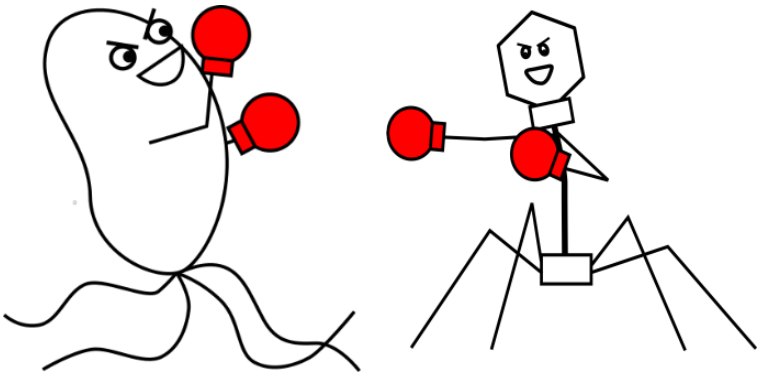
Flinders University

Decoding microbe host interactions



By
Bhavya Nalagampalli Papudeshi
Bachelor of Technology (Biotechnology)
Master of Science (Bioinformatics)

Thesis
Submitted to Flinders University
for the degree of

Doctor of Philosophy
College of Science and Engineering
October 2025

TABLE OF CONTENTS

ABSTRACT

Microbial ecosystems are intricate networks of bacteria, viruses, fungi, and archaea that influence ecosystem health. However, uncovering these interactions remains challenging due to the limited genomic frameworks and the complexity of community interactions within microbial ecosystems. In this thesis, I focus on deciphering phage-bacteria and bacteria-host interactions. I identify the genetic factors using genome-resolved bioinformatic approaches.

While bacterial genomics has made significant strides, phage biology remains relatively underexplored, especially regarding host interactions. As interest in phage therapy to combat antimicrobial-resistant infections grows, the need for standardised frameworks to name, classify, and annotate phage genomes becomes critical. To address this, I begin with a review of phage biology and bioinformatic methods used to sequence and characterise phage genomes. Building on this review, I introduce Sphae, an automated, reproducible bioinformatics toolkit that can seamlessly assemble, annotate, and classify phages. Sphae incorporates advanced tools to rapidly detect genomic features, such as integrases, toxins, and antimicrobial resistance genes — elements that may disqualify phage candidates from therapeutic applications. While Sphae's primary use case lies in the clinical evaluation of phages, it can also broadly characterise phages in other contexts and inform the current gaps in phage biology.

Building on this foundation, I characterised novel phages and explored their interactions with bacterial hosts, illustrating the kinds of questions Sphae is designed to facilitate. I focused on *Crassvirales* phages that infect *Bacteroides*, both key players in the human gut microbiome. In this work, I characterise 14 novel *Crassvirales* isolates, which were assigned to three genera across two families, despite infecting the same host, *Bacteroides cellulosilyticus*. Comparative genomics revealed a conserved tail spike protein across these phages, suggesting a role in host recognition. Using structural modelling and protein-protein interaction predictions, we demonstrate that this protein may interact with TonB-dependent receptors, suggesting convergent host attachment. These findings advance our understanding of phage-mediated modulation of gut microbiomes and highlight the potential of such phages in microbiome-based interventions.

Recognising that microbial ecosystems extend beyond simple phage-bacteria dynamics, I explored more complex interactions by investigating the tripartite relationship between the symbiont bacterium *Xenorhabdus bovienii* with its nematode host, *Steinernema*, and their joint parasitism of insect hosts. Analyses of 42 *X. bovienii* genomes revealed clustering not only by host species but also geography, underscoring the influence of ecological niches on bacterial population structure. Further, signatures of selective sweeps in genes associated with colonisation and interbacterial competition highlight host-specific and spatial drivers of microbial evolution in multipartite systems.

Together, this work elucidates how selective pressures shape microbial interactions across diverse contexts—from phage-bacteria dynamics in the gut to bacterial-eukaryote mutualisms. Through genome-resolved pipelines and structural modelling, this thesis provides both conceptual frameworks and tools to decode microbial interactions. These approaches advance our understanding of microbial community assembly and stability, while also informing future efforts to manipulate microbes or entire communities for therapeutic benefit.

DECLARATION

I certify that this thesis:

1. does not incorporate, without acknowledgement, any material previously submitted for a degree or diploma in any university
2. Research within will not be submitted for any other future degree or diploma without the permission of Flinders University
3. to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where due reference is made in the text; and
4. Generative artificial intelligence was used in sections of my thesis, solely employed to improve sentence clarity, grammar, and flow. All research content, data analysis, interpretation, and academic writing decisions remain my own.
5. The total word count of this thesis is below the required 70,000 words due to the nature of my project and the requirement to write succinctly for publications, as advised by my supervisors. This has been discussed and agreed upon with my supervisory team, and the rationale for the reduced word count has been provided in accordance with the requirements for thesis submission.

Signed: Bhavya Nalagampalli Papudeshi

Date: 15/10/2025

ACKNOWLEDGEMENTS

I acknowledge and honour the First Australians, whose ancestral lands we gather on, and extend my deep respect to their past and present Elders. I pay my respects to the Kaurna people, the traditional custodians of the land on which Flinders University stands and recognise and honour their connection to Country. I also extend this respect to all First Australians, as well as to the Traditional Owners and Custodians of the lands where my collaborators, colleagues, and conference hosts live and work, and to all Indigenous peoples whose lands I have had the privilege of engaging with through my research and academic journey.

I am deeply grateful to my principal supervisor, Professor Robert Edwards, for his unwavering support, guidance, and confidence in my work. His mentorship has been instrumental in shaping my academic journey, constantly challenging me to grow as a scientist while also opening doors to new opportunities that have significantly broadened my skill set and research horizons. I may not always see the potential in myself, but I am genuinely grateful that you do, and for the confidence you have placed in me throughout.

I also wish to acknowledge my co-supervisors, Professor Elizabeth Dinsdale and Professor Jim Mitchell, whose expertise, encouragement, and generosity have greatly enriched my research. To Professor Elizabeth Dinsdale, who was also my master's supervisor, thank you for your ongoing mentorship and for fostering collaborative connections that have been pivotal to my work; many of my co-authorships and interdisciplinary projects have emerged through your lab's support and vision. To Professor Jim Mitchell, thank you for welcoming me into your lab, for the memorable lab lunches, and your constant encouragement and enthusiasm, which consistently lifted my spirits and inspired me to keep pushing forward.

I extend my thanks to Professor Kirstin Ross and Dr. Bart Eijkelkamp, who assessed my milestones and thesis updates. Their feedback and constructive guidance not only shaped my thesis but also encouraged me to explore experimental work, science communication, and broader career avenues beyond a traditional PhD path. Following their suggestions, I complemented my computational focus with experimental experience, which has greatly enriched my understanding of microbiology. They also encouraged me to pursue media releases in addition to my publications, ultimately enhancing my communication skills. A special thanks goes to Tania Bawden from Flinders Media Communications for her invaluable assistance. Though our conversations were brief, they significantly shaped how I pitched my research and helped me recognise my strengths.

I sincerely thank my collaborators and co-authors across various projects, as well as my colleagues at Flinders Accelerator for Microbiome Exploration. Their support and discussions have been invaluable to my research journey. I also recognise their patience. I know there were times

when I committed to more than I could deliver. These experiences became valuable lessons in understanding my limitations and helping with personal growth. Beyond academic support, they regularly checked in on me, offering a space to reflect on challenges, for which I am truly thankful.

To my friends, thank you for being my sounding board, enduring my complaints, and reminding me that there is more to life than a PhD. You kept me active and engaged beyond my academic work, and as an international student, you made me feel at home. Finally, I dedicate this thesis to the family members I lost along the way, those I could not be there for in their final moments. Your absence was deeply felt, but your memories carried me through. Most of all, to my mom, dad, and brother—I do not know what I would have done without you. Your unwavering love, support, and belief in me have been my foundation. Thank you for enduring my frustrations, listening to my endless complaints, and constantly reminding me why I started this journey. When I doubted myself, you brought me back and helped me find my way again. I honestly could not have done this without you. Thank you!

Funding

I want to acknowledge the financial support that made this research possible. My PhD studies were supported by a Flinders University scholarship along with a top-up scholarship from CRC-TIME, which together provided essential funding throughout my candidature.

Additional computational resources were generously provided by the DeepThought High-Performance Cluster (HPC), the Australian Nectar Research Data Commons (ARDC) Nectar Infrastructure, the Pawsey Supercomputing Research Centre, and the National Computational Infrastructure (NCI), funded by the Australian Government. These resources were indispensable for conducting large-scale analyses presented across several chapters.

This work was further supported by funding awarded to the lab from the National Institutes of Health (NIH), the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) (RC2DK116713), the Australian Research Council (DP220102915), and the Gordon and Betty Moore Foundation (9871; Perpetual Viral Origins). Additional funding was provided by Indiana University (IU) through the CAS and OVPR CRCAF awards. These funding sources enabled critical aspects of this research, including data collection, experimental work, microscopy, and publication costs.

I am sincerely thankful to all these organisations for their investment in this research, which has allowed me to pursue and complete this work.

PUBLICATIONS

Publications within this thesis

1. Grigson, S. R., Giles, S. K., Edwards, R. A., & **Papudeshi, B**. (2023). Knowing and naming: phage annotation and nomenclature for phage therapy. Clinical Infectious Diseases, 77(Supplement_5), S352-S359. https://doi.org/10.1093/cid/ciad539
   Included in the literature review, Chapter 2
2. **Papudeshi, B.**, Roach, M. J., Mallawaarachchi, V., Bouras, G., Grigson, S. R., Giles, S. K., ... & Edwards, R. A. (2025). Sphae: an automated toolkit for predicting phage therapy candidates from sequencing data. Bioinformatics Advances, 5(1), vbaf004. https://doi.org/10.1093/bioadv/vbaf004
   Included in Chapter 3
3. **Papudeshi, B.**, Vega, A. A., Souza, C., Giles, S. K., Mallawaarachchi, V., Roach, M. J., ... & Edwards, R. A. (2023). Host interactions of novel *Crassvirales* species belonging to multiple families infecting bacterial host, *Bacteroides cellulosilyticus* WH2. Microbial Genomics, 9(9), 001100. https://doi.org/10.1099/mgen.0.001100
   Included in Chapter 4
4. **Papudeshi, B.**, Rusch, D. B., VanInsberghe, D., Lively, C. M., Edwards, R. A., & Bashey, F. (2023). Host association and spatial proximity shape but do not constrain population structure in the mutualistic symbiont *Xenorhabdus bovienii*. MBio, 14(3), e00434-23. https://doi.org/10.1128/mbio.00434-23
   Included in Chapter 5

Statement of authorship

I, Bhavya Papudeshi, am a significant contributor to each of the chapters in this thesis. Bioinformatics is an inherently interdisciplinary field, and science is a collaborative endeavour. While I took primary responsibility for the research presented in this thesis, this work was made possible through the collective efforts of many researchers who supported me in various ways—whether through data collection, training me in new methodologies, testing and validating developed code, or providing invaluable critique and feedback based on their unique expertise.

By actively seeking collaboration, I ensured that each chapter benefited from the strengths of multiple researchers, leading to higher-quality research and ultimately contributing to peer-reviewed publications. My role was not only to conduct the work but also to bring together and coordinate a team of co-authors whose contributions strengthened the impact of this thesis. The co-authors have been informed that this work will be used in my thesis, and I have included the signed co-authorship approval forms in Appendix A. The contributions of each co-author have been explicitly stated at the beginning of each chapter, and their permission to include these works

has been obtained as per Flinders University's Authorship of Research Output Procedures (Appendix A).

Other publications during candidature

5. Giles, S. K., Stroeher, U. H., **Papudeshi, B.**, Edwards, R. A., Carlson-Jones, J. A., Roach, M., & Brown, M. H. (2022). The StkSR two-component system influences colistin resistance in *Acinetobacter baumannii*. Microorganisms, 10(5), 985. https://doi.org/10.3390/microorganisms10050985
6. Hesse, R. D., Roach, M., Kerr, E. N., **Papudeshi, B.**, Lima, L. F., Goodman, A. Z., ... & Dinsdale, E. A. (2022). Phage diving: an exploration of the carcharhinid shark epidermal virome. Viruses, 14(9), 1969. https://doi.org/10.3390/v14091969
7. Goodman, A. Z., **Papudeshi, B.**, Doane, M. P., Mora, M., Kerr, E., Torres, M., ... & Dinsdale, E. (2022). Epidermal microbiomes of leopard sharks (*Triakis semifasciata*) are consistent across captive and wild environments. Microorganisms, 10(10), 2081. https://doi.org/10.3390/microorganisms10102081
8. Roach, M. J., Pierce-Ward, N. T., Suchecki, R., Mallawaarachchi, V., **Papudeshi, B.**, Handley, S. A., ... & Edwards, R. A. (2022). Ten simple rules and a template for creating workflows-as-applications. PLoS computational biology, 18(12), e1010705. https://doi.org/10.1371/journal.pcbi.1010705
9. Kerr, E. N., **Papudeshi, B.**, Haggerty, M., Wild, N., Goodman, A. Z., Lima, L. F., ... & Dinsdale, E. A. (2023). Stingray epidermal microbiomes are species-specific with local adaptations. Frontiers in Microbiology, 14, 1031711. https://doi.org/10.3389/fmicb.2023.1031711
10. McKerral, J. C., **Papudeshi, B.**, Inglis, L. K., Roach, M. J., Decewicz, P., McNair, K., ... & Edwards, R. A. (2023). The promise and pitfalls of prophages. bioRxiv. 10.1101/2023.04.20.537752
11. Pargin, E., Roach, M. J., Skye, A., **Papudeshi, B.**, Inglis, L. K., Mallawaarachchi, V., ... & Giles, S. K. (2023). The human gut virome: composition, colonization, interactions, and impacts on human health. Frontiers in Microbiology, 14, 963173. https://doi.org/10.3389/fmicb.2023.963173
12. Lima, L. F., Alker, A. T., **Papudeshi, B.**, Morris, M. M., Edwards, R. A., De Putron, S. J., & Dinsdale, E. A. (2023). Coral and seawater metagenomes reveal key microbial functions to coral health and ecosystem functioning shaped at reef scale. Microbial Ecology, 86(1), 392-407. https://doi.org/10.1007/s00248-022-02094-6
13. Doane, M. P., Reed, M. B., McKerral, J., Farias Oliveira Lima, L., Morris, M., Goodman, A. Z., **Papudeshi, B.,** ... & Dinsdale, E. A. (2023). Emergent community architecture despite distinct diversity in the global whale shark (*Rhincodon typus*) epidermal microbiome. Scientific Reports, 13(1), 12747. https://doi.org/10.1038/s41598-023-39184-5

14. Mallawaarachchi, V., Roach, M. J., Decewicz, P., **Papudeshi, B.**, Giles, S. K., Grigson, S. R., ... & Edwards, R. A. (2023). Phables: from fragmented assemblies to high-quality bacteriophage genomes. Bioinformatics, 39(10), btad586. https://doi.org/10.1093/bioinformatics/btad586
15. Bouras, G., Grigson, S. R., **Papudeshi, B.**, Mallawaarachchi, V., & Roach, M. (2024). Dnaapler: a tool to reorient circular microbial genomes. Journal of Open Source Software, 9(93), 5968. https://doi.org/10.21105/joss.05968
16. Roach, M. J., Hart, B. J., Beecroft, S. J., **Papudeshi, B.**, Inglis, L. K., Grigson, S. R., ... & Edwards, R. A. (2024). Koverage: read-coverage analysis for massive (meta) genomics datasets. Journal of Open Source Software, 9(94), 6235. https://doi.org/10.21105/joss.06235
17. Goodman, A. Z., **Papudeshi, B.**, Mora, M., Kerr, E. N., Torres, M., Moffatt, J. N., ... & Dinsdale, E. A. (2024). Elasmobranchs Exhibit Species-Specific Epidermal Microbiomes Guided by Denticle Topography. bioRxiv, 2024-04. https://doi.org/10.1101/2024.04.05.588334
18. Bouras, G., Houtak, G., Wick, R. R., Mallawaarachchi, V., Roach, M. J., **Papudeshi, B.**, ... & Vreugde, S. (2024). Hybracter: enabling scalable, automated, complete and accurate bacterial genome assemblies. Microbial Genomics, 10(5), 001244. https://doi.org/10.1099/mgen.0.001244
19. Pedersen, J. S., Carstens, A. B., Rothgard, M. M., Roy, C., Viry, A., **Papudeshi, B.**, ... & Hansen, L. H. (2024). A novel genus of Pectobacterium bacteriophages display broad host range by targeting several species of Danish soft rot isolates. Virus Research, 347, 199435. https://doi.org/10.1016/j.virusres.2024.199435
20. Mallawaarachchi, V., Wickramarachchi, A., Xue, H., **Papudeshi, B.**, Grigson, S. R., Bouras, G., ... & Edwards, R. A. (2024). Solving genomic puzzles: computational methods for metagenomic binning. Briefings in Bioinformatics, 25(5), bbae372. https://doi.org/10.1093/bib/bbae372
21. Kerr, E. N., Hesse, R. D., Carlson-Jones, J. A., **Nalagampalli Papudeshi, B.**, Butcher, P. A., Doane, M. P., & Dinsdale, E. A. (2025). Draft genomes of five bacteria isolated from Carcharodon carcharias (white shark) and Carcharhinus brachyurus (bronze whaler shark). Microbiology Resource Announcements, e00226-25. https://doi.org/10.1128/mra.00226-25

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
GENERAL INTRODUCTION

Statement of authorship: Bhavya Papudeshi wrote this chapter with editorial input from supervisor Prof. Robert A. Edwards. This work has not been published and is intended solely for inclusion in this thesis.

Statement on the Use of Generative Artificial Intelligence (AI): Generative AI tools (specifically ChatGPT by OpenAI and Grammarly) were used during the preparation of this chapter for language editing purposes, such as improving sentence clarity, grammar, and structure. These tools were not used to generate original content, perform data analysis, or contribute to reading the papers referenced. All intellectual interpretations and analytical perspectives presented in this thesis are my own, following Flinders University's policy on the responsible use of generative AI in research.

1.1 Introduction

Microbes are the unseen engineers of life, orchestrating essential processes that sustain ecosystems, drive biogeochemical cycles and influence the health of living organisms and ecosystems. Microbial communities, comprising archaea, bacteria, fungi, and viruses, are collectively termed as the microbiome (Lederberg & Mccray, 2001). Although the term microbiome was initially coined to describe the diverse, heterogeneous microbes inhabiting the human body, it is widely recognised that microbiomes are not exclusive to humans(Integrative HMP (iHMP) Research Network Consortium, 2019). Complex microbial systems have also been described in association with other animals, such as corals(Lima et al., 2020), kelp(Minich et al., 2018; Morris et al., 2016), and sharks(Doane et al., 2020, 2023; Goodman et al., 2022; Hesse et al., 2022; Kerr et al., 2023), as well as entire ecosystems(Dinsdale et al., 2008; Gilbert et al., 2018; Sunagawa et al., 2020). Reflecting on this broader perspective, Berg et al. (2020) redefined the microbiome in more comprehensive and ecologically meaningful terms, as microbial community together with their "theatre of activity". This term now encompasses structural elements, metabolites, mobile genetic elements, and interactions within their habitat. This modern definition acknowledges microbiomes as integrated ecological entities, shaping and shaped by their environments across scales from individual hosts to global ecosystems.

These microbial systems involve dynamic interactions among their members that are pivotal in driving their evolution, structuring populations, and maintaining the stability of microbial ecosystems. In many studies to date, the bacterial component of microbial communities is relatively well characterised, whereas viruses, particularly bacteriophages, remain comparatively unexplored. Most research tends to focus either on bacterial–bacterial interactions or solely on the virome, with limited attention given to how these microbial components interact and influence each other within shared environments.

1.2 Knowledge gap and rationale

These intricate interactions between bacterial populations, their phages, and eukaryotic hosts play fundamental roles in determining the host range, resistance mechanisms, and the broader co-evolutionary dynamics that underpin microbial ecosystems. They exemplify core evolutionary concepts, such as the Red Queen hypothesis(Van Valen, 1973), where continual adaptation is essential to maintain relative fitness amid biotic pressure(Brockhurst et al., 2014). In bacteria, processes such as horizontal gene transfer and selective sweeps drive diversification and ecological partitioning, enabling communities to rapidly shift their functional potential and restructure themselves in response to changing environments(Arevalo et al., 2019; VanInsberghe et al., 2020). In phages, their host interactions are studied by transferring their host range information to networks to illuminate the underlying coevolutionary processes to inform host breadth and specificity(Kauffman et al., 2022; Weitz et al., 2013). Understanding these dynamics is

essential not only for elucidating patterns of microbial diversity and stability but also for informing strategies to manipulate microbiomes for the benefit of their hosts and the environment.

Yet despite these advances, predicting how cross-taxa interactions drive mechanistic insights, structure populations, and impact ecosystems remains limited. This shortfall is compounded by a lack of reproducible and scalable genomic frameworks, which restricts our capacity to compare findings across studies robustly or to trace the evolutionary forces shaping microbial populations. Disentangling these complexities within naturally heterogeneous microbiomes poses significant challenges. This thesis addresses these knowledge gaps by isolating and studying host–microbe and microbe–microbe interactions, focusing on individual isolate genomes to enable fine-scale analysis of genomic and evolutionary mechanisms. While single-isolate approaches necessarily simplify the intricate context of whole communities, they provide critical mechanistic insights that are often obscured in metagenomic studies. By coupling this with the development of automated, reproducible bioinformatics workflows, this work not only advances our understanding of these foundational interactions but also establishes methodological standards that enable rigorous, cross-system comparisons in microbial ecology and evolution.

1.3 Research aims

To answer these questions, this thesis is structured around three research aims which have been addressed as chapters in this thesis:

Aim 1: Develop a reproducible and scalable genomic framework

Phages were selected as the focus for this aim because they are abundant and ecologically significant in shaping microbial community dynamics and show promise as alternatives to combat antibiotic-resistant infections. However, they are understudied, and there are currently no workflows available to assemble and characterise these genomes quickly. To address this gap, I first reviewed the current literature on existing methods for study phage genomes analysis (Chapter 2). Building on this foundation, I then developed an integrated, user-friendly pipeline for characterising phages in chapter 3 of this thesis.

Aim 2: Investigate the genomic markers underlying bacteriophage interactions with their bacterial hosts

Understanding how phages interact with their bacterial hosts is crucial for deciphering the selective forces that shape microbial genomes, influence host specificity, and drive ecosystem-level functional dynamics. Yet, despite the ecological prominence of phage–bacteria interactions, the genomic determinants of these relationships remain poorly understood, especially in systems like the human gut, where phage diversity is high but experimental isolates are scarce. To address this, I focus on characterising phage isolates using genomic data to identify key host-interaction genes, assess the conservation and selection pressures acting on these proteins, and model their interactions with bacterial surface receptors. This work is presented in Chapter 4 of this thesis.

Through this approach, this aim provides mechanistic insights into how phages influence bacterial adaptations and uncovers how phage diversity contributes to the structure and function of gut microbial communities.

Aim 3: Investigate the population structuring of bacterial symbionts in multipartite interactions

Although bacterial symbionts play a vital role in host health, we still lack a clear understanding of how their populations are structured across host species and environments, particularly in complex, multipartite interactions involving hosts and diverse microbial communities. In this aim, I use bacterial symbionts as a model system to explore how gene flow, recombination, and selection shape population structure and adaptation within these multipartite networks. This work is presented in chapter 5 of this thesis. By applying comparative genomic analyses to symbionts sampled across host taxa and geographic gradients, tracing patterns that underpin population cohesion and diversification. This work provides a framework for understanding how microbial communities interact and evolve within host-associated ecosystems.

Together, these aims provide a multi-scale perspective on the processes that structure microbial communities, advancing our understanding of community ecology and strategies to harness microbial interactions for therapeutic and ecological applications.

1.4 Thesis structure

This thesis is organised to gradually build from a broad conceptual and methodological foundation, leading to detailed studies of specific microbial systems. It begins with a literature review on bacteriophages in Chapter 2. This thesis begins with a focus on phages due to their crucial roles as modulators of microbial communities and their potential as tools against antibiotic-resistant infections. The review covers their biology, and the bioinformatics methods used to sequence and characterise phage genomes. It also discusses the need for improved annotation tools that accurately assign gene functions, along with standardised nomenclature for effective classification

and communication. The chapter concludes with an evaluation of phages for therapeutic applications.

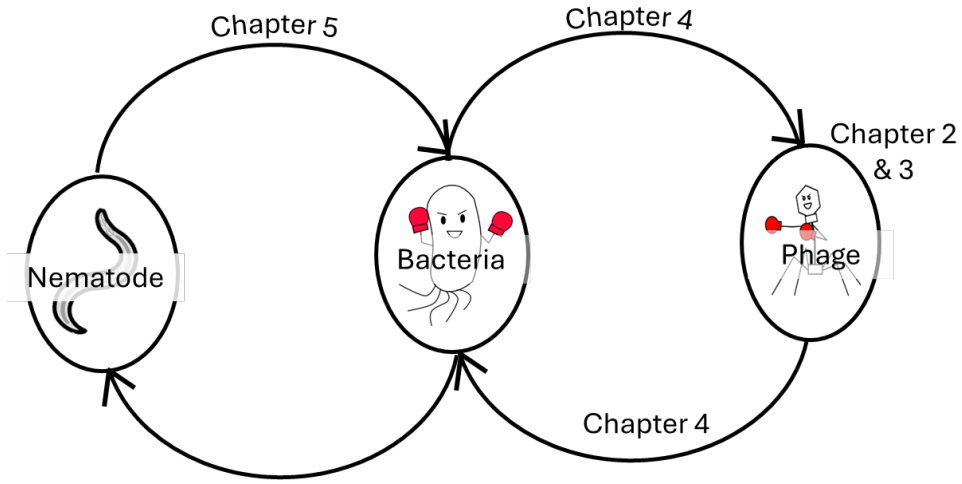


Figure 1.1: Conceptual framework of this thesis. Highlights the phage–bacteria interactions (Aim 2, Chapter 4) and the tripartite interactions among nematodes, bacterial symbionts, and their broader environment (Aim 3, Chapter 5).

Next, in Chapter 3, I address Aim 1. I start with a brief overview on why scalable bioinformatic workflows matter, they lower the entry barrier, analyses reproducible, and put powerful tools in the hands of the whole phage-research community. As no dedicated pipelines existed for phage characterisation, beginning with sequencing data to naming and functional annotations. I developed Sphae, an automated bioinformatics toolkit designed to analyse phage genomes quickly. Recognising that the major application of tis workflow would be towards phage therapy, I incorporated tools to detect genomic markers such as antimicrobial resistance and virulence genes, directly addressing the demand for scalable and reproducible computational solutions. This workflow has already been peer-reviewed and published in *Bioinformatics Advances,* and is available through GitHub, PyPI, conda and Docker package managers. This chapter walks through Sphae's architecture and demonstrates how its end-to-end design enables the discovery of therapeutic phage candidates to be truly scalable and reproducible. The toolkit is also modular, allowing users to flexibly run only the state-of-the-art tools they need, making it adaptable to a range of workflows and expertise levels.

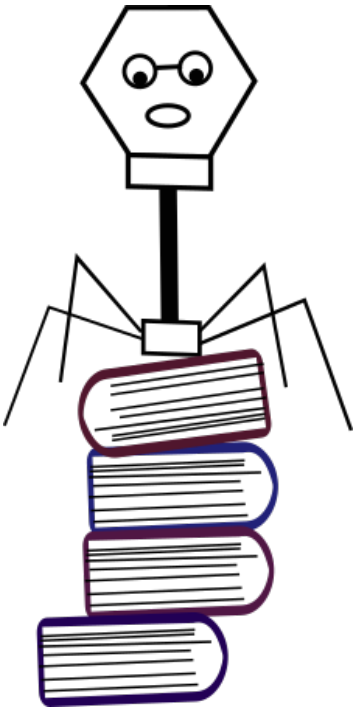
In Chapter 4, I address Aim 2, through applying the phage characterisation workflow developed in Chapter 3 to investigate the dominant gut phages belonging to the order *Crassvirales*, which infect *Bacteroides*, key microbes involved in digestion, immune modulation, and disease susceptibility. Despite their abundance, they remain largely uncharacterised, with few available isolates for experiments to study mechanisms of host interactions. In this chapter, I characterise novel *Crassvirales* isolates, exploring their genomic features and the evolutionary relationships of genes involved in host specificity and interaction through comparative genomics. I also employed

structural modelling to examine key proteins involved in host specificity, providing insights into the mechanisms by which phages interact with and adapt to their bacterial hosts. The work in this chapter has been published in *Microbial Genomics* and demonstrates a mechanism of phage-host interaction in the gut microbiome.

In Chapter 5, I address Aim 3 by investigating the factors that shape population structure in bacterial symbionts engaged in multipartite interactions with phages and other partners. In this chapter, I shift my focus from phages to bacterial symbionts that are not only interacting with each other but also their hosts. I focus on the system, *Xenorhabdus bovienii*, a mutualistic bacterium associated with nematodes, which together are parasitic to insect hosts. I employ comparative genomic analyses across host species and geographic gradients to investigate how *X. bovienii* populations navigate these evolutionary forces. This work was published in *MBio.* These findings highlight how microbial populations structure in multipartite environments, supporting a broader understanding of how host–microbe associations and ecological interactions drive the assembly, resilience, and functional capabilities of microbiomes.

Finally, in Chapter 6, I bring together the findings from each chapter into a unified discussion. I connect the development of scalable workflows, insights into phage–bacteria interactions, and patterns of bacterial symbiont population structure back to the original aims, highlighting how each contributes to addressing key knowledge gaps in the field. The discussion connects fine-scale genomic processes to broader ecological and evolutionary dynamics, offering new insights into how microbe-host interactions collectively influence the assembly, resilience, and functional potential of microbiomes. It also outlines future research directions to build on this work, addressing remaining gaps in our understanding of the forces that govern microbial community diversity and stability.

CHAPTER 2
LITERATURE REVIEW



This chapter is based on the published literature review Grigson, S. R., Giles, S. K., Edwards, R. A., & **Papudeshi, B**. (2023). Knowing and naming: phage annotation and nomenclature for phage therapy. Clinical Infectious Diseases, 77(Supplement_5), S352-S359. https://doi.org/10.1093/cid/ciad539. This article is reproduced in full under the terms of the Creative Commons Attribution License (CC BY 4.0). © The Author(s) 2023. Published by Oxford University Press on behalf of the Infectious Diseases Society of America.

This publication has been expanded and restructured with additional sections and updates to better align with the broader scope of this thesis.

Statement on the Use of Generative Artificial Intelligence (AI): A Generative AI tool, including ChatGPT and Grammarly, was used during the preparation of this chapter for language editing purposes, such as improving sentence clarity, grammar, and structure.

Preface

This chapter is based on the review article titled “Knowing and Naming: Phage Annotation and Nomenclature for Phage Therapy,” which I co-authored as part of this thesis. This paper was written to address fundamental challenges in characterising phages for phage therapy, for the audience of Clinical Infectious Diseases Journal. The article critically examined current practices and bioinformatic tools in phage genome annotation and naming, emphasising the urgent need for rigorous, standardised, genome-informed approaches that align with modern sequencing and comparative genomics capabilities. As the last author on this publication, I played a key role in shaping its scope, identifying main themes, and synthesising literature on phage taxonomy, annotation tools, and therapeutic considerations.

Since its publication, this article has been cited in 18 other publications, demonstrating its relevance to the rapidly evolving field of phage therapeutics and annotation standards. In this thesis, I have built on the content from the review article, restructuring and expanding the work to align with the narrative of this thesis. I have added sections focusing on phage diversity and biology relevant to understanding phage-bacteria interactions in Chapter 4. Moreover, where relevant, the bioinformatics tools have also been updated to include those published since 2023, to ensure this literature review is up to date.

This chapter provides a comprehensive foundation for the next two chapters, establishing the biological, ecological, and computational contexts necessary to understand the challenges and opportunities in phage research, thereby directly motivating the methodological developments and comparative analyses presented in the following chapters.

Statement of authorship

As the last author on this work, I played a senior leadership role in shaping the paper’s direction and scope. I was instrumental in identifying the key themes that needed to be addressed, curating and synthesising relevant literature, and framing the central arguments regarding phage taxonomy, gene annotation, and therapeutic suitability. My contribution included substantial input into the conceptual framing, critical analysis of existing tools and standards, and guidance on the tone and structure of the manuscript to ensure it was both accessible and rigorous.

Below is a breakdown of the author's contributions:

Author	**Contribution**
Susanna Grigson	Writing the annotation section and editing the manuscript

Sarah K. Giles	Writing the first section, including experimental techniques to isolate phages, and editing of the manuscript
Robert A. Edwards	Structuring and editing of the manuscript
Bhavya Papudeshi	Writing the naming a phage section, structuring, and editing of the manuscript

The contributions of each co-author have been explicitly stated, and their permission to include these works has been obtained as per Flinders University's Authorship of Research Output Procedures (Appendix A)

All the world's a phage

Abstract

Bacteriophages, or phages, are the most abundant biological entities on the planet and play central roles in microbial ecology through their diverse infection strategies and genomic architectures. They influence nutrient cycling, microbial population dynamics, and horizontal gene transfer, positioning them as key drivers of microbial evolution. This chapter provides an overview of phage diversity and infection dynamics, focusing on the modular and mosaic nature of phage genomes that complicates functional prediction. It reviews current approaches for genome annotation, the identification of key genomic features, and the assignment of functional labels to protein-coding sequences—essential steps to avoid the inadvertent inclusion of undesirable genes in therapeutic applications. The chapter also highlights the role of the International Committee on Taxonomy of Viruses (ICTV) in standardising phage classification and nomenclature. With growing interest in phages as therapeutic agents to combat antibiotic resistance, understanding phage biology is increasingly important. Lytic phages are preferred in therapy due to their bacteria-killing abilities, while temperate phages are often avoided because of their potential to transfer resistance or toxin genes. Selecting suitable phage candidates relies heavily on plaque morphology and genome sequencing, underscoring the need for robust annotation tools and a deep understanding of phage life cycles. Altogether, accurate annotation and consistent classification enhance our understanding of phage–host interactions, replication strategies, and evolution, advancing both basic research and the safe development of phage-based therapies.

2.1 Introduction

Bacteriophages, also known as phages, are viruses that infect bacteria. They influence microbial evolution, modulate microbial communities, and drive biogeochemical cycles(Chevallereau et al., 2022; Suttle, 2007), and maintain ecosystem stability. Their interactions with bacteria underpin critical ecological processes and have far-reaching implications for understanding microbial community dynamics across diverse environments. In their seminal work, Hendrix et al., (1999) described the extensive diversity and evolutionary relationships among bacteriophages and prophages, aptly titling their paper "All the World's a Phage." This metaphor highlights the pervasive presence and influence of phages on microbial life, serving as the title of this chapter.

Viral diversity is staggering, with a single gram of soil estimated to contain >1 billion virions(Williamson et al., 2017) and an estimated 10^{31} total virions on Earth(Hendrix et al., 1999). Despite this staggering abundance, only a fraction of viral diversity is formally recognized: as of the 2024 International Committee on Taxonomy of Viruses (ICTV) release, ~16,215 viral species are classified across all viruses. For pahges specifically, genomic resources are expanding rapidly, with 26,048 complete phage genomes and ~28,468 isolated phages reported as of September 2023 (Cook et al., 2021), still only a sliver of the vast, uncharted phage diversity. They exhibit

variations in morphology, replication strategies, genetic composition, and host specificity. Despite their ubiquity and ecological significance, the broader study and application of phages continue to face substantial challenges. Inconsistencies in phage taxonomy and classification, along with shortcomings in genome annotation(Shen & Millard, 2021; Turner et al., 2021), where the majority of predicted phage genes remain labelled as hypothetical. This limits our ability to compare phages across studies, understand their roles in microbial communities, or fully harness their potential as tools in biotechnology and medicine. These knowledge gaps hinder not only fundamental insights into phage–bacteria interactions but also practical efforts to leverage phages in areas ranging from microbiome engineering to therapeutic interventions.

This chapter provides a comprehensive overview of the biological and genomic characteristics of phages, forming a foundational framework for the thesis. It examines phage diversity and explores the modular, mosaic architecture of phage genomes. The chapter also reviews current approaches to genome annotation and standardised classification, drawing on insights from our published review and incorporating additional perspectives needed to contextualise the original research presented in the subsequent chapters.

2.2 Phage biology

Their staggering genetic variability is one of the challenges in studying these entities. Unlike bacteria and eukaryotes, phages lack universal marker genes, such as 18S rRNA genes, and often have genomes dominated by hypothetical proteins with no known homologues in databases. Their biology is unique, as they are obligate parasites that rely entirely on host cellular machinery for replication. Their evolutionary trajectories are shaped by rapid genetic turnover, modular genome architectures, and intense evolutionary pressures. These characteristics highlight why phages cannot be fully understood through frameworks developed for other cellular life, underscoring the need for specialised tools and models to understand their roles in microbial ecosystems.

2.2.1 Morphological diversity

Phage morphology(Figure 2.1A) was historically used for classification and mainly classified as tailed and non-tailed phages. Most known phages belong to the Class: *Caudoviricetes*, characterised by their tail morphology:

- Myovirus-like morphology, which possess long contractile tails,
- Siphovirus-like morphology, which possess long non-contractile tails, and
- Podovirus-like morphology, which possess short, non-contractile tails.

Phage tail structures can offer insights into their interactions with their bacterial hosts. These tails, comprising long or short tail fibres, spikes, baseplates, and contractile sheaths, serve as the sensory and mechanical interface between phages and bacteria. Tail fibres and spikes mediate highly specific, reversible binding to cell-surface receptors such as lipopolysaccharides, membrane

proteins, pili, or flagella, establishing host range and initiating infection(Nobrega et al., 2018). These tail components evolve rapidly in response to host resistance mechanisms, driving co-evolution and influencing phage specificity and fitness.

In contrast, non-tailed phage morphology is classified into:

- Inoviridae, filamentous or rod-shaped phages,
- Plasmavirinae, which includes membrane-enveloped phages exhibiting variable, pleomorphic shapes(Krupovic & ICTV Report Consortium, 2018),
- Icosahedral phages comprise icosahedral capsids that lack tail structures(Dion et al., 2020). These phages can be further grouped into
    - Microviridae, small ssDNA phages
    - Leviviricetes, ssRNA phages
    - Cystoviridae, segmented dsRNA genomes with lipid envelope around the capsid,

Tectiviridae, Corticoviridae, Sphaerolipoviridae, linear dsDNA with lipid membrane containing phages. Non-tailed phages rely on alternative structures such as coat or envelope proteins—like *pIII* in filamentous *Inoviridae* (Knezevic et al., 2021), capsid spikes in *Microviridae* (Cherwa & Fane, 2011), or envelope proteins in *Plasmaviridae* to recognise and bind to their hosts (Krupovic & ICTV Report Consortium, 2018)*.*

2.2.2 Infection strategies

Phages are also classified based on infection strategies (Figure 2.1B). They primarily fall into one of two categories. The lytic cycle involves hijacking the bacterial machinery to produce progeny and ultimately lysing the host cell. In contrast, the lysogenic lifecycle involves phages that integrate into the bacterial genome and replicate with the bacteria as prophages. In this cycle, rather than producing new viral particles immediately, they remain as prophages, within the host lineage. These prophages are a driving force of bacterial ecology and evolution within the bacterial populations, offering advantages such as superinfection exclusion (Barr et al., 2013; Bondy-Denomy et al., 2016) and contribute to metabolic functions (Bondy-Denomy & Davidson, 2014).

Other variants of infection strategies include pseudolysogeny, which describes a stalled phage infection that neither proceeds to full genome replication (as in the lytic cycle) nor integrates into the host genome (as in the lysogenic cycle). During this stage, the phage genome remains inside the host cell largely dormant and unreplicated, awaiting more favourable conditions to either initiate replication or integration(Correa et al., 2021; Łoś & Węgrzyn, 2012). In the case of P22, this pseudolysogenic state often occurs after DNA injection, when the phage has entered the cell. However, the infection process is paused due to stressors like nutrient limitation. In the other strategy, the chronic infection cycle, the phage enters the host, replicates and new virions are released continuously without host cell lysis. This lifecycle is generally observed in filamentous

phage (Inoviridae). For example, CTXΦ filamentous phage carries the cholera toxin in *Vibrio cholerae*(Waldor & Mekalanos, 1996), and Pf4 phages increase biofilm production in *Pseudomonas aeruginosa*(Gavric & Knezevic, 2022). At the population level, this can contribute to a carrier state, where the bacteria and phages coexist stably over time. Here, a subpopulation of cells undergoes productive, lytic infections (releasing phage progeny), while other cells either resist infections or maintain the phage genome non-productively. This balance enables phages to persist long-term within a bacterial culture, contributing to ongoing gene transfer and infection dynamics. Interestingly, Long-term persistence is also observed for abundant gut phages belonging to the Order: *Crassvirales* (Cortés-Martín et al., 2025; Shkoporov, Khokhlova, et al., 2021). However, these phages are obligately lytic, their persistence is thought to arise form host–phage ecological dynamics (e.g., resistant subpopulations, spatial structure, and turnover), rather than a true carrier state. Finally, some temperate phages adopt to plasmid lifecycle, where the phage genome is stably maintained as an extrachromosomal plasmid (e.g., P1, N15) (Pfeifer et al., 2022, 2021; Ravin, 2011). This represents a plasmid form of lysogeny distinct from both chromosomal integration and transient pseudolysogeny, further demonstrating that phage lifecycles do not fit neatly into rigid categories. Instead, they span a spectrum of strategies that blur traditional definitions, highlighting ongoing novelty in how phages interact with their hosts.

2.2.3 Lysogeny to lytic conversion switch based on microbial density

Environmental factors, particularly density-dependent microbial concentrations, have been shown to influence phage life cycle decision(Silveira et al., 2021). For example, in nutrient-depleted environments such as deep oceans, phages often favour lysogeny, consistent with the Piggyback-the-Loser (Refugium) model, where integration provides stability when the hosts are metabolically limited. They favour a lysogenic state when the bacterial host is in more favourable conditions(Felipe H. Coutinho et al., 2017).On the other hand, in high-density environments, such as the human gut, lysogeny prevails (M.-S. Kim & Bae, 2018). The Piggyback-the-Winner model suggests that lysogeny also becomes more prevalent at high bacterial densities, whereas the Kill-the-Winner model predicts that lytic phages dominate at intermediate densities (Knowles et al., 2016). The models illustrate how bacterial and phage densities influence the balance between lytic and lysogenic cycles, allowing them to access different ecosystems(Silveira et al., 2021).

2.2.4 Modularity and plasticity of phage genomes

These infection strategies, involving shifts between lytic and lysogenic modes in response to environmental cues and host densities, are governed by finely tuned regulatory networks encoded within phage genomes(Benler & Koonin, 2020; Feiner et al., 2015). More broadly, the architecture of phage genomes, shaped by horizontal gene transfer and recombination, equips phages with the evolutionary flexibility needed to adapt to diverse ecological niches(Botstein, 1980; Hatfull & Hendrix, 2011; Pedulla et al., 2003). Phages have a conserved architecture, organised into functional modules: host takeover genes (early), phage replication genes (middle), and structural

virion genes (late). Genomic mosaicism refers to the composition of genomes from interchangeable, semi-independent functional modules, often acquired from unrelated phages or bacterial hosts(Botstein, 1980; Westmoreland et al., 1969). This modularity facilitates the adaptation to new environments and host species, especially in key genes such as receptor-binding proteins on tail spikes or tail fibres, and endolysins (Figure 2.1C). To demonstrate genomic mosaicism, phage specificity was swapped using host-recognition modules, which not only reflects evolutionary history but can be harnessed as a tool for engineering phage function(Dunne et al., 2019; Latka et al., 2021; Smug et al., 2023).

Recombination does not occur solely within closely related phages; it can also cross significant taxonomic distances, especially in phages enriched with recombinases and transposases. In a study by Moura de Sousa et al.,(2021) showed that while phage family and host phylum can act as barriers, gene exchange still occurs, including between temperate and virulent phages. Notably, some virulent phages with large genomes contribute genes related to cell energetics, nucleotide metabolism, DNA packaging, and injection, which enhance infection efficiency and may influence host physiology.

While the genomic mosaicism paradigm has been influential in shaping our understanding of phage evolution, it likely overemphasises the role of horizontal gene transfer. More recent phylogenomic analyses indicate that vertical inheritance is the dominant evolutionary force in many phage taxa, with horizontal gene exchange contributing mainly at specific loci involved in host interactions and niche adaptation(Rohwer & Edwards, 2002). Thus, modular organisation and recombination remain important for local innovation, but the broader evolutionary trajectories of phages are best explained by stable vertical lineages rather than a purely mosaic framework.This perspective is reflected in current classification frameworks by International Consortium of Taxonomy of Viruses (ICTV), which rely on *terL* (the large terminase subunit) as a conserved phylogenetic marker across dsDNA phages, underscoring its reliable vertical signal (Lefkowitz et al., 2018).While certain regions like tail fibre genes remain hotspots for recombination and mosaic exchange, core genes like terminase (and structural virion genes) often maintain coherent vertical lineage relationships. Thus, the use of *terL* in taxonomy underscores the dominance of vertical evolutionary signal in delineating phage relationships.

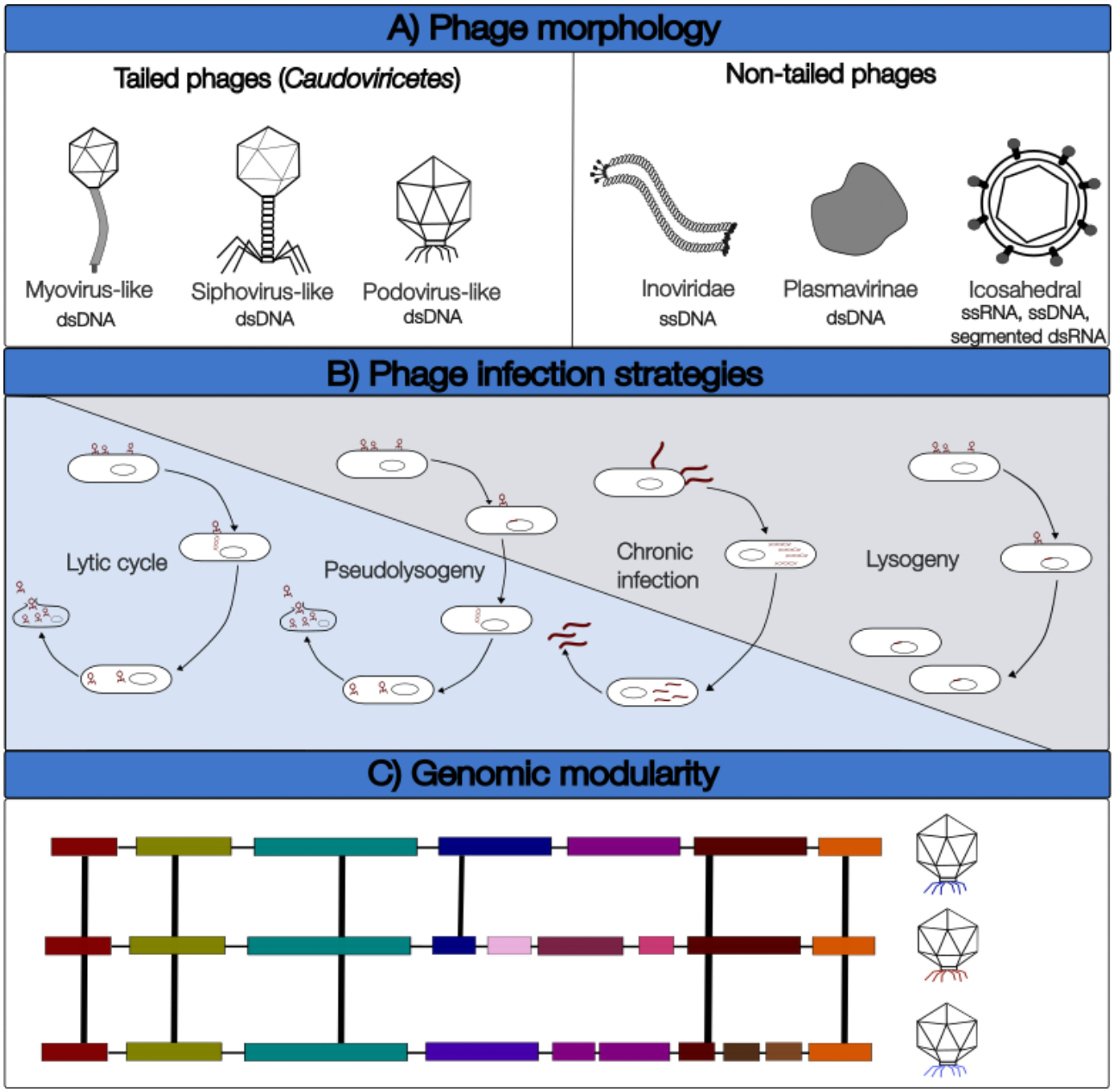


Figure 2. 1: Phage diversity A) Phage host morphology showing both tailed and non-tailed phages, and B) viral infection strategies. This image is adapted from(Valencia-Toxqui & Ramsey, 2024). Blue represents the lytic lifecycle, and grey represents the lysogenic lifecycle of the phages. This image is adapted from Correa et al., (2021). C) Genomic modularity showing how genes within the phage genomes are interchangeable and can serve in host specificity

2.2.5 Phage packaging influences horizontal gene transfer

Once the phage injects its genome into the host cell, it hijacks the bacterial machinery to replicate its DNA and synthesise new virion components. The last step of synthesising new components includes genome packaging, which plays a key role in shaping how phages interact with bacterial hosts and influence microbial communities. The mechanism by which a phage packages its genome into new viral particles can influence its potential for horizontal gene transfer (HGT), a crucial evolutionary force in microbial ecosystems.

Two of the best studied dsDNA strategies are *cos* and *pac* systems. In the *cos* site packaging model, as exemplified by phage λ, a concatemeric DNA molecule is cleaved precisely between two specific *cos* recognition sites by the terminase complex, resulting in genome-length precision with cohesive ends(Catalano & Morais, 2021; Dennehy & Abedon, 2020). Cohesive (cos) packaging ensures exact genome size, facilitating efficient circularisation and replication upon infection. The *pac* or headful packaging mechanism, common in phages like T4, SPP1, and P22, involves cleavage of a concatemeric substrate at a single pac site followed by translocation of DNA into the procapsid until it reaches capacity ("headful"). Termination is then signalled by the filled capsid, resulting in genomes that frequently exceed unit length and generate terminal redundancies(Wolput et al., 2024). These extra terminal repeats can promote circularisation, recombination, and exchange of host or phage genes, contributing to genetic diversity and lateral transduction events. Understanding these mechanistic differences is essential, cos-type phages tend to avoid generalised transduction because they require two specific cleavage sites, whereas pac-type phages can mobilise larger segments of host DNA due to their headful packaging strategy(Wolput et al., 2024).

Beyond these two well-studied dsDNA packaging mechanisms that describe how the terminase recognises, cleaves and initiates packaging, in other phages other mechanisms of recognising genome termini are defined. For instance, in cos-type phages, genome termini are precisely defined by cleavage at specific cos recognition sites, whereas in pac-type phages, termini are determined by headful packaging capacity, resulting in variable ends with terminal redundancy. Other phages define their ends differently, through replication generated repeats, terminal proteins, and nucleic acid structures which in turn shape their packaging outcomes(Casjens & Gilcrease, 2009). T7 like phages generate short terminal repeats during replication, which are recognised and cut at these fixed distances, yielding precisely terminally repeated genomes (Chung et al., 1990). T5-like phages instead pre-form long terminal redundancies during replication which are packaged without the need for concatemeric cleavage(Rhoades & Rhoades, 1972). *Bacillus* phage Φ29-like phages use a protein-primed system, in which the genome termini are defined by covalently attached terminal proteins that act as recognition signals for packaging initiation(Ito, 1978; Salas et al., 1978). In addition, ssDNA phages such as Microviridae package complete unit-length circular genomes generated via rolling-circle replication(Martin et al., 2011), while ssRNA phages (e.g., MS2, family Leviviricetes) rely on specific RNA secondary structures that interact with coat proteins to ensure selective encapsidation of full-length RNA genomes(Tars, 2020).

These diverse systems illustrate that genome termini can be defined in multiple ways. Distinction of these mechanisms is crucial not only for interpreting phage evolution but also for guiding safe therapeutic phage selection and genome assembly in comparative genomics studies.

2.3 Knowing phage genetic potential

Phage biology can be inferred directly from their genome content: the presence, absence, or variants of specific phage genes reflects infection strategies, replication mechanisms, structural components, and interactions with hosts. Here, we outline the key steps and tools used in phage isolation and detection, followed by genome assembly and the structural and functional annotation of phage genomes (Figure 2.2). Each of these processes is described in detail below, providing a comprehensive overview of how phages are recovered, sequenced, and characterised from both experimental and computational perspectives.

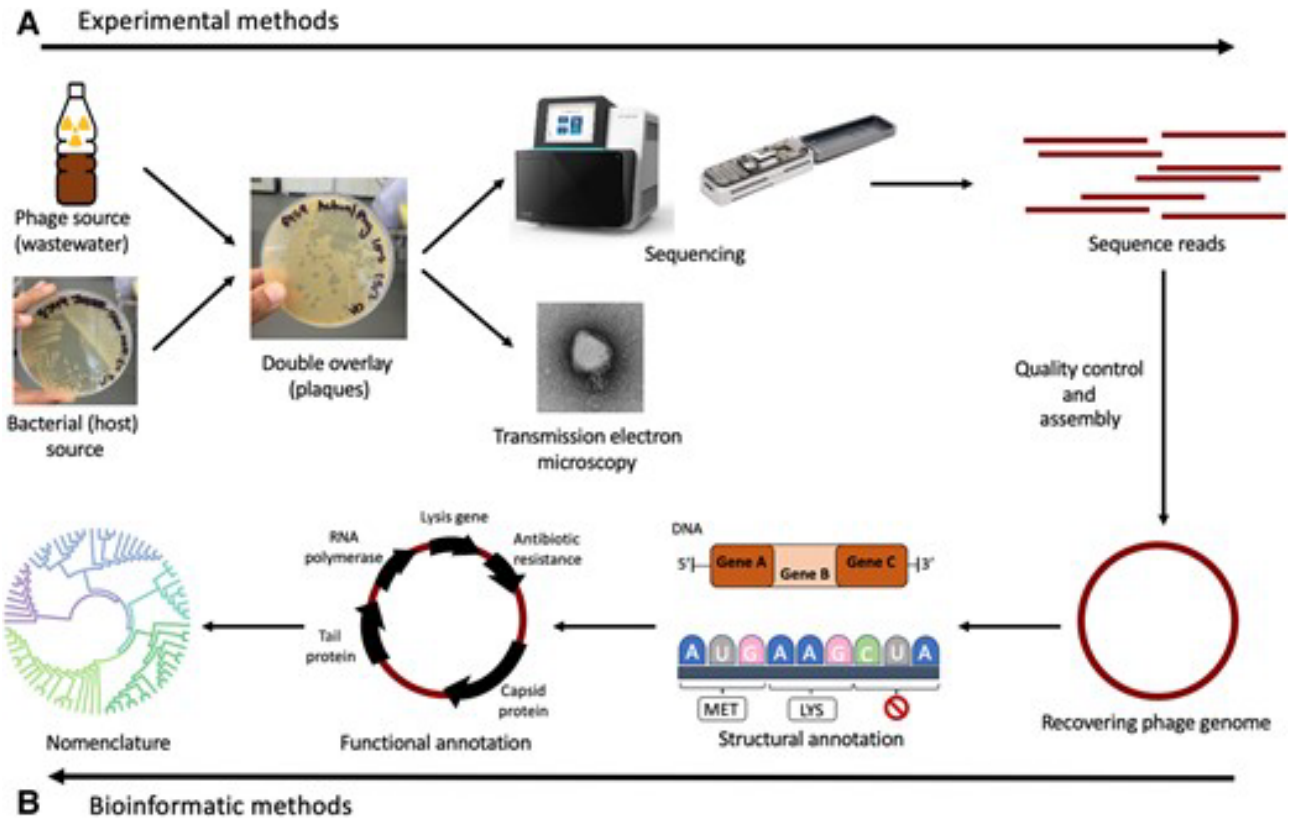


Figure 2. 2: Overview of the steps in phage isolation and characterisation. A) Experimental methods: the double overlay method facilitates plaque formation, helping isolate and select lytic phages from an environmental source. Transmission electron microscopy helps visualise the isolated phages to determine the broad taxonomic grouping. Concurrently, the process involves extracting the isolated phage's DNA and sequencing it. B) Bioinformatics methods: assembly of sequence reads allows for the recovery of complete genomes, accurate annotation, and phylogenetic classification.

Abbreviations: LYS, lysine; MET, methionine

2.3.1 Finding phages

Typically, phages are isolated from environmental samples rich in bacteria, such as soil, water, or sewage, to isolate phages. The process begins with enrichment and filtration steps to concentrate phages within a sample(Luong et al., 2020). Enrichment is achieved by co-culturing bacterial strains with environmental samples, which amplifies the population of environmental phages and results in higher plaque counts. The widely used double overlay methods(Stachurska et al., 2021) enables the visualisation, isolation, and purification of phages through plaque formation, permitting the identification of diverse morphologies. Single plaques are picked and cultured to separate distinct phage species from environmental samples. A second round of plaquing allows us to

distinguish between large (>3 mm diameter) and small (1 mm in diameter) plaque sizes (Figure 2.3 A). Other morphological features include halos around the plaque, indicating depolymerase or endolysin activity (Figure 2.3A). Interestingly, infection by a lytic phage can induce resident prophages within the same bacterial cell. These mixed infections can result in plaques where both lytic and lysogenic phage phenotypes co-occur producing complex plaque morphologies on a single, double overlay plate. Transparent plaques indicate lytic phages, while cloudy plaques indicate the lysogenic phages (Shymialevich et al., 2023; Zheng et al., 2020).

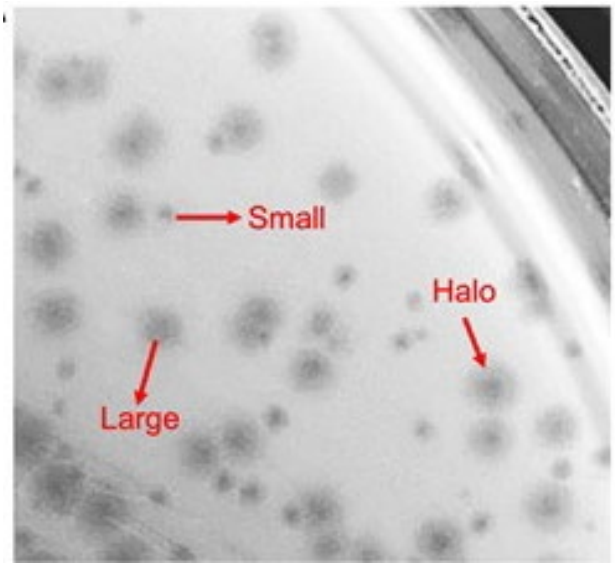


Figure 2. 3: Distinct plaque morphologies using the double overlay method. Plaques from an environmental sample display a halo and are of variable sizes, large (>3 mm in diameter) or small (1 mm in diameter), denoting different phage species.

After successive rounds of purification, phages are enumerated by serial dilution and titration. We visualise the phage by transmission electron microscopy, which reveals the structural details critical for classifying them into broad morphological groups. Subsequently, we grow phages in large quantities to extract DNA for sequencing. Although short-read sequencing technologies are more commonly used to sequence phage genomes, there has been a growing interest in using long-read sequencing. A recent study showed that long-read assemblies generated higher-quality, complete genome assemblies; however, they required polishing with short-read sequencing to correct frameshift errors (Papudeshi, Vega, et al., 2023). Recent advances in long-read platforms, especially PacBio's high-fidelity (HiFi) circular consensus sequencing (CCS), have dramatically reduced raw read error rates to below 1%(van Dijk et al., 2023). As a result, modern long-read assemblies now achieve both high completeness and accuracy, making them increasingly reliable for phage genomics and structural annotation efforts. Nonetheless, regardless of platform, achieving the right coverage depth is crucial: too low, and assemblies fragment; too high, and assembly algorithms struggle with excessive single-nucleotide polymorphisms (SNPs), leading to

misassemblies or failures. For this reason, the optimal range is typically 25× to 100× coverage, which balances completeness with assembly robustness—higher coverage can actually increase assembly errors due to overrepresented SNPs in the dataset(Turner et al., 2021).

Post-sequencing, the first step is to filter and trim the DNA sequences through quality control(Cantu et al., 2019; S. Chen et al., 2018; Tully, 2016; Wick, 2018) before assembling them to recover complete phage genomes(Antipov et al., 2020, 2022; D. Li et al., 2016). Phage genomes must be evaluated for quality because low-quality assemblies can significantly hinder accurate annotation and downstream analyses. To evaluate genome quality, reads can be mapped back to the assembled genome using MMseqs2(Steinegger & Söding, 2017). Key metrics assessed include the distance between paired ends, the mapping orientation of reads, and the evenness of read coverage across the genome (Shen & Millard, 2021; Turner et al., 2021). Genome completeness is further assessed using tools such as CheckV(Nayfach et al., 2021). Although the presence of direct terminal repeats (DTRs) can indicate terminal redundancy (Casjens & Gilcrease, 2009; Garneau et al., 2017), these signals are often obscured during the assembly process. Long-read assemblies, when combined with short-read assemblies, can help resolve phage DTRs more accurately(Elek et al., 2023). Alternatively, the Phables algorithm identifies and corrects terminal redundancy by exploring assembly graph structures(V. Mallawaarachchi et al., 2023).

After assembly, computational methods that include genome information are used to anticipate whether a phage has a virulent or temperate lifestyle. These methods encompass sequence similarity–based random forest classifiers(McNair et al., 2012), the presence of specific protein domains(Mavrich & Hatfull, 2017), and a hybrid approach that utilises both types of information(Hockenberry & Wilke, 2021).

2.3.2 Finding phage genomes from metagenomic sequencing datasets

Whole genome sequencing metagenomic datasets generated or obtained from public databases, such as the NCBI Sequence Read Archive (SRA), can be utilised to extract phages (Benler & Koonin, 2021; Levi et al., 2018). Metagenomics analysis has led to most of the phage genomes identified to date (Benler et al., 2021; Camarillo-Guerrero et al., 2021; Yutin et al., 2021). Tools like Hecatomb (M. Roach et al., 2022), VirFinder (Ren et al., 2017) and vCONTACT2(Bin Jang et al., 2019; Bolduc et al., 2017) identify viral-like sequences from metagenomes by searching for similarities against other known viral genomes.

In bacterial metagenomes, assembled contigs are clustered based on genomic signatures like GC content and read coverage, creating metagenome-assembled genomes (MAGs)(Papudeshi et al., 2017). Although traditional methods rely on clustering, recent tools like GraphBin(V. G. Mallawaarachchi et al., 2020; V. Mallawaarachchi et al., 2020) and MetaCoAG(V. G. Mallawaarachchi & Lin, 2022) have successfully utilised assembly graphs to generate high-quality

MAGs that retain information about how the reads are connected to form contigs. These steps can be transferred to viral metagenomes or viromes. Since phage genomes are small, complete phage genomes can be assembled after assembly, and the use of assembly graph information using Phables(V. Mallawaarachchi et al., 2023)) or the use of co-abundance and k-mer profile(Johansen et al., 2022; Nissen et al., 2021) can also be applied. These steps successfully resolve several phage genomes from metagenomic datasets.

Having successfully isolated, sequenced, and assembled a phage genome, or identified phage genomes from metagenomes, we start by characterising the structural features to unpack their functional properties.

2.3.3 Structural annotation

The first step involves identifying open reading frames (ORFs), which encode proteins and detect other elements such as tRNAs, noncoding RNAs, promoters, and transposons. Gene-calling algorithms use codon usage, GC content, start/stop codons, Shine–Dalgarno sequences, and short nucleotide sequences that bind to ribosomes during protein translation to identify open-reading frames(Hyatt et al., 2010; Larralde, 2022; McNair et al., 2019). The unique characteristics of phage genomes, including higher coding capacity, shorter intergenic regions, and more overlaps between coding domain sequences than bacterial genomes, require phage-specific tools(Akhter et al., 2012; Kang et al., 2017). Phage prediction tool, PHANOTATE, exploits these idiosyncrasies and improves phage gene identification(McNair et al., 2019). Beyond canonical gene structure, phages employ a range of strategies to expand their coding capacity. One such mechanism is programmed ribosomal frameshifting, which allows a single sequence to encode multiple protein products by shifting between reading frames(McNair et al., 2023). This not only increases coding density but also supports dynamic regulation of protein expression during infection.

In addition to protein-coding genes, phage genomes often carry RNA-encoding elements, such as transfer RNAs (tRNAs). Identifying these features requires a combination of sequence similarity and structural predictions. Tools like tRNAscan-SE(Chan et al., 2021; Chan & Lowe, 2019) and ARAGORN(Laslett & Canback, 2004) are commonly used to detect tRNA genes, often prior to ORF calling. Notably, some phages have evolved to reassign canonical stop codons, thereby encoding amino acids and modulating translational termination. This strategy can help regulate gene expression and delay host cell lysis(J. H. Campbell et al., 2013; Cook, Telatin, et al., 2023; Ivanova et al., 2014; Pfennig et al., 2023; Yuqian Zhang et al., 2023).

Phage genomes also exhibit chemical innovations beyond the sequence level. Many encode heavily modified or alternative nucleotides, which can help phages evade host defence systems or manipulate host metabolism for efficient replication(Bryson et al., 2015). Long-read sequencing platforms offer a distinct advantage here: both PacBio SMRT and Oxford Nanopore Technologies

are capable of detecting DNA modifications directly, providing insights into epigenetic features of phage genomes that are missed by short-read platforms(Nielsen et al., 2023).

2.3.4 Functional annotation

After identifying the protein- and RNA-encoding regions, the next step is to assign biological functions to each gene. Despite the fact that phages have smaller genomes than bacteria, it remains challenging to ascribe functions, as 65% of viral protein sequences lack known biological function(Susanna Grigson & Edwards, 2023).

Phage genes typically encode proteins in one of several categories of functions as follows: First, structural and packaging proteins include capsid, baseplate, and tail fibre proteins. These proteins form the outer capsid of the free phage. Second, phage integration, excision, and maintenance of the integrated state, including recombination and DNA binding proteins. Third, DNA replication proteins, including DNA polymerases and single-stranded binding proteins. Fourth, accessory metabolic genes often provide temporary metabolic boosts to the cell, increasing energy production while the phage is replicating. Finally, morons and genes of unknown function(Juhala et al., 2000).

Three popular databases target those clusters. PHROGs (prokaryotic virus remote homologous groups)(Terzian et al., 2021), VOGs (viral orthologous groups), and pVOGs (prokaryotic virus orthologous groups)(Grazziotin et al., 2017) build clusters of orthologous genes with shared functions. Orthologous genes have arisen through speciation (vertical transmission), and these databases attempt to disambiguate orthologs from proteins that evolved through duplication (paralogs) and thus may have similar but distinct functions. The PHROGs database is unique because it assigns each orthologous group to one of nine categories, allowing two levels of annotation. While phage genes are assigned an orthologous group, many have unknown functions; for example, 87% (33,747 out of 38,880 orthologous groups) of PHROGs lack a function. Most phage annotation pipelines use homology searches (e.g., with MMSeqs2 (Steinegger & Söding, 2017), HMMER(Eddy, 2009, 2011; Finn et al., 2011), or HHpred(Söding et al., 2005) to search each coding domain sequence with these databases.

Among the protein categories, structural proteins receive more annotations than the others due to their prominent role in the phage life cycle and direct involvement in host interactions and the external environment. Efforts to distinguish these structural proteins from those that serve alternative functions have involved various approaches, including using classifiers built with support vector machines(Charoenkwan et al., 2020; Fang et al., 2022; Manavalan et al., 2018). We further expanded on these methodologies by training an artificial neural network to discern 10 distinct classes of proteins, including primary and minor capsid proteins, based on their sequence composition(Cantu et al., 2020). Recent advancements in artificial intelligence have introduced large language models that take PHROG cluster sequences as input and derive protein functional

properties from the embedded amino acid sequences to improve protein annotations(Heinzinger et al., 2024; Kelly et al., 2023). Another approach to improve annotations has been through the prediction of the 3-dimensional protein structures and structure-based predictions(Jumper et al., 2021; Mirdita et al., 2022; Varadi et al., 2022). For instance, Say et al (2023) used Colabfold (Mirdita et al., 2022) to predict structures and used Foldseek(van Kempen et al., 2022) (van Kempen et al., 2022) to search a database. These emerging strategies improve functional annotations and deepen their biological and practical context.

Table 2. 1: Summary of the bioinformatics tools used for phage assembly and annotation described in this chapter

Category	**Name**	**Brief Overview**	**Github**	**Reference**
Quality control	Prinseq++	Quality control steps of sequenced reads	https://github.com/Adrian-Cantu/PRINSEQ-plus-plus	(Cantu et al., 2019; Schmieder & Edwards, 2011)
	Filtlong	Quality control steps of Nanopore sequenced reads	https://github.com/rrwick/Filtlong	(Wick, 2018)
	Fastp	Quality control	https://github.com/OpenGene/fastp	(S. Chen et al., 2018)
Assembly	MetaViralSPAdes	Assembling viral sequences	https://github.com/ablab/spades	(Antipov et al., 2020)
	MEGAHIT	Assembling prokaryotic sequences	https://github.com/voutcn/megahit	(D. Li et al., 2015)
	ViralFlye	Assembling viral sequences	https://github.com/Dmitry-Antipov/viralFlye	(Antipov et al., 2022)
Phage quality assessment	CheckV	Phage genome completeness	https://bitbucket.org/berkeleylab/checkv/src/master/	(Nayfach et al., 2021)

	Viralverify	Identify the viral sequences in the assembled contigs	https://github.com/ablab/viralVerify	(Raiko, 2021)
	Virsorter2	Detect virus genomes	https://github.com/jiarong/VirSorter2	(Guo et al., 2021; Roux et al., 2015)
	viralComplete	Phage genome completeness	https://github.com/ablab/viralComplete	
	MMSeqs2	Map the reads to the phage genome, to determine even genome coverage	https://github.com/soedinglab/MMseqs2	(Steinegger & Söding, 2017)
Phages from metagenomes	Phables	Identify genome termini signals from assembly graphs	https://github.com/Vini2/phables	(V. Mallawaarachchi et al., 2023)
	Hecatomb	Identify viral portion of the metagenomes	https://github.com/shandley/hecatomb	(M. Roach et al., 2022)
	vCONTACT2	Cluster viral genomes based on gene similarity	https://bitbucket.org/MAVERICLab/vcontact2	(Bin Jang et al., 2019)
	virFinder	Identify viral sequences from metagenomes	https://github.com/jessieren/VirFinder	
Phage structural annotation	PHANOTATE	Gene identification in phage genomes	https://github.com/deprekate/PHANOTATE	(McNair et al., 2019)

	Prodigal	Gene identification in prokaryotic genomes	https://github.com/hyattpd/Prodigal	(Hyatt et al., 2010)
	PRFect	Predict ribosomal frameshifting	https://github.com/deprekate/prfect	(McNair et al., 2023)
	tRNAscanSE	Detection and functional classification of transfer RNA genes	http://lowelab.ucsc.edu/tRNAscan-SE/	(Chan et al., 2021; Chan & Lowe, 2019)
	ARAGORN	Detection of mRNA and tRNA genes	http://www.ansikte.se/ARAGORN/	
	Mgcod	Accurate annotation considering stop codon reassignment	https://github.com/gatech-genemark/Mgcod	(Pfennig et al., 2023)
	pyrodigal	Gene identification in prokaryotic genomes and stop codon reassignment	https://github.com/althonos/pyrodigal	(Larralde, 2022)
Phage orthologous cluster gene database	PHROGs	Prokaryotic virus proteins database	https://phrogs.lmge.uca.fr/	(Terzian et al., 2021)
	pVOGs	Prokaryotic virus orthologous groups database	ftp://ftp.ncbi.nlm.nih.gov/pub/kristensen/pVOGs/home.html	(Grazziotin et al., 2017)
Phage functional annotation	mmseqs2	Map the proteins to a phage protein database	https://github.com/soedinglab/MMseqs2	(Steinegger & Söding, 2017)

	HMMER	Probabilistic models to predict the phage protein annotation	https://github.com/EddyRivasLab/hmmer	(Eddy, 2011)
	PVP-SVM	Prediction of phage virion proteins	www.thegleelab.org/PVP-SVM/PVP-SVM.html	(Manavalan et al., 2018)
	DeePVP	Classification of phage structural proteins	https://github.com/fangzcbio/DeePVP	(Fang et al., 2022)
	PVPred-SCM	Prediction of phage virion proteins	https://github.com/Shoombuatong/PVPred-SCM	(Charoenkwan et al., 2020)
	VirionFinder	Prediction of phage virion proteins	https://github.com/zhenchengfang/VirionFinder	(Fang & Zhou, 2021a, 2021b)
	PhANNs	Classification of phage structural proteins	https://github.com/Adrian-Cantu/PhANNs	(Cantu et al., 2020)
	Phynteny_transformer	Uses phage genome architecture, using synteny to predict function of hypothetical proteins	https://github.com/susiegriggo/Phynteny_transformer	
	AlphaFold	Protein structure prediction	https://github.com/deepmind/alphafold	(Jumper et al., 2021)
	ColabFold	Protein structure prediction	https://github.com/sokrypton/ColabFold	(Mirdita et al., 2022)

	FoldSeek	Protein structure search	https://github.com/steineggerlab/foldseek	(van Kempen et al., 2022)
	AMRFinderPlus	Antimicrobial resistance gene search	https://github.com/ncbi/amr	(Feldgarden et al., 2021)
	CARD	Antimicrobial resistance gene search	https://card.mcmaster.ca/	(Alcock et al., 2023, 2020)
	VFDB	Virulence factor database search	http://www.mgc.ac.cn/VFs/	(L. Chen et al., 2005, 2016; Liu et al., 2022)
	BACPHLIP	Bacteriophage lifestyle prediction	https://github.com/adamhockenberry/bacphlip	(Hockenberry & Wilke, 2021)
	PHACTS	Bacteriophage lifestyle prediction	https://github.com/deprekate/PHACTS	(McNair et al., 2012)
	RaFAH	Host prediction	https://sourceforge.net/projects/rafah/	(Felipe Hernandes Coutinho et al., 2021)
	Pharokka	Phage annotation pipeline	https://github.com/gbouras13/pharokka	(Bouras, Nepal, et al., 2023)
	MultiPhaTE2	Phage annotation and comparative analyses	https://github.com/carolzhou/multiPhATE2	(Ecale Zhou et al., 2021)
	PHOLD	Phage annotation using protein structures	https://github.com/gbouras13/phold	

2.3.5 Genome annotation and comparison workflows

Bioinformatics workflows include the multistep process described above to annotate phages. Pharokka(Bouras, Nepal, et al., 2023) layers the annotations, beginning with phage gene prediction, followed by similarity searches against large databases. Pharokka generates the files required to submit a new genome sequence to GenBank but also creates visualisations and provides output formats suitable for comparing genomes. In contrast, another workflow, MultiPhaTE2 can handle multiple phage genomes and perform comparative analyses(Ecale Zhou et al., 2021) in order to understand phage diversity and dynamics.

2.4 Naming a phage

2.4.1 Phage nomenclature

Most biological organisms are named using a binomial nomenclature and a standard taxonomic hierarchy. Phages were initially characterised based on the physical characteristics of the virion, including the capsid size, structure, genome size, and type of nucleic acid (double-stranded or single-stranded; DNA or RNA). This method received criticism as the genomic and proteomic information was not considered. As sequencing has become more affordable, more genomes are being sequenced and assembled to complete the genomes, allowing genomic information to be considered. Another challenge was that, unlike bacteria, which include conserved 16S rRNA gene sequences, phages do not have similar markers. Therefore, measuring phage biodiversity and taxonomic classification has had limited success. However, in 2002, Rohwer and Edwards proposed a phage proteome tree that is constructed by clustering the phage genomes based on overall protein similarity across the genomes (Rohwer & Edwards, 2002). Another approach suggested is based on the comparative genomics of the structural gene module as these genes have been observed to exhibit sequence relatedness(Asare et al., 2015; Proux et al., 2002). Other proposed methods include phage network clusters that classify phages based on gene content. Phage-phage similarities are represented as networks with each node representing a phage and the edges showing their shared similarity. Graph topology is then applied to distinguish their phylogenetic clusters and evolutionary cohesive units (Lima-Mendez et al., 2008). K-mer-based methods have also been applied where viral nucleotide usage is used to determine the ancestral relationships(Pride et al., 2006).

The first global attempt to classify viruses took place at the 1966 International Congress of Microbiology in Moscow, which established the ICTV to develop a universal taxonomic system for viruses. The International Committee on the Taxonomy of Viruses (ICTV) is the authoritative committee for classifying viruses. It delineates that classification into 15 taxonomic ranks between realm and species(Krupovic & ICTV Report Consortium, 2018). The ICTV ratifies viral nomenclature, and the Bacterial and Archaeal Viruses Subcommittee (BAVS) is responsible for phage nomenclature. The primary requirement for ratifying a new phage is depositing a complete genome sequence in 1 of the 3 International Nucleotide Sequence Database Collaboration (INSDC) member databases(Cochrane et al., 2016). By 2024, the ICTV had ratified 7 realms, 11

kingdoms, 22 phyla, 4 subphyla, 49 classes, 93 orders, 12 suborders, 368 families, 213 subfamilies, 3769 genera, 86 subgenera, and 16,215 species (https://ictv.global/taxonomy, as of May 2025).

2.4.2 Naming guidelines

The first step in naming a phage is to invent a novel name for the isolate. There are several guidelines for prescribing phages (e.g., the SEA-PHAGES program has a set of rules, https://phagesdb.org/namerules/, and members of the ICTV BAVS published an informal guide to choosing a name(Adriaenssens & Brister, 2017). The guidelines are similar: do not use an existing phage name or one like a current name, keep the name short (about 5 to 15 characters), and do not start with a number.

The next step is to determine its novelty in comparison to known phage sequences. We usually compare the phage's genome sequence to the phage genomes that are in existing databases. The Millard Lab maintains and regularly updates a list of complete phage genomes (https://millardlab.org/). The ICTV guidelines suggest that novel species are more than 5% different from existing species at the nucleotide level. In comparison, novel genera are more than 50% distinct from existing genera at the nucleotide level. Unique characteristics distinguish unknown taxonomic groups, including genome length, number of coding sequences, and phylogenetic clustering of marker genes such as portal protein, large terminase, and significant capsid genes. While electron micrographs remain useful to describe phage morphology and support taxonomic proposals, high-level classifications (eg, subfamily or family) is now entirely based on genome-based analyses, following the abolition of morphology based families (Turner et al., 2023).

When identifying a new phage, one should determine the highest possible taxonomic classification based on sequence similarity and explore additional characteristics to assign a proper name. For example, we recently described three new phages that infect *Bacteroides cellulosilyticus* (ICTV application in Appendix C). The phage's genome lengths and podovirus-like morphologies suggested they belonged to the *Crassvirales* order, and their sequence similarity to other crAss-like phages further supported this(Papudeshi, Vega, et al., 2023). Phylogenetic clustering and nucleotide similarity searches confirmed that 2 genomes belonged to known genera but were distinct enough to be considered new species. The third phage was unlike any known crAss-like phages, so we proposed it as a new genus using the templates on the ICTV website (https://ictv.global/taxonomy/templates). A complete list of guidelines for naming phages is available on the ICTV website (https://ictv.global/about/code). If the isolated phage does not fit the current known taxonomic groups genomically or morphologically, the ICTV will help define a new category.

2.5 Phages in a therapeutic context

Given their ability to specifically infect and lyse bacterial hosts, often harnessing diverse genetic machinery to enhance infection, phages have long been explored as therapeutic agents. Phages were first isolated to target pathogens and widely used to treat bacterial infections in the early twentieth century. However, in the post-war era, antibiotics' broad-spectrum activity displaced phages, and interest in them waned. In the early twenty-first century, the rampant spread of antibiotic-resistant microbes that abrogate the effects of antibiotics has led to renewed interest in using phages as therapeutic agents(Dedrick et al., 2023; Gordillo Altamirano et al., 2022; M. I. Kutateladze et al., 2009; Nir-Paz et al., 2024; Schooley et al., 2017; Teney et al., 2024).

However, not all phages are suitable for therapeutic use. Ideal candidates must have a strictly lytic lifecycle, a well-defined host range, restricted to the target bacterial species (but broad enough to cover multiple strains within that species), and lack genes encoding antimicrobial resistance or virulence factors. Temperate phages are unsuitable for phage therapy due to their potential to transfer antibiotic resistance, virulence genes, or pathogenicity islands to other bacteria(Ghequire & De Mot, 2015; Hyman, 2019; Waldor & Mekalanos, 1996). Conversely, virulent phages use lytic replication that induces the host cell's death post-infection. Therefore, obligate lytic phages are preferred for phage therapy, as they pose reduced susceptibility to confer phage resistance.

Isolated phages are first screened based on plaque morphology, aiming to select lytic phages that display distinct, clear plaques. Subsequently, these phages are sequenced to decipher their genetic makeup and identify their genes (Figure 2.2). These annotations are pivotal in determining phage interactions, replication, evolutionary dynamics, and host range. Accurate genome annotation also ensures that phages do not transfer unwanted genes, such as antimicrobial resistance genes and toxins, to new environments. Simultaneously, this process allows for selecting genomes with depolymerases that provide antimicrobial and antibiofilm activities(Calero-Cáceres et al., 2019; Colavecchio et al., 2017; Debroas & Siguret, 2019); anti–clustered, regularly interspaced, short palindromic repeats (CRISPRs); and antitoxin genes that help phages overcome bacterial immune systems(Azam & Tanji, 2019). To comprehensively assess phages' therapeutic potential and diversity, we classify them into a suitable taxonomy to place the phage in an evolutionary context.

The search for lytic phages is the same steps as phage isolation described above. Additional steps are added including plate-based methods such as cross-streaking(Hanna et al., 2012) and detecting spontaneous phage release using spot and immunity assay techniques(Gordillo Altamirano & Barr, 2021) can assist in identifying temperate phenotypes.

2.5.1 Examining phage suitability from their genomic information

In the context of phage therapy, specific genes are pivotal in determining if the phage is a potential candidate for phage therapy. For instance, "integrase" annotated genes are not preferred as they are markers for temperate lifestyles. While specific gene functions may confer phage therapy

advantages, this includes phage-encoded depolymerase genes, which degrade specific components of a bacterial surface, including extracellular polysaccharides and biofilm matrices(Knecht et al., 2019). Depolymerase activity can present as a halo around the plaques (Figure 2.3 A), and this gene activity makes the bacteria more susceptible to host infection, improving access to bacterial surfaces and reducing antibiotic dependence. Annotation pipelines may label these genes as “depolymerase,” “host-interaction protein,” “capsule degrading enzyme,” “biofilm matrix-degrading enzyme,” or similar (Hsieh et al., 2017; Shahed-Al-Mahmud et al., 2021; Wu et al., 2019).

Additionally, tail spike and tail fibre proteins include phage receptor binding proteins (RBPs), which are crucial for host interactions and indicative of host specificity. Often, the annotations list these proteins as “receptor-binding protein,” “receptor-recognising protein,” or similar variations(Boeckaerts et al., 2021, 2022, 2024). Tools such as PhageRBP have developed models based on detected RBP sequences and protein domains to visualise these proteins in the phage isolated. Experimentally, techniques such as the efficiency of plating, which quantifies plaque-forming efficiency and burst size, representing the number of progeny virions released per infected host cell, also characterise phage–host interactions(Ellis & Delbrück, 1939).

Other relevant genes include anti-host systems that phages develop to counter bacterial defence mechanisms against phage infections. Phages adapt by mutating RBPs at a high frequency in an order that is mediated by the activity of diversity-generating retroelements(Benler et al., 2018; Sharifi & Ye, 2019) to recognise the bacterial host. To bypass the restriction-modification systems in the bacterial hosts, phages alter their restriction sites using nucleotide modifications or reorientation(Egido et al., 2022). In response to bacterial immunity mechanisms, abortive infection mechanisms, which involve toxin–antitoxin action, phages inhibit the antitoxin degradation or produce antitoxin analogues to defend themselves. Finally, to evade clustered, regularly interspaced, short palindromic repeats (CRISPR/Cas), phages modify palindrome repeats or use anti-CRISPRs (Acr) proteins that hinder the system’s activity. Programs such as minCED identify CRISPR spaces in the phage genome(Bland et al., 2007; Huang et al., 2021; J. Wang et al., 2020).

Interestingly, phage resistance can develop quickly in vitro, but the results are variable in vivo(Egido et al., 2022). Unfortunately, while we can identify and examine the phage for known anti-host defence systems, several mechanisms still need to be identified. Mechanisms that underlie phage resistance are seldom studied; instead, there is more focus on clinical outcomes on phage cocktails or combination therapy (antibiotics with phages) to overcome phage resistance.

Moreover, the potential for phages to routinely transfer antibiotic resistance genes in the environment requires further clarification. Studies exhibit varying outcomes; some indicate increased antibiotic resistance gene transfer(Colavecchio et al., 2017; Modi et al., 2013), while others present opposing results(Enault et al., 2017). Similarly, many phages are known to carry

toxins, so we use virulence factor databases to identify gene products that might be involved in pathogenesis, including toxins(L. Chen et al., 2005, 2016; Liu et al., 2022). Therefore, aligning phage genes against boutique databases with highly accurate, manually curated data such as the National Centre for Biotechnology Information (NCBI) AMRFinderPlus(Feldgarden et al., 2021) and comprehensive antibiotic resistance database (CARD), virulence factor databases(L. Chen et al., 2005, 2016; Liu et al., 2022) is necessary.

2.6 Sharing phages

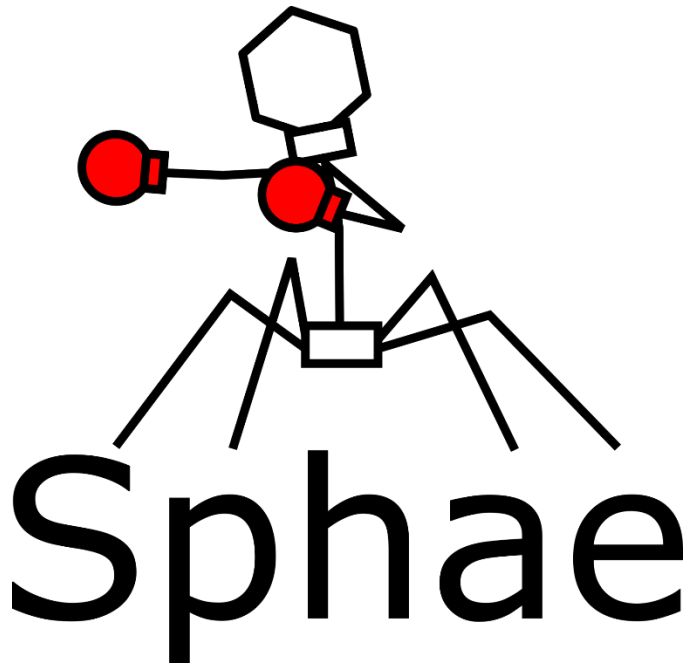
Making the bacteriophage genomes available through one of the three public INSDC repositories (National Centre for Biotechnology Information [NCBI], DNA Data Bank of Japan [DDBJ], or European Nucleotide Archive [ENA]) is good practice. It is required when publishing the work in any peer-reviewed journal. Each of the three databases has an online portal where complete, annotated genome sequences can be submitted.

Depositing the phage isolates to specialised repositories such as the American Type Culture Collection, National Collection of Industrial, Food, and Marine Bacteria, and Phage Australia facilitates access to these phages. Alternatively, local collections such as Phage Australia provide a platform for making phage information and their characteristics publicly accessible.

2.7 Conclusions

Bacteriophages stand out as some of the most diverse and ecologically influential entities on Earth, playing pivotal roles in shaping microbial evolution, structuring communities, and maintaining ecosystem function. Yet their diversity, mosaic genomic architectures, and dynamic interactions with bacterial hosts present unique challenges for systematic study. Accurate annotations are essential to navigating this complexity; they underpin our ability to precisely identify and compare phages, uncover their ecological roles, and understand the mechanisms driving their interactions and evolutionary trajectories. Concurrently, standardised nomenclature approaches not only facilitate clear communication and data integration across studies but also provide a critical foundation for advancing applied efforts, from safe and effective phage therapies to microbiome manipulation. In this way, careful genome characterisation and classification directly address the knowledge gaps, enabling us to better appreciate and harness the extraordinary biology and ecological significance of phages.

CHAPTER 3
PHAGE BIOINFORMATICS TOOLKIT



This chapter is based on the published article — **Papudeshi, B.**, Roach, M. J., Mallawaarachchi, V., Bouras, G., Grigson, S. R., Giles, S. K., ... & Edwards, R. A. (2025). Sphae: an automated toolkit for predicting phage therapy candidates from sequencing data. Bioinformatics Advances, 5(1), vbaf004. https://doi.org/10.1093/bioadv/vbaf004.

Statement on the Use of Generative Artificial Intelligence (AI): Generative AI tools—including ChatGPT by OpenAI and Grammarly—were used during the preparation of this chapter for language editing purposes, such as enhancing sentence clarity, grammar, and structure. Additionally, Microsoft Copilot was employed to assist in identifying and resolving coding errors during the development of the Sphae workflow described in this chapter. These tools were not used to generate original scientific content, perform data analysis, or contribute to the interpretation of research findings. All intellectual, analytical, and conceptual contributions presented herein are my own, in full accordance with Flinders University's policy on the responsible use of generative AI in research.

Preface

This chapter is based on the published article, “Sphae: an automated toolkit for predicting phage therapy candidates from sequencing data,” for which I am the first author. The work was published in *Bioinformatics Advances* in January 2025 and addresses the lack of reproducible, domain-specific workflows for genome assembly, annotation, and evaluation of phages, particularly in the context of phage therapy. I tailored both the tool and the original publication to highlight therapeutic screening, recognising that one of the major applications of such a tool would be in supporting phage therapy research. This workflow has been downloaded 20,000 times, underscoring its impact and adoption within the phage research community.

In developing Sphae, multiple tools were integrated into a streamlined pipeline capable of rapidly characterising phages and identifying key genomic markers. For this thesis, I have adapted and expanded the content to align with my broader research aims. The content of this chapter is adapted from the original publication, which details the development and capabilities of the Sphae toolkit. I have added a “Conclusion” section to better align this work with the broader research aims and to tie it into the overall thesis structure.

Overall, this chapter establishes the methodological foundation for the thesis. Sphae is not limited to therapeutic phages; its modular steps can equally be applied to other phages, including temperate types, making it broadly useful for diverse ecological and evolutionary studies. By providing a robust framework for phage genome characterisation, it directly supports the comparative analyses and investigations of phage–bacteria interactions explored in the subsequent chapter, ultimately advancing our understanding of microbe–host dynamics.

Statement of authorship

As the primary author of this chapter, I led the design, development, and validation of the Sphae bioinformatics workflow. I was responsible for conceptualising the toolkit architecture, integrating multiple open-source tools, and ensuring the pipeline was scalable, reproducible, and accessible to the broader research community. In developing this workflow, I sought feedback from domain experts in phage genomics and workflow engineering to refine tool selection and usability, while maintaining the codebase and documentation. I also led the manuscript writing, figure generation, and benchmarking analyses, ensuring clear presentation of both the technical aspects and biological relevance of the workflow. Although this project involved collaborative input—particularly around test datasets, comparative evaluation, and end-user testing—the work reflects my leadership in developing computational solutions tailored to the challenges of phage genome analysis. Below is a breakdown of the author's contributions:

Author	Contribution

Bhavya Papudeshi	Conceptualised and developed the Sphae workflow, performed data analysis, wrote the manuscript, and contributed to its editing
Michael J. Roach	Assisted in computational development, workflow optimisation, and data analysis.
Vijini Mallawaarachchi	Assisted in computational development, workflow optimisation, and data analysis.
George Bouras	Contributed to data analysis and manuscript editing.
Susanna R. Grigson	Contributed to data analysis and manuscript editing.
Sarah K. Giles	Assisted with data collection and validation.
Clarice M. Harker	Assisted with data collection and validation.
Abbey L. K. Hutton	Assisted with data collection and validation.
Anita Tarasenko	Assisted with data collection and validation.
Laura K. Inglis	Assisted with data collection and validation.
Alejandro A. Vega	Assisted with data collection and validation.
Cole Souza	Assisted with data collection and validation.
Lance Boling	Assisted with data collection and validation.
Hamza Hajama	Assisted with data collection and validation.
Ana Georgina Cobián Güemes	Provided insights into phage therapy applications and validation.
Anca M. Segall	Provided research supervision, guided project development, and contributed to manuscript editing
Elizabeth A. Dinsdale	Provided research supervision, guided project development, and contributed to manuscript editing
Robert A. Edwards	Provided research supervision, guided project development, and contributed to manuscript editing

The contributions of each co-author have been explicitly stated, and their permission to include these works has been obtained as per Flinders University's Authorship of Research Output Procedures (Appendix A). I also affirm that, while this research was strengthened through these collaborations, all overarching hypotheses, research aims, analyses, and interpretations presented in this thesis are my own. The coordination of interdisciplinary methods, the synthesis of findings across chapters, and the articulation of their broader significance reflect my independent intellectual contributions.

Sphae: an automated toolkit for predicting phage therapy candidates from sequencing data

Abstract

Phage therapy presents a viable alternative for treating bacterial infections amid growing concerns about antimicrobial resistance. Its success relies on selecting safe and effective phage candidates that require comprehensive genomic screening to identify potential risks. However, this process is often labour-intensive and time-consuming, hindering rapid clinical deployment. We therefore developed Sphae, an automated bioinformatics pipeline designed to streamline the assessment of the therapeutic potential of a phage in under 10 minutes. Using Snakemake workflow manager, Sphae integrates tools for quality control, assembly, genome assessment, and annotation tailored specifically for phage biology. Sphae automates the detection of key genomic markers, including virulence factors, antimicrobial resistance genes, and lysogeny indicators such as integrase, recombinase, and transposase, which could preclude therapeutic use. Among the 65 phage sequences analysed, 28 showed therapeutic potential, 8 failed due to low sequencing depth, 22 contained prophage or virulent markers, and 23 had multiple phage genomes. This workflow generates a report to quickly assess phage safety and suitability for therapy. Sphae is scalable and portable, facilitating efficient deployment across most high-performance computing and cloud platforms, accelerating the genomic evaluation process. Sphae source code is freely available at https://github.com/linsalrob/sphae, with installation supported on Conda, PyPi, Docker containers.

3.1 Introduction

With the escalating global challenge of antimicrobial resistance (AMR) comes an increasing demand for alternative treatments against bacterial infections. Bacteriophages, also known as phages, are viruses that infect bacteria and are ubiquitous in the environment. The use of phages to treat bacterial infections is being explored worldwide as an alternative to antimicrobials. In the USA, Australia, and parts of Europe, this treatment option is typically administered as a last-resort care for severely ill patients under compassionate use (J. Singh et al., 2023; Y. Wang et al., 2022). For phage therapy to be most effective, thorough safety assessments of the phage isolates must be performed before treatment. This includes experimental testing to confirm that the phage is a pure isolate and can infect the targeted pathogen variant. While phages can be co-isolated with satellites or other phages, these complex systems are more challenging to characterise and regulate, and therefore a pure isolate is generally preferred as the starting point for therapeutic development, before combining phages into cocktails. Additionally, phages are screened to specifically select lytic phages that infect, replicate, and quickly kill the bacterial host over temperate or lysogenic phages that integrate into the host genome during infection and remain stable(Bondy-Denomy et al., 2016; Gordillo Altamirano & Barr, 2019). Temperate phages are not preferred as they can protect the host by improving its fitness and may confer phage resistance through repressor-mediated immunity and/or superinfection exclusion (Bondy-Denomy et al., 2016;

Samson et al., 2013). Additionally, phages are screened for large burst sizes and short latent periods to ensure quick and sustained infectivity and high adsorption rates to ensure effectiveness at low concentrations. The presence of these qualities is essential for high virulence to overwhelm the bacteria quickly(Rohde et al., 2018).

Phages and bacteria often engage in evolutionary arms race where bacterial defence mechanisms like CRISPR-Cas systems co-evolve with phage countermeasures and can propagate throughout bacterial populations(Fineran et al., 2009; Sorek et al., 2008; Yirmiya et al., 2024). Interestingly, it has been shown that the development of phage resistance by the host often coincides with a loss of antibiotic resistance(Oromí-Bosch et al., 2023), allowing antibiotics to augment phage therapy by eliminating bacteria as they switch from an antibiotic- to a phage-resistant state. This synergy can be enhanced by using phage cocktails consisting of a range of phages with a combined specificity for a broad host range to further reduce the evolution of phage resistance within a bacterial infection. Especially if the cocktail includes phages with distinct mechanisms of host recognition and/or host factors, so that resistance to one phage does not confer resistance to all phages(Altamirano & Barr, 2021; Torres-Barceló et al., 2022; Wandro et al., 2022). Consequently, phage therapy has significant potential to be an effective treatment strategy for combating antibiotic resistance.

Efforts have been renewed to isolate phages for antibiotic-resistant bacterial pathogens in Europe, the USA, and Australia. The use of bacteriophages as therapeutic applications is subject to stringent regulatory oversight, particularly concerning toxin production and AMR genes. Ideally, phage isolates are sequenced during screening to predict their genetic potential for safety and efficacy(Cobián Güemes et al., 2023; S. R. Grigson et al., 2023; Luong et al., 2020)Bioinformatics analysis is now an indispensable component of this approach, ensuring that sequencing data is processed efficiently to guide decision-making. For time-sensitive applications, rapid and scalable computational tools are essential, especially for large-scale screening initiatives. However, current analysis workflows can be time-consuming and require manual intervention, which limits their throughput and scalability.

Phage genomes are typically small, with a median size of about 40 kb, and can usually be assembled easily into complete genomes. However, the assembly process using default assembly tools obfuscates genome termini signals(S. R. Grigson et al., 2023). The recently published Phables algorithm (V. Mallawaarachchi et al., 2023) uses the assembly graph and read coverage to identify and correctly resolve genome termini. Alternatively, the HYbrid and Poly-polish Phage Assembly method utilises long-read assemblies in combination with short-read sequencing(Elek et al., 2023). Phage genome sequences can also be contaminated with contigs from the bacterial host due to contamination during DNA extraction or the induction of host prophages, resulting in mixed phage lysates (Cobián Güemes et al., 2023). Tools such as ViralVerify(Raiko, 2021) identify

and remove putative host contigs. Additionally, phage assemblies may be split over multiple contigs. Therefore, it is important to utilise tools such as CheckV(Nayfach et al., 2021) to determine if the assembly represents a single complete phage genome, and in identification of direct terminal repeats (DTRs). In some cases, even a single phage lysate can yield multiple phage genomes, making such tools indispensable for accurate phage identification(Gendre et al., 2022).

Once assembled, genome annotation tools like Pharokka(Bouras, Nepal, et al., 2023) predict genes and assign biological functions using database searches against genes with known functions. However, assigning biological functions remains challenging, as 65% of viral proteins lack sequence homology to a protein with a known function(Susanna Grigson & Edwards, 2023). Nonetheless, specific genes that serve as markers for temperate lifestyle (such as integrase genes) or confer phage resistance, including a search for toxin, virulence factors, or AMR, are screened for. The acquisition of such genes by bacteria through infection is considered phage conversion and poses risks for therapeutic application due to the potential propagation of resistance or virulence within bacterial populations. These genes are exclusionary criteria for phage therapeutic use; however, in cases where lytic phages are unavailable, engineered phages with disabled integrase and repressor functions have been demonstrated as an option(Dedrick et al., 2019; Strathdee et al., 2023). Meanwhile, anti-CRIPSR (Acr) proteins against their host and depolymerase genes are preferred as they can be advantageous in infection(S. R. Grigson et al., 2023). However, running all these tools sequentially is time-consuming and resource intensive.

Previous studies describe step-by-step tutorials and guidelines for assembling high-quality phage genomes and best practices for predicting and annotating their genes(S. R. Grigson et al., 2023; Shen & Millard, 2021; Turner et al., 2021). The steps listed above can be intimidating for novices, so workflow managers can run all these steps under one command to make it easier.

In this chapter, I have developed Sphae, a Snakemake-powered rapid phage characterisation workflow, designed to streamline the selection of phage therapy candidates. This name is derived from “spae” which means “to foretell” with a modified spelling (s-ph-ae) denoting its specific focus on predicting a phage’s suitability for therapeutic use. This workflow enables the rapid selection of phage therapy candidates based on their genomic potential, leading to faster medical interventions and improved patient survival outcomes. We developed this workflow to ensure reproducibility and consistency in the outputs, as using different databases and software versions can influence the results. This workflow is easy to install and run, generating a final summary text file with phage characteristics that anyone can examine to determine the therapeutic potential of a phage.

3.2 Methods

3.2.1 Workflow input

Sphae requires sequencing reads in fastq format, either paired-end short reads from Illumina or MGI sequencing platforms or unpaired long reads from Oxford Nanopore sequencing platforms. Oxford Nanopore raw sequencing output is in fast5 or pod5 format, which must be basecalled using Guppy or Dorado to convert the reads to fastq format before running this workflow.

3.2.2 Snakemake workflow manager

We utilised the Snakemake workflow manager(Köster & Rahmann, 2012), which facilitates the automated installation of packages and dependencies. We also utilised Snaketool, which provides a user-friendly command line interface for Sphae to make running the pipeline as easy as possible(M. J. Roach et al., 2022). Workflow managers such as Snakemake provide scalability, reproducibility, and re-entrancy(Welivita et al., 2018), parallel processing of multiple samples, and integration for running commands and various steps on high-performance computing (HPC) systems and cloud-based environments. Therefore, we employed this template to leverage the capabilities of the Snakemake workflow manager in developing our pipeline for carrying out quality control, genome assembly, and annotation (Figure 3.1).

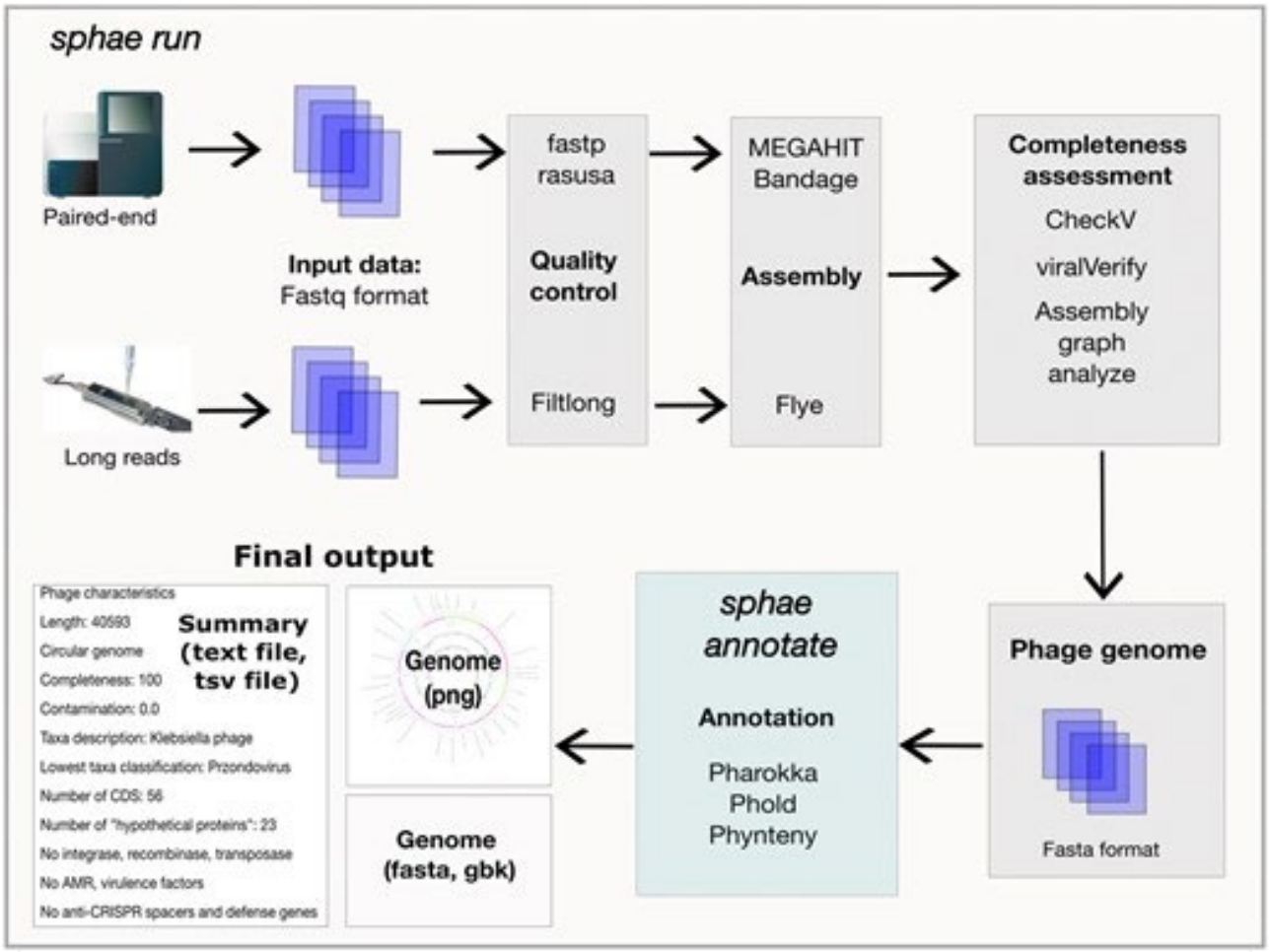


Figure 3. 1: Sphae workflow overview. The workflow processes sequencing reads from short- and/or long-read data in fastq format. The command sphae run, starts with quality control, filtering out low-quality reads and adaptor sequences. Processed reads are assembled, and the resulting assemblies are processed to confirm complete phage genomes in each sample. The phage genomes are annotated to identify the genes and assign biological functions. The final output folder contains the assembled genome (fasta format),

annotations (GenBank format), a Circos plot (PNG format), and a summary text file detailing phage characteristic.

3.2.3 Steps in workflow

1. Quality control: Fastp(S. Chen et al., 2018) and Filtlong(Wick, 2018) are run to remove low-quality reads and trim adaptor sequences to ensure only high-quality reads are retained for downstream analysis.
2. Read subsampling: Rasusa(Hall, 2022) is run to subsample up to 10 million base pairs per sample to keep an ideal 25× to 100× genome coverage for phage assembly(S. R. Grigson et al., 2023).
3. Assembly process: Paired-end short reads are assembled using MEGAHIT(D. Li et al., 2015, 2016), while long-read assemblies are conducted using Flye(Kolmogorov et al., 2019). Although recent advances in Nanopore sequencing chemistry have reduced the need for long-read polishing(Bouras, Houtak, et al., 2024), Medaka(Nanoporetech Consortium, 2022) is used to correct older, more error-prone reads.
4. Completeness assessment: Assembled contigs are classified using:
    a. ViralVerify(Raiko, 2021) to identify viral, plasmid, or bacteria origin using gene content,
    b. CheckV(Nayfach et al., 2021) to determine the completeness of the viral contigs by comparing the genomes against a database of viral genomes and identifying the conserved gene markers and regions,
    c. Custom Python script to assess contig connectivity within the assembly graph(V. Mallawaarachchi et al., 2023), and
    d. Overall, only contigs classified as viral by ViralVerify (longer than 1000 base pairs and having a completeness score of over 70%) are selected for further analysis. In cases where a sample contains multiple genomes, each genome is saved as a separate phage genome.
5. Gene annotation is performed using Pharokka(Bouras, Nepal, et al., 2023). Gene prediction is conducted using PHANOTATE(McNair et al., 2019) or Pyrodigal(Larralde, 2022), followed by functional annotation through comparison with the PHROGs database(Terzian et al., 2021). In addition, genes are also run against:
    a. AMR gene databases: CARD(Alcock et al., 2023),
    b. Virulence factor database; VFDB (Liu et al., 2022),
    c. CRISPR recognition tool; MinCED(Bland et al., 2007),
    d. anti-CRISPR (Acr) gene detection using AcrDB(Huang et al., 2021),
    e. anti-phage systems using DefenseFinder(Tesson et al., 2022), and
    f. tRNA genes using tRNAscanSE(Chan et al., 2021) and tmRNA using ARAGORN(Laslett & Canback, 2004).
6. Taxonomic assignment is performed within Pharokka, via MASH(Ondov et al., 2016) that compares the genome against the phage INPHARED database(Cook et al., 2021).

7. Hypothetical gene analysis: To address the prevalence of remaining hypothetical genes, Sphae uses:
    a. Phold applies the ProstT5(Heinzinger et al., 2024) protein language model to generate a structural representation for each gene. These are compared against a database of predicted phage protein structures using FoldSeek(van Kempen et al., 2022) to assign potential functions.
    b. The resulting GenBank files are further processed through Phynteny(Susie Grigson & Mallawaarachchi, 2023), which utilises a long short-term memory model trained with phage synteny to refine gene function predictions.
8. Phage therapy suitability: The annotated genome is systematically analysed for key markers, including integrase, recombinase, transposase, toxins, AMR, and virulence genes.

3.2.4 Workflow output

Each workflow step yields a set of files, not all of which are directly pertinent to deciding the therapeutic potential of the phage. Sphae workflow produces a “FINAL” directory containing essential summary files to streamline the output. These files include:

- assembled phage genome (.fasta)
- phage annotations (.gbk)
- genome plot (.png)
- summary table (.tsv): annotations from the three tools, tracking which tool assigned a function to the gene
- summary (.txt): phage characteristics described in Table 3.1

Table 3. 1: Phage characteristics and annotations for sample Bc01

Phage characteristic	**Value**	**Explanation**
Sample name	Bc01	Sample name
Total length of reads after QC and subsampling	5 363 156 bp	Total length of reads used for assembly to help calculate genome coverage
Length	100 743	Length of the phage genome assembled
Circular	False	Was the genome assembled to be circular, according to the information provided in the assembly graph? For more information, you can visualise the file ending in .gfa with Bandage (Wick et al., 2015).

Graph connections	0	If the assembly generated fragmented contigs due to low coverage, the graph shows potential connections, offering clues for identifying terminal repeats and low-complexity regions. Visualise the file ending in .gfa with Bandage (Wick et al., 2015).
Direct terminal repeat (DTR) found	–	Is DTR detected by CheckV (Nayfach et al., 2021) in the phage contig
Completeness	100.0	Phage completeness score from CheckV
Contamination	0.0	Contamination score from CheckV
Taxon description	*Kehishuvirus* sp.tikkala	Assigned taxon name from Pharokka (Bouras, Nepal, et al., 2023) output, comparing the phage genome against the INPHARED database (Cook et al., 2021) using Mash (Ondov et al., 2016)
Taxa result: matching hashes	972/1000	How close the phage isolated is to the assigned taxon. Results from Pharokka Mash sketch against the INPHARED database
Lowest taxon classification	*Kehishuvirus*	The lowest taxon rank assigned
Isolation host of the described taxa	*Bacteroides cellulosilyticus*	Bacterial host of the assigned taxa from the INPHARED database
Number of CDS	154	Number of genes identified in the genome from Pharokka results
Total number of CDS annotated as “hypothetical protein”	91	Counting only the genes annotated as hypothetical, which have not been assigned a biological function or have ambiguous descriptions in Phynteny output
GC content (proportion)	0.35	GC content from Pharokka result
Percent coding density	91.3	Phages generally have high coding capacity, so if the density is low, it could indicate issues with gene calling for this phage
Prophage or temperate lifestyle markers	No integrases, recombinases, or transposases found	These genes indicate the phage can have a temperate lifestyle, which would most likely exclude it from use in therapy. Results from Pharokka, Phold, and Phynteny searches
Toxin genes	No toxins found	Search for genes with “toxins” in the gene description from the final Phynteny output and from VFDB search

Virulence genes	No antimicrobial resistance (AMR) genes found; no virulence factors found	Search against the CARD (Alcock et al., 2023)and VFDB (Liu et al., 2022) databases using Pharokka and Phold results
Defence genes	No anti-CRISPR or spacers found, no defence genes found	Pharokka and Phold search the genes against ACR (Huang et al., 2021) and DefenseFinder (Tesson et al., 2022) databases

3.2.5 Phage sampling and sequencing

Escherichia coli strain CoGEN001851 (BEI Resources: Catalogue number NR-4359) was received as a glycerol stock from BEI Resources. The strain was plated on Brain-Heart Infusion medium, supplemented with 1.5% agar (w/v), MgSO4, and MgCl2 to final concentrations of 10 mM and 2 mM, respectively. The culture plates were incubated at 37°C for 24 h. The phages were isolated from untreated sewage water (influent) collected from the waste treatment plant in Cardiff, California, as described in Papudeshi et al. (2023). An isolated plaque was selected from each plate and purified further. Phage DNA was then extracted, and E.coli phages were sequenced using Oxford Nanopore MinION sequencing according to the manufacturer's instructions, using Oxford Nanopore Rapid Barcoding Sequencing Kit (SQK-RBK0004) and sequenced on Flowcell R9.4.1 (FLO-MIN106) as described in Papudeshi et al. (2023). The sequencing data were deposited to the Sequence Read Archive (SRA) in BioProject PRJNA737576. The resulting fast5 reads were run through Guppy v6.0.1 with model dna_r9.4.1_450bps_hac for the Nanopore sequenced isolates. The resulting fastq reads were then run through the Sphae workflow.

3.2.6 Datasets

The workflow was tested on phages isolated from the above commercially available *E. coli strains, as well as* with publicly available sequence reads or genomes for *Klebsiella*, *Salmonella*, and *Achromobacter* phages (Table 3.2 and Table S3.1). Additionally, we included one dataset with five samples that included mixed *Caudovirictes* phages from multiple bacterial hosts to demonstrate the potential of Sphae workflow in assembling and separating each phage (Table 3.2 and Table S3.1). The reads were downloaded from SRA using sra-tools in fastq format as input for Sphae.

Table 3. 2: Phage study summary

Study	**Number of phage samples**	**Sequencing platform**	**Bacterial host**	**Bioproject**	**Reference**

E.coli phages	14	MinION	*E.coli* strain CoGEN001851 (BEI Resources: Catalogue number, NR-4359)	PRJNA737576	This study
Klebsiella phages	20	MinION, Illumina NextSeq	*Klebsiella michiganensis, Klebsiella oxytoca, Klebsiella quasipneumoniae, Klebsiella variicola*	PRJNA914245	(Elek et al., 2023)
Salmonella phages	11	Illumina MiSeq	*Salmonella entericasubsp. enterica serovar Typhimurium* (ATCC 14028S)	PRJNA914245	(Gendre et al., 2022)
Achromobacter phages	15	Illumina MiSeq	*Achromobacter xylosoxidans* strain 19–32	PRJEB33638	(Cobián Güemes et al., 2023)
Mixed *Caudovirictes* phages	5	Illumina MiSeq		PRJNA222858	NA

3.2.7 Benchmarking

We benchmarked Sphae's performance on five published datasets with 65 samples (Table 3.2) to compare its functionality and performance. These datasets include known phages in each sample as they were experimentally isolated, assembled, and annotated to serve as reliable references. Previous studies have described guidelines(S. R. Grigson et al., 2023; Shen & Millard, 2021; Turner et al., 2021) for assembling high-quality phage genomes and annotating their genes; we have used these tutorials as a framework to develop Sphae. All programs and dependency versions used for benchmarking are listed in Table 3.3. This adaptable workflow is designed with versatility, ensuring compatibility with future updates and new software. As there are no comparable workflows, we assessed workflow performance using datasets with varying complexities, different sample numbers, and different sequencing platforms, including samples with single or multiple phages.

Table 3. 3: Programs and dependency versions used for benchmarking Sphae on the five projects

Steps	Sequencing	Tools	version
Quality control	paired-end	Fastp	v0.23.4
	paired-end	rasusa	v2.0.0
	longread	Filtlong	v0.2.1
Assembly	paired-end	MEGAHIT	v1.2.9
	longread	Flye	v2.9
	longread	medaka	v1.11.3
Assembly checks	paired-end, longread	ViralVerify	v1.1
	paired-end, longread	CheckV	v1.0.1
Annotation	paired-end, longread	Pharokka	v1.7.1
	paired-end, longread	Phold	v0.1.4
	paired-end, longread	Phynteny	v0.1.13

Running the workflow in parallel mode processes each phage genome as an individual job, thus speeding the output time. This can be set up on HPC systems using a user-provided profile.

3.2.8 Runtime performance comparison

To evaluate Sphae's runtime, we measured the wall-clock runtime on a RedHat Linux release 8.10 machine with an AMD EPYC 7551 CPU @ 2.55 GHz. We analysed sequencing data for a *Klebsiella* phage Amrap using both paired-end and long-read sequencing methods with default settings in Sphae. The analysis was conducted on six or eight threads and 32 GB of memory to mimic commonly available consumer hardware. Each paired-end, long-read with polishing, long-read without polishing, and annotate modes were executed five times with the same command, and the median wall-clock times with 8 and 16 threads were recorded.

3.2.9 Data availability

All raw data used to assess Sphae are publicly available through the NCBI Sequence Read Archive (SRA), with SRA accession numbers listed in Table S3.1. The Sphae workflow code is openly accessible on GitHub at https://github.com/linsalrob/sphae.

3.3 Results

3.3.1 Determining complete genomes from assembly

Depending on the complexity and genome coverage of the phage, assembly steps can result in different results (Figure 3.2). Ideally, the phage genomes are completely assembled into circular or linear genomes (Figures 3.2A and 3.2 B). In other cases, the DTR that plays a role in packaging cannot be resolved due to its low complexity during assembly; in this case, the code considers the longer contig as a final genome representation (Figure 3.2C). Similarly, the DTR regions can cause multiple genomes to be tangled in an assembly graph (Figure 3.2D). In this case, all the contigs identified as complete phage genomes by CheckV are considered separate phage genomes from a sample. In the final case, the assembly generates fragmented phage genomes; if the contigs are long enough to determine if they are components of a phage genome (Figure 3.2E), or they may be too fragmented, making it challenging to determine if they are viral (Figure 3.2F). In both the latter cases, the poor quality of the assembly can lead to poor annotation and, therefore, they are not considered further in the workflow.

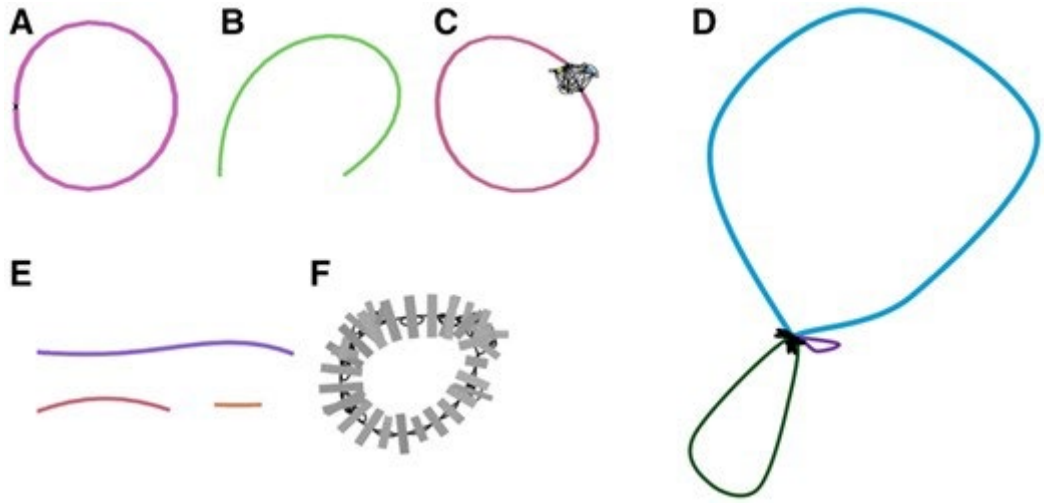


Figure 3. 2: Assembly graphs visualised using Bandage: A) complete circular phage genome, B) complete linear phage genome, C) near-complete phage genome, with terminal repeats hard to assemble, D) multiple phage genomes in one assembly, E) fragmented phage genome, likely due to low genome coverage, and F) multiple phage genomes in one assembly—in this case, there are three phages in the sample.

3.3.2 Assembly summary

We assembled 65 samples across the five datasets, described in Table 3.1, using Sphae v1.4.3 with the tools and their versions listed in Table 3.3, which resulted in the assembly of 84 phages. In the summary output (Table S3.1), we indicate if the assemblies have failed, if the assembly itself has not produced contigs, or if the assembled contigs were fragmented. This could occur due to issues with the quality of the sample or biological properties on of the phage genome.

In the *E.coli* dataset, some sequences lacked sufficient genome coverage, resulting in unassembled phage genomes (Figure S3.1). Seven of the 14 samples were assembled, four generated fragmented assemblies, and three failed during assembly (Figure S3.1B). This dataset highlighted how Sphae captures the presence of poorly sequenced samples, suggesting to the user that further sequencing data is required to generate suitable genomes for these phages.

In the case of *Klebsiella* phages, short- and long-read sequences were assembled separately, revealing differences between the two sequencing platforms. Paired-end reads generated complete, circular assemblies with assembly graphs, including one sample featuring one region with multiple contigs tangled together (Figures 3.2C and S3.2). Conversely, Nanopore read assemblies resulted in complete, linear phage genomes (Figures 3.2B and S3.3), lacking the DTR region (Table S3.2). With the *Salmonella* and *Achromobacter* phage datasets, complexity arose from samples containing multiple phage genomes. While Sphae was able to assemble phage genomes for each sample (Figures S3.4 and S3.5), two samples (Se_F6 and Salfasec_13) contained two assembled phage genomes (Figure S3.4B and S3.4J), and two samples (Se_F3 and Se_F1) contained three phage genomes (Figure S4C and E). This observation aligns with the genome characteristics outlined in the original publication(Gendre et al., 2022), confirming the presence of multiple phages in specific samples. However, three of the 11 samples were potentially contaminated with *E. coli* ϕX174, likely introduced during the sequencing process. Many Illumina sequences contain ϕX174 contamination, as it is used as a spike-in during 16S rRNA sequencing. Similarly, the *Achromobacter* phage dataset contained multiple samples with two phage genomes per isolate, with 11 out of the 15 phages having genome lengths of either 30, 40, or 70 kb. The assembly graph illustrates a structure similar to Figure 3.2D, with two phages connected by the DTR region (Figure S3.5).

We further ran Sphae on a dataset comprising five mixed *Caudoviricetes* samples (SRR8788475, SRR8869231, SRR8869234, SRR8869239, and SRR8869241), demonstrating Sphae's capacity to accurately resolve and separate multiple phages within each sample. For instance, sample SRR8788475 included four phages, and Sphae assembled all four phages (Figure S3.6B, Table S3.2); similarly, two phages in SRR8869231 were assembled (Figure S3.6C), three phages from SRR8869239 (Figure S3.6E) and SRR8869241 (Figure S3.6F). Interestingly, sample SRR8869234 was listed to include two phages, but Sphae assembled three phages, *Staphylococcus*, *Klebsiella*, and *Enterobacter* phage (Figure S3.6D). Importantly, the resulting assembly graphs across all samples were connected by short sequence fragments (Figure 3.2D), reflecting the complexity of resolving multiple phages.

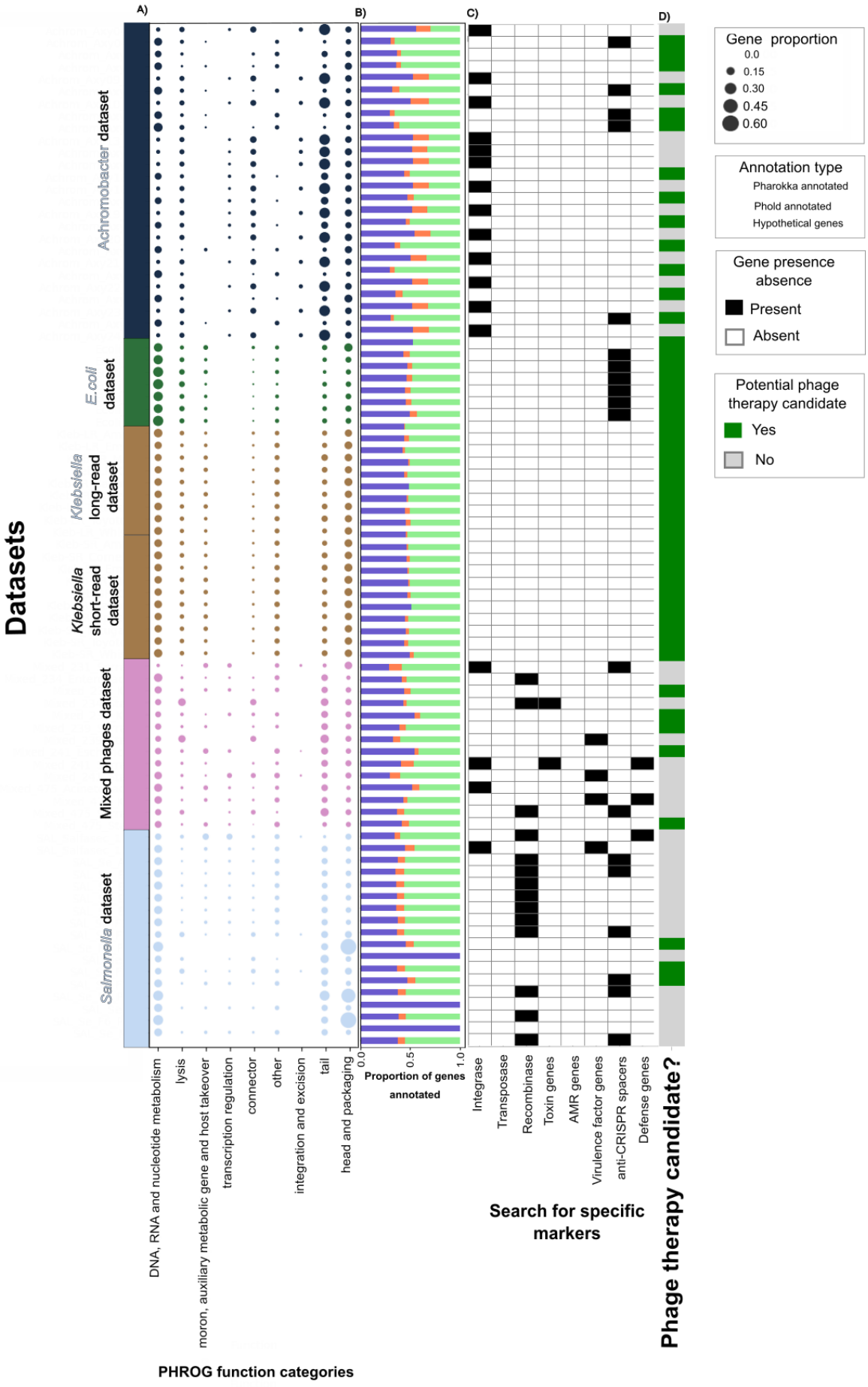

A)
B)
C)
D)
Datasets
Achromobacter dataset
E.coli dataset
Klebsiella long-read dataset
Klebsiella short-read dataset
Mixed phages dataset
Salmonella dataset
DNA, RNA and nucleotide metabolism
lysis
moron, auxiliary metabolic gene and host takeover
transcription regulation
connector
other
integration and excision
tail
head and packaging
PHROG function categories
0.0
0.5
1.0
Proportion of genes annotated
Integrase
Transposase
Recombinase
Toxin genes
AMR genes
Virulence factor genes
anti-CRISPR spacers
Defense genes
Search for specific markers
Phage therapy candidate?
Gene proportion
0.0
0.15
0.30
0.45
0.60
Annotation type
Pharokka annotated
Phold annotated
Hypothetical genes
Gene presence absence
Present
Absent
Potential phage therapy candidate
Yes
No

Figure 3. 3: Overview of phage genome characteristics across datasets. A) Proportions of genes in each PHROG function category are represented by dot sizes, with larger dots indicating higher proportions. Each row corresponds to a dataset, including *Achromobacter, E. coli, Klebsiella* (long-read and short-read), mixed phages, and *Salmonella*. B) Stacked bars display the proportion of genes annotated by three types: Pharokka annotations, Phold annotations, and hypothetical proteins, indicated by distinct fill patterns. C) Presence or absence of specific marker genes such as integrases, transposases, recombinases, toxin genes, and AMR genes is shown as filled or unfilled squares, these annotations were predicted using phage specific and specialised databases. D) The determination of phage therapy candidates is shown in the last column, where a filled square indicates a candidate, and an unfilled square indicates non-candidacy.

To identify specific genes of interest for screening these phages for potential therapeutic use, we started with the presence of integrases, which were found in 15 phages from the *Salmonella* dataset, 13 from the *Achromobacter* dataset, and 3 (*Serratia*, *Staphylococcus*, *Escherichia*) phages from the mixed phage dataset (Figure 3.3C). The presence of an integrase suggests that these phages are temperate and can persist using the lysogenic cycle. They may protect their host against other phages or express genes altering host functions. Additionally, 10 *Salmonella* phages contained recombinase genes, four phages (*Enterobacter*, 2 *Klebsiella*, and *Staphylococcus*) from the mixed phage dataset contained recombinases. In two cases (*Klebsiella* and *Escherichia* phage), we also detected genes annotated Phd-like antitoxin and Doc-like toxin, that represent a module involved in plasmid/phage maintenance system. While none of the assembled phage genomes encoded antimicrobial genes, four phages contained virulence factors. Specifically, a phage from the *Salmonella* dataset and three phages (two *Serratia* phages and an *Acinetobacter* phage) from the mixed phage dataset were identified as encoding immune-modulating virulence genes. While the specific functions of these gene products remain unknown, their presence raises concerns and would disqualify these phages from consideration for therapeutic use. Overall, these 31 phages exhibit markers indicative of a prophage lifestyle or the presence of virulence factors, suggesting they may not be suitable candidates for phage therapy (Figure 3.3D).

Among the remaining 48 phages, 12 encoded anti-CRISPR proteins: six from *E.coli*, a *Salmonella* phage, and five from *Achromobacter* phages. An *Escherichia* phage from a mixed dataset contained defence genes (Figure 3.3C). However, 19 of the 48 potential phage therapy candidates came from samples containing multiple phages, necessitating re-isolation to ensure pure cultures. This reduces the viable candidates for phage therapy to 28 phages: 7 against *E.coli*, 19 against *Klebsiella*, two against *Achromobacter*, and one against *Pseudomonas* (Figure 3.3D). No pure candidates were identified from the *Salmonella* dataset.

3.3.4 Sphae runtime performance

Sphae was executed five times on *Klebsiella* phage Amrap across various sequencing modes and thread counts to assess differences in median runtime performance. This repetition allowed for robust comparisons, highlighting the variations in efficiency between configurations. Sphae paired-end sequencing mode took a median of 42 minutes on 8 threads but dropped significantly to a

median of 9 minutes and 43 seconds on 16 threads. In long-read mode, the workflow was completed in a median of 14 minutes on 8 threads and 7 minutes and 24 seconds on 16 threads. Additionally, when Medaka polishing was omitted during the long-read mode, the median runtime increased to 17 minutes and 9 seconds on 8 threads. However, it similarly dropped to 8 minutes and 28 seconds on 16 threads. The `sphae annotate` command runs only the annotation steps of the workflow, taking a median of 6 minutes and 13 seconds on 8 threads, compared to 6 minutes and 31 seconds on 16 threads (Table 3.4). Increasing thread count significantly reduces runtime for assembly-related tasks but does not always benefit annotation steps.

Table 3. 4: Sphae runtime performance

Dataset	**Program**	**Wall time with 8 threads (h:mm:ss)**	**Wall time with16 threads (h:mm:ss)**
SRR22972379	sphae run –sequencing longreads	Median=0:14:12 Minimum=0:13:11 Maximum=0:17:06	Median=0:07:24 Minimum=0:07:12 Maximum=0:09:47
SRR22972379	sphae run –sequencing longreads –no_medaka	Median=0:17:09 Minimum=0:15:49 Maximum=0:17:23	Median=0:08:28 Minimum=0:08:21 Maximum=0:09:15
SRR22972371	sphae run –sequencing paired	Median=0:42:46 Minimum=0:35:22 Maximum=1:25:29	Median=0:09:43 Minimum=0:09:33 Maximum=0:09:51
OQ579031	sphae annotate	Median=0:06:13 Minimum=0:06:12 Maximum=0:06:13	Median=0:06:31 Minimum=0:06:24 Maximum=0:06:40

3.4 Discussion

Sphae is a reproducible workflow that automates the fundamental bioinformatics steps used in phage therapy to identify candidates for therapeutic use. By integrating 12 bioinformatics tools and nine Python scripts into a unified workflow, Sphae enables seamless execution using a single command. This workflow addresses key challenges in phage therapy by detecting induced prophages, multiple phage species in a sample, and DTRs that could influence HGT. Leveraging Snakemake's parallelisation capabilities, Sphae can process multiple phages simultaneously, often within 10 minutes on 16 threads per phage sample. This makes Sphae a user-friendly solution for clinical applications, allowing for the rapid detection of phages with therapeutic potential.

We analysed five datasets, including 65 samples, to benchmark Sphae. These datasets included both short-read and long-read sequencing data, which were used to assemble 84 phage genomes, of which 28 phages could be utilised for therapy (Figure 3.4). We found that phage samples can contain multiple phages, and Sphae reports the characteristics of these phages, making it easier to identify potential candidates for phage therapy that could be further purified if a therapeutic phage candidate is identified. In some instances, contaminants such as *E. coli ϕX174 in Illumina sequencing or phage λ in Nanopore sequencing were detected,* as they are used as sequencing controls. In other cases, induced prophages may be present, identifiable by the presence of the same or highly similar sequences across all samples, as demonstrated in the *Achromobacter* dataset in this study. Finally, in cases where the phage fails, Sphae reports at which step the sample failed, if it was during assembly or if the assembly was fragmented, as demonstrated with the *E.coli* dataset. These findings underscore the importance of thorough characterisation and identification of phages for their potential therapeutic use.

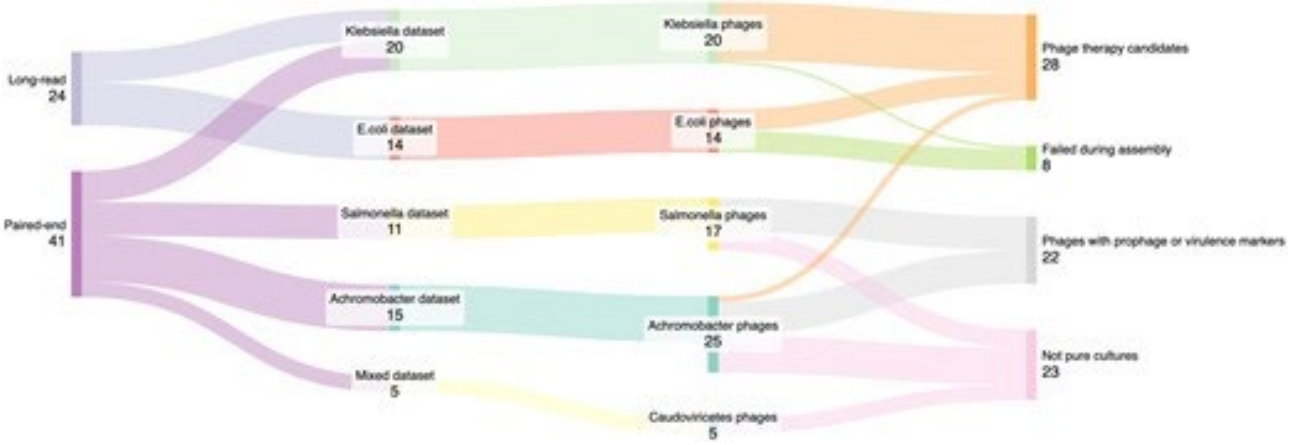


Figure 3. 4: Flowchart summarising the analysis of 65 phage samples across five datasets, detailing the number of assembled phages, therapeutic candidates, failed assemblies, impure cultures, and phages containing prophage or virulence gene markers. Diagram generated using SankeyMATIC.

3.4.1 Sphae analysis reveals genomic insights into phage biology

Phage isolation is challenging because a plaque can contain multiple phages from the environment, induced prophages, or other contaminants within a single sample. Bacterial isolates frequently contain prophages, and it has been reported that the average prophage density is 2.4%(Inglis et al., 2024; McKerral et al., 2023). The excised prophage can contaminate the therapeutic phage lysate, increasing the risk of HGT, including unwanted AMR and virulence genes(Botelho et al., 2023; Pfeifer et al., 2022). Here, we demonstrate that Sphae effectively captures prophage contamination cases and informs the user when the sample may require further purification, as shown with the *Achromobacter* dataset, allowing for the detection and exclusion of phages that could be therapeutically problematic.

In some instances, assemblies failed or produced fragmented contigs. These outcomes are more likely attributable to upstream factors such as low DNA yield, degraded input material, or phage-

specific DNA modifications that interfere with sequencing and library preparation, rather than limitations of Sphae itself. Such technical and biological challenges are well recognised in phage genomics and highlight the importance of careful sample preparation to maximise assembly success.

Sphae not only assembles and annotates phage genomes from various bacterial hosts but also identifies integrases, and recombinases—key enzymes involved in the integration and recombination of phage and bacterial DNA(S. R. Grigson et al., 2023; Turner et al., 2021). These enzymes are central to HGT, particularly in facilitating the movement of genes between phages and hosts, which has implications for phage therapy. In the 84 phages analysed, integrases were detected in 17 phages, recombinases in 14 phages (Figure 3.3C). While these three genes are associated with the temperate lifecycle, recombinases are also part of recombination systems within lytic phages, helping with DNA repair and enabling the formation of concatemers during genome packaging. Therefore, the presence of recombinase genes is not a clear indication of a temperate lifecycle; further investigation is required.

Another critical aspect of phage biology is phage genome packaging. Phage packaging mechanisms, such as cos and pac packaging, can influence the likelihood of HGT events(Borodovich et al., 2022; Catalano & Morais, 2021). For instance, cos site phages are less likely to carry out generalised transduction, while pac site or headful packaging wherein the bacterial DNA is mistakenly packaged into the phage capsids, facilitating gene transfer between the bacteria(Borodovich et al., 2022). Sphae addresses this by identifying the low-complexity DTR in genomes, typically associated with headful packaging, providing insights into the packaging processes. Sphae detected DTRs in 57 of the 84 phages. However, DTRs were detected in 83.82% of Illumina-sequenced phage genomes, while none were detected in the Nanopore assemblies. The *Klebsiella* dataset included 10 phages on both platforms, and DTRs were detected only on Illumina sequences, as noted in the original publication(Elek et al., 2023). This finding highlights two key points:(1) low-complexity regions such as DTRs are more reliable in short-read sequencing data, and (2) sequencing platforms influence the detection of packaging signals and completeness of the assembly. However, current bioinformatic tools cannot easily differentiate between the different packaging mechanisms or detect the correct copy number of repeats, as this influences completeness, which also depends on the type of phage it is(Borodovich et al., 2022).

These mechanisms are relevant for determining whether AMR genes and virulence factors in the phage can be transferred to bacterial hosts or introduced into the bacterial population. Sphae, therefore, also searches for AMR genes and virulence factors. In the datasets tested, none of the phages encoded AMR, but four genomes included virulence factors. Overall, the identification of these genes and their reporting in the summary file aim to enhance the effectiveness of detecting

phage therapy candidates. As more phages are sequenced, Sphae could serve as a valuable tool not only for identifying therapy candidates but also for advancing studies on phage evolution and host interaction dynamics.

3.4.2 Sphae follows FAIR principles

This workflow promotes adherence to the Findable, Accessible, Interoperable, Reusable, and Reproducible (FAIR) principles(Wilkinson et al., 2016). While developing this workflow, we addressed several challenges commonly associated with such workflows. This included creating comprehensive documentation with test datasets and structured output, making it easier to navigate and interpret results. While we provide the users with only pertinent outputs in the "RESULTS" directory, the intermediate files are retained so researchers can adapt their approach to resolve assembly complexities.

In instances of assembly failures, Sphae retains intermediate files that outline the steps where the breakdown occurred. For example, poor assemblies resulting from insufficient genome coverage can prompt more sequencing of the sample, if feasible. Additional adjustments, such as altering the subsampled reads or switching to alternative assemblers, could also be considered. Alternative assembler options include SPAdes(Bankevich et al., 2012), which handles a full spectrum of k-mers; Canu(Koren et al., 2017), which utilises Overlap-Layout-Consensus assemblers, or hybrid assemblies with tools like Unicycler(Wick et al., 2017) or Plassembler(Bouras, Sheppard, et al., 2023), which may be necessary to resolve assembly complexities. Cases of fragmented assemblies connected in an assembly graph can be resolved using Phables. This ensures that even when complete assemblies are not immediately achievable, researchers can refine their approach to resolve assembly complexities, especially in time-sensitive cases.

Sphae workflow also tracks the versions of the software tools used, enhancing reproducibility. We also emphasise the pre-processing steps to ensure standard execution and minimise human error while providing users with readable errors. The challenges and solutions are presented in Table 3.5.

Table 3. 5: Challenges and solutions in workflow development.

Challenges	**Solution**
Variability in tools and programming languages	Snakemake workflow manager allows the integration of tools written in multiple languages.
Lack of version, parameters documentation, and installation of multiple programs	Snakemake allows logging each step, keeping track of the tool version and the command run with the default parameters listed. Each software is automatically downloaded to its separate conda environment with dependencies or via a pre-built Docker/Singularity container.

Portability of the workflow	The workflow is available through conda, pip, pre-built containers, and source installation in GitHub or via a pre-built Docker/Singularity container.
Hardware and software dependencies	The workflow's configuration file includes resource information that the user can update for the system on which the workflow is running. Additionally, a pipeline can be configured to interact with job schedulers on high-performance computing (HPC) systems.
Error handling	Provide detailed logs with information identifying the step at which the error occurred for each rule and an overall Snakemake .log file.
Addition of new tools	New tools can be quickly added as a new rule to the workflow. This critical feature allows new and improved tools to be integrated as they are developed.

3.4.3 Sphae is a modular workflow solution

The tools were chosen based on best practices in phage genome characterisation(S. R. Grigson et al., 2023; Shen & Millard, 2021; Turner et al., 2021). The focus was on achieving high accuracy and benchmarking for low runtime results. Workflow managers offer the advantage of isolating each software in its own environment(Köster & Rahmann, 2012; M. J. Roach et al., 2022). This means that as the software is improved or new tools are published, they can be quickly added and replace outdated modules. Additionally, more samples can be added to each dataset, and the workflow will run only the new samples, with previously used tool versions if the conda environments were kept. The complete workflow, along with the individual modules, supports re-entrancy, allowing steps to be resumed in case they were interrupted.

In Sphae, we have added the option, `sphae run`, to run the entire workflow beginning with sequencing reads to generate final annotations and a summary report. However, the `sphae annotate` module has been included to allow end-users to run only the annotation steps on pre-assembled phage genomes, leveraging Sphae's approach to improving the number of annotated genes. This module was added for two reasons: first, the assembled genomes can be re-circularised to start from the large terminase subunit (*terL*) or other user-selected genes using tools like Dnaapler(Bouras, Grigson, et al., 2024) and visualised with Clinker(Gilchrist & Chooi, 2021), pyGenomeViz and Lovis4u(Egorov & Atkinson, 2025). Second, phages sometimes reassign stop codons by using alternative genetic codes(Borges et al., 2021; Cook, Telatin, et al., 2023; Peters et al., 2022; Pfennig et al., 2023) end-users can change the config file to run pyrodigal-gv(Larralde, 2022) for gene prediction in Pharokka instead of the default PHANOTATE(McNair et al., 2019). The need for changing tools can be predicted from the coding density reported in the summary.txt file. Phages generally have high coding density to minimise non-coding regions; low-density coding

regions suggest that the annotation tools may have incompletely annotated the phage genome(McNair et al., 2019).

3.4.4 Ongoing Development and Version Updates

The Sphae toolkit is actively maintained to ensure continued relevance and compatibility with evolving bioinformatic methods. Version 1.4.3 was released alongside the original publication of this work. Since then, the toolkit has undergone several updates, including adjustments to underlying tool versions, improvements to assembly and annotation modules, and the integration of a phylogeny module to facilitate comparative genomic analyses. The most recent release, Sphae v1.5.3, incorporates all changes described in this thesis.

To maintain transparency and reproducibility, a complete record of updates is provided through the *Changes.md* file available in the public GitHub repository (https://github.com/linsalrob/sphae). This changelog documents incremental improvements such as bug fixes, parameter refinements, and the incorporation of new functionalities.

The inclusion of this section highlights that Sphae is not a static toolkit but a continuously evolving resource. By maintaining an open versioning system and linking development milestones to a publicly accessible repository, this work ensures that users can both reproduce the analyses presented here and benefit from ongoing enhancements.

3.4.5 Future improvements

The ongoing isolation and analysis of phages continue to enhance our grasp of phage biology, evolution and phage-host interactions. Although short-read platforms have traditionally been used for sequencing most phages, there is a growing adoption of long-read sequencing methods such as Oxford Nanopore and PacBio sequencing. An advantage of long-read sequencing is its ability to detect phage DNA modifications, like methylation(Simpson et al., 2017; Sun et al., 2023), which may play a role in phage resistance and adaptability to microbial communities. While over 2,000 phage sequences are available in the SRA from Illumina platforms, fewer than 300 phages have been sequenced using long-read technologies, such as PacBio and Nanopore platforms (source: https://www.ncbi.nlm.nih.gov/sra). With the increasing availability of long-read sequencing data and the development of automated tools for identifying methylation in phage genomes with minimal manual intervention, we anticipate the integration of this feature into the workflow as a distinct module. Additionally, alternate codon reassignment, recently identified in phage genomes (Borges et al., 2021; Larralde, 2022; Peters et al., 2022), is now included in Sphae, offering users insights into unique coding adaptations and insights into coding adaptations relevant to host specificity. Tools like Prfect that predict programmed ribosomal frameshifts producing longer proteins(McNair et al., 2023) also present an exciting future integration. These enhancements will enable end-users to explore this specialised genome feature, as our understanding of phage biology and evolution improves. The tools and modules within Sphae will be regularly updated to accommodate these

advancements, including useful summary reports, ensuring users can easily access and interpret the latest developments in a user-friendly manner.

3.5 Conclusions

This chapter addresses the need for an end-to-end phage toolkit by presenting the development and implementation of Sphae—a dedicated bioinformatics workflow designed to rapidly assemble, annotate, and characterise phage genomes from sequencing data. The workflow is modular and user-friendly, leveraging workflow managers and explicit database configurations to ensure analyses are reproducible, scalable, and easily adaptable. This lowers the entry barrier to powerful tools and promotes wider adoption.

Building on the knowledge gaps identified in the previous chapter, this work tackles a key technical limitation in phage research by enabling more accurate and accessible characterisation of their roles in microbial systems. Understanding how phages function as both predators and vectors of gene transfer is fundamental to deciphering the forces that shape microbial adaptability and ecosystem stability. Sphae lays the groundwork for future investigations into phage–host interactions that influence host fitness and broader community dynamics. In doing so, this chapter provides essential infrastructure to support the systematic exploration of phages within complex microbial ecosystems.

3.6 Supplementary Files

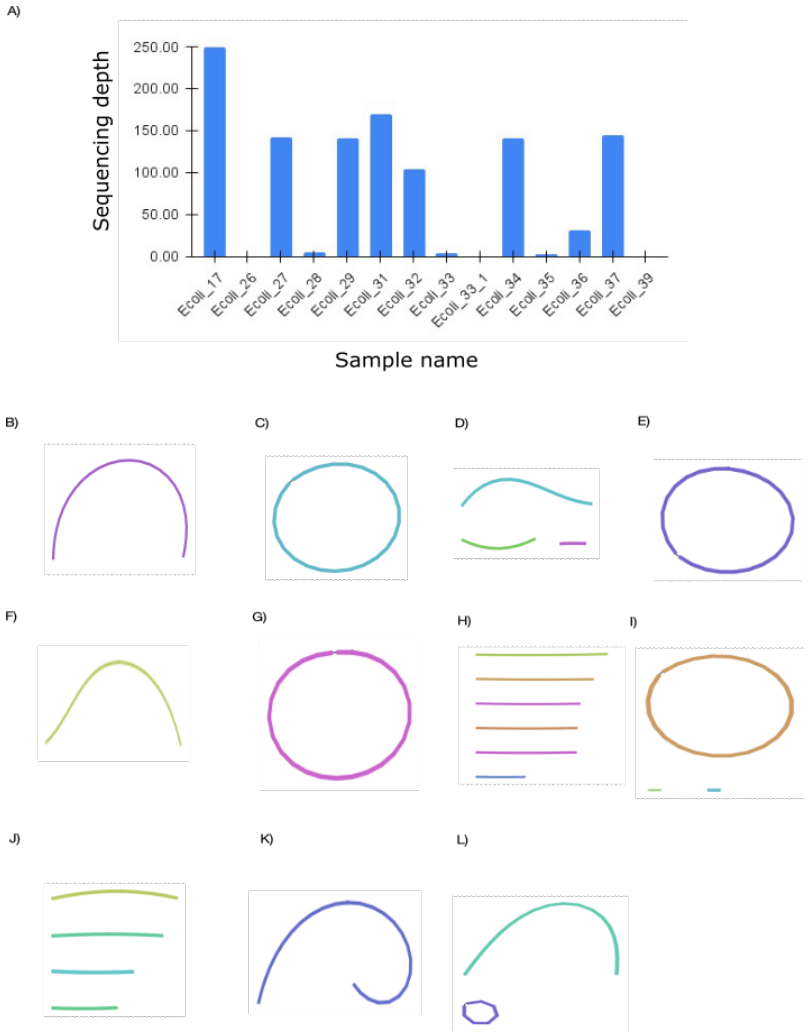


Fig S3. 1: A) Sequencing depth evaluation of *E. coli* datasets. Samples with high sequencing depth (E.coli_17, E.coli_27, E.coli_29, E.coli_31, E.coli_32, E.coli_34, E.coli_36, and E.coli_37) successfully assembled into complete phage genomes. In contrast, samples with low sequencing depth (E.coli_26, E.coli_28, E.coli_33, E.coli_33_1, E.coli_35, and E.coli_39) produced either no contigs or fragmented contigs during assembly. B-L) Bandage plots of 10 *E. coli* phages, showing assembly results for B) E.coli_17, C) E.coli_27, D) E.coli_28, E) E.coli_29, F) E.coli_31, G) E.coli_32, H) E.coli_33 (fragmented), I) E.coli_34, J) E.coli_35 (fragmented), K) E.coli_36, L) E.coli_37 (fragmented). Three samples, E.coli_33_1, E.coli_39, and E.coli_26, failed to assemble.

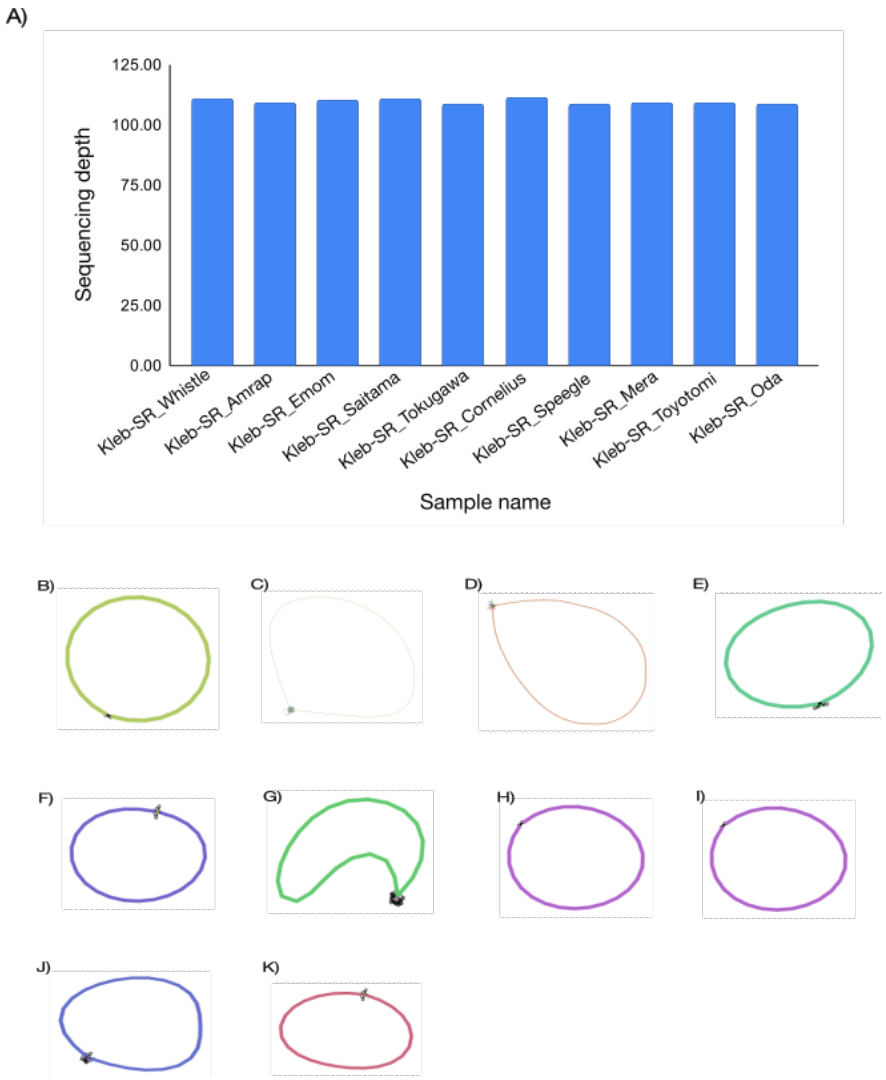


Fig S3. 2: A) Sequencing depth evaluation of *Klebsiella* short-read datasets. B-L) Bandage plot of the 10 phages; each included only one phage per sample, B) Kleb-SR_Whistle, C) Kleb-SR_Amrap, D) Kleb-SR_Emom, E) Kleb-SR_Saitama, F) Kleb-SR_Tokugawa, G) Kleb-SR_Cornelius, H) Kleb-SR_Speegle, I) Kleb-SR_Mera, J) Kleb-SR_Toyotomi, K) Kleb-SR_Oda. The width of the lines in the bandage plots is random and does not reflect genome lengths.

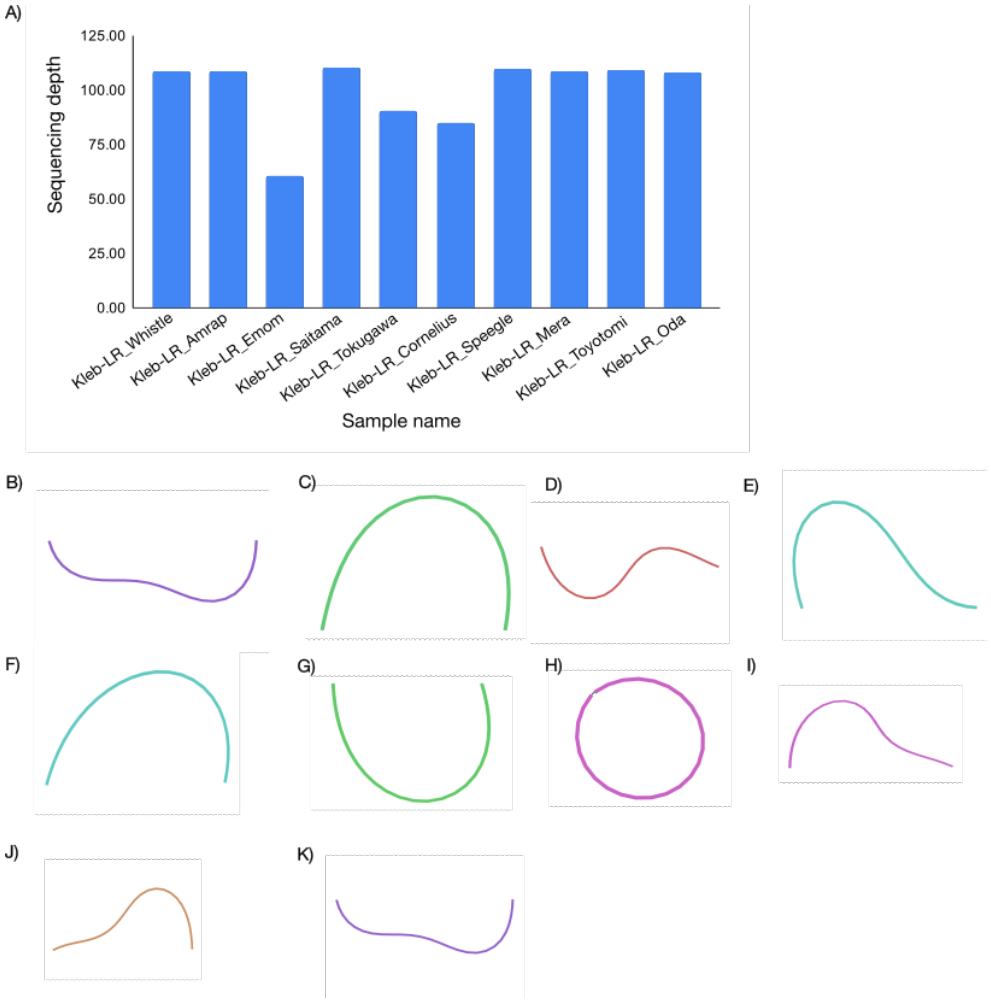


Fig S3. 3: A) Sequencing depth evaluation of *Klebsiella* long-read datasets. B-L) Bandage plot of the 10 phages; each included only one phage per sample, B) Kleb-SR_Whistle, C) Kleb-SR_Amrap, D) Kleb-SR_Emom, E) Kleb-SR_Saitama, F) Kleb-SR_Tokugawa, G) Kleb-SR_Cornelius, H) Kleb-SR_Speegle, I) Kleb-SR_Mera, J) Kleb-SR_Toyotomi, K) Kleb-SR_Oda.

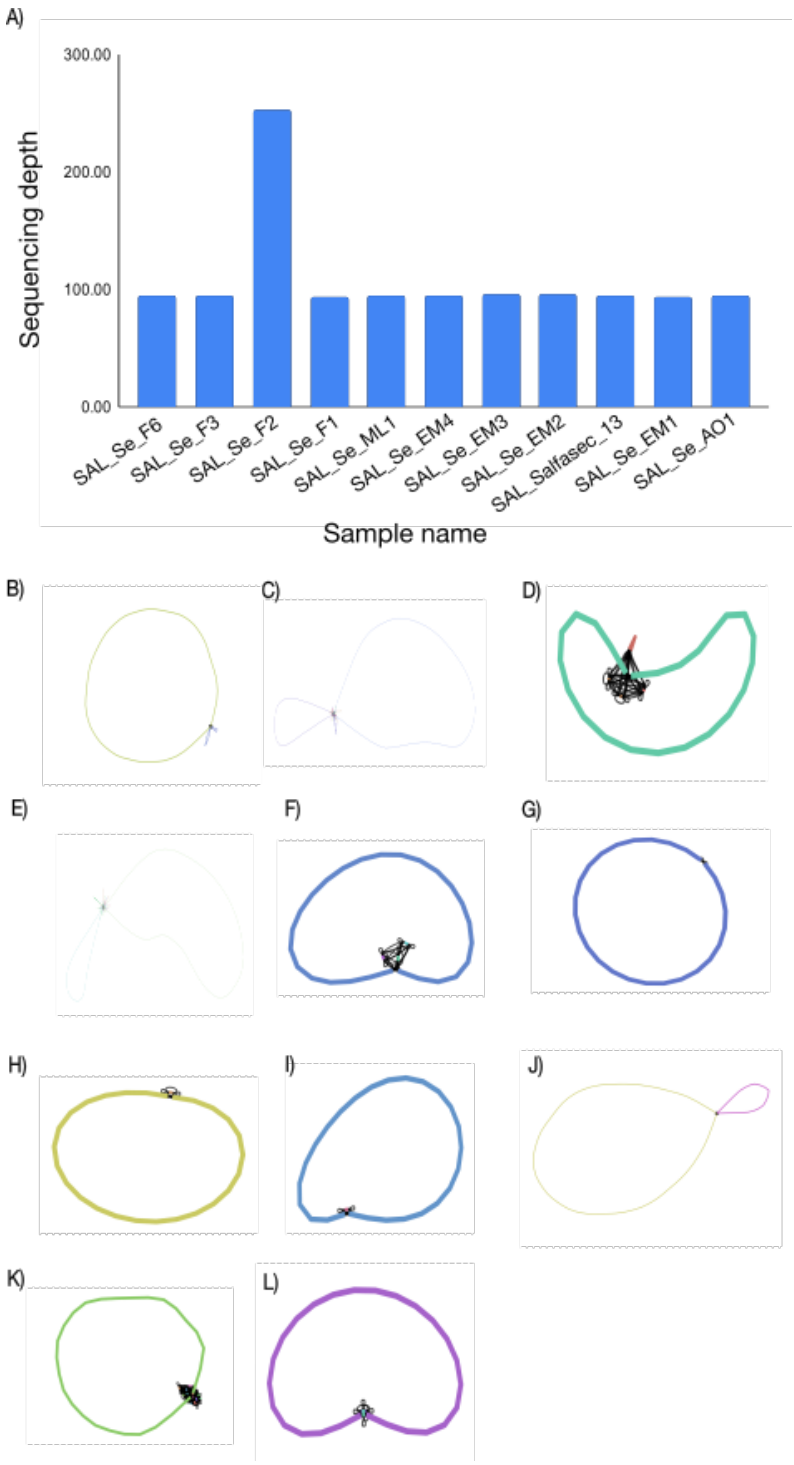


Fig S3. 4: A) Sequencing depth evaluation of *Salmonella* short-read datasets. B-L) Bandage plot of the 11 *Salmonella* phages with most samples including a single phage, except two samples, B) SAL_Se_F6 (two phages), C) SAL_Se_F3 (three phage), D) SAL_Se_F2, E) SAL_Se_F1 (three phages), F) SAL_Se_ML1, G) SAL_Se_EM4, H) SAL_Se_EM3, I) SAL_Se_EM2, J) SAL_Salfasec_13 (two phages), K) SAL_Se_EM1, L) SAL_Se_AO1. The width of the lines in the bandage plots is random and does not reflect genome lengths.

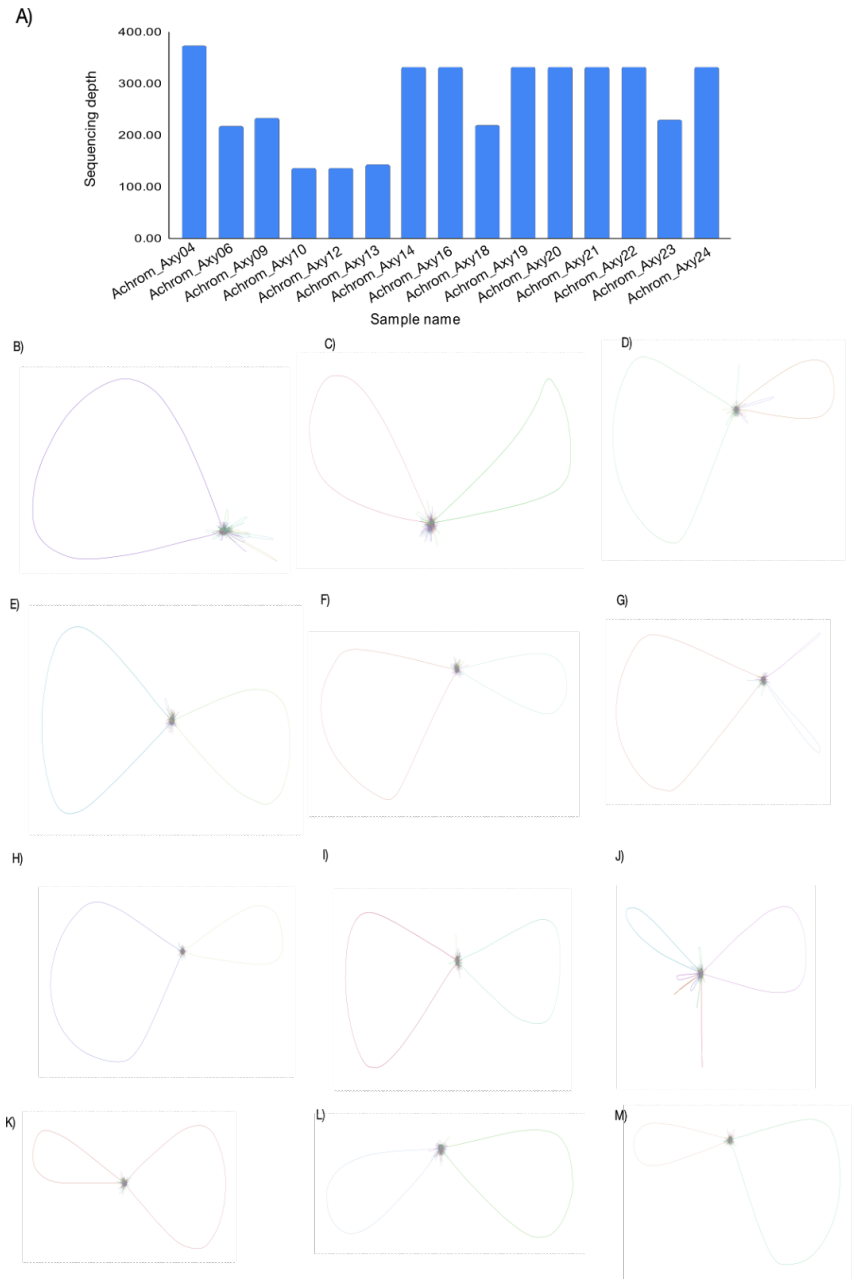


Fig S3. 5. A) Sequencing depth evaluation of 15 *Achromobacter* short-read datasets. B-M) Bandage plots of 12 of the 15 assembled *Achromobacter* phages: B) Achrom_Axy06 (one phage), C) Achrom_Axy09 (two phages), D) Achrom_Axy24 (two phages), E) Achrom_Axy23 (two phages), F) Achrom_Axy10 (two phages), G) Achrom_Axy12 (one phage), H) Achrom_Axy13 (two phages), I) Achrom_Axy21 (two phages), J) Achrom_Axy16 (one phage), K) Achrom_Axy19 (two phages), L) Achrom_Axy18 (two phages), and M) Achrom_Axy22 (two phages). Three samples are not shown, as their bandage plots were too large for display. Line widths in the bandage plots are arbitrarily scaled and do not represent actual genome lengths.

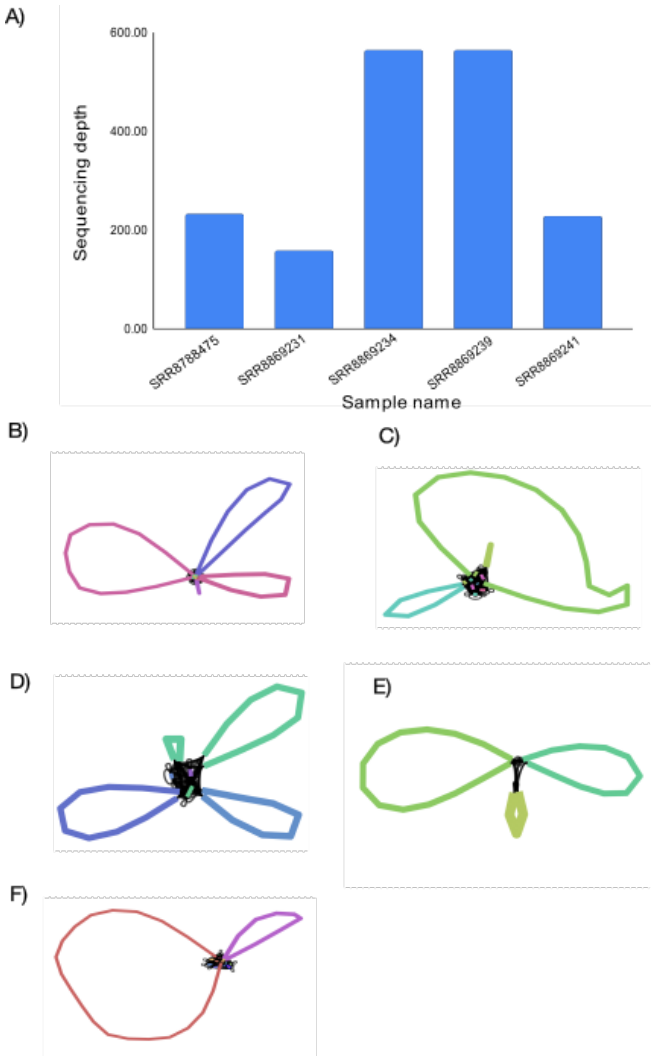


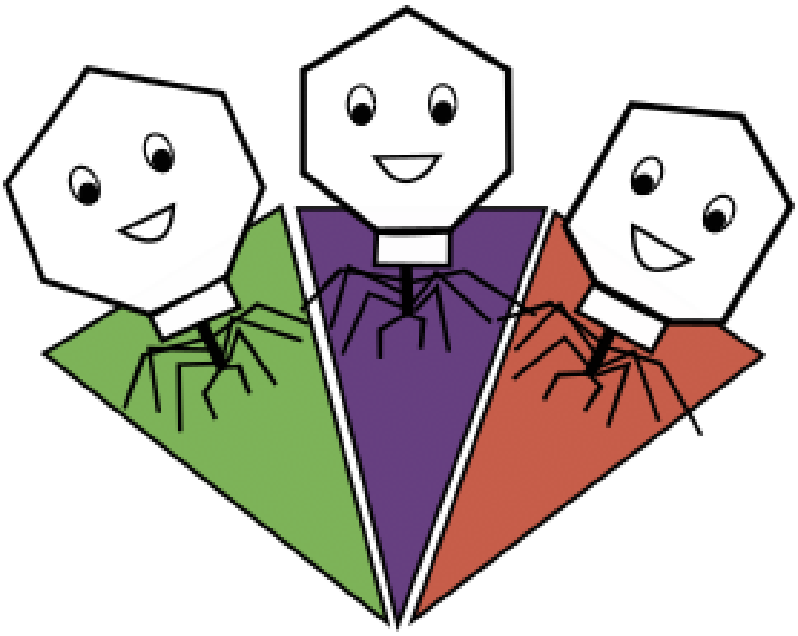
Fig S3. 6: A) Sequencing depth evaluation of the five mixed dataset phages. B-F) Bandage plots, B) SRR8788475 includes four phages, C) SRR8869231 includes two, D) SRR8869234 includes three phages, E) SRR8869239 includes three phages, F) SRR8869241 includes three phages.

Supplementary Tables available here, 10.5281/zenodo.17254792

Table S3 1: Overview of the phages analysed from different studies in this chapter

Table S3. 2: Sphae results for the 65 phages

CHAPTER 4
PHAGE-BACTERIA INTERACTIONS



This chapter is based on the published article— **Papudeshi, B.**, Vega, A. A., Souza, C., Giles, S. K., Mallawaarachchi, V., Roach, M. J., ... & Edwards, R. A. (2023). Host interactions of novel Crassvirales species belonging to multiple families infecting bacterial host, *Bacteroides cellulosilyticus* WH2. Microbial Genomics, 9(9), 001100. https://doi.org/10.1099/mgen.0.001100. This article is reproduced in full under the terms of the Creative Commons Attribution License (CC BY 4.0). © The Author(s) 2023. Published by Microbiology Society.

Statement on the Use of Generative Artificial Intelligence (AI): Generative AI tools, specifically Grammarly was used during the preparation of this chapter for language editing purposes, such as improving sentence clarity, grammar, and structure.

Preface

This chapter is based on my published research, which investigated novel *Crassvirales* phages infecting *Bacteroides cellulosilyticus* WH2, a project I led as part of this thesis. The study was motivated by the ecological significance of *Crassvirales*, which are among the most abundant viruses in the human gut. They primarily infect *Bacteroides*, a key genus within the gut microbiome. Despite their prevalence, only four *Crassvirales* isolates have been successfully cultured to date, limiting our understanding of their biology. In this work, I addressed this gap by isolating and characterising 14 new *Crassvirales* phages, substantially expanding the known diversity of this important viral group and providing one of the first detailed genomic comparisons of multiple *Crassvirales* species infecting the same bacterial host. As first author on this publication, I was responsible for designing and performing the genomic analyses, interpreting the findings, and drafting the manuscript. For this thesis, I have made minor adaptations, including relocating a few supplementary files to the main manuscript to enhance the interpretation of the data, and adding a conclusion section to this chapter to align this work with the broader research aims of understanding phage–host interactions.

Since publication, the paper has been cited in 11 other peer-reviewed articles investigating gut phage ecology, virus–host co-evolution, and the annotation of uncultured viral genomes, indicating its growing relevance in both applied and theoretical virology. In addition to its academic impact, this work gained media attention, highlighting its significance in advancing our understanding of phage–host interactions and gut viral diversity. This broader dissemination underscores the relevance of the work to both scientific and public audiences.

Statement of authorship

As the primary author of this chapter, I led the research to completion. Throughout the chapter, I actively sought input from specialists in the field, ensuring the robustness of the research while maintaining primary authorship over this work. My role extended beyond analysis, as I also led the manuscript writing, figure preparation, and critical discussion of results. This included integrating feedback from co-authors, ensuring clarity in communicating complex genomic findings, and positioning the study within the broader scientific discourse. While this project involved many collaborators, this was primarily due to the interdisciplinary nature of the work, which required expertise in phage isolation, comparative genomics, structural modelling, and evolutionary analysis. I coordinated these efforts to strengthen the study's impact and ensure a rigorous interpretation of the results.

Below is a breakdown of the author's contributions:

Author	**Contribution**
Bhavya Papudeshi	Performed data analysis, writing, and editing the manuscript

Alejandro A. Vega	Research design, data collection and sequencing
Cole Souza	Data collection and sequencing
Sarah K. Giles	Sequencing, microscopy, editing of the manuscript
Vijini Mallawaarachchi	Data analysis
Michael J. Roach	Data analysis
Michelle An	Data collection
Nicole Jacobson	Data collection
Katelyn McNair	Data analysis
Maria Fernanda Mora	Sequencing
Karina Pastrana	Sequencing
Lance Boling	Research design and data collection
Christopher Leigh	Microscopy
Clarice Harker	Data collection and sequencing
Will S. Plewa	Data collection and sequencing
Susanna R. Grigson	Data analysis
George Bouras	Data analysis
Przemyslaw Decewicz	Data analysis
Antoni Luque	Data analysis
Lindsay Droit	Sequencing
Scott A. Handley	Research design, data analysis and editing of the manuscript
David Wang	Research design, data analysis and editing of the manuscript
Anca M. Segall	Research design, data analysis and editing of the manuscript
Elizabeth A. Dinsdale	Research design, data analysis and editing of the manuscript

Robert A. Edwards	Research design, data analysis and editing of the manuscript

The contributions of each co-author have been explicitly stated, and their permission to include these works has been obtained as per Flinders University's Authorship of Research Output Procedures (Appendix A). I also affirm that, while this research was strengthened through these collaborations, all overarching hypotheses, research aims, analyses, and interpretations presented in this thesis are my own. The coordination of interdisciplinary methods, the synthesis of findings across chapters, and the articulation of their broader significance reflect my independent intellectual contributions.

Host interactions of novel *Crassvirales* species belonging to multiple families infecting the bacterial host, *Bacteroides cellulosilyticus* WH2

Abstract

Bacteroides, the prominent bacteria in the human gut, play a crucial role in degrading complex polysaccharides. Their abundance is influenced by phages belonging to the *Crassvirales* order. Despite identifying over 600 *Crassvirales* genomes computationally, only a few have been successfully isolated. Isolating additional *Crassvirales* phages in culture can provide insights into phage-host evolution and infection mechanisms. We focused on wastewater samples as potential sources of phages infecting various *Bacteroides* hosts. Sequencing, assembly, and characterisation of isolated phages revealed 14 complete genomes belonging to three novel *Crassvirales* species infecting *Bacteroides cellulosilyticus* WH2. These species, *Kehishuvirus* sp. 'tikkala' strain Bc01, *Kolpuevirus* sp. 'frurule' strain Bc03, and 'Rudgehvirus jaberico' strain Bc11, spanned two families and three genera, displaying a broad range of virion production. Upon testing all successfully cultured *Crassvirales* species and their respective bacterial hosts, we discovered that they do not exhibit co-evolutionary patterns with their bacterial hosts. Furthermore, we observed variations in gene similarity, with greater shared similarity observed within genera. However, despite belonging to different genera, the three novel species shared a unique structural gene that encodes the tail spike protein. When investigating the relationship between this gene and host interaction, we discovered evidence of purifying selection, indicating its functional importance. Moreover, our analysis predicts that this tail spike protein binds to the TonB-dependent receptors present on the bacterial host surface. Combining these observations, our findings provide insights into phage-host interactions and present three *Crassvirales* species as an ideal system for controlled infectivity experiments on one of the most dominant members of the human enteric virome.

4.1 Introduction

The intricate relationship between gut microbiomes and human health is characterised by the diverse microbial communities that help with digestion, regulate the immune system, and alter brain function (Hou et al., 2022; Integrative HMP (iHMP) Research Network Consortium, 2019; Shamash & Maurice, 2022). Metagenomics, a culture-independent technique used to capture microbial diversity in a sample (Hugenholtz et al., 1998; Pace et al., 1986), has transformed our understanding of bacteria and the corresponding bacteriophages in the environment (Anthenelli et al., 2020; Hesse et al., 2022; Inglis & Edwards, 2022; M. Roach et al., 2022). These metagenomic datasets have revealed a correlation between bacterial and bacteriophage populations, which suggests bacteriophages play a role in modulating bacterial populations (Chevallereau et al., 2022; Knowles et al., 2016). In particular, the human gut microbiome exhibits varying bacterial densities, including a high abundance of Bacteroidota (formerly Bacteroidetes) (HMP Consortium, 2012; Pargin et al., 2022; Qin et al., 2010), whose populations are thought to be influenced by phages, with *Crassvirales* being a particularly abundant group in the gut virome. These dsDNA bacteriophages have a podovirus-like morphology, genomes ranging between 100 and 200 kb, and conserved gene order (Edwards et al., 2019; Rossi et al., 2020; Yutin et al., 2021). They are widespread,

constituting a stable component of an individual's microbiome, and do not appear to be associated with human health or disease states (Edwards et al., 2019; Norman et al., 2015).

The first phage within *Crassvirales* order was computationally discovered by cross-assembly of DNA sequence reads from human gut microbiome samples(Dutilh, Cassman, et al., 2014). Since nearly 600 *Crassvirales* genomes have been identified computationally, leading to the International Committee of Taxonomy of Viruses (ICTV) formally classifying the *Crassvirales* order into four families, ten subfamilies, 42 new genera, and 72 new species (Rossi et al., 2020; Shkoporov, Stockdale, et al., 2021). The classification relied on phylogenetic analysis of conserved structural genes, including the major capsid protein (MCP), terminase large subunit (*terL*), and portal protein (portal). Additionally, the average nucleotide identity (ANI) species cutoff was set to 95% identity over 85% genome coverage.

The identification of numerous *Crassvirales* genomes has advanced our understanding of this viral order. Similar to other phages, these genomes contain three discernible regions encoding for 1) structural proteins involved in producing the capsid and tail genes, 2) transcription proteins and 3) replication proteins crucial for successful phage replication in different infection stages(Yutin et al., 2021). Gene homology analysis showed that many of the genes are highly variable when compared with other genomes from this order. Comparative genomics further displayed the unique biological characteristics of *Crassvirales* species(Dutilh et al., 2021; Walker et al., 2022), including switching DNA polymerases, alternative coding strategies(Borges et al., 2021; Crisci et al., 2021; Ivanova et al., 2014; Peters et al., 2022), and the variable intron density across lineages(Peters et al., 2022; Yutin et al., 2021). Overall, the functional annotation of Crassvirales genomes remains challenging, with many genes annotated as hypothetical proteins lacking a known biological function and exhibiting little to no similarity to sequences in reference databases. These challenges can be addressed through experimental approaches that can help elucidate the functions of these uncharacterised genes or proteins.

The first step in experimental approaches is phage isolation, which requires knowledge of their host species and the ability to culture them. This has led to only four successful isolates obtained so far, including *Kehishuvirus primarius* (crAss001) infecting *Bacteroides intestinalis* APC919/174(Shkoporov et al., 2018), *Wulfhauvirus bangladeshii* DAC15 and DAC17 from wastewater effluent infecting *Bacteroides thetaiotaomicron* VPI-5482(Hryckowian et al., 2020), and *Jahgtovirus secundus* (crAss002) infecting *Bacteroides xylanisolvens* APCS1/XY(Guerin et al., 2021). All these isolates exhibited host specialist morphotypes that can be maintained in continuous host culture, but none possess lysogeny-related genes(Shkoporov et al., 2018; Shkoporov, Khokhlova, et al., 2021). The proposed mechanism of persistence involves the bacterial host cycling between sensitive and resistant states by altering the genes encoding surface transporters and capsular polysaccharide structures (CPS) on the bacterial surface(N. T. Porter et al., 2020; Shkoporov, Khokhlova, et al., 2021). Further to improve the annotations, cryogenic-electron microscopy of *K. primarius* provided functional assignments to the virion proteins and an insight into the infection mechanism, revealing how the capsid and tail store cargo proteins aid in initial host

infection(Bayfield et al., 2023). Continued efforts to isolate more *Crassvirales* genomes can provide insights into phage-host evolution, comprehensive protein annotation, and elucidation of infection mechanisms.

Here, we present the successful isolation of 14 Crassvirales isolates from wastewater, which include three novel species belonging to the families *Steigviridae* and *Intestiviridae.* Remarkably, all these isolates infect the same host, *Bacteroides cellulosilyticus* WH2. We investigate the genes playing a role in host interaction, providing insights into the evolution of these dominant phages and how their interactions shape the gut microbiome.

4.2 Methods

4.2.1 Phage sampling

Untreated sewage water (influent) was collected from a waste treatment plant in Cardiff, CA in 1L Nalgene bottles. An aliquot of 30 mL influent was centrifuged at 5,000 × g for 5 min to pellet the debris. The supernatant was decanted and passed through a 0.22 µm pore size Sterivex filter. The filtrate was used as a phage source and stored between 2 to 8 ℃.

4.2.2 Host bacteria cultivation

Bacterial species, *B. cellulosilyticus* WH2 received as glycerol stocks from Washington University, St. Louis, *B. fragilis* NCTC 9343 (ATCC 25285), *B. stercoris* CC31F (ATCC 43183), and *B. uniformis* ATCC 8492 were received as glycerol stocks from BEI resources were used as bacterial hosts. All the bacteria were grown in brain-heart infusion media supplemented with 2 mM $MgSO_4$, and 10 mM $MgCl_2$ we denote as BHISMg. Culture plates were supplemented with 1.5 % w/v agar and incubated at 37 ℃ for 48 hrs under anaerobic conditions with 5 % H_2, 5 % CO_2, and 90 % N_2. Following incubation, an isolated colony was transferred into a 12 hrs deoxygenated BHISMg broth. Following anaerobic incubation at 37 ℃ for 24 hrs the liquid cultures were further sub-cultured into another BHISMg broth and incubated overnight.

4.2.3 Plaque assays

BHISMg plates were deoxygenated for 12 hrs in the anaerobic chamber and pre-warmed before use. For top agar plates were prepared by adding cooled molten BHISMg with 0.7 % w/v agar was inoculated with 500 µl of bacteria, and between 2 µl and 50 µl of processed phage influent. The plates were cooled for 15 min before incubating at 37 °C for up to five days. Plates were assessed daily for the development of plaques.

4.2.4 Lysate preparation

Plaque from each plate was inoculated into 200 µl of SM buffer and homogenised to diffuse the phage from the agar to the buffer. A 200 µl aliquot of the phage was added to *B. cellulosilyticus* WH2 bacteria in the log-growth phase and grown at 37 °C anaerobically, overnight. The tubes containing the bacteria and phage were manually shaken every 30 min for the first three hours of incubation. Post incubation, tubes

were centrifuged at 4500 *x* g for 5 min, and the supernatant was collected and concentrated using a 50,000 kDa MWCO Vivaspin ultrafiltration unit (Sartorius). Phage lysate was stored at 4 °C.

4.2.5 Phage tittering enumeration

Phage titres were enumerated using the molten agar overlay method described above. A 200 µl aliquot of the lysate was diluted 10-fold in sterile SM buffer, and 10 µl was spotted onto a BHISMg plate. The plates were incubated for 24-48 hrs at 37 °C. After incubation, the plates were analysed by counting the plaques obtained to determine the titre.

4.2.6 Viral DNA extraction and sequencing

Phage DNA was extracted using a Phage DNA isolation kit (Norgen) as per the manufacturer's instructions. In short, 1 ml of phage lysate was DNase I-treated, lysed, and treated with Proteinase K. The sample was added to a spin column and washed three times. DNA was eluted in 75 µl of the elution buffer. The second elution recommended by the kit was not performed. The DNA obtained was quantified using a Qubit 1x dsDNA High-Sensitivity Assay Kit (Invitrogen, Life Technologies) according to the manufacturer's instructions. Oxford Nanopore MinION sequencing was undertaken according to the manufacturer's instructions. In short, a maximum of 400 ng of sample DNA was used for library preparation using Oxford Nanopore Rapid Barcoding Sequencing Kit (SQK-RBK0004); samples were barcoded, pooled and cleaned. The pooled samples were loaded and run on a Flowcell R9.4.1 (FLO-MIN106) following the manufacturer's instructions. The Illumina sequencing libraries were prepared by extracting the total nucleic acid (RNA and DNA) using the COBAS AmpliPrep instrument (Roche), with NEBNext library construction and sequenced on Illumina MiSeq using the paired-end 2x250 bp protocol as described in (A. H. Kim et al., 2022). The sequencing data were deposited in the Sequence Read Archive in BioProject, PRJNA737576.

For the Nanopore sequenced isolates, basecalling was performed with Guppy v6.0.1 with model dna_r9.4.1_450bps_hac. The reads were then processed with Filtlong v0.2.20 (Wick, 2018) to remove reads less than 1,000 bp in length and exclude 5% of the lowest-quality reads. Similarly, Illumina sequences were processed with prinseq++ v.0.20.4 (Cantu et al., 2019), filtering reads less than 60 bp in length, reads with quality scores less than 25, and exact duplicates.

4.2.7 Genome assembly

To assemble the genomes, a pipeline based on Snakemake using Snaketool was employed. Nanopore reads were assembled using Flye v2.9(Kolmogorov et al., 2019), while Illumina reads were assembled using MEGAHIT v1.2.9(D. Li et al., 2016). These assemblers were selected as they provide assembly graphs, which are useful for completing fragmented genome assemblies(Bruce et al., n.d.; V. G. Mallawaarachchi & Lin, 2022; V. G. Mallawaarachchi et al., 2020; V. Mallawaarachchi et al., 2020; Wick et al., 2015).

To evaluate assembly quality, the resulting contigs were processed with ViralVerify v1.1 to detect viral contigs(Raiko, 2021), read coverage was calculated using CoverM v0.6.1, and the assembly graph was

examined. The assembly graph provides information on connecting unitigs (high-quality contigs) representing the longest non-branching paths joined together to form contigs.

From each assembly, unitigs meeting specific criteria were selected as complete phage genomes. These unitigs had a length greater than 90 kb, were identified as viral, exhibited the highest read coverage and were classified as complete using CheckV v1.0.1. To ensure representation, one unitig per sample was selected as the complete phage assembly. In the end, the assemblies were polished with high-coverage Illumina reads using Polca to reduce sequencing-related errors.

Among the 41 phage genomes, 14 phages infecting *B. cellulosilyticus* were identified as belonging to the *Crassvirales* order. These genomes were approximately 90 to 120 kbp in length and aligned against known *Crassvirales* genomes. Among these phages, eight samples were sequenced on both Nanopore and Illumina sequencing platforms (Bc01 to Bc03, Bc05 to Bc08), while four were sequenced only on the Nanopore platform (Bc09 to Bc11), and the remaining four were sequenced only on the Illumina platform (Bc04, Bc12 to Bc14).

4.2.8 Taxonomic and functional annotation

The isolates in this study were processed with CrassUS (https://github.com/dcarrillox/CrassUS), a specialised tool for annotating *Crassvirales* genomes, providing taxonomic and functional annotation along with direct terminal repeats (DTR), and average nucleotide identities (ANI) of similar reference genomes. Taxonomic annotations from CrassUS followed ICTV criteria for *Crassvirales* order demarcation. Phylogenetic trees were constructed using conserved genes (MCP, portal, *terL*) with MAFFT v7.49(Nakamura et al., 2018) for alignment, followed by trimal v1.4.1 for trimming, and FastTree v.2.1.10(Price et al., 2010) for inference. Trees were built using the JTT model with CAT approximation and 20 rate categories, and visualised using iTol.

To compare the predicted genes and their arrangement across species, clinker plots(Gilchrist & Chooi, 2021) were used after re-circularising the genes to start at the *terL.* This allowed for the examination of synteny across genomes. Additionally, tRNA genes encoded by the phages, which evade host translation machinery, were predicted with tRNA-scanSE (Chan & Lowe, 2019).

4.2.9 Phage-host co-phylogenetic analysis

To determine if the phage co-evolves with bacterial hosts, we performed a cophylogenetic analysis using Parafit(Legendre et al., 2002) via the ape R package. The distance matrix of the trimmed multiple sequence alignment (MSA) using MAFFT v.7.520 and trimal v1.4.1 of the seven *Crassvirales* species ('K. tikkala' Bc01, 'K.frurule' Bc03, 'R.jaberico' Bc11, *K. primarius, J. secundus, W. bangladeshii* DAC15, *W. bangladeshii* DAC17) portal gene, was generated using EMBOSS distmat v6.6.0. These steps were repeated for the associated bacterial hosts, *B. cellulosilyticus, B. intestinalis, B. thetaiotomicron* and *B. xylanisolvens.* The two distance matrices were compared using Parafit with 1000 permutations, and Cailliez

eigenvalue correction. The trimmed MSA was used to generate the two phylogenetic trees using FastTree v.2.1.10 using JTT model, CAT approximation with 20 rate categories, and visualised using iTol.

4.2.10 Transmission electron microscopy imaging

Crassvirales phages were grown using the phage overlay method described above. To prepare the phage lysates, they were diluted 1:10, and 5 µL of the diluted phage lysate was applied to a plasma-cleaned grid for two minutes at room temperature. The grids used were formvar and carbon coated 200 mesh grids, and they were plasma cleaned using the Gatan (Solarus) Advanced plasma system for 30 sec prior to use. The excess phage lysate sample was wicked off with Whatman filter paper and the grid was washed with 5 µL of water. The sample was negatively stained with 5 µL of the 2 % w/v uranyl acetate for 1 minute. The excess stain was wicked off with filter paper to dry the sample on the grid. The grid was then imaged using a Tecnai G2 Spirit TEM operated at 120kV at a magnification of 49,000x and the images were recorded on an AMT Nanosprint 15 digital camera using software v7.0.1.

Phage measurements were conducted using the ImageJ software. The capsid diameter was calculated by measuring the diameter of the circle circumscribing the capsid, such that the more distant vertices of the projected capsid contacted the circle (Figure 4.1). The length of the tail was calculated from the base of the capsid to the end of the visible tail, including the collar section of the phage structure. Tail fibres or appendages were calculated (Figure 4.1). Average measurements from 5 phages were calculated and reported. The TEM image was further edited for publication using the GNU Image Manipulation Program (GIMP).

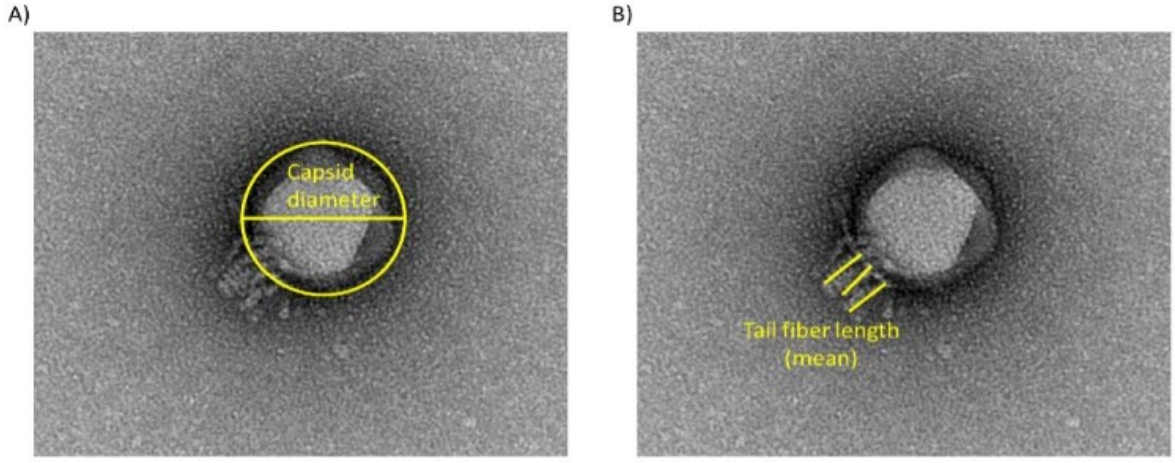


Figure 4. 1: TEM phage measurements were taken for A) Capsid diameter, by drawing a circle around the polygon with the edges within the circle. The diameter of this circle was measured and represented as the capsid diameter. B) For tail length, a line was drawn from the base of the capsid to the visible edge of the tail fibres. This was repeated over five phases of the same sample, and an average with standard deviation was calculated across all of them.

The packing genome density was predicted by correcting the measured radius from the expected capsid thickness and calculating the internal volume assuming an icosahedral model inferred from prior tailed

phage capsid studies(D. Y. Lee et al., 2022; Luque et al., 2020). The results can be reproduced using the Colab notebook available at the link, https://shorturl.at/enAKS.

4.2.11 Evolutionary analyses

The 14 *Crassvirales* isolated and assembled genomes in this study and the four reference pure culture isolates, *K. primarius*, *J. secundus, W. bangladeshii* DAC15 and DAC17 were assessed together for this analysis. Orthologous genes were identified from genes predicted from the above 18 genomes, using Orthofinder v2.5.4 default settings to determine signatures for host interactions. The default settings utilise diamond for sequence search, MAFFT for alignment, FastTree for tree inference, and the STAG species tree method. Orthogroups that included genes present only in phages from the host *B. cellulosilyticus* WH2 were examined further.

These orthogroups were aligned using Muscle codon-based multiple sequence alignment in MEGA11. To test for codon-based positive selection, we calculated the probability of rejecting the null hypothesis of strict neutrality ($d_N = d_S$) in favour of the alternative hypothesis ($d_N > d_S$). The d_N/d_S values were calculated from the MSA using MEGA v.11.0, with the Li-Wu-Luo method. The variance of the difference was computed using bootstraps, set to 100 replicates. As this analysis can be misleading in the presence of recombination breakpoints, orthogroups were run through Genetic Algorithm for Recombination Detection (GARD) analysis (Kosakovsky Pond et al., 2006), with default settings. This method utilises a combination of phylogenetic and statistical approaches to detect recombination signals.

4.2.12 Predicting proteins 3D structure and docking

The 3D structures of the proteins from ‘K. tikkala’ strain Bc01, ‘K. frurule’ strain Bc03, and ‘R. jaberico’ strain Bc11 were predicted using Colabfold version 1.4.0(Mirdita et al., 2022) on the Gadi server at the National Computational Infrastructure (NCI). To determine structural similarity, the protein structures were run through pairwise structure alignment using the Flexible Structure Alignment by Chaining Aligned fragment pairs allowing Twists (FATCAT), which allows for flexible protein structure comparison(Z. Li et al., 2020; Ye & Godzik, 2003).

To predict the phage protein interaction with the bacterial host, the previously predicted 3D protein structures of all the proteins for *B. cellulosilyticus* WH2 were downloaded from the AlphaFold Protein Structure Database via the Google Cloud Platform(Varadi et al., 2022). All protein pairs were docked using hdock-lite v1.1 on the Gadi server. The results from hdock were sorted based on the binding score (hdock-scores) in the output file to identify the highest-quality binding predictions for each phage protein. In general, lower HDock-scores are indicative of more favourable or stronger protein-protein interactions, suggesting a higher likelihood of stable complex formation. Higher HDock-scores, on the other hand, may suggest weaker or less favourable interactions. The 3D structure of the proteins was visualised using Chimera.

4.3.14 Data availability

The genomes used in this research are available on the Sequence Read Archive (SRA) within the project, PRJNA737576. *Bacteroides cellulosilyticus* WH2, *Kehishuvirus* sp. 'tikkala' strain Bc01, *Kolpuevirus* sp. 'frurule' strain Bc03, and 'Rudgehvirus jaberico' strain Bc11 are all available on GenBank with accessions NZ_CP072251.1 (*B. cellulosilyticus* WH2), OQ198717.1 (Bc01), OQ198718.1 (Bc03), and OQ198719.1 (Bc11), and we are working on making the strains available through ATCC. The 3D protein structures for the three *Crassvirales* genomes are available to download at doi.org/10.25451/flinders.21946034.

4.3 Results

4.3.1 Search for *Crassvirales* phages

We obtained a total of 41 phages from wastewater infecting four different *Bacteroides* species, *B. cellulosilyticus* WH2, *B. fragilis* NCTC 9343, *B. stercoris* CC31F, and *B. uniformis* ATCC 8492. The phages were sequenced using Oxford Nanopore or Illumina MiSeq platforms, and the resulting sequences were assembled. We performed BLASTN searches of the assembled phages against the non-redundant (nr/nt) NCBI database for taxonomic assignment. As a result, we identified 14 phages infecting *B. cellulosilyticus* WH2 that belong to the *Crassvirales* order. Each of these phages was labelled with a code ranging from Bc01 to Bc14.

4.3.2 Isolation and taxonomic classification of *Crassvirales* isolates

Crassvirales isolates formed distinct clear circular plaques with a uniform diameter of 1 mm on soft agar overlays. We performed shotgun sequencing on the 14 isolates of phages, with Bc01 to Bc03, Bc05 to Bc11 sequenced on the Oxford Nanopore platform, and Bc01 to Bc08, Bc12 to Bc14 sequenced on the Illumina platform. The assemblies produced multiple contigs, and our selection criteria to identify complete phage genomes were based on presence of viral genes, highest read coverage and unitigs (high-quality contig) size of approximately 100 kb (Table 4.1). This resulted in complete genomes for each of the 14 phages, which were polished with Illumina reads correcting for substitution, insertion, and deletion errors.

Table 4. 1: Taxonomic classification of the 14 *Crassvirales* genomes isolated from wastewater infecting *Bacteroides cellulosilyticus* WH2

Phage isolate	**Sequencing platform**	**Genome length (bp)**	**Taxonomy**	**Biosample ID**
Bc01	MinION, MiSeq	100,722	*Kehishuvirus* sp. 'tikkala' strain Bc01	SAMN20326212
Bc02	MinION, MiSeq	98,905	*Kolpuevirus* sp. 'frurule' strain Bc02	SAMN20326213

Bc03	MinION, MiSeq	99,379	*Kolpuevirus* sp. 'frurule' strain Bc03	SAMN20326214
Bc04	MiSeq	99,033	*Kolpuevirus* sp. 'frurule' strain Bc04	SAMN29929441
Bc05	MinION, MiSeq	97,832	*Kolpuevirus* sp. 'frurule' strain Bc05	SAMN20326216
Bc06	MinION, MiSeq	99,845	*Kolpuevirus* sp. 'frurule' strain Bc06	SAMN20326217
Bc07	MinION, MiSeq	98,518	*Kolpuevirus* sp. 'frurule' strain Bc07	SAMN20326218
Bc08	MinION, MiSeq	98,067	*Kolpuevirus* sp. 'frurule' strain Bc08	SAMN20326219
Bc09	MinION	98,788	*Kolpuevirus* sp. 'frurule' strain Bc09	SAMN20326220
Bc10	MinION	96,329	*Kolpuevirus* sp. 'frurule' strain Bc10	SAMN20326221
Bc11	MinION	90,458	'Rudgehvirus jaberico' strain Bc11	SAMN20326222
Bc12	MiSeq	96,952	Kolpuevirus sp. 'frurule' strain Bc12	SAMN29929442
Bc13	MiSeq	90,716	'Rudgehvirus jaberico' strain Bc13	SAMN29929443
Bc14	MiSeq	98,803	Kehishuvirus sp. 'tikkala' strain Bc14	SAMN29929444

For taxonomic classification of these isolates, we applied the ICTV report guidelines for defining taxonomy within *Crassvirales* order. Phylogenetic clustering of the conserved portal gene and average nucleotide identity (ANI) species cutoff (95% identity over 85% genome coverage) identified three distinct clusters (Figure 4.2A). We selected the highest confidence genomes: Bc01, Bc03, and Bc11 from each cluster. These three isolates were compared against all known *Crassvirales* genomes through phylogenetic

clustering of conserved genes, major capsid protein (MCP), portal, terminase large subunits (*terL*) genes to determine that Bc01 and Bc03 clusters belong to the *Steigviridae* family, and Bc11 to the *Intestiviridae* family (Figure 4.2B, Figure S4.1). Confirmation of the genus assignment was obtained through ANI and shared protein information, which identified Bc01 to *Kehishuvirus*, Bc03 to *Kolpuevirus*, and Bc11 to a novel genus group that we propose to name 'Rudgehvirus*'*.

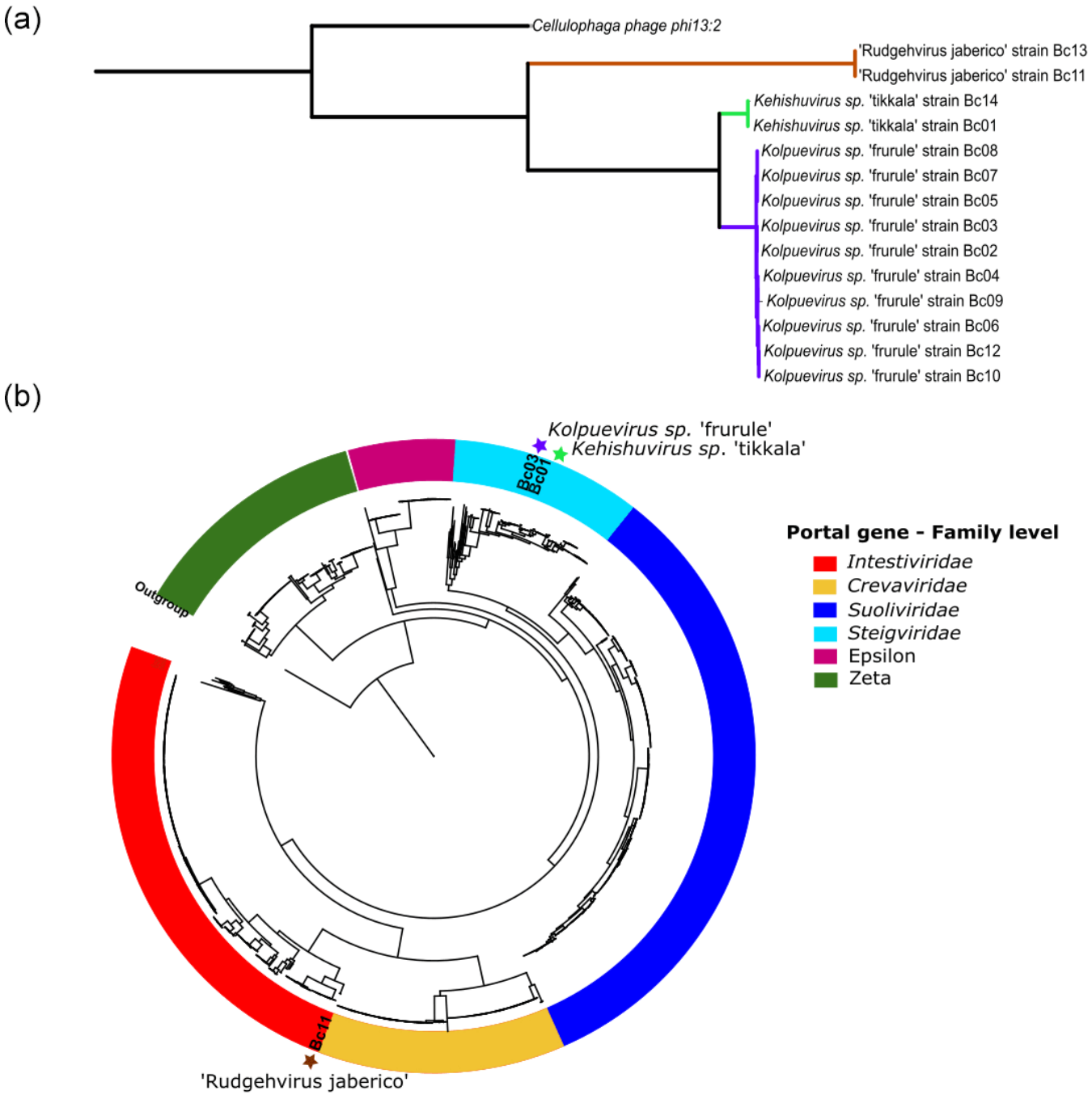


Figure 4. 2: Phylogenetic tree constructed using the portal protein using JTT model, CAT approximation with 20 rate categories and outgroup set to *Cellulophaga* phage phi13:2 A) Phylogenetic tree of the 14 *Crassvirales* isolates with the branches colour-coded to represent the three species, *Kehishuvirus* in light green, *Kolpuevirus* in purple, and 'Rudgehvirus' in brown. B) Clustering of all known *Crassvirales* genomes confirming that isolate *Kehishuvirus* sp. 'tikkala' strain Bc01 and *Kolpuevirus* sp. 'frurule' strain Bc03 belong to the family *Steigviridae* (cyan), and 'Rudgehvirus jaberico' strain Bc11 to *Intestiviridae* (red).

All three isolates represent novel species exhibiting less than 95% identity and 85% coverage to any known *Crassvirales* genomes. Bc01 is most similar to the reference genome *Kehishuvirus primarius* (Genbank ID: MH675552) with 95.5% identity across 79.1% genome coverage. Bc03 aligns with *Kolpuevirus hominis* (Genbank ID: MT774391) with 82.8% identified across 53.73% query coverage. Bc11 aligns with the reference genome *Jahgtovirus intestinalis* (Genbank ID: OGOL01000109) with 74.7% identity across only 9.9% query coverage. We proposed names for these novel species as *Kehishuvirus* sp. 'tikkala' strain Bc01, *Kolpuevirus* sp 'frurule' strain Bc03, and 'Rudgehvirus jaberico' strain Bc11.

4.3.3 Genome characteristics of the novel Crassvirales species

Kehishuvirus sp. 'tikkala' strain Bc01 is 100,841 bp, with 104 proteins, 24 tRNAs and GC content of 35.09 % (Table 4.2), which is lower than the bacterial host GC content of 42.8 %. These isolates formed clear, uniform circular spot plaques approximately 1 mm in diameter, forming $9.3*10^9$ PFU/mL (Figure 4.3A). Transmission electron microscopy (TEM) revealed they have podovirus-like morphology, displaying polyhedral capsids with a diameter of 94 ± 3 nm, tails with collar structures that were 34 ± 3 nm, with tail fibres of variable lengths (Figure 4.3B). From the calculated capsid size and genome length, we see that this phage packages its DNA at a density of 0.54 bp/nm^3. This genome lacks direct terminal repeat sequences, stop-codon reassignment, and lysogeny-related genes (Table S4.1).

Table 4. 2: Genome characteristics of the three novel *Crassvirales* species

Genome	**Length (bp)**	**GC %**	**Coding density**	**no. of CDS**	**Unknown function**	**tRNA**	**DTR**
Kehishuvirus sp. 'tikkala' strain Bc01	100,841	35.09	91.84	104	58	24	False
Kolpuevirus sp. 'frurule' strain Bc03	99,523	33.00	92.06	108	63	5	False
'Rudgehvirus jaberico' strain Bc11	90,575	29.15	87.45	84	48	0	False

Kolpuevirus sp. 'frurule' strain Bc03 shares the *Steigviridae* family with 'K. tikkala' strain Bc01. This genome is 99,523 bp long with GC content of 33%, 108 genes and four tRNA genes encoding arginine, asparagine, and tyrosine (Table 4.2). Similar to 'K. tikkala' strain Bc01, this phage also formed clear, uniform circular spot plaques, but formed $2.3*10^9$ PFU/mL (Figure 4.3A). Displaying a podovirus morphology, the virion was slightly larger than K. tikkala' strain Bc01, with capsids of diameter 97 ± 3 nm, a tail with collar structures of 33 ± 3 nm (Figure 4.3B), and packaging its DNA at a lower density of 0.48 bp/nm^3. Annotation of genes

confirmed the absence of direct terminal repeats, stop-codon reassignments, and lysogeny-related genes (Table S4.2). This is the first isolate within its genus.

‘Rudgehvirus jaberico’ strain Bc11 belongs to *Intestiviridae* family in a novel genus. This genome is 90,575 bp long, with a 29.15% GC content, and encodes 84 genes, lacking tRNA genes (Table 4.2). Unlike the above two species, this isolate formed plaques with a circular halo, indicating depolymerase activity, forming 3.75 x10^3 PFU/mL (Figure 4.3A). This isolate’s virion was relatively smaller in size compared to the other two isolates, with tails measuring 25 ± 4 nm (Figure 4.3B). Despite the smaller capsid size and genome, this phage packages its DNA at a density of 0.56 bp/nm^3, comparable to ‘K. tikkala’ strain Bc01. Similar to the other two genomes, direct terminal repeats, stop-codon reassignments, and lysogeny-related genes (Table S4.3) were absent.

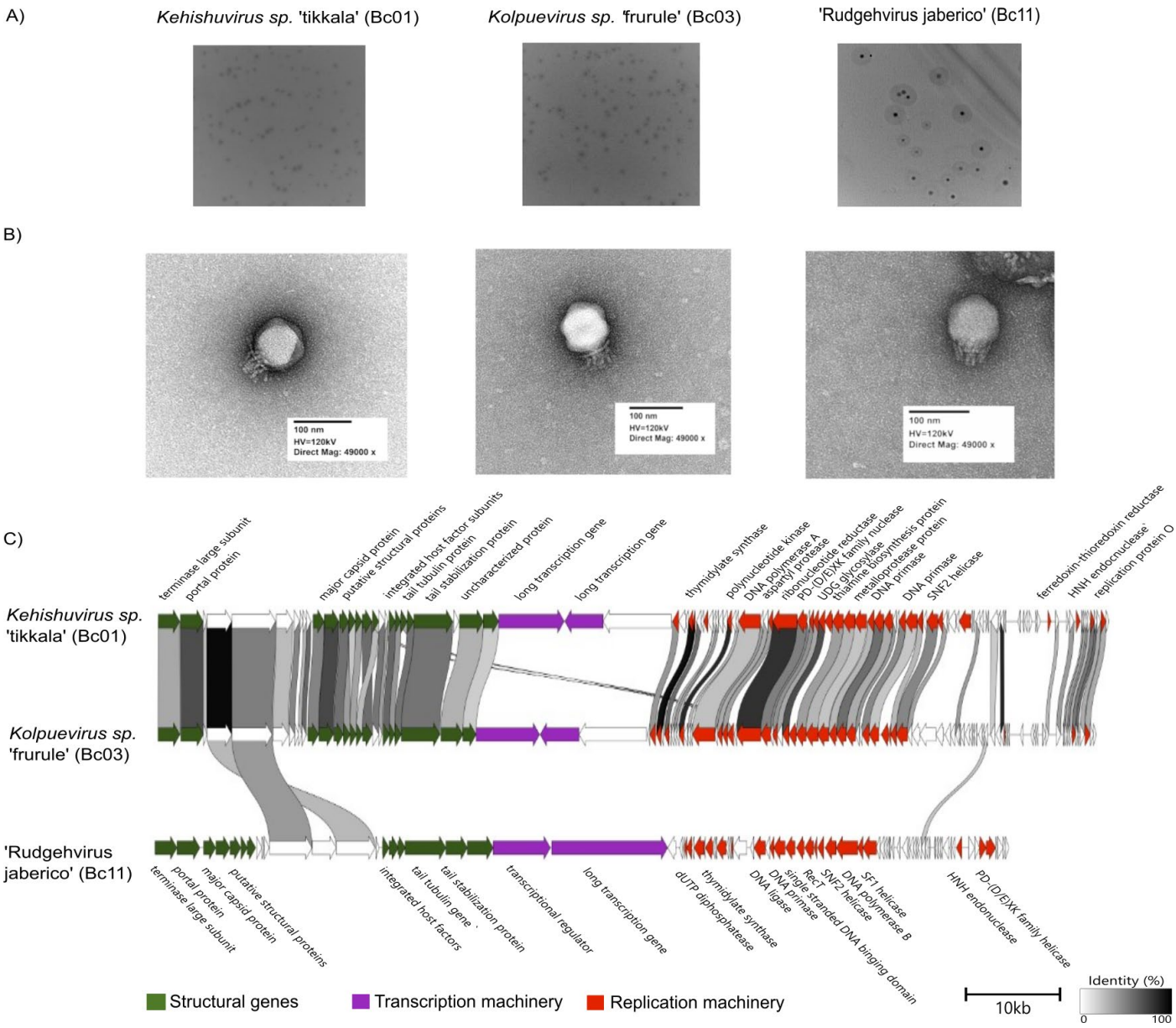


Figure 4. 3 A) Plaque morphology of three species, ‘K. tikkala’ strain Bc01, ‘K. frurule’ strain Bc03, and ‘R. jaberico’ strain Bc11 B) Transmission electron microscopy images negatively stained with uranyl acetate of the three isolates C) Gene arrangement and functional annotation of the three genomes colour-coded based on their functional modules and hypothetical genes represented in white. The direction of the arrows represents the direction of the gene read from the genome, and the arrows themselves represent individual genes. The links connecting the genes indicate amino acid sequence identity, ranging from 30% (grey) to 100% (black).

Comparative analysis across the three isolates shows that ‘K. tikkala’ strain Bc01 and ‘K. frurule’ strain Bc03, belonging to the same family (*Steigviridae*) exhibits higher gene similarity with each other. In contrast, ‘R. jaberico’ strain Bc11 from *Intestiviridae* family displays distinct gene arrangements (Figure 4.3C). Notably, all three genomes share two structural genes encoding tail spike proteins, which include a domain that encodes polysaccharide-degrading enzymes, such as glycoside hydrolases. Structural protein 1 encompassing the ‘K. tikkala’ strain Bc01 protein (WEU69744.1) shared 97 % sequence similarity with the ‘K. frurule’ strain Bc03 protein (WEY17522.1), while collectively these sequences share greater than 39 % similarity with ‘R. jaberico’ strain Bc11 protein (WEU69859.1) (Figure S4.2). Similarly, structural protein 2 encompasses ‘K. tikkala’ strain Bc01 protein (WEU69745.1) shared 59 % sequence similarity with ‘K.

frurule' strain Bc03 protein (WEY17523.1) and together share more than 46% similarity with 'R. jaberico' strain Bc11 protein (WEU69857.1) Figure S4.3).

4.3.4 Synteny across all seven *Crassvirales* species successfully isolated

The comparison of the three novel species from this study that infect the same bacterial host with the four isolate *Crassvirales* genomes that infect other *Bacteroides* hosts showed expected gene similarity based on their taxonomic assignment. Among the *Steigviridae* genomes, 'K. tikkala' strain Bc01 was most similar to *K. primarius, sharing 76 of 106 genes,* and the two genomes from *Wulfhauvirus* genus (strains DAC15 and DAC17) shared 115 of the 121 genes with greater than 30% similarity. 'K. frurule' strain Bc03 belonging to a unique genus, *Kolpuevirus* exhibited intermediate similarity, sharing 68 genes with *Kehishuvirus* and 71 genes with *Wulfhauvirus* genus (Figure 4.4A). Within the *Intestiviridae* family, 'R. jaberico' strain Bc11 was compared to the *J. secundus,* and they shared 37 genes, including 11 structural genes, three transcription genes, and 23 replication-related genes (Figure 4.4B).

The exception to the taxa-based similarity was the two structural genes, encoding tail spike proteins that were shared only among isolates infecting the same host, 'K. tikkala' strain Bc01, 'K. frurule' strain Bc03, 'R. jaberico' strain Bc11 despite belonging to different genera (Figure 4.4C).

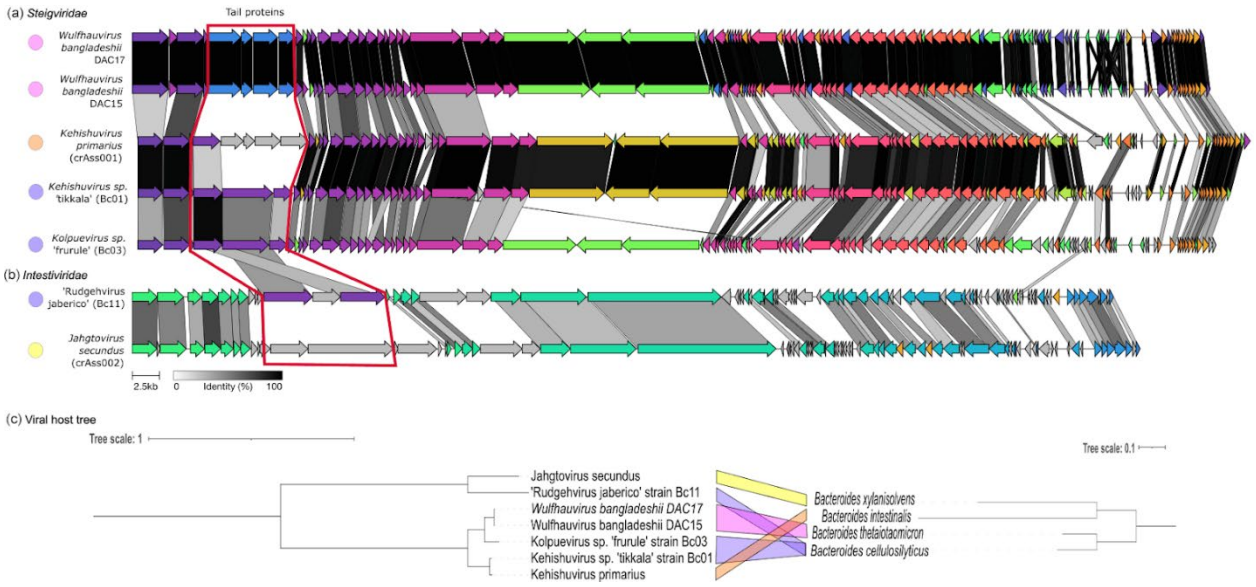


Figure 4. 4: Gene synteny across seven pure culture isolates across two *Crassvirales* families A) *Steigviridae* family comprising five isolates spanning across three genera B) *Intestiviridae* family comprising two isolates from two genera. Genes are represented as arrows, with their direction indicating the gene's direction, and their colour indicating the cluster group. Grey-coloured arrows represent unique genes that did not form any clusters. Finally, the links connecting the genes are colour-coded based on sequence similarity, ranging from grey (30%) to black (100%). The tail proteins shared among the three isolates from this study are highlighted in a red box. A dot is added next to each of the phage to represent the bacterial host, *B. thetaiotaomicron* in pink, *B. xylanisolvens* in orange, *B. cellulosilyticus* in purple, and *B. intestinalis* in pink C) Viral host-tree constructed using the portal gene for Crassvirales species and 16S rRNA gene for the bacterial hosts, with unique colours connecting the phage to its bacterial host.

As there were multiple *Crassvirales* species infecting multiple bacterial hosts (Figure 4.4C), we performed a coevolutionary test using Parafit(Legendre et al., 2002) that supported random association between *Crassvirales* phages with their bacterial hosts (Parafit Global = 3.33, p values >0.05).

4.3.5 Structural proteins playing a role in host interaction

To investigate the phage genes involved in host interaction, we compared all 1,887 genes across the 18 Crassvirales genomes, including 14 from this study that infect the bacterial host B. cellulosilyticus WH2, and the four Crassvirales isolates that infect four different bacterial hosts: *B. intestinalis, B. xylanisolvens,* and *B. thetaiotaomicron*. Together, from 18 *Crassvirales* isolates 1,766 genes were categorised into 383 orthologous groups (Figure 4.5), while the remaining 121 genes remained singletons. To reinforce the validity of this analysis, we corroborated that the species tree inferred from orthogroups (Figure 4.5) exhibits the same species-level clustering as observed in the phylogenetic tree (Figure 4.2A). There was one exception, *J. secundus,* which belongs to *Intestiviridae* family, was grouped with the *Steigviridae* isolates instead of its relative 'R. jaberico' strains, due to gene duplication or recombination events in this genome.

Following the species-level clustering, we identified 64 orthogroups (193 genes) that were specific to *Kehishuvirus,* 55 orthogroups (564 genes) specific to *Kolpuevirus,* 89 orthogroups (187 genes) specific to *Wulfhauvirus,* 73 orthogroups (148 genes) specific to 'Rudgehvirus', and 5 orthogroups (10 genes) specific to *Jahgtovirus* genera (Figure 4.5). Within these groups, only two orthogroups–OG000000 (including Bc01: WEU69745.1, Bc03: WEY17523.1, Bc11: WEU69858.1) and OG000008 (including Bc01: WEU69744.1, Bc03: WEY17522.1, Bc11: WEU69859.1) included genes only from the 14 *Crassvirales* isolates that infect the same bacterial host, *B. cellulosilyticus* WH2. However, in orthogroup OG000000, four gene duplication events occurred with at least 50% of the descendant species having retained both the gene duplicates; therefore, this orthogroup was not investigated further. Conversely, no gene duplication events were observed within OG000008.

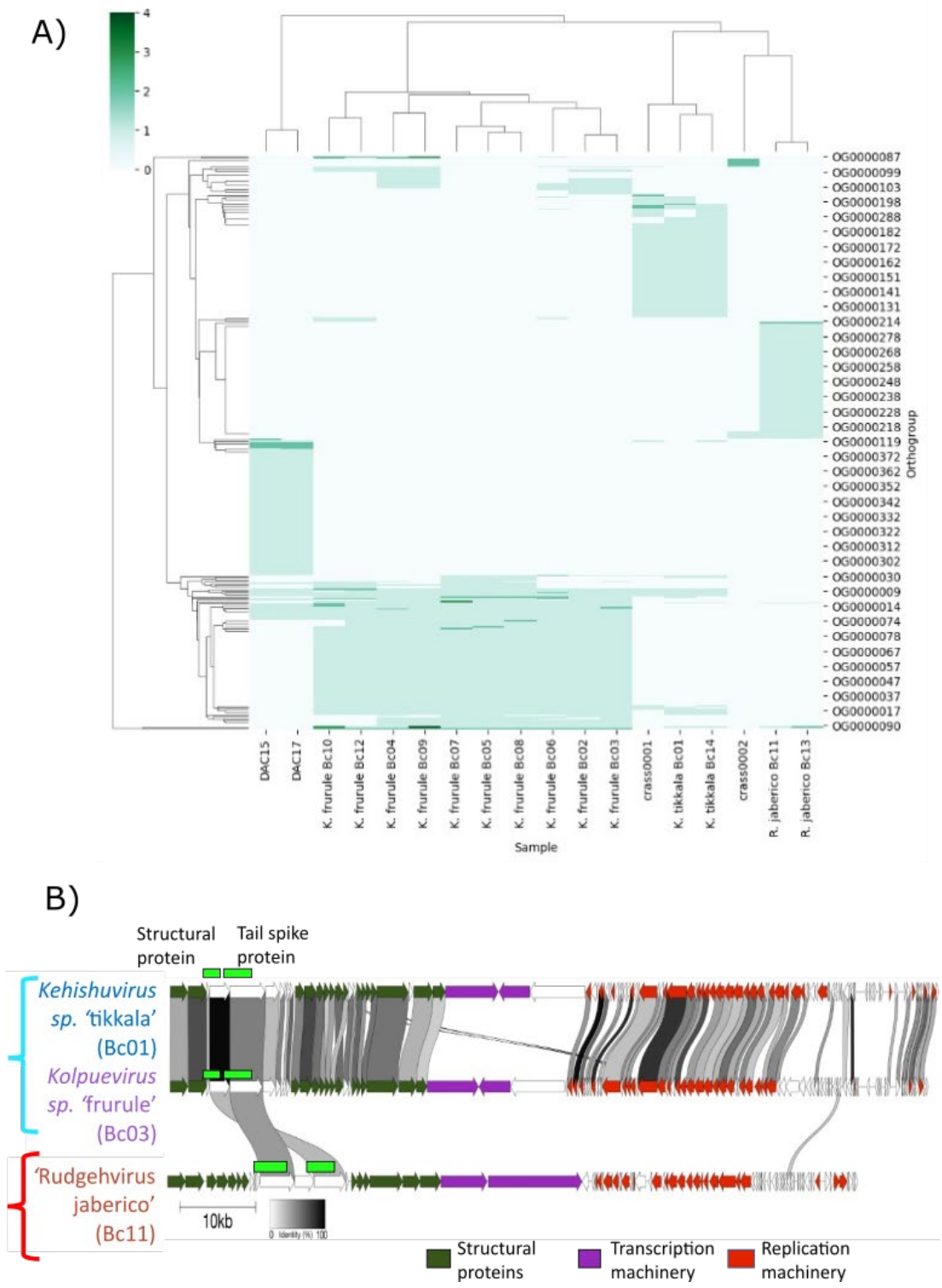


Figure 4. 5 A) Orthologous groups identified across the 18 Crassvirales isolates highlighted the two orthogroups that are present within the 14 *Crassvirales* isolates from this study, infecting the same bacterial host. B) Highlighting the two orthogroups (in green) that were identified to be undergoing selection pressure.

To determine if the genes in OG000008 are undergoing selection pressure, we calculated the number of synonymous (d_S) and non-synonymous (d_N) mutations occurring (d_N/d_S <1). Averaging all the sequence pairs, we used the codon-based z-test to identify genes under selection and found that OG000008 rejected the null hypothesis (z-score=0.56, p-value<0.001), suggesting that these genes are under purifying

selection. As recombination can impact this analysis, we ran Genetic Algorithm for Recombination Detection (GARD) to detect recombination, which identified five recombination breakpoints, none of which were significant to be detected by the genetic algorithm. We therefore investigated the tail spike protein structure and role in host interaction.

4.3.6 Tail spike protein interacts with TonB-dependent receptors

To identify the potential host interactions, we predicted the structure of all 103 proteins from 'K. tikkala' strain Bc01, 109 proteins from 'K. frurule' strain Bc03, and 83 proteins 'R. jaberico' strain Bc11 were generated using Colabfold(Mirdita et al., 2022) (Protein structures available at doi.org/10.25451/flinders.21946034). Specifically, we compared the folded structures of tail spike proteins belonging to orthogroup OG000008 (Bc01: WEU69744.1, Bc03: WEY17522.1, Bc11: WEU69859.1) using Flexible Structure Alignment by Chaining Aligned fragment pairs allowing Twists (FATCAT-rigid) method to show 'K. tikkala' strain Bc01 is similar to 'K. frurule' strain Bc03 with root-mean square deviation (RMSD) of 5.86 Å and 74 % of paired residues in the structural alignment (Figure 4.6A). On the other hand, the tail spike protein of 'K. tikkala' strain Bc01 exhibited an RMSD of 6.61 Å and 60 % identity when compared to 'R. jaberico' strain Bc11 (Figure 4.6B).

Each of the tail spike protein structures was individually docked against all 3,223 predictions from the *B. cellulosilyticus* WH2 proteome available in the AlphaFold database using hdock-lite (Table S4.4). 'K. tikkala' strain Bc01 tail spike protein (WEU69744.1) (Figure 4.6C) interacted best with TonB-dependent receptors (UniProt ID: A0A0P0GGA2, hdock-score =-700) (Figure 4.6D), 'K. frurule' strain Bc03 protein (WEY17522.1) with another TonB-dependent receptor (UniProt ID: A0A0P0GR14, hdock-score= -694), and 'R. jaberico' strain Bc11 protein (WEU69859.1) with a different TonB-dependent receptor (UniProt ID: A0A0P0FZA4, hdock-score= -574).

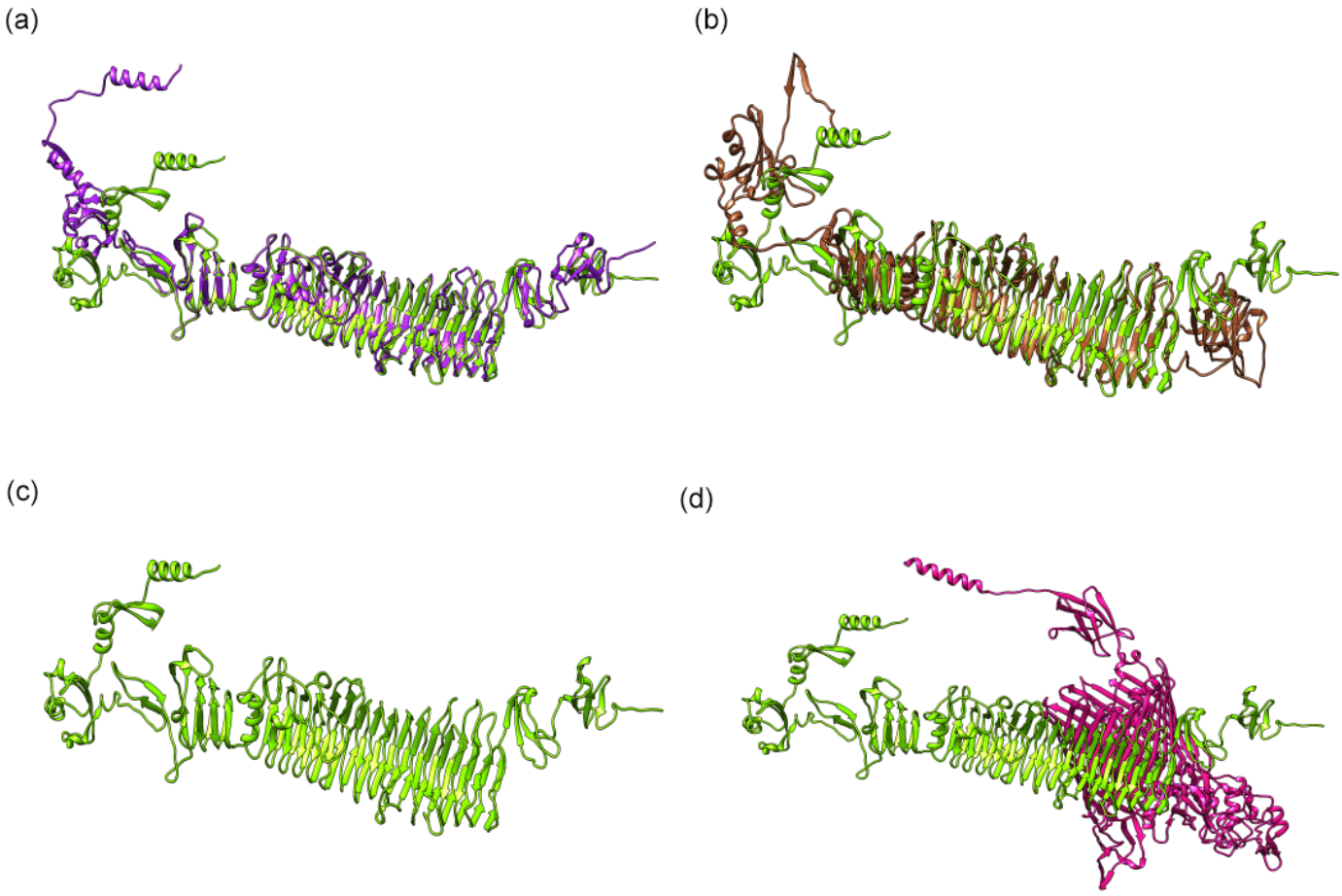


Figure 4. 6: 3D structure of tail spike proteins visualised using Chimera. A) Structural alignment of tail spike protein *Kehishuvirus* sp. 'tikkala' strain Bc01 (WEU69744.1 in green) with *Kolpuevirus* sp 'frurule' strain Bc03 (WEY17522.1 in purple). B) Structural alignment of tail spike protein *Kehishuvirus* sp. 'tikkala' strain Bc01 (WEU69744.1 in green) with 'Rudgehvirus jaberico' strain Bc11 (WEU69859.1 in brown). C) 3D Structure of *Kehishuvirus* sp. 'tikkala' strain Bc01 (WEU69744.1 in green) D) *Kehishuvirus* sp. 'tikkala' strain Bc01 (WEU69744.1 in green) docked with *Bacteroides cellulosilyticus* WH2 TonB-dependent receptor (A0A0P0GGA2 in pink).

4.4 Discussion

The role of *Crassvirales* genomes in the human gut is enigmatic, hindered by the limited number of cultured *Crassvirales* phages. Here, we address this gap by successfully isolating three novel *Crassvirales* species infecting *Bacteroides cellulosilyticus* WH2, belonging to different genera and families. This observation suggests that the phages are not co-evolving with their bacterial hosts; rather, they have a shared ability to exploit similar features in their bacterial hosts. Notably, we identified a unique tail spike protein shared among isolates infecting the same bacterial host, undergoing purifying selection and interacting with the TonB-dependent receptors on the bacterial surface.

The *Crassvirales* order is currently comprised of a vast and diverse collection of genomes. Despite this, the study of these phages has been limited due to the scarcity of pure isolates. The challenge associated with successful isolation highlights the difficulty in identifying and predicting the bacterial hosts associated with these organisms. In our study, we addressed this challenge through focusing on wastewater samples, a source for phages infecting different *Bacteroides* hosts. Employing this approach, we successfully isolated 14 novel *Crassvirales* isolates specifically infecting *B. cellulosilyticus* WH2. These isolates were sequenced

on different sequencing platforms, including Oxford Nanopore, Illumina Miseq or a combination of both. Nanopore assemblies provided high-quality and complete assemblies, however required polishing the assembly with Illumina reads to correct for frameshift errors that can fragment genes(Arumugam et al., 2019; Cook, Brown, et al., 2023; Nanoporetech Consortium, 2022). As a result, the 14 complete genomes were classified at the family and genus levels, denoted as three novel species (Figure 4.2A), while the remaining isolates were grouped as strains of the same species (Table 4.1). The highest confidence isolate was selected for each species, *Kehishuvirus* sp. 'tikkala' strain Bc01, *Kolpuevirus* sp, 'frurule' strain Bc03 and 'Rudgehvirus jaberico' strain Bc11 and examined further.

Taxonomic assignment of the three novel species showed they belong to two families. *Kehishuvirus* sp. 'tikkala' strain Bc01 and *Kolpuevirus* sp, 'frurule' strain Bc03 were assigned to the *Steigviridae* family. This family also comprised of three other *Crassvirales* phages, *Kehishuvirus primarius, Wulfhauvirus bangladeshii* DAC15, and *Wulfhauvirus bangladeshii* DAC17 infecting other *Bacteroides* hosts. Notably, the two novel isolates exhibited clear, uniform circular spot morphology distinct from the turbid plaques observed in *K. primarius*, despite their close relationship within the same family. The third novel species, 'Rudgehvirus jaberico' strain Bc11 belonged to *Intestivirdae* family, along with other *Crassvirales* species, *Jahgtovirus secundus.* 'R. jaberico' strain Bc11 presented plaques with a circular halo surrounding the cleared spot, indicating depolymerase activity to break down the polysaccharides found on the bacterial cell wall.

Furthermore, comparing the three novel species, we found that their virion production, estimated from the number of plaques formed, was correlated with the number of tRNA genes within the genome(Delesalle et al., 2016). However, it is possible there are other factors such as gene regulation and host immune responses that could also be influencing virion production. Additionally, we conducted genome density analysis in association with capsid sizes and genome lengths, revealing inconsistencies with prior studies on isolated *K. primarius* and *J. secundus* species. The capsid diameters of the new three novel *Crassvirales* species of virions (90 to 97 nm) were apparently 20% larger in size than those reported for *K. primarius* and *J. secundus* virions (77 nm) (Guerin et al., 2021; Shkoporov et al., 2018). However, considering that the reported values for *K. primarius* and *J. secundus* corresponded to the inscribed rather than the circumscribed diameters, a geometric correction of 22% that brought the genome density near 0.5 bp/nm^3. This correction aligned with a larger diameter measured in the recently published cryo-EM reconstruction of *K. primarius* (Bayfield et al., 2023). The finding highlights the importance of accurately assessing virion dimensions and genome density to ensure consistency in the classification of *Crassvirales* phages.

The addition of the three novel *Crassvirales* species spanning multiple families infecting one bacterial host, *B. cellulosilyticus* WH2 indicated these species may not be co-evolving with their bacterial hosts. We therefore tested all the successfully cultured *Crassvirales* species and their respective bacterial hosts to discover that they do not exhibit co-evolutionary patterns but rather support random association. These findings imply that the phage-host association within *Crassvirales* group are shaped by the environment

and host interactions(Legendre et al., 2002; Papudeshi, Rusch, et al., 2023). Additionally, genome comparison of the known *Crassvirales* species showed greater shared similarity within genera. However, the three *Crassvirales* species, despite belonging to three different genera, shared two unique structural genes. Evolutionary analysis confirmed one of the two structural genes, encoding tail spike protein (comprising Bc01: WEU69744.1, Bc03: WEY17522.1, Bc11: WEU69859.1) formed an orthologous group, and is undergoing purifying selection pressure. Tail spike proteins have been shown to play a crucial role in binding to specific membrane receptors on the bacteria in tailed bacteriophage(Nobrega et al., 2018). Therefore, through preserving this gene function, the phage can successfully infect and replicate within the host.

We found the tail spike proteins of the three novel *Crassvirales* species to interact with different TonB-dependent receptors on the bacterial surface, providing significant insights into the mechanism of phage-host interactions. The bacterial host, *B. cellulosilyticus* WH2 possesses a substantial repertoire of up to 112 TonB-receptors on its surface. *Bacteroides* typically use these receptors to take up starches(Pollet et al., 2021) and have been associated with phage sensitivity (N. T. Porter et al., 2020; Shkoporov, Khokhlova, et al., 2021). The tail spike protein also encodes for polysaccharide-degrading enzymes, such as glycoside hydrolase domains, that target the capsular polysaccharides on the bacterial surface, allowing for phage-host interaction and leading to infection. This interaction therefore ensures successful propagation, highlighting the evolutionary adaptation between the *Crassvirales* phage and their bacterial hosts.

Overall, our study on the three novel *Crassvirales* species infecting *Bacteroides cellulosilyticus* WH2 revealed critical insights into their evolutionary dynamics and interactions with the bacterial host. The novel phages belonging to different genera but infecting the same host provide a valuable model system for studying the interactions that occur within one of the dominant members of the gut microbiome.

4.5 Conclusions

Phages are increasingly recognised as key ecological players in the gut microbiome, where their ability to modulate bacterial populations can influence host health, microbial stability, and resilience. Despite their abundance, particularly of groups like *Crassvirales,* the specific interaction profiles between gut phages and their bacterial hosts remain poorly understood. This gap has limited our mechanistic understanding of how phages shape microbial ecosystems, beyond what metagenomics can infer. Addressing this, the work presented in this chapter provides one of the first experimental investigations of multiple *Crassvirales* families infecting a single gut bacterium, *Bacteroides cellulosilyticus*.

We identify a conserved tail spike protein used by these phages for host recognition, revealing a common mechanism of infection despite broad genomic divergence. Importantly, we also show that *Crassvirales* phages are not co-evolving with their hosts, suggesting alternative strategies for persistence in the densely populated and competitive gut environment. These findings significantly expand the known genomic diversity of this viral order and offer a functional framework for understanding phage–host specificity in complex ecosystems.

By linking genomic features to ecological function, this work provides a blueprint for studying host interactions in other phage groups and lays a foundation for integrating viral dynamics into broader models of gut microbiome structure and stability. Ultimately, this chapter highlights the need to move beyond one-to-one phage–host studies toward examining multipartite interactions, where phages, bacteria, and their surrounding community co-shape ecosystem outcomes.

4.6 Supplementary Files

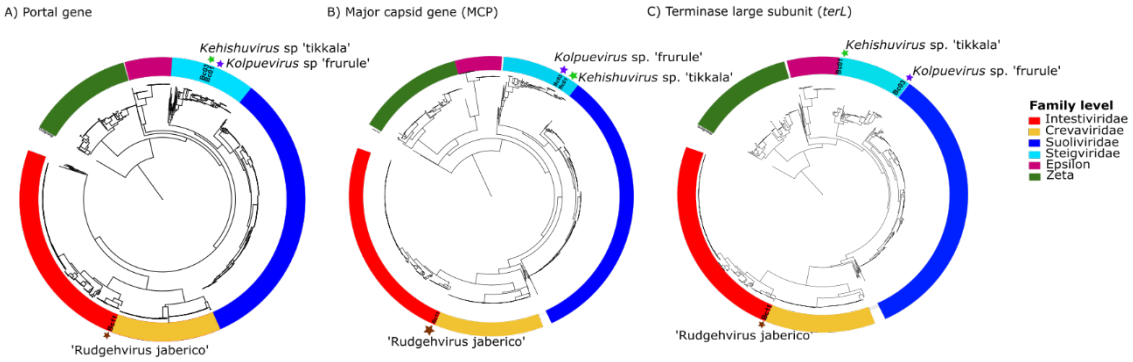


Figure S4.1: Showing the taxa classification of the three novel species remains consistent across the three conserved proteins A) portal gene, B) Major capsid protein (MCP), and C) terminase large subunit (*terL*). The outgroup across all three trees set to *Cellulophaga* phage phi13:2. The placement of the three novel species are highlighted on the tree, Bc01 belonging to *Kehishuvirus* genera (light green), Bc03 belonging to *Kolpuevirus* genera (purple), and Bc11 belonging to a novel genus named ‘Rudgehvirus’ (brown).

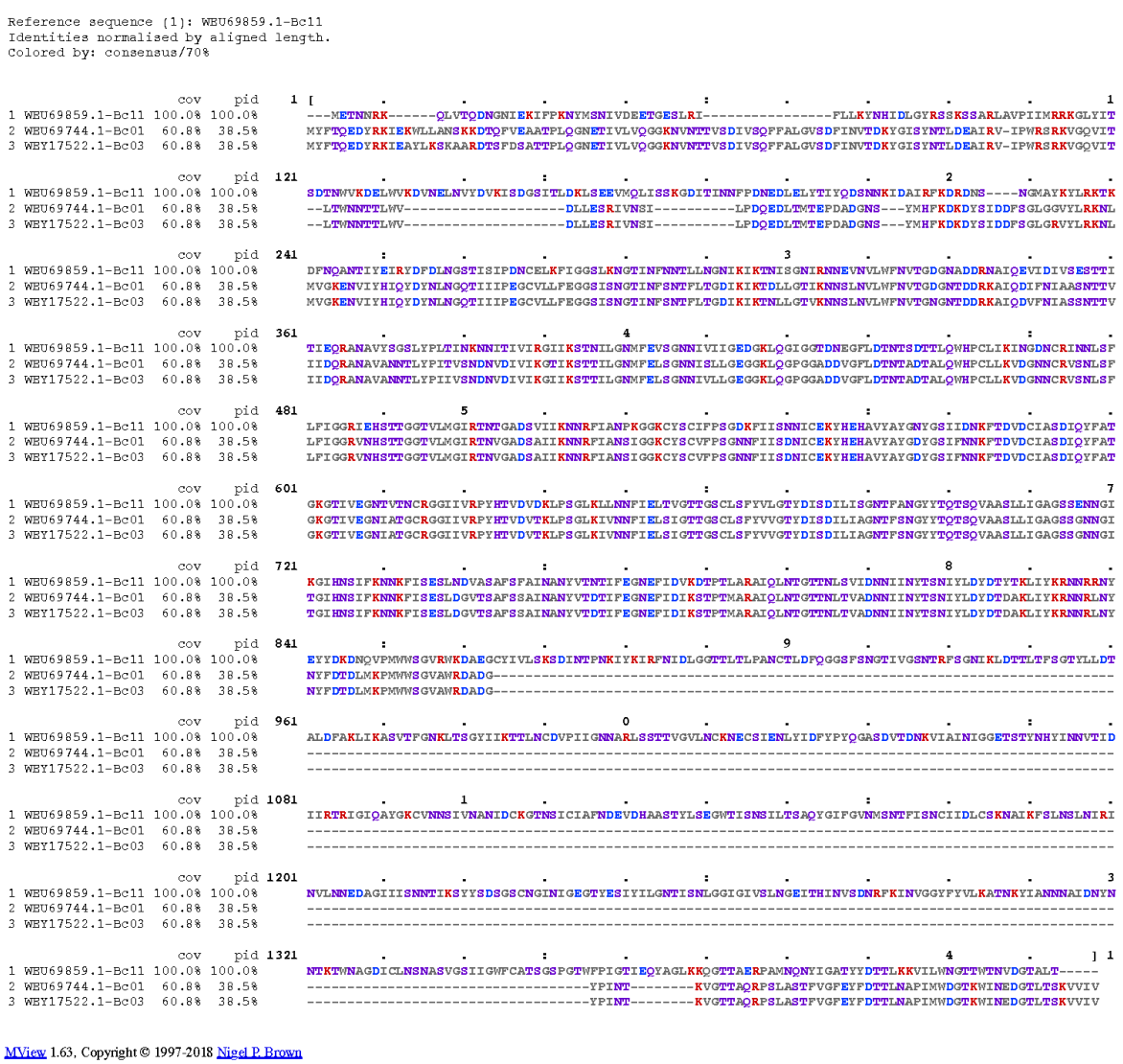


Figure S4. 2: Multiple sequence alignment of shared structural protein 1 from Figure 4.3C including K. tikkala' strain Bc01 protein, WEU69744.1, 'K. frurule' strain Bc03 protein, WEY17522.1, and 'R. jaberico' strain Bc11 WEU69859.1 (reference sequence), showcasing the sequence identity with amino acids that were not shared in grey, and the rest in a different colour, based on the amino acid group.

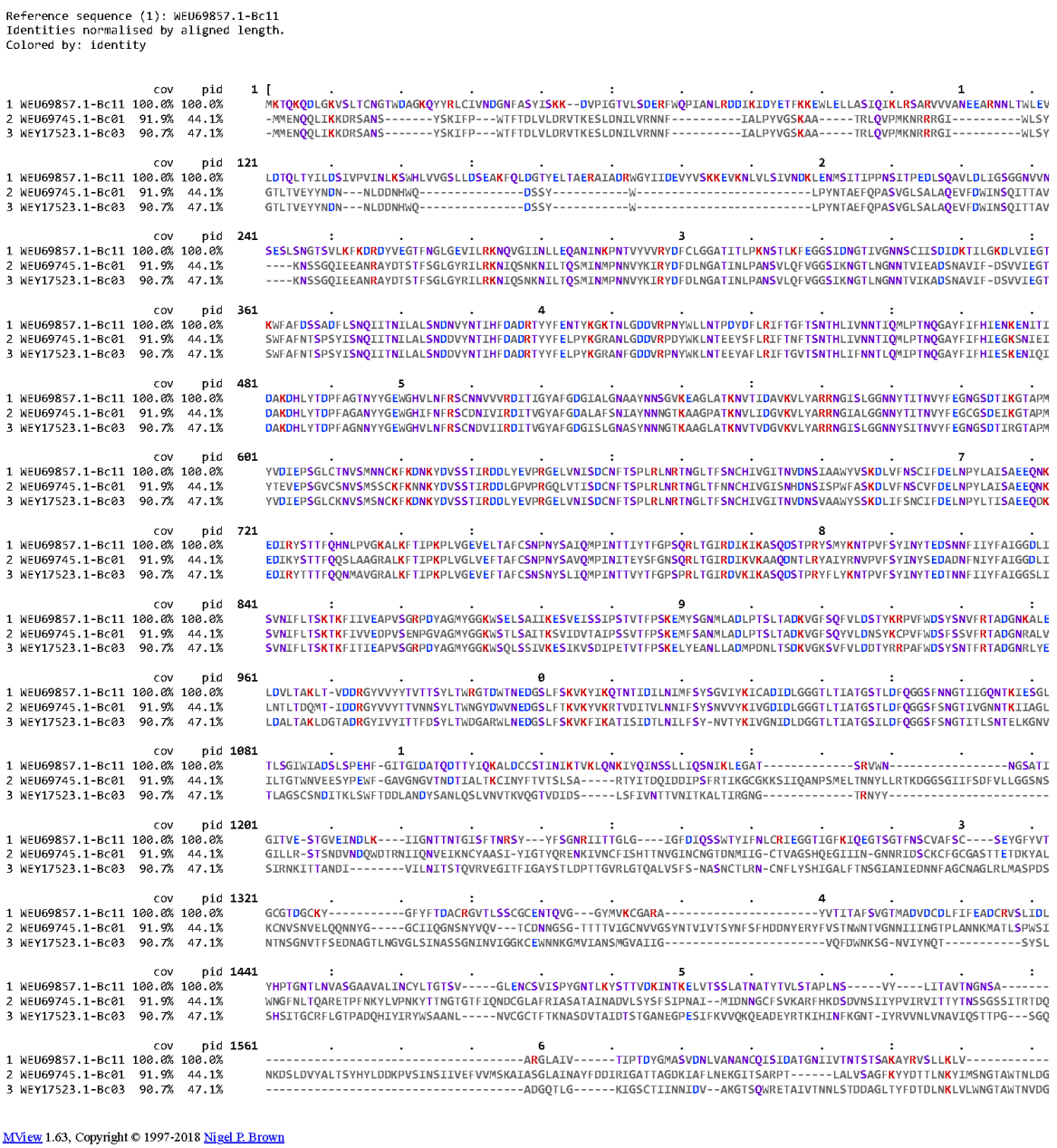


Figure S4. 3: Multiple sequence alignment of shared structural protein 2 from Figure 4.3C including ‘K. tikkala’ strain Bc01 protein, WEU69745.1, ‘K. frurule’ strain Bc03 WEY17523.1, and ‘R. jaberico’ strain Bc11 WEU69857.1 (reference sequence), showcasing the sequence identity with amino acids that were not shared in grey, and the rest in different colour, based on the amino acid group.

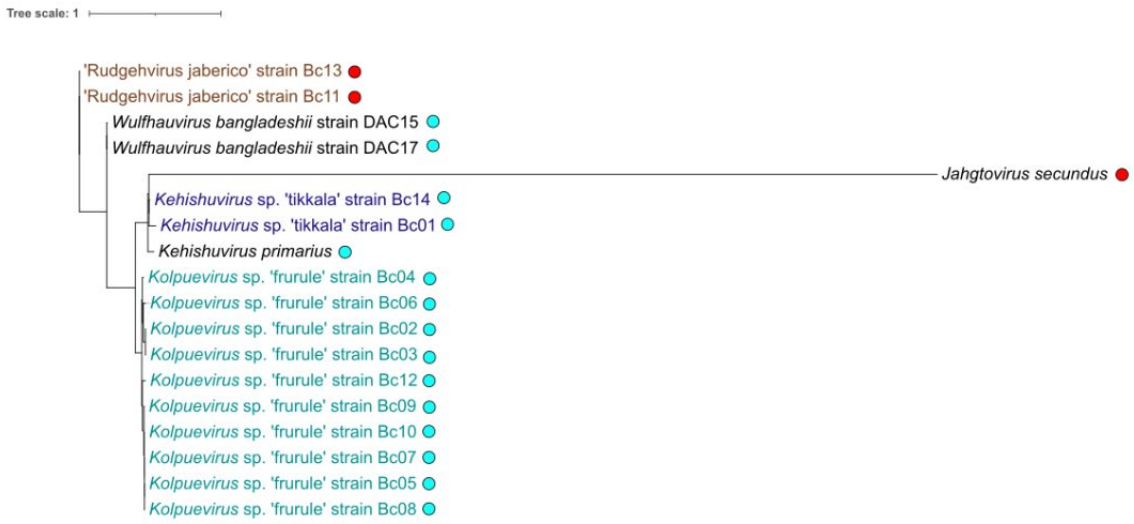


Figure S4.4. Species tree inferred from orthogroups using OrthoFinder and rooted to 'Rudgehvirus jaberico' strain Bc13 using the STRIDE algorithm. The 14 *Crassvirales* isolates are colour coded based on their species classification, 'Rudgehvirus jaberico' strains in brown, *Kehishuvirus* sp. 'tikkala' strains in dark blue, and *Kolpuevirus* sp. 'frurule' strains in light blue. The four *Crassvirales* genomes from other studies are colour coded in black. Further, to denote the family level classification, we added red dots next to *Intestiviridae* family members, and cyan dots next to *Steigviridae* family members.

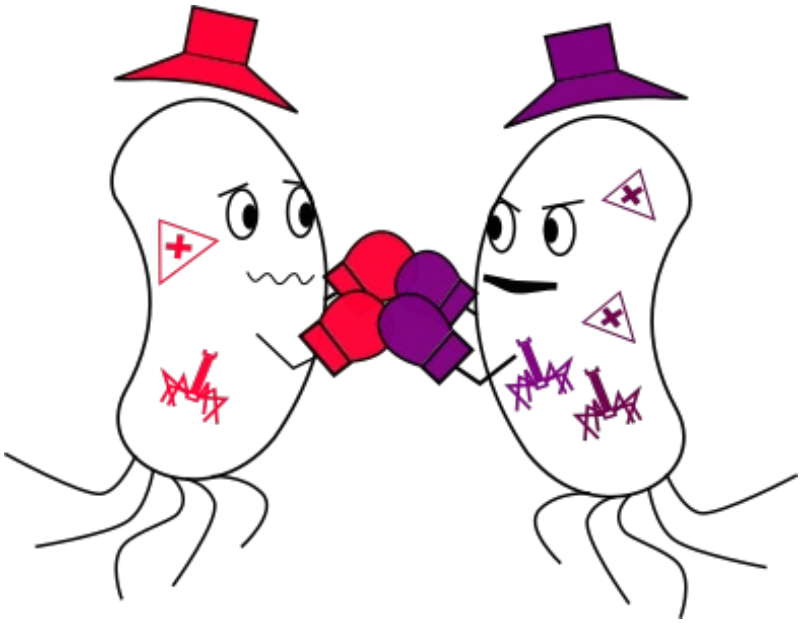
Table S4.1: Provided as Supplementary File 1,
https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.001100#supplementary_data

Table S4.2: Provided as Supplementary File 2,
https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.001100#supplementary_data

Table S4.3: Provided as Supplementary File 3,
https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.001100#supplementary_data

Table S4.4: Provided as Supplementary File 4,
https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.001100#supplementary_data

CHAPTER 5
BACTERIAL SYMBIONTS HOST ASSOCIATION



This chapter is based on the published article— **Papudeshi, B**., Rusch, D. B., VanInsberghe, D., Lively, C. M., Edwards, R. A., & Bashey, F. (2023). Host association and spatial proximity shape but do not constrain population structure in the mutualistic symbiont Xenorhabdus bovienii. MBio, 14(3), e00434-23. https://doi.org/10.1128/mbio.00434-23.

Statement on the Use of Generative Artificial Intelligence (AI): No generative AI was used to write this manuscript.

Preface

This chapter is based on the published research article titled “Host association and spatial proximity shape but do not constrain population structure in the mutualistic symbiont *Xenorhabdus bovienii,”* which I led as first author and forms a foundational part of this thesis. The study, published in mBio in 2023, explores how mutualistic bacteria evolve within host-associated environments while maintaining genetic diversity across populations. Using *Xenorhabdus bovienii*, a bacterial symbiont of *Steinernema* nematodes, we tested how host association, spatial proximity, and dispersal shape bacterial population structure and evolutionary trajectories. This research was motivated by fundamental questions in microbial ecology: how do symbionts adapt to their hosts without losing genetic variability, and how do physical and ecological constraints shape their population dynamics? While strict vertical transmission is often assumed in mutualistic systems, our findings challenge that assumption, revealing evidence of horizontal transmission and genetic recombination among lineages even within tightly host-associated contexts.

As first author, I was responsible for conceptualising the study, formulating the research questions, and designing the analytical framework. I performed population genomic analyses, integrated spatial data, conducted comparative genomics, and carried out statistical modelling to assess host-specific structuring in *X. bovienii*. I also led the interpretation of findings, manuscript writing, and figure preparation.

In the context of this thesis, this chapter presents essential evidence on how host association can drive, but not strictly determine, population structuring. It complements earlier chapters on phage-bacteria interactions by extending the theme of host influence to bacterial symbionts and sets up the comparative exploration of microbial specificity, co-evolution, and selective sweeps that continues throughout the thesis.

Authorship statement

As the first author of this study, I led the project from research design through to publication. I conceptualised the research questions, designed the analytical pipeline, conducted all core genomic and statistical analyses, and wrote the manuscript. I also coordinated the contributions of co-authors, integrating ecological and evolutionary perspectives into the final interpretation of results. Below is a breakdown of the author's contributions:

Author	**Contribution**
Bhavya Papudeshi	Research design, data analysis, writing and editing the manuscript
Douglas B. Rusch	Research design and data analysis
David VanInsberghe	Data analysis

Curtis M. Lively	Research design, data collection and editing of the manuscript
Robert A. Edwards	Data analysis and editing of the manuscript
Farrah Bashey	Research design, data collection, analysis, writing and editing of the manuscript

The contributions of each co-author have been explicitly stated, and their permission to include these works has been obtained as per Flinders University's Authorship of Research Output Procedures (Appendix A)

Host association and spatial proximity shape but do not constrain population structure in the mutualistic symbiont *Xenorhabdus bovienii*

Abstract

To what extent are generalist species cohesive evolutionary units rather than a compilation of recently diverged lineages? We examine this question in the context of host specificity and geographic structure in the insect pathogen and nematode mutualist *Xenorhabdus bovienii*. This bacterial species partners with multiple nematode species across two clades in the genus *Steinernema*. We sequenced the genomes of 42 *X. bovienii* strains isolated from four different nematode species and three field sites within a 240-km^2 region and compared them to globally available reference genomes. We hypothesised that *X. bovienii* would comprise several host-specific lineages, such that bacterial and nematode phylogenies would be largely congruent. Alternatively, we hypothesised that spatial proximity might be a dominant signal, as increasing geographic distance might lower shared selective pressures and opportunities for gene flow. We found partial support for both hypotheses. Isolates clustered largely by nematode host species but did not strictly match the nematode phylogeny, indicating that shifts in symbiont associations across nematode species and clades have occurred. Furthermore, both genetic similarity and gene flow decreased with geographic distance across nematode species, suggesting differentiation and constraints on gene flow across both factors, although no absolute barriers to gene flow were observed across the regional isolates. Several genes associated with biotic interactions were found to be undergoing selective sweeps within this regional population. The interactions included several insect toxins and genes implicated in microbial competition. Thus, gene flow maintains cohesiveness across host associations in this symbiont and may facilitate adaptive responses to a multipartite selective environment.

5.1 Introduction

Microbes live in complex and abstract microenvironments, obscuring our ability to determine what evolutionary forces structure the diversity we observe. Additionally, it is challenging to predict *a priori* the extent to which closely related isolates sampled from a specific region or habitat reflect a cohesive unit, distinct from other such units. As in macroorganisms, genetic distance can increase with geographic distance within microbial species(Chase et al., 2019; Cho & Tiedje, 2000; Edwards et al., 2019; Oda et al., 2003) and be correlated with distinct habitats(K. M. Campbell et al., 2017; McArthur et al., 1988), indicating that homogenising forces (i.e., selection, drift, and gene flow) are more likely to operate with physical and ecological proximity. However, diverse population structures are observed across bacterial species. For instance, nearly identical isolates of *Staphylococcus aureus* and *Vibrio cholerae* have been found globally(Dutilh, Thompson, et al., 2014; McAdam et al., 2012; Mutreja et al., 2011), while in other species, sympatric isolates are found to be genetically differentiated and nonrecombining(Cadillo-Quiroz et al., 2012; Chase et al., 2019; Shapiro et al., 2012), demonstrating that divergence can arise and be maintained at a small spatial scale. A key goal remains to link geographic patterns to the evolutionary forces shaping microbial populations.

Work on host-associated microbes has examined the role of hosts in governing the population structures of their symbionts. Some host specialist pathogens, such as *Mycobacterium tuberculosis*, display a long history of coevolution that can be seen by congruent phylogenies between the pathogens and their human host populations(Comas et al., 2013), while others, such as *Helicobacter pylori,* reflect more recent human migrations(Thorell et al., 2017). In contrast, host generalists, such as *Campylobacter* species and *Escherichia coli*, show little signature of host species association(Dearlove et al., 2016) and are found to be structured more by geography than host phylogeny(Matthews et al., 2015; Strachan et al., 2015). However, within some host generalist species, lineages can be found that are host specific and contain niche-adaptive genes(E. J. Richardson et al., 2018; Sheppard et al., 2010, 2013). While most research examining population structure has been done on pathogen species of human health or economic concern, it is important to study diverse species to better understand the processes shaping microbial evolution(Rocha, 2018).

Among beneficial symbionts, a range of population structures has also been observed. The well-studied mutualist *Vibrio fisheri*, associated with Hawaiian bobtail squid, shows little geographic structure or specificity to genetically distinct host populations(Bongrand et al., 2016). In contrast, vertically transmitted symbionts like *Buchnerna aphidocola* show structuring across aphid species and with host geography(Yang Zhang et al., 2018). Symbiont population structure can also be affected by host ecology(Lima et al., 2020). For example, the ant mutualist *Pseudonocardia actinobacteria* shows kilometre-scale geographic structuring within a single ant species that is correlated with its ability to inhibit a virulent fungal pathogen of its host(Caldera et al., 2019). Here, we examine the population structure of *Xenorhabdus bovienii*, a mutualistic symbiont of nematodes and a virulent insect pathogen, in a region where multiple nematode species occur in sympatry.

The bacterial genus *Xenorhabdus* is exclusively found associated with nematodes in the genus Steinernema. These nematodes depend on *Xenorhabdus* for successful colonisation and reproduction within insect hosts (Figure 5.1), while *Xenorhabdus* relies on the nematodes for survival and access to insects(S. Patricia Stock & Blair, 2008). Across the genera, there is a partial congruence between the host and symbiont phylogenies, with both co-speciation and host switching observed(M.-M. Lee & Stock, 2010). *X. bovienii* is noted within the genus for its ability to associate with multiple nematode species across two distinct clades of *Steinernema* nematodes (M.-M. Lee & Stock, 2010). Despite this broad host range, partial co-cladogenesis between *X. bovienii* and its nematode partners suggests specialisation (Murfin, Lee, et al., 2015). Furthermore, experimental pairings demonstrate that the fitness of both partners declines with phylogenetic distance from native association(Chapuis et al., 2009; Dinges et al., 2022b; McMullen et al., 2017; Murfin, Lee, et al., 2015). So, while on one hand there is evidence that *X. bovienii* can coevolve to form specialised partnerships, on the other hand, there is evidence that this species can be considered a host generalist(M.-M. Lee & Stock, 2010; Murfin, Lee, et al., 2015). To reconcile these findings, we sequenced genomes of *X. bovienii* isolated from four nematode host species across three study sites and compared them to all available genomes of this species. We hypothesised that this host generalist

symbiont would comprise multiple, largely host-specific lineages and sought to identify genetic markers of such specificity. Additionally, we hypothesised that spatial proximity would facilitate genetic similarity via shared selective pressures, neutral processes, and gene flow. Thus, we tested for evidence of recent gene flow among the isolates and whether gene flow and genetic similarity were limited by host species or geographic distance.

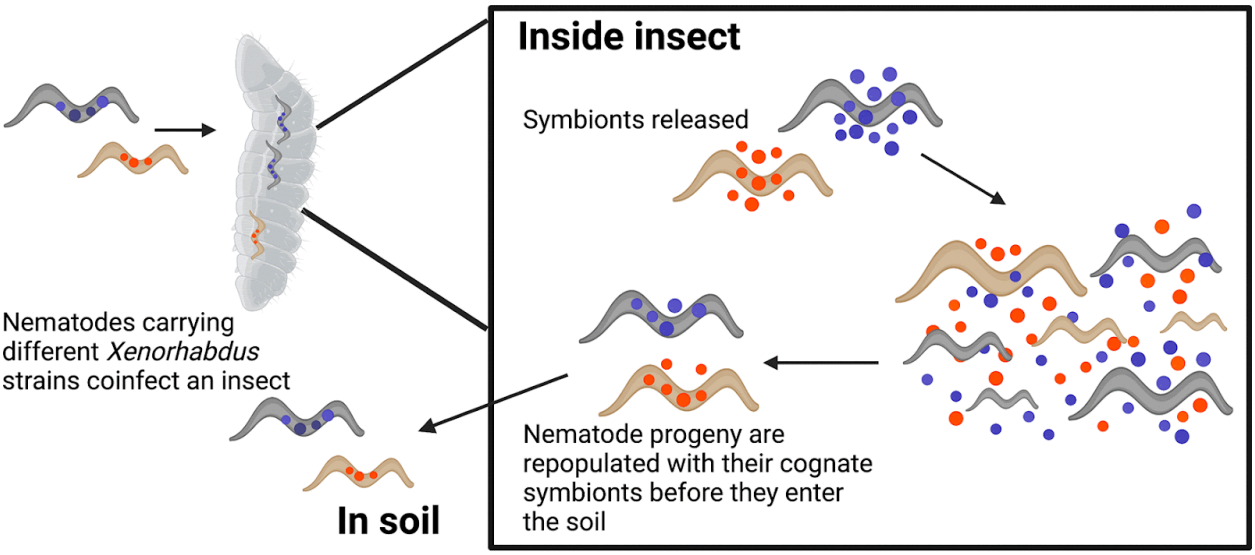


Figure 5. 1: Representation of the *Xenorhabdus-Steinernema* life cycle. Nematodes carrying different *Xenorhabdus* symbionts co-occur in the soil and coinfect an insect host. Inside the insect, nematodes release their symbionts, which replicate and produce toxins, killing the insect. The nematodes also replicate for one or more generations, producing offspring that do not carry the symbionts. When resources within the insect are depleted, nematode offspring reassociate with their cognate symbionts and nematode-symbiont pairings emerge into the soil. This image was generated using BioRender.

5.2 Materials and Methods

5.3.1 Study design

Forty-two *X. bovienii* isolates associated with four distinct nematode host species were collected from three Indiana University Research and Training Preserve sites in Indiana, USA (Figure 5.2). At each site, soil samples were collected and baited separately with insect hosts in the laboratory. Nematodes emerging from each soil-exposed insect were surface sterilised and crushed with a pestle to isolate their symbionts. The resulting supernatant was then plated onto NBTA (nutrient agar with 0.0025% bromothymol blue and 0.004% triphenyl tetrazolium chloride), and bacterial colonies were streaked for isolation to create freezer stocks as previously described(Hawlena, Bashey, & Lively, 2010). Prior work showed slight variation among bacterial symbionts within a nematode stock(Hawlena, Bashey, Mendes-Soares, et al., 2010), and therefore, only one bacterial strain was selected per nematode stock for sequencing, with one exception, LD27A and LD27B, which were isolated from the same stock. Nematode species were identified using 28S and internal transcribed spacer (ITS) genes(S. P. Stock et al., 2001).

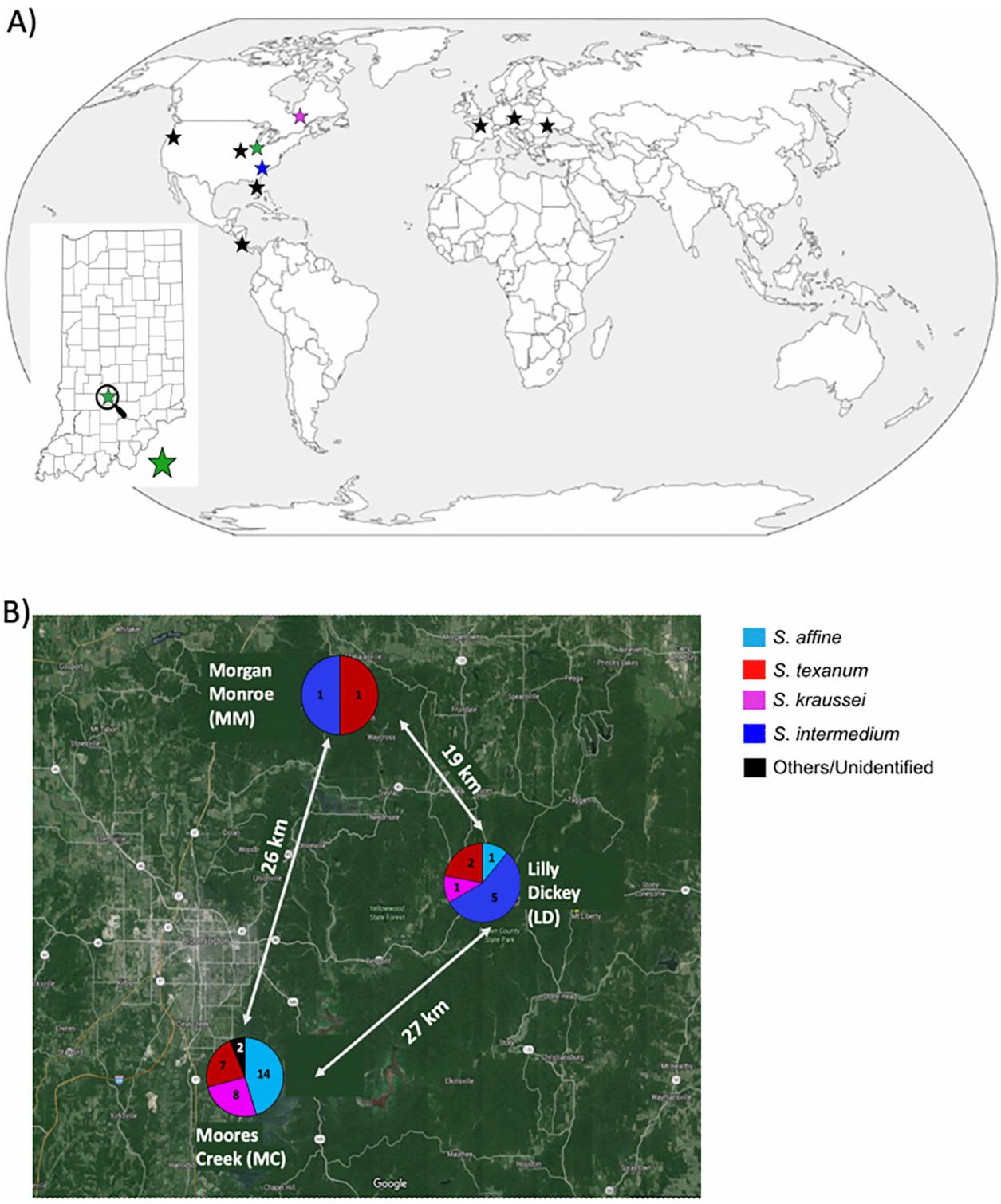


Figure 5. 2: Sample distribution of *Xenorhabdus bovienii* genomes. A) Newly sequenced isolates were collected from Indiana, USA, as depicted with a green star, while reference genomes from other studies deposited to NCBI and downloaded for this paper are represented in stars, colour-coded based on the nematode host and collection. B) Indiana isolates analysed in this paper were collected from three Indiana University Research and Teaching Preserve sites within a 240-km^2 region. Pie charts depict the relative numbers of isolates collected at each site and their nematode host associations. See Table S5.1 and NCBI BioProject accession number PRJNA700777 for information on each genome. Map outline and snapshot from Google Maps.

5.3.2 Genome sequencing, assembly and annotation

Each *X. bovienii i*solate from a freezer stock was plated on NBTA, and a single colony picked for overnight culturing in LB medium (Difco). DNA extraction was performed following the DNeasy blood and tissue kit

protocol for Gram-negative bacteria (Qiagen). Libraries with approximately 400-bp inserts were generated for each isolate and sequenced to generate paired-end reads on the Illumina NextSeq 500 platform using a 300-cycle kit. The reads were assembled with SPAdes assembler(Bankevich et al., 2012), and the contig statistics were assessed using QUAST v5.0.2(Gurevich et al., 2013). Additionally, 11 *X. bovienii* genomes (Figure 5.2), four *X. nematophila* genomes, and four *Photorhabdus* genomes were downloaded from NCBI (Appendix D, Table S1). All 61 genomes were clustered by average nucleotide identity (ANI) using FastANI(Jain et al., 2018). The ANI results were plotted in R using the ggplot2 and Heatmap packages. Next, the protein-coding genes were predicted in all 61 genomes using prodigal v2.6.3(Hyatt et al., 2010), and the resulting genes were annotated using Prokka v1.14.6(Seemann, 2014) against the *Xenorhabdus* gene database built from GenBank.

5.3.3 Phylogenetic analysis

Phylogenies were built from core regions. First, the assembled genomes were aligned using Mugsy v1r2.3(Angiuoli & Salzberg, 2011). The core genome was defined as regions found in all 61 genomes that were greater than 3,000 bp in length and with less than 50% gaps(Arevalo et al., 2019). Trees were constructed using RAxML v8.2.12(Stamatakis, 2014) using the general time reversible gamma (GTRGAMMA) model, with 100 bootstraps. *Photorhabdus* genomes were selected as the outgroup, and the resulting Newick tree was plotted using iTOL(Letunic & Bork, 2019). These steps were repeated for just the 53 *X. bovienii* isolates, with *X. bovienii* CS03 as the reference genome. We repeated the analysis for just the 42 regional isolates, with MC081 as the reference genome to ensure that the order of the genes was represented with a local sample. To detect and account for recombination across *X. bovienii* isolates, which can bias phylogenetic inference, the core gene alignment and initial phylogenetic tree were further analysed with ClonalFrameML v1.12(Didelot & Wilson, 2015). ClonalFrameML calculates the effect of recombination on the data set and generates a recombination-aware phylogenetic tree with adjusted branch lengths.

To compare the bacterial phylogeny to that of its nematode hosts, we used Parafit(Legendre et al., 2002) via the ape R package (permute = 1,000, eigen value correction = Cailliez). Pairwise distances were calculated from each tree using the cophenetic.phylo ape function. A maximum-likelihood nematode phylogeny was constructed in MEGA(Tamura et al., 2021) based on nematode 28S sequences available on GenBank (Table S5.1). We also conducted maximum-parsimony reconciliation via eMPRess to estimate host switching events(Santichaivekin et al., 2021).

5.2.4 Pangenomic analysis

To determine the flexible gene set across all the *X. bovienii* isolates, the genomes were run through a pangenomic analysis pipeline, Roary v3.13.0(Page et al., 2015). Roary was run with the minimum sequence identity set to 90%, clustering protein-coding genes from all 53 *X. bovienii* isolates. From the clustering results, the flexible and core gene sets were defined. To determine the grouping of *X. bovienii* isolates based on flexible gene sets, they were visualised using uniform manifold approximation and projection (UMAP) ordination plots in R.

5.2.5 GWAS analysis

To determine whether genetic markers could be associated with each nematode species, 53 *X. bovienii* genomes were run through treeWAS(Collins & Didelot, 2018). The input to treeWAS was the recombination-aware tree from ClonalFrameML and the core gene alignment used to build phylogenetic trees; we used the default parameters, setting the base P value to <0.05. In addition, all the SNPs were identified from the core gene alignment using SNP sites(Page et al., 2016). TreeWAS was also run with the gene presence and absence table from pangenome analysis to identify flexible genes that were significantly associated with nematode hosts. The significant traits were annotated through tracing the location of the trait to the Prokka annotations output. To determine whether the results were dependent on using globally available *Xenorhabdus* genomes, we reran this analysis using just the 42 Indiana isolates.

5.3.6 Gene flow analysis

Recent gene transfer events across all *X. bovienii* genomes were identified using PopCOGenT(Arevalo et al., 2019). The assembled genomes were provided as input to PopCOGenT, which first identifies gene flow between each pair of genomes by identifying regions of higher-than-expected similarity (termed length bias) based on a null model of clonal descent(Arevalo et al., 2019). Then, genomes connected by gene flow are grouped into populations, and clusters within populations are defined by genomes sharing relatively higher gene flow between them(Arevalo et al., 2019). As this analysis showed that all of the Indiana isolates shared gene flow with two of the reference genomes, *X. bovienii* intermedium (isolated from SC, USA) and *X. bovienii* kraussei Quebec (isolated from Canada), falling into a single population group, we repeated this analysis for just the 42 isolates to examine gene flow events that could be potential targets of selection within the region. For each cluster, selection is inferred by PopCOGenT through determining events that share low nucleotide diversity within a cluster and have distinct mutations between clusters across both core and flexible regions. The resulting gene sweeps were annotated from the corresponding output from Prokka.

5.3.7 Spatial analysis

Geographic distance between isolates was based on previously established field transects(Hawlena, Bashey, & Lively, 2010) or calculated from coordinates. The three field sites were less than 28 km apart (Figure 5.2), and within each field site, the isolates were collected less than 800 m apart from each other. Reference isolates were collected at least 370 km away from the Indiana isolates. We tested whether genetic similarity (average nucleotide identity [ANI]) and estimated gene flow (log10 length bias from the PopCoGenT analysis) were correlated with geographic distance (in log10 meters) by using Mantel tests via the vegan package in R. To test whether the nematode host species affected genetic similarity and gene flow, we classified each pair of isolates based on whether they were isolated from the same or different nematode species and then tested this effect in a full mixed-model analysis of covariance (ANCOVA), with geographic distance (in log10 meters) as a covariate and isolate identities and study sites as random effects. For each analysis, we tested all 53 *X. bovienii* isolates and then restricted the analysis to the 42 Indiana isolates, or to the isolates found at the MC and LD study sites.

5.3.8 Data availability

The genomes in this study have been deposited in GenBank under BioProject accession number PRJNA700777. In addition, the bioinformatics commands and files generated during analysis are available on GitHub (https://github.com/npbhavya/BovGenomes-analysis).

5.3 Results

5.3.1. Overview of X. bovienii genomes collected from Indiana

Genomes were obtained from 42 *X. bovienii* field isolates from four nematode host species across three study sites in a 240-km^2 region of Indiana (Figure 5.2). Each genome sequenced had an average of 42x genome coverage and was assembled to an average of 555 contigs. The least fragmented genome was 113 contigs (N50, 183,901 bp; total length, 4.56 Mbp), and the most fragmented was 4,329 contigs (N50, 42,376 bp; total length, 6.35 Mbp). On average 4,151 ± 312 proteins (mean ± standard deviation) were identified per isolate (range, 3,687 to 5,014), with 31.11% of the proteins annotated as hypothetical proteins. Comparisons of these genomes with 11 *X. bovienii* reference genomes, four *Xenorhabdus nematophila* genomes, and four *Photorhabdus* genomes show that all *X. bovienii* genomes have high nucleotide similarity (>94% average nucleotide identity [ANI], Figure 5.3) and form a monophyletic group based on a phylogeny of the core genome (100% support), distinct from other entomopathogenic bacteria (Figure S5.1).

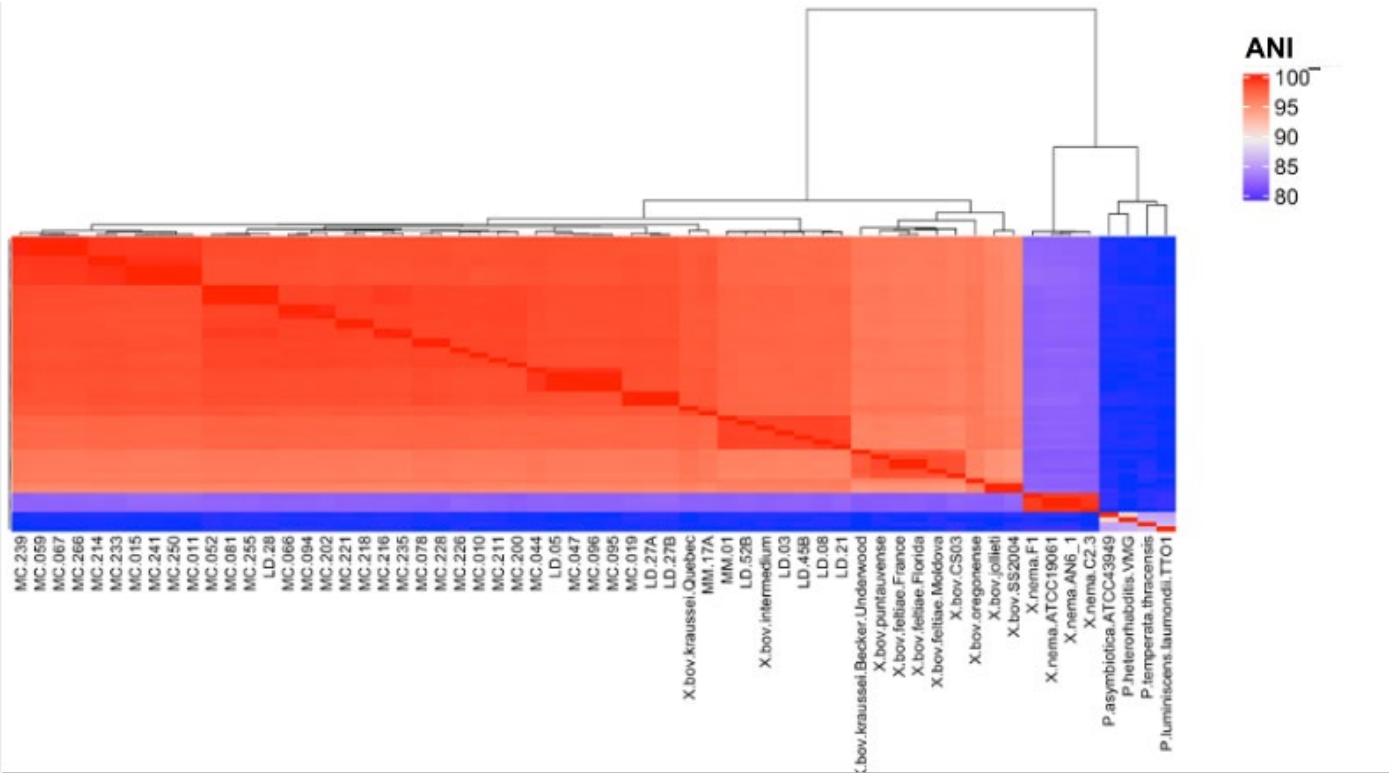


Figure 5. 3: Average nucleotide identity (ANI) of the whole genomes was compared across all 61 genomes, which include 42 *X. bovienii* Indiana isolates, 11 reference *X. bovienii* genomes, 4 *X. nematophila* spp and 4 *Photorhabdus* spp, using FastANI. Hierarchical clustering was performed on the Euclidean distance tables. The heatmap shows genome similarity, ranging from approximately 80% ANI in blue to 100% ANI in red. This figure shows *Photorhabdus* spp are equally distant to the two *Xenorhabdus* species, and that *X. nematophila* show an average of 82% ANI with *X. bovienii*. Similarity between *X. bovienii* isolates ranges from 94.34–99.99%. The Indiana isolates had a minimum of

96.9 % ANI and clustered into two distinct groups of more than 98% similarity. The first group includes 36 isolates plus the reference *X. bovienii* kraussei Quebec. The remaining six Indiana isolates were grouped with reference genome *X. bovienii* intermedium. Finally, the remaining nine reference *X. bovienii* genomes clustered together averaging 96.3% ANI.

5.3.2 Regional *X. bovienii* isolates form two distinct lineages partially based on nematode hosts

Alignment of the 53 available *X. bovienii* genomes (42 Indiana isolates and 11 reference genomes), results in a core region of 2,176,418 bp, which is 45.25% of the mean genome size. Phylogenetic analysis based on this alignment shows that 36 of the Indiana isolates group with reference genome *X. bovienii* strain kraussei Quebec forming lineage I (Figure 5.4A). Lineage I comprises all of the isolates associated with three of the nematode species: *Steinernema kraussei*, *Steinernema texanum,* and *Steinernema affine.* The remaining six Indiana isolates are all associated with *Steinernema intermedium* nematodes (Figure 5.4A, dark blue labels), and they form a monophyletic group (lineage II) with the reference *X. bovienii* strain intermedium. Thus, the bacterial phylogeny (Figure 5.4A) is not congruent with the nematode phylogeny (Figure 5.4B), where *S. intermedium* and *S. affine* group together, while *S. kraussei* and *S. texanum* belong to another clade. Notably, while all of the isolates from the Moore's Creek (MC) site are members of lineage I, isolates from the other two sites are found in both lineage I and II and cluster according to the nematode host.

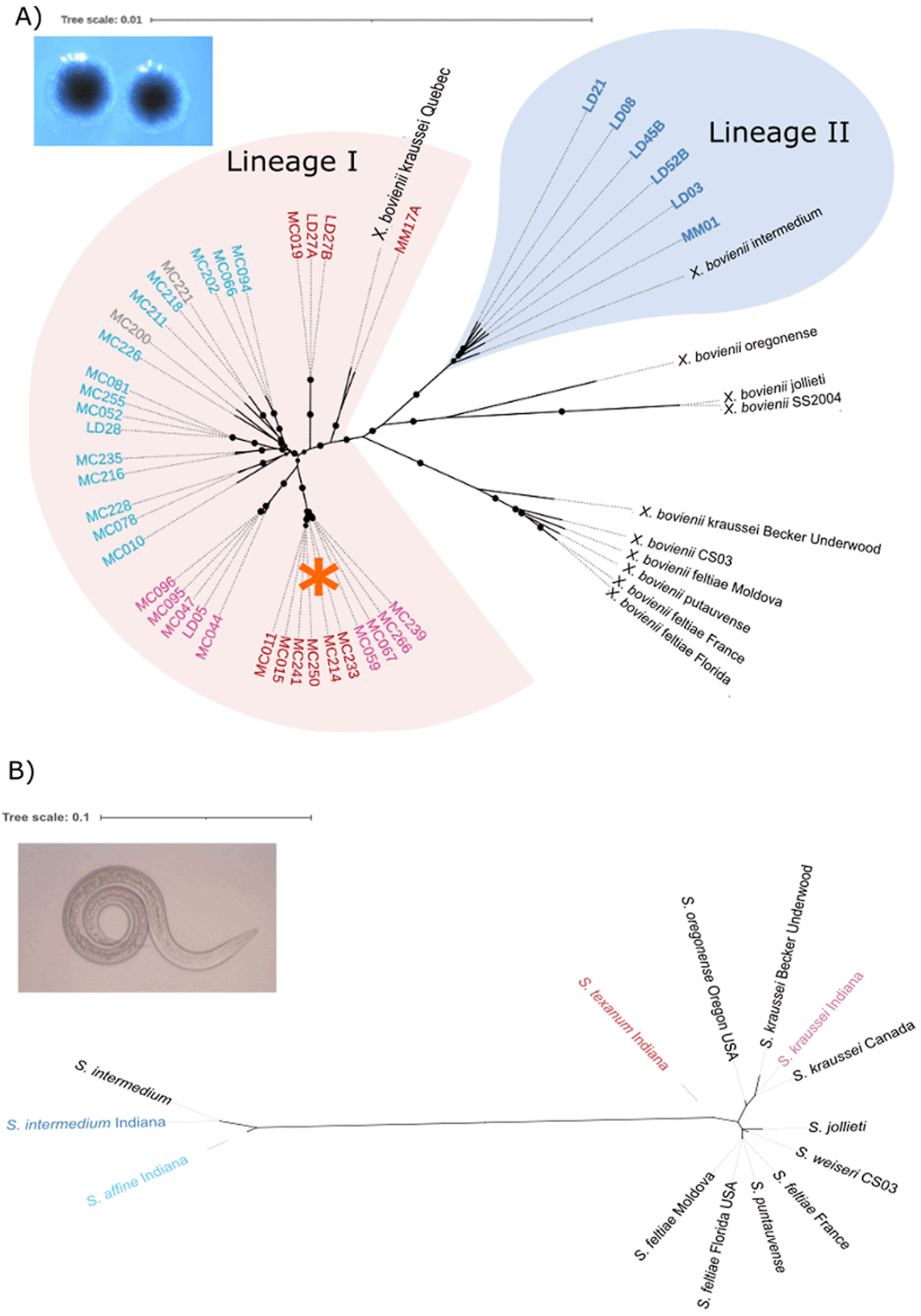


Figure 5. 4: A) Bacterial phylogeny with an image of *X. bovienii* colonies. Phylogenetic tree built using core genes from 53 *X. bovienii* genomes, with the reference genomes shown in black, the 42 Indiana isolates colour coded based on the nematode host, and two samples in grey that were isolated from an unidentified nematode host. The circles represent branches with bootstrap values ranging from 80% to 100%. The tree was built using RAxML based on the alignment of 2.18 Mb. Branch lengths have been corrected for recombination using ClonalFrameML. The orange

asterisk represents bacterial isolates from two nematode species that form a monophyletic group. B) Nematode phylogeny with an image of a *Steinernema* nematode. The nematode phylogeny was built from aligning 653 bp of the 28S rRNA gene using the general time reversible model of the maximum-likelihood method in MEGA. The nematode species are colour coded and named to match their corresponding symbionts across the two trees.

Cophylogenetic analysis using Parafit shows a non-random association between nematode species and *X. bovienii* isolates (ParafitGlobal = 0.003, P value = 0.001), supporting the clustering based on nematode host seen in Figure 5.3A. Nevertheless, this clustering is only partial. While all of the isolates associated with *S. affine* form one distinct group (Figure 5.4A, light blue labels), the other two species in lineage I do not. Isolates associated with *S. kraussei* form two distinct, well-supported clades (Figure 5.4A, pink labels) and do not form a monophyletic group with either of the two reference genomes associated with *S. kraussei*. Similarly, isolates from *S. texanum* form three distinct, well-supported clades (Figure 5.4A, labelled in red). Thus, host switching likely has occurred in this mutualism, as indicated by maximum-parsimony reconciliation (Figure S5.3).

Examination of the flexible gene content shows a pattern similar to that of the core phylogeny. Roary identified 2,147 genes as core and 15,867 genes as flexible in the *X. bovienii* pangenome. Clustering based on gene presence and absence places the *X. bovienii* isolates from *S. intermedium* (lineage II in Figure 5.5A, labelled in dark blue) in a distinct part of uniform manifold approximation and projection (UMAP) space (Figure 5.5, bottom left corner). Lineage I isolates fall along the diagonal, with *S. affine*-associated isolates (Figure 5.5, light blue labels) found more centrally, while isolates from *S. kraussei* and *S. texanum* are more dispersed. Additionally, the isolates in the monophyletic clade associated with *S. kraussei* and *S. texanum* (Figure 5.5A, labelled with an asterisk) form their distinct cluster (Figure 5.5, labelled with an asterisk). Thus, both the core and flexible genes support partial clustering based on the nematode host species, with isolates associated with *S. kraussei* and *S. texanum* showing recent host shift or gene exchange.

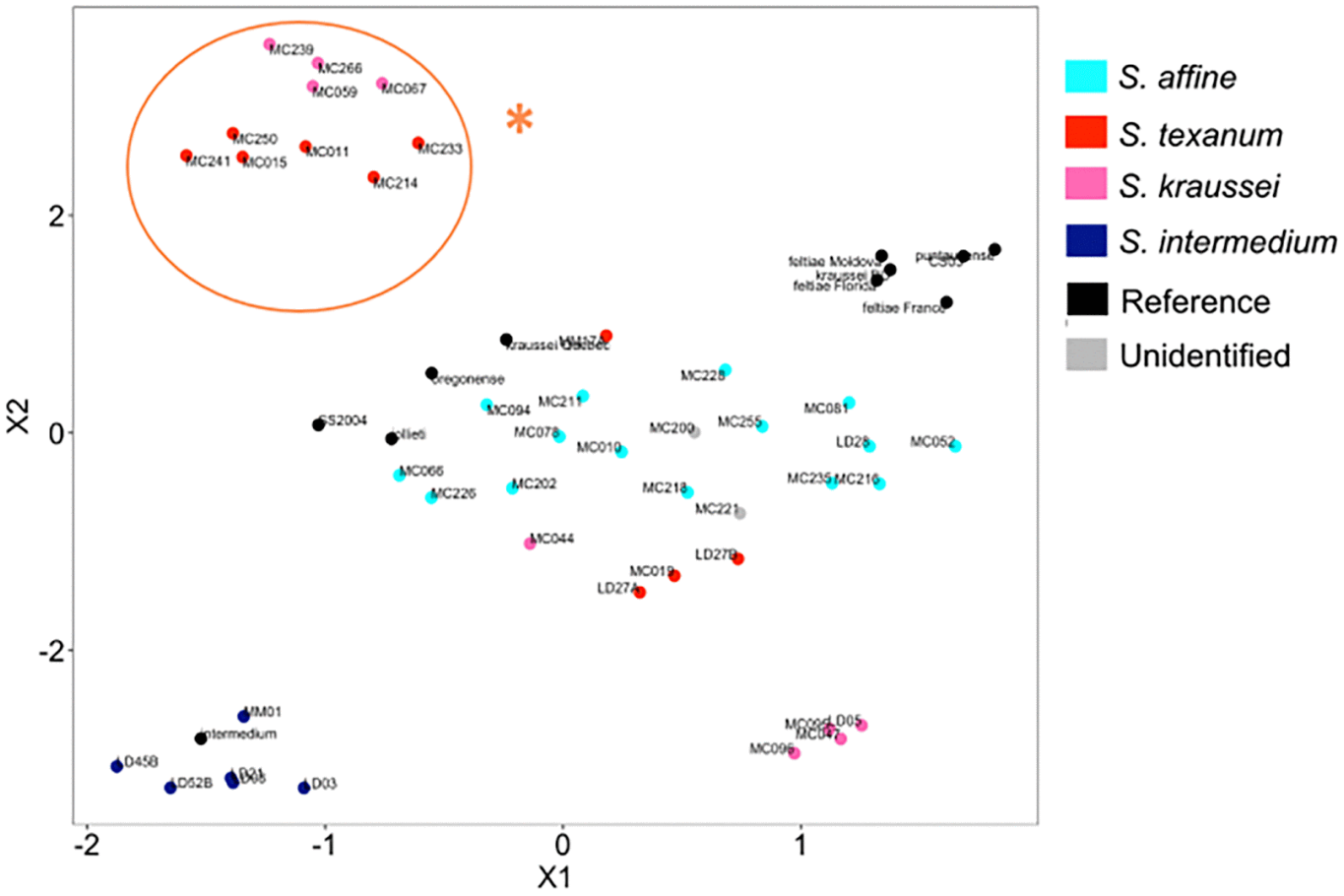


Figure 5. 5: UMAP visualisation based on gene presence and absence in the flexible gene set, with each data point representing a genome that is colour coded based on the nematode host. The orange asterisk shows isolates from two nematode species that form a monophyletic group in the core phylogeny shown in Figure 3.4B and form a distinct cluster based on the flexible genome.

5.3.3 Gene association testing across nematode hosts

To determine if any genetic markers could predict nematode host species association, we tested the null model that single-nucleotide polymorphisms (SNPs) in the core genome or flexible genes are randomly associated with respect to nematode host species by using treeWAS on all 53 *X. bovienii* genomes. Only one host, *S. texanum*, showed any significant deviations from the null model, with eight significant genes from the flexible genome. These eight genes were annotated as a putative invasin gene, a colocalised four-gene restriction modification system, a transposase gene, and two hypothetical protein genes. To determine whether this result was sensitive to genes in the global isolates, we repeated the analysis using only the Indiana isolates. Again, only *S. texanum*-associated isolates showed any significant genetic markers (Table S5.2).

5.3.4 Indiana isolates share recent gene flow

Homologous recombination among all the available *X. bovienii* genomes was assessed on the core genome using ClonalFrameML. For the 53 *X. bovienii* genomes, recombination rate (R) was half the mutation rate (θ), such that R/θ = 0.49. Although, recombination (r) had twice the effect on the core genome as mutation (m), as the average length of the recombination fragments was estimated as 200 bp, such that r/m = 2.49. Removing the global reference genomes showed similar results. Among the 42 Indiana isolates,

recombination was more frequent (R/θ = 0.57) but had a similar effect on the genome (r/m = 2.37), as the length of the recombined fragments was slightly smaller (length of recombined fragment = 167 bp). Thus, the relative impact of homologous recombination in *X. bovienii* was similar to that found in other terrestrial gammaproteobacteria(Vos & Didelot, 2009).

To better understand the extent to which gene flow is impacting the evolution of *X. bovienii,* we employed the recombination-based clustering analysis tool PopCOGenT on the 53 *X. bovienii* isolates. In contrast to ClonalFrameML, which identifies recombination in only the core genes, PopCOGenT uses pairwise alignments to test for recent genetic exchange in both the core and flexible regions and then applies network analysis to group isolates that share such exchanges. Four distinct populations with no gene flow between them were identified (Figure 5.6). Three of these populations consisted of only reference genomes. Notably, the fourth population consisted of the 42 Indiana isolates along with the reference genomes of *X. bovienii* strain intermedium (isolated from South Carolina, USA) and *X. bovienii* kraussei Quebec (isolated from Canada). Thus, recent gene flow was found to connect all the Indiana isolates.

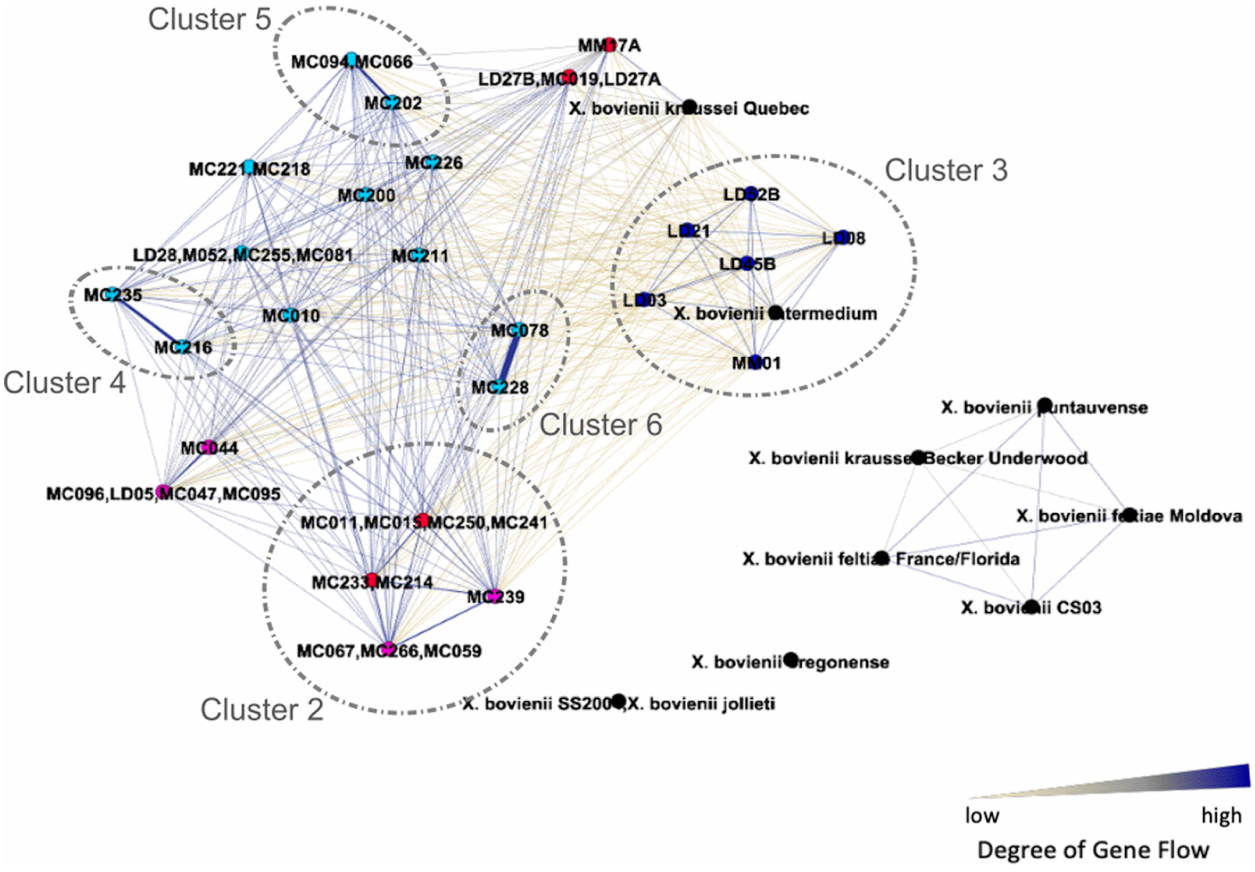


Figure 5. 6: Connectivity among the 53 *X. bovienii* isolates shows that all Indiana isolates shared gene flow with each other and with two of the reference genomes, *X. bovienii* intermedium and *X. bovienii* kraussei Quebec, while the rest of the reference genomes formed three distinct populations with no gene flow among them. Within the Indiana population, six subclusters were identified based on relative gene flow. Nodes represent the genomes and are colour coded based on the nematode host (light blue, *S. affine*; pink, *S. kraussei*; dark blue, *S. intermedium*; and red, *S. texanum*), while edges represent the degree of gene flow, with the lighter/thinner edges having lower gene flow than the darker/thicker edges. Some nodes represent multiple isolates, which are identified as clonal. Furthermore, cluster 1 is not labelled, as it was deemed to be a catchall cluster.

Focusing on the Indiana isolates, we found that some isolates shared more gene flow among them than others, such that six distinct clusters were identified (Figure 5.6). Cluster 1 is the largest cluster, including 19 genomes with isolates from three of the nematode hosts. Cluster 2 includes 10 genomes and is noteworthy as it comprises isolates from two nematode hosts, *S. kraussei* and *S. texanum*. This population cluster is also distinct in the phylogenetic and pangenomic analyses (Figures 5.4A and 5.5, marked with an asterisk in each). Cluster 3 consists of six isolates, which are all the isolates associated with *S. intermedium*. This cluster is also consistent with the grouping observed in phylogenetic and pangenomic analyses, forming its own distinct group (Figures 5.4A and 5.5). Clusters 4, 5, and 6, each consisting of two or three isolates, show the highest levels of gene flow and are from the nematode host *S. affine*.

5.3.5 Differential selection within Indiana population

PopCOGenT identifies genomic regions under selection by finding distinct genetic changes across clusters that show low nucleotide diversity within each cluster, suggesting a recent selective benefit or gene-specific selective sweep(Arevalo et al., 2019). These gene sweeps may provide insights into the traits that are adaptive in this population. For cluster 1, only one gene sweep of 1,940 bp in length (includes two genes annotated as hypothetical protein and acetyltransferase genes) was identified within the core genes, and no flexible gene sweeps were identified. Although this cluster contains the largest number of genomes, the small number of genes identified as possibly under selection suggests that this is a catchall cluster. This cluster reflects gene flow among isolates but does not show selective divergence. On the other end of the spectrum, clusters 4, 5, and 6 include fewer than three genomes each, too few to infer that recently shared genomic regions reflect selection.

In cluster 2, which contains isolates associated with two nematode host species, we identified 37 gene sweeps within the core regions, with a total length of 83.2 kb, and 34 flexible genes being swept (Table 5.1). The genes included encoded several insect toxins, antibiotics, and non-ribosomal peptide synthetases (NRPS), as well as genes conferring resistance and stress tolerance and involved in motility, biosynthesis, and transport. Similar categories of genes were identified as showing evidence of selection in cluster 3, which includes all six isolates from the nematode host S. intermedium. Additionally, genes associated with type VI secretion systems (T6SSs), siderophore (pyochelin) biosynthesis, the Mrx fimbria region, and involved in iron transport were found to be sweeping through this cluster for a total of 40 sweeps with a total length of 117 kb in the core region and 66 flexible gene sweeps (Table 5.2).

Table 5. 1 Summary of gene sweeps across population cluster 2, which includes 10 isolates from nematode hosts *S. kraussei* and *S. texanum*, within the core and flexible genes

Function	**Core gene sweeps (n = 37)**		**Flexible gene sweeps (n = 34)**	
	No. of genes	**Gene products or functions**	**No. of genes**	**Gene product(s) or functions**

Toxin	5	RtxA, Tc	1	Tc
Nonribosomal peptide synthetase	7		1	
Antimicrobial			4	Phenazine, validamycin
Resistance	1	tellurium		
Tolerance	5	DNA repair, damage-inducible protein		
Motility	2	Fimbriae, flagella		
Regulation	5	Transcriptional, translational		
Transport proteins	2	Amino acid		
Catabolic	5	Carbohydrate, fatty acid, aminopeptide, protein	3	Carbohydrate, ATP
Biosynthesis	11	Heme, fatty acid, histidine, molybdopterin, phenylalanine, ornithine, vitamin K2, vitamin B12	1	Vitamin B6
Mobile element			8	Phage, IS
Hypothetical			16	

Table 5. 2: Summary of gene sweeps across population cluster 3, which includes all isolates from the nematode host *S. intermedium*

Function	**Core gene sweeps (n = 40)**		**Flexible gene sweeps (n = 66)**	
	No. of genes	**Gene products or functions**	**No. of genes**	**Gene products or functions**

Toxin	2	RtxA, Tc	1	Hemolysin
Nonribosomal peptide synthetase	3			
Antimicrobial/anti-immune	2	NRPS dependent, membrane/LPS	18	Type VI secretion system
Resistance	8	Multidrug transports, tellurium, bleomycin, streptomycin	2	AMP, phage
Tolerance	3	DNA repair, persistence, damage inducible protein	4	DNA repair, stress response
Motility	2	Fimbriae related, oxygen sensor motility response		
Regulation	2	Transcriptional, translational		
Transport proteins	8	Siderophore, peptide, purine, zinc, potassium	6	Iron, amino acid, pigment
Catabolic	6	Lipase, phenylacetic, arginine, glycolysis, phosphatase	2	glycolysis
Biosynthesis	6	Alkaloid, heme, amino acid, folate	9	Siderophore, peptide
Hypothetical	2		24	

Gene sweeps are by definition unique to each cluster; however, in eight cases, the same region is being differentially selected across clusters (Table S5.3). For instance, a 30-kb region spanning from kilobase 680 to kilobase 720 (Figure 5.7A) encompasses three of these genes (2 NRPS genes and *fnr*). An examination of the corresponding gene trees shows that sometimes additional clusters are segregating at these regions as well. Specifically, an NRPS gene at kilobase 696 separates clusters 2, 3, 4, and 6 into monophyletic groups (Figure 5.7B), while the neighbouring gene (kilobase 701) shows that clusters 2 to 6 are all distinct (Figure 5.7C). Other regions in the core alignment showing differential sweeps contain toxin

genes (*Tc* and *rtxA*), regulatory genes (*hrpA* and *azoR*), a gene conferring tellurium resistance, and a dihydrolipoyl dehydrogenase gene.

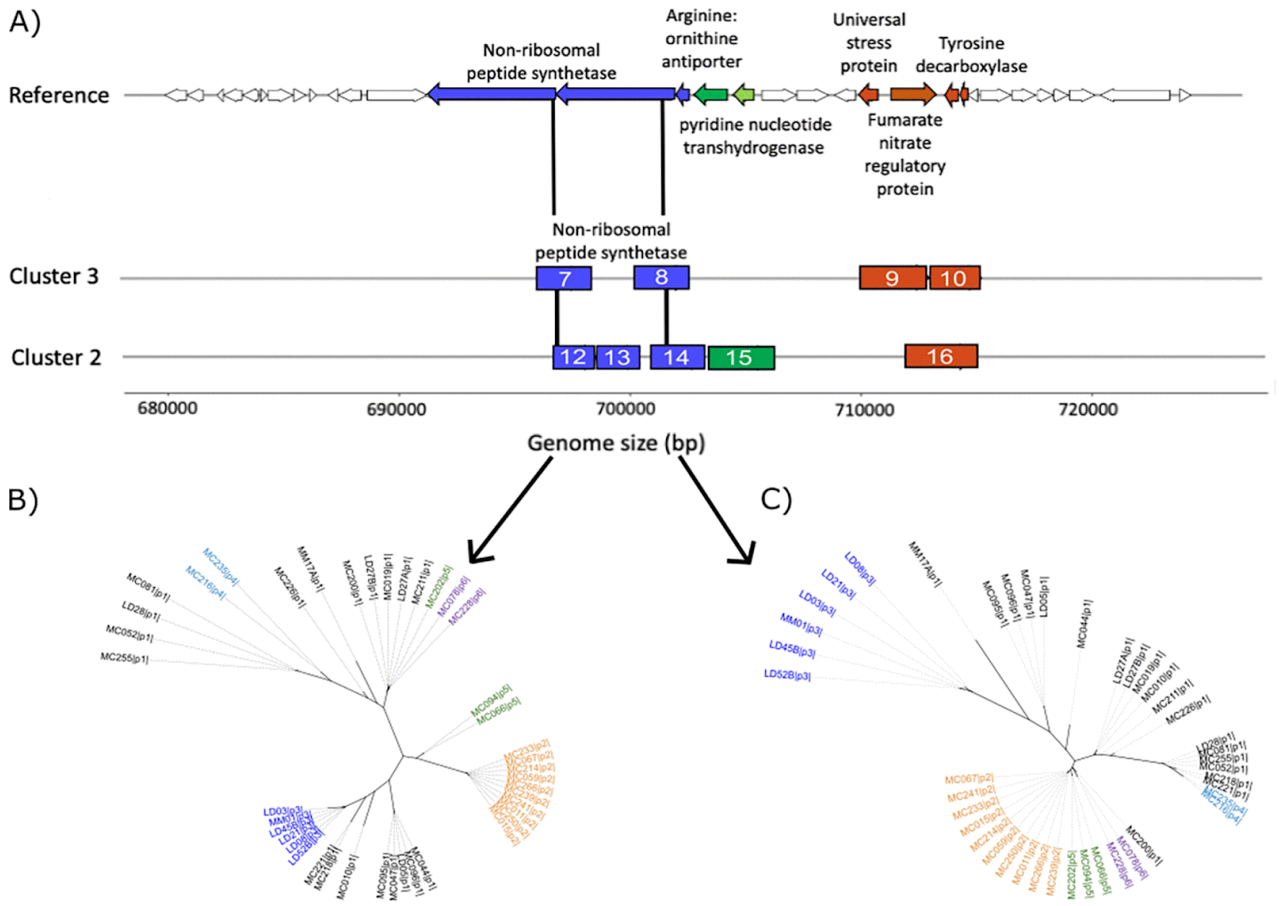


Figure 5. 7: Differential gene sweeps occurring in the same region of the core genome. A) Top, genes between 680 kb to 720 kb in the reference genome represented with block arrows showing their orientation; bottom, sweeps identified in population clusters 2 and 3 are shown as boxes. The numbers in the boxes (gene sweeps) correspond to the gene sweep identification numbers (IDs) provided by PopCOGenT analysis (Table S3). Sweep regions are unique to each cluster and identified by low nucleotide diversity; they can include only part of a gene or several genes. B) Tree of the nonribosomal peptide synthetase region (gene sweep ID 7 in cluster 3 and gene sweep ID 12 in cluster 2) showing differentiation across clusters 2, 3, 4, and 6. C) Tree of another nonribosomal peptide synthetase sweep region (gene sweep ID 8 in cluster 3 and gene sweep ID 14 in cluster 2) showing differentiation across clusters 2 to 6. Across the two trees, cluster 2 is highlighted in orange, cluster 3 in dark blue, cluster 4 in light blue, cluster 5 in green, and cluster 6 in purple.

5.3.6. Spatial proximity and shared nematode host shape population structure

Despite the somewhat loose association between *X. bovienii* and its nematode hosts shown in the above-described analyses, mixed-model analyses of covariance (ANCOVAs) show that isolates share higher genetic similarity ($F_{1,1131} = 30.33$, $P < 0.001$) and estimated gene flow ($F_{1,1131} = 78.20$, $P < 0.001$) if they are associated with the same nematode host species (Figure 5.8, Table S5.5). Moreover, both genetic similarity and gene flow between isolates decline significantly with distance, which ranges from 1 cm to 800 m for isolates collected within the same site, to 28 km across sites within Indiana, and up to 10 Mm with the reference sequences (Figure 5.8, Mantel tests given in Table S5.5). These findings remain significant if the analysis is restricted to all 42 Indiana isolates or just the MC or LD isolates (Table S5.5).

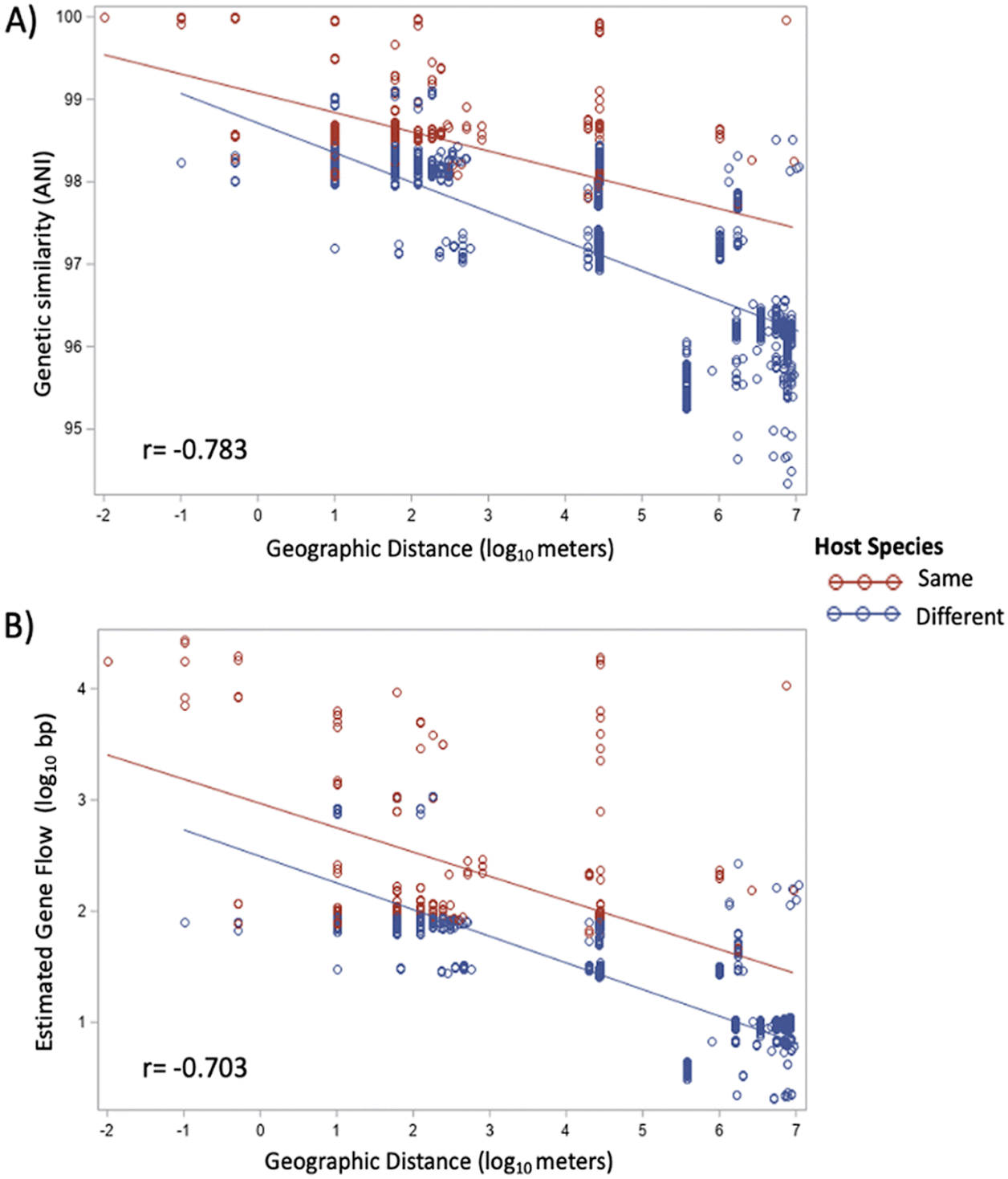


Figure 5. 8: A) Genetic similarity (ANI) and B) gene flow (estimated length bias from PopCoGenT) decrease with geographic distance (in log meters) across all 52 *X. bovienii* genomes. Each point represents a pair of *X. bovienii* isolates and is coded by whether they were isolated from the same nematode host species (red) or from two different species (blue). Correlations across all pairs are shown in the left corner of each graph and were found to be significant with Mantel tests (P = 0.001). Mixed-model analyses of covariance show that both genetic similarity ($F_{1,1131} = 30.33$, $P < 0.0001$) and gene flow ($F_{1,1131} = 78.20$, $P < 0.0001$) are significantly higher if isolates are from the same nematode host species than if they are from different host species.

5.4 Discussion

Microbial symbionts often adapt and specialise to their hosts. And yet, numerous microbial species are characterised as host generalists, able to colonise and thrive in distinct host species. How do generalists evolve through time and space? Here, we examine the population genomics of the mutualist symbiont *X. bovienii f*rom a region where four nematode host species co-occur and compare them to globally available reference genomes. We find that, despite being associated with at least 10 nematode host species across the Northern hemisphere, *X. bovienii* forms a monophyletic group. Regionally, we found two distinct lineages of *X. bovienii.* One lineage was associated exclusively with a single nematode host species, while the other lineage was associated with three other nematode host species. Even though these two lineages were distinct and well-supported, we detected recent gene flow across these lineages and among isolates from all four host species. Nevertheless, gene flow was higher if isolates shared a nematode host species and were collected from closer sites geographically. Thus, *X. bovienii* in this region can be viewed as a metapopulation, with gene flow tying this species together evolutionarily. Moreover, several genes were identified as being targets of differential selection within this population. The diverse functions of these genes, from insect toxins to antimicrobial effectors and resistance mechanisms, speak to the complex biotic environment imposing selection on these symbionts.

Xenorhabdus bacteria are specialised mutualists of nematodes, showing partial co-cladogenesis with their hosts(M.-M. Lee & Stock, 2010; Murfin, Lee, et al., 2015); although this prior work suggested that *X. bovienii* could shift to distinct nematode host species, this conclusion was based on 11 allopatrically collected isolates and so could reflect few rare events. We sampled extensively from a sympatric population and predicted that the population structure of *X. bovienii* strains would mainly reflect their nematode host associations. We found only partial support for this hypothesis. For instance, nematode phylogeny presents *S. affine* and *S. intermedium* as sister taxa, equally distant from the sister taxa *S. kraussei* and *S. texanum* (Figure 5.4B). However, the bacterial phylogeny based on core genes showed that *S. affine*-associated isolates were more closely related to isolates from *S. kraussei* and *S. texanum* than to those from *S. intermedium*. Furthermore, isolates associated with *S. kraussei* and *S. texanum* showed little structuring by nematode host in either the core or accessory genes (Fig. 3A and 4). These findings refute the hypothesis that *S. bovienii* consists of host-limited ecotypes(Sheppard et al., 2018). Instead, they suggest frequent host switching or recombination across isolates.

Based on the core phylogeny (Figure 5.4A), successful host shifts have occurred in lineage 1, which includes isolates from three nematode hosts. For a host shift to occur, lineage 1 bacteria would be carried into an insect by one species of nematode and leave with another; to persist, this novel pairing would have to outcompete the native pairs. In non-competitive laboratory experiments, wherein aposymbiotic nematodes are paired with novel bacteria, *S. affine* nematodes were not able to accept *X. bovienii* bacteria from *S. kraussei* or *S. texanum*. In contrast, *S. kraussei* nematodes could accept *S. affine*-associated *X. bovienii* bacteria, albeit at such a severe fitness cost that the pairing would be unlikely to persist in nature(Dinges et al., 2022b, 2022a). In contrast, *S. kraussei* nematodes were found to accept *S. texanum*-associated *X. bovienii* with no reduction in fitness. These empirical results match the conclusion inferred

from the phylogeny (Figure 5.4) that host shifts across nematode clades occur less frequently than those within. Despite these findings, we found no genes significantly associated with *S. affine* in our genome-wide association study (GWAS) analysis. In fact, we found significant associations for only one nematode host, *S. texanum*. Association mapping in microbes is difficult due to high levels of linkage disequilibrium and population structuring(P. E. Chen & Shapiro, 2015; Collins & Didelot, 2018), and it is possible that treeWAS is overly conservative, as PopCOGenT detected selective sweeps associated with *S. intermedium*. One sweep occurred in the *mrx* fimbria region, which is important in colonisation of the nematode host(Snyder et al., 2011). Additionally, the type VI secretion system genes sweeping in this cluster could be important for interactions with the nematode host(Logan et al., 2018; Murfin, Whooley, et al., 2015). However, within lineage 1, few host-specific markers exist, suggesting that specificity may be due to multiple mechanisms or involve epistatic interactions, and therefore not be picked up in GWAS. In fact, different *X. bovienii* isolates from *S. affine* have shown distinct pathologies on nonnative nematodes(Dinges et al., 2022b; Murfin et al., 2018).

Despite the partial structuring by nematode host species, we found no gene flow discontinuity among our regional isolates (Figure 5.6). In fact, high levels of gene flow were detected across some isolates associated with *S. kraussei* and *S. texanum*. Overall, observed recombination was higher when isolates shared a nematode host species and with geographic proximity (Figure 5.8B), likely reflecting increased opportunities for genetic exchange and shared selective pressures. Each nematode host individual likely harbours a clonal population of *X. bovienii*(Hawlena, Bashey, Mendes-Soares, et al., 2010; Martens et al., 2003); however, to successfully invade and reproduce, several nematodes, which may carry different clones, must coinfect an insect host. Thus, it is within the insect that gene flow is likely to occur as distinct *X. bovienii* strains potentially interact with each other, with other *Xenorhabdus* species, and with the insect microbiome. Most clones were isolated within a few meters of each other, although some were found across study sites and, for one pair of global reference genomes, across continents (Figure 5.8A). This pattern suggests that migration is important to the evolutionary history of *X. bovienii.* In most cases, migration will be local, driven by nematode movements, but longer-range migration could occur via erosion, predation of the insect host, or human agricultural activities. Regardless of the scale, migration has been implicated as a key factor facilitating gene sweeps through recombination(Niehus et al., 2015).

Analysis of selective sweeps in the regional isolates of *X. bovienii* identified several genes (Tables 5.1 and 5.2) that are of known importance for entomopathogens(Niehus et al., 2015). Specifically, nine toxin regions were found to be sweeping within the regional population. Two toxin genes were observed to be sweeping differentially across the clusters (Table S5.3). These sweeps may represent the ability to access additional insect species or to combat insect resistance(I.-H. Kim et al., 2017). Additionally, 11 NRPS regions (3 differentially) were also found to be undergoing selective sweeps. These regions are important in the production of secondary metabolites, some of which are key in competition with the insect microbiota(S. Singh et al., 2015). Additionally, two antibiotic-related genes were found to be sweeping in cluster 2 and several multidrug transports, a type VI secretion system, and a siderophore in cluster 3, further establishing

the dynamic competitive environment faced within the insect, as competition could come by attacking, resisting, or outgrowing a competitor(Murfin et al., 2018). In fact, in cluster 2, which contains isolates from two nematode hosts, several genes were involved in amino acid and vitamin biosynthesis, which could reflect adaptations to better support nematode reproduction that would be beneficial across nematode species. Intriguingly, the successful experimental host shifts performed between *S. kraussei* and *S. texanum*(Dinges et al., 2022b) involved isolates from this population cluster, which leaves open the question of whether the successful host shift was facilitated by these recently shared genes. Future work in this system could examine the adaptive role of the identified sweeps and possible mechanisms of gene flow. Additionally, increased sampling coupled with additional experimental host shifts could help identify the basis of host specificity in this system.

Overall, our work supports the view that gene flow in both the core and flexible genomes is important for maintaining the cohesiveness of *X. bovienii* across multiple nematode hosts. While our data suggest that host switching has occurred, it is less frequent than gene exchange, most likely due to the low fitness of newly associated pairs. This pattern contrasts with that found in the extensively studied *S. aureus*, which shows low levels of recombination in the core genome and frequent host switching, facilitated by acquiring host-specific genes from the host microbiome(Emily J. Richardson et al., 2018). The comparatively low microbial diversity in the insect host, coupled with more intense competition, may limit this pathway for host shifts in *Xenorhabdus*. In contrast, gene flow among coinfecting *Xenorhabdus* bacteria may allow beneficial alleles of genes, such as insect toxins or antimicrobials, to spread in response to local selection pressures. Thus, our results match findings in other systems that show local adaptation despite gene flow(Pérez-Carrascal et al., 2019; S. S. Porter et al., 2017) and differ from work that shows recombination barriers in sympatry(Ellegaard et al., 2013; Sheppard et al., 2014). Importantly, ours is one of only a few studies that examine the population structure and evolutionary history of a host-associated symbiont in a non-agricultural or medical setting, which increasingly enables the complex selective environments faced by microbes to become tangible.

5.5 Conclusions

Multipartite interactions, where bacterial symbionts engage with multiple hosts, pose intriguing questions about how populations are structured and evolve. Yet, we still lack a clear understanding of the evolutionary mechanisms at play. To address this, Chapter 5 focuses on mutualistic symbionts of *Xenorhabdus bovienii* associated with multiple Steinernema nematode species. Building on (Arevalo et al., 2019; VanInsberghe et al., 2020) framing, this study uses whole-genome sequencing and recombination-aware population genomics to assess gene flow and recombination across 42 regional isolates and global references. We uncover that while isolates cluster by nematode host and geography, neither factor completely restricts gene flow, and shared alleles, including insect-toxin and competition genes, move between host-associated lineages. Selective sweeps further highlight adaptation to multipartite environments.

These findings demonstrate that recombination maintains population cohesion even as symbionts shift hosts, and that gene exchange plays a key role in niche adaptation. The framework we present, which combines recombination clustering, host association, and geographic signals, can be applied to other multipartite microbiomes to better understand how gene flow influences community dynamics. Ultimately, this work enhances our understanding of microbiome dynamics by showing how symbiont population structure, adaptability, and ecosystem function emerge from complex inter-species genetic exchange.

5.6 Supplementary Files

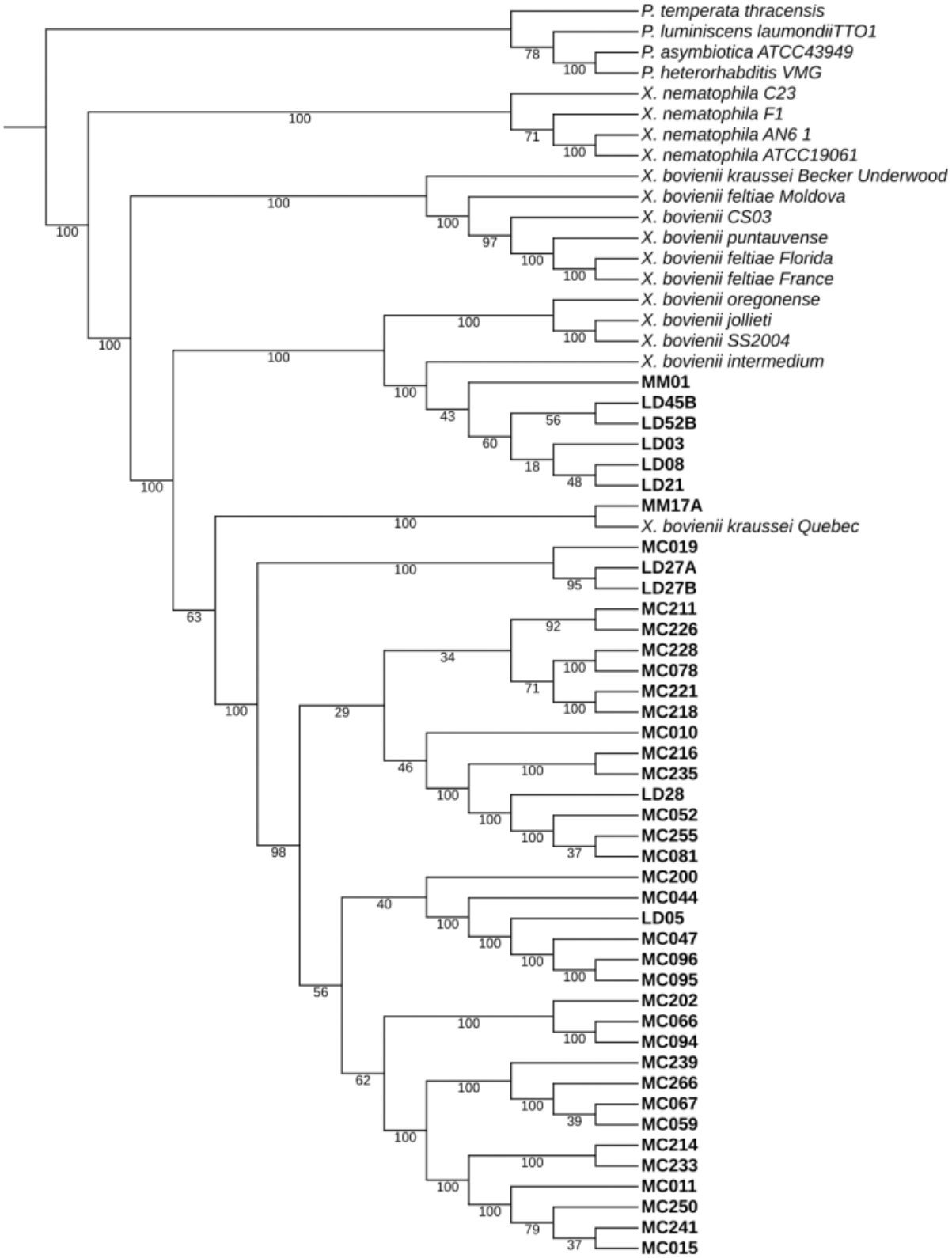


Figure S5.1: Phylogenetic tree based on the core genes identified from all 61 genomes including *Photorhabdus* spp (outgroup), *X. nematophila*, reference *X. bovienii*, and 42 *X. bovienii* Indiana isolates (in bold). The tree was built from an alignment of 273 kb using GTRGAMMA model in RAxML with bootstrapping, and the extended majority rule was used to build the above consensus tree.

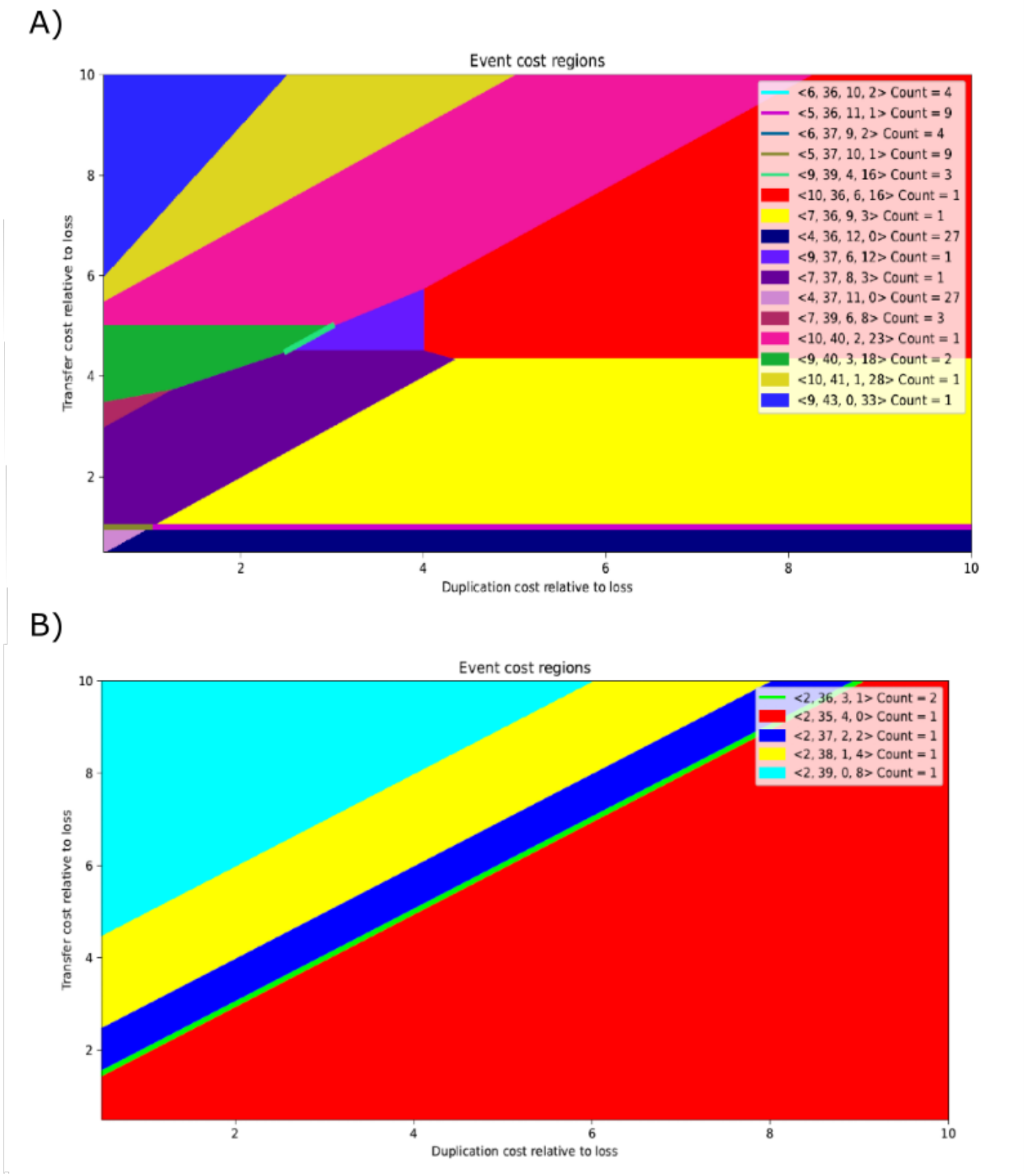


Figure S5.2: Co-phyletic analysis between *X. bovienii* and its nematode hosts. A) Possible maximum parsimony reconciliations between the 53 *X. bovienii* phylogeny and its nematode hosts as estimated by eMPRess. Reconciliations vary depending on the relative costs of different evolutionary events. For each region of the parameter space, the estimated number of co-speciation, duplication, transfer and loss events are given in the inset, with the counts representing the number of distinct mappings giving the same outcome. All but the dark blue region of the parameter space suggest host shifts (transfer events). B) Analysis based on the 42 regional isolates and their four nematode host species. Again, host shifts are predicted over the majority of the parameter space.

Supplementary Tables available with publication at Zenodo: 10.5281/zenodo.17254854.

Table S5.1: Bacterial genomes and nematode genes along with NCBI accession numbers used in this paper.

Table S5.2: Host-specific genetic markers for each of the four nematode hosts, listed by whether the marker is in the core (number of unique SNPs) or flexible (number of unique genes) genome. The first number given is from the analysis of 53 *X. bovienii* isolates, and the second number is from the analysis of only the 42 regional isolates. Markers that were found to be statistically significant using treeWAS ($P < 0.05$) are indicated with an asterisk.

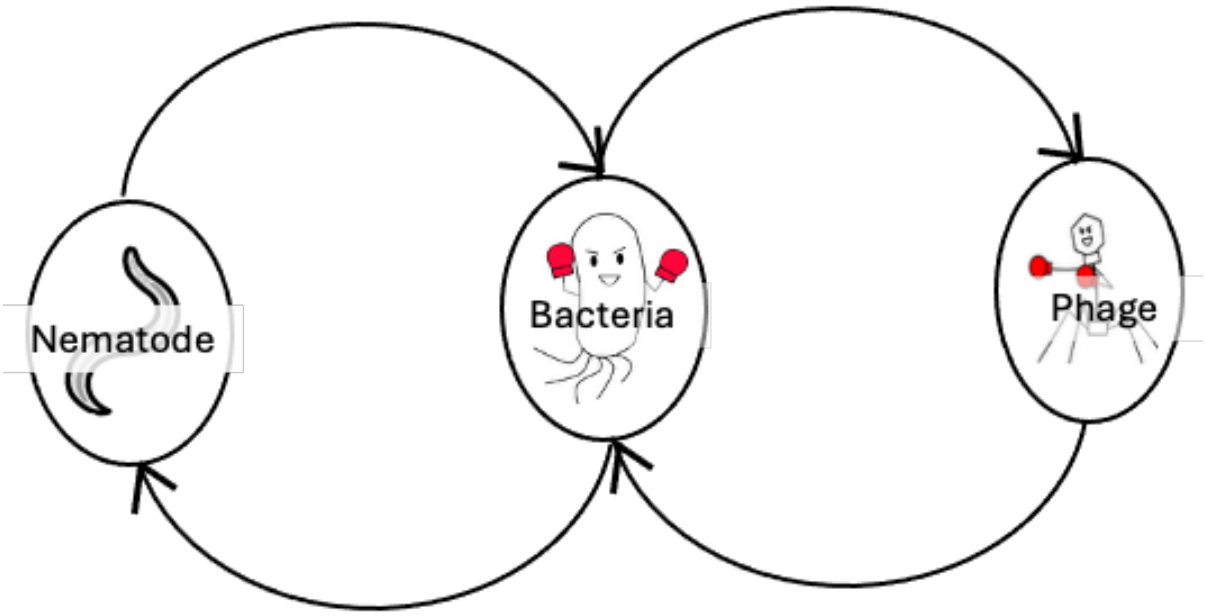
Table S5.3: Gene sweeps that are being swept through two PopCOGenT population clusters - Cluster 2 and 3. Highlighted rows mark the genes that are found in both clusters

Table S5.4: Flexible gene sweeps across clusters 2 and 3 from PopCOGenT clusters.

Table S5.5: Mixed-model analyses of variance for genetic similarity and gene flow as function of nematode host species and geographic distance. Along with Mantel tests of the correlation between genetic similarity and gene flow with distance at three spatial scales.

CHAPTER 6

DISCUSSION



Statement of authorship: Bhavya Papudeshi wrote this chapter with editorial input from supervisor Prof. Robert A. Edwards. This work has not been published and is intended solely for inclusion in this thesis.

Statement on the Use of Generative Artificial Intelligence (AI): Generative AI tools (specifically ChatGPT by OpenAI and Grammarly) were used during the preparation of this chapter for language editing purposes, such as improving sentence clarity, grammar, and structure. These tools were not used to generate original content, perform data analysis, or contribute to reading the papers referenced. All intellectual interpretations and analytical perspectives presented in this thesis are my own, in accordance with Flinders University's policy on the responsible use of generative AI in research.

6.1 Overview

Microbial systems involve dynamic interactions among their members that are pivotal in sustaining ecosystems and influencing the health of living organisms. This thesis explores these microbial interactions by dissecting the genomic mechanisms that underpin microbe–host interactions across phages and bacteria. I begin by focusing on phages, which are understudied in these systems but are of growing importance due to their therapeutic potential in combating antibiotic resistance. In Chapter 2, I review current knowledge of phage biology, best practices for genome annotation and classification, and highlight the need for reproducible and scalable genomic frameworks.

In response to the limited standardisation in phage genomics, I developed Sphae, a scalable and reproducible workflow that supports characterisation of phage genomes (Chapter 3). This toolkit is designed to be easy to download and run, lowering the entry barrier so the broader phage research community can access and use current bioinformatic tools. Recognising that phage therapy related research may represent a major part of the user base, Sphae includes modules to search for marker genes that help distinguish whether a phage may have therapeutic potential. However, the toolkit is not limited to phage therapy applications and can be applied to any phage samples. In the next chapter, using this framework, I characterised phage–host interactions in gut systems, focusing on *Crassvirales* phages infecting *Bacteroides cellulosilyticus*. I identified a conserved tail spike protein under purifying selection, and structural modelling suggests it may interact with TonB-dependent receptors on bacterial surface. Here, I present a genomic marker that likely plays a key role in how abundant gut phages interact and determine their bacterial hosts.

Extending this framework, I then focused on multipartite interactions in the mutualistic symbiont *Xenorhabdus bovienii*, collected from diverse hosts and locations. This work showed that factors such as gene flow and recombination shape population structure and their adaptive potential of these symbionts (Chapter 5). *X. bovienii* continues to exchange genes at a high rate, particularly those acting as insect toxins or antimicrobial activity, likely in response to local selective pressures. Despite forming distinct associations with different nematode species, host barriers do not fully constrain genetic exchange. However, novel host–symbiont pairings may incur fitness costs, suggesting that while host switching is possible, it may be limited by compatibility constraints.

These interactions between phage-bacteria, bacteria-hosts play fundamental roles in determining their host range and broader co-evolutionary dynamics. These results exemplify concepts such as the Red Queen hypothesis, where continuous adaptation is necessary to maintain fitness under biotic pressure. In this thesis, these dynamics were examined through genome-resolved studies of simplified systems, however the findings have broader relevance for understanding microbial community structure. They provide a transferable framework for understanding how evolutionary pressures influence microbial diversity, resilience, and function in complex ecosystems.

6.2 Bioinformatics driven advances in phage therapy

Genome characterisation and classification are essential for understanding phage biology, identifying areas that remain poorly understood, and ultimately leveraging phages therapeutic interventions. Their ability to selectively target pathogenic bacteria while sparing beneficial microbiota makes phages promising alternatives or complements to conventional antimicrobial therapies(Dedrick et al., 2023; M. Kutateladze & Adamia, 2008; Uyttebroek et al., 2022). However, to deem a phage safe for therapeutic use, they must undergo rigorous screening to ensure they exhibit strictly lytic lifecycles, have appropriate host range, and lack any genes associated with antimicrobial resistance (AMR), virulence, or genomic integration potential. These screening criteria reflect our current understanding of what constitutes a safe therapeutic phage, though they may evolve as we learn more about phage biology.

Genome-based screening for therapeutic phages is now a critical step, yet the lack of standardised, reproducible, scalable, and user-friendly tools often hampers this progress. There is a pressing need for integrated computational tools that can assist researchers in quickly triaging and characterising candidate phages. To address this need, I developed Sphae, a modular and scalable workflow designed to streamline the steps to characterise and screen phage genomes (Figure 6.1). This toolkit is easy to install and adaptable for users with varying levels of bioinformatics expertise. *Sphae* supports both large-scale screening efforts and detailed analysis, offering a robust solution to address this gap.

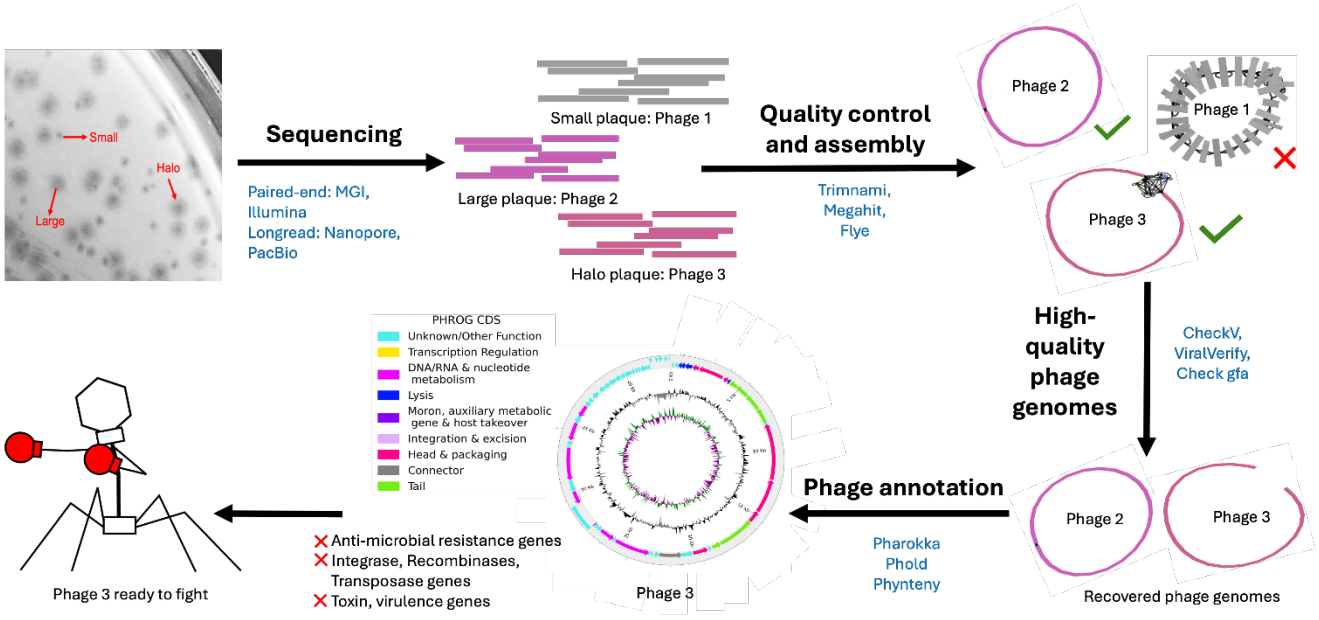


Figure 6. 1: Overview of Sphae workflow developed to characterise and screen phages for therapy

Notably, while using this toolkit, it shed light on additional biological insights that emerged on phage behaviour and risk factors in therapeutic development. One such observation was the co-existence of multiple phages within a single sample. This can occur as prophages within the bacterial host can be induced during phage isolation(Cobián Güemes et al., 2023; Livny et al., 2009). Bacterial isolates frequently contain prophages, which may be excised as a response to being infected by another lytic phage (Dieppa-Colón et al., 2025; McKerral et al., 2023). These induced prophages can therefore be co-purified

and present in the final preparations, increasing the risk of prophage transfer to a patient(Kolenda et al., 2022; Rohde et al., 2018). These findings highlight the importance of genome-based checks to avoid unintended consequences, such as the transfer of lysogenic elements or toxin genes, especially in clinical use. Sphae's ability to flag such elements makes it a key part of this quality control process.

This workflow also uncovered nuanced insights into the phage lifecycle and gene transfer mechanisms that require better understanding. For instance, identifying integrase and transposase genes signals temperate (lysogenic) potential, whereas recombinase genes require deeper contextual analysis. Recombinases can participate in DNA repair and resolve concatemers formed during genome replication in strictly lytic phages(Bobay et al., 2013). This means the presence of recombinases alone does not equate to lysogeny, and conversely, the absence of integrase genes does not guarantee a strictly lytic lifecycle(Altamirano & Barr, 2021). By flagging such genomic features, Sphae prompts for deeper analysis of the phage, either through further experiments or closer look at the genetic potential to determine if the phage is a suitable candidate. As novel mechanisms of lysogeny and horizontal gene transfer (HGT) in phages are identified, bioinformatic tools and checks can be quickly integrated into these workflows.

The use of workflow managers such as Snakemake(Köster & Rahmann, 2012) allows for a modular design that facilitates the integration of future bioinformatic tools and enriched genomic databases. Since the original publication of this workflow in Jan 2025, Sphae continues to be updated and maintained to include new version of tools, addition of phylogenetic module, improved database handling and updating the output summary to add new characteristics. Collectively, these updates have significantly improved the usability, reproducibility, and analytical depth of the workflow, aligning it more closely with the evolving needs of phage genomics research.

This toolkit supports reproducible decision-making in phage therapy. However, looking ahead, two other bottlenecks remain in the field. First, the establishment of phage banks containing readily accessible phages which is essential for enabling rapid treatment of multidrug-resistant infections. Second, a need to reduce the time and resources required for experimental validation by prioritising the most promising phage candidates. This can be achieved by bioinformatically predicting a subset of phages within the phage banks likely to infect a given pathogen, thereby narrowing the experimental focus. However, such predictive capabilities depend on robust methods for inferring host–phage interactions, highlighting a critical area for development.

6.3 Uncovering genomic mechanisms of phage-host interactions in the gut ecosystem

Phages are the most abundant and dynamic biological entities on the planet, and we know they play a role in modulating bacterial populations. Yet, despite their importance, the molecular mechanisms underpinning phage–bacteria interactions are often obscured, particularly in complex environments such as the human gut. *Bacteroides,* is a common gut microbe, with evidence of interacting with the human immune system, and playing a role in maintaining human health(Pargin et al., 2023; Shin et al., 2024). *Crassvirales* and their *Bacteroides* hosts form persistent infections, facilitated by bacterial phase-variable capsule switching,

which enables them to evade phages by altering capsular structures(Cortés-Martín et al., 2025; N. T. Porter et al., 2020; Shkoporov, Khokhlova, et al., 2021). Yet, how these phages recognise and persist with such dynamically changing hosts remains largely elusive.

In this thesis, we identified conserved tail spike protein undergoing purifying selection across the *Crassvirales* isolates. These tail spikes are known to serve as specific receptors that reversibly bind to bacterial cell receptors (Nobrega et al., 2018). Other studies have shown that these host recognition modules can be swapped to infect new hosts(Dunne et al., 2019; Latka et al., 2021), enabling rapid adaptation to new hosts in response to selective pressures. Furthermore, phylogenetic analysis revealed an absence of strict co-evolution between our *Crassvirales* isolates and their hosts. This suggests that host specificity may evolve independently of long-term host–phage pairings, potentially through modular protein innovation in tail fibre proteins or through selective sweeps that periodically favour variants with altered host ranges.

Further, in the paper we used protein docking to predict a potential interaction between the phage tail spike protein and the TonB-dependent receptor on the bacterial surface. These receptors are typically used by *Bacteroides* to take up starches (Pollet et al., 2021), and have been associated with phage sensitivity(Cortés-Martín et al., 2025; N. T. Porter et al., 2020; Shkoporov, Khokhlova, et al., 2021), offering a mechanistic hypothesis for host recognition While docking offers valuable insights into potential interaction interfaces, it is important to note that current docking algorithms have known limitations, including challenges with accurately scoring binding affinities(Shirali et al., 2025), inability to handle flexible protein regions (Harmalkar & Gray, 2020), and incorrect structural assignments. These constraints are active areas of research, and advances in deep learning–based structure prediction and integrative modelling are rapidly improving the field. Thus, while our results represent a computational hypothesis, they lay the groundwork for future experimental validation and more accurate modelling as these tools evolve.

The evidence presented here supports a model in which phages maintain evolutionary fitness not through static co-evolutionary relationships, but through flexible, modular innovations in key interaction proteins that enable rapid host switching or resistance evasion. These mechanistic insights can contribute to informing phage-host pairings to select potential phage therapy candidates quickly. By understanding which proteins are under selective constraint, and how they interact with host surface structures, we can better predict host range and therapeutic efficacy without relying solely on trial-and-error plating methods. These insights can also lay a foundation for future work exploring the ecological consequences of phage–bacteria dynamics across diverse microbiomes.

6.4 Population structuring of symbiotic bacteria

In complex and diverse microbial communities, we still lack a clear understanding of how bacterial populations are structured. To investigate this on a more manageable scale, we focused on bacterial symbionts involved in multipartite interactions, examining how their genomes reflect the evolutionary relationships with their eukaryotic hosts. Specifically, we studied *Xenorhabdus bovienii*, a mutualistic

symbiont of *Steinernema* nematodes, and a virulent insect pathogen, dissecting how biological and environmental factors, such as host specificity and spatial proximity shape bacterial population structure at fine resolution.

Population genomics analysis of *Xenorhabdus bovienii* revealed a model of partial but permeable population structure. While host species and geography contributed to population differentiation, these factors did not act as strict barriers to gene flow. Instead, *X. bovienii* populations remained genomically cohesive through frequent recombination and occasional host switching, supporting a model of generalist with flexible ecological boundaries. Recombination was particularly prominent in genes encoding insect toxins, antimicrobial effectors, and resistance mechanisms—suggesting strong selection pressure within the insect host. This points to the insect cadaver not only as a site of infection but also as a key ecological arena for inter-strain interaction and gene exchange. Using PopCOGenT, we also detected selective sweeps acting on ecologically relevant loci, such as secretion systems and multidrug transporters, underscoring the importance of localised adaptation. Importantly, while some host-specific structuring was evident, it did not align neatly with nematode phylogeny, indicating that host specificity in symbionts may be polygenic or context-dependent. These findings challenge models of strict co-divergence and instead support a more dynamic framework in which recombination, competition, and local selection jointly shape bacterial population structure in multipartite systems.

This dynamic model of population structure and gene flow in *X. bovienii* symbionts offers a useful parallel to our findings in *Crassvirales* phages in the human gut. In both systems, host association plays a role in shaping microbial diversity but does not impose strict evolutionary constraints. *Crassvirales* phages, like *X. bovienii*, persist in environments with strong selective pressures—mediated by host immune systems, microbial competition, and frequent co-infections. In both cases, we observed the conservation and apparent modularity of host interaction genes (e.g. tail spikes in *Crassvirales*; toxins and secretion systems in *X. bovienii*), which may facilitate host shifts and adaptive radiation without requiring long-term co-evolution. Methodologically, both studies leveraged comparative genomics and phylogenetics to uncover gene flow patterns and detect signatures of selection. Together, these findings argue for a broader ecological and evolutionary framework where recombination, modular gene architecture, and niche-driven selection, rather than strict phylogenetic coupling, govern the population structure of both bacteria and their phages.

In this thesis, I present a reproducible workflow to characterise phages, followed by investigations into how phages identify their bacterial hosts within gut ecosystems, and how symbiotic *Xenorhabdus bovienii* populations act as cohesive units that share genes despite host and spatial constraints. Together, these studies demonstrate how population genomics can uncover the genetic and evolutionary forces shaping microbial life and provide tools and conceptual frameworks to extend these insights into complex, real-world microbiomes.

6.5 Broader context and implications

The findings presented in this thesis have broad implications spanning both applied and theoretical microbiology. First and foremost, this work informs the development of phage therapy and microbiome interventions by providing a reproducible workflow (*Sphae*) that enables scalable, genome-based characterisation and screening of phage candidates. Through identifying conserved host-interaction markers and integrating risk assessment tools, this research advances the field toward more predictive, mechanism-driven approaches to therapeutic phage selection. In parallel, this thesis contributes to microbial population genomics frameworks by showing that recombination, modular gene exchange, and local adaptation—not just phylogenetic lineage or host specificity—underpin the structure of both bacterial and phage populations. These insights challenge conventional models of co-divergence and support a more dynamic understanding of microbial evolution. Additionally, the work highlights horizontal gene transfer as a central adaptive mechanism in symbiotic and competitive microbial contexts, particularly in loci related to host interaction, antimicrobial production, and resistance. By coupling bioinformatics with evolutionary and ecological theory, this thesis integrates mechanistic and ecological perspectives, demonstrating how molecular processes translate into community-level patterns. Finally, the tools and conceptual models developed here lay a foundation for future systems-level microbiome studies, enabling deeper insights into microbial resilience, therapeutic design, and ecosystem function in increasingly complex environments. While this thesis often frames findings in the context of phage therapy, the broader implications extend into microbial ecology and evolutionary biology. The concepts of recombination, modularity, and niche-driven selection are equally important for understanding how microbial communities maintain resilience, adapt to selective pressures, and interact with their hosts across gut and symbiotic systems.

6.6 Limitations and Future Directions

While this thesis advances our understanding of host–microbe interactions and microbial genomic dynamics, several limitations remain that present opportunities for future research. One key limitation lies in the reliance on DNA sequence data alone to predict biological function and ecological roles. There are critical aspects of phage biology, such as epigenetic DNA modifications or protein modifications that play a role in defence mechanisms(Birkholz et al., 2022; Longin et al., 2024; Mayo-Muñoz et al., 2024). These modifications play a key role in evading bacterial restriction systems and represent an important feature of phage biology. Similarly, in phages novel strategies like stop codon reassignment, allowing phages to evade bacterial defence mechanisms(Borges et al., 2021), and hyper modification that includes biochemical alterations to phage DNA, protecting it from bacterial restriction-modification systems, thus facilitating successful phage infection and replication(Hutinet et al., 2021) as key adaptive mechanisms. These remain largely inaccessible without integrated multi-omics or experimental validation and highlight the need for expanding beyond genomics alone.

Even with the current use of state-of-the-art phage annotation tools, phage annotations remain limited by the high proportion of genes categorised as "hypothetical proteins." Despite integrating structural and synteny-based annotations in Sphae, many genes remain uncharacterised. Continued database expansion

and integration of protein language models, machine learning, and functional screens will be vital to improving annotation accuracy. Similarly, phage packaging mechanisms are crucial as they directly influence horizontal gene transfer and therefore need to also be considered.

Phage packaging mechanisms are another underexplored area with significant implications. The choice between cos-site, headful, and other packaging strategies directly influences the likelihood of horizontal gene transfer (HGT), including unintended mobilization of host DNA. Integrating packaging mechanism prediction into bioinformatic workflows will allow more accurate risk assessments for phage therapy candidates.

Another major limitation in phage genomics is the vast proportion of “viral dark matter,” where the majority of predicted genes have no assigned function. Despite the use of structural modelling and comparative genomics, linking these genes to ecological roles or molecular mechanisms remains a challenge. Addressing this will require integrative approaches combining genomics with transcriptomics, proteomics, and experimental validation. These efforts will be essential to truly understand how phages shape microbial ecosystems, particularly within complex environments such as the gut or multipartite symbioses.

6.7 Conclusions

This thesis makes key contributions to microbial genomics, evolutionary biology, and bioinformatics by addressing foundational questions about how microbial populations are structured and evolve in the context of host–phage–microbe interactions. Through fine-scale genomic analyses of individual bacterial and phage isolates, it reveals the evolutionary strategies—such as generalist adaptability, recombination, and toxin evolution—that enable microbial resilience across dynamic environments. Importantly, this work also includes the development of Sphae, a reproducible and scalable toolkit designed to characterise phage genomes and support the safe application of phage therapy. By integrating evolutionary insight with computational tool development, this thesis bridges a critical gap between microbial ecology and applied microbiome research. The findings have broader implications for managing microbial ecosystems, guiding phage therapy design, and mitigating antimicrobial resistance, while also providing a framework for future research across complex, host-associated microbiomes. Ultimately, while genomics offers a powerful framework for investigating microbial interactions and phage biology, it represents only one layer of understanding. The integration of functional assays and multi-omics will be crucial to translate genomic predictions into ecological and therapeutic outcomes.

CHAPTER 7
BIBLIOGRAPHY

Adriaenssens, E., & Brister, J. R. (2017). How to Name and Classify Your Phage: An Informal Guide. *Viruses*, *9*(4). https://doi.org/10.3390/v9040070

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. https://doi.org/10.1093/nar/gks406

Alcock, B. P., Huynh, W., Chalil, R., Smith, K. W., Raphenya, A. R., Wlodarski, M. A., Edalatmand, A., Petkau, A., Syed, S. A., Tsang, K. K., Baker, S. J. C., Dave, M., McCarthy, M. C., Mukiri, K. M., Nasir, J. A., Golbon, B., Imtiaz, H., Jiang, X., Kaur, K., … McArthur, A. G. (2023). CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Research*, *51*(D1), D690–D699. https://doi.org/10.1093/nar/gkac920

Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A.-L. V., Cheng, A. A., Liu, S., Min, S. Y., Miroshnichenko, A., Tran, H.-K., Werfalli, R. E., Nasir, J. A., Oloni, M., Speicher, D. J., Florescu, A., Singh, B., … McArthur, A. G. (2020). CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Research*, *48*(D1), D517–D525. https://doi.org/10.1093/nar/gkz935

Altamirano, F. L. G., & Barr, J. J. (2021). Screening for Lysogen Activity in Therapeutically Relevant Bacteriophages. *Bio-Protocol*, *11*(8), e3997. https://doi.org/10.21769/BioProtoc.3997

Angiuoli, S. V., & Salzberg, S. L. (2011). Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* , *27*(3), 334–342. https://doi.org/10.1093/bioinformatics/btq665

Anthenelli, M., Jasien, E., Edwards, R., Bailey, B., Felts, B., Katira, P., Nulton, J., Salamon, P., Rohwer, F., Silveira, C. B., & Luque, A. (2020). Phage and bacteria diversification through a prophage acquisition ratchet. In *bioRxiv* (p. 2020.04.08.028340). https://doi.org/10.1101/2020.04.08.028340

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

Antipov, D., Rayko, M., Kolmogorov, M., & Pevzner, P. A. (2022). viralFlye: assembling viruses and identifying their hosts from long-read metagenomics data. *Genome Biology*, *23*(1), 57. https://doi.org/10.1186/s13059-021-02566-x

Arevalo, P., VanInsberghe, D., Elsherbini, J., Gore, J., & Polz, M. F. (2019). A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations. *Cell*, *178*(4), 820-834.e14. https://doi.org/10.1016/j.cell.2019.06.033

Arumugam, K., Bağcı, C., Bessarab, I., Beier, S., Buchfink, B., Górska, A., Qiu, G., Huson, D. H., & Williams, R. B. H. (2019). Annotated bacterial chromosomes from frame-shift-corrected long-read metagenomic data. *Microbiome*, *7*(1), 61. https://doi.org/10.1186/s40168-019-0665-y

Asare, P. T., Jeong, T.-Y., Ryu, S., Klumpp, J., Loessner, M. J., Merrill, B. D., & Kim, K.-P. (2015). Putative type 1 thymidylate synthase and dihydrofolate reductase as signature genes of a novel Bastille-like group of phages in the subfamily Spounavirinae. *BMC Genomics*, *16*(1), 582. https://doi.org/10.1186/s12864-015-1757-0

Azam, A. H., & Tanji, Y. (2019). Bacteriophage-host arm race: an update on the mechanism of phage resistance in bacteria and revenge of the phage with the perspective for phage therapy. *Applied Microbiology and Biotechnology*, *103*(5), 2121–2131. https://doi.org/10.1007/s00253-019-09629-x

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Barr, J. J., Auro, R., Furlan, M., Whiteson, K. L., Erb, M. L., Pogliano, J., Stotland, A., Wolkowicz, R., Cutting, A. S., Doran, K. S., Salamon, P., Youle, M., & Rohwer, F. (2013). Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(26), 10771–10776. https://doi.org/10.1073/pnas.1305923110

Bayfield, O. W., Shkoporov, A. N., Yutin, N., Khokhlova, E. V., Smith, J. L. R., Hawkins, D. E. D. P., Koonin, E. V., Hill, C., & Antson, A. A. (2023). Structural atlas of a human gut crassvirus. *Nature*, *617*(7960), 409–416. https://doi.org/10.1038/s41586-023-06019-2

Benler, S., Cobián-Güemes, A. G., McNair, K., Hung, S.-H., Levi, K., Edwards, R., & Rohwer, F. (2018). A diversity-generating retroelement encoded by a globally ubiquitous Bacteroides phage. *Microbiome*, *6*(1), 191. https://doi.org/10.1186/s40168-018-0573-6

Benler, S., & Koonin, E. V. (2020). Phage lysis-lysogeny switches and programmed cell death: Danse macabre. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *42*(12), e2000114. https://doi.org/10.1002/bies.202000114

Benler, S., & Koonin, E. V. (2021). Fishing for phages in metagenomes: what do we catch, what do we miss? *Current Opinion in Virology*, *49*, 142–150. https://doi.org/10.1016/j.coviro.2021.05.008

Benler, S., Yutin, N., Antipov, D., Rayko, M., Shmakov, S., Gussow, A. B., Pevzner, P., & Koonin, E. V. (2021). Thousands of previously unknown phages discovered in whole-community human gut metagenomes. *Microbiome*, *9*(1), 78. https://doi.org/10.1186/s40168-021-01017-w

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M.-C. C., Charles, T., Chen, X., Cocolin, L., Eversole, K., Corral, G. H., Kazou, M., Kinkel, L., Lange, L., Lima, N., Loy, A., Macklin, J. A., Maguin, E., Mauchline, T., McClure, R., … Schloter, M. (2020). Correction to: Microbiome definition re-visited: old concepts and new challenges. *Microbiome*, *8*(1), 119. https://doi.org/10.1186/s40168-020-00905-x

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

Birkholz, N., Jackson, S. A., Fagerlund, R. D., & Fineran, P. C. (2022). A mobile restriction-modification system provides phage defence and resolves an epigenetic conflict with an antagonistic endonuclease. *Nucleic Acids Research*, *50*(6), 3348–3361. https://doi.org/10.1093/nar/gkac147

Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., & Hugenholtz, P. (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, *8*, 209. https://doi.org/10.1186/1471-2105-8-209

Bobay, L.-M., Touchon, M., & Rocha, E. P. C. (2013). Manipulating or superseding host recombination functions: a dilemma that shapes phage evolvability. *PLoS Genetics*, *9*(9), e1003825. https://doi.org/10.1371/journal.pgen.1003825

Boeckaerts, D., Stock, M., Criel, B., Gerstmans, H., De Baets, B., & Briers, Y. (2021). Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. *Scientific Reports*, *11*(1), 1467. https://doi.org/10.1038/s41598-021-81063-4

Boeckaerts, D., Stock, M., De Baets, B., & Briers, Y. (2022). Identification of Phage Receptor-Binding Protein Sequences with Hidden Markov Models and an Extreme Gradient Boosting Classifier. *Viruses*, *14*(6). https://doi.org/10.3390/v14061329

Boeckaerts, D., Stock, M., Ferriol-González, C., Oteo-Iglesias, J., Sanjuán, R., Domingo-Calap, P., De Baets, B., & Briers, Y. (2024). Prediction of Klebsiella phage-host specificity at the strain level. *Nature Communications*, *15*(1), 4355. https://doi.org/10.1038/s41467-024-48675-6

Bolduc, B., Jang, H. B., Doulcier, G., You, Z.-Q., Roux, S., & Sullivan, M. B. (2017). vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ*, *5*(e3243), e3243. https://doi.org/10.7717/peerj.3243

Bondy-Denomy, J., & Davidson, A. R. (2014). When a virus is not a parasite: the beneficial effects of prophages on bacterial fitness. *The Journal of Microbiology*, *52*(3), 235–242. https://doi.org/10.1007/s12275-014-4083-3

Bondy-Denomy, J., Qian, J., Westra, E. R., Buckling, A., Guttman, D. S., Davidson, A. R., & Maxwell, K. L. (2016). Prophages mediate defense against phage infection through diverse mechanisms. *The ISME Journal*, *10*(12), 2854–2866. https://doi.org/10.1038/ismej.2016.79

Bongrand, C., Koch, E. J., Moriano-Gutierrez, S., Cordero, O. X., McFall-Ngai, M., Polz, M. F., & Ruby, E. G. (2016). A genomic comparison of 13 symbiotic Vibrio fischeri isolates from the perspective of their host source and colonization behavior. *The ISME Journal*, *10*(12), 2907–2917. https://doi.org/10.1038/ismej.2016.69

Borges, A. L., Lou, Y. C., Sachdeva, R., Al-Shayeb, B., Jaffe, A. L., Lei, S., Santini, J. M., & Banfield, J. F. (2021). Stop codon recoding is widespread in diverse phage lineages and has the potential to regulate translation of late stage and lytic genes. In *bioRxiv* (p. 2021.08.26.457843). https://doi.org/10.1101/2021.08.26.457843

Borodovich, T., Shkoporov, A. N., Ross, R. P., & Hill, C. (2022). Phage-mediated horizontal gene transfer and its implications for the human gut microbiome. *The Gastroenterology Report*, *10*, goac012. https://doi.org/10.1093/gastro/goac012

Botelho, J., Cazares, A., & Schulenburg, H. (2023). The ESKAPE mobilome contributes to the spread of antimicrobial resistance and CRISPR-mediated conflict between mobile genetic elements. *Nucleic Acids Research*, *51*(1), 236–252. https://doi.org/10.1093/nar/gkac1220

Botstein, D. (1980). A theory of modular evolution for bacteriophages. *Annals of the New York Academy of Sciences*, *354*(1), 484–490. https://doi.org/10.1111/j.1749-6632.1980.tb27987.x

Bouras, G., Grigson, S. R., Papudeshi, B., Mallawaarachchi, V., & Roach, M. J. (2024). Dnaapler: A tool to reorient circular microbial genomes. *Journal of Open Source Software*, *9*(93), 5968. https://doi.org/10.21105/joss.05968

Bouras, G., Houtak, G., Wick, R. R., Mallawaarachchi, V., Roach, M. J., Papudeshi, B., Judd, L. M., Sheppard, A. E., Edwards, R. A., & Vreugde, S. (2024). Hybracter: enabling scalable, automated, complete and accurate bacterial genome assemblies. *Microbial Genomics*, *10*(5). https://doi.org/10.1099/mgen.0.001244

Bouras, G., Nepal, R., Houtak, G., Psaltis, A. J., Wormald, P.-J., & Vreugde, S. (2023). Pharokka: a fast scalable bacteriophage annotation tool. *Bioinformatics* , *39*(1). https://doi.org/10.1093/bioinformatics/btac776

Bouras, G., Sheppard, A. E., Mallawaarachchi, V., & Vreugde, S. (2023). Plassembler: an automated bacterial plasmid assembly tool. *Bioinformatics (Oxford, England)*, *39*(7). https://doi.org/10.1093/bioinformatics/btad409

Brockhurst, M. A., Chapman, T., King, K. C., Mank, J. E., Paterson, S., & Hurst, G. D. D. (2014). Running with the Red Queen: the role of biotic conflicts in evolution. *Proceedings. Biological Sciences*, *281*(1797), 20141382. https://doi.org/10.1098/rspb.2014.1382

Bruce, T., De Wardt, R., Papudeshi, B., Robinson, B., Cuevas, D. A., Aguinaldo, K., Lopez, T., Mora, M. F., Morris, M. M., Mcnair, K., Peterson, M., Skye, A., Turnland, A. C., Cavallia, D., Haggerty, M. H., Kumudika Susinghe, A., Edwards, R. A., & Dinsdale, E. A. (n.d.). *The intersection of genotype and phenotype: Informing ecotype development*.

Bryson, A. L., Hwang, Y., Sherrill-Mix, S., Wu, G. D., Lewis, J. D., Black, L., Clark, T. A., & Bushman, F. D. (2015). Covalent modification of bacteriophage T4 DNA inhibits CRISPR-Cas9. *MBio*, *6*(3), e00648. https://doi.org/10.1128/mBio.00648-15

Cadillo-Quiroz, H., Didelot, X., Held, N. L., Herrera, A., Darling, A., Reno, M. L., Krause, D. J., & Whitaker, R. J. (2012). Patterns of gene flow define species of thermophilic Archaea. *PLoS Biology*, *10*(2), e1001265. https://doi.org/10.1371/journal.pbio.1001265

Caldera, E. J., Chevrette, M. G., McDonald, B. R., & Currie, C. R. (2019). Local Adaptation of Bacterial Symbionts within a Geographic Mosaic of Antibiotic Coevolution. *Applied and Environmental Microbiology*, *85*(24). https://doi.org/10.1128/AEM.01580-19

Calero-Cáceres, W., Ye, M., & Balcázar, J. L. (2019). Bacteriophages as Environmental Reservoirs of Antibiotic Resistance. *Trends in Microbiology*, *27*(7), 570–577. https://doi.org/10.1016/j.tim.2019.02.008

Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D., & Lawley, T. D. (2021). Massive expansion of human gut bacteriophage diversity. *Cell*, *184*(4), 1098-1109.e9. https://doi.org/10.1016/j.cell.2021.01.029

Campbell, J. H., O'Donoghue, P., Campbell, A. G., Schwientek, P., Sczyrba, A., Woyke, T., Söll, D., & Podar, M. (2013). UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(14), 5540–5545. https://doi.org/10.1073/pnas.1303090110

Campbell, K. M., Kouris, A., England, W., Anderson, R. E., McCleskey, R. B., Nordstrom, D. K., & Whitaker, R. J. (2017). Sulfolobus islandicus meta-populations in Yellowstone National Park hot springs. *Environmental Microbiology*, *19*(6), 2334–2347. https://doi.org/10.1111/1462-2920.13728

Cantu, V. A., Sadural, J., & Edwards, R. (2019). *PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets* (e27553v1). PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.27553v1

Cantu, V. A., Salamon, P., Seguritan, V., Redfield, J., Salamon, D., Edwards, R. A., & Segall, A. M. (2020). PhANNs, a fast and accurate tool and web server to classify phage structural proteins. In *bioRxiv*. bioRxiv. https://doi.org/10.1101/2020.04.03.023523

Casjens, S. R., & Gilcrease, E. B. (2009). Determining DNA Packaging Strategy by Analysis of the Termini of the Chromosomes in Tailed-Bacteriophage Virions. In M. R. J. Clokie & A. M. Kropinski (Eds.), *Bacteriophages: Methods and Protocols, Volume 2 Molecular and Applied Aspects* (pp. 91–111). Humana Press. https://doi.org/10.1007/978-1-60327-565-1_7

Catalano, C. E., & Morais, M. C. (2021). Viral genome packaging machines: Structure and enzymology. *The Enzymes*, *50*, 369–413. https://doi.org/10.1016/bs.enz.2021.09.006

Chan, P. P., Lin, B. Y., Mak, A. J., & Lowe, T. M. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research*, *49*(16), 9077–9096. https://doi.org/10.1093/nar/gkab688

Chan, P. P., & Lowe, T. M. (2019). tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods in Molecular Biology* , *1962*, 1–14. https://doi.org/10.1007/978-1-4939-9173-0_1

Chapuis, E., Emelianoff, V., Paulmier, V., Le Brun, N., Pagès, S., Sicard, M., & Ferdy, J.-B. (2009). Manifold aspects of specificity in a nematode-bacterium mutualism. *Journal of Evolutionary Biology*, *22*(10), 2104–2117. https://doi.org/10.1111/j.1420-9101.2009.01829.x

Charoenkwan, P., Kanthawong, S., Schaduangrat, N., Yana, J., & Shoombuatong, W. (2020). PVPred-SCM: Improved Prediction and Analysis of Phage Virion Proteins Using a Scoring Card Method. *Cells* , *9*(2). https://doi.org/10.3390/cells9020353

Chase, A. B., Arevalo, P., Brodie, E. L., Polz, M. F., Karaoz, U., & Martiny, J. B. H. (2019). Maintenance of Sympatric and Allopatric Populations in Free-Living Terrestrial Bacteria. *MBio*, *10*(5). https://doi.org/10.1128/mBio.02361-19

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., & Jin, Q. (2005). VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Research*, *33*(Database issue), D325-8. https://doi.org/10.1093/nar/gki008

Chen, L., Zheng, D., Liu, B., Yang, J., & Jin, Q. (2016). VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. *Nucleic Acids Research*, *44*(D1), D694-7. https://doi.org/10.1093/nar/gkv1239

Chen, P. E., & Shapiro, B. J. (2015). The advent of genome-wide association studies for bacteria. *Current Opinion in Microbiology*, *25*, 17–24. https://doi.org/10.1016/j.mib.2015.03.002

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* , *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Cherwa, J. E., & Fane, B. A. (2011). Microviridae: Microviruses and Gokushoviruses. In *eLS*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470015902.a0000781.pub2

Chevallereau, A., Pons, B. J., van Houte, S., & Westra, E. R. (2022). Interactions between bacterial and phage communities in natural environments. *Nature Reviews. Microbiology*, *20*(1), 49–62. https://doi.org/10.1038/s41579-021-00602-y

Cho, J. C., & Tiedje, J. M. (2000). Biogeography and degree of endemicity of fluorescent Pseudomonas strains in soil. *Applied and Environmental Microbiology*, *66*(12), 5448–5456. https://doi.org/10.1128/AEM.66.12.5448-5456.2000

Chung, Y. B., Nardone, C., & Hinkle, D. C. (1990). Bacteriophage T7 DNA packaging. III. A “hairpin” end formed on T7 concatemers may be an intermediate in the processing reaction. *Journal of Molecular Biology*, *216*(4), 939–948. https://doi.org/10.1016/S0022-2836(99)80012-6

Cobián Güemes, A. G., Le, T., Rojas, M. I., Jacobson, N. E., Villela, H., McNair, K., Hung, S.-H., Han, L., Boling, L., Octavio, J. C., Dominguez, L., Cantú, V. A., Archdeacon, S., Vega, A. A., An, M. A., Hajama, H., Burkeen, G., Edwards, R. A., Conrad, D. J., … Segall, A. M. (2023). Compounding Achromobacter Phages for Therapeutic Applications. *Viruses*, *15*(8). https://doi.org/10.3390/v15081665

Cochrane, G., Karsch-Mizrachi, I., & Takagi, T. (2016). International Nucleotide Sequence Database Collaboration. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res*, *44*, D48-50.

Colavecchio, A., Cadieux, B., Lo, A., & Goodridge, L. D. (2017). Bacteriophages Contribute to the Spread of Antibiotic Resistance Genes among Foodborne Pathogens of the Enterobacteriaceae Family - A Review. *Frontiers in Microbiology*, *8*, 1108. https://doi.org/10.3389/fmicb.2017.01108

Collins, C., & Didelot, X. (2018). A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Computational Biology*, *14*(2), e1005958. https://doi.org/10.1371/journal.pcbi.1005958

Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K. E., Kato-Maeda, M., Parkhill, J., Malla, B., Berg, S., Thwaites, G., Yeboah-Manu, D., Bothamley, G., Mei, J., Wei, L., Bentley, S., Harris, S. R., Niemann, S., Diel, R., Aseffa, A., … Gagneux, S. (2013). Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nature Genetics*, *45*(10), 1176–1182. https://doi.org/10.1038/ng.2744

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 (New Rochelle, N.Y.)*, *2*(4), 214–223. https://doi.org/10.1089/phage.2021.0007

Cook, R., Brown, N., Rihtman, B., Michniewski, S., Redgwell, T., Clokie, M., Stekel, D. J., Chen, Y., Scanlan, D. J., Hobman, J. L., Nelson, A., Jones, M. A., Smith, D., & Millard, A. (2023). The long and short of it: Benchmarking viromics using Illumina, Nanopore and PacBio sequencing technologies. In *bioRxiv*. https://doi.org/10.1101/2023.02.12.527533

Cook, R., Telatin, A., Bouras, G., Camargo, A. P., Larralde, M., Edwards, R. A., & Adriaenssens, E. M. (2023). Predicting stop codon reassignment improves functional annotation of bacteriophages. *BioRxiv : The Preprint Server for Biology*. https://doi.org/10.1101/2023.12.19.572299

Correa, A. M. S., Howard-Varona, C., Coy, S. R., Buchan, A., Sullivan, M. B., & Weitz, J. S. (2021). Revisiting the rules of life for viruses of microorganisms. *Nature Reviews. Microbiology*, *19*(8), 501–513. https://doi.org/10.1038/s41579-021-00530-x

Cortés-Martín, A., Buttimer, C., Maier, J. L., Tobin, C. A., Draper, L. A., Ross, R. P., Kleiner, M., Hill, C., & Shkoporov, A. N. (2025). Adaptations in gut Bacteroidales facilitate stable co-existence with their lytic bacteriophages. *Gut Microbes*, *17*(1), 2507775. https://doi.org/10.1080/19490976.2025.2507775

Coutinho, Felipe H., Silveira, C. B., Gregoracci, G. B., Thompson, C. C., Edwards, R. A., Brussaard, C. P. D., Dutilh, B. E., & Thompson, F. L. (2017). Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nature Communications*, *8*(1), 15955. https://doi.org/10.1038/ncomms15955

Coutinho, Felipe Hernandes, Zaragoza-Solas, A., López-Pérez, M., Barylski, J., Zielezinski, A., Dutilh, B. E., Edwards, R., & Rodriguez-Valera, F. (2021). RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. *Patterns (New York, N.Y.)*, *2*(7), 100274. https://doi.org/10.1016/j.patter.2021.100274

Crisci, M. A., Chen, L.-X., Devoto, A. E., Borges, A. L., Bordin, N., Sachdeva, R., Tett, A., Sharrar, A. M., Segata, N., Debenedetti, F., Bailey, M., Burt, R., Wood, R. M., Rowden, L. J., Corsini, P. M., van Winden, S., Holmes, M. A., Lei, S., Banfield, J. F., & Santini, J. M. (2021). Closely related Lak

megaphages replicate in the microbiomes of diverse animals. *IScience*, *24*(8), 102875. https://doi.org/10.1016/j.isci.2021.102875

Dearlove, B. L., Cody, A. J., Pascoe, B., Méric, G., Wilson, D. J., & Sheppard, S. K. (2016). Rapid host switching in generalist Campylobacter strains erodes the signal for tracing human infections. *The ISME Journal*, *10*(3), 721–729. https://doi.org/10.1038/ismej.2015.149

Debroas, D., & Siguret, C. (2019). Viruses as key reservoirs of antibiotic resistance genes in the environment. *The ISME Journal*, *13*(11), 2856–2867. https://doi.org/10.1038/s41396-019-0478-9

Dedrick, R. M., Guerrero-Bustamante, C. A., Garlena, R. A., Russell, D. A., Ford, K., Harris, K., Gilmour, K. C., Soothill, J., Jacobs-Sera, D., Schooley, R. T., Hatfull, G. F., & Spencer, H. (2019). Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant Mycobacterium abscessus. *Nature Medicine*, *25*(5), 730–733. https://doi.org/10.1038/s41591-019-0437-z

Dedrick, R. M., Smith, B. E., Cristinziano, M., Freeman, K. G., Jacobs-Sera, D., Belessis, Y., Whitney Brown, A., Cohen, K. A., Davidson, R. M., van Duin, D., Gainey, A., Garcia, C. B., Robert George, C. R., Haidar, G., Ip, W., Iredell, J., Khatami, A., Little, J. S., Malmivaara, K., … Hatfull, G. F. (2023). Phage therapy of Mycobacterium infections: Compassionate use of phages in 20 patients with drug-resistant Mycobacterial disease. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *76*(1), 103–112. https://doi.org/10.1093/cid/ciac453

Delesalle, V. A., Tanke, N. T., Vill, A. C., & Krukonis, G. P. (2016). Testing hypotheses for the presence of tRNA genes in mycobacteriophage genomes. *Bacteriophage*, *6*(3), e1219441. https://doi.org/10.1080/21597081.2016.1219441

Dennehy, J. J., & Abedon, S. T. (2020). Phage Infection and Lysis. In *Bacteriophages* (pp. 1–43). Springer International Publishing. https://doi.org/10.1007/978-3-319-40598-8_53-1

Didelot, X., & Wilson, D. J. (2015). ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Computational Biology*, *11*(2), e1004041. https://doi.org/10.1371/journal.pcbi.1004041

Dieppa-Colón, E., Martin, C., Kosmopoulos, J. C., & Anantharaman, K. (2025). Prophage-DB: a comprehensive database to explore diversity, distribution, and ecology of prophages. *Environmental Microbiome*, *20*(1), 5. https://doi.org/10.1186/s40793-024-00659-1

Dinges, Z. M., Phillips, R. K., Lively, C. M., & Bashey, F. (2022a). Post-association barrier to host switching maintained despite strong selection in a novel mutualism. *Ecology and Evolution*, *12*(6), e9011. https://doi.org/10.1002/ece3.9011

Dinges, Z. M., Phillips, R. K., Lively, C. M., & Bashey, F. (2022b). Pre- and post-association barriers to host switching in sympatric mutualists. *Journal of Evolutionary Biology*, *35*(7), 962–972. https://doi.org/10.1111/jeb.14028

Dinsdale, E. A., Edwards, R. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., Furlan, M., Desnues, C., Haynes, M., Li, L., McDaniel, L., Moran, M. A., Nelson, K. E., Nilsson, C., Olson, R., Paul, J., Brito, B. R., Ruan, Y., Swan, B. K., … Rohwer, F. (2008). Functional metagenomic profiling of nine biomes. *Nature*, *452*(7187), 629–632.

Dion, M. B., Oechslin, F., & Moineau, S. (2020). Phage diversity, genomics and phylogeny. *Nature Reviews. Microbiology*, *18*(3), 125–138. https://doi.org/10.1038/s41579-019-0311-5

Doane, M. P., Morris, M. M., Papudeshi, B., Allen, L., Pande, D., Haggerty, J. M., Johri, S., Turnlund, A. C., Peterson, M., Kacev, D., Nosal, A., Ramirez, D., Hovel, K., Ledbetter, J., Alker, A., Avalos, J., Baker, K., Bhide, S., Billings, E., … Dinsdale, E. A. (2020). The skin microbiome of elasmobranchs follows phylosymbiosis, but in teleost fishes, the microbiomes converge. *Microbiome*, *8*(1), 93. https://doi.org/10.1186/s40168-020-00840-x

Doane, M. P., Reed, M. B., McKerral, J., Farias Oliveira Lima, L., Morris, M., Goodman, A. Z., Johri, S., Papudeshi, B., Dillon, T., Turnlund, A. C., Peterson, M., Mora, M., de la Parra Venegas, R., Pillans, R., Rohner, C. A., Pierce, S. J., Legaspi, C. G., Araujo, G., Ramirez-Macias, D., … Dinsdale, E. A. (2023). Emergent community architecture despite distinct diversity in the global whale shark (Rhincodon typus) epidermal microbiome. *Scientific Reports*, *13*(1), 12747. https://doi.org/10.1038/s41598-023-39184-5

Dunne, M., Rupf, B., Tala, M., Qabrati, X., Ernst, P., Shen, Y., Sumrall, E., Heeb, L., Plückthun, A., Loessner, M. J., & Kilcher, S. (2019). Reprogramming bacteriophage host range through structure-guided design of chimeric receptor binding proteins. *Cell Reports*, *29*(5), 1336-1350.e4. https://doi.org/10.1016/j.celrep.2019.09.062

Dutilh, B. E., Cassman, N., McNair, K., Sanchez, S. E., Silva, G. G. Z., Boling, L., Barr, J. J., Speth, D. R., Seguritan, V., Aziz, R. K., Felts, B., Dinsdale, E. A., Mokili, J. L., & Edwards, R. A. (2014). A highly

abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature Communications*, *5*, 4498. https://doi.org/10.1038/ncomms5498

Dutilh, B. E., Thompson, C. C., Vicente, A. C. P., Marin, M. A., Lee, C., Silva, G. G. Z., Schmieder, R., Andrade, B. G. N., Chimetto, L., Cuevas, D., Garza, D. R., Okeke, I. N., Aboderin, A. O., Spangler, J., Ross, T., Dinsdale, E. A., Thompson, F. L., Harkins, T. T., & Edwards, R. A. (2014). Comparative genomics of 274 Vibrio cholerae genomes reveals mobile functions structuring three niche dimensions. *BMC Genomics*, *15*, 654.

Dutilh, B. E., Varsani, A., Tong, Y., Simmonds, P., Sabanadzovic, S., Rubino, L., Roux, S., Muñoz, A. R., Lood, C., Lefkowitz, E. J., Kuhn, J. H., Krupovic, M., Edwards, R. A., Brister, J. R., Adriaenssens, E. M., & Sullivan, M. B. (2021). Perspective on taxonomic classification of uncultivated viruses. *Current Opinion in Virology*, *51*, 207–215. https://doi.org/10.1016/j.coviro.2021.10.011

Ecale Zhou, C. L., Kimbrel, J., Edwards, R., McNair, K., Souza, B. A., & Malfatti, S. (2021). MultiPhATE2: code for functional annotation and comparison of phage genomes. *G3* , *11*(5). https://doi.org/10.1093/g3journal/jkab074

Eddy, S. R. (2009, October). A new generation of homology search tools based on probabilistic inference. *Genome Informatics 2009*. Proceedings of the 20th International Conference, Pacifico Yokohama, Japan. https://doi.org/10.1142/9781848165632_0019

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Edwards, R. A., Vega, A. A., Norman, H. M., Ohaeri, M., Levi, K., Dinsdale, E. A., Cinek, O., Aziz, R. K., McNair, K., Barr, J. J., Bibby, K., Brouns, S. J. J., Cazares, A., de Jonge, P. A., Desnues, C., Díaz Muñoz, S. L., Fineran, P. C., Kurilshikov, A., Lavigne, R., … Dutilh, B. E. (2019). Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nature Microbiology*, *4*(10), 1727–1736. https://doi.org/10.1038/s41564-019-0494-6

Egido, J. E., Costa, A. R., Aparicio-Maldonado, C., Haas, P.-J., & Brouns, S. J. J. (2022). Mechanisms and clinical importance of bacteriophage resistance. *FEMS Microbiology Reviews*, *46*(1). https://doi.org/10.1093/femsre/fuab048

Egorov, A. A., & Atkinson, G. C. (2025). LoVis4u: a locus visualization tool for comparative genomics and coverage profiles. *NAR Genomics and Bioinformatics*, *7*(1), lqaf009. https://doi.org/10.1093/nargab/lqaf009

Elek, C. K. A., Brown, T. L., Le Viet, T., Evans, R., Baker, D. J., Telatin, A., Tiwari, S. K., Al-Khanaq, H., Thilliez, G., Kingsley, R. A., Hall, L. J., Webber, M. A., & Adriaenssens, E. M. (2023). A hybrid and poly-polish workflow for the complete and accurate assembly of phage genomes: a case study of ten przondoviruses. *Microbial Genomics*, *9*(7). https://doi.org/10.1099/mgen.0.001065

Ellegaard, K. M., Klasson, L., Näslund, K., Bourtzis, K., & Andersson, S. G. E. (2013). Comparative genomics of Wolbachia and the bacterial species concept. *PLoS Genetics*, *9*(4), e1003381. https://doi.org/10.1371/journal.pgen.1003381

Ellis, E. L., & Delbrück, M. (1939). THE GROWTH OF BACTERIOPHAGE. *The Journal of General Physiology*, *22*(3), 365–384. https://doi.org/10.1085/jgp.22.3.365

Enault, F., Briet, A., Bouteille, L., Roux, S., Sullivan, M. B., & Petit, M.-A. (2017). Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *The ISME Journal*, *11*(1), 237–247. https://doi.org/10.1038/ismej.2016.90

Fang, Z., Feng, T., Zhou, H., & Chen, M. (2022). DeePVP: Identification and classification of phage virion proteins using deep learning. *GigaScience*, *11*. https://doi.org/10.1093/gigascience/giac076

Fang, Z., & Zhou, H. (2021a). Corrigendum: VirionFinder: Identification of Complete and Partial Prokaryote Virus Virion Protein From Virome Data Using the Sequence and Biochemical Properties of Amino Acids. *Frontiers in Microbiology*, *12*, 824018. https://doi.org/10.3389/fmicb.2021.824018

Fang, Z., & Zhou, H. (2021b). VirionFinder: Identification of complete and partial prokaryote virus virion protein from virome data using the sequence and biochemical properties of amino acids. *Frontiers in Microbiology*, *12*, 615711. https://doi.org/10.3389/fmicb.2021.615711

Feiner, R., Argov, T., Rabinovich, L., Sigal, N., Borovok, I., & Herskovits, A. A. (2015). A new perspective on lysogeny: prophages as active regulatory switches of bacteria. *Nature Reviews. Microbiology*, *13*(10), 641–650. https://doi.org/10.1038/nrmicro3527

Feldgarden, M., Brover, V., Gonzalez-Escalona, N., Frye, J. G., Haendiges, J., Haft, D. H., Hoffmann, M., Pettengill, J. B., Prasad, A. B., Tillman, G. E., Tyson, G. H., & Klimke, W. (2021). AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial

resistance, stress response, and virulence. *Scientific Reports*, *11*(1), 12728. https://doi.org/10.1038/s41598-021-91456-0

Fineran, P. C., Blower, T. R., Foulds, I. J., Humphreys, D. P., Lilley, K. S., & Salmond, G. P. C. (2009). The phage abortive infection system, ToxIN, functions as a protein-RNA toxin-antitoxin pair. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(3), 894–899. https://doi.org/10.1073/pnas.0808832106

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, *39*(Web Server issue), W29-37. https://doi.org/10.1093/nar/gkr367

Garneau, J. R., Depardieu, F., Fortier, L.-C., Bikard, D., & Monot, M. (2017). PhageTerm: a tool for fast and accurate determination of phage termini and packaging mechanism using next-generation sequencing data. *Scientific Reports*, *7*(1), 8292. https://doi.org/10.1038/s41598-017-07910-5

Gavric, D., & Knezevic, P. (2022). Filamentous Pseudomonas phage Pf4 in the context of therapy-inducibility, infectivity, lysogenic conversion, and potential application. *Viruses*, *14*(6), 1261. https://doi.org/10.3390/v14061261

Gendre, J., Ansaldi, M., Olivenza, D. R., Denis, Y., Casadesús, J., & Ginet, N. (2022). Genetic Mining of Newly Isolated Salmophages for Phage Therapy. *International Journal of Molecular Sciences*, *23*(16). https://doi.org/10.3390/ijms23168917

Ghequire, M. G. K., & De Mot, R. (2015). The Tailocin Tale: Peeling off Phage Tails. *Trends in Microbiology*, *23*(10), 587–590. https://doi.org/10.1016/j.tim.2015.07.011

Gilbert, J. A., Jansson, J. K., & Knight, R. (2018). Earth microbiome project and global systems biology. *MSystems*, *3*(3). https://doi.org/10.1128/mSystems.00217-17

Gilchrist, C. L. M., & Chooi, Y.-H. (2021). Clinker & clustermap.js: Automatic generation of gene cluster comparison figures. *Bioinformatics* . https://doi.org/10.1093/bioinformatics/btab007

Goodman, A. Z., Papudeshi, B., Doane, M. P., Mora, M., Kerr, E., Torres, M., Nero Moffatt, J., Lima, L., Nosal, A. P., & Dinsdale, E. (2022). Epidermal Microbiomes of Leopard Sharks (Triakis semifasciata) Are Consistent across Captive and Wild Environments. *Microorganisms*, *10*(10). https://doi.org/10.3390/microorganisms10102081

Gordillo Altamirano, F. L., & Barr, J. J. (2019). Phage Therapy in the Postantibiotic Era. *Clinical Microbiology Reviews*, *32*(2). https://doi.org/10.1128/CMR.00066-18

Gordillo Altamirano, F. L., & Barr, J. J. (2021). Unlocking the next generation of phage therapy: the key is in the receptors. *Current Opinion in Biotechnology*, *68*, 115–123. https://doi.org/10.1016/j.copbio.2020.10.002

Gordillo Altamirano, F. L., Kostoulias, X., Subedi, D., Korneev, D., Peleg, A. Y., & Barr, J. J. (2022). Phage-antibiotic combination is a superior treatment against Acinetobacter baumannii in a preclinical study. *EBioMedicine*, *80*, 104045. https://doi.org/10.1016/j.ebiom.2022.104045

Grazziotin, A. L., Koonin, E. V., & Kristensen, D. M. (2017). Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Research*, *45*(D1), D491–D498. https://doi.org/10.1093/nar/gkw975

Grigson, S. R., Giles, S. K., Edwards, R. A., & Papudeshi, B. (2023). Knowing and Naming: Phage Annotation and Nomenclature for Phage Therapy. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *77*(Supplement_5), S352–S359. https://doi.org/10.1093/cid/ciad539

Grigson, Susanna, & Edwards, R. (2023). *What the protein!? Computational methods for predicting microbial protein functions*. https://doi.org/10.31219/osf.io/jhmta

Grigson, Susie, & Mallawaarachchi, V. (2023). *susiegriggo/Phynteny: Phynteny 0.1.10*. Zenodo. https://doi.org/10.5281/ZENODO.8128917

Guerin, E., Shkoporov, A. N., Stockdale, S. R., Comas, J. C., Khokhlova, E. V., Clooney, A. G., Daly, K. M., Draper, L. A., Stephens, N., Scholz, D., Ross, R. P., & Hill, C. (2021). Isolation and characterisation of ΦcrAss002, a crAss-like phage from the human gut that infects Bacteroides xylanisolvens. *Microbiome*, *9*(1), 89. https://doi.org/10.1186/s40168-021-01036-7

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

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* , *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hall, M. (2022). Rasusa: Randomly subsample sequencing reads to a specified coverage. *Journal of Open Source Software*, *7*(69), 3941. https://doi.org/10.21105/joss.03941

Hanna, L. F., Matthews, T. D., Dinsdale, E. A., Hasty, D., & Edwards, R. A. (2012). Characterization of the ELPhiS prophage from Salmonella enterica serovar Enteritidis strain LK5. *Applied and Environmental Microbiology*, *78*(6), 1785–1793. https://doi.org/10.1128/AEM.07241-11

Harmalkar, A., & Gray, J. J. (2020). Advances to tackle backbone flexibility in protein docking. In *arXiv [q-bio.BM]*. arXiv. http://arxiv.org/abs/2010.07455

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

Hawlena, H., Bashey, F., & Lively, C. M. (2010). The evolution of spite: population structure and bacteriocin-mediated antagonism in two natural populations of xenorhabdus bacteria. *Evolution; International Journal of Organic Evolution*, *64*(11), 3198–3204. https://doi.org/10.1111/j.1558-5646.2010.01070.x

Hawlena, H., Bashey, F., Mendes-Soares, H., & Lively, C. M. (2010). Spiteful Interactions in a Natural Population of the Bacterium*Xenorhabdus bovienii*. In *The American Naturalist* (Vol. 175, Issue 3, pp. 374–381). https://doi.org/10.1086/650375

Heinzinger, M., Weissenow, K., Sanchez, J. G., Henkel, A., Mirdita, M., Steinegger, M., & Rost, B. (2024). Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, *6*(4), lqae150. https://doi.org/10.1093/nargab/lqae150

Hendrix, R. W., Smith, M. C., Burns, R. N., Ford, M. E., & Hatfull, G. F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(5), 2192–2197. https://doi.org/10.1073/pnas.96.5.2192

Hesse, R. D., Roach, M., Kerr, E. N., Papudeshi, B., Lima, L. F. O., Goodman, A. Z., Hoopes, L., Scott, M., Meyer, L., Huveneers, C., & Dinsdale, E. A. (2022). Phage Diving: An Exploration of the Carcharhinid Shark Epidermal Virome. *Viruses*, *14*(9). https://doi.org/10.3390/v14091969

HMP Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207–214. https://doi.org/10.1038/nature11234

Hockenberry, A. J., & Wilke, C. O. (2021). BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. *PeerJ*, *9*, e11396. https://doi.org/10.7717/peerj.11396

Hou, K., Wu, Z.-X., Chen, X.-Y., Wang, J.-Q., Zhang, D., Xiao, C., Zhu, D., Koya, J. B., Wei, L., Li, J., & Chen, Z.-S. (2022). Microbiota in health and diseases. *Signal Transduction and Targeted Therapy*, *7*(1), 135. https://doi.org/10.1038/s41392-022-00974-4

Hryckowian, A. J., Merrill, B. D., Porter, N. T., Van Treuren, W., Nelson, E. J., Garlena, R. A., Russell, D. A., Martens, E. C., & Sonnenburg, J. L. (2020). Bacteroides thetaiotaomicron-Infecting Bacteriophage Isolates Inform Sequence-Based Host Range Predictions. *Cell Host & Microbe*, *28*(3), 371-379.e5. https://doi.org/10.1016/j.chom.2020.06.011

Hsieh, P.-F., Lin, H.-H., Lin, T.-L., Chen, Y.-Y., & Wang, J.-T. (2017). Two T7-like Bacteriophages, K5-2 and K5-4, Each Encodes Two Capsule Depolymerases: Isolation and Functional Characterization. *Scientific Reports*, *7*(1), 4624. https://doi.org/10.1038/s41598-017-04644-2

Huang, L., Yang, B., Yi, H., Asif, A., Wang, J., Lithgow, T., Zhang, H., Minhas, F. U. A. A., & Yin, Y. (2021). AcrDB: a database of anti-CRISPR operons in prokaryotes and viruses. *Nucleic Acids Research*, *49*(D1), D622–D629. https://doi.org/10.1093/nar/gkaa857

Hugenholtz, P., Goebel, B. M., & Pace, N. R. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of Bacteriology*, *180*(18), 4765–4774. https://doi.org/10.1128/JB.180.18.4765-4774.1998

Hutinet, G., Lee, Y.-J., de Crécy-Lagard, V., & Weigele, P. R. (2021). Hypermodified DNA in Viruses of E. coli and Salmonella. *EcoSal Plus*, *9*(2), eESP00282019. https://doi.org/10.1128/ecosalplus.ESP-0028-2019

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

Hyman, P. (2019). Phages for Phage Therapy: Isolation, Characterization, and Host Range Breadth. *Pharmaceuticals* , *12*(1). https://doi.org/10.3390/ph12010035

Inglis, L. K., & Edwards, R. A. (2022). How Metagenomics Has Transformed Our Understanding of Bacteriophages in Microbiome Research. *Microorganisms*, *10*(8). https://doi.org/10.3390/microorganisms10081671

Inglis, L. K., Roach, M. J., & Edwards, R. A. (2024). Prophages: an integral but understudied component of the human microbiome. *Microbial Genomics*, *10*(1). https://doi.org/10.1099/mgen.0.001166

Integrative HMP (iHMP) Research Network Consortium. (2019). The Integrative Human Microbiome Project. *Nature*, *569*(7758), 641–648. https://doi.org/10.1038/s41586-019-1238-8

Ito, J. (1978). Bacteriophage phi29 terminal protein: its association with the 5' termini of the phi29 genome. *Journal of Virology*, *28*(3), 895–904. https://doi.org/10.1128/JVI.28.3.895-904.1978

Ivanova, N. N., Schwientek, P., Tripp, H. J., Rinke, C., Pati, A., Huntemann, M., Visel, A., Woyke, T., Kyrpides, N. C., & Rubin, E. M. (2014). Stop codon reassignments in the wild. *Science*, *344*(6186), 909–913. https://doi.org/10.1126/science.1250691

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, *9*(1), 5114. https://doi.org/10.1038/s41467-018-07641-9

Johansen, J., Plichta, D. R., Nissen, J. N., Jespersen, M. L., Shah, S. A., Deng, L., Stokholm, J., Bisgaard, H., Nielsen, D. S., Sørensen, S. J., & Rasmussen, S. (2022). Genome binning of viral entities from bulk metagenomics data. *Nature Communications*, *13*(1), 965. https://doi.org/10.1038/s41467-022-28581-5

Juhala, R. J., Ford, M. E., Duda, R. L., Youlton, A., Hatfull, G. F., & Hendrix, R. W. (2000). Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. *Journal of Molecular Biology*, *299*(1), 27–51. https://doi.org/10.1006/jmbi.2000.3729

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kang, H. S., McNair, K., Cuevas, D. A., Bailey, B. A., Segall, A. M., & Edwards, R. A. (2017). Prophage genomics reveals patterns in phage genome organization and replication. In *bioRxiv* (p. 114819). https://doi.org/10.1101/114819

Kauffman, K. M., Chang, W. K., Brown, J. M., Hussain, F. A., Yang, J., Polz, M. F., & Kelly, L. (2022). Resolving the structure of phage-bacteria interactions in the context of natural diversity. *Nature Communications*, *13*(1), 372. https://doi.org/10.1038/s41467-021-27583-z

Kelly, L., Flamholz, Z., & Biller, S. (2023). Large language models improve annotation of viral proteins. *Research Square*. https://doi.org/10.21203/rs.3.rs-2852098/v1

Kerr, E. N., Papudeshi, B., Haggerty, M., Wild, N., Goodman, A. Z., Lima, L. F. O., Hesse, R. D., Skye, A., Mallawaarachchi, V., Johri, S., Parker, S., & Dinsdale, E. A. (2023). Stingray epidermal microbiomes are species-specific with local adaptations. *Frontiers in Microbiology*, *14*, 1031711. https://doi.org/10.3389/fmicb.2023.1031711

Kim, A. H., Armah, G., Dennis, F., Wang, L., Rodgers, R., Droit, L., Baldridge, M. T., Handley, S. A., & Harris, V. C. (2022). Enteric virome negatively affects seroconversion following oral rotavirus vaccination in a longitudinally sampled cohort of Ghanaian infants. *Cell Host & Microbe*, *30*(1), 110-123.e5. https://doi.org/10.1016/j.chom.2021.12.002

Kim, I.-H., Aryal, S. K., Aghai, D. T., Casanova-Torres, Á. M., Hillman, K., Kozuch, M. P., Mans, E. J., Mauer, T. J., Ogier, J.-C., Ensign, J. C., Gaudriault, S., Goodman, W. G., Goodrich-Blair, H., & Dillman, A. R. (2017). The insect pathogenic bacterium Xenorhabdus innexi has attenuated virulence in multiple insect model hosts yet encodes a potent mosquitocidal toxin. *BMC Genomics*, *18*(1), 927. https://doi.org/10.1186/s12864-017-4311-4

Kim, M.-S., & Bae, J.-W. (2018). Lysogeny is prevalent and widely distributed in the murine gut microbiota. *The ISME Journal*, *12*(4), 1127–1141. https://doi.org/10.1038/s41396-018-0061-9

Knecht, L. E., Veljkovic, M., & Fieseler, L. (2019). Diversity and Function of Phage Encoded Depolymerases. *Frontiers in Microbiology*, *10*, 2949. https://doi.org/10.3389/fmicb.2019.02949

Knezevic, P., Adriaenssens, E. M., & Ictv Report Consortium. (2021). ICTV virus Taxonomy profile: Inoviridae. *The Journal of General Virology*, *102*(7). https://doi.org/10.1099/jgv.0.001614

Knowles, B., Silveira, C. B., Bailey, B. A., Barott, K., Cantu, V. A., Cobián-Güemes, A. G., Coutinho, F. H., Dinsdale, E. A., Felts, B., Furby, K. A., George, E. E., Green, K. T., Gregoracci, G. B., Haas, A. F., Haggerty, J. M., Hester, E. R., Hisakawa, N., Kelly, L. W., Lim, Y. W., … Rohwer, F. (2016). Lytic to temperate switching of viral communities. *Nature*, *531*(7595), 466–470. https://doi.org/10.1038/nature17193

Kolenda, C., Medina, M., Bonhomme, M., Laumay, F., Roussel-Gaillard, T., Martins-Simoes, P., Tristan, A., Pirot, F., Ferry, T., Laurent, F., & PHAGEinLYON Study Group. (2022). Phage therapy against Staphylococcus aureus: Selection and optimization of production protocols of novel broad-spectrum Silviavirus phages. *Pharmaceutics*, *14*(9), 1885. https://doi.org/10.3390/pharmaceutics14091885

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540–546. https://doi.org/10.1038/s41587-019-0072-8

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. In *Genome Research* (Vol. 27, Issue 5, pp. 722–736). https://doi.org/10.1101/gr.215087.116

Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H., & Frost, S. D. W. (2006). GARD: a genetic algorithm for recombination detection. *Bioinformatics (Oxford, England)*, *22*(24), 3096–3098. https://doi.org/10.1093/bioinformatics/btl474

Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* , *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Krupovic, M., & ICTV Report Consortium. (2018). ICTV virus taxonomy profile: Plasmaviridae. *The Journal of General Virology*, *99*(5), 617–618. https://doi.org/10.1099/jgv.0.001060

Kutateladze, M., & Adamia, R. (2008). Phage therapy experience at the Eliava Institute. *Médecine et Maladies Infectieuses*, *38*(8), 426–430. https://doi.org/10.1016/j.medmal.2008.06.023

Kutateladze, M. I., Tevdoradze, E. S., Balarjishvili, N. S., Skhirtladze, N. R., Adamia, R. S., & Tophuria, T. I. (2009). Characterization of bacterial strains isolated from the CF patients in Georgia and evaluation of the efficacy of phage treatment. *Journal of Cystic Fibrosis: Official Journal of the European Cystic Fibrosis Society*, *8*, S29. https://doi.org/10.1016/s1569-1993(09)60115-6

Larralde, M. (2022). Pyrodigal: Python bindings and interface to Prodigal, an efficient method for gene prediction in prokaryotes. *Journal of Open Source Software*, *7*(72), 4296. https://doi.org/10.21105/joss.04296

Laslett, D., & Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research*, *32*(1), 11–16. https://doi.org/10.1093/nar/gkh152

Latka, A., Lemire, S., Grimon, D., Dams, D., Maciejewska, B., Lu, T., Drulis-Kawa, Z., & Briers, Y. (2021). Engineering the modular receptor-binding proteins of Klebsiella phages switches their capsule serotype specificity. *MBio*, *12*(3). https://doi.org/10.1128/mBio.00455-21

Lederberg, J., & Mccray, A. T. (2001, April 2). `ome sweet `omics--A genealogical treasury of words. *Scientist (Philadelphia, Pa.)*, *15*, 8. https://go.gale.com/ps/i.do?id=GALE%7CA73535513&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=08903670&p=AONE&sw=w

Lee, D. Y., Bartels, C., McNair, K., Edwards, R. A., Swairjo, M. A., & Luque, A. (2022). Predicting the capsid architecture of phages from metagenomic data. *Computational and Structural Biotechnology Journal*, *20*, 721–732. https://doi.org/10.1016/j.csbj.2021.12.032

Lee, M.-M., & Stock, S. P. (2010). A multilocus approach to assessing co-evolutionary relationships between Steinernema spp. (Nematoda: Steinernematidae) and their bacterial symbionts Xenorhabdus spp. (gamma-Proteobacteria: Enterobacteriaceae). *Systematic Parasitology*, *77*(1), 1–12. https://doi.org/10.1007/s11230-010-9256-9

Lefkowitz, E. J., Dempsey, D. M., Hendrickson, R. C., Orton, R. J., Siddell, S. G., & Smith, D. B. (2018). Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Research*, *46*(D1), D708–D717. https://doi.org/10.1093/nar/gkx932

Legendre, P., Desdevises, Y., & Bazin, E. (2002). A statistical test for host-parasite coevolution. *Systematic Biology*, *51*(2), 217–234. https://doi.org/10.1080/10635150252899734

Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research*, *47*(W1), W256–W259. https://doi.org/10.1093/nar/gkz239

Levi, K., Rynge, M., Abeysinghe, E., & Edwards, R. A. (2018). Searching the Sequence Read Archive using Jetstream and Wrangler. *Proceedings of the Practice and Experience on Advanced Research Computing*, 1–7. https://doi.org/10.1145/3219104.3229278

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* , *31*(10), 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., & Lam, T.-W. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* , *102*, 3–11. https://doi.org/10.1016/j.ymeth.2016.02.020

Li, Z., Jaroszewski, L., Iyer, M., Sedova, M., & Godzik, A. (2020). FATCAT 2.0: towards a better understanding of the structural diversity of proteins. *Nucleic Acids Research*, *48*(W1), W60–W64. https://doi.org/10.1093/nar/gkaa443

Lima, L. F. O., Weissman, M., Reed, M., Papudeshi, B., Alker, A. T., Morris, M. M., Edwards, R. A., de Putron, S. J., Vaidya, N. K., & Dinsdale, E. A. (2020). Modeling of the Coral Microbiome: the Influence of Temperature and Microbial Network. *MBio*, *11*(2). https://doi.org/10.1128/mBio.02691-19

Lima-Mendez, G., Van Helden, J., Toussaint, A., & Leplae, R. (2008). Reticulate representation of evolutionary and functional relationships between phage genomes. *Molecular Biology and Evolution*, *25*(4), 762–777. https://doi.org/10.1093/molbev/msn023

Liu, B., Zheng, D., Zhou, S., Chen, L., & Yang, J. (2022). VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Research*, *50*(D1), D912–D917. https://doi.org/10.1093/nar/gkab1107

Livny, J., Larock, C. N., & Friedman, D. I. (2009). Identification and isolation of lysogens with induced prophage. *Methods in Molecular Biology (Clifton, N.J.)*, *501*, 253–265. https://doi.org/10.1007/978-1-60327-164-6_22

Logan, S. L., Thomas, J., Yan, J., Baker, R. P., Shields, D. S., Xavier, J. B., Hammer, B. K., & Parthasarathy, R. (2018). The Vibrio cholerae type VI secretion system can modulate host intestinal mechanics to displace gut bacterial symbionts. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(16), E3779–E3787. https://doi.org/10.1073/pnas.1720133115

Longin, H., Broeckaert, N., van Noort, V., Lavigne, R., & Hendrix, H. (2024). Posttranslational modifications in bacteria during phage infection. *Current Opinion in Microbiology*, *77*(102425), 102425. https://doi.org/10.1016/j.mib.2024.102425

Łoś, M., & Węgrzyn, G. (2012). Pseudolysogeny. *Advances in Virus Research*, *82*, 339–349. https://doi.org/10.1016/B978-0-12-394621-8.00019-4

Luong, T., Salabarria, A.-C., Edwards, R. A., & Roach, D. R. (2020). Standardized bacteriophage purification for personalized phage therapy. *Nature Protocols*, *15*(9), 2867–2890. https://doi.org/10.1038/s41596-020-0346-0

Luque, A., Benler, S., Lee, D. Y., Brown, C., & White, S. (2020). The Missing Tailed Phages: Prediction of Small Capsid Candidates. *Microorganisms*, *8*(12). https://doi.org/10.3390/microorganisms8121944

Mallawaarachchi, V. G., & Lin, Y. (2022). Metacoag: Binning metagenomic contigs via composition, coverage and assembly graphs. *Research in Computational Molecular Biology: ... Annual International Conference, RECOMB ...: Proceedings. International Conference on Research in Computational Molecular Biology*. https://link.springer.com/chapter/10.1007/978-3-031-04749-7_5

Mallawaarachchi, V. G., Wickramarachchi, A. S., & Lin, Y. (2020). GraphBin2: refined and overlapped binning of metagenomic contigs using assembly graphs. *WASA ...: International Conference on Wireless Algorithms, Systems, and Applications: Proceedings. WASA*. https://drops.dagstuhl.de/opus/volltexte/2020/12797/

Mallawaarachchi, V., Roach, M. J., Decewicz, P., Papudeshi, B., Giles, S. K., Grigson, S. R., Bouras, G., Hesse, R. D., Inglis, L. K., Hutton, A. L. K., Dinsdale, E. A., & Edwards, R. A. (2023). Phables: from fragmented assemblies to high-quality bacteriophage genomes. *Bioinformatics* , *39*(10). https://doi.org/10.1093/bioinformatics/btad586

Mallawaarachchi, V., Wickramarachchi, A., & Lin, Y. (2020). GraphBin: refined binning of metagenomic contigs using assembly graphs. *Bioinformatics* , *36*(11), 3307–3313. https://doi.org/10.1093/bioinformatics/btaa180

Manavalan, B., Shin, T. H., & Lee, G. (2018). PVP-SVM: Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine. *Frontiers in Microbiology*, *9*, 476. https://doi.org/10.3389/fmicb.2018.00476

Martens, E. C., Heungens, K., & Goodrich-Blair, H. (2003). Early Colonization Events in the Mutualistic Association between *Steinernema carpocapsae* Nematodes and *Xenorhabdus nematophila* Bacteria. In *Journal of Bacteriology* (Vol. 185, Issue 10, pp. 3147–3154). https://doi.org/10.1128/jb.185.10.3147-3154.2003

Martin, D. P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P., & Varsani, A. (2011). Recombination in eukaryotic single stranded DNA viruses. *Viruses*, *3*(9), 1699–1738. https://doi.org/10.3390/v3091699

Matthews, T. D., Schmieder, R., Silva, G. G. Z., Busch, J., Cassman, N., Dutilh, B. E., Green, D., Matlock, B., Heffernan, B., Olsen, G. J., Farris Hanna, L., Schifferli, D. M., Maloy, S., Dinsdale, E. A., & Edwards, R. A. (2015). Genomic Comparison of the Closely-Related Salmonella enterica Serovars Enteritidis, Dublin and Gallinarum. *PloS One*, *10*(6), e0126883. https://doi.org/10.1371/journal.pone.0126883

Mavrich, T. N., & Hatfull, G. F. (2017). Bacteriophage evolution differs by host, lifestyle and genome. *Nature Microbiology*, *2*, 17112. https://doi.org/10.1038/nmicrobiol.2017.112

Mayo-Muñoz, D., Pinilla-Redondo, R., Camara-Wilpert, S., Birkholz, N., & Fineran, P. C. (2024). Inhibitors of bacterial immune systems: discovery, mechanisms and applications. *Nature Reviews. Genetics*, *25*(4), 237–254. https://doi.org/10.1038/s41576-023-00676-9

McAdam, P. R., Templeton, K. E., Edwards, G. F., Holden, M. T. G., Feil, E. J., Aanensen, D. M., Bargawi, H. J. A., Spratt, B. G., Bentley, S. D., Parkhill, J., Enright, M. C., Holmes, A., Girvan, E. K., Godfrey, P. A., Feldgarden, M., Kearns, A. M., Rambaut, A., Robinson, D. A., & Fitzgerald, J. R. (2012). Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant Staphylococcus aureus. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(23), 9107–9112. https://doi.org/10.1073/pnas.1202869109

McArthur, J. V., Kovacic, D. A., & Smith, M. H. (1988). Genetic diversity in natural populations of a soil bacterium across a landscape gradient. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(24), 9621–9624. https://doi.org/10.1073/pnas.85.24.9621

McKerral, J. C., Papudeshi, B., Inglis, L. K., Roach, M. J., Decewicz, P., McNair, K., Luque, A., Dinsdale, E. A., & Edwards, R. A. (2023). The Promise and Pitfalls of Prophages. In *bioRxiv* (p. 2023.04.20.537752). https://doi.org/10.1101/2023.04.20.537752

McMullen, J. G., Peterson, B. F., Forst, S., Blair, H. G., & Patricia Stock, S. (2017). Fitness costs of symbiont switching using entomopathogenic nematodes as a model. In *BMC Evolutionary Biology* (Vol. 17, Issue 1). https://doi.org/10.1186/s12862-017-0939-6

McNair, K., Bailey, B. A., & Edwards, R. A. (2012). PHACTS, a computational approach to classifying the lifestyle of phages. *Bioinformatics* , *28*(5), 614–618. https://doi.org/10.1093/bioinformatics/bts014

McNair, K., Salamon, P., Edwards, R. A., & Segall, A. M. (2023). PRFect: A tool to predict programmed ribosomal frameshifts in prokaryotic and viral genomes. *Research Square*. https://doi.org/10.21203/rs.3.rs-2997217/v1

McNair, K., Zhou, C., Dinsdale, E. A., Souza, B., & Edwards, R. A. (2019). PHANOTATE: a novel approach to gene identification in phage genomes. *Bioinformatics* , *35*(22), 4537–4542. https://doi.org/10.1093/bioinformatics/btz265

Minich, J. J., Morris, M. M., Brown, M., Doane, M., Edwards, M. S., Michael, T. P., & Dinsdale, E. A. (2018). Elevated temperature drives kelp microbiome dysbiosis, while elevated carbon dioxide induces water microbiome disruption. *PloS One*, *13*(2), e0192772.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*, *19*(6), 679–682. https://doi.org/10.1038/s41592-022-01488-1

Modi, S. R., Lee, H. H., Spina, C. S., & Collins, J. J. (2013). Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature*, *499*(7457), 219–222. https://doi.org/10.1038/nature12212

Morris, M. M., Haggerty, J. M., Papudeshi, B. N., Vega, A. A., Edwards, M. S., & Dinsdale, E. A. (2016). Nearshore Pelagic Microbial Community Abundance Affects Recruitment Success of Giant Kelp,. *Frontiers in Microbiology*, *7*, 1800.

Moura de Sousa, J. A., Pfeifer, E., Touchon, M., & Rocha, E. P. C. (2021). Causes and consequences of bacteriophage diversification via genetic exchanges across lifestyles and bacterial taxa. *Molecular Biology and Evolution*, *38*(6), 2497–2512. https://doi.org/10.1093/molbev/msab044

Murfin, K. E., Ginete, D. R., Bashey, F., & Goodrich-Blair, H. (2018). Symbiont-mediated competition: Xenorhabdus bovienii confer an advantage to their nematode host Steinernema affine by killing competitor Steinernema feltiae. *Environmental Microbiology*. https://doi.org/10.1111/1462-2920.14278

Murfin, K. E., Lee, M.-M., Klassen, J. L., McDonald, B. R., Larget, B., Forst, S., Stock, S. P., Currie, C. R., & Goodrich-Blair, H. (2015). Xenorhabdus bovienii Strain Diversity Impacts Coevolution and Symbiotic

Maintenance with Steinernema spp. Nematode Hosts. *MBio*, *6*(3), e00076. https://doi.org/10.1128/mBio.00076-15

Murfin, K. E., Whooley, A. C., Klassen, J. L., & Goodrich-Blair, H. (2015). Comparison of Xenorhabdus bovienii bacterial strain genomes reveals diversity in symbiotic functions. *BMC Genomics*, *16*, 889. https://doi.org/10.1186/s12864-015-2000-8

Mutreja, A., Kim, D. W., Thomson, N. R., Connor, T. R., Lee, J. H., Kariuki, S., Croucher, N. J., Choi, S. Y., Harris, S. R., Lebens, M., Niyogi, S. K., Kim, E. J., Ramamurthy, T., Chun, J., Wood, J. L. N., Clemens, J. D., Czerkinsky, C., Nair, G. B., Holmgren, J., … Dougan, G. (2011). Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature*, *477*(7365), 462–465. https://doi.org/10.1038/nature10392

Nakamura, T., Yamada, K. D., Tomii, K., & Katoh, K. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics (Oxford, England)*, *34*(14), 2490–2492. https://doi.org/10.1093/bioinformatics/bty121

Nanoporetech Consortium. (2022). *medaka: Sequence correction provided by ONT Research*. https://github.com/nanoporetech/medaka/releases

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), 578–585. https://doi.org/10.1038/s41587-020-00774-7

Niehus, R., Mitri, S., Fletcher, A. G., & Foster, K. R. (2015). Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nature Communications*, *6*, 8924. https://doi.org/10.1038/ncomms9924

Nielsen, T. K., Forero-Junco, L. M., Kot, W., Moineau, S., Hansen, L. H., & Riber, L. (2023). Detection of nucleotide modifications in bacteria and bacteriophages: Strengths and limitations of current technologies and software. *Molecular Ecology*, *32*(6), 1236–1247. https://doi.org/10.1111/mec.16679

Nir-Paz, R., Onallah, H., Dekel, M., Gellman, Y. N., Haze, A., Ben-Ami, R., Braunstein, R., Hazan, R., Dror, D., Oster, Y., Cherniak, M., Attal, F., Barbosa, A. R., Dordio, H., Wagner, A., Jones-Dias, D., Neves, J., Barreto, M., Leandro, C., … Garcia, M. (2024). Randomized double-blind study on safety and

tolerability of TP-102 phage cocktail in patients with infected and non-infected diabetic foot ulcers. *Med (New York, N.Y.)*, 100565. https://doi.org/10.1016/j.medj.2024.11.018

Nissen, J. N., Johansen, J., Allesøe, R. L., Sønderby, C. K., Armenteros, J. J. A., Grønbech, C. H., Jensen, L. J., Nielsen, H. B., Petersen, T. N., Winther, O., & Rasmussen, S. (2021). Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology*, *39*(5), 555–560. https://doi.org/10.1038/s41587-020-00777-4

Nobrega, F. L., Vlot, M., de Jonge, P. A., Dreesens, L. L., Beaumont, H. J. E., Lavigne, R., Dutilh, B. E., & Brouns, S. J. J. (2018). Targeting mechanisms of tailed bacteriophages. *Nature Reviews. Microbiology*, *16*(12), 760–773. https://doi.org/10.1038/s41579-018-0070-8

Norman, J. M., Handley, S. A., Baldridge, M. T., Droit, L., Liu, C. Y., Keller, B. C., Kambal, A., Monaco, C. L., Zhao, G., Fleshner, P., Stappenbeck, T. S., McGovern, D. P. B., Keshavarzian, A., Mutlu, E. A., Sauk, J., Gevers, D., Xavier, R. J., Wang, D., Parkes, M., & Virgin, H. W. (2015). Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell*, *160*(3), 447–460. https://doi.org/10.1016/j.cell.2015.01.002

Oda, Y., Star, B., Huisman, L. A., Gottschal, J. C., & Forney, L. J. (2003). Biogeography of the purple nonsulfur bacterium Rhodopseudomonas palustris. *Applied and Environmental Microbiology*, *69*(9), 5186–5191. https://doi.org/10.1128/AEM.69.9.5186-5191.2003

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1), 132. https://doi.org/10.1186/s13059-016-0997-x

Oromí-Bosch, A., Antani, J. D., & Turner, P. E. (2023). Developing Phage Therapy That Overcomes the Evolution of Bacterial Resistance. *Annual Review of Virology*, *10*(1), 503–524. https://doi.org/10.1146/annurev-virology-012423-110530

Pace, N. R., Stahl, D. A., Lane, D. J., & Olsen, G. J. (1986). The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences. In K. C. Marshall (Ed.), *Advances in Microbial Ecology* (pp. 1–55). Springer US. https://doi.org/10.1007/978-1-4757-0611-6_1

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* , *31*(22), 3691–3693. https://doi.org/10.1093/bioinformatics/btv421

Page, A. J., Taylor, B., Delaney, A. J., Soares, J., Seemann, T., Keane, J. A., & Harris, S. R. (2016). *SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments*. https://doi.org/10.1101/038190

Papudeshi, B., Haggerty, J. M., Doane, M., Morris, M. M., Walsh, K., Beattie, D. T., Pande, D., Zaeri, P., Silva, G. G. Z., Thompson, F., Edwards, R. A., & Dinsdale, E. A. (2017). Optimizing and evaluating the reconstruction of Metagenome-assembled microbial genomes. *BMC Genomics*, *18*(1), 915. https://doi.org/10.1186/s12864-017-4294-1

Papudeshi, B., Rusch, D. B., David, V., Lively, C. M., Edwards, R. A., & Bashey, F. (2023). Host association and spatial proximity shape but do not constrain population structure in the mutualistic symbiont, Xenorhabdus bovienii. *MBio*, *Accepted*. https://doi.org/10.1128/mbio.00434-23

Papudeshi, B., Vega, A. A., Souza, C., Giles, S. K., Mallawaarachchi, V., Roach, M. J., An, M., Jacobson, N., McNair, K., Fernanda Mora, M., Pastrana, K., Boling, L., Leigh, C., Harker, C., Plewa, W. S., Grigson, S. R., Bouras, G., Decewicz, P., Luque, A., … Edwards, R. A. (2023). Host interactions of novel Crassvirales species belonging to multiple families infecting bacterial host, Bacteroides cellulosilyticus WH2. *Microbial Genomics*, *9*(9). https://doi.org/10.1099/mgen.0.001100

Pargin, E., Roach, M. J., Skye, A., Papudeshi, B., Inglis, L. K., Mallawaarachchi, V., Grigson, S. R., Harker, C., Edwards, R. A., & Giles, S. K. (2023). The human gut virome: composition, colonization, interactions, and impacts on human health. *Frontiers in Microbiology*, *14*, 963173. https://doi.org/10.3389/fmicb.2023.963173

Pargin, E., Roach, M., Skye, A., Edwards, R., & Giles, S. (2022). *The human gut virome: Composition, colonisation, interactions, and impacts on human health*. https://doi.org/10.31219/osf.io/s9px2

Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T., Wadsworth, C., Lewis, J. A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N. R., Brucker, W., Kumar, V., Kandasamy, J., Keenan, L., Bardarov, S., Kriakov, J., Lawrence, J. G., Jacobs, W. R., Jr, Hendrix, R. W., & Hatfull, G. F. (2003). Origins of highly mosaic mycobacteriophage genomes. *Cell*, *113*(2), 171–182. https://doi.org/10.1016/s0092-8674(03)00233-2

Pérez-Carrascal, O. M., Terrat, Y., Giani, A., Fortin, N., Greer, C. W., Tromas, N., & Shapiro, B. J. (2019). Coherence of Microcystis species revealed through population genomics. *The ISME Journal*, *13*(12), 2887–2900. https://doi.org/10.1038/s41396-019-0481-1

Peters, S. L., Borges, A. L., Giannone, R. J., Morowitz, M. J., Banfield, J. F., & Hettich, R. L. (2022). Experimental validation that human microbiome phages use alternative genetic coding. *Nature Communications*, *13*(1), 5710. https://doi.org/10.1038/s41467-022-32979-6

Pfeifer, E., Bonnin, R. A., & Rocha, E. P. C. (2022). Phage-plasmids spread antibiotic resistance genes through infection and lysogenic conversion. *MBio*, *13*(5), e0185122. https://doi.org/10.1128/mbio.01851-22

Pfeifer, E., Moura de Sousa, J. A., Touchon, M., & Rocha, E. P. C. (2021). Bacteria have numerous distinctive groups of phage-plasmids with conserved phage and variable plasmid gene repertoires. *Nucleic Acids Research*, *49*(5), 2655–2673. https://doi.org/10.1093/nar/gkab064

Pfennig, A., Lomsadze, A., & Borodovsky, M. (2023). MgCod: Gene Prediction in Phage Genomes with Multiple Genetic Codes. *Journal of Molecular Biology*, *435*(14), 168159. https://doi.org/10.1016/j.jmb.2023.168159

Pollet, R. M., Martin, L. M., & Koropatkin, N. M. (2021). TonB-dependent transporters in the Bacteroidetes: Unique domain structures and potential functions. *Molecular Microbiology*, *115*(3), 490–501. https://doi.org/10.1111/mmi.14683

Porter, N. T., Hryckowian, A. J., Merrill, B. D., Fuentes, J. J., Gardner, J. O., Glowacki, R. W. P., Singh, S., Crawford, R. D., Snitkin, E. S., Sonnenburg, J. L., & Martens, E. C. (2020). Phase-variable capsular polysaccharides and lipoproteins modify bacteriophage susceptibility in Bacteroides thetaiotaomicron. *Nature Microbiology*, *5*(9), 1170–1181. https://doi.org/10.1038/s41564-020-0746-5

Porter, S. S., Chang, P. L., Conow, C. A., Dunham, J. P., & Friesen, M. L. (2017). Association mapping reveals novel serpentine adaptation gene clusters in a population of symbiotic Mesorhizobium. In *The ISME Journal* (Vol. 11, Issue 1, pp. 248–262). https://doi.org/10.1038/ismej.2016.88

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS One*, *5*(3), e9490. https://doi.org/10.1371/journal.pone.0009490

Pride, D. T., Wassenaar, T. M., Ghose, C., & Blaser, M. J. (2006). Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. *BMC Genomics*, *7*, 8. https://doi.org/10.1186/1471-2164-7-8

Proux, C., van Sinderen, D., Suarez, J., Garcia, P., Ladero, V., Fitzgerald, G. F., Desiere, F., & Brüssow, H. (2002). The dilemma of phage taxonomy illustrated by comparative genomics of Sfi21-like

Siphoviridae in lactic acid bacteria. *Journal of Bacteriology*, *184*(21), 6026–6036. https://doi.org/10.1128/JB.184.21.6026-6036.2002

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., … Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, *464*(7285), 59–65. https://doi.org/10.1038/nature08821

Raiko, M. (2021). *viralVerify: viral contig verification tool* (1.1) [Computer software]. https://github.com/ablab/viralVerify

Ravin, N. V. (2011). N15: the linear phage-plasmid. *Plasmid*, *65*(2), 102–109. https://doi.org/10.1016/j.plasmid.2010.12.004

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

Rhoades, M., & Rhoades, E. A. (1972). Terminal repetition in the DNA of bacteriophage T5. *Journal of Molecular Biology*, *69*(2), 187–200. https://doi.org/10.1016/0022-2836(72)90224-0

Richardson, E. J., Bacigalupe, R., Harrison, E. M., Weinert, L. A., Lycett, S., Vrieling, M., Robb, K., Hoskisson, P. A., Holden, M. T. G., Feil, E. J., & Others. (2018). *Gene exchange drives the ecological success of a multi-host bacterial pathogen. Nat Ecol Evol 2: 1468--1478*.

Richardson, Emily J., Bacigalupe, R., Harrison, E. M., Weinert, L. A., Lycett, S., Vrieling, M., Robb, K., Hoskisson, P. A., Holden, M. T. G., Feil, E. J., Paterson, G. K., Tong, S. Y. C., Shittu, A., van Wamel, W., Aanensen, D. M., Parkhill, J., Peacock, S. J., Corander, J., Holmes, M., & Fitzgerald, J. R. (2018). Gene exchange drives the ecological success of a multi-host bacterial pathogen. *Nature Ecology & Evolution*, *2*(9), 1468–1478. https://doi.org/10.1038/s41559-018-0617-0

Roach, M., Beecroft, S., Mihindukulasuriya, K. A., Wang, L., Lima, L. F. O., Dinsdale, E. A., Edwards, R. A., & Handley, S. A. (2022). Hecatomb: An End-to-End Research Platform for Viral Metagenomics. In *bioRxiv* (p. 2022.05.15.492003). https://doi.org/10.1101/2022.05.15.492003

Roach, M. J., Pierce-Ward, N. T., Suchecki, R., Mallawaarachchi, V., Papudeshi, B., Handley, S. A., Brown, C. T., Watson-Haigh, N. S., & Edwards, R. A. (2022). Ten simple rules and a template for creating

workflows-as-applications. *PLoS Computational Biology*, *18*(12), e1010705. https://doi.org/10.1371/journal.pcbi.1010705

Rocha, E. P. C. (2018). Neutral Theory, Microbial Practice: Challenges in Bacterial Population Genetics. *Molecular Biology and Evolution*, *35*(6), 1338–1347. https://doi.org/10.1093/molbev/msy078

Rohde, C., Resch, G., Pirnay, J.-P., Blasdel, B. G., Debarbieux, L., Gelman, D., Górski, A., Hazan, R., Huys, I., Kakabadze, E., Łobocka, M., Maestri, A., Almeida, G. M. de F., Makalatia, K., Malik, D. J., Mašlaňová, I., Merabishvili, M., Pantucek, R., Rose, T., … Chanishvili, N. (2018). Expert Opinion on Three Phage Therapy Related Topics: Bacterial Phage Resistance, Phage Training and Prophages in Bacterial Production Strains. *Viruses*, *10*(4). https://doi.org/10.3390/v10040178

Rohwer, F., & Edwards, R. (2002). The Phage Proteomic Tree: a genome-based taxonomy for phage. *Journal of Bacteriology*, *184*(16), 4529–4535. https://doi.org/10.1128/JB.184.16.4529-4535.2002

Rossi, A., Treu, L., Toppo, S., Zschach, H., Campanaro, S., & Dutilh, B. E. (2020). Evolutionary Study of the Crassphage Virus at Gene Level. *Viruses*, *12*(9). https://doi.org/10.3390/v12091035

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

Salas, M., Mellado, R. P., Viñuela, E., & Sogo, J. M. (1978). Characterization of a protein covalently linked to the 5′ termini of the DNA of Bacillus subtilis phage φ29. *Journal of Molecular Biology*, *119*(2), 269–291. https://doi.org/10.1016/0022-2836(78)90438-2

Samson, J. E., Magadán, A. H., Sabri, M., & Moineau, S. (2013). Revenge of the phages: defeating bacterial defences. *Nature Reviews. Microbiology*, *11*(10), 675–687. https://doi.org/10.1038/nrmicro3096

Santichaivekin, S., Yang, Q., Liu, J., Mawhorter, R., Jiang, J., Wesley, T., Wu, Y.-C., & Libeskind-Hadas, R. (2021). eMPRess: a systematic cophylogeny reconciliation tool. *Bioinformatics (Oxford, England)*, *37*(16), 2481–2482. https://doi.org/10.1093/bioinformatics/btaa978

Say, H., Joris, B., Giguere, D., & Gloor, G. B. (2023). Annotating Metagenomically Assembled Bacteriophage from a Unique Ecological System using Protein Structure Prediction and Structure Homology Search. In *bioRxiv* (p. 2023.04.19.537516). https://doi.org/10.1101/2023.04.19.537516

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* , *27*(6), 863–864. https://doi.org/10.1093/bioinformatics/btr026

Schooley, R. T., Biswas, B., Gill, J. J., Hernandez-Morales, A., Lancaster, J., Lessor, L., Barr, J. J., Reed, S. L., Rohwer, F., Benler, S., Segall, A. M., Taplitz, R., Smith, D. M., Kerr, K., Kumaraswamy, M., Nizet, V., Lin, L., McCauley, M. D., Strathdee, S. A., … Hamilton, T. (2017). Development and Use of Personalized Bacteriophage-Based Therapeutic Cocktails To Treat a Patient with a Disseminated Resistant Acinetobacter baumannii Infection. *Antimicrobial Agents and Chemotherapy*, *61*(10). https://doi.org/10.1128/AAC.00954-17

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. In *Bioinformatics* (Vol. 30, Issue 14, pp. 2068–2069). https://doi.org/10.1093/bioinformatics/btu153

Shahed-Al-Mahmud, M., Roy, R., Sugiokto, F. G., Islam, M. N., Lin, M.-D., Lin, L.-C., & Lin, N.-T. (2021). Phage φAB6-Borne Depolymerase Combats Acinetobacter baumannii Biofilm Formation and Infection. *Antibiotics (Basel, Switzerland)*, *10*(3). https://doi.org/10.3390/antibiotics10030279

Shamash, M., & Maurice, C. F. (2022). Phages in the infant gut: a framework for virome development during early life. *The ISME Journal*, *16*(2), 323–330. https://doi.org/10.1038/s41396-021-01090-x

Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabó, G., Polz, M. F., & Alm, E. J. (2012). Population genomics of early events in the ecological differentiation of bacteria. *Science*, *336*(6077), 48–51. https://doi.org/10.1126/science.1218198

Sharifi, F., & Ye, Y. (2019). MyDGR: a server for identification and characterization of diversity-generating retroelements. *Nucleic Acids Research*, *47*(W1), W289–W294. https://doi.org/10.1093/nar/gkz329

Shen, A., & Millard, A. (2021). Phage Genome Annotation: Where to Begin and End. *PHAGE (New Rochelle, N.Y.)*, *2*(4), 183–193. https://doi.org/10.1089/phage.2021.0015

Sheppard, S. K., Cheng, L., Méric, G., de Haan, C. P. A., Llarena, A.-K., Marttinen, P., Vidal, A., Ridley, A., Clifton-Hadley, F., Connor, T. R., Strachan, N. J. C., Forbes, K., Colles, F. M., Jolley, K. A., Bentley, S. D., Maiden, M. C. J., Hänninen, M.-L., Parkhill, J., Hanage, W. P., & Corander, J. (2014). Cryptic ecology among host generalist Campylobacter jejuni in domestic animals. *Molecular Ecology*, *23*(10), 2442–2451. https://doi.org/10.1111/mec.12742

Sheppard, S. K., Colles, F., Richardson, J., Cody, A. J., Elson, R., Lawson, A., Brick, G., Meldrum, R., Little, C. L., Owen, R. J., Maiden, M. C. J., & McCarthy, N. D. (2010). Host association of Campylobacter genotypes transcends geographic variation. *Applied and Environmental Microbiology*, *76*(15), 5269–5277. https://doi.org/10.1128/AEM.00124-10

Sheppard, S. K., Didelot, X., Meric, G., Torralbo, A., Jolley, K. A., Kelly, D. J., Bentley, S. D., Maiden, M. C. J., Parkhill, J., & Falush, D. (2013). Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(29), 11923–11927. https://doi.org/10.1073/pnas.1305559110

Sheppard, S. K., Guttman, D. S., & Ross Fitzgerald, J. (2018). Population genomics of bacterial host adaptation. In *Nature Reviews Genetics* (Vol. 19, Issue 9, pp. 549–565). https://doi.org/10.1038/s41576-018-0032-z

Shin, J. H., Tillotson, G., MacKenzie, T. N., Warren, C. A., Wexler, H. M., & Goldstein, E. J. C. (2024). Bacteroides and related species: The keystone taxa of the human gut microbiota. *Anaerobe*, *85*(102819), 102819. https://doi.org/10.1016/j.anaerobe.2024.102819

Shirali, A., Stebliankin, V., Karki, U., Shi, J., Chapagain, P., & Narasimhan, G. (2025). A comprehensive survey of scoring functions for protein docking models. *BMC Bioinformatics*, *26*(1), 25. https://doi.org/10.1186/s12859-024-05991-4

Shkoporov, A. N., Khokhlova, E. V., Fitzgerald, C. B., Stockdale, S. R., Draper, L. A., Ross, R. P., & Hill, C. (2018). ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nature Communications*, *9*(1), 4781. https://doi.org/10.1038/s41467-018-07225-7

Shkoporov, A. N., Khokhlova, E. V., Stephens, N., Hueston, C., Seymour, S., Hryckowian, A. J., Scholz, D., Ross, R. P., & Hill, C. (2021). Long-term persistence of crAss-like phage crAss001 is associated with phase variation in Bacteroides intestinalis. *BMC Biology*, *19*(1), 163. https://doi.org/10.1186/s12915-021-01084-3

Shkoporov, A. N., Stockdale, S. R., Adriaenssens, E. M., Yutin, N., Koonin, E. V., Dutilh, B. E., Krupovic, M., Edwards, R. A., Tolstoy, I., & Hill, C. (2021). *Create one new order (Crassvirales) including four new families, ten new subfamilies, 42 new genera and 73 new species (Caudoviricetes)*. https://ictv.global/ictv/proposals/2021.022B.R.Crassvirales.zip

Shymialevich, D., Wójcicki, M., Świder, O., Średnicka, P., & Sokołowska, B. (2023). Characterization and genome study of a newly isolated temperate phage belonging to a new genus targeting Alicyclobacillus acidoterrestris. *Genes*, *14*(6). https://doi.org/10.3390/genes14061303

Silveira, C. B., Luque, A., & Rohwer, F. (2021). The landscape of lysogeny across microbial community density, diversity and energetics. *Environmental Microbiology*, *23*(8), 4098–4111. https://doi.org/10.1111/1462-2920.15640

Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., & Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, *14*(4), 407–410. https://doi.org/10.1038/nmeth.4184

Singh, J., Fitzgerald, D. A., Jaffe, A., Hunt, S., Barr, J. J., Iredell, J., & Selvadurai, H. (2023). Single-arm, open-labelled, safety and tolerability of intrabronchial and nebulised bacteriophage treatment in children with cystic fibrosis and Pseudomonas aeruginosa. *BMJ Open Respiratory Research*, *10*(1). https://doi.org/10.1136/bmjresp-2022-001360

Singh, S., Orr, D., Divinagracia, E., McGraw, J., Dorff, K., & Forst, S. (2015). Role of secondary metabolites in establishment of the mutualistic partnership between Xenorhabdus nematophila and the entomopathogenic nematode Steinernema carpocapsae. *Applied and Environmental Microbiology*, *81*(2), 754–764. https://doi.org/10.1128/AEM.02650-14

Smug, B. J., Szczepaniak, K., Rocha, E. P. C., Dunin-Horkawicz, S., & Mostowy, R. J. (2023). Ongoing shuffling of protein fragments diversifies core viral functions linked to interactions with bacterial hosts. *Nature Communications*, *14*(1), 7460. https://doi.org/10.1038/s41467-023-43236-9

Snyder, H., He, H., Owen, H., Hanna, C., & Forst, S. (2011). Role of Mrx fimbriae of Xenorhabdus nematophila in competitive colonization of the nematode host. *Applied and Environmental Microbiology*, *77*(20), 7247–7254. https://doi.org/10.1128/AEM.05328-11

Söding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, *33*(Web Server issue), W244-8. https://doi.org/10.1093/nar/gki408

Sorek, R., Kunin, V., & Hugenholtz, P. (2008). CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nature Reviews. Microbiology*, *6*(3), 181–186. https://doi.org/10.1038/nrmicro1793

Stachurska, X., Roszak, M., Jabłońska, J., Mizielińska, M., & Nawrotek, P. (2021). Double-layer agar (DLA) modifications for the first step of the phage-antibiotic synergy (PAS) identification. *Antibiotics (Basel, Switzerland)*, *10*(11), 1306. https://doi.org/10.3390/antibiotics10111306

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

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, *35*(11), 1026–1028. https://doi.org/10.1038/nbt.3988

Stock, S. P., Campbell, J. F., & Nadler, S. A. (2001). Phylogeny of Steinernema travassos, 1927 (Cephalobina: Steinernematidae) inferred from ribosomal DNA sequences and morphological characters. *The Journal of Parasitology*, *87*(4), 877–889. https://doi.org/10.1645/0022-3395(2001)087[0877:POSTCS]2.0.CO;2

Stock, S. Patricia, & Blair, H. G. (2008). Entomopathogenic nematodes and their bacterial symbionts: the inside out of a mutualistic association. *Symbiosis* . https://dalspace.library.dal.ca/bitstream/handle/10222/78367/VOLUME%2046-NUMBER%202-2008-PAGE%2065.pdf?sequence=1

Strachan, N. J. C., Rotariu, O., Lopes, B., MacRae, M., Fairley, S., Laing, C., Gannon, V., Allison, L. J., Hanson, M. F., Dallman, T., Ashton, P., Franz, E., van Hoek, A. H. A. M., French, N. P., George, T., Biggs, P. J., & Forbes, K. J. (2015). Whole Genome Sequencing demonstrates that Geographic Variation of Escherichia coli O157 Genotypes Dominates Host Association. *Scientific Reports*, *5*, 14145. https://doi.org/10.1038/srep14145

Strathdee, S. A., Hatfull, G. F., Mutalik, V. K., & Schooley, R. T. (2023). Phage therapy: From biological mechanisms to future directions. *Cell*, *186*(1), 17–31. https://doi.org/10.1016/j.cell.2022.11.017

Sun, C., Chen, J., Jin, M., Zhao, X., Li, Y., Dong, Y., Gao, N., Liu, Z., Bork, P., Zhao, X.-M., & Chen, W.-H. (2023). Long-Read Sequencing Reveals Extensive DNA Methylations in Human Gut Phagenome Contributed by Prevalently Phage-Encoded Methyltransferases. *Advancement of Science*, *10*(25), e2302159. https://doi.org/10.1002/advs.202302159

Sunagawa, S., Acinas, S. G., Bork, P., Bowler, C., Tara Oceans Coordinators, Eveillard, D., Gorsky, G., Guidi, L., Iudicone, D., Karsenti, E., Lombard, F., Ogata, H., Pesant, S., Sullivan, M. B., Wincker, P., & de Vargas, C. (2020). Tara Oceans: towards global ocean ecosystems biology. *Nature Reviews. Microbiology*, *18*(8), 428–445. https://doi.org/10.1038/s41579-020-0364-5

Suttle, C. A. (2007). Marine viruses--major players in the global ecosystem. *Nature Reviews. Microbiology*, *5*(10), 801–812. https://doi.org/10.1038/nrmicro1750

Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution*, *38*(7), 3022–3027. https://doi.org/10.1093/molbev/msab120

Tars, K. (2020). ssRNA Phages: Life Cycle, Structure and Applications. In *Biocommunication of Phages* (pp. 261–292). Springer International Publishing. https://doi.org/10.1007/978-3-030-45885-0_13

Teney, C., Poupelin, J.-C., Briot, T., Le Bouar, M., Fevre, C., Brosset, S., Martin, O., Valour, F., Roussel-Gaillard, T., Leboucher, G., Ader, F., Lukaszewicz, A.-C., Ferry, T., & PHAGEinLYON Clinic Study Group. (2024). Phage therapy in a burn patient colonized with extensively drug-resistant Pseudomonas aeruginosa responsible for relapsing ventilator-associated pneumonia and bacteriemia. *Viruses*, *16*(7), 1080. https://doi.org/10.3390/v16071080

Terzian, P., Olo Ndela, E., Galiez, C., Lossouarn, J., Pérez Bucio, R. E., Mom, R., Toussaint, A., Petit, M.-A., & Enault, F. (2021). PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genomics and Bioinformatics*, *3*(3), lqab067. https://doi.org/10.1093/nargab/lqab067

Tesson, F., Hervé, A., Mordret, E., Touchon, M., d'Humières, C., Cury, J., & Bernheim, A. (2022). Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nature Communications*, *13*(1), 2561. https://doi.org/10.1038/s41467-022-30269-9

Thorell, K., Yahara, K., Berthenet, E., Lawson, D. J., Mikhail, J., Kato, I., Mendez, A., Rizzato, C., Bravo, M. M., Suzuki, R., Yamaoka, Y., Torres, J., Sheppard, S. K., & Falush, D. (2017). Rapid evolution of distinct Helicobacter pylori subpopulations in the Americas. *PLoS Genetics*, *13*(2), e1006546. https://doi.org/10.1371/journal.pgen.1006546

Torres-Barceló, C., Turner, P. E., & Buckling, A. (2022). Mitigation of evolved bacterial resistance to phage therapy. *Current Opinion in Virology*, *53*, 101201. https://doi.org/10.1016/j.coviro.2022.101201

Tully, B. (2016). *Quality Assessment: FastQC v1*. https://doi.org/10.17504/protocols.io.fa3bign

Turner, D., Adriaenssens, E. M., Tolstoy, I., & Kropinski, A. M. (2021). Phage Annotation Guide: Guidelines for Assembly and High-Quality Annotation. *PHAGE (New Rochelle, N.Y.)*, *2*(4), 170–182. https://doi.org/10.1089/phage.2021.0013

Turner, D., Shkoporov, A. N., Lood, C., Millard, A. D., Dutilh, B. E., Alfenas-Zerbini, P., van Zyl, L. J., Aziz, R. K., Oksanen, H. M., Poranen, M. M., Kropinski, A. M., Barylski, J., Brister, J. R., Chanisvili, N., Edwards, R. A., Enault, F., Gillis, A., Knezevic, P., Krupovic, M., … Adriaenssens, E. M. (2023). Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Archives of Virology*, *168*(2), 74. https://doi.org/10.1007/s00705-022-05694-2

Uyttebroek, S., Chen, B., Onsea, J., Ruythooren, F., Debaveye, Y., Devolder, D., Spriet, I., Depypere, M., Wagemans, J., Lavigne, R., Pirnay, J.-P., Merabishvili, M., De Munter, P., Peetermans, W. E., Dupont, L., Van Gerven, L., & Metsemakers, W.-J. (2022). Safety and efficacy of phage therapy in difficult-to-treat infections: a systematic review. *The Lancet Infectious Diseases*, *22*(8), e208–e220. https://doi.org/10.1016/S1473-3099(21)00612-5

Valencia-Toxqui, G., & Ramsey, J. (2024). How to introduce a new bacteriophage on the block: a short guide to phage classification. *Journal of Virology*, *98*(10), e0182123. https://doi.org/10.1128/jvi.01821-23

van Dijk, E. L., Naquin, D., Gorrichon, K., Jaszczyszyn, Y., Ouazahrou, R., Thermes, C., & Hernandez, C. (2023). Genomics in the long-read sequencing era. *Trends in Genetics: TIG*, *39*(9), 649–671. https://doi.org/10.1016/j.tig.2023.04.006

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Gilchrist, C. L. M., Söding, J., & Steinegger, M. (2022). Foldseek: fast and accurate protein structure search. In *bioRxiv* (p. 2022.02.07.479398). https://doi.org/10.1101/2022.02.07.479398

Van Valen, L. (1973). A New evolutionary law. *Evol Theory*, *1*, 1–30. https://www.mn.uio.no/cees/english/services/van-valen/evolutionary-theory/volume-1/vol-1-no-1-pages-1-30-l-van-valen-a-new-evolutionary-law.pdf

VanInsberghe, D., Arevalo, P., Chien, D., & Polz, M. F. (2020). How can microbial population genomics inform community ecology? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *375*(1798), 20190253. https://doi.org/10.1098/rstb.2019.0253

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., … Velankar, S. (2022). AlphaFold Protein Structure Database:

massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, *50*(D1), D439–D444. https://doi.org/10.1093/nar/gkab1061

Vos, M., & Didelot, X. (2009). A comparison of homologous recombination rates in bacteria and archaea. *The ISME Journal*, *3*(2), 199–208. https://doi.org/10.1038/ismej.2008.93

Waldor, M. K., & Mekalanos, J. J. (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin [Review of *Lysogenic conversion by a filamentous phage encoding cholera toxin*]. *Science*, *272*(5270), 1910–1914. https://doi.org/10.1126/science.272.5270.1910

Walker, P. J., Siddell, S. G., Lefkowitz, E. J., Mushegian, A. R., Adriaenssens, E. M., Alfenas-Zerbini, P., Dempsey, D. M., Dutilh, B. E., García, M. L., Curtis Hendrickson, R., Junglen, S., Krupovic, M., Kuhn, J. H., Lambert, A. J., Łobocka, M., Oksanen, H. M., Orton, R. J., Robertson, D. L., Rubino, L., … Zerbini, F. M. (2022). Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). *Archives of Virology*, *167*(11), 2429–2440. https://doi.org/10.1007/s00705-022-05516-5

Wandro, S., Ghatbale, P., Attai, H., Hendrickson, C., Samillano, C., Suh, J., Dunham, S. J. B., Pride, D. T., & Whiteson, K. (2022). Phage cocktails constrain the growth of Enterococcus. *MSystems*, *7*(4), e0001922. https://doi.org/10.1128/msystems.00019-22

Wang, J., Dai, W., Li, J., Xie, R., Dunstan, R. A., Stubenrauch, C., Zhang, Y., & Lithgow, T. (2020). PaCRISPR: a server for predicting and visualizing anti-CRISPR proteins. *Nucleic Acids Research*, *48*(W1), W348–W357. https://doi.org/10.1093/nar/gkaa432

Wang, Y., Subedi, D., Li, J., Wu, J., Ren, J., Xue, F., Dai, J., Barr, J. J., & Tang, F. (2022). Phage Cocktail Targeting STEC O157:H7 Has Comparable Efficacy and Superior Recovery Compared with Enrofloxacin in an Enteric Murine Model. *Microbiology Spectrum*, *10*(3), e0023222. https://doi.org/10.1128/spectrum.00232-22

Weitz, J. S., Poisot, T., Meyer, J. R., Flores, C. O., Valverde, S., Sullivan, M. B., & Hochberg, M. E. (2013). Phage-bacteria infection networks. *Trends in Microbiology*, *21*(2), 82–91. https://doi.org/10.1016/j.tim.2012.11.003

Welivita, A., Perera, I., Meedeniya, D., Wickramarachchi, A., & Mallawaarachchi, V. (2018). Managing Complex Workflows in Bioinformatics: An Interactive Toolkit With GPU Acceleration. *IEEE Transactions on Nanobioscience*, *17*(3), 199–208. https://doi.org/10.1109/TNB.2018.2837122

Westmoreland, B. C., Szybalski, W., & Ris, H. (1969). Mapping of deletions and substitutions in heteroduplex DNA molecules of bacteriophage lambda by electron microscopy. *Science (New York, N.Y.)*, *163*(3873), 1343–1348. https://doi.org/10.1126/science.163.3873.1343

Wick, R. R. (2018). *Filtlong: Tool for filtering long reads by quality* (0.2.20) [Computer software]. https://github.com/rrwick/Filtlong/

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology*, *13*(6), e1005595. https://doi.org/10.1371/journal.pcbi.1005595

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* , *31*(20), 3350–3352. https://doi.org/10.1093/bioinformatics/btv383

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Williamson, K. E., Fuhrmann, J. J., Wommack, K. E., & Radosevich, M. (2017). Viruses in soil ecosystems: An unknown quantity within an unexplored territory. *Annual Review of Virology*, *4*(1), 201–219. https://doi.org/10.1146/annurev-virology-101416-041639

Wolput, S., Lood, C., Fillol-Salom, A., Casters, Y., Albasiony, A., Cenens, W., Vanoirbeek, K., Kerremans, A., Lavigne, R., Penadés, J. R., & Aertsen, A. (2024). Phage-host co-evolution has led to distinct generalized transduction strategies. *Nucleic Acids Research*, *52*(13), 7780–7791. https://doi.org/10.1093/nar/gkae489

Wu, Y., Wang, R., Xu, M., Liu, Y., Zhu, X., Qiu, J., Liu, Q., He, P., & Li, Q. (2019). A Novel Polysaccharide Depolymerase Encoded by the Phage SH-KP152226 Confers Specific Activity Against Multidrug-Resistant Klebsiella pneumoniae via Biofilm Degradation. *Frontiers in Microbiology*, *10*, 2768. https://doi.org/10.3389/fmicb.2019.02768

Ye, Y., & Godzik, A. (2003). Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics (Oxford, England)*, *19*(suppl_2), ii246–ii255. https://doi.org/10.1093/bioinformatics/btg1086

Yirmiya, E., Leavitt, A., Lu, A., Ragucci, A. E., Avraham, C., Osterman, I., Garb, J., Antine, S. P., Mooney, S. E., Hobbs, S. J., Kranzusch, P. J., Amitai, G., & Sorek, R. (2024). Phages overcome bacterial immunity via diverse anti-defence proteins. *Nature*, *625*(7994), 352–359. https://doi.org/10.1038/s41586-023-06869-w

Yutin, N., Benler, S., Shmakov, S. A., Wolf, Y. I., Tolstoy, I., Rayko, M., Antipov, D., Pevzner, P. A., & Koonin, E. V. (2021). Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nature Communications*, *12*(1), 1044. https://doi.org/10.1038/s41467-021-21350-w

Zhang, Yang, Su, X., Harris, A. J., Caraballo-Ortiz, M. A., Ren, Z., & Zhong, Y. (2018). Genetic Structure of the Bacterial Endosymbiont Buchnera aphidicola from Its Host Aphid Schlechtendalia chinensis and Evolutionary Implications. *Current Microbiology*, *75*(3), 309–315. https://doi.org/10.1007/s00284-017-1381-0

Zhang, Yuqian, Li, H., Shen, Y., Wang, S., Tian, L., Yin, H., Shi, J., Xing, A., Zhang, J., Ali, U., Sami, A., Chen, X., Gao, C., Zhao, Y., Lyu, Y., Wang, X., Chen, Y., Tian, Z., Wu, S.-B., & Wu, L. (2023). Widespread readthrough events in plants reveal unprecedented plasticity of stop codons. In *bioRxiv*. https://doi.org/10.1101/2023.03.20.533458

Zheng, X.-F., Yang, Z.-Q., Zhang, H., Jin, W.-X., Xu, C.-W., Gao, L., Rao, S.-Q., & Jiao, X.-A. (2020). Isolation of virulent phages infecting dominant mesophilic aerobic bacteria in cucumber pickle fermentation. *Food Microbiology*, *86*(103330), 103330. https://doi.org/10.1016/j.fm.2019.103330

CHAPTER 8
APPENDICES

Appendix A: Signed Co-authorship forms

Office of Graduate Research
Room 003, Registry Building
Bedford Park, SA 5042
GPO Box 2100, Adelaide 5001 Australia
Email: hdrexams@flinders.edu.au
Phone: (08) 8201 5961
Website: https://students.flinders.edu.au/my-course/hdr
CRICOS Provider: 00114A

CO-AUTHORSHIP APPROVALS FOR HDR THESIS FOR EXAMINATIONS

In accordance with Clause 5, 7 and 8 in the HDR Thesis Rules, a student must sign a declaration that the thesis does not contain any material previously published or written by another person except where due reference is made in the text or footnotes. There can be no exception to this rule.

a. Publications or significant sections of publications (whether accepted, submitted or in manuscript form) arising out of work conducted during candidature may be included in the body of the thesis, or submitted as additional evidence as an appendix, on the following conditions:
   I. they contribute to the overall theme of the work, are conceptually linked to the chapters before and after, and follow a logical sequence
   II. they are formatted in the same way as the other chapters (i.e. not presented as reprints unless as an appendix), whether included as separate chapters or integrated into chapters
   III. they are in the same typeface as the rest of the thesis (except for reprints included as an appendix)
   IV. published and unpublished sections of a chapter are clearly differentiated with appropriate referencing or footnotes, and
   V. unnecessary repetition in the general introduction and conclusion, and the introductions and conclusions of each published chapter, is avoided.

b. Multi-author papers may be included within a thesis, provided:
   I. the student is the primary author
   II. there is a clear statement in prose for each publication at the front of each chapter, recording the percentage contribution of each author to the paper, from conceptualisation to realisation and documentation.
   III. The publication adheres to Flinders Authorship of Research Output Procedures, and
   IV. each of the other authors provides permission for use of their work to be included in the thesis on the Submission of Thesis Form below.

c. Papers where the student is not the primary author may be included within a thesis if a clear justification for the paper's inclusion is provided, including the circumstances relating to production of the paper and the student's position in the list of authors. However, it is preferable to include such papers as appendices, rather than in the main body of the thesis.

STUDENT DETAILS

Student Name	Bhavya Nalagamapalli Papudeshi
Student ID	2247247
College	College of Science and Engineering
Degree	PhD
Title of Thesis	Decoding microbe host interactions

CO-AUTHORSHIP APPROVALS FOR HDR THESIS EXAMINATION

PUBLICATION 1

This section is to be completed by the student and co-authors. If there are more than four co-authors (student plus 3 others), only the three co-authors with the most significant contributions are required to sign below.

Please note: A copy of this page will be provided to the Examiners.

Full Publication Details	Sussana R. Grigson, Sarah K. Giles, Robert A. Edwards and Bhavya Papudeshi. "Knowing and naming: phage annotation and nomenclature for phage therapy." Clinical Infectious Diseases 77, no. Supplement_5 (2023): S352-S359.
Section of thesis where publication is referred to	Chapter 1 Literature Review

Student's contribution to the publication

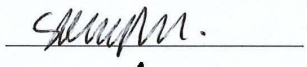
90	%	Research design
90	%	Data collection and analysis
60	%	Writing and editing

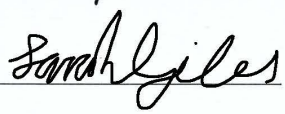
Outline your (the student's) contribution to the publication:

I contributed by writing the manuscript, structuring the manuscript, and contributing to the overall editing process.

APPROVALS

By signing the section below, you confirm that the details above are an accurate record of the students contribution to the work.

Name of Co-Author 1 Susanna R. Grigson Signed  Date 10/6/24

Name of Co-Author 2 Dr. Sarah K. Giles Signed  Date 11/6/25

Name of Co-Author 3 Prof. Robert Edwards Signed  Date 10/6/24

CO-AUTHORSHIP APPROVALS FOR HDR THESIS EXAMINATION

PUBLICATION 2

This section is to be completed by the student and co-authors. If there are more than four co-authors (student plus 3 others), only the three co-authors with the most significant contributions are required to sign below.

Please note: A copy of this page will be provided to the Examiners.

Full Publication Details	Bhavya Papudeshi, Michael J. Roach, Vijini Mallawaarachchi, George Bouras, Susanna R. Grigson, Sarah K. Giles, Clarice M. Harker, Abbey L. K. Hutton, Anita Tarasenko, Laura K. Inglis, Alejandro A. Vega, Cole Souza, Lance Boling, Hamza Hajama, Ana Georgina Cobián Güemes, Anca M. Segall, Elizabeth A. Dinsdale, and Robert A. Edwards. "Sphae: an automated toolkit for predicting phage therapy candidates from sequencing data." Bioinformatics Advances 5, no. 1 (2025): vbaf004.
Section of thesis where publication is referred to	Chapter 2

Student's contribution to the publication

80	%	Research design
70	%	Data collection and analysis
70	%	Writing and editing

Outline your (the student's) contribution to the publication:

As the first author of this study, I was responsible for conceptualising and developing the Sphae toolkit, implementing the computational workflow, performing data analysis, writing the manuscript, and contributing to its editing.

APPROVALS

By signing the section below, you confirm that the details above are an accurate record of the students contribution to the work.

Name of Co-Author 1 Dr. Vijini Mallawaarachchi Signed Vijini Date 10/06/2025

Name of Co-Author 2 Susanna R. Grigson Signed Susanna. Date 10/6/2025

Name of Co-Author 3 Prof. Robert A. Edwards Signed Rob J Edw Date 10/6/25

CO-AUTHORSHIP APPROVALS FOR HDR THESIS EXAMINATION

PUBLICATION 3

This section is to be completed by the student and co-authors. If there are more than four co-authors (student plus 3 others), only the three co-authors with the most significant contributions are required to sign below.

Please note: A copy of this page will be provided to the Examiners.

Full Publication Details	Bhavya Papudeshi, Alejandro A. Vega, Cole Souza, Sarah K. Giles, Vijini Mallawaarachchi, Michael J. Roach, Michelle An, Nicole Jacobson, Katelyn McNair, Maria Fernanda Mora, Karina Pastrana, Lance Boling, Christopher Leigh, Clarice Harker, Will S. Plewa, Susanna R. Grigson, George Bouras, Przemyslaw Decewicz, Antoni Luque, Lindsay Droit, Scott A. Handley, David Wang, Anca M. Segall, Elizabeth A. Dinsdale, and Robert A. Edwards. "Host interactions of novel *Crassvirales* species belonging to multiple families infecting bacterial host, *Bacteroides cellulosilyticus* WH2." Microbial Genomics 9, no. 9 (2023): 001100.
Section of thesis where publication is referred to	Chapter 3

Student's contribution to the publication

60	%	Research design
50	%	Data collection and analysis
70	%	Writing and editing

Outline your (the student's) contribution to the publication:

As the first author of this study, I was responsible for performing the data analysis, writing the manuscript, and contributing to its editing.

APPROVALS

By signing the section below, you confirm that the details above are an accurate record of the students contribution to the work.

Name of Co-Author 1 Dr. Vijini Mallawaarachchi Signed Vijini Date 10/06/2025

Name of Co-Author 2 Susanna R. Grigson Signed [signature] Date 10/6/2025

Name of Co-Author 3 Prof. Robert Edwards Signed [signature] Date 10/6/25

CO-AUTHORSHIP APPROVALS FOR HDR THESIS EXAMINATION

PUBLICATION 1

This section is to be completed by the student and co-authors. If there are more than four co-authors (student plus 3 others), only the three co-authors with the most significant contributions are required to sign below.

Please note: A copy of this page will be provided to the Examiners.

Full Publication Details	Papudeshi, B., Rusch, D. B., VanInsberghe, D., Lively, C. M., Edwards, R. A., & Bashey, F. (2023). Host Association and Spatial Proximity Shape but Do Not Constrain Population Structure in the Mutualistic Symbiont Xenorhabdus bovienii. mBio, e00434-23.
Section of thesis where publication is referred to	This pubication is included in chapter 1: Bacterial genome assembly and annotation. Under subsection, 1a. Factors driving population structure in Xenorhabdus bovienii symbionts This publication is chapter 4 in my final version of my thesis

Student's contribution to the publication		
50%	%	Research design
60%	%	Data collection and analysis
70%	%	Writing and editing

Outline your (the student's) contribution to the publication:

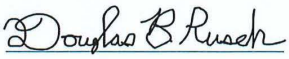
I played a significant role in the research design and data analysis phases, as well as a substantial role in the writing and editing phases of the publication.

APPROVALS

By signing the section below, you confirm that the details above are an accurate record of the students contribution to the work.

Name of Co-Author 1 Dr. Farrah Bashey Signed  Date May 12, 2023

Name of Co-Author 2 Dr. Robert Edwards Signed  Date 10/6/23

Name of Co-Author 3 Dr. Douglas Rusch Signed  Date May 30th, 2023

Appendix B : Achievements

Conference presentations

Oral Presentation

1. “Mechanisms of genetic variation in *Bacteroides* phages” - CERVAID (Computational and Experimental Resources for Virome Analysis in Inflammatory Bowel Disease), San Diego, USA, May 2022 (Chapter 4)
2. “Mechanisms of genetic variation in *Bacteroides* phages” – Nanopore Day, Adelaide, Australia, June 2022 (Chapter 4)
3. “Host specificity of abundant bacterial virus found within the human gut”- Molecular Science and Technology HDR(Higher Degree Research) Conference, Adelaide, Australia, December 2022 (Chapter 4)
4. “Novel crAssphage isolates exhibit conserved gene order and purifying selection of the host specificity protein” - Phage Bites symposium, Online, March 2023 (Chapter 4)
5. **Invited speaker:** “The crAssphage host specificity puzzle” - Australian Society for Microbiology, Perth, Australia, July 2023 (Chapter 4)
6. “Maximising bioinformatic workflow scalability with workflow managers to study microbes” - Data and Information Science Research HDR Conference, Adelaide, Australia, December 2023 (Chapter 3)
7. “Sphae: Phage assembly and annotation workflow” – Microseq, Online, September 2024
8. **Invited speaker:** “Bioinformatics and Phage Therapy” – Bioinformatics Meeting, University of Australia, Adelaide, Australia, June 2025

Poster presentations

9. **Best poster award:** “Sphae: Phage assembly and annotation workflow”- International Conference on Bacteriophage Research and Antimicrobial Resistance, Chennai, India, September 2023 (Chapter 3)
10. **Best poster award:** “Sphae: Phage assembly and annotation workflow” - MST HDR Conference, Adelaide, Australia, October 2023 (Chapter 3)
11. “Sphae: Phage assembly and annotation workflow” - ABACBS (Australian Bioinformatics and Computational Biology Society) Conference, Brisbane, Australia, December 2023 (Chapter 3)
12. “Sphae: Phage assembly and annotation workflow” – VoM(Viruses of Microbes) Conference, Cairns, Australia, July 2024 (Chapter 3)
13. “Sphae: Phage assembly and annotation workflow” - ABACBS (Australian Bioinformatics and Computational Biology Society) Conference, Sydney, Australia, November 2024 (Chapter 3)

Grant applications

Successful

1. Cooperative Research Time in Mining Economies (CRC-TIME) PhD Top-up scholarship
2. Flinders University of Student Association (FUSA) Development Award, 2022
3. Flinders University Conference Travel Grant, 2022
4. CSE Higher Degree by Research International Conference Support Scheme, 2023
5. Australian Bioinformatics and Computational biology Society (ABACBS/COMBINE) Conference Travel Grant, 2023
6. CSE Higher Degree by Research International Conference Support Scheme, 2024
7. Flinders University of Student Association (FUSA) Development Award, 2024

Unsuccessful

8. Flinders University of Student Association (FUSA) Development Award, 2023
9. Australian Society for Microbiology SA/NT Chapter, Student Awards 2023
10. StudyAdelaide International Student Awards 2023
11. College of Science and Engineering Publication Award 2023
12. HDR Leadership and Scholarly Excellence Award, 2024

Workshops

1. Taught a section of “Mobile Genetic Elements Special Interest Group: Introduction to transposable elements and bacteriophages” – Australian Society for Microbiology (ASM) Workshop, Adelaide, Australia, July 2025
2. Taught a section in “Nextflow vs Snakemake” – South Australian Genomics Centre Workshop, Adelaide, Australia, July 2025
3. Organised “Intro to Python” – Australian Bioinformatics student society (COMBINE) workshop, Adelaide 2024
4. Supported hackathon on “Analysis of Cystic Fibrosis Metagenomics”, Adelaide, 2023
5. Taught “ Getting started in Bioinformatics” – COMBINE workshop, Adelaide, 2023
6. Taught sections of “Metagenomics workshop” – South Australia Genomics Centre Workshop, Adelaide, 2022

Associations and Service

- President, COMBINE, the Australian Student Bioinformatics Society, 2024
- Flinders University, Molecular Society and Technology HDR Conference Organising committee, 2024
- Training and Events Coordinator, COMBINE student committee, 2023

Peer review

Peer-reviewed journal articles for

- One article for BMC genomics
- One article for Journal of virology
- Two articles for Royal Society Open Science

Appendix C : Phage Submissions and Naming

Crassvirales submission to ICTV April 2023

Copy of the submitted form

Part 1: TITLE, AUTHORS, APPROVALS, etc

Code assigned:	**to be assigned by ICTV officers**
Short title: Create two new species in the genus *Kehishuvirus,* and *Kolpuevirus,* and one new genus in order *Crassvirales*	

Author(s) and email address(es)

Papudeshi B, Vega AA, Souza C, Giles SK, Mallawaarachchi V, Roach MJ, An M, Jacobson N, McNair K, Mora MF, Pastrana K, Boling L, Leigh C, Harker C, Plewa WS, Grigson SR, Bouras G, Decewicz P, Luque A, Droit L, Handley SA, Wang D, Segall AM, Dinsdale EA, Edwards RA	nala0006@flinders.edu.au; alexvega619@gmail.com; colesouza017@gmail.com; sarah.giles@flinders.edu.au; mall0133@flinders.edu.au; michael.roach@flinders.edu.au; michellean92@gmail.com; njacobson@sdsu.edu; deprekate@gmail.com; moramariaf21@gmail.com; kpastrana0331@sdsu.edu; liquidgrey@gmail.com; chris.leigh@adelaide.edu.au; clarice.cram@flinders.edu.au; will.plewa@flinders.edu.au; p.decewicz@uw.edu.pl; susie.grigson@flinders.edu.au; george.bouras@adelaide.edu.au; aluque@sdsu.edu; ldroit@wustl.edu; shandley@wustl.edu; davewang@wustl.edu; asegall@sdsu.edu; elizabeth.dinsdale@flinders.edu.au; robert.edwards@flinders.edu.au

Author(s) institutional address(es) (optional)

Flinders University, Adelaide, Australia [BP, SKG, VM, MJR, CH, WSP, SRG, EAD, RAE]
San Diego State University, San Diego, USA [AAV, CS, MA, NJ, KM, MFM, KP, LB, AL, AMS]
University of Adelaide, Adelaide, Australia [CL, GB]

University of Warsaw, Warsaw, Poland [PD] Washington University School of Medicine, St. Louis, USA [LD, SH, DW]

Corresponding author

Robert A. Edwards, member of ICTV *Crassvirales* phages Study Group

List the ICTV Study Group(s) that have seen this proposal

ICTV Bacterial Viruses Subcommittee, *Crassvirales* phages Study Group

ICTV Study Group comments and response of proposer

ICTV Study Group votes on proposal

Study Group	**Number of members**		
	Votes support	**Votes against**	**No vote**

Authority to use the name of a living person

Is any taxon name used here derived from that of a living person (Y/N)	N

Taxon name	**Person from whom the name is derived**	**Permission attached (Y/N)**

Submission dates

Date first submitted to SC Chair	May, 2023
Date of this revision (if different to above)	

ICTV-EC comments and response of the proposer

Part 2: <u>NON-TAXONOMIC PROPOSAL</u>

Text of proposal

Part 3: TAXONOMIC PROPOSAL

Name of accompanying Excel module

2023.001B.Ud.v1.Crassvirales3Species.xlsx

Abstract

The isolation of *Crassvirales in vitro* remains a challenge with only four successful pure isolates since the discovery of the first *Crassvirales* species in 2014. However, over 600 *Crassvirales* phages have been identified from metagenomes. In our study, we successfully isolated three novel *Crassvirales* phages from wastewater that infect the bacterial host *Bacteroides cellulosilyticus* WH2. Following the taxonomic demarcation and naming convention proposed by the ICTV for *Crassvirales,* we propose two novel species, *Kehishuvirus tikkala* (Bc01), *Kolpuevirus frurule* (Bc03), and a new genus, *Rudgehvirus jaberico* (Bc11).

Text of proposal

The three isolates' phages were visualised using transmission electron microscopy (TEM) and their genomes were sequenced. A nucleotide BLAST search against the nr database of the assembled genomes confirm their closely related genomes are other crass-like phages. Additionally, transmission electron micrographs revealed that the three phages share a podovirus morphology and have a genome length of 100kb which is consistent with the other crass-like phages.

To determine their taxonomic assignment, we followed the *Crassvirales* order demarcation criteria. The genera within *Crassvirales* were defined based on the topology of the protein phylogenetic trees, which showed at least 80% shared orthologous groups. The species demarcation criteria included 95% nucleotide sequence identity over 85% of the complete genome length.

Following the above taxonomic classification, we confirmed the three crAss-like phages isolated represent three novel species. Here are the details of each species:

Kehishuvirus tikkala: This genome shares 80% orthologous genes within known genera, specifically *Kehishuvirus.* At the species level, the most closely related strain to this genome is *Kehishuvirus primarius,* with a sequence identity of 95.67% across 80.98% genome.

Kolpuevirus frurule: This genome also shares 80% orthologous genes and is classified within the existing genus "*Kolpuevirus"*. However, it represents a novel species as it is most similar to the genome *Kolpuevirus hominis,* with an 82.58% sequence identity across 55.01% of the genome. We propose to call this new species *Kolpuevirus frurule.*

Rudgehvirus jaberico: This genome was classified at the family level under *Intestiviridae,* with its closest related genome being *Jahgtovirus intestinalis* sharing 74.75% identity across 9.86%. Phylogenetic classification of the three conserved genes, portal protein, terminase large subunit and major capsid protein - suggests that these isolate forms a neighbouring clade to its closely related genome. Following the ICTV *Crassvirales* genus naming convention, we propose to call this genus "*Rudgehvirus*" after Ridgeback dog breed.

Supporting evidence

History: Since discovery of one of the most abundant bacteriophages in the human gut microbiome in 2014, crAssphage has been of interest [1]. Since there have been over 600 other crass-like phages that share some similarity with the crAssphage [4]. In 2021, these crass-like phages have been classified into a formal taxonomic system including a new order to represent all of them *Crassvirales* [3]. This order includes four new families, ten new subfamilies, 42 new genera and a total of 73 new species (Taxonomy Proposal 2021.021B.A.Crassvirales).

Here we are presenting the supporting evidence on how we are classifying the three novel crass-like phages isolated [2] into the *Crassvirales* order.

Podovirus morphology

All three phages isolated in Papudeshi et al., 2023 [2] were confirmed to have podovirus morphology when visualised as an electron micrograph (Figure 1).

A) *Kehishuvirus tikkala* B) *Kolpuevirus frurule* C) *Rudgehvirus jaberico*

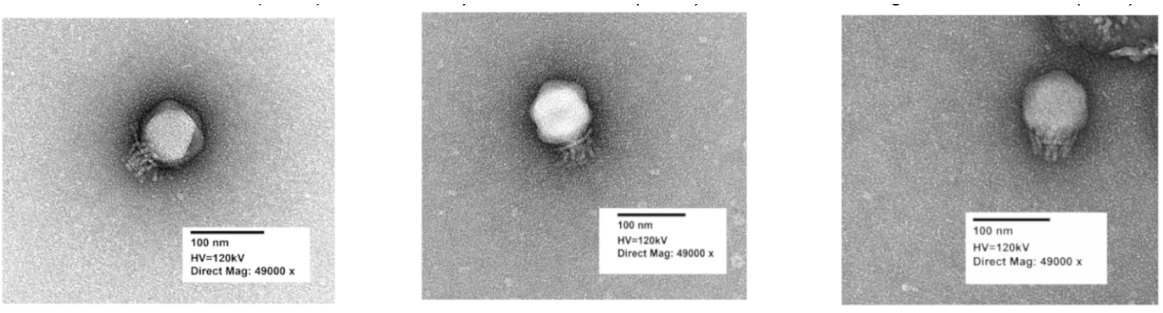


Figure 1: TEM image of negatively stained crass-like virions, under 49,000x magnification. Image taken from Papudeshi et al., 2023 [2].

Phylogenomics

Taxonomic classification of the three phages were assigned using three signals, 1) phylogeny 2) orthologous genes shared 3) nucleotide identity using a CrassUS program (https://github.com/dcarrillox/CrassUS).

Open reading frame (ORFs) were predicted Prodigal. In this study, a revised version of prodigal [4] was employed by CrassUS, to specifically detect codon reassignment within the three isolated phages. The genomes were annotated to predict the ORFs using both the standard codon table, and the codon table with TAG and TGA reassigned, as observed in some of the *Crassvirales* phages[4]. Upon analysing coding potential, the highest values were observed when using the standard codon table. These identified ORFs were subsequently utilised for taxonomic classification.

1. **Orthologous genes**
   Orthologous genes were analysed for the three phages, after predicting the open reading frames (ORFs) using revised Prodigal bioinformatic tool. Amino acid sequences of the predicted ORFs were aligned against known *Crassvirales* genome protein clusters from Yutin et al [4], using mmseqs2 v13.45. The clustered proteins were then used to build presence/absence matrix. If the genome shared 80% of its proteins with a known genus, the genome is assigned a taxon (Table 1).

   Table 1: Taxonomic classification based on shared orthologous groups when compared against known *Crassvirales* genomes.

Genome	Reference shared proteins	Most similar family	Most similar subfamily	Most similar genus
Bc01	84.0	*Steigviridae*	*Asinivirinae*	*Kehishuvirus*
Bc03	82.4	*Steigviridae*	*Asinivirinae*	*Kolpuevirus*
Bc11	50.0	*Intestiviridae*	-	-

2. **Nucleotide identity**
   Average nucleotide identity was calculated, comparing each isolate to the known *Crassvirales* genomes using BLAST alignment. From the BLAST results, the average nucleotide identity and query coverage was calculated using scripts in CrassUS (https://github.com/dcarrillox/CrassUS). If

the genome shared 95% DNA sequence identity over 85% query coverage to a complete reference genome, to assign taxonomy to species level classification (Table 2).

Table 2: Taxonomic classification based on shared nucleotide identity when compared against known *Crassvirales* genomes.

Genome	Most similar reference species	Percent identity (pid)	Query coverage
Bc01	*Kehishuvirus primarius*	95.51	79.08
Bc03	*Kolpuevirus hominis*	82.79	53.73
Bc11	*Jahgtovirus intestinalis*	74.72	9.86

3. **Phylogeny**

Three conserved proteins, large terminase subunit, portal, and major capsid proteins from the genome annotations were used for phylogenetic reconstruction. These genes are then aligned using MAFFT v7.49, the poorly aligned regions are then trimmed using trimal v1.4.1. The resulting alignment was used to infer phylogenetic relationship FastTree Version 2.1.10 that generated maximum likelihood tree using the Jones-Taylor-Thornton (JTT) model and Continuous Rate Ancestral State Reconstruction (CAT) approximation with 20 rate categories to account for heterogeneity in substitution rates across the alignment. The resulting trees were visualised using iTol (Figure 2), and the outgroup is set to *Cellulophaga phage phi13:2.*

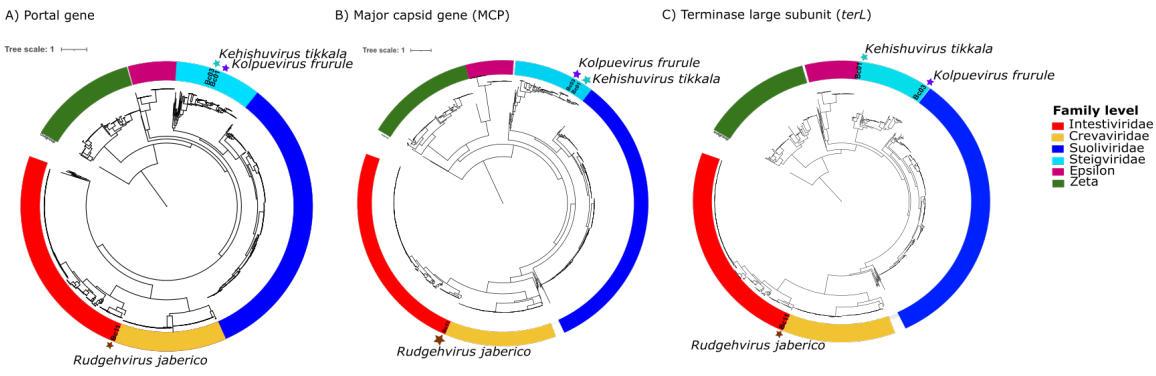


Figure 2: Maximum likelihood phylogenetic tree of known *Crassvirales* phages with *Cellulophaga phage phi13:2* set as the outgroup. The phylogenetic trees were plotted with three conserved genes A) portal, B) major capsid protein (MCP), and C) terminase large subunit (*terL*). The tips of the tree are colour coded

based on family level classification, with the three isolates highlighted in bold, along with the proposed names for the three phages. This figure was taken from Papudeshi et al., 2023 [2]

References

1. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GG, Boling L, Barr JJ, Speth DR, Seguritan V, Aziz RK, Felts B, Dinsdale EA, Mokili JL, Edwards RA (2014) A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun 5:4498
2. Papudeshi B, Vega AA, Souza C, Giles SK, Mallawaarachchi V, Roach MJ, An M, Jacobson N, McNair K, Mora MF, Pastrana K, Leigh C, Cram C, Plewa WS, Grigson SR, Bouras G, Decewicz PC, Luque A, Droit L, Handley SA, Segall AM, Dinsdale EA, Edwards RA (2023) Novel crAssphage isolates exhibit conserved gene order and purifying selection of the host specificity protein. bioRxiv
3. Turner D, Shkoporov AN, Lood C, Millard AD, Dutilh BE, Alfenas-Zerbini P, van Zyl LJ, Aziz RK, Oksanen HM, Poranen MM, Kropinski AM, Barylski J, Brister JR, Chanisvili N, Edwards RA, Enault F, Gillis A, Knezevic P, Krupovic M, Kurtboke I, Kushkina A, Lavigne R, Lehman S, Lobocka M, Moraru C, Moreno Switt A, Morozova V, Nakavuma J, Reyes Munoz A, Rumnieks J, Sarkar BL, Sullivan MB, Uchiyama J, Wittmann J, Yigang T, Adriaenssens EM (2023) Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. Arch Virol 168:74
4. Yutin N, Makarova KS, Gussow AB, Krupovic M, Segall A, Edwards RA, Koonin EV (2018) Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. Nat Microbiol 3:38-46

Proposed Taxonomy sheet

Realm	Subrealm	Kingdom	Subkingdom	Phylum	Subphylum	Class	Subclass	Order	Suborder	Family	Subfamily	Genus	Subgenus	Species	Realm	Subrealm	Kingdom	Subkingdom	Phylum	Subphylum	Class	Subclass	Order	Suborder	Family	Subfamily	Genus	Subgenus	**Species**	Exemplar GenBank Accession Number	Exemplar virus name	Virus name abbreviation	Exemplar isolate designation	*Genome coverage*	*Genome composition*	*Host/ Source*	*Change*	*ProposedRank*
															Duplodnaviria		*Heunggongvirae*		*Uroviricota*		*Caudoviricetes*		*Crassvirales*		*Steigviridae*	*Asiniviridae*	*Kehishuvirus*		*Kehishuvirustikkala*	QQ198717	*Bacteroides cellulosilyticus* phage 1	Bc01		CG	dsDNA	bacteria	Create new	species
															Duplodnaviria		*Heunggongvirae*		*Uroviricota*		*Caudoviricetes*		*Crassvirales*		*Steigviridae*	*Asiniviridae*	*Kolpuevirus*		*Kolpuevirusfruule*	QQ198718	*Bacteroides cellulosilyticus* ph	Bc03		CG	dsDNA	bacteria	Create new	species

																															age 3							
															Duplodnaviria		*Heunggongvirae*		*Uroviricota*		*Caudoviricetes*		*Crassvirales*		*Intestiviridae*		*Rudgehvirus*							CG	dsDNA	bacteria	Create new	genus
															Duplodnaviria		*Heunggongvirae*		*Uroviricota*		*Caudoviricetes*		*Crassvirales*		*Intestiviridae*		*Rudgehvirus*		*Rudgehvirus jaberico*	QQ198719	*Bacteroides cellulosilyticus phage 11*	Bc11		CG	dsDNA	bacteria	Create new	species