AN ABSTRACT OF THE DISSERTATION OF

Anh Dieu Ha for the degree of Doctor of Philosophy in Integrative Biology presented on December 12, 2019.

Title: Orbiting the Genus of False Units: Dynamics of Phage Genomes and Nematode Microbiomes Associated with *Pseudomonas* spp.

Abstract approved:

_____

Dee R. Denver

Pseudomonas is a diverse, ubiquitous, and widely studied genus of bacteria. As *Pseudomonas* species occupy a wide range of niches in the ecosystem, they have made remarkable biological impacts. Better understanding of *Pseudomonas* biology, genetic diversity, and functional interactions with other organisms and the environment will provide valuable insights into our understanding of their roles in nature, and open doors to developing practical scientific applications involving these bacteria. Chapter 1 of this dissertation provided background information about *Pseudomonas* spp., and discussed strategies for investigating these and other bacteria in the postgenomic era. The first data chapter of this dissertation, Chapter 2, focused on the genomes of phages that infect *Pseudomonas* spp. bacteria; phages directly influence the diversity and evolution of their host bacteria, yet remain an understudied component of *Pseudomonas* biology. The diversity, phage host range evolution, and the evolutionary pressures acting on *Pseudomonas* phage genes were investigated using whole genome comparative and evolutionary analysis, focusing on 130 complete phage genomes. The results revealed immense genome sequence

variation, evidence that some phage possess potential to shift between host species, and purifying selection as the dominant evolutionary force acting on phage genes. In Chapter 3, a metagenomic approach was applied to detect and characterize phage DNA, including those that infect *Pseudomonas* spp. and other bacteria, focusing on a set of eight metagenomic data sets associated with two different plant-parasitic nematode species: *Globodera pallida* (potato cyst nematode) and *Heterodera glycines* (soybean cyst nematode). Three complementary bioinformatics strategies were deployed to analyze phage DNA in this study. Phage DNA was detected in all of the metagenomes, with varying patterns observed between the different focal nematode species. Strengths and weaknesses of the three different analytical strategies were considered in this chapter. One phage DNA sequence, with fixed length and highly similar DNA sequences and gene annotations, was detected at very high DNA sequence coverage levels in each of the eight metagenomes. Chapter 4 explored the dynamics of *Pseudomonas* spp. and other bacteria in microbiomes associated with a slug-parasitic nematode, *Phasmarhabditis hermaphrodita*, before and after nematodes infect slugs in a controlled laboratory setting.16S rRNA amplicon data analysis was conducted to investigate the microbial community composition, infer the microbiome diversity and level of dissimilarity between pre- and post-infection time points. Notable shifts in microbial community composition before and after infection was detected. Four 16S rRNA sequences classified as *Pseudomonas* expanded in post-infection samples during the assay, and constitute candidate bacterial contributors that might positively interact with *P. hermaphrodita* nematodes in the slug killing process. Chapter 5 discussed future study directions that will potentially advance the results of

this work. Altogether, this dissertation provides findings that may contribute to the integrated understanding of microbes and their dynamics, with special focus on *Pseudomonas* spp. and the organisms with which these bacteria interact. The dissertation research delivered new insights, with the potential to contribute to the future development of clinical and industrial applications, and genomic exploratory tools.

Orbiting the Genus of False Units: Dynamics of Phage Genomes and Nematode Microbiomes Associated with *Pseudomonas* spp.

by
Anh Dieu Ha

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented December 12, 2019
Commencement June 2020

Doctor of Philosophy dissertation of Anh Dieu Ha presented on December 12, 2019.

APPROVED:

Major Professor, representing Integrative Biology

Head of the Department of Integrative Biology

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Anh Dieu Ha, Author

ACKNOWLEDGEMENTS

This work was made possible owing to the tremendous support that I have received throughout my graduate program. First of all, I would especially like to express my profound appreciation to my advisor, Dr. Dee Denver for his guidance, support, positivity, wisdom, and inspiration. Dee is my role model of scientific thinking and creativeness, who constantly thinks outside of the box and motivates me to do so. I have learned so much from him - so many invaluable lessons in science, and in life: he always has the best positivity and the wisest advice ready to offer. I am deeply thankful that Dee is an incredibly supportive, thoughtful, and caring mentor. Ever since that day way back at the end of my first term at OSU, when he seriously sat me down and asked how I was doing, because his 'trainee's well-being is very important', I knew that I was going to be grateful for life.

I would like to thank my committee members Drs. Barbara Taylor, David Hendrix, Joyce Loper, and John Fowler for their thoughtful feedback, support, and kindness throughout my program. I heartily thank Barb for her brilliant comments, the (many) conversations about thinking science and about life, the warm hellos down the hallway, and the much-needed assuring hugs. I am grateful for Dave's bioinformatics expertise and his unwavering encouragement, which taught me so much, constantly inspires me, and really helped me pull through the many consecutive hours sitting in front of the computer screen. Many thanks to Joyce for the profound insights into the world of *Pseudomonas*, for your thoughtful advice, the kindest words, and smiles (and the yogurt!). Thank you, John, for teaching me how to

really think critically from the beginning, for the questions that urged me to think about what I have missed in the bigger picture, and the patience to listen. Thank you professors, for being the best committee that I could ever ask for.

I owe a debt of gratitude to the Denver Lab, past and present, for their support and motivation. I look up to Dana Howe for her marvelous thoughtfulness, positivity, attention to detail, and the wise advice she gave; and I could not thank her enough for the patience and guidance with new lab techniques, and for having my back. I would like to extend special thanks to my academic twin sister, Sulochana Wasala, for being there with me from day one as a loyal friend and a wonderful sister, for constantly calming me down when I whined about trivia, and for always telling me that "it's okay". I thank Subi for her energy and cheerfulness. Dr. Amanda Brown, Dr. Emily Bellis, Dr. Ian Morelan, Dr. Riana Wernick, and Danielle Tom made me feel welcome since day one, and have since offered the best companionship.

I would like to thank Dr. Shawn O'Neil for teaching me the first code writing experience, for the patience, extreme helpfulness, and motivation. I also thank the CGRB, especially Ed Davis, for their bioinformatics assistance and sequencing support.

My friends have been wonderfully supportive, and I am ever so grateful for their friendship. My special thanks to Chi for having been tirelessly and constantly on my side, for her advice and tips, and for never giving up. I thank Đần, HH, and Sò for the humor, gossips, arguments, and lots of laughs. I am grateful to my housemates at 660, past and present: Trang, Z, Trung, Kiệt, and Jonathan for making the house my home during the last five years. I thank my VEF Fellows chị Hà, anh Quân, and Ý for

the conversations (serious talks, confiding, and gossips), cheers, and encouragements. I also want to thank chị Phụng, Khôi, and Hùng for their friendship and humor, my college friends Phan Anh, Bi, and Thắng for the regular check-ins, and my 9H friends Thảo, Duy, Tùng, Long, and Tuất for the emotional support.

Lastly, I would like to thank my father, the scientist who seeded the idea of pursuing science and has supported my every decision – no matter how unexpected they were. My mother and little sister, who have been patiently handling my ups and downs, unconditionally staying by my side every single step of the way though we are thousands miles apart. And finally, I thank the incredible Anh Ninh for being my anchor.

CONTRIBUTION OF AUTHORS

Chapter 2: Anh Ha helped design the study, performed all of the analyses, and led in the preparation of the manuscript. Dee Denver helped design the study, contributed to revision, read and approved the final manuscript. Dee Denver conceived the study and critically revised the manuscript.

Chapter 3: Anh Ha helped design the study, performed all of the analyses, and led in the preparation of the manuscript. Dana Howe, Sulochana Wasala, Cedar Hesse, Amy Peetz, and Inga Zasada contributed to nematode metagenome sequencing and assembly efforts. Melissa Mitchum provided *Heterodera glycines* nematodes for analysis. Louise-Marie Dandurand provided *Globodera pallida* nematodes for analysis. Dee Denver conceived the study and critically revised the manuscript.

Chapter 4: Anh Ha helped design the study, contributed to bacteria sample preparation, performed all of the analyses, and led in the preparation of the manuscript. Dana Howe contributed to sample preparation and DNA extraction. Andrew Colton and Rory Mc Donnell led in the slug infection assay and contributed to bacteria sample preparation. Dee Denver conceived the study and critically revised the manuscript.

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

LIST OF FIGURES (Continued)

Page

LIST OF TABLES

Table Page

LIST OF APPENDIX FIGURES

Figure Page

LIST OF APPENDIX TABLES

Table Page

LIST OF APPENDIX DOCUMENTS

Document Page

DEDICATION

This dissertation is dedicated in loving memory to my grandmother, to whom I am eternally grateful for the love throughout our time together and beyond. Thank you for always being my guardian angel, guiding me through thick and thin. And, most importantly, *thank you for always making sure I finish my food*. I miss you tremendously, all the time.

Orbiting the Genus of False Units:
Dynamics of Phage Genomes and Nematode Microbiomes Associated with *Pseudomonas* spp.

Chapter 1: Introduction

The genus *Pseudomonas,* with 255 species currently recognized (Parte, 2018), is the largest genus of Gram-negative bacteria, and has successfully colonized a wide variety of ecological habitats. First classified at the end of the 19th century, *Pseudomonas* has since been recognized as one of the most diverse, dynamic, tolerant, and widespread bacterial genera. In each of their ecological niches, *Pseudomonas* bacteria form diverse interactions with other organisms and collectively exert considerable influence across the biosphere. This chapter discusses the background overview of the genus *Pseudomonas*, their remarkable diversity, an important source of their genetic variation - bacteriophages, and a biocontrol application in *Pseudomonas*-relevant framework.

Part I: History, Diversity, Genomics, and Applications of *Pseudomonas* spp.

Bacteria of the Genus Pseudomonas*: History and Classification*

Pseudomonas bacteria were initially identified as a new genus by Walter Migula in 1894 (Migula, 1894), when he provided a general description of their phenotypic traits e.g. aerobic metabolism, rod-like shape, and motility by polar flagella. The term *Pseudomonas* was coined at the time without clarification of the etymology, and some thought that Migula intended to reflect the bacteria's resemblance to the nanoflagellate Monas (Palleroni, 2010). The genus name was officially explained as a derivation from the Greek *pseudes* ("false") and *monas* ("unit") in Bergey's Manual of Systematic Bacteriology, 7th edition (Bergey and Breed, 1957). At first, due to the genus' prevalence and the lack of a clear set of determinative taxonomic keys, numerous species were rapidly isolated and assigned to the genus by various bacteriologists - incorrectly in many

cases, and by the 1910s the number of members had reached one thousand (Cornelis, 2008).

Later, more modernized methodologies were applied to resolve the complex taxonomy of *Pseudomonas*. Efforts to revise and reclassify *Pseudomonas* bacteria had developed well beyond the initial simple observation of phenotypic appearance, using more focal criteria such as the remarkable nutritional versatility (Jong, 1926; Stephenson, 1939), physiological and biochemical properties e.g. growth factor requirements, the capability to accumulate poly-β-hydroxybutyrate, oxidase , produce extracellular enzymes, and benzene ring cleavage mechanisms (Stanier et al., 1966). Early investigation of the bacterial genome involved describing the DNA base composition of DNA in different *Pseudomonas* species (Colwell and Mandel, 1964; Mandel, 1966).

Further DNA studies focused on determining sequence similarity and comparison among *Pseudomonas* spp. genomes by nucleic acid hybridization, both DNA-DNA and DNA-RNA, which was a recommended method for distinguishing bacterial species at the time. The work using DNA-rRNA hybridization assigned *Pseudomonas* species into five large groups (Palleroni et al., 1973). Later, the five groups were found to be distributed over at least seven different genera across the class *Proteobacteria* (De Vos et al., 1989; De Vos and De Ley, 1983), and the remaining *Pseudomonas* 'sensu stricto' were grouped in the rRNA group 1, and under the subclass *Gammaproteobacteria.* 16S ribosomal RNA gene sequences were then directly employed as a taxonomic tool to cluster species, commencing the transformation of the genus' taxonomy (Anzai et al., 2000; Moore et al., 1996). Anzai and colleagues performed a detailed taxonomic revision for the genus using the 16S rRNA sequences of 128 strains then considered *Pseudomonas*, out of which 57

strains were found to be *Pseudomonas* ‘sensu stricto’ and arranged into seven clusters (Anzai et al., 2000). In the meantime, the classification approach that characterized bacterial macromolecules, compounds, and metabolites were still be applied. For example, species determination and grouping were performed based on isoelectric focusing (IEF) patterns of pyoverdines, a fluorescent siderophore pigment in iron-limiting environment (Meyer et al., 2002), the ability to produce various pigments that becomes fluorescent under UV wavelength (Palleroni and Moore, 2004), the capacity of forming biofilm, which could help escaping phagocytosis by white blood cells, and broad-range antibiotic resistance (Ryan et al., 2014).

Larger-scaled genomic approach to resolve **Pseudomonas** *taxonomy*

Although the 16S rRNA has become a classic marker for taxonomic analyses, it was suggested that the information from a single orthologous gene is not sufficiently effective to resolve taxonomy at the species level (Janda and Abbott, 2007; Mignard and Flandrois, 2006). For example, the type strains of three different *Edwardsiella* species may exhibit up to 99.81% similarity in 16S rRNA sequences, despite the clear divergence in their biochemical characteristics and DNA hybridization results that emphasized the species boundary (Janda and Abbott, 2007). Therefore, to distinguish closely related species or subspecies, and at least one additional housekeeping gene is necessary. Many other gene markers were also found to be informative for taxonomic studies, such as *gyrB, rpoD, gacA, atpD, carA,* and *recA* (Ait Tayeb et al., 2005; De Souza et al., 2003; Hilario et al., 2004; Yamamoto et al., 2000). The multilocus sequence analysis (MLSA) approach characterized bacteria and inferred phylogenetic relationships based on the DNA

sequences of the 16S rRNA gene and multiple (typically four to eight) other protein-coding genes (Frapolli et al., 2007; Gomila et al., 2015; Mulet et al., 2010, 2012) was demonstrated to be reliable for recognizing species with less defined boundaries (Almeida et al., 2010; Gevers et al., 2005). With the aid of new collective data, many species have been reclassified to other genera, for example *Delftia, Aminobacter, Burkholderia, Acidovorax, Lysobacter, Sphingomonas,* and *Ralstonia* (Denner, 1999; Loveridge et al., 2017; Miess et al., 2016; Urakami et al., 1992; Wen et al., 1999; Yabuuchi et al., 1992, 1995). The genus *Pseudomonas* was then divided into 11–13 groups (Gomila et al., 2015; Mulet et al., 2010).

As Next Generation Sequencing technologies rapidly advanced, complete genomes of many *Pseudomonas* species and strains have been sequenced, including *P. aeruginosa* PAO1 (Stover et al., 2000), *P. putida* KT2440 (Nelson et al., 2002), *P. syringae* DC3000 (Buell et al., 2003), *P. protegens* Pf-5 (Loper et al., 2007; Paulsen et al., 2005), and *P. stutzeri* A1501 (Yan et al., 2008). As of November 2019, the total number of draft and complete genomes of *Pseudomonas* spp. available in the *Pseudomonas* Genome Database (Winsor et al., 2016b) has exceeded nine thousand, providing valuable resources to study the sequence composition, evolution, and taxonomic relationships in the genus.

Complete genome sequences allow access to all housekeeping genes that are taxonomically informative, and genome-wide analysis may provide a finer resolution of species assignment compared to MLSA. Thus, classification approaches have substantially broadened to genome-wide scale and included a large number of type strains for a more systematic and robust inference. In their 2018 study, Hesse and colleagues

analyzed 166 genomes of *Pseudomonas* type strains, representing 163 species and 3 subspecies, to construct a phylogeny for the genus using the protein sequences of 100 orthologous genes (Hesse et al., 2018). Thirteen groups of *Pseudomonas* were identified from the phylogeny. They also recognized 189 clusters of *Pseudomonas* spp. available in public database that show clear distinction to the existing type strains, which may indicate novel species. In addition, sequence comparison suggested yet another re-evaluation of *Pseudomonas* species assignment, as the genomic data have pointed to several misclassified strains, as well as strains of other genera (i.e. *Azotobacter* and *Azomonas*) that may actually belong to *Pseudomonas*. Resources from genomic data are, therefore, powerful tools to advance the general knowledge of *Pseudomonas* biology, taxonomy, and relationships.

Sources of genomic diversity in* Pseudomonas *bacteria

Genome sequences indicated an enormous diversity among bacterial species of the genus *Pseudomonas*, with a core set of genes that are shared between all species accounting for only about 25% to 35% of each genome (Loper et al., 2012). It was estimated that the core genome size shared among ten strains of the *P. fluorescens* group is 2,789 genes. For five *P. syringae* and *P. putida/entomophila* strains the core genome sizes are 3,456 and 3,185 genes (Loper et al., 2012). Five *P. aeruginosa* strains examined in the work (Mathee et al., 2008) share 5,021 common CDSs, however when one strain of *P. fluorescens* were added to the comparison, the core genome of this *Pseudomonas* spp. set is reduced to 1,836 genes. (Hesse et al., 2018) found that the core genome across the 166 *Pseudomonas* type strains contain only 794 genes, and the pan genome size increases

with the number of genomes added to analysis, which is to be expected. This observation suggested the divergence of *Pseudomonas* species' genome sequences.

Among these 166 genomes, genome sizes vary from 3.03 to 7.38 Mbp, the number of putative coding sequences in each genome range from 2,877 to 7,113, and GC content vary from 48% to 68%. The variation in *Pseudomonas* genomes could be explained by many factors, including environmental conditions (e.g. species interactions, population competition, periodic mortality events, and energy input) (Mehrabi et al., 2016; Spiers et al., 2000), genetic causes (mutation), and active horizontal gene transfer (HGT) activities. Gene exchanging may be promoted by transposons, plasmids, integrons and prophages (Nakamura et al., 2004). Plasmids are common in bacterial cells, can have a broad host range, and often encode beneficial pathways, which enable their transfer and persisting in new host cells. Other genetic mobile elements are frequently reported as causes of the diversity and the ability to adapt of *Pseudomonas* bacteria (Garrido-Sanz et al., 2016; Silby et al., 2011). Prophage, as an additional accessory, adds to the diversity of the host bacteria, and prophage sequences themselves usually varied between genomes (Ceyssens and Lavigne, 2010).

Studies of* Pseudomonas *bacteriophages contribute to the understanding of their host bacteria

As the origin of prophages in bacterial genomes, *Pseudomonas* phages are considered an important factor shaping the diversity and evolution of *Pseudomonas* bacteria through many mechanisms, actively contributing to HGT events and alterations of host virulence (Hayashi et al., 1990; Sano et al., 2004). Research of phages has been applied broadly in

epidemiological studies of *Pseudomonas* diseases over the last 50 years, including the identification and sub-classification of pathogenic strains (Bergan, 1978; Bernstein-Ziv et al., 1973; Lindberg and Latta, 1974), DNA transduction to introduce genes to bacterial cells (Darzins and Casadaban, 1989; Rolain et al., 2011), and therapeutic strategy for antibiotic-resistant *Pseudomonas* infections (Hraiech et al., 2015; Pires et al., 2015; Yayan et al., 2015). Therefore, *Pseudomonas* phages provide a valuable resource to study the driving evolutionary force influencing their host bacteria, and improved understanding of these phages may contribute to a broader understanding of *Pseudomonas*.

Given the important role of *Pseudomonas* phages, many studies focused on gaining new insights into the phenotypic and genotypic characteristics of selected phage strains that are prevalent and/or of clinical and biocontrol relevance e.g. Pf1, DMS3, ϕKMV, ϕKZ, and EL infecting *P. aeruginosa,* PPpW-3 infecting *P. plecoglossicida,* and ϕIBB-PF7A infecting *P. fluorescens* (Budzik et al., 2004; Ceyssens et al., 2008; Hertveldt et al., 2005; Knezevic et al., 2015; Lavigne et al., 2003; Park et al., 2000; Sillankorva et al., 2008a). ϕIBB-PF7A belongs to the *Podoviridae* family and has a very short, noncontractile-tail, an icosahedral capsid, and a dsDNA genome of about 42 kb (Sillankorva et al., 2008a). The phage is capable of removing *P. fluorescens* biofilms and thus a promising sanitation agent to control spoilage in dairy and other food products (Sillankorva et al., 2008b). The *P. aeruginosa* phage ϕKMV is considered a *Pseudomonas* phage model species. ϕKMV also belongs to the *Podoviridae* family and shares several morphological traits and genome structure with the classic coliphage T7 (Lavigne et al., 2003). The genome of ϕKMV is linear, double-stranded DNA with

relatively high GC content at 62%. As the phage is strongly lytic towards *P. aeruginosa* isolates and have a relatively narrow host range (Ceyssens et al., 2011), it has attracted attention as a potential candidate for phage therapy. Nevertheless, within the host species, Ceyssens and colleagues found that ϕKMV could infect several different strains, which suggests the potential nucleotide diversity that the phage may facilitate between their intra-species host strains when they shift across these hosts. Likewise, by tracing the host range evolution of phages, one might be able to partly explain and/or anticipate the diversity between their host bacterial strains and even species.

The first chapter of this dissertation aimed to further understand the diversity, host range evolution and selection pressure acting on phages infecting *Pseudomonas* bacteria by investigating a set of 130 phage complete genome sequences available in public databases. The collection of *Pseudomonas* phage genomes was subsequently expanded to 306 sequences and used in the second chapter, along with other phage complete sequence databases, and complementary genomic approaches to search and examine for the presence of phage DNA signatures in nematode-derived metagenomic data sets.

Pseudomonas *bacteria occupy a broad range of ecological niches*

Mediated in part by phages' ability to diversify the genome reservoir for *Pseudomonas* spp., members of this bacterial genus have adapted to a wide variety of ecological habitats: water, soil, decaying matter and in associations with fungi, plants, animals, and humans. The most well-known *Pseudomonas* species is the opportunistic human pathogen *P. aeruginosa*, the leading cause of pneumonia and acute nosocomial infections, which poses a serious challenge to human patients with burn wounds, cancer,

and cystic fibrosis (Iglewski, 1996). Other species are pathogenic to ayu fish (*P. plecoglossicida* - Nishimori et al., 2000) and mushrooms (*P. costantinii* - Munsch et al., 2002). *P. syringae*, with more than 50 identified pathovars infecting diverse plant species, is a major, well-studied plant pathogen (Benson et al., 2007; Morris et al., 2008). *P. salomonii* and *P. palleroniana* are pathogenic on garlic and rice, respectively (Gardan et al., 2002). Meanwhile, the bacteria *P. protegens* and *P. chlororaphis* are characterized as 'Plant Growth Promoting Rhizobacteria' for their ability to protect plants against pathogens by competitive colonization (Andreolli et al., 2019), production of antifungal compounds and antibiotics (Calderón et al., 2015; Raaijmakers et al., 2010), and alternation of auxin level which benefits plant growth (Kang et al., 2006). Studies have reported the presence of *P. thermotolerans* in industrial cooking water, which could survive the temperature of 55 °C (Manaia and Moore, 2002), *P. guineae* in Antarctic soil (Ria Bozal et al., 2007), *P. marincola* in isolates from marine environments (Romanenko et al., 2008), and *P. duriflava* in desert soil (Liu et al., 2008).

Pseudomonas *as biocontrol agents*

Living in very diverse ecological niches, *Pseudomonas* spp. can be found in numerous types of biological associations, competitions or rivalries with a vast number of organisms in the ecosystem. Due to their extremely diverse interactions with different kinds of partners and hosts, it is perhaps not surprising to frequently find *Pseudomonas* names listed in biocontrol frameworks, where the relationship between two or more species is exploited for plant disease and pest management purposes. *Pseudomonas* bacteria have been utilized as a preventive method to restrict certain crop pathogens since

the 1980s (Raaijmakers et al., 1999; Schippers et al., 1987; Weller, 1988). Studies suggest that these bacteria produce pathogen-suppressing antibiotics, compete and limit the growth of pathogenic microorganisms, and promote plant resistance to diseases (Haas and Défago, 2005; Ramette et al., 2011; Weller, 2007). A considerable number of strains also engage in antifungal activities, e.g. *P. chlororaphis* (Chin-A-Woeng et al., 2000), *P. syringae* B359 (Fogliano et al., 2002), and *P. fluorescens* (Gaffney et al., 1994). The species *P. aurantiaca* has been reported to produce a metabolite with antibiotic activity against some Gram-positive bacteria (Esipov et al., 1975).

At least two *Pseudomonas* bacteria - *P. fluorescens* and *P. paucimobilis* - were found in association with the slug-parasitic nematode *Phasmarhabditis hermaphrodita* (Wilson et al., 1995b). The nematode *P. hermaphrodita* was first isolated in 1859 (Schneider, 1859), subsequently developed into a biological molluscicide against a broad range of slug families (Grewal et al., 2003; Rae et al., 2007) and have been sold commercially under the name Nemaslug® in 15 European countries since 1994 (Pieterse et al., 2017). In the production of Nemaslug®, *P. hermaphrodita* are specifically grown on the bacteria *Moraxella osloensis*; however, studies have suggested that the nematodes may associate with other bacteria (Rae et al., 2010; Wilson et al., 1995b, 1995a). *P. hermaphrodita* nematodes reared on *P. fluorescens* isolate 141 were shown to be pathogenic to grey field slugs while those cultured on several other bacteria were not (Wilson et al., 1995a). Nevertheless, the relative role of *Pseudomonas* in causing slug mortality, if there is any, is still unknown. The third chapter of this dissertation focused on *Pseudomonas* bacteria, along with two other bacterial communities, in the context of their association with nematodes in a biocontrol method against grey field slugs.

Part II: Strategies for Investigating Microbes and their DNA in the Post-Genomic Era

Next Generation Sequencing and microbial study

In the past two decades, the advent of Next Generation Sequencing (NGS) technologies has revolutionized biological research programs, bringing forth unprecedented amounts of DNA sequence data for investigators to examine and decipher. NGS platforms generate many millions of DNA sequence reads in parallel, and offer multiplexing approaches whereby many different biological samples can be analyzed in a single run. These new and continuously evolving technologies offer significantly reduced time investments and technical laboratory involvement as compared to first generation Sanger sequencing, which typically offered small numbers of reads per run (Behjati and Tarpey, 2013).

Currently, available sequencing platforms include Ion Torrent, Pacific Biosciences (PacBio), Oxford Nanopore, and Illumina (Solexa) sequencing, each requires specific DNA pre-processing procedures to serve different strategies of sequence reading (Buermans and den Dunnen, 2014; Kozińska et al., 2019). The Ion Torrent platform employs semiconductor technology, which detects the incorporation of each new nucleotide on DNA strands during synthesis not based on fluorescent signals, but on the release of protons (H+). DNAs are amplified by emulsion PCR on bead particles within proton-sensing wells, which detects the release of H+ by changes in pH (Rothberg et al., 2011). PacBio's strategy relies on single-molecule real-time sequencing, without

amplification step (Eid et al., 2009). DNA polymerases were bound to a DNA template strand, attached to the bottom of zero-mode waveguide (ZMW) wells, and allowed to perform uninterrupted DNA synthesis using fluorescently labeled nucleotides. As each new nucleotide is incorporated, its fluorescent signal is detected in real time by a sensitive optical laser system and a confocal recording system. Oxford Nanopore sequencer also produces single, long reads in real time without amplification or chemical labeling (Ashkenasy et al., 2005). The platform tunnels the DNA or RNA molecule of interest through a nano-sized protein pore that separates two compartments, and monitors the temporary changes in the electric current between these two compartments as the nucleic acid molecule is passed through the nanopore. Illumina technology performs sequencing by synthesis entirely on a flow cell surface, using fluorescently labeled nucleotides to read the sequence of the growing DNA chain (Bentley et al., 2008). In Chapters 3 and 4 of this dissertation, shotgun metagenomic data (Chapter 3) and 16S rRNA amplicon sequencing data (Chapter 4) were generated on Illumina platform.

NGS can be used for a variety of applications that include sequencing complete target genomes or a specific gene/genome section of interest, or characterizing DNA profiles in environmental samples. The versatility of the approach enables the investigation of the genetic makeup of organisms, especially microbes previously difficult to study, by facilitating the acquisition of accurate reference genomes for microbial identification and genomic analysis, detecting low abundance members of the microbiome, determining molecular pathways, and profiling of the bacterial community composition.

Whole-genome sequencing and comparative genomics

Whole-genome sequencing constructs the complete nucleotide sequence of genomes, which is widely utilized to achieve a comprehensive view of the genetic information and gene characteristics within the genomes (Hobert, 2010). As sequencing processes are constantly improved to be faster and at a lower cost, an immense amount of whole-genome data of all organisms and living entities have been continuously generated and made available on public databases. The genomic data can be used for genome-wide comparative genomics to explore the similarities and diversity among different organisms (Brüssow and Hendrix, 2002) e.g. identify syntenic regions, sequence mosaicism, and gene acquisition; and furthermore to infer their relationships and evolutionary pattern. In their 2018 work, Hesse and colleague used whole-genome comparison to demonstrate a rarefaction curve of pan genome size from sequence comparisons and evaluate the current sampling effort and predict the comprehensive diversity of the *Pseudomonas* genus (Hesse et al., 2018). In term of taxonomy, whole-genomic analysis may complement and confirm the findings of 16S rRNA gene analysis. It has been noted that the phylogenetic trees based on 16S rRNA gene and the whole genomic trees are highly similar (Bansal and Meyer, 2002). Chapter 1 of this work focuses on whole-genome analysis of 130 phage complete sequences to investigate the diversity, potential of host shifting and phage gene evolution.

16S rRNA amplicon approach

16S rRNA amplicon sequencing is a method commonly used to investigate microbiome genomics to investigate the composition, diversity and dynamics of bacterial

communities, including uncultured bacteria. The approach targets and sequences all the 16S ribosomal RNA genes in the samples of interest. The 16S rRNA marker gene is present in virtually all living organisms, evolves at a slow rate (Woese and Fox, 1977) but may still carry sufficient differences to identify the bacterial origin down to the Genus level (sometimes even Species and Subspecies levels) and calculate statistically valid diversity measures. The nucleotide divergences are mainly located in the nine "hypervariable regions" V1-V9 (Van de Peer et al., 1996) and these regions are usually flanked by sequences conserved among bacteria. Pre-designed universal primers can be annealed to these conserved regions, thus allowing PCR amplification of the target hypervariable sequences in almost all bacterial phyla (Baker et al., 2003; Lane et al., 1985; McCabe et al., 1999). Each hypervariable region may have different level of diversity among taxa (Mills et al., 2006), and it has been suggested that no single region could reliably distinguish community components (Chakravorty et al., 2007). Therefore, the simultaneous use of multiple hypervariable regions is recommended for general 16S rRNA analysis. The amplification of the V3-V4 region was standardized in the Illumina MiSeq 16S sequencing library preparation protocol (Illumina, 2013), and probably is most commonly used in 16S rRNA microbiome diversity analysis (Klindworth et al., 2013).

The 16S rRNA amplicon sequences obtained from amplification and sequencing are assigned into Operational Taxonomic Units (OTUs) i.e. clusters of sequences that share 97% similarity (Stackebrandt et al., 1994) or exact Amplicon Sequence Variants (ASVs) (Callahan et al., 2016). OTUs/ASVs are treated as bacterial taxa, and subsequent analyses of community composition and diversity are conducted based on the

characteristics of these units, e.g. the presence and/or abundance (depending on which measures are being utilized), phylogenetic distances of the OTUs/ASVs.

The method has the advantages of capturing broad levels of community composition, along with cost-effectiveness and less extensive computational analyses required. However, as the information obtained is only limited within one single gene, one could infer limited information about the functional profile and interaction network within the microbial community. Nevertheless, the method is still informative to quantify the richness, abundances of taxa components, and diversity metrics within the microbiomes. Furthermore, data achieved from 16S rRNA sequencing is capable of facilitating observations of the temporal changes and/or alterations in response to outside perturbations in microbial communities (Poretsky et al., 2014). Analysis of taxa abundances may potentially help discover the taxa that actively associate with the host activity (Warnecke et al., 2007), serving as a powerful method to infer the role of the taxa and the relationship between host and the associated microbial community. Therefore, the 16S rRNA approach was chosen for analyses in Chapter 4, which addresses questions about the alterations occurring in the bacterial communities associated with gastropod-parasitic nematodes after slug infection. The sample data sets were obtained by Illumina sequencing of the V3-V4 regions of the 16S rRNA, which was standardized by the Illumina protocol as discussed above.

Metagenomic approach

Metagenomics has emerged as a powerful tool to study the composition, evolution of microbiomes and interactions with the hosts. By directly sequencing all genomes present

in an environmental sample, metagenomics offers the independence from a limited number of phylogenetic markers e.g. 16S rRNA, 18S rRNA, and housekeeping genes. Metagenomic data can provide insights into the community's gene composition, genomic linkages between functions, functional profile of a community, and phylogeny of organisms including uncultured microbes - with a potentially higher resolution and sensitivity compared to 16S rRNA analysis (Poretsky et al., 2014). As for investigation of taxonomic composition, direct analysis of metagenomic DNA has been considered the most accurate method (von Mering et al., 2007) since it is able to provide high sampling depth and avoid biases during PCR amplification of phylogenetic markers. Other novel applications include comparative metagenomics approach where taxonomic and functional profiles were compared between two or more metagenomic datasets (Huson et al., 2009; Shi et al., 2013; Tringe et al., 2005). The approach allows the possibility to improve understanding of microbial specialization and adaptive response to the environment.

However, shotgun metagenome data is not without shortcomings. The method is costly to generate and requires more exhaustive computational resources compared to 16S rRNA amplicon sequencing. Nonetheless, given the advantages, metagenomic-based research has been conducted broadly over the past 10 years and substantially contributes to the understanding of microbial communities.

As metagenomic data capture all genetic materials in the sample of interest, one can obtain valuable information about sequences not well-characterized in existing databases, as well as sequences of low- or high-abundance. The advantage was applied in Chapter 3 of this work, which investigates eight metagenome datasets to identify and

characterize bacteriophage sequences in the microbiome of cyst plant-parasitic nematodes.

Conclusion

Collectively, this dissertation aims to address research questions regarding various organisms and biological frameworks, all of which are connected by the common thread: bacteria of the genus *Pseudomonas.* The data chapter 2 focuses on *Pseudomonas* bacteriophages, using comparative genomics to investigate their diversity, host range, and genome evolution. In Chapter 3, the set of *Pseudomonas* phage complete sequences is employed, along with other genomic tools, to identify and characterize phage DNA in cyst nematode metagenomic data. Chapter 4 explores the microbial communities associated with slug-killing nematodes and probes into the potential role of *Pseudomonas* bacteria in the nematode's pathogenicity. Finally, Chapter 5 summarizes the findings of the three data chapters and discusses future study directions. Together, this work provides new insights that help broaden the understanding of phage and bacterial community's diversity, dynamics, and evolution; also offers the potential for clinical and control applications, as well as for genomic exploratory tools development.

Chapter 2: Comparative Genomic Analysis of 130 Bacteriophages Infecting Bacteria in the Genus *Pseudomonas.*

Anh D. Ha and Dee R. Denver

Published In:

Frontiers in Microbiology - Evolutionary and Genomic Microbiology

DOI: 10.3389/fmicb.2018.01456

Accepted: 12 Jun 2018

Avenue du Tribunal Fédéral 34, CH – 1005 Lausanne, Switzerland

Abstract

Bacteria of the genus *Pseudomonas* are genetically diverse and ubiquitous in the environment. Like other bacteria, those of the genus *Pseudomonas* are susceptible to bacteriophages which can significantly affect their host in many ways, ranging from cell lysis to major changes in morphology and virulence. Insights into phage genomes, evolution, and functional relationships with their hosts have the potential to contribute to a broader understanding of *Pseudomonas* biology, and the development of novel phage therapy strategies. Here we provide a broad-based comparative and evolutionary analysis of 130 complete *Pseudomonas* phage genome sequences available in online databases. We discovered extensive variation in genome size (ranging from 3 to 316 kb), G+C percentage (ranging from 37- 66%), and overall gene content (ranging from 81-96% of genome space). Based on overall nucleotide similarity and the numbers of shared gene products, 100 out of 130 genome sequences were grouped into 12 different clusters; 30 were characterized as singletons, which do not have close relationships with other phage genomes. For 5/12 clusters, constituent phage members originated from two or more different *Pseudomonas* host species, suggesting that phage in these clusters can traverse bacterial species boundaries. An analysis of CRISPR spacers in *Pseudomonas* bacterial genome sequences supported this finding. Substantial diversity was revealed in analyses of phage gene families; out of 4,462 total families, the largest had only 39 members and there were 2,992 families with only one member. An evolutionary analysis of 72 phage gene families, based on patterns of nucleotide diversity at nonsynonymous and synonymous sites, revealed strong and consistent signals for purifying selection. Our

study revealed highly diverse and dynamic *Pseudomonas* phage genomes, and evidence for a dominant role of purifying selection in shaping the evolution of genes encoded in them.

Introduction

Bacteriophages are the most abundant, dynamic and genetically diverse forms of life in the biosphere, with an estimate existence of 10^{31} phages worldwide, continuously infecting 10^{23} to 10^{24} bacterial cells every second (Hendrix, 2010; Keen et al., 2017). That is, if we were to lay all 10^{31} phage particles in the world side by side, the length of that line would be about 200 million light years (estimated by Domingo et al., 2008), or about 1.892 x 10^{21} kilometers. The paramount number and diversity of phages, which exceed those of bacteria, are reflected in their varied habitats, reproduction cycles, infection strategies, and ability to shift hosts. Investigations of phage genomes by sequencing and metagenomics have already suggested enormous genomic diversity, complexity, and mosaic evolution (Lawrence et al., 2002), but a detailed understanding of phage genetic diversity and evolution remains a challenge.

Bacteria of the genus *Pseudomonas* are Gram-negative, aerobic gammaproteobacteria with more than 200 species identified (Özen and Ussery, 2012). *Pseudomonas* bacteria live in a wide variety of niches and actively interact with other organisms. For example, members of the genus are known to be pathogenic to human (*P. aeruginosa*) and other vertebrates (*P. mallei*, *P. plecoglossicida*), pathogenic to plants and fungi (*P. syringae*, *P. tolaasii*, *P. costantinii)*, promote plant growth (*P. chlororaphis*, *P. protegens)*, produce antibiotics (*P. aurantiaca*), and have been reported in many other

biological interactions. Altogether, *Pseudomonas* bacteria have highly important and diverse environmental, biological, and human health-related impacts.

Pseudomonas phages have been known to significantly influence the diversity and evolution of their host bacteria (Paulsen et al., 2005; Sano et al., 2004). Studies of phages have been applied extensively in *Pseudomonas*-related clinical applications, especially *Pseudomonas* disease epidemiology and therapeutic tool development (Bergan, 1978; Pires et al., 2015). Improved understanding of phages infecting *Pseudomonas* may actively contribute to a better understanding of *Pseudomonas* bacteria.

Whole genome sequencing of phage virions and prophages, facilitated by the modern advance in sequencing technology and metagenomics, has been offering an exciting new avenue for understanding phages and their impacts on host bacteria. To date, approximately 8,300 complete bacteriophage genomes have been sequenced, and about 400 were isolated from members of *Pseudomonas*, providing valuable source for investigation into the diversity and complexity of *Pseudomonas*-infecting phages. Whole-genome comparative analysis has been successfully applied in previous studies focused on mycobacteriophages (Hatfull et al., 2010), *Staphylococcus aureus* phages (Kwan et al., 2005), *Bacillus* phages (Grose et al., 2014), and *Enterobacteriaceae* phages (Grose and Casjens, 2014), the findings of which all highlighted the remarkable dynamics and diversity of phages. In each study, a large percentage of total genes identified (e.g. 47.2% of mycobacteriophage genes and 58% of *Bacillus* phage genes) were unique, lacking any clear discernible relationships to other genes. Phage genomes were commonly grouped into clusters of related phages, with outliers characterized as ‘singletons’ that lacked strong relationships with other genomes in the analysis. By identifying clusters of closely

related genomes, whole-genome analysis provides the opportunity to evaluate the diversity and complex relationship between phages, as well as exploring their dynamic host range.

Comparative analysis of phage sequences also offers an avenue to understanding selection pressures acting on phage genes, providing further insight into their history of evolution. A common population genetic metric to study the evolutionary process of protein-coding genes is the ratio of nucleotide diversity at nonsynonymous and synonymous sites (π_N and π_S, respectively). In coding regions, nucleotide changes between sequences may either lead to amino acid substitutions (nonsynonymous) or keep the protein sequences unchanged (synonymous). Assuming that synonymous sites evolve neutrally and represent 'background' mutational variation in genome, the ratio π_N/π_S provides insights into the direction and magnitude of natural selection pressure. The ratio $\pi_N/\pi_S < 1$ is considered a signature of purifying selection, which maintains genetic stability of the genes of interest. It is hypothesized that the majority of phage protein coding genes, especially those encoding structural functions and survival benefits, are under purifying selection (Domingo et al., 2008; Hartl, 2014), therefore have $\pi_N/\pi_S < 1$. Positive selection is inferred when $\pi_N/\pi_S > 1$. If nucleotide substitutions happen randomly with respect to protein-coding function, the sequences of interest are likely evolving under neutrality and have $\pi_N/\pi_S = 1$.

To further study the deep divergence and potential for host range evolution of phages infecting *Pseudomonas*, we examined genome statistics, putative gene contents, and performed whole-genome comparative analysis of 130 phage sequences available in

NCBI GenBank databases. To probe into phage gene evolution, we reported, for the first time, the pattern of selection acting on phage putative gene orthologs by comparing the nucleotide diversity patterns π_N/π_S in predicted ORF families. Our study has contributed new insights into the diversity and evolution of phages infecting *Pseudomonas* and facilitate comparison with phages infecting other bacteria.

Material and methods

Phage genome sequences

One hundred and thirty complete DNA genomes then available on the NCBI GenBank database were downloaded as fasta sequences on February 15, 2016. Only entries with the description "complete genome" in the "Definition" field and the isolation source "*Pseudomonas*" at the genus level, stated in the "host" field, were included in downstream analyses. The GenBank accession numbers of these genomes were included in Appendix Table 1. It was stated in the GenBank entries that the 130 phages selected for study were isolated from six different host species: *P. chlororaphis, P. tolaasii, P. syringae, P. putida, P. fluorescens* and *P. aeruginosa,* from varied sources (i.e. water sewage, hospital wastewater, and environmental samples) at different geographical locations.

Genome annotation.

All 130 phage genomes analyzed in this study were re-annotated to ensure annotation uniformity. Gene prediction was performed with GeneMark.hmm 3.26 (Besemer and Borodovsky, 1999; Zhu et al., 2010) using the heuristic parameters, Glimmer 3.02

(Delcher et al., 1999; Salzberg et al., 1998), and BLASTN when necessary, followed by genome annotation by Phage Rapid Annotation using Subsystem Technology (RAST) (Aziz et al., 2008; Brettin et al., 2015; Overbeek et al., 2014) and PHASTER web server (Arndt et al., 2016; Zhou et al., 2011). Gene annotation was not manually curated systematically in all genomes. While annotation using the gene calling programs GeneMarkS and Glimmer were previously shown to be accurate and suitable for phage genome analysis (Besemer and Borodovsky, 1999; Delcher et al., 1999; Mills et al., 2003), the GeneMarkS self-training version automatically excludes candidate ORFs which are shorter than 300 bp, and assign ORFs from the 5′-most ATG codon (Besemer et al., 2001). These automated rules likely resulted in slight underestimation of gene numbers and coding density in phage genomes.

The program ARAGORN (Laslett and Canback, 2004) was used to detect tRNA and tmRNA genes in phage genome sequences.

Genome mapping and assigning open reading frame (ORF) families

The open-source program Phamerator (Cresawn et al., 2011) was used to map phage genomes in linear illustration and identify similarity between sequences. Predicted ORFs were assigned into families also using the program with a ClustalW threshold of 35% amino acid identity and a BLASTP score of 1e-50. Conserved protein domains in ORFs were identified by searching for hits from the NCBI Conserved Domain Database (CDD). To probe into the possibility of ‘ORFans’ (i.e. ORFs that shared no detectable relationships with other ORFs in this analysis) having homologs in undocumented phages, we compared the protein sequences of the 2,992 ORFans in our 130 genomes

with the globally sampled virome dataset constructed by Paez-Espino and colleagues (Paez-Espino et al., 2016) using TBLASTN (Camacho et al., 2009). All the 125,842 viral genomes/contigs assembled from this large-scale metagenomic dataset were downloaded from the public FTP site http://portal.nersc.gov/dna/microbial/prokpubs/EarthVirome_DP/.

Genome clustering

Genome alignment and calculation of percentage of nucleotide identity were performed with Kalign (Lassmann and Sonnhammer, 2005). A dot plot of whole genome comparison for all 130 phage DNA sequences was generated in Gepard 1.40 (Krumsiek et al., 2007) with a sliding window of 10 nucleotides, allowing for the visualization of pairwise similarity between genomes and assigning preliminary clusters. Phage genomes were incorporated into a cluster if they shared more than 45% nucleotide identity with the cluster's genome members and there was nucleotide sequence similarity visually recognizable in the dot plot. Previous studies utilized the threshold 50% nucleotide identity (Hatfull et al., 2010; Pope et al., 2011), however, this value was applied for phages isolated from a single host (*Mycobacterium smegmatis*), while our study included phages infecting bacteria across the genus *Pseudomonas*, which potentially lead to greater variation in genome sequences. Therefore, we lowered the threshold parameter in our analysis to 45%. A second criterion for assigning members into cluster was that they share over 20 predicted protein families, to ensure detectable relatively conserved regions and synteny. Third, clusters were required to have at least three members. Phage genomes not meeting these three criteria were not included in clusters, identified as 'singletons'.

Within each cluster, the variation of host bacterial species among phage members was investigated. Information of original host species was extracted from the 'host' field of the phage GenBank entry. To confirm the potential of phages, especially members of clusters with multiple hosts to infect different *Pseudomonas* species, we looked for signatures of past phage-host infections using the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/ Cas (CRISPR-associated) spacer approach. All available spacer sequences identified in *Pseudomonas* bacteria were downloaded from the public database CRISPRdb (Grissa et al., 2007) and compared with phage genomes individually using BLASTN. Since spacer sequences, which range from 25 to 75 bp, are shorter than usual BLASTN query size, we utilized the following BLASTN parameters: maximum e-value of 0.3, word size 7, gap extension penalty 2, gap opening penalty 10, mismatch penalty 1, dust filtering off (adapted with modification from Edwards et al., 2016).

Evolutionary analysis

To investigate the genetic variation and pattern of evolution in predicted coding regions, we calculated the nucleotide diversity within selected ORF families (π_N/π_S). Since evolutionary analysis using π_N/π_S test requires high confident alignment regions of ORF families, we increased ClustalW threshold value to 50% to ensure less diversity in ORF members and facilitate better alignments. All ORF families which have ClustalW score of at least 50%, BLASTP score threshold of 1e-50 and have 15 members or more were included in evolutionary analysis. Predicted protein-coding ORF sequences of each family were aligned using the MUSCLE program in MEGA7 software (Kumar et al., 2016). π_N and π_S were calculated individually for each ORF family using the software

DNAsp v5.10.1 (Librado and Rozas, 2009). The π_N and π_S values of the ORF families were compared using the nonparametric Wilcoxon signed-rank test for paired data.

Results

Patterns of genomic variation

The basic genome metrics of the 130 phage sequences included in this study (host, genome size, G+C content, number of predicted ORFs, number of tRNA or tmRNA genes) were provided in Appendix Table 1. The analyzed *Pseudomonas* phages showed broad diversity of genome size from 3 kb to 316 kb. The majority of phages (119 out of 130 - 92%) had genome size in the range from 35 to 100 kb, distributing uniformly across this interval. Four outliers were very small at less than 8 kb, and seven were larger than 200 kb (Figure 2.1A). This pattern was mirrored in the number of predicted ORFs (Figure 2.1B), as the four phages with the smallest genomes contained only 3 to 8 predicted ORFs, whereas the seven unusually large genomes harbored more than 200 putative ORFs. The remaining 119 phages encoded 41 to 173 ORFs. The overall average of putative ORFs per genome was 93.4. G+C content also varied greatly (ranging from 36.8% to 66.4%), averaging at 56.8% (Figure 2.1A, inset).

Categorization of phage genomes into clusters

To assign phage genomes to clusters, we calculated pairwise sequence similarity between all possible pairwise phage genomes and performed whole genome dot plot analysis. The methods of clustering phages based on dot plot matrix were similar to those previously applied to mycobacteriophages (Hatfull et al., 2010; Pope et al., 2011). All 130

nucleotide genomes were concatenated into a single sequence and duplicated to form two axes, generating a dot plot matrix (Figure 2.2). If two sequences had high similarity, a diagonal would show at that location on the plot (the center diagonal line demonstrated the 100% similarity where sequences were compared to itself). The resulting dot plot matrix revealed 12 clusters (Clusters A through M) of phages sharing at least 45% nucleotide similarity; 30 genomes were not assigned to any cluster and remained singletons. Pairwise nucleotide similarity between phages of each cluster were reported in Appendix Table 2. The number of members and pairwise nucleotide similarity between phage genomes of each cluster varied from cluster to cluster, e.g. cluster L has 8 members sharing at least 58% nucleotide similarity, while this number in cluster K is 75% among the 4 members (Appendix Table 2). Some clusters (Clusters A through D, H, and L) were further divided into subclusters. Details of phage assignment to clusters and subclusters were shown in Appendix Table 1. Basic characteristics of members in each cluster including morphotype (according to ICTV classification system), host species from which the members were isolated, average genome size, number of ORFs, GC content, number of tRNAs were also provided in Appendix Table 1.

Characterization of open reading frames (ORFs) in phage genomes

The annotation process predicted a total of 12,139 putative ORFs ranging from 54 bp to 12 kb in size, with an average length of 650 bp among the 130 genomes analyzed. The predicted ORFs were assigned to groups of closely related sequences (ORF families) using Phamerator with a ClustalW threshold of 35% amino acid identity and a BLASTP score of 1e-50. In genomic maps, putative ORFs were colored and numbered according

to Phamerator assigned ORF family (white color denoted ORFs having no similarity with other ORFs, equal to the threshold BLASTP 1e-50 or smaller). A total of 4,462 ORF families were generated with an overall average number of only 2.72 members. The largest family had 39 members (family 107) and 2,992 families (67.1% of the total) only had one member (Figure 2.3). Out of these 2,992 ORFams, 365 sequences (12.2%) had significant hits (maximum e-value set at 1e-50) with the contigs in Paez-Espino et al., 2016's virome dataset, and with a more relaxed threshold (e-value 1e-25), homologs of 665 ORFans (22.22%) were found.

ORF density of all phage genomes varied from 0.83 to 2.50 per kb, averaging at 1.42 ORFs per kb. This value was highly similar between members within each cluster i.e. shown in very small deviations from the mean (Table 1.1). Members of Cluster E and F showed higher mean ORF densities among all clusters, ranging from 1.73 to 1.87 ORFs per kb. However, the highest average numbers of ORFs each kb were observed in two singletons Pf1 and phi_Pto-bp6g (1.91 and 2.49). The lowest numbers of ORFs per kb (0.83 and 0.84) were found in PP7 and PRR1, the only two Leviphages included in this study.

To investigate the mode of selection acting on phage ORFs, we calculated the nucleotide diversity at nonsynonymous sites and compared it to the diversity at synonymous sites (the π_N/π_S ratio) for a select set of 72 ORF families. π_N/π_S at or near one suggested that the ORF sequences of interest were likely evolving under neutrality; π_N/π_S values greater than one implied positive selection and values less than one indicated purifying selection. We observed a broad pattern of π_N/π_S values distribution, however,

all ORF families analyzed had π_N/π_S value under 1 (Figure 2.4A). The majority of the families had π_N/π_S ratio closer to zero (62/72 families - 86% have π_N/π_S within the range from 0.0 to 0.5) (Figure 2.4B).

Diversity of phage hosts within clusters

Of the 12 *Pseudomonas* phage clusters defined here, five contained phage members isolated from more than one host species (Clusters C, F, H, L and M) (Figure 2.5A). On the extreme end, the eight members of Cluster L were isolated from five different hosts: *P. putida, P. fluorescens, P. tolaasii, P. syringae,* and *P. plecoglossicida,* while all members shared at least 58% nucleotide similarity with others in cluster L (Appendix Table 2). When compared with the *Pseudomonas* bacterial CRISPR spacer library, all these phages showed matches with spacers of multiple *Pseudomonas* species that differed from the original hosts. Meanwhile, in clusters of phages all isolated from one sole species, especially Clusters A, E and I (all originated from *P. aeruginosa*), members tended to share more sequence similarities with spacers of *P. aeruginosa* - their host species (Figure 2.5B).

Genomic map of phages in Cluster L was shown in Figure 2.6A. Within subcluster L1 and L2, we observed long regions of violet shading indicating long conserved regions between phage genomes. Meanwhile, between subclusters, this relationship was apparently weaker with shading towards the red end of the color spectrum. Regions of high similarity and same-colored ORF blocks shown on the map indicated prevalent synteny. Breaks in synteny were also evident as interspersed white

blocks and little or no sequence similarity between genome sequences. An example of such synteny breaks was shown in figure 2.6B. Between the two phage genomes gh-1 and phiPSA2 of subcluster L1, the presence of gh-1 ORFs (gh1_80 and gh1_170) had interrupted the synteny organization.

On mapped sequences, conserved domain hits from CDD database could be found, including not only hits from *Pseudomonas* bacteria and their phages, but also from different prokaryotes and viruses. In many ORFs, multiple overlapped hits corresponding to the same portion of the ORF were observed. This pattern was more likely to be found in ORFs belonging to large families (e.g. gh1_30, gh1_150, phiPSA2_9, phiPSA2_13, Figure 2.6B). The domains identified usually involves in conserved phage functions (e.g. ORF gh1_150 contains peptidoglycan recognition protein domain, T3-like lysozyme domain, and domain of N-acetyl-anhydromuramyl-L-alanine amidase, which cleaves the amide bonds between N-acetyl-anhydromuramyl and L-amino acids in bacterial cell wall).

Discussion

Variation in Pseudomonas *phage genomes and genes.*

The 130 phage genomes analyzed show a wide variety of G+C content, from as low as 37% to 66%. Interestingly, since the average G+C percentage of bacterial *Pseudomonas sp.* genomes is in the range from 58% to 66% (Winsor et al., 2016a), the G+C content of a large number of *Pseudomonas* phage genomes in this study is much lower than that of its host (Table 1.2). Similar observation was noted in a sample of 18 *P. aeruginosa*

phages (Kwan et al., 2006), and did not agree with other findings that phage G+C content is usually similar to that of the host (see examples in Table 1.2). The low G+C percentage in *Pseudomonas* phage genomes may indicate an active history of shifting from other bacterial hosts where the phages acquired low G+C content sequences via horizontal gene transfer (HGT).

Cluster assignment based on sequence similarity and homologs is supported by highly similar properties i.e. low standard deviation of the within-cluster average, such as morphotype, genome size, GC percentage, number of predicted ORFs (Table 1.1). The proportion of singletons (30/ 130 phages – 23.1%) is notably higher compared to previous results of other comparative genomic studies: 1.3% of 627 mycobacteriophage genomes (Pope et al., 2015); 5.3 % of 337 *Enterobacteriaceae* phages (Grose and Casjens, 2014), and 18.1% of 83 *Bacillus* phages (Grose et al., 2014) were designated to be singletons. It is worth noting that the small percentage of singletons in mycobacteriophage might be due to the very large scale of survey on phages isolated from a single species (*Mycobacterium smegmatis*). As the survey broadened to higher taxonomic level yet smaller sample size, the percentage of singletons increased noticeably. The large number of singletons suggests that the current stage of discovery has revealed just part of the dynamic diversity in the world of *Pseudomonas* phages. Adding new genomes could bring a singleton into a cluster by identifying intermediate phage relatives, therefore further sampling, profiling and reassigning of clusters might be necessary to better evaluate the extent of *Pseudomonas* phage diversity.

Genome mosaicism was observed extensively in all phage genomes with the remarkable frequency of ORF modules and breaks in synteny between genomes. ORFs of the same family consistently located between different flanking ORFs. This pattern of pervasive mosaicism is well in line with previous findings in *Enterobacteriaceae* phages, *S. aureus* phages and mycobacteriophages (Grose and Casjens, 2014; Kwan et al., 2005; Pope et al., 2011). The mosaicism may suggest 1) high activities of HGT and 2) phage evolution to drop unnecessary genes to keep the genome minimal and efficient as they adapt to new purposes. Dynamic HGT also hints at a flexible host range, which is needed for more opportunities to gain access to a larger gene reservoir. Other genomic events could also contribute to the large-scale mosaicism such as transposition (Edgell et al., 2010), cleavage by endonucleases (Kristensen et al., 2013), phage recombinases acting on relaxed homology (De Paepe et al., 2014), and mistakes in genome replication and host repair mechanism during the prophage phase.

The great diversity of *Pseudomonas* phage was also indicated on the scale of genes. With a ClustalW threshold of 35% amino acid identity and a BLASTP score of 1e-50, the largest ORF family has only 39 members, which is remarkably small compared to the largest family (104 members) that (Pope et al., 2011) assigned with similar thresholds from 80 mycobacteriophage genomes. Moreover, 67.1% of the predicted ORF families have only one member, while the numbers of one-member families identified in studies of mycobacteriophages and *Bacillus* phages are much lower at 47.2% and 58%, respectively. Remarkably, a large number of presumptive insertions/ deletions in genomes are unique ORFs (displayed in genome map as white boxes), and rarely contain

known conserved domains. The vast number of unique ORFs and the diversity of ORF families suggest a large gene influx from novel bacterial hosts and/or other phages by HGT. Among the 2,992 ORFans identified, a considerable proportion (665 sequences - 22.2%) had significant TBLASTN hits with the global virome dataset despite a stringent threshold (1e-25). This result further demonstrates the largely unexplored gene reservoir of *Pseudomonas* phages, with many potential homologs with undocumented phage sequences in nature.

The number of predicted ORFs was significantly positively correlated with phage genome sizes (R^2= 0.936, $p < 0.001$ - Figure 2.1B) in our study. ORFs account for more than 80% of the total genome sequence space for all phages examined (Appendix Table 1) with an average coding percentage of 92.4%, indicating their high genetic efficiency, which is consistent with observations of previous studies, e.g. of mycobacteriophages (Rohwer et al., 2014), *Staphylococcus aureus* phages (Kwan et al., 2005). Mean ORF density was at 1.42 ORFs per kb, which is slightly less than that in mycobacteriophages and *Staphylococcus aureus* phages (1.69 and 1.67 genes per kb, respectively) (Hatfull et al., 2010; Kwan et al., 2005). The only two Leviphages included in this study, PP7 and PRR1, have the lowest ORF densities (0.83 and 0.84 ORFs per kb, respectively), while possessing the smallest genomes among the 130 phages (3,588 and 3,573 kb). This may initially seem counter-intuitive, since small phages must compress a minimum number of genes required for surviving in very small sequence space, which should result in higher number of ORFs per kb - for example, Microphage φX174 accommodates 11 genes over the length of only 5,386 bp through gene overlap involving multiple reading frames

(Rohwer et al., 2014). However, small phages also must maintain minimum sizes and extra nucleotide sequences to allow for efficient packaging, thus decrease their ORF density. This was observed in phage lambda sequence, where non-coding sites (*cos*) are essential for DNA cleavage, processing, duplex nicking, enzyme binding (Catalano et al., 1995; Feiss and Catalano, 2013). It is possible that non-ORF sequences in these small *Pseudomonas* phage genomes perform important functions that we do not yet understand.

Insights into phage host range

Eight members of Cluster L were reported to have five different host species while sharing at least 58.7% nucleotide identity. While these five hosts appear to be more related to each other than to *P. aeruginosa* (Ait Tayeb et al., 2005), interestingly, member of Cluster M, isolated from *P. aeruginosa* and *P. chlororaphis*, also share as high as at least 50.9% identity. Although it was expected that percentage of identity among Cluster M is smaller than that of clusters with members isolated from the same host species (e.g. members of Cluster K, all isolated from *P. aeruginosa* - different strains - share at least 75.2% identity), this sequence clustering is remarkable as *P. aeruginosa* and *P. chlororaphis* are much further apart in the *Pseudomonas* phylogenetic tree (Ait Tayeb et al., 2005). This pattern of clustering suggests a flexible host range not only among strains within one species, but also could expand to between species. The dynamics would allow for better adaption to their fast-changing bacterial hosts.

To further evaluate phage potential for broad host ranges beyond a single bacterial species, we computationally investigated the history of past infections in all

Pseudomonas species by comparing each phage genome with all available CRISPR spacers originating from *Pseudomonas* bacterial genomes. CRISPR modules in the genome provide bacteria with an adaptive immunity against viruses and mobile genetic elements (Horvath and Barrangou, 2010). CRISPR arrays consist of interspaced repeated sequences that are separated by short different sequences named spacers. These fragments may represent a part of phage sequences inserted into CRISPR arrays on bacterial genome during previous infections and are constantly replaced and heritable. Therefore, spacers provide a paleogenomic window into recent phage infections. Members of Cluster L showed matches to various species, 100% of which are different from their host and might suggest the ability and/or a history of attacking different species in *Pseudomonas*, and then shifting to the current host, while in *P. aeruginosa* phages of Cluster A, E and I, this variation is much less extensive (Figure 2.5B). We note that Cluster D, also consists of all *P. aeruginosa* phages, shows a lower percentage of matches to *P.aeruginosa* itself (12.5% - Figure 2.5B), however, it is also worth noting that the total number of hits is only eight, hence this low proportion might be a result of error sampling. While the spacer library in *P. aeruginosa* is expected to be better documented than that in other *Pseudomonas* species, which could in part explain the dominant proportion of *P. aeruginosa* hits of phages in Cluster A, E and I, and the lack of matches to the recorded hosts of all phages in Cluster L, the presence of similarities to species other than the original hosts does imply the capability of host shifting.

Purifying selection is prevalent among a subset of predicted ORF families analyzed.

Nucleotide diversity at nonsynonymous and synonymous sites among multi sequences, π_N/π_S, and Ka/Ks, which performs pairwise comparisons in different species, have been considered useful tools to evaluate the type of natural selection acting on coding regions (Chen et al., 2017; Howe and Denver, 2008; Zhang et al., 2016). The ratio Ka/Ks is frequently used in well-defined species, calculated by pairwise comparisons and highly time dependent (Rocha et al., 2006). Since we were comparing the diversity between ORF orthologs in multiple phages, of which the species taxonomic classification is not well established (Lawrence et al., 2002), we performed the π_N/π_S test to evaluate the selection pressure on phage putative genes.

We found that all of the 72 ORF families included have a π_N/π_S ratio less than 1.0, which implies a reduced diversity in nonsynonymous sites and a history of long term purifying selection (Wilcoxon signed-rank test, p = 8.5e-14). Eighty-six percent (62/ 72) families have π_N/π_S ratio closer to zero, pointing to relatively strong purifying selection. Out of ten families indicating weaker purifying selection ($\pi_N/\pi_S > 0.5$), six showed $\pi_S < 0.5$. As lower values of π_S points to smaller possibility of saturation at synonymous sites, these six families provided the most reliable evidence for purifying selection. Two families were annotated as housekeeping genes with structural function i.e. major capsid protein (family 1310) and putative large terminase subunit (family 5602). The other four families were unannotated and have no known domains identified. No signal of positive selection, which favors synonymous substitutions and results in π_N/π_S ratios above 1.0, was observed. This pattern agrees with previous studies of mycobacteriophages and

cyanophages. Weigele et al., 2007 measured the ratio Ka/Ks for all pairwise orthologues in 20 mycobacteriophages and 5 cyanophages and found that the values average at about 0.25 and 0.01, respectively. Based on the fact that the majority of protein coding genes have $\pi_N/\pi_S < 1$, the ratio offers the potential to assist the identification of true genes (Nekrutenko et al., 2002). Although the results of π_N/π_S and similar genetic code-based tests alone cannot absolutely confirm the accuracy of gene calling, they offer informative complementary strategies to evaluate the veracity of candidate ORFs identified by annotation tools and can be applied to any putative coding sequences.

The genomic data of 130 phages included in this study has revealed extensive gene diversity and strong purifying selection acting on genes of *Pseudomonas* phages. Nevertheless, we note that the dataset used for the present study was downloaded in February 2016, incorporating the 130 complete *Pseudomonas* phage sequences then available on the NCBI GenBank database. In the time it took to execute the present analyses and write-up, the number of available genomes has more than tripled (391 *Pseudomonas* phage genomes available in May 2018). Exhausting all available genome entries in public databases proves to be a continual problem for comparative genomic analyses, especially in this instance, as phages with small genomes are constantly added to databases at accelerating rates. However, given the high genetic diversity, cluster structures, and the abundance of unique genes with no apparent relatives, future analysis including an increased number of sequences will provide more information about the genetic diversity and evolution of the world of *Pseudomonas* phages. Expanding the scope of analysis with the ever-increasing numbers of genomes is expected to decrease

the number of ORFans as newly identified homologs in the broader gene reservoir could bring them into families; though how many genomes will be required to see substantial ORFan number reductions remains unclear.

Currently, the number of complete *Pseudomonas* phage genomes available from GenBank is heavily skewed towards *P. aeruginosa* phages. In this study, 107 out of the 130 phages analyzed were listed as isolated from *P. aeruginosa. P. aeruginosa* is a major human pathogen with the increasing ability to develop antimicrobial resistance and the potential for phage therapeutic strategy against *P. aeruginosa* infections has long been recognized. Consequently, it was anticipated that *P. aeruginosa* bacteriophages would receive preferential attention, and thus would be more frequently profiled and investigated. To achieve a more comprehensive understanding of phages infecting *Pseudomonas,* future research should include more sequences of phages isolated from other host species.

Table 2. 1. Summary of phage characteristics by clusters.

Cluster	# mem.	ICTV Family	Host(s)	# ORFs	# ORFs per kb	Genome size (bp)	GC content	# tRNAs
A	20	Siphoviridae	*P. aeruginosa*	53.6 ± 2.0	1.43 ± 0.06	37289.5 ± 1174.0	64.0 ± 0.5	0 ± 0
B	6	Siphoviridae	*P. aeruginosa*	82.0 ± 4.3	1.37 ± 0.04	59645.3 ± 1973.1	64.2 ± 0.4	0 ± 0
C	13	Podoviridae	*P. aeruginosa, P. fluorescens*	52.0 ± 2.4	1.21 ± 0.06	42945.2 ± 522.7	62.0 ± 1.0	0 ± 0
D	6	Podoviridae	*P. aeruginosa*	99.5 ± 11.4	1.36 ± 0.14	73158.5 ± 979.3	54.4 ± 0.8	0 ± 0
E	6	Myoviridae	*P. aeruginosa*	157.3 ± 3.2	1.80 ± 0.02	87273.0 ± 1534.8	54.7 ± 0.1	3 ± 0
F	10	Myoviridae	*P. aeruginosa, P. syringae*	171.1 ± 1.4	1.83 ± 0.04	93570.7 ± 1668.2	49.2 ± 0.5	15 ± 1.8
G	3	Podoviridae	*P. aeruginosa*	68.3 ± 2.9	1.52 ± 0.03	45068.7 ± 903.5	52.2 ± 0.2	3.3 ± 0.5
H	7	Podoviridae	*P. aeruginosa, P. fluorescens, P. putida*	68.0 ± 1.4	1.49 ± 0.03	45698.1 ± 291.3	52.3 ± 0.5	2.14 ± 1.5
I	14	Myoviridae	*P. aeruginosa*	91.5 ± 2.7	1.38 ± 0.02	66104.4 ± 1066.5	55.4 ± 0.3	0 ± 0
K	4	Siphoviridae	*P. aeruginosa*	54.3 ± 1.3	1.26 ± 0.03	43031.0 ± 157.3	53.9 ± 0.5	0.25 ± 0.5
L	8	Podoviridae	*P. aeruginosa, P. fluorescens, P. putida, P.*	48.1 ± 2.5	1.20 ± 0.03	40032.5 ± 1325.6	57.0 ± 0.7	0 ± 0

			plecoglossicida, P. chlororaphis					
M	3	Myoviridae	*P. aeruginosa, P. chlororaphis*	416.7 ± 52.1	1.38 ± 0.09	302072.0 ± 19192.2	43.3 ± 5.7	6.3 ± 3.1

Table 2. 2. Examples of G+C content in different phages and their bacterial host. Many *Pseudomonas* phages included in this study have notably lower G+C content than their host, in contrast with the relative similarity between phages of other bacteria.

Phage	**G+C content of phage**	**G+C content of host**
phiKZ	36.83%	*P. aeruginosa* strain PA01 (66.6%) (Labaer et al., 2004)
phiPA3	47.73%	
PA2	54.86%	
PABG	55.82%	
UFV-P2	51.47%	*P. fluorescens* (60 to 66%) (Cornelissen et al., 2012)
phi_Pto-bp6g	42.71%	*P. tolaassii* strain 6264 (60.6%) (Winsor et al., 2016a)
eiAU	55.37%	*Edwardsiella ictaluri* (57%) (Carrias et al., 2011)
eiDWF	55.54%	
eiMSLS	55.77%	
Staphylococcus aureus phages	33.7%	*Staphylococcus aureus* (32.9%) (Kwan et al., 2005)
mycobacteriophages	63.4%	*Mycobacterium smegmatis* (67.4%) (Mohan et al., 2015)
Streptococcus pneumoniae phages	39.8%	*Streptococcus pneumoniae* (39.7%) (Kwan et al., 2006)

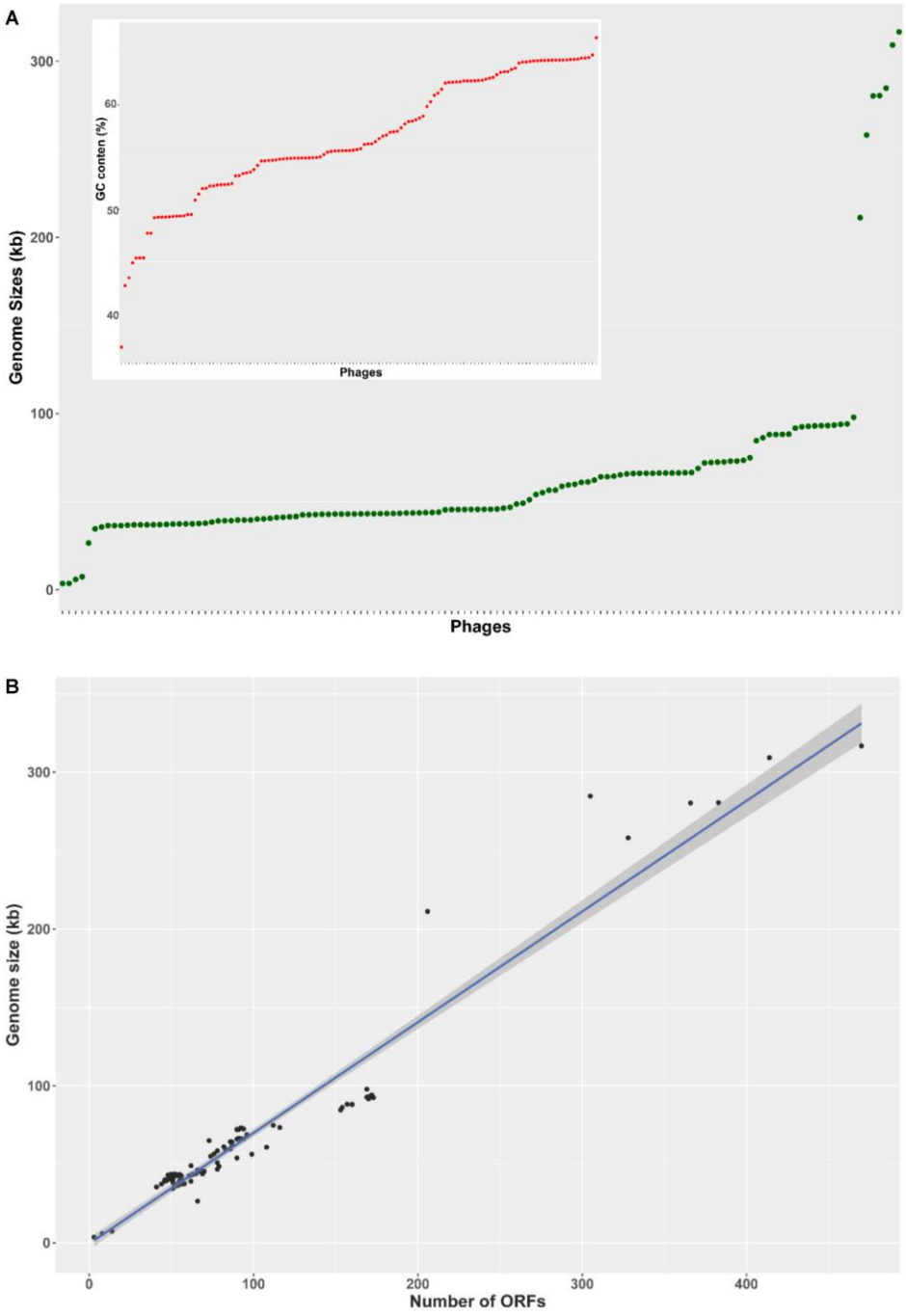


Figure 2.1. Genome characteristics of 130 Pseudomonas phages. Phages were rank-ordered on the X axis based on the property identified on the Y axis. (A) A rank-ordered plot of genomes sizes reveals a range of 3-316 kb and only a few genomes larger than 100 kb. Ranked plot of G+C content (inset) reveals a range of 37-66%. (B) The number of predicted ORFs in phage genomes showed a strong, statistically significant correlation with genome size (R2= 0.936, p< 0.001). The shading denoted 95% confident interval of the linear correlation.

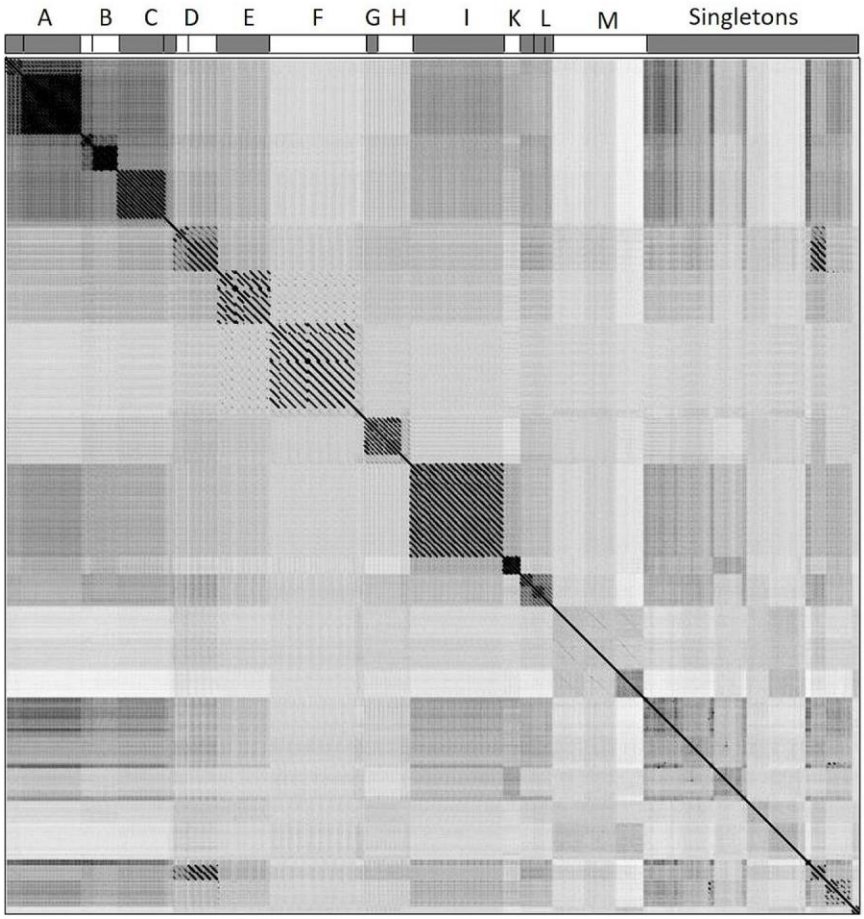


Figure 2.2. Whole-genome dot plot comparison of phage nucleotide sequences. All 130 genomes were concatenated into a single sequence, then plotted against itself with a sliding window of 10 bp and visualized by Gepard 1.40. 111 phage genomes were assigned to 12 (A-M) and 30 phage genomes remained singletons. The assignment of phages to clusters A-M is shown at the top horizontal axis.

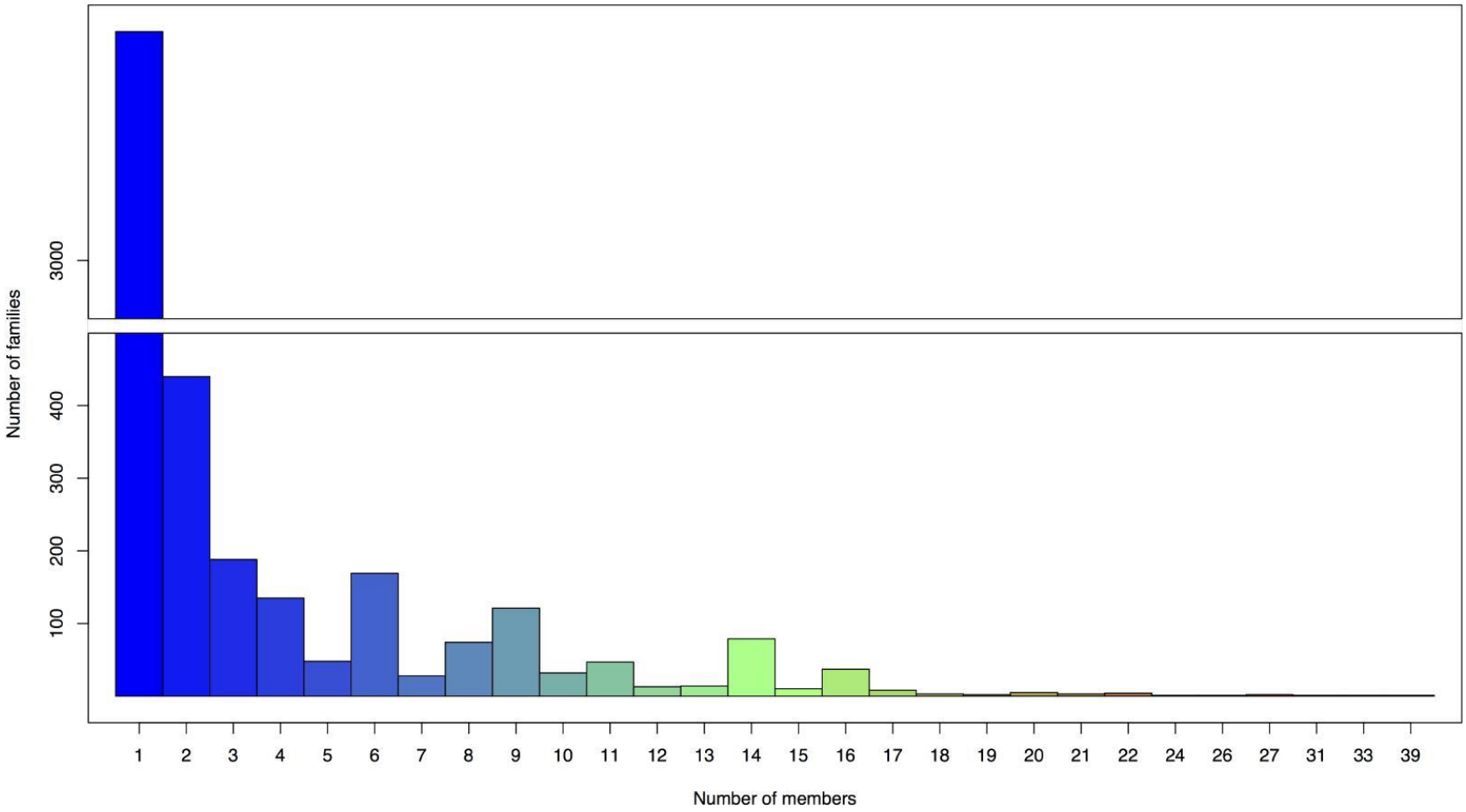


Figure 2.3. Number of members in ORFs families assigned by Phamerator. The largest family has 39 members and 2,992 families (67.1% of the total 4,462 ORF families generated) only have one member.

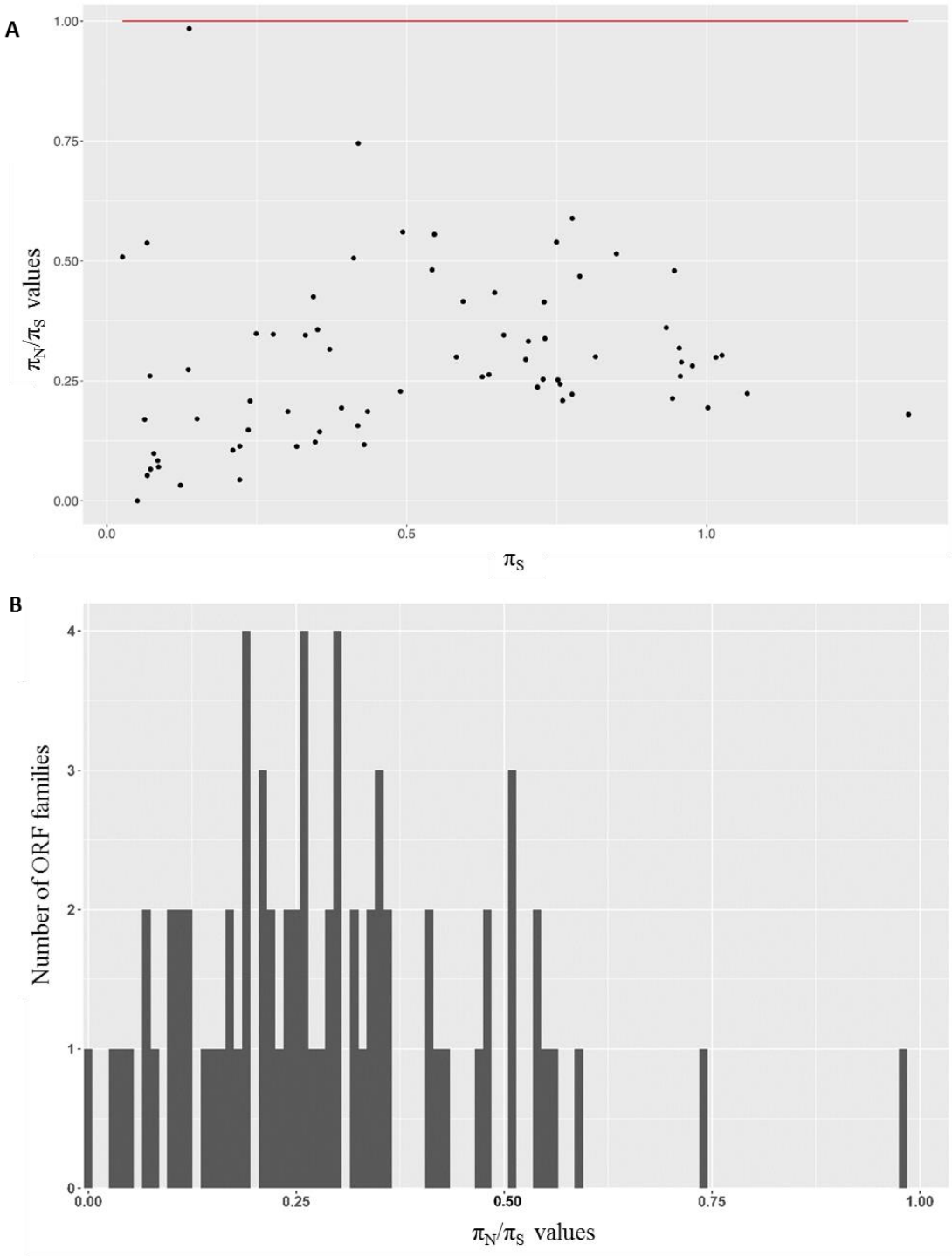


Figure 2.4. Modes of selection acting on a subset of ORF families. (A) Seventy-two families were chosen for further evolutionary analysis. The πN/πS ratio of each family are shown on the Y axis. The Ps values are shown on the X axis. The red line indicates the π_N/π_S =1. (B) A histogram of the π_N/π_S values among ORF families analyzed. The majority (62/72 - 86%) of the families included has a π_N/π_S ratio ranging from 0- 0.5.

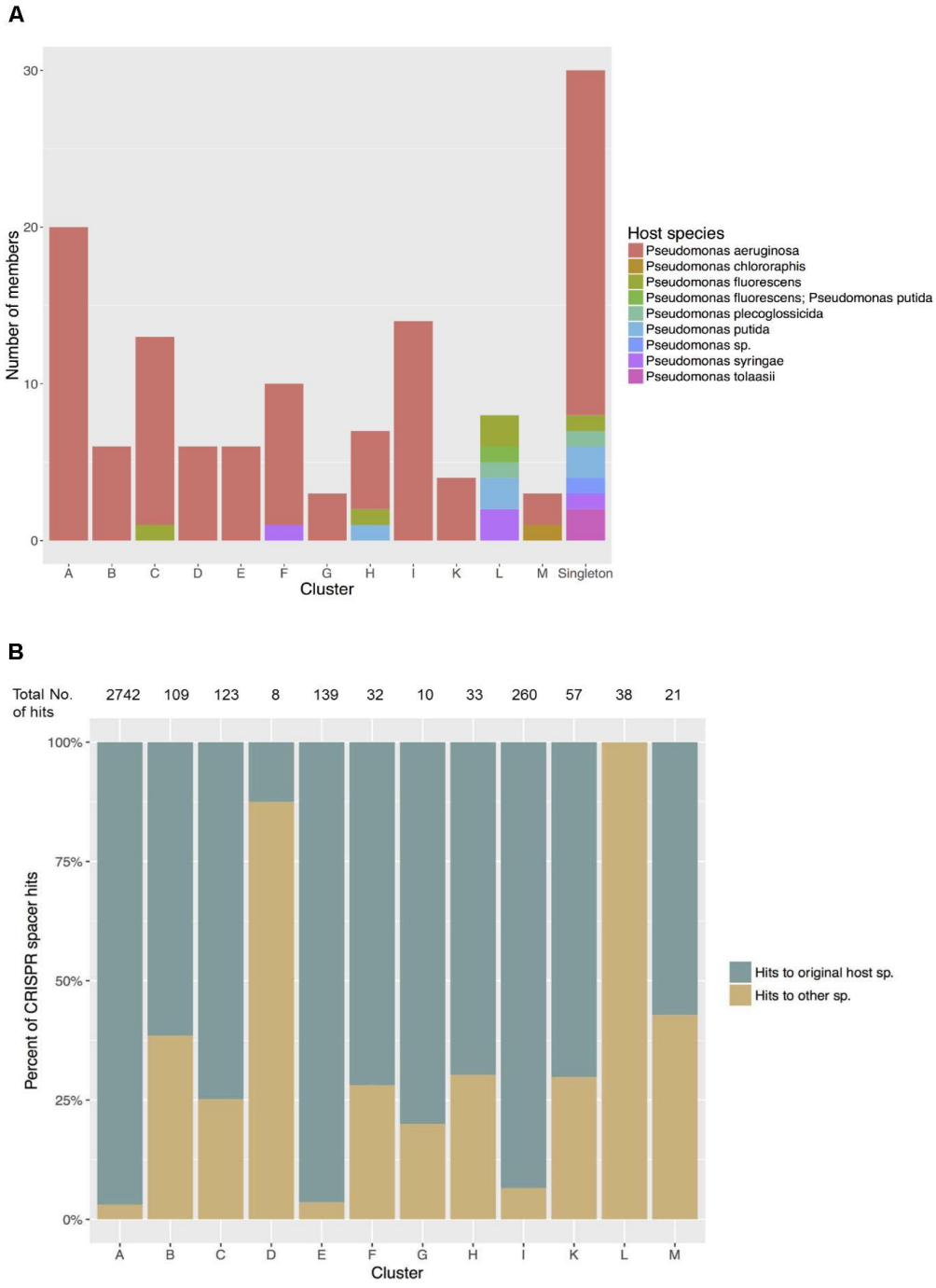
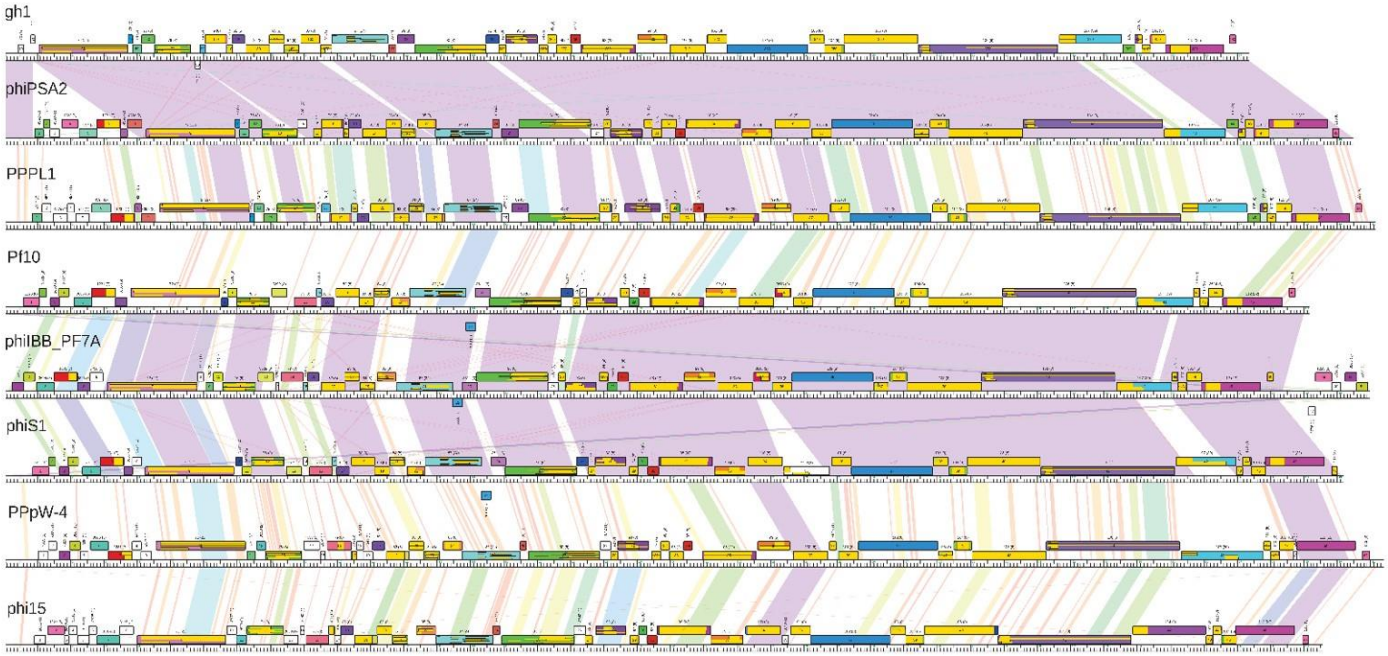


Figure 2.5. (A) Host species of phages in each cluster. Five clusters show closely related phages infecting different host species. (B). Matches between phage sequences in each cluster and CRISPR spacers in *Pseudomonas* host species. Significant matches were recorded as hits to spacers predicted in their original host species and hits to other *Pseudomonas* species. The total numbers of hits, regardless of the types found in each cluster were shown at the top.

A



HOSTS

P. putida

P. syringae pv. actinidiae

P. syringae pv. actinidiae

P. fluorescens

P. fluorescens

P. fluorescens

P. plecoglossicida

P. putida

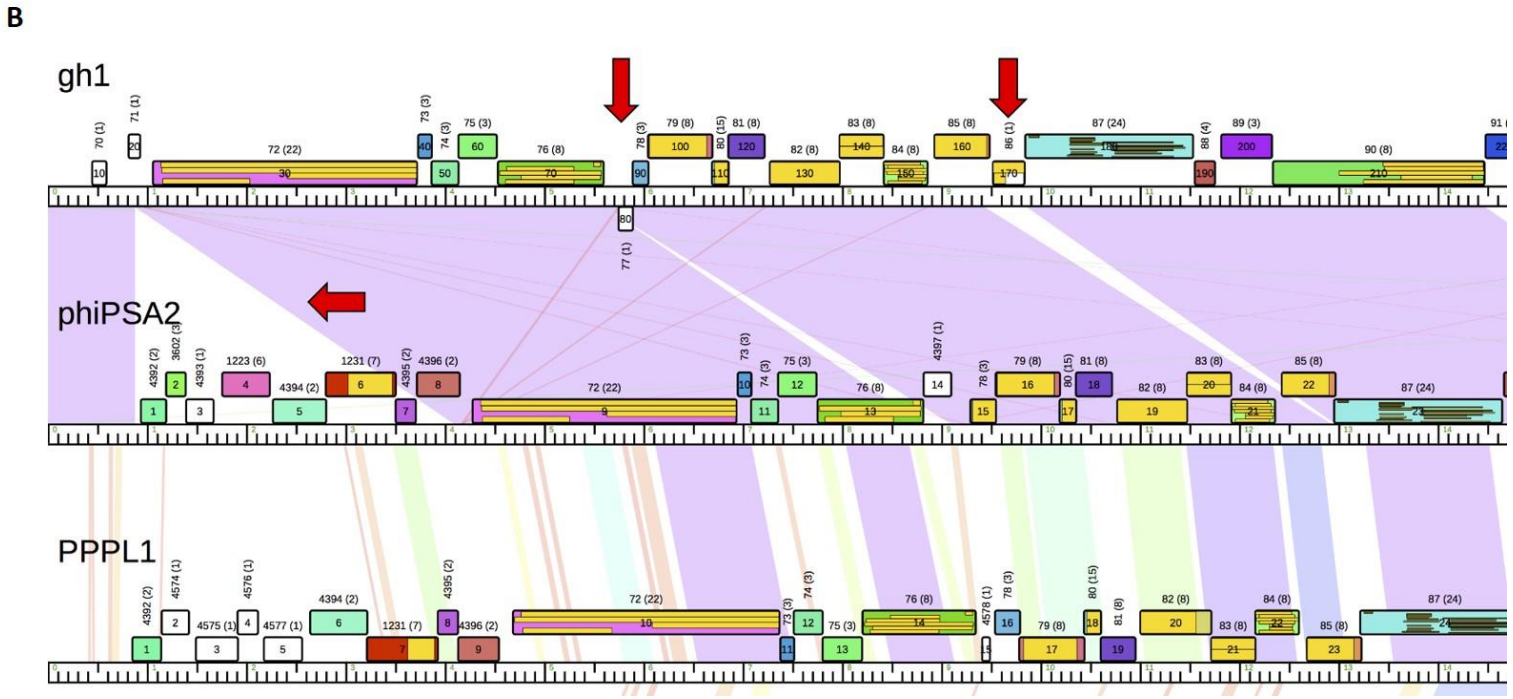


Figure 2.6. (A) Genomic map of phages in Cluster L. Phage genomes were mapped using Phamerator. Genomes were arranged in the map according to the assigned subclusters: subcluster L1 with gh-1, phiPSA2, PPPL-1, subcluster L2 with Pf-10, phi-S1, phiIBB-PF7A, and subcluster L3 with phi15 and PPpW-4. Boxes indicate predicted ORFs, numbers and colors are assigned according to predicted protein families. White boxes denote ORFs that have no similarity at an E-value 1e-50 or smaller to other predicted ORFs. Shading between genomes indicates regions of pairwise nucleotide similarity and was coded in color spectrum so that color indicates nucleotide similarity (violet representing highest similarity with an E-value of zero and red being similarity with E-value of 1e-50). (B) Close-up view of subcluster L1 map. Red arrows indicate breaks of synteny. Yellow boxes within ORFs display conserved domain hits from CDD database, separated by lines if there are multiple hits found in one ORF.

Chapter 3: Discovery and analysis of bacteriophage DNA in plant-parasitic nematode metagenomes

Anh D. Ha, Dana K. Howe, Sulochana K. Wasala, Cedar Hesse, Amy Peetz, Melissa Mitchum, Louise-Marie Dandurand, Inga A. Zasada, and Dee R. Denver

In preparation for:
PLoS One
Public Library of Science
San Francisco, California

Abstract

Bacteriophages are constant and important members of all existing microbial communities in our biosphere. As phages directly influence their bacterial hosts, they set substantial impacts on the dynamics and interactions within their home microbiomes. Therefore, research of phages as a component of metagenomes may facilitate better understanding of natural microbial community. Here we investigate eight nematode metagenomes generated from the microbiomes of cyst nematodes *G. pallida* and *H. glycines* to identify and characterize phage DNA using three approaches: 1) the software VirSorter, 2) BLAST and Blob, and 3) Sequence coverage threshold. We detected presumptive phage and prophage sequences in all of the eight data sets analyzed. We also note the constant presence of a single contig sequence with very high base coverage, a length of exactly 5,461 bp and similar gene content in all eight nematode metagenomes. The strengths and weaknesses of each viral-searching approach in this study were discussed, and we suggest the combined use of multiple discovery strategies to identify phages in metagenomic data.

Introduction

Bacteriophages are diverse and ubiquitous components of ecosystems, playing central roles in the ecological and evolutionary dynamics of all microbial communities on Earth (Clokie et al., 2011; Keen, 2015; Paez-Espino et al., 2016). Virtually every existing bacterium is thought to be infected by at least one (and most likely more than one) phage strain (Paul et al., 2002). Phages directly shape the diversity, ecology and evolution of their bacterial hosts in many ways, including altering the host's biological properties and

fitness (Roossinck, 2011a, 2011b), regulating bacterial populations and concurrently maintaining the community's diversity (Weinbauer and Rassoulzadegan, 2003), fueling the phage-host evolutionary arms races (Gómez and Buckling, 2011), and mediating horizontal gene transfers (Keen, 2015; Sano et al., 2004). Given their importance, detailed understanding of phages is necessary to explore the enormous complexity, evolutionary dynamics, and the interaction network in microbiomes.

In spite of their prevalence and influence, phages remain relatively understudied compared to their bacterial hosts. Traditionally, characterization of environmental phages largely relies on laboratory cultivation, which isolates and propagates phage particles on defined single-strain hosts (Green and Joseph, 2012). Cultured phages may then be described using quantitative techniques e.g. serology, electron microscopy, three-dimensional image reconstruction, and mass spectrometry (Ackermann, 2012; Amos and Klug, 1975; Ceyssens et al., 2006; Dowell and Rosenblum, 1962; Rosenblum and Tyrone, 1964). Nonetheless, the culturing procedure has several significant limitations, including long hands-on procedure time, suboptimal culturing conditions, and uncultivable bacterial hosts. It is estimated that only less than 5% of bacteria in the environment can be grown under laboratory conditions (Rappé and Giovannoni, 2003), and even then, the presence of phages have not been reported in these cultivated bacteria (Chen and Novick, 2009; Duhaime and Sullivan, 2012). Although cultivation-based approaches, for over a century, have provided great fundamental knowledge of phage biology and phage-host interactions, it is unlikely a comprehensive perspective of phages in natural microbial systems can be obtained using these methods.

To circumvent the culture step, various molecular technologies have been developed to characterize phages at the genetic level e.g. gene marker-based survey (Breitbart et al., 2004; Comeau and Krisch, 2008; Sullivan et al., 2006), restriction fragment length polymorphisms (RFLP) (Karama and Gyles, 2008), randomly amplified polymorphic DNA (RAPD) PCR (Winget and Wommack, 2008), and single virus genomics (Allen et al., 2011). These techniques have been applied successfully to reveal ample insights into the composition and diversity of phages in environmental samples. Nevertheless, they also face technical obstacles, including the lack of universally present regions or genes in all phages and the lack of quantitative information achieved. Therefore, a method capable of providing a broader view of all viral sequences and more data for inferences is anticipated.

In the past few years, the advance of metagenomic strategies has offered a new, efficient approach to unveil a broader picture of bacteria and their phages in microbiomes, including the unculturable components. Metagenomics provides a comprehensive picture of all the sequences in a biological sample without undergoing marker gene amplification step and its bias. This may allow details of the community's phage gene composition and possibly detecting phage genes of abundance and/or significance based on the sequence coverage in the data set. Specific protocols have been developed to specially target phages in microbiomes by including an additional enrichment step for viral particles (Bachrach and Friedmann, 1971; Shkoporov et al., 2018; Thurber et al., 2009). Yet, it is worth noting that to date, only about 4% of all present microbiome datasets were generated with this viral-targeted approach (Paez-Espino et al., 2017), and for the rest, phage sequences - though presumably present in

abundance - have usually been overlooked. Phage DNA sequences in these untargeted data sets, however, could be a useful resource to unveil more information about the prophages residing as lysogenic elements in bacterial genomes as well as phage particle-bound forms in the microbiome.

"Cyst nematode" is the name commonly given to certain species of plant-parasitic roundworms, mostly from the genera *Heterodera* and *Globodera*, which can form a structure called "cyst". Adult female cyst nematode typically enlarges her body into a spherical shape to retain eggs inside, and when she dies, the sac-like body browns to become a cyst shielding the eggs until hatching conditions (Bernard et al., 2017; Jones et al., 2013). Cyst nematodes are internationally recognized as major plant pathogens (Contina et al., 2018; Lilley et al., 2005), with the most damaging species include the potato cyst nematode *Globodera pallida* and the soybean cyst nematode *Heterodera glycines. G. pallida* could cause up to 80% yield loss of potato in severely infected areas and 9% loss of production globally (Jones et al., 2013), while *H. glycines* infestation is estimated to result in $1,29 billion yearly soybean loss - solely in the US (Bernard et al., 2017). Given the economical impact, detailed knowledge of the nematodes' biology and interactions with other organisms is required to assess their pathogenicity and to optimize control strategies.

The microbial communities associated with *H. glycines* and *G. pallida*, which can be integral to the nematode's lifestyle and development, have been attracting growing interest (Eberlein et al., 2016; Nour et al., 2003; Zhu et al., 2013). However, the identities and diversity of phages within these microbiomes have not been investigated. In this study, we investigated eight untargeted nematode metagenome datasets - four *G. pallida*

and four *H. glycines* to search for phage DNA using three independent approaches: 1) Viral contigs searching by the software VirSorter, 2) BLAST and Blob, and 3) Sequence coverage threshold. The program VirSorter (Roux et al., 2015) is a widely-used tool to detect both prophage sequences and lytic particles in assembled metagenome data sets using gene-based comparison to available viral gene databases and other types of evidence e.g. short and uncharacterized genes. The second approach relied on nucleotide sequence similarity searches against complete phage genome databases. The third approach specifically targeted lytic sequences, assuming that free particles, as the result of a lytic burst, should be present in large numbers and therefore, lytic sequences should have high sequencing coverage. Understanding of phages as a component in these nematode's microbiome should contribute to a broader understanding of the microbiome as a functioning intricate system.

Materials and Methods

Nematode sampling

A total of eight nematode samples, four potato cyst (*G. pallida*) and four soybean cyst nematodes (*H. glycines*) were collected from eight different locations. The 'field' potato cyst nematode samples (Gp_Bin25, Gp_Bin26, and Gp_Bin258) were collected from infested fields in Southern Idaho during the years from 2006 to 2014. In each location, 22.42 kg/hectare of soil were collected, from which cysts were extracted using the Fenwick flotation method (Fenwick, 1940). The 'greenhouse' sample (named Gp_GH) was collected from cysts formed by *G.pallida* in the University of Idaho's greenhouse in January 2017. This greenhouse population derived from the field sample Gp_Bin25. The

soybean cyst nematode samples (Hg_Al, Hg_Aud, Hg_Pet, and Hg_War) were collected in Alabama and Missouri, USA. Details of the samples' collection information and assembly statistics are shown in Table 3.1.

DNA extraction, metagenome sequencing and assembly

For each sample, 50 cysts were pooled for DNA extraction. Cysts were homogenized by mortar and motorized micro pestle. DNA was isolated from the homogenate using QIAmpDNA Micro kit (Qiagen, Hilden, Germany) and quantified by the Qubit dsDNA HS Assay kit.

The obtained DNA was then sheared to peak library fragment size, ~500 bp. Libraries were prepared using the NEBNext® Ultra™ II DNA Library Prep Kit for Illumina (San Diego, CA, USA). Paired-end 150 bp sequencing of barcoded DNA libraries was performed using the Illumina HiSeq 3000 at the Center for Genome Research and Biocomputing at Oregon State University (Corvallis, OR, USA). Raw reads are available from NCBI's Sequence Read Archive (SRA).

Raw reads were trimmed and quality controlled using bbduk (DOE Joint Genome Institute). Reads shorter than 50 bp and/or containing more than 20% of called bases with low Phred quality scores (Q<20) were excluded from datasets. Filtered reads were then *de novo* assembled with MetaSPAdes (Nurk et al., 2017) to generate contigs for downstream annotation. Contigs less than 300 bp in length were excluded from subsequent analyses.

Approach #1. Phage sequence identification by VirSorter

The metagenome assemblies were examined for the presence of phage sequences. Viral contigs were first predicted using the program VirSorter (Roux et al., 2015), which searched for the presence of 'hallmark' viral genes (e.g. capsid protein, spike, terminase subunit, portal protein), signatures of enrichment of virus-like, non-Caudovirales, short, or uncharacterized genes, depletion of Pfam-affiliated genes or strand switch. If a predicted region on a contig was more than 80% length of the contig, the entire contig would be classified a "phage"; if it was less than 80%, the subset of the contig would be called a "prophage". Only contigs predicted by the program as 'phage' or 'prophage' with higher confidence (categories 1, 2 for "phages" and 4, 5 for "prophages") were included in subsequent analysis. Characteristics of the contigs were summarized and visualized using the anvi'o pipeline (Eren et al., 2015), Blobtools (Laetsch and Blaxter, 2017), and R.

Predicted phage contig sequences of categories 1 and 2 were annotated using RAST subsystem (Aziz et al., 2008) and prokka (Seemann, 2014). The program Phamerator (Cresawn et al., 2011) was used to map contig sequences and assemble gene families with a ClustalW threshold of 35% amino acid identity and a BLASTP score of 1e-50.

Approach #2. BLAST and Blob - Detecting phage DNA using complete genome databases

Complete genome sequences of phages infecting bacteria of the three genera *Pseudomonas, Mycobacteria* and *Bacillus* then available on the NCBI GenBank database

were downloaded as FASTA sequences to build local phage databases. Search in the NCBI GenBank Nucleotide with keywords specific to the bacterial genus, (e.g. "*(Pseudomonas phage[Title]) AND Pseudomonas[All Fields] AND complete genome*") resulted in 306 DNA sequences of *Pseudomonas* phages, 206 sequences of *Bacillus* phages and 1,237 sequences of *Mycobacteriophages*, which were assembled to build three phage databases in February 2019. To detect signals for phages infecting each bacterial genus, all contigs in the greenhouse *G. pallida* (Gp_GH) and *H. glycines* from Alabama (Hg_Al) datasets were compared to the three phage databases using TBLASTX with an e-value threshold 1e-5 and minimum score of 50.

Approach #3. Sequence coverage threshold

We targeted metagenomic contigs displaying exceptionally high coverage patterns for focused analysis, reasoning that high-abundance lytic phage particle DNA might be present at high levels in the metagenomes. All contigs with length greater than 3 kb and outstandingly high coverage, determined by the Interquartile Range method as extreme outliers (lying more than five Interquartile Range above the third quartile of all contigs' coverage in the dataset) were selected for further analysis. The threshold 3 kb was chosen based on the length of the shortest *Pseudomonas* phage sequence reported (Ha and Denver, 2018). To examine their biological origin, the contigs of interest were annotated with the RAST annotation server (Aziz et al., 2008), prokka (Seemann, 2014), and compared to the NCBI nr/nt database using BLASTN megablast with a threshold of 95% identity and maximum E-value of 1e-25.

Comparisons of 5461-bp-long 'phage' contigs in all datasets

For each dataset, one contig with a length of 5461 bp and outstandingly high coverage was extracted. Sequences were aligned using MAFFT alignment module (Katoh, 2002). To investigate the variation within a dataset at this particular contig, we performed variant calling in Geneious Prime 2019.2.1.

Results

Approach #1. Phage DNA detected by VirSorter in nematode metagenomes

The analysis based on VirSorter detected from 2 to 8 prophages and 6 to 62 phage sequences (i.e. more than 80% of the total length was predicted to be phage genes), with an average of 4 prophages and 32 phage sequences per data set. These sequences were predicted to have originated from phage genomes with high confidence (VirSorter categories 1, 2, 4, and 5). The predicted phage sequences accounted for from 0.05% (Hg_War and Hg_Pet) to 0.16% (Hg_Al) of the total length of all sequences in the whole data set. The total number of putative phage genes found in the eight metagenomes varied greatly from 87 (data set Hg_War) to 881 (Gp_Bin26), with a mean of 388 phage genes. The shortest phage contig identified was 689 bp long, and 97,440 bp was the size of the longest one. Details of the phage contigs found by VirSorter are provided in Figure 3.1A-H and Table 3.2. The GC contents and average base sequencing coverages of predicted phage contigs in relative to other contigs in the same metagenome assembly are demonstrated in Figure 3.2A-H.

Within one data set, phage sequence features also varied considerably. For example, phage contig length in Gp_GH ranged from 1,304 to 79,137 bp, and the number of predicted phage genes varied from 2 to 118 (Figure 3.1G), having a mean length of

6,732 bp and a very large standard deviation of 11,871 bp. Fifty-four contigs of category 1 and 2, i.e. classified as complete phage contigs with high confidence (see Materials and Methods) in Gp_GH data set were selected for gene content clustering analysis. A total of 489 genes predicted in the contigs were assigned into 456 families, out of which 93.6% (427/ 456) are 'orphams' i.e. genes that shared no detectable relationships with other genes in all of the 54 contigs analyzed. No tight clustering was observed among these sequences as the content of gene families varied greatly. The relationships among these contigs based on shared gene families were illustrated in Figure 3.3.

Approach #2. Phage genome sequence search strategy

We searched for phage DNA sequences using TBLASTX comparison with three local databases built from complete genomes of phages originally isolated from three bacterial genera *Pseudomonas, Bacillus,* and *Mycobacterium.* Due to the extensive computational resource that TBLASTX similarity search method requires, we only performed the BLAST and Blob approach on two metagenome data sets *G. pallida* greenhouse (Gp_GH) and *H. glycines* Alabama sample (Hg_Al). Each of these data sets was found to contain a high number of phage sequences predicted by VirSorter in Approach #1. In the data set Gp_GH, we found 7,567 contigs spanning 17,160,241 bp that match *Bacillus* phage genomes; 4,217 (totaling 12,511,463 bp) and 2,635 contigs (9,730,094 bp) match to *Pseudomonas* phage and Mycobacteriophage databases, respectively. There were 492 contigs with a total length of 6,494,777 bp showing similarity with sequences of all three types of phages (Figure 3.4A). In the data set Hg_Al, 523 contigs (2,416,344 bp) aligned with *Pseudomonas* phage sequences, 319 contigs (1,593,218 bp) to Mycobacteriophages, and 904 contigs (3,940,898 bp) hit to *Bacillus* genomes. Seventy-two contigs (total length

882,201 bp) matched partially or completely to all three phage databases (Figure 3.4B). The GC contents and coverage of contigs that match all three databases compared to other contigs in the same metagenome assembly are illustrated in Figure 3.5.

Details of the distribution of GC content, size, sequence coverage, and the percentage of contigs that match phage DNA of each of the three groups are displayed in Appendix Figure 1.

Approach #3. Searching for putative lytic phage DNA using a coverage threshold

A five Interquartile Range coverage threshold was applied to specifically target sequences with very high base coverage, a metagenomic feature expected of sequences of phages that recently experienced a lytic burst. In the eight data sets, the number of contigs that showed homology to characterized sequences in the nr/nt database varied from 8% (Hg_Al) to 93% (Gp_Bin26) of the contigs that passed that threshold (Figure 3.6). Among these identified sequences, we detected groups that are expected to be present at high coverage, such as mitochondrial DNA, ribosomal RNA, and satellite DNA.

One circular contig with a fixed length of 5,461 bp, had very high coverage compared to other contigs in the same assembly, and similar gene annotation was detected in all of the eight meta-datasets analyzed; and was identified as phage sequence. The contigs were annotated with the following seven phage signature genes, listed in order: phage DNA replication protein, single-stranded DNA synthesis, external scaffolding protein D, DNA-binding protein, major capsid protein, major spike protein, minor spike protein (DNA pilot protein). The features of this particular contig in the dataset Gp_GH was shown in Figure 3.7A.

Little nucleotide variation was observed for this circular contig within each dataset - only one transversion SNP of frequency less than 7%, with two exceptions: a transition of 33.3% frequency and a double nucleotide substitution of 14.3% frequency in the datasets Gp_Bin25 and Gp_Bin258, respectively. Details of the variant calling results are listed in Table 3.3. Between different datasets, the contigs also showed limited variations, with pairwise percentage of identity all exceeds 97%. Visualization of comparison between these 'phage' contig sequences was demonstrated in Figure 3.7B, and their pairwise percentage of identity were shown in Table 3.4.

Discussion

Evaluation of approaches to identify and characterize phage DNA

The contribution of phage DNA to metagenomic datasets remains an underexplored question in the postgenomic era. This "viral dark matter" consists of sequences that originate from viruses but otherwise show no detectable similarity with previously characterized sequences, and are estimated to account for 40 to 90% of a dataset (Krishnamurthy and Wang, 2017). This uncertainty is largely due to limitations in existing reference viral genomes or protein databases. In addition, phage sequences are greatly diverse and often mutate at a high rate (Rohwer et al., 2014; Rose et al., 2016), which might confound sequence alignment and classification. These ongoing unresolved issues require the application of diverse and complementary bioinformatic strategies to identify and characterize phage DNA in metagenome data.

Here in this study, we applied three different strategies to identify phage DNA in cyst nematode metagenome data sets: 1) the phage prediction tool VirSorter, 2) Alignment-based similarity search using known phage complete genome databases - “BLAST and Blob”, and 3) a sequence coverage threshold-based method. While VirSorter mainly utilizes a phage gene-based comparison approach, the BLAST and Blob approach utilizes complete phage genome sequences. Coverage threshold also depends on sequence comparisons, but specially focuses analysis on high-coverage contigs. These three approaches are expected to deliver different strengths and weaknesses.

The phage-predicting tool VirSorter executes multiple evaluations in addition to gene-based similarity searches on assembled metagenomic contigs to detect phage signals, enabling more reliable elucidations of phage sequences and the semi-independence from currently available databases of viral protein sequences. It could detect both lytic and lysogenic sequences, including potentially novel viruses. However, as VirSorter assigns the categories ‘phage’ or ‘prophage’ to a contig based only on the proportion of that contig predicted to be phage, with a rather arbitrary threshold (80%), the classification might be less accurate to some extent, for example, a contig labeled “phage” could actually be part of a prophage, or a fragmented phage genome could be labeled “prophage” for having longer non-coding region, which might account for up to 19% of some viral genomes (Mahmoudabadi and Phillips, 2018). Another disadvantage of VirSorter is that the program may perform inefficiently on short (especially < 3 kb) viral contigs (Roux et al., 2015). Lastly, predictions by protein-based search strategy may be compounded by fragmented viral genes, which could be prevalent in metagenomic data, leading to false negative detection.

The second approach, BLAST and Blob, deviated from gene-based searching and instead performed comparison with complete phage genome sequences. The search tool utilized, TBLASTX, which compares the six-frame translations of metagenomic sequences against the six-frame translations of complete phage genome databases is especially suitable for fast-modifying viral sequences, lowering the chance of false negatives. Since sequences were compared directly, this approach was less sensitive to the size of contigs or fragmented viral genes. However, a potentially large number of hits found by this strategy could be false positives. For instance, bacterial contigs containing genes that were present in phage genome sequences via earlier horizontal gene transfers would be reported as a 'phage hit'. As phages constantly acquire genes or sequence fragments from their hosts (Breitbart et al., 2007; De Paepe et al., 2014), this could be significantly problematic to discern the true origin of the sequences. Another clear weakness of this method is that its sequence similarity search to known phage database could not facilitate detecting novel phage sequences.

Lastly, the coverage threshold method relied on the assumption that lytic sequences might have higher copy number and thus higher sequencing coverage compared to the majority of other contigs in the same assembly. On average, a lytic burst could release around 100 particles per host cell, although the burst size largely depends on the phage and the latent time (Heilmann et al., 2010; Wang, 2006). Using the coverage threshold, we identified sequences of varied biological origins that were expected to be present in high copy number, such as mitochondrial DNA, ribosomal RNA gene, satellite DNA, which might lend more plausibility to the method. The phage contig found was packed with phage genes and annotated with high confidence. However, this approach

was clearly biased to target phage particles and not likely to find prophages. Furthermore, similar to BLAST and Blob, the annotation step did not allow identifying novel viral sequences with no known reference in existing databases.

Complex phage signal detected in cyst nematode metagenome data sets

Phage sequences were identified by all of the three approaches in all eight cyst metagenomic datasets analyzed. The number of phage sequences varied among datasets, ranging from 12 to 64 sequences found by VirSorter. These numbers are considerably lower than those found by many previous studies using the same viral prediction tool, which are usually around hundreds of sequences (Garin-Fernandez et al., 2018; Leigh et al., 2018; Miller-Ensminger et al., 2018; Nigro et al., 2018). It may first be explained by the lack of a viral enrichment procedure when generating the untargeted nematode meta-datasets examined in this study, while the other studies all included standard steps of viral purification and enrichment. However, the consistent detection of phage sequences in all of our eight data sets suggested that generic untargeted metagenomes may be a useful resource to study phage diversity and significance.

The BLAST and Blob results of the *G. pallida* data set showed that sequences sharing similarity with *Bacillus* phage genomes were the most abundant, followed by those matching *Pseudomonas* phages, and Mycobacteriophage database had the least number of matches; this pattern was also observed in the *H. glycines* data set examined. It could be possible that the diversity of bacterial host species in each phage databases might have contributed to the match abundance observed. The *Bacillus* phages included in the local database infected a total of seventeen *Bacillus* host species with no clear skew towards a particular host. Meanwhile, only nine species of *Pseudomonas* bacterial host

species were reported among all the *Pseudomonas* phages, 75% of which were *P. aeruginosa*. And although the Mycobacteriophage database was the largest among the three, almost all of the Mycobacteriophages in the reference set were isolated from a single host species *Mycobacterium smegmatis*. The bacterial host diversity may have resulted in more diverse phage regions and increased the chance of matching with metagenomic sequences.

Signal searching with VirSorter and the TBLASTX approaches both indicated that phage sequence contigs in *G. pallida* data sets exist in larger abundance compared to those of *H. glycines.* For example, the number, individual length, and total length of phage contigs found by the tool VirSorter in Gp_GH were almost twice those of phages found in Hg_Al (Table 3.2). Likewise, TBLASTX search detected a much higher number of contigs that match phage local databases in the Gp_GH data compared to Hg_Al (Figure 3.4). The difference might be accounted for by several reasons. First, it is possible that the pattern might have resulted from analytical discrepancy. Assembly statistics of *H. glycines* data sets indicated better assembled metagenomes in general (Table 3.1). Therefore, phage contig sequences in *G. pallida* microbiomes might actually be present in the same abundance - but more fragmented or just partly complete, and thus added up to a higher total length. Second, the microbial communities of *G. pallida* might contain more bacterial hosts of the phages that we detected. This may potentially explain the higher sequencing coverage of phage contigs and number of ‘prophage’ sequences in Gp_GH in comparison with Hg_Al. For instances, the 5,461-bp phage contig in Gp_GH had a sequencing coverage of 688.7, while the coverage of its counterpart in Hg_Al was 316.9. Furthermore, the bacterial members in Gp_GH may also be susceptible to more

phage particle per host cell. Lastly, there is a possibility that the relatively low abundance of phage sequences in *H. glycines* metagenomes was simply a temporal/seasonal state, as all the four *Hg* samples were collected in January 2018, and could change substantially with time.

The coverage thresholding strategy has revealed a single phage contig that was highly conserved within and between the data sets, with very little nucleotide variation. The entire phage sequence and its particular set of genes are likely to have a significant role in cyst nematode metagenomes. Previous studies have utilized abundance information in metagenomic data to investigate environment-specific changes in the phage composition and infer the significant role of specific phages in different ecological niches (Dutilh et al., 2014; Norman et al., 2015; Roux et al., 2016). Clearly there might be sequencing bias - but over, or under representation of particular genes can yield useful information about phage biology. However, we do not exclude the possibility that the sequence is just an experimental artifact, which was introduced to the metagenomes from contaminants in the DNA extraction kit or other sources during our sample preparation. Nonetheless, we noted that the pairwise differences between these sequences, in most of the cases, were likely to be more significant than what would be observed if they all originated from a contaminant (Appendix Table 3).

Conclusion

In this chapter, we detected varied signals of phage DNA in all of our eight metagenome data sets, and noted the lower number of phage sequences identified in all *H. glycines* data sets compared to those of *G. pallida.* One specific contig with a fixed length and highly similar annotation was found in all of the eight data sets. We also discussed the

strengths and weaknesses of the three independent search approaches utilized in the work, and we suggest applying multiple complementary strategies in phage discovery from metagenomic sequences.

Table 3. 1. Collection information and de novo assembly statistics of the datasets.

Sample	Location (USA)	Date of collection	Total number of QC-passed reads	# of contigs (>= 300 bp)	Largest contig (bp)	N50	L50
Gp_GH	Green house, U of Idaho	Jan 2017	37,117,112	272,561	601,196	1,341	38,690
Gp_Bin25	Site #25, Bingham, Idaho	May 2007	27,323,936	494,453	600,871	789	87,997
Gp_Bin26	Site #26, Bingham, Idaho	May 2007	40,795,618	782,171	1,079,796	841	133,680
Gp_Bin258	Site #258, Bingham, Idaho	May 2014	29,074,518	323,949	595,308	1,009	50,718
Hg_Al	Alabama	Jan 2018	26,524,664	90,110	139,523	2,860	9,311
Hg_Aud	Pettis, Missouri	Jan 2018	57,878,600	122,813	1,533,904	3,164	10,090
Hg_Pet	Audrain, Missouri	Jan 2018	44,414,058	133,920	1,016,637	1,853	11,949
Hg_War	Warren, Missouri	Jan 2018	26,333,758	89,735	809,795	3,321	7,521

Table 3. 2. Phage contigs predicted by the tool VirSorter in each of the eight nematode metagenome data set.

Dataset	Prophage	Phage	No. of genes in phages	Total phage length	Total length of all contigs >= 300 bp	% phage in dataset	Avg. phage length	Min phage length	Max phage length
Gp_GH	6	54	489	363,567	249,932,024	0.15	6,732 ± 11,871	1,304	79,137
Gp_Bin26	2	62	881	578,508	581,324,724	0.10	9,331 ± 16,155	1,418	97,440
Gp_Bin25	2	36	471	349,861	354,249,375	0.10	9,718 ± 17,914	915	87,008
Gp_Bin258	3	45	487	316,802	256,003,752	0.12	7,040 ± 9,281	882	46,884
Hg_Al	3	29	214	189,412	116,527,583	0.16	6,531 ± 4,258	689	15,570
Hg_War	4	8	87	65,058	138,887,545	0.05	8,132 ± 8,380	3,055	28,235
Hg_Pet	8	6	119	85,053	157,290,633	0.05	14,175 ± 14,656	2,596	41,400
Hg_Aud	6	18	356	247,514	186,127,784	0.13	13,750 ± 19,080	1,246	65,170

Table 3. 3. Nucleotide variation of phage sequences observed within each metagenome.

Dataset	Name	Change	Coverage	Polymorphism Type	Variant Frequency	Variant P-Value (approximate)
Gp_GH	T	G -> T	648	SNP (transversion)	4.50%	9.10E-58
Hg_Al	A	C -> A	300	SNP (transversion)	4.00%	3.40E-24
Hg_War	T	G -> T	212	SNP (transversion)	5.20%	1.40E-23
Gp_Bin26	T	G -> T	361	SNP (transversion)	6.90%	2.20E-52
Gp_Bin25	A	C -> A	651	SNP (transversion)	5.10%	4.90E-64
	G	A -> G	39	SNP (transition)	33.30%	1.20E-15
Gp_Bin258	T	G -> T	332	SNP (transversion)	5.40%	5.60E-38
	AA	TT -> AA	7	Substitution	14.30%	0.00056
Hg_Pet	T	G -> T	692	SNP (transversion)	4.30%	2.50E-50
Hg_Aud	A	C -> A	443	SNP (transversion)	5.60%	3.60E-40

Table 3. 4. Pairwise percentage of identity between the 5,461-bp long contigs with high sequence coverage from each of the eight datasets analyzed.

	Hg_Al_NODE_3798	Gp_25_NODE_3025	Hg_Aud_NODE_4214	Gp_GH_NODE_3750	Hg_Pet_NODE_1496	Gp_258_NODE_3062	Gp_26_NODE_7217	Hg_War_NODE_3465
Hg_Al_NODE_3798		98.9	97.2	97.2	98.5	98.5	98.5	97.2
Gp_25_NODE_3025			97.2	97.2	98.5	98.5	98.5	97.2
Hg_Aud_NODE_4214				97.3	98.6	98.6	98.6	97.3
Gp_GH_NODE_3750					98.6	98.6	98.6	97.3
Hg_Pet_NODE_1496						100	100	98.6
Gp_258_NODE_3062							100	98.6
Gp_26_NODE_7217								98.6
Hg_War_NODE_3465								

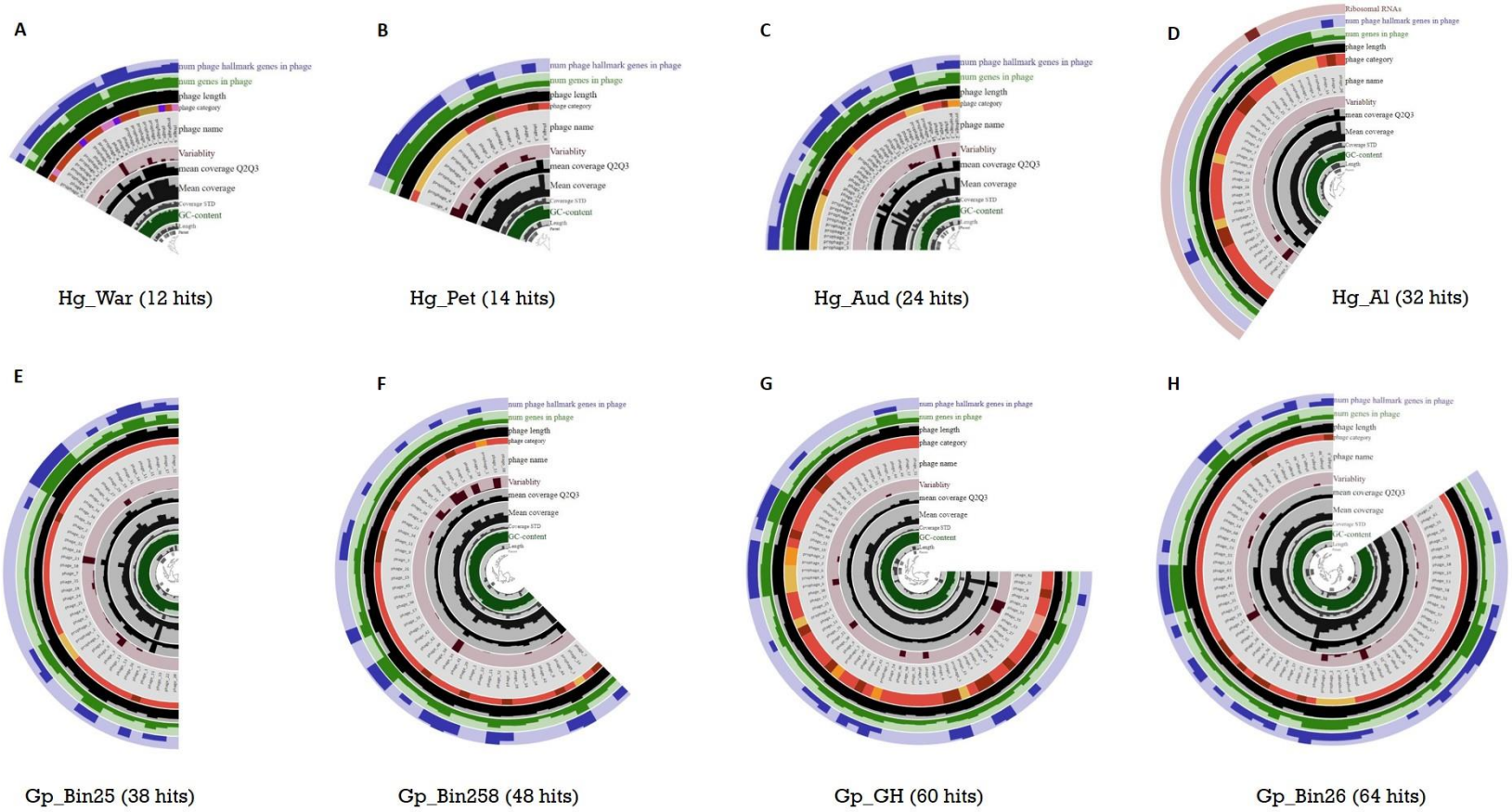
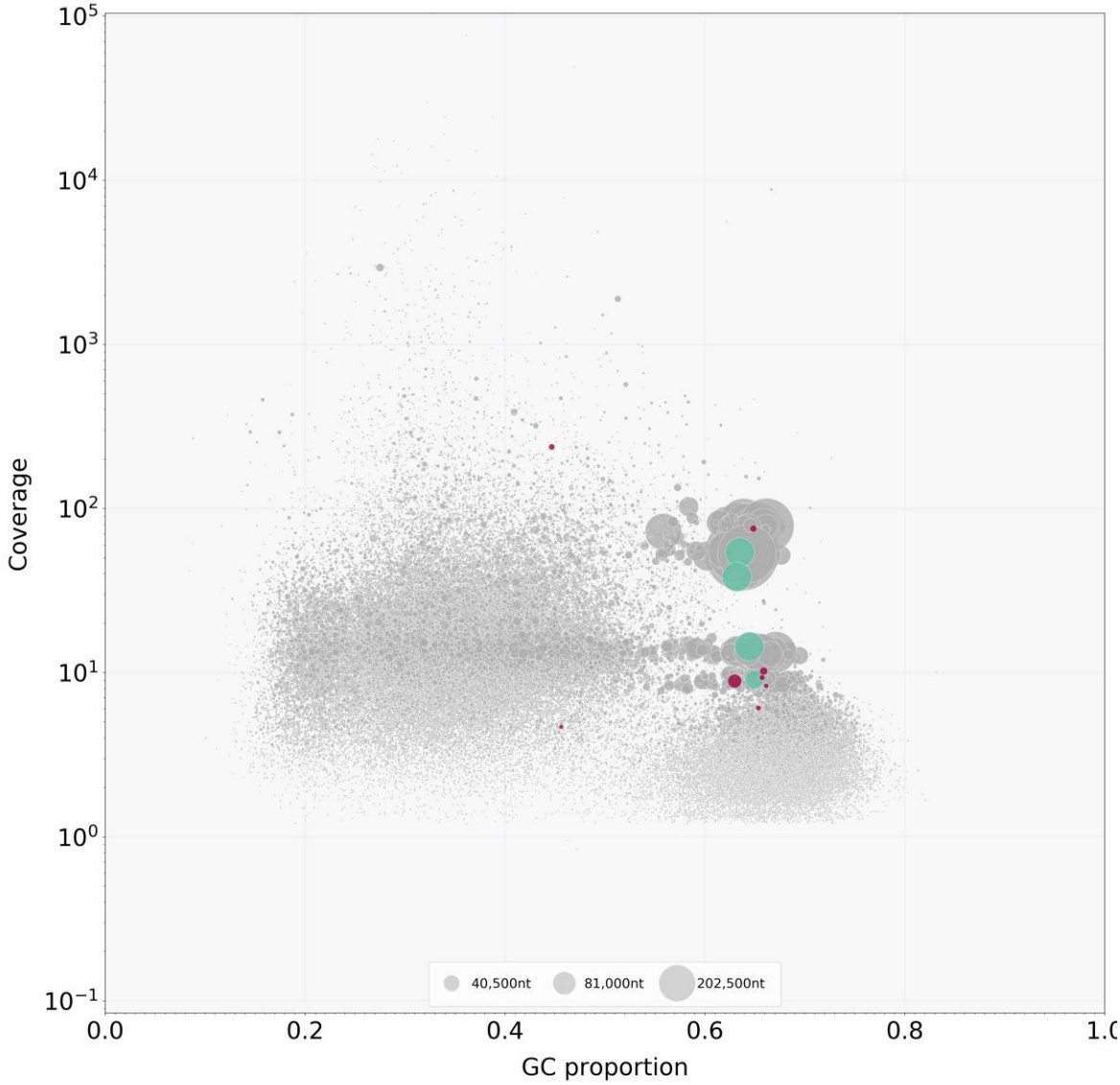
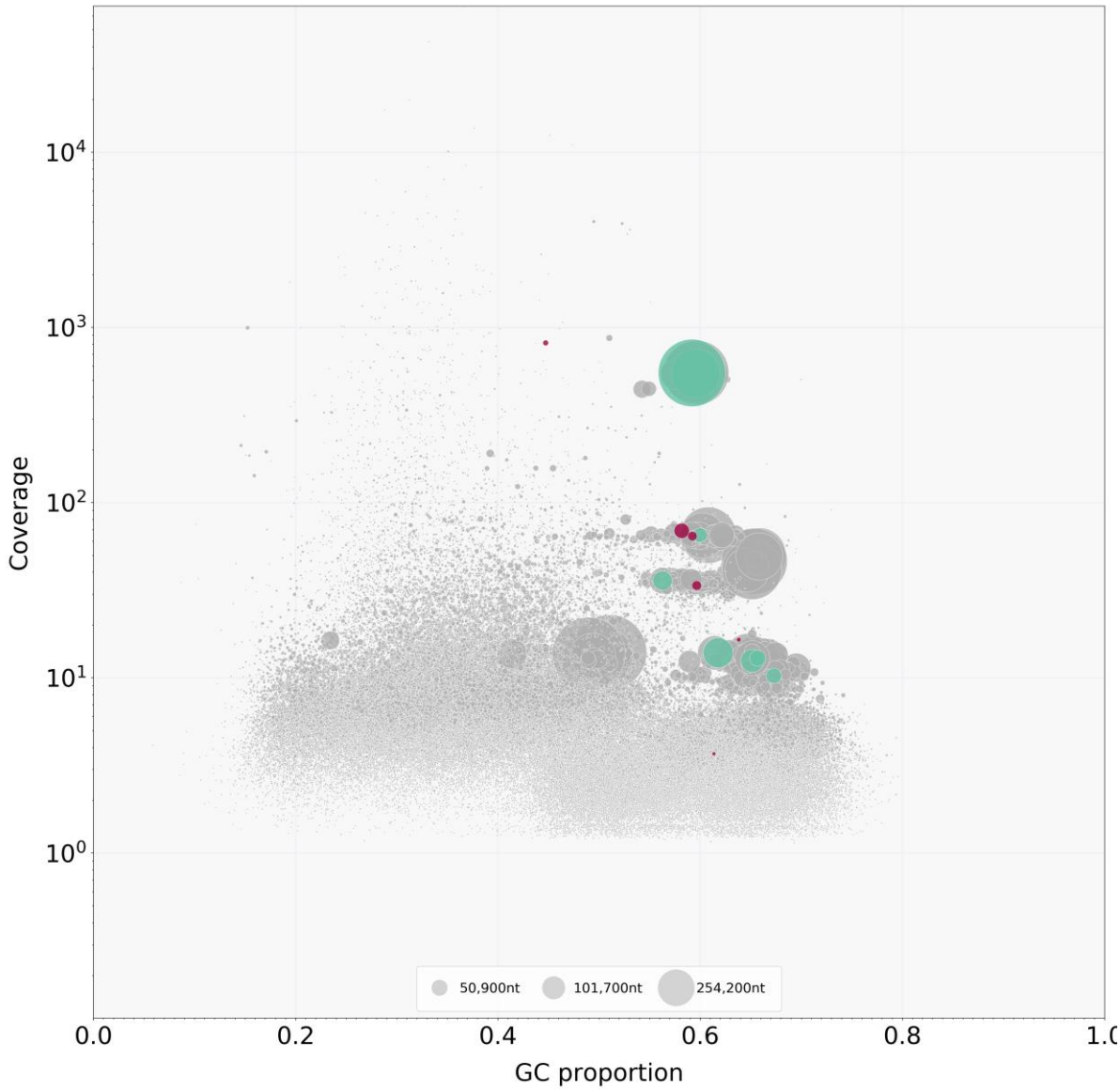


Figure 3.1. Basic characteristics of predicted phage and prophage contigs in the dataset Hg_War (A), Hg_Pet (B), Hg_Aud (C), Hg_Al (D), Gp_Bin25 (E), Gp_Bin258 (F), Gp_GH (G), and Gp_Bin 26 (H). Each ring is a basic feature of the sequences, ordered from inner to outer rings: GC content, standard deviation of sequence coverage, mean sequence coverage, mean sequence coverage Q2Q3, variability, phage name, phage category, phage length, number of genes in phage sequences, and number of hallmark genes in the contigs.

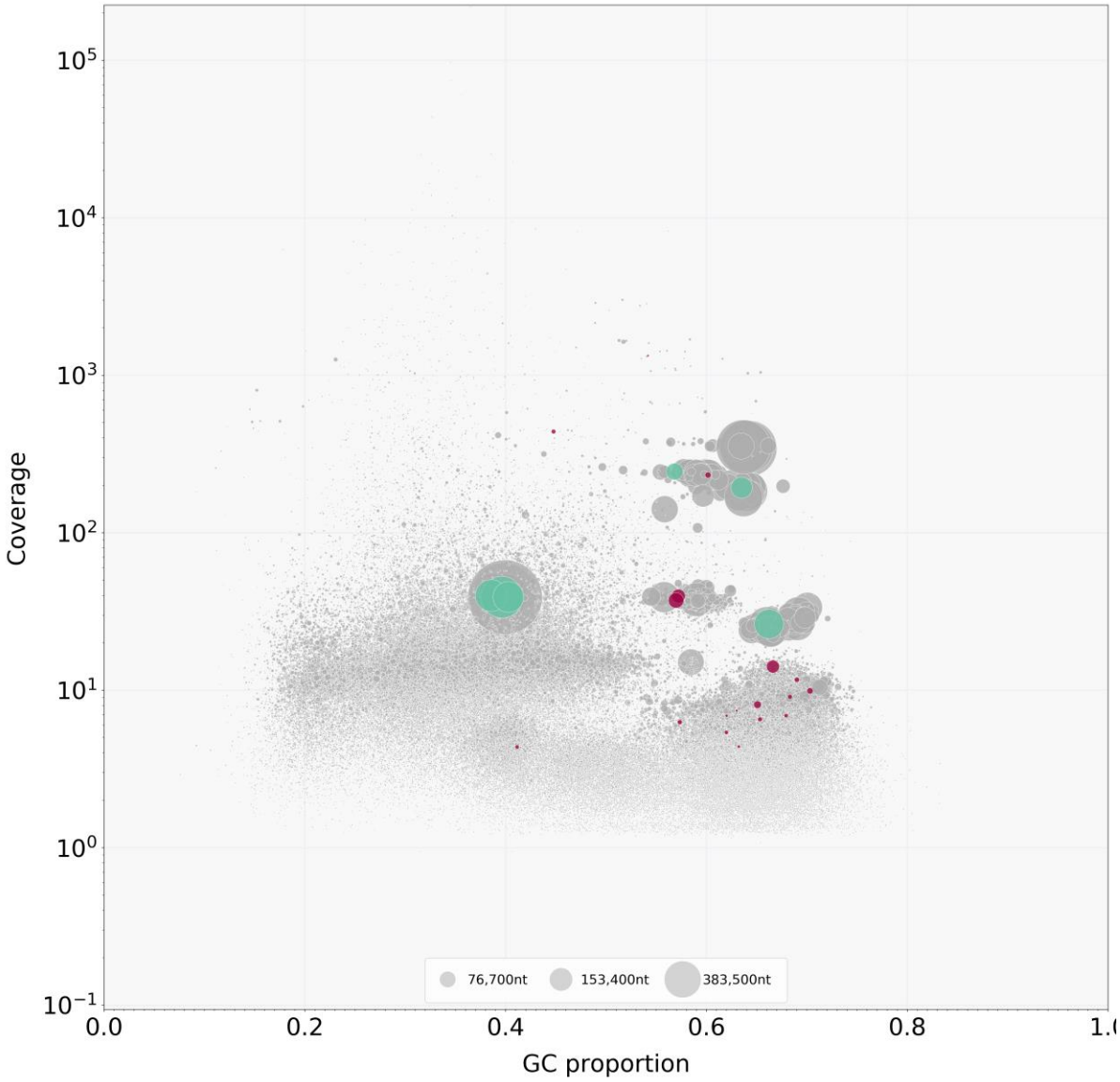
A



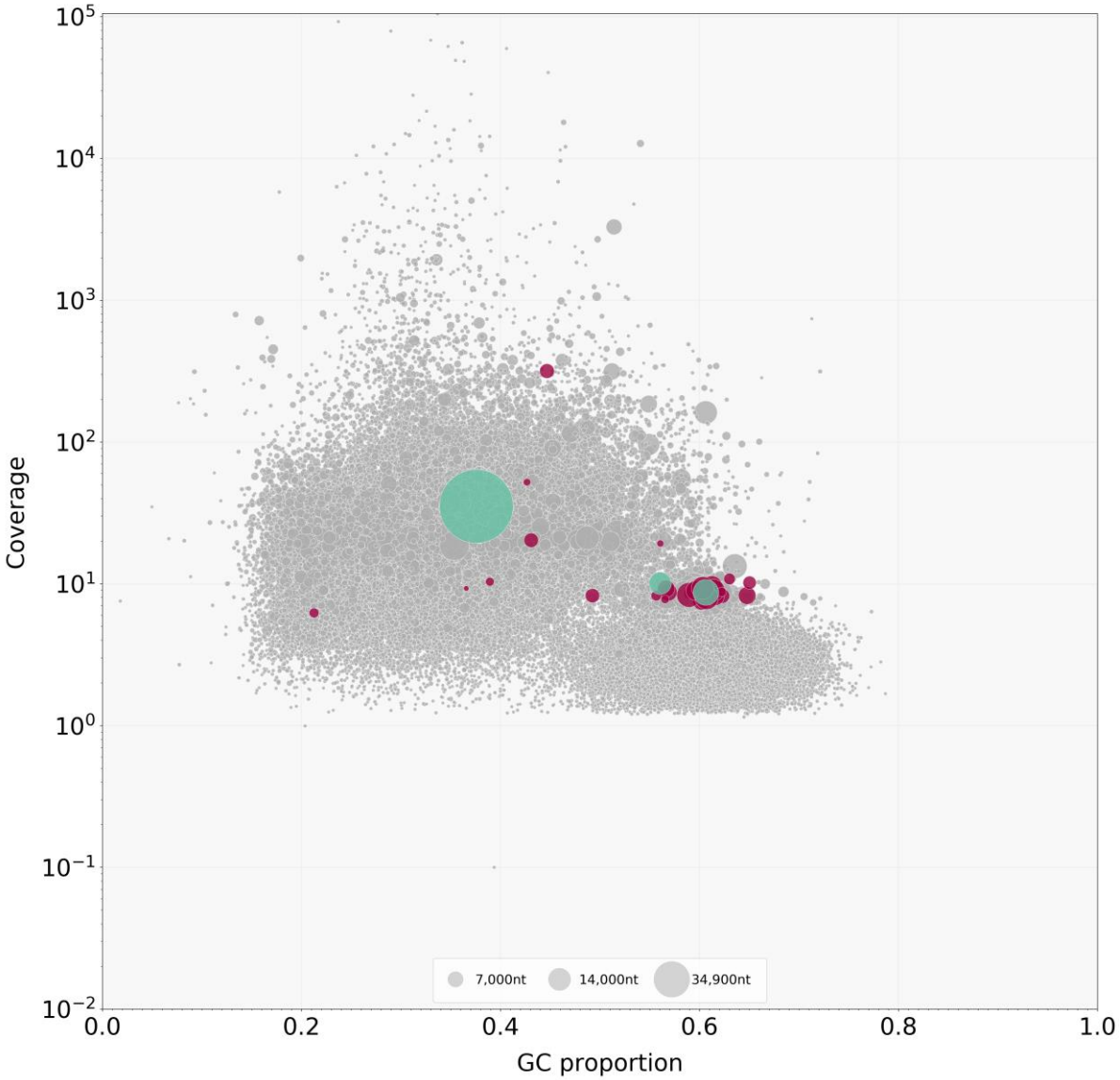
B


Coverage
10^4
10^3
10^2
10^1
10^0
50,900nt
101,700nt
254,200nt
0.0
0.2
0.4
0.6
0.8
1.0
GC proportion

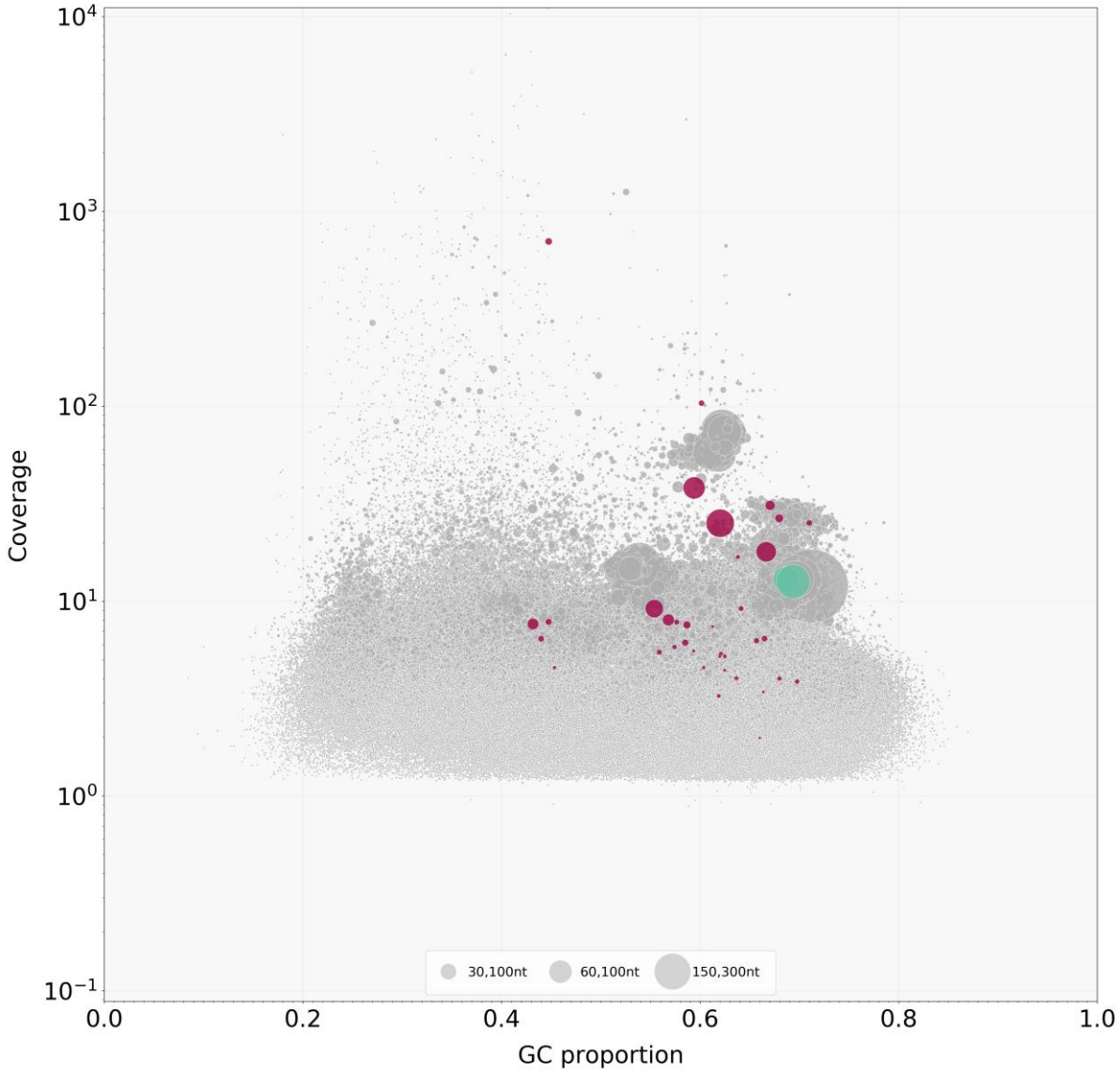
C



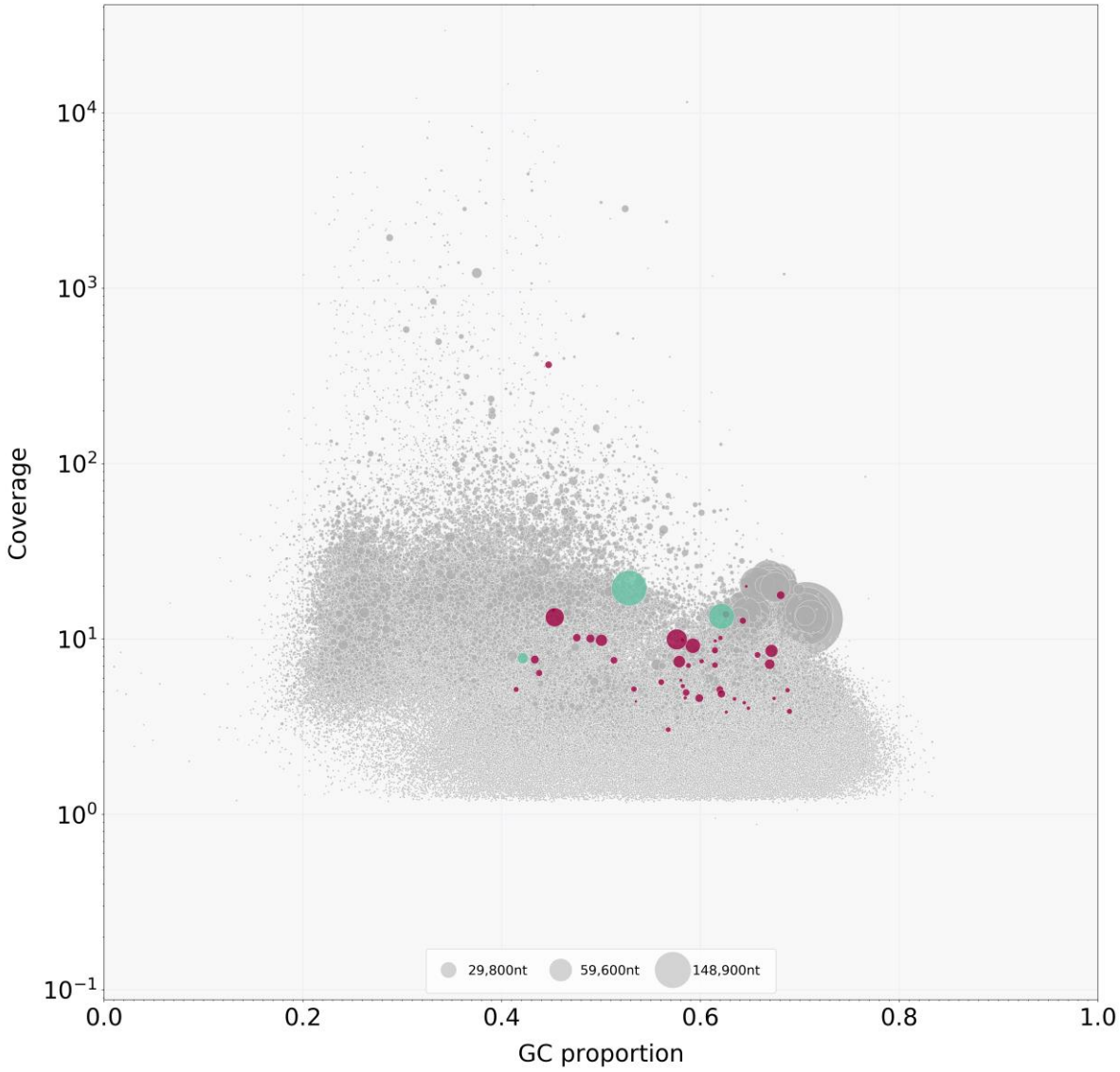
D



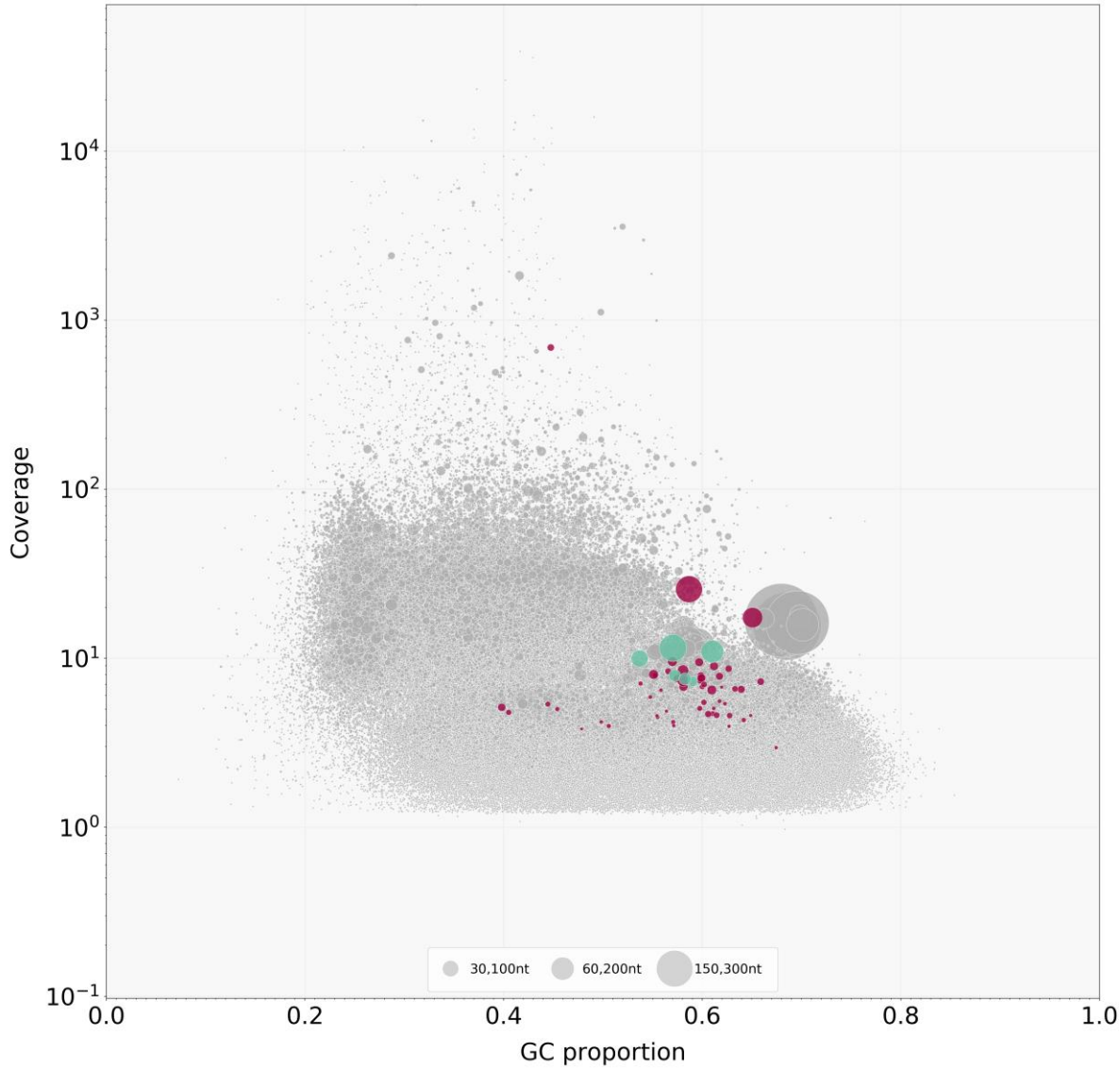
E



F


Coverage
10^4
10^3
10^2
10^1
10^0
10^{-1}
0.0
0.2
0.4
0.6
0.8
1.0
GC proportion
29,800nt
59,600nt
148,900nt

G


Coverage
10^4
10^3
10^2
10^1
10^0
10^{-1}
0.0
0.2
0.4
0.6
0.8
1.0
GC proportion
30,100nt
60,200nt
150,300nt

H

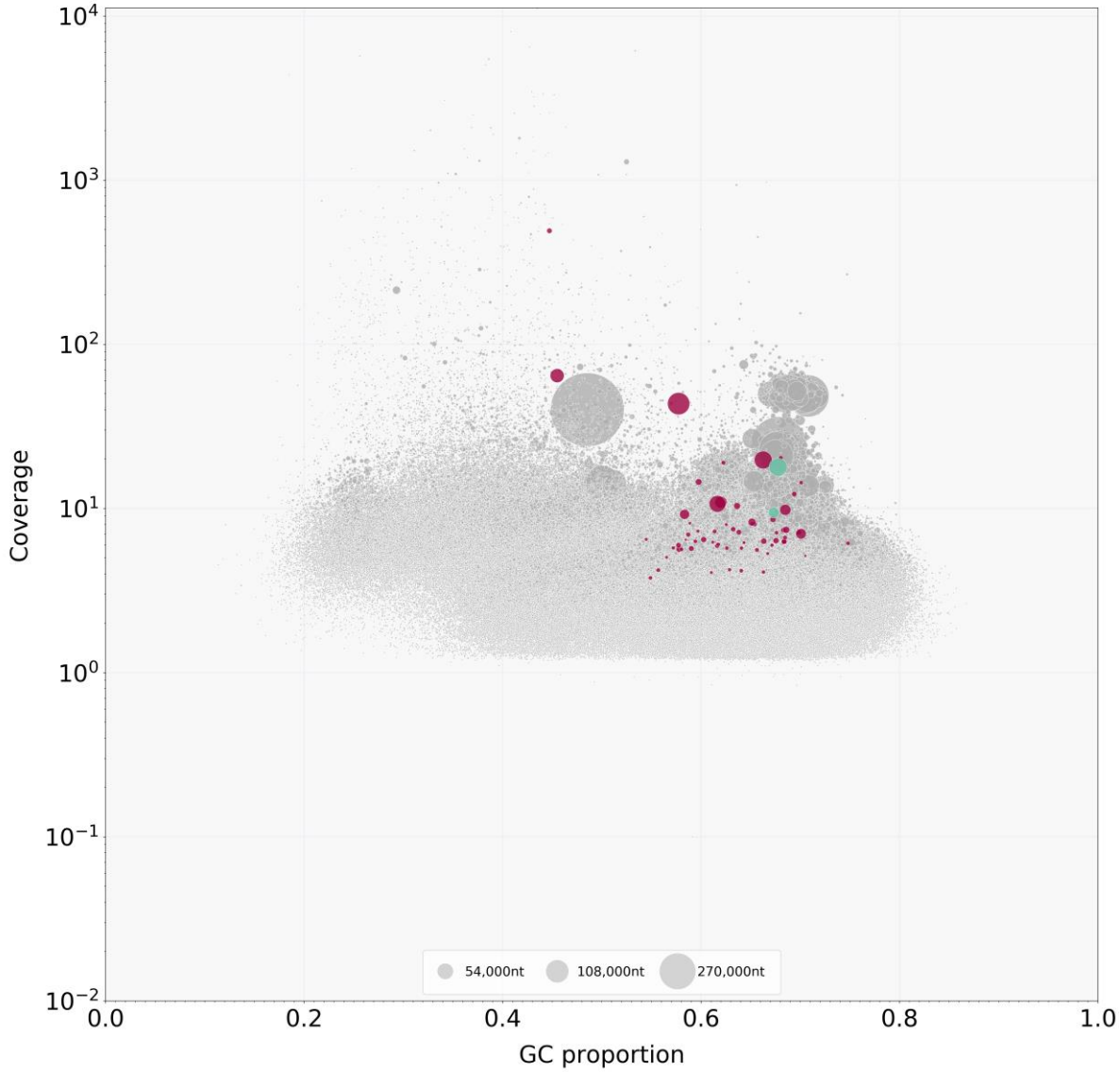


Figure 3. 2. GC content and base sequencing coverage of contigs predicted by the program VirSorter to be of phage origin in the dataset Hg_War (A), Hg_Pet (B), Hg_Aud (C), Hg_Al (D), Gp_Bin25 (E), Gp_Bin258 (F), Gp_GH (G), and Gp_Bin 26 (H). The base coverage of contigs are indicated on the Y axis. The GC content of the contigs are shown on the X axis. Each dot represents a contig, and the size of the dot correlates to contig length. “Phage” contigs are demonstrated in red and “prophage” are in green (categories were designated by VirSorter). One contig identified as ‘phage’ has a very high read coverage.

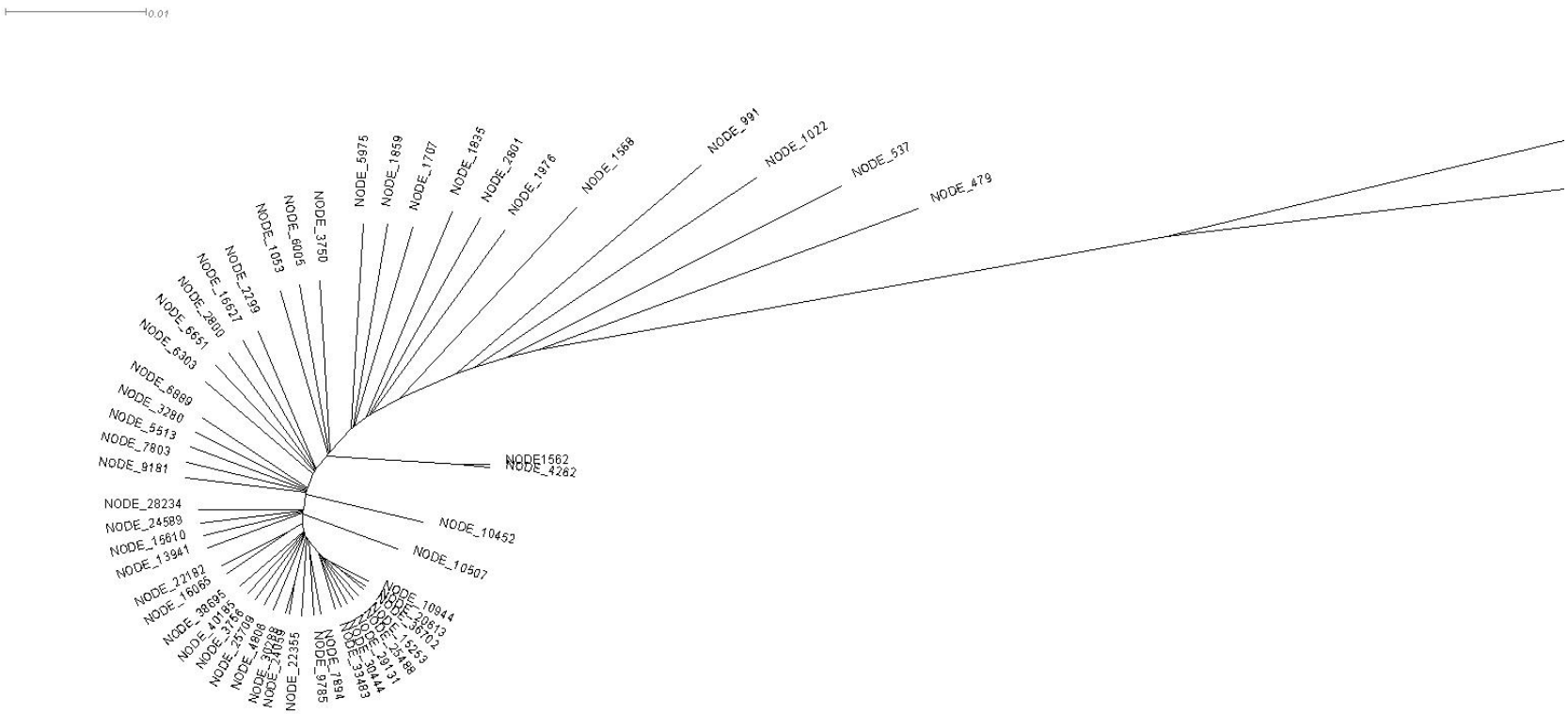


Figure 3.3. Relationship between the 54 VirSorter-predicted ‘phage’ contigs in the data set Gp_GH based on shared gene families.

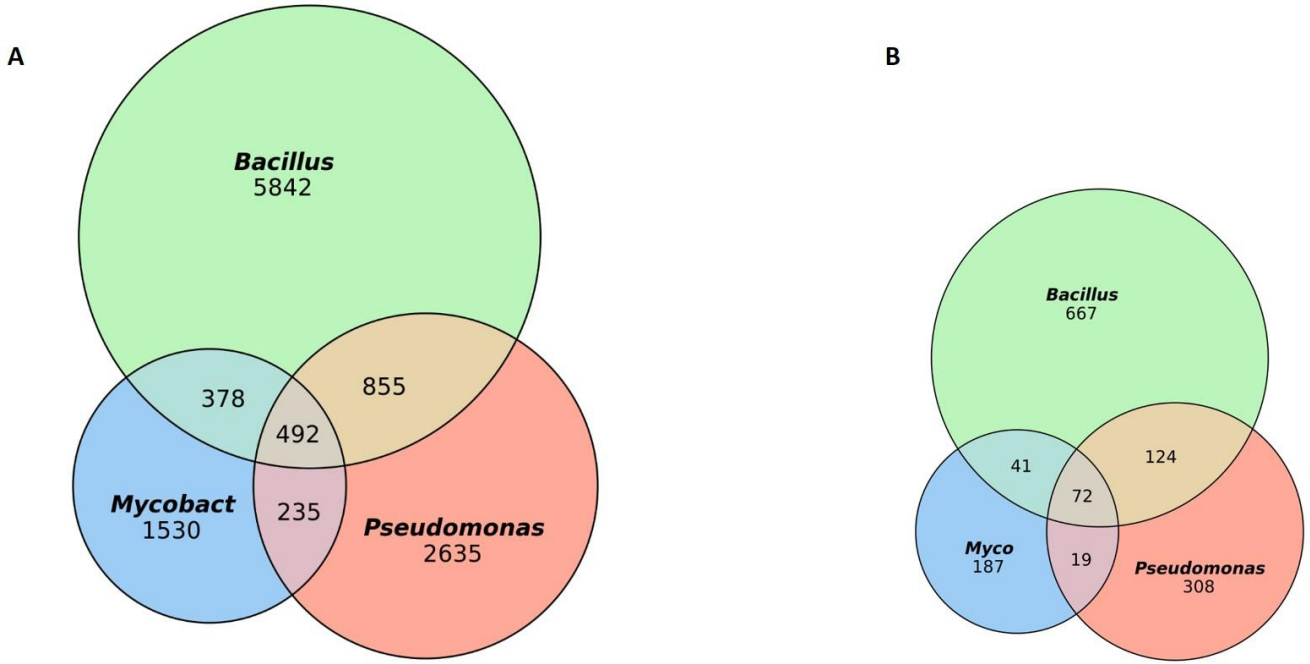
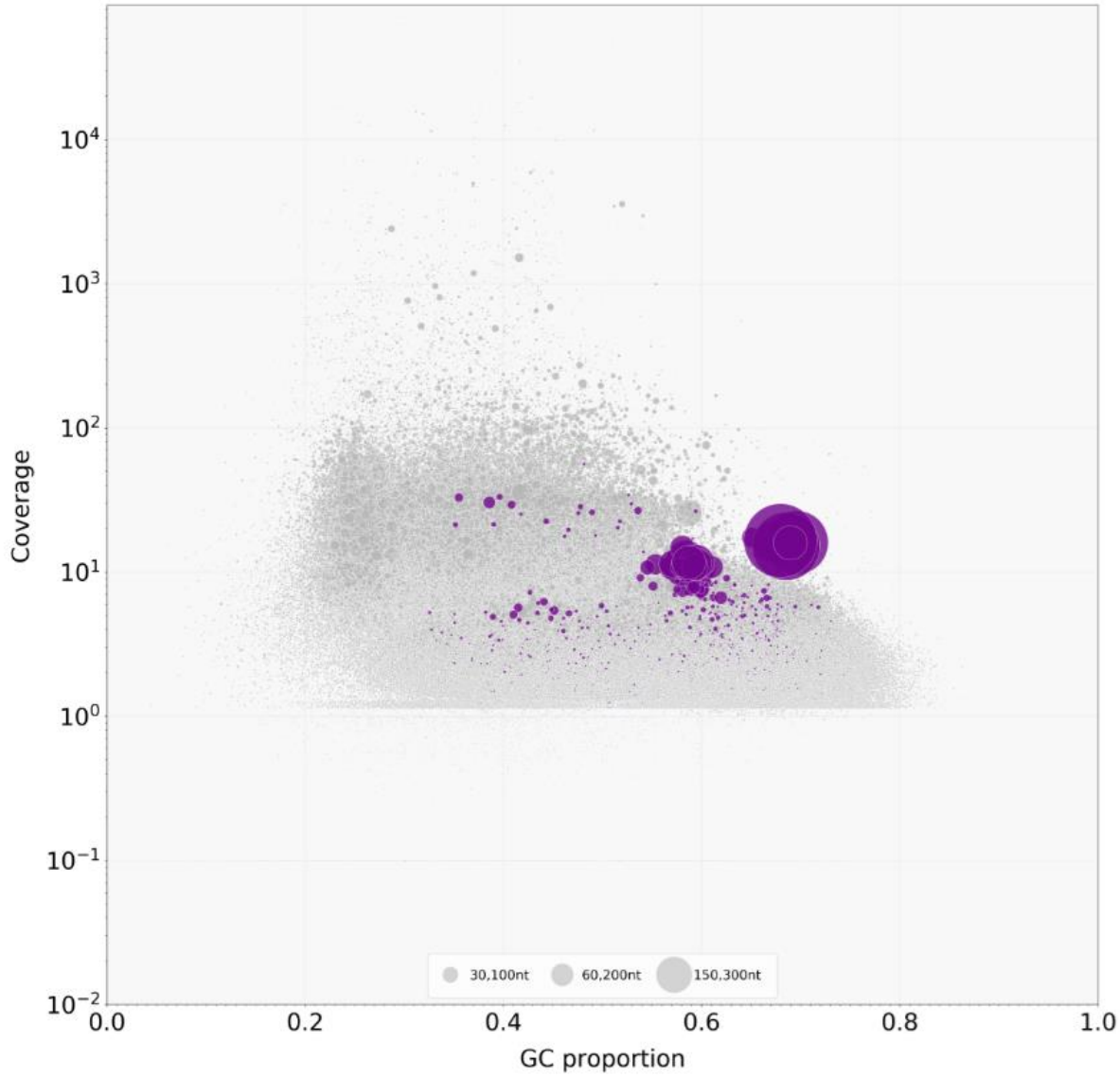


Figure 3.4. The number of contigs that matched to three phage genome databases in the dataset Gp_GH (A) and Hg_Al (B). The proportion of BLAST hits was conserved between the two data sets.

A


Coverage
10^4
10^3
10^2
10^1
10^0
10^{-1}
10^{-2}
0.0
0.2
0.4
0.6
0.8
1.0
GC proportion
30,100nt
60,200nt
150,300nt

B.

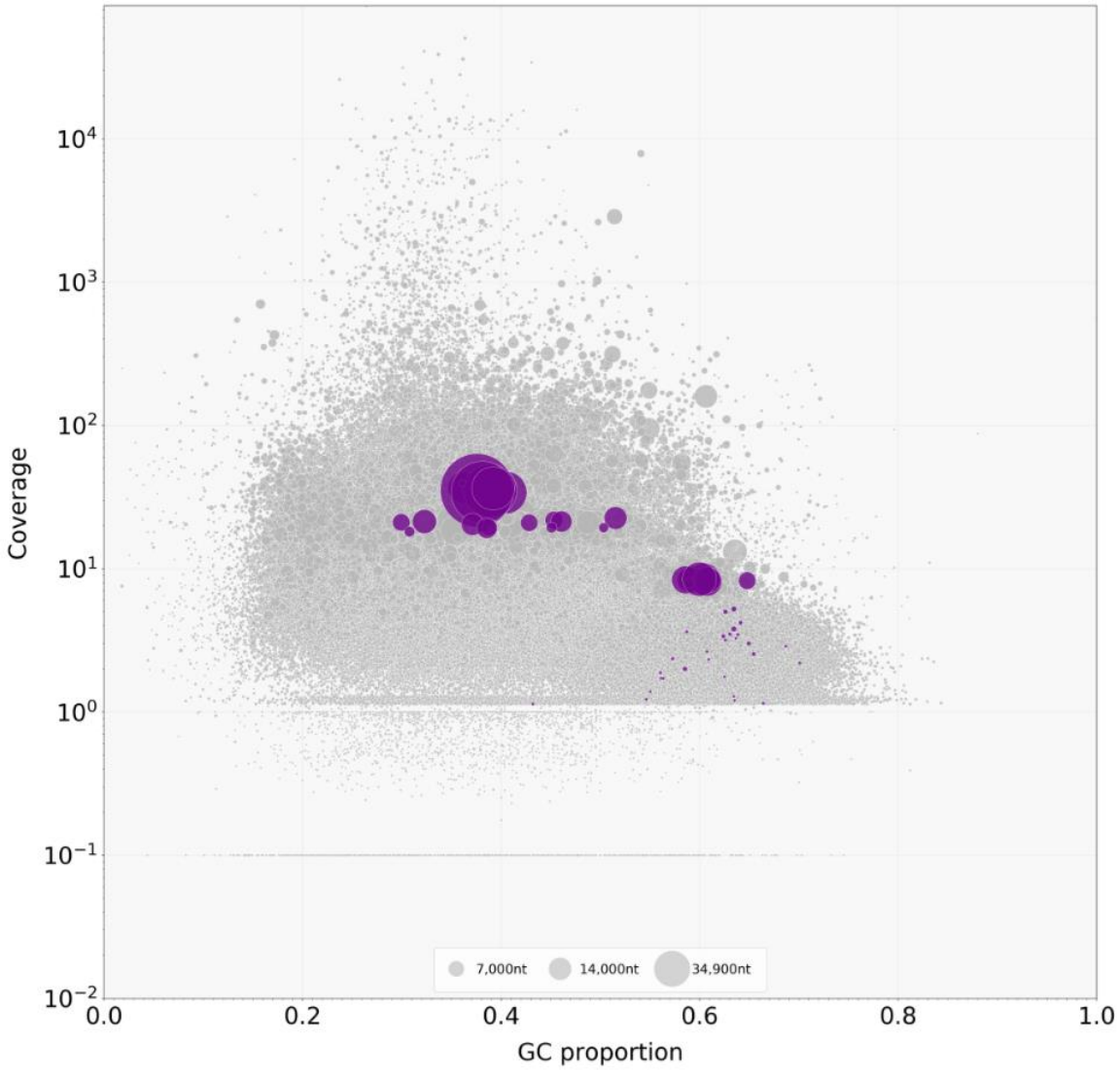
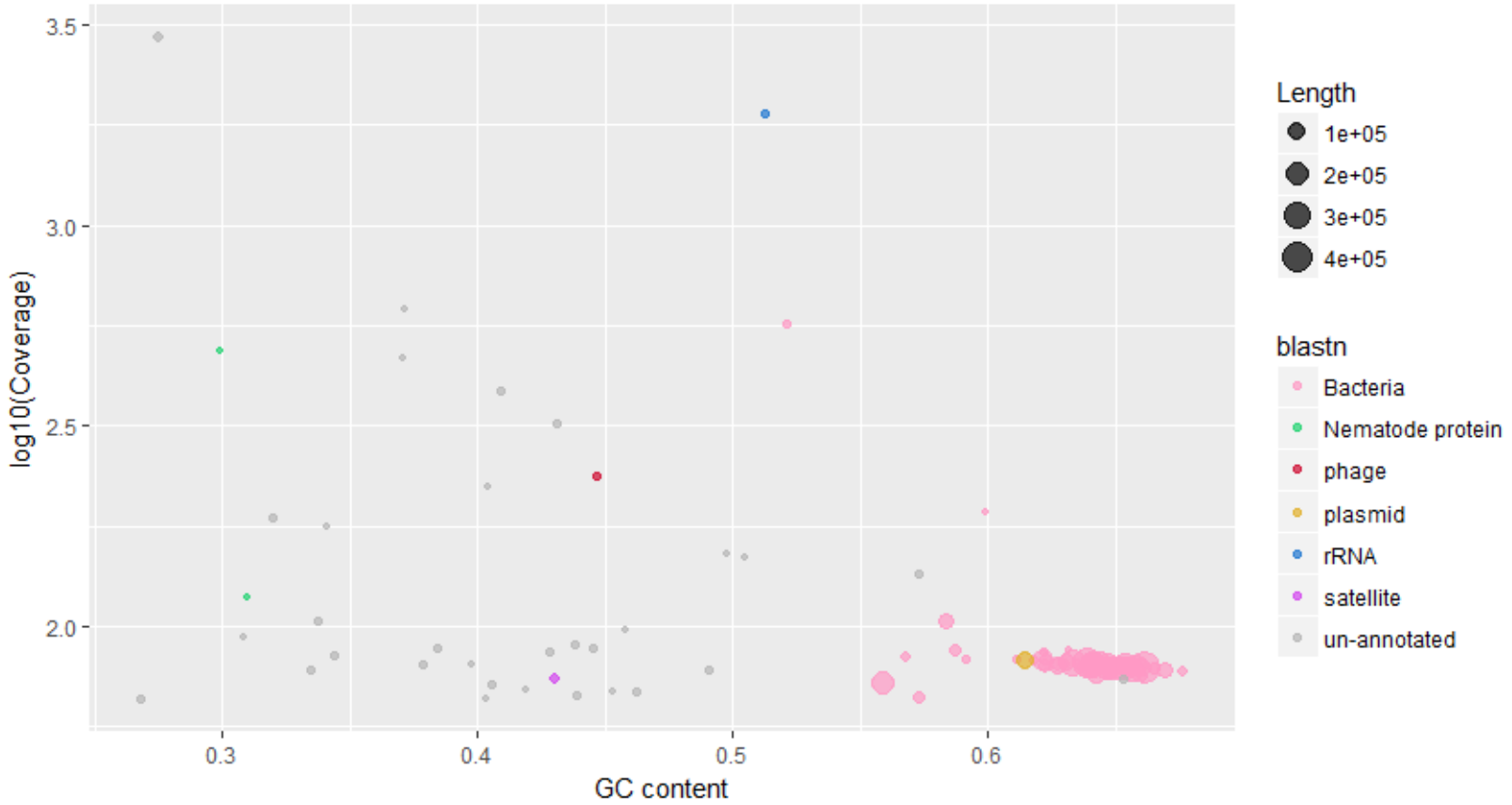
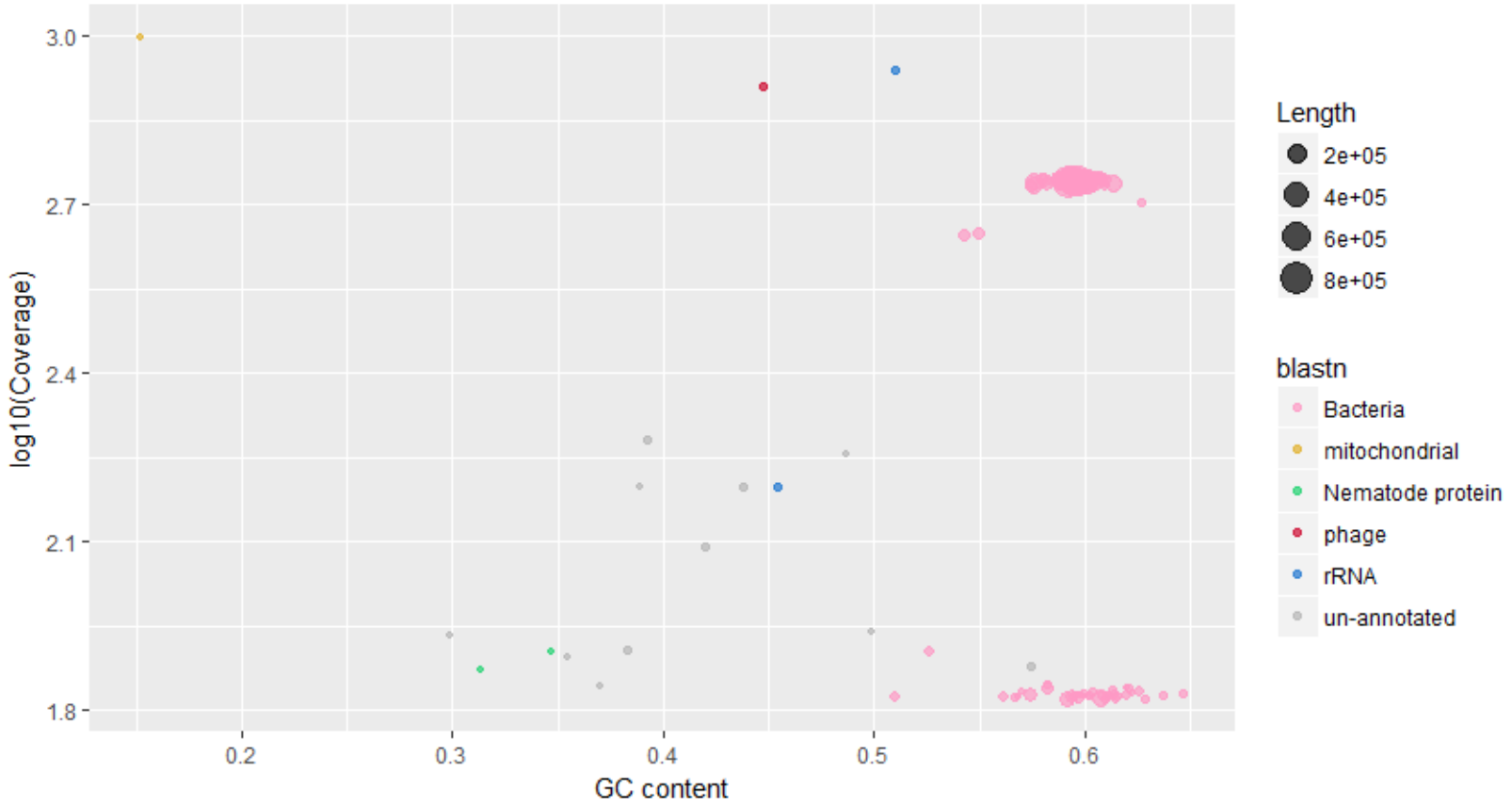


Figure 3.5. GC content, size and sequencing coverage of the contig that hit to all the three phage databases in Gp_GH (A) and Hg_Al (B). The x axis is GC content, and the y axis is sequence coverage.

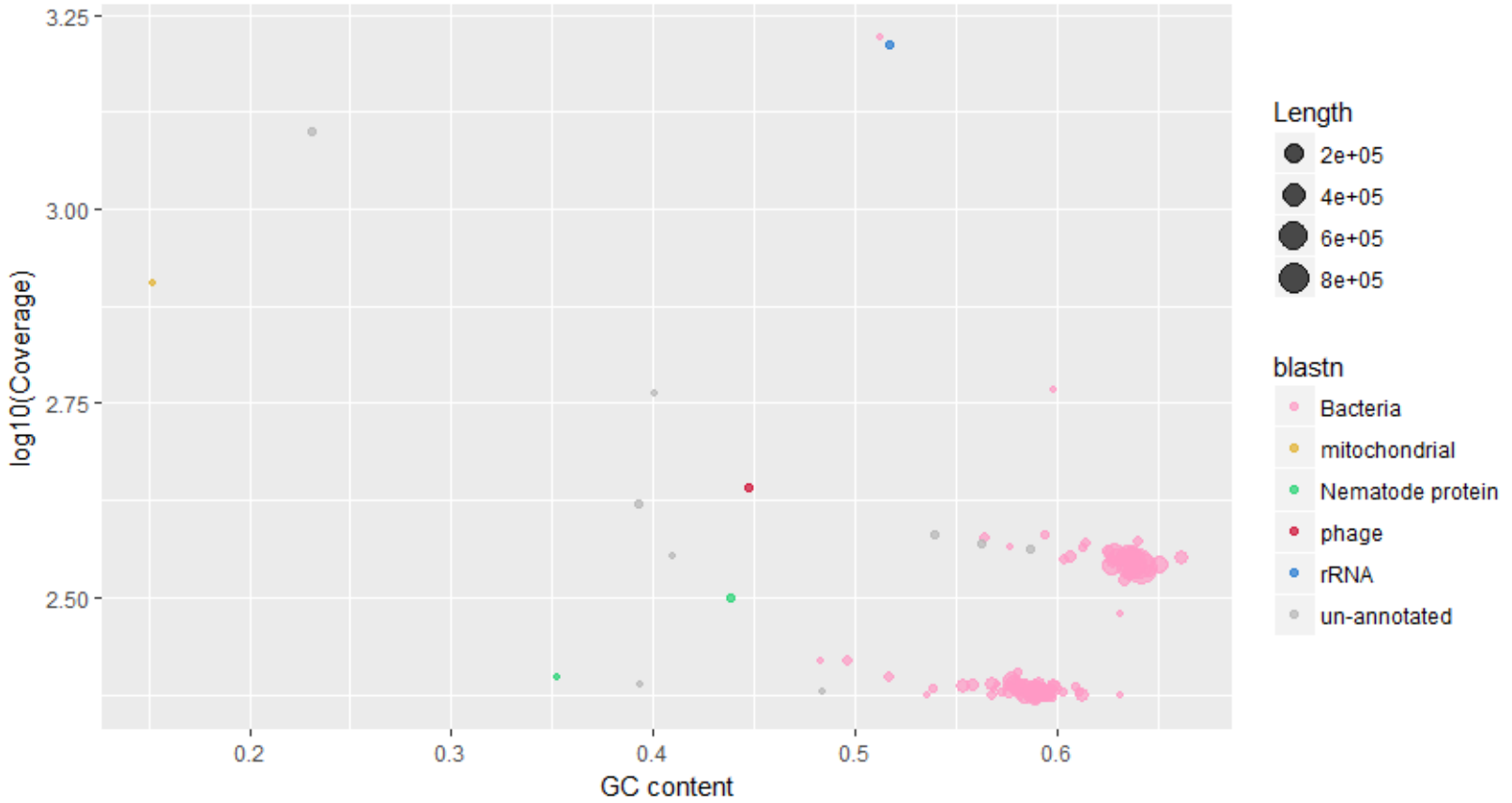
A


3.5
3.0
2.5
2.0
log10(Coverage)
0.3
0.4
0.5
0.6
GC content
Length
1e+05
2e+05
3e+05
4e+05
blastn
Bacteria
Nematode protein
phage
plasmid
rRNA
satellite
un-annotated

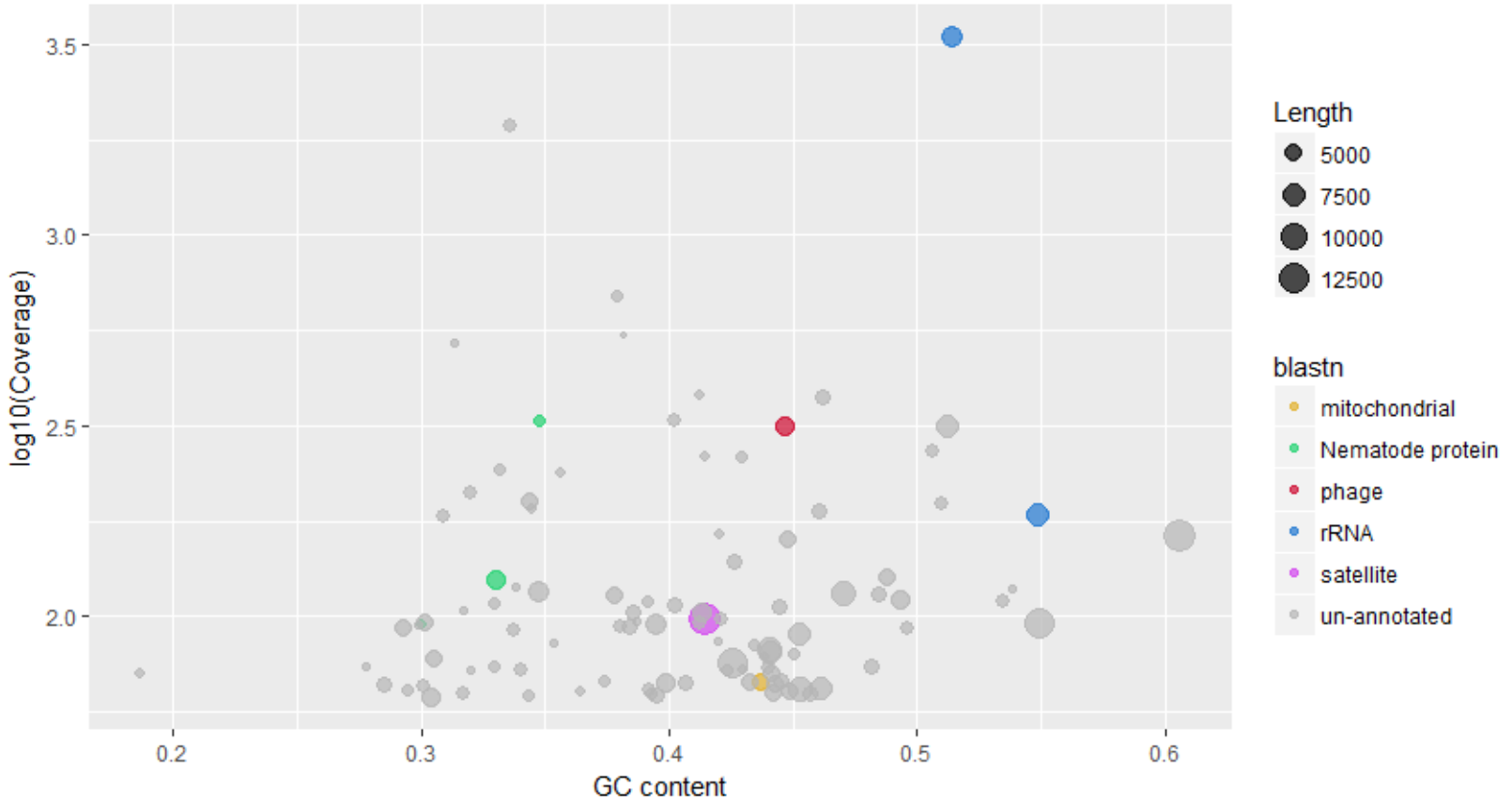
B


3.0
2.7
2.4
2.1
1.8
log10(Coverage)
0.2
0.3
0.4
0.5
0.6
GC content
Length
2e+05
4e+05
6e+05
8e+05
blastn
Bacteria
mitochondrial
Nematode protein
phage
rRNA
un-annotated

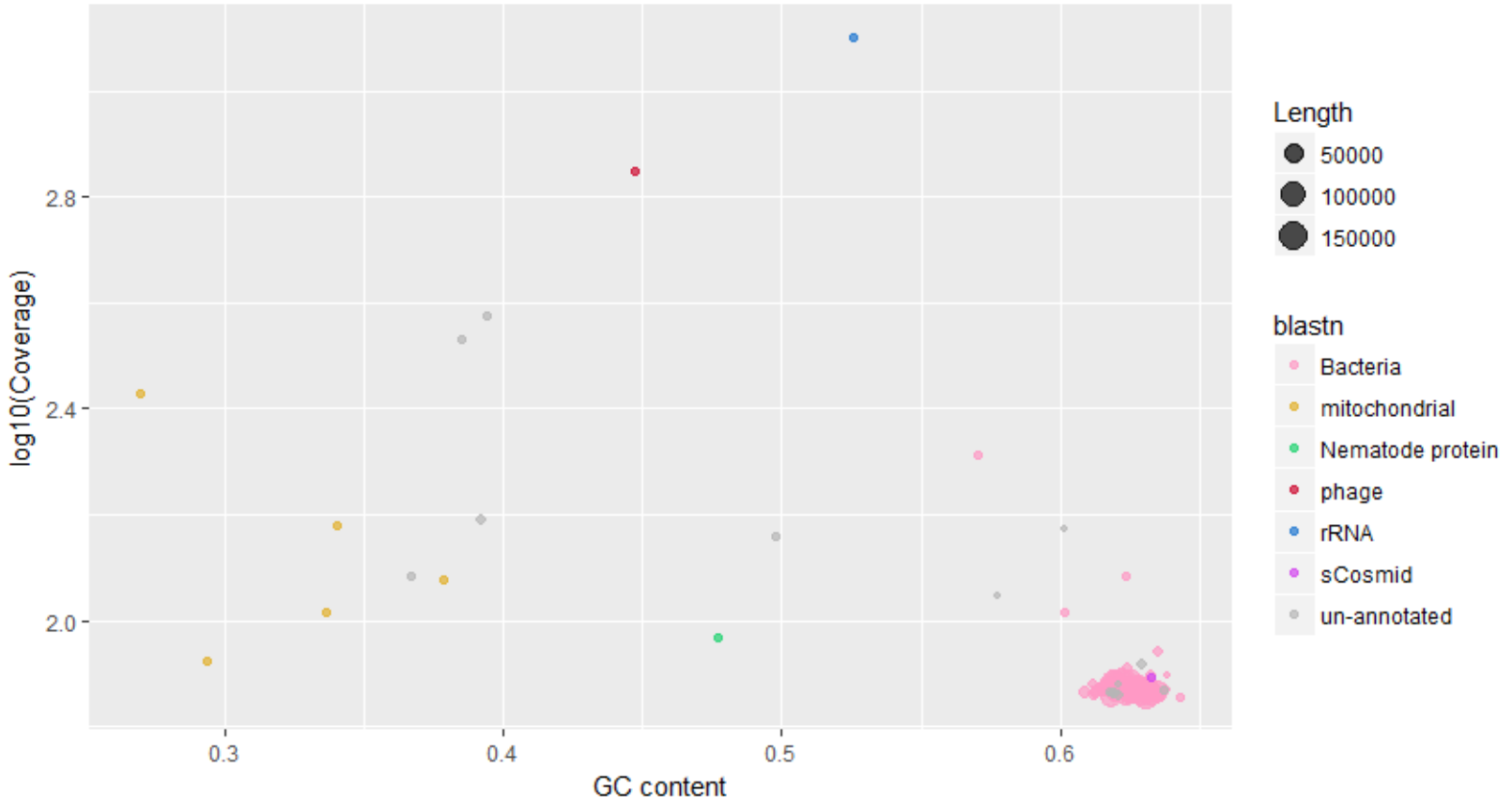
C


3.25
3.00
2.75
2.50
log10(Coverage)
0.2
0.3
0.4
0.5
0.6
GC content
Length
2e+05
4e+05
6e+05
8e+05
blastn
Bacteria
mitochondrial
Nematode protein
phage
rRNA
un-annotated

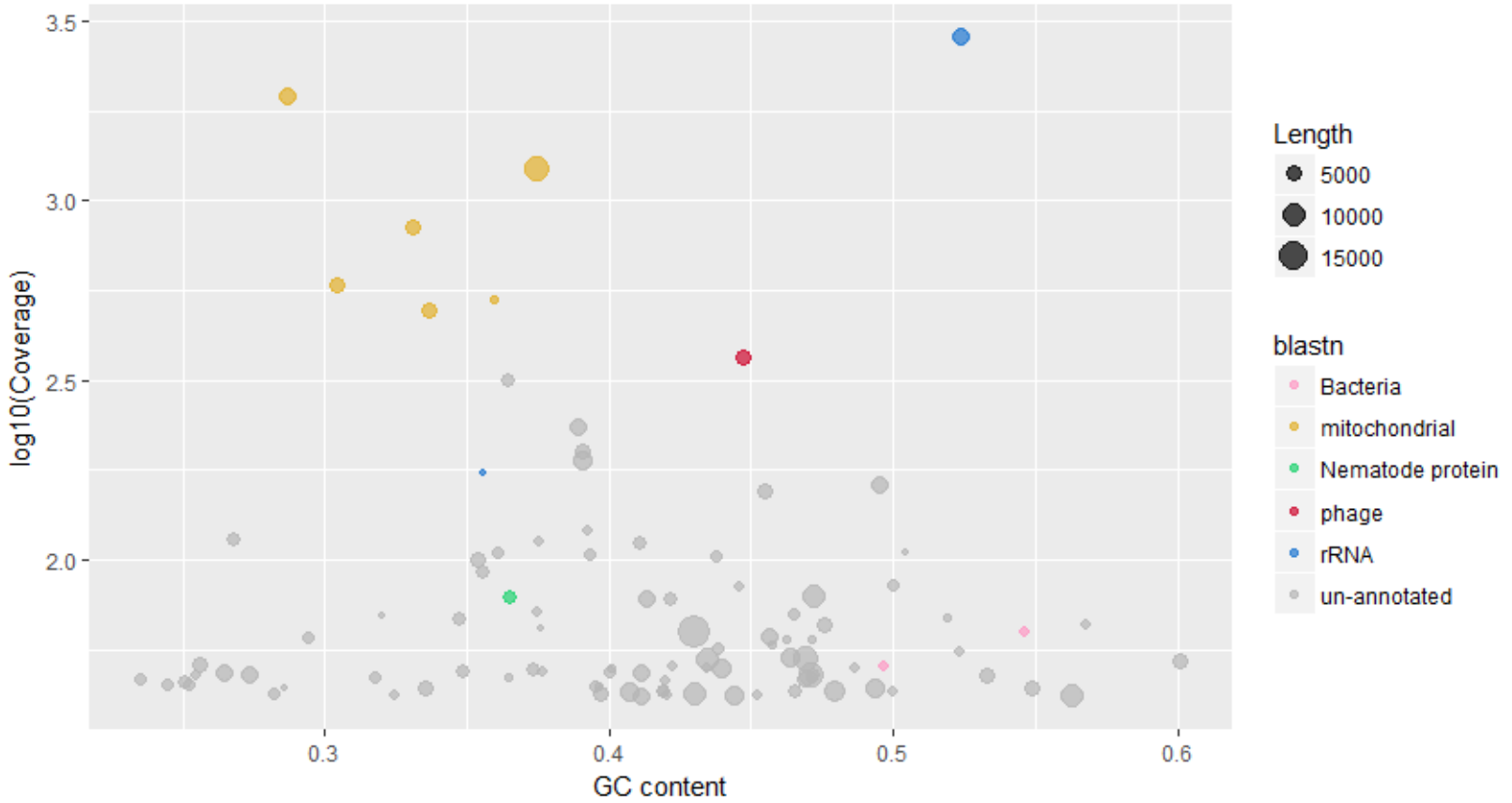
D


3.5
3.0
2.5
2.0
log10(Coverage)
0.2
0.3
0.4
0.5
0.6
GC content
Length
5000
7500
10000
12500
blastn
mitochondrial
Nematode protein
phage
rRNA
satellite
un-annotated

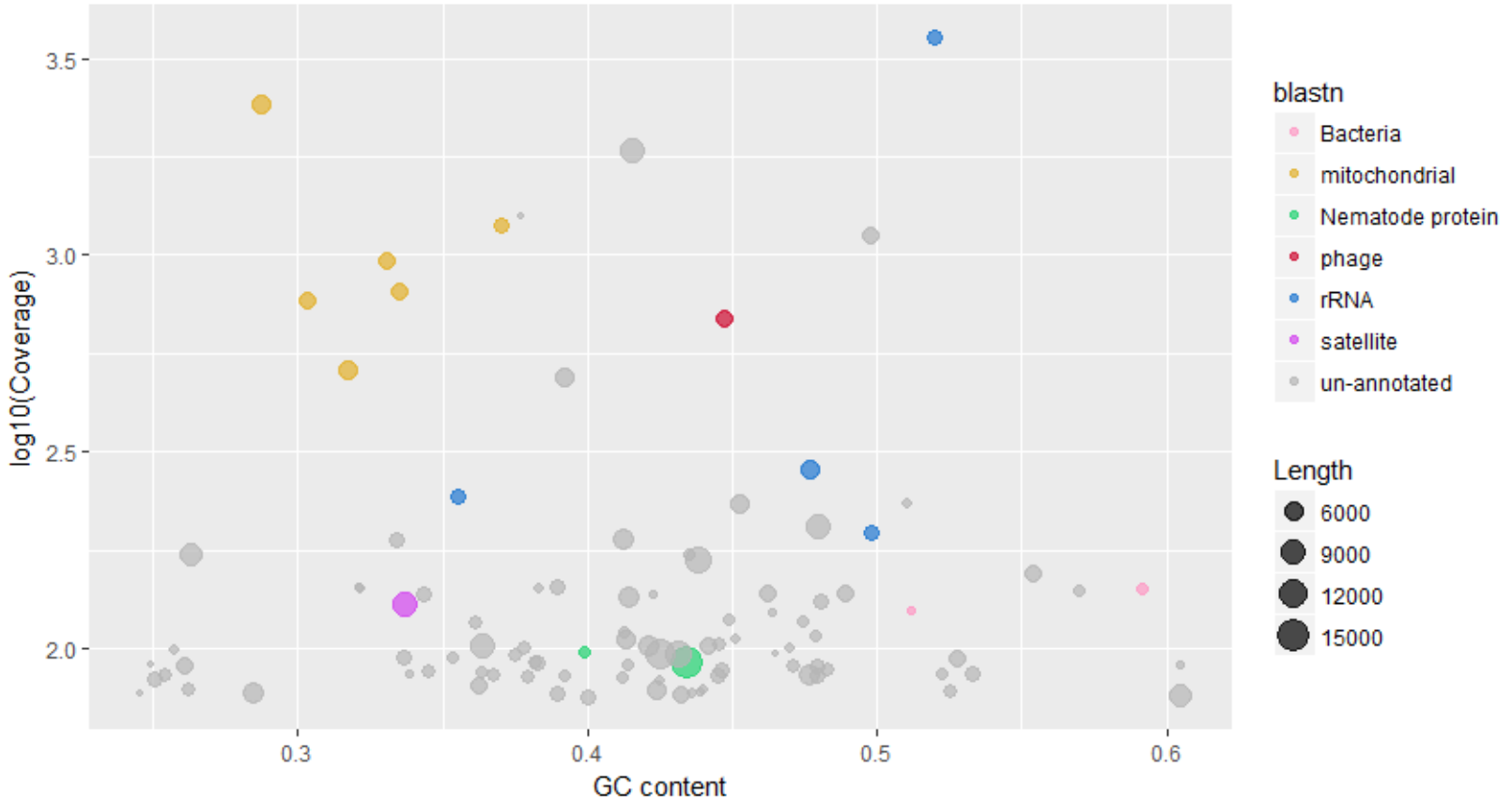
E


log10(Coverage)
2.8
2.4
2.0
0.3
0.4
0.5
0.6
GC content
Length
50000
100000
150000
blastn
Bacteria
mitochondrial
Nematode protein
phage
rRNA
sCosmid
un-annotated

F


3.5
3.0
2.5
2.0
log10(Coverage)
0.3
0.4
0.5
0.6
GC content
Length
5000
10000
15000
blastn
Bacteria
mitochondrial
Nematode protein
phage
rRNA
un-annotated

G


3.5
3.0
2.5
2.0
log10(Coverage)
0.3
0.4
0.5
0.6
GC content
blastn
Bacteria
mitochondrial
Nematode protein
phage
rRNA
satellite
un-annotated
Length
6000
9000
12000
15000

H

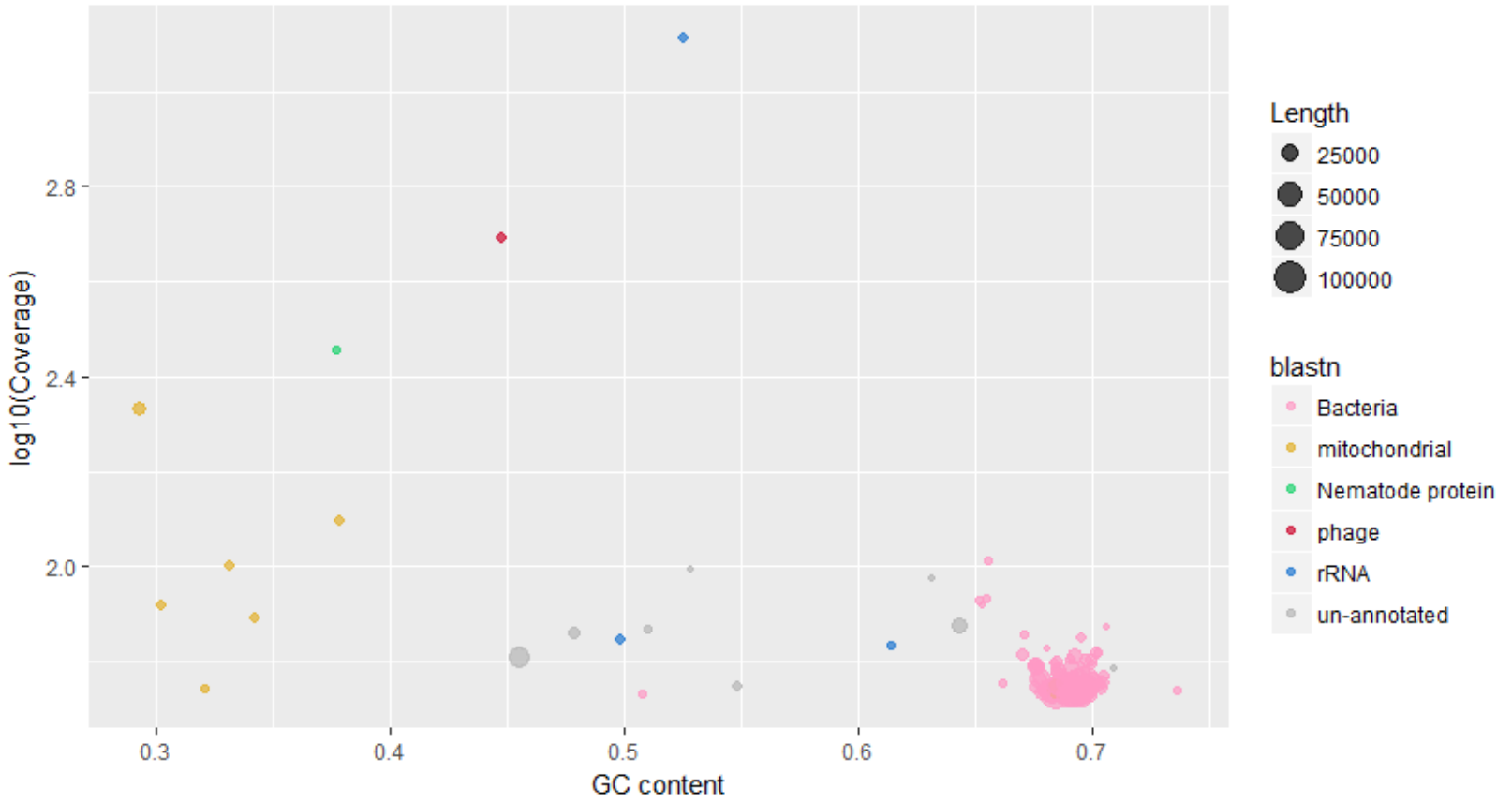
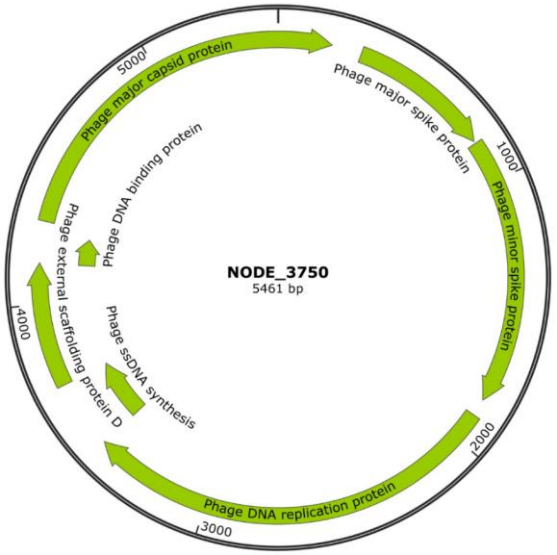


Figure 3.6. Close-up view of 100 contigs with the highest coverage in datasets Hg_War (A), Hg_Pet (B), Hg_Aud (C), Hg_Al (D), Gp_Bin25 (E), Gp_Bin258 (F), Gp_GH (G), and Gp_Bin 26 (H). Different colors denote the biological identity of the contigs based on megablast results. The 'phage' contig is shown in red.

A



B

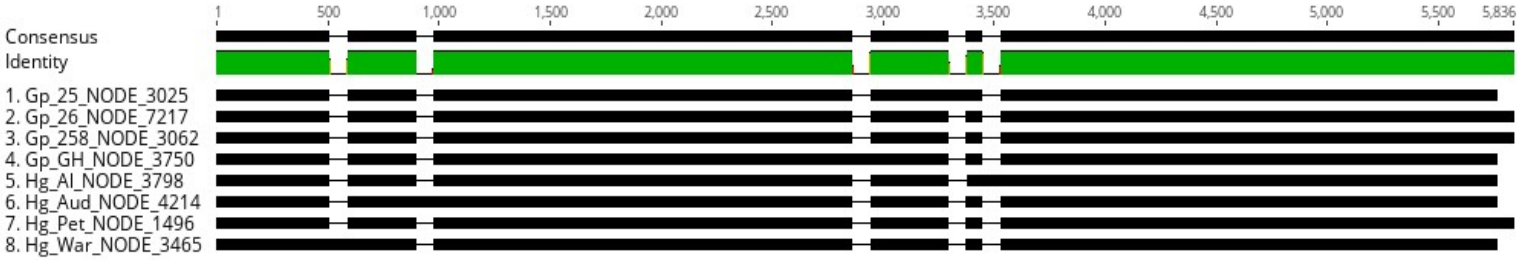


Figure 3.7. (A) Gene annotation of the contig with length 5461 bp and high coverage (contig NODE_3750) in the dataset Gp_GH. (B) Sequence alignment of all contigs with length 5461 bp and high coverage in the eight datasets analyzed.

Chapter 4: Microbiome dynamics associated with *Phasmarhabditis hermaphrodita* before and after infection of grey field slugs

Anh D. Ha, Dana K. Howe, Andrew Colton, Rory Mc Donnell, and Dee R. Denver

In preparation for:
PLoS One
Public Library of Science
San Francisco, California

Abstract

The slug-parasitic nematode *Phasmarhabditis hermaphrodita* is currently recognized as a promising candidate in biological control of slug pests. *P. hermaphrodita* closely resemble the insect control agent *Steinernema* spp.; however, unlike *Steinernema* nematodes, which form an obligate symbiosis with the specific bacteria *Xenorhabdus* spp., *P. hermaphrodita* have no specific bacterial associates reported. Nevertheless, the presence of different bacterial partners during slug infection was shown to affect the pathogenicity and efficiency of the nematodes. Therefore, characterization of *P. hermaphrodita*'s microbial community and determining the bacteria that could potentially improve the nematode's performance may assist the optimization of the sustainable biocontrol method using this nematode species. In this chapter, we examined three *P. hermaphrodita* populations cultured in three different microbial conditions: plates seeded with *Escherichia coli*, a newly discovered *Pseudomonas sp.* that co-cultured with *P. hermaphrodita*, and the original complex bacterial community associated with the nematode. For each of the three treatments, we evaluated the pathogenicity of *P. hermaphrodita*, explored the composition of their microbial communities, and investigated the changes in microbiome structure before and after slug infection using 16S rRNA amplicon sequencing. We identified a wide variety of taxa components (82 genera) in the community associated with the nematode pre-infection, most of which are of low abundance. In all bacterial treatments post-infection, the number of genera almost quadrupled and those taxa's abundance changes remarkably, although the composition of highest abundance taxa remained stable. Alpha diversity analyses showed that the community richness alter between collection time points, but the community diversity did

not significantly differ among types of bacterial treatment post-infection. Beta diversity results indicated clear clustering of the six sample sets. Finally, we observed four *Pseudomonas* ASVs that conserved and increased abundance after slug infection in the *Pseudomonas* treatment, which may suggest a role of the taxa in the infection process.

Introduction

Invasive pest slug species, such as *Deroceras reticulatum* (‘grey field slug’), are among the most widespread, damaging pests of agricultural and horticultural production, causing reductions in quality and yield loss on a broad range of crops including wheat, corn, legumes, vegetables, and fruits (Barker, 2002; Godan, 1983; Mc Donnell et al., 2009). Slug pests have also been documented to vector several plant and human pathogens e.g. *Alternaria brassicicola* (causing black spot disease in brassica), *Escherichia coli* (causing food poisoning) (Mc Donnell et al., 2009) and *Angiostrongylus cantonensis*, the causal agent of the potentially lethal eosinophilic meningitis (Centers for Disease Control and Prevention, 2019). Conventional slug control methods are largely chemical-based, using active molluscicide ingredients such as metaldehyde, iron phosphate, and carbamates (Barker, 2002). However, there is growing environmental concern about these chemical molluscicides harming non-target organisms (Aktar et al., 2009).There is also evidence that target slugs might be developing resistance to these molluscicide compounds (Barker, 2002). Therefore, current research efforts focused on combating these damaging slug pests have focused on alternative control strategies, such as biological control approaches.

Phasmarhabditis hermaphrodita is a parasitic, rhabditid nematode species that has long been known to be a natural enemy of slugs and snails. The potential of *P.*

hermaphrodita for biocontrol against slugs was realized as early as 1988, when it was confirmed to have high virulence against a wide range of pest gastropods (Rae et al., 2007; Wilson et al., 1993). The nematode has been sold commercially under the trade name Nemaslug® in Europe for over 20 years but is not commercially available in the United States due to regulatory concerns about the unknown effects of this nematode on North America-native gastropod species, such as banana slugs. Recently, the discovery of *P. hermaphrodita* in California (Lopez et al., 2014) and then later in Oregon by our research team (Howe et al. - in preparation) has provided justification for further exploring this nematode as a potential slug biocontrol agent in the United States.

Regarding the mechanism of infection, *P. hermaphrodita* shares notable similarities with the well-studied entomopathogenic nematodes (EPNs) of the genus *Steinernema*, which has been commonly used as biological control agents of insect herbivores (Smart, 1995). In both systems, the infective juvenile-stage nematodes live freely in soil and gain entrance to the targeted hosts upon exposure, usually through the mantle cavity. Once inside, infective juveniles develop into self-fertilizing hermaphroditic adults, reproduce, and the infection spreads to the entire body of the slug host, eventually leading to its death within 4-21 days. The nematodes continue to consume the slug carcass until the food source is depleted, and their new infective juveniles again move into soil to search for new susceptible hosts (Rae et al., 2007; Tan and Grewal, 2001a).

Despite sharing many similarities in the life cycle with the EPNs of the genus *Steinernema*, which engage in an obligate, specific mutualism with *Xenorhabdus* spp. bacteria (Sicard et al., 2004) required for insect host infection, *P. hermaphrodita* has

more uncertain and understudied relationship with its bacterial associates. In fact, the scientific literature focusing on the role of bacteria in the lifestyle, virulence and mollusc-killing efficacy of *P. hermaphrodita* remains quite contradicting. The Nemaslug® product consists of both *P. hermaphrodita* and the bacterium *Moraxella osloensis*, and it is often assumed that both biotic components of this product play roles in slug infection and killing. However, *P. hermaphrodita* can feed on and carry out slug infection with many other bacterial partners that include *Pseudomonas fluorescens*, and *Pseudomonas paucimobilis* (Wilson et al., 1995a, 1995b). Little is known about the nature of *P. hermaphrodita*-bacteria association and how the two partners interact, if at all, to achieve pathogenicity to gastropods.

Different bacteria have been shown to have different effects on the virulence of the nematode-bacteria combination in the *P. hermaphrodita* system (Wilson et al., 1994, 1995a). Some studies suggested that the sole agent responsible for pathogenicity was a number of particular bacterial species, most notably *M. osloensis* (Tan and Grewal, 2001b). Meanwhile, other studies suggested that *P. hermapherodita* lacks a specific obligate partner in killing the slugs, and bacteria did not have discernible influence on virulence (Rae et al., 2010). The association between *P. hermaphrodita* and bacteria is then expected to be complex, variable, and the role of each partner in slug infection is yet to be well understood, preventing the optimization of the biocontrol method using the nematode. This gap in knowledge requires further study and is necessary for the development of the prospective biocontrol strategy using the newly found *P. hermaphrodita* in the United States.

To better understand the composition and function of the natural bacteria harbored in *P. hermaphrodita*, and whether there are alterations in the bacterial community before and after slug infection, we conducted a set of slug infection trials involving nematodes reared on different bacterial communities in an effort to:

1) Evaluate the pathogenicity of *P. hermaphrodita* nematodes reared on *Escherichia coli* strain OP50, a known non-pathogenic bacterium that is the common food source for laboratory populations of the infamous model organism *Caenorhabditis elegans*.
2) Investigate the pathogenicity of *P. hermaphrodita* reared on a *Pseudomonas sp.* bacterium, which was isolated from the complex bacterial community associated with nematodes infecting grey field slugs in Oregon.
3) Evaluate the pathogenicity of *P. hermaphrodita* reared on the total complex bacterial community that co-cultured with nematodes infecting gray field slugs in Oregon.
4) For each of the three above bacterial communities, investigate how bacterial communities associated with the nematodes changed before and after slug infection experiments in the lab.

We performed a 16S rRNA-based analysis of the *P. hermaphrodita* microbiota sampled immediately before and after infection of the grey field slug *Deroceras reticulatum*. Previous research has suggested that *P. hermaphrodita* is capable of carrying out slug infection with many different bacterial partners, and that the bacteria might play a legible role in the infection process. Nevertheless, we hypothesized that microbiomes associated with P. hermaphrodita will impact the nematode's ability to infect slugs, either

through direct effects (e.g., direct involvement of bacteria in slug mortality) or through indirect means (e.g., through variable nutritional effects on nematode metabolism). We predicted changes to occur in nematode-associated microbiomes, when comparing pre-infection to post-infection nematode samples. The infection of a slug host might act as a selection pressure on the bacterial community, favoring the bacteria that can better contribute to the nematode's virulence, spread, reproduction and killing efficacy. This research provides further insights essential for the development and optimization of this nematode-based slug biocontrol strategy.

Materials and Methods

Nematode Preparation

Prior to the slug infection assay, *P. hermaphrodita* nematodes were bleached with 5% solution of sodium hypochlorite following standard nematode bleaching protocol (Sulston and Hodgkin, 1988) to remove/minimize their associated bacteria. After bleaching, nematodes were cultured on nutrient growth media (NGM) plates (Stiernagle, 2006) seeded with one of three designated bacterial food sources: *E. coli* OP50 (EC), *Pseudomonas* sp (PS)., and the complex total bacterial community (BC) associated with the nematodes. The E. coli OP50 culture was obtained from the *Caenorhabditis* Genetics Center at the University of Minnesota. The *Pseudomonas* sp. culture was previously isolated from a single colony deriving from the complex bacterial community that co-cultured with a *P. hermaphrodita* strain discovered in Oregon (isolated from a slug found on the Oregon State University Campus (McDonnell et al., 2018).

Nematode culturing and transfers were performed using 'semi-sterile' conditions that are typical in many laboratory-based nematode research programs. NGM plates were

initially poured using autoclave-sterilized media, equipment, and sterile plastic plates (VWR International, Lutterworth, UK). Bacteria cultures were grown using autoclave-sterilized flasks. NGM plates were seeded with the appropriate bacterial cultures on lab bench tops, exposing the agar media and bacterial cultures to potential environmental bacteria found in the lab environment. Nematodes were added to plates in a similar fashion, both upon initial establishment of cultures and subsequent sub-culturing. Thus, we expected to detect the presence of additional non-target bacteria (e.g., bacteria other than *E. coli* OP50 in the EC samples) in our subsequent analyses.

Nematode Sample Collection

Collecting samples at the first time point - before infection

This dissertation chapter relied on the collection of nematode samples that were part of a larger slug infectivity assay, the broader results of which are separate from the microbiome-centric focus on this chapter. The results of the broader slug assay, led by Dr. Rory Mc Donnell and his research team in the OSU Department of Crop and Soil Science, were summarized in supplementary document 1 to provide necessary context for this dissertation research chapter. We plan to ultimately combine the broader slug infectivity assay results (supplementary document 1) and the results of this microbiome-focused research chapter for a larger, collaborative peer-reviewed publication (we plan to submit this article to *PLoS Pathogens*).

We collected nematode samples in a fixed volume of 200 µL nematode suspension in sterile M9 solution, using semi-sterile lab benchtop techniques as previously described. For each of the three bacterial conditions examined in this study (EC, PS, BC), three biological replicates were collected and analyzed. Immediately

before the infection assay, three replicates of nematode suspensions (8,000 nematodes/ml) were taken out from each of the three combinations, totaling nine 'before' samples. The worms were washed once with 200 µL M9 buffer, then twice with 200 µL distilled H_2O. We also included six negative controls to assess the contamination introduced during the preparation procedure and sequencing: three samples of the M9 buffer, three samples of distilled H_2O from the Denver Lab, three samples of distilled H_2O from the McDonnell Lab where the slug infection assay was carried out, and three DNA Extraction kit blanks. The nine pre-infection nematode samples (hereafter referred to as "PreInf" sample) and twelve negative controls were stored immediately after collecting at -80 °C.

Collecting samples at the second time point - after infection

Nematode and bacteria from the high-dose nematode treatment (see Supplementary Document 1 for description of high- and low-dose treatment) were collected in three biological replicates. For each of the three bacteria treatment, we collected three slug carcasses (each from a separate container) to extract nematode populations. Immediately following the death of a slug, the fresh carcass was submerged in 200 µL of sterile M9 solution to elute nematodes. Collected nematodes were washed once with 200 µL M9 buffer and then twice with 200 µL of molecular biology grade water (VWR International, Lutterworth, UK). The nematode samples (hereafter referred to as "PostInf" samples) were then stored at -80 °C until DNA extraction. The simplified scheme for referring to the six different bacterial treatment types, used hereinafter in this manuscript, is summarized in Table 1.

DNA Isolation, 16S rRNA Amplification and Sequencing

Prior to DNA isolation, nematode samples were homogenized thoroughly by bead beating. We conducted DNA extraction from the homogenate, and from the negative control samples, using the PowerSoil DNA Extraction kit (MOBIO, Carlsbad, CA, USA). DNA extractions were then quantified using a fluorescent plate reader at the OSU Center for Genome Research and Biocomputing (CGRB).

The process of 16S library preparation, amplification and sequencing were conducted by the OSU CGRB. Amplicon libraries were prepared following the standard protocol by Illumina (16S Metagenomic Sequencing Library Preparation, 2013). Primers targeting the V3-V4 region (forward: 5'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGC WGCAG3'; reverse: 5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGT ATCTAATCC 3') were utilized for PCR amplification (Klindworth et al. 2013). Subsequently, the PCR products were proceeded through clean-up, index PCR, library quantification, normalization, and pooling. Libraries were sequenced using paired-end MiSeq v2 Nano 300 bp platform (Illumina, USA).

DNA Sequence Data Processing

DNA sequence quality control procedures began with read trimming: raw forward reads were trimmed off 17 nucleotides on the left and truncated to a length of 283 bases while reverse reads were trimmed 21 nucleotides on the right and truncated to a length of 266 bases using the DADA2 pipeline (Callahan et al., 2016). The trimmed nucleotides included the 16S amplification primers. DADA2 was also used to merge paired reads,

denoise, quality filter, and infer amplicon sequence variants (ASVs). ASVs represent exact 16S rRNA gene sequence variants resolved to the level of single-nucleotide dissimilarity; ASVs constituted the fundamental unit of taxon identity used in subsequent analyses. Compared to the commonly used operational taxonomic units (OTUs) with a threshold of nucleotide difference of 97%, ASVs have been shown to offer better reproducibility, reusability, and comprehensiveness in 16S microbiome analyses (Callahan et al., 2017).

Phylogenetic analyses and representations of the ASVs were constructed using MAFFT alignment (Katoh, 2002) and FastTree (Price et al., 2010). To assign taxonomy to the ASVs, we utilized a self-trained classifier, trained on the Silva Project's database (release 132) (Quast et al., 2013). Putative contaminant sequences were identified and filtered out of sample data using the R package decontam (Davis et al., 2018). ASVs were considered a contaminant if they were more prevalent in negative controls than in positive samples and/or their frequency significantly varied inversely with sample DNA concentration (probability threshold $p < 0.1$).

We utilized the Phylum, Genus, and ASV levels as units for subsequent compositional taxa analyses. Since different taxonomic classifiers trained on different databases may differ at lower taxonomic ranks (Balvočiūtė and Huson, 2017), the Phylum level were chosen as the unit for taxa analysis; however, the Genus and ASV level were also selected for a finer resolution and subsequent composition analysis.

Alpha Diversity Analyses

We used four metrics: Chao1, Shannon, Inverse Simpson and Faith's Phylogenetic Diversity (PD) to compare and evaluate alpha diversity from different approaches. Chao1

index is an estimator of species richness i.e. the expected total number of OTUs/ASVs in the sample given all the species were identified. It weighs the low abundance taxa (i.e. only singletons and doubletons) to infer the number of missing species (Chao, 1984). Shannon Index estimates the overall richness and also the evenness/uniformity between the taxa present in the sample (Spellerberg and Fedor, 2003). Also focusing on the species richness and evenness, however, Inverse Simpson emphasizes more on species evenness (Simpson, 1949) while Shannon index gives a greater weight to the species richness. Faith's PD metric considers the phylogeny of taxa to estimate diversity, and is proportional to how much of the OTU/ASV phylogenetic tree is covered by the taxa present (Faith, 1992).

To calculate Shannon Index and Inverse Simpson, the ASV table were rarefied ten times to a depth of 100,000 reads in order to lessen the random biases introduced by randomly subsampling. Rarefying the ASV table to a fixed number of reads per sample is necessary for calculations of Shannon and Inverse Simpson, as these two abundance-based indices could be substantially affected by the differences in the total number of reads between samples. The two indices were calculated for each of the ten rarefied table using the package phyloseq, and the mean values were used for alpha diversity analysis. As for Chao1, which depends on low-abundance taxa, the ASV table was not rarefied to retain all 'rare' taxa. The index Faith's PD was calculated using the package picante. Kruskal-Wallis tests were conducted to assess the significance of the differences among the bacteria-nematode combinations at the same time point, and Mann-Whitney U tests or paired t-test were applied, as appropriate, for the difference between the two time points before and after infection. The normality of diversity values was assessed using Shapiro-

Wilk normality test; when the p-value of Shapiro test is larger than 0.05, it is implied that the distribution could be assumed normal.

Beta Diversity Analyses

To assess the difference in microbial community between samples, we used four beta diversity metrics: two abundance-weighted beta diversity indices (Bray-Curtis and weighted UniFrac) and two presence-absence indices (Jaccard and unweighted UniFrac). While Bray-Curtis and Jaccard are non-phylogenetic metrics, weighted and unweighted UniFrac also take into account the phylogeny of the bacterial community composition to assess differences between samples. Specifically, UniFrac distances are based on the sum of branch length shared between samples on the phylogenetic tree constructed from all the 16S rRNA sequences of from all communities of interest. To test whether the groups differ significantly from each other, we conducted the permutational multivariate analysis of variance (PERMANOVA) test using the package vegan (Oksanen et al., 2019).

The beta diversity distance matrices obtained were visualized using the ordination technique Principal Coordinates Analysis (PCoA), which reduced the number of dimensions of the distance matrices and visualized the sample dissimilarities in a low-dimensional space. In the result plots, each sample was a point, and the clustering/distance between these points might indicate the similarity/dissimilarity of the samples.

Results

16S rRNA Sequence Data

After quality filtering and removing chimeric sequences, we obtained a total of 10,485,882 reads, with an average of 582,549 reads and a median depth of 349,960 reads per sample. 2,938 ASVs (i.e. taxa) were identified. The decontamination procedure removed 131 ASVs, leaving 2,807 unique ASVs for subsequent analyses. The components of the removed contaminants are shown in Appendix Figure 2. The filtered data set then contained 10,422,195 reads in total, with an average depth of 579,010 reads and a median 343,942 reads per sample. Details of the number of quality-controlled reads obtained in each sample, as well as the total number of reads, median and mean values of reads per nematode-bacterial combination were listed in Appendix Table 4.

Microbiome shifts before and after slug infection experiment

The number of different genera detected in each sample included in the analysis were listed in Table 4.1. On the Phylum level, the majority of bacterial components in all samples were identified as members of the phyla Bacteroidetes and Proteobacteria. We observed a shift in composition towards Proteobacteria in the BC-PostInf and PS-PostInf samples compared to their PreInf counterparts, while in the EC-PostInf samples there was a shift towards Bacteroidetes (Figure 4.1A). In general, the number of genera in all samples increased almost fourfold after infection. The abundance of the main genera also altered remarkably between PreInf and PostInf samples (Figure 4.1B).

In *P. hermaphrodita*'s original microbiome (i.e. the BC-PreInf community), 15 Phyla were identified with Bacteroidetes being the most abundant phylum, followed by

Proteobacteria. On the Genus level, the most abundant groups were *Pseudochrobactrum, Flavobacterium, Raoultella,* and *Pseudomonas* out of the 82 genera observed (Figure 4.1B). Details of relative taxa abundance on the Phylum and Genus level in the BC-PreInf communities were shown in Appendix Table 5.

The numbers of ASVs unique to specific sample sets and shared among sample sets, were illustrated in Figure 4.2. In general, the number of ASVs unique to a single sample set were higher than those shared between different sets. The BC-PostInf samples had the largest number of unique ASVs (i.e. 600) that were not shared with any other type of samples, followed by the EC-PostInf and PS-PostInf (451 and 337 ASVs, respectively). Comparing PreInf and PostInf samples, the BC-PreInf and BC-PostInf communities shared 82 common ASVs, while PS-PreInf and PS-PostInf shared 36 ASVs, and 44 ASVs were shared between EC-PreInf and EC-PostInf samples.

Patterns of change in alpha diversity

We evaluated alpha diversity using the four metrics Chao1, Faith's PD, Inverse Simpson, and Shannon Index (Figure 4.3A). Chao1 index had significantly higher median across all PostInf samples compared to PreInf (Mann-Whitney U test, $p = 4.114e-05$). There was suggestive evidence that the Faith's PD and Shannon indices were lower in PreInf- than PostInf-samples (Mann-Whitney U, $p = 0.06$ and $p = 0.09$, respectively).

Analysis of the EC microbiomes suggested some evidence that the mean Shannon index increased after the slug infection process (Welch two sample t-test, $p = 0.068$). The BC samples showed a moderately significant increase in Faith's PD and Chao1 value ($p = 0.029$ and $p = 0.03$) and suggestive decrease of Inverse Simpson index. The PS-PostInf samples demonstrated strong evidence of an increase in Shannon and chao1 index ($p =$

0.01 and $p < 0.0004$, respectively) compared to PS-PreInf. The changes in alpha diversity of each bacteria-nematode combination at the two time points were summarized in Figure 4.3B and detailed in Table 4.2.

There was some evidence that the diversity in all bacteria-nematode combinations prior to infection was not equal (Kruskal-Wallis, $p = 0.05$). After the infection assay, alpha diversity indices between the three bacteria-nematode combination did not differ significantly (Kruskal-Wallis test, $p = 0.3012$).

Patterns of change in beta diversity

We applied four beta diversity indices to evaluate the dissimilarity between samples of the three combinations sampled before and after the infection assay: Bray-Curtis dissimilarity, Jaccard, unweighted and Weighted UniFrac (see Materials and Methods). Bray-Curtis results showed a clear pattern of distinct sample clustering (Figure 5). The bacterial community structures were significantly different between the EC, PS, and BC sample sets (PERMANOVA, $p = 0.001$, $R2 = 0.541$). This result was robust to Jaccard and Weighted UniFrac indices (PERMANOVA, $p = 0.001$ for both), but not to UniFrac (PERMANOVA, $p = 0.06$). Between samples collected PreInf and PostInf, Bray-Curtis dissimilarity, Jaccard, and UniFrac indices indicated a significant difference (PERMANOVA, $p = 0.02$, $p = 0.01$, $p = 0.001$, respectively); however, Weighted UniFrac did not show a clear deviation ($p = 0.25$). PCoA ordination plots of the four beta diversity indices were shown in Figure 4.4.

Pseudomonas *ASV expansion after slug infection in the PS samples*

We analyzed the ASV components of the PS microbiome, with specific focus on those deriving from *Pseudomonas* spp., to evaluate the potential effects of the infection process

on this group of bacteria (Figure 4.5). The microbiome of PS-PostInf only shared a total of 36 ASVs with PS-PreInf, four of which were identified to be of the genus *Pseudomonas.* The relative abundance of these four ASVs in PS-PreInf samples was 13.23% of the total, and expanded to 59.58% in PS-PostInf samples. Out of these four ASVs, one was present in all three bacterial-nematode combinations sampled both before and after infection, and displayed an increase of 7.5%, 3.6%, and 3.2% in EC, PS, and BC communities, respectively. The three other *Pseudomonas* ASVs were only detected in the microbiomes of PS and BC nematodes, both before and after the infection assay; they were absent in all EC samples. While their abundances all increased after the course of infection in PS samples - by 6.9%, 30.5%, and 5.5%, there was a consistent decrease of 1.38%, 4.13%, and 0.73% from BC-PreInf to BC-PostInf samples. Two *Pseudomonas* ASVs only present in PS-PreInf accounted for just 0.004% abundance, whereas the 14 ASVs unique to the PS-PostInf had a total relative abundance of 1.11%.

Discussion

Complex composition of bacterial communities associated with P. hermaphrodita

The microbiome samples of all the EC, PS, and BC samples showed complex taxa (ASVs) compositions both before and after slug infection. The species richness and diversity of EC-PreInf and PS-PreInf communities were higher than our expectations, given that the nematodes had been bleached with hypochlorite, washed thoroughly with water and salt solution, and reared on pure cultures of *E. coli* or *Pseudomonas sp* prior to the infection assay. The intended bacteria were present in the samples collected before infection trials; however, *Pseudomonas sp.* and *E. coli* were a relatively minor component of the community by the time the infection experiments began - as

characterized in our 16S composition analyses. There are many likely contributors to the unexpectedly high taxonomic diversity observed in the EC and PS samples, including laboratory contamination associated with our 'semi-sterile' lab techniques, some bacteria surviving the bleaching procedures, and the possibility that there are unknown bacterial endosymbionts associated with *P. hermaphrodita* as has been reported in other nematode systems (Denver et al. *PLoS Pathogens* 2008).

Before the infection trial, fifteen different phyla and 82 genera were observed in the BC-PreInf samples, with *Pseudochrobactrum, Flavobacterium, Raoultella,* and *Pseudomonas* being the four most abundant taxa in the community. These four genera all have been reported in association or rivalry with nematodes: *Pseudochrobactrum* spp. with the EPNs *Heterorhabditidae* (Sharifi 2016), *Raoultella* spp. with the root-knot and pine wood nematodes (Wen 2015, Liu 2019), *Flavobacterium sp.* with the insect-parasitic nematode *Rhabditis blumi,* the soybean cyst nematode *Heterodera glycines*, the oriental beetle pathogenic nematode *Butlerius sp*. (Tian 2000, Ook Kim 2011, Woong Park, 2007), and *Pseudomonas* spp. with *Caenorhabditis elegans*. The high-abundance components in BC-PreInf samples is relatively similar to that of *Caenorhabditis elegans,* whose dominant genera are *Pseudomonas, Stenotrophomonas, Ochrobactrum*, *Sphingomonas*, and unclassified Enterobacteriaceae (*Raoultella* is a representative of this family) (Dirksen 2016, Samuel 2016). *Stenotrophomonas, Ochrobactrum,* and *Sphingomonas* spp. are all detected in top 10 most abundant genera in our BC-PreInf bacterial community.

The complex features of the bacterial community structure and the results of slug mortality during the infection assay (See Supplementary Document 1) indicated that *P.*

hermaphrodita can carry out slug infection with various bacteria regardless of the primary bacterial food source, albeit at different rates. This observation agrees with Rae and colleague's finding that the nematode is able to associate with non-specific and complex microbial assemblages (Rae et al., 2010). However, the slug-killing efficacy varied between different bacteria-nematode combinations. The BC-cultured nematodes were the first out the three treatments to cause significant slug mortality compared to the negative controls on Day 5, followed by the PS nematodes on Day 6, and lastly the EC nematodes, starting Day 10. Once began, the nematodes treated with *Pseudomonas* took only three days to kill slugs and were the first to cause 100% slug mortality on Day 8, followed by the BC nematodes (100% slug mortality on Day 9 - spanning five days) and EC nematodes (Day 10). The results may imply that *Pseudomonas* spp. might potentially be a bacterial partner that could help maximize the nematode's ability to kill slug hosts. This could be done by assisting *P. hermaphrodita*'s pathogenicity and/or providing a favorable food source, which may benefit the worm's growth and development for better infection. These insights could be useful from the biocontrol viewpoint - to optimize the pathogenicity and slug-killing efficacy of *P. hermaphrodita* nematodes.

Analysis of ASVs in both the PS-PreInf and PS-PostInf sample sets revealed four common ASVs that belonged to the genus *Pseudomonas* (see Results). Out of these four ASVs, we detected the presence of one particular ASV in all sample sets, of which the relative abundance consistently increased after the course of infection in all three bacterial treatment. This suggests that the specific ASV is stably retained with the nematode and could be important to the slug infection activity. The abundance of three other common ASV before and after infection also expanded, however, their absence in

all EC sample, as well as the decline in abundance observed in the BC treatment could indicate that these ASVs might originated from bacteria that are beneficial to the nematode, but not crucial to the pathogenicity and could be easily outgrown by other bacteria in the microbiome.

Bacterial community shifts after slug infection

In general, we found that alpha diversity indices increased in all of the post-infection samples, especially the PS-PostInf and BC-PostInf samples. For BC-PostInf samples, Chao1 and Faith's PD alpha diversity indices increased significantly, which may indicate the enrichment of low-abundance taxa during the course of infection and also implied that the diversity in the PostInf microbiome is not just the result of the presence of a few highly diverse taxonomic groups, but rather a broader range of phylogenetically different taxa. Indeed, for example, the total number of different genera in the BC microbiomes more than tripled after infection, however, the composition of the top 20 most abundant taxa remained stable. Likewise, for the *Pseudomonas*-enriched treatment we reported a significant increase in Chao1 and Shannon indices after the assay, indicating a spike in rare taxa post infection, although composition analysis pointed to the dominance of taxa of the genus *Pseudomonas* in the community.

Microbial community differs by type of bacterial treatment and time of collection.

We found that the difference between samples from different bacteria-nematode combinations were significant for Bray-Curtis and weighted UniFrac metrics, but not for unweighted UniFrac. This means that between bacterial treatments, the communities tend to have a common set of core ASVs (i.e. similar richness), but these ASVs have highly varied abundances (i.e. diversity). Additionally, it is likely that the higher abundance taxa

are distinct between samples from different bacterial treatments, and shared taxa are of low abundance.

In the meantime, samples collected before and after infection did not show a significant difference by the weighted UniFrac metric, indicating that only low-abundance taxa differ between the communities pre- and post-infection. This observation is in concordance with the findings of the alpha diversity analyses discussed above.

Conclusion

In this chapter, we described the microbial community associated with the slug-parasitic nematode *P. hermaphrodita*, and reported the compositional changes in the nematode's microbiome before and after slug infection. We observed a substantial increase in species richness; however, the composition of the top abundant taxa pre-infection remained mostly the same post-infection. We also detected four specific taxa of the genus *Pseudomonas*, whose relative abundance expanded remarkably after infection in the communities originally enriched with *Pseudomonas.* This finding suggests future investigations to further evaluate the possible role of *Pseudomonas sp.* in the killing of slugs, which will assist the optimization of the slug biocontrol method using *P. hermaphrodita*.

Table 4. 1. Naming scheme of samples.

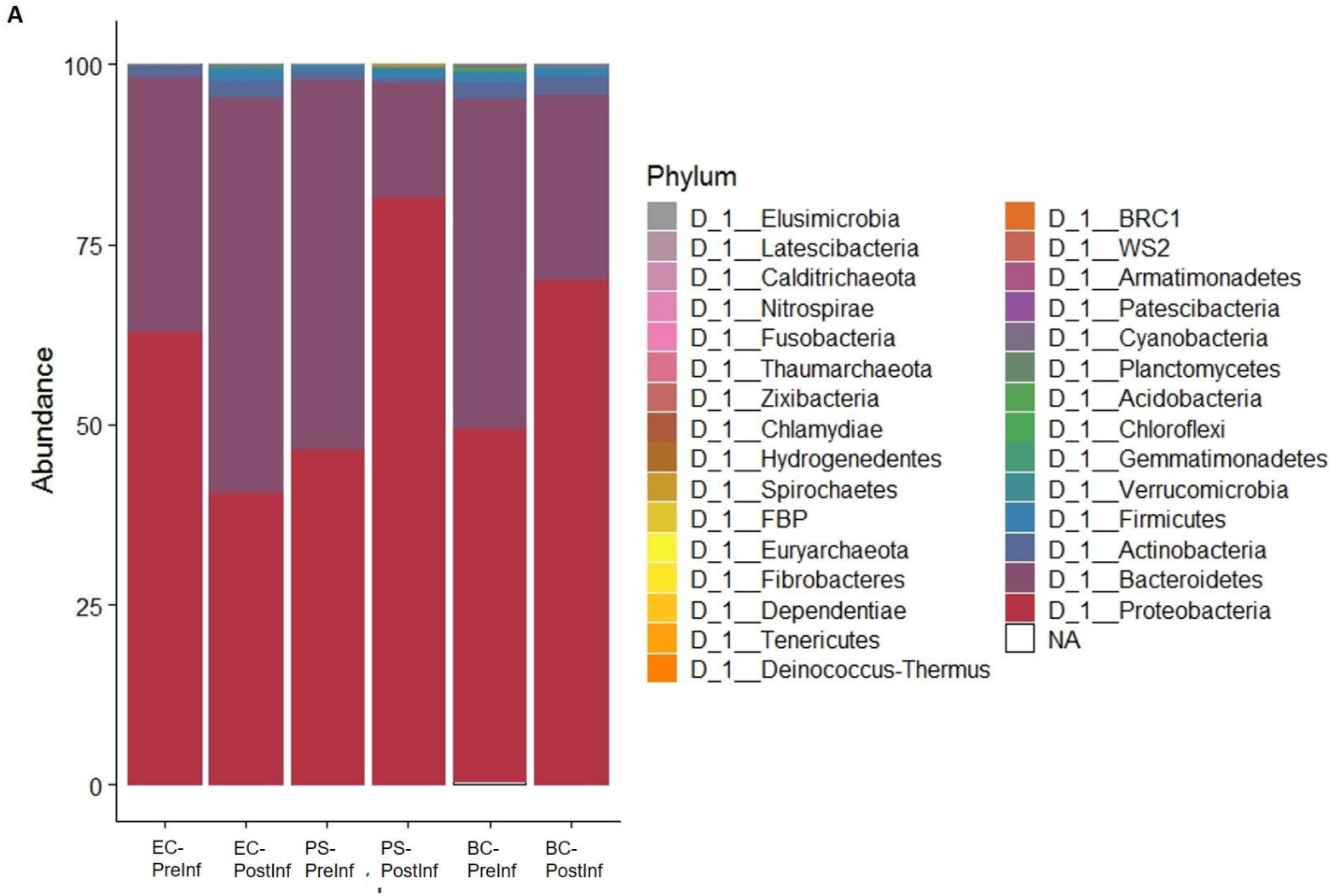
Bacterial treatment	**Time point**	**Sample name**
E. coli-enriched samples	before infection	EC-PreInf
	after infection	EC-PostInf
Pseudomonas sp. - enriched samples	before infection	PS-PreInf
	after infection	PS-PostInf
Complex bacterial community	before infection	BC-PreInf
	after infection	BC-PostInf

Table 4. 2. The number of different genera detected in each sample.

Bacterial treatment	Sample (before infection)	No. of genera	Sample (after infection)	No. of genera
E. coli	EC-PreInf-1	30	EC-PostInf-1	108
	EC-PreInf-2	31	EC-PostInf-2	185
	EC-PreInf-3	32	EC-PostInf-3	80
Pseudomonas sp.	PS-PreInf-1	22	PS-PostInf-1	108
	PS-PreInf-2	27	PS-PostInf-2	115
	PS-PreInf-3	24	PS-PostInf-3	107
Bacterial complex	BC-PreInf-1	66	BC-PostInf-1	130
	BC-PreInf-2	36	BC-PostInf-2	182
	BC-PreInf-3	35	BC-PostInf-3	144

Table 4. 3. Changes in alpha diversity indices of each bacteria-nematode combination after infection trial.

Diversity metrics	***E. coli* (EC)**	**Bacterial Complex (BC)**	***Pseudomonas sp.* (PS)**
Shannon	**Increase** (p = 0.068)	No significant difference (p = 0.25)	**Increase** (p = 0.01)
Faith's PD	No significant difference (p = 0.67)	**Increase** (p = 0.029)	No significant difference (p = 0.7)
Inverse Simpson	No significant difference (p = 0.27)	**Decrease** (p = 0.09)	No significant difference (p = 0.24)
Chao1	No significant difference (p = 0.19)	**Increase** (p = 0.03)	**Increase** (p < 0.0004)


A
100
75
50
25
0
Abundance
EC-PreInf
EC-PostInf
PS-PreInf
PS-PostInf
BC-PreInf
BC-PostInf
Phylum
D_1__Elusimicrobia
D_1__Latescibacteria
D_1__Calditrichaeota
D_1__Nitrospirae
D_1__Fusobacteria
D_1__Thaumarchaeota
D_1__Zixibacteria
D_1__Chlamydiae
D_1__Hydrogenedentes
D_1__Spirochaetes
D_1__FBP
D_1__Euryarchaeota
D_1__Fibrobacteres
D_1__Dependentiae
D_1__Tenericutes
D_1__Deinococcus-Thermus
D_1__BRC1
D_1__WS2
D_1__Armatimonadetes
D_1__Patescibacteria
D_1__Cyanobacteria
D_1__Planctomycetes
D_1__Acidobacteria
D_1__Chloroflexi
D_1__Gemmatimonadetes
D_1__Verrucomicrobia
D_1__Firmicutes
D_1__Actinobacteria
D_1__Bacteroidetes
D_1__Proteobacteria
NA

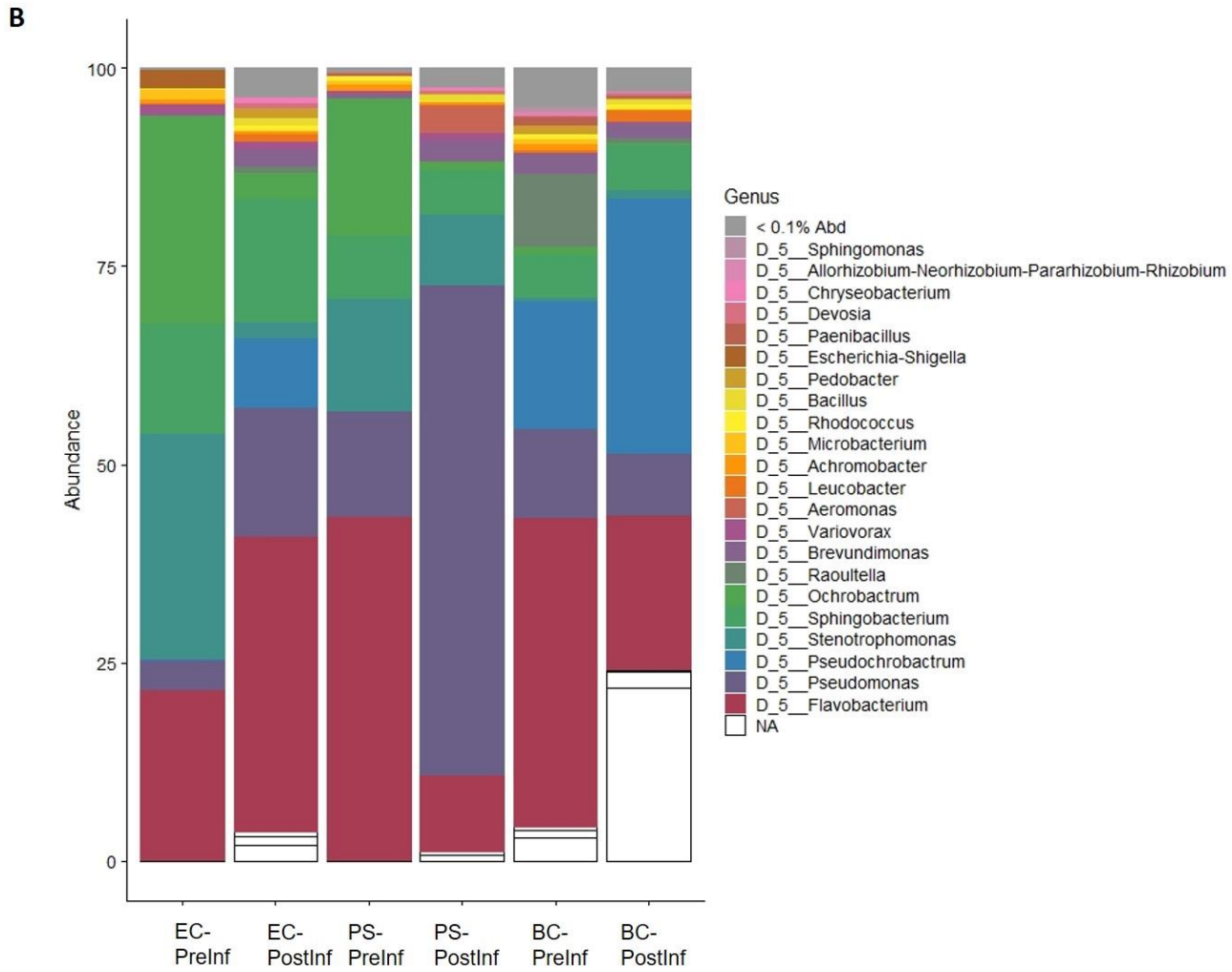


Figure 4. 1. (A). Phylum composition of communities associated with *P. hermaphrodita.* (B) Genus composition of the communities. Each stacked colored bar indicates a different genus with relative abundance of more than 0.1%.

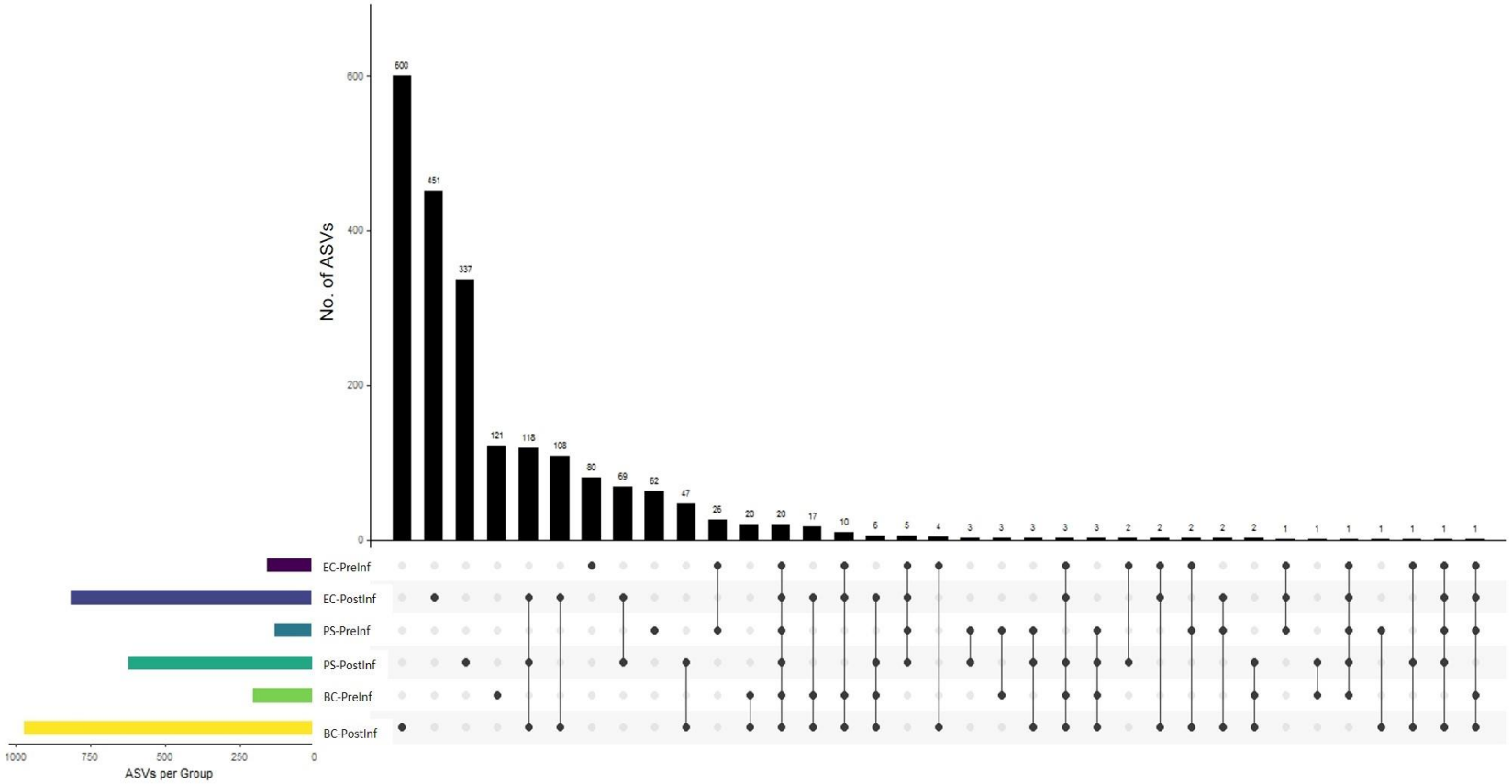
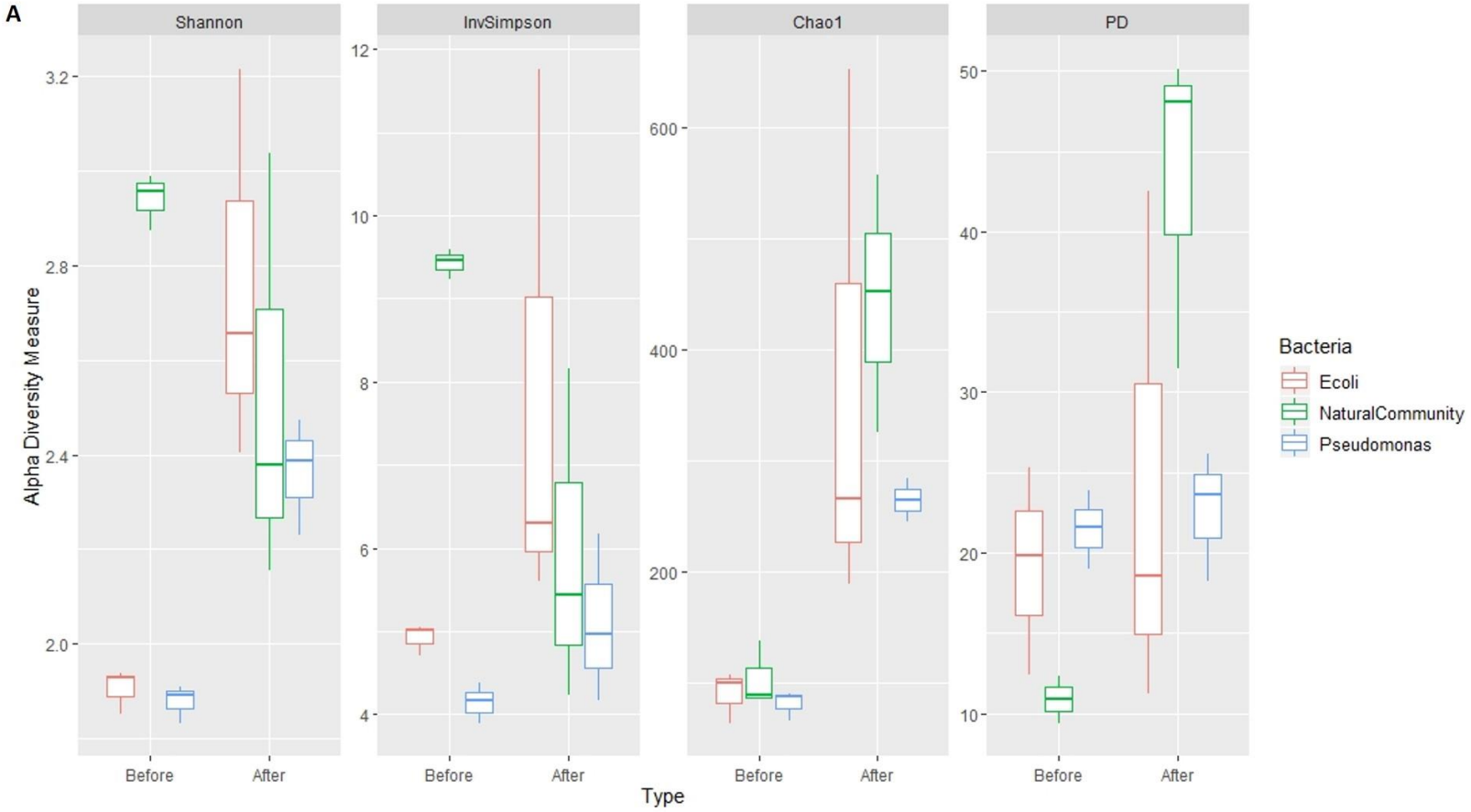


Figure 4. 2. The ASV components present in each of the sample sets. Each horizontal colored bar (left) indicates the number of unique ASVs per sample group. Each vertical bar in the plot illustrates the number of ASVs detected in the samples. The dots (bottom) correspond to the presence of those numbers of ASV in the sample sets, and co-appearance of one dot in multiple samples demonstrates the number of ASVs that are shared among those different sets.


A
Shannon
InvSimpson
Chao1
PD
Alpha Diversity Measure
Type
Before
After
Bacteria
Ecoli
NaturalCommunity
Pseudomonas

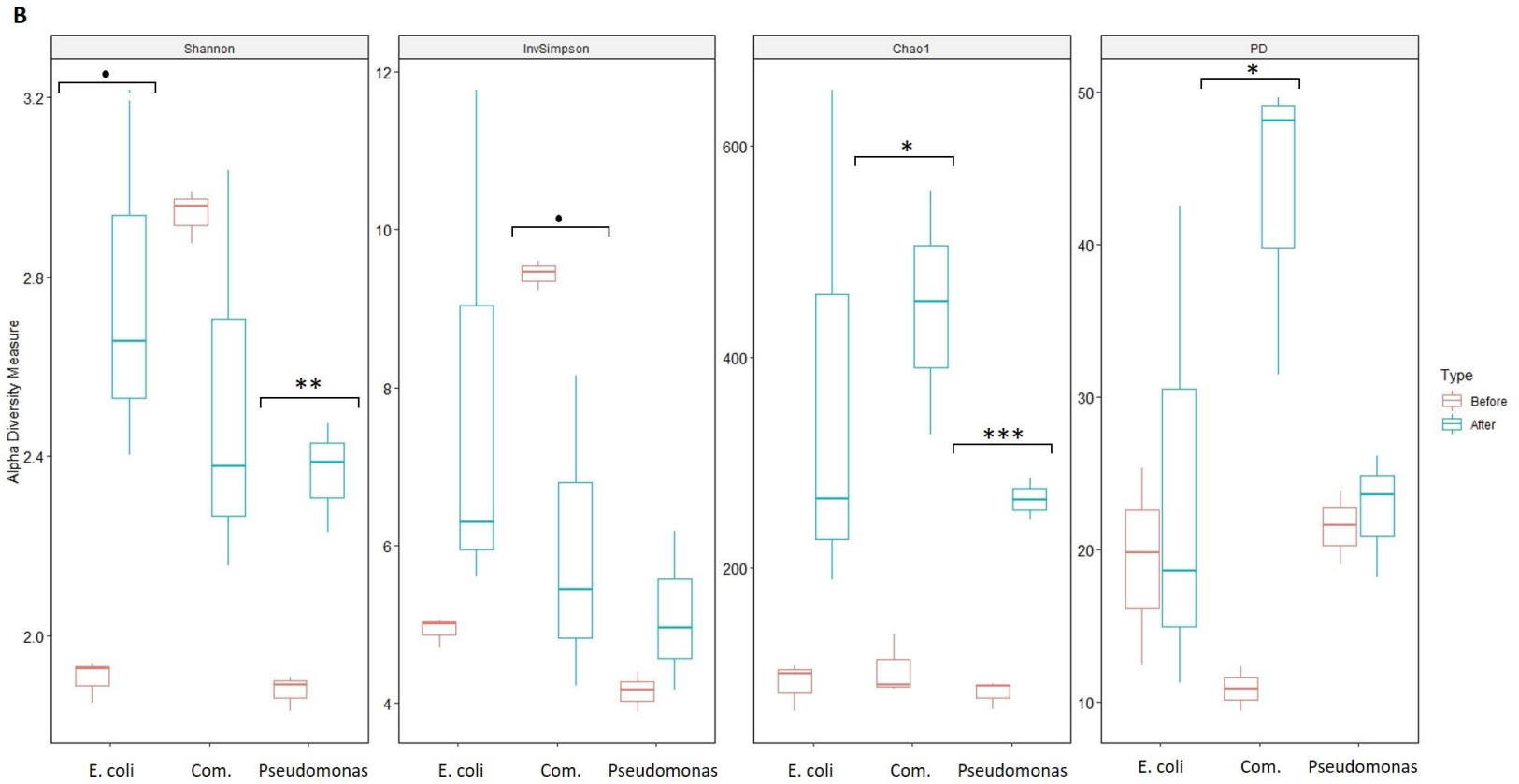


Figure 4. 3. Alpha diversity indices Shannon, Inverse Simpson, Chao1 and Faith's Diversity of samples grouped by (A) Time of collection and (B) Bacterial treatment. Significant differences of the metrics between pre- and post-infection of each bacterial treatment are indicated with the significant code: (***) 0-0.001, (**) 0.001-0.01, (*) 0.01-0.05, (.) 0.05-0.1, ( ) > 0.1.

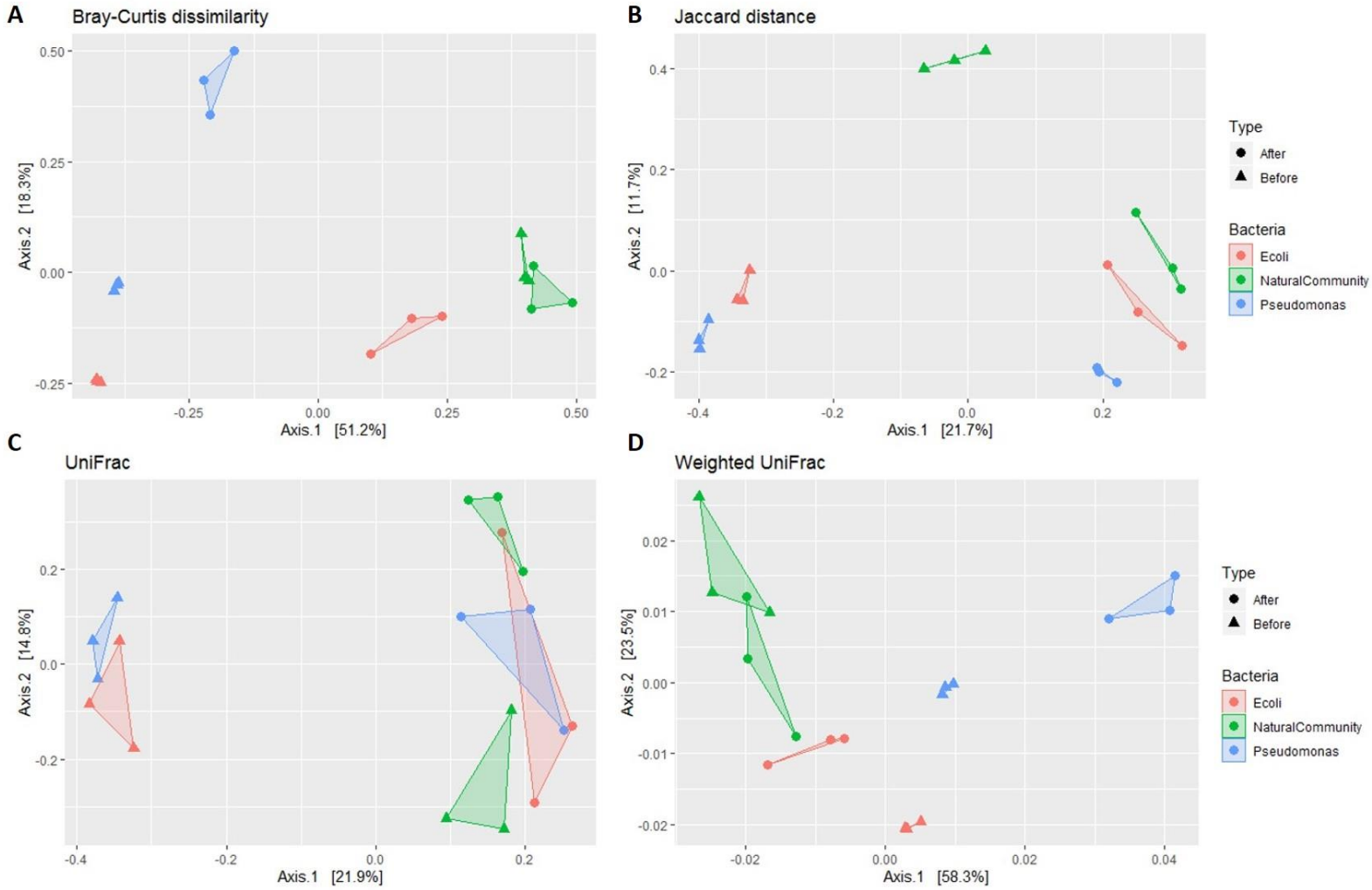


Figure 4. 4. Beta diversity indices of the microbiome samples. Principal coordinate analysis (PCoA) with a (A) Bray-Curtis dissimilarity, (B) Jaccard distance, (C) Unweighted UniFrac distance, and (D) Weighted UniFrac distance. The samples are colored by the type of bacterial treatment and shaped by the time of collection.

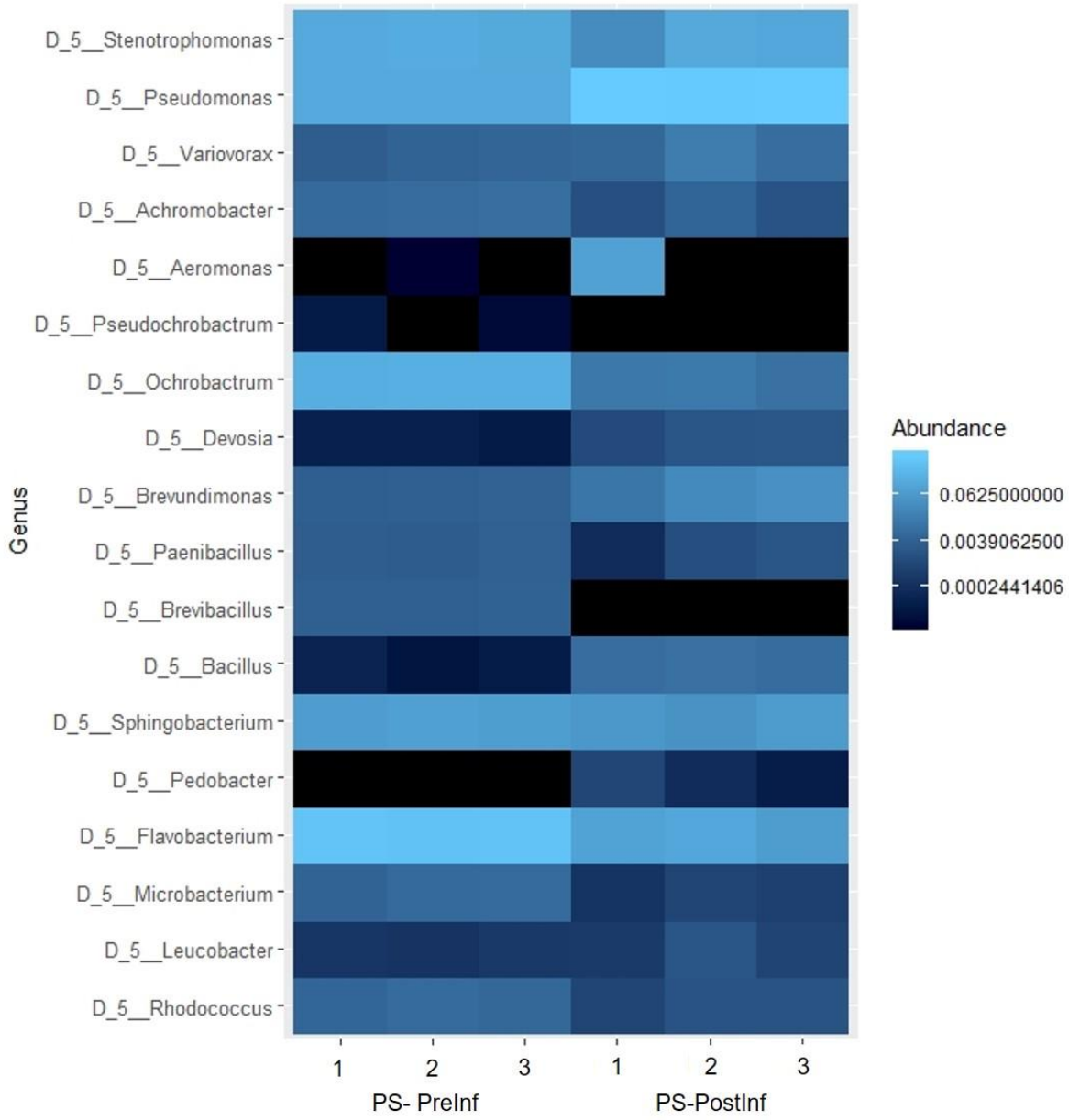


Figure 4. 5. Genus components of microbial communities in each PS biological replicate. Relative abundance of the observed genera are colored on a gradient - the darker the color, the higher abundance.

Chapter 5: Conclusion

This dissertation combined diverse evolutionary and genomic investigation methods, applied to diverse and complementary research subjects, in an effort to further understand the diversity and evolution of organisms in nature, with special focus on those that interact with *Pseudomonas* spp. bacteria. The work extended over a range of biological systems, from bacteriophages and bacterial communities to nematodes. The research analyzed different sets of data - ranging from single complete genomes to metagenomic sequences and microbiome amplicon data, and utilized different approaches such as laboratory slug infection assays, whole-genome sequencing and comparative genomics, 16S rRNA microbiome analyses, and shotgun metagenome sequencing analyses. The three data chapters are linked by a common thread: *Pseudomonas* bacteria. Chapter 2 focuses on bacteriophages infecting *Pseudomonas*, a factor that may substantially influence the bacteria's diversity. Chapter 3 progressed from the *Pseudomonas* phage complete genome dataset and transition to a broader microbial community framework, where the set of sequences were the references to search for general phage DNA in metagenomes. Chapter 4 investigated the bacterial community associated with slug-killing nematode, and began to explore the potential role of *Pseudomonas* spp. in the scheme of slug infection. This concluding chapter summarizes the findings of those data chapters, relates to the broader observations in other studies, and proposes directions for future research.

Phage genomic diversity and evolution

The specific goals of Chapter 2 were to evaluate the genomic diversity, capability to shift hosts, and selection force acting on bacteriophages infecting *Pseudomonas*. We observed enormous genomic diversity among the 130 genomes examined, with highly varied genome characteristics as well as a large number of phage singletons and ORFams. These findings agreed with results from other studies of other phage groups e.g. Mycobacteriphage (Hatfull et al., 2010; Pope et al., 2015), *Bacillus* phages (Grose et al., 2014), *Acinetobacter* phages (Turner et al., 2018), *Lactobacillus* phages (Kiliç et al., 2001), *Lactococcus lactis* phages (Murphy et al., 2016), and thus reaffirmed the commonly observed pattern of bacteriophage hyper-diversity in different environments. As previously discussed, the current database for *Pseudomonas* bacterial genome is still heavily skewed towards *P. aeruginosa* (by the time of this dissertation, 50% of the available draft and complete *Pseudomonas* genomes are *P. aeruginosa* - Winsor et al., 2011). The number of phage genomes sequenced deposited in public databases is correlated to, and much smaller than that of their own hosts. This is expected to substantially obstruct the comprehensive evaluation *Pseudomonas* phage diversity and evolution. Therefore, we consider it essential to discover and sequence novel phages, which would provide important resources for more thoroughly understanding the diversity of phage genes, and provide broader resources to evaluate the biology of these phages and their host bacteria.

Our second finding demonstrated that *Pseudomonas* phages have a considerable potential to shift hosts, as phages infecting different host trains and even species shared reportable sequence similarities, and diverse CRISPR spacer sequences from multiple

host species within a single phage were detected. While bacteriophages have been well-known for the high host-specificity observed in some strains (which is particularly suitable for the development of phage therapy), other groups of phages expressed the ability to infect more than one host species (Murphy et al., 2016). The observation suggests future study directions to explore phages' extent and rate of switching hosts, and also to further probe into the potential causes and environmental prompts that may lead to host shifting. These questions could be addressed in a microbiome theme, where the dynamic network among microbial organisms and the close interactions between phages and bacteria are likely to be important factors that influence both phage and bacterial evolution. For example, an induced disturbance of the microbial community (e.g. changing in temperature, humidity, or administration of specific antibiotics) that reduced the bacterial host population might be used to test phages' host shifting rate. By putting investigations into more practical framework, we expect the achievement of helpful insights that would unveil more details about the nature of phages and could be applied towards the development of clinical applications.

Lastly, we identified a dominating pattern of purifying selection on phage genes using the ratio of nucleotide diversity at non-synonymous (πN) and synonymous sites (πS). The evolutionary test $\pi N/\pi S$ and its sister - dN/dS have been applied to various groups of phages, and our findings are in line with previous results of other studies, which demonstrated that purifying selection are exerting on the majority of phage genes. Other studies have noted that the ratios are higher in smaller genomes in general (Kuo et al., 2009; Sela et al., 2016), which suggests further investigation into smaller phages i.e. Leviviruses and Inoviruses. Small viruses are underrepresented in our genome collection,

as only two Leviphages and two Inophages were available on GenBank database at the time. Currently, more Inoviruses and Leviviruses, including those infecting *Pseudomonas*, are being discovered and continuously added to genome databases (Roux et al., 2019). Including more phages of these families in comparative analyses may reveal novel evolutionary patterns.

Searching for phage sequences in metagenome data

Chapter 3 aimed to identify and characterize phage DNA in cyst nematode metagenome data that were not specifically generated for virome studies. We used three different, independent phage DNA search strategies, which were based on both public phage gene databases and complete genome sequences, and detected complex phage signals in all of the eight cyst nematode metagenome data sets included. We noted that using the program VirSorter, the numbers of phage DNA identified in any of our data sets were substantially lower than those in other studies employing the same software (Garin-Fernandez et al., 2018; Leigh et al., 2018; Miller-Ensminger et al., 2018; Nigro et al., 2018). This might have resulted partly from a lack of a viral enrichment procedure prior to our metagenome sequencing, while the other studies followed standard handling protocols including the enrichment step. It is not clear of the extent of phage matter that was lost during our initial prep, and how this would skew the search results. Therefore, future work might include a comparison of two data sets from a single microbiome, generated with and without the viral purification procedure. During preparation for sequencing, the microbiome sample of interest could be divided in half, one would undergo the standard viral enrichment protocol (i.e. following the viral targeted approach), and the other would have no enrichment step (i.e. untargeted approach). The difference in phage signals

detected would allow the evaluation of the effect of the viral enrichment step, and also reveal the characteristics of particular sequences that tend to be lost (or remain) without those steps. For example, we speculate that more circular sequences with signatures of phage hallmark genes, which indicates phage lytic particles, would be found in the viral targeted data; while the untargeted data would contain more prophage sequences.

Our findings also showed that among our eight data sets, the numbers of phage sequence contigs identified in *H. glycines*-associated metagenomes are generally lower than those of *G. pallida* (about one-third on average). It was unclear whether the difference was truly biological or caused by analytical discrepancy. We noted the lower quality of *G. pallida* metagenome assemblies, which raised the concern that lower-confidence assembled contigs may resulted in misidentification of phages and/or falsely adding up multiple fragmented parts of the same phage sequence, inflating the number of phages reported. Therefore, we propose that future studies include optimization of the assembling step and ensure the quality uniformity across data sets. Another suggested approach to circumvent issues regarding metagenome assembly is to develop a bioinformatic pipeline that recognizes and detects patterns of viral sequences directly from raw reads, for example aligning reads to a marker gene database and characterizing based on the alignments, which was employed for the program MetaPhlAn2 (Truong et al., 2015). This method may help avoiding technical errors of assembly, such as fragmented and incomplete sequences, chimeras (i.e. sequences from different genomes that are incorrectly assembled into one contig). Using raw reads will be especially helpful when comparing signals from different metagenomic data and provide a valuable exploratory tool to investigate phage signals in metagenomes.

Microbial communities have been known to be fluid and regularly experience natural temporary changes (Caporaso et al., 2011; Gerber, 2014; Uhr et al., 2019). As all of our four *H. glycines* samples were collected around the same time - which was different than the collection time points of the four *G. pallida* samples, the speculation that low phage biomass is in fact a temporal state remains untested. Therefore, future research might focus on testing the stability of the phage community in nematode metagenomes over time. Microbiome samples could be collected in a time series, with a set of time points ranging from shorter intervals (i.e. hours, days) to long period of time (weeks, months, years). Findings of this experiment would be able to provide a broader picture of the phage component in metagenomes and potential roles/ significance of the phage community as well as particular phage sequences. For our work in particular, more information about the sequences, especially the 5,461-bp contigs observed (e.g. whether or not they were in fact a contaminant, the temporal relative abundance, and the importance of the phage) could be further inferred.

Microbial community associated with slug-parasitic nematodes

Chapter 4 examined microbiomes associated with the slug-parasitic nematode *P. hermaphrodita* in the context of experimental slug infections in the lab. Three bacterial communities, including nematodes fed on *E. coli OP50, Pseudomonas sp.,* and the bacterial complex associated with the nematodes, were set up as media for *P. hermaphrodita* with the intention of evaluating the slug-killing efficiency of nematodes accompanied by these bacteria. The three treatments demonstrated varied in their capacity to kill slugs, which may implicate roles of the bacterial partners in the slug infection process. However, we noted the non-uniform microbial communities that were

present in the PS-PreInf and EC-PostInf samples instead of the expected sole presence of *Pseudomonas sp.* and *E. coli*, respectively. This is likely to have resulted from multiple contributing factors such as residual bacteria that was retained within the nematode, and the 'semi-sterile' laboratory technique used to handle nematodes in the infection experiment, which is the current standard practice in this field of research. This observation suggested future research focused on the role of bacteria in this system should implement improved sterile nematode culturing techniques, essentially repeat the infection assay with better starting sterilization procedures. Nematode sterilization could be tested by submerging in sterilizing agents e.g. thimerosal (Wilson et al., 1995b), Nycodenz, a gradient method which is able to separate worms and their bacterial symbionts (Woyke et al., 2006), and culturing media solution with multi antibiotics. More confident conclusions about the roles of the bacterial partners might be revealed as the nematode is grown on monoxenic bacterial cultures.

An additional future direction in this research program will be to better dissect the relative roles of the nematodes and the bacteria in killing slug hosts, and help resolve the contradicting and controversial findings in current literature (Rae et al., 2010; Tan and Grewal, 2001b). Following the optimized sterilization procedures, 'clean' nematode might be introduced in the proximity of grey field slugs and monitor slug mortality. *P. hermaphrodita*'s performance may help verify its intrinsic killing capability. Clearly, we cannot exclude the possibility that the nematode may pick up certain bacteria within the microbiome of slugs, which is non-pathogenic to the slug host until being put in an association with *P. hermaphrodita.* In that case, the nematode's unassisted pathogenicity may not be confirmed, and it would be difficult to limit the presence of the slug

microbiome without affecting their vitality. Nevertheless, this approach using sterilized nematodes i.e. without their native bacteria may provide a closer view to the nematode's nature. Regarding the role of bacteria, they may also be tested separately for killing ability. Bacterial pure cultures of different experimental concentration might be injected into the slug mantle, or pipetted directly over the opening of the dorsal integumental pouch, posterior to the mantle, where nematode infective juveniles presumably enter the slug body (Tan and Grewal, 2001a) to facilitate bacterial infection. If choosing injection, pilot study assessing the effect of injecting using syringes to slug's vitality should be conducted, in which slug negative controls (i.e. injected with sterile solution such as saline) will need to be handled and monitored carefully.

Finally, a prospective research direction could focus on testing different natural strains of *P. hermaphrodita.* It is possible that different *P. hermaphrodita* strains possess significantly varied characteristics, including pathogenicity towards slugs. Multiple strains that are genetically distinguished have recently been identified team (Howe et al. - in preparation), which could provide the nematode source for evaluating experiments. While Chapter 4 in this dissertation work has been centered on testing different bacterial partners in the association, the nematode itself should also be examined among different strain candidates to investigate any variation in slug-killing efficiency. The results of this evaluation may contribute to the development and optimization of the slug biocontrol method using *P. hermaphrodita.*

Concluding remarks

Overall, this dissertation research provides new insights into the dynamics of *Pseudomonas* bacteriophage genomes and nematode microbiomes associated with

bacteria of the genus *Pseudomonas.* We present different analysis approaches, applied on multiple types of data sets in an effort to further understand the characteristics, biology, dynamics, interactions, and evolution of various biological organisms, and contribute to the broader knowledge of the *Pseudomonas* genus and its associates.

Bibliography

Ackermann, H. W. (2012). “Bacteriophage Electron Microscopy,” in *Advances in Virus Research* (Academic Press Inc.), 1–32. doi:10.1016/B978-0-12-394621-8.00017-0.

Ait Tayeb, L., Ageron, E., Grimont, F., and Grimont, P. A. D. (2005). Molecular phylogeny of the genus Pseudomonas based on rpoB sequences and application for the identification of isolates. *Res. Microbiol.* 156, 763–773. doi:10.1016/j.resmic.2005.02.009.

Aktar, W., Sengupta, D., and Chowdhury, A. (2009). Impact of pesticides use in agriculture: Their benefits and hazards. *Interdiscip. Toxicol.* 2, 1–12. doi:10.2478/v10102-009-0001-7.

Allen, L. Z., Ishoey, T., Novotny, M. A., McLean, J. S., Lasken, R. S., and Williamson, S. J. (2011). Single virus genomics: A new tool for virus discovery. *PLoS One* 6. doi:10.1371/journal.pone.0017722.

Almeida, N. F., Shuangchun, Y., Rongman, C., Clarke, C. R., Morris, C. E., Schaads, N. W., et al. (2010). PAMDB, a multilocus sequence typing and analysis database and website for plant-associated microbes. *Phytopathology* 100, 208–215. doi:10.1094/PHYTO-100-3-0208.

Amos, L. A., and Klug, A. (1975). Three-dimensional image reconstructions of the contractile tail of T4 bacteriophage. *J. Mol. Biol.* 99, 51–64. doi:10.1016/s0022-2836(75)80158-6.

Andreolli, M., Zapparoli, G., Angelini, E., Lucchetta, G., Lampis, S., and Vallini, G. (2019). Pseudomonas protegens MP12: A plant growth-promoting endophytic bacterium with broad-spectrum antifungal activity against grapevine phytopathogens. *Microbiol. Res.* 219, 123–131. doi:10.1016/j.micres.2018.11.003.

Anzai, Y., Kim, H., Park, J. Y., Wakabayashi, H., and Oyaizu, H. (2000). Phylogenetic affiliation of the pseudomonads based on 16S rRNA sequence. *Int. J. Syst. Evol. Microbiol.* 50, 1563–1589. doi:10.1099/00207713-50-4-1563.

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

Ashkenasy, N., Sánchez-Quesada, J., Bayley, H., and Ghadiri, M. R. (2005). Recognizing a Single Base in an Individual DNA Strand: A Step Toward DNA Sequencing in Nanopores. *Angew. Chemie Int. Ed.* 44, 1401–1404. doi:10.1002/anie.200462114.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* 9, 75. doi:10.1186/1471-2164-9-75.

Bachrach, U., and Friedmann, A. (1971). Practical procedures for the purification of bacterial viruses. *Appl. Microbiol.* 22, 706–715.

Baker, G. C., Smith, J. J., and Cowan, D. A. (2003). Review and re-analysis of domain-specific 16S primers. *J. Microbiol. Methods* 55, 541–555. doi:10.1016/j.mimet.2003.08.009.

Balvočiūtė, M., and Huson, D. H. (2017). SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? *BMC Genomics* 18, 114. doi:10.1186/s12864-017-3501-4.

Bansal, A. K., and Meyer, T. E. (2002). Evolutionary analysis by whole-genome

comparisons. *J. Bacteriol.* 184, 2260–72. doi:10.1128/jb.184.8.2260-2272.2002.
Barker, G. M. (2002). *Molluscs as crop pests*. CABI Pub.
Behjati, S., and Tarpey, P. S. (2013). What is next generation sequencing? *Arch. Dis. Child. Educ. Pract. Ed.* 98, 236–238. doi:10.1136/archdischild-2013-304340.
Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2007). GenBank. *Nucleic Acids Res.* 35, D21-5. doi:10.1093/nar/gkl986.
Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59. doi:10.1038/nature07517.
Bergan, T. (1978). "Chapter V Phage Typing of Pseudomonas aeruginosa," in, 169–199. doi:10.1016/S0580-9517(08)70662-7.
Bergey, D. H., and Breed, R. S. (1957). *Bergey's manual of determinative bacteriology, by Robert S. Breed [and others]*. Baltimore,: Williams & Wilkins Co., doi:10.5962/bhl.title.10728.
Bernard, G. C., Egnin, M., and Bonsi, C. (2017). "The Impact of Plant-Parasitic Nematodes on Agriculture and Methods of Control," in *Nematology - Concepts, Diagnosis and Control* (InTech). doi:10.5772/intechopen.68958.
Bernstein-Ziv, R., Mushin, R., and Rabinowitz, K. (1973). Typing of Pseudomonas aeruginosa: comparison of the phage procedure with the pyocine technique. *J. Hyg. (Lond).* 71, 403–10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/4198203 [Accessed March 12, 2017].
Besemer, J., and Borodovsky, M. (1999). Heuristic approach to deriving models for gene finding. *Nucleic Acids Res.* 27, 3911–3920. doi:10.1093/nar/27.19.3911.
Besemer, J., Lomsadze, A., and Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* 29, 2607–18. doi:10.1093/NAR/29.12.2607.
Breitbart, M., Miyake, J. H., and Rohwer, F. (2004). Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol. Lett.* 236, 249–256. doi:10.1016/j.femsle.2004.05.042.
Breitbart, M., Thompson, L., Suttle, C., and Sullivan, M. (2007). Exploring the Vast Diversity of Marine Viruses. *Oceanography* 20, 135–139. doi:10.5670/oceanog.2007.58.
Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., et al. (2015). RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci. Rep.* 5, 8365. doi:10.1038/srep08365.
Brüssow, H., and Hendrix, R. W. (2002). Phage Genomics: Small is beautiful. *Cell* 108, 13–16. doi:10.1016/S0092-8674(01)00637-7.
Budzik, J. M., Rosche, W. A., Rietsch, A., and O'Toole, G. A. (2004). Isolation and Characterization of a Generalized Transducing Phage for Pseudomonas aeruginosa Strains PAO1 and PA14. *J. Bacteriol.* 186, 3270–3273. doi:10.1128/JB.186.10.3270-3273.2004.
Buell, C. R., Joardar, V., Lindeberg, M., Selengut, J., Paulsen, I. T., Gwinn, M. L., et al. (2003). The complete genome sequence of the Arabidopsis and tomato pathogen Pseudomonas syringae pv. tomato DC3000. *Proc. Natl. Acad. Sci. U. S. A.* 100,

10181–10186. doi:10.1073/pnas.1731982100.
Buermans, H. P. J., and den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochim. Biophys. Acta - Mol. Basis Dis.* 1842, 1932–1941. doi:10.1016/j.bbadis.2014.06.015.
Calderón, C. E., Ramos, C., de Vicente, A., and Cazorla, F. M. (2015). Comparative Genomic Analysis of *Pseudomonas chlororaphis* PCL1606 Reveals New Insight into Antifungal Compounds Involved in Biocontrol. *Mol. Plant-Microbe Interact.* 28, 249–260. doi:10.1094/MPMI-10-14-0326-FI.
Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643. doi:10.1038/ismej.2017.119.
Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi:10.1038/nmeth.3869.
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421. doi:10.1186/1471-2105-10-421.
Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., et al. (2011). Moving pictures of the human microbiome. *Genome Biol.* 12, R50. doi:10.1186/gb-2011-12-5-r50.
Carrias, A., Welch, T. J., Waldbieser, G. C., Mead, D. A., Terhune, J. S., and Liles, M. R. (2011). Comparative genomic analysis of bacteriophages specific to the channel catfish pathogen Edwardsiella ictaluri. *Virol. J.* 8, 6. doi:10.1186/1743-422X-8-6.
Catalano, C. E., Cue, D., and Feiss, M. (1995). Virus DNA packaging: the strategy used by phage ? *Mol. Microbiol.* 16, 1075–1086. doi:10.1111/j.1365-2958.1995.tb02333.x.
Centers for Disease Control and Prevention (2019). Parasitic Meningitis. Available at: https://www.cdc.gov/meningitis/parasitic.html [Accessed November 26, 2019].
Ceyssens, P.-J., Glonti, T., Kropinski, ndrew M., Lavigne, R., Chanishvili, N., Kulakov, L., et al. (2011). Phenotypic and genotypic variations within a single bacteriophage species. *Virol. J.* 8, 134. doi:10.1186/1743-422X-8-134.
Ceyssens, P.-J., and Lavigne, R. (2010). Bacteriophages of Pseudomonas. *Future Microbiol.* 5, 1041–55. doi:10.2217/fmb.10.66.
Ceyssens, P. J., Hertveldt, K., Ackermann, H. W., Noben, J. P., Demeke, M., Volckaert, G., et al. (2008). The intron-containing genome of the lytic Pseudomonas phage LUZ24 resembles the temperate phage PaP3. *Virology* 377, 233–238. doi:10.1016/j.virol.2008.04.038.
Ceyssens, P. J., Lavigne, R., Mattheus, W., Chibeu, A., Hertveldt, K., Mast, J., et al. (2006). Genomic analysis of Pseudomonas aeruginosa phages LKD16 and LKA1: Establishment of the φKMV subgroup within the T7 supergroup. *J. Bacteriol.* 188, 6924–6931. doi:10.1128/JB.00831-06.
Chakravorty, S., Helb, D., Burday, M., Connell, N., and Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J. Microbiol. Methods* 69, 330–339. doi:10.1016/j.mimet.2007.02.005.
Chao, A. (1984). Estimation of the Number of Classes in a Population Nonparametric Estimation of the Number of Classes in a Population.

Chen, J., Glémin, S., and Lascoux, M. (2017). Genetic Diversity and the Efficacy of Purifying Selection across Plant and Animal Species. *Mol. Biol. Evol.* 34, 1417–1428. doi:10.1093/molbev/msx088.

Chen, J., and Novick, R. P. (2009). Phage-mediated intergeneric transfer of toxin genes. *Science (80-.).* 323, 139–141. doi:10.1126/science.1164783.

Chin-A-Woeng, T. F. C., Bloemberg, G. V., Mulders, I. H. M., Dekkers, L. C., and Lugtenberg, B. J. J. (2000). Root colonization by phenazine-1-carboxamide-producing bacterium Pseudomonas chlororaphis PCL1391 is essential for biocontrol of tomato foot and root rot. *Mol. Plant-Microbe Interact.* 13, 1340–1345. doi:10.1094/MPMI.2000.13.12.1340.

Clokie, M. R., Millard, A. D., Letarov, A. V, and Heaphy, S. (2011). Phages in nature. *Bacteriophage* 1, 31–45. doi:10.4161/bact.1.1.14942.

COLWELL, R. R., and MANDEL, M. (1964). ADANSONIAN ANALYSIS AND DEOXYRIBONUCLEIC ACID BASE COMPOSITION OF SOME GRAM-NEGATIVE BACTERIA. *J. Bacteriol.* 87, 1412–22. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14188722 [Accessed November 22, 2019].

Comeau, A. M., and Krisch, H. M. (2008). The capsid of the T4 phage superfamily: The evolution, diversity, and structure of some of the most prevalent proteins in the biosphere. *Mol. Biol. Evol.* 25, 1321–1332. doi:10.1093/molbev/msn080.

Contina, J. B., Dandurand, L. M., and Knudsen, G. R. (2018). A Spatial Analysis of the Potato Cyst Nematode *Globodera pallida* in Idaho. *Phytopathology* 108, 988–1001. doi:10.1094/PHYTO-11-17-0388-R.

Cornelis, P. ed. (2008). *Pseudomonas: Genomics and Molecular Biology - Google Books*. Horizon Scientific Press Available at: https://books.google.com/books?id=GzOK2yZDvbEC&lpg=PR26&dq=Über ein neues System der Bakterien. Arb Bakteriol Inst Karlsruhe 1&pg=PR12#v=onepage&q=stanier&f=false.

Cornelissen, A., Hardies, S. C., Shaburova, O. V, Krylov, V. N., Mattheus, W., Kropinski, A. M., et al. (2012). Complete genome sequence of the giant virus OBP and comparative genome analysis of the diverse ΦKZ-related phages. *J. Virol.* 86, 1844–52. doi:10.1128/JVI.06330-11.

Cresawn, S. G., Bogel, M., Day, N., Jacobs-Sera, D., Hendrix, R. W., and Hatfull, G. F. (2011). Phamerator: a bioinformatic tool for comparative bacteriophage genomics. *BMC Bioinformatics* 12, 395. doi:10.1186/1471-2105-12-395.

Darzins, A., and Casadaban, M. J. (1989). Mini-D3112 bacteriophage transposable elements for genetic analysis of Pseudomonas aeruginosa. *J. Bacteriol.* 171, 3909–16. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2544562 [Accessed March 12, 2017].

Davis, N. M., Proctor, Di. M., Holmes, S. P., Relman, D. A., and Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* 6. doi:10.1186/s40168-018-0605-2.

De Paepe, M., Hutinet, G., Son, O., Amarir-Bouhram, J., Schbath, S., and Petit, M.-A. (2014). Temperate Phages Acquire DNA from Defective Prophages by Relaxed Homologous Recombination: The Role of Rad52-Like Recombinases. *PLoS Genet.* 10, e1004181. doi:10.1371/journal.pgen.1004181.

De Souza, J. T., Mazzola, M., and Raaijmakers, J. M. (2003). Conservation of the

response regulator gene gacA in Pseudomonas species. *Environ. Microbiol.* 5, 1328–1340. doi:10.1111/j.1462-2920.2003.00438.x.

DE VOS, P., and DE LEY, J. (1983). Intra- and Intergeneric Similarities of Pseudomonas and Xanthomonas Ribosomal Ribonucleic Acid Cistrons. *Int. J. Syst. Bacteriol.* 33, 487–509. doi:10.1099/00207713-33-3-487.

De Vos, P., Van Landschoot, A., Segers, P., Tytgat, R., Gillis, M., Bauwens, M., et al. (1989). Genotypic relationships and taxonomic localization of unclassified Pseudomonas and Pseudomonas-like strains by deoxyribonucleic acid: Ribosomal ribonucleic acid hybridizations. *Int. J. Syst. Bacteriol.* 39, 35–49. doi:10.1099/00207713-39-1-35.

Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* 27, 4636–41. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10556321 [Accessed March 12, 2017].

Denner, E. B. M. (1999). Reclassif ication of Pseudomonas echinoides Neumann 1962, 343AL, in the genus Sphingomonas as Sphingomonas echinoides comb. nov. *Int. J. Syst. Bacteriol.* 49, 1103–1109. doi:10.1099/00207713-49-3-1103.

DOE Joint Genome Institute BBDuk Guide. Available at: https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/ [Accessed November 26, 2019].

Domingo, E., Parrish, C. R., and Holland, J. J. (2008). *Origin and evolution of viruses.* Amsterdam ;;London : Elsevier Academic Press Available at: http://www.worldcat.org/title/origin-and-evolution-of-viruses/oclc/272383166 [Accessed June 23, 2017].

Dowell, C. E., and Rosenblum, E. D. (1962). SEROLOGY AND TRANSDUCTION IN STAPHYLOCOCCAL PHAGE. *J. Bacteriol.* 84, 1071–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16561969 [Accessed November 26, 2019].

Duhaime, M. B., and Sullivan, M. B. (2012). Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology* 434, 181–186. doi:10.1016/j.virol.2012.09.036.

Dutilh, B. E., Cassman, N., McNair, K., Sanchez, S. E., Silva, G. G. Z., Boling, L., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* 5. doi:10.1038/ncomms5498.

Eberlein, C., Heuer, H., Vidal, S., and Westphal, A. (2016). Microbial Communities in *Globodera pallida* Females Raised in Potato Monoculture Soil. *Phytopathology* 106, 581–590. doi:10.1094/PHYTO-07-15-0180-R.

Edgell, D. R., Gibb, E. A., and Belfort, M. (2010). Mobile DNA elements in T4 and related phages. *Virol. J.* 7, 290. doi:10.1186/1743-422X-7-290.

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, 258–272. doi:10.1093/femsre/fuv048.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science (80-.).* 323, 133–138. doi:10.1126/science.1162986.

Eren, A. M., Esen, O. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., et al. (2015). Anvi'o: An advanced analysis and visualization platformfor 'omics data. *PeerJ* 2015. doi:10.7717/peerj.1319.

Esipov, S. E., Adanin, V. M., Baskunov, B. P., Kiprianova, E. A., and Garagulia, A. D. (1975). [New antibiotically active fluoroglucide from Pseudomonas aurantiaca]. *Antibiotiki* 20, 1077–81. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1225181 [Accessed November 19, 2019].

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* 61, 1–10. doi:10.1016/0006-3207(92)91201-3.

Feiss, M., and Catalano, C. E. (2013). Bacteriophage Lambda Terminase and the Mechanism of Viral DNA Packaging. Available at: https://www.ncbi.nlm.nih.gov/books/NBK6485/ [Accessed December 2, 2017].

Fenwick, D. W. (1940). Methods for the recovery and counting of cysts of heterodera schachtii from soil. *J. Helminthol.* 18, 155–172. doi:10.1017/S0022149X00031485.

Fogliano, V., Ballio, A., Gallo, M., Woo, S., Scala, F., and Lorito, M. (2002). Pseudomonas lipodepsipeptides and fungal cell wall-degrading enzymes act synergistically in biological control. *Mol. Plant-Microbe Interact.* 15, 323–333. doi:10.1094/MPMI.2002.15.4.323.

Frapolli, M., Défago, G., and Moënne-Loccoz, Y. (2007). Multilocus sequence analysis of biocontrol fluorescent Pseudomonas spp. producing the antifungal compound 2,4-diacetylphloroglucinol. *Environ. Microbiol.* 9, 1939–1955. doi:10.1111/j.1462-2920.2007.01310.x.

Gaffney, T. D., Lam, S. T., Ligon, J., Gates, K., Frazelle, A., Di Maio, J., et al. (1994). Global regulation of expression of antifungal factors by a Pseudomonas fluorescens biological control strain. *Mol. Plant. Microbe. Interact.* 7, 455–63. doi:10.1094/mpmi-7-0455.

Gardan, L., Bella, P., Meyer, J.-M., Christen, R., Rott, P., Achouak, W., et al. (2002). Pseudomonas salomonii sp. nov., pathogenic on garlic, and Pseudomonas palleroniana sp. nov., isolated from rice. *Int. J. Syst. Evol. Microbiol.* 52, 2065–2074. doi:10.1099/00207713-52-6-2065.

Garin-Fernandez, A., Pereira-Flores, E., Glöckner, F. O., and Wichels, A. (2018). The North Sea goes viral: Occurrence and distribution of North Sea bacteriophages. *Mar. Genomics* 41, 31–41. doi:10.1016/j.margen.2018.05.004.

Garrido-Sanz, D., Meier-Kolthoff, J. P., Göker, M., Martín, M., Rivilla, R., and Redondo-Nieto, M. (2016). Genomic and genetic diversity within the Pseudomonas fluoresces complex. *PLoS One* 11. doi:10.1371/journal.pone.0150183.

Gerber, G. K. (2014). The dynamic microbiome. *FEBS Lett.* 588, 4131–4139. doi:10.1016/j.febslet.2014.02.037.

Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., et al. (2005). Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* 3, 733–739. doi:10.1038/nrmicro1236.

Godan, D. (1983). Pest slugs and snails. Biology and control. *Pest slugs snails. Biol. Control.* Available at: https://www.cabdirect.org/cabdirect/abstract/19840508195 [Accessed January 21, 2018].

Gómez, P., and Buckling, A. (2011). Bacteria-phage antagonistic coevolution in soil. *Science (80-.).* 332, 106–109. doi:10.1126/science.1198767.

Gomila, M., Peña, A., Mulet, M., Lalucat, J., and García-Valdés, E. (2015). Phylogenomics and systematics in Pseudomonas. *Front. Microbiol.* 6. doi:10.3389/fmicb.2015.00214.

Green, M. R., and Joseph, S. (2012). *Molecular Cloning: A Laboratory Manual*. 4th ed. Cold Spring Harbor Laboratory Press.

Grewal, S. K., Grewal, P. S., and Hammond, R. B. (2003). Susceptibility of North American native and non-native slugs (Mollusca: Gastropoda) to Phasmarhabditis hermaphrodita (Nematoda: Rhabditidae). *Biocontrol Sci. Technol.* 13, 119–125. doi:10.1080/0958315021000054449.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* 8, 172. doi:10.1186/1471-2105-8-172.

Grose, J. H., and Casjens, S. R. (2014). Understanding the enormous diversity of bacteriophages: The tailed phages that infect the bacterial family Enterobacteriaceae. *Virology* 468–470, 421–443. doi:10.1016/j.virol.2014.08.024.

Grose, J. H., Jensen, G. L., Burnett, S. H., and Breakwell, D. P. (2014). Genomic comparison of 93 Bacillus phages reveals 12 clusters, 14 singletons and remarkable diversity. *BMC Genomics* 15, 855. doi:10.1186/1471-2164-15-855.

Ha, A. D., and Denver, D. R. (2018). Comparative Genomic Analysis of 130 Bacteriophages Infecting Bacteria in the Genus Pseudomonas. *Front. Microbiol.* 9, 1456. doi:10.3389/fmicb.2018.01456.

Haas, D., and Défago, G. (2005). Biological control of soil-borne pathogens by fluorescent pseudomonads. *Nat. Rev. Microbiol.* 3, 307–319. doi:10.1038/nrmicro1129.

Hartl, D. L. (2014). *Essential genetics : a genomics perspective*. Jones & Bartlett Learning.

Hatfull, G. F., Jacobs-Sera, D., Lawrence, J. G., Pope, W. H., Russell, D. A., Ko, C.-C., et al. (2010). Comparative genomic analysis of 60 Mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. *J. Mol. Biol.* 397, 119–43. doi:10.1016/j.jmb.2010.01.011.

Hayashi, T., Baba, T., Matsumoto, H., and Terawaki, Y. (1990). Phage-conversion of cytotoxin production in Pseudomonas aeruginosa. *Mol. Microbiol.* 4, 1703–1709. doi:10.1111/j.1365-2958.1990.tb00547.x.

Heilmann, S., Sneppen, K., and Krishna, S. (2010). Sustainability of Virulence in a Phage-Bacterial Ecosystem. *J. Virol.* 84, 3016–3022. doi:10.1128/jvi.02326-09.

Hendrix, R. W. (2010). "Recoding in Bacteriophages," in (Springer New York), 249–258. doi:10.1007/978-0-387-89382-2_11.

Hertveldt, K., Lavigne, R., Pleteneva, E., Sernova, N., Kurochkina, L., Korchevskii, R., et al. (2005). Genome comparison of Pseudomonas aeruginosa large phages. *J. Mol. Biol.* 354, 536–545. doi:10.1016/j.jmb.2005.08.075.

Hesse, C., Schulz, F., Bull, C. T., Shaffer, B. T., Yan, Q., Shapiro, N., et al. (2018). Genome-based evolutionary history of *Pseudomonas* spp. *Environ. Microbiol.* 20, 2142–2159. doi:10.1111/1462-2920.14130.

Hilario, E., Buckley, T. R., and Young, J. M. (2004). Improved resolution on the phylogenetic relationships among Pseudomonas by the combined analysis of atpD, carA, recA and 16S rDNA. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.* 86, 51–64. doi:10.1023/B:ANTO.0000024910.57117.16.

Hobert, O. (2010). The impact of whole genome sequencing on model system genetics: Get ready for the ride. *Genetics* 184, 317–319. doi:10.1534/genetics.109.112938.

Horvath, P., and Barrangou, R. (2010). CRISPR/Cas, the Immune System of Bacteria and Archaea. *Science (80-.).* 327, 167–170. doi:10.1126/science.1179555.

Howe, D. K., and Denver, D. R. (2008). Muller's Ratchet and compensatory mutation in Caenorhabditis briggsae mitochondrial genome evolution. *BMC Evol. Biol.* 8, 62. doi:10.1186/1471-2148-8-62.

Howe, D. K., Ha, A., Colton, A., De Ley, I., Rae, R., Ross, J., et al. Phylogenetic evidence for the invasion of a commercialized European Phasmarhabditis hermaphrodita lineage into North America and New Zealand.

Hraiech, S., Brégeon, F., and Rolain, J.-M. (2015). Bacteriophage-based therapy in cystic fibrosis-associated Pseudomonas aeruginosa infections: rationale and current status. *Drug Des. Devel. Ther.* 9, 3653–63. doi:10.2147/DDDT.S53123.

Huson, D. H., Richter, D. C., Mitra, S., Auch, A. F., and Schuster, S. C. (2009). Methods for comparative metagenomics. *BMC Bioinformatics* 10, S12. doi:10.1186/1471-2105-10-S1-S12.

Iglewski, B. H. (1996). *Pseudomonas*. University of Texas Medical Branch at Galveston Available at: http://www.ncbi.nlm.nih.gov/pubmed/21413324 [Accessed March 12, 2017].

Illumina (2013). PCR Amplicon, PCR Clean-up, and Index PCR. *16S Metagenomic Seq. Libr. Prep.* Available at: https://www.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf [Accessed November 26, 2019].

Janda, J. M., and Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *J. Clin. Microbiol.* 45, 2761–2764. doi:10.1128/JCM.01228-07.

Jones, J. T., Haegeman, A., Danchin, E. G. J., Gaur, H. S., Helder, J., Jones, M. G. K., et al. (2013). Top 10 plant-parasitic nematodes in molecular plant pathology. *Mol. Plant Pathol.* 14, 946–961. doi:10.1111/mpp.12057.

Jong, L. de (1926). Bijdrage tot de kennis van het mineralisatieproces.

Kang, B. R., Yang, K. Y., Cho, B. H., Han, T. H., Kim, I. S., Lee, M. C., et al. (2006). Production of indole-3-acetic acid in the plant-beneficial strain Pseudomonas chlororaphis O6 is negatively regulated by the global sensor kinase GacS. *Curr. Microbiol.* 52, 473–6. doi:10.1007/s00284-005-0427-x.

Karama, M., and Gyles, C. L. (2008). Characterization of verotoxin-encoding phages from Escherichia coli O103:H2 strains of bovine and human origins. *Appl. Environ. Microbiol.* 74, 5153–5158. doi:10.1128/AEM.00723-08.

Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi:10.1093/nar/gkf436.

Keen, E. C. (2015). A century of phage research: Bacteriophages and the shaping of modern biology. *BioEssays* 37, 6–9. doi:10.1002/bies.201400152.

Keen, E. C., Bliskovsky, V. V, Malagon, F., Baker, J. D., Prince, J. S., Klaus, J. S., et al. (2017). Novel "Superspreader" Bacteriophages Promote Horizontal Gene Transfer by Transformation. *MBio* 8, e02115-16. doi:10.1128/mBio.02115-16.

Kiliç, A. O., Pavlova, S. I., Alpay, S., Kiliç, S. S., and Tao, L. (2001). Comparative study of vaginal Lactobacillus phages isolated from women in the United States and Turkey: Prevalence, morphology, host range, and DNA homology. *Clin. Diagn.*

Lab. Immunol. 8, 31–39. doi:10.1128/CDLI.8.1.31-39.2001.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., et al. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* 41. doi:10.1093/nar/gks808.

Knezevic, P., Voet, M., and Lavigne, R. (2015). Prevalence of Pf1-like (pro)phage genetic elements among Pseudomonas aeruginosa isolates. *Virology* 483, 64–71. doi:10.1016/j.virol.2015.04.008.

Kozińska, A., Seweryn, P., and Sitkiewicz, I. (2019). A crash course in sequencing for a microbiologist. *J. Appl. Genet.* 60, 103–111. doi:10.1007/s13353-019-00482-2.

Krishnamurthy, S. R., and Wang, D. (2017). Origins and challenges of viral dark matter. *Virus Res.* 239, 136–142. doi:10.1016/j.virusres.2017.02.002.

Kristensen, D. M., Waller, A. S., Yamada, T., Bork, P., Mushegian, A. R., and Koonin, E. V (2013). Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *J. Bacteriol.* 195, 941–50. doi:10.1128/JB.01801-12.

Krumsiek, J., Arnold, R., and Rattei, T. (2007). Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* 23, 1026–8. doi:10.1093/bioinformatics/btm039.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi:10.1093/molbev/msw054.

Kuo, C. H., Moran, N. A., and Ochman, H. (2009). The consequences of genetic drift for bacterial genome complexity. *Genome Res.* 19, 1450–1454. doi:10.1101/gr.091785.109.

Kwan, T., Liu, J., Dubow, M., Gros, P., and Pelletier, J. (2006). Comparative Genomic Analysis of 18 Pseudomonas aeruginosa Bacteriophages. *J. Bacteriol.* 188, 1184–1187. doi:10.1128/JB.188.3.1184–1187.2006.

Kwan, T., Liu, J., DuBow, M., Gros, P., and Pelletier, J. (2005). The complete genomes and proteomes of 27 Staphylococcus aureus bacteriophages. *Proc. Natl. Acad. Sci. U. S. A.* 102, 5174–9. doi:10.1073/pnas.0501140102.

Labaer, J., Qiu, Q., Anumanthan, A., Mar, W., Zuo, D., Murthy, T. V. S., et al. (2004). The Pseudomonas aeruginosa PA01 gene collection. *Genome Res.* 14, 2190–200. doi:10.1101/gr.2482804.

Laetsch, D. R., and Blaxter, M. L. (2017). BlobTools: Interrogation of genome assemblies. *F1000Research* 6, 1287. doi:10.12688/f1000research.12232.1.

Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L., and Pace, N. R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl. Acad. Sci. U. S. A.* 82, 6955–6959. doi:10.1073/pnas.82.20.6955.

Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* 32, 11–16. doi:10.1093/nar/gkh152.

Lassmann, T., and Sonnhammer, E. (2005). Kalign--an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* 6, 298. doi:10.1186/1471-2105-6-298.

Lavigne, R., Burkal'tseva, M. V, Robben, J., Sykilinda, N. N., Kurochkina, L. P., Grymonprez, B., et al. (2003). The genome of bacteriophage phiKMV, a T7-like

virus infecting Pseudomonas aeruginosa. *Virology* 312, 49–59. doi:10.1016/s0042-6822(03)00123-5.

Lawrence, J. G., Hatfull, G. F., and Hendrix, R. W. (2002). Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. *J. Bacteriol.* 184, 4891–905. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=135278&tool=pmcentrez&rendertype=abstract [Accessed September 29, 2015].

Leigh, B. A., Djurhuus, A., Breitbart, M., and Dishaw, L. J. (2018). The gut virome of the protochordate model organism, Ciona intestinalis subtype A. *Virus Res.* 244, 137–146. doi:10.1016/j.virusres.2017.11.015.

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452. doi:10.1093/bioinformatics/btp187.

LILLEY, C. J., ATKINSON, H. J., and URWIN, P. E. (2005). Molecular aspects of cyst nematodes. *Mol. Plant Pathol.* 6, 577–588. doi:10.1111/j.1364-3703.2005.00306.x.

Lindberg, R. B., and Latta, R. L. (1974). Phage typing of Pseudomonas aeruginosa: clinical and epidemiologic considerations. *J. Infect. Dis.* 130 Suppl, S33-42. Available at: http://www.ncbi.nlm.nih.gov/pubmed/4213792 [Accessed October 25, 2015].

Liu, R., Liu, H., Feng, H., Wang, X., Zhang, C. X., Zhang, K. Y., et al. (2008). Pseudomonas duriflava sp. nov., isolated from a desert soil. *Int. J. Syst. Evol. Microbiol.* 58, 1404–1408. doi:10.1099/ijs.0.65716-0.

Loper, J. E., Hassan, K. A., Mavrodi, D. V., Davis, E. W., Lim, C. K., Shaffer, B. T., et al. (2012). Comparative genomics of plant-associated pseudomonas spp.: Insights into diversity and inheritance of traits involved in multitrophic interactions. *PLoS Genet.* 8. doi:10.1371/journal.pgen.1002784.

Loper, J. E., Kobayashi, D. Y., and Paulsen, I. T. (2007). The Genomic Sequence of Pseudomonas fluorescens Pf-5: Insights Into Biological Control. 97. Available at: https://apsjournals.apsnet.org/doi/pdf/10.1094/PHYTO-97-2-0233 [Accessed February 3, 2018].

Lopez, S., Paine, T. D., De Ley, P., De Ley, I. T., and McDonnell, R. D. (2014). Phasmarhabditis hermaphrodita (Nematoda: Rhabditidae), a potential biocontrol agent isolated for the first time from invasive slugs in North America. *Nematology* 16, 1129–1138. doi:10.1163/15685411-00002838.

Loveridge, E. J., Jones, C., Bull, M. J., Moody, S. C., Kahl, M. W., Khan, Z., et al. (2017). Reclassification of the specialized metabolite producer Pseudomonas mesoacidophila ATCC 31433 as a member of the Burkholderia cepacia complex. *J. Bacteriol.* 199. doi:10.1128/JB.00125-17.

Mahmoudabadi, G., and Phillips, R. (2018). A Comprehensive and Quantitative Exploration of Thousands of Viral Genomes.

Manaia, C. M., and Moore, E. R. B. (2002). Pseudomonas thermotolerans sp. nov., a thermotolerant species of the genus Pseudomonas sensu stricto. *Int. J. Syst. Evol. Microbiol.* 52, 2203–2209. doi:10.1099/ijs.0.02059-0.

Mandel, M. (1966). Deoxyribonucleic acid base composition in the genus Pseudomonas. *J. Gen. Microbiol.* 43, 273–292. doi:10.1099/00221287-43-2-273.

Mathee, K., Narasimhan, G., Valdes, C., Qiu, X., Matewish, J. M., Koehrsen, M., et al.

(2008). Dynamics of Pseudomonas aeruginosa genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* 105, 3100–3105. doi:10.1073/pnas.0711982105.

Mc Donnell, R., Paine, T., and Gormally, M. (2009). Slugs: A Guide to the Invasive and Native Fauna of California. Available at: http://www.ansp.org/ [Accessed November 28, 2019].

McCabe, K. M., Zhang, Y. H., Huang, B. L., Wagar, E. A., and McCabe, E. R. B. (1999). Bacterial species identification after DNA amplification with a universal primer pair. *Mol. Genet. Metab.* 66, 205–211. doi:10.1006/mgme.1998.2795.

McDonnell, R. J., Lutz, M. S., Howe, D. K., and Denver, D. R. (2018). First report of the gastropod-killing nematode, phasmarhabditis hermaphrodita, in Oregon, U.S.A. *J. Nematol.* 50, 77–78. doi:10.21307/jofnem-2018-014.

Mehrabi, Z., McMillan, V. E., Clark, I. M., Canning, G., Hammond-Kosack, K. E., Preston, G., et al. (2016). Pseudomonas spp. diversity is negatively associated with suppression of the wheat take-all pathogen. *Sci. Rep.* 6. doi:10.1038/srep29905.

Meyer, J. M., Geoffroy, V. A., Baida, N., Gardan, L., Izard, D., Lemanceau, P., et al. (2002). Siderophore typing, a powerful tool for the identification of fluorescent and nonfluorescent pseudomonads. *Appl. Environ. Microbiol.* 68, 2745–2753. doi:10.1128/AEM.68.6.2745-2753.2002.

Miess, H., Van Trappen, S., Cleenwerck, I., De Vos, P., and Gross, H. (2016). Reclassification of Pseudomonas sp. PB-6250T as Lysobacter firmicutimachus sp. nov. *Int. J. Syst. Evol. Microbiol.* 66, 4162–4166. doi:10.1099/ijsem.0.001329.

Mignard, S., and Flandrois, J. P. (2006). 16S rRNA sequencing in routine bacterial identification: A 30-month experiment. *J. Microbiol. Methods* 67, 574–581. doi:10.1016/j.mimet.2006.05.009.

Migula, W. (1894). Über ein neues System der Bakterien. Arb Bakteriol Inst Karlsruhe. 235–238.

Miller-Ensminger, T., Garretto, A., Brenner, J., Thomas-White, K., Zambom, A., Wolfe, A. J., et al. (2018). Bacteriophages of the urinary microbiome. *J. Bacteriol.* 200. doi:10.1128/JB.00738-17.

Mills, D. K., Entry, J. A., Voss, J. D., Gillevet, P. M., and Mathee, K. (2006). An assessment of the hypervariable domains of the 16S rRNA genes for their value in determining microbial community diversity: the paradox of traditional ecological indices. *FEMS Microbiol. Ecol.* 57, 496–503. doi:10.1111/j.1574-6941.2006.00135.x.

Mills, R., Rozanov, M., Lomsadze, A., Tatusova, T., and Borodovsky, M. (2003). Improving gene annotation of complete viral genomes. *Nucleic Acids Res.* 31, 7041–55. doi:10.1093/NAR/GKG878.

Mohan, A., Padiadpu, J., Baloni, P., and Chandra, N. (2015). Complete Genome Sequences of a Mycobacterium smegmatis Laboratory Strain (MC2 155) and Isoniazid-Resistant (4XR1/R2) Mutant Strains. *Genome Announc.* 3. doi:10.1128/genomeA.01520-14.

Moore, E. R. B., Mau, M., Arnscheidt, A., Böttger, E. C., Hutson, R. A., Collins, M. D., et al. (1996). The determination and comparison of the 16S rRNA gene sequences of species of the genus Pseudomonas (sensu stricto) and estimation of the natural intrageneric relationships. *Syst. Appl. Microbiol.* 19, 478–492. doi:10.1016/S0723-2020(96)80021-X.

Morris, C. E., Sands, D. C., Vinatzer, B. A., Glaux, C., Guilbaud, C., Buffière, A., et al. (2008). The life history of the plant pathogen Pseudomonas syringae is linked to the water cycle. *ISME J.* 2, 321–334. doi:10.1038/ismej.2007.113.

Mulet, M., Gomila, M., Scotta, C., Sánchez, D., Lalucat, J., and García-Valdés, E. (2012). Concordance between whole-cell matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry and multilocus sequence analysis approaches in species discrimination within the genus Pseudomonas. *Syst. Appl. Microbiol.* 35, 455–464. doi:10.1016/j.syapm.2012.08.007.

Mulet, M., Lalucat, J., and García-Valdés, E. (2010). DNA sequence-based analysis of the Pseudomonas species. *Environ. Microbiol.* 12, 1513–1530. doi:10.1111/j.1462-2920.2010.02181.x.

Munsch, P., Alatossava, T., Marttinen, N., Meyer, J. M., Christen, R., and Gardan, L. (2002). Pseudomonas costantinii sp. nov., another causal agent of brown blotch disease, isolated from cultivated mushroom sporophores in Finland. *Int. J. Syst. Evol. Microbiol.* 52, 1973–1983. doi:10.1099/ijs.0.02090-0.

Murphy, J., Bottacini, F., Mahony, J., Kelleher, P., Neve, H., Zomer, A., et al. (2016). Comparative genomics and functional analysis of the 936 group of lactococcal Siphoviridae phages. *Sci. Rep.* 6. doi:10.1038/srep21345.

Nakamura, Y., Itoh, T., Matsuda, H., and Gojobori, T. (2004). Biased biological functions of horizontally-transferred genes in prokaryotic genomes. *Nat. Genet.* 36, 760–766. doi:10.1038/ng1381.

Nekrutenko, A., Makova, K. D., and Li, W.-H. (2002). The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.* 12, 198–202. doi:10.1101/gr.200901.

Nelson, K. E., Weinel, C., Paulsen, I. T., Dodson, R. J., Hilbert, H., Martins dos Santos, V. A. P., et al. (2002). Complete genome sequence and comparative analysis of the metabolically versatile Pseudomonas putida KT2440. *Environ. Microbiol.* 4, 799–808. doi:10.1046/j.1462-2920.2002.00366.x.

Nigro, O. D., Tully, B. J., Jungbluth, S., Huber, J. A., Rappe, M. S., and Steward, G. (2018). A Comparison of Viral Populations Inhabiting Atlantic and Pacific Oceanic Crustal Fluids. *AGUFM* 2018, V43G–0203.

Nishimori, E., Kita-Tsukamoto, K., and Wakabayashi, H. (2000). Pseudomonas plecoglossicida sp. nov., the causative agent of bacterial haemorrhagic ascites of ayu, Plecoglossus altivelis. *Int. J. Syst. Evol. Microbiol.* 50, 83–89. doi:10.1099/00207713-50-1-83.

Norman, J. M., Handley, S. A., Baldridge, M. T., Droit, L., Liu, C. Y., Keller, B. C., et al. (2015). Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* 160, 447–460. doi:10.1016/j.cell.2015.01.002.

Nour, S. M., Lawrence, J. R., Zhu, H., Swerhone, G. D. W., Welsh, M., Welacky, T. W., et al. (2003). Bacteria associated with cysts of the soybean cyst nematode (Heterodera glycines). *Appl. Environ. Microbiol.* 69, 607–15. doi:10.1128/aem.69.1.607-615.2003.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). MetaSPAdes: A new versatile metagenomic assembler. *Genome Res.* 27, 824–834. doi:10.1101/gr.213959.116.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D., et al.

(2019). “vegan”: Community Ecology Package. Available at: https://cran.r-project.org/package=vegan [Accessed November 26, 2019].

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., et al. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* 42, D206-14. doi:10.1093/nar/gkt1226.

Özen, A. I., and Ussery, D. W. (2012). Defining the Pseudomonas genus: where do we draw the line with Azotobacter? *Microb. Ecol.* 63, 239–48. doi:10.1007/s00248-011-9914-8.

Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., et al. (2016). Uncovering Earth’s virome. *Nature* 536, 425–430. doi:10.1038/nature19094.

Paez-Espino, D., Pavlopoulos, G. A., Ivanova, N. N., and Kyrpides, N. C. (2017). Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. *Nat. Protoc.* 12, 1673–1682. doi:10.1038/nprot.2017.063.

Palleroni, N. J. (2010). The Pseudomonas Story. *Environ. Microbiol.* 12, 1377–1383. doi:10.1111/j.1462-2920.2009.02041.x.

Palleroni, N. J., Kunisawa, R., Contopoulou, R., and Doudoroff, M. (1973). Nucleic acid homologies in the genus Pseudomonas. *Int. J. Syst. Bacteriol.* 23, 333–339. doi:10.1099/00207713-23-4-333.

Palleroni, N. J., and Moore, E. R. B. (2004). “Taxonomy of Pseudomonads: Experimental Approaches,” in *Pseudomonas* (Springer US), 3–44. doi:10.1007/978-1-4419-9086-0_1.

Park, S. C., Shimamura, I., Fukunaga, M., Mori, K. I., and Nakai, T. (2000). Isolation of bacteriophages specific to a fish pathogen, Pseudomonas plecoglossicida, as a candidate for disease control. *Appl. Environ. Microbiol.* 66, 1416–1422. doi:10.1128/AEM.66.4.1416-1422.2000.

Parte, A. C. (2018). LPSN - List of prokaryotic names with standing in nomenclature (Bacterio.net), 20 years on. *Int. J. Syst. Evol. Microbiol.* 68, 1825–1829. doi:10.1099/ijsem.0.002786.

Paul, J. H., Sullivan, M. B., Segall, A. M., and Rohwer, F. (2002). Marine phage genomics. in *Comparative Biochemistry and Physiology - B Biochemistry and Molecular Biology*, 463–476. doi:10.1016/S1096-4959(02)00168-9.

Paulsen, I. T., Press, C. M., Ravel, J., Kobayashi, D. Y., Myers, G. S. A., Mavrodi, D. V, et al. (2005). Complete genome sequence of the plant commensal Pseudomonas fluorescens Pf-5. *Nat. Biotechnol.* 23, 873–878. doi:10.1038/nbt1110.

Pieterse, A., Malan, A. P., and Ross, J. L. (2017). Nematodes that associate with terrestrial molluscs as definitive hosts, including Phasmarhabditis hermaphrodita (Rhabditida: Rhabditidae) and its development as a biological molluscicide. *J. Helminthol.* 91, 517–527. doi:10.1017/S0022149X16000572.

Pires, D. P., Vilas Boas, D., Sillankorva, S., and Azeredo, J. (2015). Phage Therapy: a Step Forward in the Treatment of Pseudomonas aeruginosa Infections. *J. Virol.* 89, 7449–56. doi:10.1128/JVI.00385-15.

Pope, W. H., Bowman, C. A., Russell, D. A., Jacobs-Sera, D., Asai, D. J., Cresawn, S. G., et al. (2015). Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. *Elife* 4, e06416. doi:10.7554/eLife.06416.

Pope, W. H., Jacobs-Sera, D., Russell, D. A., Peebles, C. L., Al-Atrache, Z., Alcoser, T. A., et al. (2011). Expanding the Diversity of Mycobacteriophages: Insights into Genome Architecture and Evolution. *PLoS One* 6, e16329. doi:10.1371/journal.pone.0016329.

Poretsky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D., and Konstantinidis, K. T. (2014). Strengths and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics. *PLoS One* 9, e93827. doi:10.1371/journal.pone.0093827.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One* 5. doi:10.1371/journal.pone.0009490.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* 41. doi:10.1093/nar/gks1219.

Raaijmakers, J. M., Bonsall, R. F., and Weller, D. M. (1999). Effect of population density of Pseudomonas fluorescens on production of 2,4-diacetylphloroglucinol in the rhizosphere of wheat. *Phytopathology* 89, 470–475. doi:10.1094/PHYTO.1999.89.6.470.

Raaijmakers, J. M., De Bruijn, I., Nybroe, O., and Ongena, M. (2010). Natural functions of lipopeptides from Bacillus and Pseudomonas: more than surfactants and antibiotics. *FEMS Microbiol. Rev.* 34, 1037–62. doi:10.1111/j.1574-6976.2010.00221.x.

Rae, R. G., Tourna, M., and Wilson, M. J. (2010). The slug parasitic nematode Phasmarhabditis hermaphrodita associates with complex and variable bacterial assemblages that do not affect its virulence. *J. Invertebr. Pathol.* 104, 222–226. doi:10.1016/j.jip.2010.04.008.

Rae, R., Verdun, C., Grewal, P. S., Robertson, J. F., and Wilson, M. J. (2007). Biological control of terrestrial molluscs using Phasmarhabditis hermaphrodita - Progress and prospects. *Pest Manag. Sci.* 63, 1153–1164. doi:10.1002/ps.1424.

Ramette, A., Frapolli, M., Saux, M. F. Le, Gruffaz, C., Meyer, J. M., Défago, G., et al. (2011). Pseudomonas protegens sp. nov., widespread plant-protecting bacteria producing the biocontrol compounds 2,4-diacetylphloroglucinol and pyoluteorin. *Syst. Appl. Microbiol.* 34, 180–188. doi:10.1016/j.syapm.2010.10.005.

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

Ria Bozal, N., Jesú S Montes, M., and Mercadé, E. (2007). Pseudomonas guineae sp. nov., a novel psychrotolerant bacterium from an Antarctic environment. *Int. J. Syst. Evol. Microbiol.* 57, 2609–2612. doi:10.1099/ijs.0.65141-0.

Rocha, E. P. C., Smith, J. M., Hurst, L. D., Holden, M. T. G., Cooper, J. E., Smith, N. H., et al. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J. Theor. Biol.* 239, 226–235. doi:10.1016/j.jtbi.2005.08.037.

Rohwer, F., Youle, M., and Maughan, H. (2014). *Life in Our Phage World*. Wholon Available at: https://books.google.com/books/about/Life_in_Our_Phage_World.html?id=g4HTrQEACAAJ&pgis=1 [Accessed October 14, 2015].

Rolain, J. M., Fancello, L., Desnues, C., and Raoult, D. (2011). Bacteriophages as

vehicles of the resistome in cystic fibrosis. *J. Antimicrob. Chemother.* 66, 2444–2447. doi:10.1093/jac/dkr318.

Romanenko, L. A., Uchino, M., Tebo, B. M., Tanaka, N., Frolova, G. M., Mikhailov, V. V, et al. (2008). Pseudomonas marincola sp. nov., isolated from marine environments. *Int. J. Syst. Evol. Microbiol.* 58, 706–710. doi:10.1099/ijs.0.65406-0.

Roossinck, M. (2011a). Changes in Population Dynamics in Mutualistic versus Pathogenic Viruses. *Viruses* 3, 12–19. doi:10.3390/v3010012.

Roossinck, M. J. (2011b). The good viruses: Viral mutualistic symbioses. *Nat. Rev. Microbiol.* 9, 99–108. doi:10.1038/nrmicro2491.

Rose, R., Constantinides, B., Tapinos, A., Robertson, D. L., and Prosperi, M. (2016). Challenges in the analysis of viral metagenomes. *Virus Evol.* 2, vew022. doi:10.1093/ve/vew022.

Rosenblum, E. D., and Tyrone, S. (1964). SEROLOGY, DENSITY, AND MORPHOLOGY OF STAPHYLOCOCCAL PHAGES Downloaded from. Available at: http://jb.asm.org/ [Accessed November 26, 2019].

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475, 348–352. doi:10.1038/nature10242.

Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., et al. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* 537, 689–693. doi:10.1038/nature19366.

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

Roux, S., Krupovic, M., Daly, R. A., Borges, A. L., Nayfach, S., Schulz, F., et al. (2019). Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nat. Microbiol.* 4, 1895–1906. doi:10.1038/s41564-019-0510-x.

Ryan, K. J., George, C., York, N., San, C., Athens, F., and Madrid, L. (2014). *MEDICAL MICROBIOLOGY*. Available at: www.mhprofessional.com. [Accessed November 26, 2019].

Salzberg, S. L., Delcher, A. L., Kasif, S., and White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* 26, 544–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9421513 [Accessed October 23, 2017].

Sano, E., Carlson, S., Wegley, L., and Rohwer, F. (2004). Movement of Viruses between Biomes. *Appl. Environ. Microbiol.* 70, 5842–5846. doi:10.1128/AEM.70.10.5842-5846.2004.

Schippers, B., Bakker, A. W., and Bakker, P. A. H. M. (1987). Interactions of Deleterious and Beneficial Rhizosphere Microorganisms and the Effect of Cropping Practices. *Annu. Rev. Phytopathol.* 25, 339–358. doi:10.1146/annurev.py.25.090187.002011.

Schneider, A. (1859). Über eine Nematodenlarvae und gewisse Verschiedenheiten in den Geschlechtsorganen der Nematoden. Zeitschrift für wissenschaftliche Zoologie. 176–178.

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi:10.1093/bioinformatics/btu153.

Sela, I., Wolf, Y. I., and Koonin, E. V. (2016). Theory of prokaryotic genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* 113, 11399–11407. doi:10.1073/pnas.1614083113.

Shi, W., Xie, S., Chen, X., Sun, S., Zhou, X., Liu, L., et al. (2013). Comparative genomic analysis of the microbiome [corrected] of herbivorous insects reveals eco-environmental adaptations: biotechnology applications. *PLoS Genet.* 9, e1003131. doi:10.1371/journal.pgen.1003131.

Shkoporov, A. N., Ryan, F. J., Draper, L. A., Forde, A., Stockdale, S. R., Daly, K. M., et al. (2018). Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* 6, 68. doi:10.1186/s40168-018-0446-z.

Sicard, M., Ferdy, J.-B., Pagès, S., Le Brun, N., Godelle, B., Boemare, N., et al. (2004). When mutualists are pathogens: an experimental study of the symbioses between Steinernema (entomopathogenic nematodes) and Xenorhabdus (bacteria). *J. Evol. Biol.* 17, 985–993. doi:10.1111/j.1420-9101.2004.00748.x.

Silby, M. W., Winstanley, C., Godfrey, S. A. C., Levy, S. B., and Jackson, R. W. (2011). Pseudomonas genomes: Diverse and adaptable. *FEMS Microbiol. Rev.* 35, 652–680. doi:10.1111/j.1574-6976.2011.00269.x.

Sillankorva, S., Neubauer, P., and Azeredo, J. (2008a). Isolation and characterization of a T7-like lytic phage for Pseudomonas fluorescens. *BMC Biotechnol.* 8, 80. doi:10.1186/1472-6750-8-80.

Sillankorva, S., Neubauer, P., and Azeredo, J. (2008b). Pseudomonas fluorescens biofilms subjected to phage phiIBB-PF7A. *BMC Biotechnol.* 8, 79. doi:10.1186/1472-6750-8-79.

Simpson, E. H. (1949). Measurement of diversity. *Nature* 163, 688. doi:10.1038/163688a0.

Smart, G. C. (1995). JOURNAL OF NEMATOLOGY DECEMBER 1995 NUMBER 4S Supplement to the.

Spellerberg, I. F., and Fedor, P. J. (2003). A tribute to Claude Shannon (1916-2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon-Wiener' Index. *Glob. Ecol. Biogeogr.* 12, 177–179. doi:10.1046/j.1466-822X.2003.00015.x.

Spiers, A. J., Buckling, A., and Rainey, P. B. (2000). The causes of Pseudomonas diversity. *Microbiology* 146, 2345–2350. doi:10.1099/00221287-146-10-2345.

Stackebrandt, E., Goebel, B. M., Wayne, G., Brenner, D. J., Colwell, R. R., D Grimont, P. A., et al. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16s rRNA Sequence Analysis in the Present Species Definition in Bacteriology. Available at: www.microbiologyresearch.org [Accessed November 26, 2019].

Stanier, R. Y., Palleroni, N. J., and Doudoroff, M. (1966). The aerobic pseudomonads: a taxonomic study. *J. Gen. Microbiol.* 43, 159–271. doi:10.1099/00221287-43-2-159.

Stephenson, M. (1939). Bacterial Metabolism. *Soil Sci.* 48, 356. doi:10.1097/00010694-193910000-00024.

Stiernagle, T. (2006). Maintenance of C. elegans. *WormBook*, 1–11. doi:10.1895/wormbook.1.101.1.

Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., et al. (2000). Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. *Nature* 406, 959–964. doi:10.1038/35023079.

Sullivan, M. B., Lindell, D., Lee, J. A., Thompson, L. R., Bielawski, J. P., and Chisholm, S. W. (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol.* 4, 1344–1357.

doi:10.1371/journal.pbio.0040234.
Sulston, J., and Hodgkin, J. (1988). *The nematode Caenorhabditis elegans*. Cold Spring Harbor Laboratory Press.
Tan, L., and Grewal, P. S. (2001a). Infection behavior of the rhabditid nematode Phasmarhabditis hermaphrodita to the grey garden slug Deroceras reticulatum. *J. Parasitol.* 87, 1349–54. doi:10.1645/0022-3395(2001)087[1349:IBOTRN]2.0.CO;2.
Tan, L., and Grewal, P. S. (2001b). Pathogenicity of Moraxella osloensis, a bacterium associated with the nematode Phasmarhabditis hermaphrodita, to the slug Deroceras reticulatum. *Appl. Environ. Microbiol.* 67, 5010–6. doi:10.1128/AEM.67.11.5010-5016.2001.
Thurber, R. V, Haynes, M., Breitbart, M., Wegley, L., and Rohwer, F. (2009). Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* 4, 470–83. doi:10.1038/nprot.2009.10.
Tringe, S. G., von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., et al. (2005). Comparative metagenomics of microbial communities. *Science* 308, 554–7. doi:10.1126/science.1107851.
Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., et al. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* 12, 902–903. doi:10.1038/nmeth.3589.
Turner, D., Ackermann, H. W., Kropinski, A. M., Lavigne, R., Sutton, J. M., and Reynolds, D. M. (2018). Comparative analysis of 37 Acinetobacter bacteriophages. *Viruses* 10. doi:10.3390/v10010005.
Uhr, G. T., Dohnalová, L., and Thaiss, C. A. (2019). The Dimension of Time in Host-Microbiome Interactions. *mSystems* 4. doi:10.1128/msystems.00216-18.
Urakami, T., Araki, H., Oyanagi, H., Suzuki, K. I., and Komagata, K. (1992). Transfer of Pseudomonas aminovorans (den Dooren de Jong 1926) to Aminobacter gen. nov. as Aminobacter aminovorans comb. nov. and description of Aminobacter aganoensis sp. nov. and Aminobacter niigataensis sp. nov. *Int. J. Syst. Bacteriol.* 42, 84–92. doi:10.1099/00207713-42-1-84.
Van de Peer, Y., Chapelle, S., and De Wachter, R. (1996). A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucleic Acids Res.* 24, 3381–3391. doi:10.1093/nar/24.17.3381.
von Mering, C., Hugenholtz, P., Raes, J., Tringe, S. G., Doerks, T., Jensen, L. J., et al. (2007). Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* 315, 1126–30. doi:10.1126/science.1133420.
Wang, I. N. (2006). Lysis timing and bacteriophage fitness. *Genetics* 172, 17–26. doi:10.1534/genetics.105.045922.
Warnecke, F., Luginbühl, P., Ivanova, N., Ghassemian, M., Richardson, T. H., Stege, J. T., et al. (2007). Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* 450, 560–565. doi:10.1038/nature06269.
Weigele, P. R., Pope, W. H., Pedulla, M. L., Houtz, J. M., Smith, A. L., Conway, J. F., et al. (2007). Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and *Synechococcus*. *Environ. Microbiol.* 9, 1675–1695. doi:10.1111/j.1462-2920.2007.01285.x.
Weinbauer, M. G., and Rassoulzadegan, F. (2003). Are viruses driving microbial diversification and diversity? *Environ. Microbiol.* 6, 1–11. doi:10.1046/j.1462-

2920.2003.00539.x.
Weller, D. M. (1988). Biological Control of Soilborne Plant Pathogens in the Rhizosphere with Bacteria. *Annu. Rev. Phytopathol.* 26, 379–407. doi:10.1146/annurev.py.26.090188.002115.
Weller, D. M. (2007). Pseudomonas biocontrol agents of soilborne pathogens: Looking back over 30 years. in *Phytopathology*, 250–256. doi:10.1094/PHYTO-97-2-0250.
Wen, A., Fegan, M., Hayward, C., Chakraborty, S., and Sly, L. I. (1999). Phylogenetic relationships among members of the Comamonadaceae, and description of Delftia acidovorans (den Dooren de Jong 1926 and Tamaoka et al. 1987) gen. nov., comb. nov. *Int. J. Syst. Bacteriol.* 49, 567–576. doi:10.1099/00207713-49-2-567.
Wilson, M. J., Glen, D. M., and George, S. K. (1993). The Rhabditid Nematode Phasmarhabditis hermaphrodita as a Potential Biological Control Agent for Slugs. *Biocontrol Sci. Technol.* 3, 503–511. doi:10.1080/09583159309355306.
Wilson, M. J., Glen, D. M., George, S. K., and Pearce, J. D. (1995a). Selection of a bacterium for the mass production of Phasmarhabditis hermaphrodita (Nematoda : Rhabditidae) as a biocontrol agent for slugs.
Wilson, M. J., Glen, D. M., Hughes, L. A., Pearce, J. D., and Rodgers, P. B. (1994). Laboratory tests of the potential of entomopathogenic nematodes for the control of field slugs (Deroceras reticulatum). *J. Invertebr. Pathol.* 64, 182–187. doi:10.1016/S0022-2011(94)90100-7.
Wilson, M. J., Glen, D. M., Pearee, J. D., and Rodgers, P. B. (1995b). Monoxenic culture of the slug parasite Phasmarhabditis hermaphrodita (Nematoda : Rhabditidae) with different bacteria in liquid and solid phase. *Fundam. appl. Nematol* 18, 159–166. Available at: http://horizon.documentation.ird.fr/exl-doc/pleins_textes/fan/42721.pdf [Accessed January 19, 2018].
Winget, D. M., and Wommack, K. E. (2008). Randomly amplified polymorphic DNA PCR as a tool for assessment of marine viral richness. *Appl. Environ. Microbiol.* 74, 2612–2618. doi:10.1128/AEM.02829-07.
Winsor, G. L., Griffiths, E. J., Lo, R., Dhillon, B. K., Shay, J. A., and Brinkman, F. S. L. (2016a). Enhanced annotations and features for comparing thousands of *Pseudomonas* genomes in the Pseudomonas genome database. *Nucleic Acids Res.* 44, D646–D653. doi:10.1093/nar/gkv1227.
Winsor, G. L., Griffiths, E. J., Lo, R., Dhillon, B. K., Shay, J. A., and Brinkman, F. S. L. (2016b). Enhanced annotations and features for comparing thousands of Pseudomonasgenomes in the Pseudomonas genome database. *Nucleic Acids Res.* 44, D646–D653. doi:10.1093/nar/gkv1227.
Winsor, G. L., Lam, D. K. W., Fleming, L., Lo, R., Whiteside, M. D., Yu, N. Y., et al. (2011). Pseudomonas Genome Database: improved comparative analysis and population genomics capability for Pseudomonas genomes. *Nucleic Acids Res.* 39, D596–D600. doi:10.1093/nar/gkq869.
Woese, C. R., and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5088–5090. doi:10.1073/pnas.74.11.5088.
Woyke, T., Teeling, H., Ivanova, N. N., Huntemann, M., Richter, M., Gloeckner, F. O., et al. (2006). Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* 443, 950–955. doi:10.1038/nature05192.

Yabuuchi, E., Kosako, Y., Oyaizu, H., Yano, I., Hotta, H., Hashimoto, Y., et al. (1992). Proposal of Burkholderia gen. nov. and transfer of seven species of the genus Pseudomonas homology group II to the new genus, with the type species Burkholderia cepacia (Palleroni and Holmes 1981) comb. nov. *Microbiol. Immunol.* 36, 1251–75. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1283774 [Accessed November 19, 2019].

Yabuuchi, E., Yano, I., Hotta, H., Nishiuchi, Y., and Kosako, Y. (1995). Transfer of Two Burkholderia and an Alcaligenes Species to Ralstonia Gen. Nov.: Proposal of Ralstonia pickettii (Ralston, Palleroni and Doudoroff 1973) Comb. Nov., Ralstonia solanacearum (Smith 1896) Comb. Nov. and Ralstonia eutropha (Davis 1969) Comb. Nov. *Microbiol. Immunol.* 39, 897–904. doi:10.1111/j.1348-0421.1995.tb03275.x.

Yamamoto, S., Kasai, H., Arnold, D. L., Jackson, R. W., Vivian, A., and Harayama, S. (2000). Phylogeny of the genus Pseudomonas: Intrageneric structure reconstructed from the nucleotide sequences of gyrB and rpoD genes. *Microbiology* 146, 2385–2394. doi:10.1099/00221287-146-10-2385.

Yan, Y., Yang, J., Dou, Y., Chen, M., Ping, S., Peng, J., et al. (2008). Nitrogen fixation island and rhizosphere competence traits in the genome of root-associated Pseudomonas stutzeri A1501. Available at: www.pnas.org/cgi/content/full/ [Accessed November 20, 2019].

Yayan, J., Ghebremedhin, B., Rasche, K., Hoepelman, A., Carmeli, Y., and Yadav, R. (2015). Antibiotic Resistance of Pseudomonas aeruginosa in Pneumonia at a Single University Hospital Center in Germany over a 10-Year Period. *PLoS One* 10, e0139836. doi:10.1371/journal.pone.0139836.

Zhang, J., Jia, H., Li, J., Li, Y., Lu, M., and Hu, J. (2016). Molecular evolution and expression divergence of the Populus euphratica Hsf genes provide insight into the stress acclimation of desert poplar. *Sci. Rep.* 6, 30050. doi:10.1038/srep30050.

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., and Wishart, D. S. (2011). PHAST: A Fast Phage Search Tool. *Nucleic Acids Res.* 39, W347–W352. doi:10.1093/nar/gkr485.

Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* 38, e132–e132. doi:10.1093/nar/gkq275.

Zhu, Y., Shi, F., Tian, J., Liu, J., Chen, S., Xiang, M., et al. (2013). Effect of Soybean Monoculture on the Bacterial Communities Associated with Cysts of Heterodera glycines. *J. Nematol.* 45, 228–35. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24115788 [Accessed November 26, 2019].

APPENDICES

Appendix A: Supplementary material for Chapter 2

Appendix Table 1. Basic genome metrics of the 130 Pseudomonas phage sequences included.

Phage name	Cluster	Host species	No. of ORFs	Genome size (bp)	%ORF	ORF per kb	%GC	ICTV Family	tRNAs	tmRNA
B3	A1	*P. aeruginosa*	51	38,439	92.9	1.3	63.2	Siphoviridae	0	0
JBD18	A1	*P. aeruginosa*	51	39,014	95.5	1.3	63.4	Siphoviridae	0	0
JBD25	A1	*P. aeruginosa*	55	39,552	96.4	1.4	62.5	Siphoviridae	0	0
vB_PaeS_PM105	A1	*P. aeruginosa*	55	39,593	95.2	1.4	63.1	Siphoviridae	0	0
D3112	A2	*P. aeruginosa*	55	37,611	94.4	1.5	64.3	Siphoviridae	0	0
DMS3	A2	*P. aeruginosa*	53	36,415	94.4	1.5	64.3	Siphoviridae	0	0
H70	A2	*P. aeruginosa*	55	37,359	94.7	1.5	64.2	Siphoviridae	0	0
JBD24	A2	*P. aeruginosa*	55	37,095	95.7	1.5	64.2	Siphoviridae	0	0
JBD30	A2	*P. aeruginosa*	52	36,947	93.8	1.4	64.3	Siphoviridae	0	0
JBD5	A2	*P. aeruginosa*	58	37,740	96.6	1.5	64.3	Siphoviridae	0	0
JBD88a	A2	*P. aeruginosa*	54	36,429	95.7	1.5	64.0	Siphoviridae	0	0
JD024	A2	*P. aeruginosa*	57	37,380	95.8	1.5	64.2	Siphoviridae	0	0
LPB1	A2	*P. aeruginosa*	53	36,814	95.2	1.4	64.4	Siphoviridae	0	0
MP22	A2	*P. aeruginosa*	52	36,409	96.4	1.4	64.2	Siphoviridae	0	0
MP29	A2	*P. aeruginosa*	52	36,632	95.7	1.4	64.3	Siphoviridae	0	0
MP38	A2	*P. aeruginosa*	52	36,885	95.6	1.4	64.5	Siphoviridae	0	0
MP42	A2	*P. aeruginosa*	52	36,847	94.4	1.4	64.2	Siphoviridae	0	0
MP48	A2	*P. aeruginosa*	54	36,838	95.6	1.5	64.1	Siphoviridae	0	0
PA1/KOR/2010	A2	*P. aeruginosa*	51	34,553	92.0	1.5	64.8	Siphoviridae	0	0
vB_PaeS_PAO1_Ab30	A2	*P. aeruginosa*	54	37,238	93.6	1.5	64.1	Siphoviridae	0	0
PaMx11	B1	*P. aeruginosa*	86	59,878	93.6	1.4	64.5	Siphoviridae	0	0
vB_PaeS_PAO1_Ab18	B1	*P. aeruginosa*	76	56,537	93.7	1.3	63.5	Siphoviridae	0	0

M6	B2	*P. aeruginosa*	83	59,446	94.9	1.4	64.5	Siphoviridae	0	0
MP1412	B2	*P. aeruginosa*	82	61,167	94.9	1.3	64.3	Siphoviridae	0	0
PAE1	B2	*P. aeruginosa*	87	62,181	94.8	1.4	64.2	Siphoviridae	0	0
YuA	B2	*P. aeruginosa*	78	58,663	94.3	1.3	64.3	Siphoviridae	0	0
DL62	C1	*P. aeruginosa*	52	42,508	92.8	1.2	62.2	Podoviridae	0	0
LKD16	C1	*P. aeruginosa*	50	43,200	89.8	1.2	62.3	Podoviridae	0	0
LUZ19	C1	*P. aeruginosa*	55	43,548	93.0	1.3	62.3	Podoviridae	0	0
MPK6	C1	*P. aeruginosa*	53	42,957	92.7	1.2	62.3	Podoviridae	0	0
MPK7	C1	*P. aeruginosa*	54	42,874	93.3	1.3	62.1	Podoviridae	0	0
phikF77	C1	*P. aeruginosa*	54	43,152	92.7	1.3	62.9	Podoviridae	0	0
phiKMV	C1	*P. aeruginosa*	49	42,519	91.1	1.2	62.3	Podoviridae	0	0
PT2	C1	*P. aeruginosa*	50	42,961	89.9	1.2	62.2	Podoviridae	0	0
vB_Pae-TbilisiM32	C1	*P. aeruginosa*	53	42,966	91.9	1.2	62.3	Podoviridae	0	0
vB_PaeP_PAO1_Ab05	C1	*P. aeruginosa*	50	43,639	90.0	1.1	62.3	Podoviridae	0	0
vB_PaeP_PPA-ABTNL	C1	*P. aeruginosa*	52	43,227	91.5	1.2	62.4	Podoviridae	0	0
LKA1	C2	*P. aeruginosa*	56	41,593	92.1	1.3	60.9	Podoviridae	0	0
phi-2	C2	*P. fluorescens*	48	43,144	89.9	1.1	58.9	Podoviridae	0	0
KPP21	D1	*P. aeruginosa*	116	73,420	84.6	1.6	53.5	Podoviridae	0	0
LUZ7	D1	*P. aeruginosa*	112	74,901	94.7	1.5	53.2	Podoviridae	0	0
LIT1	D2	*P. aeruginosa*	94	72,544	93.8	1.3	55.0	Podoviridae	0	0
Pa2	D2	*P. aeruginosa*	93	73,008	94.1	1.3	54.9	Podoviridae	0	0
vB_PaeP_C2-10_Ab09	D2	*P. aeruginosa*	90	72,028	94.1	1.2	54.9	Podoviridae	0	0
YH6	D2	*P. aeruginosa*	92	73,050	93.4	1.3	54.9	Podoviridae	0	0
CHA_P1	E	*P. aeruginosa*	160	88,255	91.1	1.8	54.6	Myoviridae	3	0
KPP10	E	*P. aeruginosa*	157	88,322	91.0	1.8	54.8	Myoviridae	3	0
PAK_P3	E	*P. aeruginosa*	160	88,097	91.3	1.8	54.8	Myoviridae	3	0
PAK_P5	E	*P. aeruginosa*	160	88,135	91.5	1.8	54.7	Myoviridae	3	0
vB_PaeM_PAO1_Ab03	E	*P. aeruginosa*	154	86,246	90.6	1.8	54.7	Myoviridae	3	0
vB_PaeM_PS24	E	*P. aeruginosa*	153	84,583	91.6	1.8	54.7	Myoviridae	3	0

C11	F	*P. aeruginosa*	172	94,109	88.8	1.8	49.4	Myoviridae	15	0
JG004	F	*P. aeruginosa*	172	93,017	89.6	1.8	49.3	Myoviridae	15	0
K8	F	*P. aeruginosa*	172	93,879	89.2	1.8	49.4	Myoviridae	14	0
PAK_P1	F	*P. aeruginosa*	172	93,198	89.8	1.8	49.5	Myoviridae	14	0
PAK_P2	F	*P. aeruginosa*	173	92,495	89.9	1.9	49.3	Myoviridae	14	0
PAK_P4	F	*P. aeruginosa*	170	93,147	89.7	1.8	49.3	Myoviridae	14	0
PaoP5	F	*P. aeruginosa*	172	93,464	89.0	1.8	49.5	Myoviridae	14	0
PaP1	F	*P. aeruginosa*	170	91,715	89.9	1.9	49.4	Myoviridae	15	0
phiPsa374	F	*P. syringae*	169	97,906	88.7	1.7	47.7	Myoviridae	20	0
vB_PaeM_C2-10_Ab01	F	*P. aeruginosa*	169	92,777	89.2	1.8	49.3	Myoviridae	15	0
DL54	G	*P. aeruginosa*	70	45,673	92.3	1.5	52.4	Podoviridae	3	0
PaP3	G	*P. aeruginosa*	70	45,503	91.9	1.5	52.2	Podoviridae	4	0
vB_PaeP_p2-10_Or1	G	*P. aeruginosa*	65	44,030	92.5	1.5	52.0	Podoviridae	3	0
LUZ24	H1	*P. aeruginosa*	68	45,625	90.7	1.5	52.2	Podoviridae	3	0
PhiCHU	H1	*P. aeruginosa*	70	45,626	91.4	1.5	52.0	Podoviridae	3	0
phiIBB-PAA2	H1	*P. aeruginosa*	69	45,344	91.9	1.5	52.3	Podoviridae	3	0
TL	H1	*P. aeruginosa*	67	45,696	89.5	1.5	52.4	Podoviridae	3	0
vB_PaeP_C2-10_Ab22	H1	*P. aeruginosa*	69	45,808	90.9	1.5	52.4	Podoviridae	3	0
tf	H2	*P. putida*	66	46,271	89.9	1.4	53.2	Podoviridae	0	0
UFV-P2	H2	*P. fluorescens*	67	45,517	91.5	1.5	51.5	Podoviridae	0	0
14-1	I	*P. aeruginosa*	91	66,235	92.5	1.4	55.6	Myoviridae	0	0
DL60	I	*P. aeruginosa*	90	66,103	92.5	1.4	54.9	Myoviridae	0	0
DL68	I	*P. aeruginosa*	90	66,111	91.8	1.4	55.7	Myoviridae	0	0
F8	I	*P. aeruginosa*	93	66,015	91.9	1.4	54.9	Myoviridae	0	0
JG024	I	*P. aeruginosa*	93	66,275	92.4	1.4	55.6	Myoviridae	0	0
KPP12	I	*P. aeruginosa*	87	64,144	92.0	1.4	55.6	Myoviridae	0	0
LBL3	I	*P. aeruginosa*	86	64,427	92.0	1.3	55.5	Myoviridae	0	0
LMA2	I	*P. aeruginosa*	91	66,530	91.4	1.4	55.6	Myoviridae	0	0
NH-4	I	*P. aeruginosa*	93	66,116	91.3	1.4	55.5	Myoviridae	0	0

PB1	I	*P. aeruginosa*	92	65,764	92.1	1.4	54.9	Myoviridae	0	0
SN	I	*P. aeruginosa*	92	66,390	92.0	1.4	55.6	Myoviridae	0	0
vB_Pae_PS44	I	*P. aeruginosa*	96	68,871	90.4	1.4	55.2	Myoviridae	0	0
vB_PaeM_C1-14_Ab28	I	*P. aeruginosa*	93	66,181	92.4	1.4	54.9	Myoviridae	0	0
vB_PaeM_PAO1_Ab27	I	*P. aeruginosa*	94	66,299	91.8	1.4	55.7	Myoviridae	0	0
73	K	*P. aeruginosa*	54	42,999	93.4	1.3	53.6	Siphoviridae	0	0
PaMx42	K	*P. aeruginosa*	53	43,225	91.4	1.2	54.6	Siphoviridae	1	0
vB_Pae-Kakheti25	K	*P. aeruginosa*	56	42,844	94.2	1.3	53.8	Siphoviridae	0	0
vB_PaeS_SCH_Ab26	K	*P. aeruginosa*	54	43,056	92.2	1.3	53.4	Siphoviridae	0	0
gh-1	L1	*P. putida*	44	37,359	94.1	1.2	57.4	Podoviridae	0	0
phiPSA2	L1	*P. syringae*	49	40,472	93.7	1.2	57.4	Podoviridae	0	0
PPPL-1	L1	*P. syringae*	50	41,149	93.2	1.2	57.0	Podoviridae	0	0
Pf-10	L2	*P. fluorescens; P. putida*	46	39,167	93.6	1.2	56.5	Podoviridae	0	0
Phi-S1	L2	*P. fluorescens*	47	40,192	93.8	1.2	56.2	Podoviridae	0	1
phiIBB-PF7A	L2	*P. fluorescens*	51	40,973	93.2	1.2	56.3	Podoviridae	0	0
phi15	L3	*P. putida*	47	39,562	92.7	1.2	58.2	Podoviridae	0	0
PPpW-4	L3	*P. plecoglossicida*	51	41,386	92.3	1.2	56.8	Podoviridae	0	0
201phi2-1	M	*P. chlororaphis*	470	316,674	94.7	1.5	45.3	Myoviridae	1	0
phiKZ	M	*P. aeruginosa*	366	280,334	91.9	1.3	36.8	Myoviridae	7	0
PhiPA3	M	*P. aeruginosa*	414	309,208	93.0	1.3	47.7	Myoviridae	5	0
119X	NON	*P. aeruginosa*	53	43,365	93.1	1.2	44.9	Podoviridae	0	0
AF	NON	*P. putida*	61	42,689	95.6	1.4	58.4	Podoviridae	0	0
Bf7	NON	*P. tolaasii*	46	40,058	94.7	1.1	58.4	Podoviridae	0	0
D3	NON	*P. aeruginosa*	99	56,426	89.4	1.8	57.8	Siphoviridae	4	0
DL64	NON	*P. aeruginosa*	90	72,378	93.0	1.2	55.0	Podoviridae	0	0
EL	NON	*P. aeruginosa*	206	211,215	93.1	1.0	49.3	Myoviridae	1	0
F10	NON	*P. aeruginosa*	62	39,199	94.8	1.6	62.1	Myoviridae	0	0
F116	NON	*P. aeruginosa*	73	65,195	93.5	1.1	63.2	Podoviridae	0	0

KPP25	NON	*P. aeruginosa*	87	64,113	92.1	1.4	60.3	Podoviridae	0	0
Lu11	NON	*P. putida*	383	280,538	95.8	1.4	50.9	Myoviridae	0	0
OBP	NON	*P. fluorescens*	305	284,757	94.3	1.1	43.5	Myoviridae	4	0
PaBG	NON	*P. aeruginosa*	328	258,139	94.3	1.3	55.8	Myoviridae	5	0
PAJU2	NON	*P. aeruginosa*	78	46,872	90.8	1.7	56.3	Siphoviridae	0	0
PaMx28	NON	*P. aeruginosa*	74	55,108	92.4	1.3	66.4	Siphoviridae	0	0
PaP2	NON	*P. aeruginosa*	52	43,783	93.0	1.2	45.4	Podoviridae	0	0
PaP4	NON	*P. aeruginosa*	69	43,895	94.6	1.6	52.5	Podoviridae		0
Pf1	NON	*P. aeruginosa*	14	7,349	96.4	1.9	61.5	Inoviridae	0	0
Pf3	NON	*P. aeruginosa*	8	5,833	92.0	1.4	45.4	Inoviridae	0	0
phi_Pto-bp6g	NON	*P. tolaasii*	66	26,499	80.9	2.5	42.7	unclassified	0	0
phi297	NON	*P. aeruginosa*	62	49,135	88.9	1.3	62.1	Siphoviridae	0	0
phiCTX	NON	*P. aeruginosa*	41	35,580	84.9	1.2	62.6	Myoviridae	0	0
phiPSA1	NON	*P. syringae*	78	51,090	93.4	1.5	58.6	Siphoviridae	1	0
PP7	NON	*P. aeruginosa*	3	3,588	94.7	0.8	54.2	Leviviridae	0	0
PPpW-3	NON	*P. plecoglossicida*	63	43,564	93.0	1.4	61.1	Myoviridae	1	0
PRR1	NON	*P. aeruginosa*	3	3,573	89.0	0.8	49.2	Leviviridae	0	0
PS-1	NON	*P. sp.*	79	48,666	89.4	1.6	59.8	Siphoviridae	0	0
vB_PaeP_Tr60_Ab31	NON	*P. aeruginosa*	68	45,550	93.0	1.5	57.1	unclassified	1	0
vB_PaeS_PMG1	NON	*P. aeruginosa*	90	54,024	91.2	1.7	57.5	Siphoviridae	1	0
YH30	NON	*P. aeruginosa*	91	72,192	94.3	1.3	54.9	Podoviridae	0	0
YMC11/02/R656	NON	*P. aeruginosa*	108	60,919	86.9	1.8	58.7	Siphoviridae	1	0

Appendix Table 2. Pairwise nucleotide similarity between phages of each cluster.

Cluster A1

	JBD18	**JBD25**	**JBD24**
JBD18	100	59.78	52.56
JBD25		100	60.52
JBD24			100

Cluster A2

	JBD5	**JBD88a**	**JBD30**	**vB_ PaeS_ PAO1_ Ab30**	**vB_ PaeS_ PM105**	**B3**	**PA1/ KOR/ 2010**	**DMS3**	**H70**	**JD024**	**D3112**	**MP29**	**MP42**
JBD18	57.06	52.82	51.7	52.24	59.34	61.13	57.89	51.61	56.02	55.94	56.8	57.47	51.95
JBD25	60.59	59.44	59.01	59.91	47.9	57.52	61.41	59.53	60.91	59.56	60.78	59.97	58.92
JBD24	80.95	86.64	91.48	95.13	54.26	52.16	82.77	93.62	83.17	80.87	82.95	82.2	90.87
JBD5	100	84.81	81.4	80.34	53.69	52.48	96.38	80.41	92.95	95.65	91	94.31	80.2
JBD88a		100	86.6	85.86	51.92	52.08	84.76	85.14	82.18	84.52	80.01	83.3	85.36
JBD30			100	92.72	53.48	52.28	81.65	92.23	83.57	81.97	83.17	81.34	95.11
vB_PaeS_ PAO1_ Ab30				100	53.5	52.02	81.1	93.42	83.02	80.19	83.14	81.01	91.07
vB_ PaeS_ PM105					100	58.58	54.86	55.6	54.44	53.71	54.46	54.01	54.59
B3						100	53.49	52.04	52.38	52.74	52.39	53.07	51.95

PA1/ KOR/ 2010							100	80.7	92.89	96.02	92.47	98.21	81.12
DMS3								100	83.45	80.48	83.65	79.95	93.33
H70									100	93.78	92.98	91.62	82.43
JD024										100	91.33	94.8	79.8
D3112											100	93.37	84.28
MP29												100	82.61
MP42													100

Cluster B

	PaMx11	**YuA**	**M6**	**MP1412**	**PAE1**	**vB_PaeS_PAO1_Ab18**
PaMx11	100	54	54.58	53.84	54.51	82.63
YuA		100	94.13	94.27	93.95	54.5
M6			100	95.66	96.73	53.21
MP1412				100	95.19	54.36
PAE1					100	52.59
vB_PaeS_PAO1_Ab18						100

Cluster C

	DL62	**phiKMV**	**MPK7**	**PT2**	**phikF77**	**LKD16**	**vB_ PaeP_ PPA-ABTNL**	**MPK6**	**vB_ PaeP_ PAO1_ Ab05**	**vB_Pae-TbilisiM32**	**LUZ19**
DL62	100	91.23	92.12	92.52	88.87	86.9	91.07	92.69	91.82	93.55	93.67

phiKMV		100	92.46	97.8	88.63	86.27	92.1	91.14	90.38	95.56	91.14
MPK7			100	93.09	88.28	85.4	94.55	93.54	90.96	91.2	92.6
PT2				100	88.23	86.21	92.61	91.78	90.17	94.75	92.09
phikF77					100	86.34	88.15	90.46	88.73	90.31	90.85
LKD16						100	87.08	86.77	86.81	87.23	86.62
vB_ PaeP_ PPA-ABTNL							100	92.95	89.23	91.36	91.16
MPK6								100	90.9	92.68	95.75
vB_ PaeP_ PAO1_ Ab05									100	91.84	90.11
vB_ Pae-TbilisiM32										100	92.63
LUZ19											100

Cluster D

	KPP21	**Pa2**	**LUZ7**	**vB_PaeP_C2-10_Ab09**	**LIT1**	**YH6**
KPP21	100	58.9	85.78	59.63	58.83	58.46
Pa2		100	61.35	95.32	94.5	93.2
LUZ7			100	62.07	61.08	61.57
vB_PaeP_C2-10_Ab09				100	94.17	94.3
LIT1					100	94.54
YH6						100

Cluster E

	CHA_P1	KPP10	PAK_P3	PAK_P5	vB_PaeM_PAO1_Ab03	vB_PaeM_PS24
CHA_P1	100	95.72	95.13	96.42	91.78	94.95
KPP10		100	95.29	95.98	91.24	96.16
PAK_P3			100	95.22	91.73	95.3
PAK_P5				100	91.8	95.78
vB_PaeM_PAO1_Ab03					100	90.93
vB_PaeM_PS24						100

Cluster F

	C11	JG004	PAK_P1	PAK_P2	PAK_P4	PaoP5	K8	PaP1	phiPsa374	vB_PaeM_C210_Ab01
C11	100	93.66	89.72	93.57	93.52	93.08	93.4	91.04	51.81	91.9
JG004		100	88.51	94.46	95.33	91.84	91.93	92.04	52.01	92.65
PAK_P1			100	88.57	89.24	91.76	90.08	90.37	52.08	90.78
PAK_P2				100	96.06	90.81	92.85	92.89	52.7	94.39
PAK_P4					100	90.73	93.6	93.45	52.77	94.71
PaoP5						100	90.98	90.35	51.89	91.39
K8							100	94.8	51.99	92.16
PaP1								100	51.96	92.12
phiPsa374									100	53.27
vB_PaeM_C2-10_Ab01										100

Cluster G

	DL54	PaP3	vB_PaeP_p2-10_Or1
DL54	100	79.04	79.1
PaP3		100	96.2
vB_PaeP_p2-10_Or1			100

Cluster H

	UFV-P2	vB_PaeP_C2-10_Ab22	phiIBB-PAA2	TL	PhiCHU	LUZ24	tf
UFV-P2	100	58.97	58.41	59.02	58.17	58.6	55.32
vB_PaeP_C2-10_Ab22		100	90.25	93.2	79.13	92.37	56.1
phiIBB-PAA2			100	91.93	77.83	94.06	55.98
TL				100	79.27	94.18	56.15
PhiCHU					100	78.94	55.43
LUZ24						100	56.57
tf							100

Cluster I

	LBL3	PB1	DL60	DL68	LMA2	vB_Pae_PS44	KPP12	SN	vB_PaeM_C1-14_Ab28	vB_PaeM_PAO1_Ab27	JG024	14-1	F8	NH-4
LBL3	100	91.29	44.72	45.76	92.32	93.03	92.89	92.76	91.32	93.07	92.98	92.92	91.13	92.12
PB1		100	43.69	44.62	87.79	90.36	88.85	87.71	94.29	88.6	87.92	89.35	94.69	87.99
DL60			100	80.57	44.93	44.86	44.92	45	43.57	45.07	45.02	45.03	43.88	45.02

DL68				100	46.2	46.11	46.24	46.25	44.44	46.27	46.23	46.2	44.78	46.27
LMA2					100	93.14	95.71	94.59	87.21	96.31	93.9	93.81	89.07	96.74
vB_Pae_ PS44						100	93.51	94.87	89.78	93.42	93.49	94.34	90.75	93.32
KPP12							100	94.91	88.54	96.88	94.67	94.49	89.25	95.52
SN								100	87.37	94.9	95.58	95.42	89.06	95.02
vB_ PaeM_ C1-14_ Ab28									100	88.19	87.62	88.42	94.94	87.68
vB_ PaeM_ PAO1_ Ab27										100	94.21	94.59	89.66	95.71
JG024											100	96.21	88.65	95.27
14-1												100	89.54	94.73
F8													100	89.04
NH-4														100

Cluster K

	73	**PaMx42**	**vB_Pae-Kakheti25**	**vB_PaeS_ SCH_Ab26**
73	100	75.84	95.38	94.48
PaMx42		100	75.66	76.04
vB_Pae-Kakheti25			100	94.18
vB_PaeS_SCH_Ab26				100

Cluster L

	gh-1	phiPSA2	PPPL-1	Phi-S1	phiIBB-PF7A	Pf-10	phi15	PPpW-4
gh-1	100	95.4	74.38	61.62	61.91	61.72	59.83	59.29
phiPSA2		100	72.78	60.33	60.95	60.47	58.23	57.86
PPPL-1			100	59.44	60.23	59.75	57.7	57.79
Phi-S1				100	81.12	89.57	64.35	64.98
phiIBB-PF7A					100	85.03	64.75	65.23
Pf-10						100	64.84	65.39
phi15							100	66.96
PPpW-4								100

Cluster M

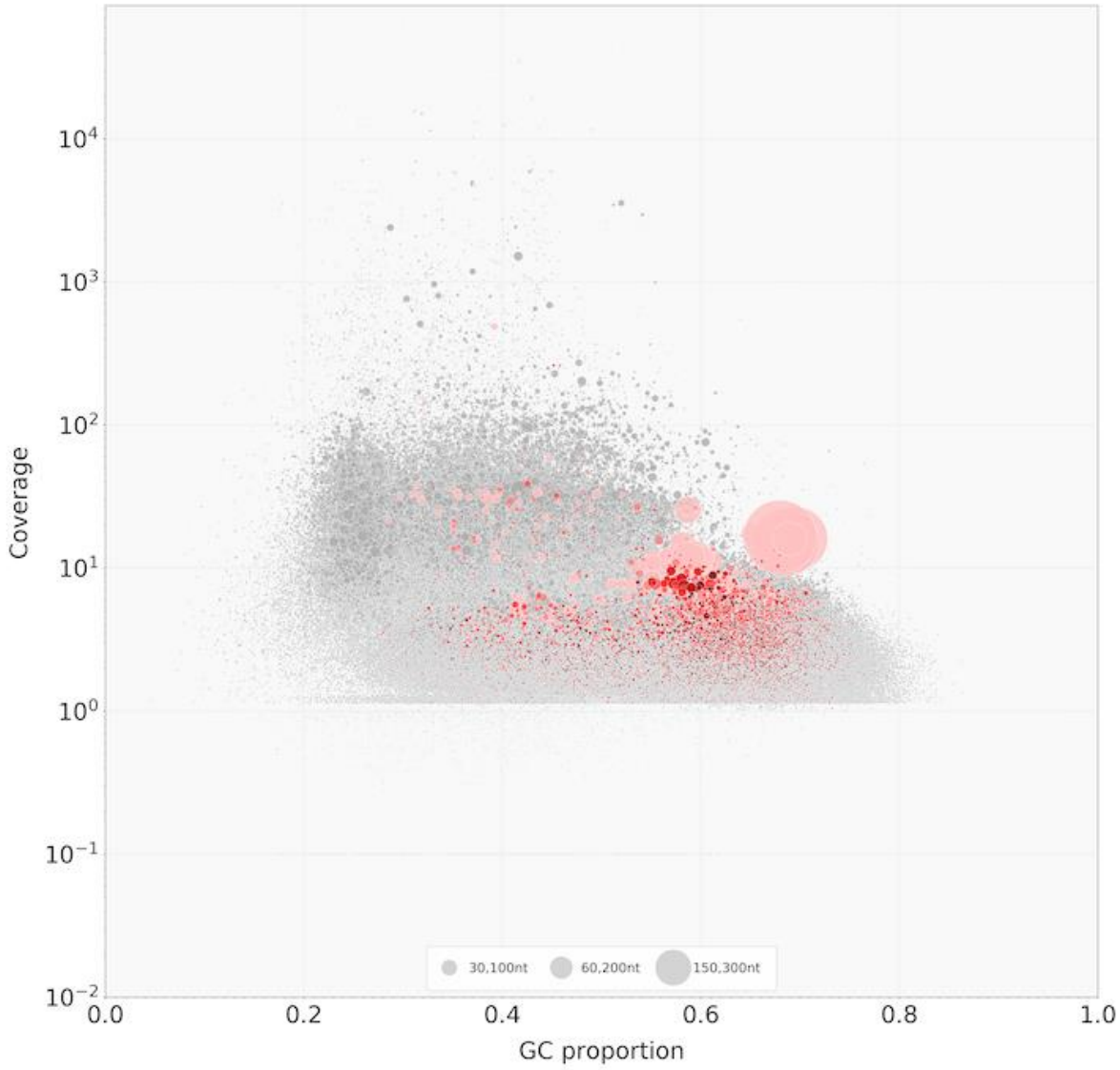
	phiKZ	PhiPA3	201phi2-1
phiKZ	100	50.89	52.67
PhiPA3		100	51.89
201phi2-1			100

Appendix B: Supplementary material for Chapter 3

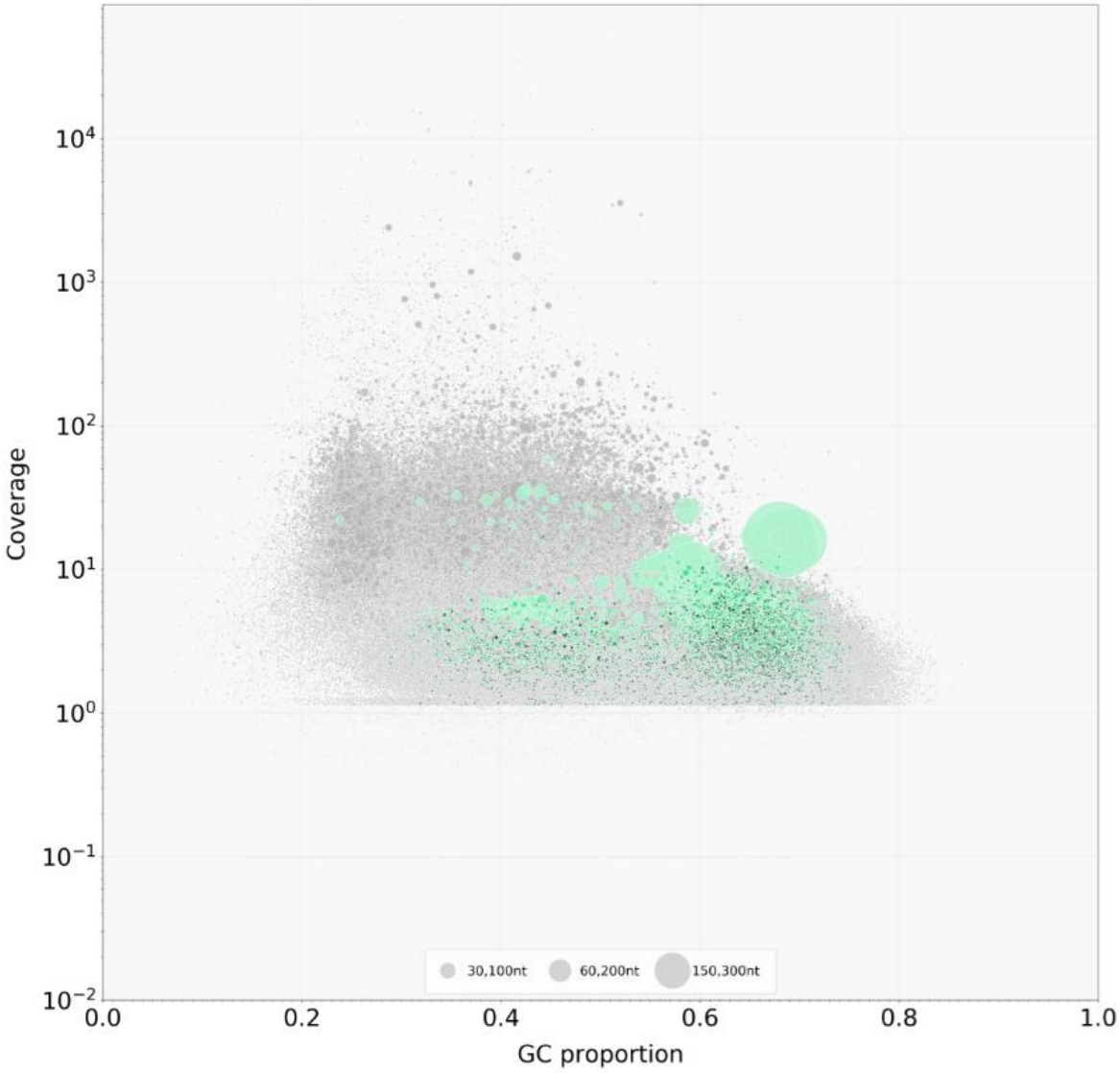
Appendix Table 3. Number of pairwise differences between contigs with length 5,461 bp and high coverage in the eight data sets.

	Hg_Al_NODE_3798	Gp_25_NODE_3025	Hg_Aud_NODE_4214	Gp_GH_NODE_3750	Hg_Pet_NODE_1496	Gp_258_NODE_3062	Gp_26_NODE_7217	Hg_War_NODE_3465
Hg_Al_NODE_3798		60	157	157	82	82	82	157
Gp_25_NODE_3025			155	155	80	80	80	155
Hg_Aud_NODE_4214				150	75	75	75	150
Gp_GH_NODE_3750					75	75	75	150
Hg_Pet_NODE_1496						0	0	75
Gp_258_NODE_3062							0	75
Gp_26_NODE_7217								75
Hg_War_NODE_3465								

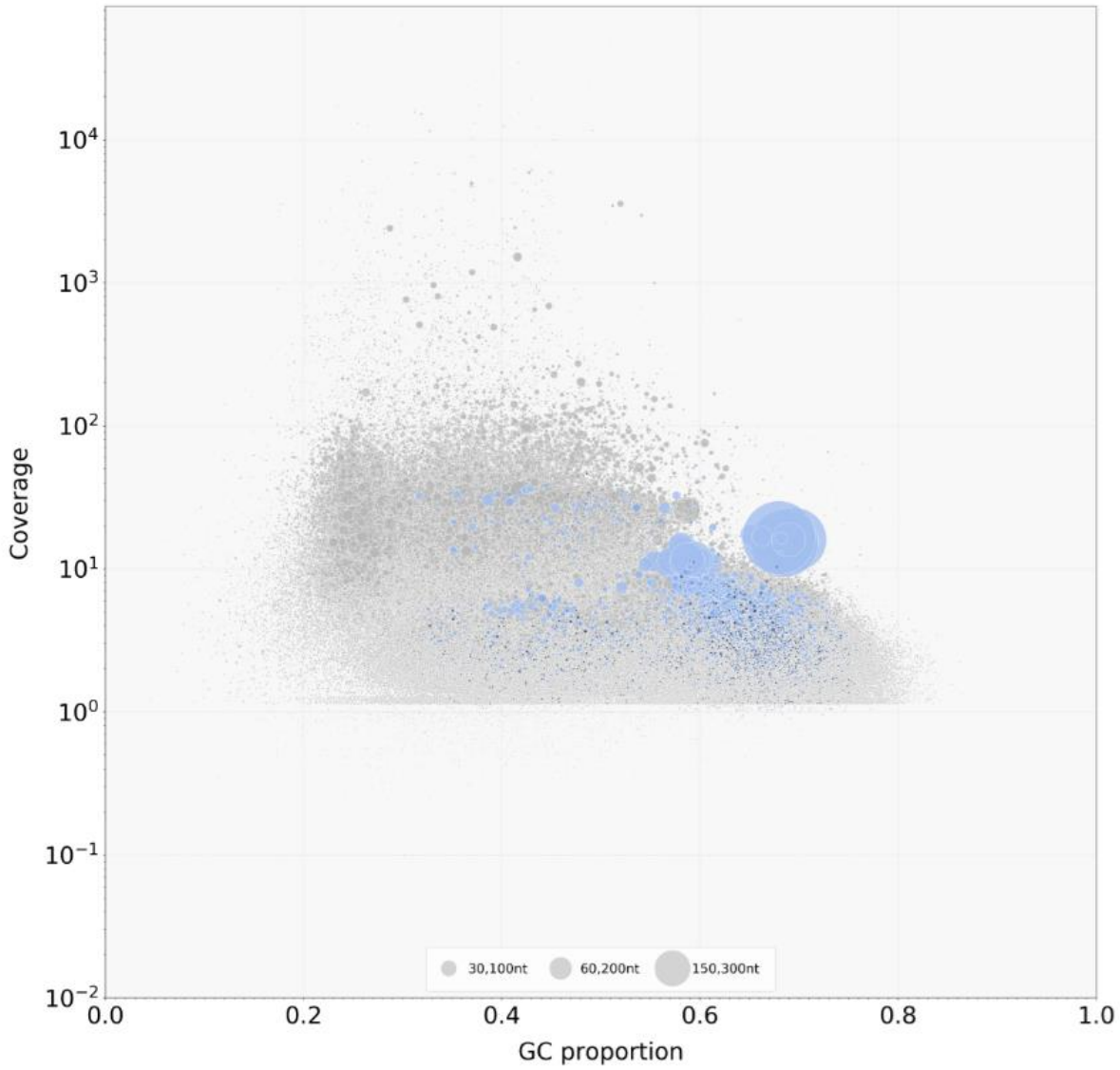
A


Coverage
10^4
10^3
10^2
10^1
10^0
10^{-1}
10^{-2}
0.0
0.2
0.4
0.6
0.8
1.0
GC proportion
30,100nt
60,200nt
150,300nt

B

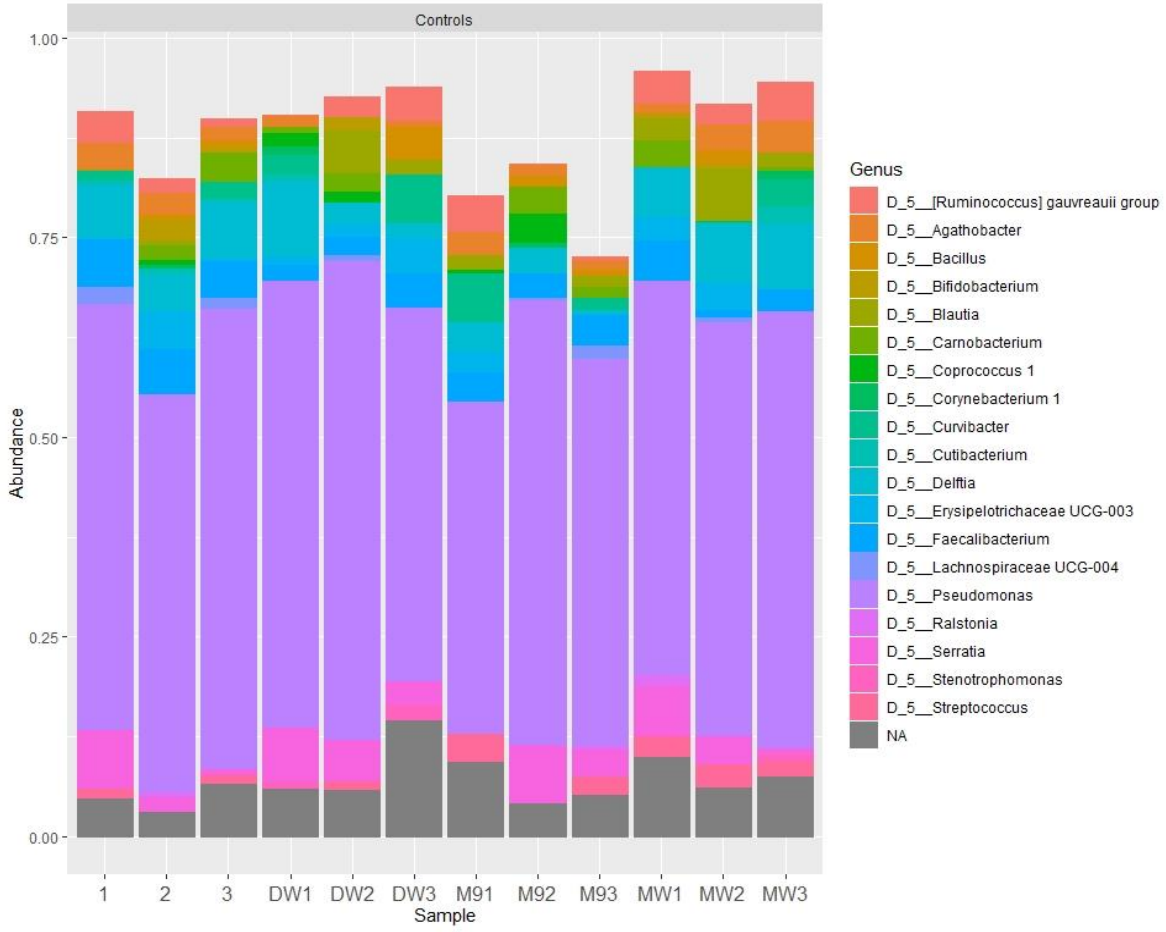

Coverage
10^4
10^3
10^2
10^1
10^0
10^{-1}
10^{-2}
0.0
0.2
0.4
0.6
0.8
1.0
GC proportion
30,100nt
60,200nt
150,300nt

C



Appendix Figure 1. TBLASTX hits to *Pseudomonas* phage (A), *Bacillus* phage (B), and Myobacteriophage (C) databases. Each dot represents a contig, and the size of the dot correlates to contig length. Contigs are shaded based on the degree of matching to phage database - the darker the color, the higher the percentage of that contig match to phage DNA.

Appendix C: Supplementary material for Chapter 4



Appendix Figure 2. Composition of ASVs contaminants removed from ASV table.

Appendix Table 4. The number of quality controlled read counts in decontaminated samples.

Sample	No. of reads	Total reads in each combination	Median value in each combination	Mean value in each combination
EA1	604,813	4,087,108	490,705	681,185
EA2	1,610,879			
EA3	1,091,611			
EB1	224,670			
EB2	376,596			
EB3	178,539			
UA1	153,427	2,572,446	231,484	428,741
UA2	153,680			
UA3	100,672			
UB1	309,288			
UB2	1,005,733			
UB3	849,646			
PA1	641,501	3,762,641	453,176	627,107
PA2	1,198,645			
PA3	1,231,796			
PB1	264,851			
PB2	250,356			
PB3	175,492			

Appendix Table 5: **(A)**. Phylum composition and relative abundance of the bacterial community associated with *P. hermaphrodita* before infection (samples BC-PreInf). Phyla were sorted from highest to lowest mean abundance in samples.

Phylum	meanRA	sdRA	minRA	maxRA
D_1__Bacteroidetes	0.022711	0.05372	1.30E-05	0.216386
D_1__Proteobacteria	0.0099	0.025887	1.99E-05	0.195276
D_1__Gemmatimonadetes	0.002883	0.004853	7.81E-05	0.010142
D_1__Firmicutes	0.002073	0.003461	1.95E-05	0.012119
D_1__Chloroflexi	0.001788	NA	0.001788	0.001788
D_1__Planctomycetes	0.001664	0.001337	3.90E-05	0.003268
D_1__Acidobacteria	0.001592	0.000449	0.001222	0.002245
NA	0.001585	0.002498	1.30E-05	0.006981
D_1__Actinobacteria	0.001375	0.001653	3.26E-05	0.006019
D_1__Verrucomicrobia	0.001323	0.001488	0.000169	0.003447
D_1__WS2	0.000864	NA	0.000864	0.000864
D_1__Armatimonadetes	0.000808	NA	0.000808	0.000808
D_1__Patescibacteria	0.000505	2.30E-05	0.000489	0.000521
D_1__Cyanobacteria	0.000192	0.000169	5.96E-05	0.000437
D_1__Deinococcus-Thermus	4.97E-05	NA	4.97E-05	4.97E-05

B. Genus composition of the natural bacterial community associated with *P. hermaphrodita.* Genera were sorted from the highest to lowest mean abundance in samples.

Genus	meanRA	sdRA	minRA	maxRA
D_5__Pseudochrobactrum	0.069449	0.085976	0.002215	0.195276
D_5__Flavobacterium	0.045221	0.077108	1.30E-05	0.216386
D_5__Raoultella	0.01917	0.020054	0.000404	0.062018
D_5__Pseudomonas	0.018958	0.027937	4.55E-05	0.110072
D_5__Brevundimonas	0.010851	0.011081	0.000467	0.029283
D_5__Sphingobacterium	0.007877	0.00789	0.000235	0.026679
D_5__Pedobacter	0.005992	0.002071	0.003795	0.009097
D_5__wb1-P19	0.005553	NA	0.005553	0.005553
D_5__Ochrobactrum	0.005453	0.003768	0.001662	0.011023
D_5__Bauldia	0.005066	NA	0.005066	0.005066
D_5__Paenibacillus	0.004858	0.004821	0.000469	0.012119
D_5__Achromobacter	0.004599	0.002172	0.001597	0.00691
D_5__Kaistia	0.004247	0.004221	0.00058	0.00886
NA	0.003932	0.007775	1.30E-05	0.033866
D_5__Sphingomonas	0.003857	0.003819	0.000111	0.009262
D_5__ADurb.Bin063-1	0.003447	NA	0.003447	0.003447
D_5__Microbacterium	0.003383	0.002822	0.000372	0.006019
D_5__Rhodococcus	0.003253	0.002015	0.000111	0.00509
D_5__Stenotrophomonas	0.003141	0.004696	1.99E-05	0.012148
D_5__Allorhizobium-Neorhizobium-Pararhizobium-Rhizobium	0.003072	0.001925	0.000891	0.005275
D_5__Rhodoferax	0.002959	NA	0.002959	0.002959

D_5__uncultured	0.002167	0.002845	3.90E-05	0.010142
D_5__Comamonas	0.002157	0.001104	0.001356	0.003416
D_5__Anaerococcus	0.002007	NA	0.002007	0.002007
D_5__Bacteroides	0.001973	0.002716	5.21E-05	0.003894
D_5__Marmoricola	0.001828	0.001888	0.000104	0.00449
D_5__uncultured soil bacterium	0.001788	NA	0.001788	0.001788
D_5__Herminiimonas	0.001429	0.000767	0.000886	0.001972
D_5__Chryseobacterium	0.001421	NA	0.001421	0.001421
D_5__Caulobacter	0.001351	0.001301	5.96E-05	0.003084
D_5__Propionivibrio	0.001297	NA	0.001297	0.001297
Ambiguous_taxa	0.001281	0.000945	0.000613	0.001949
D_5__Oikopleura dioica	0.001245	NA	0.001245	0.001245
D_5__RB41	0.001222	NA	0.001222	0.001222
D_5__Porphyrobacter	0.001162	NA	0.001162	0.001162
D_5__uncultured bacterium	0.001155	0.001066	0.000489	0.003258
D_5__Terrimonas	0.001108	NA	0.001108	0.001108
D_5__Nakamurella	0.001108	NA	0.001108	0.001108
D_5__Nesterenkonia	0.001088	NA	0.001088	0.001088
D_5__Rathayibacter	0.001036	NA	0.001036	0.001036
D_5__Crossiella	0.000939	NA	0.000939	0.000939
D_5__Leucobacter	0.000917	0.000763	0.00013	0.002781
D_5__Lachnoclostridium	0.000867	NA	0.000867	0.000867
D_5__Pir4 lineage	0.00072	0.000963	3.90E-05	0.001401
D_5__Variovorax	0.000719	0.000572	0.000137	0.001623

D_5__Friedmanniella	0.000704	NA	0.000704	0.000704
D_5__Nordella	0.000691	NA	0.000691	0.000691
D_5__Devosia	0.000677	0.000596	0.000104	0.001744
D_5__Brevibacillus	0.00062	0.000738	9.78E-05	0.001464
D_5__Phreatobacter	0.000602	0.000382	0.000332	0.000872
D_5__Shinella	0.000591	0.000626	0.000119	0.001301
D_5__Faecalibacterium	0.00055	0.000578	0.00015	0.001212
D_5__Nocardioides	0.000521	0.000516	0.000156	0.000886
D_5__Hydrogenophaga	0.000504	0.000621	6.52E-05	0.000944
D_5__Medicago truncatula	0.000437	NA	0.000437	0.000437
D_5__Arenimonas	0.00043	0.000378	0.000163	0.000697
D_5__Candidatus Udaeobacter	0.00043	NA	0.00043	0.00043
D_5__Geodermatophilus	0.000398	NA	0.000398	0.000398
D_5__Gemmatimonas	0.000365	NA	0.000365	0.000365
D_5__Streptococcus	0.000328	NA	0.000328	0.000328
D_5__Jatrophihabitans	0.000274	NA	0.000274	0.000274
D_5__Moheibacter	0.00026	NA	0.00026	0.00026
D_5__Roseburia	0.000254	0.000323	2.60E-05	0.000482
D_5__Sphingopyxis	0.000247	0.00023	8.47E-05	0.00041
D_5__uncultured Sphingomonadaceae bacterium	0.000235	NA	0.000235	0.000235
D_5__alphaI cluster	0.000235	NA	0.000235	0.000235
D_5__Coprococcus 1	0.000215	NA	0.000215	0.000215
D_5__Gaiella	0.000202	NA	0.000202	0.000202
D_5__Leptothrix	0.000182	NA	0.000182	0.000182

D_5__Prosthecobacter	0.000169	NA	0.000169	0.000169
D_5__Aeromicrobium	0.000163	NA	0.000163	0.000163
D_5__Rubellimicrobium	0.00015	NA	0.00015	0.00015
D_5__Candidimonas	0.000143	NA	0.000143	0.000143
D_5__Mesorhizobium	0.00014	0.000124	5.21E-05	0.000228
D_5__Delftia	7.81E-05	NA	7.81E-05	7.81E-05
D_5__metagenome	7.81E-05	NA	7.81E-05	7.81E-05
D_5__Rhodopseudomonas	6.52E-05	NA	6.52E-05	6.52E-05
D_5__Deinococcus	4.97E-05	NA	4.97E-05	4.97E-05
D_5__Mycobacterium	3.91E-05	NA	3.91E-05	3.91E-05
D_5__Conexibacter	3.26E-05	NA	3.26E-05	3.26E-05
D_5__Hyphomicrobium	2.60E-05	NA	2.60E-05	2.60E-05
D_5__Dorea	1.95E-05	NA	1.95E-05	1.95E-05

Appendix Document 1. Slug infection assay procedure.

Note: This document provides a summary of the slug infection assay experiment carried out in the laboratory of Rory Mc Donnell, which provided the *Phasmarhabditis hermaphrodita* nematode-associated microbiomes used for the dissertation Ch. 4 analysis. The results of dissertation Ch. 4 will be combined with the infection assay results described here for publication in a peer-reviewed scientific manuscript.

1. Nematode-Bacteria Combination Set-up

Three populations of the nematode *P. hermaphrodita* were reared separately on three types of bacteria: *E. coli, Pseudomonas sp.,* and the natural bacterial community associated with nematodes that were collected from a slug carcass discovered on OSU campus. We used the strain *E. coli* OP50, which has been grown and maintained in the Denver Lab. The bacterium *Pseudomonas sp.* were isolated from the nematode's original microbiome and confirmed the genus identity with 16S rRNA and *rpoB* gene sequencing.

Prior to the culturing step, nematodes were bleached with alkaline hypochlorite 0.25M solution following the standard worm bleaching protocol (Stiernagle, 2006) to minimize the bacteria on the surface and inside. The worms were monitored on blank NGM agar plates for 2 days to ensure no bacterial growth. Subsequently, *P. hermaphrodita* nematodes were harvested from agar plates and washed once with sterile M9 buffer then twice with molecular water.

Nematodes were then distributed onto 10 replicates per combinations, five with a higher concentration (8,000 nematodes/ml -'high dose') and five in 'low dose' (4,000 nematodes per ml). The number of nematodes were estimated by counting sub-samples under light microscopy. In total, 30 samples were set up [3 nematode-bacteria combinations x 2 doses x 5 replicates = 30 replicates].

2. Infection Assay Procedure

The assay was conducted in Mc Donnell lab (Department of Crop and Soil Science, OSU). Test slugs were collected one day prior to the experiment and maintained in growth chambers at 18C, which is the optimal temperature for slugs until the infectivity trials. Bioassay containers were also stored in the same growth chamber.

Infectivity trials were performed in 16oz plastic round containers (13.8 cm height x 11cm in diameter) containing 25g sterilized, damp topsoil. Worms were sprayed onto the soil to the final concentration of 210 worms/cm2 for low dose, or 420 worms/cm2 for high dose. Afterwards, six healthy, mature *D. reticulatum* slugs (>100 mg) were added to each of the 30 containers. In the previous pilot trial, no sign of aggression was documented with this number of slugs cohabiting in the same container. Five additional containers, each held six slugs and no nematodes were also set up as negative controls.

The slugs were monitored for symptoms of nematode infection (e.g. swelling of the mantle, emaciation of the slug body, exposure of the internal shell, and nematodes visible on the slug body) and mortality were recorded daily for 15 days. Statistical analysis was conducted using the non-parametric Kruskal-Wallis test and corrected with Bonferroni correction.

3. Results highlights

- Significant differences were between the controls and nematode/bacteria treatments.
- The BC High concentration was the first nematode/bacteria treatment to cause significantly more slug mortality than the negative controls (Day 5, $p = 0.015$).
- The first *Pseudomonas* treatment to cause significantly more mortality than the control was the High concentration (Day 6, $p = 0.01$).
- It took 10 days for the first *E. coli* treatment (i.e. the high concentration) to cause significantly more slug mortality than the controls.
- The *Pseudomonas* High rate treatment was the first to cause 100% slug mortality in all replicates (Day 8). The *Pseudomonas* Low rate and the BC High rate caused

100% slug mortality on Day 9. This occurred for BC Low and *E coli* High on Day 10, and on Day 13 for *E coli* Low.

- *Pseudomonas* High was the first treatment to cause significantly more slug mortality than any other treatment (i.e. *E. coli* Low) on Day 7 ($p = 0.004$).
- The second treatment to cause significantly more slug mortality than another treatment (*E. coli* Low) was BC High (Day 8, $p = 0.002$). The third and fourth treatments to cause significantly more slug mortality than *E. coli* Low were *Pseudomonas* Low (Day 9, $P = 0.028$) and BC Low (Day 10, $p=0.001$).

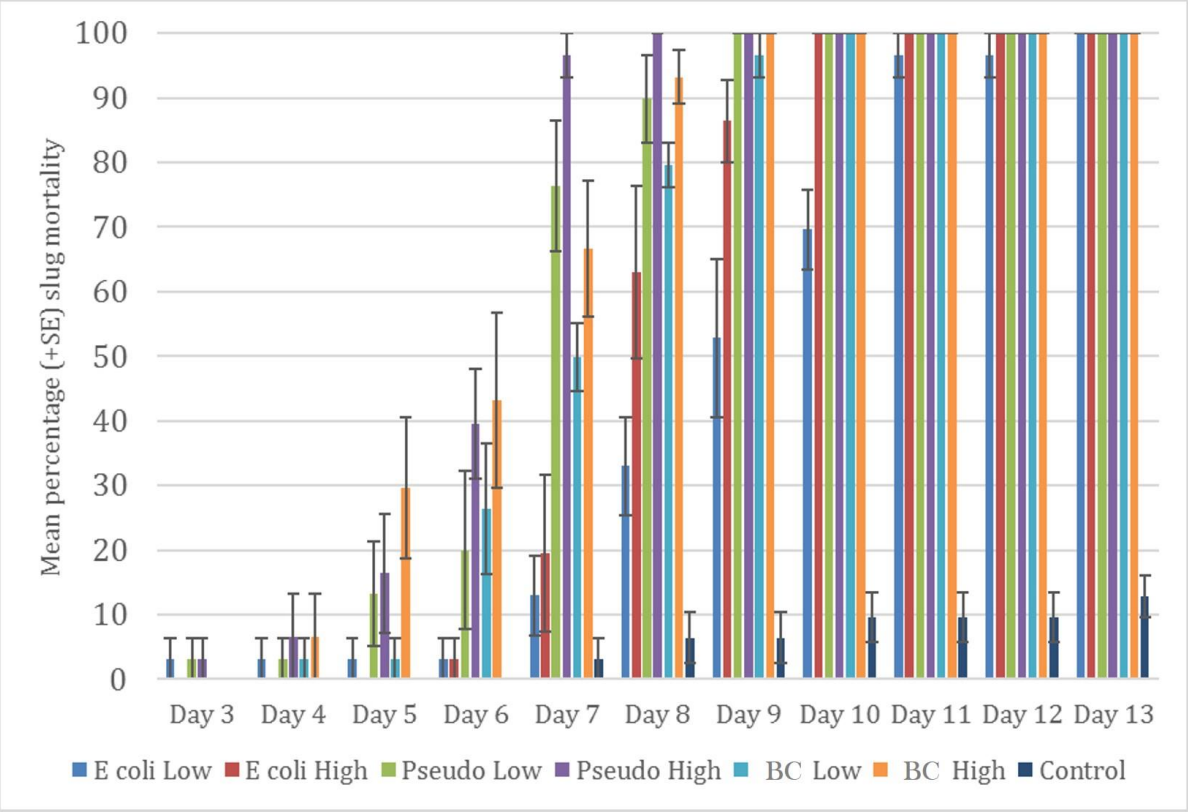


Figure S1. Mean percentage (±SE) slug mortality caused by *Phasmarhabditis hermaphrodita* cultured on its native bacterial community (BC), *Pseudomonas sp.* that co-cultured with this bacterial community, and *E. coli* OP50 over 13 days in a laboratory infection trial.