A Metagenomic Examination of Virus Diversity, Host Interactions, and Genetic Potential Across Municipal Waste Sites

by

Nikhil Anil George

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Biology

Waterloo, Ontario, Canada, 2023

© Nikhil Anil George, 2023

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Dr. Rachel Whitaker Professor, University of Illinois at Urbana-Champaign
Supervisor	Dr. Laura Hug Associate Professor, University of Waterloo
Internal Members	Dr. Kirsten Müller Professor, University of Waterloo
	Dr. Barb Butler Continuing Lecturer Emeritus, University of Waterloo
Internal-External Member	Dr. Brian Ingalls Professor, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Nikhil George was the primary author for the work described in this thesis. Chapters 1 and 5 of this thesis were not written for publication. Chapter 2 has been published as "CRISPR-resolved virus-host interactions in a municipal landfill include non-specific viruses, hyper-targeted viral populations, and interviral conflicts" (George, N.A. & Hug, L.A., Scientific Reports 2023, DOI: https://doi.org/10.1038/s41598-023-32078-6). Chapters 3 and 4 are being prepared as a single manuscript for publication. Exceptions to sole authorship denote work that includes authors in addition to N.A. George and L.A. Hug. An exception to sole authorship of material is described below.

Chapter 3

Viral genome binning, Auxiliary Metabolic Gene (AMG) curation, and megaphage validation were aided by Dr. Karthik Anantharaman and Dr. Zhichao Zhou at the University of Wisconsin-Madison. This work will be described in a manuscript in preparation for submission as "Diverse prokaryotic viruses across municipal waste sites encode novel metabolic potential and CRISPR-Cas systems (George, N.A., Zhou, Z., Anantharaman, K. & Hug, L.A.)"

Abstract

Viruses are the most abundant microbial guild on the planet, impacting microbial community structure and ecosystem services. Despite this, they have been underrepresented in reference databases compared to Bacteria and Archaea, which viruses are predicted to outnumber by at least an order of magnitude. The advent of metagenomics allowed the discovery of a multitude of viruses from a variety of environments. These metagenomic examinations illustrated that prokaryotic viruses are specifically understudied in engineered environments. In this thesis, I employ metagenomic and computational biology methods to examine the diversity, host interactions, and genetic systems of prokaryotic viruses that were predicted from 27 in-house metagenomes from samples taken at three municipal landfills across North America. I identified numerous viruses that are not represented in reference databases, including the third largest bacteriophage genome identified to date (~678 kbp), and note a cosmopolitan diversity of viruses in landfills that are distinct from viromes in other systems. Host-virus interactions were examined via host CRISPR spacer to viral protospacer mapping which captured hyper-targeted viral populations and 11 viruses predicted to infect across multiple phyla, suggesting that some viruses are far less host-specific than is typically expected. Viruses in my study encode Auxiliary Metabolic Genes (AMGs) with the potential to augment their hosts' methane, sulfur, and contaminant degradation metabolisms, including AMGs not previously reported in literature. CRISPR arrays and CRISPR-Cas systems were identified from predicted viral genomes, including the two largest bacteriophage genomes reported to contain these genetic features. Some virally encoded Cas effector proteins appear distinct relative to previously reported Cas systems, and are interesting targets for potential biotechnological tools. Virally encoded CRISPR arrays were predicted to target other viral elements, causing interviral conflicts; and CRISPR-encoding proviruses integrated into host chromosomes were latent examples of CRISPR-immunity-based superinfection exclusion. My observations indicate landfills, as heterogeneous contaminated sites with unique selective pressures, are key locations for diverse viruses and atypical virus-host dynamics, meriting further study of these ecosystems.

Acknowledgements

The work described in this thesis would not have been possible without the support of funding agencies, academic societies, collaborators, mentors, friends, and family. Throughout this research, I was primarily supported by two fellowships from the National Sciences and Engineering Research Council of Canada, an Ontario Graduate scholarship, the W.S. Rickert Graduate Student Fellowship in Science, and funding from my supervisor, Dr. Laura Hug.

I would like to express my gratitude to my supervisor, Dr. Laura Hug, for her support and trust throughout my degree. Laura allowed me to pioneer the first project on viruses in her group, giving me the opportunity to gain independence as a scientist. This instilled me with the confidence to take on new challenges, both in academia and beyond, knowing that I will figure them out. She has provided me numerous opportunities to attend scientific conferences, teach workshops, and collaborate with other researchers. These experiences have allowed me to expand my scientific network, hone my scientific communication skills, strengthen my CV, meet new friends, and have an awesome time seeing places in the world I may not have otherwise gotten a chance to. Lastly, I thank her for her kindness during the COVID-19 pandemic.

I thank all members of the Hug Lab, past and present, especially Angus Hilts for helping spark my passion for computational biology. I thank my committee members, Dr. Barb Butler and Dr. Kirsten Müller, for enduring many hours of me talking (i.e., a M.Sc. proposal defence, a Ph.D. transfer meeting, a Ph.D. proposal defence, a comprehensive examination, three committee meetings, and a pre-defence meeting). I also thank them for challenging me and for their constructive feedback throughout my degree.

Lastly, I thank my friends and family for celebrating my successes and for picking me up when things have not gone the way I planned. I would like to especially thank my mother, Neena, for her selflessness and flexibility, and my brother, Joseph, for his words of wisdom and for being my biggest supporter.

Table of Contents

Examining Committee Membership ii

Author's Declaration iii

Statement of Contributions iv

Abstract v

Acknowledgements vi

List of Figures ix

List of Tables x

List of Abbreviations xi

Chapter 1 Introduction 1

1.1 Background 1

1.1.1 Assessing viral diversity 1

1.1.2 Metagenomics for querying microbial and viral communities 3

1.1.3 Ecosystem functions of prokaryotic viruses: lysis and Auxiliary Metabolic Genes (AMGs) 4

1.1.4 Leveraging CRISPR-Cas biology to guide host identification of prokaryotic viruses 5

1.1.5 Virally encoded CRISPR-Cas systems 7

1.2 Study sites 8

1.3 Objectives 9

Chapter 2 Virus-host interactions in a municipal landfill include non-specific viruses, hyper-targeted viral populations, and interviral conflicts 11

2.1 Introduction 11

2.2 Methods 13

2.3 Results 19

2.4 Discussion 33
Chapter 3 Viral diversity, host interactions, and genetic potential across three municipal landfills 38
3.1 Introduction 38
3.2 Methods 40
3.3 Results 45
3.4 Discussion 58
Chapter 4 Virally encoded CRISPR-Cas systems 64
4.1 Introduction 64
4.2 Methods 67
4.3 Results 70
4.4 Discussion 81
Chapter 5 Research summary, study limitations, and field recommendations 87
5.1 Research summary 87
5.2 Study limitations and field recommendations 90
References 94
Appendix A Supplementary Material: Chapter 2 113
Appendix B Supplementary Material: Chapter 3 135
Appendix C Supplementary Material: Chapter 4 137

List of Figures

Figure 2.1: A: Virus-host interactions in a Southern Ontario landfill mapping 2016 hosts and 2016 viruses. 23
Figure 2.2: A: Genetic features targeted by host CRISPR arrays on a hyper-targeted viral element. . 29
Figure 2.3: Assessment of 187 spacer encoding viral scaffold. 32
Figure 3.1: A: Phylogenetic placement of newly discovered megaphages. 51
Figure 3.2: Gene-sharing network of landfill viruses with related groups from IMG/VR. 51
Figure 3.3: Gene-sharing network of landfill viruses. 52
Figure 3.4: AMG metabolism category proportions across datasets. 56
Figure 3.5: AMG counts per metabolic category across landfill datasets. 57
Figure 4.1: Classification of Class II CRISPR-Cas systems. 66
Figure 4.2: Gene-sharing network of CRISPR and non-CRISPR-encoding viruses. 72
Figure 4.3: CRISPR array and operons from a six-CRISPR-encoding viral element. 74
Figure 4.4: CRISPR array and operons from all CRISPR-encoding megaphages. 75
Figure 4.5: Type V CRISPR-associated nucleases, including virally encoded effectors from this study and reference nucleases. 78

List of Tables

Table 2.1: Metagenome statistics 22
Table 2.2: Virus-host networks summary. 30
Table 3.1: Summary statistics for metagenomes, MAGs, and viral scaffolds from the sampling sites. 46
Table 3.2: Genome statistics for megaphages identified in this study. 47
Table 3.3. Host CRISPR spacer to viral protospacer targeting within and across municipal waste sites. 54
Table 3.4: Virus-host networks summary. 54
Table 3.5: Shared CRISPR spacers within and between landfill datasets. 55
Table 3.6: Notable auxiliary metabolic genes encoded by viruses in the landfill datasets. 58

List of Abbreviations

aa	amino acid
AMG	auxiliary metabolic gene
BLAST	Basic Local Alignment Search Tool
BLASTn	BLAST-nucleotide
BLASTp	BLAST-protein
bp	base pairs
CA	California
Cas	(C)RISPR-(as)sociated
CD-HIT	Cluster Database at High Identity with Tolerance
CLC	composite leachate cistern
CONCOCT	(C)lustering c(ON)tigs with (CO)verage and (C)omposi(T)ion
CRISPR	clustered regularly interspaced short palindromic repeats
DAS Tool	Dereplication Aggregation and Scoring Tool
DGR	diversity generating retroelement
DNA	deoxyribonucleic acid
dsDNA	double-stranded DNA
dsRNA	double-stranded RNA
Gbp	gigabase pair
GTDB	Genome Taxonomy Database
GTDB-Tk	GTDB-Toolkit

GW	groundwater well
HMM	profile Hidden Markov Model
ICTV	International Committee on Taxonomy of Viruses
IMG/M	Integrated Microbial Genomes and Microbiomes
IMG/VR	Integrated Microbial Genomes and Viruses
JGI	Joint Genome Institute
kbp	kilobase pairs
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG orthology
LW	leachate well
MAG	metagenome assembled genome
Mbp	megabase pairs
MetaBAT	(Meta)genome (B)inning with (A)bundance and (T)etra-nucleotide frequencies
MGEs	mobile genetic elements
MSW	municipal solid waste
NCBI	National Centre for Biotechnology Information
NEUS	Northeastern United States
NGS	next-generation sequencing
ORF	open reading frame
PAM	protospacer adjacent motif
PCR	Polymerase Chain Reaction
PHASTER	(PHA)ge (S)earch (T)ool (E)nhanced (R)elease

psbA	photosystem-II reaction center (gene)
RefSeq	NCBI Reference Sequence Database
RNA	ribonucleic acid
SO	Southern Ontario
sp.	species
ssDNA	single-stranded DNA
ssRNA	single-stranded RNA
vConTACT	(v)iral (CONT)igs (A)utomatic (C)lustering and (T)axonomy
VIBRANT	(V)irus (I)dentification (B)y ite(R)ative (AN)no(T)ation
vMAG	viral metagenome assembled genome
VOG	Virus Orthologous Groups
VS	VirSorter

Chapter 1 Introduction

1.1 Background

1.1.1 Assessing viral diversity

Viruses are predicted to be the most abundant biological entities on the planet. This numerical dominance is driven by viruses of bacteria (bacteriophages) and archaea (archaeal viruses). Historically, viruses were classified mainly by morphology, with the order *Caudovirales*, tailed bacteriophages with double-stranded DNA (dsDNA) genomes, being the most highly represented in reference databases (Krupovic et al. 2011; Dion et al. 2020). Phage genomes can be linear or circular, and unlike cellular genomes, can extend beyond dsDNA encompassing single-stranded DNA (ssDNA), single-stranded RNA (ssRNA), and double-stranded RNA (dsRNA) genomes. These non-dsDNA genomes types are packaged in non-tailed viral particles (Dion et al. 2020). Isolated archaeal virus genomes are circular or linear and, unlike phage genomes, are exclusively composed of DNA (Krupovic et al. 2018); however, RNA viruses of archaea have been predicted from environmental sequencing of Yellowstone hot springs (Bolduc et al. 2012). Due to advances in genomics over the last twenty years, additional major lineages of tailed-dsDNA, non-tailed-dsDNA, and non-tailed-ssDNA bacteriophages have been discovered (Turner et al. 2021). These discoveries also illustrated that the three families within the *Caudovirales* are paraphyletic and do not accurately recapitulate shared evolutionary histories (Turner et al. 2023). Recent decisions by the bacterial viruses subcommittee of the International Committee for the Taxonomy of Viruses (ICTV) have put an end to morphology-based taxonomy, where taxonomic affiliations must now rely on core genes and phylogenetic placement (Turner et al. 2023). Most

notably, these changes involved the abolishment of the order *Caudovirales* and its paraphyletic families: *Myoviridae*, *Podoviridae*, and *Siphoviridae*.

As of May 1st, 2023, there are only ~13,000 viral genomes present in NCBI's RefSeq database. This count is dwarfed by that from the Bacteria, for which over 73,000 genomes are available in RefSeq. The comparatively poor representation of viruses within reference databases despite their numerical dominance (Gambelli et al. 2016) can be attributed to the fact that characterizing the virosphere has historically been a methods-limited endeavor. Hindering the discovery of new prokaryotic viruses was the challenge of culturing them in the laboratory. Most bacteria and archaea are unable to be cultured (Rappé and Giovannoni 2003), which in turn makes the cultivation of their viruses difficult given viral dependence on a host in order to propagate. To sidestep this challenge, culture independent methods, typically using specific primers to target hallmark viral genes or conserved genetic elements, have been adopted to probe viral diversity within an environment (Ansari et al. 1992; Abbaszadegan and Pepper 1993; Chen et al. 1996). This approach has its own limitations, e.g., certain primer sets are more compatible with some viral groups than others, influencing the viral diversity that is observed and characterized (Zhong and Jacquet 2014). There is also no single hallmark gene common to all viruses, unlike the 18S and 16S rRNA genes which are present in all eukaryotic and prokaryotic genomes, respectively (Koonin 2009). Thus, while culture-independent methods are the way forward for viral diversity studies, gene-targeted and primer-based approaches have shortcomings. Recent research has therefore employed metagenomics for the examination of viral communities, which has led the discovery of a multitude of viral lineages from a variety of environments (Paez-Espino et al. 2016; Suttle 2016; Dutilh et al. 2017; Roux et al. 2019; Al-Shayeb et al. 2020; Schulz et al. 2020).

1.1.2 Metagenomics for querying microbial and viral communities

Metagenomics, the sequencing of genetic material obtained from samples without prior amplification or target selection, allows microbes to be studied in much higher throughput, as this method circumvents the requirement for cultivation (Handelsman 2004). Short-read sequencing platforms generate metagenomic sequence information in the form of short fragments, called reads, which can be assembled based on sequence overlaps to generate longer contiguous elements known as contigs or scaffolds (Sangwan et al. 2016). Using paired-end sequencing, in which both ends of the target nucleic acids are sequenced, aids alignment and assembly in a variety of ways including in resolving repetitive regions (Sexton and Han 2019), detecting structural variation (Korbel et al. 2007), and bridging multiple genomic contigs together as genomic scaffolds (Kunin et al. 2008). All metagenomes examined within this thesis were paired-end sequenced using short-read sequencing technology. To reconstruct population-level genomes from the original sample that were fragmented during sequencing, genome binning is performed. Genome binning involves partitioning assembled scaffolds or contigs into bins based on parameters such as kmer frequencies and coverage (Sieber et al. 2018), to generate draft metagenome assembled genomes (MAGs) for populations of organisms (Dick et al. 2009). Viruses can be detected in both unassembled or assembled metagenomic data, usually by querying sequences for the presence of viral genes and genetic features (Roux et al. 2015; Paez-Espino et al. 2016b; Kieft et al. 2020), mapping sequences to viral reference databases (Tithi et al. 2018), and/or examining the frequency of kmers in a query sequence (Ren et al. 2017, 2020). Detected viruses can be binned into viral metagenome assembled genomes (vMAGs) (Kieft et al. 2022) and connected to their predicted hosts using several different computational approaches (Edwards et al. 2016; Roux et al. 2023). The limitations of metagenomic sequencing mainly stem from experimental protocols, sequencing depth, and from the computational workflows to

analyze metagenomics data (Wooley and Ye 2010). Nonetheless, metagenomic sequencing serves as a valuable tool to gain new insights into the microbial and viral world. The work described in this thesis leverages cultivation-independent metagenomics to elucidate the diversity, host interactions, and genetic potential of prokaryotic viruses in understudied engineered environments.

1.1.3 Ecosystem functions of prokaryotic viruses: lysis and Auxiliary Metabolic Genes (AMGs)

Viruses are major agents of mortality for bacteria and archaea. The activity of viruses can have significant impacts on the regulation of ecosystem services through their control of host mortality. The ecosystem roles of viruses have been examined in a variety of ecosystems (Roux and Brum 2019) and are mainly attributed to host lysis, lateral gene transfer, and manipulation of host metabolisms. In the case of host lysis, the best studied example is in marine systems where viral lysis has the potential to remove ~20% of prokaryotic biomass per day (Suttle 2007). This makes nutrients available for other organisms to utilize and regulates the flow of carbon within marine ecosystems, which in turn affects the concentration of carbon dioxide in the atmosphere (Suttle 2007).

While some bacteriophages and archaeal viruses strictly undergo lytic infection cycles, in which they replicate within the host cell and ultimately lyse the cell upon the exit of their virion progeny, others can undergo a lysogenic cycle in which their genome integrates within their host's genome for an indefinite period (Correa et al. 2021). If the normal integration or excision of the viral genome from the host genome is compromised in some way, fragments of the viral genome may be retained in the host genome and elements of the host genome may be excised and packaged into new virions (Chiang et al. 2019). For the host, this may impart new functionality from virally-derived genes, or may disrupt regular function/regulation of host genes

(Feiner et al. 2015). If the genetic elements lost by the virus in this process are essential to its ability to infect a new host cell, the virus will not propagate further. However, if the lost viral genetic elements are non-essential to the virus, upon infection of a new host cell, the virus can either pass the host-derived genetic elements to the new host via transduction (Chiang et al. 2019), or transiently augment the metabolism of the host cell when these new genetic elements are expressed upon infection. In the latter scenario, the virally encoded genetic elements that alter host metabolism upon infection are referred to as auxiliary metabolic genes (AMGs) (Breitbart et al. 2007). AMGs have been observed in a variety of systems, including hydrothermal vent plumes (Anantharaman et al. 2014) and organochlorine contaminated soils (Zheng et al. 2022), and are predicted to increase the metabolic efficiency of the host, which ultimately provides the virus with more resources and the means to a larger progeny pool (Luo et al. 2022). AMGs can have prominent ecosystem effects: studies showing that in marine surface waters ~60% of *psbA* (photosystem-II reaction center) genes (Sharon et al. 2007) and >50% of *psbA* transcripts (Sieradzki et al. 2019) were of viral origin indicate phage-encoded proteins likely play a key role in the regulation of ocean primary productivity.

1.1.4 Leveraging CRISPR-Cas biology to guide host identification of prokaryotic viruses

The co-evolutionary arms race between prokaryotes and their viruses has led to the accumulation of adaptations on both sides (Hampton et al. 2020). Among the most well-known prokaryotic host defenses against viral infection are Clustered Regularly Interspaced Short Palindromic Repeats-CRISPR associated (CRISPR-Cas) systems. CRISPR have been recognized as the most ubiquitous family of prokaryotic repetitive elements and are composed of multiple short repetitive elements located in arrays within bacterial genomes (Mojica et al. 2005; Pourcel et al. 2005a; Bolotin et al. 2005). These repetitive elements flank short, extra-chromosomally derived

variable regions of similar size that are referred to as spacers. CRISPR arrays, which herein refer to clusters of repeats and spacers, are often proximal to a suite of associated genes, which are termed *cas* genes (Jansen et al. 2002). *Cas* genes encode proteins which use information encoded in the CRISPR arrays to mediate host immunity against extrachromosomal elements (Mojica et al. 2005; Pourcel et al. 2005a; Bolotin et al. 2005).

In 2007, Barrangou and colleagues (Barrangou et al. 2007) showed that after a viral challenge, *Streptococcus thermophilus* strains integrated new spacers within their CRISPR arrays. These spacers, when altered or removed, resulted in loss of *S. thermophilus'* immunity against one of its viral pathogens, illustrating that the CRISPR-Cas system has a definitive role in host immunity against viruses (Barrangou et al. 2007). In brief, spacer elements within a CRISPR array are transcribed into guide RNAs, which direct Cas proteins to cleave DNA or RNA targets (Marraffini and Sontheimer 2008; Brouns et al. 2008; Makarova et al. 2020). The ability to program the target of Cas protein effectors by altering the sequence of their guide RNAs allows CRISPR-Cas to be used for precise genome editing and regulation (Cong et al. 2013; Van Der Oost and Patinios 2023). For the purposes of this thesis, the nature of CRISPR-Cas biology also creates identifiable links between bacterial or archaeal hosts and their respective viruses.

A very high degree of sequence similarity between a host CRISPR spacer and a viral genetic element is indicative of a history of infection between said host and virus (Andersson and Banfield 2008). New CRISPR spacers can be integrated into a CRISPR array in cases where a viral infection did not result in the death of the host (e.g., the virus was defective [Hynes et al. 2014], or slow replicating [Dimitriu et al. 2022]). In the spacer acquisition process, the viral genome is cleaved, often by restriction endonucleases (Dimitriu et al. 2020; Maguin et al. 2022),

and some of the resulting fragments are integrated within the host's CRISPR array. This updated CRISPR array can be vertically transmitted to daughter cells, providing heritable immunity against the specific local invading virus population. By comparing spacer sequences within a bacterial or archaeal host's CRISPR array to find matches on viral genomes (called protospacers), one can computationally determine the hosts of viruses within environmental datasets (Edwards et al. 2016). Numerous methods for computational host prediction of prokaryotic viruses exist, which include examining similarities between the abundance and oligonucleotide profiles of hosts and their viruses (Ahlgren et al. 2017; Galiez et al. 2017), and exact matching of virus elements in host genomes (Edwards et al. 2016). Matching host CRISPR spacers to viral protospacers results in comparatively few false positives, but does have limited recall (high false negative rates) (Edwards et al. 2016). Determining the hosts of prokaryotic viruses allows the ecological impacts of these viruses to be better understood.

1.1.5 Virally encoded CRISPR-Cas systems

The ability to engineer genomes is widely applicable in both academic and industrial settings. Before CRISPR-Cas systems' discovery, tools for genome engineering relied on protein-DNA interactions to modify target DNA, unlike the RNA-DNA interactions used in CRISPR-Cas technologies. Pre-CRISPR genome engineering tools were thus less specific and more tedious to design (Barrangou and Doudna 2016). CRISPR-Cas systems have been used, or are being considered, in the development of novel antimicrobials, biological controls of disease vectors, cancer therapeutics, and a wide range of other applications (Barrangou and Doudna 2016). Despite the widespread adoption of CRISPR-Cas-based methods in both academic and industrial labs, the technology still has limitations (Barrangou and Doudna 2016). These limitations include target element recognition by Cas proteins and off-target effects. Discovery

of new, divergent Cas proteins and CRISPR-Cas system assemblages is of interest to address these limitations, in hopes of broadening the current application space (Barrangou and Doudna 2016). The recent discoveries of some of the most streamlined CRISPR-Cas systems from prokaryotic viruses (Al-Shayeb et al. 2020, 2022; Pausch et al. 2020) established viruses as a potential hot spot for new CRISPR-Cas systems with biotechnological value. In this thesis, I expand the examination of virally encoded CRISPR-Cas systems to municipal landfill microbial communities.

1.2 Study sites

In one of the largest examinations of earth's virosphere to date, it was noted that samples from engineered sites have poorly characterized viral diversity (Paez-Espino et al. 2016b). My work aims to address this gap through an examination of viruses across municipal landfills. Landfills are polluted, highly engineered, heterogeneous systems that are of crucial environmental and economic importance (Köchling et al. 2015; Meyer-Dombard et al. 2020). Some of the contaminants within landfills may be transformed or metabolized by bacterial and archaeal members of the microbial communities within these environments (Song et al. 2015; Stamps et al. 2016). Prokaryotic viruses are likely important players in the landfill microbial community. Through lysis of host cells, bacteriophages and archaeal viruses may disrupt contaminant metabolism, potentially resulting in harmful reaction intermediates being released into the local environment. These viruses may also augment or modify the metabolism of their hosts which, along with host lysis, can lead to shifts in biogeochemical cycling and chemical transformations in landfills.

The research described in this thesis leverages samples taken from three municipal landfills distributed across North America. Landfill site anonymity has been preserved upon

request of site owners in all cases. The first study site is a municipal waste landfill and its adjacent aquifer in Southern Ontario, Canada (SO). This sanitary landfill was opened in 1972 and includes waste sorting and compacting on site, with the addition of daily and intermediate covers. The landfill is instrumented with over 100 leachate wells, four composite leachate cisterns, and a landfill gas collection system. The chemical composition of the leachate is monitored by quarterly sampling by contracted environmental engineers. The SO landfill was sampled twice in July 2016 and once in October 2017. The second municipal landfill is in the northeastern United States (NEUS). This site is also sanitary landfill that is equipped with leachate collection and gas collection systems. Municipal Solid Waste (MSW) is contained within six landfill cells, which are referred to as cells A, B, C, D, E, and F for the purpose of this work. The cells were sampled in February 2019, at which time the oldest cell, A, had received waste from 1980 to 1982 (39 years of waste decomposition time), whereas the youngest cell, F, began receiving waste in 2014 (5 years of waste decomposition time). Cells A, B, and C are completely closed and no longer receive waste, whereas cells D, E, and F are only partially closed. The third landfill is in southern California (CA). The CA landfill received both MSW and hazardous waste from mid 1960s to the mid 1990s and has been closed for over 20 years. The landfill is equipped with leachate and gas collection systems and remains under management for contaminant containment. The CA landfill was sampled in June of 2019.

1.3 Objectives

The objectives of the research presented in this thesis were:

1. Detect prokaryotic viruses in a municipal waste site and examine their host interactions, to determine whether the virus-host infection dynamics in this understudied environment are distinct from previously studied ecosystems (Chapter 2).

2. Characterize the diversity, host interactions, and genetic potential of prokaryotic viruses across three municipal waste sites, to clarify whether landfill viromes and their encoded functions are distinct relative to other ecosystems (Chapter 3).
3. Examine the diversity and novelty of virally encoded CRISPR-Cas systems identified from municipal waste viromes (Chapter 4).

Chapter 2[1] Virus-host interactions in a municipal landfill include non-specific viruses, hyper-targeted viral populations, and interviral conflicts

2.1 Introduction

Viruses are important and understudied members of microbial ecosystems that, with the advent of direct community shotgun sequencing, can now be studied at much higher throughput. Metagenomics-based methods have identified a multitude of new viruses from a variety of environments, but there is a recognized need for better representation from soil, plant-associated, and engineered environments (Paez-Espino et al. 2016). Through their host interactions, prokaryotic viruses can disrupt microbial ecosystem services by reducing or eliminating the hosts responsible (Rohwer et al. 2009; Roux and Brum 2019), augment host metabolism with Auxiliary Metabolic Genes (AMGs) (Lindell et al. 2004; Ignacio-Espinoza and Sullivan 2012; Hurwitz et al. 2013; Anantharaman et al. 2014; Kieft et al. 2021), stimulate gene flow within microbial populations through horizontal gene transfer (Penadés et al. 2015; Keen et al. 2017; Chen et al. 2018b), or alter host immunity through virally encoded CRISPR arrays (Seed et al. 2013; Faure et al. 2019) that can also drive interviral conflicts (Faure et al. 2019; Medvedeva et al. 2019). While prokaryotic viruses are expected to be host-specific, recent research has provided evidence for non-specific prokaryotic viral infection across taxonomic orders (Peters et al. 2015) and even across phyla (Malki et al. 2015; Paez-Espino et al. 2016; Dalcin Martins et al. 2018).

---

[1] A version of this chapter has been published as:
George, N.A. and Hug, L.A. CRISPR-resolved virus-host interactions in a municipal landfill include non-specific viruses, hyper-targeted viral populations, and interviral conflicts. *Scientific Reports* **13**, 5611 (2023). doi:10.1038/s41598-023-32078-6

Current computational methods for matching microbial viruses to their putative hosts vary in accuracy and sensitivity (Edwards et al. 2016). In comparison to other methods, matching Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) (Mojica et al. 2005; Pourcel et al. 2005; Bolotin et al. 2005) spacers to protospacers in viral elements yields conservative, relatively accurate predictions (Edwards et al. 2016). CRISPR spacer to viral protospacer matching has often been utilized in examining virus-host interactions (Dutilh et al. 2014; Edwards et al. 2016; Paez-Espino et al. 2016; Daly et al. 2016; Emerson et al. 2018; Trubl et al. 2018; Dalcin Martins et al. 2018; Bezuidt et al. 2020; Jarett et al. 2020) visualized as virus host networks (Daly et al. 2016; Trubl et al. 2018; Dalcin Martins et al. 2018; Bezuidt et al. 2020), however, there are few examples of temporal components in network architectures. In one notable example, researchers examined the dynamics between integrated Mobile Genetic Elements (iMGEs, which include viruses) in a wastewater treatment system over a year and a half, finding phage-host interactions were tied to community shifts, and observing host CRISPR array evolution in real-time (Martínez Arbas et al. 2021). The dynamics of virus-host interactions over time in other engineered systems remains an open question.

Landfills are heterogeneous, contaminated, engineered systems of growing environmental and economic importance (Köchling et al. 2015; Song et al. 2015; Stamps et al. 2016). Contaminants in landfills can be metabolized by bacteria and archaea (Song et al. 2015; Stamps et al. 2016), and, by extension, the viruses that prey on these microorganisms are likely important players in mediating contaminant cycling within landfills. From the perspective of viral diversity and host interactions, landfills remain largely unexamined. Previous work from our group has shown that bacterial and archaeal communities in these systems are diverse and spatially dynamic (Sauk and Hug 2022), but these analyses lacked temporal resolution and did

not include an examination of viral communities. Here, we identify the host-associated viral fraction from multiple sites across an active municipal waste site and model the virus-host interactions across two years to determine infection dynamics, host specificities, and interviral conflicts in this dynamic, understudied system.

2.2 Methods

*Sampling***.** This study makes use of 14 metagenomes, each of which were generated from one of 14 distinct samples collected from an active municipal landfill in Southern Ontario. The sample collection involved three trips, two of which were in July 2016 (six samples collected from five distinct sites) and the third in October 2017 (eight samples taken from eight distinct sites). Examinations of the bacterial and archaeal communities in our July 2016 samples have been previously reported, but no examinations of the virome were performed at that time (Sauk and Hug 2022). Of the samples obtained in July 2016, referred to as the 2016 samples, three were from leachate wells (LW1-3), two were from the same composite leachate cistern that was sampled on two occasions, six days apart (CLC_T1, CLC_T2), and one sample was obtained from a groundwater well adjacent to the landfill (GW1). The relevant October 2017 samples, referred to as the 2017 samples, included samples from the same five sites sampled in 2016, with the addition of samples from one leachate well (LW4), one storm water capture system (SWC1), and one groundwater well (GW3). About 5L of liquid material (leachate or groundwater) was sourced from each of the sampled sites. This material was filtered on-site through a 3 μm glass fiber filter (Pall Corporation, Ann Arbor, MI) and the resulting filtrate was subsequently filtered through a 27 mm diameter 0.1 μm polyethersulfone membrane (Pall Corporation, Ann Arbor, MI) via peristaltic pump filtration until the membrane clogged. Multiple 0.1 μm filters were taken for each site. All filters were stored on dry ice prior to long-term storage at -80ºC. All

samples were handled identically with the exception of CLC_T1, which was filtered through 0.2 μm and 0.1 μm filters sequentially instead of 3 μm and 0.1 μm. For the CLC_T1 sample, biomass on the 0.2 μm filter was used for metagenomic sequencing as there was insufficient biomass on the 0.1 μm filter.

DNA extractions. DNA was extracted from filters using the PowerMax DNA Isolation Kit (MO BIO Laboratories, Carlsbad, CA) using the default protocol outlined by the manufacturer, with the exception that the filter was cut into small pieces and added to the bead tube in place of soil. Extracted DNA was assessed for quality using the NanoDrop 1000 (Thermo Scientific, Waltham, MA) and quantity using a Qubit fluorometric method.

Metagenome sequencing, analysis, and host genome binning. Extracted DNA for all 2016 and six 2017 samples was sent to the Joint Genome Institute (JGI, Walnut Creek, CA) for paired-end 150-bp read metagenome sequencing using the HiSeq 2500 (Illumina, San Diego, CA). Read assembly and scaffold annotations, which included CRISPR array annotations employed in downstream analyses, were performed by the JGI in accordance with their most recently described standard operating pipeline (Huntemann et al. 2015). Three of the 2017 samples were sequenced by The Center for Applied Genomics (TCAG, Toronto, Canada). For these metagenomes, quality trimming and assembly followed the published JGI protocol, including bbduk (https://github.com/BioInfoTools/BBMap/blob/master/sh/bbduk.sh) and sickle (https://github.com/najoshi/sickle) for read quality trimming and spades3 (Bankevich et al. 2012) (https://cab.spbu.ru/software/spades/) under the -meta flag with kmers 33, 55, 77, 99, and 127 for assembly. Scaffold annotation for all metagenomes was performed by the JGI. Only scaffolds that were greater than or equal to 2.5 kbp were considered for subsequent analyses.

All read datasets were mapped individually to each assembled metagenome using Bowtie2 (Langmead and Salzberg 2012). Scaffolds from a single metagenome were binned using three binning algorithms: CONCOCT (Alneberg et al. 2014), MaxBin2 (Wu et al. 2016), and MetaBAT2 (Kang et al. 2019). Using DAS Tool (Sieber et al. 2018), the resulting bins were dereplicated and binned using an iterative consensus-based approach. Genome bins output from DAS Tool were supplied to CheckM (Parks et al. 2015) for quality assessment. Taxonomy was assigned to the genome bins using the Genome Taxonomy Database Toolkit (Chaumeil et al. 2019) application available on DOE-KBase (Arkin et al. 2018). Only genome bins that were >70% complete with less than 10% contamination (defined as medium quality draft Metagenome Assembled Genomes (MAGs)) or better were considered as putative host MAGs in further analyses.

Viral element identification and diversity analysis. Scaffolds not included in the putative host MAGs ("unbinned scaffolds") were supplied as input to the VirSorter (Roux et al. 2015) application on CyVerse (Merchant et al. 2016) with "Virome db" as the reference database and the virome decontamination function turned on. Only viral scaffold predictions from the highest confidence categories (1 and 2 for non-integrated viruses [omitting category 3], and 4 and 5 for provirus [omitting category 6]) were retained. Unbinned scaffolds were also supplied as input to VIBRANT (Kieft et al. 2020) using default parameters. Only scaffolds predicted as viral by both VirSorter and VIBRANT were considered for further analysis. We note that while this approach results in robust viral predictions, it potentially overlooks highly novel viruses. These predicted viral scaffolds were clustered through the use of CD-HIT-est (Li and Godzik 2006) using a global sequence identity threshold of c = 0.95, where c is the number of identical nucleotides over an alignment divided by the full length of the shorter sequence in the alignment. The

resulting clusters were considered as viral populations. Only unbinned scaffolds were used for viral detection to prevent CRISPR-array-encoding viral scaffolds that were misbinned within MAGs from having matches to themselves in the virus-host linking workflow described below. All representatives of viral populations were used as input to vConTACT2 (Bin Jang et al. 2019) in order to assess how similar viral communities are across samples and between our two timepoints. Gene-sharing networks created from vConTACT2 were visualized using Cytoscape ver. 3.8.0 (https://cytoscape.org/).

Virus-host linking. Viral elements were linked to their hosts through matching host CRISPR array spacers to protospacers found in the viral elements from the intersection of VirSorter and VIBRANT predictions. CRISPR spacers in genome bins were BLASTn (Altschul et al. 1990) searched against all viral elements, with hits retained only if they had up to one mismatch, a query coverage of $\geq 90\%$ and an E-value $\leq 10^{-4}$. The results of these BLASTn searches were visualized as protospacer-based prokaryotic virus-host networks using Cytoscape ver. 3.8.0. Recent analyses have shown that oligonucleotide and abundance profiles are less reliable for host detection relative to CRISPR- and provirus-based analyses (Edwards et al. 2016), thus we did not use co-binning of unintegrated viral scaffolds as valid links for putative virus-host linkages.

Community membership between networks. We first used dRep (Olm et al. 2017) to compare all 2016 MAGs to the MAGs reconstructed from 2017 datasets, regardless of whether MAGs were included in the networks. We then used CD-HIT-est (global sequence identity threshold of c = 0.95) to compare all 2016 viral elements to those predicted from 2017 datasets, again irrespective of whether any of these viral elements were present in the networks. To understand whether the hosts and viruses observed in the 2016 network were detected and included in the 2017 network, we used dRep and CD-HIT to dereplicate the network-associated hosts and viruses, respectively.

For any MAGs that were shared between the 2016 and 2017 networks, we examined changes in their detected CRISPR arrays using command-line tools and CRISPRCasFinder (Couvin et al. 2018).

Interviral conflicts. All high confidence viral scaffolds from VirSorter and VIBRANT, prior to clustering with CD-HIT-est, were cross-checked against the CRISPR array annotations provided by the JGI. The spacers from putatively virus-encoded CRISPR arrays were used as input into a BLASTn search (hits retained only if they had up to one mismatch, a query coverage of ≥ 90% and had an E-value ≤ 10^{-4}) against a database of all high confidence viral scaffolds agreed upon by VirSorter and VIBRANT. Matches between spacers from two distinct viral CRISPR arrays were not considered genuine interviral conflicts. To identify relevant proviruses, putative viral elements containing CRISPRs were BLASTn searched against all genome bins. In accordance with previous literature, putative provirus detection was considered with a bit score threshold of 50, an E-value threshold of 10^{-3}, and a minimum alignment length of 2,500 bp with at least 70% average nucleotide identity over the length of the alignment (Emerson et al. 2018). If 90% or more of a scaffold in a host genome bin aligned with a putative viral scaffold, this was considered to be indicative of a viral scaffold being co-binned with host scaffolds rather than being integrated in a host scaffold as a provirus (Emerson et al. 2018). All putative proviruses identified through the aforementioned approach were then confirmed using Prophage Hunter (Song et al. 2019) and PHASTER (Arndt et al. 2016) where if either tool detected a proviral element, said element was incorporated into downstream analyses. The CRISPR arrays and CRISPR-Cas systems of putative proviruses were further validated using CRISPRCasFinder. For CRISPR-encoding proviruses, we manually inspected whether the host targeted a specific viral

element via the CRISPR array provisioned by its provirus, or via a different CRISPR array within its genome.

In the case of the 187-spacer CRISPR array, we read-mapped reads from the original assembly to the scaffold to confirm connection of this array to viral marker genes. Two possible consensus sequences were identified at the 3' end of the CRISPR array, with the higher depth option connecting to the viral segment of the scaffold. The second, lower abundance sequence may represent an alternate current host or the original bacterial host for this array.

Hyper-targeted virus analyses. The CRISPR-spacer targets of all four viral elements in 2016 that were hyper-targeted by members of both the 2016 and 2017 hosts in our networks were manually examined in Geneious version 2021.1.1 (https://www.geneious.com/prime/) (Kearse et al. 2012). Detection of the Open Reading Frames (ORFs) in the viral elements and their initial annotations were performed during the JGI's annotation as described above. Predicted ORFs that were not annotated by the JGI's pipeline were manually annotated using hmmer3 (http://hmmer.org/) and Hidden Markov Models from the Virus Orthologous Groups database (https://vogdb.org/, release number vog206).

Examining the potential for viral infection across the domain and phylum levels. To further validate putative cross-phylum and cross-domain infecting viral elements, we performed a series of additional analyses: the quality of the MAGs predicted to target the implicated viral elements was considered, the length of the scaffolds encoding CRISPR arrays that targeted the implicated viral elements was examined, the CRISPR arrays from these scaffolds were compared between hosts targeting the same viral element, and the gene content and taxonomic affiliation of these scaffolds were assessed to aid chimera and misbin detection. The legitimacy of virus-host

interactions was assessed on a case-by-case basis with an emphasis on taxonomic congruence between MAGs and their CRISPR array-encoding scaffolds (Table A.1).

Computing platforms. The bulk of the computational analyses performed were run on an in-house server with an Ubuntu 18.04 operating system. CyVerse was used specifically to run VirSorter, DOEKBase was used specifically to run GTDB-tk and vConTACT2, and for three of our 2017 metagenomes that were sequenced outside the JGI, Compute Canada's Graham cluster was used for the spades3 assembly step in the metagenomics workflow.

2.3 Results

Identification of MAGs and viral elements

We sampled microbial biomass from a municipal landfill in Southern Ontario in July 2016 and October 2017. The sample sites in 2016 included three distinct leachate wells (LW1-3), a groundwater well adjacent to the landfill (GW1), and two temporal samples from the same leachate collection cistern, collected one week apart (CLC_T1 & CLC_T2). The 2017 sample series included samples from each of the five 2016 sites, with the addition of a fourth leachate well (LW4), an additional groundwater well (GW3), and a storm water catchment collection system (SWC1).

DNA extraction from sample biomass followed by metagenomic sequencing, assembly, and binning of the 2016 data yielded 688 MAGs that were >70% complete (70.13% to 100.00%, average completion 83.72%) and had <10% contamination (0.00% to 9.54%, average contamination 1.83%) (Table 2.1). Collectively, these bins are considered, at minimum, medium quality draft Metagenome Assembled Genomes (Bowers et al. 2017). The 2017 data yielded 1,064 MAGs that were >70% complete (70.09% to 100.00%, average completion 85.68%) and had <10% contamination (0.00% to 9.99%, average contamination = 2.88%). Viral detection

using VirSorter (Roux et al. 2015) and VIBRANT (Kieft et al. 2020) followed by dereplication with CD-HIT-est predicted 17,057 and 18,877 viral elements from the 2016 and 2017 data, respectively (Table 2.1). Examination of viral diversity across sample sites (Figure A.1) and years (Figure A.2) identified highly diverse viral communities with significant similarities across space and time.

Prokaryotic virus-host interactions

36% of our 2016 MAGs and 42% of our 2017 MAGs encode CRISPR arrays. CRISPR spacers were excised from CRISPR arrays and matched to viral protospacers allowing one mismatch. Virus-host interaction networks were derived for all host CRISPR spacer to virus protospacer combinations (2016 hosts – 2016 viruses; 2017 hosts – 2017 viruses; 2016 hosts – 2017 viruses; and 2017 hosts – 2016 viruses; Table 2.2, Tables A.2-A.5) to explore host-virus dynamics and adaptive immunity over time. We identified between 37-90 host MAGs implicated in a network, with up to 173 unique links between hosts and viral scaffolds (unique links = one MAG to one viral element, regardless of number of spacer connections; Table 2.2, Table A.6). Viral elements and predicted hosts were frequently from different sample sites (Figure 2.1, Figures A.3-A.6, Table A.7). Across all networks, 28-44% of host MAGs were connected to multiple viral scaffolds, targeting from two to nine predicted viral elements (see Tables A.2-A.5 for specific MAG-viral element connections). 1-10% of viral elements were predicted to infect two or more distinct hosts.

Temporal community continuity

As microbial communities in this system are poorly conserved across sites (Sauk and Hug 2022), we were interested in examining the temporal stability in the host-viral interactions as well as the microbial community as a whole. We used dRep (Olm et al. 2017) to cluster the reconstructed

MAGs into populations. In the 2016 dataset, 525/688 MAGs passed dRep's quality filtering. Of these, 232 (~44%) were represented in the 2017 MAGs dataset. For the 2017 dataset, the 232 shared MAGs comprise 28% of the 824/1064 MAGs that passed dRep filtering (see Figure A.7 for a visual summary of these overlaps). Viral elements were clustered with CD-Hit, which resulted in 3,198 viral elements represented in both datasets, comprising 18.8% and 16.9% of the viral elements annotated in the 2016 and 2017 metagenomes, respectively (Figure A.7).

We next investigated community composition between our temporally segregated networks. There were fewer conserved viral elements and MAGs within the host-virus networks compared to the overall datasets. For the viral elements, 19/59 viral elements (32.2%) from the 2016 network were represented in the entire set of 2017 virus annotations. Only 5/59 viral elements (8.5%) had near-identical viral elements in the 2017 virus-host network.

Table 2.1: Metagenome statistics

Sample name	Read Count	Scaffolds >=2.5kbp	MAGs[1]	% scaffolds in MAGs	# viral scaffolds[2]
CLC_T1 (2016)	199,293,412	92,890	71	18.25	3948
CLC_T2 (2016)	187,799,476	110,160	134	28.62	3368
GW1 (2016)	170,612,252	85,861	161	20.99	856
LW1 (2016)	177,231,300	93,614	72	12.6	7120
LW2 (2016)	200,095,936	108,918	169	31.85	2801
LW3 (2016)	199,991,448	90,376	81	14.28	5117
CLC (2017)	258,387,248	90,316	143	37.91	1532
GW1 (2017)	344,037,992	174,133	273	16.39	2919
GW3 (2017)	125,101,280	19,781	54	33.97	66
LW1 (2017)	104,844,754	59,502	54	14.36	3635
LW2 (2017)	270,424,804	90,072	160	34.45	2071
LW3 (2017)	318,700,912	184,907	149	13.9	12241
LW4 (2017)	258,469,368	89,271	171	21.36	2728
SWC1 (2017)	142,415,766	33,875	60	29.19	152

1 – Quality filtered bins with >70% completion, <10% contamination)
2 – viral scaffolds detected from unbinned scaffolds (prior to CD-HIT clustering)

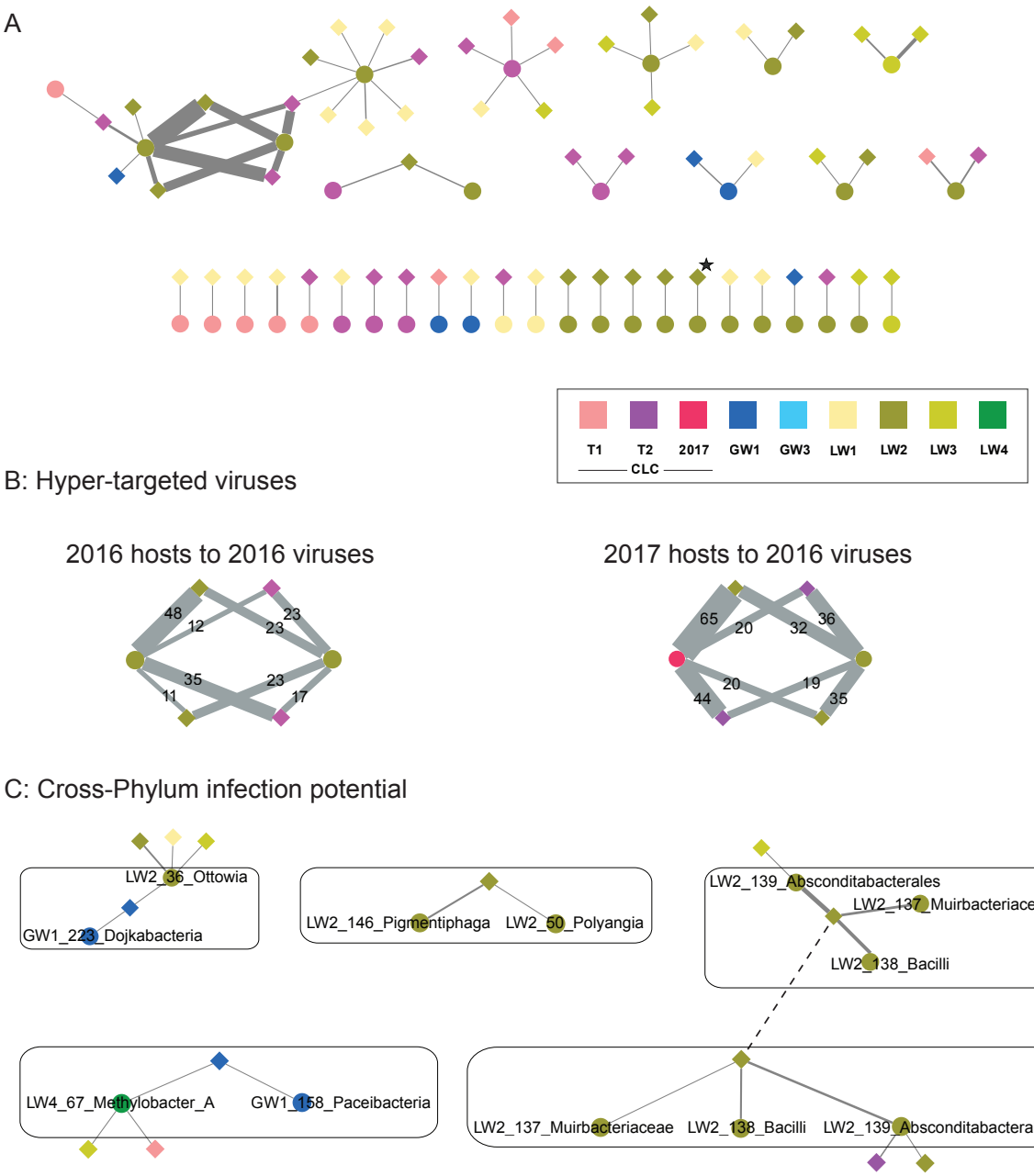


Figure 2.1: A: Virus-host interactions in a Southern Ontario landfill mapping 2016 hosts and 2016 viruses. CRISPR array spacers from the MAGs (circles) were mapped against the viral elements (diamonds) predicted by VirSorter and VIBRANT to establish virus-host linkages (connecting edges). The network is colored by the geographic sampling location from which the

MAG or viral element originated, with the width of the edges proportional to the number of spacer-protospacer matches supporting the connection. A viral element encoding a Diversity Generating Retroelement (DGR) (Roux et al. 2021) is denoted by a star. MAG identifier and lowest level of taxonomy as determined by GTDB-Tk (Chaumeil et al. 2019) are noted in Figure A.3. The other three networks are presented in full in Figures A.4-A.6. B: Hyper-targeting dynamics between 2016 (left) and 2017 (right) hosts against the 2016 viral elements, with the width of the edges proportional to the number of spacer-protospacer matches supporting the connections. C: Viral elements and hosts predicted to be involved in cross-phylum infections across all virus-host networks. The two viral elements connected by a dashed edge are predicted to be the same viral element in common between 2016 and 2017.

For the hosts, of the 36 2016-networked MAGs that passed dRep's quality filtering, 16 (~44%) were represented in the 2017 MAG dataset as a whole, and only three (~8%) were present in the 2017 network (Figure A.7).

Two virus-host pairs were shared between the 2016 and 2017 networks (i.e., viral elements clustered at 95% global sequence ID, hosts collapsed by dRep). Each of the two host MAGs in 2017 shared no spacers in their encoded CRISPR arrays with their 2016 counterparts, meaning each host targeted the same viral element with different CRISPR spacers in each year. In one of the host MAGs, the CRISPR arrays shared an exact match in a sequence flanking their respective CRISPR arrays (Figure A.8). Based on sequence similarity to the direct repeat element found in both CRISPR arrays, this flanking sequence may contain a CRISPR spacer and a partially degraded repeat, which is evidence for a shared CRISPR array origin within this population even in the current absence of shared spacer sequences. Note all subsequent work did not use dereplicated MAGs to preserve differences in CRISPR array structure.

Cross-phylum/cross-domain infection

In the networks, 14 viral elements were predicted to infect multiple hosts that are from different phyla (8) or domains (6) (Table A.1). Following curation based on length of CRISPR-array containing scaffolds, quality of MAGs, and taxonomic consistency between MAGs and CRISPR-array encoding scaffolds, we predict that four cross-phylum viral infections are legitimate, one of which was identified in two networks. No cross-domain interactions passed our quality screen (see Appendix A).

In the 2017 network, two viral elements were implicated in connections that passed subsequent quality control (Table A.1). One viral element (V2_2017) showed infection potential against hosts from three distinct phyla: the class Bacilli (LW2_138, phylum Firmicutes), the family Muirbacterieaceae (LW2_137, phylum Muirbacteria), and the order Absconditabacteriales (LW2_139, phylum Patescibacteria), all of which, including the viral element itself, were identified at the same sample site. These same three MAGs also shared a common viral element in the 2017-hosts-to-2016-viruses network (V154_2016). Based on 95% sequence identity clustering, the 2016 viral element is predicted to be the same as the 2017 viral element (V2_2017). The second 2017 cross-phylum interaction involves a member of the genus *Ottowia* (LW2_36, phylum Proteobacteria) and a member of the class Dojkabacteria (GW1_223, phylum Patescibacteria) which were both predicted to be infected by a single viral element (V102_2017).

In the 2017-host-to-2016-viruses network, we saw two additional instances of viral elements with potential to infect across phyla. One viral element (V36_2016) showed potential to infect hosts from the genus *Pigmentiphaga* (LW2_146, phylum Proteobacteria) and the class Polyangia (LW2_50, phylum Myxococcota). Finally, one viral element (V42_2016) was predicted to infect GW1_158_Paceibacteria (phylum Patescibacteria) and

LW4_67_Methylobacteria (phylum Proteobacteria). Interestingly, CRISPR arrays targeting this viral element share 56 identical spacers, suggesting a horizontal gene transfer event of this array between these two hosts.

We observed no instances of viral elements with potential to infect across phyla in the 2016-host-2017-viruses network and none of the observed interactions in the 2016 network passed quality control.

Diversity generating retroelements encoded by viruses

Diversity Generating Retroelements (DGRs) may increase a virus' breadth of infection (Roux et al. 2021). Leveraging the results of a previous meta-analysis in which a subset of our datasets were included (Roux et al. 2021), we report 92 DGR targets in our 2016 virus data and 30 from six of nine 2017 datasets. In total, eight viral elements that are involved in our networks were predicted to contain DGRs (star in Figure 2.1, Figures A.3-A.6), but none of these viral elements were predicted to infect across different phyla or domains.

Hyper-targeting of viral elements

Five viral elements were hyper-targeted by host MAG CRISPR arrays, where hyper-targeting was defined as having ≥ 20 protospacers targeted by a single host's set of CRISPR spacers. Hyper-targeted viral elements were not observed in the 2017 or 2016-host-to-2017-virus networks. One 2016 viral element was hyper-targeted by 2017 host MAGs, where the viral element had not been targeted by the 2016 hosts. The other four viral elements were hyper-targeted in both the 2016 and 2017-host-to-2016-viruses networks (V111_2016, V80_2016, V168_2016, and V65_2016), allowing examination of CRISPR arrays over time (Figure 2.1B). For these four viral elements, two host MAGs were responsible for the observed hyper-targeting

in each network, and the CRISPR arrays responsible for the hyper-targeting were highly conserved over their overlapping region between the two temporal samples. The MAGs involved in hyper-targeting had different taxonomies between the 2016 and 2017 hosts, with two members of the Acholeplasmataceae (Firmicutes) hyper-targeting the four viral elements in 2016, and two members of the Bacteroidales (Bacteroidota) hyper-targeting the same viral elements in the 2017-hosts-to-2016-viruses network (See Appendix A). It was not clear whether this is due to lateral gene transfer of CRISPR-array-encoding regions or mis-binning of CRISPR-encoding scaffolds, so we focused instead on array structure and nature of the hyper-targeting on the viral elements.

Each of the four hyper-targeted viral elements was targeted more times by the 2017 hosts (members of the Bacteriodales) relative to the 2016 hosts (members of the Acholeplasmataceae), though some of these connections were not reinforced by at least 20 host spacer to viral protospacer matches and thus were not considered hyper-targeting. The highest degree of targeting (against V111_2016) was 65 CRISPR spacer to viral protospacer connections, compared to 48 connections to the same viral element in the 2016 network. CRISPR spacers aligned to the hyper-targeted viral elements showed preferential targeting of specific regions within these elements (Figure 2.2A). The two regions most strongly targeted were tail protein coding sequences and hypothetical coding sequences (Figure 2.2A). In some instances, host targeting of viral elements appeared biased to one end of the viral elements (Figure 2.2A, Figure A.9).

We compared host CRISPR array-encoding scaffolds to assess the nature of these shared connections. For 3 of 4 host MAGs, connections to hyper-targeted viral elements stemmed from a single CRISPR array on the genomes, even in cases where multiple CRISPR arrays were

encoded. For one hyper-targeting CRISPR array, the shorter 2016-host-encoded array was a subset of the longer 2017-host-encoded array, without flanking regions to define a complete array and allow comparison of array adaptation over time. For the second pair of viral scaffolds, the 5' end and flanking region of the CRISPR arrays was captured on both scaffolds. There were no additional spacers incorporated into the CRISPR array over the ~year between sampling events, which may suggest that the array is inactive given CRISPR arrays can change over short time scales (Tyson and Banfield 2008). We note that three of the four MAGs involved in hyper-targeting are among the top ten most abundant network-associated MAGs across all metagenomes described from the SO landfill, with one (LW2_114_2017, order Bacteroidales) as the most abundant MAG in the dataset (average coverage of 877).

We predicted *Chi* sites in all viral elements that were targeted by hosts to assess if hyper-targeting was correlated with a depletion in *Chi* sites, which is associated with increased spacer recruitment (Levy et al. 2015). All hyper-targeted viral elements were among those with the lowest proportional counts of *Chi* sites (mean = 0.48 and median = 0.12 per 1000 bp for all targeted viral elements) with counts for hyper-targeted viral elements being 0.16, 0.12, 0.18, and 0 per 1000 bp [mean: 0.12, median: 0.14] Figure 2.2B).

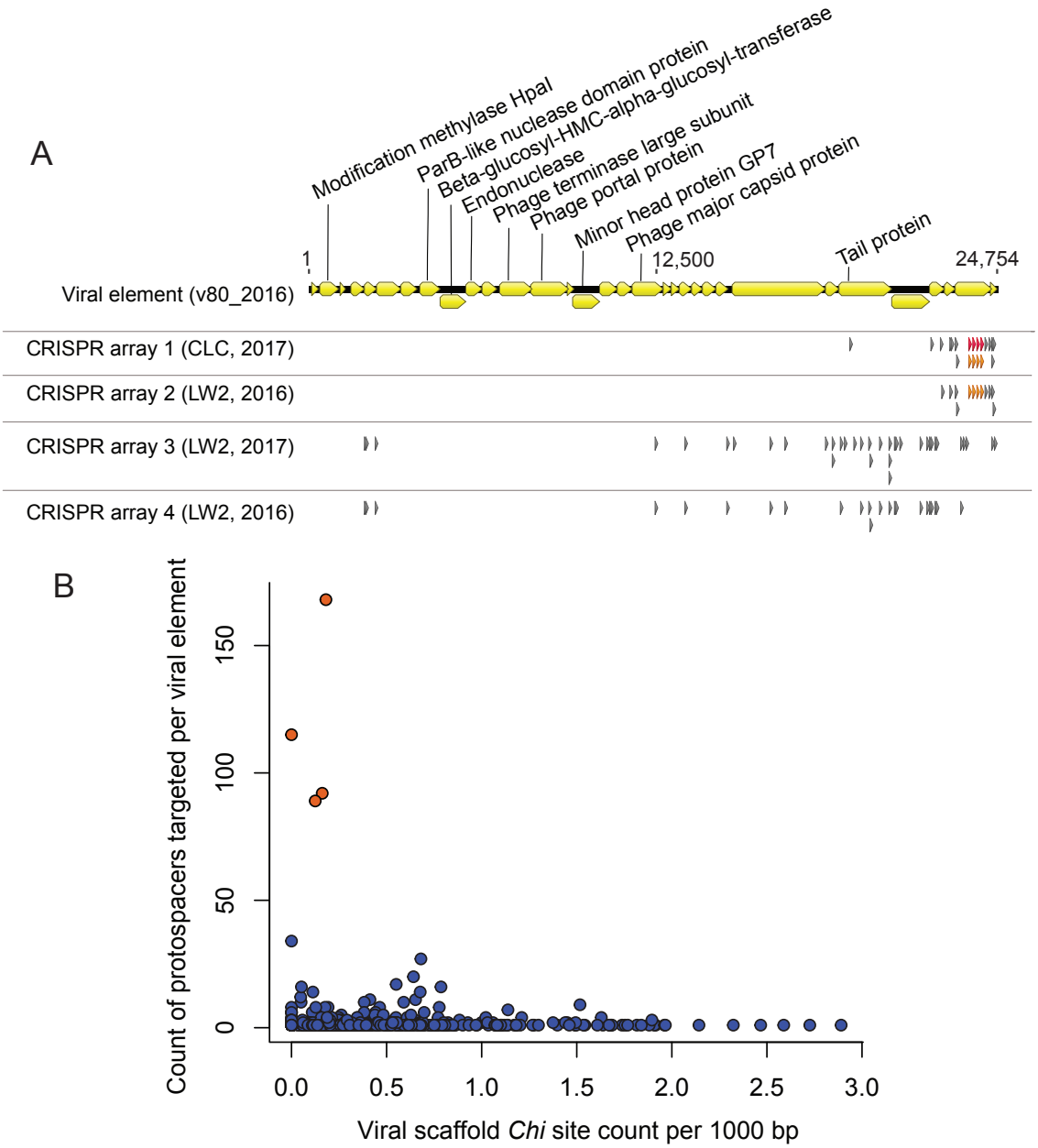


Figure 2.2: A: Genetic features targeted by host CRISPR arrays on a hyper-targeted viral element. The coding sequences' annotations for one hyper-targeted viral element (V80_2016) with corresponding CRISPR spacer mappings are displayed. In the targeting profile of a single host scaffold, a non-grey triangle represents a single spacer that had multiple matches to regions of the viral element. When shared, these single spacers are colored the same between distinct host targeting profiles. CRISPR spacer mapping and visualization was conducted using Geneious (Kearse et al. 2012). Host scaffold accessions are as follows: CRISPR array 1:

Ga0265294_10000314; CRISPR array 2: Ga0172382_10051677; CRISPR array 3: Ga0265293_10000924; CRISPR array 4: Ga0172382_10020296 & Ga0172382_10071763 (two arrays within one host MAG, only one is responsible for hyper-targeting). B: *Chi* site counts in viral elements (in *Chi* site per 1000 bp) compared to number of times a viral scaffold was targeted by a host across all virus-host networks. Hyper-targeted viral elements are highlighted in orange.

Table 2.2: Virus-host networks summary. Virus-host interactions are based on host MAG CRISPR spacers and viral element protospacer connections. Unique links = 1 MAG to 1 viral element, regardless of number of spacer connections.

Network	Host MAGs	Viral scaffolds	Unique links
2016	37	59	66
2017	90	161	173
2016-host-to-2017-viruses	39	79	80
2017-host-to-2016-viruses	87	125	139

Virally encoded CRISPR arrays convolute virus-host networks

Viruses have been shown to encode CRISPR arrays as well as Cas proteins, with hypotheses that include these systems acting as defense mechanisms preventing other viruses from infecting the same host (Faure et al. 2019; Medvedeva et al. 2019). We explored the potential for interviral conflicts using the full set of viral predictions prior to clustering with CD-HIT in order to preserve CRISPR array variation between closely related viral elements. In our 2016 data, 70 (0.41%) viral elements were predicted to encoded CRISPR arrays, with no shared CRISPR spacers between viral elements. These arrays contained between 2-187 spacers (average spacer

count per array ~7, 6 arrays contained >10 spacers). The current longest CRISPR array is 587 spacers long, from *Haliangium ochraceum* strain DSM 14365. To our knowledge, the 187-spacer viral CRISPR array is the longest viral CRISPR element that has been predicted to date. This element also contains a *cas2* gene, which is further evidence of functionality, but is uncommon, as most phage CRISPR systems do not encode spacer acquisition machinery (Al-Shayeb et al. 2020, 2022) (Figure 2.3). VIBRANT predicted that this genetic element was a non-integrated virus and Prophage Hunter did not identify any proviruses within the element. Based on read mapping, the scaffold containing the 187-spacer CRISPR array showed high heterogeneity in a region between the CRISPR array and viral marker genes (Figure 2.3). Curation of reads mapping to this region confirmed appropriate depth of coverage for the consensus sequence linking the array to the viral genes (Figure 2.3). In our 2017 data, 91 (0.48%) viral elements were predicted to encode CRISPR arrays. There were 8 instances of shared CRISPR spacers between CRISPR-encoding viral elements. The 2017 virally encoded CRISPR arrays contained between 2-21 spacers (average spacer count per array ~4, 4 arrays contained >10 spacers).

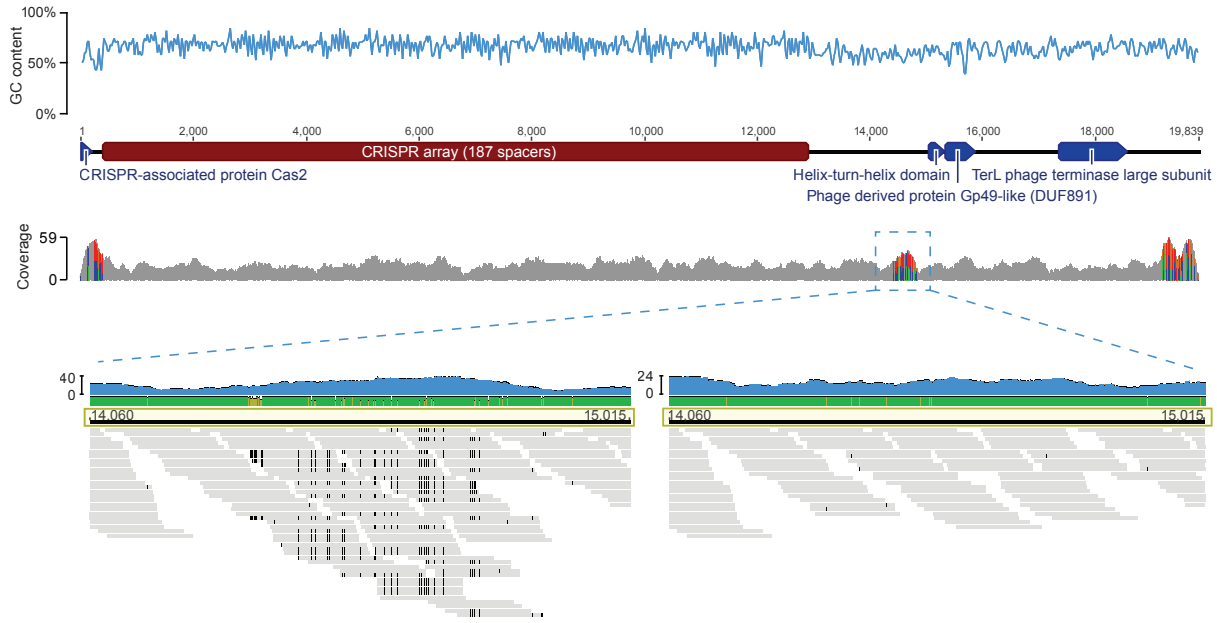


Figure 2.3: Assessment of 187 spacer encoding viral scaffold. From top to bottom: GC content along the length of the scaffold (total length 19,839 bp); annotation map of relevant genes; depth of read coverage and heterogeneity. The variable region between 14K-15K is marked with a blue box. Full read mapping on the left demonstrates two potential sequences for this region. On the right, the higher abundance path is shown. This image was made using Geneious (Kearse et al. 2012) and IGV (Robinson et al. 2011).

Spacers encoded in 2016 virus CRISPR arrays were predicted to target 32 viral elements in the unclustered 2016 virus dataset; whereas spacers encoded in 2017 virus CRISPR arrays were predicted to target 44 viral elements in the unclustered 2017 dataset. These may represent latent interviral conflicts which would need to be co-located in a cellular host for conflicts to be active.

Within our host-virus networks, viral elements encoding CRISPR arrays were predicted to be proviruses in 3 hosts (See Appendix A for details). In all cases, connections in the networks between provirus-encoded CRISPR arrays and non-integrated viral elements were based on

shared CRISPR spacers in CRISPR arrays on both viruses rather than a spacer-protospacer match. When CRISPR spacers are integrated into an array, the protospacer-adjacent motifs (PAMs) are not incorporated. As a result, CRISPR spacers are not active targets for Cas nucleases (Collias and Beisel 2021). In one case, both the host-encoded and provirus-encoded CRISPR arrays target the same non-integrated viral elements. In another case, the viral element contains two distinct CRISPR-Cas systems, both of which are shared in the provirus and non-integrated virus versions of this element.

2.4 Discussion

Using a metagenomics-based approach, we were able to resolve prokaryotic virus-host interactions and infection networks from samples collected from a municipal landfill and adjacent aquifers approximately one year apart. Cellular populations were more stable than viral elements, with ~44% and ~19% conserved between the two years, respectively. Infection networks were more dynamic, with lower continuity of MAGs and viral elements between the 2016 and 2017 networks compared to the baseline communities.

It was common for a viral element and its predicted host to be from different sample sites. Planktonic viruses were expected to pass through the 0.1 μm filters used to collect biomass, meaning most of the prokaryotic viruses detected were captured from active infections and indicate presence of the host even if it was not detected as a high-quality MAG from that site's metagenome. Hosts in our networks were connected to 1-9 viral elements, potentially representing a large infection burden for some populations. Alternatively, multiple viral elements infecting the same host may be genomic fragments of the same viral genome. In our temporal comparison, the 2017-host-to-2016-virus network was not larger than the 2017-only network,

despite a hypothesis that hosts from 2017 would have established immunity to 2016 viruses via earlier infections.

We predicted four instances where viral elements infect members of multiple phyla. Cross-phyla infection potential for viruses has been predicted *in silico* before (Paez-Espino et al. 2016; Dalcin Martins et al. 2018), and also confirmed in a single study for four bacteriophages isolated from Lake Michigan (Malki et al. 2015). There are numerous adaptations a viral element would require in order to successfully infect multiple hosts, none of which were examined here. These adaptations would be required for crucial steps of the infection process such as host target receptor recognition and evasion of host immunity, the former of which would likely lower infection efficiency on the original host (Dunne et al. 2019). While we have no evidence against four putative cross-phylum interactions, laboratory confirmation is a necessary requirement to substantiate these predictions.

Diversity generating retroelements (DGRs) are genetic elements that enhance mutation and evolution in microbes, including viruses (Roux et al. 2021). DGRs were shown to be enriched in landfills (Roux et al. 2021), possibly owing to landfills' extreme heterogeneity and concentration of solvents and other stressors increasing the evolutionary pressures on microbes, thus selecting for DGRs to enhance microbial adaptation. DGRs can also facilitate diversification of virus host-attachment proteins, where viruses that encode DGR targets within their host attachment protein encoding genes show a broader host range (Roux et al. 2021). In our networks, none of our putative broadly-infecting viral elements encode any detectable DGRs, and viral elements with DGRs were associated with only one host or, at most, two hosts from the same family. DGRs thus do not appear to be strong factors in influencing the instances of broad viral host range in the networks from this study.

A striking feature of our networks were the hyper-targeted viral elements, with up to 65 unique connections between a host's CRISPR spacers and single viral elements. Hypothetical protein coding regions on the viral elements were highly targeted, suggesting that the proteins encoded in these regions are important for viral propagation. A bias of targeting at one end of a viral element was observed in some cases which may be due to some CRISPR-Cas systems preferentially recruiting spacers from the regions of viral DNA first injected into the host cell, as this provides better immunity than spacers from late-injected regions (Modell et al. 2017). The hyper-targeting of viral elements may have evolved due to an on-going arms-race between host and virus which builds a history of immunity over time; either from a strong priming response in the CRISPR-Cas system, which results in recruitment of multiple spacers once activated (Richter et al. 2014; Künne et al. 2016); or because the viral element is deficient in *Chi* sites, which are molecular benchmarks for preventing DNA cleavage by the repair complex (Levy et al. 2015). Hyper-targeted viral elements in the 2017-hosts-to-2016-viruses network showed higher levels of hyper-targeting than in the 2016 network. The hyper-targeting of the same four viral elements by hosts in 2016 and 2017 suggests a long history of infection dynamics. The hyper-targeted viral elements had a relatively low proportion of *Chi* sites when compared to viral elements with weaker targeting (Figure 2.2B). The viral elements with the highest proportion of *Chi* site counts were not strongly targeted in our networks.

CRISPR arrays encoded on viral elements were found in a small fraction of predicted viral elements (0.45%). Virus-encoded CRISPR arrays and CRISPR-Cas systems have been identified before (and are involved in crippling host viral defense systems (Seed et al. 2013), regulating host transcription and translation (Al-Shayeb et al. 2020), CRISPR-Cas system inhibition (Faure et al. 2019) and interviral conflicts (Faure et al. 2019; Medvedeva et al. 2019)).

In the case of interviral conflicts, a CRISPR-encoding virus, integrated as a provirus, can provide its host with immunity against other viruses, akin to a superinfection exclusion system (Faure et al. 2019). Our datasets contained a putative viral element (V170_2016) that encodes a CRISPR array with 187 spacers and a *cas2* gene, the longest virally encoded CRISPR array to date. Although this viral element was not targeted by any bacterial or archaeal hosts in our networks, CRISPR spacers within its array did target two other viral elements in the 2016 data, neither of which were predicted to encode CRISPR arrays themselves. Due to *cas2* gene being the last gene on one end of the scaffold, along with limitations on the accuracy of prophage boundary detection tools, it is possible that this scaffold contains a bacterial CRISPR array adjacent to an integrated prophage.

In our networks, three proviruses containing CRISPR arrays targeted their own planktonic population of viruses. These connections were due to CRISPR spacers that are common to the arrays of both the proviruses and their non-integrated virus equivalents. Due to the requirement of a Protospacer-adjacent motif (PAM) for virtually all Cas nucleases (Collias and Beisel 2021), spacer-to-spacer matches wouldn't result in immunity unless the hypothetical nuclease involved in target cleavage was PAM agnostic or PAM permissive. Two connections in our networks were exclusively due to putative provirus CRISPR-arrays matching to their non-integrated virus equivalent's CRISPR arrays, however, the presence of the provirus itself indicates that the non-integrated virus equivalent is able to infect the host.

Within this engineered system, virus-host interactions are substantially more complicated than a 1-virus-1-host model. Cross-phylum infections, hyper-targeting, interviral conflicts, and provirus-encoded CRISPR arrays were observed in a single system, suggesting these interactions may be more common than currently understood. We hope that in future, these interactions are

curated on a case-by-case basis as we have done here, in order to expand our understanding of virus-host interactions in natural environments. We note that this wealth of insight derived from a single landfill, and suggest this category of highly heterogenous engineered environments merits further study of microbial communities and virus-host interactions.

Chapter 3 Viral diversity, host interactions, and genetic potential across three municipal landfills

3.1 Introduction

In one of the largest examinations of earth's virosphere to date, it was noted that samples from engineered sites have poorly characterized viral diversity (Paez-Espino et al. 2016b). Two recent studies of wastewater treatment plants confirmed this, as >98% of detected viral contigs were unmatched to a representative subset of the IMG/VR database in one (Chen et al. 2021) and >99% were unmatched to the NCBI virus database in the other (Martínez Arbas et al. 2021). Landfills, similar to wastewater treatment plants, are highly engineered, polluted and environmentally significant systems (Köchling et al. 2015; Song et al. 2015; Stamps et al. 2016). Viruses, particularly those that infect bacteria and archaea, have previously been predicted to have diverse infection dynamics with their hosts in these systems, including rarely seen interactions (George and Hug 2023, presented in Chapter 2).

Bacteriophages are predicted to be quite host specific. However, from recent research on both isolated phages (Malki et al. 2015) and those identified *in silico*, bacteriophages have been shown or predicted to infect across distinct bacterial phyla (Paez-Espino et al. 2016b; Dalcin Martins et al. 2018; George and Hug 2023; Hwang et al. 2023) and domains of life (Chen et al. 2021; Hwang et al. 2023). Various computational approaches to host prediction of prokaryotic viruses exist (Edwards et al. 2016; Roux et al. 2023), with CRISPR spacer to viral protospacer matches having relatively low sensitivity, but high specificity (Edwards et al. 2016). CRISPR spacers can undergo rapid turnover in arrays both in *in vitro* experiments (Paez-Espino et al. 2013) and in natural communities (Andersson and Banfield 2008; Tyson and Banfield 2008; Pride et al. 2011). Nonetheless, some spacers are selectively retained, which are predicted to

target regions of viral genomes that are less amenable to mutational escape (Paez-Espino et al. 2013; Nasko et al. 2019). Examinations of CRISPR array variation and spacer conservation have been performed across a variety of sample types (Garrett 2021), including in wastewater (Martínez Arbas et al. 2021), but these examinations have not been performed with samples from municipal waste sites. The open questions remain of whether the notable virus-host interactions observed in a single municipal waste site, such as hyper-targeted viral elements, interviral conflicts, and putative cross-phylum-infecting phages (George and Hug 2023, presented in Chapter 2), are consistent across other municipal waste sites and whether these interactions are underlain by conserved CRISPR array architectures.

Prokaryotic viral genomes are variable in size, but are canonically considered short, efficient genomic elements constrained by their requirement to package into capsids for infection. In recent years genomic approaches have increased the number of identified phage genomes ≥200 kbp in size, with phages having ≥200 kbp genomes being referred to as jumbophages and phages having ≥500 kbp genomes referred to as megaphages (Paez-Espino et al. 2016b; Yuan and Gao 2017; Devoto et al. 2019). The current record length for a circularized megaphage genome is 735 kbp (Al-Shayeb et al. 2020). These less streamlined phage genomes have more coding space, and have been shown to carry genes encoding tRNAs and associated enzymes, other translation machinery, and CRISPR-Cas systems (Al-Shayeb et al. 2020).

During infections, viruses may augment or modify the metabolism of their hosts through Auxiliary Metabolic Genes (AMGs; Lindell et al. 2004; Ignacio-Espinoza and Sullivan 2012; Hurwitz et al. 2013; Anantharaman et al. 2014; Kieft et al. 2021). The functions of these AMGs can include contaminant modification/degradation. Recent research provides evidence of viruses encoding arsenic resistance and transporter genes in arsenic-contaminated paddy soils (Gao et al.

2023), and has shown viruses aiding their hosts' survival in organochlorine contaminated soils with AMGs involved in organochlorine degradation (Zheng et al. 2022; Yuan et al. 2023). This phenomenon of contaminant modification/degradation genes being provisioned by viruses to their hosts has yet to be examined in municipal waste sites, which are highly and heterogeneously contaminated systems.

3.2 Methods

Sampling

This study makes use of 27 in-house metagenomes that were sequenced from samples from three distinct landfills. 14 metagenomes were sequenced from an active municipal landfill and adjacent aquifer in Southern Ontario (SO), sampled twice one year apart (July 2016 [SO 2016] and October 2017 [SO 2017]) and described previously (Sauk and Hug 2022; George and Hug 2023). Nine metagenomes were from a northeastern United States active municipal landfill (NEUS) in which nine distinct cells were sampled in February 2019 as described previously (Grégoire et al. 2023). At the time of sampling, the oldest cell had received waste from 1980 to 1982 (39 years of waste decomposition time), whereas the youngest cell began receiving waste in 2014 (5 years of waste decomposition time). The final four metagenomes came from a closed landfill from Southern California (CA) that was sampled in June of 2019. The CA landfill was operational from the mid 1960s to the mid 1990s and received both municipal and hazardous waste. The landfill has been closed for over 20 years but remains under management for contaminant containment. Leachate from one leachate well and the influent leachate to the leachate treatment plant were collected in sterile 500 mL containers. An additional sample of the treatment plant biofilter sludge was taken, with biomass split between planktonic and solid materials. All samples were passed through 0.2 μm Sterivex filters (Sigma Aldrich) until the

filters clogged, with a minimum of three filters taken per sample (~50-500 mL of volume processed per site).

The viral fraction of the SO landfill metagenomes was described from the perspective of viral-host interactions in Chapter 2 (George and Hug 2023). None of the other metagenomes have been analyzed for their viral fraction, and all cross-site comparisons as well as AMGs and megaphage analyses are novel to this chapter.

DNA extractions, sequencing, and host genome binning

DNA extractions for the SO landfill and adjacent aquifers and the NEUS samples have been previously described (Sauk and Hug 2022; Grégoire et al. 2023; George and Hug 2023). DNA extractions for the CA landfill were completed using the Qiagen Power Soil Pro kit following the manufacturer's instructions with the addition of the sterivex filter directly into the bead beating tube in place of soil. Metagenome sequencing, analysis, and host genome binning were performed as previously described for the SO and NEUS samples (Grégoire et al. 2023; George and Hug 2023). The CA landfill samples underwent metagenomic sequencing at The Center for Applied Genomics (Toronto), generating 2x150 bp reads on an Illumina HiSeq. For metagenome assembly and host genome binning, the same pipeline as for previously analyzed landfill metagenomes was used. In brief, read trimming was accomplished with bbduk (https://github.com/BioInfoTools/BBMap/blob/master/sh/bbduk.sh) and sickle (https://github.com/najoshi/sickle). Spades3 (Bankevich et al. 2012) (https://cab.spbu.ru/software/spades/) under the -meta flag with kmers 33, 55, 77, 99, and 127 was used for assembly. Scaffolds ≥ 2.5 kbp were retained for further analyses. Read datasets were mapped to assembled metagenomes using Bowtie2 (Langmead and Salzberg 2012). Scaffolds were binned using CONCOCT (Alneberg et al. 2014), MaxBin2 (Wu et al. 2016),

MetaBAT2 (Kang et al. 2019), and consensus bins determined using DAS Tool (Sieber et al. 2018). Resultant bins were quality assessed using CheckM (Parks et al. 2015), and taxonomically classified using the Genome Taxonomy Database Toolkit (Chaumeil et al. 2019). Genome bins that were >70% complete and less than 10% contaminated (defined as, at minimum, medium quality draft Metagenome Assembled Genomes (MAGs; Bowers et al. 2017)) were used as putative host MAGs in further analyses.

Viral identification, classification and genome binning

All scaffolds with a minimum length of 5 kbp were put through VirSorter2 (VS2; Guo et al. 2021) and CheckV (Nayfach et al. 2021) sequentially, following the protocol at dx.doi.org/10.17504/protocols.io.bwm5pc86. Since CheckV, along with all other state-of-the-art tools, may not detect proviral region boundaries accurately, all proviruses identified by CheckV were discarded to maintain the integrity of further analyses, including AMG identification. All putative non-integrated viruses that had either at least one viral gene annotated by CheckV, or had zero viral and host genes annotated by CheckV, were retained for subsequent clustering with CD-HIT-est (Li and Godzik 2006) using a global sequence identity threshold of c = 0.95. Viral genome binning was performed on the dereplicated set of viral scaffolds using vRhyme (Kieft et al. 2022). For viral taxonomic assignment, clustering, and comparisons of the landfill virome to related systems, predicted viral scaffolds with a minimum length of 10 kbp were input to vConTACT2 (Bin Jang et al. 2019) along with 24,982 viral elements from wastewater samples and 20,541 viral elements from groundwater samples (all of which were ≥ 10 kbp and obtained from IMG/VR (Paez-Espino et al. 2016a), and the vConTACT2 reference database "Prokaryotic Viral Refseq version 85 with ICTV-only taxonomy"). 10 kbp is the minimum size requirement for vConTACT2 hence viral scaffolds were size filtered accordingly. To assess the genetic

similarity of landfill viruses across distinct landfills specifically, vConTACT2 was run a second time on all ≥ 10 kbp viral elements identified from our landfill metagenomes.

In cases where we identified putative megaphages, we annotated terminase large subunit ORFs using HMMER3 (Eddy 2011) and Hidden Markov Models (HMMs) from the Virus Orthologous Groups database (https://vogdb.org/, release number vog218) and the Prokaryotic Virus Orthologous Groups (pVOGs) datatabase (Grazziotin et al. 2017). Predicted large terminase subunits were confirmed with VIBRANT (Kieft et al. 2020). Predicted large terminase subunit sequences were aligned with other giant phage large terminase subunit sequences (Al-Shayeb et al. 2020) using MUSCLE (Edgar 2004) version 5.1.linux64 and the Super5 algorithm. ModelTest-NG (Darriba et al. 2020) was used to predict the best protein evolutionary model from the MUSCLE alignment. A phylogenetic tree was inferred using RAxML v.8.2.12 (Stamatakis 2014) under the PROTGAMMAVTF model and was visualized using Geneious (Kearse et al. 2012).

Virus-host linking

Viral elements that were either at least 10 kbp in size, or at least 5 kbp in size if they were within a vMAG, were linked to their hosts through matching host CRISPR array spacers to viral protospacers. CRISPR arrays in host MAGs were detected using MINCED (https://github.com/ctSkennerton/minced) with the following parameters: -minSL [minimum spacer length] set to 25 and the -spacers flag included to output a fasta file of all identified CRISPR spacers. These CRISPR spacers were BLASTn (Altschul et al. 1990) searched against viral elements with -task set as 'blastn-short', with alignments retained only if they had at most one mismatch, no gaps, a query coverage of ≥ 90% (Chen et al. 2021) and an E-value ≤ 10^{-4}.

Cross-phylum infection curation

All host MAGs involved in putative cross-phylum interactions were manually examined to assess the legitimacy of the interaction and to reduce the likelihood of incorrectly binned scaffolds leading to predicted cross-phylum interactions. First, the CRISPR-array encoding scaffolds of the MAGs predicted to target a putative cross-phylum infecting virus were ensured to be at least 10kb in length. Next, we required all MAGs predicted to be involved in cross-phylum interactions to have a maximum contamination of 5.00% according to CheckM.

Assessment of CRISPR-Cas targeting space within and across landfills

Using the same workflow described above in "*Virus-host linking,*" we examined how often host CRISPR spacers from a site targeted viral elements from the same site and from different sites. Using a custom Python script, we next detected spacers that were encoded multiple times in a single dataset or those that were shared between data sets, with no mismatches. Reverse complements were considered matches, but hits between a query spacer and a subsequence of a larger spacer were not considered as matches, as an exact length match requirement was implemented.

Auxiliary metabolic gene (AMG) detection

AMGs were detected by inputting all dereplicated viral scaffolds into VIBRANT, which provides AMG predictions, with the '--virome' flag. Predicted AMGs with limited or no prior representation in literature were manually curated as follows: the genomic neighborhoods of these putative AMGs were examined for obvious host genes, integrase presence and position, flanking viral hallmark genes, and hypothetical viral genes to assess whether predicted AMGs were due to microbial scaffolds misannotated as viral scaffolds or microbial genes proximal to integrated prophage. To compare the incidence of rare municipal waste site-derived AMGs to

other systems, 100,000 viruses (≥ 5 kbp) were randomly sampled from IMG/VR v4 (Camargo et al. 2023) and subjected to the same AMG detection pipeline.

3.3 Results

Viral proportion and genome sizes

In the SO 2016 samples, viral scaffold prediction by VS2 and CheckV, followed by clustering with CD-hit, resulted in a total of 22,658 unique viral scaffolds (min size 5 kbp and 27 scaffolds ≥ 200 kbp). The SO 2017 samples had a total of 37,182 unique viral scaffolds (min size 5 kbp, 69 scaffolds ≥ 200 kbp), the NEUS samples had a total of 31,944 unique viral scaffolds (min size 5 kbp, 31 scaffolds ≥ 200 kbp), and the SC samples had a total of unique 5,419 viral scaffolds (min size 5 kbp, 11 scaffolds ≥ 200 kbp). Collectively, 97,203 viral scaffolds ≥ 5 kbp were identified (Table 3.1). Notably, seven of these viral scaffolds are predicted to be nearly complete or complete megaphage genomes as they are ≥ 500 kbp in length. Six out of seven of these sequences are predicted to be 100% complete according to CheckV's identification of high-confidence Direct Terminal Repeats (DTRs) in the sequences, and all seven sequences were predicted to circularize by vRhyme (Table 3.2). Large terminase subunit genes were identified in six out of seven putative megaphages, with five of these sequences placing within the Mahaphage clade of giant phages (Al-Shayeb et al. 2020) in a phylogenetic tree (Figure 3.1), and one of them being unclassified under the most recent classification scheme for giant phages (Al-Shayeb et al. 2020). The two largest putative megaphage sequences are ~641 kbp (NEUS_megaphage_1) and ~678 kbp (SO_2017_megaphage_1) in size. The 678 kbp phage is, to our knowledge, the third largest phage genome reported to date, with the largest previously reported phage genomes being 634, 636, 642, and 735 kbp from one report (Al-Shayeb et al.

2020), 656 kbp from another (Michniewski et al. 2021) and 717 kbp from a third (Chen and Banfield 2023).

Table 3.1: Summary statistics for metagenomes, MAGs, and viral scaffolds from the sampling sites.

Site	# of Metagenomes	Read count	Scaffolds ≥5 kbp	MAGs[1]	viral scaffolds ≥5 kbp[2]	% viral[3]
SO_2016	6	1,135,023,824	190,754	688	23,893	12.43
SO_2017	8	1,822,382,124	269,004	1,064	39,612	13.53
NEUS_2019	9	1,869,347,618	530,926	1,881	36,453	6.66
CA_2019	4	862,249,102	231,655	762	6,410	2.15

1 – Quality filtered bins with >70% completion, <10% contamination
2 – viral scaffolds detected (prior to CD-HIT clustering)
3 – % of total nucleotides in assembled scaffolds ≥5 kbp predicted to be of viral origin

Table 3.2: Genome statistics for megaphages identified in this study.

Megaphages from this study	Genome length (bp)	Sample site	% complete	Gene count	VIBRANT AMG count	CRISPR-encoding (spacer count)
NEUS_megaphage_4	508,676	NEUS	100	643	21	Y (32)
SO_2016_megaphage	518,809	SO_2016	100	812	20	N
NEUS_megaphage_3	589,338	NEUS	100	791	17	N
SO_2017_megaphage_2	617,251	SO_2017	100	900	11	N
NEUS_megaphage_2	623,658	NEUS	100	920	10	N
NEUS_megaphage_1	641,243	NEUS	100	1,001	14	Y (3)
SO_2017_megaphage_1	678,337	SO_2017	97.76	1,128	20	Y (10)

Genetic relatedness of viruses across ecosystems

vConTACT2 uses shared protein clusters (PCs) to delineate viral clusters (VCs) (Bin Jang et al. 2019). vContTACT2 was run against all ≥10 kbp viral elements (34,499) detected in our samples, along with a total of 45,523 viral elements from IMG/VR (Paez-Espino et al. 2016a) detected in either wastewater samples (24,982) or groundwater samples (20,541), and the "Prokaryotic Viral Refseq version 85 with ICTV-only taxonomy" reference database (2,617 reference sequences) (Figure 3.2). None of the viruses (≥10 kbp) detected in our samples were classified at the order, family, or genus level using vConTACT2's taxonomy framework. 17,164 of the landfill-derived viruses were placed into novel genus-level groupings, with unassigned taxonomy, and 3,579 of the landfill-derived viruses were singletons. The remaining 13,756 viruses were either outliers that shared some PCs with viruses in VCs, viruses that overlapped between two VCs, or viruses that were ambiguously considered clustered or singletons. While some viral elements across landfills, wastewater, and groundwater showed similarity (seen mainly in the top half of the network, Figure 3.2), there were numerous groups that only included viruses from the same sample type (mainly in the bottom half of the network; Figure 3.2). We next used vConTACT2 to examine how similar viral communities are across the different landfill samples specifically (Figure 3.3). Compared to the gene-sharing network that included wastewater and groundwater viruses (Figure 3.2), the gene-sharing network including just landfill viruses showed a much higher level of viral clustering (Figure 3.3) i.e., Figure 3.2 contains 13,163 VCs and Figure 3.3 contains only 6,463 VCs. The lower level of clustering observed in Figure 3.2 compared to

Figure 3.3, however, is at least partially influenced by some viral elements from outgroups (i.e., wastewater and groundwater) in Figure 3.2 sharing more PCs with viruses that were loosely linked in the large bundle of viruses observed in Figure 3.3, resulting in fragmentation of that group.

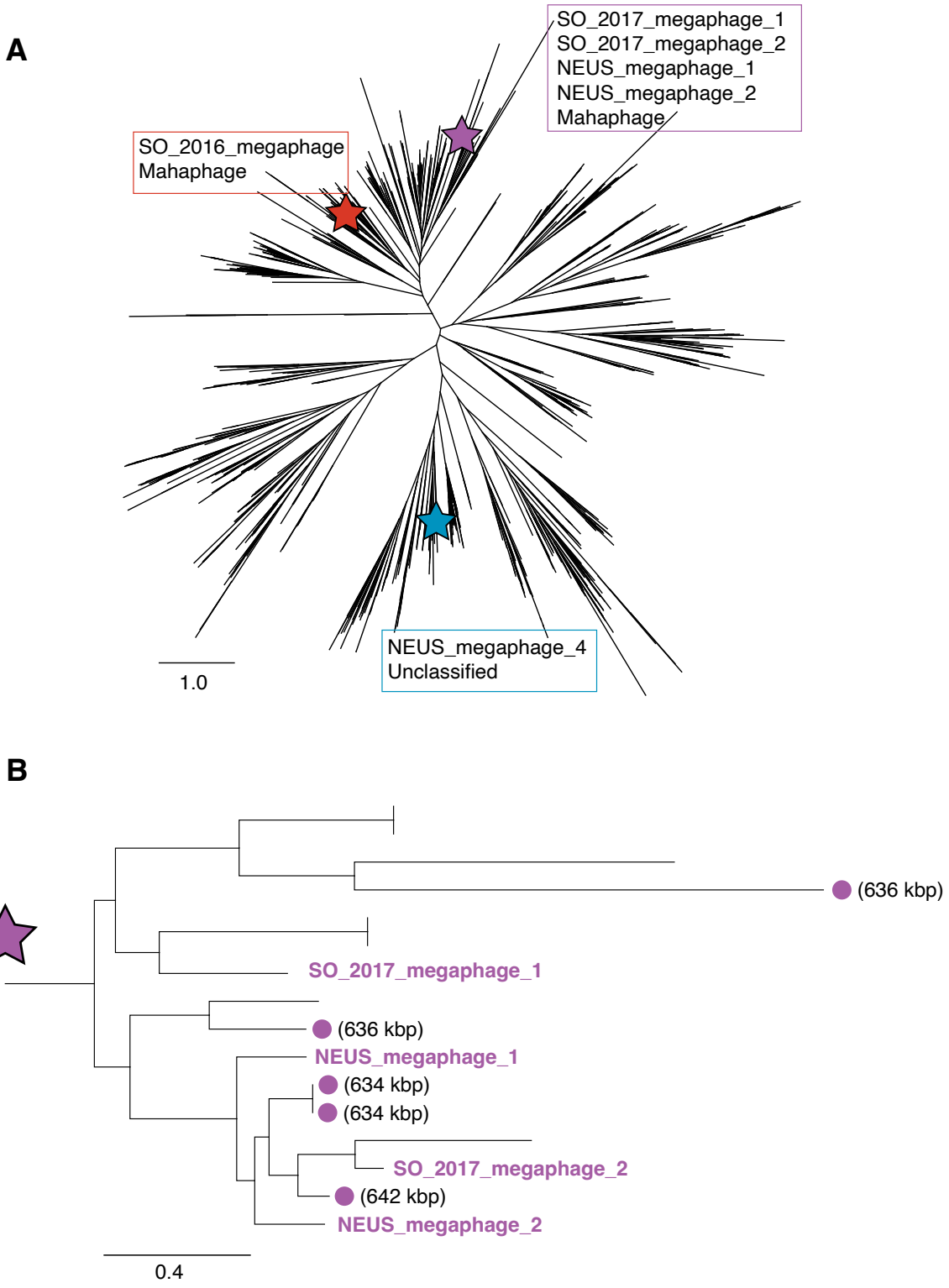

A
SO_2017_megaphage_1
SO_2017_megaphage_2
NEUS_megaphage_1
NEUS_megaphage_2
Mahaphage
SO_2016_megaphage
Mahaphage
NEUS_megaphage_4
Unclassified
1.0
B
(636 kbp)
SO_2017_megaphage_1
(636 kbp)
NEUS_megaphage_1
(634 kbp)
(634 kbp)
SO_2017_megaphage_2
(642 kbp)
NEUS_megaphage_2
0.4

Figure 3.1: A: Phylogenetic placement of newly discovered megaphages. Phylogenetic reconstruction was based on an alignment of large terminase subunit encoding sequences from this study (n=6) and those from a reference study (n=1,256, Al-Shayeb et al. 2020). Stars indicate the clades in which megaphages identified in this study were placed (Al-Shayeb et al. 2020). B: Placement of four newly discovered megaphages in a clade specific to huge phages. References phages in this clade include some of the largest phages identified from a previous study (Al-Shayeb et al. 2020). The sizes of reference phage genomes are indicated if they are ≥600 kbp.

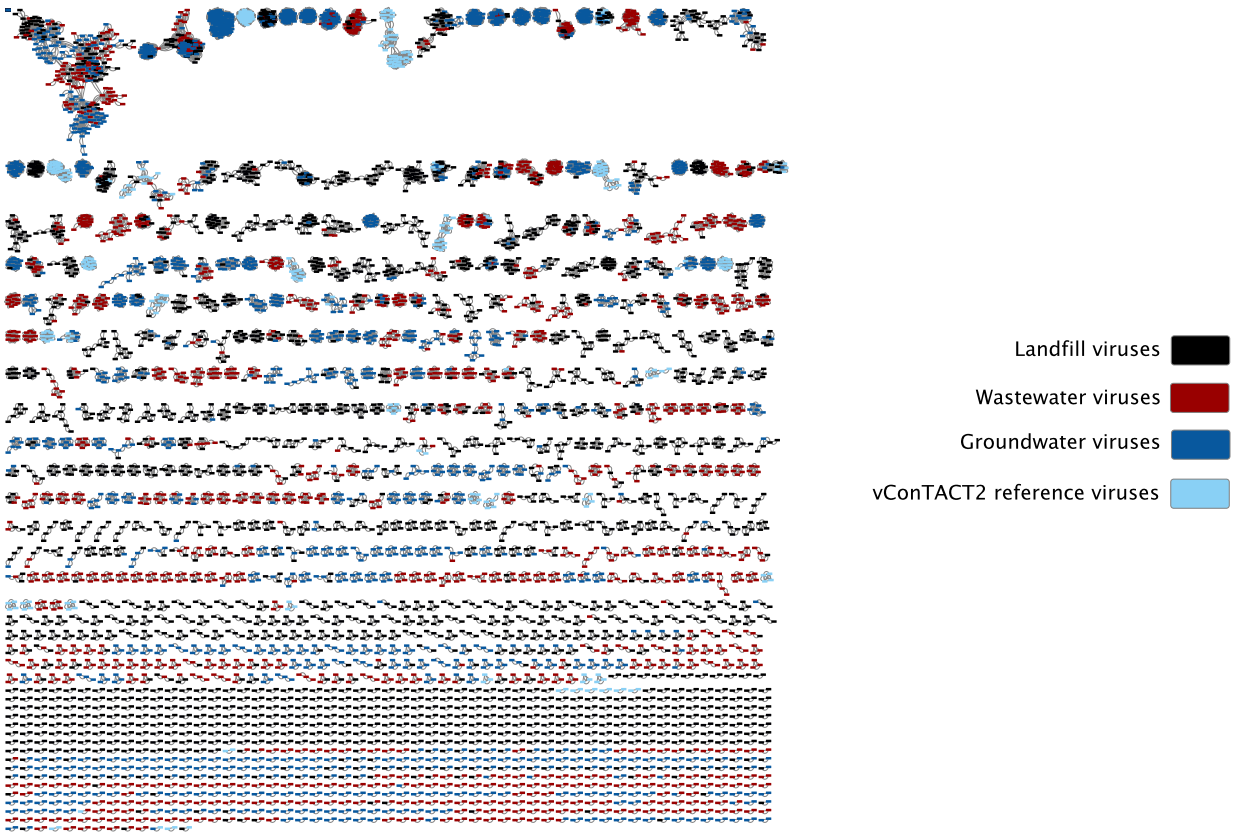


Figure 3.2: Gene-sharing network of landfill viruses with related groups from IMG/VR. Nodes represent viral elements and are coloured by sample site as summarized in the legend. Nodes connected by edges represent viral elements that share protein clusters. Network was generated using vConTACT2. The vConTACT2 reference database used was Prokaryotic Viral Refseq version 85 with ICTV-only taxonomy**.**

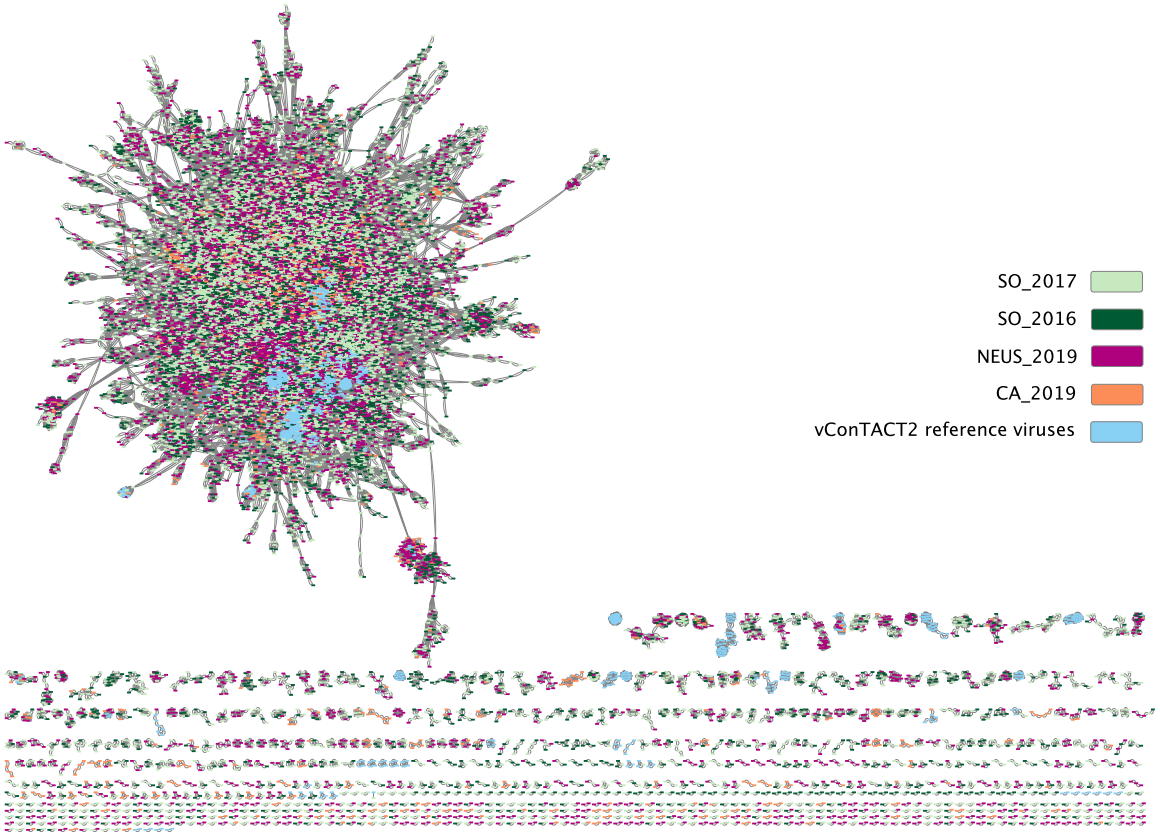


Figure 3.3: Gene-sharing network of landfill viruses. Viral nodes are coloured by the landfill they were identified in. Nodes connected by edges represent viral elements that share protein clusters. Network was generated using vConTACT2. The vConTACT2 reference database used was Prokaryotic Viral Refseq version 85 with ICTV-only taxonomy**.**

Host-virus connections within and across landfills

We examined CRISPR spacer to viral protospacer targeting both within a particular landfill site and across distinct landfills/times. SO_2016 showed the strongest targeting to NEUS_2019, having more spacer to protospacer matches than NEUS_2019 did to SO_2016, 327 vs. 262, despite NEUS_2019 having almost four times the number of spacers as SO_2016 (Table 3.3). SO_2017 showed strongest targeting to its preceding temporal dataset, SO_2016. NEUS_2019 spacers targeted viruses in its own dataset most strongly. Lastly,

CA_2019 was targeted least often by spacers from other datasets. Despite having, at minimum, nearly four times fewer viral elements than the other datasets, CA_2019 spacer to protospacer matches were the second highest out of all 16 comparisons (Table 3.3). The SO_2016 networks showed a low diversity of both hosts and viral elements, but two host MAGs, LW2_3 and LW2_22, had 55 and 32 CRISPR spacer matches, respectively, to a single vMAG. Notably, this vMAG contained two out of four viruses that were hyper-targeted, as described in Chapter 2. All sample sites showed instances of hyper-targeted viral elements and/or vMAGs, which we previously defined as viral elements (or vMAGs) targeted at least 20 times by a single host's complement of CRISPR spacers (George and Hug 2023, presented in Chapter 2). The maximum number of spacer-to-protospacer matches between a host MAG and a viral element (or vMAG) per dataset are 55 (SO_2016), 24 (SO_2017), 32 (CA_2019), and 22 (NEUS_2019). After curation, across all networks, a total of five vMAGs and one viral element were predicted to infect across distinct phyla (Table B.1), with one vMAG from NEUS_2019 predicted to infect three distinct bacterial phyla: Bacteroidota, Cloacimonadota, and Firmicutes_B. The SO_2016 network was the only network that didn't include putative cross-phylum infections. Lastly, considering the relatively low viral scaffold count from the CA_2019 dataset (Table 3.1), it showed relatively high host and viral diversity in its network (Table 3.4).

Table 3.3. Host CRISPR spacer to viral protospacer targeting within and across municipal waste sites. Host CRISPR spacers and viral elements were not dereplicated for this analysis. VE = viral elements.

	# of spacer-protospacer matches (# VE)			
Dataset (spacer count)	**to SO_2016 (23,893 VE)**	**to SO_2017 (39,612 VE)**	**to NEUS_2019 (36,453 VE)**	**to CA_2019 (6,410 VE)**
SO_2016 (5,446)	266	186	327	12
SO_2017 (12,578)	461	451	443	21
NEUS_2019 (20,236)	262	654	1,595	79
CA_2019 (9,299)	119	39	121	853

Table 3.4: Virus-host networks summary. Virus-host interactions are based on host MAG CRISPR spacers and viral element, or vMAG if element was binned, protospacer connections. Unique links = 1 MAG to 1 viral element, regardless of number of spacer-protospacer connections.

Network	**Host MAGs**	**Viral scaffolds or vMAGs**	**Unique links**
SO_2016	13	20	22
SO_2017	73	100	115
NEUS_2019	169	206	321
CA_2019	68	86	127

CRISPR spacer conservation within and across waste sites

We next examined conservation of specific CRISPR spacers within and between landfill samples (Table 3.5). Generally, SO_2016 and SO_2017 had the most spacers in common, which was expected given they are temporal samples from the same location, and with overlapping sampling sites. The CA_2019 spacer set had very limited overlap with the other landfill sites.

Table 3.5: Shared CRISPR spacers within and between landfill datasets. Each dataset's host spacers were dereplicated by CD-Hit at 100% identity for this analysis. Shared spacers required exact matches. Reverse complement matches were included in this analysis.

Dataset	Number of host spacers	Spacers shared with SO_2016	Spacers shared with SO_2017	Spacers shared with NEUS_2019	Spacers shared with CA_2019
SO_2016	5,361	N/A	1,252	162	6
SO_2017	11,812	1,252	N/A	174	10
NEUS_2019	17,581	162	174	N/A	17
CA_2019	7,223	6	10	17	N/A

Viral auxiliary metabolic genes

Across all landfills, a total of 582 unique AMGs were identified (Table B.2), representing functions across all major cellular processes (Figure 3.4). AMG profiles across different landfills and compared to the IMG/VR dataset were quite similar, with the exception of CA_2019 (Figure 3.4). CA_2019 was the only landfill to receive industrial waste in addition to municipal waste and had relatively high counts of AMGs associated with lipid, amino acid, and aromatic compound metabolism, and comparatively low counts of AMGs associated with carbohydrate and terpenoids/polyketide metabolism. AMGs for processes that had not previously been associated with viruses or those that had very limited representation in literature were manually curated (Table 3.7). AMGs of high interest that passed manual curation included those predicted to be involved in carbohydrate metabolism, methane metabolism, and sulfur metabolism. The CA_2019 dataset included an AMG predicted to have a role in degradation of organochlorine compounds, which has been observed in previous studies (Zheng et al. 2022; Yuan et al. 2023). Numerous AMGs were also predicted in the putative megaphages - full details of VIBRANT-predicted AMGs encoded by these putative megaphages are available in Table B.3.

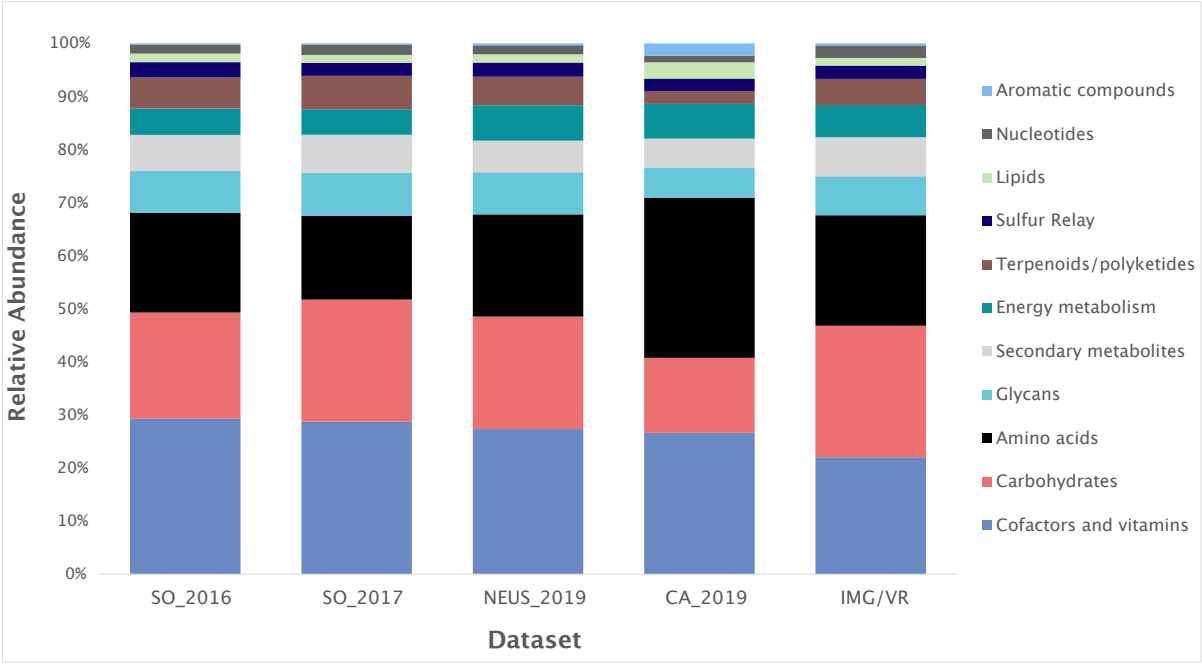


Figure 3.4: AMG metabolism category proportions across datasets. AMGs were detected using VIBRANT. The IMG/VR dataset was used as a reference and was generated by randomly sampling 100,000 viruses from IMG/VR4 that were ≥5 kbp in size. The legend denotes AMG metabolism categories assigned by VIBRANT. Counts of viral elements (≥5 kbp) per in-house dataset are as follows: SO_2016 - 23,893, SO_2017 - 39,612, NEUS_2019 - 36,453, CA_2019 - 6,410.

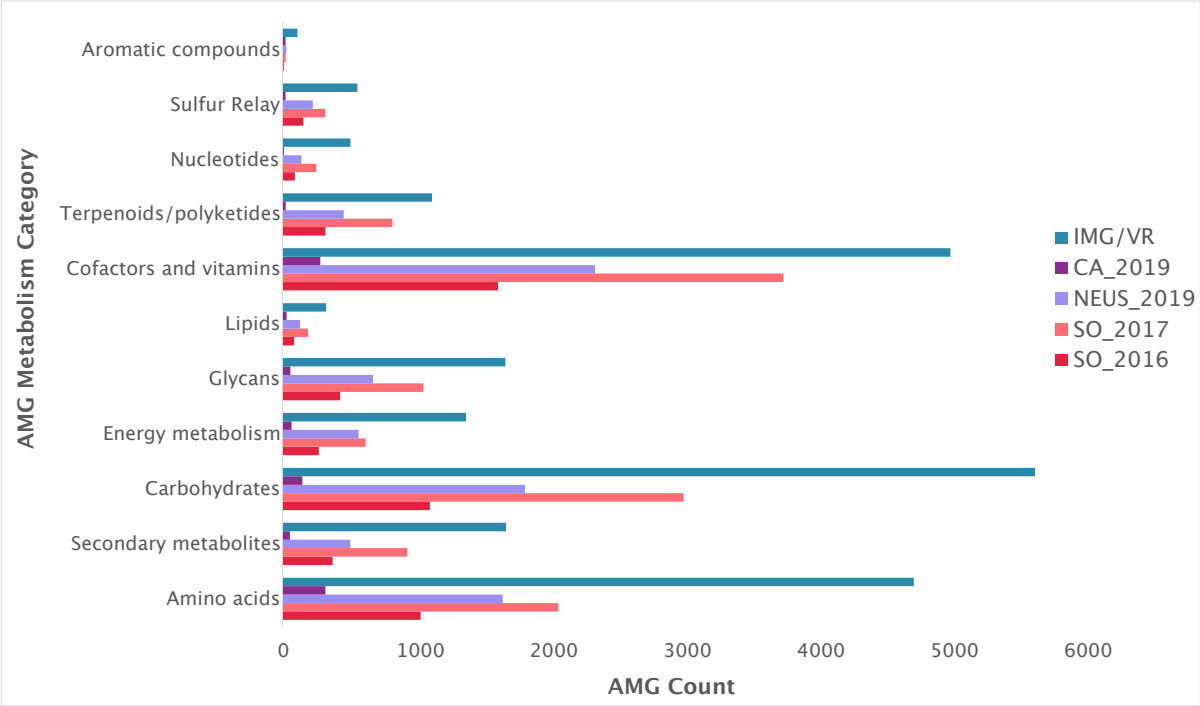


Figure 3.5: AMG counts per metabolic category across landfill datasets. AMGs were detected using VIBRANT. The legend denotes the different datasets in which AMGs were detected. The IMG/VR dataset was used as a reference and was generated through randomly sampling 100,000 viruses from IMG/VR4 that were ≥5 kbp in size. Counts of viral elements (≥5 kbp) per in-house dataset are as follows: SO_2016 - 23,893, SO_2017 - 39,612, NEUS_2019 - 36,453, CA_2019 - 6,410.

Table 3.6: Notable auxiliary metabolic genes encoded by viruses in the landfill datasets. Yellow highlighted entries represent functions also observed in 100k randomly selected IMG/VR prokaryotic viruses, but which are not widely reported in the literature. Functions are presented in alphabetical order according to AMG name.

AMG name	Sites where identified	Times identified per site	AMG function
aprB; adenylylsulfate reductase, subunit B [EC:1.8.99.2]	SO_2017	SO_2017: 1	Sulfur metabolism
dehH; haloacetate dehalogenase [EC:3.8.1.3]	CA	CA:1	Chlorocyclohexane and chlorobenzene degradation, Chloroalkane and chloroalkene degradation
E3.2.1.1A; alpha-amylase [EC:3.2.1.1]	NEUS	NEUS: 1	Polysaccharide metabolism
frhB; coenzyme F420 hydrogenase subunit beta [EC:1.12.98.1]	SO_2016, SO_2017	SO_2016: 1 SO_2017: 8	Methane metabolism
hdrA2; heterodisulfide reductase subunit A2 [EC:1.8.7.3 1.8.98.4 1.8.98.5 1.8.98.6] / 4Fe-4S binding domain	NEUS	NEUS: 1 SO_2017: 1	Methanogenesis
mcrD; methyl-coenzyme M reductase subunit D	CA	CA:1	Methanogenesis, anaerobic methane metabolism
mddA; methanethiol S-methyltransferase [EC:2.1.1.334]	CA, SO_2017	CA:1, SO_2017:1	Methionine metabolism to dimethyl sulfide
phoD; alkaline phosphatase D [EC:3.1.3.1]	CA, NEUS	CA:3 NEUS:1	Alkaline phosphatase
rbcL; ribulose-bisphosphate carboxylase large chain [EC:4.1.1.39]	NEUS, SO_2016, SO_2017	NEUS:1 SO_2016:1 SO_2017:3	D-ribulose 1,5-bisphosphate carboxylation, carbon dioxide fixation
soxX; L-cysteine S-thiosulfotransferase [EC:2.8.5.2]	CA	CA:1	Thiosulfate oxidation

3.4 Discussion

Here we examined the diversity of viruses in municipal waste sites, contaminated engineered systems with limited prior exploration of microbial community dynamics beyond microbial taxonomic diversity. Limited overlap between viral communities from wastewater and our landfills was observed (Figure 3.2). This suggests that, while engineered sites as a whole are underexamined from a viral community perspective, two distinct contaminated engineered sites,

i.e., landfills and wastewater, can vary considerably in their viromes. When we compared our landfill communities relative to each other, we observed multiple Viral Clusters (VCs) that either included viral elements from more than one landfill site or were closely connected to VCs of viral elements from another landfill site (Figure 3.3). This suggests that the landfill virome is conserved, but distinct relative to the wastewater virome. None of the landfill-derived viruses were able to be taxonomically classified in a comparison to reference viruses using vConTACT2. Coincident with writing this thesis, but after performing our vConTACT2 analyses, the International Committee on Taxonomy of Viruses (ICTV) radically changed their taxonomic classification of prokaryotic viruses (Zerbini et al. 2023). The latest Virus Metadata Resource from the ICTV is available (https://ictv.global/vmr). Many of the vConTACT2 reference viruses used here had changes to their taxonomy under this new classification, mainly due to the abolishment of the order Caudovirales (Turner et al. 2023). This would not change the taxonomic assignment of the landfill-derived viruses, as they were all unassigned, which indicates that the diversity of landfills viruses is not well-represented by the NCBI viral RefSeq database used in vConTACT2. This is further evidenced by the absence of viral RefSeq reference genomes in nearly 99% (6,023/6,116) of the VCs that include our landfill viral elements. We note that while metagenomic discovery of prokaryotic viruses often outpaces taxonomic classification (Turner et al. 2023), new tools are required to efficiently classify viruses under the most recently released taxonomic frameworks (Turner et al. 2023; Zerbini et al. 2023).

Across all landfill datasets, we identified a total of 138 jumbophage genomes (i.e., those ≥200 kbp in size) with seven predicted to be complete or nearly complete megaphage genomes (i.e., ≥500 kbp). One of these putative megaphage genomes, SO_2017_megaphage_1, is ~678 kbp long, which is, to our knowledge, the third largest phage genome reported to date. It is

noteworthy that the two largest megaphage genomes reported in literature (735 and 717 kbp) were described in a report that examined 13 ecosystems (Al-Shayeb et al. 2020), and a report that examined 231 freshwater metagenomes (Chen and Banfield 2023), respectively – these reports are also over three years apart. This suggests that bacteriophage genomes of this size are very rare in nature, yet despite this, we have identified several of these genomes from only two ecosystems (landfills and their adjacent groundwater wells). The number of genes encoded by the seven megaphages ranges from 643 to 1,128, with 10-20 predicted AMGs (Table B.3). Three of the seven megaphages encoded CRISPR arrays, with the 641 kbp and 678 kbp phage genomes now being the largest reported phage genomes to encode CRISPR arrays (exceeding the previous length of 620 kbp reported by Al-Shayeb et al. 2022). None of the seven megaphages were targeted within our virus-host networks, nor did the CRISPR arrays encoded by these viruses target any other viruses in our data. We were able to identify the large terminase subunit genes for six of the megaphages, which we used to place them in a reference tree of huge phages using reference sequences from Al-Shayeb and colleagues (2020) (Figure 3.1). Four of our megaphage genomes that were >600 kbp grouped together in a clade that included some the largest phage genomes described in literature (Al-Shayeb et al. 2020; Chen and Banfield 2023), all of which are considered as “Mahaphages” and predicted to infect members of the phylum Bacteroidetes (Al-Shayeb et al. 2020; Chen and Banfield 2023). This suggests that while huge phages are taxonomically diverse (Al-Shayeb et al. 2020), the phages with the largest genomes have shared ancestry across unconnected environments.

In our virus-host networks, consistent with our previous work (George and Hug 2023), we observed hyper-targeted viral elements and viral elements predicted to infect across distinct phyla. Our consistent observation of putative cross-phylum infecting viruses in landfills, despite

such viruses being reported in a limited number of other studies (Malki et al. 2015; Paez-Espino et al. 2016b; Dalcin Martins et al. 2018; Hwang et al. 2023), suggests that landfills may be enriched for these broadly infecting viruses. Recent work showed that these very broad host range viruses are most often observed in syntrophic biofilms found in densely populated ecosystems (Hwang et al. 2023). Landfills are typically anaerobic systems where labile carbon is utilized fully through microbial handoffs and syntrophy, as evidenced by consistent and long-term methanogenesis. Syntrophy is predicted to be widespread in landfills (Meyer-Dombard et al. 2020), and biofilm modes of life likely dominate in solid waste repositories, meaning the ecosystem characteristics of landfills are consistent with those predicted to select for broadly infecting viruses (Hwang et al. 2023). While we manually curated our putative cross-phylum interactions, and what we observe is consistent with existing literature, we note that because we solely utilize CRISPR spacer-viral protospacer matches to identify these interactions, these observations could stem from incorrectly binned CRISPR arrays. *In-vitro* evidence is necessary to substantiate these predictions.

We identified conserved CRISPR spacers between our different landfill samples to test whether landfills with higher similarity in their viral communities, and higher levels of virus-host cross-targeting between sites, also shared similar protospacer targets in viral genomes. Consistent with our examinations of viral community similarity, and cross-targeting, both SO_2016 and SO_2017 shared >160 spacers with NEUS_2019, and SO_2016, SO_2017, and NEUS_2019 all shared <20 spacers with CA_2019. Omitting shared spacers between SO_2016 and SO_2017, as they are temporal datasets from the same landfill, ~1-3% of spacers were shared between SO samples and NEUS_2019 with 100% identity and equal length. This suggests that landfills that share more similar viruses are more likely to share host CRISPR spacers,

however, these shared spacers alone do not account for the cross-targeting observed between these sites.

From examining 27 metagenomes from three geographically distinct landfills across North America, we expand our previous work on landfill viromes (Chapter 2), showing that landfill viral diversity is relatively conserved, but distinct relevant to similar environment types such as wastewater and groundwater. We detected several instances of viruses predicted to infect across distinct phyla, AMGs with limited reports in literature, and huge phage genomes. Consistent with our previous observations (Chapter 2), we detected phages with the potential to infect across distinct phyla in three geographically distinct waste sites. The observation of phages with such a broad host range contrasts with the highly specific isolate phages used in phage therapy applications. With limited phages isolates to choose from, researchers with interest in phage therapy have turned to isolating phages from environments such as sewage and wastewater in order to combat drug resistant infections (Alharbi and Ziadi 2021). Similar to the phages predicted in our study, these environmental phages could have a host range beyond the target pathogen, which may not be readily apparent through conventional assays. We therefore suggest computational tools (Edwards et al. 2016; Roux et al. 2023) be used in combination with conventional assays in order to clarify phage host ranges for therapeutic applications. Some AMGs detected in this study have limited instances reported in literature, expanding our knowledge of genes that can be laterally transferred within microbial communities through Mobile Genetic Elements. These include genes involved in organochlorine and methane metabolism, both of which are relevant to landfills as contaminated sites that are globally significant producers of methane (Grégoire et al. 2023). Our consistent observation of some of the largest bacteriophage genomes described to date in landfill metagenomes, despite phage

genomes of this size being rarely reported in the literature, suggests that landfills may be a hot spot for these viruses. Further examination of these huge phages and other landfill viromes may expand our knowledge of the cellular genes harboured in huge phage genomes, which may include novel genes encoding CRISPR-Cas systems that can be used in biotechnological applications (Al-Shayeb et al. 2020, 2022).

Chapter 4 Virally encoded CRISPR-Cas systems

4.1 Introduction

CRISPR-Cas systems are broadly classified into Class I and Class II systems, each with three major types and multiple subtypes, all of which employ RNA-directed targeting of either DNA or RNA (Makarova et al. 2020). Class II systems are of interest for biotechnological applications as they are more streamlined in size, containing a single multi-domain effector nuclease (Van Der Oost and Patinios 2023). In 2014, using metagenomic data, researchers were able to identify new variants of Cas9 and novel Cas proteins that weren't previously reported in literature (Zhang et al. 2014). In 2017, Burstein and colleagues identified some of the most streamlined Class II CRISPR-Cas systems seen to date, CRISPR-CasX (Cas12e) and CRISPR-CasY (Cas12d) from metagenomic data (Burstein et al. 2017). Unlike in previous predictive work, these systems were artificially synthesized and expressed in *E. coli* to prove they facilitate RNA-guided DNA cleavage (Burstein et al. 2017). This discovery illustrated the strong potential of metagenomic methods to augment discovery of novel CRISPR-Cas systems not represented in cultivated bacteria and archaea. Subsequent metagenomic and biochemical studies described additional minimal CRISPR-Cas systems from microbes, details of which are well-summarized in two works (Makarova et al. 2020; Van Der Oost and Patinios 2023). These "minimal" systems contain smaller effector proteins, some of which are not dependent on accessory genes, or genes previously thought to be essential.

CRISPR-Cas systems and their evolutionary progenitors are not exclusively encoded in prokaryote genomes: multiple Mobile Genetic Elements (MGEs) including plasmids, bacteriophages, and insertion sequences encode CRISPR-Cas systems and their predecessors (Faure et al. 2019b; Koonin et al. 2020). Indeed, Class II system effector nucleases that target

DNA (see Figure 4.1 for the current nomenclature) are predicted to have evolved multiple times from proteins encoded by insertion sequences. These insertion sequence-encoded proteins include IscB, the likely ancestor of Cas9, and TnpB, the likely ancestor of Cas12 (Altae-Tran et al. 2021), both of which Cas descendants have been repurposed as genetic engineering tools (Altae-Tran et al. 2021; Karvelis et al. 2021). The subtype V-U systems are predicted to be evolutionary intermediates between TnpB and full-fledged type V effectors (Figure 4.1, Makarova et al. 2020). The shared ancestry of Class II Cas effector nucleases, IscB, and TnpB is apparent through a common nuclease domain known as RuvC (Makarova et al. 2020), which is depicted in light green in Figure 4.1.

Virus-encoded CRISPR arrays and CRISPR-Cas systems have been identified from a variety of environments (Minot et al. 2011, 2013; Seed et al. 2013; Hargreaves et al. 2014; Chénard et al. 2016; Faure et al. 2019; Medvedeva et al. 2019; Al-Shayeb et al. 2020, 2022; George and Hug 2023) and are involved in crippling host viral defense systems (Seed et al. 2013); regulating host transcription and translation (Al-Shayeb et al. 2020); CRISPR-Cas system inhibition (Faure et al. 2019b); and interviral conflicts (Faure et al. 2019; Medvedeva et al. 2019; George and Hug 2023). In the case of interviral conflicts, a CRISPR-encoding virus, integrated as a provirus, can provide its host with immunity against other viruses, akin to a superinfection exclusion system (Faure et al. 2019).

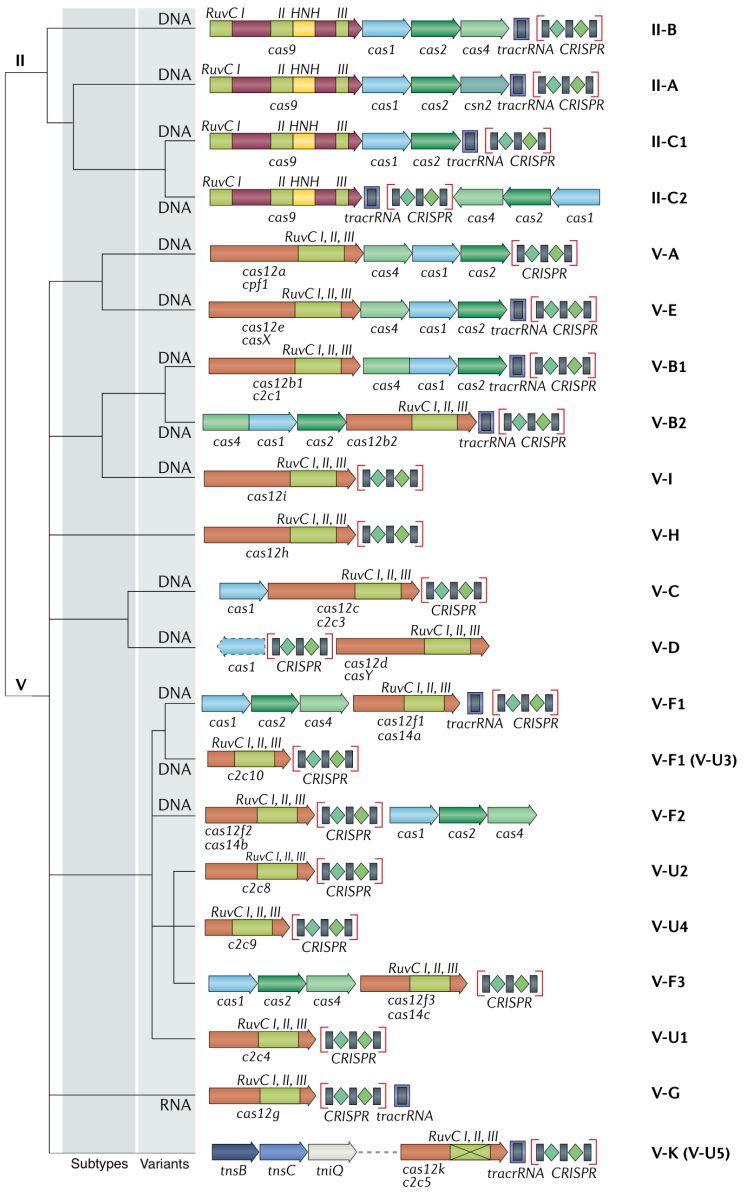


Figure 4.1: Classification of Class II CRISPR-Cas systems. This figure was adapted from Makarova et al. 2020, "Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants." Representative CRISPR-Cas loci for Class II subtypes and selected variants are displayed. The dendrogram depicts the likely evolutionary relationships between loci. Commonly used legacy names are denoted below systematic names where relevant. Dispensable modules are in dashed lines. Effector modules/nucleases are those containing RuvC domains (light green). Subtypes V-U(1-5) have not been experimentally validated.

A large-scale investigation into giant phages by Al-Shayeb and colleagues (Al-Shayeb et al. 2020), and a follow-up study validating the genome-editing potential of a virus-encoded Cas protein, Cas12Φ, (Pausch et al. 2020), highlighted virally encoded CRISPR-Cas systems as an avenue for discovery of new CRISPR-Cas systems with divergent properties. Potentially owing to the strong evolutionary pressure on viruses to maintain compact genomes, virally encoded CRISPR-Cas systems are quite minimal, making some of them excellent candidates for development as biotechnological tools (Pausch et al. 2020). A survey of a more complete set of phage genomes, i.e., not just those of giant phages, illustrated a wide diversity of virally encoded CRISPR-Cas systems, and an enrichment in streamlined Class II CRISPR-Cas systems. This study was the first to report a second Cas12 system exclusively encoded by viruses, the Cas12λ (Cas12L), representatives from which were also validated for genome-editing capabilities (Al-Shayeb et al. 2022). Considering that some of the newest and most streamlined additions to the CRISPR-Cas toolkit have been identified from viral metagenomics data, investigations into virally encoded CRISPR-Cas systems in other metagenomic datasets is a worthwhile endeavor. In thesis Chapters 2 and 3, CRISPR arrays were detected in diverse viral populations from municipal landfills. Here, we examine the Cas effector nuclease genes in the vicinity of these viral CRISPR arrays to assess the encoded diversity of Cas systems in landfill prokaryotic viruses, and to determine if any virally encoded CRISPR-Cas systems detected are novel Class II systems.

4.2 Methods

Viruses encoding CRISPR arrays and interviral conflicts

Viruses were previously detected in 27 metagenomes using VirSorter2 and CheckV as described in Chapter 3, section 3.2. All predicted viral elements were screened for CRISPR arrays using

MINCED (https://github.com/ctSkennerton/minced) with the following parameters: -minSL [minimum spacer length] set to 25 and the -spacers flag included to output a fasta file of all identified CRISPR spacers. To examine whether virally encoded CRISPR arrays targeted other viral elements, the spacer sequences from viral CRISPR arrays were BLASTn searched against all non-CRISPR-array-encoding viruses identified in this study (-task set as 'blastn-short', with alignments retained only if they had at most one mismatch, no gaps, a query coverage of ≥ 90% and an E-value ≤ 10^{-4}). Any hits were annotated using the latest release of the Virus Orthologous Groups database (https://vogdb.org/, release vog217). vConTACT2 (Bin Jang et al. 2019) was employed to determine whether viruses predicted to encode CRISPR systems were taxonomically distinct relative to viral elements that were not predicted to encode CRISPR systems.

Cas effector genes proximal to viral CRISPR arrays

CRISPRCasTyper (Russel et al. 2020) was used for automated detection of CRISPR-Cas operons in viral elements predicted to encode CRISPR arrays. Manual detection of Class II effector nucleases was performed as follows. All 10 kbp regions directly upstream or downstream from CRISPR arrays were examined for Open Reading Frames (ORFs) using prodigal (Hyatt et al. 2010) (flags -m -p meta -q). All predicted ORFs that were 300-800 amino acids (aa) in length were retained for further analyses to identify streamlined Class II proteins. This 800 aa limit was set to exclude Cas protein effectors that were not smaller in size than the current most streamlined Cas effector subtype, Cas12Φ/Cas12j, whose members are ~800 aa in size (Al-Shayeb et al. 2020; Pausch et al. 2020). Class II effector-encoding *cas* genes proximal to CRISPR arrays were predicted by querying CRISPR-array-adjacent ORFs using profile Hidden Markov Models (HMMs) (Eddy 1998) obtained from previous studies (Burstein et al. 2017;

Russel et al. 2020) and HMMER v.3.2.1 (Eddy 2011). Additionally, BLASTp (Altschul et al. 1990) searches were performed using multiple protein families that have been previously classified, and in some cases experimentally validated, as queries: TnpB (the predicted ancestor of Cas12 nucleases [Shmakov et al. 2017; Makarova et al. 2020]); subtypes of Cas12; V-U nucleases (some of which have not been validated for interference activity); and Cas14 (Harrington et al. 2018; Makarova et al. 2020; Al-Shayeb et al. 2020; Wu et al. 2022); along with the recently identified and experimentally validated phage-encoded Cas12Φ (Cas12j) and Cas12λ (Cas12L). The screen did not include type VI effector proteins as queries, which are RNA-guided RNA targeting systems, as previous work demonstrates these systems aren't nearly as common in viruses compared to the other Class II systems, i.e., type II and type V (Al-Shayeb et al. 2022), and I was specifically interested in systems capable of DNA manipulation.

Identification of putatively novel Cas protein effectors

Examinations were focused on type II and type V nucleases, as diverse nucleases of these types were previously identified in phages (Al-Shayeb et al. 2022). All ORFs that hit to type II nucleases were manually examined for the presence of RuvC and HNH motifs. ORF hits to type V nucleases were gathered. Due to some type V nucleases having very low identity to other subtypes (Al-Shayeb et al. 2022), we also generated a separate dataset of all ORFs proximal to viral CRISPR arrays that were not identified as matches using our profile HMM or BLASTp searches ("Proximal ORFs"). The Proximal ORFs were combined with all reference type V nucleases and TnpB sequences, and aligned using MUSCLE version 5.1.linux64 and the super5 algorithm. ModelTest-NG (Darriba et al. 2020) was used to predict the best protein evolutionary model from the MUSCLE alignment. A phylogenetic tree inferred using RAxML v.8.2.12 (Stamatakis 2014) with -m PROTGAMMAVTF was visualized using Geneious. Since the

Proximal ORFs/proteins that were not hit by reference HMMs could include *cas* genes that don't encode effector nucleases, or genes unrelated to CRISPR-Cas systems, putative effector proteins in this set were identified through BLASTp searching representatives of each clade of sequences on the phylogenetic tree against NCBI's non redundant protein database (nr). A clade was retained only if the proteins/domains hit in nr were hypothetical proteins, TnpB, or endonucleases (excluding restriction endonucleases). Each retained clade was manually examined for RuvC motifs, and those with a detectable RuvC signature that also clustered close to known type V effector proteins and TnpB in the phylogenetic tree, were retained as putative Cas effector proteins. This subset of the Proximal ORFs dataset was then combined with the predicted effector ORFs that were identified through the HMM and/or BLASTp searches. This final aggregated dataset was aligned using MUSCLE version 5.1.linux64 and the PPP algorithm, due to smaller input size. ModelTest-NG (Darriba et al. 2020) was used to predict the best protein evolutionary model from the MUSCLE alignment (Edgar 2004). A phylogenetic tree was inferred using RAxML v.8.2.12 (Stamatakis 2014) with -m PROTGAMMAVTF and was visualized using Geneious.

4.3 Results

Virally encoded CRISPR arrays and interviral conflicts

A total of 106,371 unclustered landfill-derived viral elements that were ≥5 kbp were included in examinations of virally encoded CRISPR arrays and novel Cas protein identification. Only 0.43% (454) of these viral elements encoded CRISPR arrays, a set which had an average length of ~47 kbp. Both viral CRISPR-array prevalence and scaffold lengths were in line with a previous examination of CRISPR arrays and CRISPR-Cas systems in viral elements (Al-Shayeb et al. 2022). Only 9/454 CRISPR array-encoding viruses were targeted in the virus-host networks

described in Chapter 3, with predicted hosts from seven different phyla. vConTACT2 requires ≥10 kbp scaffolds; 315/454 CRISPR-encoding viruses were long enough to be included in this analysis. Comparison between CRISPR-encoding and non-CRISPR encoding viruses did not show many distinct viral clusters between the two groups (Figure 4.2). Viral arrays encoded 2-102 spacers with an average spacer count of 8, with 15 viral arrays encoding at least 30 spacers. Notably, the five longest CRISPR arrays were encoded on relatively short scaffolds (5,027-7,697bp) that did not contain any detectable viral genes according to CheckV, or any detectable host (bacterial/archaeal) genes. These five scaffolds were conservatively considered false positive viral elements, most probably deriving from other Mobile Genetic Elements (MGEs), such as plasmids. Spacers from predicted virally encoded CRISPR arrays targeted 164 of the 105,917 viral elements lacking CRISPR arrays (0.15%). The regions targeted were annotated to 9 Virus Orthologous Groups (VOGs), and included collagen-like proteins, a virion structural protein, a baseplate wedge protein, and a Type II methyltransferase.

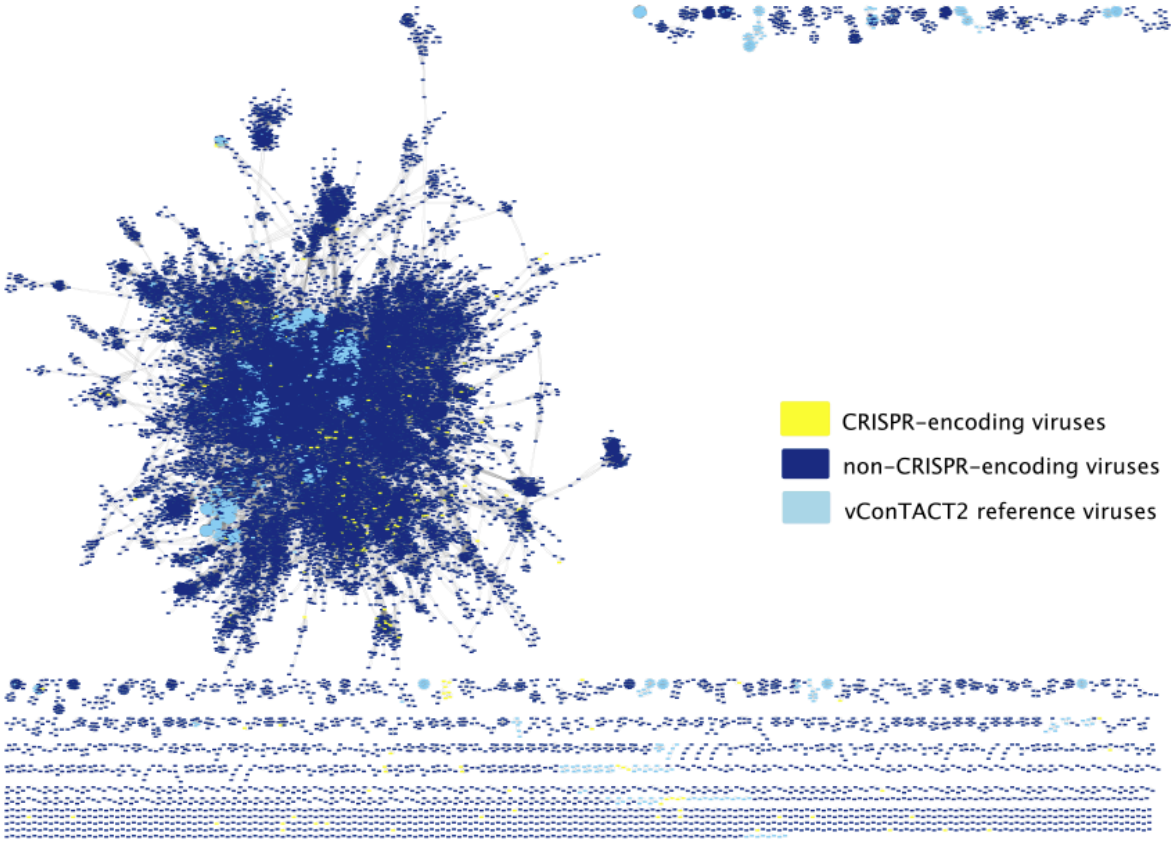


Figure 4.2: Gene-sharing network of CRISPR and non-CRISPR-encoding viruses. Nodes represent viral elements and nodes connected by edges indicate shared protein clusters between viral elements. The vConTACT2 reference database used was Prokaryotic Viral Refseq version 85 with ICTV-only taxonomy**.**

The longest, high-confidence virally encoded CRISPR array encodes 55 spacers. The longest putative viral elements that encoded CRISPR arrays containing 30+ spacers included a 364 kbp jumbophage (55 spacers) and a 508 kbp megaphage (32 spacers), both of which are predicted to be complete viral genomes according to CheckV. Multiple CRISPR arrays were less common, with 48 viral elements predicted to encode two or more CRISPR arrays (10.6% of CRISPR-array-encoding viral elements), with a maximum of six predicted CRISPR arrays for a

single element. In total, 519 CRISPR arrays were associated with the 454 array-encoding elements.

The viral element with six CRISPR arrays is a predicted jumbo phage that is 421,673 bp in length and >95% complete, according to CheckV. All six arrays were confirmed using CRISPR-CasTyper and five of these arrays were confirmed with CRISPR-CasFinder (Couvin et al. 2018), the latter of which assigns evidence levels of 1-4 (4 being the strongest) to its predictions. CRISPR-CasFinder predicted a total of 9 CRISPRs in this element, but 3/9 of those predictions were low confidence (evidence level 1) and were not identified by our workflow/CRISPR-CasTyper, one had an evidence level of three and was not detected by our workflow/CRISPR-CasTyper, and five had an evidence level of four and were detected both by our workflow and CRISPR-CasTyper. CRISPR-CasTyper also predicted a variety of CRISPR-Cas types and subtypes from this six-CRISPR-encoding viral element, which included the annotation of putative effector nucleases (Figure 4.3). Three of the megaphages described in Chapter 3 (678, 641, and 508 kbp sizes) were predicted to encode CRISPR systems (Figure 4.4). Two of these megaphages exceed the ~620 kbp genome size of the previously largest circularized CRISPR-encoding phage (Al-Shayeb et al. 2022).

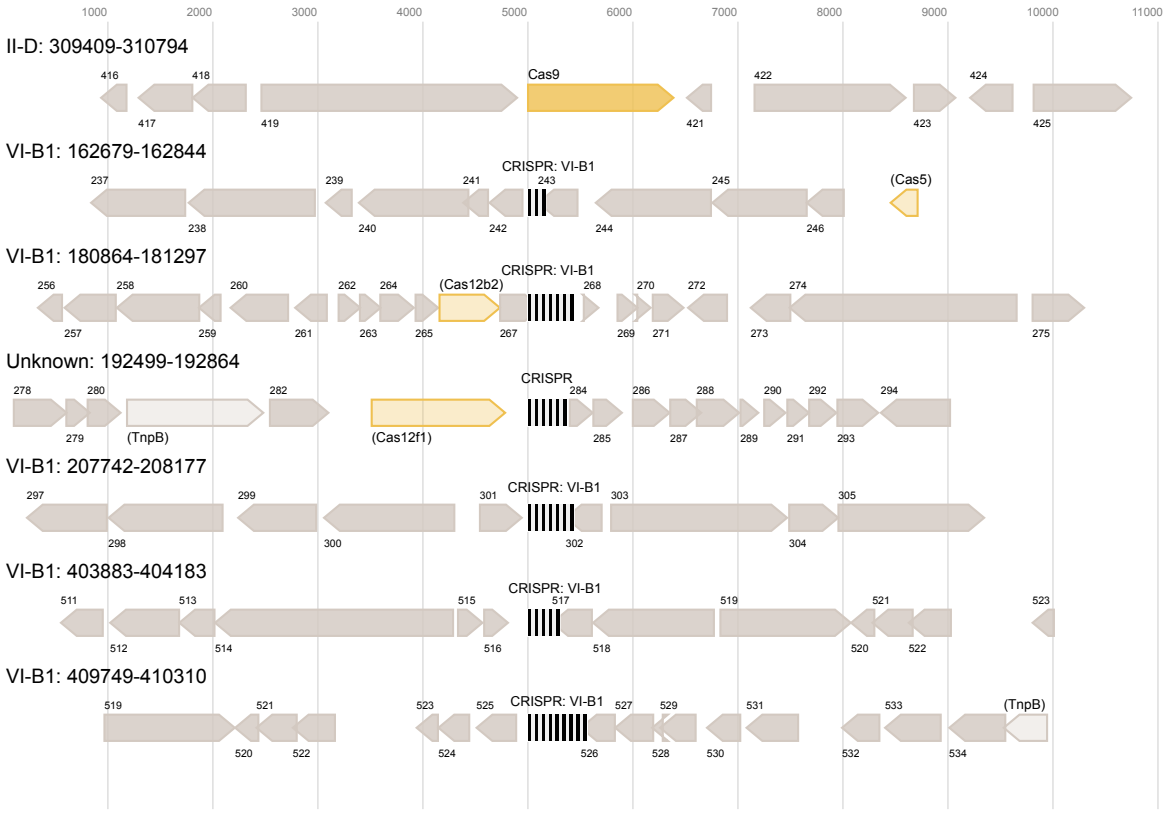


Figure 4.3: CRISPR array and operons from a six-CRISPR-encoding viral element. The most closely related Cas subtype classification is noted in each case, followed by the coordinates of each CRISPR array on the jumbo phage genome. ORFs proximal to predicted CRISPR arrays are numbered in order of occurrence within a single genetic element. In some cases, CRISPR arrays/CRISPR-Cas operons and their surrounding regions overlap with one another, resulting in ORFs or CRISPR arrays that are depicted more than once in the figure. Cas protein interference/effector modules are noted in yellow and CRISPR arrays are denoted by clusters of vertical black lines. Unknown genes are dark grey, and putative TnpB homologs are light grey. Cas genes with more speculative annotations (Russel et al. 2020) have reduced opacity and parentheses around the predicted protein name. This image was generated using CRISPR-CasTyper and processed using Affinity Designer.

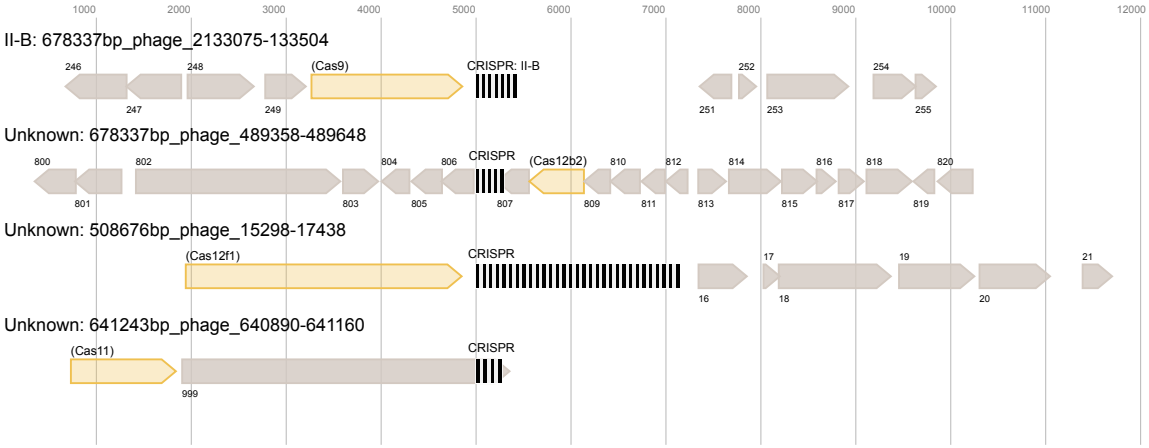


Figure 4.4: CRISPR array and operons from all CRISPR-encoding megaphages. The most closely related Cas subtype classification is noted in each case, followed by the coordinates of each CRISPR array on each megaphage genome. ORFs proximal to predicted CRISPR arrays are numbered in order of occurrence within a single genetic element. In some cases, CRISPR arrays/CRISPR-Cas operons and their surrounding regions overlap with one another, resulting in ORFs or CRISPR arrays that are depicted more than once in the figure. Cas protein interference/effector modules are noted in yellow and CRISPR arrays are denoted by clusters of vertical black lines. Unknown genes are dark grey. All predicted interference/effector modules were considered speculative annotations (Russel et al. 2020) and are indicated by reduced opacity and parentheses around the predicted protein name. This image was generated using CRISPR-CasTyper and processed using Affinity Designer.

Virally encoded Class II effector nucleases

To identify putative Class II effector nucleases, a dataset of 1,494 ORFs ≥300 and ≤800 amino acids was identified from the 10 kbp regions upstream or downstream of the 519 putative virally encoded CRISPR arrays, including any potentially encoded on non-viral MGEs. DNA-targeting class II effectors were predicted from these ORFs using profile HMMs and blastp searches. Type II nuclease HMM searches identified six relatively short (418-686 aa) ORFs which best hit to IscB in the NCBI nr database. Previous work has suggested that IscB is the predicted ancestor of the Cas9 protein and is rarely found proximal to CRISPR arrays (Altae-Tran et al. 2021). Cas9

proteins range in size between 1000-1400 aa (Pausch et al. 2020), while IscB proteins are typically 400 aa in length (Altae-Tran et al. 2021). As the predicted ORFs demonstrated unexpected characteristics for both IscB and Cas9, they were manually examined for the presence of tri-split RuvC motifs and an HNH motif, both of which are characteristic of type II systems (Makarova et al. 2020). Partial HNH motifs were identified in all six sequences (Figure C.1), as was weak similarity to RuvC motifs, but none of these ORFs contained all three expected RuvC motifs. The detected proteins depicted many deletions compared to Cas9 (Figure C.2). 5/6 of these ORFs were similar to reference IscB proteins in length and sequence composition and considered to be IscB variants. The last sequence (putative II-Z/Y, Figure C.1) was, at 638 aa, considerably longer than reference IscB sequences. Based on its sequence length, proximity to CRISPR arrays, and similarity to Cas9, this protein is predicted to be a variant of IscB and more similar to the Type II subtypes Z and Y which were recently identified in bacteriophages (Al-Shayeb et al. 2022). Sequences of these subtypes from the cited study are not available for comparison at this time as the sequences have not been deposited in public databases and have not been made available after requests to the authors.

For type V effector nucleases, a total of 161 putative type V nuclease ORFs were identified after dereplication. These ORFs were placed in a reference tree that included known type V nucleases and TnpB sequences (Figure 4.5). We observed no virally encoded nucleases that grouped with Cas12 subtypes a, b1, b2, c, d, e, f1, g, h, and i, likely because the majority of these protein families, with the exception of f1, b2, and g, have an average length that exceeds the 800 aa limit set for our ORFs of interest. As noted in the methods section, this 800 aa limit was set to exclude Cas protein effectors that were not streamlined in size, i.e., not smaller in size relative to the currently most streamlined Cas effector, Cas12j/Cas12Φ, which is ~800 aa in size

(Al-Shayeb et al. 2020; Pausch et al. 2020). We observed multiple nucleases that associated on the tree with other subtypes of Cas12 (i.e., f2 (1), f3 (1), j (1), and m (24)) and Cas14 (e (1), i (2), j (7), and k (9)), each of which average less than 800 aa in size per subtype. We also observed two predicted TnpB sequences encoded by viruses. These putative TnpB sequences have all three RuvC motifs of TnpB and are 401aa and 433aa in length, respectively, in line with TnpB lengths of 350-550 aa (Nety et al. 2023). These sequences are both encoded proximal to CRISPR arrays, consistent with the hypothesis that TnpB associated with CRISPR multiple times to give rise to different type V CRISPR-Cas systems, and the observation that TnpB is frequently encoded by phages (Al-Shayeb et al. 2022). In total, 114/454 (25%) of viral elements that encode CRISPR arrays were predicted to encode a type V homolog, consistent with the recently-observed enrichment of these hypercompact systems in phage (Al-Shayeb et al. 2022) compared to their limited presence in bacterial and archaeal genomes (Makarova et al. 2020). Of these 114 viral elements, 10 (8.7%) are predicted to encode 2-3 unique type V effector proteins. These ten viral elements had an average length of 155,173 bp (15,118 – 421,673 bp) with 6 being larger than 120 kbp, whereas the average length of all viruses that encoded at least one type V system was 57,638 bp. One of these ten viral elements encodes three distinct predicted V-U4 nucleases, and all ten of these viral elements was predicted to encode at least one subtype V-U nuclease, or a nuclease that is most closely related to a V-U subtype nuclease. Lastly, we observed multiple nodes/clades on the Type V phylogeny that only include sequences from our study (Figure 4.5), representing additional unexplored novelty within this family that are obvious targets for nuclease activity characterization.

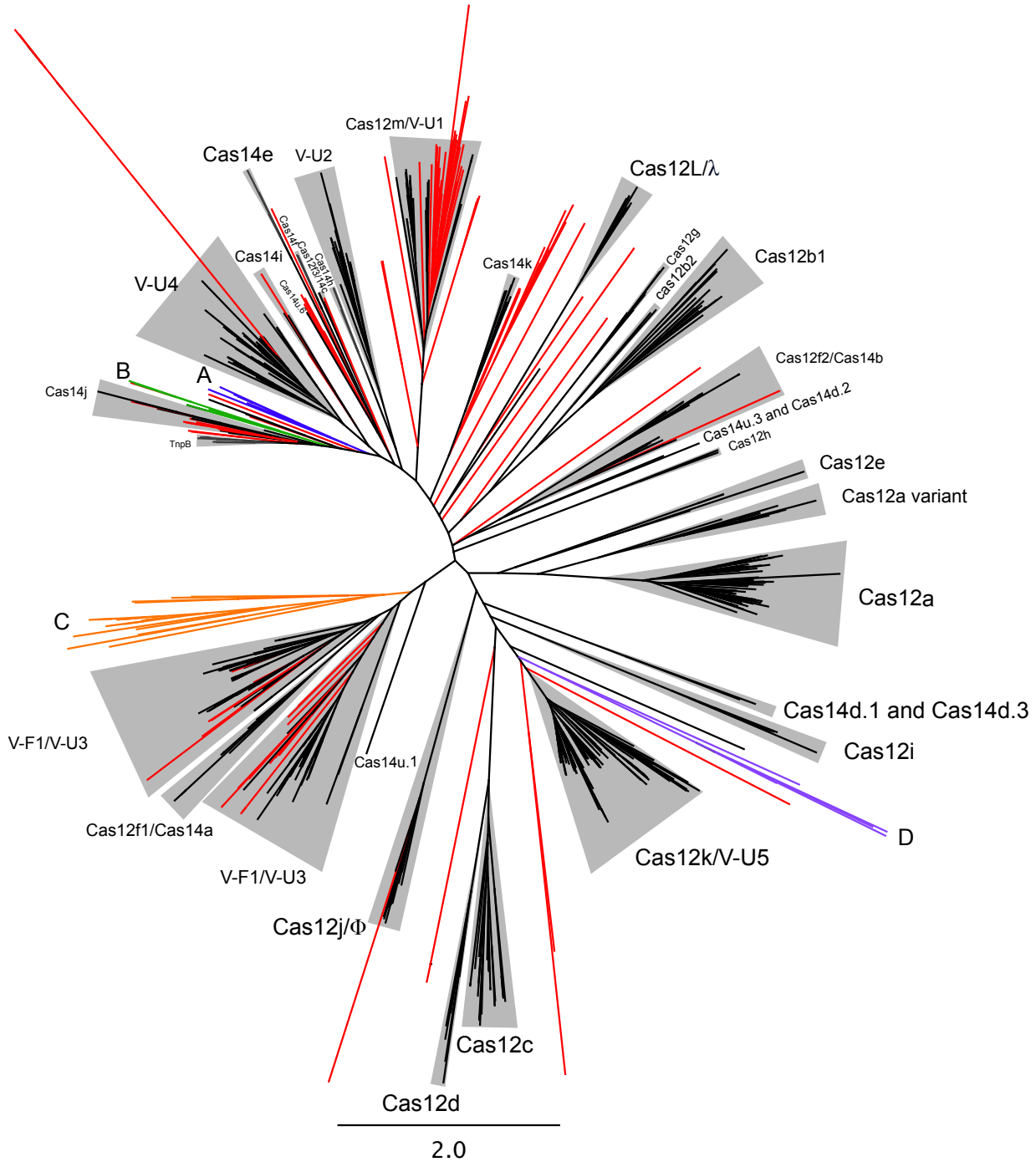


Figure 4.5: Type V CRISPR-associated nucleases, including virally encoded effectors from this study and reference nucleases. Putative nucleases from this study are coloured red, except for clades of further interest denoted by other colours and a letter (A-D). Reference nucleases are in black. Names assigned to clades are those of (putative) effector nucleases, regardless of whether they have been experimentally characterized. In cases where the Type V subtype of a

predicted nuclease has been classified, but the predicted nuclease has not been assigned an official Cas protein name, the predicted effector nucleases (e.g., c2c9; Makarova et al. 2020), is denoted in the tree by its subtype classification (VU-2). Legacy names of nucleases that have been renamed during reclassifications are included for clarity, as are the subtypes of recently characterized nucleases, even if they have received an official Cas protein name (e.g., Cas12m/V-U1) as nuclease/subtype. Singletons and landfill-exclusive clusters not associated with a specific named clade are unlabeled.

Similarity to previously identified virally encoded CRISPR-Cas systems

We observed several instances where our predicted effector sequences were clustered with or were most closely related to previously identified virally encoded CRISPR-Cas systems (i.e., Cas14j (7), Cas14k (9), Cas14i (2), Cas12j (1), and Cas12L (1), Figure 4.5), of which only Cas12j and Cas12L have been experimentally validated for function (Pausch et al. 2020; Al-Shayeb et al. 2022). Each of our relevant sequences was examined for RuvC motifs and assessed for similarity to the virally encoded nuclease it clustered most closely with.

Seven sequences ranging from 373-441aa clustered with previously identified Cas14j sequences (378-451aa; Figure 4.5, 10 o'clock). Our sequences showed high sequence similarity to Cas14j sequences and contained all three RuvC motifs (RuvCI-III). Six additional sequences clustered proximal but distinct to the Cas14J cluster and are described in the next section. We detected two sequences of lengths 402 and 509 aa that clustered with Cas14i (Figure 4.5, 11 o'clock). Both of these sequences as well as reference Cas14i proteins had detectable RuvCI and RuvCIII motifs but very weak, if detectable, RuvCII motifs. We detected one sequence clustering with Cas12j (Figure 4.5, 7 o'clock). This sequence lacked the RuvCIII motif, as did Cas12j6, one of the 10 reference Cas12j nucleases (Al-Shayeb et al. 2020). Our putative Cas12j sequence was also missing key residues in the RuvCII domain, which Cas12j6 lacked entirely, and key residues in the RuvCI motif, all of which were present in Cas12j6. Only three orthologs

of Cas12j(1-3) have been experimentally confirmed for function (Pausch et al. 2020). The comparisons made to Cas12j6 add confidence to the assignment of our query sequence as a Cas12j ortholog, despite its lack of key catalytic residues. Our putative Cas12j is the shortest within the clade, at 346 aa compared to 441 (Cas12j6 from a giant phage (Al-Shayeb et al. 2020)) and 708-813aa for the remaining 9 Cas12j proteins. While our sequence branches within the Cas12j clade, its activity is less confidently predicted based on the aberrant characteristics described above. Notably, our putative Cas12j sequence is encoded by a predicted plasmid, the second time a Cas12j-like protein was identified on a plasmid (Pinilla-Redondo et al. 2022). None of the phages encoding the putative effectors highlighted on Figure 4.5 were predicted to encode the spacer acquisition machinery essential to adding new spacers to CRISPR arrays. This is consistent with previous studies showing that nearly all CRISPR-encoding phages lack adaptation machinery (Al-Shayeb et al. 2020, 2022).

Potentially novel Class II effector nucleases.

All clades within the virally-encoded Type V effector phylogeny that included only our sequences (Figure 4.5) were examined on a case-by-case basis for motifs of interest. We identified 6 sequences (300-399 aa) that grouped proximal, but distinct to Cas14j (Figure 4.5, A/blue at 10 o'clock). These sequences had very weak RuvCI motifs, weak but stronger RuvCII motifs, and only the three longer sequences (385, 398, and 399 aa) had detectable RuvCIII motifs. We identified 6 additional sequences (372-534 aa) that were not associated with a specific Cas protein clade on the tree (Figure 4.5, B/green, 9 o'clock). All 6 had all 3 detectable RuvC motifs (i.e., RuvC[I,II,III]), but only two sequences had strong signatures of all RuvC motifs. We identified a clade of 19 sequences (Figure 4.5, C/orange, 8 o'clock) that, from sequence divergence, was predicted to be a new subtype of Type V effector protein, however,

manual examination of these 19 sequences revealed no RuvC motifs. Five sequences were identified proximal to the Cas12k/VU-5 clade (Figure 4.5, D/purple, 4 o'clock), but only one of these sequences had even a partial RuvC motif, specifically RuvCIII. Numerous singleton effectors were present on the tree, with most containing zero or one identifiable RuvC motif (in all cases, RuvCIII). None of the phages encoding putatively novel Class II effectors were predicted to encode spacer acquisition machinery.

4.4 Discussion

From a total of 106,371 viral elements identified in thesis Chapter 3, 454 viral elements (0.43%) were predicted to encode CRISPR arrays, a subset that has an average length of ~47 kbp compared to an average length of ~13 kbp for viruses lacking CRISPR arrays. These observations are remarkably consistent with a previous study examining virally encoded CRISPR systems in which only 0.4% of phages encoded CRISPR-Cas systems and had an average length of 52 kbp (Al-Shayeb et al. 2022). CRISPR-encoding viruses in this study did not show any infection bias to a particular host taxon nor did they cluster distinctly in viral gene-sharing networks (Figure 4.2), suggesting that the presence of CRISPR in virus genomes is not tied to their taxonomy. Spacers from viral arrays targeted <0.2% of viral elements lacking CRISPR arrays. A total of 228/4,019 spacers (<6%), 196 spacers after dereplication with CD-hit, from viral CRISPR arrays targeted viral elements lacking CRISPR arrays. Given the viral arrays are expected to target viruses that are in competition for their same host populations (Faure et al. 2019; Medvedeva et al. 2019) and that, with some rare exceptions (Malki et al. 2015; Paez-Espino et al. 2016; Dalcin Martins et al. 2018; Chen et al. 2021; George and Hug 2023; Hwang et al. 2023), most viruses are quite host specific, this low percentage makes sense. We would not expect these viral arrays to broadly target other viruses.

Where virally encoded CRISPR spacers did target viral genomes, only the targets of 5/196 unique spacers could be annotated using the Virus Orthologous Groups database, resulting in 9 unique annotations (one spacer had five distinct targets, one spacer had two distinct targets, and two spacers had one target each). These 9 annotations included three collagen-like proteins, three hypothetical/uncharacterized proteins/regions, two structural proteins, and one genome polyprotein. Spacers in CRISPR arrays are selected over time for those that are most effective at blocking viral replication. Previous research has shown that phage genes that encode structural proteins, and proteins that interact with DNA, are among those most often targeted by host CRISPR-Cas systems, likely because mutational escape in these regions leads to fitness costs in viruses (Paez-Espino et al. 2013; George and Hug 2023). This makes one of our structural protein-encoding targets, the Baseplate wedge protein gp10 encoding gene, an ideal target as its product is both a structural protein and also a key determinant in host range in some phages (Sun et al. 2023); however we did not observe any enrichment of spacers targeting this or any of our other annotated regions. The targets for the vast majority of the virally encoded spacers in this study could not be annotated, which suggests the presence of viral dark matter, i.e., regions within viral genomes that have not been annotated but are likely important in preventing viral propagation, evidenced by CRISPR spacer targeting. This is consistent with what we observed in Chapter 2 (Figure 2.2 and Figure A.9), in which spacers targeting hyper-targeted viral elements most often targeted a tail protein encoding gene and a hypothetical coding region. We note that selection over time for targets preventing viral propagation could occur in prokaryote genomes or in MGEs before the CRISPR arrays were horizontally transferred to viral elements, or once in the elements themselves.

A jumbophage with an approximately 421 kbp genome was predicted to encode six CRISPR arrays, five of which were annotated by two state-of-the-art CRISPR prediction tools (Couvin et al. 2018; Russel et al. 2020). According to CRISPRCasTyper, only four of these six CRISPR arrays were proximal (i.e., within 5 kbp upstream or downstream) to any *cas* gene or gene predicted to encode a TnpB-like protein. In total, this jumbophage contains two genes predicted to encode TnpB-like proteins, one gene predicted to encode Cas9, one gene predicted to encode Cas12b2, and a gene predicted to encode Cas12f1 (Figure 4.3)., The predicted *cas9* gene was not proximal to any CRISPR arrays (Figure 4.3). The disjunction of arrays and *cas* genes is potentially due to the guide RNAs from the two orphan CRISPR arrays directing effectors encoded elsewhere in the viral genome, or in the host genome (Shmakov et al. 2020). The orphan arrays could also be explained by the presence of proximal effector genes that are too divergent to be detected by current models of *cas* genes. The potential presence of multiple class II effector encoding genes in a single viral genome, along with TnpB encoding genes, which may themselves be hypercompact DNA targeting systems (Karvelis et al. 2021), could aid these elements against systems that can inhibit one or more of their Cas proteins, such as anti-CRISPR (Bondy-Denomy et al. 2013). Since CRISPR-Cas systems also have roles in processes such as gene expression and signal transduction (Faure et al. 2019a), and viral CRISPR-Cas systems can target genetic systems of the host organism that regulate immunity (Seed et al. 2013) and host gene regulation (Al-Shayeb et al. 2020), it is possible that the expansive CRISPR-Cas repertoire of viruses with large genomes may allow these phage to exert greater control over their host organism during infection, resulting in the creation of a "virocell" (Forterre 2013; Howard-Varona et al. 2020).

The average length of the ten viral elements encoding more than one predicted type V system was ~155 kbp, which is nearly three times larger than the average length of all viral elements that encode at least one type V system (~58 kbp). Combined with the observation that the average length of CRISPR-encoding viruses was more than three times that of viruses that were not predicted to encode CRISPR, this study's results suggest that the evolutionary costs of CRISPR-Cas systems (Vale et al. 2015) may be more manageable for viruses with larger genomes. This manageable cost can be explained by jumbophages undergoing substantial coevolution of their capsid and genomes in which mutations can expand capsid volume which in turn supports genome expansion (Hua et al. 2017). Whether this capsid and genome coevolution extends beyond jumbophages and whether it can drive the acquisition of CRISPR-Cas systems remain open questions.

We detected virally encoded effector proteins that clustered with multiple subtypes of Cas12 (*i.e.,* f2, f3, j, and m), and Cas14 (e, i, j, and k), but most of our virally encoded type V-like effector proteins were located proximal to or with TnpB, the phage encoded systems Cas14j and Cas14k, and type V-U1, V-U5, and V-F1/VU-3 systems. All of these systems are either among the most streamlined type V or the progenitor (i.e., TnpB) systems. The subtype V-U systems are predicted as intermediates between TnpB and full-fledged type V effectors (Makarova et al. 2020). It is thus possible that viruses act as vectors for these intermediate systems, allowing an evolutionary path distinct from host-encoded systems, and with the potential to develop unique systems nearly exclusive to MGEs such as phages (Al-Shayeb et al. 2020, 2022), and to MGE-associated systems such as Tn7-like transposons (Strecker et al. 2019; Faure et al. 2019). We identified 18 clades comprised of sequences exclusively from our study that were substantially divergent from previously described Type V effectors (Figure 4.5). Some

of these proteins may represent novel type V nucleases, especially the six for which we were able to identify all RuvC motifs. The full or partial absence of identified RuvC motifs in the other proteins could be due to substantial divergence of these motifs from those in reference enzymes, as is the case for phage-encoded Cas12L (Al-Shayeb et al. 2022), or absence of canonical catalytic sites within the RuvC domain, as is the case for Cas12m (Wu et al. 2022). Absent or divergent RuvC motifs do not necessarily indicate loss of function or alternate function for these Type V effector-related proteins. Even if these catalytic motifs are entirely absent, these proteins may still demonstrate interference activity through binding to, but not cleaving, their target nucleic acid, a function that has been seen for dCas9 (Gilbert et al. 2013) and Cas12m (Wu et al. 2022). Both dCas9 and Cas12m have been repurposed as biotechnological tools, as alternate interactions with nucleic acids beyond targeted cleavage have downstream uses. For the putatively novel effector nucleases in this study to be considered as biotechnological tools, they would need to be validated *in-vivo* for interference and/or cleavage activities.

In summary, we identified over 400 CRISPR-encoding viral elements from samples taken from three municipal landfills across North America. These virus-encoded arrays were predicted to target a small fraction of other viral elements in interviral conflicts, and some of the genomic regions targeted encode predicted structural proteins essential to virion formation and host specificity. In some cases, these virally encoded CRISPRs were flanked by progenitor or full-fledged type V effector nucleases, including new branches of this protein family identified for the first time within this study. While there are several hundred papers pertaining to the discovery and characterization of CRISPR-Cas systems in literature, to our knowledge, the study described in this thesis chapter is one of very few that aimed to identify virally encoded

CRISPR-Cas systems from metagenomic data. All of the putative type V effectors we identified in this study had a maximum length of 800 amino acids. If proven functional, these nucleases would help circumvent the size restrictions of certain viral delivery vectors such as those of Adeno-associated Viruses (AAVs) (Asmamaw Mengstie 2022).

Novel, and rare Cas proteins could theoretically result in reduced immunogenicity if introduced into mammalian systems such as mice and humans. Currently, the two most commonly used Cas9 orthologs, from *Streptococcus pyogenes* (SpCas9) and *Staphylococcus aureus* (SaCas9), trigger immune responses in 78% (Charlesworth et al. 2019) and 58-96% (Charlesworth et al. 2019; Wagner et al. 2019) of individuals examined, respectively. Considering the novelty of the systems identified in this study, and the non-pathogenic, exclusively microbial interactions they are involved in, these novel predicted effectors may elicit lowered immune responses. Finally, we posit that deeper examinations of virally encoded CRISPR-Cas systems across more environments will identify new CRISPR-Cas systems and improve our understanding of the ecological relevance and evolution of these ubiquitous and versatile genetic systems.

Chapter 5 Research summary, study limitations, and field recommendations

5.1 Research summary

Inspired by a relatively low representation of prokaryotic viruses from engineered environments (Paez-Espino et al. 2016b) and the absence of prior metaviromic examinations of municipal landfills, the work described in this thesis examined the diversity, host interactions, metabolic potential, and CRISPR-Cas systems of prokaryotic viruses from municipal waste sites. From the work described in Chapter 2, we noted that viral elements from a municipal waste site in Southern Ontario are distinct relative to reference viruses, however, limited variation in viral diversity between sites within the landfill was observed. Some viral elements were hyper-targeted by host CRISPR-Cas machinery, suggestive of both the importance of these specific viruses in this system and a strong coevolutionary dynamic between them and their predicted host populations. A very small percentage of viruses (<0.5%) encoded their own CRISPR systems, which were predicted to be involved in interviral conflicts. Lastly, we provided one of the first computational examinations of putative cross-phylum infecting viruses, providing a framework by which future researchers can assess the legitimacy of these observations in cases where *in vitro* substantiation is not possible. This work was the first metaviromic examination of a municipal waste site, reporting multiple rare observations from a single system, suggesting further examination of municipal waste sites could lead to the discovery of additional novel viruses and host interactions.

In Chapter 3, we extended our analyses to include two additional municipal waste sites, one from California and one from the Northeastern United States. Examining the viral diversity across all three of our sampled landfills showed that there is strong conservation of the landfill

virome, but this virome also has some overlap with the wastewater virome. The conserved diversity of viruses between landfills was further confirmed with the detection of spacers from host CRISPR arrays that had strong cross-targeting between geographically distinct landfills. As in thesis Chapter 2 with our Southern Ontario landfill, hyper-targeted viral elements were also observed in both the Northeastern United States landfill and the California landfill. The use of state-of-the-art viral detection tools (Nayfach et al. 2021; Guo et al. 2021; Kieft et al. 2022) allowed us to identify and curate seven megaphage genomes, including a 678 kbp one that is the third largest bacteriophage genome identified to date. We examined the Auxiliary Metabolic Genes (AMGs) encoded by all predicted viruses from our study, and after manually curating AMGs of interest, we found that some viruses in our study encode AMGs that are rare or have not been previously reported in literature. Notably, these genes include those involved in contaminant and methane metabolism, both of which are enriched and essential functions in landfills. Lastly, after an even more stringent level of curation relative to that used in thesis Chapter 2, we still observed viral elements predicted to infect across distinct phyla in all three of our landfill datasets. This evidence, along with recent *in vitro* work confirming environmental viruses infecting across broad ranges (Hwang et al. 2023), including at the phylum and domain level, provide credence to our inference of broad infection potential of prokaryotic viruses in Chapter 2, and indicate this broad infection potential may be widely distributed within natural and engineered environments.

Finally, in Chapter 4, inspired both by recent discoveries of virally encoded CRISPR-Cas systems (Al-Shayeb et al. 2020, 2022) and the interviral conflicts we detected in Chapter 2, we screened our metagenomes for virally encoded CRISPR-Cas systems. We focused our analyses on Class II CRISPR-Cas systems as these systems are most amenable to adaptation as

biotechnological tools. We noted that some viral elements encode multiple CRISPR-Cas systems that include various subtypes of streamlined CRISPR-Cas systems. Some of these systems included predicted effector-protein-encoding genes not represented in our reference set. These putatively novel effector ORFs were manually examined for key catalytic motifs and for accessory *cas* genes. We identified some groups that were putatively novel and had detectable catalytic residues, but we also identified other groups that while putatively novel, lacked key regions considered to be necessary for catalysis/target cleavage. While these regions may have been absent, or unresolvable with our methods due to high divergence, these proteins could still be repurposed as biotechnological tools, as has been the case with other Cas proteins lacking key catalytic residues (Gilbert et al. 2013; Wu et al. 2022).

In summary, municipal waste sites harbor distinct viromes, and these viruses have rare or unique interactions with their hosts. Some of these viruses are predicted to broadly infect across the bacterial domain, which is consistent with recent, but limited literature. Increased awareness around broad host-range viruses in natural environments is relevant for phage therapy applications as scientists are turning to environmental phage isolates for use in antibiotic resistant bacterial infections (Alharbi and Ziadi 2021). These environmental isolates may have a host range that well-exceeds the target pathogen, leading to off-target effects. The addition of computational host prediction to conventional *in vitro* host range prediction pipelines may aid in preventing off-target effects.

Some of the bacteriophage genomes uncovered in this work are among the largest identified to date, sharing common evolutionary origins with the largest phage genomes described in literature. The discovery of seven of these phages from only 27 landfill metagenomes is in stark contrast to limited detection of these viruses even in studies that explore

several hundred metagenomes (Al-Shayeb et al. 2020; Chen and Banfield 2023). Landfills may be a hot spot for discovering more of these "megaphages", which may aid our understanding of the evolution of these phages and expand our knowledge of microbial genes that have been co-opted by giant phages. The discovery of AMGs in landfills that are relevant to landfill function but have rarely been observed elsewhere, supports that genes mobilized within a microbial community are influenced by the selective pressures imposed by the environment. Lastly, the discovery of putatively novel Cas proteins and CRISPR-Cas systems that are virally encoded suggests the need for further examination of both viral genomes and understudied environments in order to increase the rate of biotechnological discovery.

5.2 Study limitations and field recommendations

The sample collection, DNA extraction, sequencing, and computational workflows employed in this thesis each have limitations. I will begin by describing the limitations of our study, followed by how we tried to address them, and lastly, suggestions I have for future studies. Beginning first with sample collection, 0.1 μm or 0.2 μm filters were the final filters used for filtration and DNA was extracted from these filters. Since many prokaryotic virions are smaller than the pore size of these filters, these viral particles would be lost in filtrate that was not ultracentrifuged or concentrated for viral particle enrichment, which in some cases was due to a restriction on moving leachate off site. This means that the viruses detected in our study are not representative of the true breadth of viruses from our samples. To address this, for our samples from the Northeastern United States landfill, for which we were granted permission to move leachate off site, we filtered our leachate through 0.2 μm filters and subsequently through 0.01 μm filters. DNA was extracted from both filters. While this may provide a more complete snapshot of the viral community, DNA extractions from the 0.01 μm filters did not yield sufficient DNA for

sequencing. Our group is in the process of using Multiple Displacement Amplification of the DNA from these 0.01 μm filters. While that data is not a part of this thesis, it will provide a reference of the viral diversity we can capture in our sites with the use of filters with an even smaller pore size targeting viral capsids. To improve retrieval of viral particles and their DNA on relative large pore size filters in future work, we suggest the use of iron chloride flocculation (Poulos et al. 2018), although this technique's efficiency varies by the sample origin (Langenfeld et al. 2021) and may need substantial troubleshooting to be employed in the metal-rich leachate environment.

Once the DNA from our samples was extracted, it was sent for paired-end sequencing of 150 bp reads using the Illumina HiSeq platform (Illumina, San Diego, CA). These relatively short reads were computationally assembled. Metagenomic assembly is particularly problematic for viral genomes (Sutton et al. 2019), often resulting in fragmented assemblies, owing to the mosaicism and repetitive nature of viral genomes. It is possible that many viral elements in this study do not represent complete viral genomes, but rather, parts of a whole. CRISPR array assembly is also a challenge due to the repetitive nature of these elements and their natural variation between populations (Skennerton et al. 2013). This could mean that our detected CRISPR arrays are often incomplete or improperly assembled, blurring or misrepresenting the true history of infection recapitulated by these genetic elements. To combat this issue, we have two suggestions: employ CRISPR array detection tools that leverage machine learning for an added level of curation and false positive reduction (Mitrofanov et al. 2021), and detect CRISPR arrays with multiple tools and retain the arrays agreed upon by some predetermined fraction of those tools. We note that the latter suggestion may not be scalable to large metagenomes, in this case, this approach can be reserved for CRISPR arrays that are implicated in rare observations,

such as targeting viruses predicted to infect across broad taxonomic levels. Tools like the CRISPR comparison toolkit (CCTK) (Collins and Whitaker 2023) can be used to compare between multiple CRISPR array representations predicted by different tools. Poor assembly of CRISPR arrays may also compound the issues of viral assembly leading to viruses encoding CRISPR arrays being fragmented more often, thus leading to lower estimates of their average length. To combat assembly issues that lead to fragmented viral elements, we employed vRhyme, a tool built in our collaborator Dr. Karthik Anantharaman's lab, for the work in thesis Chapters three and four. vRhyme specifically groups viral assemblies (i.e., contigs or scaffolds) into distinct clusters based on certain parameters, and then provides quality estimates of the resulting Viral Metagenome Assembled Genomes (vMAGs). This aided our analyses in a variety of ways. For example, in thesis Chapter 2, we did not use vRhyme, and we saw evidence of four viral elements that were hyper-targeted by hosts from both temporal samples of our Southern Ontario landfill data. In thesis Chapter 3, we used vRhyme with the Southern Ontario landfill data included in the datasets examined. Notably, two of the hyper-targeted viral elements from Southern Ontario were clustered into the same vMAG. Therefore, vRhyme aided in clarifying whether any of the viral elements predicted to infect a host MAG are from the same population of viruses. While these utilities were not examined in this thesis, vRhyme may also aid assessments of AMGs, i.e., some viral elements may encode AMGs responsible for only part of a pathway, however, the vMAG those viral elements are included in may include other viral elements that encode AMGs that improve the completion of the metabolic pathway. Lastly, vRhyme may aid our ability to understand whether a single viral population encodes multiple CRISPR-Cas systems that are spread across distinct viral elements due to fragmented sequencing and assembly.

While vRhyme helps address challenges post assembly, we propose other solutions that address the root cause of viral genome assembly issues: using viral-specific assembly algorithms and/or using long-read sequencing. Using viral-specific assembly algorithms such as Metaviral SPAdes allow researchers to circumvent some of the challenges of viral assembly, leading to improved detection of viruses in standard pipelines (Antipov et al. 2020). Using long-read technologies has been shown to increase recovery of viral genomes, however, the accuracy of the viral sequences is compromised, most notably for sequences with low coverage (Kolmogorov et al. 2020). While long-read sequencing technologies are continually improving, we suggest the combined use of short-read and long-read sequencing to leverage the accuracy of the former and the improvements in assembly size and quality of the latter.

Overall, the continued examination of viromes as a core component of microbial ecology studies is likely to further expand our knowledge of the diversity, ecological interactions, and evolutionary history of prokaryotic viruses, and inspire new applications for viruses and their genetic systems. In landfills specifically, it is clear additional diversity, and important host dynamics are likely to be identified from surveying more than three of the tens of thousands of landfills globally. Exciting times lie ahead in uncovering the vast reservoir of genetic diversity harbored by prokaryotic viruses in natural environments.

References

Abbaszadegan, M., and Pepper, I., L. 1993. Detection of Enteroviruses in Groundwater with the Polymerase Chain Reaction. Appl. Environ. Microbiol. **59**: 7.

Ahlgren, N.A., Ren, J., Lu, Y.Y., Fuhrman, J.A., and Sun, F. 2017. Alignment-free d2* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Res. **45**(1): 39–53. doi:10.1093/nar/gkw1002.

Alharbi, N.M., and Ziadi, M.M. 2021. Wastewater as a fertility source for novel bacteriophages against multi-drug resistant bacteria. Saudi J. Biol. Sci. **28**(8): 4358–4364. doi:10.1016/j.sjbs.2021.04.025.

Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Loman, N.J., Andersson, A.F., and Quince, C. 2014. CONCOCT: clustering contigs on coverage and composition. Nat. Methods **11**: 1144–1146.

Al-Shayeb, B., Sachdeva, R., Chen, L.-X., Ward, F., Munk, P., Devoto, A., Castelle, C.J., Olm, M.R., Bouma-Gregson, K., Amano, Y., He, C., Méheust, R., Brooks, B., Thomas, A., Lavy, A., Matheus-Carnevali, P., Sun, C., Goltsman, D.S.A., Borton, M.A., Sharrar, A., Jaffe, A.L., Nelson, T.C., Kantor, R., Keren, R., Lane, K.R., Farag, I.F., Lei, S., Finstad, K., Amundson, R., Anantharaman, K., Zhou, J., Probst, A.J., Power, M.E., Tringe, S.G., Li, W.-J., Wrighton, K., Harrison, S., Morowitz, M., Relman, D.A., Doudna, J.A., Lehours, A.-C., Warren, L., Cate, J.H.D., Santini, J.M., and Banfield, J.F. 2020. Clades of huge phages from across Earth's ecosystems. Nature **578**(7795): 425–431. doi:10.1038/s41586-020-2007-4.

Al-Shayeb, B., Skopintsev, P., Soczek, K.M., Stahl, E.C., Li, Z., Groover, E., Smock, D., Eggers, A.R., Pausch, P., Cress, B.F., Huang, C.J., Staskawicz, B., Savage, D.F., Jacobsen, S.E., Banfield, J.F., and Doudna, J.A. 2022. Diverse virus-encoded CRISPR-Cas systems include streamlined genome editors. Cell **185**(24): 4574-4586.e16. doi:10.1016/j.cell.2022.10.020.

Altae-Tran, H., Kannan, S., Demircioglu, F.E., Oshiro, R., Nety, S.P., McKay, L.J., Dlakić, M., Inskeep, W.P., Makarova, K.S., Macrae, R.K., Koonin, E.V., and Zhang, F. 2021. The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases. Science **374**(6563): 57–65. doi:10.1126/science.abj6856.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. J. Mol. Biol. **215**(3): 403–410.

Anantharaman, K., Duhaime, M.B., Breier, J.A., Wendt, K.A., Toner, B.M., and Dick, G.J. 2014. Sulfur Oxidation Genes in Diverse Deep-Sea Viruses. Science **344**(6185): 757–760. doi:10.1126/science.1252229.

Anderson, I.C., Campbell, C.D., and Prosser, J.I. 2003. Potential bias of fungal 18S rDNA and internal transcribed spacer polymerase chain reaction primers for estimating fungal biodiversity in soil. Environ. Microbiol. **5**(1): 36–47.

Andersson, A.F., and Banfield, J.F. 2008. Virus Population Dynamics and Acquired Virus Resistance in Natural Microbial Communities. Science **320**(5879): 1047–1050. doi:10.1126/science.1157358.

Ansari, S.A., Farrah, S.R., and Chaudhry, R.G. 1992. Presence of Human Immunodeficiency Virus Nucleic Acids in Wastewater and Their Detection by Polymerase Chain Reaction. Appl. Environ. Microbiol. **58**: 7.

Antipov, D., Raiko, M., Lapidus, A., and Pevzner, P.A. 2020. MetaviralSPAdes: assembly of viruses from metagenomic data. Bioinformatics **36**(14): 4126–4129. doi:10.1093/bioinformatics/btaa490.

Arkin, A.P., Cottingham, R.W., Henry, C.S., Harris, N.L., Stevens, R.L., Maslov, S., Dehal, P., Ware, D., Perez, F., Canon, S., Sneddon, M.W., Henderson, M.L., Riehl, W.J., Murphy-Olson, D., Chan, S.Y., Kamimura, R.T., Kumari, S., Drake, M.M., Brettin, T.S., Glass, E.M., Chivian, D., Gunter, D., Weston, D.J., Allen, B.H., Baumohl, J., Best, A.A., Bowen, B., Brenner, S.E., Bun, C.C., Chandonia, J.-M., Chia, J.-M., Colasanti, R., Conrad, N., Davis, J.J., Davison, B.H., DeJongh, M., Devoid, S., Dietrich, E., Dubchak, I., Edirisinghe, J.N., Fang, G., Faria, J.P., Frybarger, P.M., Gerlach, W., Gerstein, M., Greiner, A., Gurtowski, J., Haun, H.L., He, F., Jain, R., Joachimiak, M.P., Keegan, K.P., Kondo, S., Kumar, V., Land, M.L., Meyer, F., Mills, M., Novichkov, P.S., Oh, T., Olsen, G.J., Olson, R., Parrello, B., Pasternak, S., Pearson, E., Poon, S.S., Price, G.A., Ramakrishnan, S., Ranjan, P., Ronald, P.C., Schatz, M.C., Seaver, S.M.D., Shukla, M., Sutormin, R.A., Syed, M.H., Thomason, J., Tintle, N.L., Wang, D., Xia, F., Yoo, H., Yoo, S., and Yu, D. 2018. KBase: The United States Department of Energy Systems Biology Knowledgebase. Nat. Biotechnol. **36**(7): 566–569. doi:10.1038/nbt.4163.

Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D.S. 2016. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. **44**(W1): W16–W21. doi:10.1093/nar/gkw387.

Asmamaw Mengstie, M. 2022. Viral Vectors for the in Vivo Delivery of CRISPR Components: Advances and Challenges. Front. bioeng. biotechnol. **10**. Available from https://www.frontiersin.org/articles/10.3389/fbioe.2022.895713 [accessed 22 May 2023].

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., and Pevzner, P.A. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. **19**(5): 455–477. doi:10.1089/cmb.2012.0021.

Barrangou, R., and Doudna, J.A. 2016. Applications of CRISPR technologies in research and beyond. Nature Biotechnology: 933–941. doi:10.1038/nbt.3659.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. 2007. CRISPR Provides Acquired Resistance against Viruses in Prokaryotes. Science **315**(5819): 1709–1712.

Bergmann, G.T., Bates, S.T., Eilers, K.G., Lauber, C.L., Caporaso, J.G., Walters, W.A., Knight, R., and Fierer, N. 2011. The under-recognized dominance of Verrucomicrobia in soil bacterial communities. Soil Biol. Biochem. **43**(7): 1450–1455. doi:10.1016/j.soilbio.2011.03.012.

Bezuidt, O.K.I., Lebre, P.H., Pierneef, R., León-Sobrino, C., Adriaenssens, E.M., Cowan, D.A., Van de Peer, Y., and Makhalanyane, T.P. 2020. Phages Actively Challenge Niche Communities in Antarctic Soils. mSystems **5**(3). doi:10.1128/mSystems.00234-20.

Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J.H., Roux, S., Adriaenssens, E.M., Brister, J.R., Kropinski, A.M., Krupovic, M., Lavigne, R., Turner, D., and Sullivan, M.B. 2019. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat. Biotechnol. **37**(6): 632–639. doi:10.1038/s41587-019-0100-8.

Bolduc, B., Shaughnessy, D.P., Wolf, Y.I., Koonin, E.V., Roberto, F.F., and Young, M. 2012. Identification of Novel Positive-Strand RNA Viruses by Metagenomic Analysis of Archaea-Dominated Yellowstone Hot Springs. Virol. J. **86**(10): 5562–5573. doi:10.1128/JVI.07196-11.

Bolotin, A., Quinquis, B., Sorokin, A., and Ehrlich, D.S. 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology **151**(8): 2551–2561. doi:10.1099/mic.0.28048-0.

Bondy-Denomy, J., Pawluk, A., Maxwell, K.L., and Davidson, A.R. 2013. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. Nature **493**(7432): 429–432. doi:10.1038/nature11723.

Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., Tringe, S.G., Ivanova, N.N., Copeland, A., Clum, A., Becraft, E.D., Malmstrom, R.R., Birren, B., Podar, M., Bork, P., Weinstock, G.M., Garrity, G.M., Dodsworth, J.A., Yooseph, S., Sutton, G., Glöckner, F.O., Gilbert, J.A., Nelson, W.C., Hallam, S.J., Jungbluth, S.P., Ettema, T.J.G., Tighe, S., Konstantinidis, K.T., Liu, W.-T., Baker, B.J., Rattei, T., Eisen, J.A., Hedlund, B., McMahon, K.D., Fierer, N., Knight, R., Finn, R., Cochrane, G., Karsch-Mizrachi, I., Tyson, G.W., Rinke, C., Lapidus, A., Meyer, F., Yilmaz, P., Parks, D.H., Murat Eren, A., Schriml, L., Banfield, J.F., Hugenholtz, P., and Woyke, T. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled

genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. **35**(8): 725–731. doi:10.1038/nbt.3893.

Breitbart, M., Thompson, L.R., Suttle, C.A., and Sullivan, M.B. 2007. Exploring the Vast Diversity of Marine Viruses. Oceanography **20**(2): 135–139.

Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L., Dickman, M.J., Makarova, K.S., Koonin, E.V., and van der Oost, J. 2008. Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. Science **321**(5891): 960–964. doi:10.1126/science.1159689.

Burstein, D., Harrington, L.B., Strutt, S.C., Probst, A.J., Anantharaman, K., Thomas, B.C., Doudna, J.A., and Banfield, J.F. 2017. New CRISPR–Cas systems from uncultivated microbes. Nature **542**(7640): 237–241. doi:10.1038/nature21059.

Camargo, A.P., Nayfach, S., Chen, I.-M.A., Palaniappan, K., Ratner, A., Chu, K., Ritter, S.J., Reddy, T.B.K., Mukherjee, S., Schulz, F., Call, L., Neches, R.Y., Woyke, T., Ivanova, N.N., Eloe-Fadrosh, E.A., Kyrpides, N.C., and Roux, S. 2023. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. Nucleic Acids Res. **51**(D1): D733–D743. doi:10.1093/nar/gkac1037.

Charlesworth, C.T., Deshpande, P.S., Dever, D.P., Camarena, J., Lemgart, V.T., Cromer, M.K., Vakulskas, C.A., Collingwood, M.A., Zhang, L., Bode, N.M., Behlke, M.A., Dejene, B., Cieniewicz, B., Romano, R., Lesch, B.J., Gomez-Ospina, N., Mantri, S., Pavel-Dinu, M., Weinberg, K.I., and Porteus, M.H. 2019. Identification of preexisting adaptive immunity to Cas9 proteins in humans. Nat. Med. **25**(2): 249–254. doi:10.1038/s41591-018-0326-x.

Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics **36**: 1925–1927. doi:10.1093/bioinformatics/btz848.

Chen, F., Suttle, C.A., and Short, S.M. 1996. Genetic Diversity in Marine Algal Virus Communities as Revealed by Sequence Analysis of DNA Polymerase Genes. Appl. Environ. Microbiol. **62**: 6.

Chen, J., Quiles-Puchalt, N., Chiang, Y.N., Bacigalupe, R., Fillol-Salom, A., Chee, M.S.J., Fitzgerald, J.R., and Penadés, J.R. 2018. Genome hypermobility by lateral transduction. Science **362**(6411): 207–212. doi:10.1126/science.aat5867.

Chen, L., and Banfield, J.F. 2023. COBRA improves the quality of viral genomes assembled from metagenomes. bioRxiv. doi:https://doi.org/10.1101/2023.05.30.542503.

Chen, Y., Wang, Y., Paez-Espino, D., Polz, M.F., and Zhang, T. 2021. Prokaryotic viruses impact functional microorganisms in nutrient removal and carbon cycle in wastewater treatment plants. Nat. Commun. **12**(1): 5398. doi:10.1038/s41467-021-25678-1.

Chénard, C., Wirth, J.F., and Suttle, C.A. 2016. Viruses Infecting a Freshwater Filamentous Cyanobacterium (*Nostoc* sp.) Encode a Functional CRISPR Array and a Proteobacterial DNA Polymerase B. mBio **7**(3). doi:10.1128/mBio.00667-16.

Chiang, Y.N., Penadés, J.R., and Chen, J. 2019. Genetic transduction by phages and chromosomal islands: The new and noncanonical. PLOS Pathog. **15**(8): e1007878. doi:10.1371/journal.ppat.1007878.

Collias, D., and Beisel, C.L. 2021. CRISPR technologies and the search for the PAM-free nuclease. Nat. Commun. **12**(1): 555. doi:10.1038/s41467-020-20633-y.

Collins, A.J., and Whitaker, R.J. 2023. CRISPR Comparison Toolkit: Rapid Identification, Visualization, and Analysis of CRISPR Array Diversity. CRISPR J. doi:10.1089/crispr.2022.0080.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., and Zhang, F. 2013. Multiplex Genome Engineering Using CRISPR/Cas Systems. Science **339**(6121): 819–823. doi:10.1126/science.1231143.

Correa, A.M.S., Howard-Varona, C., Coy, S.R., Buchan, A., Sullivan, M.B., and Weitz, J.S. 2021. Revisiting the rules of life for viruses of microorganisms. Nat. Rev. Microbiol. doi:10.1038/s41579-021-00530-x.

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E.P.C., Vergnaud, G., Gautheret, D., and Pourcel, C. 2018. CRISPRCasFinder, an update of CRISPRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res. **46**(W1): W246–W251. doi:10.1093/nar/gky425.

Dalcin Martins, P., Danczak, R.E., Roux, S., Frank, J., Borton, M.A., Wolfe, R.A., Burris, M.N., and Wilkins, M.J. 2018. Viral and metabolic controls on high rates of microbial sulfur and carbon cycling in wetland ecosystems. Microbiome **6**(1): 138. doi:10.1186/s40168-018-0522-4.

Daly, R.A., Borton, M.A., Wilkins, M.J., Hoyt, D.W., Kountz, D.J., Wolfe, R.A., Welch, S.A., Marcus, D.N., Trexler, R.V., MacRae, J.D., Krzycki, J.A., Cole, D.R., Mouser, P.J., and Wrighton, K.C. 2016. Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales. Nat. Microbiol. **1**(10): 16146. doi:10.1038/nmicrobiol.2016.146.

Darriba, D., Posada, D., Kozlov, A.M., Stamatakis, A., Morel, B., and Flouri, T. 2020. ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. Mol. Biol. Evol. **37**(1): 291–294. doi:10.1093/molbev/msz189.

Devoto, A.E., Santini, J.M., Olm, M.R., Anantharaman, K., Munk, P., Tung, J., Archie, E.A., Turnbaugh, P.J., Seed, K.D., Blekhman, R., Aarestrup, F.M., Thomas, B.C., and Banfield, J.F. 2019. Megaphages infect *Prevotella* and variants are widespread in gut microbiomes. Nat. Microbiol. **4**(4): 693–700. doi:10.1038/s41564-018-0338-9.

Dick, G.J., Andersson, A.F., Baker, B.J., Simmons, S.L., Thomas, B.C., Yelton, A.P., and Banfield, J.F. 2009. Community-wide analysis of microbial genome sequence signatures. Genome Biol. **10**(8): R85. doi:10.1186/gb-2009-10-8-r85.

Dimitriu, T., Kurilovich, E., Łapińska, U., Severinov, K., Pagliara, S., Szczelkun, M.D., and Westra, E.R. 2022. Bacteriostatic antibiotics promote CRISPR-Cas adaptive immunity by enabling increased spacer acquisition. Cell Host Microbe **30**(1): 31-40.e5. doi:10.1016/j.chom.2021.11.014.

Dimitriu, T., Szczelkun, M.D., and Westra, E.R. 2020. Evolutionary Ecology and Interplay of Prokaryotic Innate and Adaptive Immune Systems. Curr. Biol. **30**(19): R1189–R1202. doi:10.1016/j.cub.2020.08.028.

Dion, M., Labrie, S., Shah, S., and Moineau, S. 2018. CRISPRStudio: A User-Friendly Software for Rapid CRISPR Array Visualization. Viruses **10**(11): 602. doi:10.3390/v10110602.

Dion, M.B., Oechslin, F., and Moineau, S. 2020. Phage diversity, genomics and phylogeny. Nat. Rev. Microbiol. **18**(3): 125–138. doi:10.1038/s41579-019-0311-5.

Dunne, M., Rupf, B., Tala, M., Qabrati, X., Ernst, P., Shen, Y., Sumrall, E., Heeb, L., Plückthun, A., Loessner, M.J., and Kilcher, S. 2019. Reprogramming Bacteriophage Host Range through Structure-Guided Design of Chimeric Receptor Binding Proteins. Cell Rep. **29**(5): 1336-1350.e4. doi:10.1016/j.celrep.2019.09.062.

Dutilh, B.E., Cassman, N., McNair, K., Sanchez, S.E., Silva, G.G.Z., Boling, L., Barr, J.J., Speth, D.R., Seguritan, V., Aziz, R.K., Felts, B., Dinsdale, E.A., Mokili, J.L., and Edwards, R.A. 2014. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat. Commun. **5**(1): 4498. doi:10.1038/ncomms5498.

Dutilh, B.E., Reyes, A., Hall, R.J., and Whiteson, K.L. 2017. Editorial: Virus Discovery by Metagenomics: The (Im)possibilities. Front. Microbiol. **8**: 1710. doi:10.3389/fmicb.2017.01710.

Eddy, S.R. 1998. Profile hidden Markov models. Bioinformatics **14**(9): 755–763. doi:10.1093/bioinformatics/14.9.755.

Eddy, S.R. 2011. Accelerated Profile HMM Searches. PLoS Comput. Biol. **7**(10): e1002195. doi:10.1371/journal.pcbi.1002195.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**(5): 1792–1797. doi:10.1093/nar/gkh340.

Edwards, R.A., McNair, K., Faust, K., Raes, J., and Dutilh, B.E. 2016. Computational approaches to predict bacteriophage–host relationships. FEMS Microbiol. Rev. **40**(2): 258–272. doi:10.1093/femsre/fuv048.

Emerson, J.B., Roux, S., Brum, J.R., Bolduc, B., Woodcroft, B.J., Jang, H.B., Singleton, C.M., Solden, L.M., Naas, A.E., Boyd, J.A., Hodgkins, S.B., Wilson, R.M., Trubl, G., Li, C., Frolking, S., Pope,

P.B., Wrighton, K.C., Crill, P.M., Chanton, J.P., Saleska, S.R., Tyson, G.W., Rich, V.I., and Sullivan, M.B. 2018. Host-linked soil viral ecology along a permafrost thaw gradient. Nat. Microbiol. **3**(8): 870–880. doi:10.1038/s41564-018-0190-y.

Faure, G., Makarova, K.S., and Koonin, E.V. 2019a. CRISPR-Cas: Complex Functional Networks and Multiple Roles beyond Adaptive Immunity. J. Mol. Biol. **431**(1): 3–20. doi:10.1016/j.jmb.2018.08.030.

Faure, G., Shmakov, S.A., Yan, W.X., Cheng, D.R., Scott, D.A., Peters, J.E., Makarova, K.S., and Koonin, E.V. 2019b. CRISPR–Cas in mobile genetic elements: counter-defence and beyond. Nat. Rev. Microbiol. **17**(8): 513–525. doi:10.1038/s41579-019-0204-7.

Feiner, R., Argov, T., Rabinovich, L., Sigal, N., Borovok, I., and Herskovits, A.A. 2015. A new perspective on lysogeny: prophages as active regulatory switches of bacteria. Nat. Rev. Microbiol. **13**(10): 641–650. doi:10.1038/nrmicro3527.

Forterre, P. 2013. The virocell concept and environmental microbiology. ISME J. **7**(2): 233–236. doi:10.1038/ismej.2012.110.

Galiez, C., Siebert, M., Enault, F., Vincent, J., and Söding, J. 2017. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. Bioinformatics **33**(19): 3113–3114. doi:10.1093/bioinformatics/btx383.

Gambelli, L., Cremers, G., Mesman, R., Guerrero, S., Dutilh, B.E., Jetten, M.S.M., Op den Camp, H.J.M., and van Niftrik, L. 2016. Ultrastructure and Viral Metagenome of Bacteriophages from an Anaerobic Methane Oxidizing *Methylomirabilis* Bioreactor Enrichment Culture. Front. Microbiol. **7**. doi:10.3389/fmicb.2016.01740.

Gao, R., Ma, B., Hu, M., Fang, L., Chen, G., Zhang, W., Wang, Y., Song, X., and Li, F. 2023. Ecological drivers and potential functions of viral communities in flooded arsenic-contaminated paddy soils. Sci. Total Environ. **872**: 162289. doi:10.1016/j.scitotenv.2023.162289.

Garrett, S.C. 2021. Pruning and Tending Immune Memories: Spacer Dynamics in the CRISPR Array. Front. Microbiol. **12**: 664299. doi:10.3389/fmicb.2021.664299.

Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. 2012. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proceedings of the National Academy of Sciences **109**(39): E2579–E2586. doi:10.1073/pnas.1208507109.

George, N.A., and Hug, L.A. 2023. CRISPR-resolved virus-host interactions in a municipal landfill include non-specific viruses, hyper-targeted viral populations, and interviral conflicts. Sci. Rep. **13**(1): 5611. doi:10.1038/s41598-023-32078-6.

Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., Lim, W.A., Weissman, J.S., and Qi, L.S. 2013. CRISPR-

Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. Cell **154**(2): 442–451. doi:10.1016/j.cell.2013.06.044.

Grazziotin, A.L., Koonin, E.V., and Kristensen, D.M. 2017. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res. **45**: D491–D498. doi:10.1093/nar/gkw975.

Grégoire, D.S., George, N.A., and Hug, L.A. 2023, January 21. Microbial methane cycling in a landfill on a decadal time scale. bioRxiv. doi:10.1101/2023.01.20.524919.

Guo, J., Bolduc, B., Zayed, A.A., Varsani, A., Dominguez-Huerta, G., Delmont, T.O., Pratama, A.A., Gazitúa, M.C., Vik, D., Sullivan, M.B., and Roux, S. 2021. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome **9**(1): 37. doi:10.1186/s40168-020-00990-y.

Hampton, H.G., Watson, B.N.J., and Fineran, P.C. 2020. The arms race between bacteria and their phage foes. Nature **577**(7790): 327–336. doi:10.1038/s41586-019-1894-8.

Handelsman, J. 2004. Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiol. Mol. Biol. Rev. **68**(4): 669–685. American Society for Microbiology. doi:10.1128/mmbr.68.4.669-685.2004.

Hargreaves, K.R., Flores, C.O., Lawley, T.D., and Clokie, M.R.J. 2014. Abundant and Diverse Clustered Regularly Interspaced Short Palindromic Repeat Spacers in *Clostridium difficile* Strains and Prophages Target Multiple Phage Types within This Pathogen. mBio **5**(5). doi:10.1128/mBio.01045-13.

Harrington, L.B., Burstein, D., Chen, J.S., Paez-Espino, D., Ma, E., Witte, I.P., Cofsky, J.C., Kyrpides, N.C., Banfield, J.F., and Doudna, J.A. 2018. Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. Science **362**(6416): 839–842. doi:10.1126/science.aav4294.

Howard-Varona, C., Lindback, M.M., Bastien, G.E., Solonenko, N., Zayed, A.A., Jang, H., Andreopoulos, B., Brewer, H.M., Glavina del Rio, T., Adkins, J.N., Paul, S., Sullivan, M.B., and Duhaime, M.B. 2020. Phage-specific metabolic reprogramming of virocells. ISME J. **14**(4): 881–895. doi:10.1038/s41396-019-0580-z.

Hua, J., Huet, A., Lopez, C.A., Toropova, K., Pope, W.H., Duda, R.L., Hendrix, R.W., and Conway, J.F. 2017. Capsids and Genomes of Jumbo-Sized Bacteriophages Reveal the Evolutionary Reach of the HK97 Fold. mBio **8**(5): 10.1128/mbio.01579-17. doi:10.1128/mbio.01579-17.

Huntemann, M., Ivanova, N.N., Mavromatis, K., Tripp, H.J., Paez-Espino, D., Palaniappan, K., Szeto, E., Pillay, M., Chen, I.-M.A., Pati, A., Nielsen, T., Markowitz, V.M., and Kyrpides, N.C. 2015. The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4). Stand. Genom. Sci. **10**(1): 86. doi:10.1186/s40793-015-0077-y.

Hurwitz, B.L., Hallam, S.J., and Sullivan, M.B. 2013. Metabolic reprogramming by viruses in the sunlit and dark ocean. Genome Biol. **14**(11): R123. doi:10.1186/gb-2013-14-11-r123.

Hwang, Y., Roux, S., Coclet, C., Krause, S.J.E., and Girguis, P.R. 2023. Viruses interact with hosts that span distantly related microbial domains in dense hydrothermal mats. Nat. Microbiol. **8**: 946–957. doi:10.1038/s41564-023-01347-5.

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform. **11**(1): 119. doi:10.1186/1471-2105-11-119.

Hynes, A.P., Villion, M., and Moineau, S. 2014. Adaptation in bacterial CRISPR-Cas immunity can be driven by defective phages. Nat. Commun. **5**(1): 4399. doi:10.1038/ncomms5399.

Ignacio-Espinoza, J.C., and Sullivan, M.B. 2012. Phylogenomics of T4 cyanophages: lateral gene transfer in the 'core' and origins of host genes: Molecular evolution of T4 myoviruses. Environ. Microbiol. **14**(8): 2113–2126. doi:10.1111/j.1462-2920.2012.02704.x.

Jansen, Ruud., Embden, Jan.D.A. van, Gaastra, Wim., and Schouls, Leo.M. 2002. Identification of genes that are associated with DNA repeats in prokaryotes. Mol. Microbiol. **43**(6): 1565–1575. doi:10.1046/j.1365-2958.2002.02839.x.

Jarett, J.K., Džunková, M., Schulz, F., Roux, S., Paez-Espino, D., Eloe-Fadrosh, E., Jungbluth, S.P., Ivanova, N., Spear, J.R., Carr, S.A., Trivedi, C.B., Corsetti, F.A., Johnson, H.A., Becraft, E., Kyrpides, N., Stepanauskas, R., and Woyke, T. 2020. Insights into the dynamics between viruses and their hosts in a hot spring microbial mat. ISME J. **14**(10): 2527–2541. doi:10.1038/s41396-020-0705-4.

Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ **7**: e7359. doi:10.7717/peerj.7359.

Karvelis, T., Druteika, G., Bigelyte, G., Budre, K., Zedaveinyte, R., Silanskas, A., Kazlauskas, D., Venclovas, Č., and Siksnys, V. 2021. Transposon-associated TnpB is a programmable RNA-guided DNA endonuclease. Nature **599**(7886): 692–696. doi:10.1038/s41586-021-04058-1.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., and Drummond, A. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics **28**(12): 1647–1649. doi:10.1093/bioinformatics/bts199.

Keen, E.C., Bliskovsky, V.V., Malagon, F., Baker, J.D., Prince, J.S., Klaus, J.S., and Adhya, S.L. 2017. Novel "Superspreader" Bacteriophages Promote Horizontal Gene Transfer by Transformation. mBio **8**(1). doi:10.1128/mBio.02115-16.

Kieft, K., Adams, A., Salamzade, R., Kalan, L., and Anantharaman, K. 2022. vRhyme enables binning of viral genomes from metagenomes. Nucleic Acids Res. **50**(14): e83. doi:10.1093/nar/gkac341.

Kieft, K., Zhou, Z., and Anantharaman, K. 2020. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome **8**(1): 90. doi:10.1186/s40168-020-00867-0.

Kieft, K., Zhou, Z., Anderson, R.E., Buchan, A., Campbell, B.J., Hallam, S.J., Hess, M., Sullivan, M.B., Walsh, D.A., Roux, S., and Anantharaman, K. 2021. Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. Nat. Commun. **12**(1): 3503. doi:10.1038/s41467-021-23698-5.

Köchling, T., Sanz, J.L., Gavazza, S., and Florencio, L. 2015. Analysis of microbial community structure and composition in leachates from a young landfill by 454 pyrosequencing. Appl. Microbiol. Biotechnol. **99**(13): 5657–5668. doi:10.1007/s00253-015-6409-4.

Kolmogorov, M., Bickhart, D.M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S.B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T.P.L., and Pevzner, P.A. 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat. Methods **17**(11): 1103–1110. doi:10.1038/s41592-020-00971-x.

Koonin, E.V. 2009. On the Origin of Cells and Viruses: Primordial Virus World Scenario. Ann. N. Y. Acad. Sci. **1178**(1): 47–64. doi:10.1111/j.1749-6632.2009.04992.x.

Koonin, E.V. 2010. The wonder world of microbial viruses. Expert Rev. Anti Infect. Ther. **8**(10): 1097–1099. doi:10.1586/eri.10.96.

Koonin, E.V., Makarova, K.S., Wolf, Y.I., and Krupovic, M. 2020. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. Nat. Rev. Genet. **21**(2): 119–131. doi:10.1038/s41576-019-0172-9.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., Taillon, B.E., Chen, Z., Tanzer, A., Saunders, A.C.E., Chi, J., Yang, F., Carter, N.P., Hurles, M.E., Weissman, S.M., Harkins, T.T., Gerstein, M.B., Egholm, M., and Snyder, M. 2007. Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. Science **318**(5849): 420–426. doi:10.1126/science.1149504.

Krupovic, M., Cvirkaite-Krupovic, V., Iranzo, J., Prangishvili, D., and Koonin, E.V. 2018. Viruses of archaea: Structural, functional, environmental and evolutionary genomics. Virus Res. **244**: 181–193. doi:10.1016/j.virusres.2017.11.025.

Krupovic, M., Prangishvili, D., Hendrix, R.W., and Bamford, D.H. 2011. Genomics of Bacterial and Archaeal Viruses: Dynamics within the Prokaryotic Virosphere. Microbiol. Mol. Biol. Rev. **75**(4): 610–635. doi:10.1128/MMBR.00011-11.

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., and Hugenholtz, P. 2008. A Bioinformatician's Guide to Metagenomics. Microbiol. Mol. Biol. Rev. **72**(4): 557–578. doi:10.1128/mmbr.00009-08.

Künne, T., Kieper, S.N., Bannenberg, J.W., Vogel, A.I.M., Miellet, W.R., Klein, M., Depken, M., Suarez-Diez, M., and Brouns, S.J.J. 2016. Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. Mol. Cell **63**(5): 852–864. doi:10.1016/j.molcel.2016.07.011.

Langenfeld, K., Chin, K., Roy, A., Wigginton, K., and Duhaime, M.B. 2021. Comparison of ultrafiltration and iron chloride flocculation in the preparation of aquatic viromes from contrasting sample types. PeerJ **9**: e11111. doi:10.7717/peerj.11111.

Langmead, B., and Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**(4): 357–359. doi:10.1038/nmeth.1923.

Levy, A., Goren, M.G., Yosef, I., Auster, O., Manor, M., Amitai, G., Edgar, R., Qimron, U., and Sorek, R. 2015. CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature **520**(7548): 505–510. doi:10.1038/nature14302.

Li, W., and Godzik, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22**(13): 1658–1659. doi:10.1093/bioinformatics/btl158.

Lindell, D., Sullivan, M.B., Johnson, Z.I., Tolonen, A.C., Rohwer, F., and Chisholm, S.W. 2004. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proc. Natl. Acad. Sci. **101**(30): 11013–11018. doi:10.1073/pnas.0401526101.

Luo, X.-Q., Wang, P., Li, J.-L., Ahmad, M., Duan, L., Yin, L.-Z., Deng, Q.-Q., Fang, B.-Z., Li, S.-H., and Li, W.-J. 2022. Viral community-wide auxiliary metabolic genes differ by lifestyles, habitats, and hosts. Microbiome **10**(1): 190. doi:10.1186/s40168-022-01384-y.

Maguin, P., Varble, A., Modell, J.W., and Marraffini, L.A. 2022. Cleavage of viral DNA by restriction endonucleases stimulates the type II CRISPR-Cas immune response. Mol. Cell **82**(5): 907-919.e7. doi:10.1016/j.molcel.2022.01.012.

Makarova, K.S., Wolf, Y.I., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J., Charpentier, E., Cheng, D., Haft, D.H., Horvath, P., Moineau, S., Mojica, F.J.M., Scott, D., Shah, S.A., Siksnys, V., Terns, M.P., Venclovas, Č., White, M.F., Yakunin, A.F., Yan, W., Zhang, F., Garrett, R.A., Backofen, R., van der Oost, J., Barrangou, R., and Koonin, E.V. 2020. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nat. Rev. Microbiol. **18**(2): 67–83. doi:10.1038/s41579-019-0299-x.

Malki, K., Kula, A., Bruder, K., Sible, E., Hatzopoulos, T., Steidel, S., Watkins, S.C., and Putonti, C. 2015. Bacteriophages isolated from Lake Michigan demonstrate broad host-range across several bacterial phyla. Virol. J. **12**(1). doi:10.1186/s12985-015-0395-0.

Marraffini, L.A., and Sontheimer, E.J. 2008. CRISPR interference limits horizontal gene transfer in Staphylococci by targeting DNA. Science **322**: 4.

Martínez Arbas, S., Narayanasamy, S., Herold, M., Lebrun, L.A., Hoopmann, M.R., Li, S., Lam, T.J., Kunath, B.J., Hicks, N.D., Liu, C.M., Price, L.B., Laczny, C.C., Gillece, J.D., Schupp, J.M., Keim, P.S., Moritz, R.L., Faust, K., Tang, H., Ye, Y., Skupin, A., May, P., Muller, E.E.L., and Wilmes, P. 2021. Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics. Nat. Microbiol. **6**(1): 123–135. doi:10.1038/s41564-020-00794-8.

Medvedeva, S., Liu, Y., Koonin, E.V., Severinov, K., Prangishvili, D., and Krupovic, M. 2019. Virus-borne mini-CRISPR arrays are involved in interviral conflicts. Nat. Commun. **10**(1): 5204. doi:10.1038/s41467-019-13205-2.

von Meijenfeldt, F.A.B., Arkhipova, K., Cambuy, D.D., Coutinho, F.H., and Dutilh, B.E. 2019. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. Genome Biol. **20**(1): 217. doi:10.1186/s13059-019-1817-x.

Merchant, N., Lyons, E., Goff, S., Vaughn, M., Ware, D., Micklos, D., and Antin, P. 2016. The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. PLoS Biol. **14**(1): e1002342. doi:10.1371/journal.pbio.1002342.

Meyer-Dombard, D.R., Bogner, J.E., and Malas, J. 2020. A Review of Landfill Microbiology and Ecology: A Call for Modernization With "Next Generation" Technology. Front. Microbiol. **11**: 1127. doi:10.3389/fmicb.2020.01127.

Michniewski, S., Rihtman, B., Cook, R., Jones, M.A., Wilson, W.H., Scanlan, D.J., and Millard, A. 2021. A new family of "megaphages" abundant in the marine environment. ISME. Commun. **1**(1): 1–4. doi:10.1038/s43705-021-00064-6.

Minot, S., Bryson, A., Chehoud, C., Wu, G.D., Lewis, J.D., and Bushman, F.D. 2013. Rapid evolution of the human gut virome. Proc. Natl. Acad. Sci. **110**(30): 12450–12455.

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S.A., Wu, G.D., Lewis, J.D., and Bushman, F.D. 2011. The human gut virome: Inter-individual variation and dynamic response to diet. Genome Res. **21**(10): 1616–1625. doi:10.1101/gr.122705.111.

Mitrofanov, A., Alkhnbashi, O.S., Shmakov, S.A., Makarova, K.S., Koonin, E.V., and Backofen, R. 2021. CRISPRidentify: identification of CRISPR arrays using machine learning approach. Nucleic Acids Res. **49**(4): e20. doi:10.1093/nar/gkaa1158.

Modell, J.W., Jiang, W., and Marraffini, L.A. 2017. CRISPR–Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. Nature **544**(7648): 101–104. doi:10.1038/nature21719.

Mojica, F.J.M., Diez-Villasenor, C., Garcia-Martinez, J., and Soria, E. 2005. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. J. Mol. Evol. **60**(2): 174–182. doi:10.1007/s00239-004-0046-3.

Nasko, D.J., Ferrell, B.D., Moore, R.M., Bhavsar, J.D., Polson, S.W., and Wommack, K.E. 2019. CRISPR Spacers Indicate Preferential Matching of Specific Virioplankton Genes. mBio **10**(2): 10.1128/mbio.02651-18. doi:10.1128/mbio.02651-18.

Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N.C. 2021. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat. Biotechnol. **39**(5): 578–585. doi:10.1038/s41587-020-00774-7.

Nety, S.P., Altae-Tran, H., Kannan, S., Demircioglu, F.E., Faure, G., Hirano, S., Mears, K., Zhang, Y., Macrae, R.K., and Zhang, F. 2023. The Transposon-Encoded Protein TnpB Processes Its Own mRNA into ωRNA for Guided Nuclease Activity. CRISPR J. **6**(3): 232–242. doi:10.1089/crispr.2023.0015.

Olm, M.R., Brown, C.T., Brooks, B., and Banfield, J.F. 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. **11**(12): 2864–2868. doi:10.1038/ismej.2017.126.

Paez-Espino, D., Chen, I.-M.A., Palaniappan, K., Ratner, A., Chu, K., Szeto, E., Pillay, M., Huang, J., Markowitz, V.M., Nielsen, T., Huntemann, M., K. Reddy, T.B., Pavlopoulos, G.A., Sullivan, M.B., Campbell, B.J., Chen, F., McMahon, K., Hallam, S.J., Denef, V., Cavicchioli, R., Caffrey, S.M., Streit, W.R., Webster, J., Handley, K.M., Salekdeh, G.H., Tsesmetzis, N., Setubal, J.C., Pope, P.B., Liu, W.-T., Rivers, A.R., Ivanova, N.N., and Kyrpides, N.C. 2016a. IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. Nucleic Acids Res. **45**(D1): gkw1030. doi:10.1093/nar/gkw1030.

Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N.N., and Kyrpides, N.C. 2016b. Uncovering Earth's virome. Nature **536**(7617): 425–430. doi:10.1038/nature19094.

Paez-Espino, D., Morovic, W., Sun, C.L., Thomas, B.C., Ueda, K., Stahl, B., Barrangou, R., and Banfield, J.F. 2013. Strong bias in the bacterial CRISPR elements that confer immunity to phage. Nat. Commun. **4**(1): 1430. doi:10.1038/ncomms2440.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. **25**(7): 1043–1055. doi:10.1101/gr.186072.114.

Pausch, P., Al-Shayeb, B., Bisom-Rapp, E., Tsuchida, C.A., Li, Z., Cress, B.F., Knott, G.J., Jacobsen, S.E., Banfield, J.F., and Doudna, J.A. 2020. CRISPR-CasΦ from huge phages is a hypercompact genome editor. Science **369**(6501): 333–337. doi:10.1126/science.abb1400.

Penadés, J.R., Chen, J., Quiles-Puchalt, N., Carpena, N., and Novick, R.P. 2015. Bacteriophage-mediated spread of bacterial virulence genes. Curr. Opin. Microbiol. **23**: 171–178. doi:10.1016/j.mib.2014.11.019.

Peters, D.L., Lynch, K.H., Stothard, P., and Dennis, J.J. 2015. The isolation and characterization of two *Stenotrophomonas maltophilia* bacteriophages capable of cross-taxonomic order infectivity. BMC Genom. **16**(1): 664. doi:10.1186/s12864-015-1848-y.

Pinilla-Redondo, R., Russel, J., Mayo-Muñoz, D., Shah, S.A., Garrett, R.A., Nesme, J., Madsen, J.S., Fineran, P.C., and Sørensen, S.J. 2022. CRISPR-Cas systems are widespread accessory elements across bacterial and archaeal plasmids. Nucleic Acids Res. **50**(8): 4315–4328. doi:10.1093/nar/gkab859.

Poulos, B.T., John, S.G., and Sullivan, M.B. 2018. Iron Chloride Flocculation of Bacteriophages from Seawater. Methods Mol. Biol. **1681**: 49–57. doi:10.1007/978-1-4939-7343-9_4.

Pourcel, C., Salvignol, G., and Vergnaud, G. 2005. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology **151**(3): 653–663. doi:10.1099/mic.0.27437-0.

Pride, D.T., Sun, C.L., Salzman, J., Rao, N., Loomer, P., Armitage, G.C., Banfield, J.F., and Relman, D.A. 2011. Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time. Genome Res. **21**(1): 126–136. doi:10.1101/gr.111732.110.

Rappé, M.S., and Giovannoni, S.J. 2003. The Uncultured Microbial Majority. Annu. Rev. Microbiol. **57**(1): 369–394. doi:10.1146/annurev.micro.57.030502.090759.

Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., and Sun, F. 2017. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome **5**(1): 69. doi:10.1186/s40168-017-0283-5.

Ren, J., Song, K., Deng, C., Ahlgren, N.A., Fuhrman, J.A., Li, Y., Xie, X., Poplin, R., and Sun, F. 2020. Identifying viruses from metagenomic data using deep learning. Quant. Biol. **8**(1): 64–77. doi:10.1007/s40484-019-0187-4.

Richter, C., Dy, R.L., McKenzie, R.E., Watson, B.N.J., Taylor, C., Chang, J.T., McNeil, M.B., Staals, R.H.J., and Fineran, P.C. 2014. Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. Nucleic Acids Res. **42**(13): 8516–8526. doi:10.1093/nar/gku527.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. 2011. Integrative genomics viewer. Nat. Biotechnol. **29**(1): 24–26. doi:10.1038/nbt.1754.

Rohwer, F., Prangishvili, D., and Lindell, D. 2009. Roles of viruses in the environment. Environ. Microbiol. **11**(11): 2771–2774. doi:10.1111/j.1462-2920.2009.02101.x.

Roux, S., and Brum, J.R. 2019. A viral reckoning: viruses emerge as essential manipulators of global ecosystems. Environ. Microbiol. Rep. **11**(1): 3–8. doi:10.1111/1758-2229.12700.

Roux, S., Camargo, A.P., Coutinho, F.H., Dabdoub, S.M., Dutilh, B.E., Nayfach, S., and Tritt, A. 2023. iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. PLOS Biol. **21**(4): e3002083. doi:10.1371/journal.pbio.3002083.

Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. 2015. VirSorter: mining viral signal from microbial genomic data. PeerJ **3**: e985. doi:10.7717/peerj.985.

Roux, S., Krupovic, M., Daly, R.A., Borges, A.L., Nayfach, S., Schulz, F., Sharrar, A., Matheus Carnevali, P.B., Cheng, J.-F., Ivanova, N.N., Bondy-Denomy, J., Wrighton, K.C., Woyke, T., Visel, A., Kyrpides, N.C., and Eloe-Fadrosh, E.A. 2019. Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. Nat. Microbiol. **4**(11): 1895–1906. doi:10.1038/s41564-019-0510-x.

Roux, S., Paul, B.G., Bagby, S.C., Nayfach, S., Allen, M.A., Attwood, G., Cavicchioli, R., Chistoserdova, L., Gruninger, R.J., Hallam, S.J., Hernandez, M.E., Hess, M., Liu, W.-T., McAllister, T.A., O'Malley, M.A., Peng, X., Rich, V.I., Saleska, S.R., and Eloe-Fadrosh, E.A. 2021. Ecology and molecular targets of hypermutation in the global microbiome. Nat. Commun. **12**(1): 3076. doi:10.1038/s41467-021-23402-7.

Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S.A., and Sørensen, S.J. 2020. CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. CRISPR J. **3**(6): 462–469. doi:10.1089/crispr.2020.0059.

Sangwan, N., Xia, F., and Gilbert, J.A. 2016. Recovering complete and draft population genomes from metagenome datasets. Microbiome **4**(1). doi:10.1186/s40168-016-0154-5.

Sauk, A.H., and Hug, L.A. 2022. Substrate-restricted methanogenesis and limited volatile organic compound degradation in highly diverse and heterogeneous municipal landfill microbial communities. ISME Commun. **2**(1): 58. doi:10.1038/s43705-022-00141-4.

Schulz, F., Roux, S., Paez-Espino, D., Jungbluth, S., Walsh, D.A., Denef, V.J., McMahon, K.D., Konstantinidis, K.T., Eloe-Fadrosh, E.A., Kyrpides, N.C., and Woyke, T. 2020. Giant virus diversity and host interactions through global metagenomics. Nature **578**(7795): 432–436. doi:10.1038/s41586-020-1957-x.

Seed, K.D., Lazinski, D.W., Calderwood, S.B., and Camilli, A. 2013. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. Nature **494**(7438): 489–491. doi:10.1038/nature11927.

Sexton, C.E., and Han, M.V. 2019. Paired-end mappability of transposable elements in the human genome. Mob. DNA **10**(1): 29. doi:10.1186/s13100-019-0172-5.

Sharon, I., Tzahor, S., Williamson, S., Shmoish, M., Man-Aharonovich, D., Rusch, D.B., Yooseph, S., Zeidner, G., Golden, S.S., Mackey, S.R., Adir, N., Weingart, U., Horn, D., Venter, J.C., Mandel-Gutfreund, Y., and Béjà, O. 2007. Viral photosynthetic reaction center genes and transcripts in the marine environment. ISME J. **1**(6): 492–501. doi:10.1038/ismej.2007.67.

Shmakov, S., Smargon, A., Scott, D., Cox, D., Pyzocha, N., Yan, W., Abudayyeh, O.O., Gootenberg, J.S., Makarova, K.S., Wolf, Y.I., Severinov, K., Zhang, F., and Koonin, E.V. 2017. Diversity and evolution of class 2 CRISPR–Cas systems. Nat. Rev. Microbiol. **15**(3): 169–182. doi:10.1038/nrmicro.2016.184.

Shmakov, S.A., Utkina, I., Wolf, Y.I., Makarova, K.S., Severinov, K.V., and Koonin, E.V. 2020. CRISPR Arrays Away from *cas* Genes. CRISPR J. **3**(6): 535–549. doi:10.1089/crispr.2020.0062.

Sieber, C.M.K., Probst, A.J., Sharrar, A., Thomas, B.C., Hess, M., Tringe, S.G., and Banfield, J.F. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat. Microbiol. **3**(7): 836–843. doi:10.1038/s41564-018-0171-1.

Sieradzki, E.T., Ignacio-Espinoza, J.C., Needham, D.M., Fichot, E.B., and Fuhrman, J.A. 2019. Dynamic marine viral infections and major contribution to photosynthetic processes shown by spatiotemporal picoplankton metatranscriptomes. Nat. Commun. **10**(1): 1169. doi:10.1038/s41467-019-09106-z.

Skennerton, C.T., Imelfort, M., and Tyson, G.W. 2013. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. Nucleic Acids Res. **41**(10): e105–e105. doi:10.1093/nar/gkt183.

Song, L., Wang, Y., Zhao, H., and Long, D.T. 2015. Composition of bacterial and archaeal communities during landfill refuse decomposition processes. Microbiol. Res. **181**: 105–111. doi:10.1016/j.micres.2015.04.009.

Song, W., Sun, H.-X., Zhang, C., Cheng, L., Peng, Y., Deng, Z., Wang, D., Wang, Y., Hu, M., Liu, W., Yang, H., Shen, Y., Li, J., You, L., and Xiao, M. 2019. Prophage Hunter: an integrative hunting tool for active prophages. Nucleic Acids Res. **47**(W1): W74–W80. doi:10.1093/nar/gkz380.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**(9): 1312–1313. doi:10.1093/bioinformatics/btu033.

Stamps, B.W., Lyles, C.N., Suflita, J.M., Masoner, J.R., Cozzarelli, I.M., Kolpin, D.W., and Stevenson, B.S. 2016. Municipal Solid Waste Landfills Harbor Distinct Microbiomes. Front. Microbiol. **7**. doi:10.3389/fmicb.2016.00534.

Strecker, J., Ladha, A., Gardner, Z., Schmid-Burgk, J.L., Makarova, K.S., Koonin, E.V., and Zhang, F. 2019. RNA-guided DNA insertion with CRISPR-associated transposases. Science **365**(6448): 48–53. doi:10.1126/science.aax9181.

Sun, L., You, J., Li, D., Zhang, Z., Qin, X., Pang, W., Li, P., Han, Q., Li, Y., Huang, Z., Zhang, X., Gong, M., and Yang, H. 2023. Variants of a putative baseplate wedge protein extend the host range of *Pseudomonas* phage K8. Microbiome **11**(1): 18. doi:10.1186/s40168-022-01459-w.

Suttle, C.A. 2007. Marine viruses — major players in the global ecosystem. Nat. Rev. Microbiol. **5**(10): 801–812. doi:10.1038/nrmicro1750.

Suttle, C.A. 2016. Environmental microbiology: Viral diversity on the global stage. Nat. Microbiol. **1**(11): 16205. doi:10.1038/nmicrobiol.2016.205.

Sutton, T.D.S., Clooney, A.G., Ryan, F.J., Ross, R.P., and Hill, C. 2019. Choice of assembly software has a critical impact on virome characterisation. Microbiome **7**(1): 12. doi:10.1186/s40168-019-0626-5.

Tithi, S.S., Aylward, F.O., Jensen, R.V., and Zhang, L. 2018. FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data. PeerJ **6**: e4227. doi:10.7717/peerj.4227.

Trubl, G., Jang, H.B., Roux, S., Emerson, J.B., Solonenko, N., Vik, D.R., and Solden, L.M. 2018. Soil Viruses Are Underexplored Players in Ecosystem Carbon Processing. mSystems **3**(5). [accessed 7 January 2019].

Turner, D., Kropinski, A.M., and Adriaenssens, E.M. 2021. A Roadmap for Genome-Based Phage Taxonomy. Viruses **13**(3): 506. doi:10.3390/v13030506.

Turner, D., Shkoporov, A.N., Lood, C., Millard, A.D., Dutilh, B.E., Alfenas-Zerbini, P., van Zyl, L.J., Aziz, R.K., Oksanen, H.M., Poranen, M.M., Kropinski, A.M., Barylski, J., Brister, J.R., Chanisvili, N., Edwards, R.A., Enault, F., Gillis, A., Knezevic, P., Krupovic, M., Kurtböke, I., Kushkina, A., Lavigne, R., Lehman, S., Lobocka, M., Moraru, C., Moreno Switt, A., Morozova, V., Nakavuma, J., Reyes Muñoz, A., Rūmnieks, J., Sarkar, B.L., Sullivan, M.B., Uchiyama, J., Wittmann, J., Yigang, T., and Adriaenssens, E.M. 2023. Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. Arch. Virol. **168**(2): 74. doi:10.1007/s00705-022-05694-2.

Tyson, G.W., and Banfield, J.F. 2008. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environ. Microbiol. **10**(1): 200–207. doi:10.1111/j.1462-2920.2007.01444.x.

Vale, P.F., Lafforgue, G., Gatchitch, F., Gardan, R., Moineau, S., and Gandon, S. 2015. Costs of CRISPR-Cas-mediated resistance in *Streptococcus thermophilus*. Proc. R. Soc. B: Biol. Sci. **282**(1812): 20151270. doi:10.1098/rspb.2015.1270.

Van Der Oost, J., and Patinios, C. 2023. The genome editing revolution. Trends Biotechnol. **41**(3): 396–409. doi:10.1016/j.tibtech.2022.12.022.

Wagner, D.L., Amini, L., Wendering, D.J., Burkhardt, L.-M., Akyüz, L., Reinke, P., Volk, H.-D., and Schmueck-Henneresse, M. 2019. High prevalence of *Streptococcus pyogenes* Cas9-reactive T cells within the adult human population. Nat. Med. **25**(2): 242–248. doi:10.1038/s41591-018-0204-6.

Wooley, J.C., and Ye, Y. 2010. Metagenomics: Facts and Artifacts, and Computational Challenges. J. Comput. Sci. Technol. **25**(1): 71–81. doi:10.1007/s11390-010-9306-4.

Wu, W.Y., Mohanraju, P., Liao, C., Adiego-Pérez, B., Creutzburg, S.C.A., Makarova, K.S., Keessen, K., Lindeboom, T.A., Khan, T.S., Prinsen, S., Joosten, R., Yan, W.X., Migur, A., Laffeber, C., Scott, D.A., Lebbink, J.H.G., Koonin, E.V., Beisel, C.L., and Van Der Oost, J. 2022. The miniature CRISPR-Cas12m effector binds DNA to block transcription. Mol. Cell **82**(23): 4487-4502.e7. doi:10.1016/j.molcel.2022.11.003.

Wu, Y.-W., Simmons, B.A., and Singer, S.W. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics **32**(4): 605–607. doi:10.1093/bioinformatics/btv638.

Yuan, S., Friman, V.-P., Balcazar, J.L., Zheng, X., Ye, M., Sun, M., and Hu, F. 2023. Viral and Bacterial Communities Collaborate through Complementary Assembly Processes in Soil to Survive Organochlorine Contamination. Appl. Environ. Microbiol. **89**(3): e01810-22. doi:10.1128/aem.01810-22.

Yuan, Y., and Gao, M. 2017. Jumbo Bacteriophages: An Overview. Front. Microbiol. **8**: 403. doi:10.3389/fmicb.2017.00403.

Zerbini, F.M., Siddell, S.G., Lefkowitz, E.J., Mushegian, A.R., Adriaenssens, E.M., Alfenas-Zerbini, P., Dempsey, D.M., Dutilh, B.E., García, M.L., Hendrickson, R.C., Junglen, S., Krupovic, M., Kuhn, J.H., Lambert, A.J., Łobocka, M., Oksanen, H.M., Robertson, D.L., Rubino, L., Sabanadzovic, S., Simmonds, P., Smith, D.B., Suzuki, N., Van Doorslaer, K., Vandamme, A., and Varsani, A. 2023. Changes to virus taxonomy and the ICTV Statutes ratified by the International Committee on Taxonomy of Viruses (2023). Arch. Virol. **168**(7): 175. doi:10.1007/s00705-023-05797-4.

Zhang, Q., Doak, T.G., and Ye, Y. 2014. Expanding the catalog of *cas* genes with metagenomes. Nucleic Acids Res. **42**(4): 2448–2459. doi:10.1093/nar/gkt1262.

Zheng, X., Jahn, M.T., Sun, M., Friman, V.-P., Balcazar, J.L., Wang, J., Shi, Y., Gong, X., Hu, F., and Zhu, Y.-G. 2022. Organochlorine contamination enriches virus-encoded metabolism and pesticide degradation associated auxiliary genes in soil microbiomes. ISME J. **16**(5): 1397–1408. doi:10.1038/s41396-022-01188-w.

Zhong, X., and Jacquet, S. 2014. Contrasting diversity of phycodnavirus signature genes in two large and deep western European lakes: Phycodnavirus diversity in alpine lakes. Environ. Microbiol. **16**(3): 759–773. doi:10.1111/1462-2920.12201.

Appendix A

Supplementary Material: Chapter 2

Supplementary text

Supplemental Results and Discussion

Prokaryotic virus-host interactions

The 2016 host-virus network included 37 unique host genome bins with 66 unique links to their predicted prokaryotic viral elements (Figure A.3). In the 2017 network, 90 unique host MAGs had a total of 173 unique links to predicted prokaryotic viral elements (Figure A.4). In our 2016-host-to-2017-virus network (Figure A.5) we observed 80 unique links between prokaryotic viral elements and their predicted hosts (compared to 66 in the 2016 network) and 39 unique host genome bins (compared to 37 in the 2016 network). In general, this network was simpler than the 2016-2016 or 2017-2017 networks. In the 2017-host-to-2016-virus network (Figure A.6) we observed 139 unique links between prokaryotic viral elements and their predicted hosts (compared to 173 in the 2017 network) and 87 unique host genome bins (compared to 90 in the 2017 network). The virus-interacting MAGs in the 2016 network ranged in completion from 70.46% to 100.00% (average completion 91.39%) and ranged in contamination from 0.00% to 9.47% (average contamination 2.45%). The virus-interacting MAGs in the 2017 network ranged in completion from 70.30% to 100.00% (average completion 90.98%) and ranged in contamination from 0.00% to 9.99% (average contamination 3.25%). The virus-interacting MAGs in the 2016-host-to-2017-virus network ranged in completion from 70.36% to 100.00% (average completion 89.61%) and ranged in contamination from 0.00% to 7.27% (average contamination 2.43%). The virus-interacting MAGs in the 2017-host-to-2016-virus network ranged in completion from 70.53% to

100.00% (average completion 89.45%) and ranged in contamination from 0.00% to 9.99% (average contamination 3.45%).

In the 2016 network, six of 59 viral elements were predicted to infect two or more distinct hosts. In the 2017 network, eleven of 161 viral elements were predicted to infect two or more distinct hosts. In the 2016-host-to-2017-viruses network, one of 79 viral elements were predicted to infect two or more distinct hosts. In the 2017-hosts-to-2016-viruses, thirteen of 125 viral elements were predicted to infect two or more distinct hosts.

Curation of cross-phylum and cross-domain virus-host interactions

14 viral elements were predicted to infect multiple hosts that are from different phyla (8) or domains (6), in a total of 10 interactions (Table A.1). Curation took the following requirements into consideration:

1) Host genome quality – following our baseline thresholds of >70 completion and <10 contamination, all 10 interactions pass.
2) CRISPR-encoding scaffolds' taxonomic annotations from JGI scaffold annotations and/or the Contig Annotation Tool (von Meijenfeldt et al. 2019) were required to be consistent with GTDB-tk-based MAG taxonomy (Chaumeil et al. 2019) – six of the remaining interaction sets had erratic scaffold annotations (four remaining).
3) CRISPR-encoding scaffolds were required to be minimum 4,000 bp long – no additional interactions were eliminated. Four interactions pass quality curation.

Only in the case of GW1_158 (class Paceibacteria) and LW4_67 (genus *Methylobacter*) did two MAGs share any spacers in the CRISPR array responsible for the targeting of their mutually shared

virus. Notably, GW1_158's array contained 57 spacers and LW4_67's contained 78, but 56 of these spacers were shared with 100% nucleotide identity. The scaffold of LW4_67 relevant to this analysis is the longer of the two (15,46 includes an intact Type I CRISPR-Cas system (Makarova et al. 2020), and numerous coding sequences with predicted species-level taxonomy that is in accordance with the GTDB-tk annotated taxonomy of this MAG. While the scaffold of GW1_158 relevant to this analysis only has a length of 4,111 bp, this scaffold also has predicted species-level taxonomy, according to the Contig Annotation Tool, that is in accordance with the GTDB-tk annotated taxonomy of this MAG. We thus predict the shared spacers are the result of a possible horizontal gene transfer of a CRISPR array via a mobile genetic element (Faure et al. 2019b), but could also be due to a misbin, a possibility that cannot be refuted for such a short fragment. In all other cases, no spacers were shared, reducing the likelihood that the predicted shared viral targeting is underlaid by lateral gene transfer or mis-binning (Table A.1).

Three of the refuted cross-domain interactions all include the same Halobacterota bin (CLC_33_Methanoculleus_2017), and three different bacterial partners – two Cloacimonadota and one Spirochaetota (CLC_82_Cloacimonadaceae_2017, LW3_15_Cloacimonadaceae_2017, and CLC_92_Sphaerochaeata_2017). CLC_33_Methanoculleus_2017 contains a CRISPR array on a relatively short scaffold (3,667 bp) that targets two different viral scaffolds, each of which is targeted by a distinct Cloacimonadota MAG. This same *Methanoculleus* MAG encodes a second CRISPR array on another short scaffold (3,710) that targets a different viral scaffold that is also targeted by a Sphaerochaeata MAG. This may be a legitimate signal, or the result of multiple orphaned CRISPR arrays or mobile elements binning into this genome. We note the observation as worth monitoring in other systems and future samples from this system.

Within the cross-phylum interactions, there was no consistent pattern of phyla involved, including connections between Bacteroidota-Firmicutes, Chloroflexota-Firmicutes, Patescibactera-Proteobacteria, and others (Table A.1).

We anticipate that, as computational methods for associating hosts and viral elements improve, there will be more instances of viruses that show potential to infect across relatively high taxonomic levels. We hope that rather than defaulting to these interactions being workflow artifacts, or considering them as ground truths, they are examined critically on a case-by-case basis.

Hyper-targeting of viral elements

In the 2016 network, a member of the family Acholeplasmataceae_A (LW2_3) showed the strongest targeting, with 48 CRISPR spacer matches to a single viral element (v111_2016) (Figure 2.1B, Figure A.3). In the 2017 host-to-2016-viruses network, while the number of unique virus-host links was less than in the 2017-only network (139 compared to 173 in our 2017 network), we now also observed 2017 hosts with the ability to hypertarget viral elements (Figure 2.1B). It is notable that these four viral elements are not only hyper-targeted, but are also predicted to infect members of different phyla (phylum Bacteroidota and phylum Firmicutes). Members of these two phyla (LW2_63_Bacteroidales, and LW2_3_ Acholeplasmataceae_A) also shared a viral element in the 2016 network (v80_2016, Figure A.3), but only LW2_3_ Acholeplasmataceae_A hyper-targeted the viral element in that case.

We posit that the shared hyper-targeting across multiple phyla is a result of LGT or misbinning of CRISPR-encoding scaffolds. CLC_107_Bacteroidales_2017 scaffold Ga0265294_10000314 (length of 63,905bp, 147 spacers encoded) has a six mismatch, 4,167 bp alignment with LW2_3_Acholeplasmataceae_A_2016 scaffold Ga0172382_10051677 (length of 4,167bp, 63 spacers encoded) and thus mutual hyper-targeting between distinct phyla is likely due to a misbin

of scaffold Ga0172382_10051677. Furthermore, LW2_114_Bacteroidales_2017 scaffold Ga0265293_10000924 (length 67,619 bp, 182 spacers encoded) has a single mismatch 3,322bp alignment with LW2_22_Acholeplasmataceae_2016 scaffold Ga0172382_10071763 (length 3,322bp, 34 spacers encoded) and thus mutual hyper-targeting between distinct phyla is likely due to a misbin of scaffold Ga0172382_10071763. LW2_22_Acholeplasmataceae_2016 and LW2_114_Bacteroidales_2017 also each encode two other CRISPR arrays that share five and two spacers, respectively, between MAGs.

Despite the striking differences between their viral element targeting profiles, CLC_107_Bacteroidales_2017 and LW2_114_Bacteroidales_2017 are predicted to have an Average Nucleotide Identity (ANI) of >98.5 according to dRep (Olm et al. 2017). It is noteworthy to mention that despite their high genetic identity, CLC_107_Bacteroidales_2017 and LW2_114_Bacteroidales_2017 do not share a single spacer between any of their CRISPR arrays, however, CLC_107_Bacteroidales_2017 shares six spacers with LW2_22_Acholeplasmatceae_2016. Lastly, LW2_3_Acholeplasmataceae_A_2016 shares one spacer with LW2_114_Bacteroidales_2017 and five spacers with LW2_22_Acholeplasmataceae_2016.

Virally encoded CRISPR arrays convolute virus-host networks

While linking hosts to their putative viruses using CRISPR spacer matches to protospacers, we also came across instances of viral elements that encoded CRISPR arrays and CRISPR-Cas systems. Virus-encoded CRISPR arrays and CRISPR-Cas systems have been identified before (Minot et al. 2011, 2013; Seed et al. 2013; Hargreaves et al. 2014; Chénard et al. 2016; Faure et al. 2019b; Medvedeva et al. 2019; Al-Shayeb et al. 2020, 2022) and have been shown to be involved in crippling host viral defense systems (Seed et al. 2013) and are suggested to play roles

in regulating host transcription and translation (Al-Shayeb et al. 2020), CRISPR-Cas system inhibition (Faure et al. 2019b) and interviral conflicts (Faure et al. 2019b; Medvedeva et al. 2019). In the case of interviral conflicts, a CRISPR-encoding virus, integrated as a provirus, can provide its host with immunity against other viruses, akin to a superinfection exclusion system (Faure et al. 2019b). For this work, we used the full set of viral predictions prior to clustering with CD-HIT (Li and Godzik 2006) in order to preserve CRISPR array variation between closely-related viral elements.

Spacers encoded in viral CRISPR arrays from 2016 were predicted to target 89 viral elements in the unclustered 2016 virus data set. Omitting spacer matches between viral CRISPR arrays, spacers encoded in 2016 virus CRISPR arrays were predicted to target 32 viral elements in the unclustered 2016 virus dataset.

We next wanted to assess if any of the viral elements targeted by viral CRISPR arrays were also targeted by host MAGs in any of our networks. Wherever this was the case, we wanted to determine if the CRISPR immunity against the viral element came from a CRISPR-encoding provirus within the host MAG. For this, we screened relevant host MAGs for proviruses using a homology-based approach (Emerson et al. 2018) (see section 2.2 for details), followed by Prophage Hunter (Song et al. 2019), which in select cases was also followed by PHASTER (Arndt et al. 2016). CRISPR arrays in putative provirus elements were annotated using CRISPRCasFinder (Couvin et al. 2018).

In the 2016 network, only one viral element (v30_2016) shown previously to be targeted by viral CRISPR arrays, was targeted by a host MAG (CLC_T2_12_Fibrobacteraceae). Notably, v30_2016 itself was one of the viral elements predicted to encode its own CRISPR array. In the case of CLC_T2_12_Fibrobacteraceae, no prophages were detected. In the 2017 host to 2016 virus

network, 11 of 89 viral elements targeted by viral CRISPR arrays were implicated in the network, targeted by a total of 10 unique hosts. Screening first for proviruses using the homology-based approach predicted six provirus-containing host genomes. Prophage hunter agreed with 4/6 of these predictions. 2/4 of these provirus regions had high confidence CRISPR arrays annotated within them by CRISPRCasFinder.

LW1_9_Paceibacterales (completion: 73.56%, contamination: 5.17%) encodes a putative prophage of v33_2016, a virus that contains a CRISPR array that targets its free virus equivalent via six spacers in our 2017-host-to-2016-virus network. LW1_9_Paceibacterales also contains a second CRISPR array, not within a prophage, that targets v33_2016 two additional times. Second, LW3_16_Pacearchaeales (completion: 71.88%, contamination: 0.00%) encodes a provirus of v169_2016. Notably, the scaffold that contains the putative provirus, Ga0378408_0001874, is predicted to encode five CRISPRs and a total of 40 spacers according to CRISPRCasFinder, all of which are encoded within the predicted proviral region. This host's provirus also targets its free virus equivalent, v169_2016, in the 2017 host to 2016 virus network 27 times, which stem from the provirus-encoded array. Unlike in the case of LW1_9_Paceibacterales, LW3_16_Pacearchaeales does not encode CRISPR immunity against v169_2016 outside of its provirus. Lastly, LW2_146_Pigmentiphaga (completion: 97.56%, contamination: 0.30%), contains a prophage of v36_2016 which encodes two Cas gene clusters according to CRISPRCasFinder: a Cas-TypeIV cluster containing *Csf2_1_IV*, and *Csf3_1_IV*, and a Cas-TypeIA cluster containing *Cas6_0_IA*. In this same host, ~37 kbp upstream of the Cas gene clusters, but outside the prophage boundary predicted by Prophage Hunter, a mini-CRISPR array (Medvedeva et al. 2019) (containing two spacers according to CRISPR-CasFinder, three spacers according to the JGI's standard operating pipeline) was annotated. When the free viral element,

v36_2016 (predicted as circular by VirSorter) was put through CRISPR-Cas Finder, it was predicted to contain the same Cas gene clusters and CRISPR array as LW2_146_Pigmentiphaga's putative prophage-containing scaffold. This could suggest that the prophage introduced CRISPR-Cas systems into its host's genome upon integration or that the CRISPR-Cas systems of this host were co-opted by its virus after excision from the host chromosome. The CRISPR array being outside the prophage boundary predicted by Prophage Hunter is likely due to partial degradation of the prophage over time. LW2_146_Pigmentiphaga has three connections to v36_2016 in the 2017-host-to-2016-virus network, all of which are predicted to be due to prophage-mediated immunity.

We next repeated the workflow described above to examine CRISPR arrays in the viral elements identified in our 2017 data - 91 (0.48%) were predicted to encode CRISPR arrays. Unlike what was observed with our 2016 viral elements, here there were 8 instances of shared CRISPR spacers between CRISPR-encoding viral elements.

Spacers encoded in 2017 viral CRISPR arrays were predicted to target 117 viral elements in the unclustered 2017 virus data set. Accounting for matches between viral CRISPR arrays, spacers encoded in 2017 virus CRISPR arrays were predicted to target 44 viral elements in the unclustered 2017 dataset. Four viral elements target by viral CRISPR arrays are implicated in our 2016-hosts-to-2017 viruses network and are targeted by a total of three hosts. None of these three host MAGs was predicted to contain a CRISPR-encoding provirus.

In our 2017 network, 10 viral elements targeted by viral CRISPR arrays were within the network and were targeted by a total of 9 hosts. Only one of the hosts targeting any of these 11 viral elements was predicted to contain a prophage: LW1_30_Synergistaceae. Homology-based methods and PHASTER, but not Prophage Hunter, identified putative prophage elements from a

single scaffold in LW1_30_ Synergistaceae. Though CRISPRCasFinder did not identify a CRISPR in these putative prophage regions, manual inspection of the JGI CRISPR annotations revealed a CRISPR within the scaffold from positions 658:788. The two spacers in this region are also responsible for the targeting of three viral elements, v159_2017, v91_2017, and v60_2017, by LW1_30_Synergistaceae in our 2017 host to 2017 virus network. Notably, v159_2017 is predicted to be the same as the viral element integrated within the MAG as a prophage. This CRISPR array, however, only had partial overlap with the 779:10068 prophage region predicted by PHASTER. Since provirus boundary prediction methods are still very limited in their accuracy, we conservatively refute that a CRISPR-encoding prophage is within the LW1_30 scaffold.

In summary, for the 2016 viruses in the 2016 network, no provirus-mediated interviral conflicts were identified. In the 2017-host-to-2016-virus network however, while no genuine interviral conflicts were identified, shared CRISPR spacers between proviruses and their free virus equivalents convoluted our virus-host networks. In the case of the 2017 viruses, both in the 2016-host-to-2017-virus network and 2017-host-to-2017-virus networks, no provirus-mediated interviral conflicts were found. We have shown that the presence of CRISPR-encoding viral elements can convolute virus-host networks: if a MAG contains a provirus that encodes a CRISPR array and the non-integrated equivalent of said provirus is also present in the metagenomic assemblies, then provirus CRISPR spacer to free virus CRISPR spacer matches could be what draws a host-virus connection. This could be further confounded by two distinct viral elements or other mobile genetic elements sharing a spacer between their CRISPR arrays.

Supplemental Tables and Figures

Table A.1: Putative cross-phylum and cross-domain interaction based on host MAG CRISPR spacers and viral element protospacer connections, with categories for curation listed. Grey shading indicates categories where a given interaction failed curation.

Putative hosts	Host MAG phylum (GTDB-tk)	Host CRISPR scaffold phylum (JGI, Contig Annotation Tool)	Host completion, contamination (%)	Host CRISPR scaffold lengths (bp)	# shared CRISPR spacers	Predicted viral element(s)	Overall
Cross-Phylum Infection Predictions							
LW2_138_Bacilli_2017	Firmicutes	Proteobacteria, Muirbacteria	72.12, 5.33	5,475	0	v2_2017 v154_2016	pass
LW2_137_Muirbacterieaceae_2017	Muirbacteria	Firmicutes, Muirbacteria	93.26, 4.56	50,129			
LW2_139_Absconditabacteriales_2017	Patescibacteria	Planctomycetes, Bacteria	70.53, 3.61	16,695			
LW2_36_Ottowia_2017	Proteobacteria	Unknown	95.56, 3.26	7,569	0	v102_2017	pass
GW1_223_Dojkabacteria_2017	Patescibacteria	Viral/mixed bacteria/Euryarchaeota, cellular organisms/no rank	80.02, 4.08	78,174			
LW2_146_Pigmentiphaga_2017	Proteobacteria	Proteobacteria, Bacteria	97.56, 0.30	89,231	0	v36_2016	pass
LW2_50_Polyangia_2017	Myxococcota	Proteobacteria	91.20, 4.52	87,942			
GW1_158_Paceibacteria_2017	Patescibacteria	Unknown, Candidatus Riflebacteria	82.89, 1.88	4,111	56	v42_2016	pass
LW4_67_Methylobacteria_2017	Proteobacteria	Proteobacteria	97.62, 3.88	15,464			
LW2_63_Bacteroidales_2016	Bacteroidota	Bacteroidetes, no rank	91.94, 0.90	21,376	0	v80_2016	fail
LW2_22_Acholeplasmataceae_2016	Firmicutes	Bacteroidetes, Bacteroidetes/Chlorobi	94.48, 3.01	3,322			
LW2_3_Acholeplasmataceae_A_2016	Firmicutes	Unknown	73.36, 0.69	4,167			
LW2_30_Dehalococcoida_2017	Chloroflexota	Proteobacteria, Bacteria	86.44, 5.02	4,005	0	v35_2017 v123_2017	fail
LW2_58_Limnochordia_2017	Firmicutes	Bacteroidetes/Proteobacteria, cellular organisms/no rank	82.80, 9.32	5,396			
Cross-Domain Infection Predictions							

CLC_82_Cloacimonadaceae_2017 CLC_33_Methanoculleus_2017	Cloacimonadota Halobacterota	Mixed bacteria/Euryarchaeota, Cloacimonetes Unknown	95.05, 3.09 96.62, 3.27	6,990 3,667	0	v165_2016	fail
LW3_15_Cloacimonadaceae_2017 CLC_33_Methanoculleus_2017	Cloacimonadota Halobacterota	Cloacimonetes Unknown	97.74, 6.19 96.62, 3.27	40,641 3,667	0	v178_2017 v147_2017	fail
CLC_92_Sphaerochaeata_2017 CLC_33_Methanoculleus_2017	Spirochaetota Halobacterota	Unknown Unknown	87.04, 4.21 96.62, 3.27	3,829 3,710	0	v172_2017 v181_2017	fail
CLC_110_Methanoculleus_2017 CLC_67_Massilibacteroides_2017	Halobacterota Bacteroidota	Euryarchaeota Euryarchaeota	76.43, 9.31 70.44, 9.71	15,438 10,142	0	v39_2017	fail

Tables A.2-A.7 are available as a single .xlsx file for download online (https://github.com/nikhilg123/Thesis_Chapter_two_supplemental_tables.git)

Table A.2: Host MAGs and predicted viral elements included in the 2016 host to 2016 viruses network, including number of spacer matches and coverages for both partners.

Table A.3: Host MAGs and predicted viral elements included in the 2017 host to 2017 viruses network, including number of spacer matches and coverages for both partners.

Table A.4: Host MAGs and predicted viral elements included in the 2016 host to 2017 viruses network, including number of spacer matches and coverages for both partners.

Table A.5: Host MAGs and predicted viral elements included in the 2017 host to 2016 viruses network, including number of spacer matches and coverages for both partners.

Table A.6: Networked MAGs and their predicted taxonomy, genome statistics, and NCBI accessions.

Table A.7: Connection between VirSorter scaffold ID, simplified name used in article text, and IMG accession for key predicted viral elements.

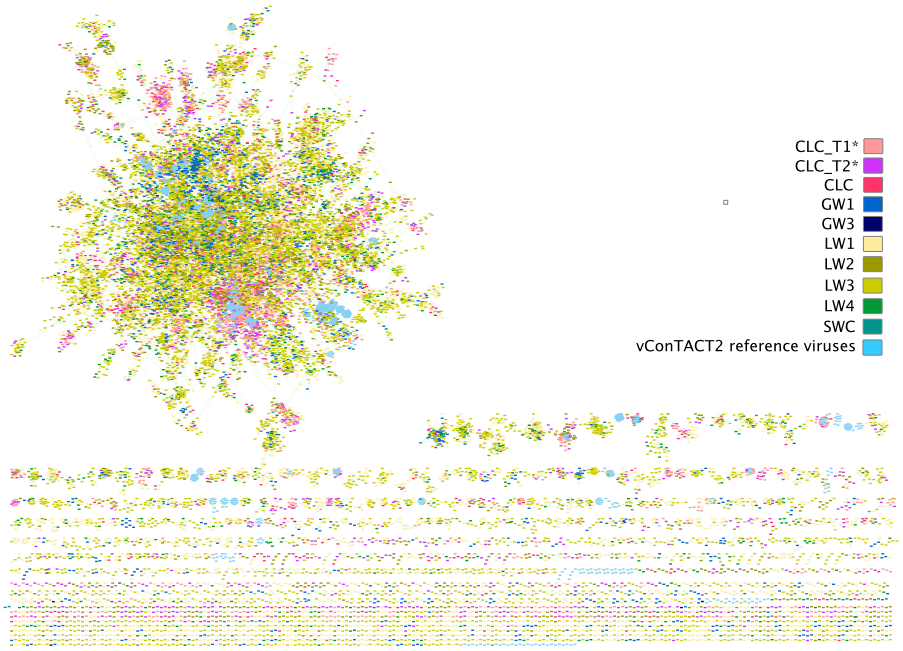


Figure A.1:Viral diversity across distinct samples. vConTACT2 was used to generate gene-sharing networks for all 2016 and 2017 viral elements. Edges connecting nodes are not shown. CLC samples have non-overlapping colors to distinguish the two 2016 time-points. The vConTACT2 reference database used was Prokaryotic Viral Refseq version 85 with ICTV-only taxonomy**.**

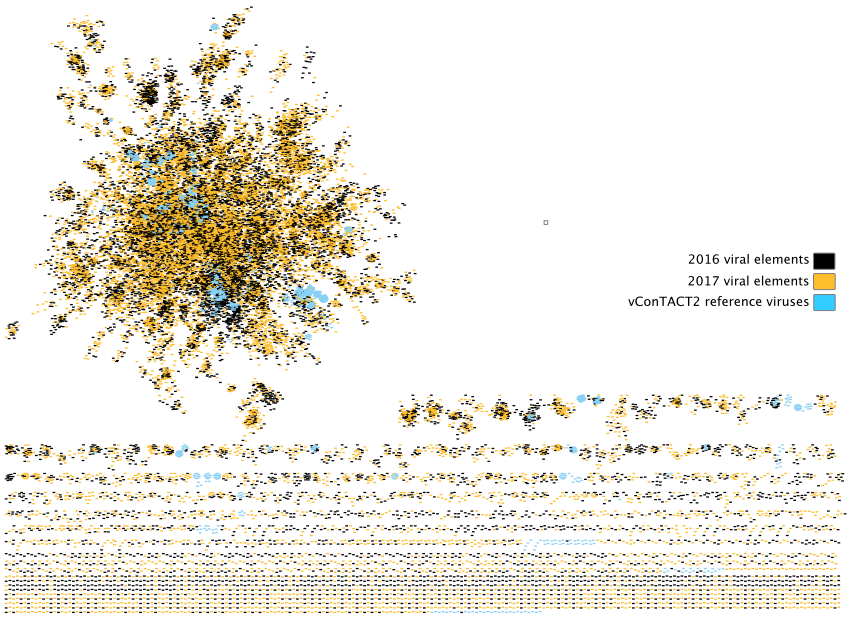


Figure A.2: Viral diversity across distinct timepoints. vConTACT2 was used to generate gene-sharing networks for all 2016 and 2017 viral elements. Edges connecting nodes are not shown. The vConTACT2 reference database used was Prokaryotic Viral Refseq version 85 with ICTV-only taxonomy**.**

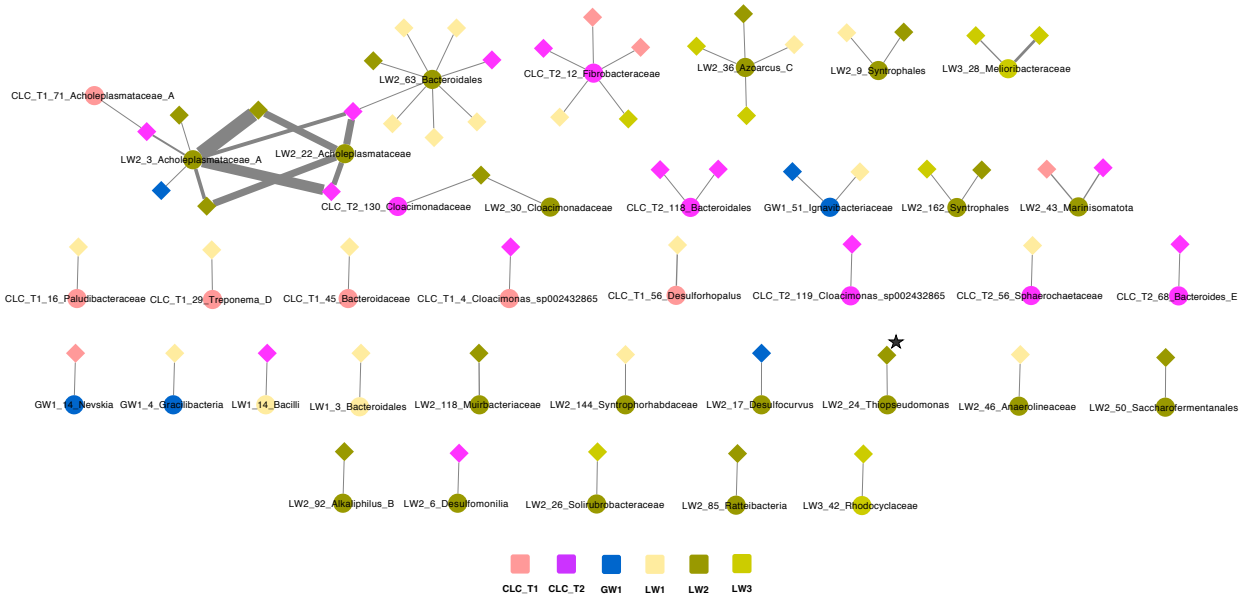


Figure A.3: Virus-host interactions in a Southern Ontario landfill (2016) with host MAG taxonomy. CRISPR array spacers from the MAGs (circles) were mapped against the viral elements (diamonds) predicted by VirSorter and VIBRANT to establish virus-host linkages (connecting edges). The network is colored by the geographic sampling location from which the MAG or viral element originated, with the width of the edges proportional to the number of spacer-protospacer matches supporting the connection. MAG identifier and lowest level of taxonomy as determined by GTDB-Tk are noted for networked MAGs. A viral element encoding a Diversity Generating Retroelement (DGR) is denoted by a star.

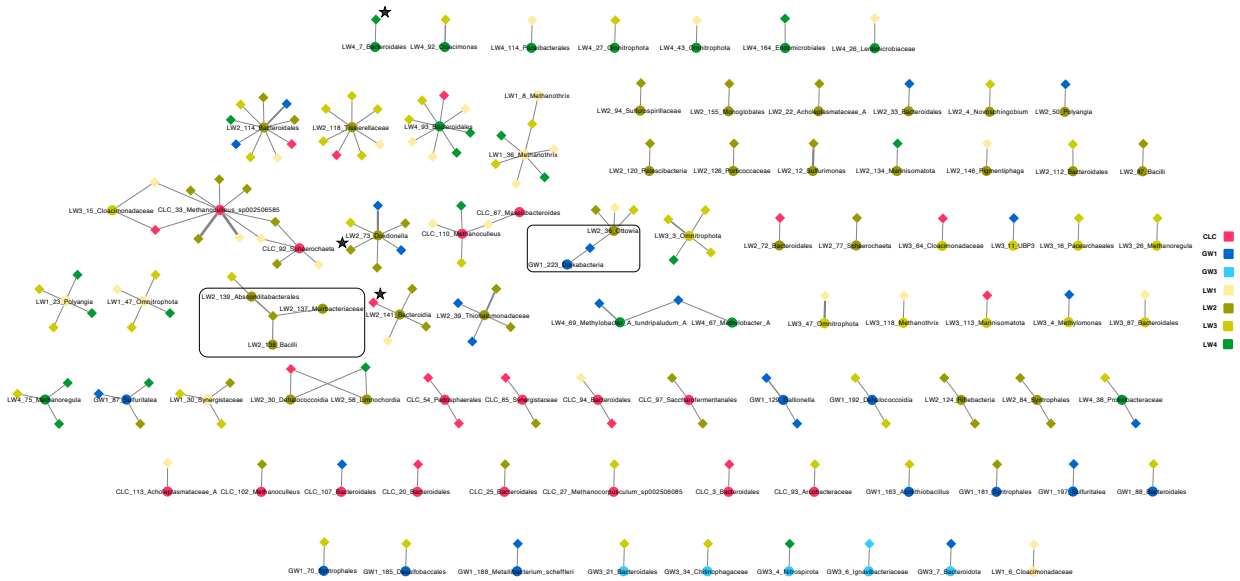


Figure A.4: Virus-host interactions in a Southern Ontario landfill (2017). CRISPR array spacers from the MAGs (circles) were mapped against the viral elements (diamonds) predicted by VirSorter and VIBRANT to establish virus-host linkages (connecting edges). The network is colored by the geographic sampling location from which the MAG or viral element originated, with the width of the edges proportional to the number of spacer-protospacer matches supporting the connection. MAG identifier and lowest level of taxonomy as determined by GTDB-Tk are noted for networked MAGs. Three viral elements encoding Diversity Generating Retroelements (DGRs) are denoted by stars. Viral elements and hosts involved in putative cross-phylum connections are highlighted with boxes.

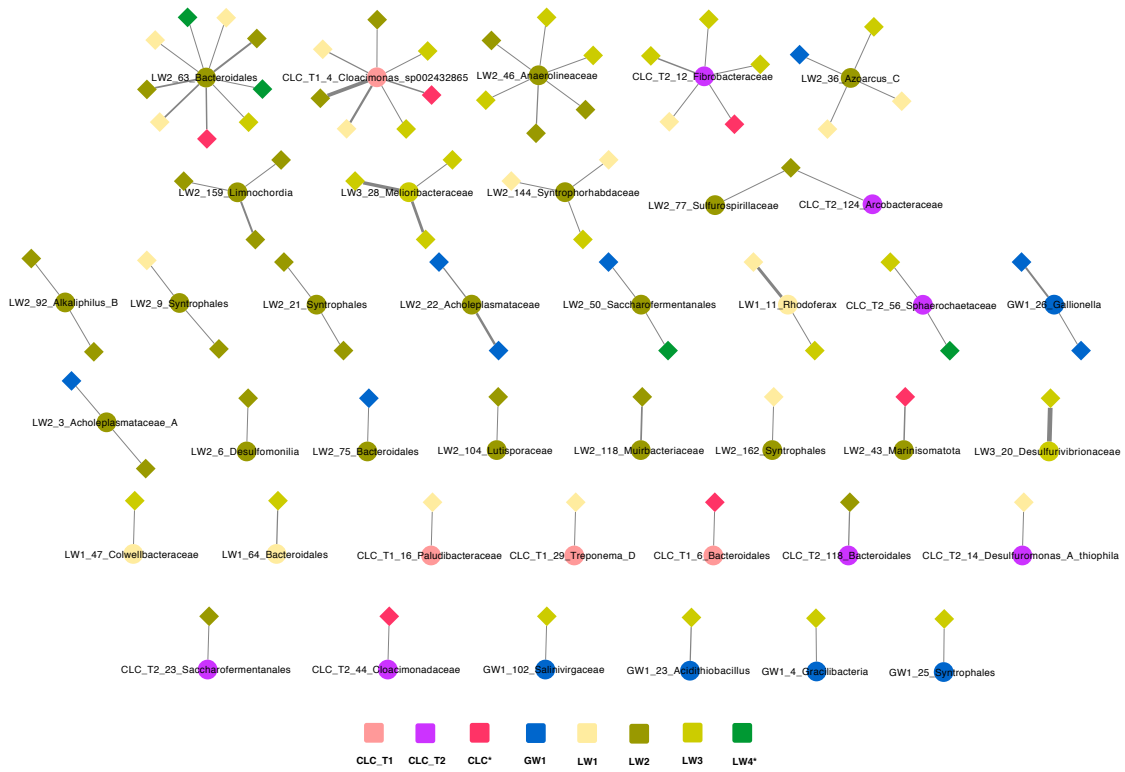


Figure A.5: Virus-host interactions in a Southern Ontario landfill (2016 hosts to 2017 viruses). CRISPR array spacers from the MAGs (circles) were mapped against the viral elements (diamonds) predicted by VirSorter and VIBRANT to establish virus-host linkages (connecting edges). The network is colored by the geographic sampling location from which the MAG or viral element originated, with the width of the edges proportional to the number of spacer-protospacer matches supporting the connection. MAG identifier and lowest level of taxonomy as determined by GTDB-Tk are noted for networked MAGs.

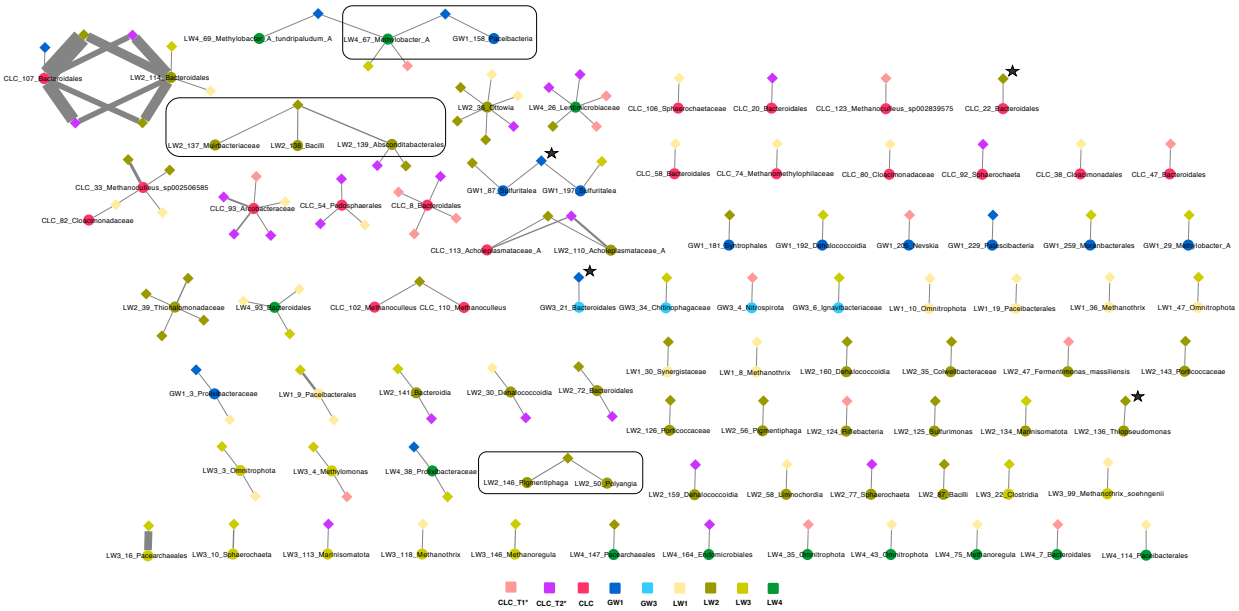


Figure A.6: Virus-host interactions in a Southern Ontario landfill (2017 hosts to 2016 viruses). CRISPR array spacers from the MAGs (circles) were mapped against the viral elements (diamonds) predicted by VirSorter and VIBRANT to establish virus-host linkages (connecting edges). The network is colored by the geographic sampling location from which the MAG or viral element originated, with the width of the edges proportional to the number of spacer-protospacer matches supporting the connection. MAG identifier and lowest level of taxonomy as determined by GTDB-Tk are noted for networked MAGs. CLC samples have non-overlapping colors to distinguish the two 2016 time-points. Four viral elements encoding Diversity Generating Elements (DGRs) are denoted by stars. All viral elements and hosts involved in putative cross-phylum connections are within boxes.

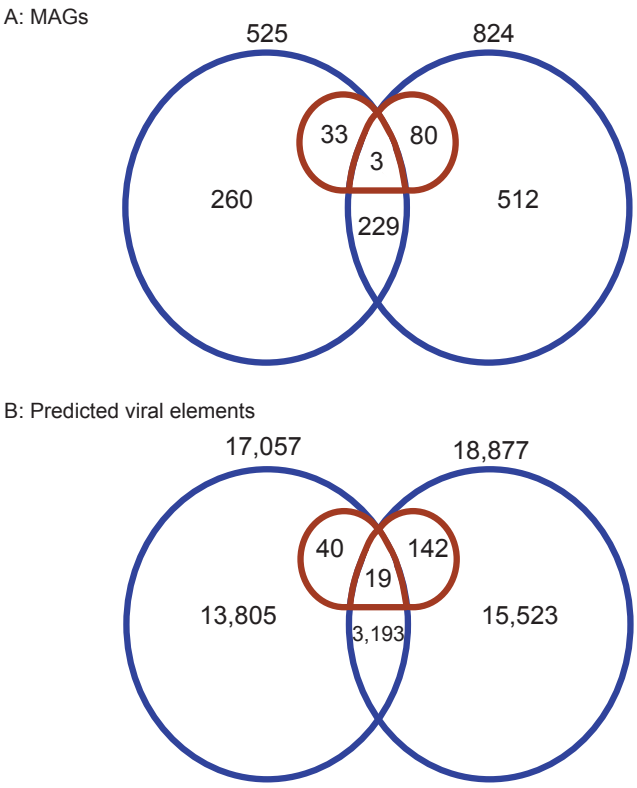


Figure A.7: Venn diagrams of A: all reconstructed MAGs from the 2016 (left) and 2017 (right) datasets, with dataset overlaps between all MAGs (blue) and networked MAGs (brown) indicated by the overlapping central region. B: all predicted viral elements from the 2016 (left) and 2017 (right) datasets, with dataset overlaps between all viral elements (blue) and networked viral elements (brown) indicated by the overlapping central region.



Figure A.8: CRISPR element targeting the same viral element between 2016 and 2017 showing changes in CRISPR array spacers (colored blocks) and one conserved flanking sequence (black rectangle). The direct repeat consensus sequence shared between the two arrays is shown under the flanking sequence and is underlined in purple. The similarity between a subsequence of the flanking sequence and the direct repeat consensus suggests that the flanking sequence may contain a partially-degraded repeat along with a 28 bp spacer upstream of it. Omitting the possible spacer within the flanking sequence, none of the spacers between the two arrays are shared despite near-identical host genomes from the CLC in 2016 and 2017. The backbone of this figure was generated through the use of CRISPRStudio (Dion et al. 2018).

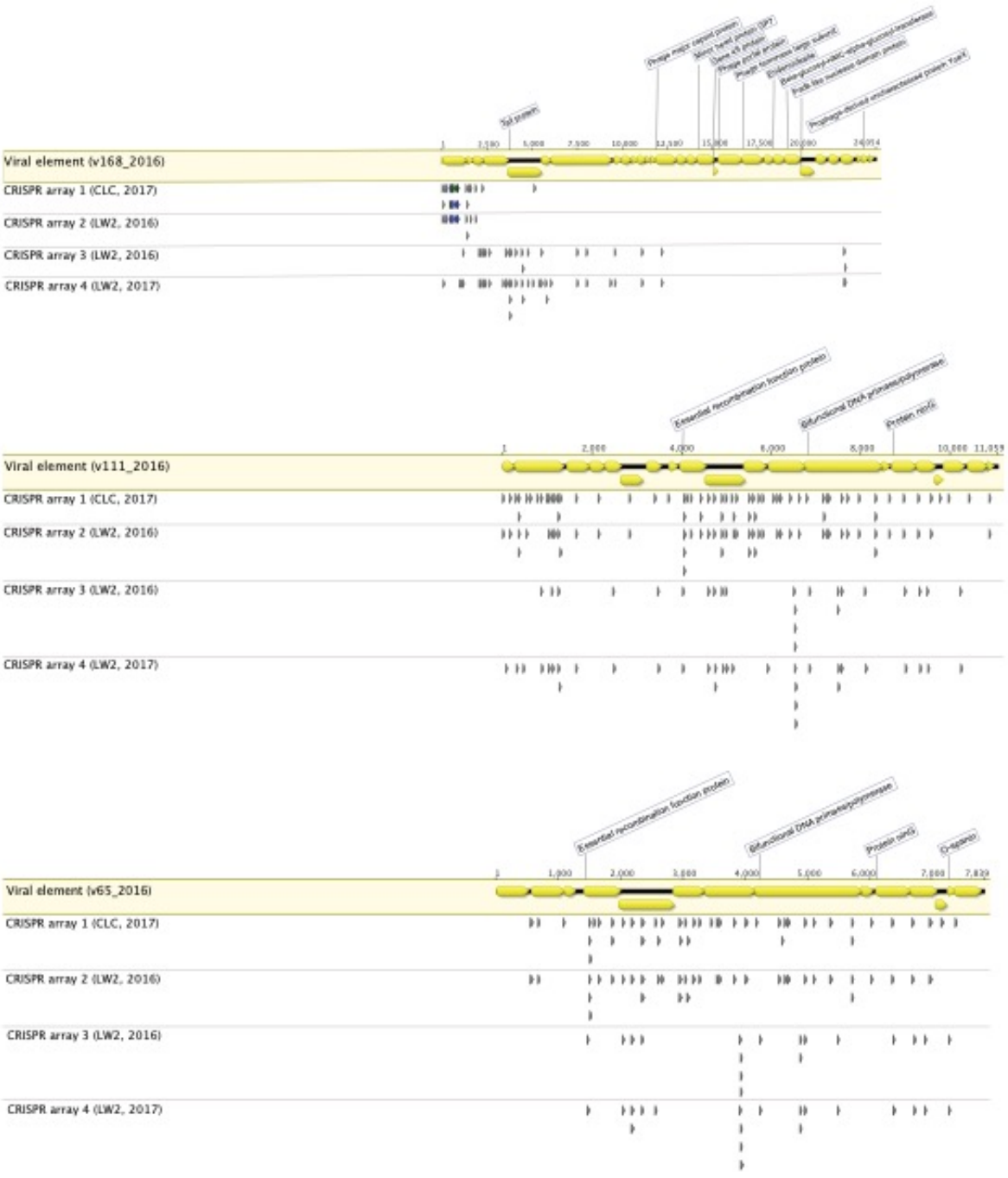


Figure A.9: Genetic features targeted by host CRISPR arrays on hyper-targeted viral elements. The coding sequences' annotations for the hyper-targeted viral elements are displayed. CRISPR spacers were mapped to regions of the viral element, with the mapping

and visualization performed in Geneious (Kearse et al. 2012). In the targeting profile of a single host, a non-grey triangle represents a single spacer that had multiple matches to regions of the viral element. When shared, these single spacers are coloured the same between distinct host targeting profiles.

Appendix B

Supplementary Material: Chapter 3

Table B.1: Putative cross-phylum host-virus interactions.

Sample set	Putative hosts	Host MAG phylum (GTDB-tk)	Host completion, contamination (%)	Number of host spacer to viral protospacer matches	Predicted viral element(s)
CA_2019	TPIn_75 TPBF_198	Desulfobacterota Proteobacteria	99.41, 0.00 90.75, 0.63	1 7	vMAG_518
NEUS_2019	STF2_137 STCSWMC_88	Bacteroidota Firmicutes_A	95.56, 3.26 80.02, 4.08	1 2	vMAG_1257
NEUS_2019	STCSWMC_93 STF1_64 STF2_19 STD2_245 STCSWMC_50 STCSWMC_25	Bacteroidota Bacteroidota Bacteroidota Bacteroidota Cloacimonadota Firmicutes_B	94.35, 0.27 95.43, 0.27 95.43, 0.00 88.71, 2.42 98.90, 2.20 90.15, 4.60	15 16 1 5 1 3	vMAG_3146
NEUS_2019	STF2_137 STCSWMC_88	Bacteroidota Firmicutes_A	95.56, 3.26 80.02, 4.08	1 2	vMAG_910
NEUS_2019	STF2_144 STCSWMC_50 STF2_148 STCSWMC_25	Cloacimonadota Cloacimonadota Cloacimonadota Firmicutes_B	100.00, 1.10 98.90, 2.20 95.54, 1.10 90.15, 4.60	1 2 2 2	NODE_3233_length_30680_cov_384.260596_Sandtown_F2||full
SO_2017	LW2_137 LW2_139	Muirbacteria Patescibacteria	93.26, 4.56 70.53, 3.61	2 4	vMAG_2310

Tables B.2 and B.3 are available as a single .xlsx file for download online

(https://github.com/nikhilg123/Thesis_supplemental_tables.git)

Table B.2: AMGs encoded across all datasets.

Table B.3: AMGs encoded by megaphage genomes.

Appendix C

Supplementary Material: Chapter 4

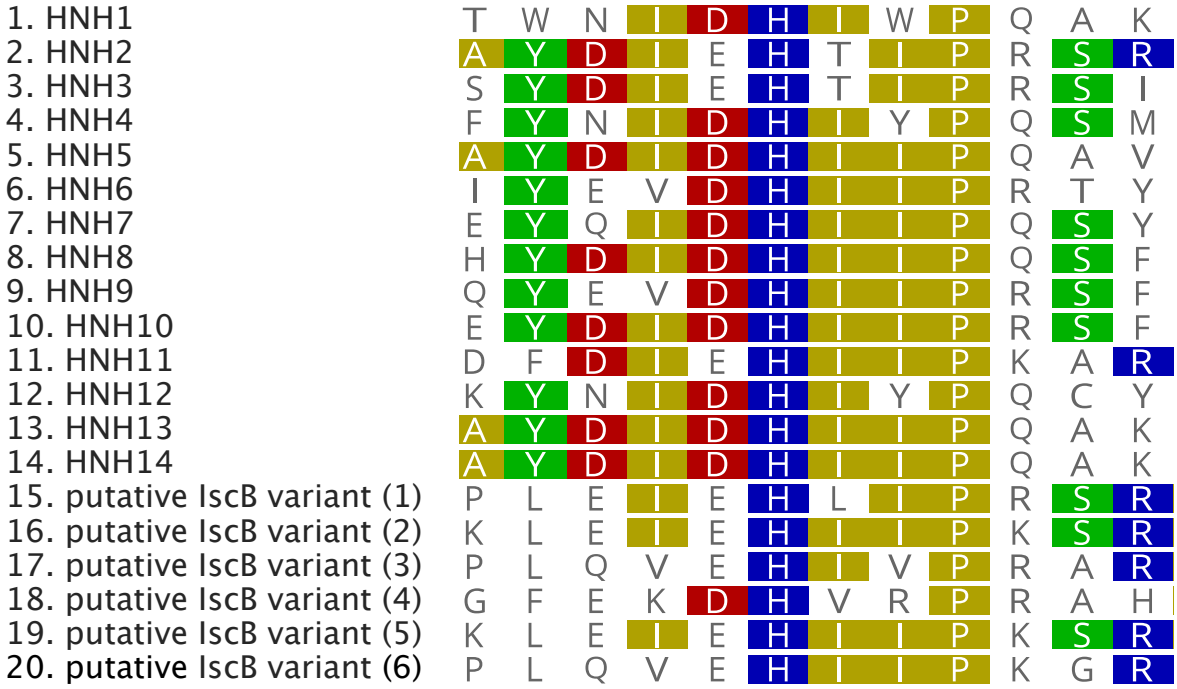


Figure C. 1: Partial HNH motifs in putative IscB variants. Sequences 1-14 are cas9 HNH motifs collected from a reference study (Al-Shayeb et al. 2020). Sequences 15-20 are predicted partial HNH motifs identified in variants of IscB or type II effector nucleases.

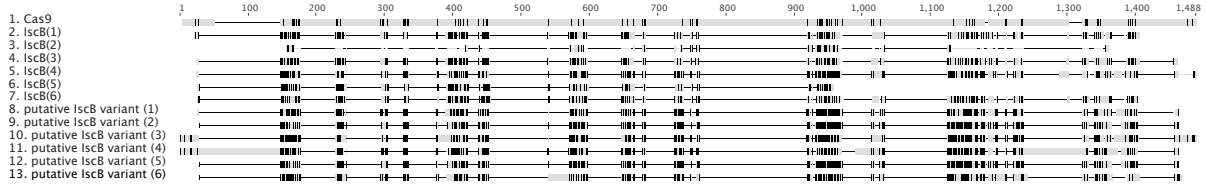


Figure C. 2: Predicted IscB variants from this study aligned with IscB and Cas9. *Streptococcus pyogenes* Cas9 is sequence number 1 (accession WP_011528583.1) and sequences 2-7 (6 sequences) are different variants of reference IscB proteins obtained from NCBI (accessions WP_059314693.1, WP_115451136.1, WP_155749208.1,

WP_173816280.1, and WP_224675911.1, respectively). Sequences 8-13 (6 sequences) were identified in this study based their scores against profile HMMs of Cas9 are predicted to be IscB variants.