University of California Merced

The phages among us: Revealing the role of bacteriophages in biological ecosystems through whole genome sequence analysis

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Quantitative and Systems Biology
By
Tyrome S. Sweet Jr.

Committee in charge:

Dr. Michael Beman, Chair of advisory committee

Dr. Miriam Barlow

Dr. Paul Smaldino

Dr. Mark Sistrom

Dr. Suzanne S. Sindi

The dissertation of Tyrome Sweet Jr., titled, "*The phages among us: Revealing the role of bacteriophages in biological ecosystems through whole genome sequence analysis*", is approved, and is acceptable in quality and form for publication on microfilm and electronically:

_____ Date _____

Committee member Dr. Miriam Barlow

_____ Date _____

Committee member Dr. Paul Smaldino

_____ Date _____

Chair Dr. Michael Beman

_____ Date _____

Co Advisor Dr. Mark Sistrom

_____ Date _____

Co Advisor Dr. Suzanne S. Sindi

Dedications

This dissertation is dedicated to my mothers Pettry Barr, and Almisher Sweet-Little Moore and all of my family, friends and colleagues who had supported me along the way. I would also like to thank the University of California, The National GEM Consortium/GEMi4, The DOE Joint Genome Institute and The National Science Foundation (NSF) for helping keep my promise of crossing the finish line.

Table of contents

X. **Chapter 4: Conclusion**
    A. Limitations to computational pipelines
    B. Future work
XI. **Chapter 1 and 4 References**

List of Abbreviations

Chapter 1

- NGS - Next-generation sequencing
- Q-PCR - Quantitative polymerase chain reaction
- WGS - whole genome sequence
- NCBI - National Center for BioTechnology Information

Chapter 2

- MRSA - Methicillin Resistant *Staphylococcus aureus*
- ORFs - Open Reading Frames
- PBPs - Penicillin-binding proteins

Chapter 3

- NPA - non plant-associated
- PA - plant-associated
- PGPR - plant growth promoting rhizobacterium
- VPF - viral protein families

List of Figures

Chapter 1

Chapter 2

Chapter 3

List of Tables

Chapter 2

Chapter 3

Acknowledgments

Multi-Environment Computer for Exploration and Discovery (MERCED) cluster at UC Merced, funded by National Science Foundation Grant No. ACI-1429783

National Science Foundation –National Research Traineeship in Intelligent Adaptive Systems (NRT-IAS) (Award No. 1633722)

National GEM Consortium/Georgia Tech Research Institute - GEM PhD Engineering and Science Fellowship

Curriculum Vitae

Tyrome S Sweet Jr.

linkedin.com/in/tyromesweet | https://github.com/Cerebro409

Summary

- Interdisciplinary scientist and GEM Fellow with skills and experience in Virology, genomics, and machine learning
- Deep understanding of genomic data analysis and visualization
- Self-motivated, problem-solving and collaborative scientist with excellent communication skills

Publications

- Sweet, T., Sindi, S., & Sistrom, M. (2021). Going through phages: A Computational approach to Revealing the role of prophage in Staphylococcus aureus. bioRxiv.

Technical Summary and Skills

- Scrum Foundation Professional Certificate (SFPC): 46972857, 09/24/2020
- **Project Management tools & Methodologies:** Jira, Clickup, Asana, Trello, Gitkraken, Github, SCRUM, Kanban
- **Programming and Scripting:**Python, JavaScript, R, Bash, Perl, SQL, Golang, C++,C
- **Tools & Packages:** VScode, Jupyter Notebook, Anaconda, Docker, R studio, Jetbrains (PyCharm, Datagrip, Clion, Webstorm), Figma, Ansible Playbook/tower, Jenkins, postman, CircleCi, terraform, kubernetes
- **Cloud Platforms & Databases:** AWS Cloud Platform, Google Cloud Platform, IBM Cloud Platform, Microsoft Azure Cloud Platform PostgreSQL, MongoDB, Cassandra, DB2, DB2 z/OS

Teaching and mentoring experience

- 2022 - Center for Advancing Diversity in Engineering (CADE) Instructor for SPARK 10 Inclusive Innovation Seminar
- 2021, 2019, 2018 - Graduate Teaching Assistant for Calculus
- 2020 - Graduate Teaching Assistant for General Virology
- 2017 - Graduate Teaching Assistant for Statistics for Scientific Data Analysis

Professional Experience

PhageNet **Merced , CA**
Founder/CEO **12/2021 – Present**

- Determining and implementing strategic direction for product management and development processes.
- Leading a team of 7 to implementing various big data/machine and deep learning pipelines for sequence analysis (phageNet)
- Working with stakeholders to effectively manage the product roadmap, release, Infrastructure (Web, Analytics, AWS Cloud) Architecture, development and implementation of a biological data analysis platform MVP

MINWO **Dover, DE**
Lead Technical Architect **07/2019 – present**

- Managed product roadmap, release, Infrastructure (Web, Analytics, AWS Cloud) Architecture, development and implementation of all software platforms that support MINWO
- Lead, trained, and managed a growing development team of 10 on key performance indicators, data mining, and data analysis for data-driven decision making under the SCRUM methodology

- Designed, developed and scaled production ready multi-user web applications from UI/UX wireframes using ReactJS, HTML, JavaScript, and AWS/Google Cloud platforms for minority-owned start up initial product release and scale

DOE Joint Genome Institute - Lawrence Berkeley National Labs **Merced, CA**
Computational Biology Intern (Plant Microbial Group) **06/2022 – 08/2022**

- Aggregated Bacterial whole genome sequences using in Python and bash and mined them for bacteriophage (viruses) using Virsorter2
- Developed data-driven workflow using Jupyter, Bash, Python, and Biopython to perform exploratory and statistical analysis to see if plant and non plant based bacteriophage have similarities
- Visualized results and analysis several groups were interested in through both talks and poster presentations

IBM **Austin, TX**
Site Reliability Engineer - Hybrid Cloud **10/2020– 10/2021**

- Designed, developed and scaled enterprise web applications from UI/UX wireframes to production using VueJS, JavaScript and golang that support IBM Hybrid Cloud internal and external users
- Developed a microservice for reading/writing large volume of data(millions) from Mongo, PostgreSQL, DB2 and DB2 z/os databases using Golang
- Used Ansible playbooks and Ansible Tower to automate repetitive tasks, quickly deploys critical applications, and perform system health checks

Method Data Science LLC **Irvine, CA**
Computer Vision Intern **08/2020 – 09/2020**

- Consulted several clients in different industries for machine learning/deep learning projects for Object detection in images and videos (data analysis, model design, performance optimization, transfer learning)
- Lead, train, and manage a growing development team of 5 on key performance indicators, data mining, and data analysis for data-driven decision making
- Integrated object detection models into client software systems using AWS Cloud platform

Georgia Tech Research Institute - CIPHER Labs **Atlanta, GA**
Data Science Intern - Malware Analysis **05/2019 – 07/2019**

- Aggregated dataset of 500K proprietary company malware files using python
- Analyzed malware by implementing an opensource PyTorch model in python, C++, and bash with an aggregated dataset of over 200K malware files
- Communicated results to departmental stakeholders on how the model should be an addition to the system, and ways the dataset could be used to explore threat detection in machine learning models

IHS Markit **Houston, TX**
Software Engineer **08/2016 – 07/2017**

- Maintained backend/data warehouses for web applications used by analyst for price forecasting
- Updated tools and macros used by analyst for data forecasting
- Converted legacy code in Access DB, VB/VBA into SQL server and C# applications

Samsung **Austin, TX**
Systems Engineer **11/2015 – 08/2016**

- Automated a SharePoint workflow to expedite processes for the Facility and Systems Engineering Dept.
- Maintained custom software for logistics project management in SharePoint, MS Access, MS Excel, VBA, SQL Server
- Initiated/planned SharePoint mobile access conversion to see maintenance documentation in the field

Raytheon **Dulles, VA**
Software Engineer **05/2015 – 11/2015**

- Redesigned dropdown menu content and layout of a tactical control system (TCS) for unmanned aerial vehicles (UAV) in JavaScript, C and C++ to increase operator efficiency in sending preferred automated commands
- Performed Vulnerability testing with ACAS on the TCS and documented/resolved weak points according to DISA (STIGs)
- Performed automated test and retest (ATRT) on the UI/UX operator controls of the TCS

Total System Services (TSYS) **Columbus, GA**
Programmer Analyst **10/2014 – 05/2015**

- Maintained Banking subsystem that automated transactions for pin debit and credit cards in JCL, COBOL, DB2 z/OS and SharePoint
- Performed maintenance and created new applications for the implementation of a new banking system for TSYS's onboarding clients
- Performed weekly production and testing region support to pin debit system during graveyard shifts

Blue Cross Blue Shield **Columbia, SC**
Application Developer **05/2012 – 11/2014**

- Maintained Healthcare subsystem that automated the reporting and pricing of healthcare claims
- Created documentation that mapped the entire Automated Medical Management System (AMMS) and its dependencies
- Planned and executed business continuity/disaster recovery exercises by restructuring health claim data backup/migration process

Education

University of California Merced	PhD – Quantitative and Systems Biology	Expected Dec 2022
University of California Merced	M.S – Quantitative and Systems Biology	May 2021
Benedict College	B.S. Computer Engineering, magna cum laude	May 2012

Grants, Awards & Honors

- 2022 JGI User Meeting best early career poster - DOE Joint Genome Institute/Lawrence Berkeley National Labs
- 2022 GEM Annual Board Meeting and Conference Student Award - The National GEM Consortium
- 2022 NSF I-Corp Regional Summer Fellows Program - The National GEM Consortium/Cornell Tech University
- Center for Advancing Diversity in Engineering (CADE) Teaching Fellow - UC Merced
- UC-HBCU Initiative grant - 78K (Mentor) 2021/2022 - UC Merced, Benedict College
- 2021 AWS CTO Fellowship - Amazon Web Services
- 2021 NSF I-Corp Regional Summer Fellows Program - The National GEM Consortium/University of Toledo
- 2021 Cox Enterprises Social Impact Accelerator Powered by Techstars (120K) - Techstars, Cox Enterprises
- Berkeley Skydeck Incubator program - UC Berkeley
- GEM Full Fellow (16K) - The National GEM Consortium
- NSF Research Fellow (34K) – National Science Foundation (Intelligent adaptive systems)
- NSF Research Fellow (15K) – National Science Foundation (Interdisciplinary Computational Graduate Education)

Memberships & Organizations

- Society for the Advancement of Biology Education Research (SABER)
- Society for the Study of Evolution (SSE)
- International Society for Computational Biology (ISCB)
- American Society for Microbiology (ASM)
- National Society of Black Engineers (NSBE) - UC Merced

Conference Presentations

- University of California Irvine Beall Applied Innovation Born in California demo/talk 2022 - University of California Irvine
- University of California Merced Quantitative and Systems Biology Seminar - DOE JGI Experience - University of California Merced
- University of California Merced Quantitative and Systems Biology Seminar - Department Current research 2022 - University of California Merced
- DOE Joint Genome Institute 2022 Annual Meeting Talk and Poster Presentation - DOE JGI, Lawrence Berkeley National Lab
- DOE Joint Genome Institute Final Symposium 2022 - DOE JGI, Lawrence Berkeley National Lab
- Finding Your Inner Modeler V 2022 - University of Illinois at Chicago/National Science Foundation
- DOE Joint Genome Institute Workforce Development and Education Poster Presentation - DOE JGI, Lawrence Berkeley National
- Lab
- DOE Joint Genome Institute Midterm Symposium - DOE JGI, Lawrence Berkeley National Lab
- University of California Merced Quantitative and Systems Biology Retreat 2022 - University of California Merced
- Industry Strategy Symposium 2022 oral/poster presenter - SEMI North America
- Benedict College UC-HBCU Recruitment Talk - Benedict College
- Evolution 2019 speaker - Society for the Study of Evolution

Abstract

The phages among us: Revealing the role of bacteriophages in biological ecosystems through whole genome sequence analysis

Doctor of Philosophy

in

Quantitative and Systems Biology

by

Tyrome Steven Sweet Jr.

University of California, Merced

2022

Chair of the Advisory Committee: Dr. Michael Beman

Bacteriophages (phages) are viruses that target and infect bacteria. The impact of phages on a biological ecosystem could result in devastation to the system or augmentation. T/., For example, in the bacterium *Staphylococcus aureus*, phages have important roles in virulence, antibiotic resistance, and genome evolution. In agricultural ecosystems, bacteria and phages offer the host plant protection from pathogens, and provide resilience against stressful environments. Determining the presence of virulence and beneficial genes helps us uncover more about the relationship between the host, bacteria and phages.

Identifying and analyzing phages in bacterial genome sequences through experimentation can be costly in both time and resources. Rapid growth in the number of sequenced bacterial genomes allows for an investigation of prophage sequences at an unprecedented scale. Computational pipelines and systems can be used to explore how phages impact the host and ecosystem through techniques such as: (1) alignment-based methods that leverage sequence homology and sequence similarity, (2) alignment-free methods centered around sequence composition and genomic features, and (3) machine-learning-based methods. In this dissertation, I leverage the above techniques and others to explore the role of phages in biological ecosystems through bacterial genome sequences.

Chapter 1: Introduction

Biological ecosystems

A biological ecosystem is a network of both living and nonliving components that co-exist **[44]**. Abiotic components are the non living components in the system which includes water, soil, and the atmosphere **[44]**. Biotic components are the living components of the system which are the plants, animals, and microorganisms found in the system. Interactions between biotic and abiotic components are complex and determine the stability of an ecosystem **[38,44]**. Bacteriophages are able to affect both abiotic and biotic components making them a force of biodiversity in ecosystems **[29,41]**. For example, In agricultural ecosystems plant-microbe interactions help maximize crop yields and decrease crop losses due to biotic or abiotic stressors **[17]**.

Bacteriophages

Bacteriophages are viruses that infect and replicate in bacteria. Phages outnumber bacteria by 10 to 1 with an estimated global population of 10^{31} and are believed to be the most abundant self-replicating organisms on earth **[27,34]**. Their genome size averages between 24-200 nm in length **[5,22]**. Bacteriophages currently have one order and 10 families that they are classified into **[1,18,22]**.

Bacteriophages are diverse in shape and genetic information they carry **[29]**. Their structure consists of a capsid (Shell that protects their DNA), a tail, and tail fibers that attach to the host's cell membrane **[26,29]**. Bacteriophages can be found in natural environments such as the human body, marine environments, and soil, as well as artificial environments such as wastewater treatment systems, industrial applications, and laboratory-based techniques (genetic engineering of phages) **[6]**. It is hypothesized that prophage sequences that confer a selective advantage to their host are more likely to be conserved in the bacterial genomes than those that are neutral or deleterious to their hosts **[19]**. The resultant expectation is that prophage sequences will contain an elevated quantity of genes conferring adaptive functions to host bacteria **[2,6,14]**.

Bacteriophages have adaptive replication cycles, lytic and lysogenic. Lytic phages replicate inside the host and cause host lysis in order to enter the external environment, thus causing the release of host organic matter and new viral particles **[12]**. Temperate bacteriophages, bacteriophages whose genome is incorporated into the host bacterium, can switch from the lysogenic to lytic replication cycle **[12,16]**, the mechanisms of which are currently still unknown.

Bacteriophages drive evolution in biological ecosystems

Bacteriophages play key roles in bacterial evolution, governing abundance, adaptation, and diversity of bacterial communities **[29]**. The impacts of phages on a biological ecosystem could result in devastation to the system or augmentation **[6,26,29]**. Through transduction (infection) they introduce several genes from previous hosts and the environment to the current infected host **[26,27]**. During horizontal gene transfer several genetic elements are introduced to the host such as capsid proteins, tail proteins and genes that are potentially beneficial **[8,23,40]**. Phages are able to cause the host to express different phenotypes due to the genes introduced during the lysogenic life cycle

[16,23]. The more beneficial the gene is to the host, the more likely the host will accept that particular phage into their genome **[12,19,23]**.
Bacteria are capable of evolving to improve survival capabilities in different environments **[50,51,52]**. In the bacterium staphylococcus aureus, temperate bacteriophages housing the mecA gene were able to drive S. aureus into evolving by making it resistant to methicillin **[45]**. Historically it has been believed that Methicillin-resistant S. aureus (MRSA) has a single point of origin **[57]**. For example, Kreiswirth et al. observed 472 isolates using chromosomal transposon Tn554 to conclude that the mecA gene introduced to S. aureus may have a single origin **[57]**. In agricultural ecosystems bacteria is capable of providing beneficial genes to plants [51]. For example, certain bacteria can develop resistance to metals found in the environment which helps it survive in an environment that's toxic **[53/54]**. In this dissertation, I explore several sequences showing genes providing beneficial genes, and explore how different identified phages are associated.
Bacteria have been observed to make connections to other organisms and phages but currently the relationship between bacteria, viruses and a host organism shares **[54,55,56]**. Bacteria can promote plant growth by providing beneficial genes **[56]**. Phages can transduce mercury resistance genes to bacteria **[55]**. The effects both instances have on each other are currently not fully understood.

Experimental procedures cost time and money

Bacteriophage experimental protocols such as isolation, purification, amplification, microscopy, DNA extraction, and characterization can take weeks and costs hundreds of thousands for equipment and supplies **[31,32, 33,37].**
Experimental protocols yield the most accuracy, however they tend to be limited in discovering the genetic processes that take place **[20]**. Experiments require precision in order to observe specific biological processes, which limits the possibility of exploratory analysis for discovering and identifying bacteriophages **[2]**. Next-generation sequencing aids this process by allowing scientists to capture samples and results from their experiments as a digital file **[32,33,37]**.
The advancement of sequencing technology and computational methods have advanced our ability to observe processes during experiments **[15,24,25]**. For instance, computational methods can detect trends in omics data but are limited to the available data from experiments to train and validate models **[24]**. Endy et al. used experimental data to simulate Bacteriophage T7 wild-type development and was able to observe and compare the simulated results to *E. coli* BL21 experiments **[15]**. Experimental procedures yield more results when both sequencing and computational methods are leveraged **(See Figure 1)[15,24,25]**.

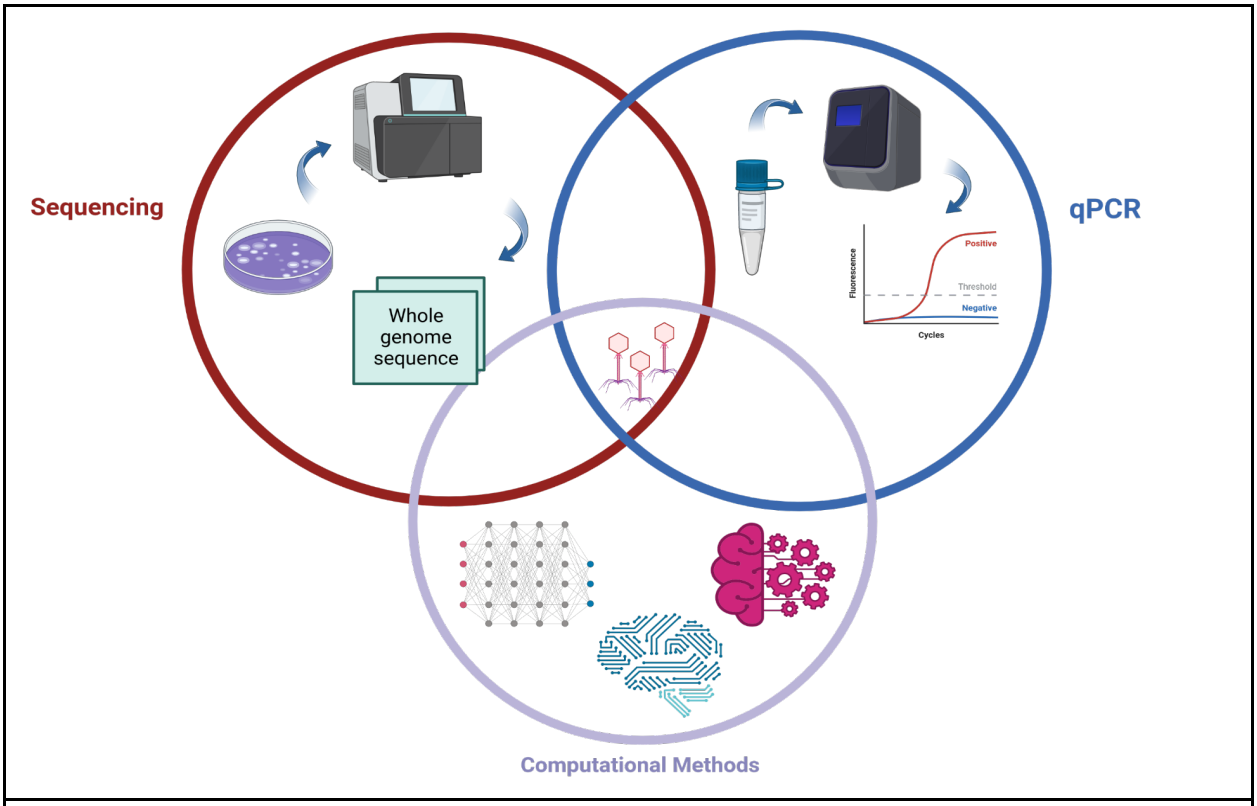


Figure 1: Bacteriophage identification and quantification methods. One approach to identifying and characterizing bacteriophages is extremely limiting. Multiple approaches through sequencing, experimental procedures and computational methods combined offer the most precision on detecting and enumerating bacteriophages **[2]**. Sequencing captures bacteria and phage genomes as digital files, allowing computational analysis. Quantitative polymerase chain reaction (Q-PCR) is a method by which the amount of the PCR product can be determined, in real-time, which was used to quantify bacteriophages M13 and T7 by investigating gene expression **[2]**. Computational methods leverage data science and mathematical methods to detect patterns and trends in genomic data. A combination of all 3 methods are leveraged to discover and quantify phages.

Computational methods for sequence analysis

Next-generation sequencing (NGS) captures genetic information in the sample's genome as a whole genome sequence (WGS)**[32]**. Viral discovery has been revolutionized by metagenomics, which allows computational identification of viral genome sequences without experimentation **[14,35].** Several online repositories offer whole genome sequences that were experimentally annotated. One example is The IMG/M system which supports the annotation, analysis and distribution of microbial genome and microbiome datasets sequenced at The Department of Energy (DOE) Joint Genome Institute (JGI) and from other contributing labs and scientists from around the world **[10]**. Another example is the National Center for BioTechnology Information (NCBI) GenBank is a comprehensive database that contains publicly available nucleotide sequences for almost 260,000 formally described bacterial species that were collected from laboratories and

large-scale sequencing projects **[7].** Both repositories contained bacterial and viral sequences, but there are more viral specific sources. ViruSITE is a database of viral genomes and genes. ViruSITE comprises all genomes from viruses, viroids and satellites published in NCBI Reference Sequence Database by computationally extracting from numerous resources (NCBI RefSeq, UniProtKB, GO, ViralZone, PubMed) and integrating under human supervision **[39]**. ViruSite has a total of 11,620 viral sequences, 14,813 genome sequences and 597,210 genes detected from the total 26,433 combined viral and genome sequences **[39]**. Overall, repositories offer the foundation for development of computational methods.

Computational focused approaches such as (1) alignment-based methods that leverage sequence homology and sequence similarity, (2) alignment-free methods centered around sequence composition and genomic features, and (3) machine-learning-based methods **[43]**. These different approaches are needed because unlike bacteria, viruses are not currently considered living organisms and they require different approaches to understand the extent of phage global diversity **[11]**.

Several applications such as PhiSpy, MUSCLE, VirSorter2 and checkV are constrained to analyzing a limited number of WGS which makes bulk analysis of sequences difficult **[3,13,21,30]**. Pipelines offer a variety of genome analysis and can be used to analyze multiple sequences **[14,28,42]**.

Exploring Bacteriophages impacts on ecosystems using computational techniques

Determining the presence of virulence and beneficial genes helps us uncover more about the relationship between the host, bacteria and phages. Computational techniques through a novel computational pipeline can demonstrate potential novel plant-host interactions that an experiment hasn't been designed for. Furthermore, Identifying and analyzing phages in bacterial genome sequences through experimentation can be costly in both time and resources. Rapid growth in the number of sequenced bacterial genomes allows for an investigation of prophage sequences at an unprecedented scale. Computational pipelines and systems can be used to explore how phages impact the host and ecosystem through techniques such as: (1) alignment-based methods that leverage sequence homology and sequence similarity, (2) alignment-free methods centered around sequence composition and genomic features, and (3) machine-learning-based methods **[3,9,21,30,36,43]**. Alignment-based methods that leverage sequence homology and sequence similarity such as BLAST **[4]** and Phirbo **[47]** give insight on sequence similarity, and taxonomy **[43]**. Alignment-free methods centered around sequence composition and genomic features can be used for viral phylogeny **[43,46]**. Machine-learning-based methods like virSorter2 and PhiSpy can predict phage regions in bacterial sequences that may not have been identified in experiments **[3,21,43]**. In this dissertation, I leverage the above techniques and others to explore the role of phages in biological ecosystems through bacterial genome sequences.The primary focus of this dissertation is to understand the role bacteriophages play in biological ecosystems, and whether or not they promote competition between microorganisms in a host's system.

Chapter 2: Going through phages: A Computational approach to Revealing the role of prophage in Staphylococcus aureus

Tyrome Sweet[1a], Suzanne Sindi[2b], Mark Sistrom[3a]
a Department of Life and Environmental Sciences, University of California, Merced, California, USA
b Department of Applied Mathematics, University of California, Merced, California, USA

Abstract

Prophages have important roles in virulence, antibiotic resistance, and genome evolution in Staphylococcus aureus. Rapid growth in the number of sequenced S. aureus genomes allows for an investigation of prophage sequences at an unprecedented scale. We developed a novel computational pipeline for phage discovery and annotation. We combined PhiSpy, a phage discovery tool, with VGAS and PROKKA, genome annotation tools to detect and analyze prophage sequences in nearly 10,011 S. aureus genomes, discovering thousands of putative prophage sequences with genes encoding virulence factors and antibiotic resistance. To our knowledge, this is the first large-scale application of PhiSpy on a large-scale set of genomes (10,011 S. aureus). Determining the presence of virulence and resistance encoding genes in prophage has implications for the potential transfer of these genes/functions to other bacteria via transduction and thus can provide insight into the evolution and spread of these genes/functions between bacterial strains. While the phage we have identified may be known, these phages were not necessarily known or characterized in S. aureus and the clustering and comparison we did for phage based on their gene content is novel. Moreover, the reporting of these genes with the S. aureus genomes is novel.

Impact statement

Bacteriophages (phage) play key roles in bacterial evolution, governing abundance, adaptation, and diversity of bacterial communities. Temperate phage can facilitate bacterial adaptation via transduction of novel genes. This study takes advantage of the unprecedented quantity of genomic sequencing in public repositories to analyze viral genes in 10,011 *Staphylococcus aureus* genomes. We found 196,727 predicted prophage genome sequences, with an estimated total of 129,935 genes. We determined the function of these genes, identifying a large quantity of novel genes that benefit the host such as beta-lactamase, enterotoxins and cytotoxins. These results will inform studies of bacterial evolution and adaptation, by identifying the mechanism of horizontal transfer of genes that confer adaptive traits to bacteria, especially in the context of antibiotic resistance.

Introduction

The ecological importance of viruses is now widely recognized, yet our limited knowledge of viral sequence space and virus–host interactions precludes accurate prediction of their roles and impacts **[65]**. Bacteriophages, viruses that infect and replicate in bacteria, are the most abundant self-replicating organisms on earth. Phages

outnumber bacteria by 10 to 1 with an estimated global population of 10^{31} **[1]**. The increase in antibiotic resistance has sparked the development of bacteriophage agents for several applications in agriculture, biotechnology, and medicine **[66]**. Before we can truly understand how to apply bacteriophage agents, we must first understand the relationship between bacteriophages and their hosts, as well as other species that could potentially be affected.
Methicillin Resistant *Staphylococcus aureus* (MRSA) is one of the major causes of antibiotic resistant clinical infections. Between 1999 and 2005, hospitalizations for *S. aureus* increased from 294,570 patients to 477, 927. Moreover, MRSA was responsible for 127,036 patients in 1999 increasing to 278,203 by 2005 **[6]**.

S. aureus has a mesh-like cell wall composed of cross-linked polymer peptidoglycans (PG). Penicillin-binding proteins (PBPs), mediate the final stages of PG synthesis **[8]**. Methicillin is a *β-lactam* antibiotic that inhibits the transpeptidation domain of PBPs, which weakens the cell wall **[9]**. MRSA produces PBP2A due to the *mecA* gene that encodes it. Furthermore, this *mecA* gene is transducible by prophage **[5]**.

Through transduction, horizontal gene transfer, bacteriophages could cause *Staphylococcus aureus* to become Methicillin Resistant through the *mecA* gene. A well-studied example of an adaptive trait conferred by transduction by lysogenic phage is the *mecA* gene transduced by the phage *Staphylococcus sciuri* **[4].** Transduction of this temperate phage into the *Staphylococcus aureus* genome confers resistance to broad spectrum beta-lactam antibiotics **[5]**.

Bacteriophages impact host evolution

Temperate bacteriophages, bacteriophages whose genome is incorporated into the host bacterium, can switch between the lytic and lysogenic life cycle **[1].** This can be triggered by environmental stressors such as toxic chemicals and low nutrient conditions. The lytic cycle destroys the host, but if the phage stays lysogenic it provides several benefits. One benefit is protection from secondary phage attacks from another prophage. Temperate phages can lose their switching ability if there are mutations in the attachment sites. Changes to the gene that encode the recombinase responsible for the excision of phage can result in 'grounding' of the phage **[7]**. Grounded phage offers the host benefits, without the risk of entering the lytic cycle.

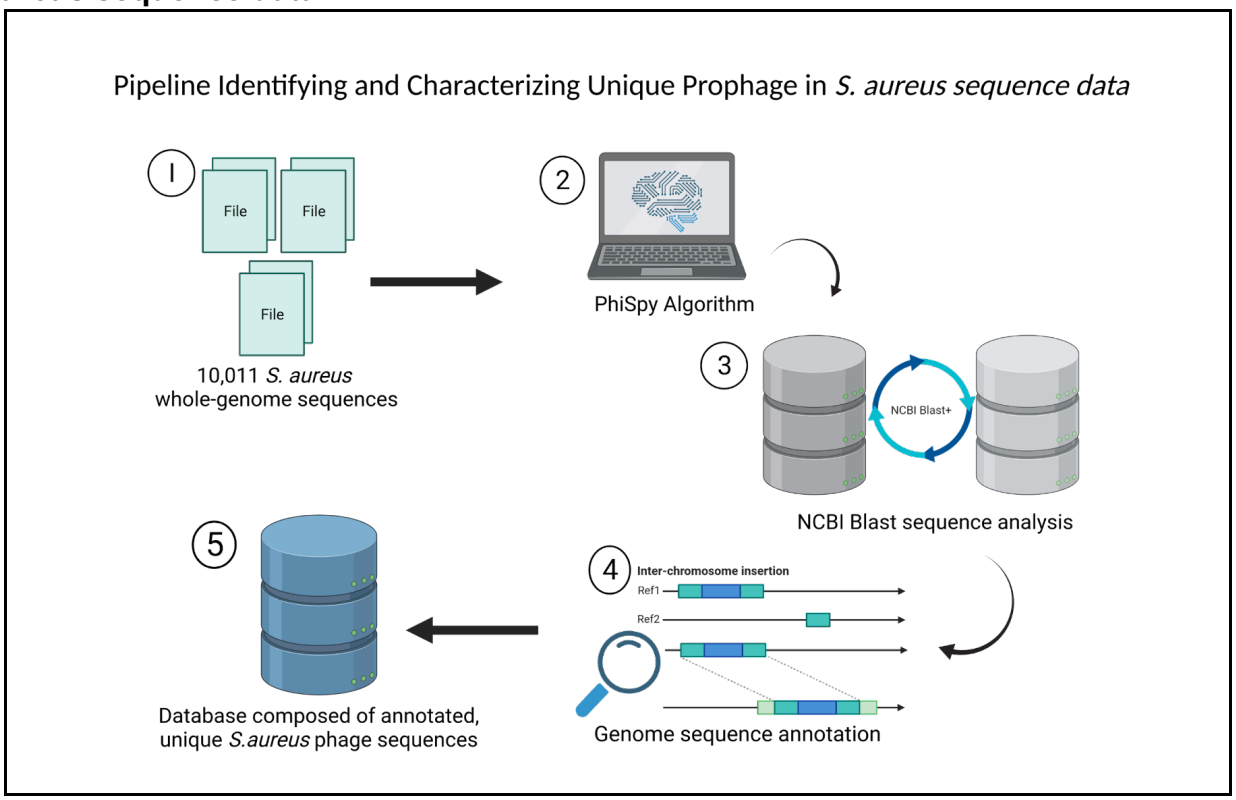
Lysogenic phage are transduced into the host bacterial genome as prophage sequences and can have a range of selectional impacts on the host, spanning the breadth of the mutualism-parasitism continuum **[2].** It is hypothesized that prophage sequences that confer a selective advantage to their host are more likely to be conserved in the bacterial genomes than those that are neutral or deleterious to their hosts **[3]**. The resultant expectation is that prophage sequences will contain an elevated quantity of genes conferring adaptive functions to host bacteria.

Computational advances for Whole Genome Sequence (WGS) analysis

The number of sequenced and annotated phage genomes is relatively small with 40,981 phage genome sequences, and 266,129 prokaryotic genome sequences **[10]** on August 18th, 2018. Given the exponential increase in the number of genome sequences

deposited in public repositories, it is timely to take advantage of these sequences to analyze them for novel functions. In this study we analyze 10,011 *S. aureus* genomes downloaded from NCBI in 2018 for prophage sequences and determine their functions. The total number of genome sequences for all organisms numbered 528,859 for 1 online repository **[12]**. Advances in computational techniques for the analysis of large data sets have advanced the omics field by enabling researchers to analyze larger datasets at lower costs **[13]**.
In this study, we developed a computational pipeline to detect and analyze prophage sequences in nearly 10,011 *S. aureus* genomes. To our knowledge, this is the first large-scale application of PhiSpy on a large-scale set of genomes (10,011 *S. aureus*). We discovered thousands of putative prophage sequences with genes encoding virulence factors and antibiotic resistance. We found genes encoding *mecA*, genes encoding toxins/antitoxins and clusters of prophage sequences that had genes in common. Our results, and methods developed, will facilitate similar studies for other bacterial species and promise to be a useful tool in the study of prophage host evolution. While most genes we identified were known, the clustering and comparison we did for phage based on their gene content is novel. Moreover, the reporting of these genes with the *S. aureus* genomes is novel **(Figure 1)**.

Chapter 1 Figure 1: Pipeline Identifying and Characterizing Unique Prophage in *S. aureus* sequence data.



A visualization of the workflow used to identify unique prophage sequences. 1) 10,011 *S. aureus* genome sequences were downloaded from the National Center for Biotechnology information (NCBI). 2) The sequences were analyzed by PhiSpy. 3) The fasta files for each predicted prophage were compared against each other using NCBI Blast nucleotide alignment tool. Prophage sequences that had 90% similarity along their full length were counted the same. 4) Phage sequences were annotated using two independent methods (VGAS, Prokka). 5) The resulting database of annotated, unique phage sequence allows for the identification of gene function encoded within prophage in *S. aureus*. **(See materials and methods section for more information)**

Methods

S. aureus Genomes

S. aureus genomes were obtained from the National Center for Biotechnology Information NCBI's Genbank repository on August 18, 2018 **[10]**. All available genome sequences (n=10,011 including complete and partial assemblies) were downloaded for this study. The sequences were collected from a variety of backgrounds that include: hospital environments, lab strains and animals. **(Accession numbers are provided in Supplemental Data).**

Viral Detection

Putative prophage sequences were detected using PhiSpy, Version 3.2 **[14]**. PhiSpy uses a random forest algorithm that has been trained on seven distinct features of prophage: protein length, transcription strand directionality, AT and GC skew, the abundance of unique phage words (unique sequence of length 12 base pairs), phage insertion points and the similarity of phage proteins. PhiSpy has 49 available training sets to increase accuracy for specific genomes. We used the *S. aureus* training dataset (option 24) and identified 196,727 phage regions in our 10,011 *S. aureus* genomes.

Prophage Clustering

Prophage sequences identified by PhiSpy were unique within a genome, but highly redundant between genomes. We identified highly similar prophages between genomes through a reciprocal **[15]** search. We increased the max_target_seqs to 12,000 (higher than our total number of *S. aureus* genomes) to ensure we captured all possible matches. We also used a custom output format which provided additional information on the alignment.

We then grouped prophages by using an undirected graph approach with nodes of the form: Genome *i*, Prophage *j*. Edges were added between nodes if they had a blast alignment which exceeded 90% similarity and 90% coverage of both source and target based on the Blastn reports. We then identified genomes sharing the same prophage by determining the connected components, resulting in 191 unique phage clusters.

Cluster Validation

Each of the 191 phage clusters were aligned with Muscle v3.8.1551 **[16]** and ClustalW v2.1 **[17]** to ensure each phage was similar. A score of 0.0000 indicates that the undirected graph script formed accurate phage clusters.

Genome Annotation

One representative was selected from each of the 191 phage clusters and analyzed with 2 different tools for gene annotation: VGAS **[18]**, and Prokka **[19]**. VGAS and PROKKA identified ORFS in each of the phage genome sequences. VGAS identifies ORFs through an enhanced version of the ZCurve algorithm **[20]** that was customized by adding 13 additional identifying variables (45 total) for the classification model, and BLASTP **[21]** searches for gene prediction. The all ORFs were annotated by all both tools with default settings. The combination of annotation tools served as a quality check. The genes identified by both tools were manually reviewed and the highest percentage, and the tool that gave the highest number of matches to known databases was selected for the phages annotation. **(Annotation reports and accession numbers are provided in Supplemental Data).**

Pairwise Sequence analysis

We identified shared genes between phage through a reciprocal blast search using the annotated phage sequences. We constructed a new undirected graph with the nodes being the phage genome and the edges representing genes shared between phages. The output was a .csv file that listed each of the 191 phage with the genes shared with other phages.

Jaccard Index

We used the layout_with_mds option for the layout function of the R package Igraph **[23]** to visualize the phages with shared genes using the pairwise count matrix for both PROKKA and VGAS. The Jaccard Index **[24]** was calculated using a modified version of the Jaccard index function in R **[25]** to compare the Prokka and VGAS networks. **(See Table 2 in the results section)**

Quality assessment of predicted phage sequences with CheckV

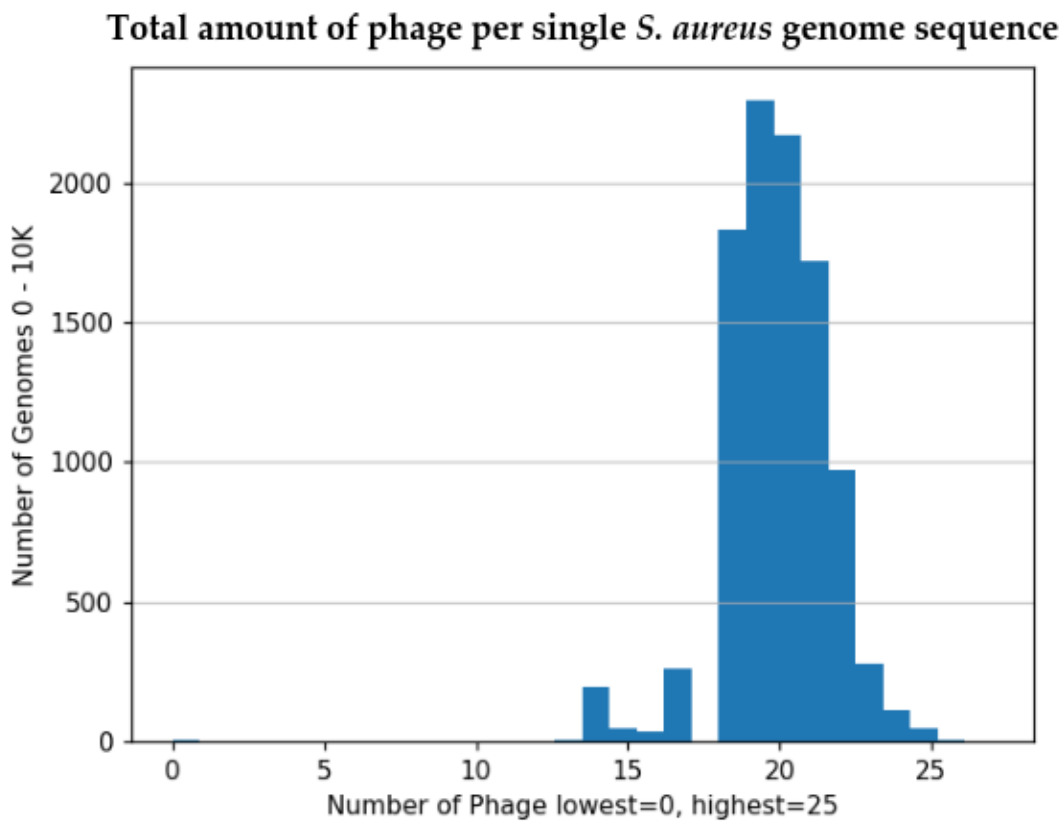
CheckV is an automated pipeline for identifying closed viral genomes, estimating the completeness of genome fragments and removing flanking host regions from integrated proviruses **[48]**. CheckV compares to Virus Orthologous Groups (VOGDB), DOE Joint Genome Institute's IMG/VR, Reference Viral DataBase (RVDB), KEGG Orthology, Pfam A, Pfam B and TIGRFAM databases [**48**]. CheckV also reports on potential viral and host genes and uses hmmsearch v.3.1b2 and CheckM to determine the quality of the viral sequences **[48]**. All 191 unique prophage sequences were analyzed with checkV using default settings **(see checkv_quality_summary in Supplemental Data).**

Results

Of the 10,011 genomes initially analyzed, 11 were not annotated completely and did not pass the conversion to SEED **[26]** due to missing locus tags **[27]**. A further 5 were too short for PhiSpy to detect phage regions, resulting in a total of 9,995 genomes which

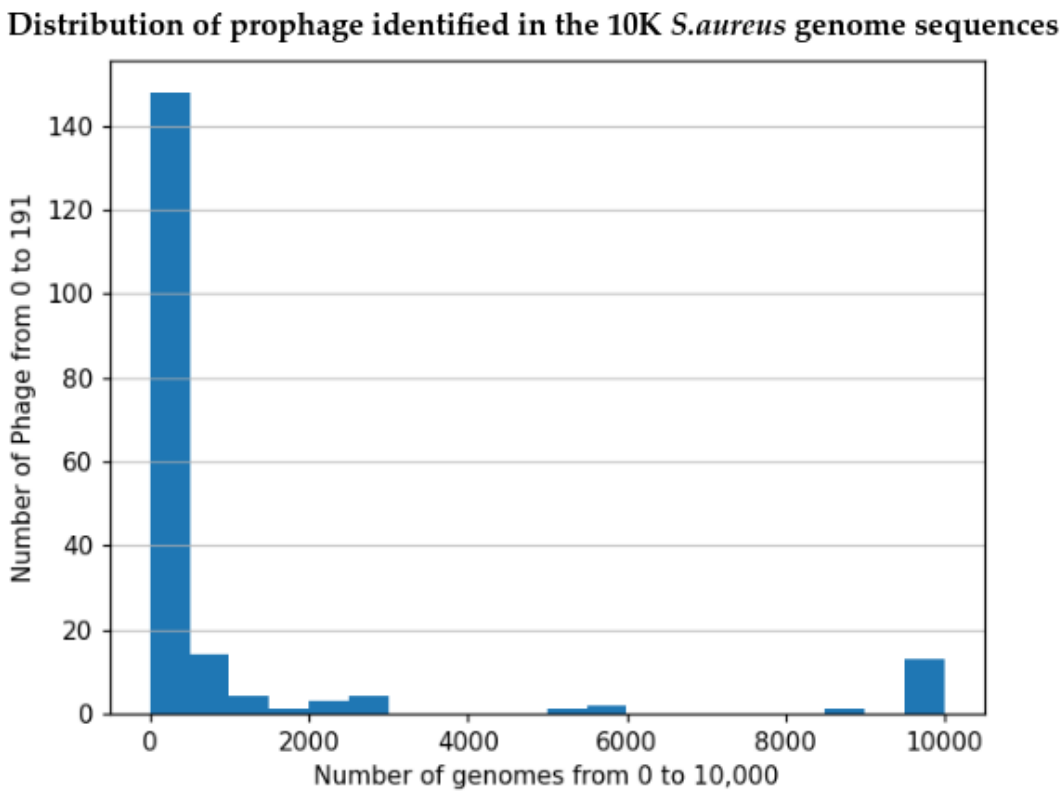
were used for subsequent analysis. Within these, we detected a total of 196,727 prophage sequences across the 10,011 genomes, with an average of 19.68 (standard deviation = 1.78) prophage sequences per genome (**Figures 2 and 3**).

Chapter 1 Figure 2: Total amount of phage per single *S. aureus* genome sequence.



This figure shows the distribution of the phage genome sequences detected by PhiSpy. A total of 196,727 prophage sequences across the 10,011 *S. aureus* genomes. The x-axis reflects the number of phage sequences per *S. aureus* genome sequence (y-axis). There is an average of 19.68 (standard deviation = 1.78) prophage sequences per *S. aureus* genome. 45 *S. aureus* genome sequences had 25 phage regions present, and 5 *S. aureus* genome sequences had 0 phage sequences detected. **(See Analysis Uncovers 191 Unique Prophage Sequences section for more information).**

Chapter 1 Figure 3: Distribution of prophage identified in the 10K *S. aureus* genome sequences.



This figure shows the distribution of 191 unique prophage sequences. PhiSpy detected phage genome sequences in nearly every *S. aureus* genome studied. The detected phage genome sequences were grouped by using an undirected graph approach **(see Methods)**. 1 representative phage from each cluster was selected, totaling 191 unique prophage sequences. The y-axis reflects the exact totals of each of the 191 phage genome sequences that were detected in the S. aureus genome sequences (x-axis). **(See Analysis Uncovers 191 Unique Prophage Sequences section for more information).**

Analysis Uncovers 191 Unique Prophage Sequences

Reciprocal BLAST analysis coupled with undirected graph analysis (see Methods) found that the 196,727 prophage sequences corresponded to 191 unique prophage sequences. Each unique prophage sequence appeared in an average of 1024 host genomes (standard deviation = 2581.33) (**Figure 3**). Each prophage contained an average of 16.83 putative coding regions, resulting in a total of 3,207 (VGAS) and 3,205 (Prokka) unique open reading frames (ORFs) (**Table 1**). One phage appeared in all 9,995 genome sequences, while 42 of the 191 distinct phages were found in only a single genome sequence.

Analysis Detects Thousands of ORFs with Potential Gene Function

One representative prophage sequence was selected from each of the 191 phage clusters and analyzed with two different tools for gene annotation: VGAS **[18]**, and

Prokka **[19]**. VGAS identified 3,207 genes, and PROKKA detected 3,205 genes (**Table 1**). For the PROKKA results, 1,155 ORFs did not have an identified function. 806 predicted ORFs corresponded to known ORFs with accession numbers matching known databases ISfinder **[28]**, NCBI **[29]**, UniProtKB **[30]**. 2041 genes had a predicted gene function. VGAS predicted 2935 ORFs, 361 of which corresponded to known accession numbers matching databases Swissprot and refseq **[18,20]** and 307 other predicted ORFs had predicted gene functions.(**Table 1**).

Chapter 1 Table 1: PROKKA and VGAS predict gene functions in 191 unique phage sequences

Tool	Total Detected ORFS	ORFs with Gene Function	ORFs with No gene function	ORFs that match known databases
PROKKA	3205	2040	45	806
VGAS	3207	307	2846	361
PROKKA and VGAS both identified several Open reading frames (ORFs). PROKKA determined there were 3205 ORFs for all 191 unique phages, while VGAS determined 3207. VGAS determined that only 307 of the ORFs had gene function, while PROKKA determined 2040 did. PROKKA had roughly 45 ORFS that did not have any gene function identified. This excludes hypothetical or predicted function. (**see PROKKA and VGAS reports in Supplemental Data**)				

Analysis Shows Shared ORFs between Unique Prophage Sequences

In order to understand how similar the prophage were, for each annotation (PROKKA and VGAS) we created a graph representing genes shared between the distinct prophage. More specifically, approach outlined in the "**prophage clustering**" section with nodes of the form: Genome *i*, identified gene *j*. Edges were added between nodes if they had a matching identified gene. We then Compared the edges produced by both tools PROKKA and VGAS with each other.

We found a total of 1,335 shared edges defined by PROKKA and VGAS. The lowest number of shared edges between phage sequences was 1, and the highest was 73 (**Table 2**). There were 1,306 shared edges between PROKKA and VGAS, and 28 shared edges unique to PROKKA (**Table 2**) out of the total 1,335 (**Table 2**). In the 28 unique PROKKA the numbers of shared edges between each node ranged from 1 to 22. VGAS defined a total of 1,334 connected components. The lowest number of genes shared between phage sequences was 1, and the highest was 75. There were 27 shared edges unique to VGAS (**Table 2**) out of the total 1334 (**Table 2**). The 27 unique VGAS shared edges ranged from 1 to 22 as well.

Chapter 1 Table 2: Jaccard index shows connections between PROKKA and VGAS Undirected Graphs

Tool	Total amount of genes shared	Shared genes between both tools	Unique shared genes	Highest # of shared genes in cluster	Lowest # of shared genes in cluster
PROKKA	1363	1335	28	73	1
VGAS	1362	1335	27	75	1

This table shows the relationship between phage genomes by their gene content. Specifically, the nodes represent the 191 phage genome sequences, and the edges between nodes indicate the two phages share a gene (as annotated by Prokka and VGAS). We determined that there were 1335 connected components between the 191 unique phage genome sequences. The total number of shared genes between the 191 unique phage sequences ranged from 1 shared gene to 73 shared genes for PROKKA and 1 shared gene to 75 shared genes for VGAS (2 more edges than the total identified by PROKKA). PROKKA had a total of 1363 connections compared to VGAS 1362. (See Analysis Shows Shared ORFs between Unique Prophage Sequences section for more information and Table 2).

Genes Encoding *mecA* Found in 2 of the 191 Unique Prophage

There were several traces of antimicrobial resistance found in the 191 phage clusters. The *mecA* ancestral gene specifically was identified in 2 sequences. The first sequence, accession number ASM900v1 **[10]**, cluster group has 1023 phage, 10% of the total *S. aureus* genomes. ASM900v1, or RF122 (ET3-1) provides a framework for the identification of specific factors associated with host specificity in this major human and animal pathogen **[32]**. RF122 (ET3-1) has several genes involved in host colonization, toxin production, iron metabolism, antibiotic resistance, and gene regulation **[33]**. ASM323779v1 **[34]** is the only phage in the cluster, making it individually unique compared to the 196,727 total detected. It is a part of 184 *S. aureus* isolates collected from 135 patients over a timespan of 3 years at an Italian pediatric hospital **[35]**.

48 Unique Gene Functions appear in several phage genome sequences

48 unique encoding traces of Antimicrobial Resistance (**Shared_genes table in supplemental data**). 4 genes stuck out the most due to the number of clusters they appeared in. GDAEFEPF_00005 Staphylococcal complement inhibitor, a gene found in ASM2514v1 appeared in 10 **[36]**. GHDFECEE_00007 Superantigen-like protein 13 was found in ASM17451v1 and appeared in 8 clusters **[37]**. ASM17451 also contained GHDFECEE_00008 Superantigen-like protein 13 which appeared in 7 clusters. GAIDFPLK_00004 Superantigen-like protein 13 was found in ASM1150v1 and was identified in 7 clusters **[38]**.

4 Genes Showing Traces of Toxin/Antitoxin (TA) System

Toxin/Antitoxin (TA) systems encode toxin proteins that interfere with vital cellular functions and are counteracted by antitoxins. There are 6 different types of TA systems **[39]**. *S. aureus* has genes identified showing types I, II and III **[40]**. Type I toxin-antitoxin systems have the base-pairing of antitoxin RNA with the toxin mRNA **[41]** Type III systems toxic proteins and an RNA antitoxin have a direct iteration where the toxic proteins are neutralized by the RNA gene **[42]**.
Type II, the most studied TA system, has proteic antitoxin that tightly binds and inhibits the activity of a stable toxin **[43]**. The TA system yoeB-yefM has been detected as genes MBJHDCJA_00021 Toxin YoeB and MBJHDCJA_00022 Antitoxin YefM in ASM900v1 **[32,33]**. yoeB inhibits bacterial growth and translation by cleavage of mRNA molecules and is repressed by antitoxin yefM **[40]**. Enterotoxin Type A causes food poisoning and was identified in 3 genome sequences **[44]**. M1022 (NCTC 8325) was identified in 2 genome sequences **[31]**. CAFLMJIC_00063 Enterotoxin type A was identified in 1 genome sequence **[32,33]**. **(See Shared_genes, Frequent_gene_Functions and Least_Frequent_Gene_Functions tables in supplemental data)**

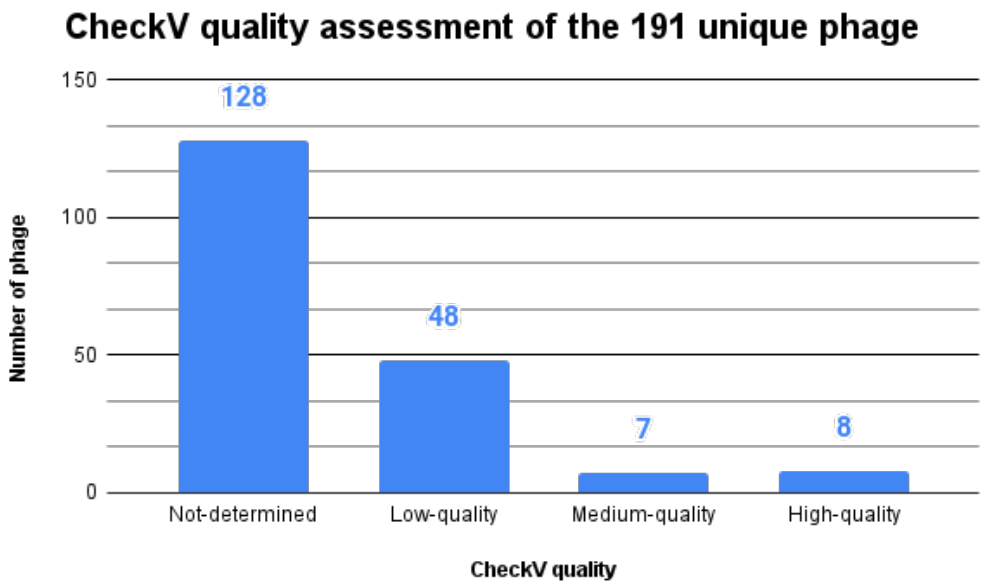
13 Most Shared Genes in the 191 Unique Phage

4 genes that stand out the most due to the amount of phage they were found in (**Frequent_gene_Functions table in supplemental data**). KHDAMHGJ_00009 Chorismate synthase, found in M0471 **[31]**, was identified in 17 phage clusters. Its gene function is shikimate pathway, which shows signs of AMR in plants **[45]**. EOLKNJBM_00007 Nucleoside diphosphate kinase in ASM1150v1_genomic.gbff_pp18.ffn **[38]** was found in 16 phage clusters. MIIMDJNA_00002 Heptaprenyl diphosphate synthase component 2 in ASM24879 **[46]** was identified in 15 clusters. HGDEFLKI_00006 3-dehydroquinate synthase in M0877_V1_genomic.gbff_pp18.ffn **[31]** was identified in 14 phage clusters.

CheckV identifies 63 phages of quality

CheckV analysis determined that there 63 phages that were of quality and 128 that could not be determined (**Figure 4**). There were 3277 total genes detected and 310 were viral genes determined by checkV. The High and medium quality phages all had viral genes detected. The low quality phages had a mix of 23 phages with viral genes detected and 25 without. **(see checkv_quality_summary in Supplemental Data).**

Chapter 1 Figure 4: CheckV quality assessment of the 191 unique phage.



CheckV determined that 63 phages out of the 191 unique phages were of quality. 48 phages were of low quality, 7 phages were of medium quality and 8 were high quality. The X axis shows the quality of phage determined by checkV. The Y axis shows the number of phages. The totals are shown above each bar. (**see CheckV identifies 63 phages of quality in the results section and PhiSpy_checkv_quality_summary in Supplemental Data**)

Discussion

Determining the presence of virulence and resistance encoding genes in prophage has implications for the potential horizontal transfer of these genes and the functions encode to other bacterial taxa via transduction, and thus can provide insight into the evolution and dissemination of virulence and resistance mechanisms of clinical importance. This knowledge can be useful when creating disease models and novel therapeutics.
The scope of this project is purely computational and determining the functionality of the genes detected would require experimentation. The genome sequences obtained from NCBI may not be representative of the complete diversity of *S. aureus* in nature. *Staphylococcus aureus subsp. aureus strain NCTC 8325* is referenced several times throughout the dataset. It was used as a propagating strain for bacteriophage 47 of the international typing set of bacteriophages and is considered the prototypical strain for most genetic research on *S. aureus* **[31]**. These limitations need to be considered in the interpretation of our results.

CheckV analysis identifies 128 potential false positives

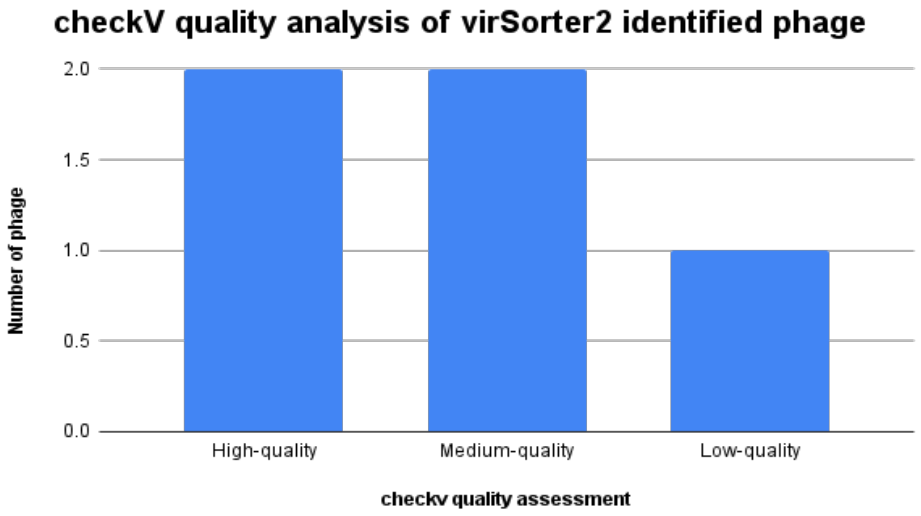
The checkV analysis determined that there 63 phages that were of quality and 128 that could not be determined (**Figure 5**) showing that there may be potential false positives. All available *S. aureus* genome sequences were downloaded from NCBI **[10]** which includes complete genome sequences, and partial sequences or contigs**.** PhiSpy uses a window size of 40 base pairs and does not rely on known homologues to identify phage regions. The identified prophage sequences appeared multiple times in a *S. aureus* sequence. The combination of PhiSpy identifying the same phages throughout the *S. aureus* sequences that were complete and partial are potentially why so many phages were identified. This is further shown where the 197,727 identified sequences were clustered into 191 unique groups. (**checkv_quality_summary in Supplemental Data**)

Analyzing 191 unique phages with virSorter 2

Prophage detection tools have significant problems with false positives and false negatives. PhiSpy identified an average of 20 phages per genome sequence which is a higher number compared to other studies. Deghorain and Van Melderen identified between 1-4 phage per genome **[51]** and Nepal et al. found an average of 3.6 phages per genome **[50]**. CheckV gave a quality assessment, but further analysis with virSorter2 **[57]** was done to see if phiSpy, virSorter2 and checkV agreed on the high and medium quality phage sequences.

Each of the 191 unique phage sequences were analyzed with virSorter2 **[57]** following a protocol from (Guo et al., 2021) **[67]**. VirSorter2 determined that 5 of the 191 unique identified phages by PhiSpy **[14]** were indeed phage sequences. The 5 virSorter2 **[57]** identified sequences were analyzed with checkv showing that all 5 phage were of quality (**Figure 5**). The 5 virSorter 2 phages were determined to be quality phage sequences by 3 different tools showing that the remaining 186 phage sequences were potential false positives identified by PhiSpy **[14]**.

Chapter 1 Figure 5: CheckV quality assessment of the virSorter2 identified phage.



CheckV determined that all 5 virSorter2 identified phage were of quality. 1 phage was low quality, 2 phages were of medium quality and 2 were high quality. The X axis shows the quality of phage determined by checkV. The Y axis shows the number of phages. (**see Analyzing 191 unique phages with virSorter 2**
in the results section and virSorter2_checkv_quality_summary in Supplemental Data)

Databases constrains limit PROKKA and VGAS annotations

There is a large possibility for novel functions to be conferred to bacterial hosts by transduction by lysogenic phage **[5]**; A significant proportion of the genes encoded by both free living and prophage sequences are of unknown function **[11]**. There were several virulence factors and toxins identified in the 191 unique prophage representatives, 1% of the total 196,727 phage detected. This is reflected through VGAS which predicted 2846 genes with no known function, and PROKKA with 45 predicted genes with no known function. PROKKA leverages UniProt **[30]**, RefSeq **[59]**, Pfam **[60]**, and TIGRFAMs **[61]** databases. VGAS uses RefSeq and SwissProt **[62]** databases. A third tool MOSGA **[69/70]** was used to analyze the 191 unique phage sequences. MOSGA **[69/70]** uses EggNog 5 **[71]**, SILVA **[72]** and SwissProt **[62]** databases. Only 34 genes were identified which was lower than both PROKKA and VGAS. PROKKA and VGAS used more databases in combination compared to MOSGA which increases the chances of finding a matching gene function.
Databases that scientists are updating with gene functions from experiments conducted serves a better foundation for gene annotation tools. The databases are limited to what scientists discover in genomics overall and this puts a major constraint on the databases. This could introduce a level of bias in the tools that are using the same databases. (**see MOSGA_annotation_analysis in Supplemental Data**)

Conclusion

We developed a novel computational pipeline for phage discovery and annotation and applied this pipeline to approximately 10,000 *S. aureus* genomes. In doing so, we discovered 191 unique clusters of putative prophage sequences with genes encoding virulence factors and antibiotic resistance. This computational pipeline consists of first identifying phage genome sequences, grouping them into clusters of identical (or nearly identical) phage, and then identifying genes within these phages. These results will be useful to those interested in bacterial evolution and adaptation, by identifying the mechanism of horizontal transfer of genes that confer adaptive traits to bacteria, especially in the context of antibiotic resistance like the *mecA* gene found in 2 of out 191 unique phage clusters. This database and pipeline can help guide future experiments by identifying phages and genes of interest.
The immediate next step is to expand the computational pipeline to leverage more tools for phage identification, gene annotation and to show the relationship between phage genome sequences using gene co-occurrence networks **[47].** *S. aureus* genome sequences will be collected from the National Center for BioTechnology Information genbank **[52]**, JGI IMG/M **[53]**, the DNA Data Bank of Japan **[54]** and phage repositories: ViruSite **[55]** and inphared **[56]** to gather more diverse *S. aureus* and *S. aureus* phage sequences. Ultimately the goal is to identify quality phage sequences computationally, and to find and test each identified phage to see if any could potentially turn lytic.

Author statements

Authors and contributors

TS was responsible for the conceptualization, methodology, formal analysis, data curation, writing of both the original draft preparation and editing. MS provided resources, writing (editing), and supervision. SS provided resources, writing (editing), and supervision.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Funding information

Multi-Environment Computer for Exploration and Discovery (MERCED) cluster at UC Merced, funded by National Science Foundation Grant No. ACI-1429783
National Science Foundation –National Research Traineeship in Intelligent Adaptive Systems (NRT-IAS) (Award No. 1633722)
National GEM Consortium/Georgia Tech Research Institute - GEM PhD Engineering and Science Fellowship

Ethical approval

All whole genome sequences were obtained from NCBI.

Acknowledgements

Ali Heydari – Pairwise Sequence Analysis

Chapter 2 References

1. Liu, L. (2014). Fields Virology, 6th Edition. *Clinical Infectious Diseases*, *59*(4), 613–613. https://doi.org/10.1093/cid/ciu346
2. Blair, J., Webber, M. A., Baylay, A. J., Ogbolu, D. O., & Piddock, L. J. (2015). Molecular mechanisms of antibiotic resistance. *Nature reviews microbiology*, *13*(1), 42-51.
3. Gandon, S. (2016). Why be temperate: lessons from bacteriophage λ. *Trends in microbiology*, *24*(5), 356-365.
4. Zeman, M., Mašlaňová, I., Indráková, A., Šiborová, M., Mikulášek, K., Bendíčková, K., ... & Pantůček, R. (2017). Staphylococcus sciuri bacteriophages double-convert for staphylokinase and phospholipase, mediate interspecies plasmid transduction, and package mecA gene. *Scientific reports*, *7*(1), 1-11.
5. Scharn, C. R., Tenover, F. C., & Goering, R. V. (2013). Transduction of staphylococcal cassette chromosome mec elements between strains of Staphylococcus aureus. *Antimicrobial agents and chemotherapy*, *57*(11), 5233-5238.
6. Klein, E., Smith, D. L., & Laxminarayan, R. (2007). Hospitalizations and deaths caused by methicillin-resistant Staphylococcus aureus, United States, 1999–2005. *Emerging infectious diseases*, *13*(12), 1840.
7. Ramisetty, B. C. M., & Sudhakari, P. A. (2019). Bacterial 'grounded'prophages: hotspots for genetic renovation and innovation. *Frontiers in genetics*, *10*, 65.
8. Scheffers, D. J., & Pinho, M. G. (2005). Bacterial cell wall synthesis: new insights from localization studies. *Microbiology and molecular biology reviews*, *69*(4), 585-607.
9. Fishovitz, J., Hermoso, J. A., Chang, M., & Mobashery, S. (2014). Penicillin-binding protein 2a of methicillin-resistant Staphylococcus aureus. *IUBMB life*, *66*(8), 572-577.
10. *Staphylococcus aureus (ID 154)—Genome—NCBI*. (n.d.). Retrieved October 2, 2020, from https://www.ncbi.nlm.nih.gov/genome/?term=Staphylococcus%20aureus[Organism]&cmd=DetailsSearch
11. Moura, A., Criscuolo, A., Pouseele, H., Maury, M. M., Leclercq, A., Tarr, C., ... & Brisse, S. (2016). Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes. *Nature microbiology*, *2*(2), 1-10.
12. *Genome List—Genome—NCBI*. (n.d.). Retrieved August 1, 2021, from https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/
13. Krassowski, M., Das, V., Sahu, S. K., & Misra, B. B. (2020). State of the field in multi-omics research: From computational needs to data mining and sharing. *Frontiers in Genetics*, 1598.
14. Akhter, S., Aziz, R. K., & Edwards, R. A. (2012). PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. *Nucleic acids research*, *40*(16), e126-e126.
15. Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic acids research*, *36*(suppl_2), W5-W9.

16. Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, *32*(5), 1792-1797.
17. Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, *22*(22), 4673-4680.
18. Zhang, K. Y., Gao, Y. Z., Du, M. Z., Liu, S., Dong, C., & Guo, F. B. (2019). Vgas: a viral genome annotation system. *Frontiers in microbiology*, *10*, 184.
19. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068-2069.
20. Guo, F. B., Ou, H. Y., & Zhang, C. T. (2003). ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic acids research*, *31*(6), 1780-1789.
21. Mahram, A., & Herbordt, M. C. (2015). NCBI BLASTP on high-performance reconfigurable computing systems. *ACM Transactions on Reconfigurable Technology and Systems (TRETS)*, *7*(4), 1-20.
22. Carvalho, T. (2020, June 29). *Heatmap Basics with Python's Seaborn*. Medium. https://towardsdatascience.com/heatmap-basics-with-pythons-seaborn-fb92ea280a6c
23. *Igraph – Network analysis software*. (n.d.). Retrieved June 7, 2021, from https://igraph.org/
24. Leydesdorff, L. (2008). On the normalization and visualization of author co-citation data: Salton's Cosine versus the Jaccard index. *Journal of the American Society for Information Science and Technology*, *59*(1), 77-85.
25. Bevelander, K. E., Smit, C. R., van Woudenberg, T. J., Buijs, L., Burk, W. J., & Buijzen, M. (2018). Youth's social network structures and peer influences: Study protocol MyMovez project – Phase I. *BMC Public Health*, *18*(1), 504. https://doi.org/10.1186/s12889-018-5353-5
26. Aziz, R. K., Devoid, S., Disz, T., Edwards, R. A., Henry, C. S., Olsen, G. J., Olson, R., Overbeek, R., Parrello, B., Pusch, G. D., Stevens, R. L., Vonstein, V., & Xia, F. (2012). SEED Servers: High-Performance Access to the SEED Genomes, Annotations, and Metabolic Models. *PLOS ONE*, *7*(10), e48053. https://doi.org/10.1371/journal.pone.0048053
27. *Prokaryotic Genome Annotation Guide*. (n.d.). Retrieved June 11, 2021, from https://www.ncbi.nlm.nih.gov/genbank/genomesubmit_annotation/
28. Siguier, P., Pérochon, J., Lestrade, L., Mahillon, J., & Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. *Nucleic acids research*, *34*(suppl_1), D32-D36.
29. *National Database of Antibiotic Resistant Organisms (NDARO)—Pathogen Detection—NCBI*. (n.d.). Retrieved July 26, 2021, from https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/
30. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., & Bairoch, A. (2007). UniProtKB/Swiss-Prot. In D. Edwards (Ed.), *Plant Bioinformatics: Methods and Protocols* (pp. 89–112). Humana Press. https://doi.org/10.1007/978-1-59745-535-0_4

31. *Staphylococcus aureus subsp. Aureus NCTC 8325 (ID 57795)—BioProject—NCBI*. (n.d.). Retrieved July 18, 2021, from https://www.ncbi.nlm.nih.gov/bioproject/57795
32. Herron, L. L., Chakravarty, R., Dwan, C., Fitzgerald, J. R., Musser, J. M., Retzel, E., & Kapur, V. (2002). Genome sequence survey identifies unique sequences and key virulence genes with unusual rates of amino Acid substitution in bovine *Staphylococcus aureus*. *Infection and Immunity*, *70*(7), 3978–3981. https://doi.org/10.1128/iai.70.7.3978-3981.2002
33. Herron-Olson, L., Fitzgerald, J. R., Musser, J. M., & Kapur, V. (2007). Molecular correlates of host specialization in Staphylococcus aureus. *PloS One*, *2*(10), e1120. https://doi.org/10.1371/journal.pone.0001120
34. *Staphylococcus aureus (ID 400143)—BioProject—NCBI*. (n.d.). Retrieved June 20, 2021, from https://www.ncbi.nlm.nih.gov/bioproject/PRJNA400143/
35. Manara, S., Pasolli, E., Dolce, D., Ravenni, N., Campana, S., Armanini, F., Asnicar, F., Mengoni, A., Galli, L., Montagnani, C., Venturini, E., Rota-Stabelli, O., Grandi, G., Taccetti, G., & Segata, N. (2018). Whole-genome epidemiology, characterisation, and phylogenetic reconstruction of Staphylococcus aureus strains in a pediatric hospital. *Genome Medicine*, *10*(1), 82. https://doi.org/10.1186/s13073-018-0593-7
36. Nübel, U., Dordel, J., Kurt, K., Strommenger, B., Westh, H., Shukla, S. K., Žemličková, H., Leblois, R., Wirth, T., Jombart, T., Balloux, F., & Witte, W. (2010). A Timescale for Evolution, Population Expansion, and Spatial Spread of an Emerging Clone of Methicillin-Resistant Staphylococcus aureus. *PLOS Pathogens*, *6*(4), e1000855. https://doi.org/10.1371/journal.ppat.1000855
37. Cameron, D. R., Ward, D. V., Kostoulias, X., Howden, B. P., Moellering, R. C., Jr, Eliopoulos, G. M., & Peleg, A. Y. (2012). Serine/Threonine Phosphatase Stp1 Contributes to Reduced Susceptibility to Vancomycin and Virulence in Staphylococcus aureus. *The Journal of Infectious Diseases*, *205*(11), 1677–1687. https://doi.org/10.1093/infdis/jis252
38. Holden, M. T. G., Feil, E. J., Lindsay, J. A., Peacock, S. J., Day, N. P. J., Enright, M. C., Foster, T. J., Moore, C. E., Hurst, L., Atkin, R., Barron, A., Bason, N., Bentley, S. D., Chillingworth, C., Chillingworth, T., Churcher, C., Clark, L., Corton, C., Cronin, A., … Parkhill, J. (2004). Complete genomes of two clinical Staphylococcus aureus strains: Evidence for the rapid evolution of virulence and drug resistance. *Proceedings of the National Academy of Sciences*, *101*(26), 9786–9791. https://doi.org/10.1073/pnas.0402521101
39. *Enterotoxin type A Staphylococcus aureus Antibody (F12)*. (n.d.). Novus Biologicals. Retrieved July 19, 2021, from https://www.novusbio.com/products/enterotoxin-type-a-staphylococcus-aureus-antibody-f12_nb100-73021
40. Schuster, C. F., & Bertram, R. (2016). Toxin-Antitoxin Systems of Staphylococcus aureus. *Toxins*, *8*(5), E140. https://doi.org/10.3390/toxins8050140
41. Fozo, E. M., Hemm, M. R., & Storz, G. (2008). Small Toxic Proteins and the Antisense RNAs That Repress Them. *Microbiology and Molecular Biology Reviews: MMBR*, *72*(4), 579–589. https://doi.org/10.1128/MMBR.00025-08

42. Labrie, S. J., Samson, J. E., & Moineau, S. (2010). Bacteriophage resistance mechanisms. *Nature Reviews Microbiology*, *8*(5), 317–327. https://doi.org/10.1038/nrmicro2315
43. Hayes, F. (2003). Toxins-Antitoxins: Plasmid Maintenance, Programmed Cell Death, and Cell Cycle Arrest. *Science*, *301*(5639), 1496–1499. https://doi.org/10.1126/science.1088157
44. Ono, H. K., Nishizawa, M., Yamamoto, Y., Hu, D.-L., Nakane, A., Shinagawa, K., & Omoe, K. (2012). Submucosal mast cells in the gastrointestinal tract are a target of staphylococcal enterotoxin type A. *FEMS Immunology & Medical Microbiology*, *64*(3), 392–402. https://doi.org/10.1111/j.1574-695X.2011.00924.x
45. *Comprehensive Natural Products II | ScienceDirect*. (n.d.). Retrieved July 20, 2021, from https://www.sciencedirect.com/referencework/9780080453828/comprehensive-natural-products-ii
46. *Staphylococcus aureus subsp. Aureus CIG1612 (ID 60683)—BioProject—NCBI*. (n.d.). Retrieved July 18, 2021, from https://www.ncbi.nlm.nih.gov/bioproject/PRJNA60683
47. Shapiro, J. W., & Putonti, C. (2018). Gene co-occurrence networks reflect bacteriophage ecology and evolution. MBio, 9(2), e01870-17.
48. Nayfach, S., Camargo, A.P., Schulz, F. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol 39, 578–585 (2021). https://doi.org/10.1038/s41587-020-00774-7
49. Kwan, T., Liu, J., DuBow, M., Gros, P., & Pelletier, J. (2005). The complete genomes and proteomes of 27 Staphylococcus aureus bacteriophages. Proceedings of the National Academy of Sciences, 102(14), 5174-5179.
50. Nepal, R., Houtak, G., Shaghayegh, G., Bouras, G., Shearwin, K., Psaltis, A. J., ... & Vreugde, S. (2021). Prophages encoding human immune evasion cluster genes are enriched in Staphylococcus aureus isolated from chronic rhinosinusitis patients with nasal polyps. Microbial genomics, 7(12).
51. Deghorain, M., & Van Melderen, L. (2012). The Staphylococci phages family: an overview. Viruses, 4(12), 3316-3335.
52. Benson, D., Lipman, D. J., & Ostell, J. (1993). GenBank. Nucleic Acids Research, 21(13), 2963-2965.
53. Chen, I. M. A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., ... & Kyrpides, N. C. (2019). IMG/M v. 5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic acids research, 47(D1), D666-D677.
54. Jun Mashima, Yuichi Kodama, Takatomo Fujisawa, Toshiaki Katayama, Yoshihiro Okuda, Eli Kaminuma, Osamu Ogasawara, Kousaku Okubo, Yasukazu Nakamura, Toshihisa Takagi, DNA Data Bank of Japan, Nucleic Acids Research, Volume 45, Issue D1, January 2017, Pages D25–D31, https://doi.org/10.1093/nar/gkw1001
55. Stano, M., Beke, G., & Klucar, L. (2016). viruSITE—integrated database for viral genomics. Database, 2016.
56. Ryan Cook, Nathan Brown, Tamsin Redgwell, Branko Rihtman, Megan Barnes, Martha Clokie, Dov J. Stekel, Jon Hobman, Michael A. Jones, and Andrew

Millard (2021). Infrastructure for a Phage Reference database: identification of large-scale biases in the current collection of cultured phage genomes. PHAGE, 2(4), 214-223.
57. Guo, J., Bolduc, B., Zayed, A.A. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome 9, 37 (2021). https://doi.org/10.1186/s40168-020-00990-y
58. Michael Shaffer, Mikayla A Borton, Bridget B McGivern, Ahmed A Zayed, Sabina Leanti La Rosa, Lindsey M Solden, Pengfei Liu, Adrienne B Narrowe, Josué Rodríguez-Ramos, Benjamin Bolduc, M Consuelo Gazitúa, Rebecca A Daly, Garrett J Smith, Dean R Vik, Phil B Pope, Matthew B Sullivan, Simon Roux, Kelly C Wrighton, DRAM for distilling microbial metabolism to automate the curation of microbiome function, Nucleic Acids Research, Volume 48, Issue 16, 18 September 2020, Pages 8883–8900, https://doi.org/10.1093/nar/gkaa621
59. Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, Kim D. Pruitt, Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation, Nucleic Acids Research, Volume 44, Issue D1, 4 January 2016, Pages D733–D745, https://doi.org/10.1093/nar/gkv1189
60. Alex Bateman, Lachlan Coin, Richard Durbin, Robert D. Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik L. L. Sonnhammer, David J. Studholme, Corin Yeats, Sean R. Eddy, The Pfam protein families database, Nucleic Acids Research, Volume 32, Issue suppl_1, 1 January 2004, Pages D138–D141, https://doi.org/10.1093/nar/gkh121
61. Daniel H. Haft, Jeremy D. Selengut, Owen White, The TIGRFAMs database of protein families, Nucleic Acids Research, Volume 31, Issue 1, 1 January 2003, Pages 371–373, https://doi.org/10.1093/nar/gkg128
62. Amos Bairoch, Rolf Apweiler, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, Nucleic Acids Research, Volume 28, Issue 1, 1 January 2000, Pages 45–48, https://doi.org/10.1093/nar/28.1.45
63. Minoru Kanehisa, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, Shuichi Kawashima, Shujiro Okuda, Toshiaki Tokimatsu, Yoshihiro Yamanishi, KEGG for linking genomes to life and the environment, Nucleic Acids Research, Volume 36, Issue suppl_1, 1 January 2008, Pages D480–D484, https://doi.org/10.1093/nar/gkm882
64. Baris E. Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, Cathy H. Wu, UniRef: comprehensive and non-redundant UniProt reference clusters,

Bioinformatics, Volume 23, Issue 10, 15 May 2007, Pages 1282–1288, https://doi.org/10.1093/bioinformatics/btm098
65. Roux, S., Hallam, S. J., Woyke, T., & Sullivan, M. B. (2015). Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. elife, 4, e08490.
66. Versoza, C. J., & Pfeifer, S. P. (2022). Computational Prediction of Bacteriophage Host Ranges. Microorganisms, 10(1), 149.
67. Guo, J., Pratama, A. A., Roux, S., & Sullivan, M. (2021, July 19). Viral sequence identification SOP with virsorter2. protocols.io. Retrieved September 12, 2022, from https://dx.doi.org/10.17504/protocols.io.bwm5pc86
68. Paul Terzian, Eric Olo Ndela, Clovis Galiez, Julien Lossouarn, Rubén Enrique Pérez Bucio, Robin Mom, Ariane Toussaint, Marie-Agnès Petit, François Enault, PHROG: families of prokaryotic virus proteins clustered using remote homology, NAR Genomics and Bioinformatics, Volume 3, Issue 3, September 2021, lqab067, https://doi.org/10.1093/nargab/lqab067
69. Roman Martin, Thomas Hackl, Georges Hattab, Matthias G Fischer, Dominik Heider (2020). MOSGA: Modular Open-Source Genome Annotator. Bioinformatics. 36(22-23). 5514–5515. doi: 10.1093/bioinformatics/btaa1003.
70. Roman Martin, Hagen Dreßler, Georges Hattab, Thomas Hackl, Matthias G Fischer, Dominik Heider (2021). MOSGA 2: Comparative genomics and validation tools. Computational and Structural Biotechnology Journal. 19. 5504-5509. doi: 10.1016/j.csbj.2021.09.024.
71. Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, Lars J Jensen, Christian von Mering, Peer Bork, eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses, Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D309–D314, https://doi.org/10.1093/nar/gky1085
72. Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, Frank Oliver Glöckner, The SILVA ribosomal RNA gene database project: improved data processing and web-based tools, Nucleic Acids Research, Volume 41, Issue D1, 1 January 2013, Pages D590–D596, https://doi.org/10.1093/nar/gks1219

Chapter 3: Wanderlust Phage: A comparative meta-analysis of plant-associated and non-plant-associated bacteriophages

Tyrome Sweet[1], Jonelle Basso[2]

[1] Department of Life and Environmental Sciences, University of California, Merced, California, USA
[2] Department of Energy Joint Genome Institute, Lawrence Berkeley National Lab, Berkeley, California, USA

Abstract

The complex ecosystem of the plant microbiome includes beneficial microbiota that can interact with the plant host to offer protection from pathogens, and provide resilience against abiotic stress. Historically, investigations that have been crucial for the elucidation of the impact of specific plant-microbe interactions in the rhizosphere have focused mostly on the bacterial and fungal components. However, viruses, which are the most abundant biological entities on the planet, continue to be grossly understudied in the rhizosphere, and their influence largely negated in these systems. We recently discovered a pair of resident bacteriophage genes found in the plant growth promoting rhizobacterium (PGPR), *Pseudomonas simiae* wcs417 that are predicted to have functional potential to modulate the ability of their bacterial host to colonize *Arabidopsis thaliana* roots. This finding led us to ask what quantity and diversity of phages exist among plant-associated bacteria (PAB), and how these compare to their non-plant-associated (NPAB) counterparts. Determining host range experimentally is a very efficient approach, but can be time consuming and have a great cost burden. Computational techniques help improve time and cost efficiency. In this study, I seek to leverage computational techniques through a novel computational pipeline demonstrating potential novel plant-microbe interactions in the rhizosphere.

Importance

Bacteriophages, viruses that infect bacteria, play key biogeochemical roles in agricultural ecosystems. Like marine viruses, those found in soil systems are greatly abundant and pervasive, but the prevalence of phage-host interactions and the mechanism of interaction with plant roots remains greatly understudied. As plants are carbon sinks, it is imperative to understand how bacteriophages can influence this nutrient flux, as well as manipulate their bacterial hosts as part of this multipartite relationship. This work focuses on the impact of phages on novel plant-microbe interactions in the rhizosphere.

Acknowledgments

Multi-Environment Computer for Exploration and Discovery (MERCED) cluster at UC Merced, funded by National Science Foundation Grant No. ACI-1429783
DOE JGI/University of California Internship Program

Introduction

In agricultural ecosystems, associated microorganisms can be found in or on plants, as well as in the soil in general **[4,8,57]**. Soil microbiota include bacteria, fungi, viruses, archaea and protists **[23]**. Plant associated microbes need to be able to use available nutrients, evade host defense systems and outcompete other microbes in the environment in order to have favorable fitness **[14]**. The interactions between plants and microbes could be negative or positive in specific contexts, and may modulate ways in which microbes can be beneficial **[14,57]**. Potential advantages of plant-microbe interactions include maximization of crop yields and a decrease of crop losses due to biotic or abiotic stressors **[23]**.

Bacteriophages are viruses that infect and replicate in bacteria. They play key roles in bacterial evolution, governing abundance, adaptation, and diversity of bacterial communities **[46]**. Historically, there has been a focus on bacterial and fungal components of the rhizosphere microbiome, but much remains unknown about bacteriophages in these systems **[44]**. Current literature has shown that lytic phages can impact bacterial abundance and composition during colonization of host plant leaves **[44]**. In this study, field-grown tomato plants were compared to juvenile plants grown under sterile conditions. They determined that the presence of bacteriophages affected overall bacterial abundance during colonization of new host plants **[44]**. Bacteriophages were capable of impacting plant leaves, but it is currently unknown whether they also impact other parts of the plant. Before we can truly understand the relationship between bacteriophages and their hosts, as well as other species that could potentially be affected, we must first understand how to identify phage regions inside of the host's genome **[12,35,64]**.

Bacteriophages mediate horizontal gene transfer

The long-term associations between phage-host interactions could lead to mutual benefits **[5]**. For example, research from Zhan et al. demonstrated that in the marine ecosystem lytic bacteriophages belonging to the podoviridae and siphoviridae families had mutually beneficial relationships with members of the roseobacter clade of marine bacteria due to shared genes between them **[69]**. Bacteriophages have an adaptive replication process where they can enter a lytic or lysogenic state in the instance of temperate phages, in comparison to their lytic counterparts that can typically only enter a lytic cycle **[22,28] (See Figure 1)**. In Basso et al., temperate bacteriophages phi-A and phi-D in *Sulfitobacter sp.* were observed where they found that phi-A and phi-D carry proteins that could be involved with the mediation of the lysogenic-lytic phage switch in CB2047 **[5]**. Lytic phages replicate inside the host and cause host lysis in order to enter the external environment, thus causing the release of host organic matter and new viral particles **[18]**. Temperate phages are capable of integrating their genetic information with that of their infecting host, therefore replicating and persisting along with the host **[18,22,28]**. Temperate bacteriophages are capable of lysogenic to lytic switching **[18,22**, the mechanisms of which are currently still unknown. Both cycles alter the release of nutrients into the environment which impacts the micro ecosystem through natural selection **[5,28,69]**. **(See Figure 1)**

During horizontal gene transfer several genetic elements are introduced to the host such as capsid proteins, tail proteins and genes that are potentially beneficial **[9,28,63].** Phages are able to cause the host to express different phenotypes due to the genes introduced during the lysogenic life cycle **[22,28]**. In the bacterium *staphylococcus aureus*, temperate bacteriophages housing the *mecA* gene were able to drive *S. aureus* into evolving by making it resistant to methicillin **[68]**. The more beneficial the gene is to the host, the more likely the host will accept that particular phage into their genome **[18,24,28]**.

Through sequencing technologies, we can identify genes computationally **[60]**. Transposon mutagenesis is a powerful means of producing randomized gene mutations in bacterial genomes **[6]**. Through next generation sequencing and transposon mutagenesis, we are able to detect genes present in bacteria and bacteriophages **[6,60]**.

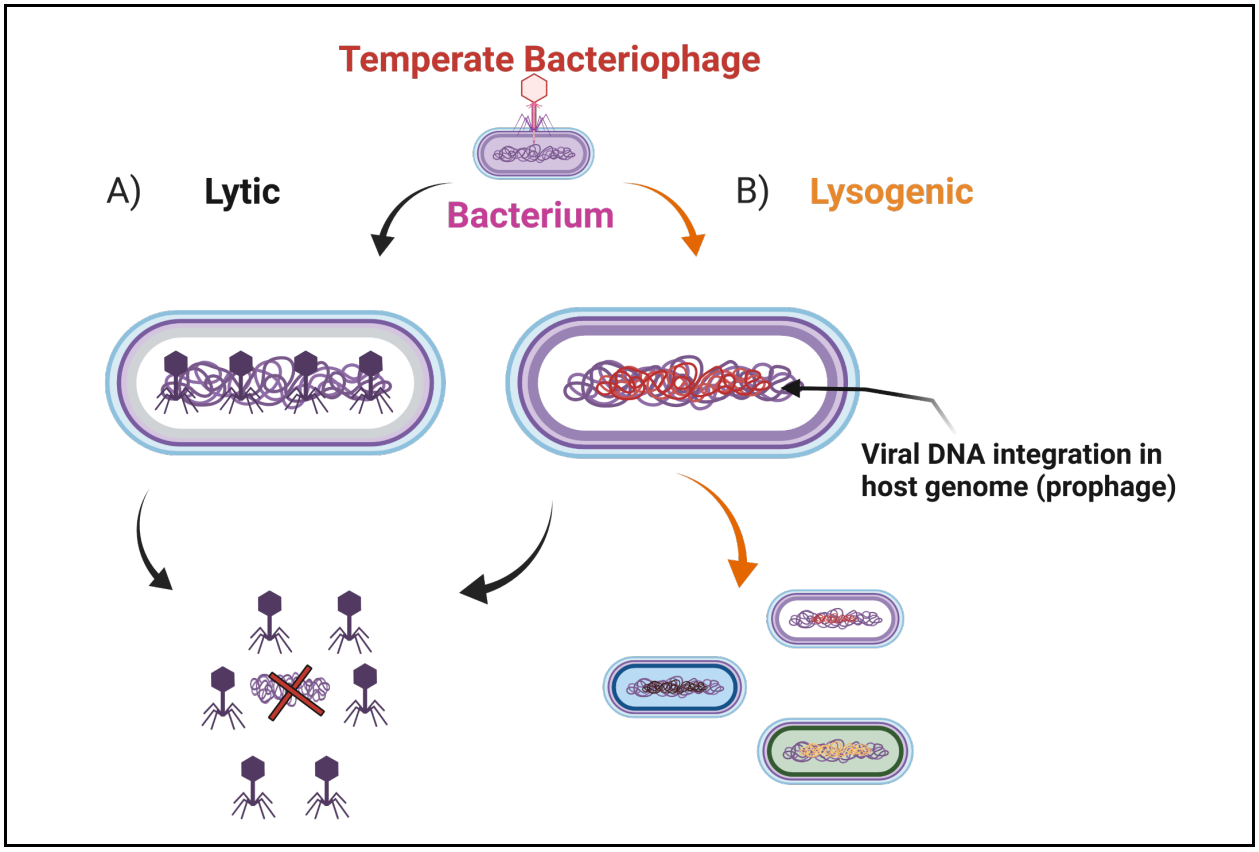


Figure 1: Temperate phages release genes to the host and environment

Temperate phages have adaptive replication cycles. A) During the lytic cycle, bacteriophages attach to the cell membrane and release genetic information into the host. The lytic phages then replicate in the host and lyses through releasing several lytic phages into the environment. B) During the lysogenic cycle viral DNA is integrated into the host genome. As the host reproduces, the lysogenic phage replicates with it. This process can cause the host to express different phenotypes. Lysogenic phages are also capable of turning into lytic phages **[22,24,28].**

Use of *A. thaliana* as a model organism

A. thaliana is widely used in plant research such as: 1) Testing mutation frequency where Monroe et al. studied epigenome-associated mutation bias in *A. thaliana* **[43]** , 2) The role of DOG1 and abscisic acid in *A. thaliana* seed dormancy **[31]**, 3) Evolutionary molecular mechanisms that contribute to *A. thaliana* flower formation **[50]**. It is diploid and has a small genome (135 Mb; ~ 27,000 genes), which makes it a favorable model organism for sequencing and gene mapping, even though it has little direct agricultural importance **[34,56]**.

A thaliana has been used to understand biological processes in other plants such as the tomato **[45,48]**. In this study, Mysore et al. expressed tomato genes in *A. thaliana* to analyze their potential function. In the past few decades *A. thaliana* has provided a wealth of information for disease resistance and pathogen susceptibility making it a great model in understanding fundamental plant-microbe interactions **[13,40,48]**. For example, in Cole et al. where they were focused on understanding how *P. simiae* impacts *A. thaliana* root colonization **[13]**.

Bacteria comprise one component of the soil microbiota. Interactions between plants and bacteria could be negative or positive, though there is evidence for possible positive impact between plants and their associated bacterial strains **[4,8,25]**. Several species of bacteria such as, *Pseudomonas fluorescens BSP53a*, *Rhizobia spp, Frankia spp.* and *P. simiae* promote plant growth **[25,38]**. In particular, *P. simiae* WCS417 is a root-colonizing bacteria that can offer *A. thaliana* well-established plant-beneficial effects upon colonization **[67]**. Some researchers sought to better connect gene and gene function of *P. simiae* WCS417 under specific conditions (surface-sterilized, stratified in the dark for 2 to 3 days at 4℃ and grown upright in aPercival incubator), an RB-TnSeq study was conducted **[13]**. Here, they illustrated that a set of 115 genes were deemed essential for root colonization of *A. thaliana* by this strain of plant growth promoting bacteria (PGPB). These examples show the importance of this plant-microbe interaction in agricultural systems.

Multipartite relationship between *P. simiae*, *A. thaliana* roots and phages

Subsequent research conducted by Cole et al. **[13]** phage specific analysis demonstrated that a gene category labeled “other” included *PS417_10145* (phage related hypothetical protein) which was included in the initial 115 genes that were deemed to be essential for root colonization. Reanalysis of these data showed that *PS417_10145* and *PS417_10150* (phage tail tape measure protein) are both predicted to be essential **(Basso.**

Unpublished data). These data provided evidence to suggest that phages may be modulating their hosts as it relates to root colonization, and hints at a possible multipartite relationship between the plant, bacteria and phages **[15]**. *P. simiae* WCS417 has several genes which have been shown to be transducible by bacteriophages, which include Membrane carboxypeptidase (penicillin-binding protein), Flagellar biosynthesis/type III secretory pathway protein FliH, and Flagellar biosynthesis/type III secretory pathway protein FliH **[67]**.

WCS417-colonized roots during the onset of induced systemic resistance (ISR) up-regulates a substantial set of genes that are also up-regulated in roots when plants are grown under conditions of iron deficiency. MYB72 and MYB10 are essential for survival of *Arabidopsis* plants growing in alkaline soils, and are both up-regulated during WCS417 root colonization **[52]**. The main focus of this project was to determine whether select plant associated (PA) and non plant associated (NPA) bacteria share phage homology **(Figure 1)**.

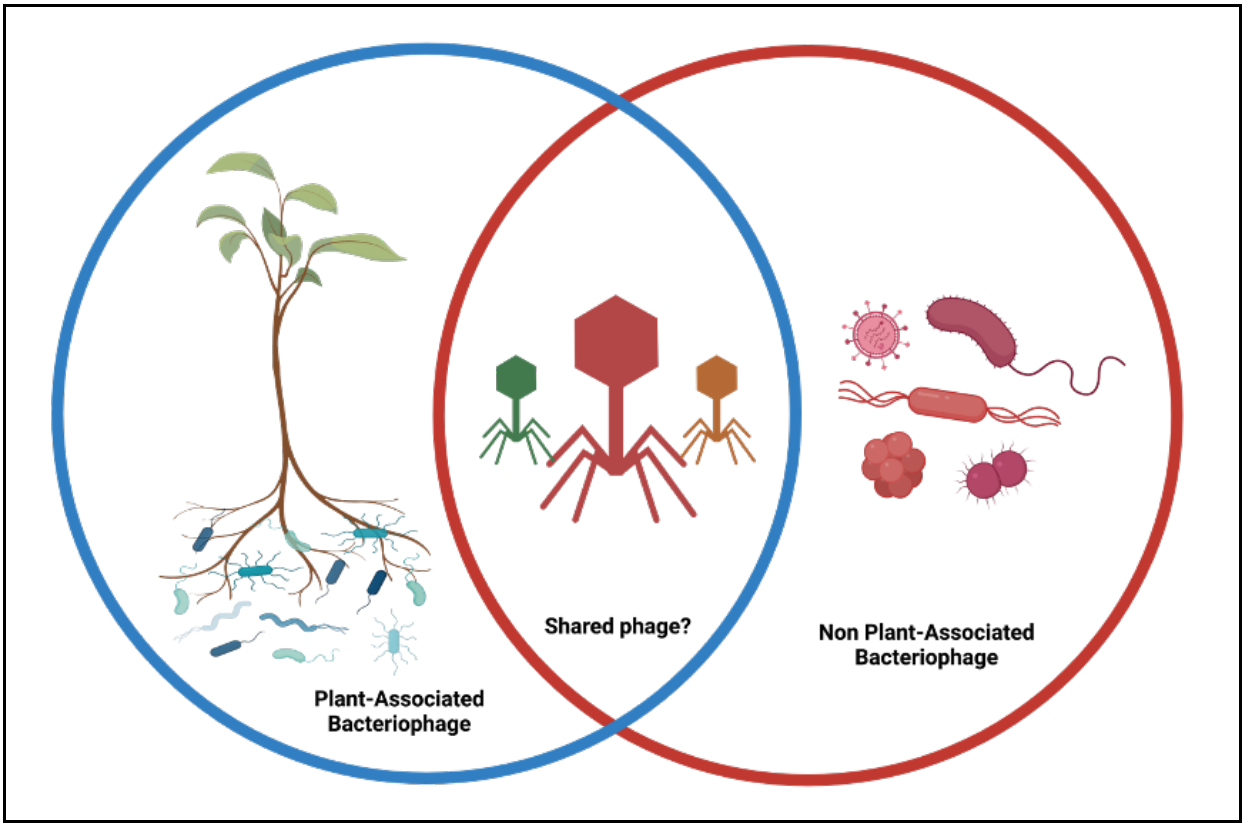


Figure 2: Comparing plant associated and non plant associated bacteriophages

The core of this project can be summarized as the following: 1) Identify phages in select plant-associated (PA) and non-plant-associated (NPA) bacterial WGS. 2) Identify potential impacts of resident bacteriophages in PA and NPA bacteria. 3) Identifying phage homologues.

Whole genome Sequences Reveal Genetic Information

Viral discovery has been revolutionized by metagenomics, which allows computational identification of viral genome sequences without experimentation **[20,58].** Next-generation sequencing (NGS) captures genetic information in the sample's genome as a whole genome sequence **[54]**. As an example, for both the influenza and ebola viruses, scientists were able to track the spread and evolution of these viruses using sequence data **[55,58]**. The computational approaches to predicting putative bacteriophage host ranges can be broadly classified into three categories: alignment-based methods based on sequence homology and sequence similarity, alignment-free methods based on sequence composition and genomic features, and machine-learning-based methods **[65]**. In this study we developed a novel computational pipeline that leverages the above techniques for phage discovery, annotation and phylogeny in order to understand ways phages affect agricultural ecosystems **(Figure 2)**.

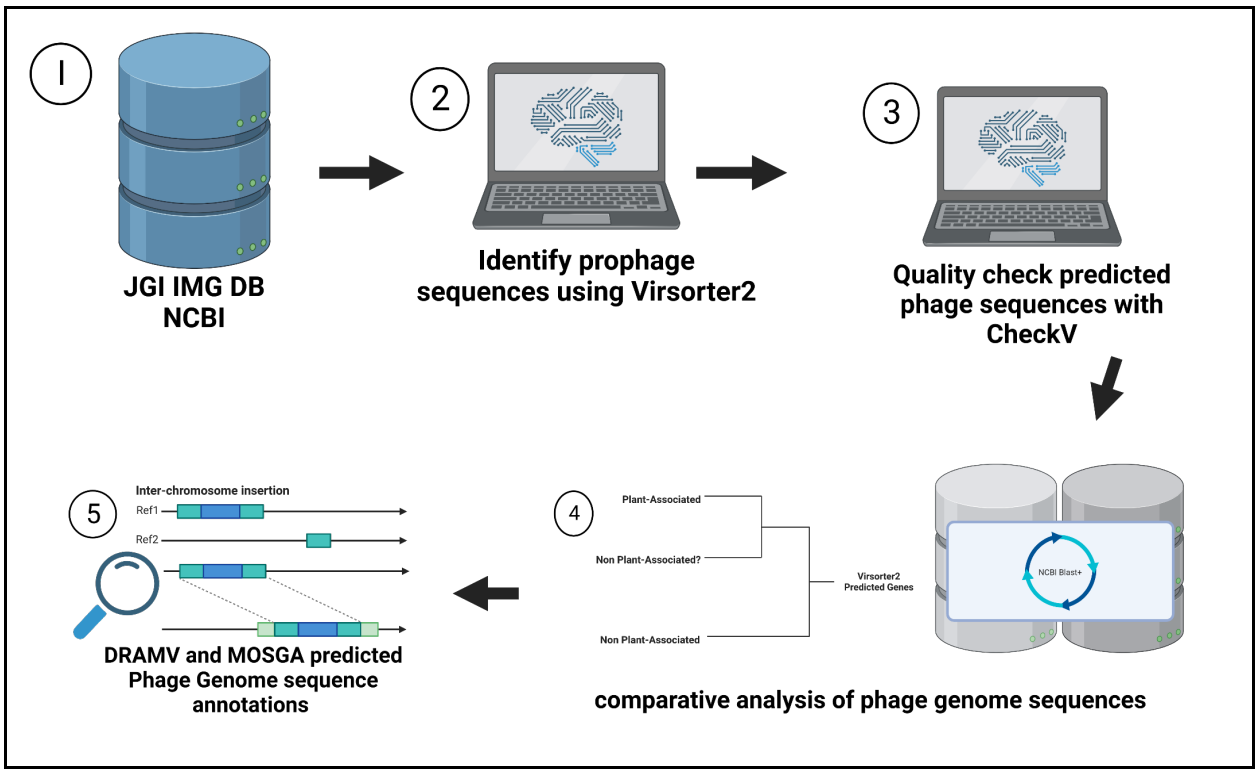


Figure 3: Computational pipeline for phage genome sequence identification

1) Aggregated 20 genome sequences from JGI IMG DB and NCBI 2) Identifying phage in whole-genome sequences with VirSorter2, 3) Quality check predicted phage sequences with CheckV, 4) Phylogenetic comparison using iTOL, Clustal Omega and MUSCLE and 5) Gene annotation with MOSGA **(See Materials and Methods section)**

Materials and Methods

Data acquisition

10 plant associated (PA) and 10 non plant associated (NPA) bacterial whole genome sequences were obtained from the JGI Integrated Microbial Genomes & Microbiomes (IMG/M) **[11]** and The National Center for BioTechnology Information (NCBI) **[7]** databases **(See table 1 and WGS_Accession_Numbers in supplemental data)**. The IMG/M system supports the annotation, analysis and distribution of microbial genome and microbiome datasets sequenced at The Department of Energy (DOE) Joint Genome Institute (JGI) and from other contributing labs and scientists from around the world **[11]**. The National Center for BioTechnology Information (NCBI) GenBank is a comprehensive database that contains publicly available nucleotide sequences for almost 260,000 formally described bacterial species that were collected from laboratories and large-scale sequencing projects **[7].**

Table 1: PA and NPA bacterial genome sequences obtained from IMG database

10 plant associated (PA) and 10 non plant associated bacterial genome sequences were obtained from the JGI IMG and NCBI databases. The plant associated sequences were collected from plant roots, leaves and from the rhizosphere from a diverse set of plants such as wheat, rice and soybeans. The non plant associated bacterial sequences were collected from water columns, human, animal and wastewater facilities. **(WGS_Accession_Numbers in supplemental data)**

Plant Associated	Non Plant Associated
Pseudomonas aeruginosa E2	*Staphylococcus aureus* RF122
Pseudomonas syringae BRIP39023	*Microbacterium luticocti* DSM19459
Acinetobacter sp. UNC436CL71CviS28	*Escherichia coli* K12
Pseudomonas fluorescens A506	*Curtobacterium flaccumfaciens* UCD-AKU
Acinetobacter pittii WP19	*Pseudomonas aeruginosa* PA01
Bradyrhizobium huanghuaihaiense CGMCC 1.10948	*Bartonella henselae* BM1374163
Pseudomonas SP Root9	*Pseudomonas putida* PC9

Pseudomonas simiae WCS417	*Bacillus cereus* G9241
Pseudomonas umsongensis UNC430CL58Col	*Ornithinimicrobium pekingense* DSM21552
Pseudomonas putida KT2440	*Acinetobacter sp. Hugh* 2212, NCTC 10304

Bioinformatic phage detection using VirSorter2

Putative prophage sequences were detected in both PA and NPA genome sets using VirSorter2 **[27]**. VirSorter2 identifies phages by using a random forest classifier that was trained with a viral hmm database **[27]**. The database consists of viral protein families (VPF) of JGI earth's virome project **[51]** and the Xfams database. The Xfams database consists of phylogenetic trees generated from a large collection of viral sequences from the Global Ocean Viromes 2.0 (GOV 2.0) **[26]** and the Stordalen Mire Viromes (SMV) **[21]**.

Quality check of predicted phage sequences with CheckV

CheckV is an automated pipeline for identifying closed viral genomes, estimating the completeness of genome fragments and removing flanking host regions from integrated proviruses **[47]**. CheckV compares to Virus Orthologous Groups (VOGDB), DOE Joint Genome Institute's IMG/VR, Reference Viral DataBase (RVDB), KEGG Orthology, Pfam A, Pfam B and TIGRFAM databases **[47]**.

ViruSITE

ViruSITE is a database of viral genomes and genes. ViruSITE comprises all genomes from viruses, viroids and satellites published in NCBI Reference Sequence Database by computationally extracting from numerous resources (NCBI RefSeq, UniProtKB, GO, ViralZone, PubMed) and integrating under human supervision **[61]**. ViruSite has a total of 11,620 viral sequences, 14,813 genome sequences and 597,210 genes detected from the total 26,433 combined viral and genome sequences. Each of the virSorter2 predicted phage genome sequences were compared to viruSite **[61]** on 2022-11-01.**(See viruSITE and viruSITE_sample table in the supplemental data section).**

Phylogenetic comparison using iTOL, Clustal Omega and MUSCLE

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences **[59]**. Clustal Omega was used to generate a tree guide that was visualized with iTOL. The Interactive Tree Of Life (iTOL) is an online tool for the display, annotation and management of phylogenetic and other trees **[36]**.

MUSCLE

MUSCLE is a tool used for sequence alignment at a quicker speed than clustal **[19]**. VirSorter2 predicted phages that were of high or medium quality determined by checkV,

and the 4 phages that clustal determined had association were analyzed further using MUSCLE **[36]**.

Gene annotation with MOSGA

MOSGA is a tool used for sequence annotation **[41,42]**. MOSGA was used to annotate the 79 virSorter2 identified phages. MOSGA uses EggNog 5 **[29]**, SILVA **[53]** and SwissProt **[3]** databases.

Results

VirSorter2 Identified 79 phages in the 10 plant associated and non plant associated bacterial sequences **(Table 2)**. All identified phages were deemed dsDNA phages by virSorter2. There were 4 whole genome sequences (WGS) that had 0 phages detected. *Acinetobacter sp. Hugh* 2212, NCTC 10304 had a total of 21 phages detected by virSorter2 **(Table 2)**.

Table 2: Identified phage regions in PA and NPA bacterial genome sequences

10 plant associated (PA) and 10 non plant associated bacterial genome sequences were analyzed with virSorter2, revealing 79 predicted phages. 4 bacterial sequences did not have any phages detected. *Acinetobacter sp. Hugh* 2212, NCTC 10304 had the most phages detected for both plant and non plant associated bacterial sequences at 21 **(See figure 6)**.

Plant Associated	#	Non Plant Associated	#
Pseudomonas aeruginosa E2	0	*Staphylococcus aureus* RF122	4
Pseudomonas syringae BRIP39023	0	*Microbacterium luticocti* DSM19459	1
Acinetobacter sp. UNC436CL71CviS28	8	*Escherichia coli* K12	5
Pseudomonas fluorescens A506	3	*Curtobacterium flaccumfaciens* UCD-AKU	2
Acinetobacter pittii WP19	7	*Pseudomonas aeruginosa* PA01	1
Bradyrhizobium huanghuaihaiense CGMCC 1.10948	6	*Bartonella henselae* BM1374163	5
Pseudomonas SP Root9	4	*Pseudomonas putida* PC9	0
Pseudomonas simiae WCS417	3	*Bacillus cereus* G9241	1

Pseudomonas umsongensis UNC430CL58Col	3	*Ornithinimicrobium pekingense* DSM21552	0
Pseudomonas putida KT2440	5	*Acinetobacter sp. Hugh* 2212, NCTC 10304	21

VirSorter2 determines confidence levels for predicted phages

VirSorter2 showed different confidence levels in the identified phage sequences **(Figure 3)**. Each sequence is scored independently using a set of classifiers customized for individual viral groups, and these scores are aggregated into a single prediction as the max score **[27]**. The default score cutoff (0.5) works well known viruses from refSeq **[49]** and was used to analyze all 20 phages. 48 phages had high confidence, meaning the max score > 0.9. VirSorter2 is limited to viral identification only, and is not reliable for taxonomic classification of predicted phages **[27]**.**(See Figure 3)**

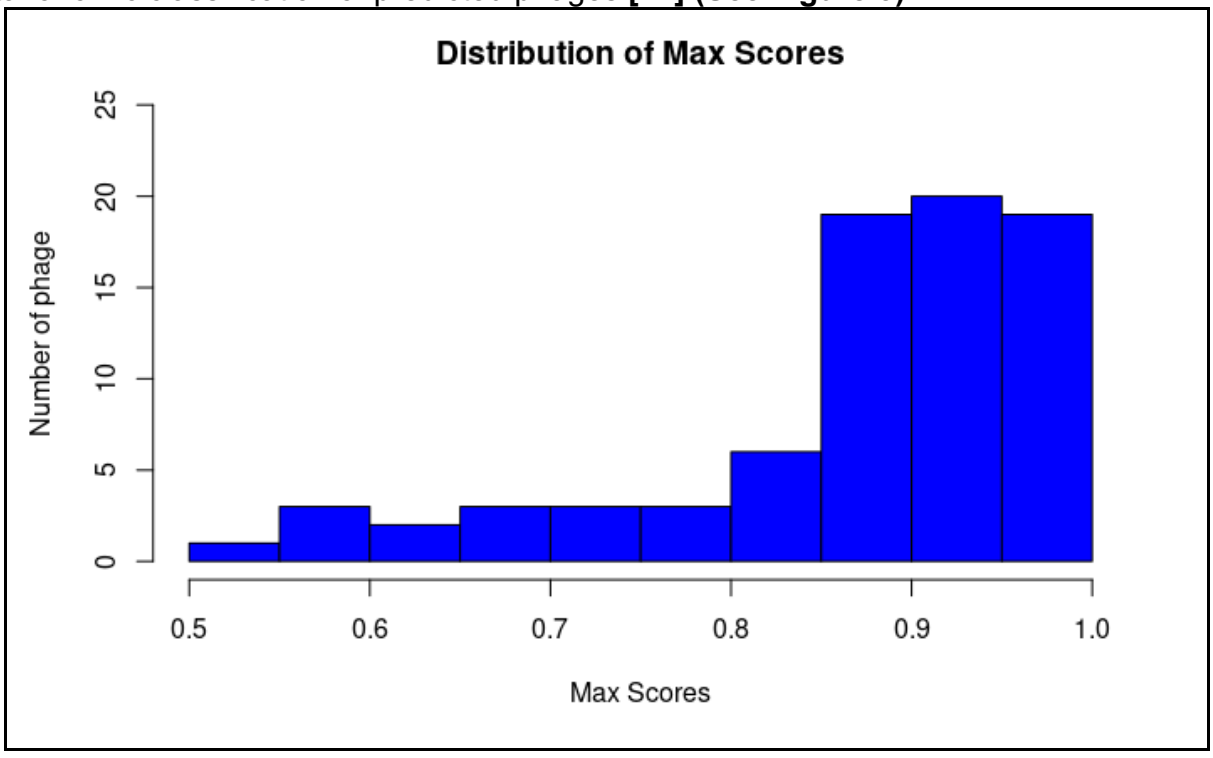


Figure 4: VirSorter2 determines confidence levels for predicted phage sequences

Virsorter2 provided confidence levels for identified phage. The lowest max score is 0.447 and the highest was 1. This shows a mix of low confidence and high confidence in the identified phages. The x axis describes max scores and on the y axis are the total amount of phages **(See Virsorter2_analysis in supplemental materials)**.

CheckV analysis determines 70 phages are of quality

CheckV **[47]** determined that 70 out of the 79 VirSorter2 identified phages were of quality **(Figure 4)**. CheckV calculates alignment scores between each contig and each complete genome in the reference databases **[47]**; The genome length (before fragmentation) of each contig was then grouped based on alignment score and contig length to find the medium relative unassigned error **[47]**. Three confidence levels: high confidence (0–5% median unsigned error), medium confidence (5–10% median unsigned error) and low confidence (>10% median unsigned error) **[47]**. 18 out of 79 predicted phages were deemed high quality, meaning it had 0–5% median unsigned error **(Figure 4)**.

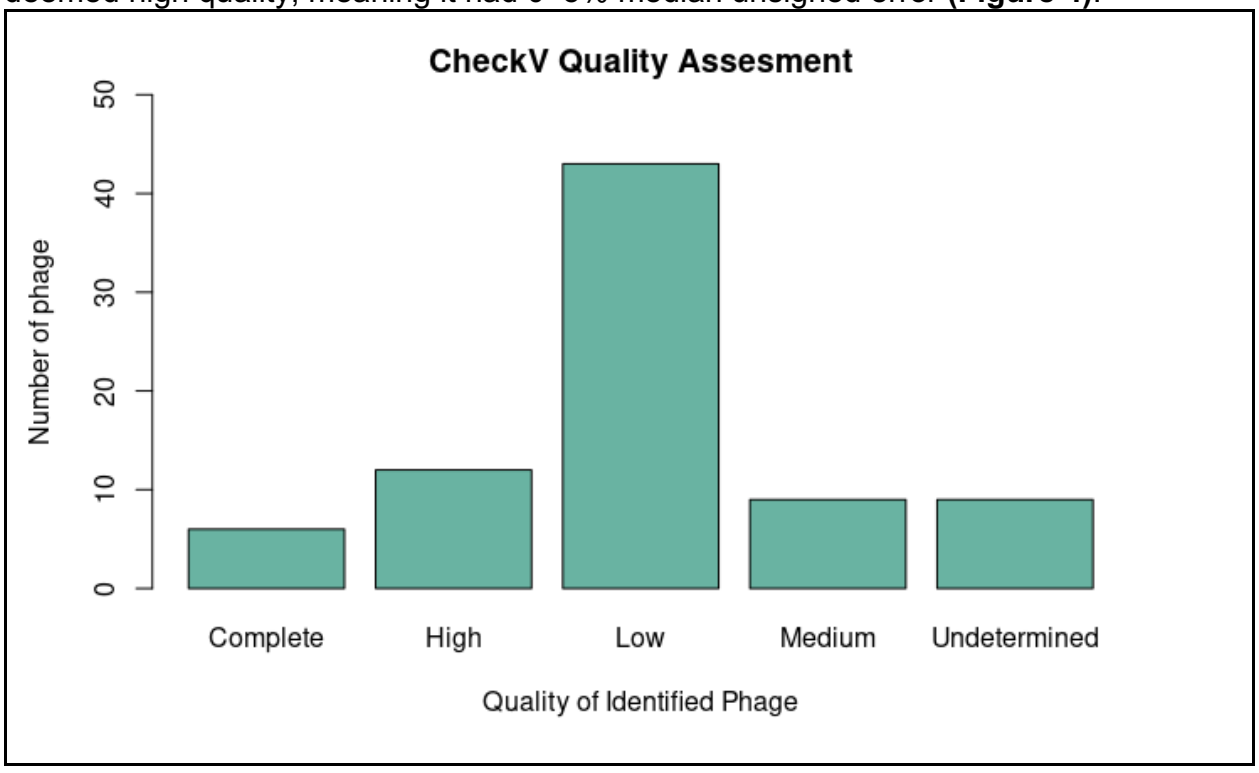


Figure 5: CheckV quality analysis of virSorter2 predicted phage

Using checkV, the quality of the identified phage sequences were checked. There were 4 sequences that did not have any phage detected, and 6 that could not have their quality determined by checkV. There were 6 complete and 12 high quality phages. Out of the 79 detected phages 9 were of medium quality and 43 were low quality phages **(See CheckV_analysis in supplemental materials)**.

ViruSITEand VirSorter2 predicted phages blast analysis

A blast database was created using predicted phage sequences from viruSITE **[61]** and Inphared **[16]**. All predicted phages were compared to ViruSITE and Inphared to see if there were any matches. A total of 18 out of the total 79 virSorter2 identified phages with identity scores greater than 95 were sampled and analyzed closer. There were a total of 2595 matches to viruSITE total **(See viruSITE and viruSITE_sample in supplemental materials)**.

Phylogenetic analysis reveals potential association between sets of phages

Clustal Omega **[59]** and iTOL **[36]** showed association between three sets of phages identified in the 20 bacterial sequences. Two phages identified in PA and NPA bacterial sequences: *A. pittii* WP19 (PA) and *A. sp.* Hugh 2212 NCTC 10304 (NPA), and two plant associated phages *P. sp* Root9 and *P. fluorescens* A506 **(Figure 5)**.

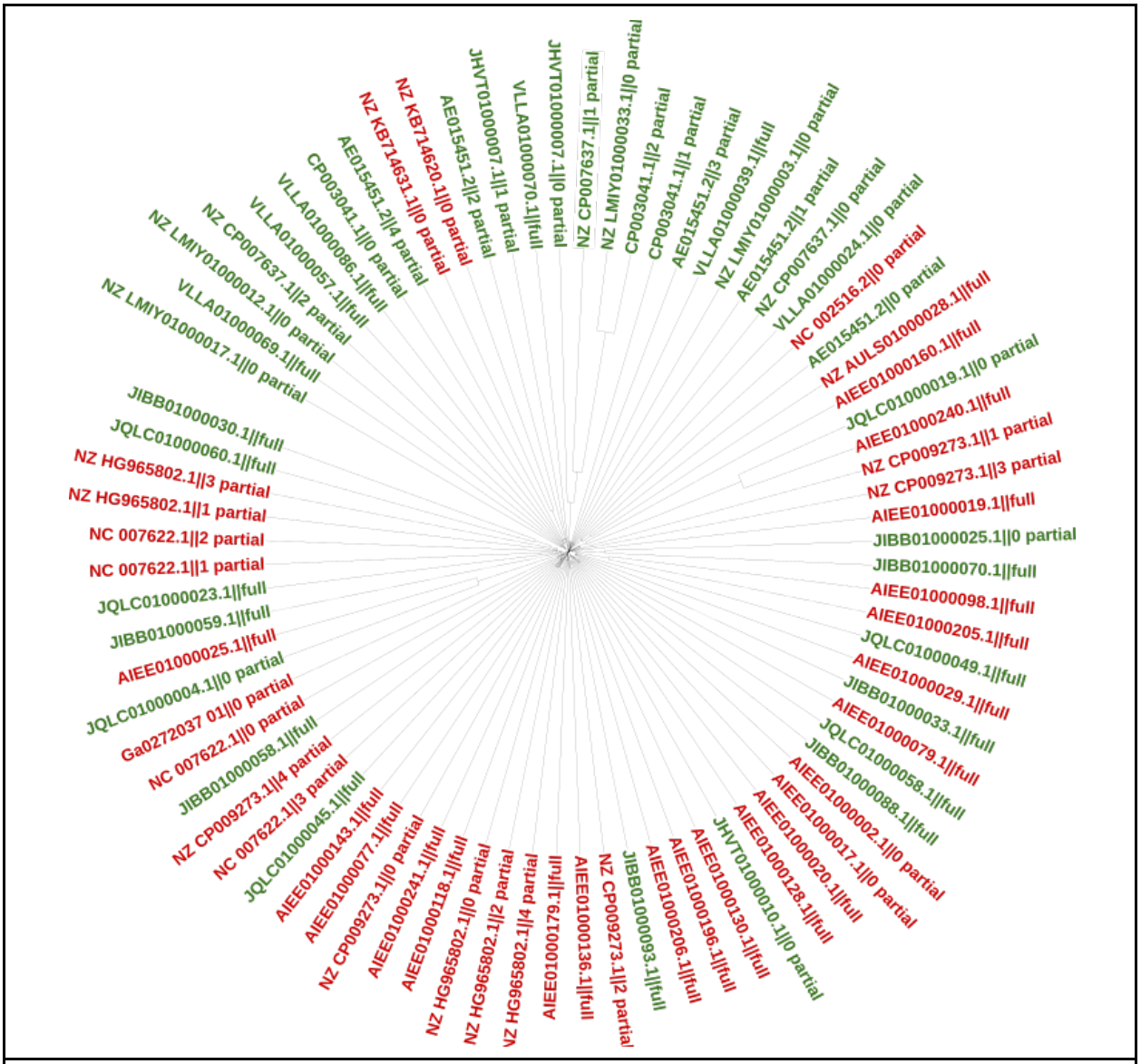


Figure 6: Phylogenetic analysis shows potential association

The mined phage were combined into a single file and processed with clustal omega. A tree guide was produced and uploaded to iTOL illustrating possible similarities between three sets of phages identified in the 20 bacterial sequences:two phages identified in *A. pittii* WP19 (PA) and *A. sp. Hugh* 2212 NCTC 10304 (NPA), and two plant associated phages *P. sp* Root9 and *P. fluorescens* A506 **(see Virsorter2_analysis in supplemental materials)**. Green denotes plant associated phages, while red denotes non plant associated phages.

MOSGA detected 97 total genes in the 79 identified phages

MOSGA **[41,42]** detected a total of 1102 genes. There were 975 repeating regions, 97 protein coding genes, and 30 tRNA genes that make up the total 1102 detected by MOSGA **(See MOSGA_overall_summary in supplemental materials)**. 21 unique

predicted gene functions detected, such as P03764 (Tail fiber protein) showing the presence of a potential phage region in *E. coli* K12, and Q9I1X7(Multifunctional non-homologous end joining protein). **(See MOSGA_detected_genes in supplemental materials)**. MOSGA detected 2766864bp, with the shortest phage contig being 1105bp and the longest at 162996bp.**(See MOSGA_overall_summary in supplemental materials)**.

Discussion

Determining the presence of prophages and exploring the potential association between identified phages in PA and NPA can provide insight into the impacts phages play in root colonization by *P. simiae* and on *A. thalia* growth and development. In this study, 10 plant associated (PA) and 10 non plant associated bacterial genome sequences were analyzed using virSorter2 which predicted 79 phages **(See Figure 6)**.

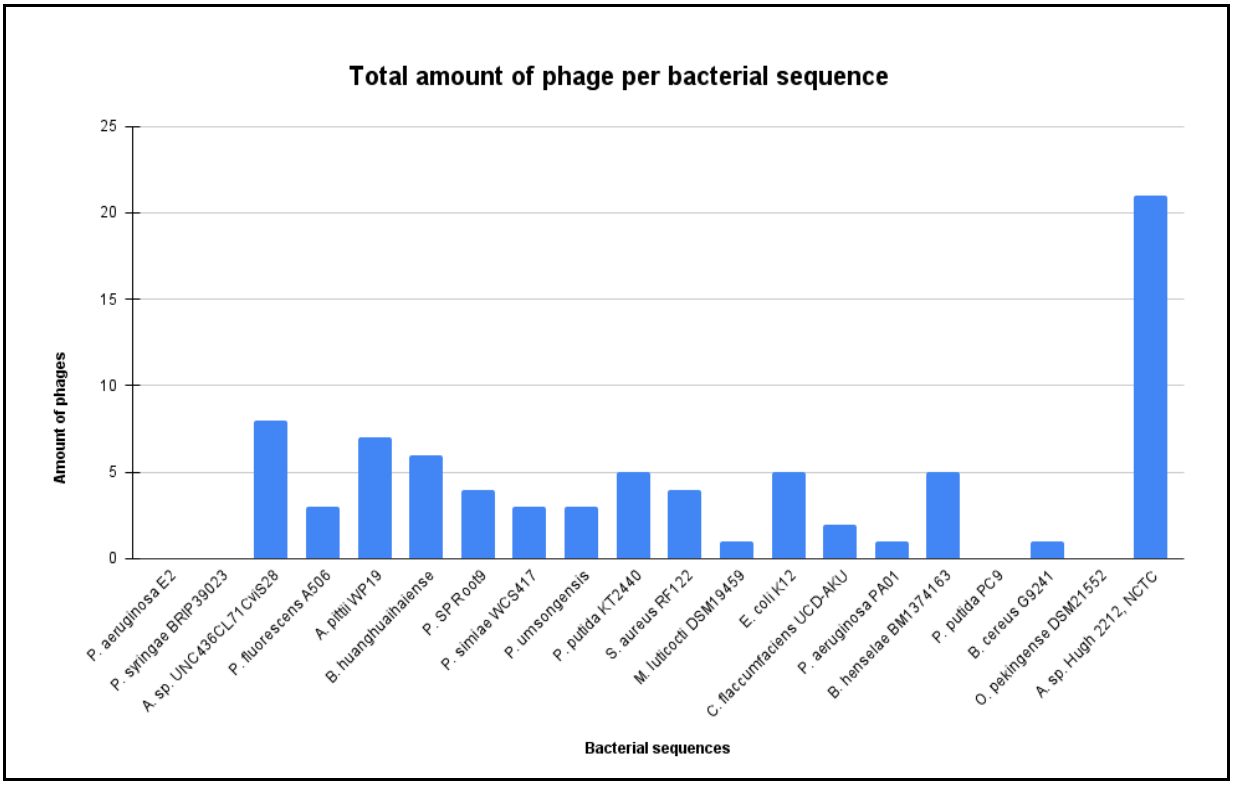


Figure 7: Total amount of phages detected per bacterial species

10 plant associated (PA) and 10 non plant associated bacterial genome sequences were analyzed with virSorter2, revealing 79 predicted phages. The bacterial species is on the x-axis, while the number of phages are along the y-axis. 4 bacterial sequences did not have any phages detected. *Acinetobacter sp. Hugh* 2212, NCTC 10304 had the most phages detected for both plant and non-plant associated bacterial sequences at 21 **(See table 2 and phages_total_per_genome in the supplemental materials)**.

Acinetobacter sp. Hugh 2212, NCTC 10304 has 21 predicted phages

Acinetobacter sp. Hugh 2212, NCTC 10304 is a rod-shaped, gram negative species found in freshwater, host, human skin environments **[10]**. It has a genome size of 4157209 bp with 4463 genes **[10]**. This sequence was sampled from the conjunctiva in 1962 obtained from UK, German and Belgium culture collections **[10]**.
In Davies et al. they observed that in *P. aeruginosa* biofilms, populations that evolved with phages possessed a higher degree of parallel evolution and faster selective sweeps than those without phages present **[17]**. Bacteriophages potentially offer the host beneficial genes that could improve its fitness in biological ecosystems. For example Kittinger et al. performed susceptibility tests on several *Acinetobacter sp.* collected from the Danube River (ICPDR), Vienna where they found several samples were antibiotic resistant **[70]**. In *S. aureus* lysogenic phage are able to transfer the *mecA* gene making it resistant to methicillin based antibiotics **[68]**. Bacteriophages are also able to introduce resistance to *Acinetobacter sp.* **[37]** which gives Hugh 2212, NCTC 10304 reason to integrate so many phages in its genome.
Further analysis was done withMUSCLE, MOSGA which showed potential genes in Hugh 2212, NCTC 10304 such as: CNRA_CUPMC (Nickel and cobalt resistance protein), ACR3_ALKMQ (Arsenical-resistance protein), MERR_PSEAI (Mercuric resistance operon regulatory protein), MERA_STRLI (Mercuric reductase) and TIPJ_LAMBD (Tip attachment protein) detected in virSorter2 predicted phage sequences **(See Hugh_2212_NCTC_10304_MOSGA_genes in supplemental materials)**.
Muscle **[19]** was used to further analyze *Acinetobacter sp.* Hugh *2212*, NCTC 10304 predicted phages **(See Figure 8)**. The average sequence length was 14757bps with the smallest length being 2401bp and the largest at 43291 bp. There were 32 genes identified in Hugh 2212, NCTC 10304, with 13 present in the 21 virSorter2 identified phages **(See Hugh_2212_NCTC_10304_MOSGA_genes in supplemental materials)**.

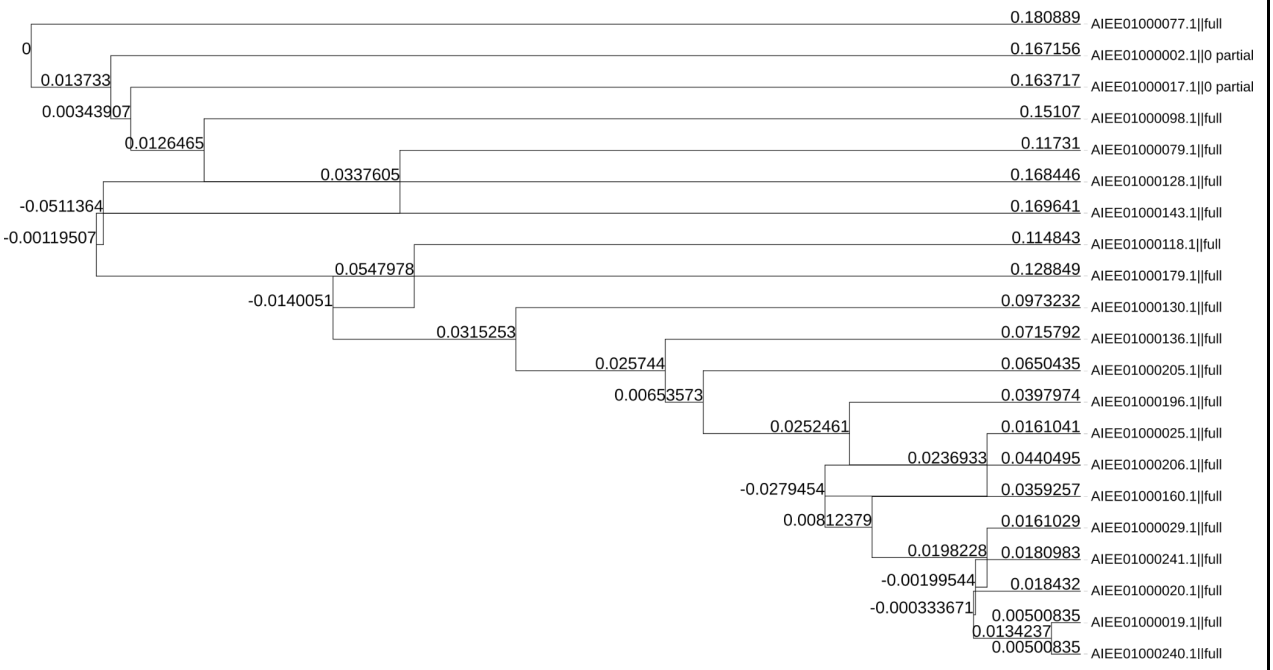


Figure 8: Phylogenetic analysis of predicted phages in NCTC 10304

Acinetobacter sp. Hugh 2212, NCTC 10304 had the most phages detected for both plant and non plant associated bacterial sequences at 21. All 21 phages were analyzed with MUSCLE **[19]** and visualized with iTOL. We found potential associations between a few phages AIEE01000128.1 and AIEE01000143.1, AIEE01000019.1 and AIEE01000240.1.

VirSorter2 phages quality assessment with ViruSITE

The virSorter2 identified phages were compared to ViruSITE **[61]** to do further quality assessment to reduce false positives. There were 2595 total matches to the viruSITE database. A small subset was selected randomly to further analyze **(see Figure 8 and viruSITE, viruSITE_sample in supplemental materials)**. There were 4 bacteria sequences that were identified to have potential association between phages: *A. pittii* WP19 (PA),*A. sp. Hugh 2212* NCTC 10304 (NPA) and *P. sp* Root9 and *P. fluorescens* A506

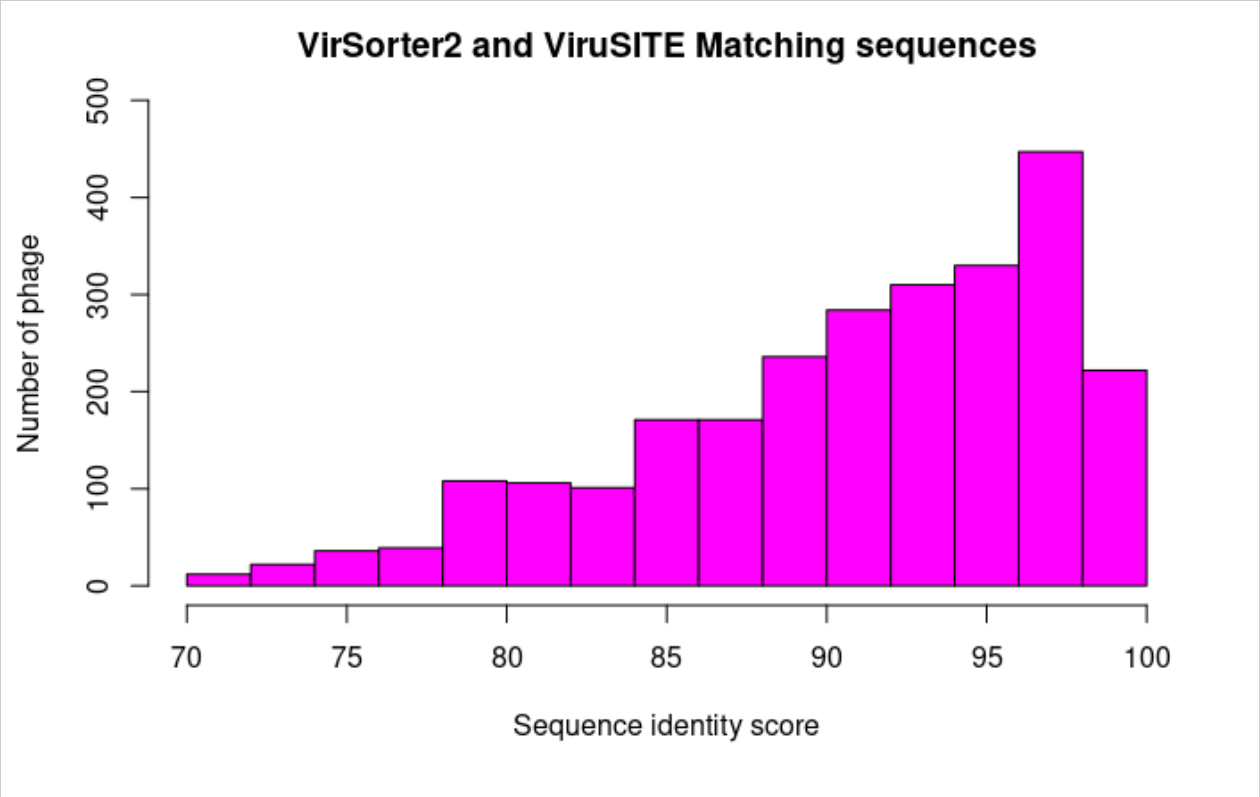


Figure 9: Subset of virSorter2 identified phages with highest identity match to viruSITE

A blast analysis was used to determine identity scores for virSorter2 identified phages that matched viruSITE. The X axis are the identity scores determined by the blast analysis. The Y axis are the number of matching phage sequences to viruSITE. There were a total of 2595 matches in which 383 had an identity score of 98 or above, while 192 had an identity score of 79 or lower. **(see viruSITE and viruSITE_sample in supplemental materials)**

Phylogenetic analysis of predicted phages

Using 1 method for phylogenetic analysis could cause bias in the association between phages. In this study, 2 different methods were used to explore potential association between PA and NPA phages and the bacterial genome sequences they were mined from. Clustal **[59]** and MUSCLE **[19]** are two well-known tools used for sequence alignment and association between sequences. Clustal identified 2 sets of potentially associated phages **(See Figure 6 and Figure 7)**. The 2 sets of potentially associated phages were further analyzed with MUSCLE **[19]**.

A. pittii WP19 (PA) and *A. sp.* Hugh 2212 NCTC 10304 (NPA) shared genes

The *Acinetobacter* species can be found in a wide range of environments such as water, soil and humans. It is used as a model organism for studying ecology in the microbial community due to its ubiquitous presence in nature **[10,32]**. For example,

Indiragandhi et al. found that *Acinetobacter sp.* PSGB04 promoted canola plant root growth through increased root length, seedling vigor, and dry biomass **[30]**. Although *A. pittii* WP19 and *A. sp*. Hugh 2212 NCTC 10304 were collected from different environments, there is potential for them to share phages since they are from the *Acinetobacter* species **[32,37]**.
In *A. pittii* WP19 (PA) and *A. sp.* Hugh 2212 NCTC 10304 (NPA) we found that phage sequences JQLC01000019.1||0_partial and AIEE01000240.1||full were also determined to be associated by MUSCLE **[19](See Figure 11)**. MOSGA analysis showed that phage sequence JQLC01000019.1||0_partial had 4 repeating regions and 1 hypothetical protein, while AIEE01000240.1||full had 2 repeating regions without any genes detected.**(see Virsorter2_analysis in supplemental materials)**

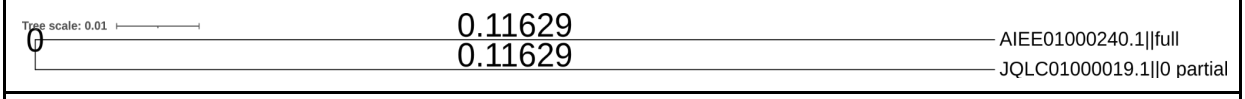


Figure 10: *A. pittii* WP19 and *A. sp.* Hugh 2212 NCTC 10304 phylogenetic analysis

Both phage sequences were analyzed with MUSCLE **[19]** and visualized with iTOL **[36]**. JQLC01000019.1||0_partial and AIEE01000240.1||full both had a score of 0.11629 which shows that MUSCLE predicted both phages are associated.

Plant associated phages *P. sp* Root9 and *P. fluorescens* A506

CheckV **[47]**, clustal omega **[59]**, MUSCLE **[19]** and iTOL **[36]** were used to explore the phages associated with *P. sp* Root9 **[2]** and *P. fluorescens* A506 **[39,62]**. Phage sequences LMIY01000033.1 || 0 partial and CP003041.1||2_partial were also analyzed with MUSCLE **[19]** to see if the results were similar to clustal omega analysis **[59]**. Both phages received an identical score of 0.06299 **(See Figure 10).** Genes SYM_PSEFS (Methionine--tRNA ligase) and LOGL8_ORYSJ (Probable cytokinin riboside 5'-monophosphate phosphoribohydrolase) were identified in LMIY01000033.1 || 0 partial, where it was determined to have a total of 4 genes detected while CP003041.1||2_partial had 4 genes detected as well which included MUTS_ALKEH (DNA mismatch repair protein) and Q47319 (TAPT_ECOLI tRNA-uridine amino carboxypropyl transferase). **(see Virsorter2_analysis in supplemental materials)**

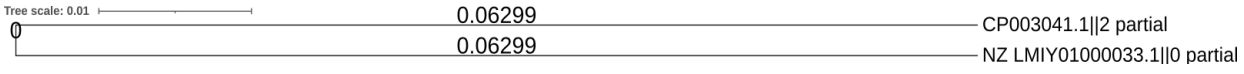


Figure 11: *P. sp* Root9 and *P. fluorescens* A506 Phylogenetic analysis

P. sp Root9 and *P. fluorescens* A506 phage sequences were analyzed with MUSCLE **[19]** and visualized with iTOL **[36]**.Similar to **Figure 10**, MUSCLE predicted that phages LMIY01000033.1 || 0 partial and CP003041.1||2_partial are associated as well. They both share a score of 0.06299.

Conclusion

We developed a novel computational pipeline for phage discovery and analysis and applied this pipeline to compare 10 PA and 10 NPA bacterial genome sequences. In this study, we identified potential association between 2 plant associated phages along with 1 plant associated and 1 non plant associated phages. The next steps are to validate the association between these phages through experimentation. The core goal is to understand if any of the identified genes that are associated with root colonization (or other plant related process) are transducible and if there is really a multipartite relationship between *A. thaliana*, bacteria and bacteriophages in the rhizosphere.

Limitations to computational pipelines

Computational methods offer valuable insight to biological processes but are limited to being predictions. Computational methods were never meant to replace experimental procedures. To fully understand the relationships between phage,bacteria and host we must leverage the insight and guidance from computational methods while performing the experiment.

Computational methods heavily rely on experimental progress. Databases such as NCBI's Genbank **[7]**, JGI IMG DB **[11]**, ViruSite **[61]** are created from the results and data scientists share from their experiments. This study serves as a pilot for leveraging computational pipelines to understand more about the relationship between phages, bacteria and hosts in biological ecosystems. To further leverage computational techniques, a lot more data is collected from samples that have validated annotations through experiments. 20 genome sequences were analyzed closely, which can show potential associations, but to definitively make connections, more sequences need to be analyzed.

Future work

The next steps are a combination of improved computational methods and experimental validation. In regard to computational methods, I plan to : 1) Analyze thousands of bacterial sequences that are known to be associated with plant growth and development such as *P. simiae*; 2) Compare predicted phages to phages identified bacterial species that are able to impact multiple organisms such as *Pseudomonas aeruginosa*, which can impact plants and humans **[66]**; 3) Optimize the pipeline with additional databases, annotation tools and multiple viral detection methods to increase the amount and quality of predicted phages. Having a system that can analyze large volumes of data would assist in leveraging the amount of new sequence data being generated.

In addition to improved computational methods, experimental validation is essential to truly understanding the potential multipartite relationship. The phages that are identified would need to go through experimental methods such as: plaque assays and quantitative PCR (qPCR)s **[5]**, of both bacterial species to see if the phage can infect the host and if the shared genes could be transduced **[1]**.

Chapter 3 References

1. Ács, N., Gambino, M., & Brøndsted, L. (2020). Bacteriophage Enumeration and

Detection Methods. Frontiers in Microbiology, 11, 594868. https://doi.org/10.3389/fmicb.2020.594868
2. Bai, Y., Müller, D. B., Srinivas, G., Garrido-Oter, R., Potthoff, E., Rott, M., Dombrowski, N., Münch, P. C., Spaepen, S., Remus-Emsermann, M., Hüttel, B., McHardy, A. C., Vorholt, J. A., & Schulze-Lefert, P. (2015). Functional overlap of the Arabidopsis leaf and root microbiota. Nature, 528(7582), 364–369. https://doi.org/10.1038/nature16192
3. Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Research, 28(1), 45–48. https://doi.org/10.1093/nar/28.1.45
4. Bais, H. P., Weir, T. L., Perry, L. G., Gilroy, S., & Vivanco, J. M. (2006). The role of root exudates in rhizosphere interactions with plants and other organisms. Annual Review of Plant Biology, 57, 233–266. https://doi.org/10.1146/annurev.arplant.57.032905.105159
5. Basso, J. T. R., Ankrah, N. Y. D., Tuttle, M. J., Grossman, A. S., Sandaa, R.-A., & Buchan, A. (2020). Genetically similar temperate phages form coalitions with their shared host that lead to niche-specific fitness effects. The ISME Journal, 14(7), Article 7. https://doi.org/10.1038/s41396-020-0637-z
6. Beaurepaire, C., & Chaconas, G. (2007). Topology-dependent transcription in linear and circular plasmids of the segmented genome of Borrelia burgdorferi. Molecular Microbiology, 63(2), 443–453. https://doi.org/10.1111/j.1365-2958.2006.05533.x
7. Benson, D., Lipman, D. J., & Ostell, J. (1993). GenBank. Nucleic Acids Research, 21(13), 2963–2965. https://doi.org/10.1093/nar/21.13.2963
8. Buée, M., De Boer, W., Martin, F., van Overbeek, L., & Jurkevitch, E. (2009). The rhizosphere zoo: An overview of plant-associated communities of microorganisms, including phages, bacteria, archaea, and fungi, and of some of their structuring factors. Plant and Soil, 321(1–2), 189–212. https://doi.org/10.1007/s11104-009-9991-3
9. Burmeister, A. R. (2015). Horizontal Gene Transfer. Evolution, Medicine, and Public Health, 2015(1), 193–194. https://doi.org/10.1093/emph/eov018
10. Chan, J. Z.-M., Halachev, M. R., Loman, N. J., Constantinidou, C., & Pallen, M. J. (2012). Defining bacterial species in the genomic era: Insights from the genus Acinetobacter. BMC Microbiology, 12(1), 302. https://doi.org/10.1186/1471-2180-12-302
11. Chen, I.-M. A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Huntemann, M., Varghese, N., White, J. R., Seshadri, R., Smirnova, T., Kirton, E., Jungbluth, S. P., Woyke, T., Eloe-Fadrosh, E. A., Ivanova, N. N., & Kyrpides, N. C. (2019). IMG/M v.5.0: An integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic Acids Research, 47(D1), D666–D677. https://doi.org/10.1093/nar/gky901
12. Clark, J. R., & March, J. B. (2006). Bacteriophages and biotechnology: Vaccines, gene therapy and antibacterials. Trends in Biotechnology, 24(5), 212–218. https://doi.org/10.1016/j.tibtech.2006.03.003
13. Cole, B. J., Feltcher, M. E., Waters, R. J., Wetmore, K. M., Mucyn, T. S., Ryan, E. M., Wang, G., Ul-Hasan, S., McDonald, M., Yoshikuni, Y., Malmstrom, R. R.,

Deutschbauer, A. M., Dangl, J. L., & Visel, A. (2017). Genome-wide identification of bacterial plant colonization genes. PLOS Biology, 15(9), e2002860. https://doi.org/10.1371/journal.pbio.2002860
14. Coleman-Derr, D., & Tringe, S. G. (2014). Building the crops of tomorrow: Advantages of symbiont-based approaches to improving abiotic stress tolerance. Frontiers in Microbiology, 5. https://www.frontiersin.org/articles/10.3389/fmicb.2014.00283
15. Cong, W., Yu, J., Feng, K., Deng, Y., & Zhang, Y. (2021). The Coexistence Relationship Between Plants and Soil Bacteria Based on Interdomain Ecological Network Analysis. Frontiers in Microbiology, 12. https://www.frontiersin.org/articles/10.3389/fmicb.2021.745582
16. Cook, R., Brown, N., Redgwell, T., Rihtman, B., Barnes, M., Clokie, M., Stekel, D. J., Hobman, J., Jones, M. A., & Millard, A. (2021). INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. PHAGE, 2(4), 214–223. https://doi.org/10.1089/phage.2021.0007
17. Davies, E. V., James, C. E., Williams, D., O'Brien, S., Fothergill, J. L., Haldenby, S., Paterson, S., Winstanley, C., & Brockhurst, M. A. (2016). Temperate phages both mediate and drive adaptive evolution in pathogen biofilms. Proceedings of the National Academy of Sciences, 113(29), 8266–8271. https://doi.org/10.1073/pnas.1520056113
18. Du Toit, A. (2017). The language of phages. Nature Reviews Microbiology, 15(3), Article 3. https://doi.org/10.1038/nrmicro.2017.8
19. Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics, 5(1), 113. https://doi.org/10.1186/1471-2105-5-113
20. Edwards, R. A., McNair, K., Faust, K., Raes, J., & Dutilh, B. E. (2016). Computational approaches to predict bacteriophage-host relationships. FEMS Microbiology Reviews, 40(2), 258–272. https://doi.org/10.1093/femsre/fuv048
21. Emerson, J. B., Roux, S., Brum, J. R., Bolduc, B., Woodcroft, B. J., Jang, H. B., Singleton, C. M., Solden, L. M., Naas, A. E., Boyd, J. A., Hodgkins, S. B., Wilson, R. M., Trubl, G., Li, C., Frolking, S., Pope, P. B., Wrighton, K. C., Crill, P. M., Chanton, J. P., … Sullivan, M. B. (2018). Host-linked soil viral ecology along a permafrost thaw gradient. Nature Microbiology, 3(8), Article 8. https://doi.org/10.1038/s41564-018-0190-y
22. Erez, Z., Steinberger-Levy, I., Shamir, M., Doron, S., Stokar-Avihail, A., Peleg, Y., Melamed, S., Leavitt, A., Savidor, A., Albeck, S., Amitai, G., & Sorek, R. (2017). Communication between viruses guides lysis-lysogeny decisions. Nature, 541(7638), 488–493. https://doi.org/10.1038/nature21049
23. Fabian, B. K., Tetu, S. G., & Paulsen, I. T. (2020). Application of Transposon Insertion Sequencing to Agricultural Science. Frontiers in Plant Science, 11. https://www.frontiersin.org/articles/10.3389/fpls.2020.00291
24. Gandon, S. (2016). Why Be Temperate: Lessons from Bacteriophage λ. Trends in Microbiology, 24(5), 356–365. https://doi.org/10.1016/j.tim.2016.02.008
25. Glick, B. R. (2012). Plant growth-promoting bacteria: Mechanisms and applications. Scientifica, 2012, 963401. https://doi.org/10.6064/2012/963401

26. Gregory, A. C., Zayed, A. A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., Dimier, C., Domínguez-Huerta, G., Ferland, J., Kandels, S., Liu, Y., Marec, C., Pesant, S., Picheral, M., Pisarev, S., … Sullivan, M. B. (2019). Marine DNA Viral Macro- and Microdiversity from Pole to Pole. Cell, 177(5), 1109-1123.e14. https://doi.org/10.1016/j.cell.2019.03.040
27. Guo, J., Bolduc, B., Zayed, A. A., Varsani, A., Dominguez-Huerta, G., Delmont, T. O., Pratama, A. A., Gazitúa, M. C., Vik, D., Sullivan, M. B., & Roux, S. (2021). VirSorter2: A multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome, 9(1), 37. https://doi.org/10.1186/s40168-020-00990-y
28. Howard-Varona, C., Hargreaves, K. R., Abedon, S. T., & Sullivan, M. B. (2017). Lysogeny in nature: Mechanisms, impact and ecology of temperate phages. The ISME Journal, 11(7), 1511–1520. https://doi.org/10.1038/ismej.2017.16
29. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research, 47(D1), D309–D314. https://doi.org/10.1093/nar/gky1085
30. Indiragandhi, P., Anandham, R., Madhaiyan, M., & Sa, T. M. (2008). Characterization of Plant Growth–Promoting Traits of Bacteria Isolated from Larval Guts of Diamondback Moth Plutella xylostella (Lepidoptera: Plutellidae). Current Microbiology, 56(4), 327–333. https://doi.org/10.1007/s00284-007-9086-4
31. Iwasaki, M., Penfield, S., & Lopez-Molina, L. (2022). Parental and Environmental Control of Seed Dormancy in Arabidopsis thaliana. Annual Review of Plant Biology, 73(1), 355–378. https://doi.org/10.1146/annurev-arplant-102820-090750
32. Jung, J., & Park, W. (2015). Acinetobacter species as model microorganisms in environmental microbiology: Current state and perspectives. Applied Microbiology and Biotechnology, 99(6), 2533–2548. https://doi.org/10.1007/s00253-015-6439-y
33. Kittinger, C., Kirschner, A., Lipp, M., Baumert, R., Mascher, F., Farnleitner, A. H., & Zarfel, G. E. (2018). Antibiotic Resistance of Acinetobacter spp. Isolates from the River Danube: Susceptibility Stays High. International Journal of Environmental Research and Public Health, 15(1), Article 1. https://doi.org/10.3390/ijerph15010052
34. Krämer, U. (2015). Planting molecular functions in an ecological context with Arabidopsis thaliana. ELife, 4, e06100. https://doi.org/10.7554/eLife.06100
35. Kutter, E., & Sulakvelidze, A. (Eds.). (2004). Bacteriophages: Biology and Applications. CRC Press. https://doi.org/10.1201/9780203491751
36. Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. Nucleic Acids Research, 49(W1), W293–W296. https://doi.org/10.1093/nar/gkab301
37. Leungtongkam, U., Thummeepak, R., Kitti, T., Tasanapak, K., Wongwigkarn, J., Styles, K. M., Wellington, E. M. H., Millard, A. D., Sagona, A. P., & Sitthisak, S. (2020). Genomic analysis reveals high virulence and antibiotic resistance

amongst phage susceptible Acinetobacter baumannii. Scientific Reports, 10, 16154. https://doi.org/10.1038/s41598-020-73123-y
38. Levy, A., Salas Gonzalez, I., Mittelviefhaus, M., Clingenpeel, S., Herrera Paredes, S., Miao, J., Wang, K., Devescovi, G., Stillman, K., Monteiro, F., Rangel Alvarez, B., Lundberg, D. S., Lu, T.-Y., Lebeis, S., Jin, Z., McDonald, M., Klein, A. P., Feltcher, M. E., Rio, T. G., … Dangl, J. L. (2018). Genomic features of bacterial adaptation to plants. Nature Genetics, 50(1), Article 1. https://doi.org/10.1038/s41588-017-0012-9
39. Loper, J. E., Hassan, K. A., Mavrodi, D. V., Davis, E. W., Lim, C. K., Shaffer, B. T., Elbourne, L. D. H., Stockwell, V. O., Hartney, S. L., Breakwell, K., Henkels, M. D., Tetu, S. G., Rangel, L. I., Kidarsa, T. A., Wilson, N. L., van de Mortel, J. E., Song, C., Blumhagen, R., Radune, D., … Paulsen, I. T. (2012). Comparative genomics of plant-associated Pseudomonas spp.: Insights into diversity and inheritance of traits involved in multitrophic interactions. PLoS Genetics, 8(7), e1002784. https://doi.org/10.1371/journal.pgen.1002784
40. Lundberg, D. S., Lebeis, S. L., Paredes, S. H., Yourstone, S., Gehring, J., Malfatti, S., Tremblay, J., Engelbrektson, A., Kunin, V., Rio, T. G. del, Edgar, R. C., Eickhorst, T., Ley, R. E., Hugenholtz, P., Tringe, S. G., & Dangl, J. L. (2012). Defining the core Arabidopsis thaliana root microbiome. Nature, 488(7409), Article 7409. https://doi.org/10.1038/nature11237
41. Martin, R., Dreßler, H., Hattab, G., Hackl, T., Fischer, M. G., & Heider, D. (2021). MOSGA 2: Comparative genomics and validation tools. Computational and Structural Biotechnology Journal, 19, 5504–5509. https://doi.org/10.1016/j.csbj.2021.09.024
42. Martin, R., Hackl, T., Hattab, G., Fischer, M. G., & Heider, D. (2020). MOSGA: Modular Open-Source Genome Annotator. Bioinformatics, 36(22–23), 5514–5515. https://doi.org/10.1093/bioinformatics/btaa1003
43. Monroe, J. G., Srikant, T., Carbonell-Bejerano, P., Becker, C., Lensink, M., Exposito-Alonso, M., Klein, M., Hildebrandt, J., Neumann, M., Kliebenstein, D., Weng, M.-L., Imbert, E., Ågren, J., Rutter, M. T., Fenster, C. B., & Weigel, D. (2022). Mutation bias reflects natural selection in Arabidopsis thaliana. Nature, 602(7895), Article 7895. https://doi.org/10.1038/s41586-021-04269-6
44. Morella, N. M., Gomez, A. L., Wang, G., Leung, M. S., & Koskella, B. (2018). The impact of bacteriophages on phyllosphere bacterial abundance and composition. Molecular Ecology, 27(8), 2025–2038. https://doi.org/10.1111/mec.14542
45. Mysore, K. S., Tuori, R. P., & Martin, G. B. (2001). Arabidopsis genome sequence as a tool for functional genomics in tomato. Genome Biology, 2(1), reviews1003.1. https://doi.org/10.1186/gb-2001-2-1-reviews1003
46. Naureen, Z., Dautaj, A., Anpilogov, K., Camilleri, G., Dhuli, K., Tanzi, B., Maltese, P. E., Cristofoli, F., Antoni, L. D., Beccari, T., Dundar, M., & Bertelli, M. (2020). Bacteriophages presence in nature and their role in the natural selection of bacterial populations. Acta Biomedica Atenei Parmensis, 91(13-S), Article 13-S. https://doi.org/10.23750/abm.v91i13-S.10819
47. Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., & Kyrpides, N. C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nature Biotechnology, 39(5), Article 5.

https://doi.org/10.1038/s41587-020-00774-7
48. Nishimura, M. T., & Dangl, J. L. (2010). Arabidopsis and the plant immune system. The Plant Journal, 61(6), 1053–1066. https://doi.org/10.1111/j.1365-313X.2010.04131.x
49. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Research, 44(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189
50. Ó'Maoiléidigh, D. S., Graciet, E., & Wellmer, F. (2014). Gene networks controlling Arabidopsis thaliana flower development. New Phytologist, 201(1), 16–30. https://doi.org/10.1111/nph.12444
51. Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N. N., & Kyrpides, N. C. (2016). Uncovering Earth's virome. Nature, 536(7617), Article 7617. https://doi.org/10.1038/nature19094
52. Palmer, C. M., Hindt, M. N., Schmidt, H., Clemens, S., & Guerinot, M. L. (2013). MYB10 and MYB72 are required for growth under iron-limiting conditions. PLoS Genetics, 9(11), e1003953. https://doi.org/10.1371/journal.pgen.1003953
53. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research, 41(Database issue), D590-596. https://doi.org/10.1093/nar/gks1219
54. Quer, J., Colomer-Castell, S., Campos, C., Andrés, C., Piñana, M., Cortese, M. F., González-Sánchez, A., Garcia-Cehic, D., Ibáñez, M., Pumarola, T., Rodríguez-Frías, F., Antón, A., & Tabernero, D. (2022). Next-Generation Sequencing for Confronting Virus Pandemics. Viruses, 14(3), Article 3. https://doi.org/10.3390/v14030600
55. Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., Bore, J. A., Koundouno, R., Dudas, G., Mikhail, A., Ouédraogo, N., Afrough, B., Bah, A., Baum, J. H., Becker-Ziaja, B., Boettcher, J.-P., Cabeza-Cabrerizo, M., Camino-Sanchez, A., Carter, L. L., … Carroll, M. W. (2016). Real-time, portable genome sequencing for Ebola surveillance. Nature, 530(7589), 228–232. https://doi.org/10.1038/nature16996
56. Rafique, F., Lauersen, K. J., Chodasiewicz, M., & Figueroa, N. E. (2022). A New Approach to the Study of Plastidial Stress Granules: The Integrated Use of Arabidopsis thaliana and Chlamydomonas reinhardtii as Model Organisms. Plants, 11(11), Article 11. https://doi.org/10.3390/plants11111467
57. Schippers, B., Bakker, A. W., & Bakker, P. A. H. M. (1987). Interactions of Deleterious and Beneficial Rhizosphere Microorganisms and the Effect of Cropping Practices. Annual Review of Phytopathology, 25(1), 339–358. https://doi.org/10.1146/annurev.py.25.090187.002011
58. Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance, 22(13), 30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494

59. Sievers, F., & Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. Protein Science, 27(1), 135–145. https://doi.org/10.1002/pro.3290
60. Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. Current Protocols in Molecular Biology, 122(1), e59. https://doi.org/10.1002/cpmb.59
61. Stano, M., Beke, G., & Klucar, L. (2016). ViruSITE-integrated database for viral genomics. Database: The Journal of Biological Databases and Curation, 2016, baw162. https://doi.org/10.1093/database/baw162
62. Stockwell, V. O., Davis, E. W., Carey, A., Shaffer, B. T., Mavrodi, D. V., Hassan, K. A., Hockett, K., Thomashow, L. S., Paulsen, I. T., & Loper, J. E. (2013). PA506, a conjugative plasmid of the plant epiphyte Pseudomonas fluorescens A506. Applied and Environmental Microbiology, 79(17), 5272–5282. https://doi.org/10.1128/AEM.01354-13
63. Stone, E., Campbell, K., Grant, I., & McAuliffe, O. (2019). Understanding and Exploiting Phage-Host Interactions. Viruses, 11(6), E567. https://doi.org/10.3390/v11060567
64. Sulakvelidze, A., Alavidze, Z., & Morris, J. G. (2001). Bacteriophage Therapy. Antimicrobial Agents and Chemotherapy, 45(3), 649–659. https://doi.org/10.1128/AAC.45.3.649-659.2001
65. Versoza, C. J., & Pfeifer, S. P. (2022). Computational Prediction of Bacteriophage Host Ranges. Microorganisms, 10(1), Article 1. https://doi.org/10.3390/microorganisms10010149
66. Walker, T. S., Bais, H. P., Déziel, E., Schweizer, H. P., Rahme, L. G., Fall, R., & Vivanco, J. M. (2004). Pseudomonas aeruginosa-Plant Root Interactions. Pathogenicity, Biofilm Formation, and Root Exudation. Plant Physiology, 134(1), 320–331. https://doi.org/10.1104/pp.103.027888
67. Yu, K., Stringlis, I. A., van Bentum, S., de Jonge, R., Snoek, B. L., Pieterse, C. M. J., Bakker, P. A. H. M., & Berendsen, R. L. (2021). Transcriptome Signatures in Pseudomonas simiae WCS417 Shed Light on Role of Root-Secreted Coumarins in Arabidopsis-Mutualist Communication. Microorganisms, 9(3), 575. https://doi.org/10.3390/microorganisms9030575
68. Zeman, M., Mašlaňová, I., Indráková, A., Šiborová, M., Mikulášek, K., Bendíčková, K., Plevka, P., Vrbovská, V., Zdráhal, Z., Doškař, J., & Pantůček, R. (2017). Staphylococcus sciuri bacteriophages double-convert for staphylokinase and phospholipase, mediate interspecies plasmid transduction, and package mecA gene. Scientific Reports, 7(1), 46319. https://doi.org/10.1038/srep46319
69. Zhan, Y., Huang, S., Voget, S., Simon, M., & Chen, F. (2016). A novel roseobacter phage possesses features of podoviruses, siphoviruses, prophages and gene transfer agents. Scientific Reports, 6(1), Article 1. https://doi.org/10.1038/srep30372
70. Kittinger, C., Kirschner, A., Lipp, M., Baumert, R., Mascher, F., Farnleitner, A. H., & Zarfel, G. E. (2018). Antibiotic Resistance of Acinetobacter spp. Isolates from the River Danube: Susceptibility Stays High. International Journal of Environmental Research and Public Health, 15(1), Article 1. https://doi.org/10.3390/ijerph15010052

Chapter 4: Conclusion

Bacteriophages are able to affect both abiotic and biotic components making them a force of biodiversity in ecosystems **[29,41]**. In this dissertation, I was able to leverage novel computational pipelines to identify bacteriophages in WGS, determine the genes present in the identified phages, and performed phylogenetic analysis to see if the phages had any potential association.

In host ecosystems bacteriophages impact the host through the genes that they could potentially introduce. I've identified 191 unique predicted PhiSpy in 10,011 *S. aureus* genome sequences, with 3205 genes detected. Bacterial hosts have been observed to accept phages into their genome when there are genes that benefit the host **[6,19]**. In my database there were 2 phages that had signs of resistance for *S. aureus*. In agricultural ecosystems, there were phages that had genes that offered host resistance to toxic metals in the environment. Bacteria can gain several beneficial genes from bacteriophages in the ecosystem. The genes introduced by phages in the environment can offer several evolutionary traits needed to improve fitness **[50,51,52]**.

Genes can be beneficial to both bacterial hosts, and other host organisms such as plants by helping them improve their fitness in their environment **[50,51]**. These genes have the potential to be spread by the bacteria, phages and plants found in the environment. In this dissertation we've identified 16 phages that were of association in agricultural ecosystems. Several phages that were associated were identified in bacteria that were from the same species. The identified phages found in the *Acinetobacter* species were both plant associated and non plant associated. This shows that a species that is commonly found in several different ecosystems can potentially be spreading phages carrying genes from various backgrounds **[51]**. Genes that are beneficial to biotic components have the ability to be shared by similarly associated phages. In agricultural ecosystems, plant associated bacterial species had several identified phages associated with each other. The bacterial species could potentially be selecting for phages that offer genes that are beneficial to both it and the plant to increase fitness in the ecosystem. In different ecosystems, *E. coli* and *S. aureus* are two different bacterial species that have been shown to share phages **[52]**. The combination of my pipeline and database can be leveraged with experimental analysis to uncover more of these potential associations and track how genes are being passed in different ecosystems.

The database produced by my pipeline allows exploration of the origin for identified phages, potential susceptibility of bacterial species being analyzed and prevalence of genes commonly shared between the host and phages in an ecosystem. Alignment tools such as MUSCLE and Clustal Omega give insight on how similar sequences are to each other **[13,36]**. This information, along with annotated genes identified in the phage sequences can be used to explore the ancestry of the different identified phages. Chapter 3 introduces this method when exploring the possibility of phages sharing genes in an ecosystem. The genes most commonly found for bacterial species and phages in different ecosystems uncovers the role of the identified phages that could be modulating bacterial fitness **[50]**. One example in agricultural ecosystems, are the toxic metals that can be found in the rhizosphere. The *Acinetobacter* phages identified in my database had genes associated with mercury and arsenic resistance showing that the samples collected from a pear tree, and the rhizosphere could have been sampled from an environment suffering from metal toxicity. The computational pipeline and database produced can be leveraged

to explore the above areas computationally without having to perform experiments initially to make the potential connections and associations.

Limitations to computational pipelines

A significant proportion of the genes encoded by both free living and prophage sequences are of unknown function **[48]**. Databases that scientists are updating with gene functions from experiments conducted serves a better foundation for gene annotation tools. The databases are limited to what scientists discover in genomics overall and this puts a major constraint on the databases. This could introduce a level of bias in the tools that are using the same databases. To create tools that are able to identify more phages and genes, we will need to continue increasing the amount of experimentally validated sequences.

Computational methods offer valuable insight to biological processes but are limited to being predictions. To fully understand the relationships between phage,bacteria and host we must leverage the insight and guidance from computational methods while performing the experiment. This dissertation serves as a pilot for leveraging computational pipelines to understand more about the relationship between phages, bacteria and hosts in biological ecosystems. To further leverage computational techniques, a lot more data is collected from samples that have validated annotations through experiments. I've analyzed 10,011 sequences for S. aureus and 20 for plant associated and non plant associated genome sequences, which showed potential associations, but to definitively make connections, more sequences need to be analyzed.

Computational methods will not be able to replace experimental procedures. They help uncover insight experiments may have overlooked. Relying on one method is limiting, while multiple methods combined reduces the chance of bias **(See Chapter 1: Figure 1)**. Experimental procedures yield more results when both sequencing and computational methods are leveraged **(See Figure 1)[15,2425]**.

Future work

Future work includes a combination of improved computational methods and experimental validation. In regard to computational methods: 1) Expand the computational pipeline to leverage more tools for phage identification, gene annotation and phylogenetic analysis; 2) Optimize the pipeline with additional databases similar to viruSITE **[39]** and INPHRED **[49]**. Having a system that can analyze large volumes of data would assist in leveraging the amount of new sequence data being generated.

In addition to improved computational methods, experimental validation is essential to truly understanding the potential multipartite relationship. The phages that are identified would need to go through experimental methods such as: plaque assays and quantitative PCR (qPCR)s **[50]**, of both bacterial species to see if the phage can infect the host and if the shared genes could be transduced **[2]**. Exploring the impacts of bacteriophages in different ecosystems will require the strengths of sequencing and computational methods to assist in observing more processes during experiments.

References

1. Ackermann, H. W. (2011). Bacteriophage taxonomy. Microbiology Australia, 32(2), 90. https://doi.org/10.1071/MA11090
2. Ács, N., Gambino, M., & Brøndsted, L. (2020). Bacteriophage Enumeration and Detection Methods. Frontiers in Microbiology, 11, 594868. https://doi.org/10.3389/fmicb.2020.594868
3. Akhter, S., Aziz, R. K., & Edwards, R. A. (2012). PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Research, 40(16), e126–e126. https://doi.org/10.1093/nar/gks406
4. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
5. Bacteriophage. (n.d.). Retrieved November 19, 2022, from https://www.microbiologybook.org/mayer/phage.htm
6. Batinovic, S., Wassef, F., Knowler, S. A., Rice, D. T. F., Stanton, C. R., Rose, J., Tucci, J., Nittami, T., Vinh, A., Drummond, G. R., Sobey, C. G., Chan, H. T., Seviour, R. J., Petrovski, S., & Franks, A. E. (2019). Bacteriophages in Natural and Artificial Environments. Pathogens, 8(3), 100. https://doi.org/10.3390/pathogens8030100
7. Benson, D., Lipman, D. J., & Ostell, J. (1993). GenBank. Nucleic Acids Research, 21(13), 2963–2965. https://doi.org/10.1093/nar/21.13.2963
8. Burmeister, A. R. (2015). Horizontal Gene Transfer. Evolution, Medicine, and Public Health, 2015(1), 193–194. https://doi.org/10.1093/emph/eov018
9. Camacho, C., Madden, T., Coulouris, G., Ma, N., Tao, T., Agarwala, R., & Morgulis, A. (n.d.). BLAST Command Line Applications User Manual. 37.
10. Chen, I.-M. A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Huntemann, M., Varghese, N., White, J. R., Seshadri, R., Smirnova, T., Kirton, E., Jungbluth, S. P., Woyke, T., Eloe-Fadrosh, E. A., Ivanova, N. N., & Kyrpides, N. C. (2019). IMG/M v.5.0: An integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic Acids Research, 47(D1), D666–D677. https://doi.org/10.1093/nar/gky901
11. Dion, M. B., Oechslin, F., & Moineau, S. (2020). Phage diversity, genomics and phylogeny. Nature Reviews Microbiology, 18(3), Article 3. https://doi.org/10.1038/s41579-019-0311-5
12. Du Toit, A. (2017). The language of phages. Nature Reviews Microbiology, 15(3), Article 3. https://doi.org/10.1038/nrmicro.2017.8
13. Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics, 5(1), 113. https://doi.org/10.1186/1471-2105-5-113
14. Edwards, R. A., McNair, K., Faust, K., Raes, J., & Dutilh, B. E. (2016). Computational approaches to predict bacteriophage-host relationships. FEMS Microbiology Reviews, 40(2), 258–272. https://doi.org/10.1093/femsre/fuv048
15. Endy, D., You, L., Yin, J., & Molineux, I. J. (2000). Computation, prediction, and experimental tests of fitness for bacteriophage T7 mutants with permuted genomes. Proceedings of the National Academy of Sciences, 97(10), 5375–

5380. https://doi.org/10.1073/pnas.090101397
16. Erez, Z., Steinberger-Levy, I., Shamir, M., Doron, S., Stokar-Avihail, A., Peleg, Y., Melamed, S., Leavitt, A., Savidor, A., Albeck, S., Amitai, G., & Sorek, R. (2017). Communication between viruses guides lysis-lysogeny decisions. Nature, 541(7638), 488–493. https://doi.org/10.1038/nature21049
17. Fabian, B. K., Tetu, S. G., & Paulsen, I. T. (2020). Application of Transposon Insertion Sequencing to Agricultural Science. Frontiers in Plant Science, 11. https://www.frontiersin.org/articles/10.3389/fpls.2020.00291
18. Fermin, G. (2018). Host Range, Host–Virus Interactions, and Virus Transmission. Viruses, 101–134. https://doi.org/10.1016/B978-0-12-811257-1.00005-X
19. Gandon, S. (2016). Why Be Temperate: Lessons from Bacteriophage λ. Trends in Microbiology, 24(5), 356–365. https://doi.org/10.1016/j.tim.2016.02.008
20. Glonti, T., & Pirnay, J.-P. (2022). In Vitro Techniques and Measurements of Phage Characteristics That Are Important for Phage Therapy Success. Viruses, 14(7), 1490. https://doi.org/10.3390/v14071490
21. Guo, J., Bolduc, B., Zayed, A. A., Varsani, A., Dominguez-Huerta, G., Delmont, T. O., Pratama, A. A., Gazitúa, M. C., Vik, D., Sullivan, M. B., & Roux, S. (2021). VirSorter2: A multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome, 9(1), 37. https://doi.org/10.1186/s40168-020-00990-y
22. Hatfull, G. F., & Hendrix, R. W. (2011). Bacteriophages and their Genomes. Current Opinion in Virology, 1(4), 298–303. https://doi.org/10.1016/j.coviro.2011.06.009
23. Howard-Varona, C., Hargreaves, K. R., Abedon, S. T., & Sullivan, M. B. (2017). Lysogeny in nature: Mechanisms, impact and ecology of temperate phages. The ISME Journal, 11(7), 1511–1520. https://doi.org/10.1038/ismej.2017.16
24. Krassowski, M., Das, V., Sahu, S. K., & Misra, B. B. (2020). State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. Frontiers in Genetics, 11, 1598. https://doi.org/10.3389/fgene.2020.610798
25. Krysiak-Baltyn, K., Martin, G. J. O., & Gras, S. L. (2018). Computational Modeling of Bacteriophage Production for Process Optimization. In J. Azeredo & S. Sillankorva (Eds.), Bacteriophage Therapy: From Lab to Clinical Practice (pp. 195–218). Springer. https://doi.org/10.1007/978-1-4939-7395-8_16
26. Kutter, E., & Sulakvelidze, A. (Eds.). (2004). Bacteriophages: Biology and Applications. CRC Press. https://doi.org/10.1201/9780203491751
27. Liu, L. (2014). Fields Virology, 6th Edition. Clinical Infectious Diseases, 59(4), 613–613. https://doi.org/10.1093/cid/ciu346
28. McNair, K., Aziz, R. K., Pusch, G. D., Overbeek, R., Dutilh, B. E., & Edwards, R. (2018). Phage Genome Annotation Using the RAST Pipeline. Methods in Molecular Biology (Clifton, N.J.), 1681, 231–238. https://doi.org/10.1007/978-1-4939-7343-9_17
29. Naureen, Z., Dautaj, A., Anpilogov, K., Camilleri, G., Dhuli, K., Tanzi, B., Maltese, P. E., Cristofoli, F., Antoni, L. D., Beccari, T., Dundar, M., & Bertelli, M. (2020). Bacteriophages presence in nature and their role in the natural selection of bacterial populations. Acta Biomedica Atenei Parmensis, 91(13-S), Article 13-S. https://doi.org/10.23750/abm.v91i13-S.10819
30. Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., & Kyrpides,

N. C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nature Biotechnology, 39(5), Article 5. https://doi.org/10.1038/s41587-020-00774-7
31. Phage Discovery Guide. Howard Hughes Medical Institute. (n.d.). SEA-PHAGES | SEA-PHAGES Phage Discovery Guide is Online! Retrieved November 19, 2022, from https://seaphages.org/blog/2018/09/05/sea-phages-phage-discovery-guide-online/
32. Quer, J., Colomer-Castell, S., Campos, C., Andrés, C., Piñana, M., Cortese, M. F., González-Sánchez, A., Garcia-Cehic, D., Ibáñez, M., Pumarola, T., Rodríguez-Frías, F., Antón, A., & Tabernero, D. (2022). Next-Generation Sequencing for Confronting Virus Pandemics. Viruses, 14(3), Article 3. https://doi.org/10.3390/v14030600
33. Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. Genomics, Proteomics & Bioinformatics, 13(5), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002
34. Roux, S., Hallam, S. J., Woyke, T., & Sullivan, M. B. (2015). Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. ELife, 4, e08490. https://doi.org/10.7554/eLife.08490
35. Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance, 22(13), 30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494
36. Sievers, F., & Higgins, D. G. (2014). Clustal omega. Current Protocols in Bioinformatics, 48, 3.13.1-16. https://doi.org/10.1002/0471250953.bi0313s48
37. Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. Current Protocols in Molecular Biology, 122(1), e59. https://doi.org/10.1002/cpmb.59
38. Stability and Diversity of Ecosystems | Science. (n.d.). Retrieved November 19, 2022, from https://www.science.org/doi/full/10.1126/science.1133258?casa_token=WPoy5GXRwDcAAAAA%3ABue5Zd2wYaV4_yuX4TmAfWZ2X53UQ--V0LW_4lo7t9mh1o9Nv2orj3KxK_JQlZtc0jHKdttukmA
39. Stano, M., Beke, G., & Klucar, L. (2016). ViruSITE-integrated database for viral genomics. Database: The Journal of Biological Databases and Curation, 2016, baw162. https://doi.org/10.1093/database/baw162
40. Stone, E., Campbell, K., Grant, I., & McAuliffe, O. (2019). Understanding and Exploiting Phage-Host Interactions. Viruses, 11(6), E567. https://doi.org/10.3390/v11060567
41. Suttle, C. A. (2007). Marine viruses—Major players in the global ecosystem. Nature Reviews. Microbiology, 5(10). https://doi.org/10.1038/nrmicro1750
42. Sweet, T., Sindi, S., & Sistrom, M. (2021). Going through phages: A Computational approach to Revealing the role of prophage in Staphylococcus aureus. BioRxiv, 2021.11.10.468171. https://doi.org/10.1101/2021.11.10.468171
43. Versoza, C. J., & Pfeifer, S. P. (2022). Computational Prediction of Bacteriophage Host Ranges. Microorganisms, 10(1), Article 1. https://doi.org/10.3390/microorganisms10010149

44. Vix. (2020, August 27). Ecosystem—Definition and Examples—Biology Online Dictionary. Biology Articles, Tutorials & Dictionary Online. https://www.biologyonline.com/dictionary/ecosystem
45. Zeman, M., Mašlaňová, I., Indráková, A., Šiborová, M., Mikulášek, K., Bendíčková, K., Plevka, P., Vrbovská, V., Zdráhal, Z., Doškař, J., & Pantůček, R. (2017). Staphylococcus sciuri bacteriophages double-convert for staphylokinase and phospholipase, mediate interspecies plasmid transduction, and package mecA gene. Scientific Reports, 7(1), 46319. https://doi.org/10.1038/srep46319
46. Zhang, Q., Jun, S.-R., Leuze, M., Ussery, D., & Nookaew, I. (2017). Viral Phylogenomics Using an Alignment-Free Method: A Three-Step Approach to Determine Optimal Length of k-mer. Scientific Reports, 7, 40712. https://doi.org/10.1038/srep40712
47. Zielezinski, A., Barylski, J., & Karlowski, W. M. (2021). Taxonomy-aware, sequence similarity ranking reliably predicts phage–host relationships. BMC Biology, 19(1), 223. https://doi.org/10.1186/s12915-021-01146-6
48. Moura, A., Criscuolo, A., Pouseele, H., Maury, M. M., Leclercq, A., Tarr, C., ... & Brisse, S. (2016). Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes. Nature microbiology, 2(2), 1-10.
49. Cook, R., Brown, N., Redgwell, T., Rihtman, B., Barnes, M., Clokie, M., Stekel, D. J., Hobman, J., Jones, M. A., & Millard, A. (2021). INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. PHAGE, 2(4), 214–223. https://doi.org/10.1089/phage.2021.0007
50. Basso, J. T. R., Ankrah, N. Y. D., Tuttle, M. J., Grossman, A. S., Sandaa, R.-A., & Buchan, A. (2020). Genetically similar temperate phages form coalitions with their shared host that lead to niche-specific fitness effects. The ISME Journal, 14(7), Article 7. https://doi.org/10.1038/s41396-020-0637-z
51. Jung, J., & Park, W. (2015). Acinetobacter species as model microorganisms in environmental microbiology: Current state and perspectives. Applied Microbiology and Biotechnology, 99(6), 2533–2548. https://doi.org/10.1007/s00253-015-6439-y
52. Prabhakaran, R., Chithambaram, S., & Xia, X. (2015). Escherichia coli and Staphylococcus phages: effect of translation initiation efficiency on differential codon adaptation mediated by virulent and temperate lifestyles. The Journal of general virology, 96(Pt 5), 1169–1179. https://doi.org/10.1099/vir.0.000050
53. Khesin, R. B., & Karasyova, E. V. (1984). Mercury-resistant plasmids in bacteria from a mercury and antimony deposit area. Molecular & general genetics : MGG, 197(2), 280–285. https://doi.org/10.1007/BF00330974
54. Bazzi, W., Abou Fayad, A. G., Nasser, A., Haraoui, L. P., Dewachi, O., Abou-Sitta, G., ... & Matar, G. M. (2020). Heavy metal toxicity in armed conflicts potentiates AMR in A. baumannii by selecting for antibiotic and heavy metal co-resistance mechanisms. Frontiers in microbiology, 11, 68.
55. Dempsey, W. B., McIntire, S. A., Willetts, N., Schottel, J., Kinscherf, T. G., Silver, S., & Shannon Jr, W. A. (1978). Properties of lambda transducing bacteriophages carrying R100 plasmid DNA: mercury resistance genes. Journal of bacteriology, 136(3), 1084-1093.

56. Glick, B. R. (2012). Plant growth-promoting bacteria: mechanisms and applications. Scientifica, 2012.
57. Kreiswirth, B., Kornblum, J., Arbeit, R. D., Eisner, W., Maslow, J. N., McGeer, A., ... & Novick, R. P. (1993). Evidence for a clonal origin of methicillin resistance in Staphylococcus aureus. Science, 259(5092), 227-230.