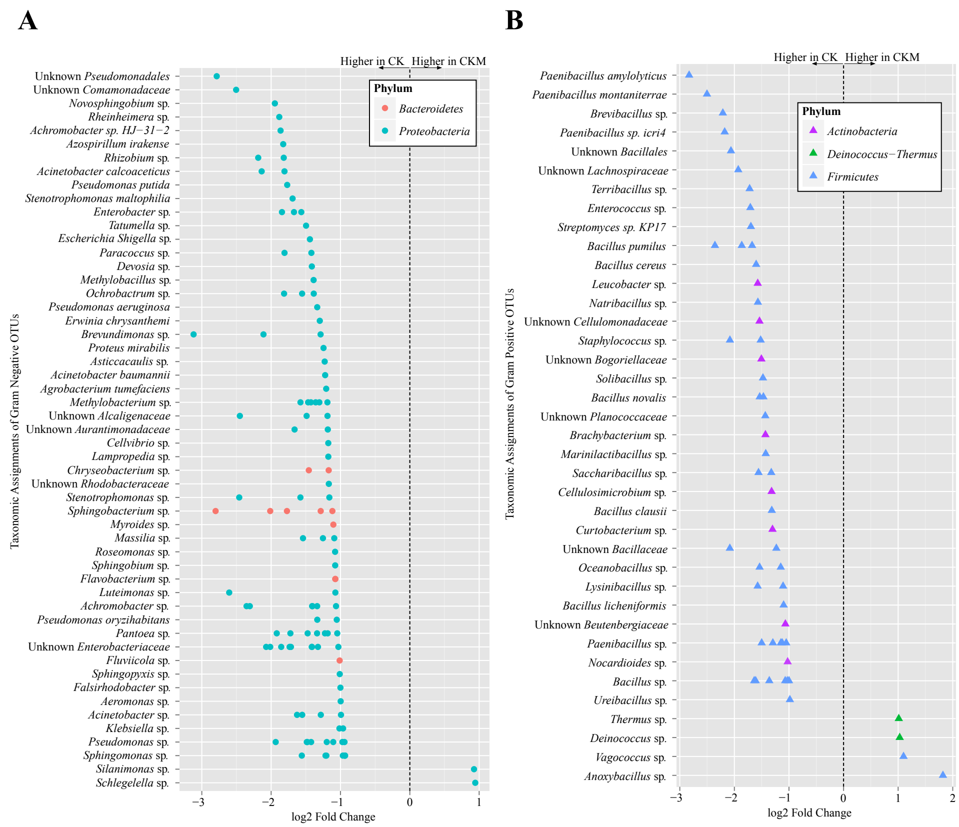

Figure A.6: Overview of relative abundances of bacterial OTUs that were statistically significantly different ($p \leq 0.05$) between custom-mentholated Camel Kings (CKM) and non-mentholated Camel Kings (CK). OTUs are colored by Phylum and differentiated by Gram negative (a) and Gram positive (b) classification. A positive log2-fold change value denotes an OTU that is significantly higher in custom-mentholated Camel Kings, while a negative log2-fold change indicates an OTU that is significantly higher in non-mentholated Camel Kings. The dotted line and arrows highlight the conversion in log2-fold change from negative to positive values.

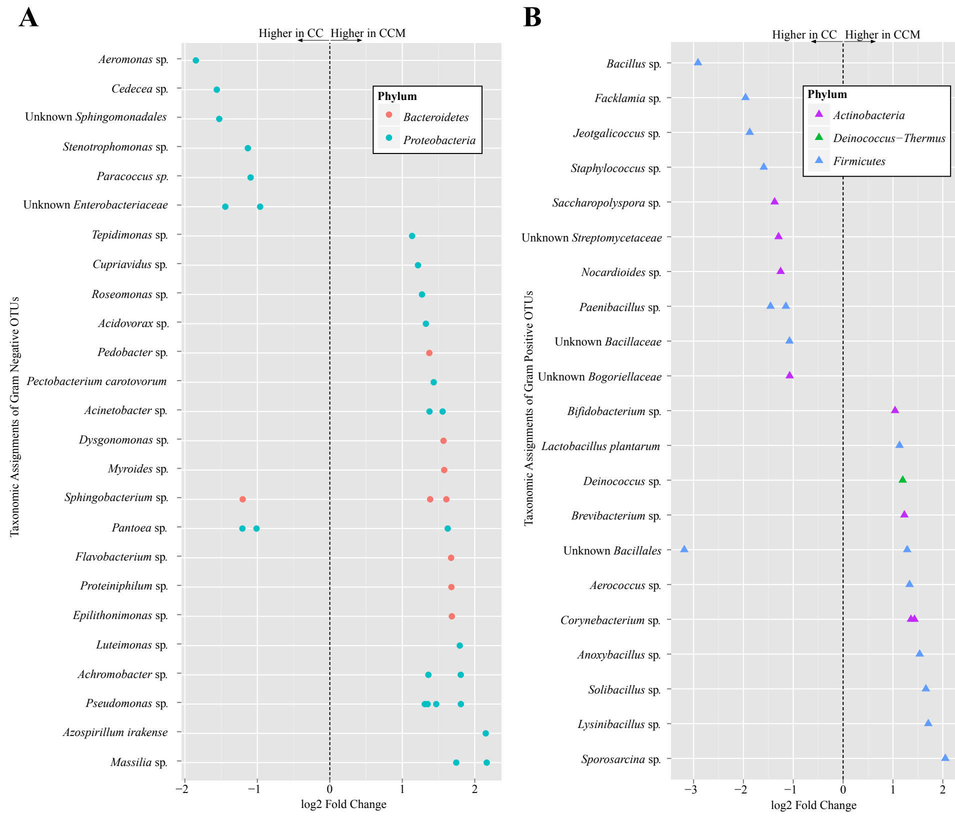

Figure A.7: Overview of relative abundances of OTUs that were statistically significantly different ($p \leq 0.05$) between mentholated Camel Crush (CCM) and non-mentholated Camel Crush (CC). OTUs are colored by Phylum and differentiated by Gram negative (a) and Gram positive (b) classification. A positive log2-fold change value denotes an OTU that is significantly higher in mentholated Camel Crush, while a negative log2-fold change indicates an OTU that is significantly higher in non-mentholated Camel Crush. The dotted line and arrows highlight the conversion in log2-fold change from negative to positive values.

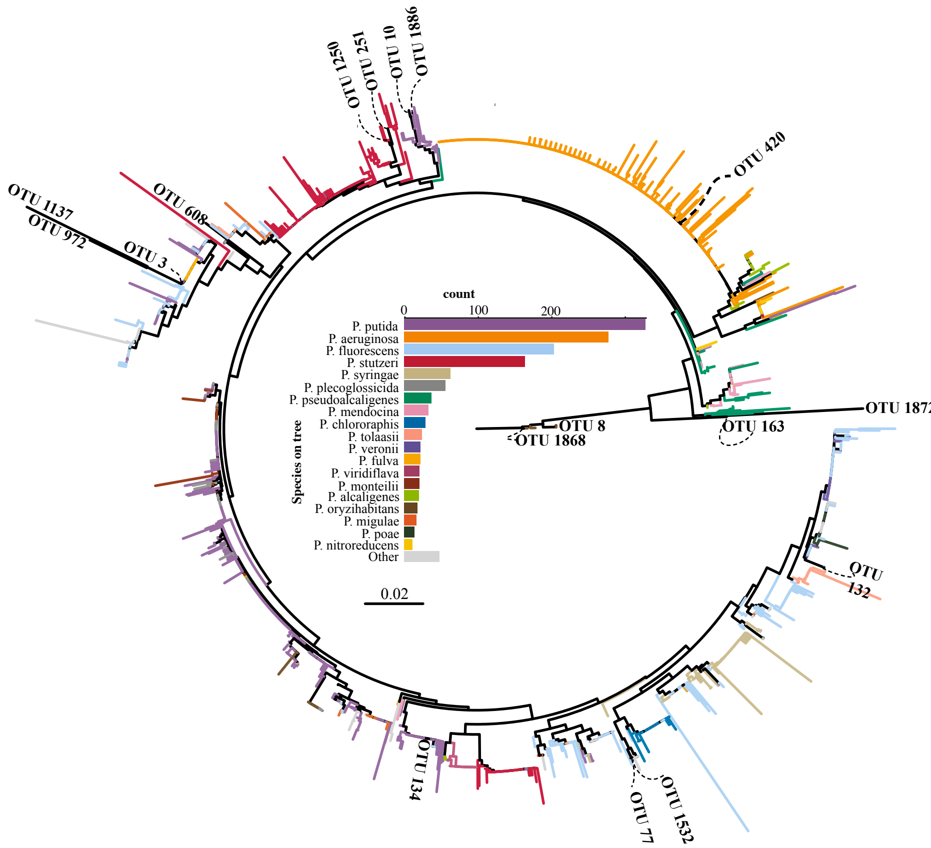

Figure A.8: *Pseudomonas* phylogenetic tree.

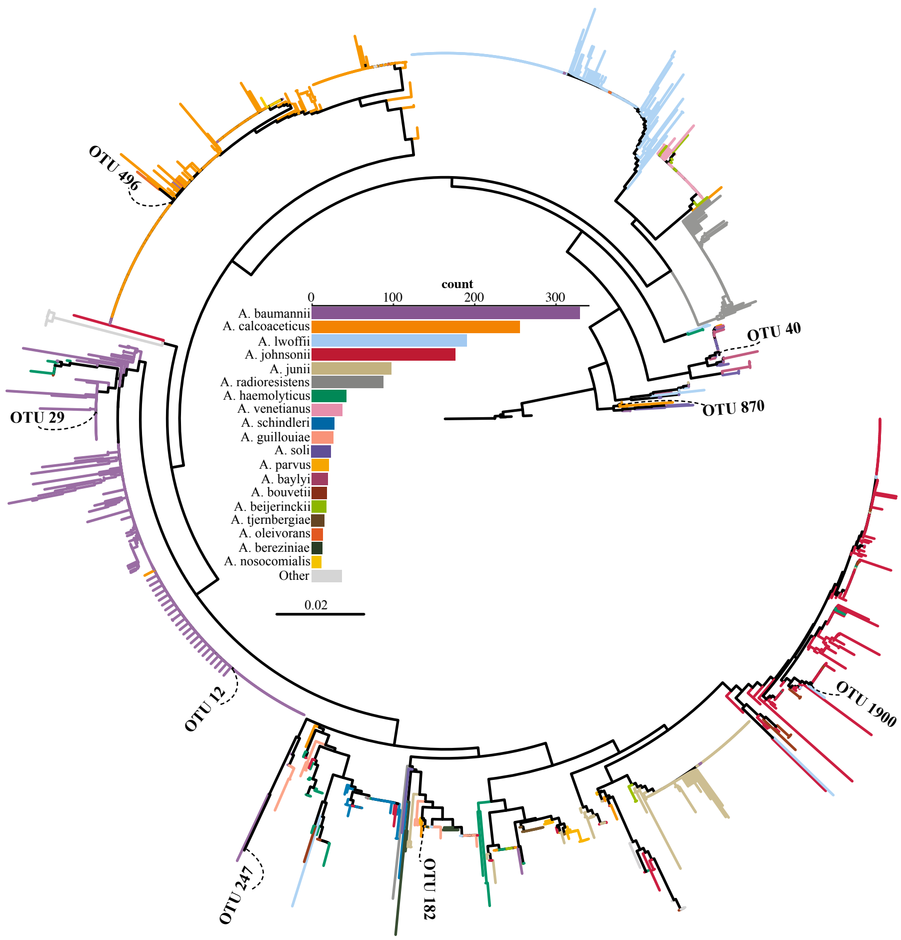

Figure A.9: *Acinetobacter* phylogenetic tree.

.

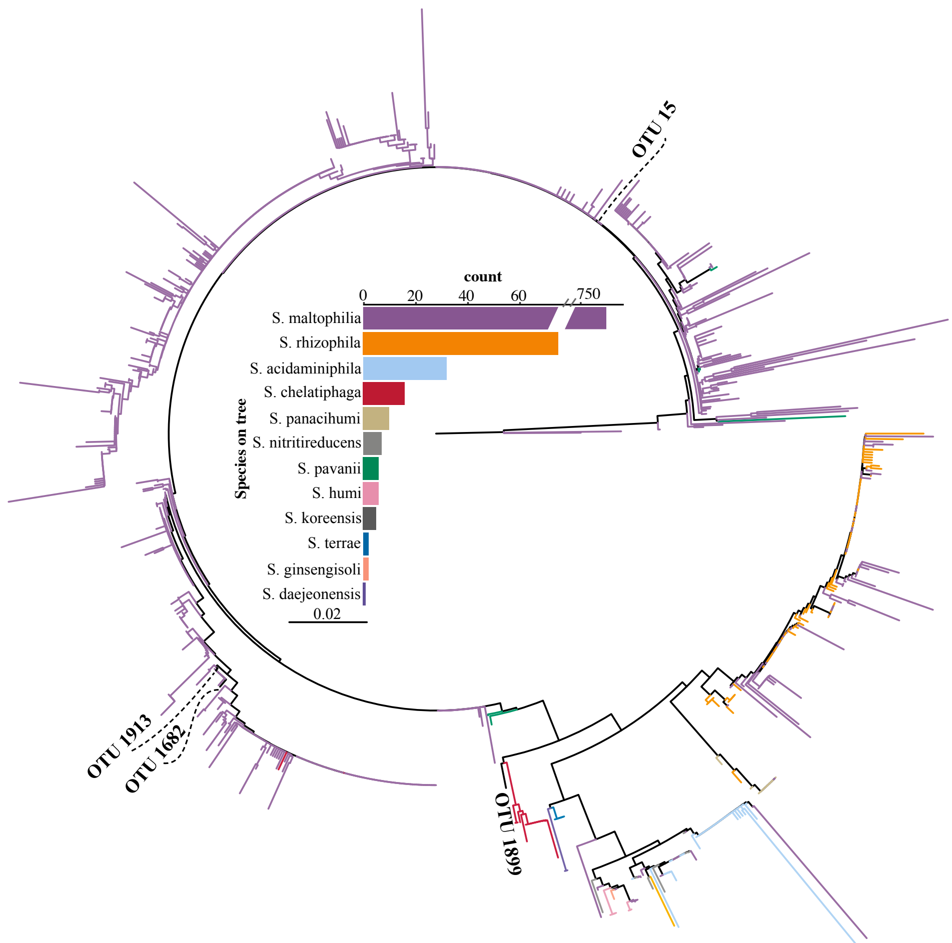

Figure A.10: *Stenotrophomonas* phylogenetic tree.

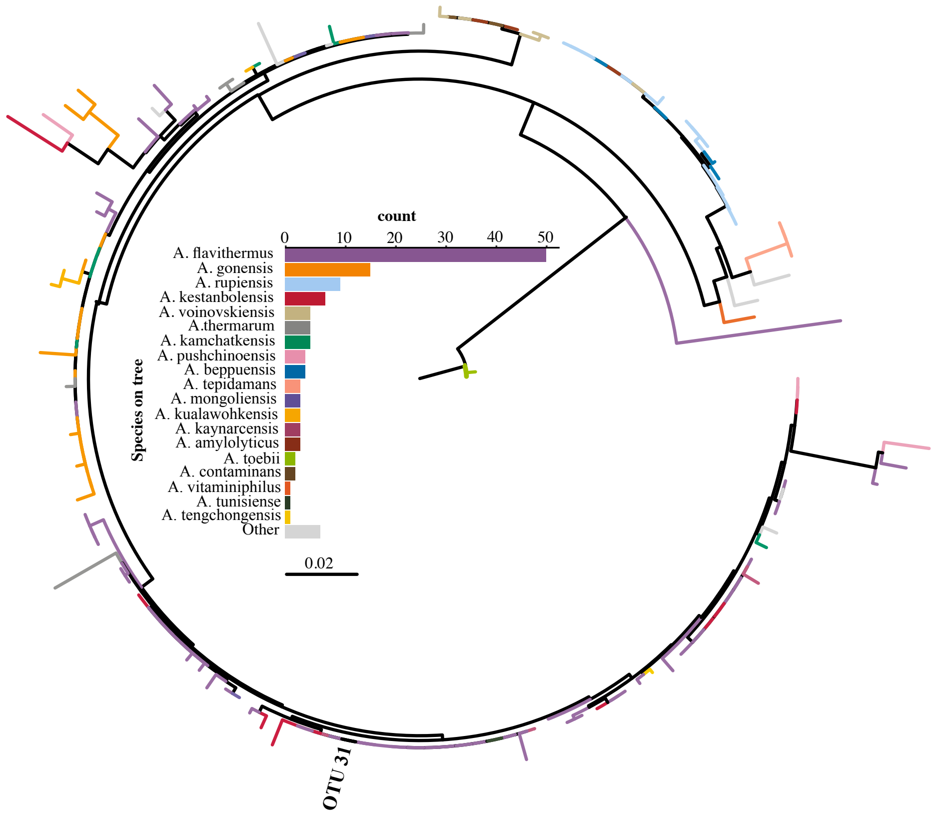

Figure A.11: *Anoxybacillus* phylogenetic tree.

.

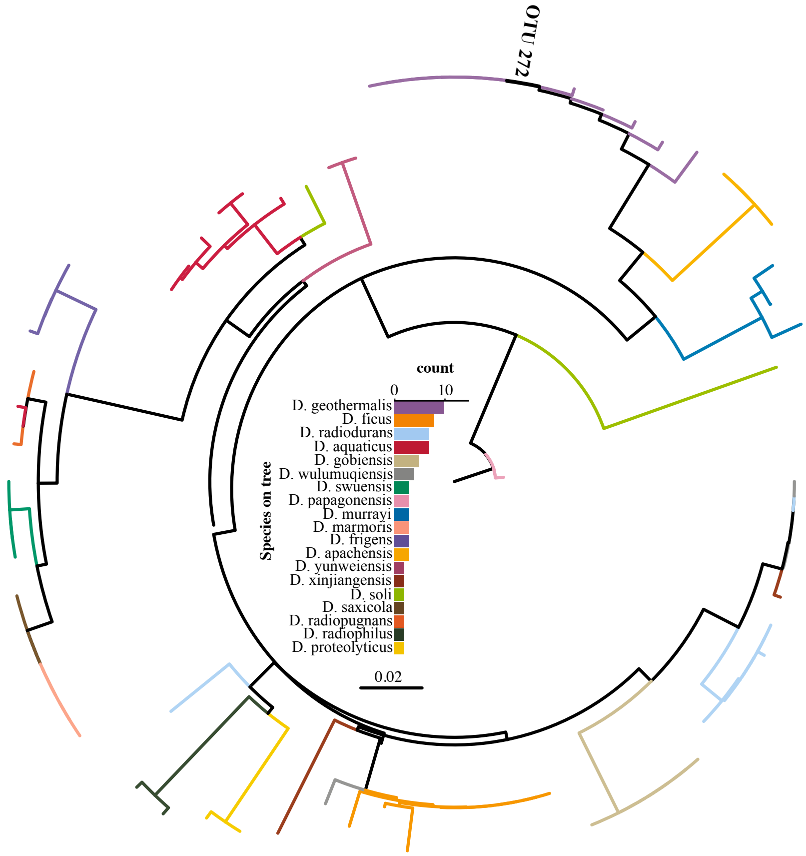

Figure A.12: *Deinococcus* phylogenetic tree.

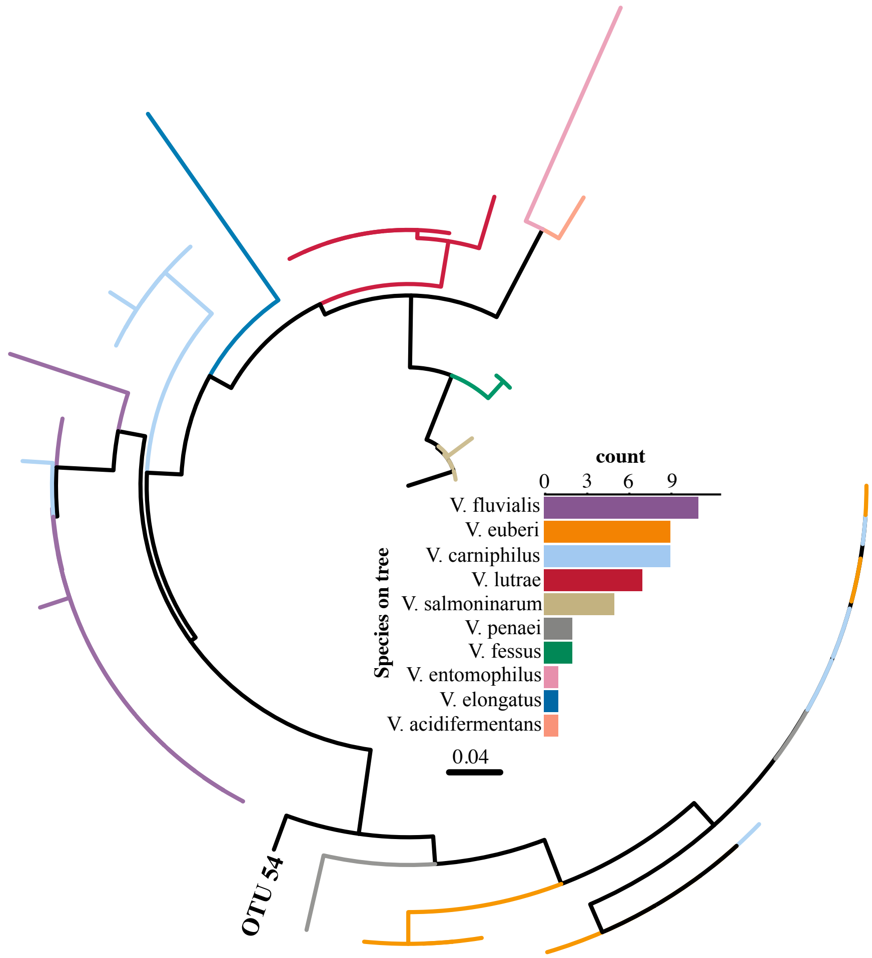

Figure A.13: *Vagococcus* phylogenetic tree.

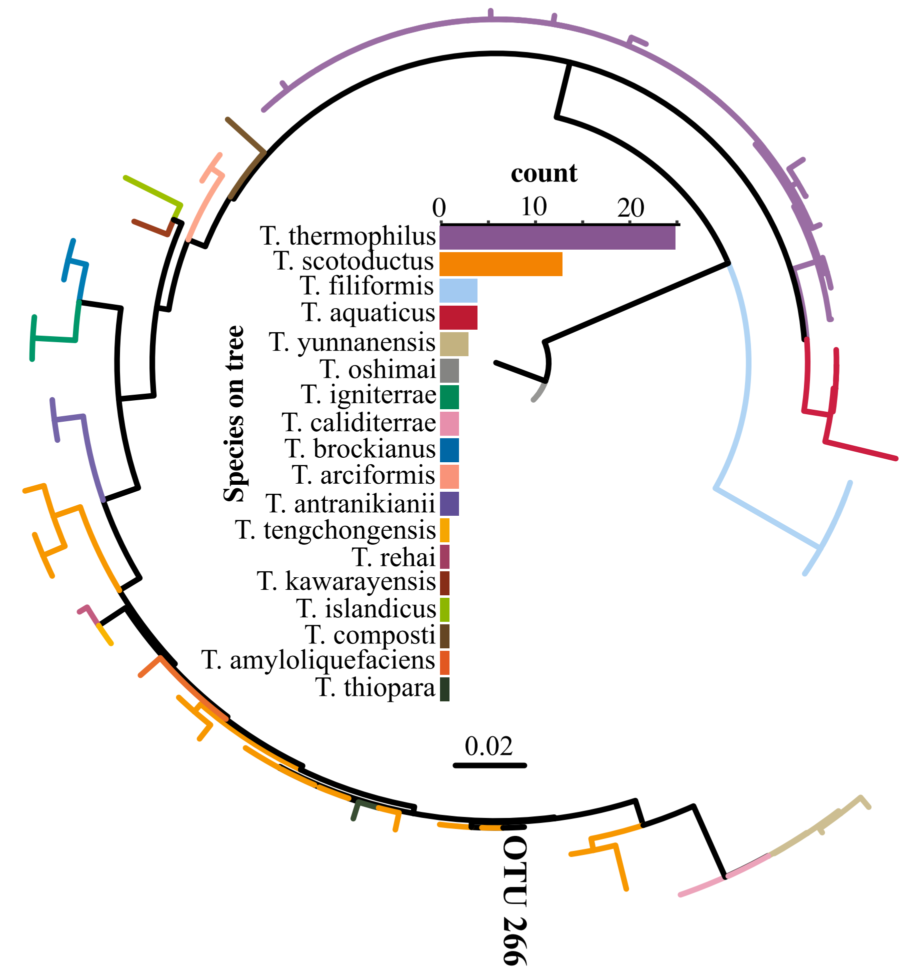

Figure A.14: *Thermus* phylogenetic tree.

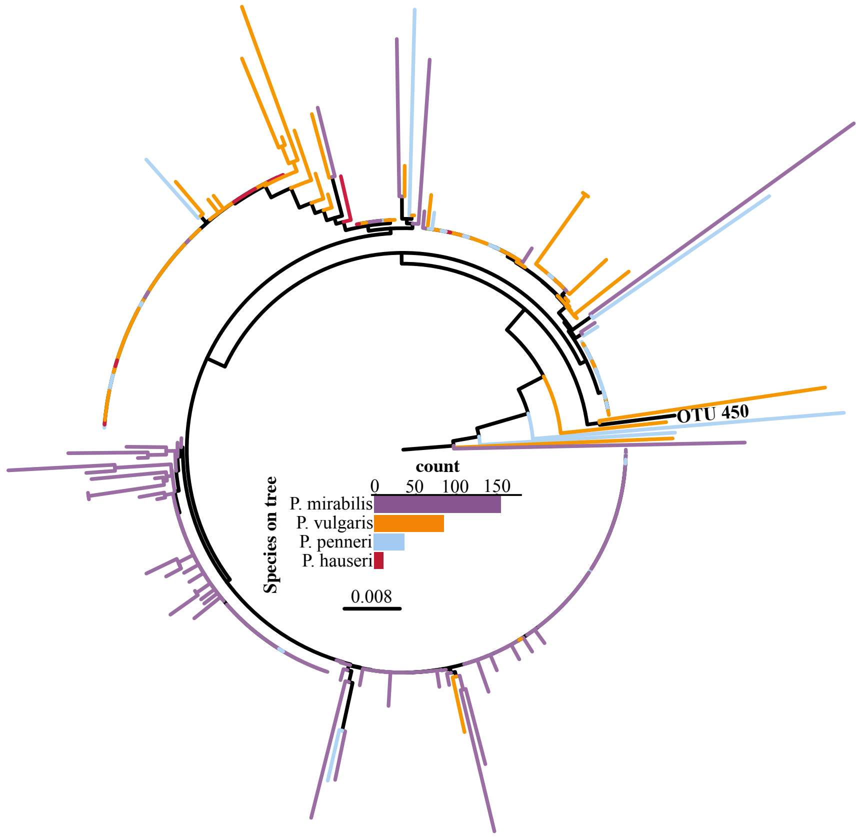

Figure A.15: *Proteus* phylogenetic tree.

Tables

Table A.1: Descriptions of cigarette products tested.

Cigarette product	**Menthol status**	**Abbreviation**
Camel King filters	Non-menthol	CK
Camel King filters	Mentholated (custom)[a]	CKM
Camel Crush	Non-menthol[b]	CC
Camel Crush	Mentholated (user)[c]	CCM
Newport Menthol Box	Mentholated (manufacturer)[d]	NMB

[a]Mentholated at The Battelle Public Health Center for Tobacco Research
[b]Camel Crush capsule within the filter was not crushed
[c]Camel Crush capsule within the filter was crushed in the laboratory prior to DNA extraction
[d]Commercially mentholated by the manufacturer

Appendix B: Temporal Variations in Cigarette Tobacco Bacterial Community Composition & Tobacco Specific Nitrosamine Content are Influenced by Brand and Storage Conditions

Jessica Chopyk, Suhana Chattopadhyay, Prachi Kulkarni, Eoghan M. Smyth, Lauren E. Hittle, Joseph N. Paulson, Mihai Pop, Stephanie S. Buehler, Pamela I. Clark, Emmanuel F. Mongodin and Amy R. Sapkota. Temporal variations in cigarette tobacco bacterial community composition and tobacco-specific nitrosamine content are influenced by brand and storage conditions. *Frontiers in Microbiology*, 8, 2017.

Abstract

Tobacco products, specifically cigarettes, are home to microbial ecosystems that may play an important role in the generation of carcinogenic tobacco-specific nitrosamines (TSNAs), as well as the onset of multiple adverse human health effects associated with the use of these products. Therefore, we conducted time-series experiments with five commercially available brands of cigarettes that were either commercially mentholated, custom-mentholated, user-mentholated, or non-

mentholated. To mimic user storage conditions, the cigarettes were incubated for 14 days under three different temperatures and relative humidities (i.e., pocket, refrigerator, and room). Overall, 360 samples were collected over the course of 2 weeks and total DNA was extracted, PCR amplified for the V3V4 hypervariable region of the 16S rRNA gene and sequenced using Illumina MiSeq. A subset of samples ($n = 32$) was also analyzed via liquid chromatography with tandem mass spectrometry for two TSNAs: N-nitrosonornicotine (NNN) and 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK). Comparative analyses of the five tobacco brands revealed bacterial communities dominated by *Pseudomonas*, *Pantoea*, and *Bacillus*, with *Pseudomonas* relatively stable in abundance regardless of storage condition. In addition, core bacterial operational taxonomic units (OTUs) were identified in all samples and included *Bacillus pumilus*, *Rhizobium* sp., *Sphingomonas* sp., unknown *Enterobacteriaceae*, *Pantoea* sp., *Pseudomonas* sp., *Pseudomonas oryzihabitans*, and *P. putida*. Additional OTUs were identified that significantly changed in relative abundance between day 0 and day 14, influenced by brand and storage condition. In addition, small but statistically significant increases in NNN levels were observed in user- and commercially mentholated brands between day 0 and day 14 at pocket conditions. These data suggest that manufacturing and user manipulations, such as mentholation and storage conditions, may directly impact the microbiome of cigarette tobacco as well as the levels of carcinogens.

Introduction

The tobacco microenvironment within cigarettes is home to complex mixtures of chemicals, metals, salts, trace pesticides, alkaloids, and commercial additives (e.g., menthol and sweeteners; [496, 497]. In fact, over 5,000 components have been identified in tobacco and over 6,000 in tobacco smoke, many of which are carcinogenic toxins [497, 498]. Among the potentially harmful constituents of tobacco are bacteria, fungi, and their microbially derived toxins [496, 499, 500]. Multiple studies have shown that bacteria can not only survive the low moisture content of tobacco but also withstand the harsh smoking process [444, 449, 450]. Specifically, species of *Bacillus*, *Kurthia*, and *Mycobacterium* have been successfully recovered *in vitro* from cigarette filters, smoked filters, paper, and tobacco microparticulates [444, 449, 450].

In addition, molecular techniques to assay the bacterial diversity of tobacco products have identified hundreds of bacterial species present in cured tobacco leaves [501–503], cigarettes [451], and smokeless tobacco brands [504]. These comprise species from the families *Pseudomonadaceae*, *Staphylococcaceae*, *Lactobacillaceae*, *Enterobacteriaceae*, *Enterococcaceae*, *Aerococcaceae*, *Corynebacteriaceae*, among others, and include potential human and respiratory pathogens [451, 501–503]. Furthermore, tobacco and tobacco smoke have been shown to harbor microbial derived toxins and secondary metabolites [500]. For instance, lipopolysaccharide, a potent inflammatory endotoxin of gram-negative bacteria, was identified as a bioactive component of cigarette smoke and a suggested cause of respiratory diseases among smokers [448, 500, 505]. These microbial components of the cigarette may be inhaled during use and deposited into the lung and oral cavity, where they may directly impact the health of the user.

Prior to packaging within the cigarette wrapper, tobacco is influenced heavily by bacteria. This occurs largely during the curing process, a necessary part of cigarette production, whereby tobacco leaves are dried generally by flue (e.g., Virginia tobacco), air (e.g., Burley tobacco), or sun (e.g., Oriental tobacco) to improve their color, flavor, and aroma [506]. During the curing stage, the amount of tobacco specific nitrosamines (TSNAs), carcinogens derived from the nitrosation of tobacco alkaloids, increases significantly [507]. This is suggested to be, in part, due to certain nitrate and nitrite reducing bacterial species present on or in the tobacco leaves [508]. High temperatures and relative humidities have been shown to be key factors that contribute to increasing levels of TSNAs throughout curing [509, 510] and storage [511, 512] of tobacco. TSNA levels in smokeless tobacco brands have also been shown to be influenced by storage conditions, with high levels of TSNAs associated with storage for 4 weeks at room and high temperatures ($> 37\,^{\circ}$C), but not low temperature ($4\,^{\circ}$C [513]). This may be due to changing bacterial diversity within these products.

Microbial populations are often dynamic and influenced by surrounding environmental conditions [501, 514]. For instance, changes in temperature, pH, and nutrient availability throughout the Toscano cigar tobacco fermentation cycle were shown to be associated with changes in the bacterial community composition of these products [501]. In addition, storage conditions have also been found to influence microbial exposures of tobacco users. For example, cigarettes kept at high humidity have been characterized by increased levels of fungi [495]. However, to our knowledge there is no literature describing the longitudinal effects of varying storage conditions (e.g., temperature and relative humidity) on the bacterial diversity of cigarettes. Therefore, this study aimed to utilize high throughput 16S rRNA gene sequencing to investigate the bacterial community composition of five cigarette brands over 14 days at average room, refrigerator, and pocket conditions to identify potential trends in overall bacterial diversity and in specific operational taxonomic units (OTUs). In addition, a subset of samples was tested for levels of two TSNAs [N-nitrosonornicotine (NNN) and 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK)] at pocket and refrigerator conditions over time.

Materials and Methods

Sample Collection and Treatment

Five different cigarette brands (including three distinct lots per brand) were analyzed in this study. Camel Crush, regular, fresh (CC; R.J. Reynolds Tobacco Co., Winston-Salem, NC, USA) and Newport Menthols (NMB; Lorillard Tobacco Co., Greensboro, NC, USA) were purchased from tobacco stores in College Park, MD, USA. CC cigarettes, where the capsule within the filter was not crushed, were considered non-mentholated, while those where the capsule was crushed to release a menthol-containing solution into the cigarette filter were considered user-mentholated (CCM). Camel full flavor, hard pack, king (CK; R.J. Reynolds Tobacco Co., Winston-Salem, NC, USA) were provided by our collaborators at The Battelle

Public Health Center for Tobacco Research (Columbus, OH, USA) along with a custom mentholated version (CKM) as described in MacGregor et al. [452]. To reflect normal user storage conditions cigarettes were subjected to 14 days of three different experimental storage conditions: pocket (25C and 30% relative humidity), refrigerator (5 °C and 18% relative humidity), and room (20 °C and 50% relative humidity). Subsets of cigarettes ($n = 6$) were sampled from each brand for DNA extraction and 16S rRNA amplification prior to onset of the experimental condition (day 0), after 5 days, after 9 days, and after 14 days for each condition (Table B.1).

DNA extraction

Total DNA extraction was adapted from procedures previously published [453, 454]. Cigarettes were dissected separately under sterile conditions and 0.2 g of tobacco was removed and aseptically placed in Lysing Matrix B tubes (MP Biomedicals, Solon, OH, USA). To achieve an effective enzymatic lysis, 1 mL of ice cold 1X molecular grade PBS buffer (Gibco by Life Technologies, Grand Island, NY, USA), 5 μl lysozyme from chicken egg white (10 mg/ml, Sigma-Aldrich, St. Louis, MO, USA), 5 μl lysostaphin from *Staphylococcus staphylolyticus* (5 mg/ml, Sigma-Aldrich, St. Louis, MO, USA) and 15 μl of mutanolysin from *Streptomyces globisporus* ATCC 21553 (1 mg/ml, Sigma-Aldrich, St. Louis, MO, USA) was added to the tubes containing cigarette tobacco and lysing matrix. Tubes were then incubated at 37 °C for 30 min followed by the addition of a second enzymatic cocktail consisting of 10 μl Proteinase K (20 mg/ml, Invitrogen by Life Technologies, Grand

Island, NY, USA) and 50 μl of SDS (10% w/v, BioRad). Incubation was repeated at 55 °C for 45 min. Samples were then subjected to mechanical lysis via the FastPrep Instrument FP-24 (MP Biomedicals, Santa Ana, CA, USA) at 6.0 m/s for 40 s followed by centrifugation for 3 min at 10,000 rcf. Subsequent DNA was purified using the QIAmp DSP DNA mini kit 50, v2 (Qiagen, Valencia, CA, USA), according to the manufacturer's protocol. Negative extraction controls were included to ensure that no exogenous DNA contaminated the samples during extraction. DNA quality control/quality assurance was performed using spectrophotometric measurements on a NanoDrop (Thermo Scientific, Waltham, MA, USA), as well as gel electrophoresis.

16S rRNA gene PCR amplification and sequencing

Using a dual-indexing strategy for multiplexed sequencing developed at the Institute for Genome Sciences and described in detail elsewhere [241], the V3V4 hypervariable region of the 16S rRNA gene was PCR-amplified and sequenced on the Illumina MiSeq (Illumina, San Diego, CA, USA). PCR reactions were set-up in 96-well microtiter plates using the 319F (ACTCCTACGGGAGGCAGCAG) and 806R (GGACTACHVGGGTWTCTAAT) universal primers, each with a linker sequence required for Illumina MiSeq 300 bp paired-ends sequencing, and a 12-bp heterogeneity-spacer index sequence to minimize biases associated with low-diversity amplicons sequencing [241, 515]. Reactions were performed with Phusion High-Fidelity DNA polymerase (Thermo Fisher, Waltham, MA, USA) and 2 ng of template DNA in a total reaction volume of 25 μl. In addition, due to the presence of

PCR inhibitors, an additional 0.375 μl of bovine serum albumin (BSA; 20 mg/ml, Sigma) was added to the PCR reactions. Negative controls without DNA template were performed for each primer pair. A DNA Engine Tetrad 2 thermo cycler (Bio-Rad, USA) was used under the following cycling parameters: 30 s at 98 °C, followed by 30 cycles of 10 s at 98 °C, 15 s at 66 °C, and 15 s at 72 °C, with a final step of 10 min at 72 °C. Successful amplification was confirmed using gel electrophoresis. This was followed by cleanup and normalization via the SequalPrep Normalization Plate kit (Invitrogen Inc., Carlsbad, CA, USA) with 25 ng of 16S PCR amplicons from each sample prior to pooling and 16S rRNA sequencing using the Illumina MiSeq (Illumina, San Diego, CA, USA) according to the manufacturer's protocol.

TSNA analysis

Concentrations of two TSNAs (NNN and NNK) in the unused product were determined for a subset of cigarette samples (n = 32). The subset included two samples taken at day 0 and two samples taken at day 14 at pocket conditions for all five brands. In addition, two samples taken at day 0 and two samples taken at day 14 at refrigerator conditions for CK, CKM, and NMB were included. Samples were stored at -80 °C until analysis. Prior to extraction, the tobacco and the outer wrapper (cut into small pieces) were removed, weighed separately, and then combined for analysis. Filters and the paper encasing them were removed and discarded.

Samples were extracted using a method adopted from those previously published for smokeless tobacco products [516, 517]. Each sample was spiked with deuterated internal standards (NNN-d4 and NNK-d4) and extracted in 30 mL of ammonium acetate on a rotary shaker for 1 h at 250 rpm. Each extract was then filtered with a 0.45 mm syringe filter. Quality control samples, including matrix spikes, were prepared with each batch of samples using 3R4F cigarettes. Extracts were analyzed using liquid chromatography with tandem mass spectrometry (LC-MS/MS). The method detection limit based on average sample tobacco weights was 0.002 mg/g. Matrix spike recoveries averaged 113 $\pm$ 23% for NNN and 110 $\pm$ 9% for NNK.

Sequence quality filtering

After screening 16S rRNA gene reads for low quality bases and short read lengths [241] paired-end read pairs were then assembled using PANDAseq [242], de-multiplexed, trimmed of artificial barcodes and primers, and assessed for chimeras using UCHIME in de-novo mode implemented in Quantitative Insights Into Microbial Ecology (QIIME; release v. 1.9; [243]). The resulting quality trimmed sequences were then clustered de-novo into OTUs with the SILVA 16S database [244] in QIIME [243], with a minimum confidence threshold of 0.97 for the taxonomic assignments. All sequences taxonomically assigned to chloroplasts were removed. To account for uneven sampling depth and to ensure less biases than the standard approach (total sum normalization), data were normalized with metagenomeSeq's cumulative sum scaling when appropriate [251].

Data analysis

Taxonomic assignments of genera were obtained through QIIME [243]. After removing genera whose maximum relative abundance was less than 1%, a heatmap was created and visualized with R version 3.2.2 and vegan heatplus [246]. The core tobacco bacterial microbiome was defined as OTUs present at a minimum fraction of 100% in all tested products with QIIME's compute core microbiome.py script [243] and visualized with Cytoscape [300].

Beta diversity for all brands at all times and conditions was calculated using the Bray-Curtis dissimilarity and compared using Analysis of similarities (ANOSIM) on normalized data (999 permutations) through the R packages: biomformat [249], vegan [248], ggplot2 [183], phyloseq [249]. Beta diversity was also calculated as described above for samples separated by brand.

Diversity was estimated for samples pooled by brand, time point, and condition using the Shannon Index [300] through the R packages: Bioconductor [247], metagenomeSeq [245], vegan [248], phyloseq [249], and fossil [250]. Significance was assessed through Tukey's test at $p \leq 0.05$. To account for uneven sampling depth, diversity was measured with and data rarefied to a minimum sampling depth.

Determination of statistically significant differences ($p \leq 0.001$) in OTU abundance was performed using DESeq2 [458] to compare the NMB brand between day 0 and day 14 at room, pocket, and refrigerator conditions. The significant OTUs ($p \leq 0.001$) were visualized with R version 3.2.2 and R packages ggplot2 [183], vegan [248], and phyloseq [249]. This was repeated for the remaining brands (CC, CCM, CK, CKM), as well as, by product lot.

Results

Sequencing

DNA extraction and sequencing was performed on 360 cigarette samples (Table B.1), with a total of 2,172,847 sequences and an average sequence per sample of 6,262 ($\pm$3,433 SD). A total of 1,985 different bacterial OTUs (97% identity) were identified at an average of 185 OTUs per sample ($\pm$46 SD).

Taxonomic analysis of all cigarette brands

After samples were pooled by brand (CC, CCM, CK, CKM, and NMB), time point (day 0, day 5, day 0, and day 14), and condition (pocket, room, and refrigerator), *Pseudomonas* had the highest relative abundance in all instances, ranging from 7.05 to 11.24%. This was followed by either *Pantoea* (3.58-8.44%) or *Bacillus* (4.58-9.38%) (Figure B.1). These three encompassed the furthest clade to the left of the cladogram (Figure B.1). The second most abundant clade of bacterial genera consisted of *Acinetobacter* (2.16-4.84%), *Enterobacter* (3.09-5.27%), Unknown *Enterobacteriaceae* (2.53-4.76%), and *Sphingomonas* (2.97-5.13%) (Figure B.1).

When samples were pooled by brand (Figure B.2) *Pseudomonas* was significantly ($p \leq 0.05$) higher in relative abundance in CCM compared to CC, CK, and NMB. CCM also had a significantly higher relative abundance of *Pantoea* than CC and a significantly lower relative abundance of *Bacillus* than CC, CK, CKM, and NMB. Furthermore, CKM had significantly higher relative abundance of *Pseudomonas* than CK. NMB had a significantly higher relative abundance of *Pantoea* than CC, CK, CKM.

Within brand condition was also a prominent factor impacting the temporal dynamics of the most abundant genera (Figures B.2). Experimental condition seemed to have little significant effect on the relative abundance of *Pseudomonas* over time. In fact, *Pseudomonas* only significantly changed in one brand, CKM, in which it decreased between day 0 and day 14 at room conditions. The relative abundance of *Bacillus* was only affected by condition in NMB at pocket conditions and CKM at room conditions. For CKM there was a significant increase in *Bacillus* between day 0 and day 9 at room conditions, followed by a decrease between day 9 and day 14 (Figure B.2). For NMB, *Bacillus* decreased in relative abundance between day 0 and day 5 and then stayed relatively unchanged for the remainder of the study.

The relative abundance of *Pantoea* appeared to be more affected by condition, whereas changes in the relative abundance occurred in CC at pocket and room conditions (Figure B.2), in CCM at room conditions, and in NMB at pocket and refrigerator conditions (Figure B.2). Specifically, for NMB there was a significant increase in the relative abundance of *Pantoea* between day 0 and day 14 and between day 0 and day 5 at pocket conditions, with an oscillation downward at day 9. In addition, there was a significant increase in the relative abundance of *Pantoea* between day 0 and day 5 at refrigerator conditions for the same brand.

For CC, the relative abundance of *Pantoea* significantly fluctuated between day 0 and day 5, day 5 and day 9, and day 9 and day 14 at pocket conditions. There was also a significant decrease in *Pantoea* between day 0 and day 9 for CC at room conditions (Figure B.2). This is in contrast to CCM in which there was a significant increase in *Pantoea* between those same times at the same condition (Figure B.2).

The core microbiome, defined for each brand, comprised 26 bacterial OTUs for CC, 24 for CK, 22 for NMB, 20 for CKM, and 16 for CCM (Figure B.3). A comparative analysis of these bacterial OTUs revealed that 11 OTUs were shared among all samples regardless of brand, time, and experimental condition at relative abundances between 1.26% (*Pseudomonas putida*, OTU #3) and 0.83% (*Rhizobium* sp., OTU #11). A comparative analysis of these bacterial OTUs revealed that 11 OTUs were shared among all samples regardless of brand, time, and experimental condition at relative abundances between (*Rhizobium* sp., OTU #11). These included: *B. pumilus* (OTU #5), *Rhizobium* sp. (OTU #11), *Sphingomonas* sp. (OTU #2), unknown *Enterobacteriaceae* (OTU #1969 and #1885), *Pantoea* sp. (OTU #398 and #1904), *Pseudomonas* sp. (OTU #1886), *Pseudomonas oryzihabitans* (OTU #1868 and #8), and *P. putida* (OTU #3) (Figure B.3)

Two OTUs were unique to the core of NMB, *Brevibacterium* sp. (OTU #42) and *Staphylococcus* sp. (OTU #143). Similarly two OTUs were unique to the core of CC, *Novosphingobium* sp. (OTU #27) and unknown *Pseudomonadales* (OTU #13). Only one OTU was unique to CCM, unknown *Enterobacteriaceae* (OTU #2018), and there were no OTUs in the core microbiome unique to CKM and CK. The largest degree of overlap was between NMB, CC, CK, and CKM, which had an additional four OTUs in common amongst their core microbiomes: *Sphingomonas* sp. (OTU #1850), *Methylobacterium* (OTU #28 and #18), and unknown *Aurantimonadaceae* (OTU #23). The two non-mentholated brands (CC and CK) both had *Enterobacter aerogenes* (OTU #1932) amongst their core microbiomes. *Enterobacter* sp. (OTU #4) and *Pseudomonas* sp. (OTU #134) were a part of the core in all brands except the custom mentholated Camel Kings (CKM). The custom mentholated and non-mentholated Camel Kings (CKM and CK) along with the NMB each had *Staphylococcus* sp. (OTU #7). *Terribacillus* sp. (OTU #6) and *Enterobacter* sp. (OTU #107) were a part of the core microbiomes of all brands except commercially mentholated NMB. CKM, CK, and CC all had *B. clausi* (OTU #9), whereas *Pseudomonas* (OTU #10) and *Sphingomonas* (OTU #1287) were in the core microbiomes of NMB, CK, and CC. In addition, *Methylobacterium* (OTU #36) was present in CC and CKM.

Beta and alpha diversity of all brands

PCoA plots of the Bray-Curtis computed beta diversity for all brands revealed the largest significant clustering by brand (R = 0.25, p = 0.001) followed by lot (R = 0.21, p = 0.001) (Figure B.4, B.5), with NMB observed clustering away from the other brands. There was no significant clustering by time point or condition (Figure B.5). When separated into distinct brands, each had minimum clustering by time point and lot (Figure B.6), particularly for CK (R = 0.1762, p = 0.001), CKM (R = 0.1703, p = 0.001), and NMB (R = 0.198, p = 0.001) lots.

All brands appeared to have fluctuating bacterial diversity, assessed through Shannon indices, during the length of the experiment (day 0, day 5, day 9, and day 14; Figure B.7). However, the only significant change in Shannon indices was between day 0 and day 9 in NMB at pocket conditions in which diversity increased ($p \leq 0.05$) (Figure B.7).

Comparative analysis of OTUs by condition between day 0 and day 14

Within the experimental conditions tested, non-mentholated CC had the greatest amount of OTUs (19 OTUs) that were significantly different in relative abundance between day 0 and day 14 at refrigerator conditions (Figure B.8). Of these, 61% (11 OTUs) were at higher relative abundance at day 14 and the rest (8 OTUs) were at higher relative abundance at day 0. This was followed by pocket conditions, which had 15 OTUs significantly different between day 0 and day 14, with 73% (11 OTUs) at higher relative abundance at day 0. Room conditions had the least amount of significantly different OTUs (nine OTUs) between time points, of which 55% (five OTUs) were at higher relative abundance at day 14.

In contrast to its non-mentholated counterpart, CCM had the greatest number of OTUs (20 OTUs) that were significantly different between day 0 and day 14 at room conditions (Figure B.8), with 70% (14 OTUs) at higher abundance at day 0 compared to day 14. Refrigerator conditions had the second largest amount of OTUs (eight OTUs) that were significantly different between day 0 and day 14 for

CCM, all of which had higher relative abundance at day 0. At pocket conditions there were only three OTUs that were significantly different between time points. Two were at higher abundance at day 0 and one was at a higher abundance at day 14.

Similar to CCM, non-mentholated Camel Kings (CK) had the largest amount of OTUs at significantly different relative abundances (34 OTUs) between day 0 and day 14 at room conditions (Figure B.9). However, unlike CCM, 67% of the OTUs (23 OTUs) were at higher relative abundance at day 14. The second condition that produced the most OTUs with significantly different relative abundances (24 OTUs) between time points was refrigerator conditions; 54% of OTUs (13 OTUs) at higher relative abundance at day 0. Pocket conditions had the smallest amount of OTUs at significantly different relative abundances (14 OTUs). Of these, 57% (eight OTUs) were at higher relative abundance at day 14.

Custom-mentholated Camel Kings (CKM) at pocket conditions had the most OTUs (43 OTUs) that were significantly different in relative abundance between day 0 and day 14 (Figure B.9). However, only one of these OTUs was at higher relative abundance at day 0, *Bacillus* (189). The remaining 98% (42 OTUs) were at higher relative abundance at day 14. Room conditions had 38 OTUs at significantly different relative abundance between time points, all at higher relative abundance at day 14. Finally, refrigerator conditions had the least number of OTUs (11 OTUs) that were significantly different in relative abundance between day 0 and day 14, all of which were higher at day 14.

There were only five OTUs at statistically significantly different ($p \leq 0.001$)

relative abundances between day 0 and day 14 among the different conditions for NMB (Figure B.10). Pocket and refrigerator conditions each had two OTUs that were at significantly different relative abundance between time points. In both conditions one of the OTUs was at higher relative abundance at day 0 and one higher at day 14. Room conditions had only one OTU, significantly higher at day 14.

Comparison of OTUs significantly different in relative abundance between day 0 and day 14 in CC and CCM

Interestingly, there were some OTUs that were shared between CC and CCM. For instance, *Massilia* (OTU #2052) was at higher relative abundance at day 0 in CCM at room temperature, however the same OTU was at higher relative abundance at day 14 in CC at the same condition. Additionally, *Olivibacter* (OTU #162) was at a higher relative abundance at day 14 at room temperature in CCM and higher relative abundance at day 0 at refrigerator conditions in CC. *Pantoea* (OTU #253) was at lower relative abundance at day 14 at both room temperature and refrigerator conditions in CCM, but was at higher relative abundance at day 14 for CC at refrigerator conditions.

Comparison of OTUs significantly different in relative abundance between day 0 and day 14 in CK and CKM

Several OTUs significantly different in relative abundance between day 0 and day 14 were shared between CK and CKM including, *Nesterenkonia* (OTU #288) and *Acinetobacter* (OTU #1900), which were at a higher relative abundance at day 14 at room temperature and refrigerator conditions, respectively. *Acinetobacter calcoaceticus* (OTU #40) was higher in both brands at day 14 at pocket conditions. *Achromobacter* (OTU #49) was higher at day 14 for CKM at room temperature, but higher at day 0 for CK at refrigerator conditions. *Sphingobacterium* (OTU #88) was at higher relative abundance at day 0 in CK at refrigerator conditions, but was higher at day 14 in CKM at room temperature. *Sphingobacterium* (OTU #161) was also at higher relative abundance at day 0 in CK at refrigerator conditions, but was higher at day 14 in CKM at pocket conditions.

Comparative analysis of significantly different OTUs by lot

Because there was clustering by lot for CK, CKM, and NMB (Figure B.11), we determined the OTUs that were statistically significantly different between lots, regardless of condition or time point. For NMB, lot 4K01 clustered away from lot 4C03 and lot 4C17, therefore 4K01 was compared with 4C03 and 4C17. There were 11 OTUs at statistically significantly different ($p \leq 0.001$) relative abundances between 4K01 and 4C03 (Figure B.11). Of these 11 OTUs, 5 from the phylum *Actinobacteria* had a higher relative abundance in 4C03 compared to 4K01 including *Ochrobactrum* sp. (OTU #110), *Marinactinospora* sp. (OTU #113), *Brevibacterium* sp. (OTU #153), *Saccharopolyspora* sp. (OTU #281), and *Enteractinococcus* sp. (OTU #157). The remaining 6 from the phylum *Firmicutes* had a higher relative abundance in 4K01: *Caldalkalibacillus* sp. (OTU #261), *Kurthia* sp. (OTU #180), *Lactobacillus mucosae* (OTU #234), *Lactobacillus* sp. AB5262 (OTU #118), *Lactobacillus fermentum* (OTU #227), and *Pediococcus* sp. (OTU #50). These OTUs were also significantly higher in relative abundance in 4K01 compared to 4C17 along with *Pantoea* sp. (OTU #725 and #229), *Bacillus coagulans* (OTU #99), *Geobacillus* sp. (OTU #119), *Lactobacillus* sp. (OTU #301), *Paenibacillus* sp. (OTU #196), and *Streptomyces* sp. KP17 (OTU #52).

For CK and CKM, lot A4 clustered away from lots Il and L1, therefore A4 was compared with Il and L1. There were 21 OTUs at statistically significantly different ($p \leq 0.001$) relative abundances between A4 and L1 for CK, of which 8 were at higher abundance in L1 and 13 were at higher abundance in A4 (Figure B.12). Six of the OTUs of higher abundance in A4 were also at higher abundance when comparing lot A4 with lot Il, *Tistrella* (OTU #439), *Azospirillum irakense* (OTU #167), *Pseudomonas* (OTU #177), *Pectobacterium* (OTU #25), *Wautersiella* (OTU #192), *Alcaligenes* (OTU #117), *Pediococcus* (OTU #50), and *Rheinheimera* (OTU #137). There were 33 OTUs at statistically significantly different ($p \leq 0.001$) relative abundances between lots A4 and L1 for CKM with 22 at higher relative abundance in lot L1 and 11 at higher relative abundance in lot A4 (Figure B.13). Of those at higher relative abundance at A4, 10 OTUs were shared with those at higher abundance in lot A4 when compared to lot 1l: *Enteractinococcus* (OTU #157), *Arthrobacter* (OTU #69), *Pseudomonas* (OTU #10), *Aeromonas* (OTU #237), *Rhizobium* (OTU #198), *Pectobacterium carotovorum* (OTU #48), *Achromobacter* sp. HJ-31-2 (OTU #16), *Pseudomonas* (OTU #245), *Cloacibacterium* (OTU #72), *Pediococcus* (OTU #50). Additionally, of those with higher abundance in 1l two OTUs were shared with those at higher abundance in L1, *Anoxybacillus* (OTU #31) and *Planococcaceae* (OTU #152). Additionally, *Pediococcus* (OTU #50) was at higher relative abundance in A4 for both CKM and CK.

Analysis of TSNA content

N-nitrosonornicotine levels were significantly higher ($p \leq 0.05$) at pocket conditions from day 0 to day 14 for NMB and CCM (Figure B.14). NNK levels increased as well from day 0 to day 14 for NMB and CCM; however, these results were not statistically significant (Figure B.14). Only CKM, CK, and NMB were tested for these TSNAs at refrigerator conditions (Figure B.15). NNN tended to increase in all brands from day 0 to day 14, while NNK levels tended to decrease in all brands over the same time. However, these differences were also not statistically significant.

Discussion

Fresh tobacco leaves are colonized by a variety of microorganisms [448] that can be altered by tobacco-processing methods following harvest, such as curing and fermentation [501, 518]. However, the effect of storage conditions on the bacterial constituents of tobacco after packaging within a cigarette was previously unknown. Here, we showed that the dominant bacterial genera, specific OTUs, and the concentration of TSNAs are related to the cigarette brand and the storage condition.

Pseudomonas was the most abundant bacterial genera detected in all brands, time points, and conditions (Figure B.1). This corroborates with previous findings suggesting that *Pseudomonas* was a dominant bacterial genera on aged and unaged flue-cured tobacco leaves [502, 503]. In addition, storage condition seemed to have little significant effect on the relative abundance of *Pseudomonas* over time, whereas *Pantoea* appeared more sensitive to storage condition (Figure B.2). This may be indicative of differing colonization strategizes between the two genera [519].

In addition, OTUs of *Pseudomonas* and *Pantoea* were both defined as members of the core microbiome of all products (Figure B.3). *Pseudomonas* and *Pantoea* are gram-negative, which may contribute to the high levels of lipopolysaccharide found in cigarette tobacco and smoke [500]. Both genera also contain species that are associated with disease in humans [520–522]. These include *P. putida* and *P. oryzihabitans*, which are generally considered opportunist pathogens [473], particularly *P. oryzihabitans* which has been linked to bacteremia, peritonitis, and pneumonia [523].

Many of the members of the core microbiome were also present in the core microbiome defined for air-cured burley tobacco including *Pantoea*, *Pseudomonas*, *Sphingomonas*, and *Bacillus* [510]. Despite this agreement in core members between products, our results showed there was some divergence in bacterial community composition between brands of cigarettes. For instance, NMB had a larger degree of the genera *Staphylococcus* (Figures B.1, B.2). A well known pathogenic species of *Staphylococcus*, *S. aureus*, has been found to have higher nasal carriage rates in smokers [524, 525]. This bacteria has also been shown to increase biofilm formation and host cell adherence in the presence of cigarette smoke [526].

In addition, levels of the TSNA NNN were found in this study to increase significantly between day 0 an day 14 at pocket conditions for NMB and CCM, a potential public health concern given that carcinogen exposure has been found to correlate with the levels of TSNAs in smokeless tobacco products. Specifically, it has been reported that NNK and NNN nitrosamine biomarkers in the urine of smokeless tobacco users increased 32 and 12%, respectively, for every one-unit (μg/g wet wt) increase in NNK and NNN levels within their smokeless tobacco products [527]. In tobacco, bacteria have been identified that are capable of reducing nitrate to nitrite for the formation of TSNAs, including species of *Bacillus*, *Staphylococcus*, and *Corynebacterium* [501, 528]. However, we are unable to determine with these data whether the OTUs present in our samples have such capabilities or were responsible for the observed increases in TSNAs levels. In addition, the type of tobacco and the subsequent nitrate availability, may factor into the ecology of TSNA production. For example, flue-cured and sun-cured tobaccos have been reported to have lower nitrate levels than air-cured [529, 530]. The different tobacco varietals are also blended in various assortments by commercial manufacturers, often with additives (e.g., menthol), thereby resulting in varied nitrate levels and potentially different arrangements of the microbial community compositions [531]. Keeping these variables in mind, more work is necessary to explore the potential connections between nitrate reducers in tobacco, such as *Lactobacillus fermentum* [532], and increasing levels of TSNAs.

Several studies have suggested that smoking tobacco products can alter the microbiome of the user by disrupting commensal bacterial populations, enabling the invasion of pathogens in an otherwise occupied niche [533–535]. However, the relationship between the microbiome of the products and the user is just beginning to be explored. Here, we present evidence that cigarette tobacco is a dynamic microenvironment, with significant changes in members of the dominant bacterial genera, specific OTUs, and the concentration of TSNAs dependent on brand, storage conditions, and time. In addition, bacterial genera present at high abundance in these products are also those common to respiratory infections among smokers [520, 521, 524, 525]. Although the capabilities of bacterial growth in cigarette filters post-smoking have been demonstrated [450], our data currently cannot ascertain whether the bacteria found in the cigarette tobacco are capable of colonizing the oral and/or lung cavities of the user. Despite this uncertainty, their potential role in TSNA and toxin production makes them a potentially appropriate target for intervention.

Figures

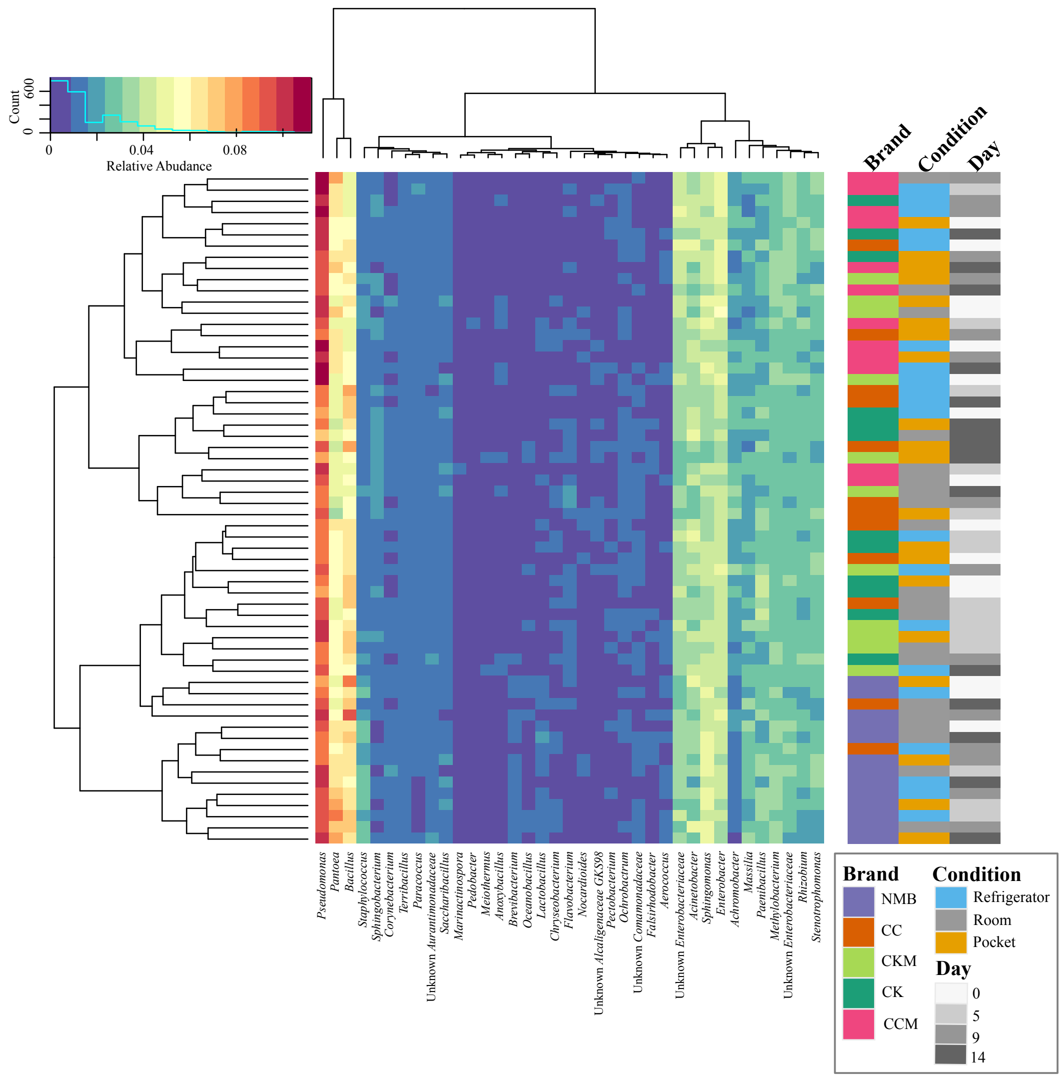

Figure B.1: Bacterial community composition of cigarette products over time and differing storage conditions. Heat map showing the relative abundances of the most dominant bacterial genera identified (> 1%) in cigarette products pooled by brand (CK, CKM, CC, CCM, and NMB), time point (day 0, day 5, day 9, and day 14), and experimental storage condition (room, pocket, refrigerator) and denoted by colored rectangles. Clustering using Manhattan distance of the pooled samples represented by the dendrograms.

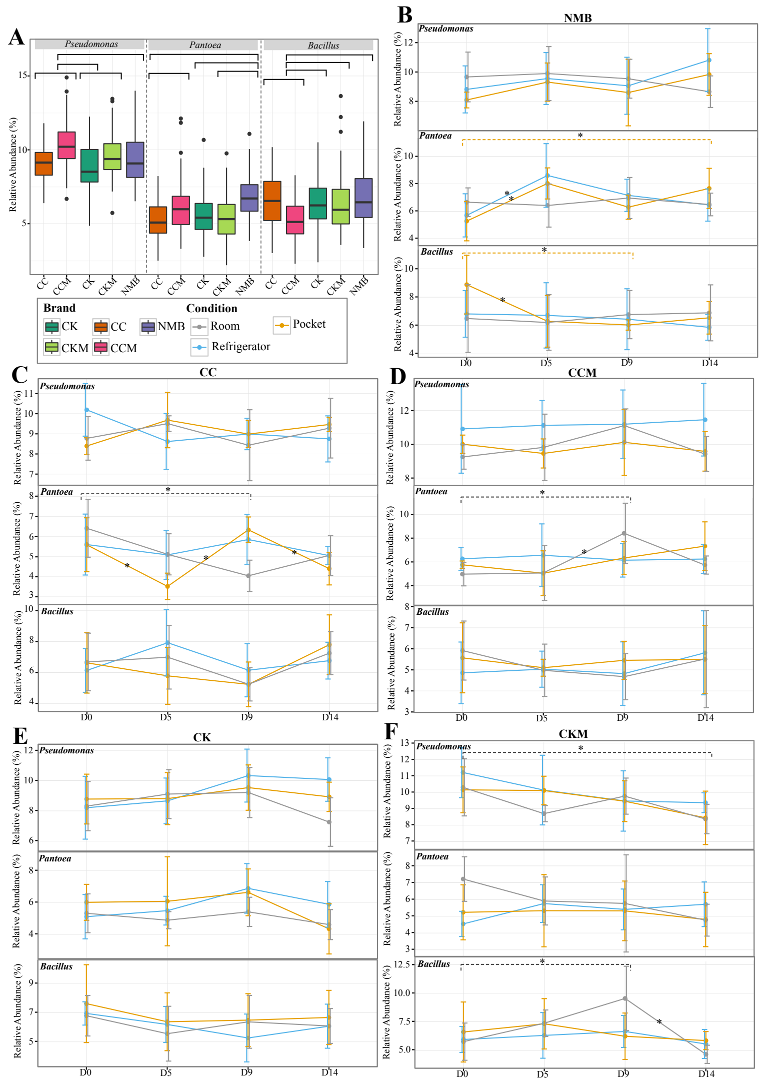

Figure B.2: Comparison of the relative abundance of the most dominant genera. (A) Boxplot of the relative abundance of the most dominant genera (*Pseudomonas*, *Pantoea*, *Bacillus*) in each brand (CC, CCM, CK, CKM, NMB). Brands are colored as follows: CC (dark orange), CCM (pink), NMB (purple), CKM (light green), CK (dark green). Line graphs with standard deviations of relative abundances of the same genera within brand (B) NMB, (C) CC, (D) CCM, (E) CK, (F) CKM over time and experimental storage condition. Experimental storage condition denoted by color as follows: room (gray), pocket (orange), and refrigerator (light blue). Asterix on lines and dashed brackets represent significant changes between time points. Significance determined by an alpha level of 0.05.

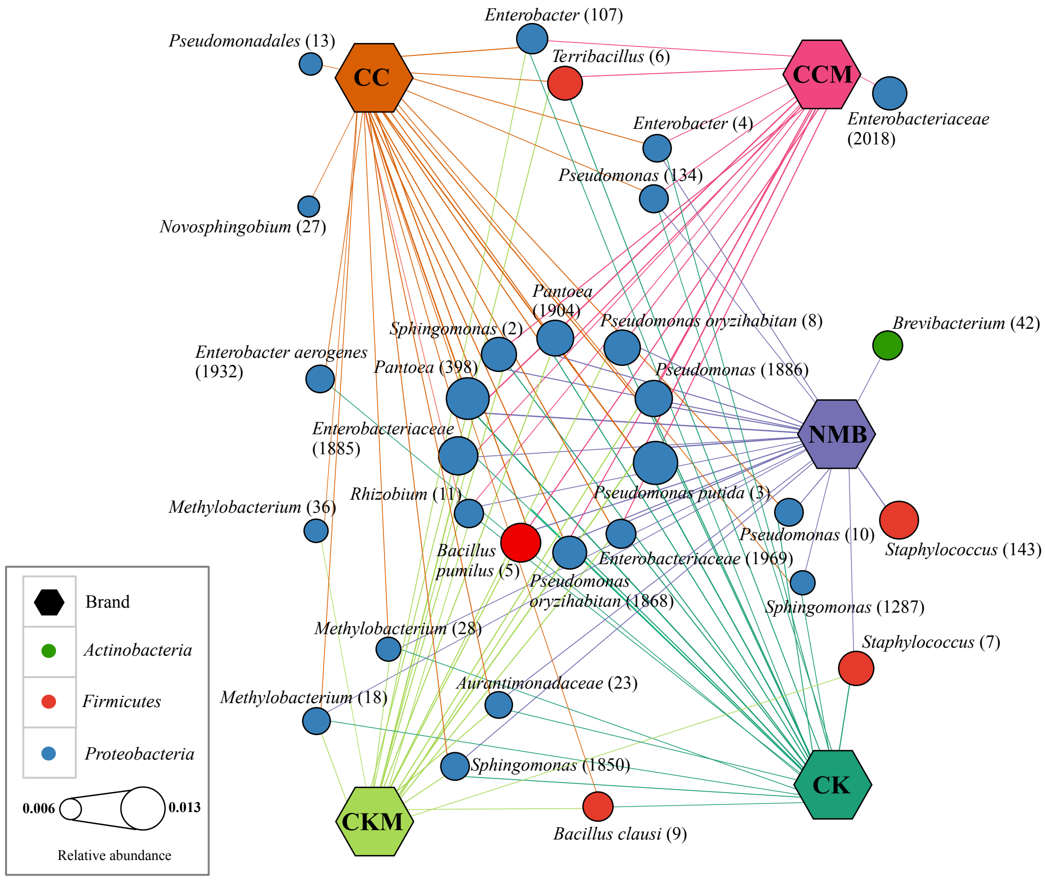

Figure B.3: Network of core bacterial operational taxonomic units (OTUs) in each brand. Cytoscape network visualizing the OTUs (circles) that occur in 100% of samples (hexagon) pooled by brand and labeled with taxonomic ID and assigned OTU number in parantheses. Brands are colored as follows: CC (dark orange), CCM (pink), NMB (purple), CKM (light green), CK (dark green). Size of OTU nodes represent relative abundance and are colored by phylum: *Actinobacteria* (green), *Firmicutes* (red), and *Proteobacteria* (blue). Inner circle of nodes represent eleven core bacterial OTUs that occur in all samples, regardless of brand, time point or condition.

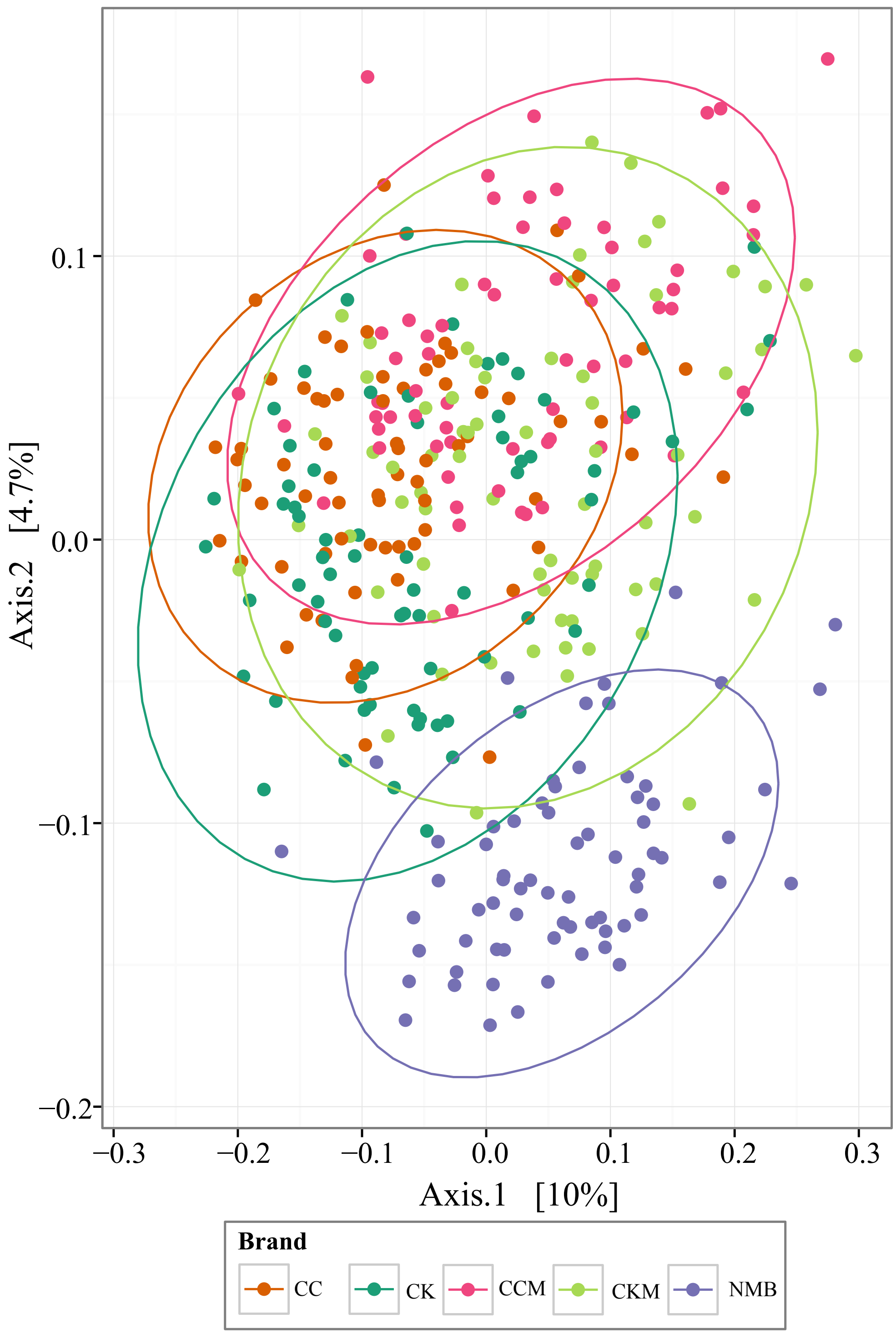

Figure B.4: PCoA analysis plots of Bray-Curtis computed distances between cigarette products. Colored by brand and tested with ANOSIM (R = 0.28, p = 0.001). Ellipses are drawn at 95% confidence intervals for product brand.

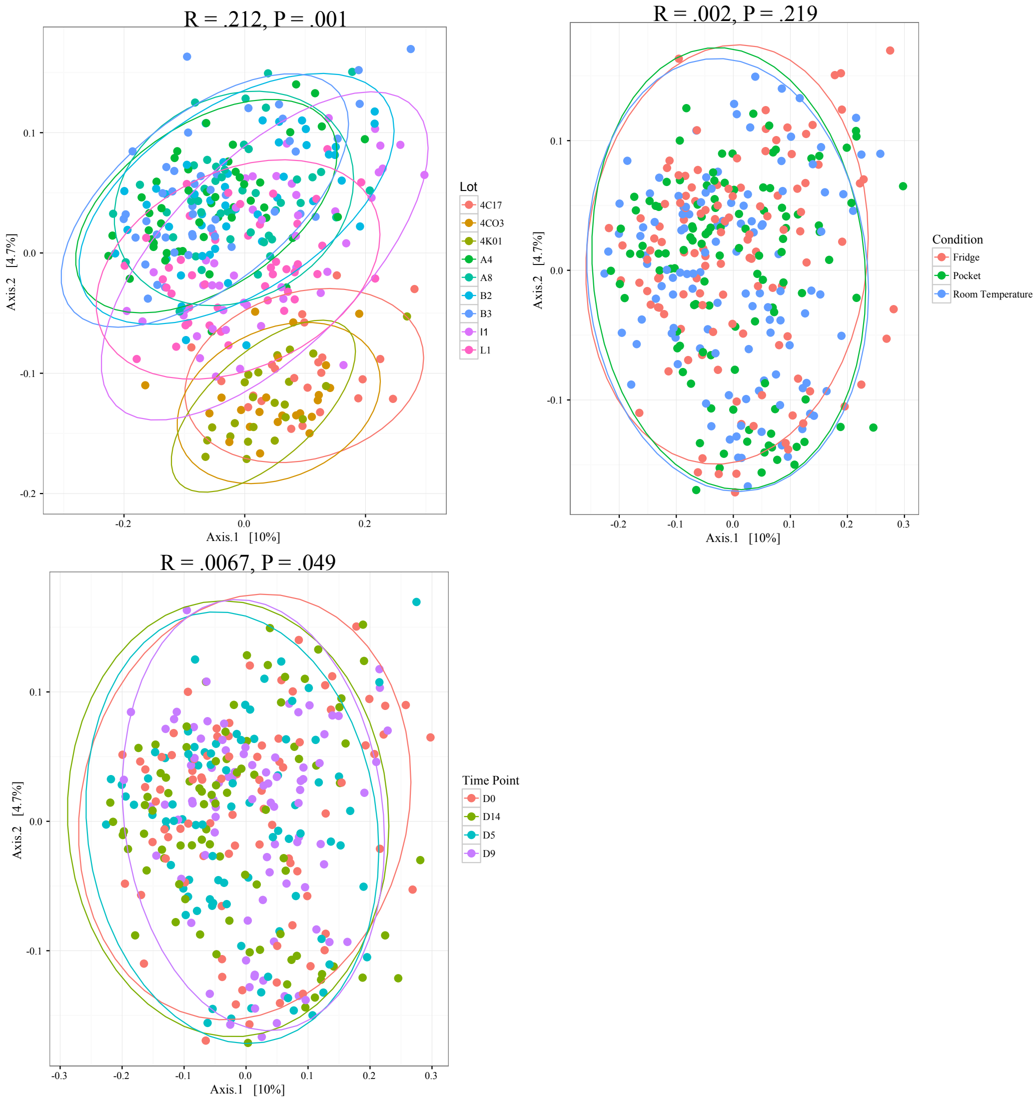

Figure B.5: PCoA analysis plots of Bray-Curtis computed distances between cigarette products. Colored by lot, condition, and time point and tested with ANOSIM. Ellipses are drawn at 95% confidence intervals.

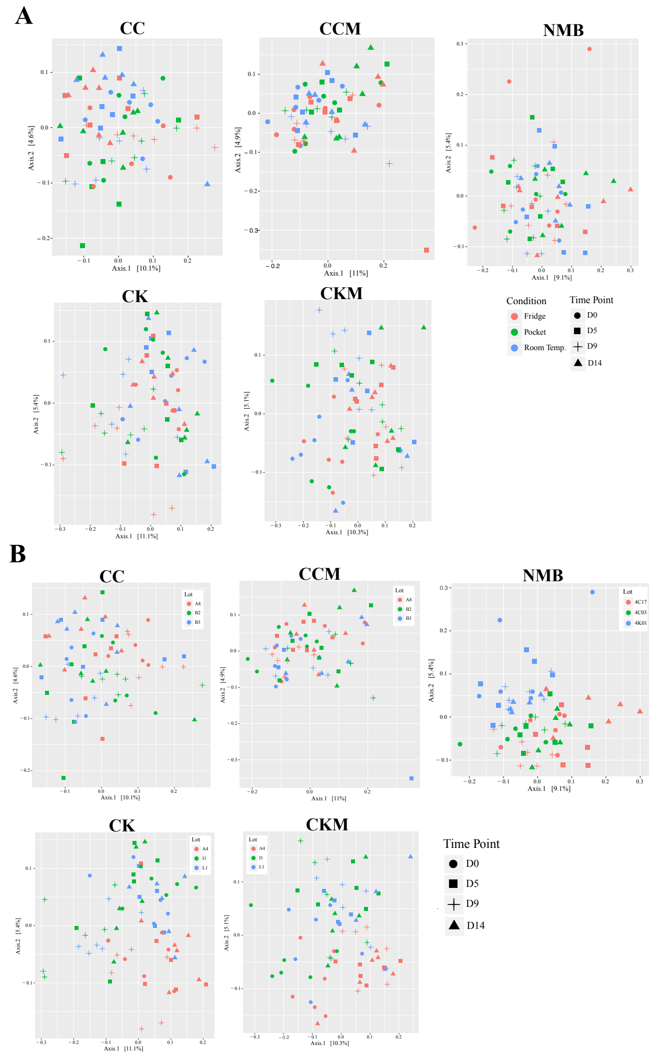

Figure B.6: PCoA analysis plots of Bray-Curtis computed distances between individual cigarette products CC, CCM, CK and CKM. Colored by (A) condition or (B) lot. Shapes represent time points day 0 (circle), day 5 (square), day 9 (plus sign), and day 14 (triangle). Tested with ANOSIM on individual variables: CC by time point (R=0.06607, p=0.002) and lot (R=0.06454, p=0.001); CCM by time point (R= 0.08513, p=0.001) and lot (R=0.06454, p=0.001); NMB by time point (R = .062, p = .001), condition (R = .062, p =.002), and lot (R = .198, p = .001); CK by time point (R = 0.1007, p =.002) and lot (R = 0.1762, p =.001); and CKM by lot (R = 0.1703, p =.001) and time point (R = 0.1865, p =.001).

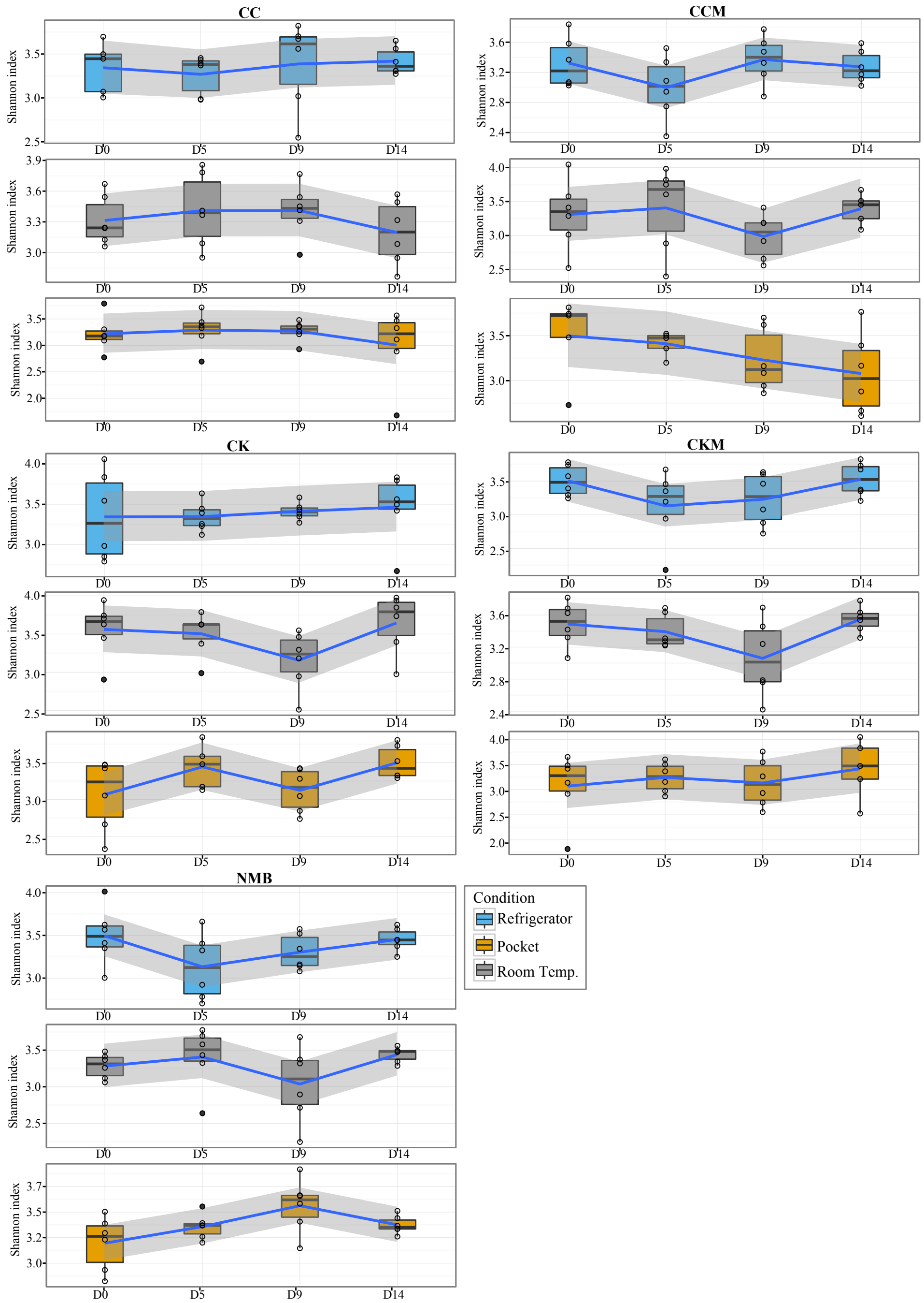

Figure B.7: Alpha diversity comparison by brand, condition, and time point. Box plots showing Shannon diversity index for mentholated Camel Crush, Camel Crush, mentholated Camel Kings, Camel Kings, and Newport Menthols over time point (Day: D0, D5, D9, D14) and experimental storage condition (pocket, room temperature and refrigerator). The blue line represents a locally estimated scatterplot-smoothed (LOESS) calibration curve with the grey areas representing 95% confidence intervals.

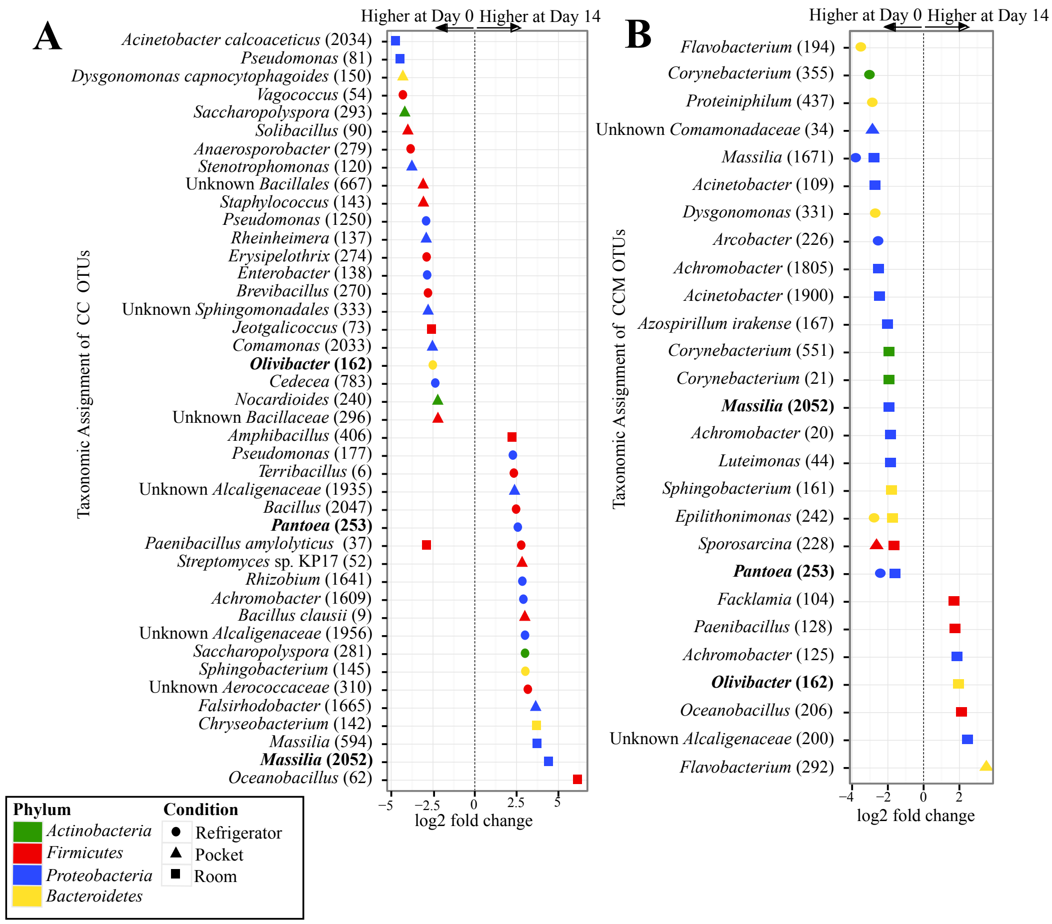

Figure B.8: Overview of relative abundances of bacterial OTUs that were statistically significantly different ($p \leq 0.001$) between day 0 and day 14 for refrigerator (circle), room (square), and pocket (triangle) conditions for (A) non-mentholated Camel Crush (CC) and (B) mentholated Camel Crush (CCM). OTUs are colored by Phylum and shaped by experimental condition. A positive log2-fold change value denotes an OTU that is significantly higher at day 14, while a negative log2-fold change indicates an OTU that is significantly higher at day 0. The dotted line and arrows highlight the conversion in log2-fold change from negative to positive values. Bolded text refers to OTUs that occur in both (A,B).

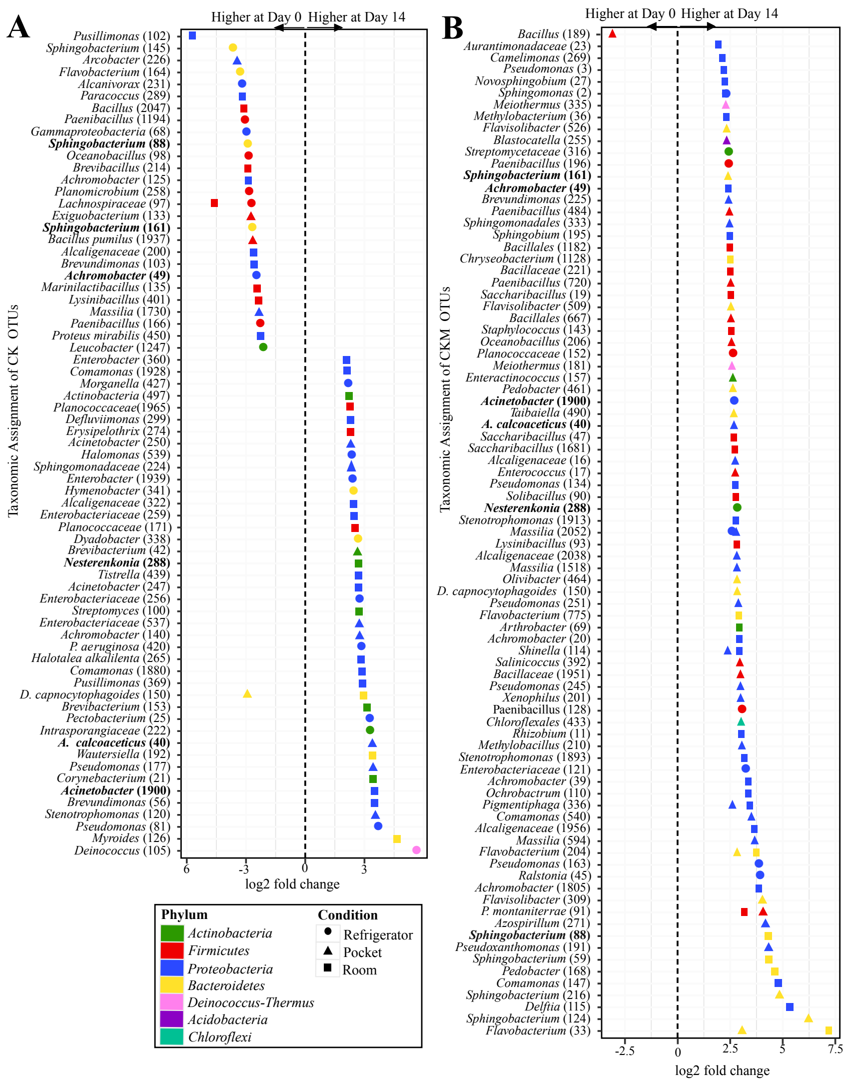

Figure B.9: Overview of relative abundances of bacterial OTUs that were statistically significantly different ($p \leq 0.001$) between day 0 and day 14 for refrigerator (circle), room temperature (square), and pocket (triangle) conditions for (A) non-mentholated Camel Kings (CK) and (B) mentholated Camel Kings (CKM). OTUs are colored by Phylum and shaped by experimental condition. A positive log2-fold change value denotes an OTU that is significantly higher at day 14, while a negative log2-fold change indicates an OTU that is significantly higher at day 0. The dotted line and arrows highlight the conversion in log2-fold change from negative to positive values. Bolded text refers to OTUs that occur in both (A,B).

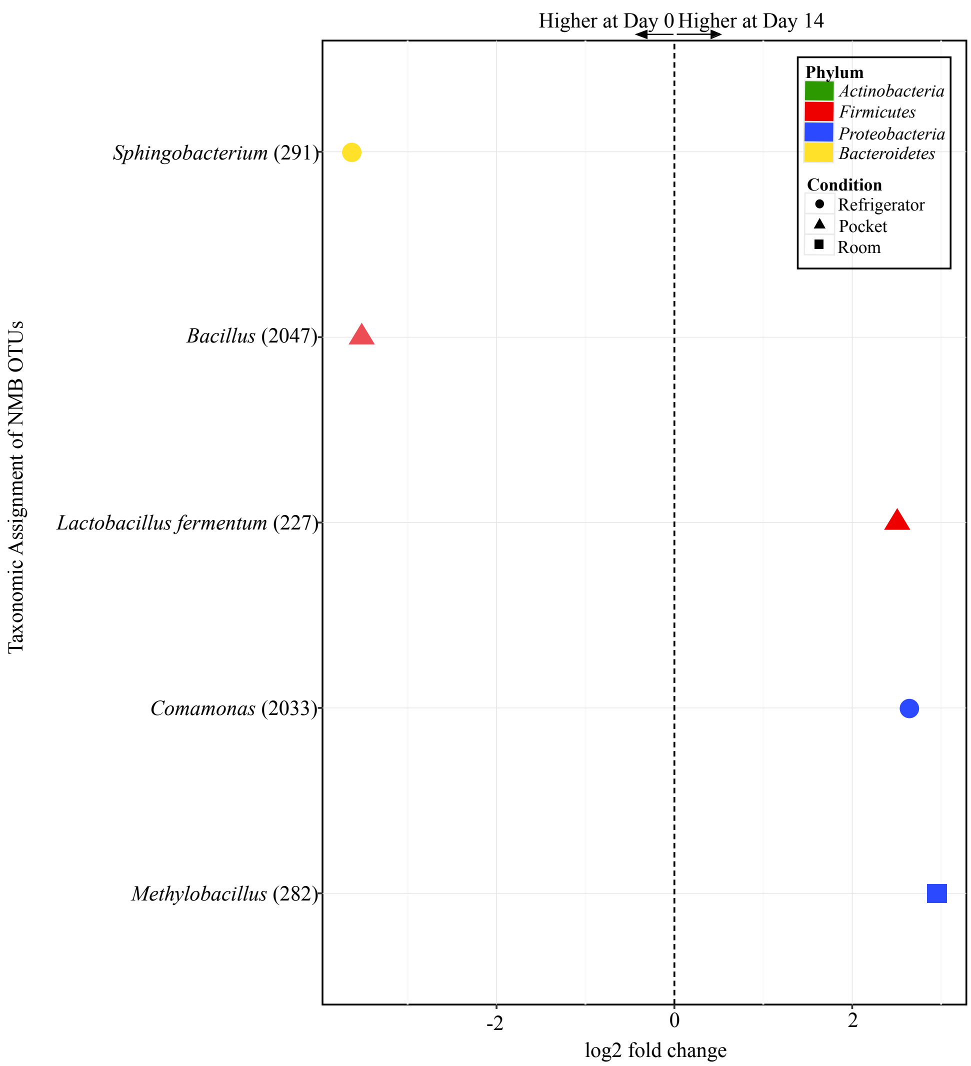

Figure B.10: Overview of relative abundances of bacterial OTUs that were statistically significantly different ($p \leq 0.001$) between day 0 and day 14 for refrigerator (circle), room temperature (square), and pocket (triangle) conditions for Newport Menthols (NMB). OTUs are colored by Phylum and shaped by experimental condition. A positive log2-fold change value denotes an OTU that is significantly higher at day 14, while a negative log2-fold change indicates an OTU that is significantly higher at day 0. The dotted line and arrows highlight the conversion in log2-fold change from negative to positive values.

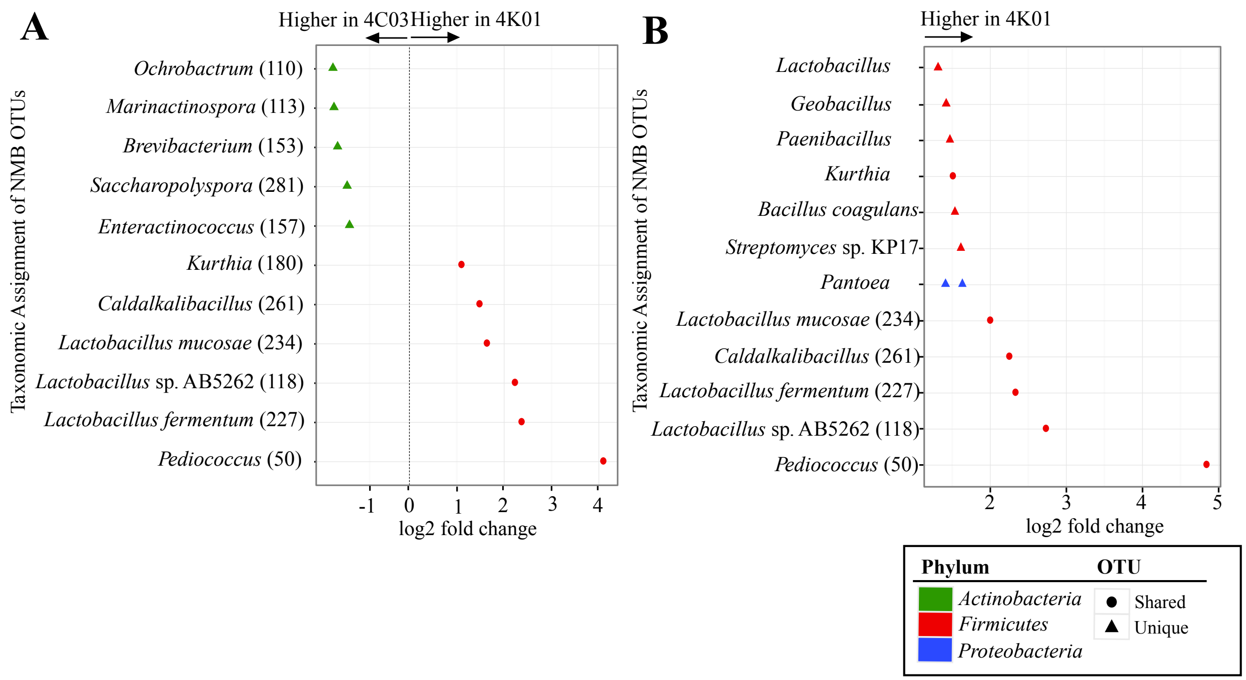

Figure B.11: Overview of relative abundances of bacterial OTUs that were statistically significantly different ($p \leq 0.001$) between lots (A) AC03 and 4K01 and lots (B) 4C17 and 4K01 for Newport Menthols (NMB). The dotted line highlights the conversion in log2-fold change from negative to positive values.

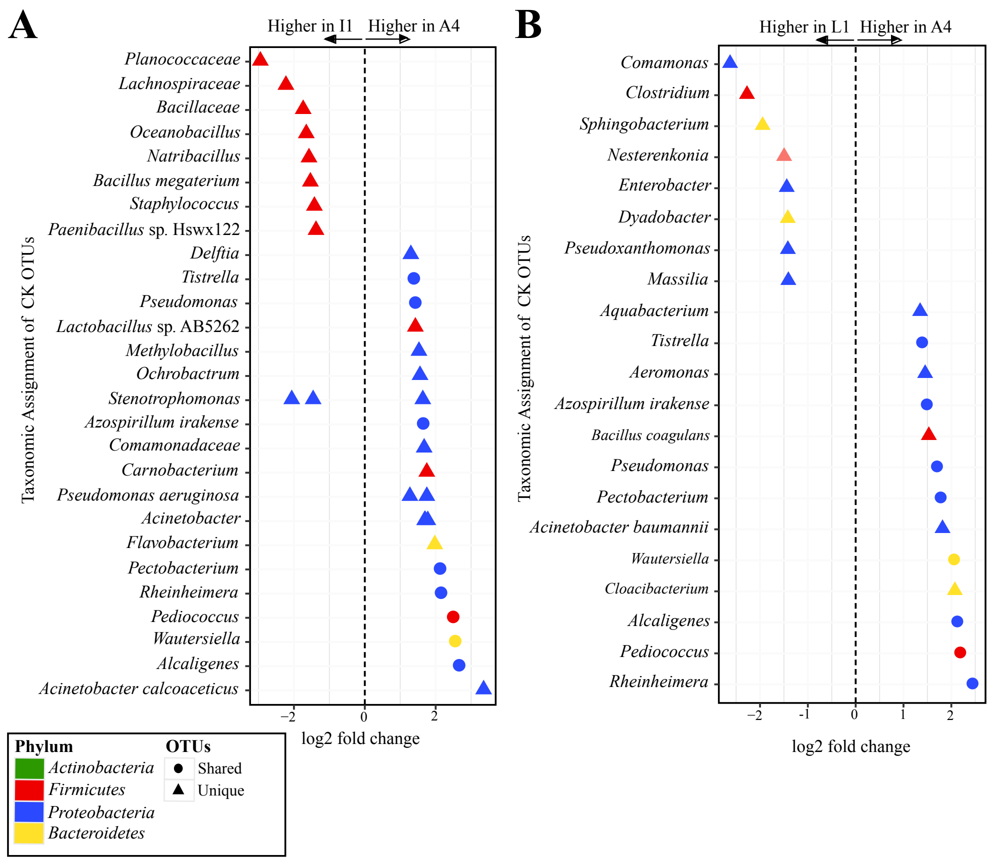

Figure B.12: Overview of relative abundances of bacterial OTUs that were statistically significantly different ($p \leq 0.001$) between lots (A) Il and A4 and lots (B) L1 and A4 for Camel Kings (CK). The dotted line highlights the conversion in log2-fold change from negative to positive values.

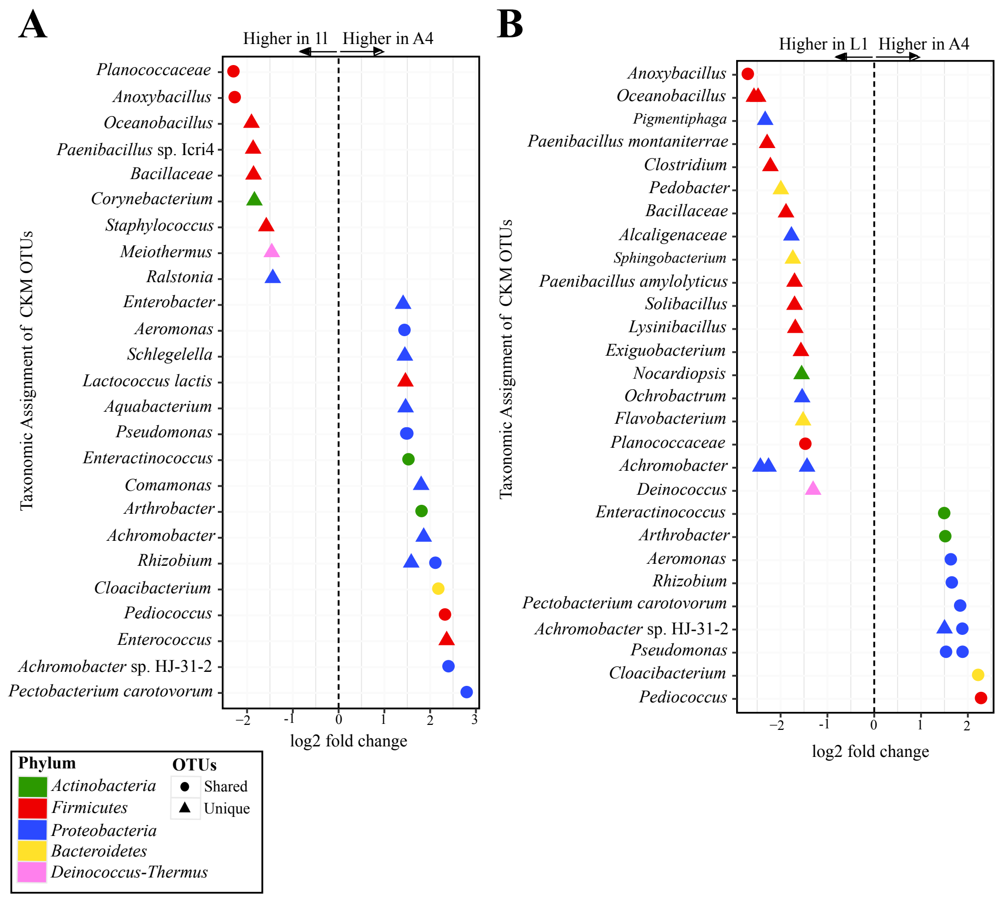

Figure B.13: Overview of relative abundances of bacterial OTUs that were statistically significantly different ($p \leq 0.001$) between lots (A) Il and A4 and lots (B) L1 and A4 for mentholated Camel Kings (CKM). The dotted line highlights the conversion in log2-fold change from negative to positive values.

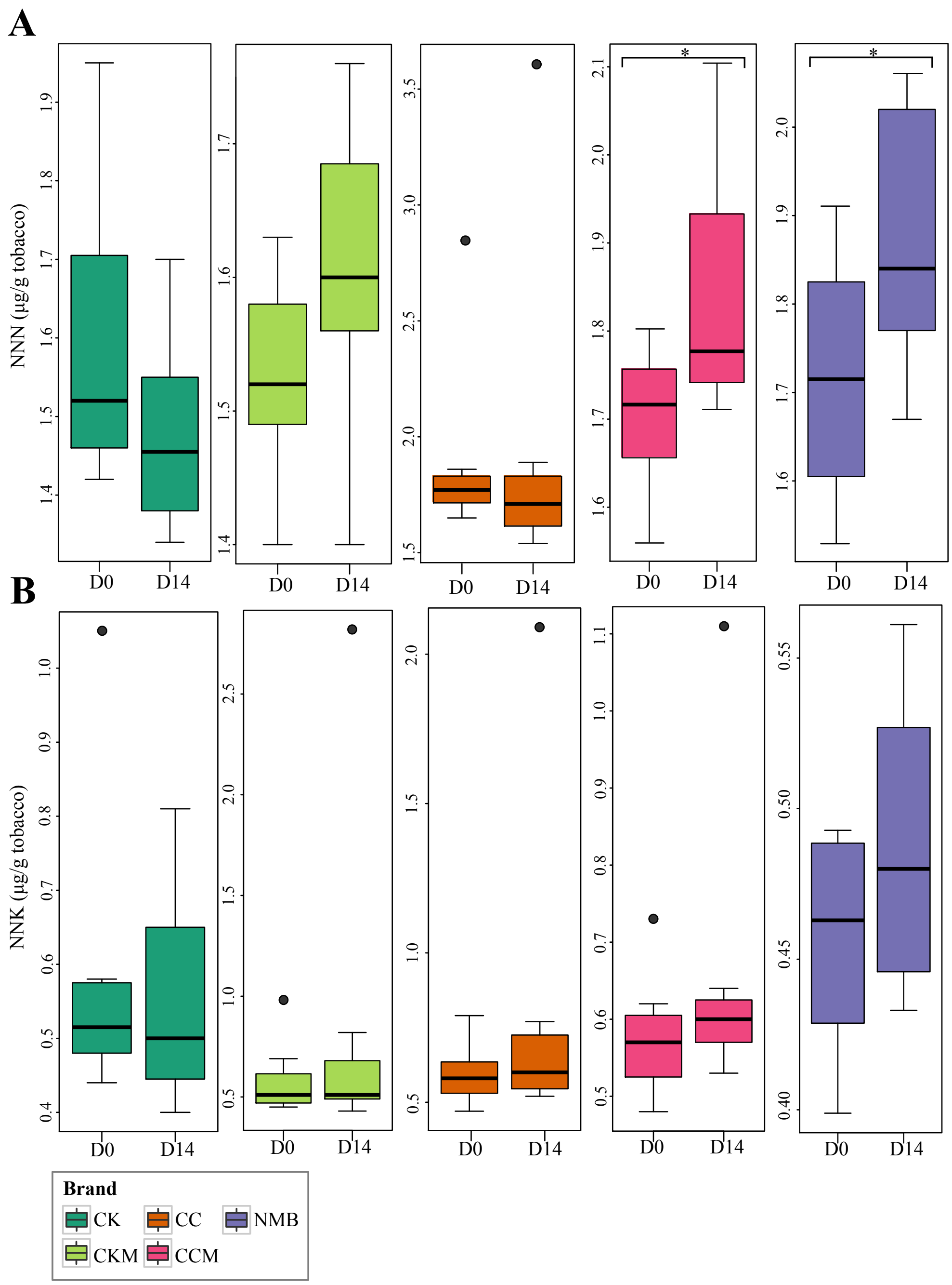

Figure B.14: Tobacco-specific nitrosamine levels over time at pocket conditions. Comparison of (A) N-nitrosonornicotine (NNN) and (B) Nicotine-derived nitrosamine ketone (NNK) levels in all brands at day 0 (D0) and day 14 (D14) at pocket conditions. Significance at $p \leq 0.05$ shown by brackets at the top of the plot.

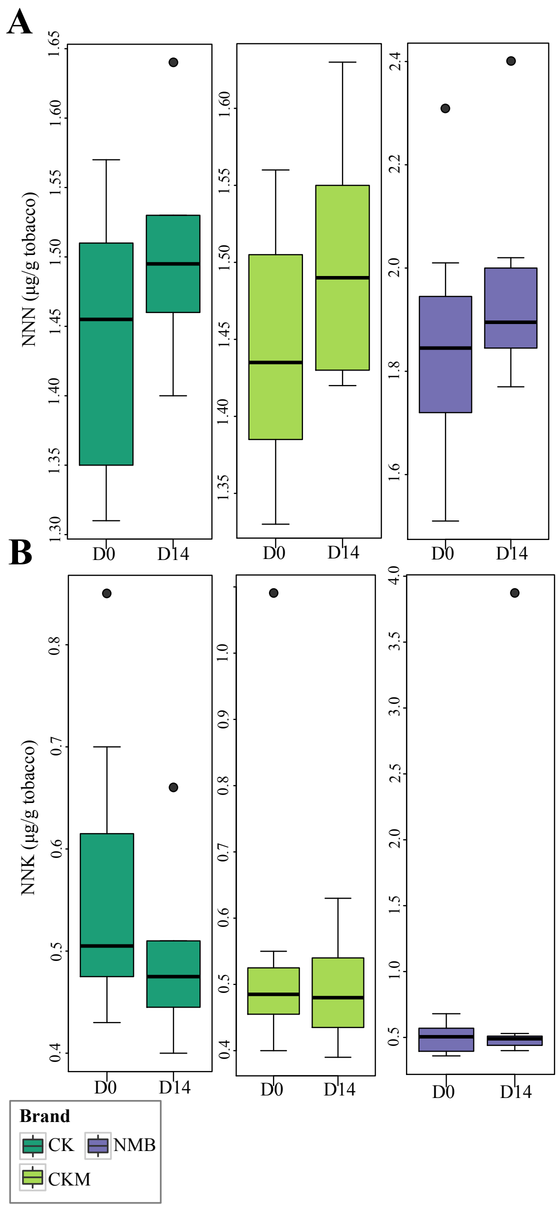

Figure B.15: TSNA levels in mg/g of tobacco over time at refrigerator conditions. Comparison of (A) N-nitrosonornicotine (NNN) and (B) Nicotine-derived nitrosamine ketone (NNK) levels in all brands at day 0 (D0) and day 14 (D14) at pocket conditions. No significant differences (at an alpha level of 0.05) were found between D0 and D14 for a given brand.

Tables

Table B.1: Descriptions of cigarette products tested at three different experimental conditions (pocket, room, and refrigerator) over time (day 0, 5, 9 and 14)

Brands	**Lot**	**Condition**	**Day 0**	**Day 5**	**Day 9**	**Day 14**
NMB	4C17	Pocket	2	2	2	2
	4C17	Room	2	2	2	2
	4C17	Fridge	2	2	2	2
	4CO3	Pocket	2	2	2	2
	4CO3	Room	2	2	2	2
	4CO3	Fridge	2	2	2	2
	4K01	Pocket	2	2	2	2
	4K01	Room	2	2	2	2
	4K01	Fridge	2	2	2	2
CC	A8	Pocket	2	2	2	2
	A8	Room	2	2	2	2
	A8	Fridge	2	2	2	2
	B2	Pocket	2	2	2	2
	B2	Room	2	2	2	2
	B2	Fridge	2	2	2	2
	B3	Pocket	2	2	2	2
	B3	Room	2	2	2	2
	B3	Fridge	2	2	2	2
CCM	A8	Pocket	2	2	2	2
	A8	Room	2	2	2	2
	A8	Fridge	2	2	2	2
	B2	Pocket	2	2	2	2
	B2	Room	2	2	2	2
	B2	Fridge	2	2	2	2
	B3	Pocket	2	2	2	2
	B3	Room	2	2	2	2
	B3	Fridge	2	2	2	2
CK	A8	Pocket	2	2	2	2
	A8	Room	2	2	2	2
	A8	Fridge	2	2	2	2
	B2	Pocket	2	2	2	2
	B2	Room	2	2	2	2
	B2	Fridge	2	2	2	2
	B3	Pocket	2	2	2	2
	B3	Room	2	2	2	2
	B3	Fridge	2	2	2	2
CKM	A8	Pocket	2	2	2	2
	A8	Room	2	2	2	2
	A8	Fridge	2	2	2	2
	B2	Pocket	2	2	2	2
	B2	Room	2	2	2	2
	B2	Fridge	2	2	2	2
	B3	Pocket	2	2	2	2
	B3	Room	2	2	2	2
	B3	Fridge	2	2	2	2

Bibliography

[1] JM Faures, A Eliasson, J Hoogeveen, and D Vallee. Aquastat-fao's information system on water and agriculture. *GRID-Magazine of the IPTRID Network (FAO/United Kingdom)*, 2001.

[2] Kenneth Strzepek and Brent Boehlert. Competition for water for the food system. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2010.

[3] Stefan Siebert, Jacob Burke, Jean-Marc Faures, Karen Frenken, Jippe Hoogeveen, Petra Dll, and Felix Theodor Portmann. Groundwater use for irrigation-a global inventory. *Hydrology and Earth System Sciences*, 14(10):1863–1880, 2010.

[4] Karen Frenken and Virginie Gillet. Irrigation water requirement and water withdrawal by country. *Food and Agriculture Organization of the United Nations*, 2012.

[5] Cheryl A Dieter, Molly A Maupin, Rodney R Caldwell, Melissa A Harris, Tamara I Ivahnenko, John K Lovelace, Nancy L Barber, and Kristin S Linsey. Estimated use of water in the United States in 2015. Report 141134233X, US Geological Survey, 2018.

[6] National Agricultural Statistics Service U.S. Department of Agriculture. Farm and ranch irrigation survey. *Accessed Online: http://www.agcensus.usda.gov/Publications/2012/Online_Resources/Farm_and_Ranch_Irrigation_Survey/*, 2013.

[7] Megan Stubbs. Irrigation in us agriculture: On-farm technologies and best management practices. *Washington DC: Congressional Research Service.*, 9:2016, 2015.

[8] Leonard F Konikow and Eloise Kendy. Groundwater depletion: A global problem. *Hydrogeology Journal*, 13(1):317–320, 2005.

[9] Leonard F Konikow. Long-term groundwater depletion in the United States. *Groundwater*, 53(1):2–9, 2015.

[10] Alexandra S Richey, Brian F Thomas, Min-Hui Lo, John T Reager, James S Famiglietti, Katalyn Voss, Sean Swenson, and Matthew Rodell. Quantifying renewable groundwater stress with grace. *Water Resources Research*, 51(7):5217–5238, 2015.

[11] Paul M Barlow. *Ground Water in fresh water-salt water environments of the Atlantic*, volume 1262. Geological Survey (USGS), 2003.

[12] Paul M Barlow and Eric G Reichard. Saltwater intrusion in coastal regions of north america. *Hydrogeology Journal*, 18(1):247–260, 2010.

[13] Christopher B Field. *Climate change 2014-Impacts, adaptation and vulnerability: Regional aspects*. Cambridge University Press, 2014.

[14] Olivier Dubois. *The state of the world's land and water resources for food and agriculture: managing systems at risk*. Earth Scan, 2011.

[15] Food and AQUASTAT agricultural organization. Irrigation area visualizations. *Accessed Online at, http://www.fao.org/nr/water/aquastat/irrigation drainage/treemap/index.stm*, 2015.

[16] Toshio Sato, Manzoor Qadir, Sadahiro Yamamoto, Tsuneyoshi Endo, and Ahmad Zahoor. Global, regional, and country level need for data on wastewater generation, treatment, and use. *Agricultural Water Management*, 130:1–13, 2013.

[17] National Risk Management Research Laboratory United States Environmental Protection Agency and U.S. Agency for International Development. 2012 guidelines for water reuse. *Accessed Online at, https://www3.epa.gov/region1/npdes/merrimackstation/pdfs/ar/AR-1530.pdf*, 2012.

[18] State Water Resources Control Board (SWRCB) and Department of Water Resources (DWR. Results, challenges, and future approaches to California's municipal wastewater recycling survey. *National Academies Press*, 2012.

[19] United States. Environmental Protection Agency. Office of Wastewater Management. Municipal Support Division, National Risk Management Research Laboratory. Technology Transfer, and Support Division. *Guidelines for water reuse*. US Environmental Protection Agency, 2004.

[20] Channah Rock, Jean E McLain, and Daniel Gerrity. Water recycling FAQs. *The University of Arizona, College of Agriculture and Life Sciences*, 2012.

[21] Beat Oertli, Dominique Auderset Joye, Emmanuel Castella, Raphalle Juge, Anthony Lehmann, and Jean-Bernard Lachavanne. Ploch: a standardized method for sampling and assessing the biodiversity in ponds. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 15(6):665–679, 2005.

[22] William H Renwick, Richard O Sleezer, Robert W Buddemeier, and Steven V Smith. Small artificial ponds in the United States: impacts on sedimentation and carbon budget. In *Proceedings of the Eighth Federal Interagency Sedimentation Conference*, pages 738–44, 2006.

[23] Jeremy Biggs, Penny Williams, Mericia Whitfield, Pascale Nicolet, and Anita Weatherby. 15 years of pond assessment in britain: results and lessons learned from the work of pond conservation. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 15(6):693–714, 2005.

[24] Luc De Meester, Steven Declerck, Robby Stoks, Gerald Louette, Frank Van De Meutter, Tom De Bie, Erik Michels, and Luc Brendonck. Ponds and pools as model systems in conservation biology, ecology and evolutionary biology. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 15(6):715–725, 2005.

[25] Rgis Crghino, J Biggs, B Oertli, and S Declerck. The ecology of european ponds: defining the characteristics of a neglected freshwater habitat. *Hydrobiologia*, 597(1):1–6, 2008.

[26] Noriko Takamura. *Status of biodiversity loss in lakes and ponds in Japan*, pages 133–148. Springer, 2012.

[27] Matthew J Hill, Christopher Hassall, Beat Oertli, Lenore Fahrig, Belinda J Robson, Jeremy Biggs, Michael J Samways, Nisikawa Usio, Noriko Takamura, and Jagdish Krishnaswamy. New policy directions for global pond conservation. *Conservation Letters*, 11(5):e12447, 2018.

[28] John A Downing, Jonathan J Cole, Jack J Middelburg, Robert G Striegl, Carlos M Duarte, Pirkko Kortelainen, Yves T Prairie, and KA Laube. Sediment organic carbon burial in agriculturally eutrophic impoundments over the last century. *Global Biogeochemical Cycles*, 22(1), 2008.

[29] Emily V Davies, Craig Winstanley, Joanne L Fothergill, and Chloe E James. The role of temperate bacteriophages in bacterial infection. *FEMS microbiology letters*, 363(5), 2016.

[30] Thomas C O'Keefe, Scott R Elliott, and Robert J Naiman. Introduction to watershed ecology. *Printed Lecture Note, University of Washington, USA*, 2012.

[31] Medagam Thirupathi Reddy, Natarajan Sivaraj, Venkateswaran Kamala, Someswara Rao Pandravada, Neelam Sunil, and Nilamani Dikshit. Classification, characterization and comparison of aquatic ecosystems in the landscape of Adilabad District, Telangana, Deccan Region, India. *Open Access Library Journal*, 5(04):1, 2018.

[32] Penny Williams, Mericia Whitfield, Jeremy Biggs, Simon Bray, Gill Fox, Pascale Nicolet, and David Sear. Comparative biodiversity of rivers, streams, ditches and ponds in an agricultural landscape in southern england. *Biological Conservation*, 115(2):329–341, 2004.

[33] John V Walther. *Earth's natural resources*. Jones and Bartlett Publishers, 2013.

[34] BR Davies, J Biggs, PJ Williams, JT Lee, and S Thompson. *A comparison of the catchment sizes of rivers, streams, ponds, ditches and lakes: implications for protecting aquatic biodiversity in an agricultural landscape*, pages 7–17. Springer, 2007.

[35] Nikolai Friberg. Impacts and indicators of change in lotic ecosystems. *Wiley Interdisciplinary Reviews: Water*, 1(6):513–531, 2014.

[36] Mieke Uyttendaele, Lee-Ann Jaykus, Philip Amoah, Alessandro Chiodini, David Cunliffe, Liesbeth Jacxsens, Kevin Holvoet, Lise Korsten, Mathew Lau, and Peter McClure. Microbial hazards in irrigation water: Standards, norms, and testing to manage use of water in fresh produce primary production. *Comprehensive Reviews in Food Science and Food Safety*, 14(4):336–356, 2015.

[37] Hanseok Jeong, Hakkwan Kim, and Taeil Jang. Irrigation water quality standards for indirect wastewater reuse in agriculture: a contribution toward sustainable wastewater reuse in south korea. *Water*, 8(4):169, 2016.

[38] US Food and Drug Administration. Standards for the growing, harvesting, packing and holding of produce for human consumption, a proposed rule. *FDA Food Safety Modernization Act (FSMA)*, 2013.

[39] Birger Rasmussen, Ian R Fletcher, Jochen J Brocks, and Matt R Kilburn. Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature*, 455(7216):1101, 2008.

[40] James F Kasting and Janet L Siefert. Life and the evolution of Earth's atmosphere. *Science*, 296(5570):1066–1068, 2002.

[41] Lynn J Rothschild and Rocco L Mancinelli. Life in extreme environments. *Nature*, 409(6823):1092–1101, 2001.

[42] David L Kirchman. *Processes in microbial ecology*. Oxford University Press, 2018.

[43] Byung C Cho and Farooq Azam. Biogeochemical significance of bacterial biomass in the ocean's euphotic zone. *Marine Ecology Progress Series.*, 63(2):253–259, 1990.

[44] James N Galloway, Alan R Townsend, Jan Willem Erisman, Mateete Bekunda, Zucong Cai, John R Freney, Luiz A Martinelli, Sybil P Seitzinger, and Mark A Sutton. Transformation of the nitrogen cycle: recent trends, questions, and potential solutions. *Science*, 320(5878):889–892, 2008.

[45] Y Wang and TA McAllister. Rumen microbes, enzymes and feed digestion-a review. *Asian-Australasian Journal of Animal Sciences*, 15(11):1659–1676, 2002.

[46] Ilseung Cho and Martin J Blaser. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*, 13(4):260, 2012.

[47] Kurt E Williamson, Mark Radosevich, and K Eric Wommack. Abundance and diversity of viruses in six delaware soils. *Applied and Environmental Microbiology*, 71(6):3119–3125, 2005.

[48] Sharath Srinivasiah, Jaysheel Bhavsar, Kanika Thapar, Mark Liles, Tom Schoenfeld, and K Eric Wommack. Phages across the biosphere: contrasts of viruses in soil and aquatic environments. *Research in Microbiology*, 159(5):349–357, 2008.

[49] Martha RJ Clokie, Andrew D Millard, Andrey V Letarov, and Shaun Heaphy. Phages in nature. *Bacteriophage*, 1(1):31–45, 2011.

[50] Steven W Wilhelm and Curtis A Suttle. Viruses and nutrient cycles in the sea viruses play critical roles in the structure and function of aquatic food webs. *Bioscience*, 49(10):781–788, 1999.

[51] Curtis A Suttle. Marine viruses–major players in the global ecosystem. *Nature Reviews: Microbiology*, 5(10):801, 2007.

[52] Osana Bonilla-Findji, Andrea Malits, Dominique Lefvre, Emma Rochelle-Newall, Rodolphe Leme, Markus G Weinbauer, and Jean-Pierre Gattuso. Viral effects on bacterial respiration, production and growth efficiency: consistent trends in the Southern Ocean and the Mediterranean Sea. *Deep Sea Research Part II: Topical Studies in Oceanography*, 55(5-7):790–800, 2008.

[53] Harald Brssow, Carlos Canchaya, and Wolf-Dietrich Hardt. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiology and Molecular Biology Reviews*, 68(3):560–602, 2004.

[54] Alejandro Reyes, Matthew Haynes, Nicole Hanson, Florent E Angly, Andrew C Heath, Forest Rohwer, and Jeffrey I Gordon. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*, 466(7304):334–338, 2010.

[55] Dana Willner, Mike Furlan, Robert Schmieder, Juris A Grasis, David T Pride, David A Relman, Florent E Angly, Tracey McDole, Ray P Mariella, and Forest Rohwer. Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4547–4553, 2011.

[56] Xiaoxue Wang, Younghoon Kim, Qun Ma, Seok Hoon Hong, Karina Pokusaeva, Joseph M Sturino, and Thomas K Wood. Cryptic prophages help bacteria cope with adverse environments. *Nature Communications*, 1:147, 2010.

[57] Centers for Disease Control and Prevention. Burden of foodborne illness: findings. *Accessed Online at, https://www.cdc.gov/foodborneburden/2011-foodborne-estimates.html*, 2011.

[58] Bradd J Haley, Dana J Cole, and Erin K Lipp. Distribution, diversity, and seasonality of waterborne salmonellae in a rural watershed. *Applied and Environmental Microbiology*, 75(5):1248–1255, 2009.

[59] Baoguang Li, George Vellidis, Huanli Liu, Michele Jay-Russell, Shaohua Zhao, Zonglin Hu, Anita Wright, and Christopher A Elkins. Diversity and antimicrobial resistance of *Salmonella enterica* isolated from surface water in southeastern us. *Applied and Environmental Microbiology*, pages AEM. 02063–14, 2014.

[60] Rachel McEgan, Jeffrey C Chandler, Lawrence D Goodridge, and Michelle D Danyluk. Diversity of *Salmonella* isolated from central Florida surface waters. *Applied and Environmental Microbiology*, pages AEM. 02191–14, 2014.

[61] Rebecca L Bell, Jie Zheng, Erik Burrows, Sarah Allard, Charles Y Wang, Christine E Keys, David C Melka, Errol Strain, Yan Luo, and Marc W Allard. Ecological prevalence, genetic diversity, and epidemiological aspects of *Salmonella* isolated from tomato agricultural regions of the virginia eastern shore. *Frontiers in Microbiology*, 6:415, 2015.

[62] Mahbub Islam, Michael P Doyle, Sharad C Phatak, Patricia Millner, and Xiuping Jiang. Persistence of enterohemorrhagic *Escherichia coli* O157:H7 in soil and on leaf lettuce and parsley grown in fields treated with contaminated manure composts or irrigation water. *Journal of Food Protection*, 67(7):1365–1370, 2004.

[63] Gabriel Mootian, Wen-Hsuan Wu, and Karl R Matthews. Transfer of *Escherichia coli* O157:H7 from soil, water, and manure contaminated with low numbers of the pathogen to lettuce plants. *Journal of Food Protection*, 72(11):2308–2312, 2009.

[64] Marilyn C Erickson, Cathy C Webb, Juan Carlos Diaz-Perez, Sharad C Phatak, John J Silvoy, Lindsey Davey, Alison S Payton, Jean Liao, Li Ma, and

Michael P Doyle. Surface and internalized *Escherichia coli* O157:H7 on field-grown spinach and lettuce treated with spray-contaminated irrigation water. *Journal of Food Protection*, 73(6):1023–1029, 2010.

[65] Guy Kisluk and Sima Yaron. Presence and persistence of *Salmonella enterica* serotype typhimurium on the phyllosphere and rhizosphere of spray irrigated parsley. *Applied and Environmental Microbiology*, pages AEM. 00087–12, 2012.

[66] SK Greene, ER Daly, EA Talbot, LJ Demma, S Holzbauer, NJ Patel, TA Hill, MO Walderhaug, RM Hoekstra, and MF Lynch. Recurrent multistate outbreak of *Salmonella* Newport associated with tomatoes from contaminated fields, 2005. *Epidemiology and Infection*, 136(2):157–165, 2008.

[67] Michele T Jay, Michael Cooley, Diana Carychao, Gerald W Wiscomb, Richard A Sweitzer, Leta Crawford-Miksza, Jeff A Farrar, David K Lau, Janice O'Connell, and Anne Millington. *Escherichia coli* O157:H7 in feral swine near spinach fields and cattle, central California coast. *Emerging Infectious Diseases*, 13(12):1908, 2007.

[68] Centers for Disease Control and Prevention. Multistate outbreak of E. coli O157:H7 infections linked to romaine lettuce. *Accessed Online at, https://www.cdc.gov/ecoli/2018/o157h7-11-18/index.html*, 2018.

[69] Alison D O'Brien, John W Newland, Steven F Miller, Randall K Holmes, H Williams Smith, and Samuel B Formal. Shiga-like toxin-converting phages from *Escherichia coli* strains that cause hemorrhagic colitis or infantile diarrhea. *Science*, 226(4675):694–696, 1984.

[70] Heather E Allison. Stx-phages: drivers and mediators of the evolution of STEC and STEC-like pathogens. *Future Microbiology*, 2007.

[71] Patrick L Wagner and Matthew K Waldor. Bacteriophage control of bacterial virulence. *Infection and Immunity*, 70(8):3985–3993, 2002.

[72] Angela HAM Van Hoek, Dik Mevius, Beatriz Guerra, Peter Mullany, Adam Paul Roberts, and Henk JM Aarts. Acquired antibiotic resistance genes: an overview. *Frontiers in Microbiology*, 2:203, 2011.

[73] Anna Colavecchio, Brigitte Cadieux, Amanda Lo, and Lawrence D Goodridge. Bacteriophages contribute to the spread of antibiotic resistance genes among foodborne pathogens of the *Enterobacteriaceae* family-a review. *Frontiers in Microbiology*, 8:1108, 2017.

[74] Horst Schmieger and Petra Schicklmaier. Transduction of multiple drug resistance of *Salmonella enterica* serovar typhimurium dt104. *FEMS Microbiology Letters*, 170(1):251–256, 1999.

[75] Bradley L Bearson, Heather K Allen, Brian W Brunelle, In Soo Lee, Sherwood R Casjens, and Thaddeus B Stanton. The agricultural antibiotic carbadox induces phage-mediated gene transfer in *Salmonella*. *Frontiers in Microbiology*, 5:52, 2014.

[76] Raymond Schuch and Vincent A Fischetti. Detailed genomic analysis of the w-beta and gamma phages infecting *Bacillus anthracis*: implications for evolution of environmental fitness and antibiotic resistance. *Journal of Bacteriology*, 188(8):3037–3051, 2006.

[77] Jakob Haaber, Jrgen J Leisner, Marianne T Cohn, Arancha Catalan-Moreno, Jesper B Nielsen, Henrik Westh, Jos R Penads, and Hanne Ingmer. Bacterial viruses enable their host to acquire antibiotic resistance genes from neighbouring cells. *Nature Communications*, 7:13333, 2016.

[78] Yifan Zhang and Jeffrey T LeJeune. Transduction of blacmy-2, *tet(A)*, and *tet(B)* from *Salmonella enterica* subspecies enterica serovar heidelberg to *S.* typhimurium. *Veterinary Microbiology*, 129(3-4):418–425, 2008.

[79] Maryury Brown-Jaque, Lirain Rodriguez-Oyarzun, Thais Cornejo Snchez, Maria Teresa Martin-Gomez, Silvia Gartner, Javier de Gracias, Sandra Rovira, Antonio Alvarez, Juan Jofre, and Juan Jose Gonzalez-Lopez. Evaluating bacteriophage particles containing antibiotic resistance genes in the sputum of cystic fibrosis patients. *Frontiers in Microbiology*, 9:856, 2018.

[80] E Marti, E Variatza, and JL Balczar. Bacteriophages as a reservoir of extended-spectrum betalactamase and fluoroquinolone resistance genes in the environment. *Clinical Microbiology and Infection*, 20(7):O456–O459, 2014.

[81] Larissa C Parsley, Erin J Consuegra, Kavita S Kakirde, Andrew M Land, Willie F Harper, and Mark R Liles. Identification of diverse antimicrobial resistance determinants carried on bacterial, plasmid, or viral metagenomes from an activated sludge microbial assemblage. *Applied and Environmental Microbiology*, 76(11):3753–3757, 2010.

[82] Stefano Colombo, Stefania Arioli, Eros Neri, Giulia Della Scala, Giorgio Gargari, and Diego Mora. Viromes as genetic reservoir for the microbial communities in aquatic environments: a focus on antimicrobial-resistance genes. *Frontiers in Microbiology*, 8:1095, 2017.

[83] Franois Enault, Arnaud Briet, La Bouteille, Simon Roux, Matthew B Sullivan, and Marie-Agns Petit. Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *The ISME Journal*, 11(1):237, 2017.

[84] M Karamanou, E Poulakou-Rebelakou, M Tzetis, and G Androutsos. Anton van Leeuwenhoek (1632-1723): father of micromorphology and discoverer of spermatozoa. *Revista Argentina de microbiologia*, 42(4), 2010.

[85] Annie Rompre, Pierre Servais, Julia Baudart, Marie-Rene De-Roubin, and Patrick Laurent. Detection and enumeration of coliforms in drinking water: current methods and emerging approaches. *Journal of Microbiological Methods*, 49(1):31–54, 2002.

[86] Michael S Rapp and Stephen J Giovannoni. The uncultured microbial majority. *Annual Reviews in Microbiology*, 57(1):369–394, 2003.

[87] Frederick Sanger, Steven Nicklen, and Alan R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.

[88] Michael L Metzker. Sequencing technologies-the next generation. *Nature Reviews Genetics*, 11(1):31, 2010.

[89] Erwin L Van Dijk, Hlne Auger, Yan Jaszczyszyn, and Claude Thermes. Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9):418–426, 2014.

[90] Carl R Woese and George E Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977.

[91] Benjamin J Callahan, Paul J McMurdie, and Susan P Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, 11(12):2639, 2017.

[92] Zarraz May-Ping Lee, Carl Bussema III, and Thomas M Schmidt. rrn db: Documenting the number of rRNA and tRNA genes in bacteria and archaea. *Nucleic Acids Research*, 37(suppl1):D489–D493, 2008.

[93] Alessandro Vezzi, S Campanaro, M D'angelo, F Simonato, N Vitulo, FM Lauro, A Cestaro, G Malacrida, B Simionati, and N Cannata. Life at depth: *Photobacterium profundum* genome sequence and expression analysis. *Science*, 307(5714):1459–1461, 2005.

[94] Emma Hambly, Francoise Ttart, Carine Desplats, William H Wilson, Henry M Krisch, and Nicholas H Mann. A conserved genetic module that encodes the major virion components in both the coliphage T4 and the marine cyanophage s-pm2. *Proceedings of the National Academy of Sciences*, 98(20):11411–11416, 2001.

[95] Alberto Lpez-Bueno, Javier Tamames, David Velzquez, Andrs Moya, Antonio Quesada, and Antonio Alcam. High diversity of the viral community from an Antarctic lake. *Science*, 326(5954):858–861, 2009.

[96] Mya Breitbart, Jon H Miyake, and Forest Rohwer. Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiology Letters*, 236(2):249–256, 2004.

[97] K Eric Wommack, Daniel J Nasko, Jessica Chopyk, and Eric G Sakowski. Counts and sequences, observations that continue to change our understanding of viruses in nature. *Journal of Microbiology*, 53(3):181–192, 2015.

[98] Mya Breitbart, Peter Salamon, Bjarne Andresen, Joseph M Mahaffy, Anca M Segall, David Mead, Farooq Azam, and Forest Rohwer. Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences*, 99(22):14250–14255, 2002.

[99] David Paez-Espino, Emiley A Eloe-Fadrosh, Georgios A Pavlopoulos, Alex D Thomas, Marcel Huntemann, Natalia Mikhailova, Edward Rubin, Natalia N Ivanova, and Nikos C Kyrpides. Uncovering Earth's virome. *Nature*, 536(7617):425, 2016.

[100] Duy Tin Truong, Eric A Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12(10):902, 2015.

[101] C Titus Brown and Luiz Irber. sourmash: a library for minhash sketching of DNA. *The Journal of Open Source Software*, 1(5), 2016.

[102] Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, 2014.

[103] Deanna M Church, Valerie A Schneider, Tina Graves, Katherine Auger, Fiona Cunningham, Nathan Bouk, Hsiu-Chuan Chen, Richa Agarwala, William M McLaren, and Graham RS Ritchie. Modernizing reference genome assemblies. *PLoS Biology*, 9(7):e1001091, 2011.

[104] Hideki Noguchi, Jungho Park, and Toshihisa Takagi. Metagene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Research*, 34(19):5623–5630, 2006.

[105] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[106] Baris E Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. Uniref: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.

[107] Andrew G McArthur, Nicholas Waglechner, Fazmin Nizam, Austin Yan, Marisa A Azad, Alison J Baylay, Kirandeep Bhullar, Marc J Canova, Gianfranco De Pascale, and Linda Ejim. The comprehensive antibiotic resistance database. *Antimicrobial Agents and Chemotherapy*, pages AAC. 00419–13, 2013.

[108] SL McLellan, SM Huse, SR Mueller-Spitz, EN Andreishcheva, and ML Sogin. Diversity and population structure of sewage-derived microorganisms in wastewater treatment plant influent. *Environmental Microbiology*, 12(2):378–392, 2010.

[109] Orin C Shanks, Ryan J Newton, Catherine A Kelty, Susan M Huse, Mitchell L Sogin, and Sandra L McLellan. Comparison of microbial community structure in untreated wastewater from different geographic locales. *Applied and Environmental Microbiology*, pages AEM. 03448–12, 2013.

[110] Lin Ye and Tong Zhang. Bacterial communities in different sections of a municipal wastewater treatment plant revealed by 16s rDNA 454 pyrosequencing. *Applied Microbiology and Biotechnology*, 97(6):2681–2690, 2013.

[111] Bo Zhang, Xiangyang Xu, and Liang Zhu. Structure and function of the microbial consortia of activated sludge in typical municipal wastewater treatment plants in winter. *Scientific Reports*, 7(1):17930, 2017.

[112] Prachi Kulkarni, Nathan D Olson, Joseph N Paulson, Mihai Pop, Cynthia Maddox, Emma Claye, Rachel E Rosenberg Goldstein, Manan Sharma, Shawn G Gibbs, and Emmanuel F Mongodin. Conventional wastewater treatment and reuse site practices modify bacterial community structure but do not eliminate some opportunistic pathogens in reclaimed water. *Science of The Total Environment*, 639:1126–1137, 2018.

[113] Hideyuki Tamaki, Rui Zhang, Florent E Angly, Shota Nakamura, Pei-Ying Hong, Teruo Yasunaga, Yoichi Kamagata, and Wen-Tso Liu. Metagenomic analysis of DNA viruses in a wastewater treatment plant in tropical climate. *Environmental Microbiology*, 14(2):441–452, 2012.

[114] Karyna Rosario, Christina Nilsson, Yan Wei Lim, Yijun Ruan, and Mya Breitbart. Metagenomic analysis of viruses in reclaimed water. *Environmental Microbiology*, 11(11):2806–2820, 2009.

[115] Mya Breitbart, Ben Felts, Scott Kelley, Joseph M Mahaffy, James Nulton, Peter Salamon, and Forest Rohwer. Diversity and population structure of a near-shore marine-sediment viral community. *Proceedings of the Royal Society of London B: Biological Sciences*, 271(1539):565–574, 2004.

[116] Robert A Edwards and Forest Rohwer. Opinion: viral metagenomics. *Nature Reviews. Microbiology*, 3(6):504, 2005.

[117] Mya Breitbart, Ian Hewson, Ben Felts, Joseph M Mahaffy, James Nulton, Peter Salamon, and Forest Rohwer. Metagenomic analyses of an uncultured viral community from human feces. *Journal of Bacteriology*, 185(20):6220–6223, 2003.

[118] Gabriel Zwart, Byron C Crump, Miranda P Kamst-van Agterveld, Ferry Hagen, and Suk-Kyun Han. Typical freshwater bacteria: an analysis of available 16s rRNA gene sequences from plankton of lakes and rivers. *Aquatic Microbial Ecology*, 28(2):141–155, 2002.

[119] Ryan J Newton, Stuart E Jones, Alexander Eiler, Katherine D McMahon, and Stefan Bertilsson. A guide to the natural history of freshwater lake bacteria. *Microbiology and Molecular Biology Reviews*, 75(1):14–49, 2011.

[120] Josefa Antn, Ramn Rossell-Mora, Francisco Rodrguez-Valera, and Rudolf Amann. Extremely halophilic bacteria in crystallizer ponds from solar salterns. *Applied and Environmental Microbiology*, 66(7):3052–3057, 2000.

[121] Beat Oertli, Dominique Auderset Joye, Emmanuel Castella, Raphalle Juge, Diana Cambin, and Jean-Bernard Lachavanne. Does size matter? The relationship between pond area and biodiversity. *Biological Conservation*, 104(1):59–70, 2002.

[122] Jeffrey A Kimbrel, Nicholas Ballor, Yu-Wei Wu, Maude M David, Terry C Hazen, Blake A Simmons, Steven W Singer, and Janet K Jansson. Microbial community structure and functional potential along a hypersaline gradient. *Frontiers in Microbiology*, 9, 2018.

[123] Sophie Crevecoeur, Warwick F Vincent, Jrme Comte, and Connie Lovejoy. Bacterial community structure across environmental gradients in permafrost thaw ponds: methanotroph-rich ecosystems. *Frontiers in Microbiology*, 6:192, 2015.

[124] CR Arias, JW Abernathy, and Z Liu. Combined use of 16s ribosomal DNA and automated ribosomal intergenic spacer analysis to study the bacterial community in catfish ponds. *Letters in Applied Microbiology*, 43(3):287–292, 2006.

[125] Ya Qin, Jie Hou, Ming Deng, Quansheng Liu, Chongwei Wu, Yingjie Ji, and Xugang He. Bacterial abundance and diversity in pond water supplied with different feeds. *Scientific Reports*, 6:35232, 2016.

[126] Frank Oliver Glckner, Evgeny Zaichikov, Natalia Belkova, Ludmilla Denissova, Jakob Pernthaler, Annelie Pernthaler, and Rudolf Amann. Comparative 16s rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of *Actinobacteria*. *Applied and Environmental Microbiology*, 66(11):5053–5065, 2000.

[127] AC Yannarell, AD Kent, GH Lauster, TK Kratz, and EW Triplett. Temporal patterns in bacterial communities in three temperate lakes of different trophic status. *Microbial Ecology*, 46(4):391–405, 2003.

[128] Eva S Lindstrm, Miranda P Kamst-Van Agterveld, and Gabriel Zwart. Distribution of typical freshwater bacterial groups is associated with ph, temperature, and lake water retention time. *Applied and Environmental Microbiology*, 71(12):8201–8206, 2005.

[129] Kilian P Hennes, Curtis A Suttle, and Amy M Chan. Fluorescently labeled virus probes show that natural virus populations can control the structure of marine microbial communities. *Applied and Environmental Microbiology*, 61(10):3623–3627, 1995.

[130] Cheryl-Emiliane T Chow, Diane Y Kim, Rohan Sachdeva, David A Caron, and Jed A Fuhrman. Top-down controls on bacterial community structure: microbial network analysis of bacteria, T4-like viruses and protists. *The ISME Journal*, 8(4):816–829, 2014.

[131] Curtis A Suttle. Marine viruses' major players in the global ecosystem. *Nature Reviews Microbiology*, 5(10):801–812, 2007.

[132] Forest Rohwer and Rebecca Vega Thurber. Viruses manipulate the marine environment. *Nature*, 459(7244):207, 2009.

[133] Daniel Aguirre de Crcer, Alberto Lpez-Bueno, David A Pearce, and Antonio Alcam. Biodiversity and distribution of polar freshwater DNA viruses. *Science Advances*, 1(5):e1400127, 2015.

[134] Simon Roux, Francois Enault, Agns Robin, Viviane Ravet, Sbastien Personnic, Sbastien Theil, Jonathan Colombet, Tlesphore Sime-Ngando, and Didier Debroas. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PloS One*, 7(3):e33641, 2012.

[135] Appolinaire Djikeng, Ryan Kuzmickas, Norman G Anderson, and David J Spiro. Metagenomic analysis of RNA viruses in a fresh water lake. *PloS One*, 4(9):e7264, 2009.

[136] Jasmin C Green, Faraz Rahman, Matthew A Saxton, and Kurt E Williamson. Metagenomic assessment of viral diversity in Lake Matoaka, a temperate, eutrophic freshwater lake in southeastern virginia, usa. *Aquatic Microbial Ecology*, 75(2):117–128, 2015.

[137] Mohammad Mohiuddin and Herb Schellhorn. Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Frontiers in Microbiology*, 6:960, 2015.

[138] Siobhan C Watkins, Neil Kuehnle, C Anthony Ruggeri, Kema Malki, Katherine Bruder, Jinan Elayyan, Kristina Damisch, Naushin Vahora, Paul O'Malley, and Brieanne Ruggles-Sage. Assessment of a metaviromic dataset generated from nearshore Lake Michigan. *Marine and Freshwater Research*, 67(11):1700–1708, 2016.

[139] Timofey Skvortsov, Colin de Leeuwe, John P Quinn, John W McGrath, Christopher CR Allen, Yvonne McElarney, Catherine Watson, Ksenia Arkhipova, Rob Lavigne, and Leonid A Kulakov. Metagenomic characterisation of the viral community of Lough Neagh, the largest freshwater lake in ireland. *PloS One*, 11(2):e0150361, 2016.

[140] Ching-Hung Tseng, Pei-Wen Chiang, Fuh-Kwo Shiah, Yi-Lung Chen, Jia-Rong Liou, Ting-Chang Hsu, Suhinthan Maheswararajah, Isaam Saeed, Saman Halgamuge, and Sen-Lin Tang. Microbial and viral metagenomes of a subtropical freshwater reservoir subject to climatic disturbances. *The ISME Journal*, 7(12):2374–2386, 2013.

[141] Madina Saparbaevna Alexyuk, Aizhan Sabirzhanovna Turmagambetova, Pavel Gennadievich Alexyuk, Andrey Pavlinovich Bogoyavlenskiy, and Vladimir Eleazarovich Berezin. Comparative study of viromes from freshwater samples of the Ile-Balkhash region of Kazakhstan captured through metagenomic analysis. *Virus Disease*, 28(1):18–25, 2017.

[142] Xiaoqiong Gu, Qi Xiang Martin Tay, Shu Harn Te, Nazanin Saeidi, Shin Giek Goh, Ariel Kushmaro, Janelle R Thompson, and Karina Yew-Hoong Gin. Geospatial distribution of viromes in tropical freshwater ecosystems. *Water Research*, 137:220–232, 2018.

[143] K Eric Wommack and Rita R Colwell. Virioplankton: viruses in aquatic ecosystems. *Microbiology and Molecular Biology Reviews*, 64(1):69–114, 2000.

[144] Stphan Jacquet, Takeshi Miki, Rachel Noble, Peter Peduzzi, and Steven Wilhelm. Viruses in aquatic ecosystems: important advancements of the last 20 years and prospects for the future in the field of microbial oceanography and limnology. *Advances in Oceanography and Limnology*, 1(1):97–141, 2010.

[145] Laura Fancello, Sbatien Trape, Catherine Robert, Mickal Boyer, Nikolay Popgeorgiev, Didier Raoult, and Christelle Desnues. Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. *The ISME Journal*, 7(2):359–369, 2013.

[146] Lucie Zinger, Anglique Gobet, and Thomas Pommier. Two decades of describing the unseen majority of aquatic microbial diversity. *Molecular Ecology*, 21(8):1878–1896, 2012.

[147] Bonnie L Brown, Rebecca V LePrell, Rima B Franklin, Maria C Rivera, Francine M Cabral, Hugh L Eaves, Vicki Gardiakos, Kevin P Keegan, and Timothy L King. Metagenomic analysis of planktonic microbial consortia from a non-tidal urban-impacted segment of James River. *Standards in Genomic Sciences*, 10(1):65, 2015.

[148] Bradley Drury, Emma Rosi-Marshall, and John J Kelly. Wastewater treatment effluent reduces the abundance and diversity of benthic bacterial communities

in urban and suburban rivers. *Applied and Environmental Microbiology*, pages AEM. 03527–12, 2013.

[149] Abasiofiok Mark Ibekwe, Menu B Leddy, Richard M Bold, and Alexandria K Graves. Bacterial community composition in low-flowing river water with different sources of pollutants. *FEMS Microbiology Ecology*, 79(1):155–166, 2012.

[150] Christopher Staley, Trevor J Gould, Ping Wang, Jane Phillips, James B Cotner, and Michael J Sadowsky. Bacterial community structure is indicative of chemical inputs in the upper Mississippi River. *Frontiers in Microbiology*, 5:524, 2014.

[151] Anne E Bernhard, Debbie Colbert, James McManus, and Katharine G Field. Microbial community dynamics based on 16s rRNA gene profiles in a Pacific Northwest estuary and its tributaries. *FEMS Microbiology Ecology*, 52(1):115–128, 2005.

[152] Thea Van Rossum, Michael A Peabody, Miguel I Uyaguari-Diaz, Kirby I Cronin, Michael Chan, Jared R Slobodan, Matthew J Nesbitt, Curtis A Suttle, William WL Hsiao, and Patrick KC Tang. Year-long metagenomic study of river microbiomes across land use and water quality. *Frontiers in Microbiology*, 6:1405, 2015.

[153] Lanlan Cai, Rui Zhang, Ying He, Xiaoyuan Feng, and Nianzhi Jiao. Metagenomic analysis of virioplankton of the subtropical Jiulong River estuary, China. *Viruses*, 8(2):35, 2016.

[154] Lisa M Dann, Stephanie Rosales, Jody McKerral, James S Paterson, Renee J Smith, Thomas C Jeffries, Rod L Oliver, and James G Mitchell. Marine and giant viruses as indicators of a marine microbial community in a riverine system. *Microbiology Open*, 5(6):1071–1084, 2016.

[155] X Fernandez-Cassi, N Timoneda, E Gonzales-Gustavson, JF Abril, S Bofill-Mas, and R Girones. A metagenomic assessment of viral contamination on fresh parsley plants irrigated with fecally tainted river water. *International Journal of Food Microbiology*, 257:80–90, 2017.

[156] S de O Bruno, Felipe H Coutinho, Gustavo B Gregoracci, Luciana Leomil, Louisi S de Oliveira, Adriana Fres, Diogo Tschoeke, Ana Carolina Soares, Anderson S Cabral, and Nicholas D Ward. Virioplankton assemblage structure in the lower river and ocean continuum of the Amazon. *mSphere*, 2(5):e00366–17, 2017.

[157] Jessica Chopyk, Sarah Allard, Daniel J Nasko, Anthony Bui, Emmanuel F Mongodin, and Amy Rebecca Sapkota. Agricultural freshwater pond supports diverse and dynamic bacterial and viral populations. *Frontiers in Microbiology*, 9:792, 2018.

[158] Jessica Chopyk, Suhana Chattopadhyay, Prachi Kulkarni, Emma Claye, Kelsey R Babik, Molly C Reid, Eoghan M Smyth, Lauren E Hittle, Joseph N Paulson, Raul Cruz-Cano, Mihai Pop, Stephanie S Buehler, Pamela I Clark, Amy R Sapkota, and Emmanuel F Mongodin. Mentholation affects the cigarette microbiota by selecting for bacteria resistant to harsh environmental conditions and selecting against potential bacterial pathogens. *Microbiome*, 5(1):22, 2017.

[159] Jessica Chopyk, Suhana Chattopadhyay, Prachi Kulkarni, Eoghan M Smyth, Lauren E Hittle, Joseph N Paulson, Mihai Pop, Stephanie S Buehler, Pamela I Clark, Emmanuel F Mongodin, and Amy R Sapkota. Temporal variations in cigarette tobacco bacterial community composition and tobacco-specific nitrosamine content are influenced by brand and storage conditions. *Frontiers in Microbiology*, 8, 2017.

[160] Jelle Bruinsma. The resource outlook to 2050: By how much do land, water and crop yields need to increase by 2050. In *Expert meeting on how to feed the world*, volume 2050, pages 24–26, 2009.

[161] Yiwen Lin, Dan Li, Siyu Zeng, and Miao He. Changes of microbial composition during wastewater reclamation and distribution systems revealed by high-throughput sequencing analyses. *Frontiers of Environmental Science and Engineering*, 10(3):539–547, 2016.

[162] Aneta Luczkiewicz, Ewa Kotlarska, Wojciech Artichowicz, Katarzyna Tarasewicz, and Sylwia Fudala-Ksiazek. Antimicrobial resistance of *Pseudomonas* spp. isolated from wastewater and wastewater-impacted marine coastal zone. *Environmental Science and Pollution Research*, 22(24):19823–19834, 2015.

[163] Rafael Szczepanowski, Burkhard Linke, Irene Krahn, Karl-Heinz Gartemann, Tim Guetzkow, Wolfgang Eichler, Alfred Phler, and Andreas Schlueter. Detection of 140 clinically relevant antibiotic-resistance genes in the plasmid metagenome of wastewater treatment plant bacteria showing reduced susceptibility to selected antibiotics. *Microbiology*, 155(7):2306–2319, 2009.

[164] Tong Zhang and Bing Li. Occurrence, transformation, and fate of antibiotics in municipal wastewater treatment plants. *Critical Reviews in Environmental Science and Technology*, 41(11):951–998, 2011.

[165] Mariya Munir, Kelvin Wong, and Irene Xagoraraki. Release of antibiotic resistant bacteria and genes in the effluent and biosolids of five wastewater utilities in Michigan. *Water Research*, 45(2):681–693, 2011.

[166] TV Suslow, MP Oria, LR Beuchat, EH Garrett, ME Parish, LJ Harris, JN Farber, and FF Busta. Production practices as risk factors in microbial food safety of fresh and fresh-cut produce. *Comprehensive Reviews in Food Science and Food Safety*, 2:38–77, 2003.

[167] Julia D Miller, Jon E Schoonover, Karl WJ Williard, and Charnsmorn R Hwang. Whole catchment land cover effects on water quality in the lower Kaskaskia River watershed. *Water, Air, and Soil Pollution*, 221(1-4):337, 2011.

[168] Nikki Shariat and Edward G Dudley. CRISPRs: molecular signatures used for pathogen subtyping. *Applied Environmental Microbiology*, 80(2):430–439, 2014.

[169] Mina Rho, Yu-Wei Wu, Haixu Tang, Thomas G Doak, and Yuzhen Ye. Diverse CRISPRs evolving in human microbiomes. *PLoS Genetics*, 8(6):e1002441, 2012.

[170] Philippe Horvath, Dennis A Romero, Anne-Claire Cot-Monvoisin, Melissa Richards, Hlene Deveau, Sylvain Moineau, Patrick Boyaval, Christophe Fremaux, and Rodolphe Barrangou. Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus. Journal of Bacteriology*, 190(4):1401–1412, 2008.

[171] John F Heidelberg, William C Nelson, Thomas Schoenfeld, and Devaki Bhaya. Germ warfare in a microbial mat community: CRISPR provide insights into the co-evolution of host and viral genomes. *PloS One*, 4(1):e4169, 2009.

[172] Jamie C Snyder, Mary M Bateson, Matthew Lavin, and Mark J Young. Use of cellular CRISPR (clusters of regularly interspaced short palindromic repeats) spacer-based microarrays for detection of viruses in environmental samples. *Applied and Environmental Microbiology*, 76(21):7251–7258, 2010.

[173] Anna Lopatina, Sofia Medvedeva, Sergey Shmakov, Maria D Logacheva, Vjacheslav Krylenkov, and Konstantin Severinov. Metagenomic analysis of bacterial communities of Antarctic surface snow. *Frontiers in Microbiology*, 7:398, 2016.

[174] Laura S Frost, Raphael Leplae, Anne O Summers, and Ariane Toussaint. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, 3(9):722, 2005.

[175] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, pages 2114–2120, 2014.

[176] Tanja Magoc and Steven L Salzberg. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963, 2011.

[177] Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, and Pavel Pevzner. metaspades: a new versatile *de novo* metagenomics assembler. *Genome Research*, 27.5:824–834, 2017.

[178] UniProt Consortium. The universal protein resource (UniProt) in 2010. *Nucleic Acids Research*, 38(suppl1):D142–D148, 2009.

[179] Rachael P Huntley, Tony Sawford, Prudence Mutowo-Meullenet, Aleksandra Shypitsyna, Carlos Bonilla, Maria J Martin, and Claire O'donovan. The GOA database: gene ontology annotation updates for 2015. *Nucleic Acids Research*, 43(D1):D1057–D1063, 2014.

[180] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[181] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Micha Wojciech Szczeniak, Daniel J Gaffney, Laura L Elo, and Xuegong Zhang. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1):1, 2016.

[182] Raivo Kolde and Maintainer Raivo Kolde. Package "pheatmap". *R Package*, 1.7, 2018.

[183] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer Science and Business Media, 2009.

[184] Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

[185] Ea Zankari, Henrik Hasman, Salvatore Cosentino, Martin Vestergaard, Simon Rasmussen, Ole Lund, Frank M Aarestrup, and Mette Voldby Larsen. Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*, 67(11):2640–2644, 2012.

[186] Kazutaka Katoh, Kazuharu Misawa, Keiichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, 2002.

[187] Charles Bland, Teresa L Ramsey, Fareedah Sabree, Micheal Lowe, Kyndall Brown, Nikos C Kyrpides, and Philip Hugenholtz. CRISPR recognition tool (crt): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, 8(1):209, 2007.

[188] Tonia Korves and Marc E Colosimo. Controlled vocabularies for microbial virulence factors. *Trends in Microbiology*, 17(7):279–285, 2009.

[189] Ann SG Lee, Irene HK Lim, Lynn LH Tang, Amalio Telenti, and Sin Yew Wong. Contribution of kasa analysis to detection of isoniazid-resistant *Mycobacterium tuberculosisin* Singapore. *Antimicrobial Agents and Chemotherapy*, 43(8):2087–2089, 1999.

[190] S Sulochana, S Narayanan, CN Paramasivan, C Suganthi, and PR Narayanan. Analysis of fluoroquinolone resistance in clinical isolates of *Mycobacterium tuberculosis* from India. *Journal of Chemotherapy*, 19(2):166–171, 2007.

[191] Koen AL De Smet, Karen E Kempsell, Alex Gallagher, Ken Duncan, and Douglas B Young. Alteration of a single amino acid residue reverses fosfomycin resistance of recombinant *MurA* from *Mycobacterium tuberculosis*. *Microbiology*, 145(11):3177–3184, 1999.

[192] Marie Ballif, Paul Harino, Serej Ley, Mireia Coscolla, Stefan Niemann, Robyn Carter, Christopher Coulter, Sonia Borrell, Peter Siba, and Suparat Phuanukoonnon. Drug resistance-conferring mutations in *Mycobacterium* tuberculosis from Madang, Papua New Guinea. *BMC Microbiology*, 12(1):191, 2012.

[193] A Brzostek, A Sajduda, T Sliwinski, E Augustynowicz-Kopec, A Jaworski, Z Zwolska, and J Dziadek. Molecular characterisation of streptomycin-resistant *Mycobacterium tuberculosis* strains isolated in Poland. *The International Journal of Tuberculosis and Lung Disease*, 8(8):1032–1035, 2004.

[194] Alexey Kubanov, Denis Vorobyev, Aleksandr Chestkov, Arvo Leinsoo, Boris Shaskolskiy, Ekaterina Dementieva, Viktoria Solomka, Xenia Plakhova, Dmitry Gryadunov, and Dmitriy Deryabin. Molecular epidemiology of drug-resistant neisseria gonorrhoeae in russia (current status, 2015). *BMC Infectious Diseases*, 16(1):389, 2016.

[195] Anne-Marie Zuurmond, Lian N Olsthoorn-Tieleman, J Martien de Graaf, Andrea Parmeggiani, and Barend Kraal. Mutant EF-TU species reveal novel features of the enacyloxin iia inhibition mechanism on the ribosome1. *Journal of Molecular Biology*, 294(3):627–637, 1999.

[196] Anna A Gogleva, Mikhail S Gelfand, and Irena I Artamonova. Comparative analysis of CRISPR cassettes from the human gut metagenomic contigs. *BMC Genomics*, 15(1):202, 2014.

[197] ke Hagstrm, Thomas Pommier, Forest Rohwer, Karin Simu, Willem Stolte, Dominika Svensson, and Ulla Li Zweifel. Use of 16s ribosomal DNA for delineation of marine bacterioplankton species. *Applied and Environmental Microbiology*, 68(7):3628–3633, 2002.

[198] Marcin Golebiewski, Joanna Calkiewicz, Simon Creer, and Kasia Piwosz. Tideless estuaries in brackish seas as possible freshwater-marine transition zones for bacteria: the case study of the Vistula river estuary. *Environmental Microbiology Reports*, 9(2):129–143, 2017.

[199] Kasia Piwosz, Michaela M Salcher, Michael Zeder, Anetta Ameryk, and Jakob Pernthaler. Seasonal dynamics and activity of typical freshwater bacteria in brackish waters of the Gulf of Gdansk. *Limnology and Oceanography*, 58(3):817–826, 2013.

[200] Yojiro Anzai, Hongik Kim, Ju-Young Park, Hisatsugu Wakabayashi, and Hiroshi Oyaizu. Phylogenetic affiliation of the pseudomonads based on 16s rRNA

sequence. *International Journal of Systematic and Evolutionary Microbiology*, 50(4):1563–1589, 2000.

[201] J Lee, CS Lee, KM Hugunin, CJ Maute, and RC Dysko. Bacteria from drinking water supply and their fate in gastrointestinal tracts of germ-free mice: a phylogenetic comparison study. *Water Research*, 44(17):5050–5058, 2010.

[202] Vincent Delafont, Amlie Brouke, Didier Bouchon, Laurent Moulin, and Yann Hchard. Microbiome of free-living amoebae isolated from drinking water. *Water Research*, 47(19):6958–6965, 2013.

[203] Tatsuya Nakayama, Tran Thi Tuyet Hoa, Kazuo Harada, Minae Warisaya, Megumi Asayama, Atsushi Hinenoya, Joon Won Lee, Tran Minh Phu, Shuhei Ueda, and Yoshinori Sumimura. Water metagenomic analysis reveals low bacterial diversity and the presence of antimicrobial residues and resistance genes in a river containing wastewater from backyard aquacultures in the Mekong Delta, Vietnam. *Environmental Pollution*, 222:294–306, 2017.

[204] JA Humphries, A MH Ashe, JA Smiley, and CG Johnston. Microbial community structure and trichloroethylene degradation in groundwater. *Canadian Journal of Microbiology*, 51(6):433–439, 2005.

[205] Kevin M Posman, Christopher M DeRito, and Eugene L Madsen. Benzene degradation by a *Variovorax* species within a coal tar-contaminated groundwater microbial community. *Applied and Environmental Microbiology*, 83(4):e02658–16, 2017.

[206] David A Lipson, Russell K Monson, Steven K Schmidt, and Michael N Weintraub. The trade-off between growth rate and yield in microbial communities and the consequences for under-snow soil respiration in a high elevation coniferous forest. *Biogeochemistry*, 95(1):23–35, 2009.

[207] Karolien Bers, Kristel Sniegowski, Pieter Albers, Philip Breugelmans, Larissa Hendrickx, Ren De Mot, and Dirk Springael. A molecular toolbox to estimate the number and diversity of *Variovorax* in the environment: application in soils treated with the phenylurea herbicide linuron. *FEMS Microbiology Ecology*, 76(1):14–25, 2011.

[208] Catarina Ferreira, Ana R Soares, Pedro Lamosa, Manuel A Santos, and Milton S Da Costa. Comparison of the compatible solute pool of two slightly halophilic *Planctomycetes* species, *Gimesia maris* and *Rubinisphaera brasiliensis*. *Extremophiles*, 20(6):811–820, 2016.

[209] Linda Tonk, Kim Bosch, Petra M Visser, and Jef Huisman. Salt tolerance of the harmful cyanobacterium *Microcystis aeruginosa*. *Aquatic Microbial Ecology*, 46(2):117–123, 2007.

[210] JR Crush, LR Briggs, JM Sprosen, and SN Nichols. Effect of irrigation with lake water containing microcystins on microcystin content and growth of ryegrass, clover, rape, and lettuce. *Environmental Toxicology*, 23(2):246–252, 2008.

[211] Nicole L Fahrenfeld, Yanjun Ma, Maureen O'Brien, and Amy Pruden. Reclaimed water as a reservoir of antibiotic resistance genes: distribution system and irrigation implications. *Frontiers in Microbiology*, 4:130, 2013.

[212] Ivone Vaz-Moreira, Olga C Nunes, and Clia M Manaia. Bacterial diversity and antibiotic resistance in water habitats: searching the links with the human microbiome. *FEMS Microbiology Reviews*, 38(4):761–778, 2014.

[213] Vanessa M D'Costa, Christine E King, Lindsay Kalan, Mariya Morar, Wilson WL Sung, Carsten Schwarz, Duane Froese, Grant Zazula, Fabrice Calmels, and Regis Debruyne. Antibiotic resistance is ancient. *Nature*, 477(7365):457, 2011.

[214] Marc W Van Goethem, Rian Pierneef, Oliver KI Bezuidt, Yves Van De Peer, Don A Cowan, and Thulani P Makhalanyane. A reservoir of "historical" antibiotic resistance genes in remote pristine Antarctic soils. *Microbiome*, 6(1):40, 2018.

[215] Heather K Allen, Justin Donato, Helena Huimi Wang, Karen A Cloud-Hansen, Julian Davies, and Jo Handelsman. Call of the wild: antibiotic resistance genes in natural environments. *Nature Reviews Microbiology*, 8(4):251, 2010.

[216] Wilhelm Paulander, Sophie Maisnier-Patin, and Dan I Andersson. The fitness cost of streptomycin resistance depends on *rpsL* mutation, carbon source and rpos (sigmas). *Genetics*, 183(2):539–546, 2009.

[217] D Barrie Johnson and Kevin B Hallberg. The microbiology of acidic mine waters. *Research in Microbiology*, 154(7):466–473, 2003.

[218] Matthew B Stott, Michelle A Crowe, Bruce W Mountain, Angela V Smirnova, Shaobin Hou, Maqsudul Alam, and Peter F Dunfield. Isolation of novel bacteria, including a candidate division, from geothermal soils in New Zealand. *Environmental Microbiology*, 10(8):2030–2041, 2008.

[219] Tendai Walter Sanyika, William Stafford, and Don A Cowan. The soil and plant determinants of community structures of the dominant *Actinobacteria* in marion island terrestrial habitats, sub-Antarctica. *Polar Biology*, 35(8):1129–1141, 2012.

[220] Jianhua Guo, Jie Li, Hui Chen, Philip L Bond, and Zhiguo Yuan. Metagenomic analysis reveals wastewater treatment plants as hotspots of antibiotic resistance genes and mobile genetic elements. *Water Research*, 123:468–478, 2017.

[221] Centers for Disease Control and Prevention. Measuring 0utpatient antibiotic prescriptions- United States. *Accessed Online at, https://www.cdc.gov/antibiotic-use/community/programs-measurement/measuring-antibiotic-prescribing.html*, 2014.

[222] Baofeng Jia, Amogelang R Raphenya, Brian Alcock, Nicholas Waglechner, Peiyao Guo, Kara K Tsang, Briony A Lago, Biren M Dave, Sheldon Pereira, and Arjun N Sharma. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, page gkw1004, 2016.

[223] Antti Karkman, Thi Thuy Do, Fiona Walsh, and Marko PJ Virta. Antibiotic-resistance genes in waste water. *Trends in Microbiology*, 2017.

[224] Christa S McArdell, Eva Molnar, Marc J-F Suter, and Walter Giger. Occurrence and fate of macrolide antibiotics in wastewater treatment plants and in the Glatt Valley Watershed, Switzerland. *Environmental Science and Technology*, 37(24):5479–5486, 2003.

[225] Edina Szekeres, Andreea Baricz, Cecilia Maria Chiriac, Anca Farkas, Ocsana Opris, Maria-Loredana Soran, Adrian-Stefan Andrei, Knut Rudi, Jose Luis Balczar, and Nicolae Dragos. Abundance of antibiotics, antibiotic resistance genes and bacterial community composition in wastewater effluents from different romanian hospitals. *Environmental Pollution*, 225:304–315, 2017.

[226] T Iwane, T Urase, and K Yamamoto. Possible impact of treated wastewater discharge on incidence of antibiotic resistant bacteria in river water. *Water Science and Technology*, 43(2):91–99, 2001.

[227] NA Sabri, H Schmitt, B Van der Zaan, HW Gerritsen, T Zuidema, HHM Rijnaarts, and AAM Langenhoff. Prevalence of antibiotics and antibiotic resistance genes in a wastewater effluent-receiving river in the Netherlands. *Journal of Environmental Chemical Engineering*, 2018.

[228] Sotaro Kuno, Takashi Yoshida, Takakazu Kaneko, and Yoshihiko Sako. Intricate interactions between the bloom-forming cyanobacterium *Microcystis aeruginosa* and foreign genetic elements, revealed by diversified clustered regularly interspaced short palindromic repeat (CRISPR) signatures. *Applied and Environmental Microbiology*, 78(15):5353–5360, 2012.

[229] Victor Kunin, Shaomei He, Falk Warnecke, S Brook Peterson, Hector Garcia Martin, Matthew Haynes, Natalia Ivanova, Linda L Blackall, Mya Breitbart, and Forest Rohwer. A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome Research*, 18(2):293–297, 2008.

[230] Susanne M Scheierling, John B Loomis, and Robert A Young. Irrigation water demand: A meta-analysis of price elasticities. *Water Resources Research*, 42(1), 2006.

[231] Lawrence R Parsons, Bahman Sheikh, Robert Holden, and David W York. Reclaimed water as an alternative water source for crop irrigation. *Hort Science*, 45(11):1626–1629, 2010.

[232] Mar Ortega-Reig, Guillermo Palau-Salvador, Maria Josep Cascant i Sempere, Javier Benitez-Buelga, David Badiella, and Paul Trawick. The integrated use of surface, ground and recycled waste water in adapting to drought in the traditional irrigation system of valencia. *Agricultural Water Management*, 133:55–64, 2014.

[233] Martin Sndergaard, Erik Jeppesen, and Jens Peder Jensen. Pond or lake: does it make any difference? *Fundamental and Applied Limnology*, 162(2):143–165, 2005.

[234] Toni Roiha, Sari Peura, Mathieu Cusson, and Milla Rautio. Allochthonous carbon is a major regulator to bacterial growth and community composition in subarctic freshwaters. *Scientific Reports*, 6:34456, 2016.

[235] Roxane Maranger and David F Bird. Viral abundance in aquatic systems: a comparison between marine and fresh waters. *Marine Ecology Progress Series*, 121:217–226, 1995.

[236] Steven W Wilhelm and Audrey R Matteson. Freshwater and marine virioplankton: a brief overview of commonalities and differences. *Freshwater Biology*, 53(6):1076–1089, 2008.

[237] Sylvie Doublie, Stanley Tabor, Alexander M Long, Charles C Richardson, and Tom Ellenberger. Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 resolution. *Nature*, 391(6664):251–258, 1998.

[238] Helen F Schmidt, Eric G Sakowski, Shannon J Williamson, Shawn W Polson, and KEric Wommack. Shotgun metagenomics indicates novel family a DNA polymerases predominate within marine virioplankton. *The ISME Journal*, 8(1):103–114, 2014.

[239] Matthew A Saxton, Nuha S Naqvi, Faraz Rahman, Charleton P Thompson, Randolph M Chambers, James M Kaste, and Kurt E Williamson. Site-specific environmental factors control bacterial and viral diversity in stormwater retention ponds. *Aquatic Microbial Ecology*, 77(1):23–36, 2016.

[240] Seth G John, Carolina B Mendez, Li Deng, Bonnie Poulos, Anne Kathryn M Kauffman, Suzanne Kern, Jennifer Brum, Martin F Polz, Edward A Boyle, and Matthew B Sullivan. A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environmental Microbiology Reports*, 3(2):195–202, 2011.

[241] Douglas W Fadrosh, Bing Ma, Pawel Gajer, Naomi Sengamalay, Sandra Ott, Rebecca M Brotman, and Jacques Ravel. An improved dual-indexing approach

for multiplexed 16s rRNA gene sequencing on the Illumina Miseq platform. *Microbiome*, 2(1):1–7, 2014.

[242] A. P. Masella, A. K. Bartram, J. M. Truszkowski, D. G. Brown, and J. D. Neufeld. Pandaseq: paired-end assembler for illumina sequences. *BMC Bioinformatics*, 13:31, 2012.

[243] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Pena, Julia K Goodrich, and Jeffrey I Gordon. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, 2010.

[244] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jrg Peplies, and Frank Oliver Glckner. The silva ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, pages D590–D596, 2012.

[245] Joseph Nathaniel Paulson, Mihai Pop, and Hector Corrada Bravo. metagenomeseq: Statistical analysis for sparse high-throughput sequencing. *Accessed Online at, http://bioconductor.jp/packages/3.0/bioc/vignettes/metagenomeSeq/inst/doc/metagenomeSeq.pdf*, 2013.

[246] Alexander Ploner. Heatplus: heatmaps with row and/or column covariates and colored clusters. *R package version*, 2(0), 2012.

[247] Wolfgang Huber, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, and Thomas Girke. Orchestrating high-throughput genomic analysis with bioconductor. *Nature Methods*, 12(2):115–121, 2015.

[248] Jari Oksanen, Roeland Kindt, Pierre Legendre, Bob O-Hara, M Henry H Stevens, and Maintainer Jari Oksanen. The vegan package. *Community ecology package*, pages 631–637, 2007.

[249] Paul J McMurdie and Susan Holmes. phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data. *PloS One*, 2013.

[250] Matthew J Vavrek. Fossil: palaeoecological and palaeogeographical analysis tools. *Palaeontologia Electronica*, 14(1):16, 2011.

[251] Joseph N Paulson, O Colin Stine, Hctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, 10(12):1200–1202, 2013.

[252] Edward W Beals. Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data. *Advances in Ecological Research*, 14:1–55, 1984.

[253] K.R. Clarke. Non-parametric multivariate analyses of changes in community structure. *Austral Ecology*, 18(1):117–143, 1993.

[254] Catherine Lozupone, Manuel E Lladser, Dan Knights, Jesse Stombaugh, and Rob Knight. UniFrac: an effective distance metric for microbial community comparison. *The ISME Journal*, 5(2):169, 2011.

[255] Brian D Ondov, Nicholas H Bergman, and Adam M Phillippy. Interactive metagenomic visualization in a web browser. *BMC Bioinformatics*, 12(1):385, 2011.

[256] Ross Overbeek, Tadhg Begley, Ralph M Butler, Jomuna V Choudhuri, Han-Yu Chuang, Matthew Cohoon, Valrie de Crcy-Lagard, Naryttza Diaz, Terry Disz, and Robert Edwards. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, 33(17):5691–5702, 2005.

[257] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[258] Aron Marchler-Bauer, Shennan Lu, John B Anderson, Farideh Chitsaz, Myra K Derbyshire, Carol DeWeese-Scott, Jessica H Fong, Lewis Y Geer, Renata C Geer, and Noreen R Gonzales. CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Research*, 39(suppl 1):D225–D229, 2011.

[259] Matthew Kearse, Richard Moir, Amy Wilson, Steven Stones-Havas, Matthew Cheung, Shane Sturrock, Simon Buxton, Alex Cooper, Sidney Markowitz, and Chris Duran. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649, 2012.

[260] Stphane Guindon and Olivier Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704, 2003.

[261] Siddharth R Krishnamurthy and David Wang. Origins and challenges of viral dark matter. *Virus Research*, 2017.

[262] Carolina Megumi Mizuno, Francisco Rodriguez-Valera, Nikole E Kimes, and Rohit Ghai. Expanding the marine virosphere using metagenomics. *PLoS Genetics*, 9(12):e1003987, 2013.

[263] Weiqi Kuang, Jie Li, Si Zhang, and Lijuan Long. Diversity and distribution of *Actinobacteria* associated with reef coral Porites lutea. *Frontiers in Microbiology*, 6:1094, 2015.

[264] Martin Allgaier and Hans-Peter Grossart. Diversity and seasonal dynamics of *Actinobacteria* populations in four lakes in northeastern germany. *Applied and Environmental Microbiology*, 72(5):3489–3497, 2006.

[265] Angelika Rieck, Daniel PR Herlemann, Klaus Jrgens, and Hans-Peter Grossart. Particle-associated differ from free-living bacteria in surface waters of the baltic sea. *Frontiers in Microbiology*, 6, 2015.

[266] Toshio Sakamoto and Donald A Bryant. Growth at low temperature causes nitrogen limitation in the cyanobacterium *Synechococcus* sp. pcc 7002. *Archives of Microbiology*, 169(1):10–19, 1997.

[267] Katleen Van der Gucht, Tom Vandekerckhove, Nele Vloemans, Sylvie Cousin, Koenraad Muylaert, Koen Sabbe, Moniek Gillis, Steven Declerk, Luc De Meester, and Wim Vyverman. Characterization of bacterial communities in four freshwater lakes differing in nutrient load and food web structure. *FEMS Microbiology Ecology*, 53(2):205–220, 2005.

[268] Martin W Hahn, Vojt-ch Kasalick, Jan Jezbera, Ulrike Brandt, and Karel imek. *Limnohabitans australis* sp. nov., isolated from a freshwater pond, and emended description of the genus *Limnohabitans*. *International Journal of Systematic and Evolutionary Microbiology*, 60(12):2946–2950, 2010.

[269] Martin W Hahn, Matthias Pckl, and Qinglong L Wu. Low intraspecific diversity in a *Polynucleobacter* subcluster population numerically dominating bacterioplankton of a freshwater pond. *Applied and Environmental Microbiology*, 71(8):4539–4547, 2005.

[270] Karel imek, Vojtch Kasalick, Jan Jezbera, Jitka Jezberov, Josef Hejzlar, and Martin W Hahn. Broad habitat range of the phylogenetically narrow r-bt065 cluster, representing a core group of the betaproteobacterial genus *Limnohabitans*. *Applied and Environmental Microbiology*, 76(3):631–639, 2010.

[271] Emiley A Eloe, Christine N Shulse, Douglas W Fadrosh, Shannon J Williamson, Eric E Allen, and Douglas H Bartlett. Compositional differences in particle-associated and free-living microbial assemblages from an extreme deep-ocean environment. *Environmental Microbiology Reports*, 3(4):449–458, 2011.

[272] Bibiana G Crespo, Thomas Pommier, Beatriz Fernndez-Gmez, and Carlos Pedrs-Ali. Taxonomic composition of the particle-attached and free-living bacterial assemblages in the Northwest Mediterranean Sea analyzed by pyrosequencing of the 16s rRNA. *Microbiology Open*, 2(4):541–552, 2013.

[273] Martin W Hahn and Manfred G Hfle. Grazing of protozoa and its effect on populations of aquatic bacteria. *FEMS Microbiology Ecology*, 35(2):113–121, 2001.

[274] Maria W Smith, Lisa Zeigler Allen, Andrew E Allen, Lydie Herfort, and Holly M Simon. Contrasting genomic properties of free-living and particle-attached microbial assemblages within a coastal ecosystem. *Frontiers in Microbiology*, 4, 2013.

[275] Brian Palik, Darold P Batzer, Richard Buech, Dale Nichols, Kory Cease, Leanne Egeland, and Dwight E Streblow. Seasonal pond characteristics across a chronosequence of adjacent forest ages in northern Minnesota, USA. *Wetlands*, 21(4):532–542, 2001.

[276] Curtis A Suttle. Viruses in the sea. *Nature*, 437(7057):356–361, 2005.

[277] L McDaniel, LA Houchin, SJ Williamson, and JH Paul. Plankton blooms: Lysogeny in marine *Synechococcus*. *Nature*, 415(6871):496–496, 2002.

[278] Jose Luis Balcazar. Bacteriophages as vehicles for antibiotic resistance genes in the environment. *PLoS Pathogens*, 10(7):e1004219, 2014.

[279] Florent E Angly, Ben Felts, Mya Breitbart, Peter Salamon, Robert A Edwards, Craig Carlson, Amy M Chan, Matthew Haynes, Scott Kelley, and Hong Liu. The marine viromes of four oceanic regions. *PLoS Biology*, 4(11):e368, 2006.

[280] Shellie R Bench, Thomas E Hanson, Kurt E Williamson, Dhritiman Ghosh, Mark Radosovich, Kui Wang, and K Eric Wommack. Metagenomic characterization of Chesapeake Bay virioplankton. *Applied and Environmental Microbiology*, 73(23):7629–7641, 2007.

[281] Beat Oertli, Jeremy Biggs, Rgis Crghino, Patrick Grillas, Pierre Joly, and Jean-Bernard Lachavanne. Conservation and monitoring of pond biodiversity: introduction. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 15(6):535–540, 2005.

[282] Scott Taylor, Peter Gilbert, David Cooke, Michael Deary, and Mike Jeffries. High carbon burial rates by small ponds in the landscape. *Frontiers in Ecology and the Environment*, 2018.

[283] Jeremy Biggs, S Von Fumetti, and M Kelly-Quinn. The importance of small waterbodies for biodiversity and ecosystem services: implications for policy makers. *Hydrobiologia*, 793(1):3–39, 2017.

[284] Wen-Wen Chou, Soen-Han Lee, and Chen-Fa Wu. Evaluation of the preservation value and location of farm ponds in Yunlin County, Taiwan. *International Journal of Environmental Research and Public Health*, 11(1):548–572, 2014.

[285] Dani Boix, Jeremy Biggs, Rgis Crghino, Andrew P Hull, Thomas Kalettka, and Beat Oertli. Pond research and management in europe: Small is beautiful. *Hydrobiologia*, 689(1):1–9, 2012.

[286] Lynsey R Harper, Andrew S Buxton, Helen C Rees, Kat Bruce, Rein Brys, David Halfmaerten, Daniel S Read, Hayley V Watson, Carl D Sayer, and Eleanor P Jones. Prospects and challenges of environmental DNA (eDNA) monitoring in freshwater ponds. *Hydrobiologia*, pages 1–17, 2018.

[287] Francey M. *WSUD Engineering Procedures: Stormwater: Stormwater*. CSIRO Publishing, 2005.

[288] Hans W Paerl and Jef Huisman. Blooms like it hot. *Science*, 320(5872):57–58, 2008.

[289] Igor Mrdjen, Siobhan Fennessy, Alex Schaal, Richard Dennis, Joan L Slonczewski, Seungjun Lee, and Jiyoung Lee. Tile drainage and anthropogenic land use contribute to harmful algal blooms and microbiota shifts in inland water bodies. *Environmental Science and Technology*, 52(15):8215–8223, 2018.

[290] Kyung Hwa Cho, YA Pachepsky, Joon Ha Kim, AK Guber, DR Shelton, and R Rowland. Release of *Escherichia coli* from the bottom sediment in a first-order creek: Experiment and reach-specific modeling. *Journal of Hydrology*, 391(3-4):322–332, 2010.

[291] David A Chin. Linking pathogen sources to water quality in small urban streams. *Journal of Environmental Engineering*, 136(2):249–253, 2009.

[292] Ganyu Gu, Zhiyao Luo, Juan M Cevallos-Cevallos, Paige Adams, George Vellidis, Anita Wright, and Ariena HC van Bruggen. Factors affecting the occurrence of *Escherichia coli* o157 contamination in irrigation ponds on produce farms in the suwannee river watershed. *Canadian Journal of Microbiology*, 59(3):175–182, 2012.

[293] Pramod K Pandey, Philip H Kass, Michelle L Soupir, Sagor Biswas, and Vijay P Singh. Contamination of water resources by pathogenic bacteria. *AMB Express*, 4(1):51, 2014.

[294] Songhe Zhang, Si Pang, PeiFang Wang, Chao Wang, Nini Han, Bin Liu, Bing Han, Yi Li, and Kwaku Anim-Larbi. Antibiotic concentration and antibiotic-resistant bacteria in two shallow urban lakes after stormwater event. *Environmental Science and Pollution Research*, 23(10):9984–9992, 2016.

[295] Justin D Brookes, Jason Antenucci, Matthew Hipsey, Michael D Burch, Nicholas J Ashbolt, and Christobel Ferguson. Fate and transport of pathogens in lakes and reservoirs. *Environment International*, 30(5):741–759, 2004.

[296] Thomas J Sharpton. An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, 5:209, 2014.

[297] Katherine Bruder, Kema Maiki, Alexandria Cooper, Emily Sible, Jason W Shapiro, Siobhan C Watkins, and Catherine Putonti. Freshwater metaviromics

and bacteriophages: A current assessment of the state of the art in relation to bioinformatic challenges: Supplementary issue: Bioinformatics methods and applications for big metagenomics data. *Evolutionary Bioinformatics*, 12:EBO. S38549, 2016.

[298] Torbjrn Rognes, Tom Flouri, Ben Nichols, Christopher Quince, and Frdric Mah. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, 2016.

[299] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31(10):1674–1676, 2015.

[300] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.

[301] Daniel J Nasko, Jessica Chopyk, Eric G Sakowski, Barbra D Ferrell, Shawn William Polson, and K Eric Wommack. Family a DNA polymerase phylogeny uncovers diversity and replication gene organization in the virioplankton. *Frontiers in Microbiology*, 9:3053, 2018.

[302] Andrew Rambaut. Figtree version 1.4. 0. *Accessed Online at http://tree.bio .ed.ac.uk/software/figtree*, 2012.

[303] Daniel Joseph Nasko, Barbra D Ferrell, Ryan M Moore, Jaysheel Bhavsar, Shawn W Polson, and K Eric Wommack. CRISPR spacers indicate preferential matching of specific virioplankton genes. *mBio*, 10(2):1–16, 2019.

[304] Kevin E Trenberth. What are the seasons? *Bulletin of the American Meteorological Society*, 64(11):1276–1282, 1983.

[305] Anna Sajduda, Anna Brzostek, Marta Popawska, Ewa Augustynowicz-Kope, Zofia Zwolska, Stefan Niemann, Jarosaw Dziadek, and Doris Hillemann. Molecular characterization of rifampin-and isoniazid-resistant *Mycobacterium tuberculosis* strains isolated in Poland. *Journal of Clinical Microbiology*, 42(6):2425–2431, 2004.

[306] Catherine Vilcheze, Torin R Weisbrod, Bing Chen, Laurent Kremer, Manzour H Hazbn, Feng Wang, David Alland, James C Sacchettini, and William R Jacobs. Altered nadh/nad+ ratio mediates coresistance to isoniazid and ethionamide in mycobacteria. *Antimicrobial Agents and Chemotherapy*, 49(2):708–720, 2005.

[307] Ruud Jansen, Jan Embden, Wim Gaastra, and Leo Schouls. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology*, 43(6):1565–1575, 2002.

[308] Rotem Sorek, Victor Kunin, and Philip Hugenholtz. CRISPR–a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nature Reviews. Microbiology*, 6(3):181, 2008.

[309] David Dudgeon, Angela H Arthington, Mark O Gessner, ZenIchiro Kawabata, Duncan J Knowler, Christian Lvque, Robert J Naiman, Anne Hlne Prieur Richard, Doris Soto, and Melanie LJ Stiassny. Freshwater biodiversity: importance, threats, status and conservation challenges. *Biological Reviews*, 81(2):163–182, 2006.

[310] Robert I McDonald, Pamela Green, Deborah Balk, Balazs M Fekete, Carmen Revenga, Megan Todd, and Mark Montgomery. Urban growth, climate change, and freshwater availability. *Proceedings of the National Academy of Sciences*, 108(15):6312–6317, 2011.

[311] Marieke Beaulieu, Frances Pick, and Irene Gregory-Eaves. Nutrients and water temperature are significant predictors of cyanobacterial biomass in a 1147 lakes data set. *Limnology and Oceanography*, 58(5):1736–1746, 2013.

[312] YoonKyung Cha, Kyung Hwa Cho, Hyuk Lee, Taegu Kang, and Joon Ha Kim. The relative importance of water temperature and residence time in predicting cyanobacteria abundance in regulated rivers. *Water Research*, 124:11–19, 2017.

[313] Walter K Dodds, Dolly A Gudder, and Dieter Mollenhauer. The ecology of *Nostoc*. *Journal of Phycology*, 31(1):2–18, 1995.

[314] B Oudra, M Dadi-El Andaloussi, and VM Vasconcelos. Identification and quantification of microcystins from a *Nostoc muscorum* bloom occurring in oukameden river (high-atlas mountains of marrakech, morocco). *Environmental Monitoring and Assessment*, 149(1-4):437–444, 2009.

[315] Rainer Kurmayer. The toxic *Cyanobacterium Nostoc* sp. strain 152 produces highest amounts of microcystic and nostophycon under stress conditions 1. *Journal of Phycology*, 47(1):200–207, 2011.

[316] United States Environmental Protection Agency. Nutrient policy and data: Health and ecological effects. *Accessed Online at, https://www.epa.gov/nutrient-policy-data/health-and-ecological-effects*, 2017.

[317] Yoshio Ueno, Satoshi Nagata, Tomoaki Tsutsumi, Akihiro Hasegawa, Mariyo F Watanabe, Ho-Dong Park, Gong-Chao Chen, Gang Chen, and Shun-Zhang Yu. Detection of microcystins, a blue-green algal hepatotoxin, in drinking water sampled in haimen and fusui, endemic areas of primary liver cancer in china, by highly sensitive immunoassay. *Carcinogenesis*, 17(6):1317–1321, 1996.

[318] Zhou Lun, Yu Hai, and Chen Kun. Relationship between microcystin in drinking water and colorectal cancer. *Biomedical and Environmental Sciences*, 15(2):166–171, 2002.

[319] Alexander Eiler and Stefan Bertilsson. *Flavobacteria* blooms in four eutrophic lakes: linking population dynamics of freshwater bacterioplankton to resource availability. *Applied Environmental Microbiology*, 73(11):3511–3518, 2007.

[320] Nikea Ulrich, Abigail Rosenberger, Colin Brislawn, Justin Wright, Collin Kessler, David Toole, Caroline Solomon, Steven Strutt, Erin McClure, and Regina Lamendella. Restructuring of the aquatic bacterial community by hydric dynamics associated with superstorm sandy. *Applied and Environmental Microbiology*, 82(12):3525–3536, 2016.

[321] Jinjun Kan. Storm events restructured bacterial community and their biogeochemical potentials. *Journal of Geophysical Research: Biogeosciences*, 123(7):2257–2269, 2018.

[322] L Fiksdal, JS Maki, SJ LaCroix, and JT Staley. Survival and detection of *Bacteroides* spp., prospective indicator bacteria. *Applied Environmental Microbiology*, 49(1):148–150, 1985.

[323] Alice Layton, Larry McKay, Dan Williams, Victoria Garrett, Randall Gentry, and Gary Sayler. Development of *Bacteroides* 16s rRNA gene taqman-based real-time pcr assays for estimation of total, human, and bovine fecal pollution in water. *Applied Environmental Microbiology*, 72(6):4214–4224, 2006.

[324] Linda K Dick, Anne E Bernhard, Timothy J Brodeur, Jorge W Santo Domingo, Joyce M Simpson, Sarah P Walters, and Katharine G Field. Host distributions of uncultivated fecal *Bacteroidales* bacteria reveal genetic markers for fecal source identification. *Applied Environmental Microbiology*, 71(6):3184–3191, 2005.

[325] Stuart B Levy and Bonnie Marshall. Antibacterial resistance worldwide: causes, challenges and responses. *Nature Medicine*, 10(12s):S122, 2004.

[326] Elisabet Marti, Eleni Variatza, and Jose Luis Balcazar. The role of aquatic ecosystems as reservoirs of antibiotic resistance. *Trends in Microbiology*, 22(1):36–41, 2014.

[327] Nadine Czekalski, Radhika Sigdel, Julia Birtel, Blake Matthews, and Helmut Brgmann. Does human activity impact the natural antibiotic resistance background- abundance of antibiotic resistance genes in 21 Swiss lakes. *Environment International*, 81:45–55, 2015.

[328] Yuyi Yang, Wenjuan Song, Hui Lin, Weibo Wang, Linna Du, and Wei Xing. Antibiotics and antibiotic resistance genes in global lakes: A review and meta-analysis. *Environment International*, 116:60–73, 2018.

[329] Milind G Watve, Rashmi Tickoo, Maithili M Jog, and Bhalachandra D Bhole. How many antibiotics are produced by the genus *Streptomyces*? *Archives of Microbiology*, 176(5):386–390, 2001.

[330] Kirandeep Bhullar, Nicholas Waglechner, Andrew Pawlowski, Kalinka Koteva, Eric D Banks, Michael D Johnston, Hazel A Barton, and Gerard D Wright. Antibiotic resistance is prevalent in an isolated cave microbiome. *PloS One*, 7(4):e34953, 2012.

[331] S Egan, P Wiener, D Kallifidas, and EMH Wellington. Phylogeny of *Streptomyces* species and evidence for horizontal transfer of entire and partial antibiotic gene clusters. *Antonie Van Leeuwenhoek, Journal of Microbiology*, 79(2):127–133, 2001.

[332] Maulik N Thaker, Wenliang Wang, Peter Spanogiannopoulos, Nicholas Waglechner, Andrew M King, Ricardo Medina, and Gerard D Wright. Identifying producers of antibacterial compounds by screening for antibiotic resistance. *Nature Biotechnology*, 31(10):922, 2013.

[333] Motoshi Suzuki, Shonen Yoshida, Elinor T Adman, A Blank, and Lawrence A Loeb. *Thermus aquaticus* DNA polymerase I mutants with altered fidelity interacting mutations in the o-helix. *Journal of Biological Chemistry*, 275(42):32728–32735, 2000.

[334] Magali Leroy, Magali Prigent, Murielle Dutertre, Fabrice Confalonieri, and Michael Dubow. Bacteriophage morphotype and genome diversity in seine river sediment. *Freshwater Biology*, 53(6):1176–1185, 2008.

[335] Cristina Howard-Varona, Katherine R Hargreaves, Stephen T Abedon, and Matthew B Sullivan. Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *The ISME Journal*, 11(7):1511, 2017.

[336] SJ Williamson, LA Houchin, L McDaniel, and JH Paul. Seasonal variation in lysogeny as depicted by prophage induction in Tampa Bay, Florida. *Applied and Environmental Microbiology*, 68(9):4307–4314, 2002.

[337] Jrme P Payet and Curtis A Suttle. To kill or not to kill: the balance between lytic and lysogenic viral infection is driven by trophic status. *Limnology and Oceanography*, 58(2):465–474, 2013.

[338] Elena Lara, Dolors Vaqu, Elisabet Laia S, Julia A Boras, Ana Gomes, Encarna Borrull, Cristina Dez-Vives, Eva Teira, Massimo C Pernice, and Francisca C Garcia. Unveiling the role and life strategies of viruses from the surface to the dark ocean. *Science Advances*, 3(9):e1602565, 2017.

[339] B Knowles, CB Silveira, BA Bailey, K Barott, VA Cantu, AG Cobin-Gemes, FH Coutinho, EA Dinsdale, B Felts, and KA Furby. Lytic to temperate switching of viral communities. *Nature*, 531(7595):466, 2016.

[340] Cynthia B Silveira and Forest L Rohwer. Piggyback-the-winner in host-associated microbial communities. *NPJ Biofilms and Microbiomes*, 2:16010, 2016.

[341] Ricardo Costeira, Rory Doherty, Christopher CR Allen, Michael J Larkin, and Leonid A Kulakov. Analysis of viral and bacterial communities in groundwater associated with contaminated land. *Science of The Total Environment*, 656:1413–1426, 2019.

[342] Sandra Chibani-Chennoufi, Anne Bruttin, Marie-Lise Dillmann, and Harald Brssow. Phage-host interaction: an ecological perspective. *Journal of Bacteriology*, 186(12):3677–3686, 2004.

[343] Paul Hyman and Stephen T Abedon. *Bacteriophage host range and bacterial resistance*, volume 70, pages 217–248. Elsevier, 2010.

[344] Anders F Andersson and Jillian F Banfield. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science*, 320(5879):1047–1050, 2008.

[345] Joanne B Emerson, Karen Andrade, Brian C Thomas, Anders Norman, Eric E Allen, Karla B Heidelberg, and Jillian F Banfield. Virus-host and CRISPR dynamics in archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia. *Archaea*, 2013, 2013.

[346] Yakov Pachepsky, Daniel R Shelton, Jean ET McLain, Jitendra Patel, and Robert E Mandrell. *Irrigation waters as a source of pathogenic microorganisms in produce: a review*, volume 113, pages 75–141. Elsevier, 2011.

[347] United States Environmental Protection Agency. National rivers and streams assessment 2008-2009: a collaborative survey, 2013.

[348] Robert G Wetzel. *Limnology: Lake and River Ecosystems*. Gulf Professional Publishing, 2001.

[349] Clara Ruiz-Gonzlez, Juan Pablo Nio-Garca, and Paul A del Giorgio. Terrestrial origin of bacterial communities in complex boreal freshwater networks. *Ecology Letters*, 18(11):1198–1206, 2015.

[350] Lydia H Zeglin. Stream microbial diversity in response to environmental changes: review and synthesis of existing research. *Frontiers in Microbiology*, 6:454, 2015.

[351] Feng-Hua Wang, Min Qiao, Jian-Qiang Su, Zheng Chen, Xue Zhou, and Yong-Guan Zhu. High throughput profiling of antibiotic resistance genes in urban park soils with reclaimed water irrigation. *Environmental Science and Technology*, 48(16):9079–9085, 2014.

[352] Christopher Staley, Trevor J Gould, Ping Wang, Jane Phillips, James B Co-Pacifictner, and Michael J Sadowsky. Species sorting and seasonal dynamics primarily shape bacterial communities in the upper Mississippi River. *Science of the Total Environment*, 505:435–445, 2015.

[353] Juan Pablo Nio-Garca, Clara Ruiz-Gonzlez, and Paul A del Giorgio. Interactions between hydrology and water chemistry shape bacterioplankton biogeography across boreal freshwater networks. *The ISME journal*, 10(7):1755, 2016.

[354] Daniel S Read, Hyun S Gweon, Michael J Bowes, Lindsay K Newbold, Dawn Field, Mark J Bailey, and Robert I Griffiths. Catchment-scale biogeography of riverine bacterioplankton. *The ISME Journal*, 9(2):516, 2015.

[355] Wenguang Xiong, Yongxue Sun, Xueyao Ding, Yiming Zhang, and Zhenling Zeng. Antibiotic resistance genes occurrence and bacterial community composition in the Liuxi River. *Frontiers in Environmental Science*, 2:61, 2014.

[356] Sara Rodriguez-Mozaz, Sara Chamorro, Elisabet Marti, Belinda Huerta, Meritxell Gros, Alexandre Snchez-Melsi, Carles M Borrego, Dami Barcel, and Jose Luis Balczar. Occurrence of antibiotics and antibiotic resistance genes in hospital and urban wastewaters and their impact on the receiving river. *Water Research*, 69:234–242, 2015.

[357] Yael Negreanu, Zohar Pasternak, Edouard Jurkevitch, and Eddie Cytryn. Impact of treated wastewater irrigation on antibiotic resistance in agricultural soils. *Environmental Science and Technology*, 46(9):4800–4808, 2012.

[358] Karl Fent, Anna A Weston, and Daniel Caminada. Ecotoxicology of human pharmaceuticals. *Aquatic Toxicology*, 76(2):122–159, 2006.

[359] Milan Kolr, Karel Urbnek, and Tom Ltal. Antibiotic selective pressure and development of bacterial resistance. *International Journal of Antimicrobial Agents*, 17(5):357–363, 2001.

[360] Mya Breitbart, Chelsea Bonnain, Kema Malki, and Natalie A Sawaya. Phage puppet masters of the marine microbial realm. *Nature Microbiology*, 2018.

[361] Sarah M Allard, Mary Theresa Callahan, Anthony Bui, Angela Marie C Ferelli, Jessica Chopyk, Suhana Chattopadhyay, Emmanuel F Mongodin, Shirley A Micallef, and Amy R Sapkota. Creek to table: Tracking fecal indicator bacteria, bacterial pathogens, and total bacterial communities from irrigation water to kale and radish crops. *Science of The Total Environment*, 2019.

[362] Lihong Chen, Jian Yang, Jun Yu, Zhijian Yao, Lilian Sun, Yan Shen, and Qi Jin. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Research*, 33(suppl_1):D325–D328, 2005.

[363] Mei Hu, Sobhan Nandi, Christopher Davies, and Robert A Nicholas. High-level chromosomally mediated tetracycline resistance in *Neisseria gonorrhoeae* results from a point mutation in the *rpsJ* gene encoding ribosomal protein s10 in combination with the *mtrR* and *penB* resistance determinants. *Antimicrobial Agents and Chemotherapy*, 49(10):4327–4334, 2005.

[364] T Hogg, JR Mesters, and R Hilgenfeld. Inhibitory mechanisms of antibiotics targeting elongation factor tu. *Current Protein Neisseria gonorrhoeae and Peptide Science*, 3(1):121–131, 2002.

[365] Evelien M Adriaenssens, Johannes Wittmann, Jens H Kuhn, Dann Turner, Matthew B Sullivan, Bas E Dutilh, Ho Bin Jang, Leonardo J van Zyl, Jochen Klumpp, and Malgorzata Lobocka. Taxonomy of prokaryotic viruses: 2017 update from the ICTV bacterial and archaeal viruses subcommittee. *Archives of Virology*, 163(4):1125–1129, 2018.

[366] Meredith AJ Hullar, Louis A Kaplan, and David A Stahl. Recurring seasonal dynamics of microbial communities in stream habitats. *Applied and Environmental Microbiology*, 72(1):713–722, 2006.

[367] Byron C Crump, Bruce J Peterson, Peter A Raymond, Rainer MW Amon, Amanda Rinehart, James W McClelland, and Robert M Holmes. Circumpolar synchrony in big river bacterioplankton. *Proceedings of the National Academy of Sciences*, page pnas. 0906149106, 2009.

[368] Hirohito Tsurumaru, Takashi Okubo, Kazuyuki Okazaki, Megumi Hashimoto, Kaori Kakizaki, Eiko Hanzawa, Hiroyuki Takahashi, Noriyuki Asanome, Fukuyo Tanaka, and Yasuyo Sekiyama. Metagenomic analysis of the bacterial community associated with the taproot of sugar beet. *Microbes and Environments*, 30(1):63–69, 2015.

[369] Julien Crovadore, Ali Asaff Torres, Ral Rodrguez Heredia, Bastien Cochard, Romain Chablais, and Franois Lefort. Metagenomes of soil samples from an established perennial cropping system of asparagus treated with biostimulants in southern france. *Genome Announcements*, 5(24):e00511–17, 2017.

[370] Luis H Orellana, Joanne C Chee-Sanford, Robert A Sanford, Frank E Lffler, and Konstantinos T Konstantinidis. Year-round shotgun metagenomes reveal stable microbial communities in agricultural soils and novel ammonia oxidizers responding to fertilization. *Applied and Environmental Microbiology*, 84(2):e01646–17, 2018.

[371] Elizabeth M Bach, Ryan J Williams, Sarah K Hargreaves, Fan Yang, and Kirsten S Hofmockel. Greatest soil microbial diversity found in micro-habitats. *Soil Biology and Biochemistry*, 118:217–226, 2018.

[372] Joseph Nesme, Sbastien Ccillon, Tom O Delmont, Jean-Michel Monier, Timothy M Vogel, and Pascal Simonet. Large-scale metagenomic-based study of

antibiotic resistance in the environment. *Current Biology*, 24(10):1096–1100, 2014.

[373] R Canten. Antibiotic resistance genes from the environment: a perspective through newly identified antibiotic resistance mechanisms in the clinical setting. *Clinical Microbiology and Infection*, 15:20–25, 2009.

[374] Kevin J Forsberg, Alejandro Reyes, Bin Wang, Elizabeth M Selleck, Morten OA Sommer, and Gautam Dantas. The shared antibiotic resistome of soil bacteria and human pathogens. *Science*, 337(6098):1107–1111, 2012.

[375] Alinne P de Castro, Gabriel da R Fernandes, and Octvio L Franco. Insights into novel antimicrobial compounds and antibiotic resistance genes from soil metagenomes. *Frontiers in Microbiology*, 5:489, 2014.

[376] Chao Niu, Dong Yu, Yuelan Wang, Hongguang Ren, Yuan Jin, Wei Zhou, Beiping Li, Yiyong Cheng, Junjie Yue, and Zhixian Gao. Common and pathogen-specific virulence factors are different in function and structure. *Virulence*, 4(6):473–482, 2013.

[377] Louis S Ates, EN Houben, and Wilbert Bitter. Type VII secretion: a highly versatile secretion system. *Microbiology Spectrum*, 4(1), 2016.

[378] Swati Shah and Volker Briken. Modular organization of the ESX-5 secretion system in *Mycobacterium tuberculosis*. *Frontiers in Cellular and Infection Microbiology*, 6:49, 2016.

[379] JoAnn M Tufariello, Jessica R Chapman, Christopher A Kerantzas, Ka-Wing Wong, Catherine Vilchze, Christopher M Jones, Laura E Cole, Emir Tinaztepe, Victor Thompson, and David Feny. Separable roles for *Mycobacterium tuberculosis* ESX-3 effectors in iron acquisition and virulence. *Proceedings of the National Academy of Sciences*, 113(3):E348–E357, 2016.

[380] Bryan Coburn, Inna Sekirov, and B Brett Finlay. Type III secretion systems and disease. *Clinical Microbiology Reviews*, 20(4):535–549, 2007.

[381] Konstantin V Korotkov, Maria Sandkvist, and Wim GJ Hol. The type II secretion system: biogenesis, molecular architecture and mechanism. *Nature Reviews Microbiology*, 10(5):336, 2012.

[382] Joachim Frey and Francesco C Origgi. Type III secretion system of *Aeromonas salmonicida* undermining the host's immune response. *Frontiers in Marine Science*, 3:130, 2016.

[383] Isoken H Igbinosa, Ehimario U Igumbor, Farhad Aghdasi, Mvuyo Tom, and Anthony I Okoh. Emerging *Aeromonas* species infections and their significance in public health. *The Scientific World Journal*, 2012, 2012.

[384] Hiroshi Hirotani, Chiaki Sese, and Hisanori Kagawa. Correlations of *Aeromonas hydrophila* with indicator bacteria of water quality and environmental factors in a mountain stream. *Water Environment Research*, 71(2):132–138, 1999.

[385] Patrick K Jjemba, Lauren A Weinrich, Wei Cheng, Eugenio Giraldo, and Mark W LeChevallier. Regrowth of potential opportunistic pathogens and algae in reclaimed-water distribution systems. *Applied and Environmental Microbiology*, 76(13):4169–4178, 2010.

[386] Andrew K Leight, Byron C Crump, and Raleigh R Hood. Assessment of fecal indicator bacteria and potential pathogen co-occurrence at a shellfish growing area. *Frontiers in Microbiology*, 9:384, 2018.

[387] C Douglas Hershberger, Rick W Ye, Matthew R Parsek, Zhi-Dong Xie, and AM Chakrabarty. The *algT* (*algU*— gene of *Pseudomonas aeruginosa*, a key regulator involved in alginate biosynthesis, encodes an alternative sigma factor (sigma e). *Proceedings of the National Academy of Sciences*, 92(17):7941–7945, 1995.

[388] Jeff G Leid, Carey J Willson, Mark E Shirtliff, Daniel J Hassett, Matthew R Parsek, and Alyssa K Jeffers. The exopolysaccharide alginate protects *Pseudomonas aeruginosa* biofilm bacteria from ifn-gamma-mediated macrophage killing. *The Journal of Immunology*, 175(11):7512–7518, 2005.

[389] Jost Wingender and Hans-Curt Flemming. Biofilms in drinking water and their role as reservoir for pathogens. *International Journal of Hygiene and Environmental Health*, 214(6):417–423, 2011.

[390] Jane Segobola, Evelien Adriaenssens, Tsepo Tsekoa, Konanani Rashamuse, and Don Cowan. Exploring viral diversity in a unique South African soil habitat. *Scientific Reports*, 8(1):111, 2018.

[391] Britt Koskella and Sean Meaden. Understanding bacteriophage specificity in natural microbial communities. *Viruses*, 5(3):806–823, 2013.

[392] Kema Malki, Alex Kula, Katherine Bruder, Emily Sible, Thomas Hatzopoulos, Stephanie Steidel, Siobhan C Watkins, and Catherine Putonti. Bacteriophages isolated from Lake Michigan demonstrate broad host-range across several bacterial phyla. *Virology Journal*, 12(1):164, 2015.

[393] James I Prosser. Replicate or lie. *Environmental Microbiology*, 12(7):1806–1810, 2010.

[394] PC Chiu. Applications of zero-valent iron (ZVI) and nanoscale ZVI to municipal and decentralized drinking water systems-a review. *Novel Solutions to Water Pollution*, pages 237–249, 2013.

[395] Franz-Georg Simon and Tamas Meggyes. Removal of organic and inorganic pollutants from groundwater using permeable reactive barriers: part 1. treatment processes for pollutants. *Land Contamination and Reclamation*, 8(2):103–116, 2000.

[396] Paul G Tratnyek, Michelle M Scherer, Timothy L Johnson, and Leah J Matheson. Permeable reactive barriers of iron and other zero-valent metals. *Environmental Science and Pollution Control Series*, pages 371–422, 2003.

[397] Youwen You, Jie Han, Pei C Chiu, and Yan Jin. Removal and inactivation of waterborne viruses using zerovalent iron. *Environmental Science and Technology*, 39(23):9263–9269, 2005.

[398] Changha Lee, Jee Yeon Kim, Won Il Lee, Kara L Nelson, Jeyong Yoon, and David L Sedlak. Bactericidal effect of zero-valent iron nanoparticles on *Escherichia coli*. *Environmental Science and Technology*, 42(13):4927–4933, 2008.

[399] DT Ingram, Mary Theresa Callahan, Sean Ferguson, DG Hoover, DR Shelton, PD Millner, MJ Camp, JR Patel, KiE Kniel, and M Sharma. Use of zero-valent iron biosand filters to reduce *Escherichia coli* O157:H12 in irrigation water applied to spinach plants in a field setting. *Journal of Applied Microbiology*, 112(3):551–560, 2012.

[400] Chunjian Shi, Jie Wei, Yan Jin, Kalmia E Kniel, and Pei C Chiu. Removal of viruses and bacteriophages from drinking water using zero-valent iron. *Separation and Purification Technology*, 84:72–78, 2012.

[401] Fenglian Fu, Dionysios D Dionysiou, and Hong Liu. The use of zero-valent iron for groundwater remediation and wastewater treatment: a review. *Journal of Hazardous Materials*, 267:194–205, 2014.

[402] Adrienne EH Shearer and Kalmia E. Kniel. Enhanced removal of norovirus surrogates, murine norovirus and tulane virus, from aqueous systems by zero-valent iron. *Journal of Food Protection*, 81.9(2):1432–1438, 2018.

[403] Dana Willner, Mike Furlan, Matthew Haynes, Robert Schmieder, Florent E Angly, Joas Silva, Sassan Tammadoni, Bahador Nosrat, Douglas Conrad, and Forest Rohwer. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PloS One*, 4(10):e7370, 2009.

[404] L Rizzo, C Manaia, C Merlin, T Schwartz, C Dagot, MC Ploy, I Michael, and D Fatta-Kassinos. Urban wastewater treatment plants as hotspots for antibiotic resistant bacteria and genes spread into the environment: a review. *Science of the Total Environment*, 447:345–360, 2013.

[405] Prachi Kulkarni, Greg Raspanti, Anthony Q Bui, Rhodel N Bradshaw, Kalmia E Kniel, Pei C Chiu, Manan Sharma, Amir Sapkota, and Amy R Sapkota. Zerovalent iron-sand filtration can reduce the concentration of multiple antimicrobials in conventionally treated reclaimed water. *Environmental Research*, 172:301–309, 2019.

[406] Triple Quest. Hydraid (r) biosand water filter installation manual. *C. Engineering (Ed.), Grand rapids*, pages 1–20, 2010.

[407] David H Manz. Slow sand filter for use with intermittently flowing water supply and method of use thereof. *U.S. Patent*, 6,123,858, 2000.

[408] Center for Affordable Water Technology and Sanitation. How was the biosand filter pore volume determined? *Accessed Online at, https://www.biosand filters.info/faq/da59051b/how-was-the-biosand-filter-pore-volume-determined*, 2015.

[409] Kurt E Williamson, K Eric Wommack, and Mark Radosevich. Sampling natural viral communities from soil for culture-independent analyses. *Applied and Environmental Microbiology*, 69(11):6628–6633, 2003.

[410] U.S. Environmental Protection Agency. Potable reuse compendium. *Accessed Online at, https://www.epa.gov/sites/production/files/2018-01/documents /potablereusecompendium_3.pdf*, 2017.

[411] Joseph N Ryan, Ronald W Harvey, David Metge, Menachem Elimelech, Theresa Navigato, and Ann P Pieper. Field and laboratory investigations of inactivation of viruses (PRD1 and MS2) attached to iron oxide-coated quartz sand. *Environmental Science and Technology*, 36(11):2403–2413, 2002.

[412] B Michen and T Graule. Isoelectric points of viruses. *Journal of Applied Microbiology*, 109(2):388–397, 2010.

[413] Johanna M Rinta-Kanto, Markku J Lehtola, Terttu Vartiainen, and Pertti J Martikainen. Rapid enumeration of virus-like particles in drinking water samples using SYBR green i-staining. *Water Research*, 38(10):2614–2618, 2004.

[414] Hanting Wang, Takashi Narihiro, Anthony P Straub, Charles R Pugh, Hideyuki Tamaki, Johnathan F Moor, Ian M Bradley, Yoichi Kamagata, Wen-Tso Liu, and Thanh H Nguyen. MS2 bacteriophage reduction and microbial communities in biosand filters. *Environmental Science and Technology*, 48(12):6702–6709, 2014.

[415] Eun-Jin Park, Kyoung-Ho Kim, Guy CJ Abell, Min-Soo Kim, Seong Woon Roh, and Jin-Woo Bae. Metagenomic analysis of the viral communities in fermented foods. *Applied and Environmental Microbiology*, 77(4):1284–1291, 2011.

[416] David M DeMarini and B Kay Lawrence. Prophage induction by DNA topoisomerase II poisons and reactive-oxygen species: role of DNA breaks. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 267(1):1–17, 1992.

[417] Arun M Nanda, Kai Thormann, and Julia Frunzke. Impact of spontaneous prophage induction on the fitness of bacterial populations and host-microbe interactions. *Journal of Bacteriology*, 197(3):410–419, 2015.

[418] Qian Zhang, Guo-Qing Shi, Guang-Peng Tang, Zhi-Tin Zou, Guang-Hai Yao, and Guang Zeng. A foodborne outbreak of *Aeromonas hydrophila* in a college, Xingyi City, Guizhou, China, 2012. *Western Pacific Surveillance and Response Journal*, 3(4):39, 2012.

[419] Sren Molin and Tim Tolker-Nielsen. Gene transfer occurs with enhanced efficiency in biofilms and induces enhanced stabilisation of the biofilm structure. *Current Opinion in Biotechnology*, 14(3):255–261, 2003.

[420] HT Solheim, Camilla Sekse, Anne Margrete Urdahl, Yngvild Wasteson, and Live Lingaas Nesse. Biofilm as an environment for dissemination of stx genes by transduction. *Applied and Environmental Microbiology*, 79(3):896–900, 2013.

[421] World Health Organization. *Guidelines for drinking-water quality*, volume 1. World Health Organization, 2004.

[422] Gary Bentrup. *Conservation buffers: design guidelines for buffers, corridors, and greenways*. US Department of Agriculture, Forest Service, Southern Research Station Asheville, NC, 2008.

[423] Rolf D Vinebrooke and Peter R Leavitt. Phytobenthos and phytoplankton as potential indicators of climate change in mountain lakes and ponds: a hplc-based pigment approach. *Journal of the North American Benthological Society*, 18(1):15–33, 1999.

[424] The U.S. Global Change Research Program. 2014 national climate assessment. *Accessed Online at, https://nca2014.globalchange.gov/report*, 2014.

[425] Maxwell JC. The Maxwell Report: Year end and fourth quarter 2014 cigarette industry. Report, 2014.

[426] Ralph S Caraballo and Katherine Asman. Epidemiology of menthol cigarette use in the United States. *Tobacco Induced Diseases*, 9(Suppl 1):S1, 2011.

[427] Christopher R McCurdy and Stephen S Scully. Analgesic substances derived from natural products (natureceuticals). *Life Sciences*, 78(5):476–484, 2005.

[428] K Schfer, HA Braun, and C Isenberg. Effect of menthol on cold receptor activity. analysis of receptor processes. *The Journal of General Physiology*, 88(6):757–776, 1986.

[429] Charyn D Sutton and Robert G Robinson. The marketing of menthol cigarettes in the United States: populations, messages, and channels. *Nicotine and Tobacco Research*, 6(Suppl 1):S83–S91, 2004.

[430] JR Reid. A history of mentholated cigarettes: This spud's for you. *Recent Advances in Tobacco Sciences*, 19:71–84, 1993.

[431] Jiu Ai, Kenneth M Taylor, Joseph G Lisko, Hang Tran, Clifford H Watson, and Matthew R Holman. Menthol content in us marketed cigarettes. *Nicotine and Tobacco Research*, page ntv162, 2015.

[432] Geoffrey Ferris Wayne and Gregory N Connolly. Application, function, and effects of menthol in cigarettes: a survey of tobacco industry documents. *Nicotine and Tobacco Research*, 6(Suppl 1):S43–S54, 2004.

[433] Phillip S Gardiner. The African Americanization of menthol cigarette use in the United States. *Nicotine and Tobacco Research*, 6(Suppl 1):S55–S65, 2004.

[434] NL Benowitz, A Blum, RL Braithwaite, and FG Castro. Tobacco use among us racial/ethnic minority groups-African Americans, American Indians and Alaska natives, Asian Americans and Pacific islanders, and Hispanics: a report of the surgeon general. *A Report of the Surgeon General*, 1998.

[435] Pamela I Clark, Shiva Gautam, and Lowell W Gerson. Effect of menthol cigarettes on biochemical markers of smoke exposure among black and white smokers. *CHEST Journal*, 110(5):1194–1198, 1996.

[436] ME Jarvik, DP Tashkin, NH Caskey, WJ McCarthy, and MR Rosenblatt. Mentholated cigarettes decrease puff volume of smoke and increase carbon monoxide absorption. *Physiology and Behavior*, 56(3):563–570, 1994.

[437] SM Gordon, MC Brinkman, RQ Meng, GM Anderson, JC Chuang, RR Kroeger, IL Reyes, and PI Clark. Effect of cigarette menthol content on mainstream smoke emissions. *Chemical Research in Toxicology*, 24(10):1744–1753, 2011.

[438] Joshua E Muscat, Gang Chen, Ashley Knipe, Steven D Stellman, Philip Lazarus, and John P Richie. Effects of menthol on tobacco smoke exposure, nicotine dependence, and NNAL glucuronidation. *Cancer Epidemiology Biomarkers and Prevention*, 18(1):35–41, 2009.

[439] Daniel R Brooks, Julie R Palmer, Brian L Strom, and Lynn Rosenberg. Menthol cigarettes and risk of lung cancer. *American Journal of Epidemiology*, 158(7):609–616, 2003.

[440] Robert P Murray, John E Connett, Melissa A Skeans, and Donald P Tashkin. Menthol cigarettes and health risks in lung health study data. *Nicotine and Tobacco Research*, 9(1):101–107, 2007.

[441] Stephen Sidney, Irene S Tekawa, Gary D Friedman, Marianne C Sadler, and Donald P Tashkin. Mentholated cigarette use and lung cancer. *Archives of Internal Medicine*, 155(7):727–732, 1995.

[442] Samuel Garten and R Victor Falkner. Role of mentholated cigarettes in increased nicotine dependence and greater risk of tobacco-attributable disease. *Preventive Medicine*, 38(6):793–798, 2004.

[443] Gkalp Iscan, NESE KIrimer, M-ne Krkcoglu, Hsn Can Baser, and FATIh DEMIrci. Antimicrobial screening of mentha piperita essential oils. *Journal of Agricultural and Food Chemistry*, 50(14):3943–3946, 2002.

[444] JL Pauly, JD Waight, and GM Paszkiewicz. Tobacco flakes on cigarette filters grow bacteria: a potential health risk to the smoker? *Tobacco Control*, 17(Suppl1):i49–i52, 2008.

[445] T Tabuchi. Microbial degradation of nicotine and nicotinic acid. i. isolation of nicotine-decomposing bacteria and their morphological and physiological properties. *Journal of the Agricultural Chemical Society of Japan*, 28:806–810, 1954.

[446] Russell L Stedman. Chemical composition of tobacco and tobacco smoke. *Chemical Reviews*, 68(2):153–207, 1968.

[447] VP Kurup, A Resnick, SL Kagen, SH Cohen, and JN Fink. Allergenic fungi and actinomycetes in smoking materials and their health implications. *Mycopathologia*, 82(1):61–64, 1983.

[448] Lennart Larsson, Bogumila Szponar, Beston Ridha, Christina Pehrson, Jacek Dutkiewicz, E Krysinska-Traczyk, and Jolanta Sitkowska. Identification of bacterial and fungal components in tobacco and tobacco smoke. *Tobacco Induced Diseases*, 4(4):1–8, 2008.

[449] Alejandro P Rooney, James L Swezey, Donald T Wicklow, and Matthew J McAtee. Bacterial species diversity in cigarettes linked to an investigation of severe pneumonitis in us military personnel deployed in operation iraqi freedom. *Current Microbiology*, 51(1):46–52, 2005.

[450] Twilla Eaton, JO Falkinham 3rd, and C Fordham von Reyn. Recovery of *Mycobacterium avium* from cigarettes. *Journal of Clinical Microbiology*, 33(10):2757, 1995.

[451] Amy R Sapkota, Sibel Berger, and Timothy M Vogel. Human pathogens abundant in the bacterial metagenome of cigarettes. *Environmental Health Perspectives*, 118(3):351, 2010.

[452] Ian C MacGregor, Stephen B Stanfill, Sydney M Gordon, Douglas J Turner, Jenny M Butler, Elizabeth A Hanft, Hyoshin Kim, Robyn R Kroeger, Marielle C Brinkman, and Margaret E Tefft. Custom mentholation of commercial cigarettes for research purposes. *Toxicology Reports*, 1:1068–1075, 2014.

[453] Hope T Jackson, Emmanuel F Mongodin, Katherine P Davenport, Claire M Fraser, Anthony D Sandler, and Steven L Zeichner. Culture-independent evaluation of the appendix and rectum microbiomes in children with and without appendicitis. *PloS One*, 9(4):e95414, 2014.

[454] Margaret L Zupancic, Brandi L Cantarel, Zhenqiu Liu, Elliott F Drabek, Kathleen A Ryan, Shana Cirimotich, Cheron Jones, Rob Knight, William A Walters, and Daniel Knights. Analysis of the gut microbiota in the old order amish and its relation to the metabolic syndrome. *PloS One*, 2012.

[455] Gregory R Warnes, Ben Bolker, L Bonebakker, R Gentleman, W Huber, A Liaw, T Lumley, M Maechler, A Magnusson, and S Moeller. gplots: Various R programming tools for plotting data. *R package version*, 2(4), 2009.

[456] Erich Neuwirth. Rcolorbrewer: Colorbrewer palettes. *R Package*, 1(5), 2011.

[457] Anne Chao. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, pages 265–270, 1984.

[458] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. *Genome Biology*, 15(12):550, 2014.

[459] Sophie J Weiss, Zhenjiang Xu, Amnon Amir, Shyamal Peddada, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R Zaneveld, Yoshiki Vazquez-Baeza, and Amanda Birmingham. Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data. Report 2167-9843, PeerJ PrePrints, 2015.

[460] Stphane Guindon, Frdric Delsuc, Jean-Franois Dufayard, and Olivier Gascuel. Estimating maximum likelihood phylogenies with phyml. *Bioinformatics for DNA Sequence Analysis*, pages 113–137, 2009.

[461] Catherine Lozupone and Rob Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71(12):8228–8235, 2005.

[462] Juhi Bagaitkar, Donald R Demuth, and David A Scott. Tobacco use increases susceptibility to bacterial infection. *Tobacco Induced Diseases*, 4(1):1, 2008.

[463] John R Erb-Downward, Deborah L Thompson, Meilan K Han, Christine M Freeman, Lisa McCloskey, Lindsay A Schmidt, Vincent B Young, Galen B Toews, Jeffrey L Curtis, and Baskaran Sundaram. Analysis of the lung microbiome in the “healthy” smoker and in COPD. *PloS One*, 6(2):e16384, 2011.

[464] Anthony A Fodor, Erich R Klem, Deirdre F Gilpin, J Stuart Elborn, Richard C Boucher, Michael M Tunney, and Matthew C Wolfgang. The adult cystic fibrosis airway microbiota is stable over time and infection type, and highly resilient to antibiotic treatment of exacerbations. *PloS One*, 2012.

[465] Matilde Fernndez, Mario Porcel, Jess de la Torre, Mara Antonia Molina-Henares, Abdelali Daddaoua, Mara A Llamas, Amalia Roca, Victor Carriel, Ingrid Garzn, and Juan L Ramos. Analysis of the pathogenic potential of nosocomial *Pseudomonas putida* strains. *Frontiers in Microbiology*, 6, 2015.

[466] Hongjuan Li, Xuemei Li, Yanqing Duan, Ke-Qin Zhang, and Jinkui Yang. Biotransformation of nicotine by microorganism: the case of *Pseudomonas* spp. *Applied Microbiology and Biotechnology*, 86(1):11–17, 2010.

[467] KD Sun, CJ Zhu, WH Zhong, JM Chen, ZJ Ye, PJ Liu, and Qiang Zhou. Isolation and characterization of a high nicotine-degrading bacterium, *Pseudomonas* sp. strain zutskd. *Acta Sci Circum*, 28:1294–1301, 2008.

[468] Hailei Wei, Liping Lei, Zhenyuan Xia, Shuo Liu, Peigui Liu, and Xingzhong Liu. Characterisation of a novel aerobic nicotine-biodegrading strain of *Pseudomonas putida*. *Annals of Microbiology*, 58(1):41–45, 2008.

[469] Shu Ning Wang, Zhen Liu, Hong Zhi Tang, Jing Meng, and Ping Xu. Characterization of environmentally friendly nicotine degradation by *Pseudomonas putida* biotype a strain s16. *Microbiology*, 153(5):1556–1565, 2007.

[470] H Wan, HG Zhao, JZ Song, JB Rae, and JH Li. Screening, identification and degradation characteristics of high concentration nicotine degradation bacterium strain. *Tobaco ScienceTechnology*, 4:50–53, 2009.

[471] Marisa Almuzara, Marcela Radice, Natalia de Grate, Alejandra Kossman, Arabela Cuirolo, Gisela Santella, Angela Famiglietti, Gabriel Gutkind, and Varolos Vay. Vim-2-producing *Pseudomonas putida*, Buenos Aires. *Emerging Infectious Diseases*, 13(4):668, 2007.

[472] Gianluigi Lombardi, Francesco Luzzaro, Jean-Denis Docquier, Maria Letizia Riccio, Mariagrazia Perilli, Alessandra Col, Gianfranco Amicosante, Gian Maria Rossolini, and Antonio Toniolo. Nosocomial infections caused by multidrug-resistant isolates of *Pseudomonas putida* producing vim-1 metallo— lactamase. *Journal of Clinical Microbiology*, 40(11):4051–4055, 2002.

[473] Ching-Huei Yang, Tzuuguang Young, Ming-Yieh Peng, and Mieng-Chang Weng. Clinical spectrum of *Pseudomonas putida* infection. *Journal of the Formosan Medical Associationi*, 95(10):754–761, 1996.

[474] Lzaro Molina, Zulema Udaondo, Estrella Duque, Matilde Fernndez, Carlos Molina-Santiago, Amalia Roca, Mario Porcel, Jess de la Torre, Ana Segura, and Patrick Plesiat. Antibiotic resistance determinants in a *Pseudomonas putida* strain isolated from a hospital. *PloS One*, 9(1):e81604, 2014.

[475] Anton Y Peleg, Harald Seifert, and David L Paterson. *Acinetobacter baumannii*: emergence of a successful pathogen. *Clinical Microbiology Reviews*, 21(3):538–582, 2008.

[476] Lenie Dijkshoorn, Alexandr Nemec, and Harald Seifert. An increasing threat in hospitals: multidrug-resistant *Acinetobacter baumannii*. *Nature Reviews Microbiology*, 5(12):939–951, 2007.

[477] JM Cisneros and J Rodriguez-Bano. Nosocomial bacteremia due to *Acinetobacter baumannii*: epidemiology, clinical features and treatment. *Clinical Microbiology and Infection*, 8(11):687–693, 2002.

[478] Pierre Edouard Fournier, Herv Richet, and Robert A Weinstein. The epidemiology and control of *Acinetobacter baumannii* in health care facilities. *Clinical Infectious Diseases*, 42(5):692–699, 2006.

[479] Carina Maciel da Silva-Boghossian, Renata Martins do Souto, Ronir R Luiz, and Ana Paula Vieira Colombo. Association of red complex, *A. actinomycetemcomitans* and non-oral bacteria with periodontal diseases. *Archives of Oral Biology*, 56(9):899–906, 2011.

[480] Yi-Ju Chou, Shih-Yi Sheu, Der-Shyan Sheu, Jih-Terng Wang, and Wen-Ming Chen. Schlegelella aquatica sp. nov., a novel thermophilic bacterium isolated from a hot spring. *International Journal of Systematic and Evolutionary Microbiology*, 56(12):2793–2797, 2006.

[481] Timothy F Murphy, Aimee L Brauer, Karen Eschberger, Phyllis Lobbins, Lori Grove, Xueya Cai, and Sanjay Sethi. *Pseudomonas aeruginosa* in chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine*, 177(8):853–860, 2008.

[482] Anna G Drannik, Mahmoud A Pouladi, Clinton S Robbins, Susanna I Goncharova, Sussan Kianpour, and Martin R Stampfli. Impact of cigarette smoke on clearance and inflammation after *Pseudomonas aeruginosa* infection. *American Journal of Respiratory and Critical Care Medicine*, 170(11):1164–1171, 2004.

[483] Valerie Mattimore and John R Battista. Radioresistance of deinococcus radiodurans: functions necessary to survive ionizing radiation are also necessary to survive prolonged desiccation. *Journal of Bacteriology*, 178(3):633–637, 1996.

[484] Jimmy H Saw, Bruce W Mountain, Lu Feng, Marina V Omelchenko, Shaobin Hou, Jennifer A Saito, Matthew B Stott, Dan Li, Guang Zhao, and Junli Wu. Encapsulated in silica: genome, proteome and physiology of the thermophilic bacterium *Anoxybacillus flavithermus— WK1. Genome Biology*, 9(11):1, 2008.

[485] Ana Cristina Ferreira, M Fernanda Nobre, Fred A Rainey, Manuel T Silva, Robin Wait, Jutta Burghardt, Ana Paula Chung, and Milton S Da Costa.

Deinococcus geothermalis sp. nov. and deinococcus murrayi sp. nov., two extremely radiation-resistant and slightly thermophilic species from hot springs. *International Journal of Systematic and Evolutionary Microbiology*, 47(4):939–947, 1997.

[486] Bing Tian, Zhenjian Xu, Zongtao Sun, Jun Lin, and Yuejin Hua. Evaluation of the antioxidant effects of carotenoids from deinococcus radiodurans through targeted mutagenesis, chemiluminescence, and DNA damage analyses. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1770(6):902–911, 2007.

[487] Dalal Asker, Teruhiko Beppu, and Kenji Ueda. Unique diversity of carotenoid-producing bacteria isolated from Misasa, a radioactive site in Japan. *Applied Microbiology and Biotechnology*, 77(2):383–392, 2007.

[488] Nozomi Ihara, Jin Sakamoto, Munehiro Yoshida, and Tetsuaki Tsuchido. Killing effect of peppermint vapor against pink-slime forming microorganisms. *Biocontrol Science*, 20(2):91–97, 2015.

[489] Eugene Rosenberg. *The Family Deinococcaceae*, pages 613–615. Springer, 2014.

[490] Eun Mi Lee, Che Ok Jeon, Inpyo Choi, Kyu-Seob Chang, and Chang-Jin Kim. Silanimonas lenta gen. nov., sp. nov., a slightly thermophilic and alkaliphilic gammaproteobacterium isolated from a hot spring. *International journal of systematic and evolutionary microbiology*, 55(1):385–389, 2005.

[491] Jakob K Kristjnsson, Sigrdur Hjrleifsdttir, Vigg Th Marteinsson, and Gudni A Alfredsson. Thermus scotoductus, sp. nov., a pigment-producing thermophilic bacterium from hot tap water in iceland and including thermus sp. x-1. *Systematic and applied microbiology*, 17(1):44–50, 1994.

[492] TL Kieft, JK Fredrickson, TC Onstott, YA Gorby, HM Kostandarithes, TJ Bailey, DW Kennedy, SW Li, AE Plymale, and CM Spadoni. Dissimilatory reduction of fe (iii) and other electron acceptors by a thermus isolate. *Applied and Environmental Microbiology*, 65(3):1214–1221, 1999.

[493] Sigurlaug Skirnisdottir, Gudmundur O Hreggvidsson, Olle Holst, and Jakob K Kristjansson. Isolation and characterization of a mixotrophic sulfur-oxidizing *Thermus scotoductus*. *Extremophiles*, 5(1):45–51, 2001.

[494] Domenico Trombetta, Francesco Castelli, Maria Grazia Sarpietro, Vincenza Venuti, Mariateresa Cristani, Claudia Daniele, Antonella Saija, Gabriela Mazzanti, and Giuseppe Bisignano. Mechanisms of antibacterial action of three monoterpenes. *Antimicrobial Agents and Chemotherapy*, 49(6):2474–2478, 2005.

[495] Lennart Larsson, Bogumila Szponar, and Christina Pehrson. Tobacco smoking increases dramatically air concentrations of endotoxin. *Indoor Air*, 14(6):421–424, 2004.

[496] John L Pauly and Geraldine Paszkiewicz. Cigarette smoke, bacteria, mold, microbial toxins, and chronic lung inflammation. *Journal of Oncology*, 2011, 2011.

[497] Alan Rodgman and Thomas A Perfetti. *The chemical components of tobacco and tobacco smoke*. CRC press, 2013.

[498] Reinskje Talhout, Thomas Schulz, Ewa Florek, Jan Van Benthem, Piet Wester, and Antoon Opperhuizen. Hazardous compounds in tobacco smoke. *International Journal of Environmental Research and Public Health*, 8(2):613–628, 2011.

[499] Harold E Pattee. Production of aflatoxins by aspergillus flavus cultured on flue-cured tobacco. *Applied microbiology*, 18(5):952, 1969.

[500] Jeffrey D Hasday, Rebecca Bascom, Joseph J Costa, Thomas Fitzgerald, and Wendy Dubin. Bacterial endotoxin is an active component of cigarette smoke. *CHEST Journal*, 115(3):829–835, 1999.

[501] Michele Di Giacomo, Marianna Paolino, Daniele Silvestro, Giovanni Vigliotta, Francesco Imperi, Paolo Visca, Pietro Alifano, and Dino Parente. Microbial community structure and dynamics of dark fire-cured tobacco fermentation. *Applied and Environmental Microbiology*, 73(3):825–837, 2007.

[502] Jingwen Huang, Jinkui Yang, Yanqing Duan, Wen Gu, Xiaowei Gong, Wei Zhe, Can Su, and Ke-Qin Zhang. Bacterial diversities on unaged and aging flue-cured tobacco leaves estimated by 16s rRNA sequence analysis. *Applied Microbiology and Biotechnology*, 88(2):553–562, 2010.

[503] Can Su, Wen Gu, Wei Zhe, Ke-Qin Zhang, Yanqing Duan, and Jinkui Yang. Diversity and phylogeny of bacteria on zimbabwe tobacco leaves estimated by 16s rRNA sequence analysis. *Applied Microbiology and Biotechnology*, 92(5):1033–1044, 2011.

[504] Robert E Tyx, Stephen B Stanfill, Lisa M Keong, Angel J Rivera, Glen A Satten, and Clifford H Watson. Characterization of bacterial communities in selected smokeless tobacco products using 16s rDNA analysis. *PloS One*, 11(1), 2016.

[505] Kevin J Wendell and Sidney H Stein. Regulation of cytokine production in human gingival fibroblasts following treatment with nicotine and lipopolysaccharide. *Journal of Periodontology*, 72(8):1038–1044, 2001.

[506] JC Leffingwell. Ba basic chemical constituents of tobacco leaf and differences among tobacco types. *Tobacco: Production, Chemistry, And Technology*, 1999.

[507] Anna Wiernik, Alex Christakopoulos, Lennart Johansson, and I Wahlberg. Effect of air-curing on the chemical composition of tobacco. *Recent Advances in Tobacco Science*, 21:39–80, 1995.

[508] SE Atawodi and E Richter. Bacterial reduction of n-oxides of tobacco-specific nitrosamines (TSNA). *Human and Experimental Toxicology*, 15(4):329–334, 1996.

[509] Harold R Burton, Lowell P Bush, and Mirjana V Djordjevic. Influence of temperature and humidity on the accumulation of tobacco-specific nitrosamines in stored burley tobacco. *Journal of Agricultural and Food Chemistry*, 37(5):1372–1377, 1989.

[510] Audrey D Law, Colin Fisher, Anne Jack, and Luke A Moe. Tobacco, microbes, and carcinogens: Correlation between tobacco cure conditions, tobacco-specific nitrosamine content, and cured leaf microbial community. *Microbial Ecology*, pages 1–10, 2016.

[511] Harold R Burton, George H Childs Jr, Roger A Andersen, and Pierce D Fleming. Changes in chemical composition of burley tobacco during senescence and curing. 3. tobacco-specific nitrosamines. *Journal of Agricultural and Food Chemistry*, 37(2):426–430, 1989.

[512] Hongzhi Shi, Ruiyun Wang, Lowell P Bush, Jun Zhou, Huijuan Yang, Neil Fannin, and Ruoshi Bai. Changes in TSNA contents during tobacco storage and the effect of temperature and nitrate level on tsna formation. *Journal of Agricultural and Food Chemistry*, 61(47):11588–11594, 2013.

[513] Mirjana V Djordjevic, Jingrun Fan, Lowell P Bush, Klaus D Brunnemann, and Dietrich Hoffann. Effects of storage conditions on levels of tobacco-specific n-nitrosamines and n-nitrosamino acids in us moist snuff. *Journal of Agricultural and Food Chemistry*, 41(10):1790–1794, 1993.

[514] Linda L Kinkel. Microbial population dynamics on leaves. *Annual Review of Phytopathology*, 35(1):327–347, 1997.

[515] J Gregory Caporaso, Christian L Lauber, William A Walters, Donna Berg-Lyons, James Huntley, Noah Fierer, Sarah M Owens, Jason Betley, Louise Fraser, and Markus Bauer. Ultra-high-throughput microbial community analysis on the Illumina Hiseq and Miseq platforms. *The ISME Journal*, 6(8):1621–1624, 2012.

[516] Stephen B Stanfill, Gregory N Connolly, Liqin Zhang, Lily T Jia, Jack E Henningfield, Patricia Richter, Tameka S Lawler, Olalekan A Ayo-Yusuf, David L Ashley, and Clifford H Watson. Global surveillance of oral tobacco products: total nicotine, unionised nicotine and tobacco-specific n-nitrosamines. *Tobacco Control*, page tc. 2010.037465, 2010.

[517] Tameka S Lawler, Stephen B Stanfill, Liqin Zhang, David L Ashley, and Clifford H Watson. Chemical characterization of domestic oral tobacco products: total nicotine, pH, unprotonated nicotine and tobacco-specific n-nitrosamines. *Food and Chemical Toxicology*, 57:380–386, 2013.

[518] Mingqin Zhao, Baoxiang Wang, Fuxin Li, Liyou Qiu, Fangfang Li, Shumin Wang, and Jike Cui. Analysis of bacterial communities on aging flue-cured tobacco leaves by 16s rDNA pcr-dgge technology. *Applied Microbiology and Biotechnology*, 73(6):1435–1440, 2007.

[519] Siva Sabaratnam and Gwyn A Beattie. Differences between *Pseudomonas syringae* pv. syringae b728a and *Pantoea agglomerans* brt98 in epiphytic and endophytic colonization of leaves. *Applied and Environmental Microbiology*, 69(2):1220–1228, 2003.

[520] Todd F Hatchette, Rekha Gupta, and Thomas J Marrie. *Pseudomonas aeruginosa* community-acquired pneumonia in previously healthy adults: case report and review of the literature. *Clinical Infectious Diseases*, 31(6):1349–1356, 2000.

[521] Daniel M Musher. *Pseudomonas* pneumonia in smokers. *Clinical Infectious Diseases*, 33(3):415–415, 2001.

[522] Jacek Dutkiewicz, Barbara Mackiewicz, Marta Kinga Lemieszek, Marcin Golec, and Janusz Milanowski. *Pantoea agglomerans*: a mysterious bacterium of evil and good. part III. deleterious effects: infections of humans, animals and plants. *Annals of Agricultural and Environmental Medicine*, 23(2):197–205, 2016.

[523] Rong-Dih Lin, Po-Ren Hsueh, Jen-Chyi Chang, Lee-Jene Teng, Shan-Chwen Chang, Shen-Wu Ho, Wei-Chuan Hsieh, and Kwen-Tay Luh. *Flavimonas oryzihabitans* bacteremia: clinical features and microbiological characteristics of isolates. *Clinical infectious diseases*, 24(5):867–873, 1997.

[524] R Durmaz, MS Tekerekolu, T Kalciolu, and O Ozturan. Nasal carriage of methicillin-resistant *Staphylococcus aureus* among smokers and cigarette factory workers. *The New Microbiologica*, 24(2):143–147, 2001.

[525] CS Choi, CS Yin, Afra Abu Bakar, Zamberi Sakewi, Nyi Nyi Naing, Farida Jamal, and Norlijah Othman. Nasal carriage of *Staphylococcus aureus* among healthy adults. *Journal of Microbiology, Immunology, and Infection*, 39(6):458–464, 2006.

[526] Ritwij Kulkarni, Swati Antala, Alice Wang, Fbio E Amaral, Ryan Rampersaud, Samuel J LaRussa, Paul J Planet, and Adam J Ratner. Cigarette smoke increases *Staphylococcus aureus* biofilm formation via oxidative stress. *Infection and Immunity*, 80(11):3804–3811, 2012.

[527] Dorothy K Hatsukami, Irina Stepanov, Herb Severson, Joni A Jensen, Bruce R Lindgren, Kimberly Horn, Samir S Khariwala, Julia Martin, Steven G Carmella, and Sharon E Murphy. Evidence supporting product standards for carcinogens in smokeless tobacco products. *Cancer Prevention Research*, 8(1):20–26, 2015.

[528] Michael T Fisher, Cliff B Bennett, Alec Hayes, Yahya Kargalioglu, Brandy L Knox, Dongmei Xu, Raheema Muhammad-Kah, and Charles L Gaworski. Sources of and technical approaches for the abatement of tobacco specific nitrosamine formation in moist smokeless tobacco products. *Food and Chemical Toxicology*, 50(3):942–948, 2012.

[529] Fischer Sophia, Bertold Spiegelhalder, and Rudolf Preussmann. Preformed tobacco-specific nitrosamines in tobacco-role of nitrate and influence of tobacco type. *Carcinogenesis*, 10(8):1511–1517, 1989.

[530] Centers for Disease, Prevention, National Center for Chronic Disease Prevention, Health Promotion, Office on Smoking, and Health. *How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General.* US Government Printing Office, 2010.

[531] Yan S Ding, Liqin Zhang, Ram B Jain, Ntasha Jain, Richard Y Wang, David L Ashley, and Clifford H Watson. Levels of tobacco-specific nitrosamines and polycyclic aromatic hydrocarbons in mainstream smoke from different tobacco varieties. *Cancer Epidemiology Biomarkers and Prevention*, 17(12):3366–3371, 2008.

[532] Jianlin Xu and Willy Verstraete. Evaluation of nitric oxide production by lactobacilli. *Applied Microbiology and Biotechnology*, 56(3-4):504–507, 2001.

[533] Sergio Bizzarro, Bruno G Loos, Marja L Laine, Wim Crielaard, and Egija Zaura. Subgingival microbiome in smokers and non-smokers in periodontitis: an exploratory study using traditional targeted techniques and a next-generation sequencing. *Journal of Clinical Periodontology*, 40(5):483–492, 2013.

[534] Andrew M Thomas, Frederico O Gleber-Netto, Gustavo R Fernandes, Maria Amorim, Luisa F Barbosa, Ana LN Francisco, Arthur Guerra de Andrade, Joo C Setubal, Luiz P Kowalski, and Diana N Nunes. Alcohol and tobacco consumption affects bacterial richness in oral cavity mucosa biofilms. *BMC Microbiology*, 14(1):250, 2014.

[535] Jing Wu, Brandilyn A Peters, Christine Dominianni, Yilong Zhang, Zhiheng Pei, Liying Yang, Yingfei Ma, Mark P Purdue, Eric J Jacobs, and Susan M Gapstur. Cigarette smoking and the oral microbiome in a large study of american adults. *The ISME Journal*, 2016.