Computational Genomics of DNA Viruses:

Novel Insights into Bacteriophage and Human Cytomegalovirus Evolution

by

Abigail Ann Howell

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved October 2023 by the
Graduate Supervisory Committee:

Susanne Pfeifer, Chair
Kerry Geiler-Samerotte
Jeffrey Jensen
Noah Snyder-Mackler

ARIZONA STATE UNIVERSITY

December 2023

ABSTRACT

Viruses are the most abundant biological entities on Earth, infecting all types of cellular organisms. Yet less than 1% of the virosphere on our planet has been characterized to date. Viruses are both an important driver of bacterial evolution and have significant implications for human health, therefore understanding the relative contributions of various evolutionary forces in shaping their genomic landscapes is of critical importance both mechanistically as well as clinically. In my thesis I use computational genomic approaches to gain novel insights into bacteriophage and human cytomegalovirus evolution. In my first two chapters and associated appendices I characterized the complete genomes of the Cluster P bacteriophage Phegasus and Cluster DR bacteriophage BiggityBass, whose isolation hosts were *Mycobacterium smegmatis* mc²155 and *Gordonia terrae* CAG3, respectively. I also determined the bacteriophages' phylogenetic placement and computationally inferred their putative host ranges. For my fourth chapter I assessed the performance of several of these computational host range prediction tools using a dataset of bacteriophages whose host ranges have been experimentally validated. Finally, in my fifth chapter I reviewed the key parameters for developing an evolutionary baseline model of another virus, human cytomegalovirus.

DEDICATION

I would like to dedicate this dissertation to all my loved ones who have supported me on this journey.

ACKNOWLEDGMENTS

To my parents, Wendy and David Howell, thank you not only for the last six years but the lifetime of love and support you've given me. Thank you for always encouraging me to continue my education. I love you Mom and Dad. I wouldn't be here without you.

To my grandparents Gail and Doc Stevens, I love you both so very much. Thank you Gangy and Doc for always taking care of me and loving me. It means everything to have you here.

To my siblings, Emma, Maddie, Ben, and Nick, you all inspire me every day in your commitment to your work and education and I love each of you so much. It has been such a special experience going to the same college as you all, and having siblings in the same field I can share ideas with. I wouldn't trade it for the world.

To my grandparents Stephen and Ann Howell, I love you and wish you could be here to share this moment.

To my best friend Jenny, thank you for your constant encouragement, and thank you for always making time for me during the busiest time of both of our lives.

To my lab friends old and new, Ziqi, Juan, Courtney, Mark, and Cy, and honorary lab member Steph, the friends I made during this experience are just as valuable as the knowledge I gained.

To my loving partner Alec, thank you for your support and always attentively listening to me ramble about population genetics, evolution, or whatever paper I read that week.

To my mentor Mr. McKelvy, thank you for always believing in me. You make science fun.

Thank you to my committee Drs Susanne Pfeifer, Jeffrey Jensen, Noah Snyder-Mackler, and Kerry Geiler-Samerotte for helping me develop my research and for providing invaluable guidance and encouragement.

To our postdocs Viv and Terbot, thank you for your kindness and patience.

And a heartfelt thank you to my PI, Susanne, for being a wonderful mentor and giving me a second chance at my doctorate.

TABLE OF CONTENTS

APPENDIX Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Viral evolution is shaped by a myriad of factors, from the immune response and co-evolution of their hosts to their own genomic architecture and infection strategies; as well as basic evolutionary forces of drift, admixture, and demography (Spielman et. al 2019; Szpara 2021). These selective forces are particularly strong in viruses due to their short, coding dense genomes and large populations sizes. Viral evolution occurs on both short (a single round of infection) and longer evolutionary timeframes (Simmonds et. al 2019). Life history strategy can also influence evolutionary trajectories, i.e. lytic (lysis of the host after replication) or temperate lifecycles (periods of lysogeny in which the viral genome is integrated into the host genome followed by a transition into the lytic cycle). Molecular mechanisms of viral evolution include a spectrum of mutations (single nucleotide changes, tandem repeat fluctuations, insertions, deletions, and duplication), recombination, and horizontal gene transfer. Higher polymerase fidelity, error correction, and lysogeny are additional factors primarily associated with, but not exclusive to, DNA viruses, that can also influence their evolution (Szpara 2021). The switch to a temperate lifecycle in viruses is predicted to be evolutionarily advantageous under conditions of oscillating population dynamics and periodic environmental collapse, so that when host cells are limited the viral strain that can maintain growth in the lowest number of cells outcompetes those with the higher growth rate (Wahl et. al 2019).

Bacteriophages, viruses that infect bacteria, are an important part of the virosphere, and may perhaps be the most abundant organisms on Earth (Comeau et. al 2008). Phages have been used as a model organism in pioneering genetics research since the 1930's, from the Luria-Delbruck mutation rate experiments to the Hershey-Chase experiments establishing DNA as the hereditary material of life (Keen 2015).

Additionally, phages are important drivers of bacterial evolution through selective pressures and gene transfer through transduction (Chevallereau et. al 2022). The community context of bacterial hosts has also been shown to have important ecological and evolutionary effects on phage-hosts systems, with many questions still outstanding (Blazanin and Turner 2021). Interest in mycobacteriophages, viruses that infect mycobacterial hosts, emerged from the work of Jacobs et al.(1987) and Jacobs (2000), where they used mycobacteriophages to deliver foreign DNA into bacteria. Mycobacteriophages can be utilized as genomic tools to further our understanding of their pathogenic hosts, including *Mycobacterium tuberculosis* and *Mycobacterium leprae*, the causative agents of human tuberculosis and leprosy. Mycobacteriophages that infect close relatives of these pathogenic bacteria, e.g., Phegasus and its isolation host *Mycobacterium smegmatis*, may be possible candidates for phage therapy applications to combat antibiotic resistance. Additionally, BiggityBass and other phages that infect bacteria of the genus *Gordonia* can potentially be used as biocontrol agents for wastewater treatment (Goodfellow et. al 1998). To effectively guide the use of bacteriophages for biocontrol and phage therapy, the host range of these phages must be determined either experimentally or computationally. While experimental validation is the gold standard in elucidating phage-host interactions, these methodologies are laborious, time-intensive, costly, and limited by the number of microbial hosts able to be cultivated in the lab (Wade 2002; Edwards and Rohwer 2005). In response, various tools to computationally predict host ranges have been developed that utilize alignment-based or machine learning-based models (see review of Versoza and Pfeifer 2022).

Previously, mycobacteriophages were organized by morphology and host range, however these groupings were inconsistent with genomic sequence similarity (Lima-Mendez et. al 2008). Mycobacteriophage genomes have been described as mosaic

(Hendrix, 2002; Hendrix et al., 1999, 2000; Pedulla et al., 2003), where large sections of the genome have been exchanged horizontally through homologous recombination, site specific recombination, transposon-mediated gene transfer, or non-homologous illegitimate recombination. This mosaicism makes the construction of whole genome phylogenies of mycobacteriophages difficult, as their evolution is fundamentally reticulate (Lawrence et al., 2002; Lima-Mendez et al., 2008). In light of this, a cluster classification approach for mycobacteriophages has arisen, which assigns phages to a given cluster based on a nucleotide sequence similarity that spans more than 50% of the genome length with one or more other genomes (Hatfull 2010). Clusters therefore do not represent hierarchical lineages but reflect recent evolutionary events within a subcluster. To identify homologues that diverged longer ago, individual genes are grouped into "phamilies" based on pairwise comparisons using Clustal and BlastP searches (Cresawn 2011; Hatfull et al., 2006; Pope et al., 2011). Through comparative analysis of closely related mycobacteriophages, we can elucidate individual mutational steps of phage evolution that lead to phenotypic changes in phages.

Other DNA viruses have their own distinct evolutionary mechanisms, selective pressures, and evolutionary constraints. Human Cytomegalovirus (HCMV) is a β-herpesvirus in the Herpesviridae family with a relatively large double-stranded (ds) DNA genome of ~235 kb in size, including between 164-167 open reading frames (ORFs) (Dolan et al. 2004). HCMV is the leading cause of infection-related birth defects and contributes significantly to solid organ transplant failure and opportunistic infections in immunocompromised individuals (Balfour 1979; Suárez et al. 2019, 2020). HCMV and other Herpesviruses are characterized by lifelong persistence in their hosts through latency, in which the virus remains episomal in the host nucleus, a process distinct from lysogeny in that the viral genome does not integrate into the host genome. Latency is

achieved through HCMV's ability to evade the host immune system, which includes strategies such as strain polymorphism, epitope competition to mislead humoral responses, endocytosis, and glycan shielding (Hu et. al 2022). Previous studies have demonstrated that over 50% of HCMV's open reading frames can be deleted without impairing replication in fibroblasts, indicating that a majority of gene function is dedicated to immune modulating functions (Dunn et. al 2003; Yu et. al 2003). This repertoire of immunomodulatory functions is likely the result of HCMV's extended co-evolution alongside the human innate and adaptive immune system (McGeoch et al. 2008). This is further supported by evidence in the herpesviruses and mammalian host phylogenies that the diversification of hosts drives diversification of the virus (McGeoch, Rixon, and Davison 2006). In contrast, antiviral medications represent a recent selective pressure on HCMV (Hakki and Chou 2011).

In order to accurately detect recent responses to selective pressure, such as antiviral resistance mutations, genomic scans for positive selection should be evaluated within the context of demography, which can confound signals of adaption (Johri et al. 2020, 2021). Different approaches for inferring selection (outlier, two-step, and simultaneous inference approaches) deal with demography in increasingly sophisticated manners. In an outlier approach, loci under selection are identified through an increase in population differentiation, which is assumed to be distinguishable from differentiation that arises through neutral processes. However, studies have shown that certain patterns of migration and mutation within subpopulations can create false positives (Nei and Maruyama 1975). In a two-step approach the demographic history is inferred from putatively neutral sites (intergenic regions, synonymous mutations, the third base in codons) and then these parameters are fixed when inferring selection. The caveat of this approach is that it assumes all sites are independent and unliked, which is particularly

problematic in coding-dense viral genomes (Ewing and Jensen 2016; Johri et al. 2021). The final approach of simultaneous inference aims to develop new statistics and analytical expressions that encapsulate the effects of both neutral and selective processes on a site. These new statistics include a method to describe the SFS at neutral sites experiencing linked BGS (Cvijovic et al. 2018), a method of describing the SFS under linkage disequilibrium (LD) through a system of ordinary differential equations Friedlander and Steinrücken (2022), and an Approximate Bayesian Computation approach that utilizes a new statistic describing decay of BGS effects away from the targets of selection (Johri 2020). It is evident that understanding the relative contributions of various evolutionary forces (mutation rate, recombination, the distribution of fitness effects, admixture, and genetic drift) in shaping observed levels and patterns of variation is important for improving statistical power and reducing false-positive rates when scanning for adaptive mutations. For HCMV in particular, special consideration should be given to the level of progeny skew, bottleneck severity during infection and re-infection, and the degree of compartmental admixture.

Overview of Dissertation Chapters

This thesis represents a contribution to a larger, ongoing effort to characterize newly discovered bacteriophages through the phylogenetic placement of new strains and investigation of novel gene functions. In my first two chapters and associated appendices we annotated the genomes of bacteriophages Phegasus and BiggityBass using GLIMMER (Delcher et. al 1999) and GeneMark (Lukashin and Borodovsky 1998) to determine gene location and number, predicted gene function with NCBI BLAST (Altschul et. al 1990) and HHpred (Söding et. al 2005), and identified tRNAs using tRNAscan-SE (Lowe and Eddy 1997). For each phage genome, we investigated a

unique gene system through comparative analysis. In Phegasus, we identified an integration-dependent immunity system, which regulates the switch between lytic and lysogenic life cycles, as well as the integration attachments sites in the Cluster P bacteriophages and three putative host genomes. This indicates that these hosts are at risk of incorporating virulence factors from bacteriophages and therefore are not suitable candidate for antibacterial therapeutics. In BiggityBass, we identified a toxin/antitoxin (TA) system that allows it to inactivate bacteria-encoded toxins (Otsuka and Yonesaki 2012; Wei et. al 2016), which was homolgous to the hicA TA system present in *Burkholderia pseudomallei*, *E. coli*, and *Pseudomonas aeruginosa* (Yamaguchi and Inouye 2011; Butt et. al 2014; Shen et. al 2016). Much like the CRISPR-Cas9 system, the toxin/antitoxin system points towards a shared evolutionary history between phage and host through the development of a viral defense system that potentially can also be exploited as a genomic tool. In addition to characterizing the genomes of these phages, we also computationally predicted their host ranges using the tool WIsH. The host range prediction results of my second and third chapter inspired the work of chapter 4, in which we investigate 11 host range prediction tools using 4 experimentally validated polyvalent (broad-range) phage datasets. This work introduces a new classification scheme for host range prediction tools as either confirmatory or exploratory and provides tool recommendations based on user availability of host strains and desire for sensitivity vs specificity. This chapter also highlights a significant issue that many bacterial strain-specific genomes are not publicly available and the implications of this for computationally predicting host ranges. Finally, in my fifth chapter we reviewed the key parameters for developing an evolutionary baseline model of another virus, human cytomegalovirus. In this review we identify special considerations for HCMV when developing an evolutionary baseline model, including the ability to detect low frequency

variants, as well as the level of progeny skew, bottleneck severity during infection and re-infection, and the degree of compartmental admixture. This work lays the foundation for the development of an evolutionary baseline model of HCMV, which is critical to understanding how and when diversity in the HCMV genome is generated and has important implications for vaccine development as well as antiviral therapy. Taken together this thesis represents a collection of novel insights into several DNA viruses using computational genomics approaches.

CHAPTER 2

PHYLOGENOMIC ANALYSES AND HOST RANGE PREDICTION OF CLUSTER P MYCOBACTERIOPHAGES

(Previously published as A.A. Howell*, C.J. Versoza*, G. Cerna, T. Johnston, S. Kakde, K. Karuku, M. Kowal, J. Monahan, J. Murray, T. Nguyen, A. Sanchez Carreon, A. Streiff, B. Su, F. Youkhana, S. Munig, Z. Patel, M. So, M. Sy, S. Weiss, S.P. Pfeifer. 2022. Phylogenomic analyses and host range prediction of cluster P mycobacteriophages. *G3 (Bethesda)*, 12(11).)

(Associated appendix previously published as A.A. Howell*, C.J. Versoza*, G. Cerna, T. Johnston, S. Kakde, K. Karuku, M. Kowal, J. Monahan, J. Murray, T. Nguyen, A. Sanchez Carreon, E. Song, A. Streiff, B. Su, F. Youkhana, S. Munig, Z. Patel, M. So, M. Sy, S. Weiss, Y. Zhou, S.P. Pfeifer. 2022. Complete genome sequence of the cluster P mycobacteriophage Phegasus. *Microbiol. Resour. Announc*. e00540-22.)

* contributed equally

Abstract

Bacteriophages, infecting bacterial hosts in every environment on our planet, are a driver of adaptive evolution in bacterial communities. At the same time, the host range of many bacteriophages—and thus one of the selective pressures acting on complex microbial systems in nature—remains poorly characterized. Here, we computationally inferred the putative host ranges of 40 cluster P mycobacteriophages, including members from 6 subclusters (P1–P6). A series of comparative genomic analyses revealed that mycobacteriophages of subcluster P1 are restricted to the *Mycobacterium* genus, whereas mycobacteriophages of subclusters P2–P6 are likely also able to infect other genera, several of which are commonly associated with human disease. Further genomic analysis highlighted that the majority of cluster P mycobacteriophages harbor a conserved integration-dependent immunity system, hypothesized to be the ancestral state of a genetic switch that controls the shift between lytic and lysogenic life cycles—a temperate characteristic that impedes their usage in antibacterial applications.

Introduction

Less than 1% of the virosphere on our planet has been characterized to date (Geoghegan and Holmes 2017). An important part of this virosphere is bacteriophages (i.e. bacteria-infecting viruses), which are impacting bacterial genome evolution and community dynamics in every environment (Howard-Varona et al. 2017).

Bacteriophages can establish lytic or lysogenic infections—the former leading to cell destruction while the latter being "dormant," with bacteriophages replicating as prophages within the host without the production of virions (Howard-Varona et al. 2017). Temperate bacteriophages can switch between lytic and lysogenic life cycles, for example through the usage of integration-dependent immunity systems that establish lysogeny by suppressing lytic growth through an interplay between 3 proteins: integrase (Int), repressor (Rep), and Cro [for an in-depth discussion on these and other genetic switches, see the commentary by Broussard and Hatfull (2013)]. In integration-dependent immunity systems, the decision on whether lytic or lysogenic growth will take place depends by and large on the activity of Int as modulated by targeted proteolysis (Broussard et al. 2013). Under conditions where integrases are broken down (i.e. in the presence of a C-terminal ssrA-like protease degradation tag in Int), integration fails to occur. Instead, the viral form of Rep is generated and subsequently degraded due to the presence of its own C-terminal ssrA-like tag. The lytic protein Cro is freely expressed and stops repressor function (Hochschild et al. 1986). Conversely, when integrases escape proteolysis due to either decreased levels of proteases (such as ClpXP) or high multiplicity of infection (i.e. a high ratio of bacteriophages to infection targets), integration of bacteriophage genetic material will occur. This leads to the expression of an active (truncated) form of Rep that lacks the ssrA-like tag, causing a downregulation of Cro expression, which ultimately leads to lysogenic establishment and prophage induction.

Thereby, the integration into the host genome is mediated by recombination between the bacteriophage attachment site (attP) and the bacterial attachment site (attB) in the host genome. Attachment sites are recognized by Int—an integral part of the attP–Int cassette required for integrase-mediated site-specific recombination (Singh et al. 2013). Thereby, Int is either a tyrosine recombinase (which requires additional host cofactors such as the one present in *Mycobacterium smegmatis*; Pedulla et al. 1996; Peña et al. 1999; Lewis and Hatfull 2003; Chen et al. 2019) or a serine recombinase (which functions without any cofactors but recognizes shorter attP sequences than the tyrosine recombinase; Groth and Calos 2004).

Mycobacteriophages are a group of both lytic and temperate bacteriophages that infect mycobacterial hosts—including the causative agents for several human diseases such as tuberculosis (*M. tuberculosis*) or leprosy (*M. leprae*), separated into 31 clusters (A–Z and AA–AE) based on their nucleotide similarity and genomic architecture (Pope et al. 2011). Out of these, temperate cluster P bacteriophages are of particular interest to the scientific community to, for example study the evolution of genetic switches as several members of this cluster have been shown to harbor an unusual switch in which the bacteriophage attachment site is located within the repressor gene (e.g. Broussard et al. 2013; Doyle et al. 2017).

Interestingly, many mycobacteriophages have the ability to broaden their host range to infect either different strains or completely new mycobacterial species (Jacobs-Sera et al. 2012). In contrast to lytic bacteriophages, which are frequently exploited as antimicrobial agents (Sharma et al. 2017), the life cycle of temperate bacteriophages often impedes their usage, particularly with regard to bacteriophage therapy, due to the risk of transferring virulence factors through genomic pathogenicity islands (Malachowa and Deleo 2010; Xia and Wolz 2014). Thus, host ranges of many temperate

bacteriophages remain poorly characterized, despite their important impact on bacterial evolution. To advance our knowledge on the topic, and as part of a course-based undergraduate research experience at Arizona State University, we analyzed the genomes and computationally inferred the host ranges of 40 cluster P mycobacteriophages.

Materials and methods

Comparative genomic analyses

A multiple sequence alignment of 40 cluster P mycobacteriophages previously isolated in M. smegmatis mc2155 (Supplementary Table 1) was generated via MAFFT v.7.407 (Katoh and Standley 2013) and subsequently used to construct a neighbor-joining tree in MEGA X (Kumar et al. 2018) using a bootstrap test of phylogeny with 10,000 replicates. Additional whole-genome and gene-specific trees were generated, including 16 bacteriophages from clusters G1, I1, and N for which integration-dependent immunity systems had previously been identified (either experimentally or through the computational identification of an attP site within the repressor gene; Supplementary Table 2). Trees were visualized using FigTree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/; last accessed 2022 April 24) and the Interactive Tree Of Life (Letunic and Bork 2019). Sequence relatedness was determined using pairwise average nucleotide identity scores calculated using the DNA Master “Genome Comparison” tool v.5.23.6 and plotted using the ggplot2 function (Wickham 2016) in R v.4.0.2. All software were executed using default settings.

Identification of attP and attB sites

Following Pham et al. (2007), NCBI BLASTn (Altschul et al. 1990) was used to compare the 300-bp region surrounding the 5′-end of the immunity repressor gene in each cluster P mycobacteriophage (Supplementary Table 1) against the genomes of 14

putative mycobacterial host species (Supplementary Table 3) to determine the plausibility of attP/attB sites. In addition, Tandem Repeats Finder v.4.09 (Benson 1999) was used to search for integrase binding sites near the attP common core.

Host prediction

Following the best practices suggested by Versoza and Pfeifer (2022), both exploratory and confirmatory methods were used to computationally predict host ranges for 40 closely related cluster P mycobacteriophages (Supplementary Table 1). First, the exploratory tool PHERI v.0.2 (Baláž et al. 2020) was used to predict bacterial host genera. Among the currently available exploratory host range prediction tools, PHERI was the most user-friendly and well-documented, making it ideally suited for course-based undergraduate research experiences. Next, WIsH v.1.1 (Galiez et al. 2017)—a bacterial host range predictor that compares virus and host sequence composition—was used to estimate the likelihood of these 40 cluster P bacteriophages to infect 14 putative mycobacterial host species with particular relevance to human health and disease (Supplementary Table 3). WIsH was selected as the representative for confirmatory host range prediction tools as it was an easily applicable alternative to alignment-based tools which frequently underpredict phage–host interactions (Zielezinski et al. 2021). Lastly, following Crane et al. (2021), PHASTER (Arndt et al. 2016) was used to search the genome of these putative host species for prophages to determine whether cluster P mycobacteriophages might be able to integrate into the host.

Results and Discussion

Comparative genomic analyses between 40 cluster P mycobacteriophages (32 subcluster P1, 1 subcluster P2, 1 subcluster P3, 2 subcluster P4, 2 subcluster P5, and 1 subcluster P6; Supplementary Table 1) demonstrated a close relatedness at the sequence level (Fig. 1a), with cluster assignments supported by pairwise average

nucleotide identities between the bacteriophages (Supplementary Fig. 1). With the exception of Tortellini (P2), Xavia (P3), and ThulaThula (P5), cluster P bacteriophage genomes harbor a conserved integration-dependent immunity system, comprised of an immunity repressor flanked by a tyrosine integrase, an excise gene, and an antirepressor (Supplementary Fig. 2) that governs the transition from the lytic to lysogenic state by binding and inactivating the lysogenic repressor (Lemire et al. 2011; Kim and Ryu 2013). It has previously been hypothesized that conserved integration-dependent immunity systems form the ancestral state of more complex genetic switches (Broussard and Hatfull 2013), such as those present in λ bacteriophages (Oppenheim et al. 2005). Interestingly, a neighbor-joining tree generated from whole-genome sequences of 16 cluster G1, I1, and N bacteriophages containing an integration-dependent immunity system (Supplementary Table 2) places cluster P4–P6 bacteriophages as sister taxa to the G1, I1, and N subclusters (Fig. 1b)—a tree topology supported by the gene-specific tree based on the immunity repressor sequences (Fig. 1c).

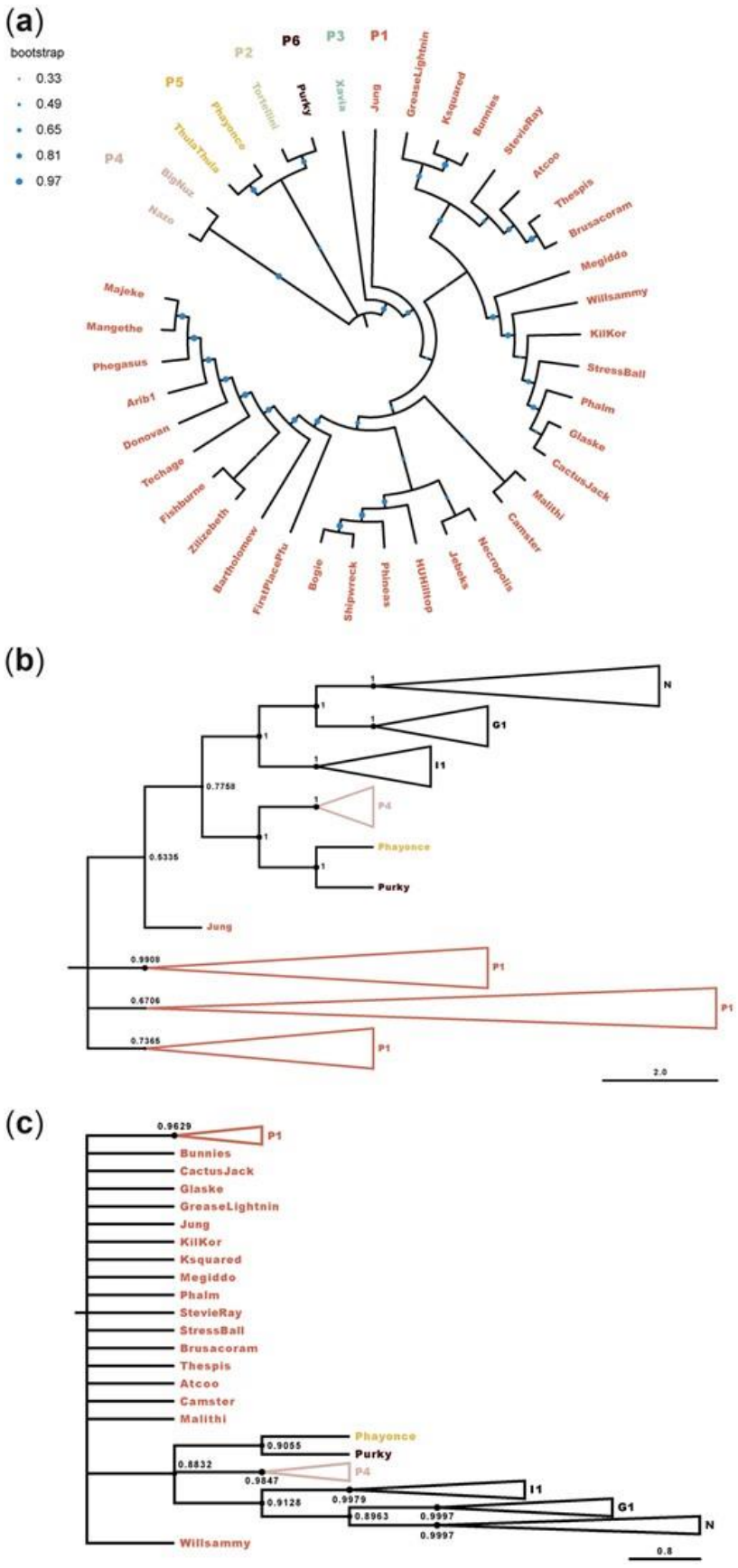

(a)
bootstrap
0.33
0.49
0.65
0.81
0.97
P6
P3
P1
P2
P5
P4
Purky
Xavia
Jung
GreaseLightnin
Ksquared
Bunnies
StevieRay
Atcoo
Thespis
Brusacoram
Megiddo
Willsammy
KillKor
StressBall
Phaim
Glaske
CactusJack
Malithi
Camster
Necropolis
Jebeks
HUHilltop
Phineas
Shipwreck
Bogie
FirstPlacePlu
Bartholomew
Zilizabeth
Fishburne
Techage
Donovan
Arib1
Phegasus
Mangethe
Majeke
Nazo
BigNaz
ThulaThula
Phayonce
Tortellini
(b)
N
G1
I1
P4
Phayonce
Purky
Jung
P1
P1
P1
1
0.7758
0.5335
0.9909
0.6706
0.7365
2.0
(c)
P1
Bunnies
CactusJack
Glaske
GreaseLightnin
Jung
KillKor
Ksquared
Megiddo
Phaim
StevieRay
StressBall
Brusacoram
Thespis
Atcoo
Camster
Malithi
Phayonce
Purky
P4
I1
G1
N
Willsammy
0.9629
0.9055
0.8832
0.9847
0.9128
0.9979
0.8963
0.9997
0.9997
0.8

Fig. 1. Neighbor-joining trees. Neighbor-joining trees generated in MAFFT (Katoh and Standley 2013) using the multiple-sequence alignment of (a) 40 cluster P mycobacteriophages (Supplementary Table 1) and (b) 16 cluster G1, I1, and N bacteriophages with a previously identified integration-dependent immunity system (Supplementary Table 2), with 10,000 bootstrap replicates. c) Gene-specific tree based on the immunity repressor sequences of the bacteriophages included in (b). Colors highlight membership in subclusters P1–P6.

To explore the impact of cluster P mycobacteriophages on bacterial communities, their host ranges were computationally predicted using a combination of exploratory and confirmatory tools, together with 14 putative mycobacterial host species relevant to human health and disease. Using the exploratory method, all but 1 P1 bacteriophages (Donovan) appear restricted to the Mycobacterium genus (Table 1). In contrast, bacteriophages of subclusters P2–P6 are likely also able to infect the nonpathogenic microbes *Gordonia* and *Rhizobium* as well as hosts of the genera *Clostridiodes, Clostridium, and Corynebacterium*, frequently associated with human disease, including diphtheria (*Corynebacterium diphtheriae*) as well as several hospital-acquired infections (see reviews by Bernard 2012 and Mangutov et al. 2021). As the ability to bind to new receptors is a key step in host-range evolution (Meyer et al. 2012), mutations within tail protein genes might explain the predicted expanded host range of subclusters P2–P6. At the species level, confirmatory results (Fig. 2) suggest that, in addition to *M. smegmatis* mc2155 used to isolate the bacteriophages, subcluster P1 mycobacteriophages are likely able to infect *Mycobacterium fortuitum*—which can cause infections in the skin, lymph nodes, and joints of immunocompromised individuals (Sethi et al. 2014), as well as *Mycobacterium gilvum*, and *Mycobacterium intracellulare*—which can cause pulmonary infections and lymphadenitis in immunocompromised individuals (Han et al. 2005). In contrast, bacteriophages of subclusters P2–P6 displayed low likelihoods of infection for all tested hosts.

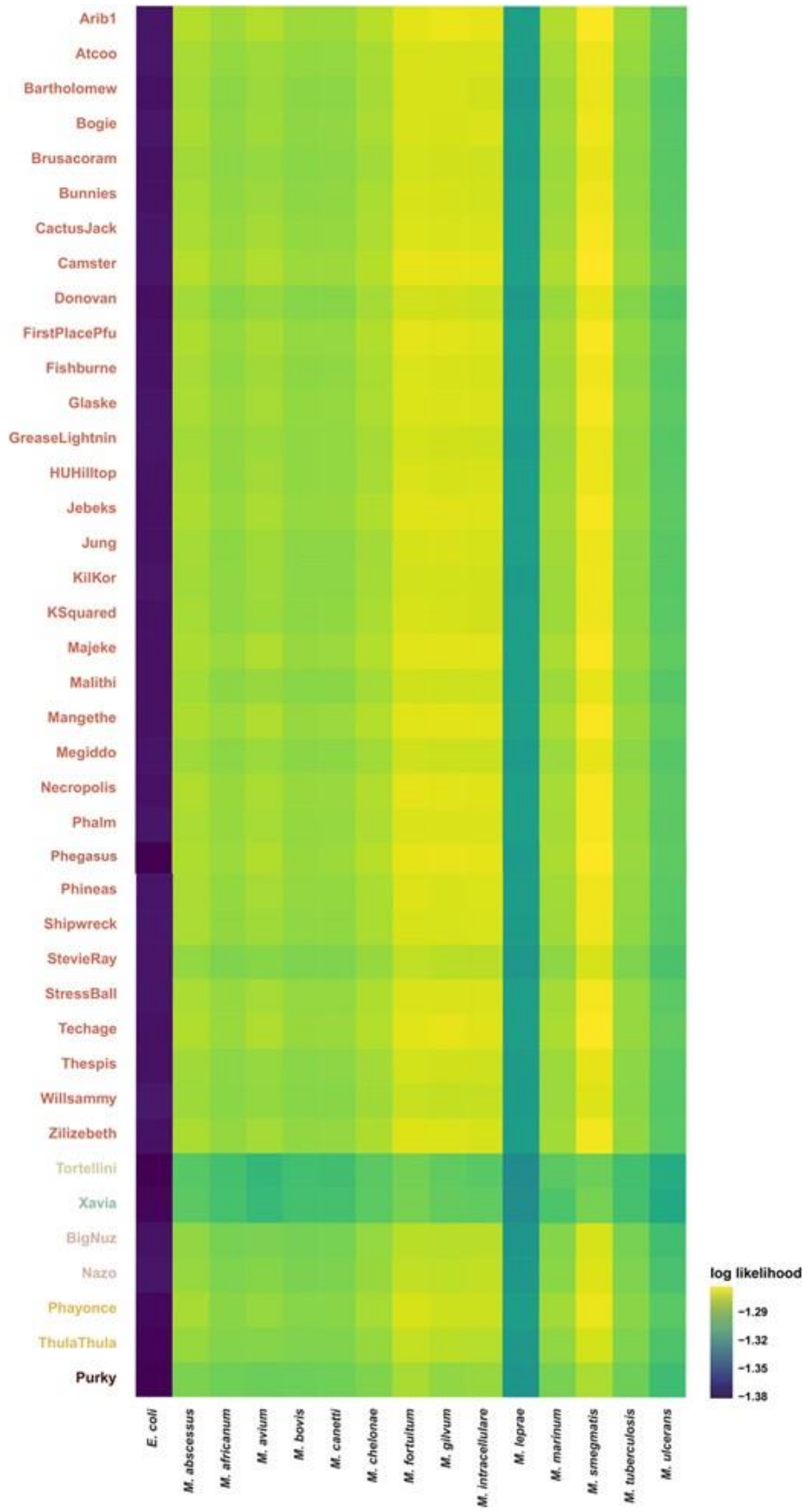

Arib1
Atcoo
Bartholomew
Bogie
Brusacoram
Bunnies
CactusJack
Camster
Donovan
FirstPlacePfu
Fishburne
Glaske
GreaseLightnin
HUHilltop
Jebeks
Jung
KilKor
KSquared
Majeke
Malithi
Mangethe
Megiddo
Necropolis
Phalm
Phegasus
Phineas
Shipwreck
StevieRay
StressBall
Techage
Thespis
Willsammy
Zilizebeth
Tortellini
Xavia
BigNuz
Nazo
Phayonce
ThulaThula
Purky
E. coli
M. abscessus
M. africanum
M. avium
M. bovis
M. canetti
M. chelonae
M. fortuitum
M. gilvum
M. intracellulare
M. leprae
M. marinum
M. smegmatis
M. tuberculosis
M. ulcerans
log likelihood
−1.29
−1.32
−1.35
−1.38

Fig. 2. Confirmatory host range prediction. Putative bacteriophage–host interactions as predicted by WIsH (Galiez et al. 2017), using 40 cluster P mycobacteriophages (Supplementary Table 1), together with 14 potential bacterial hosts and *Escherichia coli* as a negative control (Supplementary Table 2). The higher the reported value, the more likely a bacteriophage is able to infect a putative host.

Table 1. Exploratory host range prediction.

Phage	Subcluster	*Mycobacterium*	*Gordonia*	*Clostridioides*	*Corynebacterium*	*Rhizobium*	*Clostridium*
Arib1	P1	✓					
Atcoo	P1	✓					
Bartholomew	P1	✓					
Bogie	P1	✓					
Brusacoram	P1	✓					
Bunnies	P1	✓					
CactusJack	P1	✓					
Camster	P1	✓					
Donovan	P1	✓	✓				
FirstPlacePfu	P1	✓					
Fishburne	P1	✓					
Glaske	P1	✓					
GreaseLightnin	P1	✓					
HUHilltop	P1	✓					
Jebeks	P1	✓					
Jung	P1	✓					
KilKor	P1	✓					
Ksquared	P1	✓					
Majeke	P1	✓					
Malithi	P1	✓		✓			
Mangethe	P1	✓					
Megiddo	P1	✓					
Necropolis	P1	✓					
Phalm	P1	✓					
Phegasus	P1	✓					
Phineas	P1	✓					
Shipwreck	P1	✓					
StevieRay	P1	✓					
StressBall	P1	✓					
Techage	P1	✓					
Thespis	P1	✓					
Willsammy	P1	✓					
Zilizebeth	P1	✓					
Tortellini	P2	✓	✓	✓	✓		
Xavia	P3	✓	✓	✓		✓	
BigNuz	P4	✓	✓				
Nazo	P4	✓	✓				
Phayonce	P5	✓	✓				
ThulaThula	P5	✓		✓			✓
Purky	P6	✓	✓				

Table 1. Exploratory host range prediction. Putative host genera of the 40 cluster P bacteriophages included in this study (Supplementary Table 1) as predicted by PHERI (Baláž et al. 2020).

To investigate the temperate nature of cluster P mycobacteriophages, prophage sequences were computationally predicted within the putative host genomes. Three putative hosts (*Mycobacterium abscessus, Mycobacterium marinum, and M. smegmatis*) contain intact prophages—however, none of them correspond to prophages that stem from the integration of cluster P mycobacteriophages. In addition, incomplete prophages from the integration of cluster P mycobacteriophages were detected in both *M.*

abscessus and *M. marinum* (Fig. 3)—2 opportunistic pathogens known to inflict pulmonary (Winthrop and Roy 2020) and cutaneous (Aubry et al. 2000) infections in humans—indicating that these hosts are at risk of incorporating virulence factors from these bacteriophages. Interestingly, the 2 partial prophages within *M. abscessus* and *M. marinum* were predicted to stem from the integration of 2 (out of only 3) cluster P bacteriophages that lack an integration-dependent immunity system (ThulaThula and Xavia, respectively).

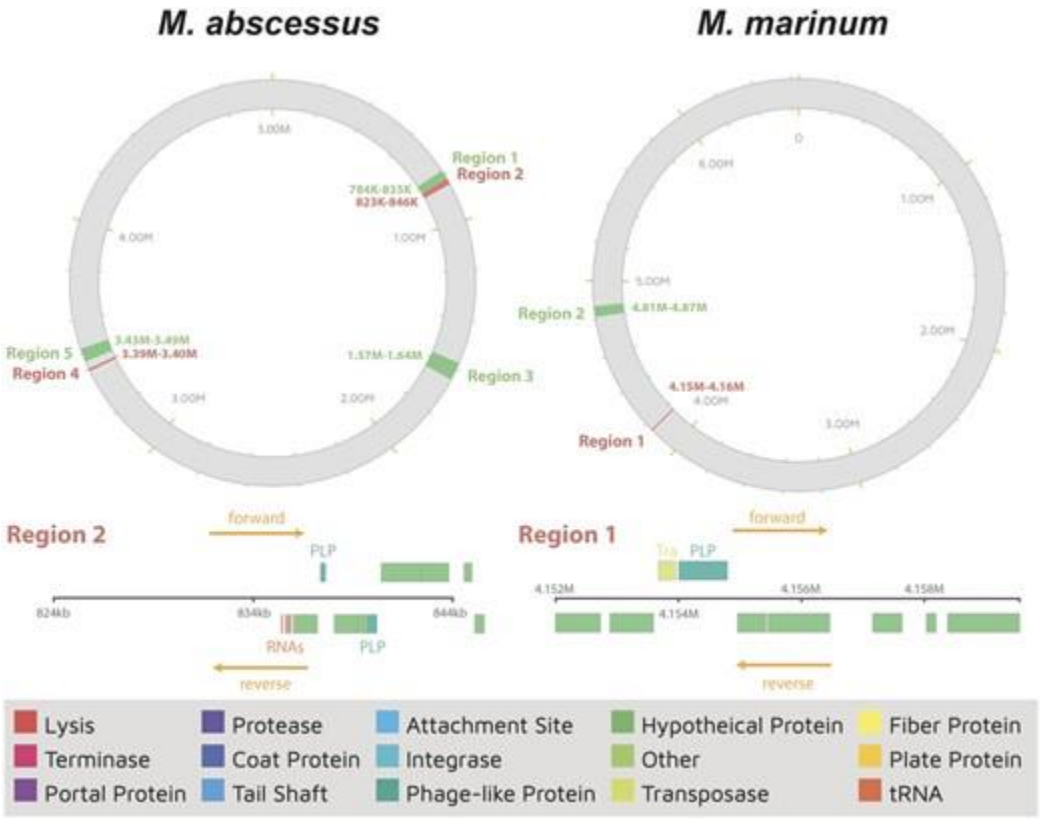


Fig. 3. Prophage prediction. Complete (green) and incomplete (red) prophages from the integration of bacteriophages were detected in both *M. abscessus* (left) and *M. marinum* (right). Incomplete prophages from the integration of cluster P mycobacteriophages are displayed at the bottom (region 2 in *M. abscessus* and region 1 in *M. marinum*), together with the protein-coding genes contained in these regions. Phage-like proteins on forward and reverse strands (indicated by orange arrows) are displayed above and below the ruler for each region, respectively.

For temperate bacteriophages, the risk of transfer of virulence factors depends (at least in part) on the presence of an attP region in the bacteriophage as well as a corresponding attB attachment site in the host genome (Pham et al. 2007). Putative attP sites in cluster P bacteriophages are similar in length to those previously reported in other mycobacteriophages (Pham et al. 2007; Morris et al. 2008) and the lack of arm-type integrase binding sites flanking the attP common core—known to be present in nonintegration-dependent immunity system bacteriophages such as λ (Landy 1989) and L5 (Peña et al. 1997) but notably absent in integration-dependent immunity system bacteriophages (Broussard et al. 2013)—is further evidence of a functional integration-dependent immunity system in these bacteriophages. To identify putative attachment sites, attP sites were compared against the genomes of 14 mycobacteria. Out of the 14 mycobacterium species tested, only 3 (*M. smegmatis, Mycobacterium chelonae, and Mycobacterium leprae*) contained a homologous attB bacterial attachment site, overlapping with the 3′-end of a tRNAThr gene (Supplementary Table 4), indicating that these hosts are at risk of incorporating virulence factors from bacteriophages that utilize tyrosine integrases in their integration-dependent immunity systems. Yet, despite the presence of an attB attachment site, 2 out of these 3 species (*M. chelonae and M. leprae*) were not predicted as potential hosts for any cluster P bacteriophage. However, it is important to note that WIsH evaluates host likelihood on the basis of oligonucleotide frequency similarity between the virus and host genomes. Consequently, more sophisticated approaches that rely on several distinct genomic features to predict the success of phage infection (such as advanced machine learning-based methods) may be able to provide a more complete picture of the putative host ranges.

Taken together, our computational predictions indicate that cluster P bacteriophages harboring a conserved integration-dependent immunity system likely

exhibit similar host ranges. An important future endeavor will be the experimental validation of the presented computational results by phenotypic studies in order to lend further credence to the hypothesis that the type of genetic switch used to induce lysogeny plays an important role in host range evolution.

Data Availability

Genomic data for all 40 cluster P mycobacteriophages, 16 cluster G1, I1, and N bacteriophages with a previously identified integration-dependent immunity system, and 14 putative bacterial host species can be downloaded from the NCBI Sequence Read Archive using the accession numbers provided in Supplementary Tables 1–3, respectively. Supplementary Table 4 lists the mycobacteriophage integration systems and putative integration sites of cluster P mycobacteriophages in *M. chelonae, M. leprae, and M. smegmatis*. Supplementary Fig. 1 displays the pairwise average nucleotide identities of the 40 cluster P bacteriophages. Supplementary Fig. 2 displays the Phamerator map of the regions encoding the tyrosine integrase, immunity repressor, and excise genes in cluster P mycobacteriophages.

Supplementary Materials

Phage	Subcluster	Length (bp)	GC-content	# ORFs	# tRNAs	IDIS*	Accession #	Reference
Arib1	P1	46,732	67.5%	78	0	yes	NC_051736.1	unpublished
Atcoo	P1	49,075	67.0%	78	0	yes	NC_051729.1	unpublished
Bartholomew	P1	46,484	67.2%	77	0	yes	NC_051734.1	Doyle *et al.* 2018
Bogie	P1	48,639	66.9%	81	0	yes	MF133446.1	Doyle *et al.* 2018
Brusacoram	P1	47,618	67.0%	78	0	yes	NC_028747.1	Hatfull *et al.* 2016
Bunnies	P1	48,822	67.1%	81	1	yes	MN096356.1	unpublished
CactusJack	P1	48,222	67.3%	79	0	yes	MN892484.1	unpublished
Camster	P1	47,149	67.2%	80	0	yes	MW055902.1	unpublished
Donovan	P1	47,162	67.2%	78	0	yes	KF841477.1	Pope *et al.* 2015
FirstPlacePfu	P1	45,680	67.3%	82	0	yes	NC_051735.1	unpublished
Fishburne	P1	47,109	67.3%	77	0	yes	NC_021302.1	Hatfull *et al.* 2013
Glaske	P1	48,222	67.3%	78	0	yes	MN807250.1	unpublished
GreaseLightnin	P1	48,424	67.1%	80	0	yes	NC_051731.1	unpublished
HUHilltop	P1	46,896	67.2%	81	0	yes	MN010757.1	Pope *et al.* 2015
Jebeks	P1	45,580	67.3%	77	0	yes	NC_041969.1	Pope *et al.* 2015
Jung	P1	46,561	67.1%	77	0	yes	NC_051730.1	Van *et al.* 2020
KilKor	P1	48,916	67.2%	79	0	yes	NC_053209.1	unpublished
Ksquared	P1	48,699	67.1%	80	0	yes	NC_051732.1	Doyle *et al.* 2018
Majeke	P1	47,612	67.4%	81	0	yes	NC_051737.1	unpublished
Malithi	P1	46,870	67.1%	79	0	yes	KP027200.1	Pope *et al.* 2015
Mangethe	P1	47,612	67.4%	81	1	yes	MK016499.1	unpublished
Megiddo	P1	48,783	67.1%	78	0	yes	NC_051728.1	unpublished
Necropolis	P1	46,263	62.9%	76	0	yes	MK937604.1	unpublished
Phalm	P1	48,213	67.3%	79	0	yes	MN807248.1	unpublished
Phegasus	P1	47,578	67.4%	81	0	yes	ON637760	Howell, Versoza *et al.* 2022
Phineas	P1	47,229	67.2%	77	0	yes	NC_051733.1	Pope *et al.* 2015
Shipwreck	P1	48,670	66.9%	81	0	yes	NC_031261.1	Pope *et al.* 2015
StevieRay	P1	48,815	66.9%	81	1	yes	MF373843.1	unpublished
StressBall	P1	47,915	67.3%	78	0	yes	MN908683.1	unpublished
Techage	P1	47,094	67.4%	79	0	yes	MK919480.1	unpublished
Thespis	P1	47,618	67.0%	78	0	yes	MG198785.1	Bushhouse *et al.* 2017
Willsammy	P1	48,399	67.0%	80	0	yes	NC_051727.1	unpublished
Zilizebeth	P1	48,056	67.3%	83	1	yes	MK524508.1	unpublished
Tortellini	P2	49,658	65.8%	76	0	no	NC_041888.1	Doyle *et al.* 2018
Xavia	P3	49,808	65.9%	71	0	no	NC_051740.1	unpublished
BigNuz	P4	48,984	66.7%	82	0	yes	NC_023692.1	Pope *et al.* 2015
Nazo	P4	48,870	66.8%	83	0	yes	KX641262	unpublished
Phayonce	P5	49,203	66.7%	77	0	yes	KR080195	Pope *et al.* 2015
ThulaThula	P5	50,415	66.5%	80	0	no	MN234172	Wada *et al.* 2017
Purky	P6	50,513	66.4%	84	0	yes	MN096355.1	Pope *et al.* 2015

* integration-dependent immunity system, comprised of an immunity repressor flanked by an integrase, an excise gene, and an anti-repressor

Table S1. Mycobacterium cluster P bacteriophages included in the comparative analyses. Bacteriophages for which integration-dependent immunity systems had previously been identified through the computational identification of an attP site within the repressor gene are highlighted in blue.

Phage Name	Subcluster	Length (bp)	GC-content	# ORFs	# tRNAs	Accession #	Reference
BPs	G1	41,901	66.6%	63	0	EU568876	Sampson *et al.* 2009
Cedarsite	G1	41,901	66.6%	63	0	KT355472	Hatfull *et al.* 2016
Halo	G1	42,289	66.7%	64	0	NC_008202.2	Sampson *et al.* 2009
Island3	I1	47,287	66.8%	76	0	HM152765.1	Pope *et al.* 2011
Babsiella	I1	48,420	67.1%	78	0	NC_023697.1	Hatfull *et al.* 2012
Brujita	I1	47,057	66.8%	74	0	NC 011291.1	Hatfull *et al.* 2010
Charcharodon	N	43,680	66.2%	71	0	KM588359	Hatfull *et al.* 2016
Charlie	N	43,036	66.3%	69	0	NC_023729.1	Hatfull *et al.* 2012
MichelleMyBell	N	42,240	66.0%	70	0	KF986246	Hatfull *et al.* 2016
Panchino	N	43,516	65.9%	66	0	KU935727	Hatfull *et al.* 2016
Phrann	N	44,872	66.3%	67	0	KU935731	Hatfull *et al.* 2016
Pipsqueaks	N	43,679	66.3%	73	0	KU935730	Hatfull *et al.* 2016
Redi	N	42,594	66.1%	70	0	NC_023730.1	Hatfull *et al.* 2012
SkinnyPete	N	43,478	66.4%	67	0	KU935729	Hatfull *et al.* 2016
Xeno	N	42,395	66.8%	69	0	KU935728	Hatfull *et al.* 2016
Xerxes	N	43,698	66.3%	72	0	KU935726	Hatfull *et al.* 2016

Table S2. Bacteriophages included in the comparative analyses for which integration-dependent immunity systems had previously been identified (green: experimentally validated; blue: computationally predicted).

Mycobacterium	# Genes	Length (kb)	GC-content	Accession #	Reference
M. abscessus	4,957	4,618	64.2%	CP004374	Kim *et al.* 2013
M. africanum	4,069	4,493	65.1%	CP014617	Hurtado *et al.* 2016
M. avium	3,935	3,981	69.3%	AE016958	Li *et al.* 2005
M. bovis	3,952	3,972	65.6%	AM408590	Brosch *et al.* 2007
M. canetti	4,139	4,482	65.6%	HE572590	Bentley *et al.* 2012
M. chelonae	4,943	5,061	64.0%	CP050145	Gu *et al.* 2020
M. fortuitum	6,023	6,255	66.2%	CP011269	Costa *et al.* 2015
M. gilvum	5,139	5,077	67.9%	CP002385	Kallimanis *et al.* 2011
M. intracellulare	5,143	4,936	68.1%	CP003322	Kim *et al.* 2012
M. leprae	1,604	1,620	57.8%	AL450380	Cole *et al.* 2001
M. marinum	5,422	5,973	65.2%	CP000854	Stinear *et al.* 2008
M. smegmatis	6,692	6,508	67.4%	CP001663	Deshayes *et al.* 2007
M. tuberculosis	3,935	3,981	65.6%	AL123456	Cole *et al.* 1998
M. ulcerans	4,159	4,074	65.4%	CP000325	Stinear *et al.* 2007

Table S3. Mycobacteria included in the comparative analyses.

	BLASTn percent identity		
Phage	***M. chelonae***	***M. leprae***	***M. smegmatis***
Arib1	38/40 (95%)	40/40 (100%)	39/39 (100%)
Atcoo	38/40 (95%)	40/40 (100%)	41/42 (98%)
Bartholomew	38/40 (95%)	40/40 (100%)	39/39 (100%)
Bogie	38/40 (95%)	40/40 (100%)	39/39 (100%)
Brusacoram	38/40 (95%)	40/40 (100%)	41/42 (98%)
Bunnies	38/40 (95%)	40/40 (100%)	41/42 (98%)
CactusJack	38/40 (95%)	40/40 (100%)	41/42 (98%)
Camster	38/40 (95%)	40/40 (100%)	41/42 (98%)
Donovan	38/40 (95%)	40/40 (100%)	39/39 (100%)
FirstPlacePfu	38/40 (95%)	40/40 (100%)	39/39 (100%)
Fishburne	38/40 (95%)	40/40 (100%)	39/39 (100%)
Glaske	38/40 (95%)	40/40 (100%)	41/42 (98%)
GreaseLightnin	38/40 (95%)	40/40 (100%)	41/42 (98%)
HUHilltop	38/40 (95%)	40/40 (100%)	39/39 (100%)
Jebeks	38/40 (95%)	40/40 (100%)	39/39 (100%)
Jung	38/40 (95%)	40/40 (100%)	41/42 (98%)
KilKor	38/40 (95%)	40/40 (100%)	41/42 (98%)
Ksquared	38/40 (95%)	40/40 (100%)	41/42 (98%)
Majeke	38/40 (95%)	40/40 (100%)	39/39 (100%)
Malithi	38/40 (95%)	40/40 (100%)	41/42 (98%)
Mangethe	38/40 (95%)	40/40 (100%)	39/39 (100%)
Megiddo	38/40 (95%)	40/40 (100%)	41/42 (98%)
Necropolis	38/40 (95%)	40/40 (100%)	39/39 (100%)
Phalm	38/40 (95%)	40/40 (100%)	41/42 (98%)
Phegasus	38/40 (95%)	40/40 (100%)	39/39 (100%)
Phineas	38/40 (95%)	40/40 (100%)	39/39 (100%)
Shipwreck	38/40 (95%)	40/40 (100%)	39/39 (100%)
StevieRay	38/40 (95%)	40/40 (100%)	41/42 (98%)
StressBall	38/40 (95%)	40/40 (100%)	41/42 (98%)
Techage	38/40 (95%)	40/40 (100%)	39/39 (100%)
Thespis	38/40 (95%)	40/40 (100%)	41/42 (98%)
Willsammy	38/40 (95%)	40/40 (100%)	41/42 (98%)
Zilizebeth	38/40 (95%)	40/40 (100%)	39/39 (100%)
BigNuz	38/40 (95%)	40/40 (100%)	39/39 (100%)
Nazo	38/40 (95%)	40/40 (100%)	39/39 (100%)
Phayonce	42/45 (93%)	44/45 (98%)	43/44 (98%)
Purky	39/41 (95%)	41/41 (100%)	42/43 (98%)

Table S4. Mycobacteriophage integration systems and putative integration sites of Mycobacterium cluster P bacteriophages in *M. chelonae* Myco3a (attB location: tRNAThr ; 447,412–447,737 bp), *M. leprae* TN (attB location: tRNAThr ; 271,936–271,975 bp), and *M. smegmatis* mc2 (attB location: tRNAThr ; 6,222,599–6,222,637 bp).

Supplementary References

Bentley SD, Comas I, Bryant JM, Walker D, Smith NH, et al. 2012. The genome of Mycobacterium africanum West African 2 reveals a lineage-specific locus and genome erosion common to the M. tuberculosis complex. PLoS Negl. Trop. Dis. 6(2): e1552.

Brosch R, Gordon SV, Garnier T, Eiglmeier K, Frigui W, et al. 2007. Genome plasticity of BCG and impact on vaccine efficacy. Proc. Natl. Acad. Sci. U. S. A. 104(13): 5596–5601. Bushhouse DZ, Bowen EK, Wolyniak MJ. 2017. Isolation and genome annotation of mycobacteriophage Thespis. H-SC Journal of Sciences. 6.

Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, et al. 2001. Massive gene decay in the leprosy bacillus. Nature. 409(6823): 1007–1011.

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, et al. 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature. 393(6685): 537–544.

Costa KC, Bergkessel M, Saunders S, Korlach J, Newman DK. 2015. Enzymatic degradation of phenazines can generate energy and protect sensitive organisms from toxicity. mBio. 6(6): e01520-15.

Deshayes C, Perrodou E, Gallien S, Euphrasie D, Schaeffer C, et al. 2007. Interrupted coding sequences in Mycobacterium smegmatis: authentic mutations or sequencing errors? Genome Biol. 8(2): R20.

Doyle EL, Fillman CL, Reyna NS, Tobiason DM, Westholm DE, et al. 2018. Genome sequences of four cluster P mycobacteriophages. Genome Announc. 6(2): e01101-17. Gu CH, Zhao C, Hofstaedter C, Tebas P, Glaser L, et al. 2020. Investigating hospital Mycobacterium chelonae infection using whole genome sequencing and hybrid assembly. PLoS One. 15(11): e0236533.

Hatfull GF, Jacobs-Sera D, Lawrence JG, Pope WH, Russell DA, et al. 2010. Comparative genomic analysis of 60 mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. J. Mol. Biol. 397(1): 119–143.

Hatfull GF, Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) Program, KwaZulu-Natal Research Institute for Tuberculosis and HIV (K-RITH) Mycobacterial Genetics Course, University of California – Los Angeles Research Immersion Laboratory in Virology, Phage Hunters Integrating Research and Education (PHIRE) Program. 2016. Complete genome sequences of 61 mycobacteriophages. Genome Announc. 4(4): e00389-16.

Hatfull GF, Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) Program, KwaZulu-Natal Research Institute for Tuberculosis and HIV (K-RITH) Mycobacterial Genetics Course, University of California – Los Angeles Research Immersion Laboratory in Virology, Phage Hunters Integrating Research and Education (PHIRE) Program. 2013. Complete genome sequences of 63 mycobacteriophages. Genome Announc. 1(6): e00847-13.

Hatfull GF, Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Science Program, KwaZulu-Natal Research Institute for Tuberculosis and HIV Mycobacterial Genetics Course Students, Phage Hunters Integrating Research and Education Program. 2012. Complete genome sequences of 138 mycobacteriophages. J. Virol. 86(4): 2382-4.

Howell* AA, Versoza* CJ, Cerna G, Johnston T, Kakde S, Karuku K, Kowal M, Monahan J, Murray J, Nguyen T, Sanchez Carreon A, Song E, Streiff A, Su B, Youkhana F, Munig S, Patel Z, So M, Sy M, Weiss S, Zhou Y, Pfeifer SP. 2022. Complete genome sequence of the cluster P mycobacteriophage Phegasus. Microbiol. Resour. Announc. e00540-22.

Hurtado UA, Solano JS, Rodriguez A, Robledo J, Rouzaud F. 2016. Draft genome sequence of a Mycobacterium africanum clinical isolate from Antioquia, Colombia. Genome Announc. 4(3): e00486-16.

Kallimanis A, Karabika E, Mavromatis K, Lapidus A, Labutti KM, et al. 2011. Complete genome sequence of Mycobacterium sp. strain (Spyr1) and reclassification to Mycobacterium gilvum Spyr1. Stand Genomic Sci. 5(1): 144–153.

Kim BJ, Kim BR, Hong SH, Seok SH, Kook YH, et al. 2013. Complete genome sequence of Mycobacterium massiliense clinical strain Asan 50594, belonging to the type II genotype. Genome Announc. 1(4): e00429-13.

Kim BJ, Choi BS, Lim JS, Choi IY, Lee JH, et al. 2012. Complete genome sequence of Mycobacterium intracellulare strain ATCC 13950(T). J. Bacteriol. 194(10): 2750.
Li L, Bannantine JP, Zhang Q, Amonsin A, May BJ, et al. 2005. The complete genome sequence of Mycobacterium avium subspecies paratuberculosis. Proc. Natl. Acad. Sci. U. S. A. 102(35): 12344–12349.

Pope WH, Bowman CA, Russell DA, Jacobs-Sera D, Asai DJ, et al. 2015. Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. eLife. 4: e06416.

Pope WH, Jacobs-Sera D, Russell DA, Peebles CL, Al-Atrache Z, et al. 2011. Expanding the diversity of mycobacteriophages: insights into genome architecture and evolution. PLoS One. 6(1): e16329.

Pope WH, Jacobetz E, Johnson CA, Kihle BL, Sobeski MA, et al. 2015. Genome sequence of mycobacteriophage Phayonce. Genome Announc. 3(3): e00598-15.

Sampson T, Broussard GW, Marinelli LJ, Jacobs-Sera D, Ray M, et al. 2009. Mycobacteriophages BPs, Angel and Halo: comparative genomics reveals a novel class of ultra-small mobile genetic elements. Microbiology. 155(Pt 9): 2962.

Stinear TP, Seemann T, Harrison PF, Jenkin GA, Davies JK, et al. 2008. Insights from the complete genome sequence of Mycobacterium marinum on the evolution of Mycobacterium tuberculosis. Genome Res. 18(5): 729–741.

Stinear TP, Seemann T, Pidot S, Frigui W, Reysset G, et al. 2007. Reductive evolution and niche adaptation inferred from the genome of Mycobacterium ulcerans, the causative agent of Buruli ulcer. Genome Res. 17(2): 192–200.

Van R, Nie W, Abdela F, Eivazi B, Kickbusch D, et al. 2020. Complete genome sequences of cluster P1 and cluster C1 Mycobacterium smegmatis phages Jung and Ronan. Microbiol. Resour. Announc. 9(34): e00678-20.

Wada T, Hijikata M, Maeda S, Hang NTL, Thuong PH, et al. 2017. Complete genome sequence of a Mycobacterium tuberculosis strain belonging to the east African-Indian family in the indo-oceanic lineage, isolated in Hanoi, Vietnam. Genome Announc. 5(24): e00509-17.

CHAPTER 3

COMPARATIVE GENOMICS OF CLOSELY-RELATED GORDONIA CLUSTER DR BACTERIOPHAGES

(Previously published as C.J. Versoza*, A.A. Howell*, T. Aftab, M. Blanco, A. Brar, E. Chaffee, N. Howell, W. Leach, J. Lobatos, M. Luca, M. Maddineni, R. Mirji, C. Mitra, M. Strasser, S. Munig, Z. Patel, M. So, M. Sy, S. Weiss, S.P. Pfeifer. 2022. Comparative genomics of closely-related Gordonia terrae cluster DR bacteriophages. Viruses, 14(8): 1647.)

(Associated appendix previously published C.J. Versoza*, A.A. Howell*, T. Aftab, M. Blanco, A. Brar, E. Chaffee, N. Howell, W. Leach, J. Lobatos, M. Luca, M. Maddineni, R. Mirji, C. Mitra, M. Strasser, S. Munig, Z. Patel, M. So, M. Sy, S. Weiss, C.D. Herren, M. Smith Caldas, S.P. Pfeifer. 2022. Complete genome sequence of the Gordonia bacteriophage BiggityBass. Microbiol. Resour. Announc. e00469-22.)

* contributed equally

Abstract

Bacteriophages infecting bacteria of the genus *Gordonia* have increasingly gained interest in the scientific community for their diverse applications in agriculture, biotechnology, and medicine, ranging from biocontrol agents in wastewater management to the treatment of opportunistic pathogens in pulmonary disease patients. However, due to the time and costs associated with experimental isolation and cultivation, host ranges for many bacteriophages remain poorly characterized, hindering a more efficient usage of bacteriophages in these areas. Here, we perform a series of computational genomic inferences to predict the putative host ranges of all *Gordonia* cluster DR bacteriophages known to date. Our analyses suggest that BiggityBass (as well as several of its close relatives) is likely able to infect host bacteria from a wide range of genera—from *Gordonia* to *Nocardia* to *Rhodococcus,* making it a suitable candidate for future phage therapy and wastewater treatment strategies.

Introduction

Bacteriophages are one of the most abundant organisms on Earth, infecting a wide range of host bacteria present in almost any environment from common garden soil to volcanic substrates and from freshwater streams to oceans (Rohwer 2003). Among these hosts, members of the order *Corynebacteriales*—including *Gordonia*, *Mycobacterium*, *Nocardia*, and *Rhodococcus*—are of particular importance to agriculture, biotechnology, and medicine as the outer membrane of their bacterial cells, which consists of long-chain hydroxylated mycolic acids, frequently leads to complications during the prevention, treatment, and cure of opportunistic pathogens (Dyson et. al 2015). Moreover, due to the hydrophobic nature of this "mycomembrane", *Corynebacteriales* often cause severe problems during wastewater treatment as they can stabilize foams on the surface of aeration tanks during the activated sludge phase (Petrovski et. al 2011), which not only complicates sludge management and increases maintenance costs but also poses a health hazard to wastewater treatment plant workers in their aerosolized form (Pal and Kumar 2014).

Owing to the growing scarcity of clean water across the globe, treated wastewater serves as an important alternative to freshwater for many nations with more than 35% of agricultural irrigation, 17% of landscape irrigation, and 12% of groundwater recharge in the United States stemming from treated wastewater (Kesari et. al 2021). However, microbial hazards, such as multi-drug resistant bacterial pathogens, are frequently discharged into sewage systems due to the common usage of antibiotics in animal farms and on crop fields. Consequently, effective wastewater treatment strategies are indispensable to combat environmental and health concerns for farmers and consumers alike (Dang et. al 2019).

Due to their host specificity, lytic bacteriophages have been proposed as promising and environmentally-friendly bacterial treatment and control agents to remove harmful (or otherwise problematic) bacteria—such as gram-positive *Gordonia* which are associated with both systemic infections in immunocompromised and local infections in immunocompetent individuals (Arenskötter et. al 2004; Grisold et. al 2007) as well as sludge foaming (De los Reyes et. al 1998; Kragelund et. al 2007)—while maintaining desirable microorganisms in the wastewater. To effectively guide these biological control strategies, bacteriophages and their host ranges (i.e., the bacterial genera and species a bacteriophage is able to infect) must be well-characterized—yet, the diversity of *Gordonia* bacteriophages remains largely unexplored.

As part of a course-based undergraduate research experience at Arizona State University, we computationally inferred putative host ranges of all *Gordonia* cluster DR bacteriophages known to date to aid the design and improvement of future wastewater treatment strategies.

Materials and Methods

Genomic data for *Gordonia* cluster DR bacteriophages (Supplementary Table S1) were explored using Phamerator (Cresawn et. al 2011) and phylogenetic relationships characterized together with representative *Microbacterium*, *Mycobacterium*, and *Streptomyces* bacteriophages as outgroups (Supplementary Table S2). Specifically, MAFFT v.7 (Katoh and Standley 2013) embedded within the EMBL-EBI Bioinformatics Toolkit (Zimmerman et. al 2018; Gabler et. al 2020) was used to generate a multiple-sequence alignment between the bacteriophages. The resulting alignment was then used to generate a neighbor-joining tree in MEGA X (Kumar et. al 2018) using a phylogeny test with 10,000 bootstrap replicates. Nucleotide sequence relatedness was assessed using Gepard v.2.1.0 (Krumsiek et. al 2007). Pairwise average nucleotide

identities (ANIs) were calculated using the “Genome Comparison” tool embedded within DNA Master v.5.23.6 and plotted using the ggplot2 package (Wickham 2009) in R v.4.1.0.

Following suggested best practices by Versoza and Pfeifer (2022), a combination of exploratory and confirmatory methods was utilized to computationally predict host ranges of the closely-related *Gordonia* cluster DR bacteriophages. Specifically, putative host ranges were predicted using two machine-learning based prediction tools—CHERRY (Shang and Sun 2022) and PHERI v.0.2 (Baláž et. al 2020)—as well as the alignment-free prediction tool WIsH v.1.1 (Galiez et. al 2017) together with genomic data from ten putative bacterial host species spanning three genera—*Gordonia*, *Nocardia*, *Rhodococcus*, and, as a negative control, *Escherichia* (Supplementary Table S3). All software was executed using default settings.

Results

To confirm cluster membership, the genomes of *Gordonia* cluster DR bacteriophages were investigated. They show a high level of sequence similarity with the left arm of the genomes mostly encoding well-conserved structural and assembly proteins (including a terminase, portal protein, capsid maturation protein as well as major capsid hexamer and pentamer proteins, a head-to-tail adaptor, tail assembly protein, tape measure protein, minor tail protein subunits, lysin A, lysin B, and several genes responsible for integration into the host). Thereby, the RuvC-like resolvase (Supplementary Figure S1), a Holliday junction resolving enzyme that is a distant relative of the RuvC proteins present in gram-negative bacteria such as *Escherichia coli* (Lilley and White 2001) is of particular interest. It closely resembles the RuvC-like endonucleases found in select *Siphoviridae* and *Myoviridae* bacteriophages

infecting *Streptococcus* and *Lactococcus* hosts (Bidnenko 2002; Curtis et. al 2004), which may hint at a shared evolutionary history. The right arm of the genomes contains non-structural genes (including an exonuclease, DNA helicase, DNA polymerase, and HNH endonuclease). Notably, several cluster DR bacteriophages exhibit a partial toxin/antitoxin (TA) system (Supplementary Figure S2). Prevalent in many archaea and bacteria, TA systems encode a toxin protein and a corresponding antitoxin in the form of a protein or non-coding RNA that serves as a defense mechanism against invading bacteriophages (Unterholzner et. al 2013; Song and Wood 2020). As bacteriophages co-evolve with their bacterial hosts (Stern and Sorek 2011), adaptations to such defense mechanisms are common (Rauch et. al 2017) to allow bacteriophages to inactivate bacteria-encoded toxins (Otsuka and Yonesaki 2012; Wei et. al 2016). Indeed, the TA system of the cluster DR bacteriophages is homologous to the *hicA* TA system frequently present in *Burkholderia pseudomallei*, *E. coli*, and *Pseudomonas aeruginosa* (Yamaguchi and Inouye 2011; Butt et. al 2014; Shen et. al 2016).

To elucidate phylogenetic relationships, comparative analyses were performed between all *Gordonia* cluster DR bacteriophages known to date (Supplementary Table S1). Following Pope and colleagues (Pope et. al 2017), clustering was based on nucleotide similarity and shared gene content, with bacteriophages sharing at least 35% of genes being grouped into clusters. A neighbor-joining tree confirmed membership in the DR cluster (Supplementary Figure S3a)—an assignment that was further supported by both the dot plot analyses (Supplementary Figure S4) as well as the pairwise average nucleotide identities (Supplementary Figure S5). Interestingly, gene trees of the RuvC-like resolvase (Supplementary Figure S3b) and the *hicA*-like toxin (Supplementary Figure S3c) do not recapitulate the whole genome phylogeny—however, it is unclear whether this is due to inconsistent resampling during bootstrapping caused by the short

sequence length (Lawrence et. al 2002) or the mosaic architecture of the genome caused by horizontal gene transfer by illegitimate recombination (Ford et. al 1998; Hatfull et. al 2006; Pedulla et. al 2003). Compared to temperate bacteriophages, both gene acquisition and gene loss, in lytic bacteriophages is less well understood (Moura de Sousa et. al 2021). However, there have been previous reports of gene transfers in T4-like and T7-like bacteriophages (Filée et. al 2006; Dekel-Bird et. al 2013) and lytic bacteriophages with large genomes have been suggested to have acquired genes from donor genomes (Mesyanzhinov et. al 2002).

Due to their bactericidal nature, bacteriophages are frequently used for a variety of agricultural, biotechnological, and medical applications (Sharma et. al 2017). To effectively guide the usage of bacteriophages in these areas, their host ranges have to first be determined (see discussion in Versoza and Pfeifer 2022). To investigate the host ranges of the closely related cluster DR bacteriophages, a combination of exploratory and confirmatory prediction tools was utilized together with a dataset of ten putative bacterial host species and *E. coli* as a negative control (Supplementary Table S3). Specifically, the tested host dataset spans the three genera of the *Corynebacteriales* order—*Gordonia*, *Nocardia*, and *Rhodococcus*—that have been implicated in activated sludge foaming in wastewater treatment plants (Goodfellow et. al 1998).

Using the exploratory method PHERI (Baláž et. al 2020), seven out of nine cluster DR bacteriophages were predicted to infect hosts under the *Gordonia* genus (Table 1), with the exception of bacteriophages AnClar and Yago84. To make host range predictions for newly encountered bacteriophages, PHERI utilizes a decision tree classifier of annotated protein clusters of bacteriophages with known hosts. Consequently, bacteriophages will only be predicted to infect a particular host if their

protein profile closely matches that of another bacteriophage known to infect that host. As minor tail proteins play an essential role in bacteriophage infection (Jacobs-Sera et. al 2012), the lack of similarity in the minor tail protein profiles of AnClar and Yago84 compared to those bacteriophages known to infect *Gordonia* hosts might explain why neither were predicted to infect the *Gordonia* genus, despite having been isolated in *G. terrae* (Supplementary Table S1). In fact, the clades observed within the gene tree of the minor tail protein shared across all cluster DR bacteriophages (Supplementary Figure S3d) reflects the clustering of the bacteriophages with respect to host range, reiterating the importance of tail proteins for host infection. Using the exploratory method CHERRY (Shang and Sun 2022) —a graph convolutional encoder and decoder that relies on a broader range of features including protein organization, sequence similarity, and *k*-mer frequency to predict host ranges—highlights *M. smegmatis*, *G. terrae*, and *R. hoagie* as the three most likely host candidates for all cluster DR bacteriophages (though the latter two scoring predictions fell below the recommended confidence threshold of 0.9). Conversely, the confirmatory method WIsH (Galiez et. al 2017)—based on a Markov model that determines the *k*-mer similarity between bacteriophage and host genomes—predicted *G. hydrophobica*, *G. malaquae*, *G. rubripertincta*, and *G. terrae* as potential hosts for all nine cluster DR bacteriophages relative to the negative control, *E. coli* (Figure 1). Moreover, log likelihood values for putative *Nocardia* and *Rhodococcus* hosts were comparable to those of *Gordonia*, suggesting the potential for a much broader host range. Interestingly, BiggityBass exhibits the broadest predicted host range among all cluster DR bacteriophages, spread across five different phyla (Table 1), making it an appealing agent to explore for future wastewater treatment strategies (Ross et. al, 2016).

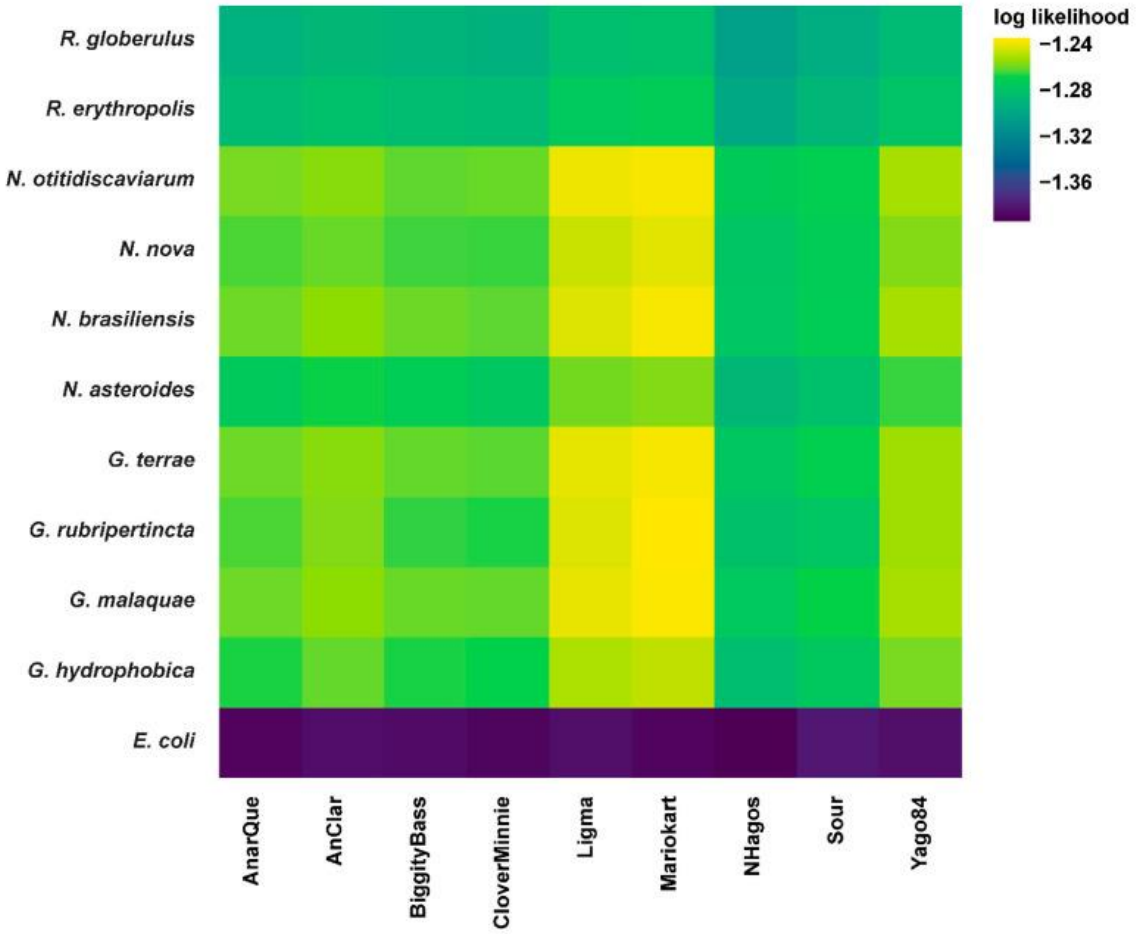


Fig. 1. Putative host ranges as predicted by WIsH. Heatmap of log-likelihoods of bacteriophage-host pairs—including nine *Gordonia* cluster DR bacteriophages (Supplementary Table S1) as well as ten potential bacterial hosts and *E. coli* as a negative control (Supplementary Table S3)—generated by the host prediction tool WIsH (Galiez et. al 2017). Higher values correspond to more likely interactions.

	Gordonia	*Arthrobacter*	*Aeromonas*	*Staphylococcus*	*Shigella*	*Corynebacterium*	*Stenotrophomonas*
AnarQue	√	√	√				
AnClar		√	√				
BiggityBass	√	√	√	√	√		
CloverMinnie	√	√	√				
Ligma	√	√	√			√	
Mariokart	√	√	√				
NHagos	√	√	√				
Sour	√	√	√				√
Yago84		√	√				

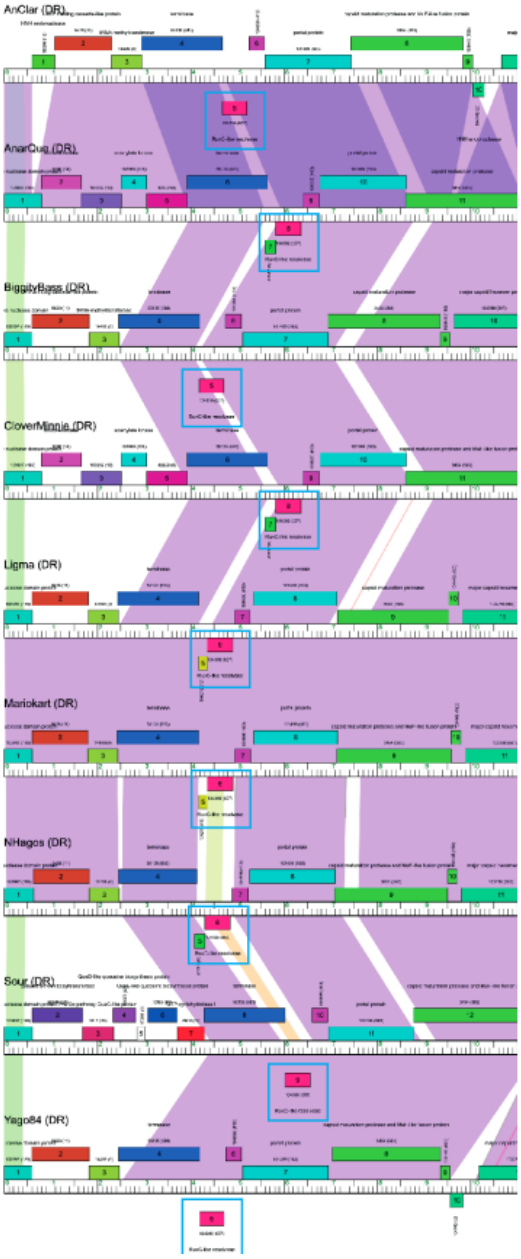
Table 1. Putative host ranges as predicted by PHERI. Putative hosts of the nine Gordonia cluster DR bacteriophages included in this study (Supplementary Table S1) predicted by PHERI (Baláž et. al 2020).

In conclusion, computational methods can offer a first glimpse into the putative host ranges of newly discovered bacteriophages—yet, it is important to remember that these methods are predictive by their very nature. Thereby, each computational method exhibits their own advantages and limitations. For example, tools that rely solely on k-mer-based models can lead to an overprediction of host ranges if convergent evolution resulted in similar nucleotide frequency patterns (Ahlgren et. al 2016), whereas tools that rely on machine-learning are inherently limited in their predictions by the bacteriophage-host datasets available for training (Versoza and Pfeifer 2022). Experimental validation through bacteriophage isolation and cultivation still remains the “gold standard” in determining bacteriophage host ranges—however, it certainly is not without its own limitations as not all microbial hosts are amendable to cultivation in the laboratory and, even if they are, results may depend on the conditions under which the experiments were performed (Versoza and Pfeifer 2022). Given the ever growing knowledge of bacteriophage diversity across the globe, it is our hope that future computational and experimental research will go hand in hand to further explore polyvalent bacteriophages

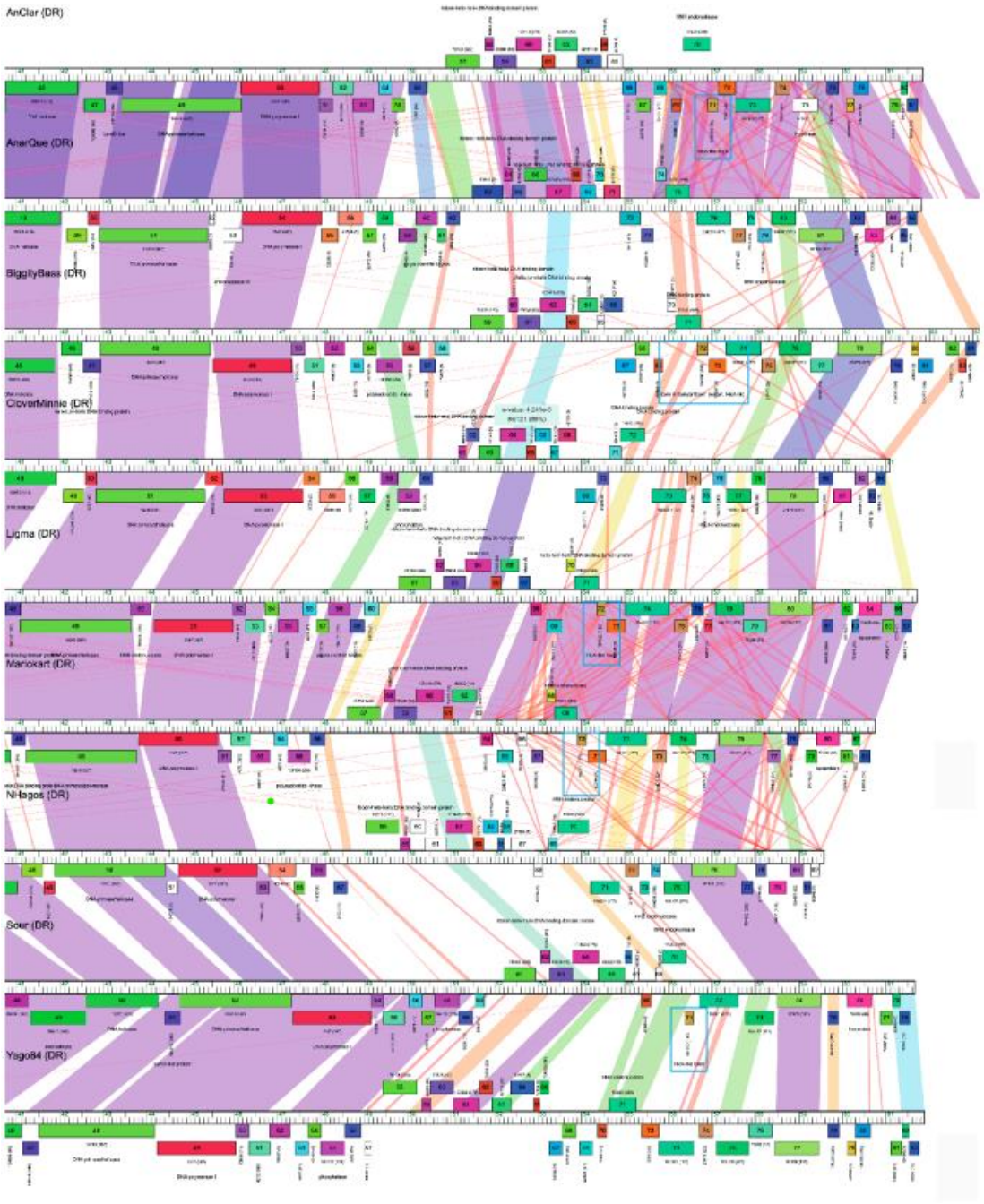
as an interesting study system to gain a better understanding of the molecular and genetic determinants underlying host range.

Supplementary Materials

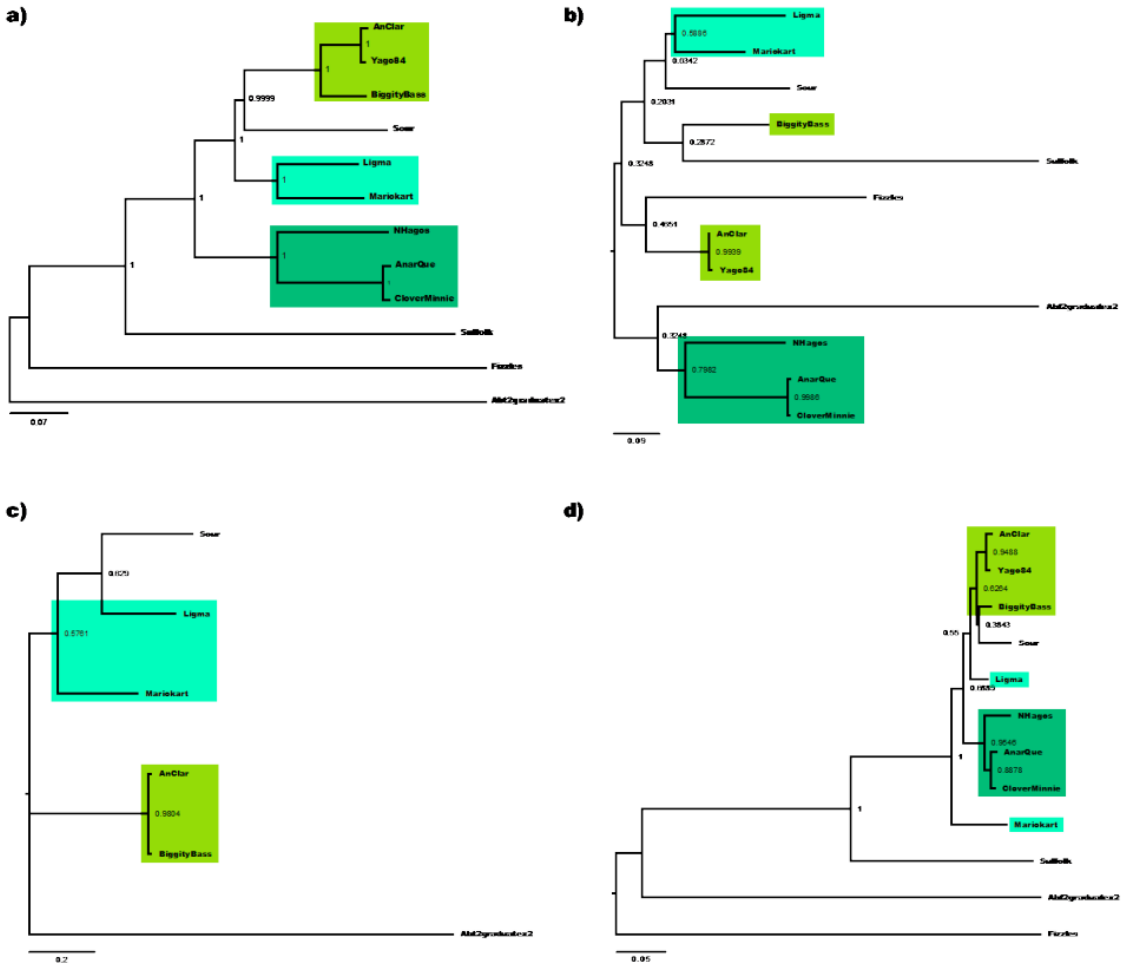
The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/v14081647/s1, Figure S1: Phamerator map of the RuvC-like resolvase gene; Figure S2: Phamerator map of the hicA-like toxin gene; Figure S3: Neighbor-joining trees; Figure S4: Dot plots; Figure S5: Average nucleotide identities; Table S1: Gordonia cluster DR bacteriophages included in the comparative analyses; Table S2: Bacteriophages included as outgroups in the comparative analyses; Table S3: Host bacteria included in the comparative analyses



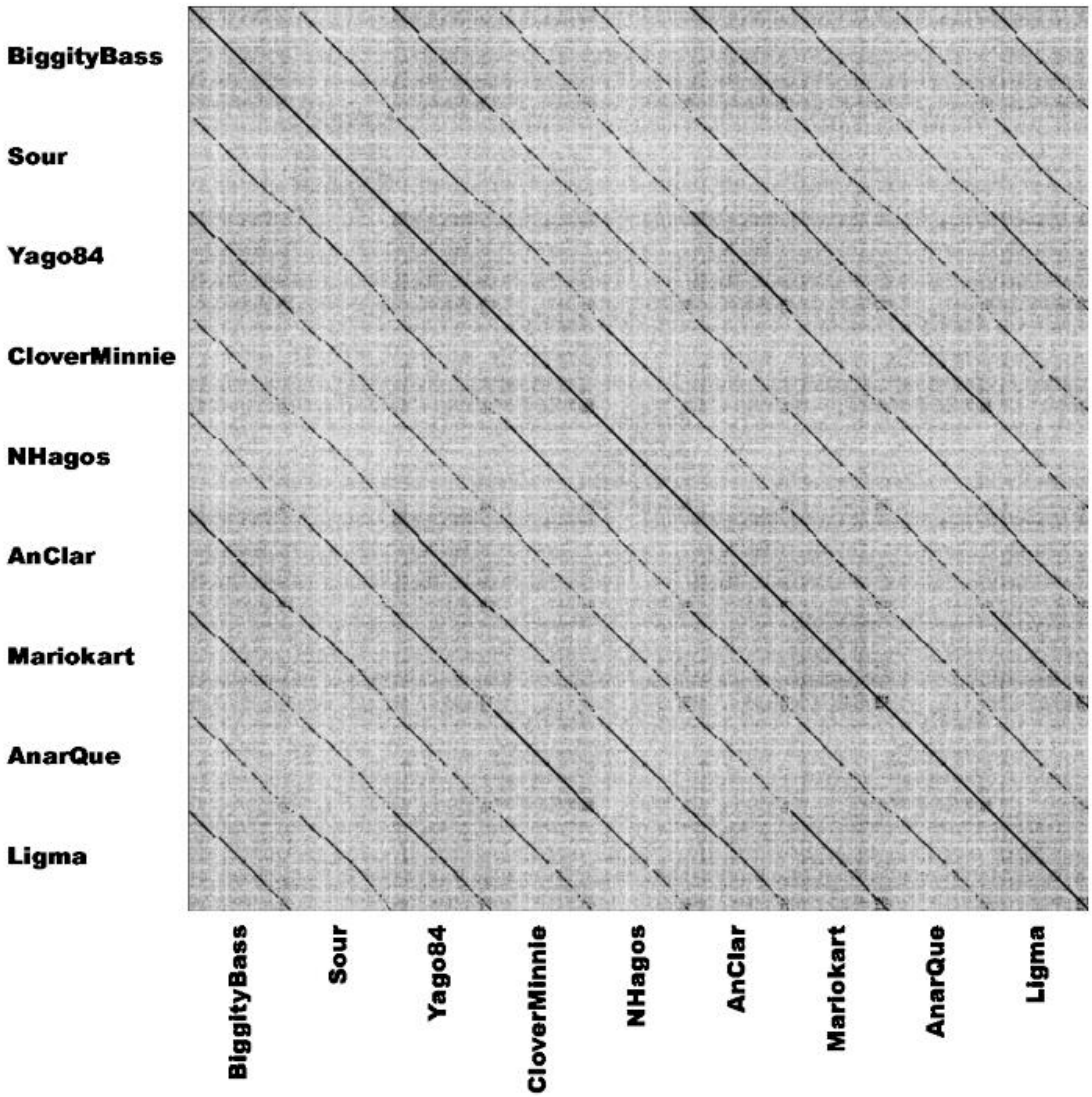
Supplementary Figure S1. Phamerator map of the RuvC-like resolvase gene of closely related *Gordonia* cluster DR bacteriophages (Supplementary Table S1). In this Phamerator map, protein-coding genes with their putative functional assignments (if available) are displayed above or below a ruler, signifying genes on forward or reverse strands, respectively. The numbers shown above each gene indicate the protein family (pham) and, in parenthesis, the number of members in the pham family. Coloring between genomes represents nucleotide similarity with areas of highest similarity shown in purple (BLAST e-value = 0), followed by red (BLAST e-value of ~10-4) and white (no significant similarity).



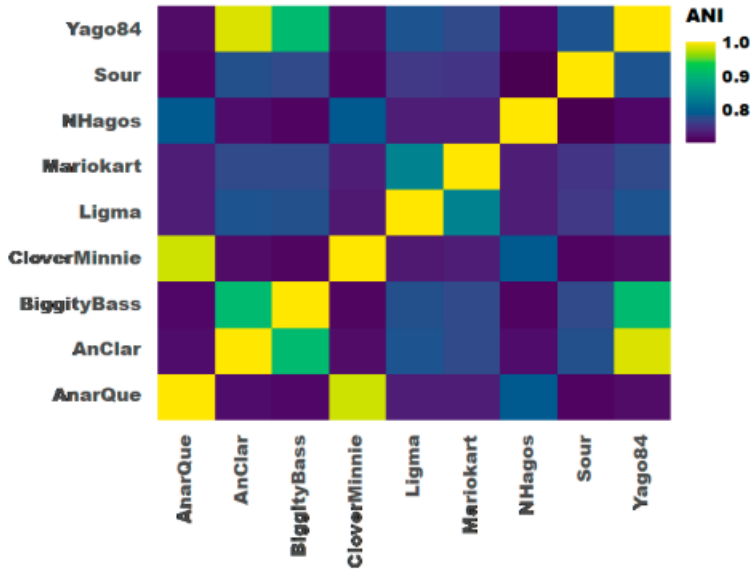
Supplementary Figure S2. Phamerator map of the hicA-like toxin gene of closely related *Gordonia* cluster DR bacteriophages (Supplementary Table S1). In this Phamerator map, protein-coding genes with their putative functional assignments (if available) are displayed above or below a ruler, signifying genes on forward or reverse strands, respectively. The numbers shown above each gene indicate the protein family (pham) and, in parenthesis, the number of members in the pham family. Coloring between genomes represents nucleotide similarity with areas of highest similarity shown in purple (BLAST e-value = 0), followed by red (BLAST e-value of ~10-4) and white (no significant similarity).



Supplementary Figure S3. Neighbor-joining trees generated in MAFFT using the multiple-sequence alignment of (a) nine *Gordonia* cluster DR bacteriophage genomes (Supplementary Table S1) and their corresponding (b) RuvC-like resolvase, (c) hicA-like toxin gene, and (d) minor tail protein with 10,000 bootstrap replicates. Representative *Microbacterium, Mycobacterium, and Streptomyces* bacteriophages were included as outgroups (Supplementary Table S2).



Supplementary Figure S4. Dot plots of closely-related *Gordonia* cluster DR bacteriophages (Supplementary Table S1).



Supplementary Figure S5. Average nucleotide identities (ANIs) of closely-related *Gordonia* cluster DR bacteriophages (Supplementary Table S1).

Phage Name	Isolation host	Length (bp)	GC-content	# ORFs	# tRNAs	Accession #	Reference
AnarQue	*G. rubripertincta* NRRL B-16540	61,822	68.8	86	0	OK216879	Curran *et al.* 2022
AnClar	*G. terrae* 3612	61,856	69.8	81	1	MN908693	unpublished
BiggityBass	*G. terrae* CAG3	63,202	69.4	83	0	ON260813	Versoza, Howell *et al.*
CloverMinnie	*G. terrae* 3612	61,098	68.7	84	0	MN234196	unpublished
Ligma	*G. terrae* NRRL B-16283	61,714	70.2	87	0	OM105886	unpublished
Mariokart	*G. terrae* NRRL B-16283	60,762	70.5	83	0	MT657335.1	unpublished
NHagos	*G. rubripertincta* NRRL B-16540	59,580	68.2	82	0	MN369758.1	Harrington *et al.* 2020
Sour	*G. terrae* NRRL B-16283	61,670	68.0	79	1	NC_042132.1	unpublished
Yago84	*G. terrae* 3612	61,890	70.0	83	0	MK801725.1	Pope *et al.* 2020

Supplementary Table S1. *Gordonia* cluster DR bacteriophages included in the comparative analyses. For detailed information on each bacteriophage, please visit the Howard Hughes Medical Institute (HHMI) – Science Education Alliance (SEA) Phage Hunters Advancing Genomics and Evolutionary Science (PHAGES) website at http://phagesdb.org.

Phage Name	Isolation host	Length (bp)	GC-content	# ORFs	# tRNAs	Accession #	Reference
Abt2graduatex2	*S. griseus* ATCC 10137	57,385	69.2	71	0	MF975638.1	Erill & Caruso 2018
Fizzles	*M. foliorum* NRRL B-24224	62,078	68.2	104	0	MW924638.1	unpublished
Suffolk	*M. smegmatis* mc^2 155	68,262	66.6	97	0	KF713485.1	Pope *et al.* 2015

Supplementary Table S2. Bacteriophages included as outgroups in the comparative analyses. Abt2graduatex2 contains a hicA-like toxin (pham 34446) whereas Fizzles and Suffolk contain a RuvC-like resolvase (pham 34304). For detailed information on each bacteriophage, please visit the Howard Hughes Medical Institute (HHMI) – Science Education Alliance (SEA) Phage Hunters Advancing Genomics and Evolutionary Science (PHAGES) website at http://phagesdb.org.

Host	# Genes	Length (kb)	GC-content	Accession #	Reference
E. coli	4,091	4,509	50.6	CP028765	Kang *et al.* 2021
G. hydrophobica	3,962	4,579	67.5	JAFBGB010000001	unpublished
G. malaque	4,349	4,714	66.2	FNRZ01000008	unpublished
G. rubripertincta	4,740	5,104	67.5	CP059694	Han *et al.* 2020
G. terrae	4,979	5,709	67.8	CP029604	unpublished
N. asteroides	6,341	6,987	71.8	CP089214	Sichtig *et al.* 2019
N. brasiliensis	7,949	8,936	68.2	CP022088	Sichtig *et al.* 2019
N. nova	7,450	8,349	67.8	CP006850	Luo *et al.* 2014
N. otitidiscaviarum	6,735	7,688	69.1	CP041695	unpublished
R. erythropolis	6,025	6,509	62.5	CP032403	unpublished
R. globerulus	6,013	6,740	61.7	CP079698	Lozano-Andrade *et al.* 2021

Supplementary Table S3. Host bacteria included in the comparative analyses.

Supplementary References

Curran, E., S. E. Callaway, R. R. Dumanlang, A. V. Harshaw, P. N. Palacio, et al. 2022. Genome sequences of Gordonia rubripertincta bacteriophages AnarQue and Figliar. Microbiol. Resourc. Announc. 11(1): e01085-21.

Erill, I., S. M. Caruso. 2018. Complete genome sequence of Streptomyces bacteriophage abt2graduateex2. Genome Announc. 6(3): e01480-17.

Han, S. S., H. K. Kang, B. Y. Jo, B. G. Ryu, H. M. Jin, et al. 2020. Complete genome sequence of Gordonia rubripertincta SD5, a soil bacterium isolated from a Di-(2-Ethylhexyl) Phthalate-degrading enrichment culture. Microbiol. Resourc. Announc. 9(45): e01087-20.

Harrington, D. A. L., J. L. Stevens, M. J. Johnson, S. J. Pochiro, M. M. Moriarty, et al. 2020. Genome sequences of Gordonia rubripertincta bacteriophages Jellybones and NHagos. Microbiol. Resourc. Announc. 9(40): e00935-20.

Kang, Y., L. Yuan, X. Shi, Y. Chu, Z. He, et al. 2021. A fine-scale map of genome-wide recombination in divergent Escherichia coli population. Brief. Bioinform. 22(4): bbaa335.

Lozano-Andrade, C. N., M. L. Strube, Á. T. Kovács. 2021. Complete genome sequences of four soil-derived isolates for studying synthetic bacterial community assembly. Microbiol. Resourc. Announc. 10(46): e00848-21.

Luo, Q., S. Hiessl, A. Poehlein, R. Daniel, A. Steinbüchel. 2014. Insights into the microbial degradation of rubber and gutta-percha by analysis of the complete genome of Nocardia nova SH22a. Appl. Environ. Microbiol. 80(13): 3895–3907.

Pope, W. H., C. A. Bowman, D. A. Russell, D. Jacobs-Sera, D. J. Asai, et al. 2015. Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. Elife. 4: e06416.

Pope, W. H., K. A. Butela, R. A. Garlena, D. Jacobs-Sera, D. A. Russell, et al. 2020. Genome sequences of 20 bacteriophages isolated on Gordonia terrae. Microbiol. Resourc. Announc. 9(3): e01489-19.

Sichtig, H., T. Minogue, Y. Yan, C. Stefan, A. Hall, et al. 2019 FDA-ARGOS is a database with public quality-controlled reference genomes for diagnostic use and regulatory science. Nat. Commun. 10(1), 1–13.

Versoza, C.J., A.A. Howell, T. Aftab, M. Blanco, A. Brar, E. Chaffee, N. Howell, W. Leach, J. Lobatos, M. Luca, et al. The complete genome sequence of the Gordonia bacteriophage BiggityBass. Accepted, Microbiol. Resourc. Announc

CHAPTER 4

EVALUATING THE PERFORMANCE OF HOST RANGE PREDICTION TOOLS FOR POLYVALENT BACTERIOPHAGES

(Currently in review as A.A. Howell*, C.J. Versoza*, S.P. Pfeifer. Computational host range prediction – the good, the bad and the ugly.)

* contributed equally

Abstract

The rapid emergence and spread of antimicrobial resistance across the globe has prompted the usage of bacteriophages (i.e., viruses that infect bacteria) in a variety of applications ranging from agriculture to biotechnology and medicine. In order to effectively guide the application of bacteriophages in these multifaceted areas, information about their host ranges – that is the bacterial strains or species that a bacteriophage can successfully infect and kill – is essential. Utilizing 16 broad-spectrum (polyvalent) bacteriophages with experimentally validated host ranges, we here benchmark the performance of 11 recently developed computational host range prediction tools that provide a promising and highly scalable supplement to traditional, but laborious, experimental procedures. We show that machine- and deep-learning approaches offer the highest levels of accuracy and precision – however, their predominant predictions at the species- or genus-level render them ill-suited for applications outside of an ecosystems metagenomics framework. In contrast, only moderate sensitivity (<80%) could be reached at the strain-level, albeit at low levels of precision (<40%). Taken together, these limitations demonstrate that there remains room for improvement in the active scientific field of *in silico* host prediction to combat the challenge of guiding experimental designs to identify the most promising bacteriophage candidates for any given application.

INTRODUCTION

Due to the rise of antimicrobial resistance – projected to lead to an estimated 10 million deaths per year (Furfaro et al. 2018) and an economic loss of US$100 trillion by 2050 across the globe (Manesh et al. 2021) – bacteriophages (i.e., viruses that infect, and replicate within, bacteria) are now being routinely used in a wide variety of fields as alternative to antibiotics for combating bacterial infections. Specifically, their applications range from agriculture (e.g., as biopesticides to combat plant pathogens in crops or biocontrol agents to manage bacterial infections in aquaculture or livestock on organic farms; Kuek et al. 2022), to food safety, production, and processing (e.g., to prevent or eliminate bacterial contaminations responsible for foodborne illnesses such as those caused by *Escherichia coli*, *Listeria*, and *Salmonella* bacteria; Oh and Park 2017; Moye et al. 2018; López-Cuevas et al. 2021), to biotechnology (e.g., as biosensing devices to detect specific bacterial strains; Harada et al. 2018), and to wastewater treatment (e.g., to regulate bacteria that negatively impact water quality, cause environmental problems, or affect industrial processes; Petrovski et al. 2011a,b). More recently, bacteriophages have also been rediscovered as agents in medical applications, including diagnostics to detect pathogenic bacteria (Monk et al. 2010), bacteriophage therapy to treat multi-drug-resistant bacterial infections (Sulakvelidze et al. 2011; Nobrega et al. 2015), bacteriophage display to discover antibodies, peptides, or proteins that bind to, for example, cancer cells (Pande et al. 2010), as well as gene therapy, drug design, and delivery (Vaks and Benhar 2011; Omidfar and Daneshpour 2015). In addition, bacteriophages are an important tool in scientific research, in particular for the study of bacterial evolution, antibiotic resistance, as well as the genetic and evolutionary mechanisms underlying viral infectious diseases (Koskella and Brockhurst 2014). In order to effectively guide the usage of bacteriophages in these multifaceted areas, a firm

understanding of their host specificity as well as their efficacy in combating bacterial pathogens must first be established – knowledge which remains largely elusive.

As natural predators of bacteria, identifying the most suitable bacteriophage for any given application requires an understanding of its host range, i.e., the bacterial strains or species that a bacteriophage can successfully hijack and kill (lyse). For example, a collection of bacteriophages with different, often overlapping, host ranges (so-called “bacteriophage cocktails”) is frequently harnessed to treat antibiotic-resistant bacterial pathogens without impacting the microorganisms beneficial to a patient (Dedrick et al. 2021; Little et al. 2022; Nick et al. 2022; Dedrick et al. 2023; and see review of Hatfull et al. 2022) or to target and control the spread of bacterial pathogens in food production without impacting consumer safety (Soffer et al. 2017; Zhang et al. 2019). To identify host-specific bacteriophages, traditional experimental procedures remain the gold standard; these techniques comprise of bacteriophage display libraries or assays that rely on plaque formation on agar plates (spot and plaque assays), optical density fluctuations in liquid cultures (liquid assays), and fluorescent labeling (viral tagging and bacteriophage fluorescence *in situ* hybridization) (for detailed information, see Box 1 of Edwards et al. 2016). However, experimental host-range determinations are, by their very nature, restricted to bacteriophages and microbial hosts that can be successfully cultivated in the laboratory under simplified growth conditions – in particular with regards to growth media, temperature, pH, and UV light – which may not fully capture the complexity of natural environments. Moreover, culturing bacteriophages and performing host assays remains a laborious, time-consuming, and expensive process, thus limiting its potential for scalable high-throughput screening (Wade 2002; Edwards and Rohwer 2005; Coutinho et al. 2019). As a consequence, several bioinformatic software packages have recently been developed to predict bacteriophage-host ranges

in silico, aiding the prioritization of experimental efforts by identifying the most promising bacteriophage candidates suitable for lysing a specific bacterial strain that may then be further studied in the laboratory.

Many such bacteriophage host range prediction tools have been developed in recent years (see review of Versoza and Pfeifer 2022). They can broadly be grouped into three categories: (a) alignment-based methods relying on sequence homology and/or sequence similarity between bacteriophages and their bacterial hosts originating from integrated prophages, short viral DNA sequences incorporated into the clustered regularly interspaced short palindromic repeat (CRISPR) loci of the host genome, tRNA genes, and/or genomic segments shared by horizontal gene transfer (with frequently used tools including Phirbo [Zielezinski et al. 2021], PHIST [Zielezinski et al. 2022], and VPF-Class [Pons et al. 2021]), (b) alignment-free methods based on sequence composition such as oligonucleotide or *k*-mer (i.e., nucleotide sequences of length *k*) frequencies that may result, for example, from shared patterns of codon usage as bacteriophages corrupt the host's replication machinery for protein synthesis (Carbone 2008) or protein clustering associated with host recognition and binding (e.g., VirHostMatcher [Ahlgren et al. 2017], and WIsH [Galiez et al. 2017]), and (c) machine- / deep-learning-based methods trained on experimentally validated datasets of bacteriophage-host interactions to develop predictive statistical models that often incorporate multiple features (e.g., nucleotide and amino acid sequence and properties, protein interactions, and/or structural characteristics such as capsid proteins or tail fibers that can contribute to host specificity) to predict bacteriophage host ranges (e.g., CHERRY [Shang and Sun 2022], HostG [Shang and Sun 2021], Prokaryotic virus Host Predictor [Lu et al. 2021], RaFAH [Coutinho et al. 2021], VirHostMatcher-Net [Wang et al. 2020], and vHULK [Amgarten et al. 2022]).

Due to the complexity and diversity of bacteriophage-host interactions, the computational prediction of host ranges based on genomic data is a challenging task and the power of recently developed methodologies is often not well-established. Further complicating this issue, a lack of standardized evaluation criteria is hindering systematic assessments as well as consistent performance benchmarking across different approaches. The limited comparisons currently available (e.g., Edwards et al. 2016; Ahlgren et al. 2017; Baláž et al. 2023) have taken advantage of bacteriophage-host pairs available to the research community through public databases such as the genomic resources maintained by the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/), the European Bioinformatics Institute (EMBL-EBI; https://www.ebi.ac.uk/), and the Actinobacteriophage database (phagesdb; https://phagesdb.org/) – not all entries of which have been experimentally validated. In addition, while these databases allow developers to assess both "true positives" (that is a bacteriophage-host interaction was computationally predicted and the available data suggested that the bacteriophage can infect the host) and "false negatives" (that is no bacteriophage-host interaction was predicted although the data suggested that the bacteriophage can infect the host), the almost complete absence of experimentally validated data that can attest to a bacteriophage not being able to infect a specific bacterial strain makes it impossible to assess "false positives" and "true negatives". Making matters worse, without experimental validation, the absence of a bacteriophage-host pair from these databases is usually taken as evidence that a bacteriophage is not able to infect a bacterial strain, thus confounding previously reported levels of precision and specificity. Lastly, these comparisons often implicitly assume that a bacteriophage can only infect a single bacterial host, despite some bacteriophages showing much broader natural host ranges (see discussion in Edwards et al. 2016).

Polyvalent (or broad-spectrum) bacteriophages are a particularly interesting study system in this regard as they are able to recognize common cell-surface receptors, allowing them to infect and lyse several different bacterial strains or species – sometimes from across multiple genera – that share these receptor characteristics. Due to their broad host range, they provide a unique opportunity for testing the sensitivity and specificity of host range prediction tools. Utilizing three polyvalent *E. coli* bacteriophages and 13 polyvalent *Gordonia* bacteriophages with experimentally validated host ranges, we here assess the performance of 11 computational host range prediction tools and discuss important factors to consider when implementing these computational methods.

Materials and Methods

Experimental Data

Computational host range prediction tools were evaluated using three polyvalent *E. coli* bacteriophages – HY01 (Lee et al. 2016), KFS-EC3 (Kim et al. 2021), and SFP10 (Park et al. 2012) – as well as 13 polyvalent *Gordonia* bacteriophages – GTE2 (Petrovski et al. 2011a), GTE7 (Petrovski et al. 2011b), GTE5 and GRU1 (Petrovski et al. 2012), as well as GMA2–GMA7, GRU3, GTE6, and GTE8 (Dyson et al. 2015) – whose host ranges were previously determined experimentally (for details, see Supplementary Tables S1 and S2, respectively). In brief, genome assemblies for all bacteriophages were downloaded from NCBI (using the accession numbers provided in Supplementary Tables S1 and S2). Genome assemblies of experimentally validated *E. coli* bacteriophage host and non-host strains were downloaded from the American Type Culture Collection (ATCC; https://www.atcc.org/) and NCBI (Supplementary Table S1) whereas genomes of experimentally validated *Gordonia* bacteriophage host and non-host strains were newly sequenced and *de novo* assembled as described below.

DNA Isolation, Library Preparation, and Long-Read Sequencing. High molecular-weight genomic DNA from five *Gordonia* strains – *Gordonia hydrophobica* DSM 44015, *Gordonia malaquae* DSM 44454, *Gordonia malaquae* DSM 44464, *Gordonia rubripertincta* DSM 43197, and *Gordonia terrae* DSM 43249 – was isolated using the QIAGEN Genomic-tip 100 / G Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. A barcoded sequencing library was prepared using the Oxford Nanopore Ligation Sequencing Kit (SQK-LSK109) together with the PCR-free Native Barcoding Expansion Kit (EXP-NBD114; Oxford Nanopore Technologies, Oxford, UK) and sequenced on an R9.4.1 FLO-MIN106 flow cell on the GridION X5 Mk1 platform for 72 hours. Reads were base-called in high-accuracy mode, validated using fastQValidator v.0.1.1a (https://github.com/statgen/fastQValidator), and quality controlled using pycoQC v.2.5.2 (Leger and Leonardi 2019).

De Novo Genome Assembly. High quality bacterial genome assemblies were generated for the five sequenced *Gordonia* strains. Prior to the assembly, genome size, repeat content, and coverage were estimated based on *k*-mer frequencies observed in the long read data using GenomeScope2.0 (Vurture et al. 2017; Ranallo-Benavidez et al. 2020) together with Jellyfish v.2.3.0 (Marçais and Kingsford 2011) (Supplementary Table S3). Reads were then *de novo* assembled using Flye v.2.9.2-b1786 (Kolmogorov et al. 2019) and one round of polishing was performed using Medaka v.1.7.2 (https://github.com/nanoporetech/medaka) to improve accuracy. To assess the completeness of the genome assemblies, BUSCO v.5.4.7 (Manni et al. 2021) was used, together with the actinobacteria database "actinobacteria_class_odb10" (for additional details, see Supplementary Table S4). All software was executed using default settings.

Computational Host Range Prediction

Computational host range prediction tools can be divided into two groups: (i) confirmatory methods that utilize a set of bacterial genomes provided by the user to infer the likelihood of a bacteriophage-host interaction and (ii) exploratory methods that predict bacteriophage-host interactions based on a set of bacteriophage genomes provided by the user and an internal database of putative host genomes. Bacteriophage host ranges were computationally predicted using the confirmatory tools Phirbo v.1.0 (Zielezinski et al. 2021), PHIST v.1.1 (Zielezinski et al. 2022), Prokaryotic virus Host Predictor (PHP) v.1.0 (Lu et al. 2021), VirHostMatcher v.1.0 (Ahlgren et al. 2017), and WIsH v.1.1 (Galiez et al. 2017), as well as the exploratory tools CHERRY v.1.0 (Shang and Sun 2022), HostG v.1.0 (Shang and Sun 2021), Random Forest Assignment of Hosts (RaFAH) v.1.0 (Coutinho et al. 2021), viral Host UnveiLing Kit (vHULK) v.2.0 (Amgarten et al. 2022), VirHostMatcher-Net v.1.0 (Wang et al. 2020), and VPF-Class v.1.0 (Pons et al. 2021). For the confirmatory tools (Phirbo, PHIST, PHP, VirHostMatcher, and WIsH), performance was evaluated based on the experimentally validated host and non-host bacterial strains (Supplementary Tables S5 and S6). Out of the five confirmatory tools, WIsH required the construction of a null model consisting of bacteriophage genomes known not to infect the bacterial strain(s) to compute the likelihood for a particular bacteriophage-host pair under a trained homogeneous Markov chain model for the host genome. To test the potential impact of null model construction on predictions, four different null models were tested based on bacteriophage genomes available in the Actinobacteriophage database (Supplementary Table S7). The first two models consisted of bacteriophage genomes expected not to infect any of the tested host strains: (1) a null model based on a large, diverse set of *Alteromonas*, *Cellulophage*, *Cyanophage*, *Lactobacillus*, *Mycobacterium*, *Oenococcus*, *Pelagibacter*, *Prochlorococcus*, *Rhizobium*, *Synechococcus*, and *Thermus* bacteriophage genomes

and (2) a null model based on a small set of *Synechococcus* bacteriophage genomes only (i.e., genomes of bacteriophages known to infect an unrelated bacterial genus). In addition, two model misspecifications were tested by including bacteriophage genomes known to infect host strains included in this study: (3) a null model based on a large, diverse set of *Alteromonas*, *Cellulophage*, *Cyanophage*, *Escherichia coli, Lactobacillus*, *Mycobacterium*, *Oenococcus*, *Pelagibacter*, *Prochlorococcus*, *Rhizobium*, *Synechococcus*, and *Thermus* bacteriophage genomes and (4) a null model based on a small set of *Escherichia coli* bacteriophages only. In contrast, exploratory tools predict bacteriophage-host interactions based on inbuilt databases either at the species-level (CHERRY and VirHostMatcher-Net) or genus-level (HostG, RaFAH, vHULK, and VPF-Class) and their performance was evaluated based on these databases (Supplementary Tables S5 and S8). All software was executed using default settings with recommended tool-specific thresholds (as indicated in Supplementary Table S5).

Comparative Genomic Analyses

Pairwise average nucleotide identities (ANIs) between (i) the three *E. coli* bacteriophages HY01, KFS-EC3, and SFP10, as well as the 13 *Gordonia* bacteriophages GMA2-7, GRU1, GRU3, GTE2, and GTE5-8 (Supplementary Figure S1) and (ii) the experimentally validated host and non-host genomes as well as genomes of closely-related bacterial strains included in the exploratory tool databases (Supplementary Figures S2 and S3 for *E. coli* and *Gordonia*, respectively) were calculated using anvi'o v.7.1 (Eren et al. 2015). Additionally, to gain information about the putative causes of exploratory tool mis-predictions, PHASTER (Arndt et al. 2016) was used to search the genome of mis-predicted hosts for integrated prophages (Supplementary Figure S4).

Results and Discussion

The performance of 11 computational host prediction tools was evaluated using three polyvalent *E. coli* bacteriophages and 13 polyvalent *Gordonia* bacteriophages for which host ranges were previously experimentally validated (for details, see Supplementary Tables S1 and S2). Out of the 11 computational prediction methods, three were alignment-based (Phirbo [Zielezinski et al. 2021], PHIST [Zielezinski et al. 2022], and VPF-Class [Pons et al. 2021]), two alignment-free (VirHostMatcher [Ahlgren et al. 2017] and WIsH [Galiez et al. 2017]), and six machine- or deep-learning-based (CHERRY [Shang and Sun 2022], PHP [Lu et al. 2021], HostG [Shang and Sun 2021], RaFAH [Coutinho et al. 2021], vHULK [Amgarten et al. 2022], and VirHostMatcher-Net [Wang et al. 2020]).

Confirmatory Tools

The five confirmatory tools – Phirbo (Zielezinski et al. 2021), PHIST (Zielezinski et al. 2022), PHP (Lu et al. 2021), VirHostMatcher (Ahlgren et al. 2017), and WIsH (Galiez et al. 2017) – require a set of candidate bacterial genomes provided by the user to infer the likelihood of a bacteriophage-host interaction. Thus, in order to predict putative host ranges for the 16 bacteriophages included in this study, datasets consisting of genome assemblies of all experimentally tested bacterial strains (that is infected and non-infected) were provided to the confirmatory tools. As well-studied model organism, such genomic datasets were readily available for experimentally validated *E. coli* bacteriophage host and non-host strains from the public ATCC and NCBI databases (using accession numbers provided in Supplementary Table S1). In contrast, genomes of five experimentally tested *Gordonia* strains – *Gordonia hydrophobica* DSM 44015,

Gordonia malaquae DSM 44454, *Gordonia malaquae* DSM 44464, *Gordonia rubripertincta* DSM 43197, and *Gordonia terrae* DSM 43249 (Supplementary Table S2) – were newly sequenced to approximately 160-fold to 360-fold coverage per strain (Supplementary Table S3) using long-read nanopore sequencing. Following the Oxford Nanopore Technologies Best Practices (https://nanoporetech.com/sites/default/files/s3/literature/microbial-genome-assembly-workflow.pdf), reads were *de novo* assembled using Flye (Kolmogorov et al. 2019) and polished using Medaka (https://github.com/nanoporetech/medaka) to improve accuracy. The resulting single-scaffold genome assemblies ranged from 4,468,569 bp (*Gordonia malaquae* DSM 44454) to 5,701,739 bp (*Gordonia terrae* DSM 43249) in size, with a GC-content of 66.2%–67.8% (Supplementary Table S4). Highly conserved single-copy orthologous actinobacteria genes (BUSCOs) demonstrated that these *Gordonia* assemblies are nearly complete, containing between 98.0% (*Gordonia rubripertincta* DSM 43197) and 99.4% (*Gordonia malaquae* DSM 44454) of BUSCOs (Supplementary Table S4).

Out of the confirmatory tools, PHP – which uses a Gaussian mixture model of differences in 4-mer sequence composition between bacteriophage and bacterial genomic sequences to predict putative hosts (i.e., bacterial strains with the lowest oligonucleotide dissimilarity) – exhibited the highest sensitivity (77.4%) (Table 1, and see Supplementary Tables S5 and S6 for additional details regarding the predicted bacteriophage-host interactions that passed recommended tool-specific thresholds). Based on a more specific 6-mer approach, VirHostMatcher's background-subtracting d_2^* similarity measure yielded a much lower sensitivity (12.9%); only WIsH's stringent 8-mer approach exhibited a lower recall (0.0%), identifying none of the genuine host strains of the 16 polyvalent bacteriophages. At the same time, the usage of longer *k*-mers also

increased specificity, from 55.3% in PHP to 83.5% and 90.6% in WIsH and VirHostMatcher, respectively. Notably, none of the predictions of VirHostMatcher and WIsH passed the recommended tool-specific thresholds for any of the *E. coli* and *Gordonia* bacteriophages, respectively (Figure 1). More generally, fewer results were observed for *Gordonia* bacteriophages, with PHP and VirHostMatcher only yielding predictions for GMA4, GMA7 and the closely-related GTE7 (PHP only), as well as GRU1 and the closely-related GTE5 and GTE8 (for pairwise average nucleotide identities between the bacteriophages, see Supplementary Figure S1), likely due to the fact that *E. coli* is a more widely studied model organism than *Gordonia*.

		tool (threshold)	sensitivity	specificity	precision	accuracy
confirmatory	*strain-level*	Phirbo (highest rank-based overlap)	19.4%	88.2%	37.5%	**69.8%**
		PHP (log(P(host))[1]: 1442)	**77.4%**	55.3%	**38.7%**	61.2%
		VirHostMatcher (distance / dissimilarity: 0.175)	12.9%	**90.6%**	33.3%	**69.8%**
		WIsH (p-value < 0.06)	0.0%	83.5%	0.0%	61.2%
exploratory	*species-level*	CHERRY (P(graph convolutional encoder): 0.9)	**47.6%**	97.4%	**60.6%**	**93.6%**
		VHMN (prediction score[2]: 0.95)	10.0%	**98.1%**	28.6%	91.7%
	genus-level	HostG (SoftMax value: 0.94)	31.3%	**100.0%**	**100.0%**	91.2%
		RaFAH (prediction score[3]: 0.14)	**88.9%**	96.9%	88.9%	**95.1%**
		vHULK (alignment significance score: 0.8)	52.2%	**100.0%**	**100.0%**	91.7%
		VPF-Class (membership: 0.3, confidence: 0.5)	35.3%	97.7%	75.0%	87.6%

Table 1. Performance of computational host range prediction tools. Performance of the confirmatory tools Phirbo, Prokaryotic virus Host Predictor (PHP), VirHostMatcher, and WIsH as well as the species-level exploratory tools CHERRY and VirHostMatcher-Net [VHMN] and the genus-level exploratory tools HostG, Random Forest Assignment of Hosts [RaFAH], viral Host UnveiLing Kit [vHULK], and VPF-Class. All tools were executed using default settings with recommended tool-specific thresholds (shown in brackets). The sensitivity / recall, specificity, precision, and accuracy of each tool was evaluated based on experimentally validated bacteriophage-host interactions (see Supplementary Tables S1 and S2 as well as Tables 1 in Park et al. 2012, Dyson et al. 2015, Lee et al. 2016, and Kim et al. 2021). Additional details about predicted bacteriophage-host interactions that passed recommended tool-specific thresholds is provided in Supplementary Tables S5, S6, and S8).

[1] log(P(host) = log probability of being a viral host under a Gaussian *k*-mer frequency model
[2] under a Markov random field framework
[3] under a multi-class random forest model

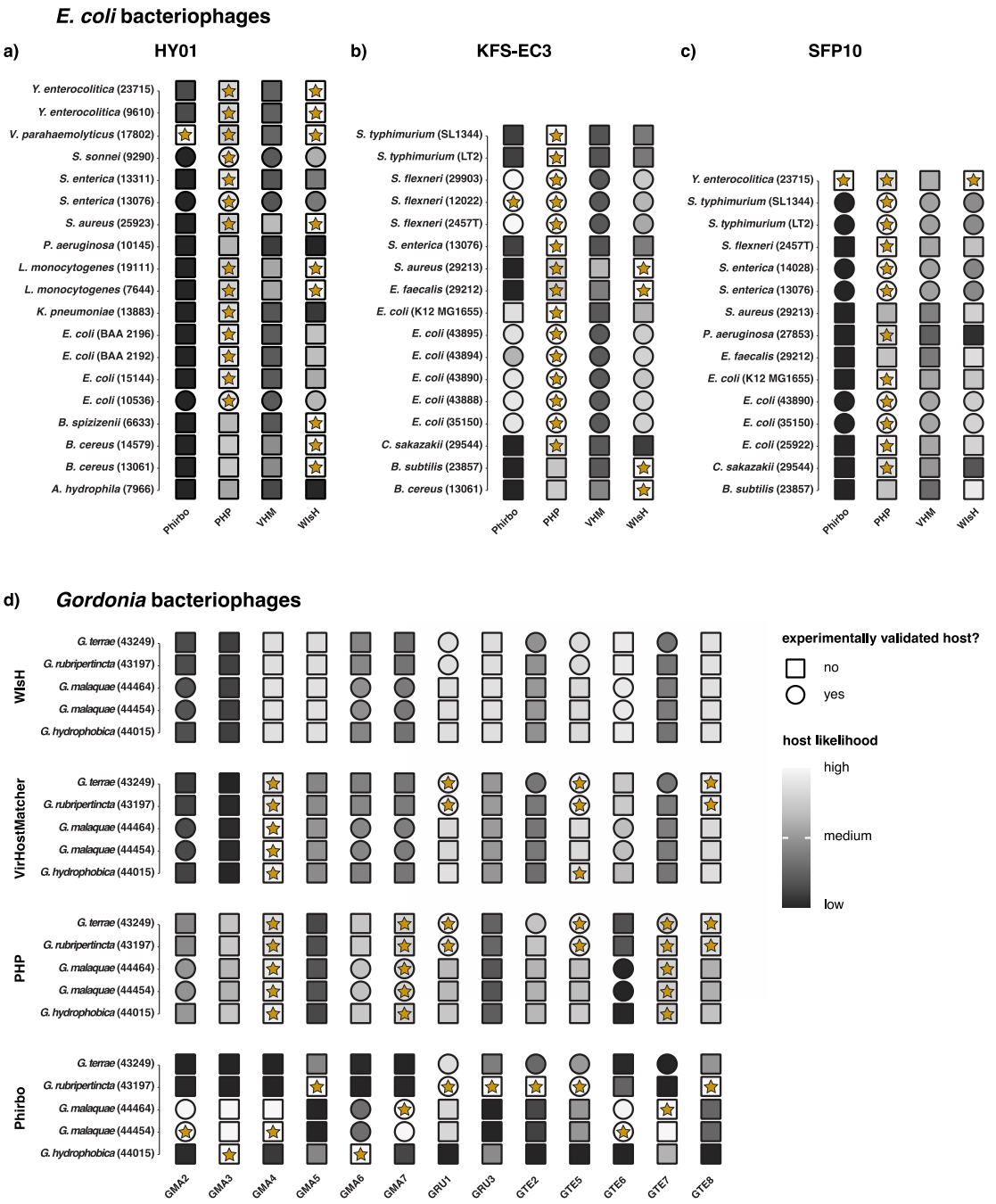


Figure 1. Computational host predictions for three *E. coli* bacteriophages – (a) HY01, (b) KFS-EC3, and (c) SFP10 – and (d) 13 *Gordonia* bacteriophages – GMA2-7, GRU1, GRU3, GTE2, and GTE5-8 – for a set of experimentally validated host and non-host strains (Supplementary Tables S1 and S2) obtained using the confirmatory tools Phirbo, Prokaryotic Host Predictor (PHP), VirHostMatcher (VHM), and WIsH. Predicted bacteriophage-host interactions passing recommended tool-specific thresholds are indicated by a star (for additional details, see Supplementary Table S6).

In contrast to PHP and VirHostMatcher, WIsH requires a null model based on bacteriophage genomes known not to infect the bacterial strain(s) to train a homogeneous Markov model and compute the likelihood (in form of a *p*-value based on the Gaussian null-distribution of the Markov model) for a particular bacteriophage-host pair. However, such data attesting to bacteriophages not being able to infect specific bacterial strains is often not readily available to researchers (i.e., this information is generally not reported in public databases). To test the potential impact of null model construction on predictions, four different null models were tested, including two models consisting of (1) a large, diverse and (2) a small set of bacteriophage genomes expected not to infect any of the tested host strains as well as two model misspecifications consisting of (3) a large, diverse and (4) a small set of bacteriophage genomes containing some known to infect host strains included in this study (for details, see Materials and Methods). Only the null model consisting of a small set of dissimilar bacteriophages (model #2) identified any (all) of the genuine host strains (Supplementary Table S7) – however, this sensitivity came at the expense of the lowest specificity (18.8%) and accuracy (31.6%) out of any tested model. Perhaps counterintuitively, the null model consisting of the much larger set of diverse bacteriophages (model #1) performed amongst the worst in all categories (sensitivity: 0.0%, specificity: 43.8%, precision: 0.0%, and accuracy: 36.8%), likely due to null bacteriophages being more dissimilar to a true negative than a true positive in the dataset, thus biasing the results towards the most dissimilar candidate hosts from among the included null bacteriophages.

The taxonomy-aware BLAST-extension Phirbo ranked in-between these *k*-mer based approaches, with 19.4% sensitivity and 88.2% specificity. As an alignment-based method that relies on sequence homology via a rank-based overlap scoring system of

sequence matches between bacteriophage and bacterial genomes, Phirbo's large number of false negatives likely results from its limited predictive power for bacteriophages that do not share any sequence homology or similarity with their host(s). Specifically, alignment-based methods tend to exhibit a bias towards predicting hosts that carry a genetic mark of a bacteriophage; for example in form of an existing CRISPR spacer or an integrated prophage. However, only ~42% of bacteria encode CRISPR viral defense systems (Makarova et al. 2020) and even fewer will contain spacers for the bacteriophage in question (or a close relative). Furthermore, only two bacteriophages included in this study, GMA5 and GRU3, were temperate; the remaining 14 bacteriophages were obligatorily lytic, thus leaving no genetic trace in the host as they do not integrate into the host genome. Despite this, Phirbo always returned a host prediction, independent of whether a genuine host was included in the provided candidates (e.g., see GMA3 in Figure 1d).

Rather than exploring potential host ranges, the alignment-based tool PHIST only returns a single, highest-scoring host prediction (or, in case of a tie, predictions) based on the number of exact *k*-mer matches between the bacteriophage and the host – a limitation that makes this method less well-suited for broad-spectrum bacteriophages such as the ones tested here. For eight bacteriophages, PHIST predicted one or more hosts (correctly predicted bacteriophage / host pairs: (1) GMA2 / *G. malaquae* 44464, (2) HY01 / *S. flexneri* 12022, (3) KFS-EC3 / *E. coli* 10536, (4) KFS-EC3 / *S. sonnei* 9290; incorrectly predicted bacteriophage/host pairs: (1) GMA4 / *G. malaquae* 44464, (2) GMA5 / *G. malaquae* 44464, (3) GRU3 / *G. malaquae* 44464, (4) GTE6 / *G. hydrophobica* 44015, (5) GTE8 / *G. malaquae* 44454, (6) GTE8 / *G. malaquae* 44464, (7) KFS-EC3 / *E. coli* 15144, (8) KFS-EC3 / *E. coli* BAA-2196, (9) SFP10 / *Y.*

enterocolitica 23715); for the remaining eight bacteriophages (GMA3, GMA6-7, GRU1, GTE2, GTE5-7), PHIST returned no prediction.

The performance of confirmatory host range prediction tools observed in this study is in agreement with earlier work by Edwards and colleagues (2016) who utilized a set of bacteriophages with known isolation hosts to demonstrate that alignment-free methods (such as PHP, VirHostMatcher, and WIsH) exhibit higher recall rates than alignment-based methods (such as Phirbo and PHIST) as their *k*-mer approaches do not rely on the availability of closely-related bacteriophage or host genomes. Overall accuracy in this study ranged from 61.2% (PHP and WIsH) to 69.8% (Phirbo and VirHostMatcher) – similar to the level of accuracy previously observed for these tools (~20%-60% prediction accuracy at the genus-level for alignment-based methods [Edwards et al. 2016; Ahlgren et al. 2017; Zielezinski et al. 2021] and ~30%-70% for alignment-free methods [Ahlgren et al. 2017; Galiez et al. 2017]; and see review of Coclet and Roux 2021). In contrast, the precision of all confirmatory tools was relatively low, ranging from 0% for WIsH (which did not identify any true positives) to 33.3%, 37.5%, and 38.7% for VirHostMatcher, Phirbo, and PHP, respectively (Table 1 and Supplementary Table S5). Thereby, the large number of false positives in the *k*-mer based methods is likely driven by the convergent evolution of oligonucleotide similarity profiles between distantly related bacteriophages and hosts (see Supplementary Figures S1–S3). Notably, most genuine hosts were only identified by a single tool – the machine-learning based PHP trained on a large set of virus-host interactions – with a limited number identified by multiple tools (Figure 2).

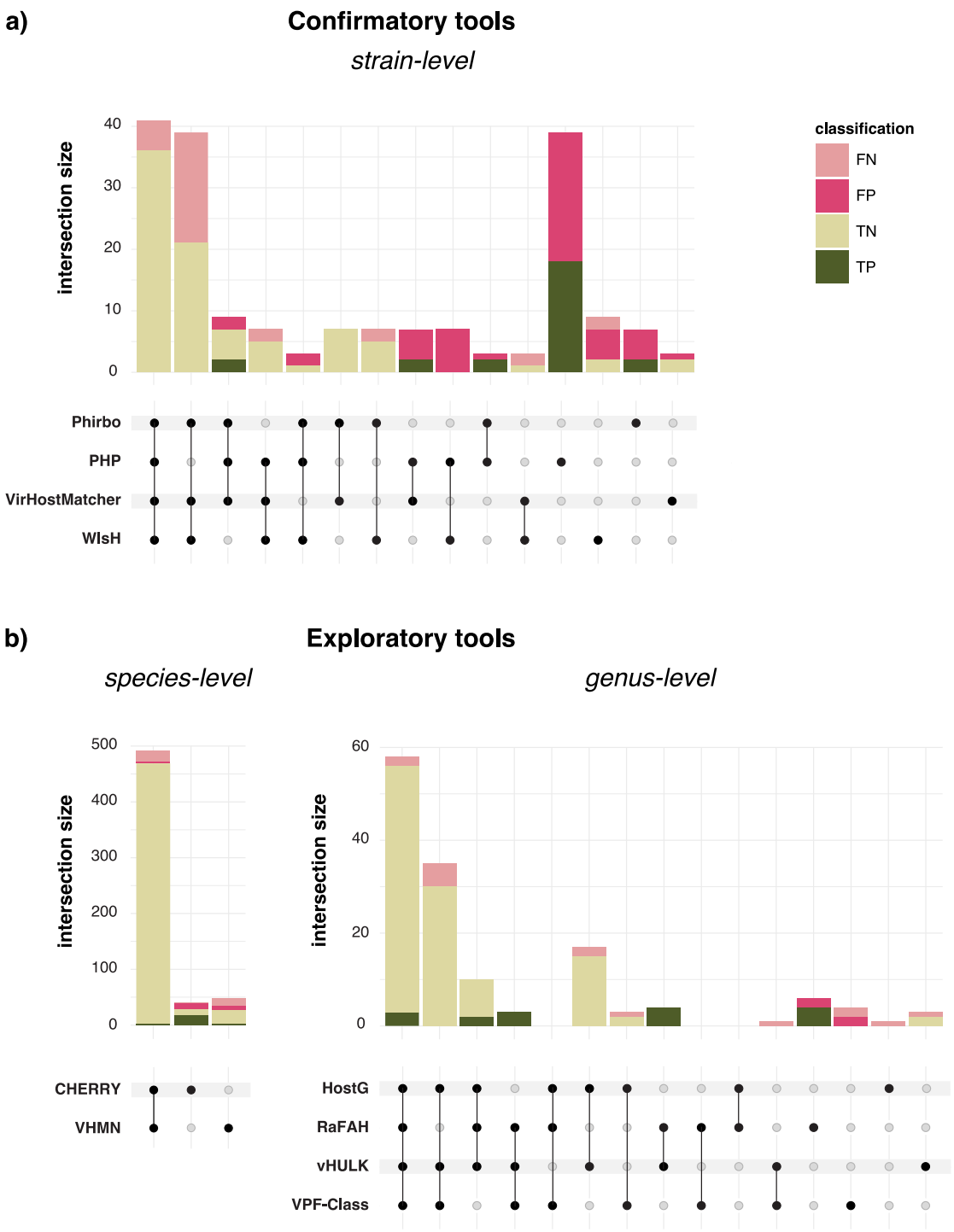

a)
Confirmatory tools
strain-level
intersection size
40
30
20
10
0
classification
FN
FP
TN
TP
Phirbo
PHP
VirHostMatcher
WIsH
b)
Exploratory tools
species-level
genus-level
intersection size
500
400
300
200
100
0
60
40
20
0
CHERRY
VHMN
HostG
RaFAH
vHULK
VPF-Class

Figure 2. Performance of 11 computational host range prediction tools based on experimentally validated bacteriophage-host interactions. (a) The confirmatory tools Phirbo, Prokaryotic virus Host Predictor (PHP), VirHostMatcher, and WIsH utilize a set of provided bacterial genomes to infer the likelihood of strain-specific bacteriophage-host interactions. Exploratory tools predict bacteriophage-host interactions based on an internal database of putative host genomes either at the (b) species-level (CHERRY and VirHostMatcher-Net [VHMN]) or (c) genus-level (HostG, Random Forest Assignment of Hosts [RaFAH], viral Host UnveiLing Kit [vHULK], and VPF-Class). True positives (TP) are shown in green, true negatives (TN) in olive, false positives (FP) in pink, and false negatives (FN) in rose color.

Exploratory Tools

In contrast to confirmatory tools which are generally based on a single type of information (such as exact sequence matches or *k*-mer profiles), the exploratory tools included in this study – CHERRY (Shang and Sun 2022), HostG (Shang and Sun 2021), RaFAH (Coutinho et al. 2021), vHULK (Amgarten et al. 2022), VirHostMatcher-Net (Wang et al. 2020), and VPF-Class (Pons et al. 2021) – utilize multiple bacteriophage-bacteriophage, bacteriophage-host, and/or host-host features to predict interactions based on comparisons of bacteriophage genomes to an internal database of genetic markers of putative host genomes.

Out of the six exploratory tools, two predict hosts at the species-level: (i) CHERRY – a semi-supervised learning model with an underlying multimodal graph that integrates several DNA and protein sequence features (such as information on alignment-based and alignment-free sequence similarity between bacteriophages and bacteria as well as shared protein organization and CRISPR spacers) – and (ii) VirHostMatcher-Net – a network-based support vector machine and random forest framework that integrates both alignment-based information (such as sequence matches between bacteriophage and putative bacterial host genomes or the presence of shared virus-host CRISPR spacers) as well as alignment-free similarity measures (such as WIsH's prediction score and the similarity measure $s_2^* = 1 - 2d_2^*$, where d_2^* is

VirHostMatcher's background-subtracting d_2^* dissimilarity score) with information about virus–host co-abundance across environments to predict bacteriophage-host interactions. Due to its usage of protein sequence information in addition to sequence similarity, CHERRY outperformed VirHostMatcher-Net in terms of specificity (47.6% vs 10.0%), precision (60.6% vs 28.6%), and accuracy (93.6% vs 91.7%) at a similar level of specificity (97.4% vs 98.1%) (Table 1, and see Supplementary Tables S5 and S8).

The remaining four exploratory tools predict hosts at the genus-level: (i) HostG – a semi-supervised learning method based on a graph convolutional network that utilizes information about bacteriophage-host as well as host-host similarities (such as gene sharing and local sequence similarity) to predict the host genus, (ii) RaFAH – a random forest algorithm that classifies bacteriophages according to their putative host genus by comparing protein content in the bacteriophage of interest to protein clusters in a custom-built database of hidden Markov model profiles of other bacteriophages, (iii) vHULK – a deep neutral network that utilizes alignment significance scores between predicted bacteriophage protein sequences and protein families contained within the Prokaryotic Virus Orthologous Group (pVOGs) database (Grazziotin et al. 2017) to infer the host genus, and (iv) VPF-Class – an approach that utilizes predicted protein sequences in the bacteriophage to infer the putative host genus based on a set of previously classified Viral Protein Families (VPFs) from the IMG/VR database (Paez-Espino et al. 2016). At the genus-level, RaFAH exhibited the highest recall (88.9%) and accuracy (95.1%) (Table 1) – higher than the ~60% genus-level accuracy previously reported (see Figure 1 in Coutinho et al. 2021) – correctly predicting *Escherichia* as a host genus for two out of the three *E. coli* bacteriophages and *Gordonia* as a host genus for all 13 *Gordonia* bacteriophages (Figure 3). In comparison, HostG, vHULK, and VPF-Class showed a sensitivity ranging from 31.3% (HostG) to 52.2% (vHULK) and an

accuracy ranging from 87.6% (VPF-Class) – similar to the 86.4% genus-level accuracy reported by the developers (see Table 5 in Pons et al. 2021) – to 91.7% (vHULK). However, RaFAH's sensitivity came at a cost of a slightly worse specificity (RaFAH: 96.9%; VPF-Class: 97.7%; HostG: 100.0%; vHULK: 100.0%). Moreover, both HostG and vHULK were more precise (100% each) than RaFAH (88.9%) and VPF-Class (75.0%). Similar to the confirmatory tools, few genuine hosts were identified by multiple species-level exploratory tools (Figure 2).

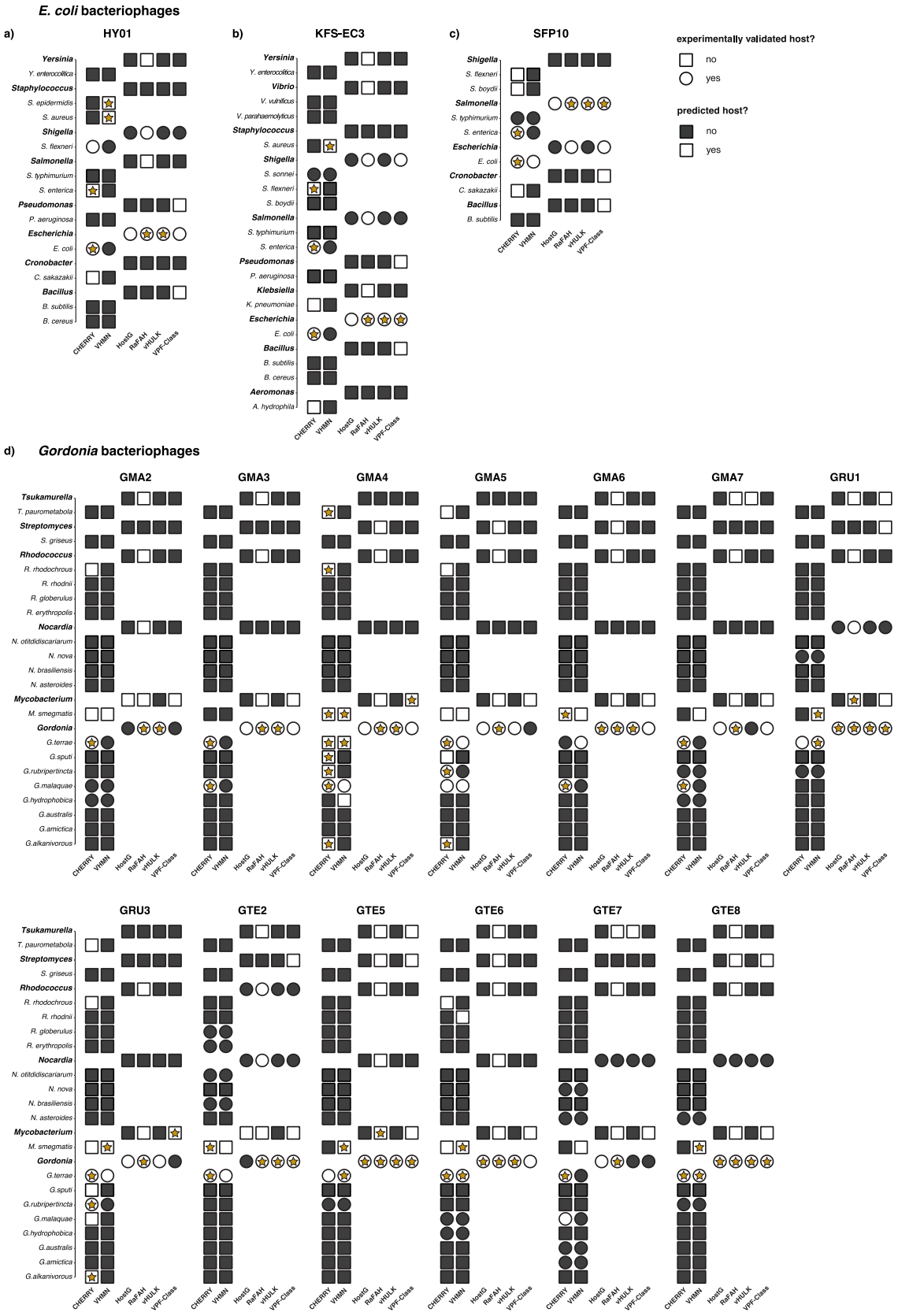

E. coli bacteriophages
a)
HY01
Yersinia
Y. enterocolitica
Staphylococcus
S. epidermidis
S. aureus
Shigella
S. flexneri
Salmonella
S. typhimurium
S. enterica
Pseudomonas
P. aeruginosa
Escherichia
E. coli
Cronobacter
C. sakazakii
Bacillus
B. subtilis
B. cereus
CHERRY
VHMN
HostG
RaFAH
vHULK
VPF-Class
b)
KFS-EC3
Yersinia
Y. enterocolitica
Vibrio
V. vulnificus
V. parahaemolyticus
Staphylococcus
S. aureus
Shigella
S. sonnei
S. flexneri
S. boydii
Salmonella
S. typhimurium
S. enterica
Pseudomonas
P. aeruginosa
Klebsiella
K. pneumoniae
Escherichia
E. coli
Bacillus
B. subtilis
B. cereus
Aeromonas
A. hydrophila
CHERRY
VHMN
HostG
RaFAH
vHULK
VPF-Class
c)
SFP10
Shigella
S. flexneri
S. boydii
Salmonella
S. typhimurium
S. enterica
Escherichia
E. coli
Cronobacter
C. sakazakii
Bacillus
B. subtilis
CHERRY
VHMN
HostG
RaFAH
vHULK
VPF-Class
experimentally validated host?
no
yes
predicted host?
no
yes
d)
Gordonia bacteriophages
GMA2
GMA3
GMA4
GMA5
GMA6
GMA7
GRU1
GRU3
GTE2
GTE5
GTE6
GTE7
GTE8
Tsukamurella
T. paurometabola
Streptomyces
S. griseus
Rhodococcus
R. rhodochrous
R. rhodnii
R. globerulus
R. erythropolis
Nocardia
N. otitidiscariarum
N. nova
N. brasiliensis
N. asteroides
Mycobacterium
M. smegmatis
Gordonia
G.terrae
G.sputi
G.rubripertincta
G.malaquae
G.hydrophobica
G.australis
G.amictica
G.alkanivorous
CHERRY
VHMN
HostG
RaFAH
vHULK
VPF-Class

Figure 3. Computational host predictions for three *E. coli* bacteriophages – (a) HY01, (b) KFS-EC3, and (c) SFP10 – and (d) 13 *Gordonia* bacteriophages – GMA2-7, GRU1, GRU3, GTE2, and GTE5-8 – for a set of experimentally validated host and non-host strains (Supplementary Tables S1 and S2 as well as Tables 1 in Park et al. 2012, Dyson et al. 2015, Lee et al. 2016, and Kim et al. 2021) obtained using the species-level exploratory tools CHERRY and VirHostMatcher-Net [VHMN] as well as the genus-level exploratory tools HostG, Random Forest Assignment of Hosts [RaFAH], viral Host UnveiLing Kit [vHULK], and VPF-Class. Predicted bacteriophage-host interactions passing recommended tool-specific thresholds are indicated by a star (for additional details, see Supplementary Table S8). Experimentally validated non-host strains that were correctly predicted as such by all tools were excluded from this figure.

A general pattern that emerged was that all exploratory tools underpredicted genuine bacteriophage host ranges. For instance, genus-level exploratory tools failed to predict *Shigella* as a host genus for HY01, *Shigella* and *Salmonella* for KFS-EC3, and *Escherichia* for SFP10 (Figure 3), suggesting that *Escherichia* might be the primary host genus for HY01 and KFS-EC3 and *Salmonella* for SFP10. Similarly, *Nocardia* was missed as an additional host genus for the *Gordonia* bacteriophages GRU1, GTE2, GTE7, and GTE8. At the same time, the genus-level predictions of HostG, RaFAH, vHULK, and VPF-Class contained few false positives, with only *Mycobacterium* being mis-predicted as a host genus for the *Gordonia* bacteriophages GMA4 and GRU3 (VPF-Class) as well as GRU1 and GTE 5 (RaFAH). In fact, *Mycobacterium smegmatis* was also frequently mis-predicted as a host for the *Gordonia* bacteriophages at the species-level, likely due to the fact that the *M. smegmatis* genome contains remnants of a prophage originating from the closely-related temperate *Gordonia* bacteriophage Curcubita (Supplementary Figure S4). Such mis-predictions are likely further elevated by dissimilarities between the genomes of the experimentally validated host strains and those available in the tools' pre-built databases (see Supplementary Figures S2 and S3). In general, the performance of machine- or deep-learning based methods depends strongly on the datasets available for training, in particular the information available on bacteriophages with similar sequence features that infect the same bacterial host

species or genuses. Limited knowledge and sparse representation of the full spectrum of the global viral and bacterial diversity remains a major challenge in this regard as many public databases are biased towards well-studied model organisms (though note that metagenomic studies recently started to address this issue; see review of Inglis and Edwards 2022). Relatedly, the robustness of predictions also depend on the accuracy of viral and bacterial genomes as well as the experimental validation of bacteriophage-host interactions reported in the databases (in our study, one out of 22 *Gordonia* and 24 out of 300 *E. coli* database entries were suspended due to misreported information; for an example, see Supplementary Figure S3). Complicating this issue further is the almost entire absence of information about negative bacteriophage-host pairs, preventing the construction of well-balanced training datasets for machine- and deep-learning based methods.

Lastly, although many authors have evaluated their developed methodology against a set of previously published approaches, no genuinely independent benchmark yet exists for exploratory tools and their reported performances are likely an overestimation due to an overfitting caused by the similarity of the test data with the training data (see also the discussion in Coclet and Roux 2021). Moreover, these studies did not include experimentally validated negative bacteriophage-host pairs (true negatives), hampering the reliable assessment of specificity and accuracy. For example, based on a dataset of known virus-host interactions, the developers of HostG reported prediction accuracies between ~35% (for the confirmatory tools WIsH and PHP) and ~60% (for the exploratory tools HostG; RaFAH, vHULK, and VirHostMatcher-Net; see Figure 6 in Shang and Sun 2021). In a follow-up study, the same authors developed CHERRY and demonstrated prediction accuracies ranging from less than 20% (for the alignment-based PHIST) to ~40% (vHULK and VirHostMatcher-Net) to almost 80%

(CHERRY) at the species-level and from ~35%-40% (PHIST, PHP, VPF-Class, and WIsH) to ~60%-70% (HostG, RaFAH, VirHostMatcher-Net, and vHULK) to more than 80% (CHERRY) at the genus-level (see Figure 4B in Shang and Sun 2022). The authors of vHULK self-reported accuracies of 95.2% and 99.1% for *E. coli* and *G. terrae* at the genus-level, with 81.9% and 90.1% sensitivity and 97.1% and 99.8% specificity, respectively (see Table 3 in Amgarten et al. 2022) – much higher than the sensitivity observed in our study (52.2%). In contrast, their reported genus-level accuracies for VirHostMatcher-Net (31.1%) and RaFAH (71.3%) (see Figure 6 in Amgarten et al. 2022) were much lower than those observed here (91.7% and 95.1%, respectively) – a difference that may be caused by the low diversity of taxa investigated.

Conclusion

Gaining a better understanding of bacteriophage host ranges is vitally important to improve their usage as antimicrobial agents. Highly scalable computational host range prediction tools are a valuable supplement to gold standard (but laborious) experimental procedures in this regard. Our benchmarking study of 11 computational host range prediction tools demonstrated that machine- and deep-learning based methods generally outperform more traditional alignment-based and alignment-free methods due to their combined usage of multiple types of information. However, although important to gain a better understanding of the viral ecology in different environments, many of these recently developed approaches are ill-suited for real-world applications (such as phage therapy) as predictions are provided at the species- or genus-level rather than at the strain-level. An additional limitation in adopting these tools is the lack of genomic resources for many bacterial strains of interest (confirmatory tools) as well as the disparity between those strains and the ones included in the tools' internal databases

(exploratory tools) which, given our limited knowledge of viral and bacterial communities in different ecosystems, remain biased towards well-studied, easily culturable model organisms. Moreover, many factors important for successful bacteriophage infection and lysis – such as the recognition of specific host receptors, the ability to overcome bacterial restriction-modification and abortive systems, as well as the compatibility of transcription and translational machinery – remain neglected in computational frameworks. Hence, whenever possible, we recommend incorporating the model sophistication of exploratory tools with the flexibility of strain-specific confirmatory tools in order to aid in the prioritization of experimental efforts to identify the most suitable bacteriophage(s) for any given application.

Data Availability

The data underlying this article are available in ATCC at https://www.atcc.org/ and NCBI at https://www.ncbi.nlm.nih.gov/, and can be accessed under BioProject X (*de novo* assemblies of *Gordonia* strains) and with the accession numbers provided in Supplementary Tables S1 and S2 (bacteriophage and *E. coli* assemblies). Analysis scripts are available at https://github.com/PfeiferLab/host_range_prediction.

Supplementary Material

Supplementary Table S1. Experimentally validated host ranges of three *E. coli* bacteriophages, HY01 (Lee et al. 2016), SFP10 (Park et al. 2012), and KFS-EC3 (Kim et al. 2021). ATCC and NCBI accession numbers are shown in brackets.

	experimentally validated host ranges	
bacteriophage (accession number)	**bacterial strains infected** (accession number)	**bacterial strains not infected** (accession number)
HY01 (KF925357.1)	*Escherichia coli* (ATCC 35150) *Escherichia coli* (ATCC 43888) *Escherichia coli* (ATCC 43890) *Escherichia coli* (ATCC 43894) *Escherichia coli* (ATCC 43895) *Shigella flexneri* (2457T) *Shigella flexneri* (ATCC 12022) *Shigella flexneri* (ATCC 29903)	*Bacilius cereus* (ATCC 13061) *Bacilius subtilis* (ATCC 23857) *Cronobacter sakazakii* (ATCC 29544) *Enterococcus faecalis* (ATCC 29212) *Escherichia coli* (K12MG1655) *Salmonella enterica* (ATCC 13076) *Salmonella typhimurium* (LT2) *Salmonella typhimurium* (SL1344) *Staphylococcus aureus* (ATCC 29213)
KFS-EC3 (MZ065353.1)	*Escherichia coli* (ATCC 10536) *Salmonella enterica* (ATCC 13076) *Shigella sonnei* (ATCC 9290)	*Aeromonas hydrophila* (ATCC 7699) *Bacilius cereus* (ATCC 13061) *Bacilius cereus* (ATCC 14579) *Bacilius spizizenii* (ATCC 6633) *Escherichia coli* (ATCC 15144) *Escherichia coli* (ATCC BAA-2192) *Escherichia coli* (ATCC BAA-2196) *Klebsiella pneumoniae* (ATCC 13883) *Listeria monocytogenes* (ATCC 7644) *Listeria monocytogenes* (ATCC 19111) *Pseudomonas aeruginosa* (ATCC 10145) *Salmonella enterica* (ATCC 13311) *Staphylococcus aureus* (ATCC 25923) *Vibrio parahaemolyticus* (ATCC 17802) *Yersenia enterocolitica* (ATCC 9610) *Yersenia enterocolitica* (ATCC 23715)
SFP10 (HQ259103.1)	*Escherichia coli* (ATCC 35150) *Escherichia coli* (ATCC 43890) *Salmonella enterica* (ATCC 13076) *Salmonella enterica* (ATCC 14028) *Salmonella typhimurium* (LT2) *Salmonella typhimurium* (SL1344)	*Bacilius subtilis* (ATCC 23857) *Cronobacter sakazakii* (ATCC 29544) *Escherichia coli* (ATCC 25922) *Escherichia coli* (K12MG1655) *Enterococcus faecalis* (ATCC 29212) *Pseudomonas aeruginosa* (ATCC 27853) *Shigella flexneri* (2457T) *Staphylococcus aureus* (ATCC 29213) *Yersenia enterocolitica* (ATCC 23715)

Supplementary Table S2. Experimentally validated host ranges of 13 *Gordonia* bacteriophages, GTE2 (Petrovski et al. 2011a), GTE7 (Petrovski et al. 2011b), GTE5 and GRU1 (Petrovski et al. 2012), as well as GMA2–GMA7, GRU3, GTE6, and GTE8 (Dyson et al. 2015). DSMZ and NCBI accession numbers are shown in brackets.

	experimentally validated host ranges	
bacteriophage (accession number)	**bacterial strains infected** (accession number)	**bacterial strains not infected** (accession number)
GMA2 (KR063281.1) **GMA6** (KR063280.1) **GMA7** (KR063278.1) **GTE6** (KR053200.1)	*Gordonia malaquae* (DSM 44454) *Gordonia malaquae* (DSM 44464)	*Gordonia hydrophobica* (DSM 44015) *Gordonia rubripertincta* (DSM 43197) *Gordonia terrae* (DSM 43249)
GMA3 (KR063279.1) **GMA4** (KR063199.1) **GMA5** (KR063198.1) **GRU3** (KR053197.1) **GTE8** (KR053201.1)	–	*Gordonia hydrophobica* (DSM 44015) *Gordonia malaquae* (DSM 44454) *Gordonia malaquae* (DSM 44464) *Gordonia rubripertincta* (DSM 43197) *Gordonia terrae* (DSM 43249)
GRU1 (JF923797.1) **GTE5** (JF923796.1)	*Gordonia rubripertincta* (DSM 43197) *Gordonia terrae* (DSM 43249)	*Gordonia hydrophobica* (DSM 44015) *Gordonia malaquae* (DSM 44454) *Gordonia malaquae* (DSM 44464)
GTE2 (HQ403646.1) **GTE7** (JN035618.1)	*Gordonia terrae* (DSM 43249)	*Gordonia malaquae* (DSM 44454) *Gordonia malaquae* (DSM 44464) *Gordonia hydrophobica* (DSM 44015) *Gordonia rubripertincta* (DSM 43197)

Supplementary Table S3. Estimates of genome size, repeat content, and coverage based on *k*-mer frequencies observed in the long read data. Estimates were obtained using GenomeScope2.0 (Vurture et al. 2017; Ranallo-Benavidez et al. 2020) together with Jellyfish v.2.3.0 (Marçais and Kingsford 2011). DSMZ accession numbers are shown in brackets.

bacterial strain (accession number)	**haploid length (bp)**		**repeat content (bp)**		**coverage**
	minimum	**maximum**	**minimum**	**maximum**	
Gordonia hydrophobica (DSM 44015)	4,419,546	4,428,462	138,580	138,860	313 X
Gordonia malaquae (DSM 44454)	4,244,128	4,251,149	4,485	4,493	164 X
Gordonia malaquae (DSM 44464)	4,339,177	4,348,911	197,518	197,961	161 X
Gordonia rubripertincta (DSM 43197)	4,916,596	4,927,204	71,066	71,219	275 X
Gordonia terrae (DSM 43249)	5,354,377	5,365,402	24,311	24,361	359 X

Supplementary Table S4. Summary statistics of the five *Gordonia de novo* genome assemblies. DSMZ accession numbers are shown in brackets. To assess the completeness of the genome assemblies, BUSCO v.5.4.7 (Manni et al. 2021) was used, together with the actinobacteria database "actinobacteria_class_odb10" (for additional details, see Supplementary Table S4).

bacterial strain (accession number)	**sequence length (bp)**	**GC-content**	**complete BUSCOs[1]**
Gordonia hydrophobica (DSM 44015)	4,632,241	67.45%	353 (99.2%)
Gordonia malaquae (DSM 44454)	4,468,569	66.37%	354 (99.4%)
Gordonia malaquae (DSM 44464)	4,523,876	66.23%	352 (99.2%)
Gordonia rubripertincta (DSM 43197)	5,174,650	67.31%	349 (98.0%)
Gordonia terrae (DSM 43249)	5,701,739	67.81%	353 (99.2%)

[1] based on the *actinobacteria_class_odb10* dataset (containing 356 BUSCOs)

Supplementary Table S5. Performance of computational host range prediction tools. The confirmatory tools Phirbo, Prokaryotic virus Host Predictor (PHP), VirHostMatcher, and WIsH utilize a set of provided bacterial genomes to infer the likelihood of strain-specific bacteriophage-host interactions. Exploratory tools predict bacteriophage-host interactions based on an internal database of putative host genomes either at the species-level (CHERRY and VirHostMatcher-Net [VHMN]) or at the genus-level (HostG, Random Forest Assignment of Hosts [RaFAH], viral Host UnveiLing Kit [vHULK], and VPF-Class). True positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) were determined based on experimentally validated bacteriophage-host interactions (see Supplementary Tables S1 and S2 for details). High confidence results passed the recommended tool-specific thresholds (shown in brackets); low confidence results were below the recommended threshold.

		tool (threshold)	**TP (high / low)**	**FP (high / low)**	**TN**	**FN**	**unvalidated predictions**
confirmatory	*strain-level*	Phirbo (highest rank-based overlap)	6	10	75	25	–
		PHP (log(P(host))[1]: 1442)	24	38	47	7	–
		VirHostMatcher (distance / dissimilarity: 0.175)	4	8	77	27	–
		WIsH (p-value < 0.06)	0	14	71	31	–
exploratory	*species-level*	CHERRY (P(graph convolutional encoder): 0.9)	20 / 5	13 / 19	491	22	33
		VHMN (prediction score[2]: 0.95)	4 / 7	10 / 8	505	36	139
	genus-level	HostG (SoftMax value: 0.94)	5 / 9	0 / 2	109	11	0
		RaFAH (prediction score[3]: 0.14)	16 / 7	2 / 47	62	2	90
		vHULK (alignment significance score: 0.8)	12 / 2	0 / 2	109	11	0
		VPF-Class (membership: 0.3, confidence: 0.5)	6 / 8	2 / 23	86	11	121

[1] log(P(host) = log probability of being a viral host under a Gaussian *k*-mer frequency model

[2] under a Markov random field framework

[3] under a multi-class random forest model

Supplementary Table S6. Bacteriophage-host interactions predicted by confirmatory tools. Predicted bacteriophage-host interactions that passed recommended confirmatory tool-specific thresholds (see Supplementary Table S5 for details). True positives (TP) and false positives (FP) were determined based on experimentally validated bacteriophage-host interactions (see Supplementary Tables S1 and S2 for details).

group	bacteriophage	tool	predicted host	prediction score (*p*-value[1])	category
E. coli	**HY01**	Phirbo	*S. flexneri* 12022	0.347545549	TP
		PHP	*E. coli* 43894	1458.509432	TP
			E. coli 43888	1458.402294	TP
			E. coli 43890	1458.400863	TP
			E. coli 35150	1458.360591	TP
			E. coli 43895	1458.333301	TP
			E. coli K12MG1655	1458.130161	FP
			S. flexneri 12022	1458.121672	TP
			S. flexneri 29903	1458.121332	TP
			S. flexneri 2457T	1458.080758	TP
			S. typhimurium SL1344	1457.638538	FP
			S. typhimurium LT2	1457.379504	FP
			S. enterica 13076	1457.245979	FP
			C. sakazakii 29544	1451.410058	FP
			S. aureus 29213	1445.921933	FP
			E. faecalis 29212	1443.424835	FP
		WIsH	*B. subtilis* 23857	-1.36966 (0.00449395)	FP
			S. aureus 29213	-1.34754 (0.01361500)	FP
			E. faecalis 29212	-1.35551 (0.01410700)	FP
			B. cereus 13061	-1.34737 (0.01756230)	FP
	KFS-EC3	Phirbo	*V. parahaemolyticus* 17802	3.97E-05	FP
		PHP	*S. sonnei* 9290	1452.045298	TP
			E. coli BAA-2196	1451.979889	FP
			E. coli BAA-2192	1451.776139	FP
			E. coli 15144	1451.678476	FP
			E. coli 10536	1451.657987	TP
			S. enterica 13311	1451.498755	FP
			S. enterica 13076	1451.490817	TP
			K. pneumoniae 13883	1447.668585	FP
			Y. enterocolitica 23715	1445.688084	FP

			Y. enterocolitica 9610	1445.546051	FP
			L. monocytogenes 7644	1444.480204	FP
			L. monocytogenes 19111	1443.383630	FP
			S. aureus 25923	1442.291869	FP
			V. parahaemolyticus 17802	1442.206717	FP
		WIsH	*B. spizizenii* 6633	-1.37294 (0.00351026)	FP
			Y. enterocolitica 23715	-1.38138 (0.00645586)	FP
			Y. enterocolitica 9610	-1.38178 (0.00697053)	FP
			L. monocytogenes 7644	-1.35207 (0.01111870)	FP
			L. monocytogenes 19111	-1.35378 (0.01143930)	FP
			S. aureus 25923	-1.34755 (0.01351530)	FP
			B. cereus 13061	-1.34890 (0.01887000)	FP
			B. cereus 14579	-1.34830 (0.01889940)	FP
			V. parahaemolyticus 17802	-1.37788 (0.02473530)	FP
	SFP10	Phirbo	*Y. enterocolitica* 23715	0.011554849	FP
		PHP	*S. typhimurium* SL1344	1456.655394	TP
			S. typhimurium LT2	1456.468581	TP
			S. enterica 13076	1456.373383	TP
			S. enterica 14028	1456.150993	TP
			E. coli K12MG1655	1455.288787	FP
			E. coli 25922	1455.209664	FP
			E. coli 35150	1454.988970	TP
			S. flexneri 2457T	1454.951719	FP
			E. coli 43890	1454.782297	TP
			Y. enterocolitica 23715	1449.898739	FP
			C. sakazakii 29544	1449.721021	FP
			P. aeruginosa 27853	1444.714408	FP
		WIsH	*Y. enterocolitica* 23715	-1.39075 (0.03924090)	FP
Gordonia	**GMA2**	Phirbo	*G. malaquae* 44454	0.549356099	TP
	GMA3	Phirbo	*G. hydrophobica* 44015	0.001598421	FP
	GMA4	Phirbo	*G. malaquae* 44454	0.675331453	FP
		PHP	*G. hydrophobica* 44015	1454.488830	FP
			G. malaquae 44464	1453.730246	FP
			G. malaquae 44454	1453.546489	FP
			G. rubripertincta 43197	1449.411706	FP
			G. terrae 43249	1448.018764	FP

		VHM	*G. malaquae* 44464	0.145640	FP
			G. malaquae 44454	0.147630	FP
			G. hydrophobica 44015	0.156868	FP
			G. rubripertincta 43197	0.163904	FP
			G. terrae 43249	0.172697	FP
	GMA5	Phirbo	*G. rubripertincta* 43197	0.218286850	FP
	GMA6	Phirbo	*G. hydrophobica* 44015	0.057686227	FP
	GMA7	Phirbo	*G. malaquae* 44464	0.028573288	TP
		PHP	*G. malaquae* 44464	1445.884769	TP
			G. rubripertincta 43197	1445.396714	FP
			G. malaquae 44454	1445.169766	TP
			G. terrae 43249	1444.389261	FP
			G. hydrophobica 44015	1443.481661	FP
	GRU1	Phirbo	*G. rubripertincta* 43197	0.148506734	TP
		PHP	*G. terrae* 43249	1452.958304	TP
			G. rubripertincta 43197	1451.781375	TP
		VHM	*G. rubripertincta* 43197	0.161947	TP
			G. terrae 43249	0.163551	TP
	GRU3	Phirbo	*G. rubripertincta* 43197	0.216302610	FP
	GTE2	Phirbo	*G. rubripertincta* 43197	0.059869213	FP
	GTE5	Phirbo	*G. rubripertincta* 43197	0.233969180	TP
		PHP	*G. terrae* 43249	1453.441435	TP
			G. rubripertincta 43197	1452.366588	TP
		VHM	*G. rubripertincta* 43197	0.158620	TP
			G. terrae 43249	0.160427	TP
			G. hydrophobica 44015	0.174834	FP
	GTE6	Phirbo	*G. malaquae* 44454	0.034698823	TP
	GTE7	Phirbo	*G. malaquae* 44464	0.004342390	FP
		PHP	*G. rubripertincta* 43197	1445.190106	FP
			G. malaquae 44464	1444.552530	FP
			G. terrae 43249	1444.205542	TP
			G. malaquae 44454	1443.803905	FP
			G. hydrophobica 44015	1442.369440	FP

	GTE8	Phirbo	*G. rubripertincta* 43197	0.248009283	FP
		PHP	*G. terrae* 43249	1449.772436	FP
			G. rubripertincta 43197	1447.896144	FP
		VHM	*G. rubripertincta* 43197	0.169169	FP
			G. terrae 43249	0.170116	FP

[1] only reported by WIsH

Supplementary Table S7. The impact of WIsH null model choice on bacteriophage KFS-EC3 host predictions. Null models #1 and #2 consist of bacteriophage genomes expected not to infect any of the tested host strains. Model misspecifications (null models #3 and #4) were tested by including *Escherichia coli* bacteriophage genomes known to infect host strains included in this study. True positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) were determined based on experimentally validated bacteriophage-host interactions (see Supplementary Table S1 for details).

null model	bacteriophages	TP	FP	TN	FN	sensitivity	specificity	precision	accuracy
#1	*Alteromonas Cellulophage Cyanophage Lactobacillus Mycobacterium Oenococcus Pelagibacter Prochlorococcus Rhizobium Synechococcus Thermus*	0	9	7	3	0.0%	43.8%	0.0%	36.8%
#2	*Synechococcus*	3	13	3	0	**100.0%**	18.8%	**18.8%**	31.6%
#3	*Alteromonas Cellulophage Cyanophage Escherichia coli* Lactobacillus Mycobacterium Oenococcus Pelagibacter Prochlorococcus Rhizobium Synechococcus Thermus*	0	6	10	3	0.0%	**62.5%**	0.0%	**52.6%**
#4	*Escherichia coli**	0	7	9	3	0.0%	56.3%	0.0%	47.4%

Supplementary Table S8. Bacteriophage-host interactions predicted by exploratory tools. Exploratory tools predict bacteriophage-host interactions based on an internal database of putative host genomes either at the species-level (CHERRY and VirHostMatcher-Net [VHMN]) or genus-level (HostG, Random Forest Assignment of Hosts [RaFAH], viral Host UnveiLing Kit [vHULK], and VPF-Class). Predictions were limited to the top 10 results per tool, with bacteriophage-host interactions that passed recommended exploratory tool-specific confidence thresholds shown with an asterisk (see Supplementary Table S5 for details). True positives (TP) and false positives (FP) were determined based on experimentally validated bacteriophage-host interactions (for details, see Supplementary Tables S1 and S2 as well as Tables 1 in Park et al. 2012, Dyson et al. 2015, Lee et al. 2016, and Kim et al. 2021); predictions for which no experimentally data was available are shown in gray.

group	bacteriophage	tool	predicted host	prediction score (membership ratio[1])	category
E. coli	**HY01**	CHERRY	*Escherichia coli*	1.00*	TP
			Salmonella enterica	0.91*	FP
			Shigella flexneri	0.87	TP
			Shigella boydii	0.51	–
			Aeromonas salmonicida	0.32	–
			Edwardsiella ictaluri	0.25	–
			Citrobacter rodentium	0.17	–
			Cronobacter sakazakii	0.13	FP
			Klebsiella oxytoca	0.02	–
			Enterobacter cloacae	0.01	–
		VHMN	*Staphylococcus aureus*	0.9774*	FP
			Lactococcus lactis subsp. lactis	0.9737*	–
			Staphylococcus epidermidis	0.9725*	–
			Clostridium tetani	0.9687*	–
			Lactobacillus sp.	0.9679*	–
			Clostridium tetani	0.9665*	–
			Clostridium tetani	0.9663*	–
			Lactococcus lactis subsp. cremoris	0.9651*	–
			Megamonas rupellensis	0.9628*	–
			Megamonas rupellensis	0.9622*	–
		HostG	*Escherichia*	0.5282155	TP
		RaFAH	*Escherichia*	0.823*	TP
			Shigella	0.100	TP
			Yersinia	0.026	–
			Citrobacter	0.009	–
			Serratia	0.006	–
			Edwardsiella	0.005	–
			Enterobacter	0.005	–
			Klebsiella	0.004	–
			Salmonella	0.003	FP
			Stenotrophomonas	0.003	–
		VPF-Class	*Mycobacterium*	0.8306640672 (2.35E-01)	–
			Escherichia	0.8306640672 (2.23E-01)	TP
			Ralstonia	0.8306640672 (1.19E-01)	–

			Bacillus	0.8306640672 (1.03E-01)	FP
			Oenococcus	0.8306640672 (9.37E-02)	–
			Pseudomonas	0.8306640672 (2.60E-02)	FP
			Mannheimia	0.8306640672 (2.43E-02)	–
			Streptococcus	0.8306640672 (1.68E-02)	–
			Acinetobacter	0.8306640672 (1.49E-02)	–
			Streptomyces	0.8306640672 (1.47E-02)	–
		vHULK	*Escherichia*	0.9665705*	TP
	KFS-EC3	CHERRY	*Escherichia coli*	1.00*	TP
			Aeromonas salmonicida	1.00*	–
			Edwardsiella ictaluri	0.96*	–
			Shigella flexneri	0.94*	FP
			Salmonella enterica	0.91*	TP
			Klebsiella pneumoniae	0.89	FP
			Cronobacter sakazakii	0.76	–
			Aeromonas hydrophila	0.61	FP
			Acinetobacter baumannii	0.46	–
			Klebsiella oxytoca	0.39	–
		VHMN	*Staphylococcus aureus*	0.9823*	FP
			Clostridium tetani	0.9783*	–
			Clostridium tetani	0.9773*	–
			Clostridium tetani	0.9768*	–
			Staphylococcus epidermidis	0.9760*	–
			Lactococcus lactis subsp. lactis	0.9744*	–
			Lactobacillus sp.	0.9688*	–
			Megamonas rupellensis	0.9679*	–
			Lactococcus lactis subsp. cremoris	0.9679*	–
			Megamonas rupellensis	0.9671*	–
		HostG	*Escherichia*	0.5276613	TP
		RaFAH	*Escherichia*	0.766*	TP
			Shigella	0.118	TP
			Yersinia	0.035	FP

			Salmonella	0.016	TP
			Citrobacter	0.009	–
			Acinetobacter	0.008	–
			Klebsiella	0.008	FP
			Serratia	0.005	–
			Vibrio	0.005	FP
			Stenotrophomonas	0.004	–
		VPF-Class	*Escherichia*	0.9416975882 (5.84E-01)*	TP
			Mycobacterium	0.9416975882 (1.17E-01)	–
			Bacillus	0.9416975882 (1.07E-01)	FP
			Oenococcus	0.9416975882 (4.87E-02)	–
			Ralstonia	0.9416975882 (2.55E-02)	–
			Shigella	0.9416975882 (2.28E-02)	TP
			Rhodothermus	0.9416975882 (1.66E-02)	–
			Pseudomonas	0.9416975882 (9.42E-03)	FP
			Streptomyces	0.9416975882 (7.30E-03)	–
			Bombyx	0.9416975882 (7.26E-03)	–
		vHULK	*Escherichia*	0.92034554*	TP
	SFP10	CHERRY	*Salmonella enterica*	0.99*	TP
			Escherichia coli	0.99*	TP
			Cronobacter sakazakii	0.54	FP
			Shigella flexneri	0.13	FP
			Burkholderia cenocepacia	0.12	–
			Aeromonas media	0.07	–
			Burkholderia thailandensis	0.07	–
			Aggregatibacter actinomycetemcomitans	0.06	–
			Aeromonas hydrophila	0.06	–
			Shigella boydii	0.05	FP
		VHMN	*Pectobacterium atrosepticum*	0.9409	–
			Pectobacterium versatile	0.9320	–
			Pectobacterium atrosepticum	0.9218	–
			Escherichia coli	0.8719	–

			Escherichia coli	0.8644	–
			Escherichia coli	0.8520	–
			Escherichia coli	0.8493	TP
			Serratia sp.	0.8448	–
			Escherichia coli	0.8411	–
			Shigella sonnei	0.8347	TP
		HostG	*Salmonella*	0.50561905	TP
		RaFAH	*Salmonella*	0.991*	TP
			Escherichia	0.008	TP
			Serratia	0.001	–
		VPF-Class	*Salmonella*	0.992381477 (5.92E-01)*	TP
			Escherichia	0.992381477 (1.18E-01)	TP
			Bacillus	0.992381477 (5.44E-02)	FP
			Cellulophaga	0.992381477 (4.52E-02)	–
			Cronobacter	0.992381477 (2.96E-02)	FP
			Ralstonia	0.992381477 (2.17E-02)	–
			Synechococcus	0.992381477 (2.05E-02)	–
			Streptococcus	0.992381477 (1.40E-02)	–
			Sulfolobus	0.992381477 (9.52E-03)	–
			Mycobacterium	0.992381477 (8.38E-03)	–
		vHULK	*Salmonella*	0.93978465*	TP
Gordonia	**GMA2**	CHERRY	*Gordonia terrae*	0.97*	TP
			Rhodococcus hoagii	0.04	–
			Mycolicibacterium smegmatis	0.02	FP
			Rhodococcus rhodochrous	0.01	FP
		VHMN	*Spiribacter salinus*	0.3583	–
			Acidithiobacillus caldus	0.3287	–
			Cutibacterium acnes	0.3102	–
			Cutibacterium acnes	0.3058	–
			Micrococcales bacterium	0.300	–
			Demequina sediminicola	0.2954	–
			Gamma proteobacterium	0.2944	–

			Halomonas utahensis	0.2929	–
			Mycolicibacterium smegmatis	0.2908	–
			Halovibrio sp.	0.2904	–
		HostG	*Mycolicibacterium*	0.78684735	FP
		RaFAH	*Gordonia*	0.740*	TP
			Mycolicibacterium	0.065	FP
			Rhodococcus	0.056	FP
			Mycobacterium	0.041	FP
			Tsukamurella	0.023	FP
			Corynebacterium	0.011	–
			Nocardia	0.010	FP
			Escherichia	0.006	–
			Salmonella	0.006	–
			Rhodopseudomonas	0.005	–
		VPF-Class	*Bacillus*	0.9567264423 (1.54E-01)	–
			Mycobacterium	0.9567264423 (1.44E-01)	FP
			Vibrio	0.9567264423 (1.44E-01)	–
			Nitrincola	0.9567264423 (7.61E-02)	–
			Pantoea	0.9567264423 (3.94E-02)	–
			Staphylococcus	0.9567264423 (3.44E-02)	–
			Pseudomonas	0.9567264423 (3.38E-02)	–
			Lactobacillus	0.9567264423 (3.28E-02)	–
			Lactococcus	0.9567264423 (3.24E-02)	–
			Clostridium	0.9567264423 (2.13E-02)	–
		vHULK	*Gordonia*	0.996979*	TP
	GMA3	CHERRY	*Gordonia malaquae*	1.00*	TP
			Gordonia terrae	0.98*	TP
			Rhodococcus hoagii	0.86	–
			Mycolicibacterium phlei	0.09	–
		VHMN	*Agrobacterium sp.*	0.4966	–
			Rhizobium sp.	0.4627	–
			Agrobacterium fabrum	0.4300	–

			Brucella inopinata	0.4142	–
			Sodalis glossinidius str. morsitans	0.3576	–
			Pseudomonas syringae pv. actinidiae	0.3560	–
			Brucella abortus	0.3498	–
			Pseudomonas syringae pv. avii	0.3465	–
			Nitrosospira sp.	0.3446	–
			Cronobacter sakazakii	0.3400	–
		HostG	*Gordonia*	0.6929956	TP
		RaFAH	*Gordonia*	0.892*	TP
			Mycolicibacterium	0.032	TP
			Rhodococcus	0.014	FP
			Escherichia	0.012	–
			Mycobacterium	0.008	FP
			Tsukamurella	0.006	FP
			Rhodopseudomonas	0.004	–
			Lactobacillus	0.003	–
			Salmonella	0.003	–
			Candidatus Ruthia	0.002	–
			Dorea	0.002	–
			Faecalibacterium	0.002	–
			Yersinia	0.002	–
		VPF-Class	*Bacillus*	0.9451088996 (1.42E-01)	–
			Gordonia	0.9451088996 (9.83E-02)	TP
			Mycobacterium	0.9451088996 (8.87E-02)	FP
			Pseudomonas	0.9451088996 (5.98E-02)	–
			Aeromonas	0.9451088996 (5.62E-02)	–
			Acinetobacter	0.9451088996 (5.38E-02)	–
			Lactococcus	0.9451088996 (4.71E-02)	–
			Streptococcus	0.9451088996 (3.80E-02)	–
			Clostridium	0.9451088996 (3.67E-02)	–
			Staphylococcus	0.9451088996 (3.54E-02)	–

		vHULK	*Gordonia*	0.9991523*	TP
	GMA4	CHERRY	*Gordonia terrae*	1.00*	FP
			Rhodococcus hoagii	1.00*	–
			Gordonia rubripertincta	0.99*	FP
			Gordonia alkanivorans	0.99*	FP
			Gordonia malaquae	0.98*	TP
			Mycolicibacterium smegmatis	0.97*	FP
			Gordonia sputi	0.96*	FP
			Rhodococcus rhodochrous	0.96*	FP
			Gordonia neofelifaecis	0.94*	–
			Tsukamurella paurometabola	0.94*	FP
		VHMN	*Mycolicibacterium smegmatis*	0.9963*	–
			Gordonia terrae	0.9539*	FP
			Gordonia terrae	0.9394	FP
			Mycolicibacterium smegmatis	0.9369	–
			Gordonia malaquae	0.8464	TP
			Gordonia malaquae	0.8416	–
			Gordonia shandongensis	0.7964	–
			Gordonia phthalatica	0.7957	–
			Gordonia westfalica	0.7927	–
			Gordonia hydrophobica	0.7801	–
		HostG	*Gordonia*	0.49579906	TP
		RaFAH	*Gordonia*	0.570*	TP
			Corynebacterium	0.117	–
			Mycobacterium	0.053	FP
			Streptomyces	0.041	FP
			Rhodococcus	0.033	FP
			Actinomyces	0.021	–
			Cutibacterium	0.019	–
			Bifidobacterium	0.017	–
			Rothia	0.014	–
			Thermomonospora	0.014	–

		VPF-Class	*Mycobacterium*	0.825986255 (3.64E-01)*	FP
			Homo	0.825986255 (1.43E-01)	–
			Gordonia	0.825986255 (8.88E-02)	TP
			Aeromonas	0.825986255 (8.38E-02)	–
			Bacillus	0.825986255 (6.49E-02)	–
			Clostridium	0.825986255 (6.32E-02)	–
			Pseudomonas	0.825986255 (3.83E-02)	–
			Flavobacterium	0.825986255 (2.27E-02)	–
			Synechococcus	0.825986255 (2.06E-02)	–
			Nitrincola	0.825986255 (1.52E-02)	–
		vHULK	*Gordonia*	0.9263134*	TP
	GMA5	CHERRY	*Gordonia terrae*	1.00*	TP
			Gordonia neofelifaecis	0.98*	–
			Gordonia rubripertincta	0.96*	TP
			Gordonia alkanivorans	0.94*	FP
			Rhodococcus hoagii	0.93*	–
			Gordonia malaquae	0.84	TP
			Mycolicibacterium smegmatis	0.83	FP
			Tsukamurella paurometabola	0.82	FP
			Gordonia sputi	0.66	FP
			Rhodococcus rhodochrous	0.52	FP
		VHMN	*Mycolicibacterium smegmatis*	0.9091	–
			Gordonia terrae	0.6024	TP
			Gordonia terrae	0.5754	TP
			Streptomyces coelicolor	0.4854	–
			Mycolicibacterium smegmatis	0.4697	–
			Microbacterium foliorum	0.4258	–
			Streptomyces venezuelae	0.4094	–
			Pseudomonas aeruginosa	0.3758	–
			Glycomyces paridis	0.3675	–
			Gordonia malaquae	0.3601	–
		HostG	*Gordonia*	0.33900467	TP

		RaFAH	*Gordonia*	0.686*	TP
			Rhodococcus	0.066	FP
			Mycolicibacterium	0.059	FP
			Mycobacterium	0.034	FP
			Microbacterium	0.033	–
			Pseudopropionibacterium	0.010	–
			Bifidobacterium	0.009	–
			Arthrobacter	0.008	–
			Stigmatella	0.007	–
			Faecalibacterium	0.004	–
			Streptomyces	0.004	FP
		VPF-Class	*Mycobacterium*	0.7123720216 (2.93E-01)	FP
			Bacillus	0.7123720216 (1.75E-01)	–
			Mus	0.7123720216 (1.33E-01)	–
			Achromobacter	0.7123720216 (5.63E-02)	–
			Pseudomonas	0.7123720216 (4.93E-02)	–
			Cellulophaga	0.7123720216 (4.13E-02)	–
			Polaribacter	0.7123720216 (2.40E-02)	–
			Salmonella	0.7123720216 (2.30E-02)	–
			Burkholderia	0.7123720216 (1.59E-02)	–
			Riemerella	0.7123720216 (1.32E-02)	–
		vHULK	*Gordonia*	0.72836643	TP
	GMA6	CHERRY	*Gordonia malaquae*	1.00*	TP
			Mycolicibacterium smegmatis	1.00*	FP
		VHMN	*Mycolicibacterium smegmatis*	0.5131	–
			Rhizobium leguminosarum bv. viciae	0.2497	–
			Rhizobium sp.	0.2149	–
			Agrobacterium sp.	0.2071	–
			Gordonia terrae	0.2048	TP
			Azospirillum brasilense	0.2007	–
			Streptomyces coelicolor	0.1971	–
			Pseudomonas aeruginosa	0.1833	–

			Prochlorococcus marinus str.	0.1800	–
			Spongiibacter tropicus	0.1774	–
		HostG	*Gordonia*	1.0000000*	TP
		RaFAH	*Gordonia*	0.640*	TP
			Mycolicibacterium	0.138	FP
			Rhodococcus	0.096	FP
			Mycobacterium	0.061	FP
			Tsukamurella	0.023	FP
			Actinomyces	0.004	–
			Corynebacterium	0.004	–
			Haemophilus	0.003	–
			Blautia	0.002	–
			Porphyrobacter	0.002	–
			Pseudopropionibacterium	0.002	–
			Ruminococcus	0.002	–
			Streptococcus	0.002	–
			Streptomyces	0.002	FP
			Yersinia	0.002	–
		VPF-Class	*Bacillus*	0.9245044056 (1.16E-01)	–
			Mycobacterium	0.9245044056 (1.07E-01)	FP
			Clostridium	0.9245044056 (8.17E-02)	–
			Ralstonia	0.9245044056 (6.88E-02)	–
			Escherichia	0.9245044056 (6.73E-02)	–
			Cronobacter	0.9245044056 (6.21E-02)	–
			Gordonia	0.9245044056 (5.36E-02)	TP
			Corynebacterium	0.9245044056 (4.26E-02)	–
			Vibrio	0.9245044056 (3.90E-02)	–
			Pseudomonas	0.9245044056 (3.60E-02)	–
		vHULK	*Gordonia*	0.99143773*	TP
	GMA7	CHERRY	*Gordonia malaquae*	1.00*	TP
			Gordonia terrae	0.98*	TP
			Rhodococcus hoagii	0.86	–

			Mycolicibacterium phlei	0.09	–
		VHMN	*Mycolicibacterium smegmatis*	0.6141	–
			Olsenella umbonata	0.5738	–
			Olsenella sp.	0.5668	–
			Olsenella umbonata	0.5660	–
			Streptomyces coelicolor	0.4864	–
			Olsenella sp.	0.4854	–
			Olsenella sp.	0.4755	–
			Olsenella sp.	0.4754	–
			Bacterium	0.4745	–
			Bacterium	0.4745	–
		HostG	*Gordonia*	0.6929956	TP
		RaFAH	*Gordonia*	0.816*	TP
			Tsukamurella	0.076	FP
			Mycolicibacterium	0.061	FP
			Mycobacterium	0.030	FP
			Rhodococcus	0.004	FP
			Corynebacterium	0.002	–
			Cutibacterium	0.002	–
			Bacillus	0.001	–
			Blautia	0.001	–
			Clostridium	0.001	–
			Coprococcus	0.001	–
			Frankia	0.001	–
			Lactobacillus	0.001	–
			Prevotella	0.001	–
			Pseudopropionibacterium	0.001	–
			Rothia	0.001	–
		VPF-Class	*Mycobacterium*	0.93894772 (2.46E-01)	FP
			Gordonia	0.93894772 (1.53E-01)	TP
			Cellulophaga	0.93894772 (1.36E-01)	–
			Synechococcus	0.93894772 (6.63E-02)	–

			Bacillus	0.93894772 (6.22E-02)	–
			Pseudomonas	0.93894772 (5.27E-02)	–
			Clostridium	0.93894772 (2.60E-02)	–
			Sinorhizobium	0.93894772 (2.49E-02)	–
			Apis	0.93894772 (2.06E-02)	–
			Prochlorococcus	0.93894772 (1.96E-02)	–
		vHULK	*Tsukamurella*	0.5079881	FP
	GRU1	CHERRY	*Gordonia terrae*	0.22	TP
			Rhodococcus hoagii	0.01	–
		VHMN	*Mycolicibacterium smegmatis*	0.9990*	–
			Mycolicibacterium smegmatis	0.9806*	–
			Gordonia terrae	0.9674*	TP
			Gordonia terrae	0.9669*	TP
			Streptomyces coelicolor	0.9438	–
			Pseudomonas aeruginosa	0.8705	–
			Pseudomonas aeruginosa	0.8623	–
			Mycobacterium sp.	0.8521	–
			Pseudomonas aeruginosa	0.8508	–
			Mycobacterium sp.	0.8504	–
		HostG	*Gordonia*	1.0000000*	TP
		RaFAH	*Gordonia*	0.721*	TP
			Mycolicibacterium	0.197*	FP
			Mycobacterium	0.052	FP
			Corynebacterium	0.013	–
			Tsukamurella	0.013	FP
			Rhodococcus	0.003	FP
			Nocardia	0.001	TP
		VPF-Class	*Gordonia*	0.9993212076 (8.12E-01)*	TP
			Mycobacterium	0.9993212076 (5.23E-02)	FP
			Tsukamurella	0.9993212076 (3.85E-02)	FP
			Bacillus	0.9993212076 (1.86E-02)	–
			Pseudomonas	0.9993212076 (1.18E-02)	–

			Streptomyces	0.9993212076 (8.61E-03)	FP
			Synechococcus	0.9993212076 (7.56E-03)	–
			Vibrio	0.9993212076 (6.95E-03)	–
			Cellulophaga	0.9993212076 (5.00E-03)	–
			Corynebacterium	0.9993212076 (4.21E-03)	–
		vHULK	*Gordonia*	0.9989813*	TP
	GRU3	CHERRY	*Gordonia rubripertincta*	1.00*	TP
			Gordonia terrae	1.00*	TP
			Gordonia neofelifaecis	0.98*	–
			Gordonia alkanivorans	0.94*	FP
			Rhodococcus hoagii	0.93*	–
			Gordonia malaquae	0.84	FP
			Mycolicibacterium smegmatis	0.83	FP
			Tsukamurella paurometabola	0.82	FP
			Gordonia sputi	0.66	FP
			Rhodococcus rhodochrous	0.52	FP
		VHMN	*Mycolicibacterium smegmatis*	0.9647*	–
			Gordonia terrae	0.6517	TP
			Mycolicibacterium smegmatis	0.6392	–
			Gordonia terrae	0.6342	TP
			Streptomyces coelicolor	0.5669	–
			Streptomyces venezuelae	0.3843	–
			Streptomyces avermitilis	0.3799	–
			Nocardia ignorata	0.3766	–
			Nocardia ignorata	0.3765	–
			Nocardia coubleae	0.3630	–
		HostG	*Gordonia*	0.33900467	TP
		RaFAH	*Gordonia*	0.744*	TP
			Mycolicibacterium	0.058	FP
			Rhodococcus	0.031	FP
			Mycobacterium	0.013	FP
			Microbacterium	0.011	FP

			Pseudopropionibacterium	0.009	–
			Arthrobacter	0.008	–
			Bifidobacterium	0.007	–
			Actinomyces	0.006	–
			Nitrolancea	0.006	–
			Streptococcus	0.006	–
		VPF-Class	*Mycobacterium*	0.5454845125 (4.02E-01)*	FP
			Mus	0.5454845125 (2.32E-01)	–
			Sulfolobus	0.5454845125 (1.27E-01)	–
			Ralstonia	0.5454845125 (6.24E-02)	–
			Aureococcus	0.5454845125 (5.29E-02)	–
			Homo	0.5454845125 (4.21E-02)	–
			Mannheimia	0.5454845125 (2.81E-02)	–
			Vibrio	0.5454845125 (1.79E-02)	–
			Acinetobacter	0.5454845125 (1.62E-02)	–
			Bacillus	0.5454845125 (1.22E-02)	–
		vHULK	*Gordonia*	0.5453266	TP
	GTE2	CHERRY	*Gordonia terrae*	1.00*	TP
			Mycolicibacterium smegmatis	0.94*	FP
			Clavibacter michiganensis	0.03	–
		VHMN	*Mycolicibacterium smegmatis*	0.7345	–
			Gordonia terrae	0.4333	TP
			Gordonia terrae	0.3420	TP
			Mycolicibacterium smegmatis	0.2973	–
			Smaragdicoccus niigatensis	0.2701	–
			Smaragdicoccus niigatensis	0.2695	–
			Timonella senegalensis	0.2536	–
			Rhodococcus kunmingensis	0.2511	–
			Coriobacteriaceae bacterium	0.2474	–
			Coriobacteriaceae bacterium	0.2443	–
		HostG	*Mycolicibacterium*	0.38070312	FP
		RaFAH	*Gordonia*	0.768*	TP

			Mycobacterium	0.135	FP
			Mycolicibacterium	0.065	FP
			Rhodococcus	0.010	TP
			Tsukamurella	0.006	FP
			Actinomyces	0.004	–
			Nocardia	0.002	TP
			Rothia	0.002	–
			Arthrobacter	0.001	–
			Bifidobacterium	0.001	–
			Corynebacterium	0.001	–
			Methylocaldum	0.001	–
			Microbacterium	0.001	–
			Ruminococcus	0.001	–
			Selenomonas	0.001	–
			Vibrio	0.001	–
		VPF-Class	*Gordonia*	0.9975927809 (7.32E-01)*	TP
			Mycobacterium	0.9975927809 (1.37E-01)	FP
			Streptomyces	0.9975927809 (2.75E-02)	FP
			Sinorhizobium	0.9975927809 (2.29E-02)	–
			Homo	0.9975927809 (1.28E-02)	–
			Arthrobacter	0.9975927809 (1.24E-02)	–
			Bacillus	0.9975927809 (9.89E-03)	–
			Pseudomonas	0.9975927809 (5.45E-03)	–
			Burkholderia	0.9975927809 (5.18E-03)	–
			Vibrio	0.9975927809 (4.49E-03)	–
		vHULK	*Gordonia*	0.98930323*	TP
	GTE5	CHERRY	*Gordonia terrae*	0.22	TP
			Rhodococcus hoagii	0.01	–
		VHMN	*Mycolicibacterium smegmatis*	0.9988*	–
			Mycolicibacterium smegmatis	0.9757*	–
			Gordonia terrae	0.9626*	TP
			Gordonia terrae	0.9625*	TP

			Streptomyces coelicolor	0.9469	–
			Pseudomonas aeruginosa	0.8571	–
			Microbacterium foliorum	0.8545	–
			Streptomyces avermitilis	0.8495	–
			Pseudomonas aeruginosa	0.8454	–
			Pseudomonas aeruginosa	0.839	–
		HostG	*Gordonia*	1.0000000*	TP
		RaFAH	*Gordonia*	0.720*	TP
			Mycolicibacterium	0.194*	FP
			Mycobacterium	0.055	FP
			Corynebacterium	0.014	–
			Tsukamurella	0.010	FP
			Rhodococcus	0.005	FP
			Nocardia	0.001	FP
			Streptomyces	0.001	FP
		VPF-Class	*Gordonia*	0.9992930149 (8.14E-01)*	TP
			Mycobacterium	0.9992930149 (5.74E-02)	FP
			Tsukamurella	0.9992930149 (4.01E-02)	FP
			Streptomyces	0.9992930149 (1.71E-02)	FP
			Bacillus	0.9992930149 (1.41E-02)	–
			Nitrincola	0.9992930149 (5.41E-03)	–
			Corynebacterium	0.9992930149 (4.42E-03)	–
			Acinetobacter	0.9992930149 (4.19E-03)	–
			Campylobacter	0.9992930149 (4.05E-03)	–
			Cellulophaga	0.9992930149 (3.81E-03)	–
		vHULK	*Gordonia*	0.9981748*	TP
	GTE6	CHERRY	*Gordonia terrae*	1.00*	TP
			Mycolicibacterium smegmatis	0.01	FP
			Rhodococcus rhodochrous	0.01	FP
		VHMN	*Mycolicibacterium smegmatis*	0.9971*	–
			Gordonia terrae	0.9516*	TP
			Gordonia terrae	0.9337	TP

			Mycolicibacterium smegmatis	0.9335	–
			Burkholderia cenocepacia	0.8845	–
			Burkholderia cenocepacia	0.8427	–
			Rhodococcus rhodnii	0.8111	FP
			Rhodococcus rhodnii	0.8049	–
			Rhodococcus rhodnii	0.8041	–
			Rhodococcus zopfii	0.8008	–
		HostG	*Gordonia*	1.0000000*	TP
		RaFAH	*Gordonia*	0.914*	TP
			Rhodococcus	0.037	FP
			Mycolicibacterium	0.019	FP
			Nocardia	0.009	FP
			Mycobacterium	0.007	FP
			Tsukamurella	0.007	FP
			Corynebacterium	0.002	–
			Streptomyces	0.002	FP
			Arthrobacter	0.001	–
			Frankia	0.001	–
			Ruminiclostridium	0.001	–
		VPF-Class	*Gordonia*	0.9621979862 (2.96E-01)	TP
			Mycobacterium	0.9621979862 (2.49E-01)	FP
			Bacillus	0.9621979862 (9.27E-02)	–
			Haloarcula	0.9621979862 (7.79E-02)	–
			Microcystis	0.9621979862 (3.38E-02)	–
			Ralstonia	0.9621979862 (2.98E-02)	–
			Aureococcus	0.9621979862 (2.16E-02)	–
			Homo	0.9621979862 (1.87E-02)	–
			Synechococcus	0.9621979862 (1.64E-02)	–
			Acidianus	0.9621979862 (1.61E-02)	–
		vHULK	*Gordonia*	0.9995683*	TP
	GTE7	CHERRY	*Gordonia terrae*	1.00*	TP
			Rhodococcus hoagii	0.86	–

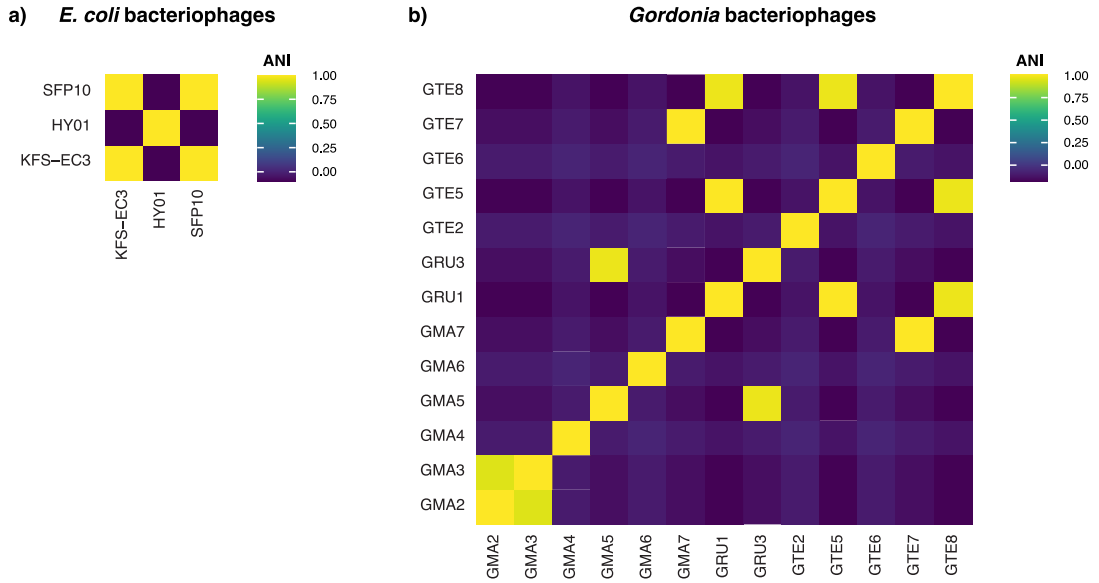
			Gordonia malaquae	0.64	TP
			Mycolicibacterium phlei	0.09	–
		VHMN	*Mycolicibacterium smegmatis*	0.6180	–
			Olsenella umbonata	0.5581	–
			Olsenella sp.	0.5511	–
			Olsenella umbonata	0.5500	–
			Streptomyces coelicolor	0.4940	–
			Olsenella sp.	0.4673	–
			Olsenella sp.	0.4645	–
			Olsenella sp.	0.4549	–
			Olsenella sp.	0.4548	–
			Bacterium	0.4530	–
			Bacterium	0.4530	–
		HostG	*Gordonia*	0.6929956	TP
		RaFAH	*Gordonia*	0.821*	TP
			Tsukamurella	0.072	FP
			Mycolicibacterium	0.059	FP
			Mycobacterium	0.034	FP
			Rhodococcus	0.004	FP
			Butyricicoccus	0.001	–
			Coprococcus	0.001	–
			Corynebacterium	0.001	–
			Frankia	0.001	–
			Lactobacillus	0.001	–
			Oenococcus	0.001	–
			Prevotella	0.001	–
			Pseudomonas	0.001	–
			Pseudopropionibacterium	0.001	–
			Rothia	0.001	–
		VPF-Class	*Escherichia*	0.9427797525 (2.68E-01)	–
			Mycobacterium	0.9427797525 (1.80E-01)	FP
			Synechococcus	0.9427797525 (1.62E-01)	–

			Bacillus	0.9427797525 (5.23E-02)	–
			Clostridium	0.9427797525 (4.99E-02)	–
			Lactobacillus	0.9427797525 (3.31E-02)	–
			Vibrio	0.9427797525 (2.99E-02)	–
			Pseudomonas	0.9427797525 (2.65E-02)	–
			Sinorhizobium	0.9427797525 (2.30E-02)	–
			Halorubrum	0.9427797525 (1.57E-02)	–
		vHULK	*Tsukamurella*	0.7927325	TP
	GTE8	CHERRY	*Gordonia terrae*	1.00*	TP
			Rhodococcus hoagii	0.01	–
		VHMN	*Mycolicibacterium smegmatis*	0.9986*	–
			Mycolicibacterium smegmatis	0.9723*	–
			Gordonia terrae	0.9704*	TP
			Gordonia terrae	0.9687*	TP
			Streptomyces coelicolor	0.9637*	–
			Streptomyces venezuelae	0.9072	–
			Microbacterium foliorum	0.9023	–
			Streptomyces avermitilis	0.8984	–
			Mycobacterium sp.	0.8942	–
			Mycobacterium sp.	0.8942	–
		HostG	*Gordonia*	1.0000000*	TP
		RaFAH	*Gordonia*	0.815*	TP
			Mycolicibacterium	0.121	FP
			Mycobacterium	0.042	FP
			Tsukamurella	0.013	FP
			Corynebacterium	0.004	–
			Rhodococcus	0.003	FP
			Streptomyces	0.001	FP
			Thermomonospora	0.001	–
		VPF-Class	*Gordonia*	0.997462661 (8.95E-01)*	TP
			Mycobacterium	0.997462661 (1.92E-02)	FP
			Streptomyces	0.997462661 (1.10E-02)	FP

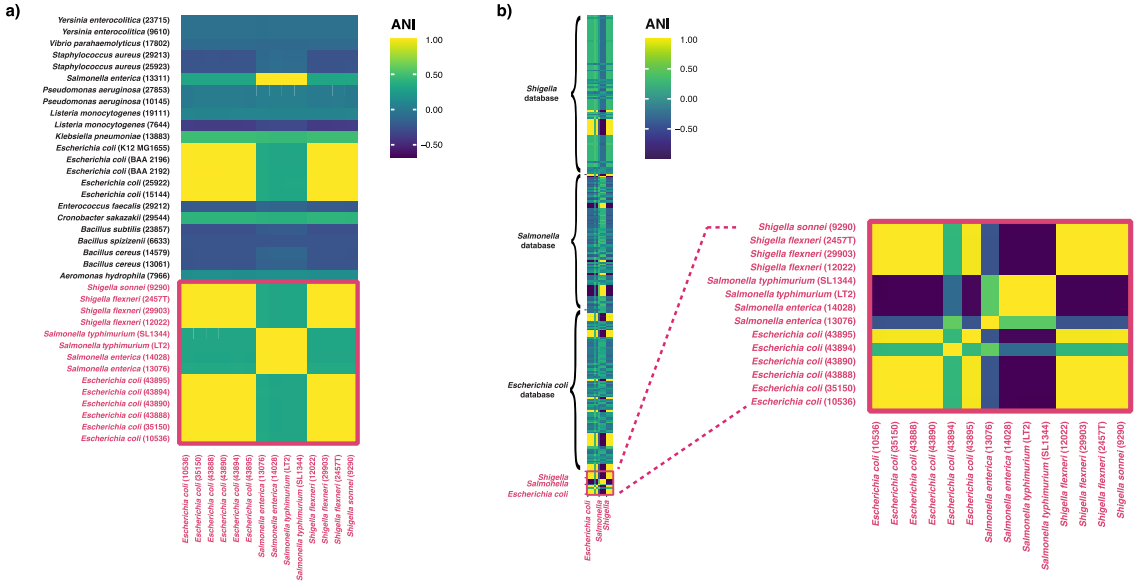
			Corynebacterium	0.997462661 (1.01E-02)	–
			Escherichia	0.997462661 (9.18E-03)	–
			Bacillus	0.997462661 (9.00E-03)	–
			Propionibacterium	0.997462661 (8.61E-03)	–
			Lactococcus	0.997462661 (7.19E-03)	–
			Riemerella	0.997462661 (6.17E-03)	–
			Cellulophaga	0.997462661 (4.98E-03)	–
		vHULK	*Gordonia*	0.999315*	TP

[1]only reported by VPF-Class

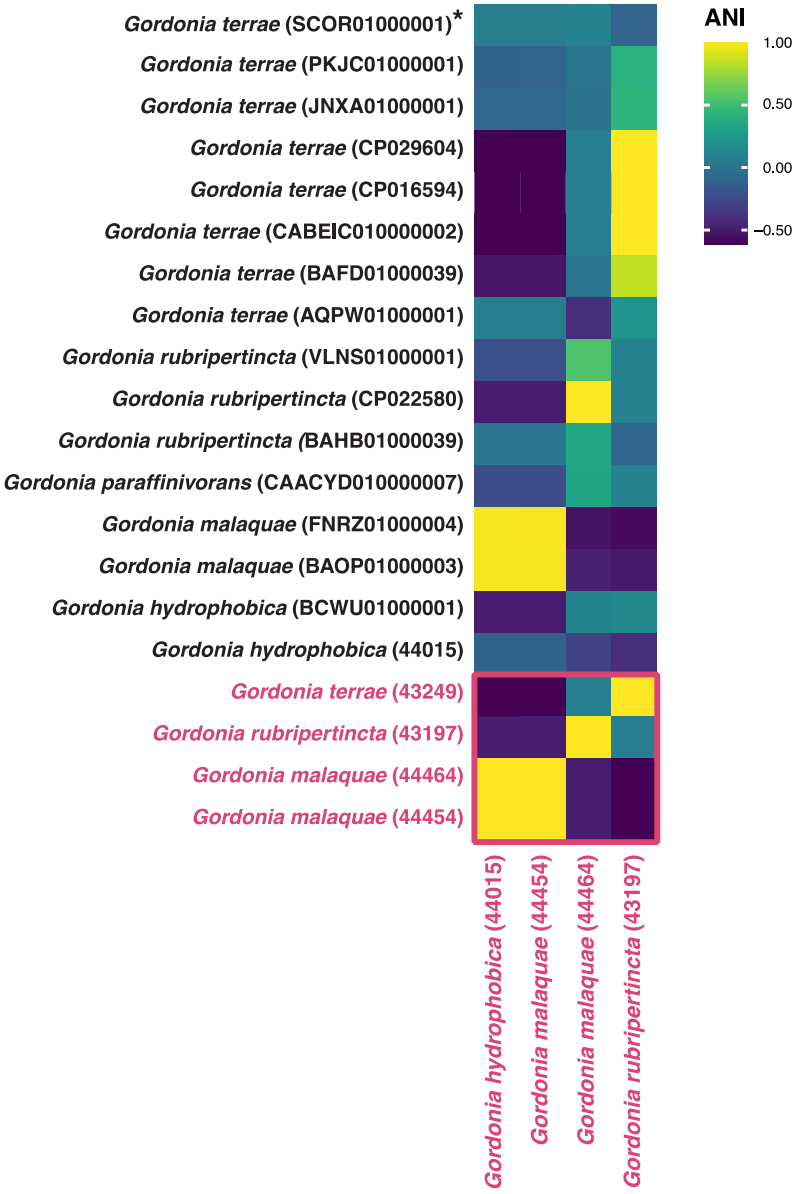
Supplementary Figure S1. Average nucleotide identity (ANI) of a) *Escherichia coli* and b) *Gordonia* bacteriophage genomes included in the analysis. Accession numbers are provided in Supplementary Tables S1 and S2, respectively.



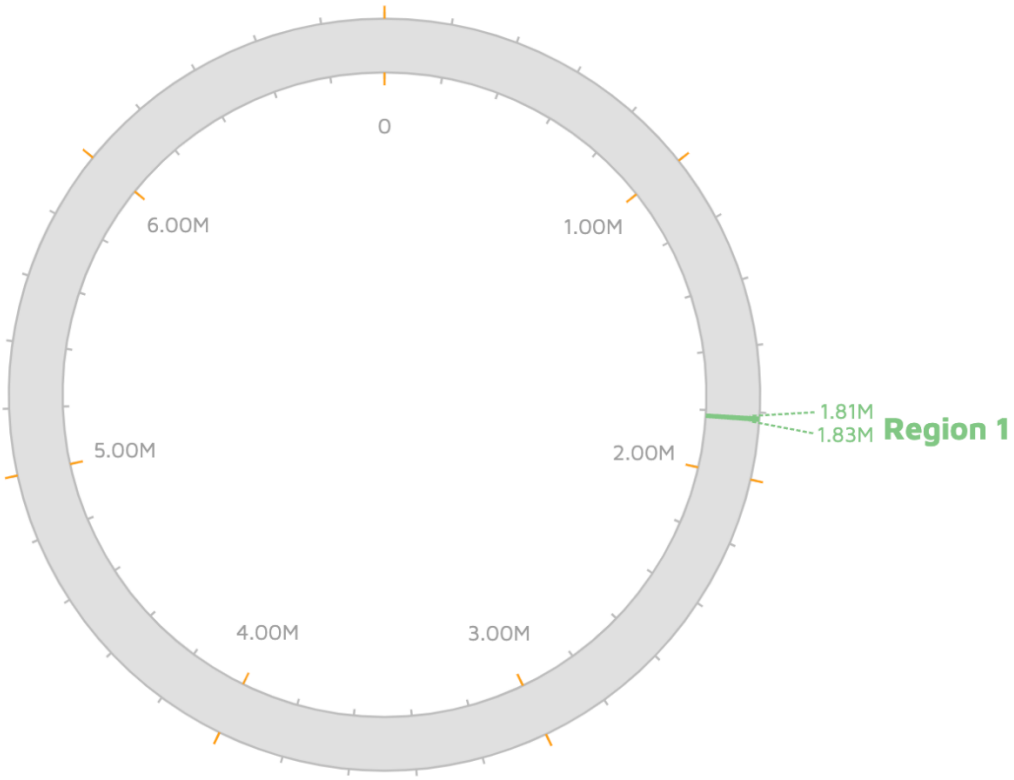
Supplementary Figure S2. Average nucleotide identity (ANI) between experimentally-validated host (shown in pink) and a) non-host (black) genomes of the three *Escherichia coli* bacteriophages HY01, KFS-EC3, and SFP10 as well as b) genomes of *E. coli*, *Salmonella*, and *Shigella* strains included in the exploratory tool databases. ATCC and NCBI accession numbers are shown in brackets.



Supplementary Figure S3. Average nucleotide identity (ANI) between experimentally-validated host (shown in pink) and non-host (black) genomes of the 13 *Gordonia* bacteriophages GMA2-7, GRU1, GRU3, GTE2, and GTE5-8 as well as genomes of closely-related *Gordonia* strains included in the exploratory tool databases. DSMZ and NCBI accession numbers are shown in brackets. * Note that, as of August 2023, the NCBI record for *Gordonia terrae* strain K (accession number: SCOR01000001) was suspended due to being from an unverified source organism.



Supplementary Figure S4. PHASTER prediction of prophages detected in *Mycobacterium smegmatis* mc^2 155.
Region 1 (shown in green) contains a BLAST hit against bacteriophage Cucurbita (e-value 2.71e-15) at position 1,822,790 bp to 1,823,125 bp, indicating the integration of a prophage.



CHAPTER 5

DEVELOPING AN APPROPRIATE EVOLUTIONARY BASELINE MODEL FOR THE STUDY OF HUMAN CYTOMEGALOVIRUS

(Previously published as A.A. Howell, J. Terbot II, V. Soni, P. Johri, J.D. Jensen*, and S.P. Pfeifer*. 2023. Developing an appropriate evolutionary baseline model for the study of human cytomegalovirus. GBE 15: evad059.)

Abstract

Human cytomegalovirus (HCMV) represents a major threat to human health, contributing to both birth defects in neonates as well as organ transplant failure and opportunistic infections in immunocompromised individuals. HCMV exhibits considerable interhost and intrahost diversity, which likely influences the pathogenicity of the virus. Therefore, understanding the relative contributions of various evolutionary forces in shaping patterns of variation is of critical importance both mechanistically and clinically. Herein, we present the individual components of an evolutionary baseline model for HCMV, with a particular focus on congenital infections for the sake of illustration—including mutation and recombination rates, the distribution of fitness effects, infection dynamics, and compartmentalization—and describe the current state of knowledge of each. By building this baseline model, researchers will be able to better describe the range of possible evolutionary scenarios contributing to observed variation as well as improve power and reduce false-positive rates when scanning for adaptive mutations in the HCMV genome.

Significance

Human cytomegalovirus (HCMV) infection is a major cause of birth defects and can lead to severe effects in immunosuppressed and immunonaïve individuals. Pathogenicity is likely driven by multiple factors, including the genetic diversity of the

virus itself. Furthermore, the accurate identification of genomic loci underlying viral adaptation relies on an appropriate baseline model that accounts for constantly operating evolutionary processes shaping this genetic diversity. With this overview of the current understanding of these processes in HCMV, we provide the necessary details for researchers to implement such a baseline model for their own genomic analysis of patient samples.

Introduction

As the leading cause of infection-related birth defects—including cognitive and hearing impairments—human cytomegalovirus (HCMV) remains a major threat to global health, with a seroprevalence of more than 90% outside of the developed world (e.g., Boppana et al. 2013; Swanson and Schleiss 2013; Dreher et al. 2014). HCMV is also a primary cause of solid organ transplant failure (Balfour 1979) and often results in opportunistic infections in immunocompromised individuals or those with immature immune systems (e.g., Suárez et al. 2019, 2020). Additionally, primary infection or reactivation is implicated in a wide variety of health complications (Griffiths et al. 2015), and recent studies suggest that HCMV may play an active role in glioma pathogenesis in individuals with glioblastoma (Cobbs et al. 2002; Abdelaziz et al. 2019). Moreover, along with human immunodeficiency virus type 1 (HIV-1), HCMV is the most common viral agent transmitted from mother to offspring and may itself contribute to the vertical transmission of HIV-1 (Johnson et al. 2015; Girsch et al. 2022).

HCMV is a β-herpesvirus in the Herpesviridae family with a relatively large double-stranded (ds) DNA genome of ∼235 kb in size, including between 164 and 167 open reading frames (ORFs) (Dolan et al. 2004). Lytic infection is initiated by the expression of genes in a flow cascade, and DNA replication initiates 1–3 days postinfection (Weekes et al. 2014). The genome contains two unique regions—the

unique long (U_L) and unique short (U_S) region—that are internally and externally flanked by repeats. The U_L region contains ORFs encoding gene products associated with latency and reactivation (Revello and Gerna 2010; Li et al. 2014); in laboratory passaged strains, cultures have been shown to accumulate large deletions in this region compared with clinically isolated viruses, likely owing to the relaxed selection in laboratory environments (Cha et al. 1996). In contrast, ORFs within the U_L region that encode envelope glycoproteins thought to be important for pathogenesis have been found to evolve under considerable constraint (He et al. 2006; Ji et al. 2006; Heo et al. 2008).

Multiple studies have suggested a link between pathogenesis and genomic variability (Meyer-König, Vogelberg, et al. 1998; Renzette et al. 2014; Wang et al. 2021), with high levels of diversity and multiple-strain infection found to be associated with higher viral loads (Pang et al. 2008; Sowmya and Madhavan 2009; Puchhammer-Stöckl and Görzer 2011). Furthermore, variation in the glycoproteins gO and gB, potentially generated through recombination (Meyer-König, Vogelberg, et al. 1998), has been proposed to influence cell tropism and dissemination (Hahn et al. 2004). Gaining a better understanding of the evolutionary forces that shape viral diversity is thus of critical importance both mechanistically and clinically. During the last decade, many efforts have been made to understand the relative contributions of admixture, positive and purifying selection, and infection-related bottlenecks in shaping HCMV interhost and intrahost variation (Renzette et al. 2013, 2015, 2017; Pokalyuk et al. 2017). Relatedly, numerous efforts have focused on elucidating key evolutionary parameters including the underlying mutation and recombination rates, as well as the selective effects of newly arising mutations (the distribution of fitness effects [DFE]; Renzette et al. 2015, 2017; Morales-Arce et al. 2022).

Importantly, recent studies focused upon evolutionary inference procedures have simultaneously demonstrated the value of jointly estimating parameters of natural selection with population history, as a neglect of one to infer the other will often result in serious misinference (Johri et al. 2020, 2021). Moreover, only by first accounting for the constantly acting evolutionary processes of genetic drift (as shaped by the infection bottleneck and subsequent viral population growth, as well as the genetic structure associated with compartmentalization) and purifying and background selection (owing to the pervasive input of deleterious mutations) may one develop a meaningful baseline model of expected levels and patterns of genomic variation. This baseline model is critical for accurately detecting and quantifying rarer and episodic evolutionary processes, such as positive selection potentially leading to viral adaptation (Johri, Aquadro, et al. 2022; Johri, Eyre-Walker, et al. 2022). More specifically, owing to overlapping patterns between neutral and selective evolutionary processes (Jensen 2009; Bank et al. 2014), this baseline model is essential for defining rates of true positives and false positives associated with the detection of rare or episodic effects in any given population and for any given data set.

As such an evolutionary baseline model has yet to be fully described for HCMV, we here outline important components of such a model and review the current state of knowledge pertaining to each: mutation rates, recombination rates, the distribution of fitness effects, infection dynamics, and compartmentalization. We close with a series of recommendations for improving evolutionary inference in this important human pathogen and highlight key areas in need of further investigation.

Mutation Rate

The mutation rate quantifies the frequency at which spontaneous (*de novo*) mutations arise in a genome, as caused by a variety of factors including DNA replication

errors and spontaneous DNA damage (see review of Pfeifer 2020). This rate is distinct from the substitution rate—that is, the rate at which mutations become fixed in a population—which is influenced not only by the *de novo* mutation rate but also by natural selection, genetic drift, as well as multiple other factors. However, for strictly neutral mutations, the rate of mutational input is equal to the rate of substitution (Kimura 1968), leading to a clock-like accumulation of mutations over time. Using a molecular clock (divergence)-based approach, recent studies have reported substitution rates of approximately 3.0×10^{-9} substitutions per nucleotide per year in HCMV (McGeoch et al. 2000)—one to two orders of magnitudes lower than the rate reported for a closely related virus, herpes simplex virus (HSV-1), which exhibits 3.0×10^{-8} (Sakaoka et al. 1994) and 1.4×10^{-7} (Kolb et al. 2013) substitutions per nucleotide per year. Mutation rates of both HCMV and HSV-1 have also been studied *in vitro*. For example, by scoring null mutations in the *tk* gene using ganciclovir, mutation rates in HSV-1 have similarly been estimated to range from 5.9×10^{-8} (Hwang et al. 2002; Drake and Hwang 2005) to 1.0×10^{-7} (Hall and Almy 1982) substitutions per nucleotide per cell infection, where cell infection is an estimate of a viral generation.

It is necessary here to highlight the various units being reported when comparing between the results described in different studies, with rates reported as substitutions per nucleotide per generation (s/n/g), substitutions per nucleotide per year (s/n/y), substitutions per nucleotide per cell infection (s/n/c), or substitutions per nucleotide per round of copying (s/n/r), if the mode of replication is known. The mode of replication of dsDNA viruses is likely limited to semiconservative replication, although RNA viruses by comparison are known to use a “stamping machine” model, where a single template is used for all progeny strands (Luria 1951). To compare between estimates using substitutions per nucleotide per cell infection and estimates using substitutions per

nucleotide per year, we have used the number of viral cycles per year as a conversion factor (table 1). Specifically, conversion factors of 181.87 to 362.48 viral cycles per year were chosen to span lower and upper estimates for HCMV, while 1,946.67 viral cycles per year were used for closely related HSV-1 for comparison. These estimates are based on internalization times of 10 min (Bodaghi et al. 1999; Hetzenecker et al. 2016) and 30 min (Zheng et al. 2014), as well as eclipse times of 24–48 h (Jean et al. 1978) and 4 h (Nishide et al. 2019), for HCMV and HSV-1, respectively. Importantly, these conversions highlight the discrepancy between divergence and *in vitro* estimates of the substitution rate, demonstrating that molecular clock-based estimates primarily provide information about the rate of neutral and nearly neutral mutation, rather than estimating full mutational spectra (as discussed in the below section). Additionally, the further analysis of future patient samples would be of great value in better characterizing the interhost variance in these rates.

Virus	Approach	Original Unit[a]	Estimated Rate/Cycle	Reference
HCMV	*In vitro*	s/n/c	2.0×10^{-7}	Renzette et al. 2015
HCMV	Divergence	s/n/y	1.6×10^{-11} / 8.2×10^{-12}	McGeoch et al. 2000
HSV-1	Divergence	s/n/y	7.1×10^{-11}	Kolb et al. 2013
HSV-1	Divergence	s/n/y	4.1×10^{-11}	Sakaoka et al. 1994
HSV-1	*In vitro*	s/n/c	1.0×10^{-7}	Hall and Almy 1982
HSV-1	*In vitro*	s/n/c	5.9×10^{-8}	Hwang et al. 2002; Drake and Hwang 2005

Table 1. In Vitro- and Divergence-Based Estimates of De Novo Mutation Rates in HCMV Compared with the Closely Related HSV-1. Note.—To compare between estimates using substitutions per nucleotide per cell infection (s/n/c) and estimates using substitutions per nucleotide per year (s/n/y), we have used conversion factors of either 181.87 or 362.48 viral cycles per year to span uncertainty in HCMV, and 1,946.67 viral cycles per year for HSV-1. [a]s = substitutions; n = nucleotide; c = cell infection; y = year.

Notably, these experimental and empirical measurements of the mutation rate based on genome-wide population genetic data neglect the substantial proportion of lethal and deleterious mutations that are removed from the population via purifying selection. Owing to this neglect, measurements obtained using these methods are likely an underestimate of the genuine genome-wide mutation rate (Peck and Lauring 2018). Mutation accumulation experiments provide a valuable (and less biased) alternative by subjecting a viral population to a series of bottlenecks that reduces the effective population size, thus minimizing the efficacy of selection. A similar strategy can be applied to natural, longitudinal population data. Using this approach, the mutation rate of

HCMV was estimated by Renzette et al. (2015) as 2.0×10^{-7} mutations per nucleotide per generation using longitudinal samples obtained from 18 patients, where mutations were called if absent in earlier samples and present in all later samples. Importantly, however, evaluating such longitudinal data in the context of a mutation accumulation study comes with the qualification that selective pressures are expected to be much stronger in patient samples relative to traditional experimental mutation accumulation lines. In addition, the presence of a reinfection event during the longitudinal sampling—if not identified—would be expected to upwardly bias these estimates. It is also important to note that rate estimates of this sort are further complicated by practical limitations of clinical sampling. Specifically, previous studies have shown that deep sequencing through the use of polymerase chain reaction amplicons requires rare variants to be present at >1% frequency in order to be reliably detected (Fonager et al. 2015; Kyeyune et al. 2016)—though newer methods that utilize target enrichment protocols may improve upon this threshold (Hage et al. 2017). Given that the vast majority of variants are expected to be rare, such detection thresholds may be of considerable significance.

Mutation rates in viruses may evolve through both mutator and antimutator alleles, the fixations of which are thought to be governed by genome size and effective population size (Lynch et al. 2016). When effective population sizes are small, selection is weak and may be unable to prevent mutator alleles from fixing. To date, one hypermutator has been identified in HCMV (Chou et al. 2016). Mutator alleles are a double-edged sword for viruses, having important implications for the rate of adaption (Taddei et al. 1997; Travis and Travis 2002), but more significantly also create the possibility of mutational meltdown (Crotty et al. 2001; Beaucourt et al. 2011; Bank et al. 2016; Matuszewski et al. 2017; Ormond et al. 2017). Indeed, owing to interference between the greater input of deleterious mutations with the minor input of beneficial

mutations, higher mutation rates may slow or stop the rate of adaptation (Pénisson et al. 2017; Jensen and Lynch 2020; Jensen et al. 2020). Other molecular determinants of viral mutation rates include postreplicative repair through interaction with DNA damage response pathways (Weitzman et al. 2010; Luftig 2014)—a particularly relevant mechanism for HCMV as herpesviruses are known to induce DNA damage responses (Xiaofei and Kowalik 2014).

As HCMV has been observed to be quite diverse compared with other DNA viruses—on the order of certain RNA viruses (Wang et al. 2002; Jerzak et al. 2005)—one formal possible explanation for the high levels of nucleotide diversity observed in HCMV is an exceptionally high mutation rate (i.e., as levels of neutral variation are expected to be a factor of the effective population size as well as the underlying mutation rate). This hypothesis was recognized as unlikely by Renzette et al. (2011), owing, among other reasons, to the proofreading activity of HCMV's DNA polymerase (Nishiyama et al. 1983). Although Cudini et al. (2019) recently rediscussed this possibility (and see the response of Jensen and Kowalik 2020), there appears to be general agreement that RNA virus-like levels of variation in HCMV are not due to RNA virus-like mutation rates. Specifically, following multiple studies on HCMV interhost and intrahost variation (Renzette et al. 2013, 2015, 2017; Pokalyuk et al. 2017; and see the below sections), it has been demonstrated that observed diversity is likely generated by a combination of mutation, recombination, reinfection, compartmentalization, selection, and infection population size histories (Jensen 2021)—with a mutation rate of 2.0×10^{-7} mutations per nucleotide per generation appearing consistent with the data (Renzette et al. 2015). More specifically, the observed high levels of variation appear to more likely be related to the population dynamics related to compartmentalization, gene flow, and reinfection, rather than to particularly elevated rates of mutation (e.g., Pokalyuk

et al. 2017; Jensen and Kowalik 2020). Renzette et al. additionally identified a weak but highly significant positive correlation between estimated mutation rates and single nucleotide polymorphism (SNP) density across the HCMV genome, as may be expected. Heterogeneity in mutation rates across the genome was additionally proposed as a contributing factor underlying the observed correlations between intraspecies variation and recombination rates, as well as of that between variation and divergence (Renzette et al. 2016).

Recombination Rate

Recombination not only contributes genetic variation through the generation of novel genotypic combinations, but it may also improve the efficacy of selection through the reduction of interference effects between and among beneficial and deleterious variants (Hill and Robertson 1966; Felsenstein 1974; Lynch et al. 1995; Pénisson et al. 2017). Studies examining the intergenic variability of HCMV glycoprotein loci (Meyer-König, Haberland, et al. 1998; Haberland et al. 1999; Yan et al. 2008) provided the initial evidence for homologous recombination in the HCMV genome. Nearly two decades later, Renzette et al. (2015) estimated a genome-wide recombination map using a population genetic approach, reporting a mean recombination rate of ~0.23 crossover events per genome per generation, based on observed patterns of linkage disequilibrium (LD) (i.e., by assessing the extent to which observed haplotype distributions may be explained by variable rates of recombination; and see the review of Stumpf and McVean (2003) for a discussion on estimating recombination rates from population genetic data). The authors further reported a correlation between recombination rate and SNP density, consistent with widespread purifying selection, as has been observed in multiple diverse species (e.g., Begun and Aquadro 1992; Pfeifer and Jensen 2016; Renzette et al. 2017; and see the review of Charlesworth and Jensen 2021). However, as with mutation rates,

recombination rate estimates can also be misinferred, for example, due to unaccounted for progeny skew, which is known to increase levels of LD in highly skewed populations relative to standard Wright–Fisher expectations (and as such may downwardly bias recombination rate estimation if unaccounted for; Eldon and Wakeley 2008; Birkner et al. 2013). This observation highlights the need for further computational method development of mutation and recombination rate estimators for the type of generalized progeny skew distributions applicable to viruses and other human pathogens (Morales-Arce et al. 2020; Sabin et al. 2022).

In addition to LD–based approaches, studies have also characterized recombination in the HCMV genome using a combination of phylogenetic and population-level analyses. By constructing “phylogenetic trees” for each gene in the HCMV genome and correcting for recombination breakpoints with the genetic algorithm GARD, Kosakovsky Pond et al. (2006) found that the majority of loci showed no consistent phylogenetic patterns, indicating that recombination occurs often enough that whole genomes can behave as “gene-scale mosaics.” In other words, what certain authors refer to as variable phylogenetic trees are in fact better described as variable coalescent histories. Further, like the Renzette et al. studies, Sijmons et al. (2015) also observed a correlation between recombination rate and nucleotide diversity using a phylogenetic approach. However, phylogenetic-based approaches are generally poorly suited for the study of recombination compared with the coalescent-based approaches utilized in population genetics—and multiple studies suffer from these limitations when trying to distinguish between recombination and competing evolutionary processes in a phylogenetic framework (e.g., Houldcroft et al. 2016; Cudini et al. 2019). Specifically, coalescent theory provides a sophisticated framework for the study of variable gene genealogies owing to recombination (Wakeley 2009) and avoids the pretense of

searching for a single (and nonexistent) “phylogenetic tree’ to describe within-population variation (e.g., Cudini et al. 2019; and see Rosenberg and Nordborg 2002 for a discussion).

The Distribution of Fitness Effects (DFE)

HCMV is characterized by a large genome relative to other human viruses. Although the set of protein-coding genes in HCMV experiences constant revision, there are 45 core genes that are conserved across all herpesviruses and ~117 noncore genes that are more specific to the CMVs, many of which are still being functionally characterized (Van Damme and Van Loock 2014; Mozzi et al. 2020). Although it is clear that protein-coding regions occupy the majority of the HCMV genome, these uncertainties mean that the precise fraction of the genome that experiences direct purifying selection is not yet fully defined—though roughly 25% of the genome has been observed to be nearly devoid of variation, potentially suggesting strong constraint (Renzette et al. 2015). Interestingly, within-patient nucleotide diversity in noncoding regions of the genome has generally been observed to be on the same order as less-constrained coding regions (Renzette et al. 2011), suggesting the presence of functionally important regions interspersed across the genome and/or widespread background selection effects (Renzette et al. 2016). This combination of factors renders the identification of neutrally evolving sites challenging.

Previous studies have used comparisons of sequence evolution at nonsynonymous versus synonymous sites at various evolutionary scales to quantify selective forces acting on protein-coding regions in the HCMV genome. A comparative genomic analysis across multiple CMV species found pervasive purifying selection in most protein-coding regions (as indicated by low levels of d_N/d_S; Mozzi et al. 2020), as would be expected. Similarly, comparisons of sequence polymorphism within hosts to

the divergence among hosts (i.e., using the McDonald and Kreitman 1991 test) also indicated the action of widespread purifying selection (Renzette et al. 2011). In contrast, evidence for positive selection was limited to specific regions, including the glycoproteins (Renzette et al. 2013). Thus, although glycoproteins and their linked regions will likely be additionally impacted by recurrent selective sweeps, the majority of the genome is expected to be largely affected by the direct and linked effects of purifying selection.

As selection against harmful mutations at functionally important sites in the genome can affect patterns of variation at linked neutral alleles (i.e., background selection; Charlesworth et al. 1993) and as this effect has been suggested to be a primary determinant of genomic variation in HCMV (Renzette et al. 2016), it is important to characterize the DFE of newly arising mutations across the genome. A recent study by Morales-Arce et al. (2022) used an approximate Bayesian computation (ABC) framework to infer the DFE of deleterious mutations from a within-patient sample of HCMV. This study accounted for the specific demographic history of the within-patient population as associated with viral infection dynamics (as previously inferred by Renzette et al. 2013), non-Wright–Fisher replication dynamics, as well as background selection. They inferred that roughly 50% of all new mutations were effectively neutral ($-1<2Nes\leq 0$), 24% were mildly deleterious ($-10<2Nes\leq -1$), 12% were moderately deleterious ($-100<2Nes\leq -10$), and 13% were strongly deleterious ($2Nes\leq -100$), where Ne refers to the effective population size and *s* to the selection coefficient against the homozygote (fig. 1*AA*). As these estimates were obtained for all sites comprising the functional region (i.e., the inference was not restricted to nonsynonymous sites) and ~30% of all sites in coding regions are likely to have little or no fitness costs upon mutation (e.g., synonymous changes), the DFE at functionally important sites in HCMV is probably closer to 30% effectively neutral, 34% weakly deleterious, 17% moderately

deleterious, and 19% lethal mutations (fig. 1*BB*). Importantly, although such a correction naturally depends on the fraction of synonymous sites that are behaving neutrally, these estimates are in fact quite consistent with multiple previous random mutagenesis studies that measured the proportion of lethal mutations in DNA viruses to be ~20% (e.g., Sanjuán 2010). While Morales-Arce et al. (2022) accounted for a number of factors that add complexity to within-patient populations of HCMV (including an extremely strong bottleneck corresponding to the infection), they simulated only a single population of HCMV. As there is strong evidence of HCMV populations being structured within patients (Pokalyuk et al. 2017; Sackman et al. 2018; and see the section on Compartmentalization below), current estimates of the deleterious DFE might still be biased, and future inference incorporating both compartmentalization and reinfection will be important in this regard.

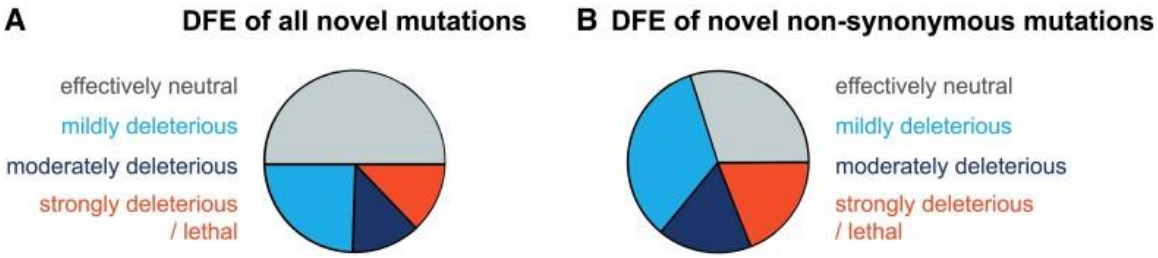


Fig.1. Distribution of fitness effects (DFE) of all new and new nonsynonymous mutations. (*A*) Using an approximate Bayesian framework to account for the specific demographic history of their within-patient population, Morales-Arce et al. (2022) inferred the DFE of all new mutations in human cytomegalovirus as roughly 50% effectively neutral ($-1<2Nes\leq 0$; gray), 24% mildly deleterious ($-10<2Nes\leq -1$; light blue), 12% moderately deleterious ($-100<2Nes\leq -10$; dark blue), and 13% strongly deleterious/lethal ($2Nes\leq -100$; red), where Ne refers to the effective population size and *s* to the selection coefficient against the homozygote. (*B*) Assuming that ~30% of all sites in coding regions likely have little or no fitness costs upon mutation, the DFE at functionally important sites corresponds to roughly 30% effectively neutral, 34% mildly deleterious, 17% moderately deleterious, and 19% strongly deleterious/lethal mutations.

Infection Dynamics

The demographic history of a population is an important determinant of both genetic variation and potential selective outcomes and therefore an appropriate starting point for evolutionary analysis, particularly in light of the high levels of HCMV diversity observed within patients (Drew et al. 1984; Spector et al. 1984; Haberland et al. 1999; Faure-Della Corte 2010; Renzette et al. 2011, 2013, 2015, 2016, 2017; Hage et al. 2017; Pokalyuk et al. 2017). The expected intrahost population dynamics involve a strong population bottleneck (a temporary reduction in population size) at the point of infection, followed by rapid population expansion (see review of Jensen 2021). The level of intrahost genetic variation that is present at the point of infection will in part be determined by the severity of the bottleneck. If the transmission bottleneck is wide, then there may be numerous virions founding the initial infection, resulting in greater genetic variation and an increased probability that beneficial variants may be transferred from the founding population. Conversely, a narrow bottleneck can result in a severe loss of genetic variation, with low-frequency variants being eliminated regardless of their fitness effects. This process is known as a founder effect (see Zwart and Elena 2015, for a discussion of this effect in viral populations).

In the case of congenital infections, demographic modeling approaches have shown support for a population bottleneck associated with the initial transplacental infection (transmission of virions from the maternal compartment to the fetal plasma compartment), followed by additional bottlenecks associated with compartmental infections (fig. 2; and Renzette et al. 2013; for a detailed discussion regarding the population structure dynamics between compartments, see the section below). Importantly, the initial bottleneck was shown to involve potentially hundreds of unique HCMV genomes, which helps to explain the relatively high levels of genetic diversity

observed at the point of infection, as compared with certain RNA viruses in which a single (or very few) virions are thought to be involved in infection (Keele et al. 2008; Fischer et al. 2010; Renzette et al. 2013, 2014). Furthermore, Renzette et al. (2013) found support for gene flow between urine and plasma compartments (the two compartments sampled in that study). Their results further suggested that plasma may serve as a “route” for gene flow within the host, with preliminary evidence indicating that it carries compartment-specific variants from other compartments; this process may thus also be an important determinant of within-host variation.

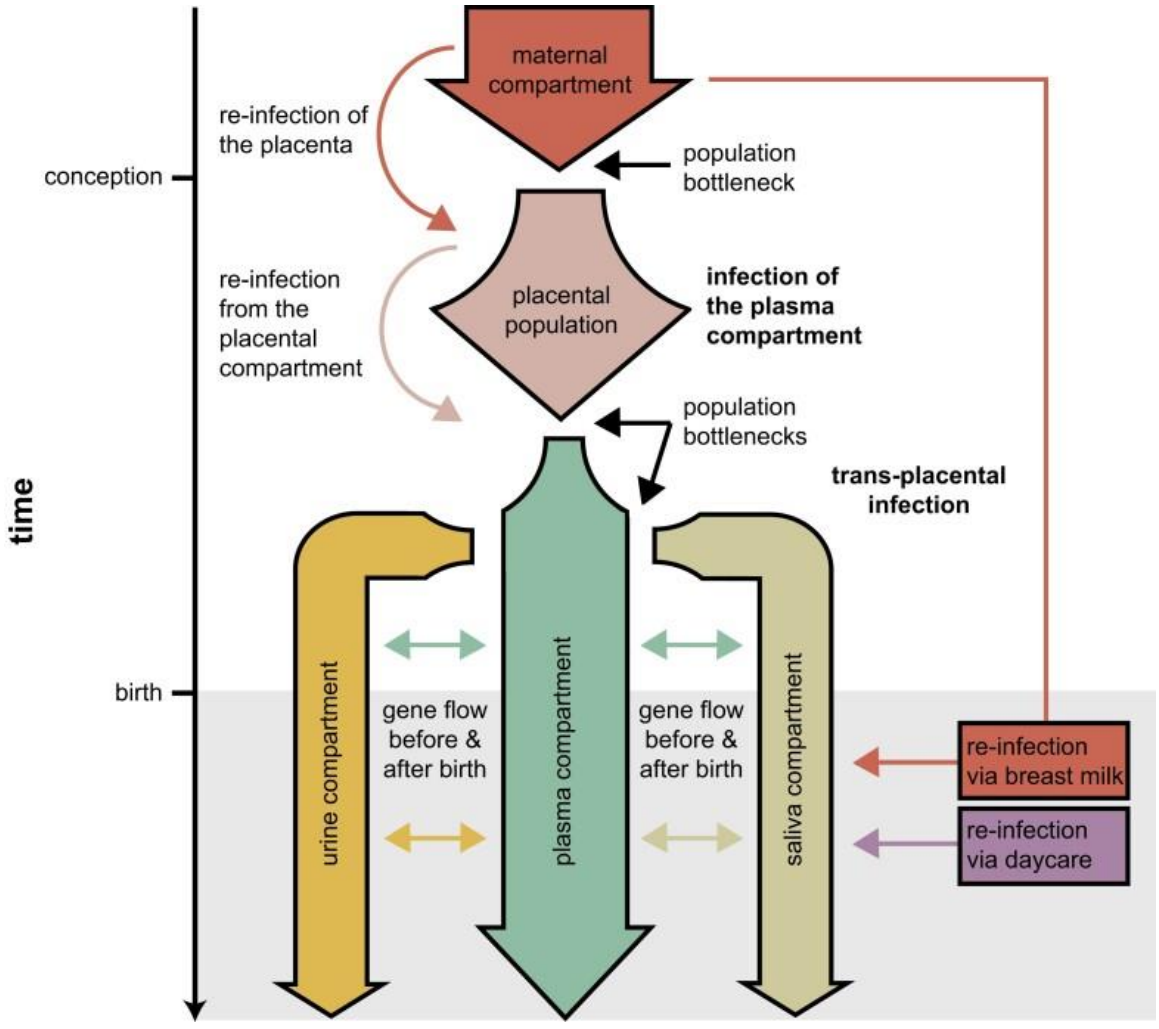


Fig. 2. Demographic dynamics of congenital human cytomegalovirus (HCMV) infection. Demographic scenarios of infection and reinfection in HCMV likely contributing to the high levels of observed interhost and intrahost diversity, including a population bottleneck associated with the initial transplacental infection (transmission of virions from the maternal compartment [red]/plasma [pink] to the fetal plasma compartment [green]), followed by additional bottlenecks associated with compartmental infections (urine [yellow] and saliva [olive]), as well as gene flow between compartments and reinfection of compartments during pregnancy and after birth (e.g., via breast milk [red] and/or daycare [purple]).

Further evidence for admixture between compartments (this time including plasma, urine, and saliva compartments) was found by Pokalyuk et al. (2017), suggesting that reinfection postbirth is possible via, for instance, breast milk (Numazaki 1997; Enders et al. 2011; also see the review of Bardanzellu et al. 2019). In other words, maternal compartment-specific variants appeared to be transmitted to the infant postbirth. Although the above examples are focused upon congenital infections, related work has similarly highlighted the importance of multistrain infections in immunosuppressed adults and particularly the relationship between this infection status and the emergence of antiviral resistance mutations (e.g., in transplant recipients; Suárez et al. 2019, 2020).

To date no method exists to prevent maternal–fetal transmission or to reduce the severity of fetal infection (Britt 2017). Therefore, the characterization of population dynamics is likely to be integral to future therapeutic strategies. For example, clinically imposing a more severe population bottleneck during pregnancy may reduce genetic variation in the HCMV infecting population, limiting the pool of variation on which natural selection may subsequently act, thereby potentially improving treatment outcomes. Finally, it has been shown that host immune suppression can reactivate dormant viruses, restarting production of viral progeny; this switch from latent to productive life cycles can induce temporary or sustained CMV replication (Porter et al. 1985; Dupont and Reeves 2016).

Demographic inference in HCMV is inherently challenging due to the genome-wide impact of selection (see the DFE section above), which will in turn bias common demographic estimators which are based on neutrality (see the discussion of Ewing and Jensen 2016; Pouyet et al. 2018). Namely, neutral demographic estimators require sufficiently large nonfunctional regions and high rates of recombination, such that assumptions of strict neutrality hold (Gutenkunst et al. 2009; Excoffier et al. 2013; Kelleher et al. 2019; Steinrücken et al. 2019). These criteria ensure that variants can be chosen that are not experiencing background selection. For example, Renzette et al. (2013) utilized *∂a∂i*, a neutral demographic inference approach based on the site frequency spectrum (Gutenkunst et al. 2009), to build and parameterize HCMV infection models (and see Sackman et al. 2018; Jensen and Kowalik 2020).

This inference problem of estimating demography in the presence of selection is indeed somewhat circular, as the estimation of selection will also be biased by unaccounted for demographic dynamics (Rousselle et al. 2018; Johri et al. 2020). This fact highlights the importance of performing joint, simultaneous inference of selection with demography, rather than taking the more common stepwise approach of first estimating one and then the other (see review of Johri, Eyre-Walker, et al. 2022). Recently proposed ABC approaches that jointly estimate population history and the DFE of deleterious mutations perform such joint inference and importantly do not require the *a priori* identification of neutrally evolving sites (Johri et al. 2020). Explicitly accounting for viral infection dynamics, Morales-Arce et al. (2020) incorporated progeny skew into the joint ABC inference scheme of Johri et al. (2020)—an important extension to this framework as the assumption of small progeny distributions utilized by a majority of population genetic inference approaches is likely violated in many pathogens, as noted above (see reviews of Tellier and Lemaire 2014; Irwin et al. 2016). The authors

demonstrated that their tailoring of this ABC inference approach specifically to viral populations avoided misinference resulting from a neglect of this consideration. Other recent inference approaches have also relaxed the assumption of small progeny skew, demonstrating an ability to coestimate parameters related to the biology of progeny skew together with those of demographic and selective histories (e.g., Matuszewski et al. 2018; Sackman et al. 2019).

Compartmentalization

The final consideration of note impacting intrahost population dynamics of viral infections is population structure between different areas of infection, commonly referred to as compartmentalization (Zárate et al. 2007). Compartmentalization may be relevant for any virus not localized to a single organ or cell type (Di Liberto et al. 2006; Zárate et al. 2007; Renzette et al. 2014; Sackman et al. 2018)—including HCMV, known to infect several cells and organs throughout the body.

As a long-studied virus, HCMV has been well documented to infect a wide variety of cells including the epithelial cells of gland and mucosal tissue, smooth muscle cells, fibroblasts, macrophages, dendritic cells, hepatocytes, and vascular endothelial cells (Sinzger et al. 2008; Jean Beltran and Cristea 2014). Unsurprisingly given this broad cellular tropism, evidence of infection in specific organs is similarly extensive and includes the brain and peripheral nerves, the eyes, the placenta, the lungs, the gastrointestinal tract from the esophagus to the colon, the liver, the lymph nodes, the heart, the peripheral blood, and the kidneys (Plachter et al. 1996). Of these areas, viral shedding from salivary glands, the ductal epithelium of mammary glands and the kidney, and the syncytiotrophoblasts (placenta) is thought to be critical to interhost transmission (Mocarski 2004; Kinzler and Compton 2005). However, because of potential gene flow

between compartments within a host, other sites of infection are nonetheless important for understanding the intrahost dynamics of this virus.

Another necessary consideration is the location of regions that can harbor the latent stage—these areas are likely important for the maintenance of genetic diversity that may otherwise be lost in actively replicating lineages (Chou 1989; Frange et al. 2013). While infections can occur across the body, the latent, and importantly nonreproducing, stage of the virus seems to be limited in cell tropism. Specifically, HCMV has been found to use endothelial and select myeloid lineages as well as monocytes, macrophages, and their progenitors (i.e., cells found in the circulating plasma population) as latency sites (Jarvis and Nelson 2002; Yatim and Albert 2011).

Given the wide range of potential sites of infection, it is crucial to resolve observed levels of intrahost population structuring that are indicative of compartmentalization. Several studies have observed considerable genomic diversity (Renzette et al. 2011, 2013; Mayer et al. 2017; Pokalyuk et al. 2017; Cudini et al. 2019; Pang et al. 2020), while others have found intrahost populations to be comparatively invariant (Hage et al. 2017). The comparison of patients with single- versus multiple-infection histories is likely one important source of disparity in these observed levels of variation (Mayer et al. 2017; Pokalyuk et al. 2017; Sackman et al. 2018; Cudini et al. 2019; Jensen and Kowalik 2020; Houldcroft et al. 2020; Pang et al. 2020). It should also be noted that the importance of multiple infections in shaping intrahost diversity of infants may still rely on compartmentalization within the maternal infection (e.g., with primary infections arising from the cervical population and secondary infections being associated with the mammary gland population; Sackman et al. 2018; Pang et al. 2020).

Compartmentalization has also been implicated as a clinically important factor in the development of a multidrug resistant lineage within the chronic infections of immunocompromised patients (Frange et al. 2013; Renzette et al. 2014; Suárez et al. 2019, 2020). Furthermore, multiple population genetic studies using longitudinally sampled patient data concluded that compartmentalization is an important factor in explaining intrahost diversity of fetal and infant infections (Renzette et al. 2013, 2015). Models developed from these studies focused on three subpopulations corresponding to source sites of samples: salivary glands/saliva, blood/plasma, and kidney/urine (Renzette et al. 2014, 2015; Pokalyuk et al. 2017; Sackman et al. 2018). Generally, these models attribute plasma as the circulating population that serves as an intermediary for spread between the distal compartments of salivary glands and kidney (fig. 2). Of particular note, levels of genetic divergence between compartments of a single patient were found to be as great as those observed between the same compartment sampled from unrelated patients (Renzette et al. 2013), suggesting limited between-compartment gene flow within a single host. Yet, the extent to which these considerable levels of differentiation are attributable to localized, compartment-specific adaptation, or simply the constant operation of neutral evolutionary processes, remains unresolved—and this continues to stand as one of the most pressing and interesting evolutionary questions in the HCMV system.

Closing Thoughts

When developing an evolutionary baseline model of HCMV, special consideration should be given to the demographic processes that shape genetic diversity and the sampling methods that generate clinical data sets, including the ability to detect low-frequency variants, as well as the level of progeny skew, bottleneck severity during infection and reinfection, and the degree of compartmental admixture.

Correctly modeling these processes and accounting for various ascertainment biases will allow researchers to better describe the relative contributions of each evolutionary force in shaping observed levels and patterns of variation, as well as quantify uncertainty in model choice and in the identification of adaptive loci. In addition, gaining a better understanding of when and how HCMV diversity is generated has important implications for vaccine development as well as antiviral therapy, both for determining the timing of drug delivery and for combating resistance evolution.

CHAPTER 6

CONCLUSION

We characterized the genomes of two newly identified bacteriophages, Phegasus and BiggityBass, and phylogenetically placed them within their respective clusters in chapters 2 and 3 and their associated appendices. In Phegasus, we identified an integration-dependent immunity system, which regulates the switch between lytic and lysogenic life cycles. Computationally inferring host ranges for Phegasus, we identified three putative hosts (M. smegmatis, Mycobacterium chelonae, and Mycobacterium leprae) that contain the attachment site motif necessary for lysogenic infection by bacteriophages with an integration-dependent immunity system (Broussard et. al 2013). This indicates that these hosts are at risk of incorporating virulence factors from bacteriophages that utilize tyrosine integrases in their integration-dependent immunity systems (Pham et. al 2007), and that for these particular hosts Phegasus is not a suitable candidate for antibacterial therapeutics. In BiggityBass, we identified a toxin/antitoxin (TA) system that allows it to inactivate bacteria-encoded toxins (Otsuka and Yonesaki 2012; Wei et. al 2016). We showed that the gene tree of the hicA-like toxin does not recapitulate the whole genome phylogeny, which may be due to the mosaic architecture of the genome caused by horizontal gene transfer, or could be an artifact of inconsistent resampling during bootstrapping caused by the short sequence length (Lawrence et. al 2002).

Further exploring the host range prediction tools used in the study of Cluster P and Cluster DR bacteriophages, we assessed the performance of ten computational host range prediction tools using a dataset of bacteriophages whose host ranges have been experimentally validated in chapter 4. Our results demonstrated that the confirmatory tool PHP and the exploratory tool CHERRY have the highest rates of true-

positive predictions, but at the cost of having the highest rates of false-positives. While PHP, WIsH, and VHM all use kmer frequency as their prediction metric, VHM's background kmer subtracting strategy and WIsH's overly specific 8-mer Markov model likely contributes to their high rate of false negatives compared to PHP's high rate of false positives. Phirbo underpredicts due its alignment-based method, which is biased towards predicting hosts which have an existing CRISPR spacer (yet only 40%–70% prokaryotes encode a CRISPR system at all (Edwards et. al 2016)) or lysogenic phages which leave a genetic mark in the host. For the exploratory tools, the features each of them are trained on and the type of machine-learning model used are directly related to the accuracy of the tool's results. Features such as kmer frequency and CRISPR/prophage sequences alone (VHMN) are less accurate than using them in combination with protein clustering (CHERRY), as demonstrated in this study and others that use non-polyvalent phages in their benchmarking (Shang and Sun 2022). For genus-level exploratory tools, while the expected order of improving performance would be vpf-class, RaFAH, vHULK, and HostG with increasing model sophistication, HostG and vHULK only return one genus-level prediction per phage, which represents a major drawback when predicting polyvalent phage host range rather than the best virus-host pair. Therefore, between vpf-class and RaFAH, RaFAH has the most true-positive predictions and fewer false negatives. The results of this evaluation study highlight that for polyvalent phages there are still challenges to accurately predicting the true inter-genus and intra-genus host range, and that even strain-specific differences may influence virus-host compatibility. Additional factors determining the success of phage infection, including recognition of specific host receptors, ability to overcome bacterial Restriction-Modification (RM) and abortive (Abi) systems, and compatibility of transcription and translational machinery could be considered to more accurately

determine host range. CHERRY presents a promising framework for integrating these features through its multimodal graph model. For exploratory tools, one of the primary limitations in adopting these tools is the disparity between strains in each tool's internal database and the strains used in experimental validation. We recommend incorporating the model sophistication of the exploratory tools with the flexibility of the confirmatory tools to evaluate the likelihood of phage-host interaction with strains researchers have available to them.

Finally, the work of chapter 5 describes the current state of knowledge of the components of an evolutionary baseline model for Human Cytomegalovirus (HCMV), including the mutation rate, recombination rate, the distribution of fitness effects, infection dynamics, and compartmentalization of the virus. The significance of this work is that HCMV is a major cause of birth defects and can lead to severe effects in immunosuppressed and immunonaive individuals. The accurate identification of genomic loci underlying viral adaptation relies on an appropriate baseline model that accounts for constantly operating evolutionary processes shaping this genetic diversity. From our review we conclude that special consideration should be given to the ability to detect low frequency variants, the level of progeny skew, the bottleneck severity during infection and re-infection, and the degree of compartmental admixture when modeling HCMV evolutionary scenarios. By providing an overview of the current understanding of the components of an evolutionary baseline model of HCMV, we provide the necessary details for researchers to implement such a baseline model for their own genomic analysis of patient samples.

REFERENCES

Abdelaziz, M. O., Ossmann, S., Kaufmann, A. M., Leitner, J., Steinberger, P., Willimsky, G., Raftery, M. J., & Schönrich, G. (2019). Development of a Human Cytomegalovirus (HCMV)-Based Therapeutic Cancer Vaccine Uncovers a Previously Unsuspected Viral Block of MHC Class I Antigen Presentation. *Frontiers in Immunology*, 10, 1776. https://doi.org/10.3389/FIMMU.2019.01776.

Ahlgren, N.A., Ren, J., Young, Lu Y., Fuhrman, J.A., Sun, F. (2016). Alignment-free d2* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.* 45:39–53. doi: 10.1093/nar/gkw1002.

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

Amgarten, D., Iha, B.K.V., Piroupo, C.M., Da Silva,A.M. and Setubal,J.C. (2022). vHULK, a new tool for bacteriophage host prediction based on annotated genomic features and neural networks. *PHAGE (New Rochelle)*, 3, 204–212.

Arenskötter M., Bröker D., Steinbüchel A. (2004). Biology of the metabolically diverse genus Gordonia. *Appl. Environ. Microbiol.* 70:3195–3204. doi: 10.1128/AEM.70.6.3195-3204.2004.

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

Aubry A, Jarlier V, Escolano S, Truffot-Pernot C, Cambau E. (2000). Antibiotic susceptibility pattern of Mycobacterium marinum. *Antimicrob Agents Chemother.* 44(11):3133–3136.

Baláž,A., Kajsik,M., Budiš,J., Szemes,T. and Turňa,J. (2023) PHERI – Phage Host ExploRation Pipeline. *Microorganisms,* 11, 1398.

Balfour HH. (1979). Cytomegalovirus: the troll of transplantation. *Arch Intern Med.* 139(3):279–280.

Bank C, et al..(2016). An experimental evaluation of drug induced mutational meltdown as an antiviral strategy. *Evolution* 70(11):2470–2484.

Bank C, Foll M, Ferrer-Admetlla A, Ewing G, Jensen JD. (2014). Thinking too positive? Revisiting current methods in population genetic statistical inference. *Trends Genet.* 30(12):540–546.

Bardanzellu F, Fanos V, Reali A. (2019). Human breast milk-acquired cytomegalovirus infection: certainties, doubts and perspectives. *Curr Pediatr Rev.* 15(1):30–41.

Beaucourt S, et al. (2011). Isolation of fidelity variants of RNA viruses and characterization of virus mutation frequency. *J Vis Exp.* 52:2953.

Begun DJ, Aquadro CF. (1992). Levels of naturally occurring DNA polymorphism correlation with recombination rates in D. melanogaster. Nature 356(6369):519–520.
Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 1999;27(2):573–580.

Bernard K. (2012). The genus Corynebacterium and other medically relevant coryneform-like bacteria. *J Clin Microbiol*. 50(10):3152–3158.

Bidnenko E., Ehrlich S.D., Chopin M.-C. (2002). Lactococcus lactis phage operon coding for an endonuclease homologous to RuvC. *Mol. Microbiol*. 28:823–834. doi: 10.1046/j.1365-2958.1998.00845.x.

Birkner M, Blath J, Eldon B. (2013). An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics* 193(1):255–290.

Blazanin M, Turner PE. (2021). Community context matters for bacteria-phage ecology and evolution. *The ISME Journal.* 15:11. 15:3119–3128. doi: 10.1038/s41396-021-01012-x.

Bodaghi B, et al.(1999). Entry of human cytomegalovirus into retinal pigment epithelial and endothelial cells by endocytosis. *Investig Ophthalmol* Vis Sci. 40(11):2598–2607.

Boppana SB, Ross SA, Fowler KB. (2013). Congenital cytomegalovirus infection: clinical outcome. *Clin Infect Dis.* 57(Suppl 4):S178–S181.

Britt WJ. (2017). Congenital human cytomegalovirus infection and the enigma of maternal immunity. *J Virol.* 91(15):e02392-16.

Broussard GW, Hatfull GF. (2013). Evolution of genetic switch complexity. *Bacteriophage.* 3(1):e24186.

Broussard GW, Oldfield LM, Villanueva VM, Lunt BL, Shine EE, Hatfull GF. (2013). Integration-dependent bacteriophage immunity provides insights into the evolution of genetic switches. *Mol Cell.* 49(2):237–248.

Butt A., Higman V.A., Williams C., Crump M.P., Hemsley C.M., Harmer N., Titball R.W. (2014). The hicA toxin from Burkholderia pseudomallei has a role in persister cell formation. *Biochem. J.* 459:333–344. doi: 10.1042/BJ20140073.

Carbone,A. (2008) Codon bias is a major factor explaining phage evolution in translationally biased hosts. *J Mol Evol*, 66, 210–223.

Cha T-A, et al..(1996). Human cytomegalovirus clinical isolates carry at least 19 genes not found in laboratory strains. *J Virol.* 70(1):78–83.

Charlesworth B, Jensen JD. (2021). Effects of selection at linked sites on patterns of genetic variability. *Annu Rev Ecol Evol.* 52:177–197.

Charlesworth B, Morgan MT, Charlesworth D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* 134(4):1289–1303.

Chen Y, Zhan Z, Zhang H, Bi L, Zhang X-E, Fu YV.(2019). Kinetic analysis of DNA compaction by mycobacterial integration host factor at the single-molecule level. *Tuberculosis (Edinb).* 119:101862.

Chevallereau A, Pons BJ, van Houte S, Westra ER. (2021). Interactions between bacterial and phage communities in natural environments. *Nature Reviews Microbiology* 20:1. 20:49–62. doi: 10.1038/s41579-021-00602-y.

Chou S, Ercolani RJ, Lanier ER. (2016). Novel cytomegalovirus UL54 DNA polymerase gene mutations selected in vitro that confer brincidofovir resistance. *Antimicrob Agents Chemother.* 60(6):3845–3848.

Chou SW. (1989). Reactivation and recombination of multiple cytomegalovirus strains from individual organ donors. *J Infect Dis.* 160(1):11–15.

Cobbs CS, et al.. (2002). Human cytomegalovirus infection and expression in human malignant glioma 1. *Cancer Res*. 62(12):3347–3350.

Coclet, C. and Roux, S. (2021). Global overview and major challenges of host prediction methods for uncultivated phages. Curr Opin Virol, 49, 117–126.

Comeau A, Hatfull G, Krisch H, Lindell D, et al. (2008). Exploring the prokaryotic virosphere. *Res Microbiol*. 159:306–13.

Coutinho,F.H., Edwards,R.A. and Rodríguez-Valera,F. (2019) Charting the diversity of uncultured viruses of Archaea and Bacteria. *BMC Biol*, 17, 1–16.

Coutinho,F.H., Zaragoza-Solas,A., López-Pérez,M., Barylski,J., Zielezinski,A., Dutilh,B.E., Edwards,R. and Rodriguez-Valera,F. (2021) RaFAH: host prediction for viruses of Bacteria and Archaea based on protein content. *Patterns* (NY), 2, 100274.

Crane A, Versoza CJ, Hua T, Kapoor R, Llyod L, Mehta R, Menolascino J, Morais A, Munig S, Patel Z, et al. (2021). Phylogenetic relationships and codon usage bias amongst cluster K mycobacteriophages. *G3 (Bethesda).* 11(11):jkab291.

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

Crotty S, Cameron CE, Andino R. (2001). RNA Virus error catastrophe: direct molecular test by using ribavirin. *Proc Natl Acad Sci U S A*. 98(12):6895–6900.

Cudini J, et al. (2019). Human cytomegalovirus haplotype reconstruction reveals high diversity due to superinfection and evidence of within-host recombination. *Proc Natl Acad Sci U S A.* 116(12):5693–5698.

Curran E., Callaway S.E., Dumanlang R.R., Harshaw A.V., Palacio P.N., Nakamura Y., Kimberley K.W., Theoret J.R., Yoon E.J., Windsor E.J., et al. (2022) Genome sequences of Gordonia rubripertincta bacteriophages AnarQue and Figliar. *Microbiol. Resourc. Announc.* 11:e01085-21. doi: 10.1128/mra.01085-21.

Curtis F.A., Reed P., Sharples G.J. (2004). Evolution of a phage RuvC endonuclease for resolution of both Holliday and branched DNA junctions. *Mol. Microbiol*. 55:1332–1345. doi: 10.1111/j.1365-2958.2004.04476.x.

Cvijovic I, Good BH, Desai MM. (2018). The effect of strong purifying selection on genetic diversity. *Genetics* 209:1235–1278.

Dang Q., Tan W., Zhao X., Li D., Li Y., Yang T., Li R., Zu G., Xi B. (2019) Linking the response of soil microbial community structure in soils to long-term wastewater irrigation and soil depth. *Sci. Total Environ.* 688:26–36. doi: 10.1016/j.scitotenv.2019.06.138.

De los Reyes M.F., de los Reyes F.L., Hernandez M., Raskin L. (1998) Quantification of Gordona amarae strains in foaming activated sludge and anaerobic digester systems with oligonucleotide hybridization probes. *Appl. Environ. Microbiol*. 64:2503–2512. doi: 10.1128/AEM.64.7.2503-2512.1998.

Dedrick,R.M., Freeman,K.G., Nguyen,J.A., Bahadirli-Talbott,A., Smith,B.E., Wu,A.E., Ong,A.S., Lin,C.T., Ruppel,L.C., Parrish,N.M., Hatfull,G.F. and Cohen,K.A. (2021) Potent antibody-mediated neutralization limits bacteriophage treatment of a pulmonary Mycobacterium abscessus infection. *Nat Med,* 27, 1357–1361.

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., McMullan,B.J., Michalik,D.E., Moscatelli,A., Nick,J.A., Tupayachi Ortiz,M.G., Polenakovik,H.M., Robinson,P.D., Skurnik,M., Solomon,D.A., Soothill,J., Spencer,H., Wark,P., Worth,A., Schooley,R.T. and Benson,C.A., Hatfull,G.F. (2023) Phage therapy of Mycobacterium infections: compassionate use of phages in 20 patients with drug-resistant mycobacterial disease. *Clin Infect Dis,* 76, 103–112.

Dekel-Bird N.P., Avrani S., Sabehi G., Pekarsky I., Marston M.F., Kirzner S., Lindell D. (2013). Diversity and evolutionary relationships of T7-like podoviruses infecting marine cyanobacteria. *Environ. Microbiol*. 15:1476–1491. doi: 10.1111/1462-2920.12103.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 27:4636–4641.

Di Liberto G, et al.(2006). Clinical and therapeutic implications of hepatitis C virus compartmentalization. *J Gastroenterol.* 131(1):76–84.

Dolan A, et al..(2004). Genetic content of wild-type human cytomegalovirus. *J Gen Virol.* 85(Pt 5):1301–1312.

Doyle EL, Fillman CL, Reyna NS, Tobiason DM, Westholm DE, Askins JL, Backus BP, Baker AC, Ballard HS, Bisesi PJ, et al. (2017). Genome sequences of four cluster P mycobacteriophages. *Genome Announc*. 6:e01101-17.

Drake JW, Hwang CBC. (2005). On the mutation rate of herpes simplex virus type 1. *Genetics* 170(2):969–970.

Dreher AM, et al.(2014). Spectrum of disease and outcome in children with symptomatic congenital cytomegalovirus infection. *J Pediatr*.164(4):855–859.

Drew WL, Sweet ES, Miner RC, Mocarski ES. (1984). Multiple infections by cytomegalovirus in patients with acquired immunodeficiency syndrome: documentation by Southern blot hybridization. *J Infect Dis*. 150(6):952–953.

Dunn W., Chou C., Li H., Hai R., Patterson D., Stolc V., Zhu H., Liu F. (2003) Functional profiling of a human cytomegalovirus genome. *Proc. Natl. Acad. Sci. USA*. 100:14223–14228.

Dupont L, Reeves MB. (2016). Cytomegalovirus latency and reactivation: recent insights into an age old problem. *Rev Med Virol*. 26(2):75–89.

Dyson,Z.A., Tucci,J., Seviour,R.J. and Petrovski,S. (2015) Lysis to kill: evaluation of the lytic abilities, and genomics of nine bacteriophages infective for Gordonia spp. and their potential use in activated sludge foam biocontrol. *PLoS ONE*, 10, e0134512.

Edwards,R.A. and Rohwer,F. (2005) Viral metagenomics. *Nat Rev Microbiol*, 3, 504–510.

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

Eldon B, Wakeley J. (2008). Linkage disequilibrium under skewed offspring distribution among individuals in a population. *Genetics* 178(3):1517–1532.

Enders G, Daiminger A, Bäder U, Exler S, Enders M. (2011). Intrauterine transmission and clinical outcome of 248 pregnancies with primary cytomegalovirus infection in relation to gestational age. *J Clin Virol*. 52(3):244–246.

Eren,A.M., Esen,Ö.C., Quince,C., Vineis,J.H., Morrison,H.G., Sogin,M.L. and Delmont,T.O. (2015) Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ,* 3, e1319.

Erill I., Caruso S.M. (2018). Complete genome sequence of Streptomyces bacteriophage abt2graduateex2. *Genome Announc*. 6:e01480-17. doi: 10.1128/genomeA.01480-17.

Ewing GB, Jensen JD. (2016). The consequences of not accounting for background selection in demographic inference. *Mol Ecol.* 25(1):135–141.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genet*. 9(10):e1003905.

Faure-Della Corte M, et al..(2010). Variability and recombination of clinical human cytomegalovirus strains from transplantation recipients. *J Clin Virol*. 47(2):161–169.

Felsenstein J. (1974). The evolutionary advantage of recombination. *Genetics* 78(2):737–756.

Filée J., Bapteste E., Susko E., Krisch H.M. (2006). A selective barrier to horizontal gene transfer in the T4-type bacteriophages that has preserved a core genome with the viral replication and structural genes. *Mol. Biol. Evol*. 23:1688–1696. doi: 10.1093/molbev/msl036.

Fischer W, et al.(2010). Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS One* 5(8):e12303.

Fonager J, et al.(2015). Identification of minority resistance mutations in the HIV-1 integrase coding region using next generation sequencing. *J Clin Virol*. 73:95–100.

Ford M.E., Sarkis G.J., Belanger A.E., Hendrix R.W., Hatfull G.F. (1998) Genome structure of Mycobacteriophage D29: Implications for phage evolution 1. Edited by J. Karn. *J. Mol. Biol.* 279:143–164. doi: 10.1006/jmbi.1997.1610.

Frange P, et al.. (2013). Temporal and spatial compartmentalization of drug-resistant cytomegalovirus (CMV) in a child with CMV meningoencephalitis: implications for sampling in molecular diagnosis. *J Clin Microbiol*. 51(12):4266–4269.

Friedlander E, Steinrücken M. (2022). A numerical framework for genetic hitchhiking in populations of variable size. *Genetics* 220:iyac012.

Furfaro,L.L., Payne,M.S. and Chang,B.J. (2018) Bacteriophage therapy: clinical trials and regulatory hurdles. *Front Cell Infect Microbiol*, 8, 376.

Gabler F., Nam S., Till S., Mirdita M., Steinegger M., Söding J., Lupas A.N., Alva V. (2020). Protein sequence analysis using the MPI Bioinformatics Toolkit. *Curr. Protoc. Bioinform.* 72:e108. doi: 10.1002/cpbi.108.

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

Geoghegan JL, Holmes EC. (2017). Predicting virus emergence amid evolutionary noise. *Open Biol*. 7(10):170189.

Girsch JH, et al..(2022). Host-viral interactions at the maternal-fetal interface. What we know and what we need to know*. Front Virol*. 2:16.

Goodfellow M., Alderson G., Chun J. (1998). Rhodococcal systematics: Problems and developments. Antonie Leeuwenhoek. 74:3–20. doi: 10.1023/A:1001730725003.

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

Griffiths P, Baraniak I, Reeves M. (2015). The pathogenesis of human cytomegalovirus. *J Pathol*. 235(2):288–297.

Grisold A.J., Roll P., Hoenigl M., Feierl G., Vicenzi-Moser R., Marth E. (2007). Isolation of Gordonia terrae from a patient with catheter-related bacteraemia. *J. Med. Microbiol.* 56:1687–1688. doi: 10.1099/jmm.0.47388-0.

Groth AC, Calos MP. (2004). Phage integrases: biology and applications. *J Mol Biol.* 335(3):667–678.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*. 5(10):e1000695.

Haberland M, Meyer-König U, Hufert FT. (1999). Variation within the glycoprotein B gene of human cytomegalovirus is due to homologous recombination. *J Gen Virol.* 80(Pt 6):1495–1500.

Hage E, et al. (2017). Characterization of human cytomegalovirus genome diversity in immunocompromised hosts by whole-genome sequencing directly from clinical specimens. *J Infect Dis*. 215(11):1673–1683.

Hahn G, et al. (2004). Human cytomegalovirus UL131-128 genes are indispensable for virus growth in endothelial cells and virus transfer to leukocytes. *J Virol*. 78(18):10023–10033.

Hakki M., Chou S. (2011). The biology of cytomegalovirus drug resistance. *Curr. Opin. Infect. Dis*. 2011;24:605–611. doi: 10.1097/QCO.0b013e32834cfb58.

Hall JD, Almy RE. (1982). Evidence for control of herpes simplex virus mutagenesis by the viral DNA polymerase. *Virology* 116(2):535–543.

Han S.-S., Kang H.K., Jo B.Y., Ryu B.-G., Jin H.M., Chung E.J., Jung J.Y. (2020). Complete genome sequence of Gordonia rubripertincta SD5, a soil bacterium isolated from a Di-(2-Ethylhexyl) Phthalate-degrading enrichment culture. *Microbiol. Resourc. Announc*. 2020;9:e01087-20. doi: 10.1128/MRA.01087-20.

Han XY, Tarrand JJ, Infante R, Jacobson KL, Truong M. (2005). Clinical significance and epidemiologic analyses of Mycobacterium avium and Mycobacterium intracellulare among patients without AIDS. *J Clin Microbiol*. 43(9):4407–4412.

Harada,L.K., Silva,E.C., Campos,W.F., Del Fiol,F.S., Vila,M., Dąbrowska,K., Krylov,V.N. and Balcão,V.M. (2018) Biotechnological applications of bacteriophages: state of the art. *Microbiol Res*, 212-213, 38-58.

Harrington D.A.L., Stevens J.L., Johnson M.J., Pochiro S.J., Moriarty M.M., Robertson M.E., Sanchez A., Whitby O.G., Kimberley K.W., McKenna C.C., et al. Genome sequences of Gordonia rubripertincta bacteriophages Jellybones and NHagos. Microbiol. Resourc. Announc. 2020;9:e00935-20. doi: 10.1128/MRA.00935-20.

Hatfull G.F., Pedulla M.L., Jacobs-Sera D., Cichon P.M., Foley A., Ford M.E., Gonda R.M., Houtz J.M., Hryckowian A.J., Kelchner V.A., et al. (2006). Exploring the

mycobacteriophage metaproteome: Phage genomics as an educational platform. *PLoS Genet.* 2:e92. doi: 10.1371/journal.pgen.0020092.

Hatfull, G. F. (2010). Mycobacteriophages: Genes and genomes. *Annu. Rev. Microbiol.* 64:331–356.

Hatfull,G.F., Dedrick,R.M. and Schooley,R.T. (2022). Phage therapy for antibiotic-resistant bacterial infections. *Annu Rev Med*, 73, 197–211.

He R, et al..(2006). Sequence variability of human cytomegalovirus UL146 and UL147 genes in low-passage clinical isolates. I*ntervirology* 49(4):215–223.

Hendrix, R. W. (2002). Bacteriophages: Evolution of the majority. *Theor. Popul. Biol.* 61(4): 471–480.

Hendrix, R. W., Lawrence, J. G., Hatfull, G. F., and Casjens, S. (2000). The origins and ongoing evolution of viruses. *Trends Microbiol*. 8(11):504–508.

Hendrix, R. W., Smith, M. C., Burns, R. N., Ford, M. E., and Hatfull, G. F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proc. Natl. Acad. Sci*. USA 96(5):2192–2197.

Heo J, et al..(2008). Polymorphisms within human cytomegalovirus chemokine (UL146/UL147) and cytokine receptor genes (UL144) are not predictive of sequelae in congenitally infected children. *Virology* 378(1):86–96.

Hetzenecker S, Helenius A, Krzyzaniak MA. (2016). HCMV Induces macropinocytosis for host cell entry in fibroblasts. *Traffic* 17(4):351–368.

Hill WG, Robertson A. (1966). The effect of linkage on limits to artificial selection. *Genet Res.* 8:269–294.

Hochschild A, Douhan J3rd, Ptashne M. (1986) How lambda repressor and lambda Cro distinguish between OR1 and OR3. *Cell.* 47(5):807–816.

Houldcroft CJ, Cudini J, Goldstein RA, Breuer J. (2020). Reply to Jensen and Kowalik: consideration of mixed infections is central to understanding HCMV intrahost diversity. *Proc Natl Acad Sci U S A*. 117(2):818–819.

Houldcroft CJ, et al..(2016). Detection of low frequency multi-drug resistance and novel putative maribavir resistance in immunocompromised pediatric patients with cytomegalovirus. *Front Microbiol*. 7:1317.

Howard-Varona C, Hargreaves KR, Abedon ST, Sullivan MB. (2017) Lysogeny in nature: mechanisms, impact, and ecology of temperate phages. *ISME* J. 2017;11(7):1511–1520.

Hu X, Wang HY, Otero CE, Jenks JA, Permar SR. (2022). Lessons from Acquired Natural Immunity and Clinical Trials to Inform Next-Generation Human Cytomegalovirus Vaccine Development. https://doi.org/10.1146/annurev-virology-100220-010653. 9:491–520.

Hwang YT, Liu BY, Hwang CBC. (2002). Replication fidelity of the supF gene integrated in the thymidine kinase locus of herpes simplex virus type 1. *J Virol*. 76(8):3605–3614.

Inglis,L.K. and Edwards,R.A. (2022) How metagenomics has transformed our understanding of bacteriophages in microbiome research. Microorganisms, 10, 1671. Irwin KK, et al.. 2016. On the importance of skewed offspring distributions and background selection in virus population genetics. *Heredity* 117:393–399.

Jacobs, W. R., Jr. (2000). Mycobacterium tuberculosis: A once genetically intractable organism. In ‘‘Molecular Genetics of the Mycobacteria’’ (G. F. Hatfull and W. R. Jacobs, Jr., eds.), pp. 1–16. *ASM Press*, Washington, DC.

Jacobs, W. R., Jr., Tuckman, M., and Bloom, B. R. (1987). Introduction of foreign DNA into mycobacteria using a shuttle phasmid. *Nature* 327(6122):532–535.

Jacobs-Sera D., Marinelli L.J., Bowman C., Broussard G.W., Guerrero Bustamante C., Boyle M.M., Petrova Z.O., Dedrick R.M., Pope W.H. (2012). Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) et al. On the nature of mycobacteriophage diversity and host preference. *Virology.* 2012;434:187–201. doi: 10.1016/j.virol.2012.09.026.

Jarvis MA, Nelson JA. (2002). Mechanisms of human cytomegalovirus persistence and latency. *Front Biosci*. 7:d1575–d1582.

Jean Beltran PM, Cristea IM. (2014). The life cycle and pathogenesis of human cytomegalovirus infection: lessons from proteomics. Expert Rev Proteomics 11(6):697–711.

Jean JH, Yoshimura N, Furukawa T, Plotkin SA. (1978). Intracellular forms of the parental human cytomegalovirus genome at early stages of the infective process. *Virology* 86(1):281–286.

Jensen JD, Kowalik TF. (2020). A consideration of within-host human cytomegalovirus genetic variation. *Proc Natl Acad Sci*. 117(2):816–817.

Jensen JD, Lynch M. (2020). Considering mutational meltdown as a potential SARS-CoV-2 treatment strategy. *Heredity* 124:619–620.

Jensen JD, Stikeleather RA, Kowalik TF, Lynch M. (2020). Imposed mutational meltdown as an antiviral strategy. *Evolution* 74(12):2549–2559.

Jensen JD. (2009). On reconciling single and recurrent hitchhiking models. *Genome Biol Evol.* 1:320–324.

Jensen JD. (2021). Studying population genetic processes in viruses: from drug-resistance evolution to patient infection dynamics. 4th ed. *Encyclopedia of Virology* 5:227–232.

Jerzak G, Bernard KA, Kramer LD, Ebel GD. (2005). Genetic variation in West Nile virus from naturally infected mosquitoes and birds suggests quasispecies structure and strong purifying selection. *J Gen Virol*. 86(Pt 8):2175–2183.

Ji YH, et al. (2006). Polymorphisms of human cytomegalovirus UL148A, UL148B, UL148C, UL148D genes in clinical strains. *J Clin Virol*. 37(4):252–257.

Johnson EL, et al..(2015). Cytomegalovirus upregulates expression of CCR5 in central memory cord blood mononuclear cells, which may facilitate in utero HIV type 1 transmission. *J Infect Dis.* 211(2):187–196.

Johri P, Aquadro CF, et al.. (2022). Recommendations for improving statistical inference in population genomics*. PLoS Biol*. 20:e3001669.

Johri P, Charlesworth B, Jensen JD. (2020). Toward an evolutionarily appropriate null model: jointly inferring demography and purifying selection. *Genetics* 215(1):173–192.

Johri P, et al..(2021). The impact of purifying and background selection on the inference of population history: problems and prospects. *Mol Biol Evol*. 38(7):2986–3003.

Johri P, Eyre-Walker A, Gutenkunst RN, Lohmueller KE, Jensen JD. (2022). On the prospect of achieving accurate joint estimation of selection with population history. *Genome Biol Evol.* 14(7):evac088.

Kang Y., Yuan L., Shi X., Chu Y., He Z., Jia X., Lin Q., Ma Q., Wang J., Xiao J., et al. A fine-scale map of genome-wide recombination in divergent Escherichia coli population. *Brief. Bioinform.* (2021);22:bbaa335. doi: 10.1093/bib/bbaa335.

Katoh K., Standley D.M. (2013) MAFFT Multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol*. 2013;30:772–780. doi: 10.1093/molbev/mst010.

Keele BF, et al..(2008). Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci U S A.* 105(21):7552–7557.

Keen EC. (2015). A century of phage research: Bacteriophages and the shaping of modern biology. BioEssays : news and reviews in molecular, cellular and developmental biology. 37:6. doi: 10.1002/BIES.201400152.

Kelleher J, et al..(2019). Inferring whole-genome histories in large population datasets. *Nat Genet.* 51(9):1330–1338.

Kesari K.K., Soni R., Jamal Q.M.S., Tripathi P., Lal J.A., Jha N.K., Siddiqui M.H., Kumar P., Tripathi V., Ruokolainen J. (2021) Wastewater treatment and reuse: A review of its applications and health implications. *Water Air Soil Pollut*. 232:208. doi: 10.1007/s11270-021-05154-8.

Kim M, Ryu S.(2013). Antirepression system associated with the life cycle switch in the temperate Podoviridae phage SPC32H. *J Virol*. 87(21):11775–11786.

Kim,S.-H., Adeyemi,D.E. and Park,M.-K. (2021) Characterization of a new and efficient polyvalent phage infecting E. coli O157:H7, Salmonella spp., and Shigella sonnei. *Microorganisms,* 9, 2105.

Kimura M. (1968). Evolutionary rate at the molecular level. Nature 217:624–626.
Kinzler ER, Compton T. 2005. Characterization of human cytomegalovirus glycoprotein-induced cell-cell fusion. *J Virol.* 79(12):7827–7837.

Kolb AW, Ané C, Brandt CR. (2013). Using HSV-1 genome phylogenetics to track past human migrations. *PLoS One* 8(10):e76267.

Kolmogorov,M., Yuan,J., Lin,Y. and Pevzner,P.A. (2019) Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*, 37, 540–546.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. (2006). GARD: a genetic algorithm for recombination detection. *Bioinformatic*s 22(14):3096–3098.

Koskella,B. and Brockhurst,M.A. (2014) Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol Rev,* 38, 916–931.

Kragelund C., Remesova Z., Nielsen J.L., Thomsen T.R., Eales K., Seviour R., Wanner J., Nielsen P.H. (2007) Ecophysiology of mycolic acid-containing Actinobacteria (Mycolata) in activated sludge foams. *FEMS Microbiol. Ecol.* 61:174–184. doi: 10.1111/j.1574-6941.2007.00324.x.

Krumsiek J., Arnold R., Rattei T. (2007). Gepard: A rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics.* 23:1026–1028. doi: 10.1093/bioinformatics/btm039.

Kuek,M., McLean,S.K. and Palombo,E.A. (2022). Application of bacteriophages in food production and their potential as biocontrol agents in the organic farming industry. *Biol Control,* 165, 104817.

Kumar S., Stecher G., Li M., Knyaz C., Tamura K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol. Biol. Evol.* 2018;35:1547–1549. doi: 10.1093/molbev/msy096.

Kyeyune F, et al.. (2016). Low-frequency drug resistance in HIV-infected Ugandans on antiretroviral treatment is associated with regimen failure. *Antimicrob Agents Chemother.* 60(6):3380–3397.

Landy A. (1989) Dynamic, structural, and regulatory aspects of lambda site-specific recombination. *Annu Rev Biochem*. 58:913–949.

Lawrence J.G., Hatfull G.F., Hendrix R.W. (2002) Imbroglios of viral taxonomy: Genetic exchange and failings of phenetic approaches. *J. Bacteriol*. 184:4891–4905. doi: 10.1128/JB.184.17.4891-4905.2002.

Lee,H., Ku,H.-J., Lee,D.-H., Kim,Y.-T., Shin,H., Ryu,S. and Lee,J.-H. (2016) Characterization and genomic study of the novel bacteriophage HY01 infecting both Escherichia coli O157:H7 and Shigella flexneri: potential as a biocontrol agent in food. *PLoS ONE,* 11, e0168985.

Leger,A. and Leonardi,T. (2019) pycoQC, interactive quality control for oxford nanopore sequencing. *J Open Source Softw*, 4, 1236.

Lemire S, Figueroa-Bossi N, Bossi L. (2011) Bacteriophage crosstalk: coordination of prophage induction by trans-acting antirepressors. *PLoS Genet*. 7(6):e1002149.

Letunic I, Bork P. (2019) Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47(W1):W256–W259.

Lewis JA, Hatfull GF. (2003) Control of directionality in L5 integrase-mediated site-specific recombination. *J Mol Biol.* 2003;326(3):805–821.

Li G, et al..(2014). An epistatic relationship between the viral protein kinase UL97 and the UL133-UL138 latency locus during the human cytomegalovirus lytic cycle. *J Virol.* 88(11):6047–6060.

Li G., Shen M., Lu S., Le S., Tan Y., Wang J., Zhao X., Shen W., Guo K., Yang Y., et al. (2016) Identification and characterization of the hicAB toxin-antitoxin system in the opportunistic pathogen Pseudomonas aeruginosa. *Toxins.* 8:113. doi: 10.3390/toxins8040113.

Lilley D.M.J., White M.F. (2001). The junction-resolving enzymes. *Nat. Rev. Mol. Cell.* Biol. 2:433–443. doi: 10.1038/35073057x.

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*. 25:762–777. doi: 10.1093/MOLBEV/MSN023.

Little,J.S., Dedrick,R.M., Freeman,K.G., Cristinziano,M., Smith,B.E., Benson,C.A., Jhaveri,T.A., Baden,L.R., Solomon,D.A. and Hatfull,G.F. (2022) Bacteriophage treatment of disseminated cutaneous Mycobacterium chelonae infection. *Nat Commun,* 13, 2313.

López-Cuevas,O., Medrano-Félix,J.A., Castro-Del Campo,N. and Chaidez,C. (2021) Bacteriophage applications for fresh produce food safety. *Int J Environ Health Res*, 31, 687–702.

Lowe TM, Eddy SR. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25:955–964.

Lozano-Andrade C.N., Strube M.L., Kovács T. (2021) Complete genome sequences of four soil-derived isolates for studying synthetic bacterial community assembly. *Microbiol. Resourc. Announc*;10:e00848-21. doi: 10.1128/MRA.00848-21.

Lu,C., Zhang,Z., Cai,Z., Zhu,Z., Qiu,Y., Wu,A., Jiang,T., Zheng,H. and Peng,Y. (2021) Prokaryotic virus Host Predictor: a gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC Biol*, 19, 5.

Luftig MA. (2014). Viruses and the DNA damage response: activation and antagonism. *Annu Rev Virol*. 1(1):605–625.

Lukashin AV, Borodovsky M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* 26:1107–1115.

Luo Q., Hiessl S., Poehlein A., Daniel R., Steinbüchel A. Steinbüchel. Insights into the microbial degradation of rubber and gutta-percha by analysis of the complete genome of Nocardia nova SH22a. Appl. *Environ. Microbiol.* 2014;80:3895–3907. doi: 10.1128/AEM.00473-14.

Luria SE. (1951). The frequency distribution of spontaneous bacteriophage mutants as evidence for the exponential rate of phage reproduction. *Cold Spring Harb Symp Quant* Biol. 16:463–470.

Lynch M, Conery J, Burger R. (1995). Mutation accumulation and the extinction of small populations. *Am Nat*. 146(4):489–518.

Lynch M, et al..(2016). Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet.* 17(11):704–714.

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

Malachowa N, Deleo FR. (2010). Mobile genetic elements of Staphylococcus aureus. *Cell Mol Life Sci.* 67(18):3057–3071.

Manesh,A. and Varghese,G.M. (2021) Rising antimicrobial resistance: an evolving epidemic in a pandemic. *Lancet Microbe*, 2, e419–e420.

Mangutov EO, Georgievna Kharseeva G, Alutina EL. (2021). Coryneebacterium spp.—problematic pathogens of the human respiratory tract. *Klin Lab Diagn*. 66(8):502–508.

Manni,M., Berkeley,M.R., Seppey,M., Simão,F.A. and Zdobnov,E.M. (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol*, 38, 4647–4654.

Marçais,G. and Kingsford,C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27, 764–770.

Matuszewski S, Hildebrandt ME, Achaz G, Jensen JD. (2018). Coalescent processes with skewed offspring distributions and non-equilibrium demography. *Genetics* 208(1):323–338.

Matuszewski S, Ormond L, Bank C, Jensen JD. (2017). Two sides of the same coin: a population genetics perspective on lethal mutagenesis and mutational meltdown. *Virus Evol.* 3(1):vex004.

Mayer BT, et al..(2017). Transient oral human cytomegalovirus infections indicate inefficient viral spread from very few initially infected cells. J Virol. 91(12):e00380-17. McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351(6328):652–654.

McGeoch D. J. et al. (2008), 'Molecular Evolution of the Herpesvirales', in *Origin and Evolution of Viruses,* 2nd edn Amsterdam: Elsevier, pp. 447–75.

McGeoch D. J., Rixon F. J., Davison A. J. (2006) 'Topics in Herpesvirus Genomics and Evolution', *Virus Research*, 117: 90–104.

McGeoch DJ, Dolan A, Ralph AC. (2000). Toward a comprehensive phylogeny for mammalian and avian herpesviruses. *J Virol.* 74(22):10401–10406.

Mesyanzhinov V.V., Robben J., Grymonprez B., Kostyuchenko V.A., Bourkaltseva M.V., Sykilinda N.N., Krylov V.N., Volckaert G. (2002). The genome of bacteriophage ΦKZ of Pseudomonas aeruginosa. *J. Mol. Biol*. 317:1–19. doi: 10.1006/jmbi.2001.5396.

Meyer JR, Dobias DT, Weitz JS, Barrick JE, Quick RT, Lenski RE. Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science.* 2012;335(6067):428–432.

Meyer-König U, Haberland M, Von Laer D, Haller O, Hufert FT. (1998). Intragenic variability of human cytomegalovirus glycoprotein B in clinical strains. J Infect Dis. 177(5):1162–1169.

Meyer-König U, Vogelberg C, et al..(1998). Glycoprotein B genotype correlates with cell tropism in vivo of human cytomegalovirus infection. J Med Virol. 55(1):75–81.

Mocarski ES. (2004). Immune escape and exploitation strategies of cytomegaloviruses: impact on and imitation of the major histocompatibility system. *Cell Microbiol*. 6(8):707–717.

Monk,A.B., Rees,C.D., Barrow,P., Hagens,S. and Harper,D.R. (2010) Bacteriophage applications: where are we now? *Lett Appl Microbiol*, 51, 363–369.

Morales-Arce AY, Harris RB, Stone AC, Jensen JD. (2020). Evaluating the contributions of purifying selection and progeny-skew in dictating within-host Mycobacterium tuberculosis evolution. *Evolution* 74(5):992–1001.

Morales-Arce AY, Johri P, Jensen JD. (2022). Inferring the distribution of fitness effects in patient-sampled and experimental virus populations: two case studies. *Heredity* 128(2):79–87.

Morris P, Marinelli LJ, Jacobs-Sera D, Hendrix RW, Hatfull GF. (2008). Genomic characterization of mycobacteriophage Giles: evidence for phage acquisition of host DNA by Illegitimate recombination. *J Bacteriol*. 190(6):2172–2182.

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. *Mol. Biol. Evol.* 2021;38:2497–2512. doi: 10.1093/molbev/msab044.

Moye,Z.D., Woolston,J. and Sulakvelidze,A. (2018) Bacteriophage applications for food production and processing. *Viruses*, 10, 205.

Mozzi A, et al..(2020). Past and ongoing adaptation of human cytomegalovirus to its host. *PLoS Pathog*. 16(5):e1008476.

Nei M, Maruyama T. (1975). Letters to the editors: Lewontin-Krakauer test for neutral genes. *Genetics* 80(2):395

Nick,J.A., Dedrick,R.M., Gray,A.L., Vladar,E.K., Smith,B.E., Freeman,K.G., Malcolm,K.C., Epperson,L.E., Hasan,N.A., Hendrix,J., Callahan,K., Walton,K., Vestal,B., Wheeler,E., Rysavy,N.M., Poch,K., Caceres,S., Lovell,V.K., Hisert,K.B., de Moura,V.C., Chatterjee,D., De,P., Weakly,N., Martiniano,S.L., Lynch,D.A., Daley,C.L., Strong,M., Jia,F., Hatfull,G.F. and Davidson,R.M. (2022) Host and pathogen response to bacteriophage engineered against Mycobacterium abscessus lung infection. *Cell*, 185, 1860–1874.

Nishide M, et al. (2019). Antiviral and virucidal activities against herpes simplex viruses of umesu phenolics extracted from Japanese apricot. *Microbiol Immunol.* 63(9):359–366.

Nishiyama Y, Maeno K, Yoshida S. 1983. Characterization of human cytomegalovirus-induced DNA polymerase and the associated 3′-to-5′, exonuclease. *Virology* 124(2):221–231.

Nobrega,F.L., Costa,A.R., Kluskens,L.D. and Azeredo,J. (2015) Revisiting phage therapy: new applications for old resources. *Trends Microbiol*, 23, 185–191.

Numazaki K. 1997. Human cytomegalovirus infection of breast milk. *FEMS Microbiol Immunol.* 18(2):91–98.

Oh,J.H. and Park,M.K. (2017) Recent trends in Salmonella outbreaks and emerging technology for biocontrol of Salmonella using phages in food: a review. *J Microbiol Biotechnol,* 27, 2075–2088.

Omidfar,K. and Daneshpour,M. (2015) Advances in phage display technology for drug discovery. *Expert Opin Drug Discov*, 10, 651–669.

Oppenheim AB, Kobiler O, Stavans J, Court DL, Adhya S. (2005). Switches in bacteriophage lambda development. *Annu Rev Genet*. 39:409–429.

Ormond L, et al..(2017). The combined effect of oseltamivir and favipiravir on influenza A virus evolution. *Genome Biol Evol*. 9(7):1913–1924.

Otsuka Y., Yonesaki T. (2012). Dmd of Bacteriophage T4 functions as an antitoxin against Escherichia coli LsoA and RnlA toxins*. Mol. Microbiol*. 83:669–681. doi: 10.1111/j.1365-2958.2012.07975.x.

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

Pal P., Kumar R. (2014) Treatment of coke wastewater: A critical review for developing sustainable management strategies. *Sep. Purif. Rev.* 43:89–123. doi: 10.1080/15422119.2012.717161.

Pande,J., Szewczyk,M.M. and Grover,A.K. (2010) Phage display: concept, innovations, applications and future. *Biotechnol Adv*, 28, 849–858.

Pang J, et al.(2020). Mixed cytomegalovirus genotypes in HIV-positive mothers show compartmentalization and distinct patterns of transmission to infants. eLife 9:e63199.

Pang X, Humar A, Preiksaitis JK. (2008). Concurrent genotyping and quantitation of cytomegalovirus gB genotypes in solid-organ-transplant recipients by use of a real-time PCR assay. *J Clin Microbiol*. 46(12):4004–4010.

Park,M., Lee,J.-H., Shin,H., Kim,M., Choi,J., Kang,D.-H., Heu,S. and Ryu,S. (2012) Characterization and comparative genomic analysis of a novel bacteriophage, SFP10, simultaneously inhibiting both Salmonella enterica and Escherichia coli O157:H7. *Appl Environ Microbiol,* 78, 58–69.

Peck KM, Lauring AS. (2018). Complexities of viral mutation rates. *J Virol.* 92(14):e01031-17.

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., et al (2003). Origins of highly mosaic mycobacteriophage genomes. *Cell.* 113:171–182. doi: 10.1016/S0092-8674(03)00233-2.

Pedulla ML, Lee MH, Lever DC, Hatfull GF. (1996) A novel host factor for integration of mycobacteriophage L5. *Proc Natl Acad Sci U S A.* 93(26):15411–15416.

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., et al. (2003). Origins of highly mosaic mycobacteriophage genomes. Cell 113(2):171–182.

Peña CEA, Kahlenberg JM, Hatfull GF (1999). Protein-DNA complexes in mycobacteriophage L5 integrative recombination. *J Bacteriol.* 1999;181(2):454–461.

Peña CEA, Lee MH, Pedulla ML, Hatfull GF. (1997). Characterization of the mycobacteriophage L5 attachment site. *J Mol Biol.* 266(1):76–92.

Pénisson S, Singh T, Sniegowski P, Gerrish P. (2017). Dynamics and fate of beneficial mutations under lineage contamination by linked deleterious mutations. *Genetics* 205(3):1305–1318.

Petrovski,S., Seviour,R.J. and Tillett,D. (2011a) Characterization of the genome of the polyvalent lytic bacteriophage GTE2, which has potential for biocontrol of Gordonia-, Rhodococcus-, and Nocardia-stabilized foams in activated sludge plants. *Appl Environ Microbiol,* 77, 3923–3929.

Petrovski,S., Seviour,R.J. and Tillett,D. (2011b) Prevention of Gordonia and Nocardia stabilized foam formation by using bacteriophage GTE7. *Appl Environ Microbiol*, 77, 7864–7867.

Petrovski,S., Tillett,D. and Seviour,R.J. (2012) Genome sequences and characterization of the related Gordonia phages GTE5 and GRU1 and their use as potential biocontrol agents. *Appl Environ Microbiol*, 78, 42–47.

Pfeifer SP, Jensen JD. (2016). The impact of linked selection in chimpanzees: a comparative study. *Genome Biol Evol*. 8(10):3202–3208.

Pfeifer SP. (2020). Spontaneous mutation rates. In: Ho, S.Y.W. (eds), The molecular evolutionary clock. Theory and practice. Cham: Springer Nature. p. 35–44. 10.1007/978-3-030-60181-2_3

Pham TT, Jacobs-Sera D, Pedulla ML, Hendrix RW, Hatfull GF. (2007). Comparative genomic analysis of mycobacteriophage Tweety: evolutionary insights and construction of compatible site-specific integration vectors for mycobacteria. *Microbiology (Reading).* 153(Pt 8):2711–2723.

Plachter B, Sinzger C, Jahn G. (1996). Cell types involved in replication and distribution of human cytomegalovirus. *Adv Virus Res*. 46:195–261.

Pokalyuk C, et al..(2017). Characterizing human cytomegalovirus reinfection in congenitally infected infants: an evolutionary perspective. *Mol Ecol*. 26(7):1980–1990.

Pons,J.C., Paez-Espino,D., Riera,G., Ivanova,N., Kyrpides,N.C. and Llabrés,M. (2021) VPF-Class: taxonomic assignment and host prediction of uncultivated viruses based on viral protein families. *Bioinformatics*, 37, 1805–1813.

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

Pope W.H., Butela K.A., Garlena R.A., Jacobs-Sera D., Russell D.A., Warner M.H., Hatfull G.F., (2020). University of Pittsburgh SEA-PHAGES Genome sequences of 20

bacteriophages isolated on Gordonia terrae. *Microbiol. Resourc. Announc*. 9:e01489-19. doi: 10.1128/MRA.01489-19.

Pope W.H., Mavrich T.N., Garlena R.A., Guerrero-Bustamante C.A., Jacobs-Sera D., Montgomery M.T., Russell D.A., Warner M.H., Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) Hatfull G.F. Bacteriophages of Gordonia spp. display a spectrum of diversity and genetic relationships. mBio. 2017;8:e01069-17. doi: 10.1128/mBio.01069-17.

Pope WH, Jacobs-Sera D, Russell DA, Peebles CL, Al-Atrache Z, Alcoser TA, Alexander LM, Alfano MB, Alford ST, Amy NE, et al. (2011). Expanding the diversity of mycobacteriophages: insights into genome architecture and evolution. *PLoS One.* 6(1):e16329.

Porter KR, Starnes DM, Hamilton JD. (1985). Reactivation of latent murine cytomegalovirus from kidney. *Kidney Int*. 28(6):922–925.

Pouyet F, Aeschbacher S, Thiéry A, Excoffier L.(2018). Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *eLife* 7:e36317.

Puchhammer-Stöckl E, Görzer I. (2011). Human cytomegalovirus: an enormous variety of strains and their possible clinical significance in the human host. *Future Virol.* 6(2):259–271.

Ranallo-Benavidez,T.R., Jaron,K.S. and Schatz,M.C. (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*, 11, 1432.

Rauch B.J., Silvis M.R., Hultquist J.F., Waters C.S., McGregor M.J., Krogan N.J., Bondy-Denomy J. (2017). Inhibition of CRISPR-Cas9 with bacteriophage proteins. *Cell.* 2017;168:150–158.e10. doi: 10.1016/j.cell.2016.12.009.

Renzette N, Bhattacharjee B, Jensen JD, Gibson L, Kowalik TF. (2011). Extensive genome-wide variability of human cytomegalovirus in congenitally infected infants. *PLoS Pathog.* 7(5):e1001344.

Renzette N, et al.. (2013). Rapid intrahost evolution of human cytomegalovirus is shaped by demography and positive selection. *PLoS Genet*. 9(9):e1003735.

Renzette N, et al.. (2015). Limits and patterns of cytomegalovirus genomic diversity in humans. *Proc Natl Acad Sci U S A*. 112(30):E4120–E4128.

Renzette N, Gibson L, Jensen JD, Kowalik TF. (2014). Human cytomegalovirus intrahost evolution—a new avenue for understanding and controlling herpesvirus infections. *Curr Opin Virol.* 8:109–115.

Renzette N, Kowalik TF, Jensen JD. (2016). On the relative roles of background selection and genetic hitchhiking in shaping human cytomegalovirus genetic diversity. *Mol Ecol*. 25(1):403–413.

Renzette N, Pfeifer SP, Matuszewski S, Kowalik TF, Jensen JD.(2017). On the analysis of intrahost and interhost viral populations: human cytomegalovirus as a case study of pitfalls and expectations. *J Virol.* 91(5):1976–1992.

Revello MG, Gerna G. (2010). Human cytomegalovirus tropism for endothelial/epithelial cells: scientific background and clinical implications. Rev Med Virol. 20(3):136–155.

Rohwer F. (2003). Global Phage Diversity. *Cell*. 2003;113:141. doi: 10.1016/S0092-8674(03)00276-9.

Rosenberg N, Nordborg M. (2002). Genealogical trees, coalescent theory, and the analysis of genetic polymorphisms. *Nat Rev Genet.* 3(5):380–390.

Ross A., Ward S., Hyman P. More is better: Selecting for broad host range bacteriophages. *Front. Microbiol*. (2016);7:1352. doi: 10.3389/fmicb.2016.01352.

Rousselle M, Mollion M, Nabholz B, Bataillon T, Galtier N. (2018). Overestimation of the adaptive substitution rate in fluctuating populations. *Biol Lett.* 14(15):20180055.

Sabin S, Morales-Arce AY, Pfeifer SP, Jensen JD. (2022). The impact of frequently neglected model violations on bacterial recombination rate estimation: a case study in Mycobacterium canettii and Mycobacterium tuberculosis. *G3 (Bethesda)* 12(5):jkac055.

Sackman AM, Harris RB, Jensen JD. (2019). Inferring demography and selection in organisms characterized by skewed offspring distributions. *Genetics* 211(3):1019–1028.

Sackman AM, Pfeifer SP, Kowalik TF, Jensen JD. (2018). On the demographic and selective forces shaping patterns of human cytomegalovirus variation within hosts. *Pathogens* 7(1):16.

Sakaoka H, et al..(1994). Quantitative analysis of genomic polymorphism of herpes simplex virus type 1 strains from six countries: studies of molecular evolution and molecular epidemiology of the virus. *J Gen Virol*. 75(Pt 3):513–527.

Sanjuán R. (2010). Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies. *Philos Trans R Soc Lond B Biol Sci.* 365(1548):1975–1982.

Sethi S, Arora S, Gupta V, Kumar S. (2014). Cutaneous Mycobacterium fortuitum infection: successfully treated with Amikacin and Ofloxacin combination. *Indian J Dermatol.* 2014;59(4):383–384.

Shang J., Sun Y. (2022). CHERRY: A Computational MetHod for AccuratE PRediction of Virus-PRokarYotic interactions using a graph encoder-decoder model. *Brief. Bioinform.* 2022:bbac182. doi: 10.1093/bib/bbac182.

Shang,J. and Sun,Y. (2021) Predicting the hosts of prokaryotic viruses using GCN-based semi-supervised learning. *BMC Biol*, 19, 250.

Sharma S, Chatterjee S, Datta S, Prasad R, Dubey D, Prasad RK, Vairale MG. (2017). Bacteriophages and its applications: an overview. *Folia Microbiol.* (Praha). 62(1):17–55.

Sichtig H., Minogue T., Yan Y., Stefan C., Hall A., Tallon L., Sadzewicz L., Nadendla S., Klimke W., Hatcher E., et al. (2019). FDA-ARGOS is a database with public quality-controlled reference genomes for diagnostic use and regulatory science. *Nat. Commun.* 2019;10:3313. doi: 10.1038/s41467-019-11306-6.

Sijmons S, et al.. (2015). High-throughput analysis of human cytomegalovirus genome diversity highlights the widespread occurrence of gene-disrupting mutations and pervasive recombination. *J Virol*. 89(15):7673–7695.

Simmonds P, Aiewsakun P, Katzourakis A. (2018). Prisoners of war — host adaptation and its constraints on virus evolution. *Nature Reviews Microbiology* 2018 17:5. 17:321–328. doi: 10.1038/s41579-018-0120-2.

Singh S, Ghosh P, Hatfull GF. (2008) Attachment site selection and identity in Bxb1 serine integrase-mediated site-specific recombination*. PLoS Genet.* 2013;9(5):e1003490.

Sinzger C, Digel M, Jahn G. (2008). Cytomegalovirus cell tropism. *Curr Top Microbiol Immunol.* 325:63–83.

Söding J, Biegert A, Lupas AN. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 33:W244–W248.

Soffer,N., Woolston,J., Li,M., Das,C. and Sulakvelidze,A. (2017). Bacteriophage preparation lytic for Shigella significantly reduces Shigella sonnei contamination in various foods. *PLoS One*, 12, e0175256.

Song S., Wood T.K. (2020) A primary physiological role of toxin/antitoxin systems is phage inhibition. *Front. Microbiol*. 2020;11:1895. doi: 10.3389/fmicb.2020.01895.

Sowmya P, Madhavan HN. (2009). Analysis of mixed infections by multiple genotypes of human cytomegalovirus in immunocompromised patients. *J Med Virol.* 81(5):861–869.

Spector SA, Hirata KK, Neuman TR. (1984). Identification of multiple cytomegalovirus strains in homosexual men with acquired immunodeficiency syndrome. *J Infect Dis.* 150(6):953–956.

Spielman SJ et al. (2019). Evolution of viral genomes: Interplay between selection, recombination, and other forces. *Methods in Molecular Biology*. 1910:427–468. doi: 10.1007/978-1-4939-9074-0_14/FIGURES/5.

Steinrücken M, Kamm J, Spence JP, Song YS. (2019). Inference of complex population histories using whole-genome sequences from multiple populations. *Proc Natl Acad Sci* U S A. 116(34):17115–17120.

Stern A., Sorek R. (2011) The phage-host arms race: Shaping the evolution of microbes. *Bioessays*. 33:43–51. doi: 10.1002/bies.201000071.

Stumpf MPH, McVean GAT. (2003). Estimating recombination rates from population-genetic data. *Nat Rev Genet.* 4(12):959–968.

Suárez NM, et al.. (2019). Human cytomegalovirus genomes sequenced direction from clinical material: variation, multiple-strain infection, recombination, and gene loss. *J Infect Dis.* 220(5):781–791.

Suárez NM, et al..(2020). Whole-genome approach to assessing human cytomegalovirus dynamics in transplant patients undergoing antiviral therapy. Front Cell *Infect Microbiol.* 10:267.

Sulakvelidze,A., Alavidze,Z. and Morris Jr,J.G. (2001) Bacteriophage therapy. *Antimicrob Agents Chemother*, 45, 649–659.

Swanson EC, Schleiss MR. (2013). Congenital cytomegalovirus infection: new prospects for prevention and therapy: for pediatric clinics of North America: advances in evaluation, diagnosis and treatment of pediatric infectious disease. *Pediatr Clin North Am.* 60(2):335–349.

Szpara ML, Van Doorslaer K. (2021). Mechanisms of DNA Virus Evolution. *Encyclopedia of Virology*. 1–5:71. doi: 10.1016/B978-0-12-809633-8.20993-X.

Taddei F, et al..(1997). Role of mutator alleles in adaptive evolution. *Nature* 387(6634):700–702.

Tellier A, Lemaire C. (2014). Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Mol Ecol.* 23(11):2637–2652.

Travis JMJ, Travis ER. (2002). Mutator dynamics in fluctuating environments. *Philos Trans R Soc Lond B Biol Sci.* 269(1491):591–597.

Unterholzner S.J., Poppenberger B., Rozhon W. (2013) Toxin–antitoxin systems: Biology, identification, and application. *Mob. Genet. Elem*. 2013;3:e26219. doi: 10.4161/mge.26219.

Vaks,L. and Benhar,I. (2011) In vivo characteristics of targeted drug-carrying filamentous bacteriophage nanomedicines. *J Nanobiotechnology*, 9, 58.

Van Damme E, Van Loock M. 2014. Functional annotation of human cytomegalovirus gene products: an update. *Front Microbiol*. 5:218.

Versoza C.J., Howell A.A., Aftab T., Blanco M., Brar A., Chaffee E., Howell N., Leach W., Lobatos J., Luca M., et al. (2022). The complete genome sequence of the Gordonia bacteriophage BiggityBass. *Microbiol. Resour. Announc.*

Versoza C.J., Pfeifer S.P. (2022). Computational prediction of bacteriophage host ranges. *Microorganisms.* 10:149. doi: 10.3390/microorganisms10010149.

Vurture,G.W., Sedlazeck,F.J., Nattestad,M., Underwood,C.J., Fang,H., Gurtowski,J. and Schatz,M.C. (2017) GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, 33, 2202–2204.

Wade,W. (2002) Unculturable bacteria – the uncharacterized organisms that cause oral infections. *J R Soc Med*, 95, 81–83.

Wahl LM, Betti MI, Dick DW, Pattenden T, Puccini AJ. (2019). Evolutionary stability of the lysis-lysogeny decision: Why be virulent? Evolution; international journal of organic evolution. 73:92–98. doi: 10.1111/EVO.13648.

Wakeley J.(2009). Coalescent theory: an introduction. Greenwood Village: Roberts & Company Publishers.

Wang HY, et al.(2021). Common polymorphisms in the glycoproteins of human cytomegalovirus and associated strain-specific immunity. *Viruses* 13(6):1106.

Wang W-K, Lin S-R, Lee C-M, King C-C, Chang S-C. (2002). Dengue type 3 virus in plasma is a population of closely related genomes: quasispecies. *J Virol*. 76(9):4662–4665.

Wang,W., Ren,J., Tang,K., Dart,E., Ignacio-Espinoza,J.C., Fuhrman,J.A., Braun,J., Sun,F. and Ahlgren,N.A. (2020) A network-based integrated framework for predicting virus–prokaryote interactions. *NAR Genom Bioinform*, 2, lqaa044.

Weekes MP, et al.(2014). Quantitative temporal viromics: an active approach to investigate host-pathogen interaction. *Cell* 157:1460–1472.

Wei Y., Gao Z., Zhang H., Dong Y. (2016). Structural characterizations of phage antitoxin Dmd and its interactions with bacterial toxin RnlA. *Biochem. Biophys. Res. Commun.* 472:592–597. doi: 10.1016/j.bbrc.2016.03.025.

Weitzman MD, Lilley CE, Chaurushiya MS. (2010). Genomes in conflict: maintaining genome integrity during virus infection. *Annu Rev Microbiol*. 64:61–81.

Wickham H. (2016). ggplot2: elegant graphics for data analysis. New York (NY): Springer-Verlag; ISBN 978-3-319–24277-4.

Winthrop KL, Roy EE. Mycobacteria and immunosuppression. In: Atzeni F, Galloway JB, Gomez-Reino JJ, Galli M, editors. (2020). *Handbook of Systemic Autoimmune Diseases,* Vol. 16. Elsevier Ltd. p. 83–107.

Xia G, Wolz C. (2014). Phages of Staphylococcus aureus and their impact on host evolution*. Infect Genet Evol* 21:593–601.

Xiaofei E, Kowalik TF. (2014). The DNA damage response induced by infection with human cytomegalovirus and other viruses. *Viruses* 6(5):2155–2185.

Yamaguchi Y., Inouye M. (2011). Regulation of growth and death in Escherichia coli by toxin–antitoxin systems. *Nat. Rev. Microbiol.* 9:779–790. doi: 10.1038/nrmicro2651.

Yan H, et al..(2008). Genetic linkage among human cytomegalovirus glycoprotein N (gN) and gO genes, with evidence for recombination from congenitally and post-natally infected Japanese infants. *J Gen Virol*. 89(Pt 9):2275–2279.

Yatim N, Albert ML. (2011). Dying to replicate: the orchestration of the viral life cycle, cell death pathways, and immunity. Immunity 35(4):478–490.

Yu D., Silva M.C., Shenk T. (2003). Functional map of human cytomegalovirus AD169 defined by global mutational analysis. Proc. Natl. Acad. Sci. USA. 100:12396–12401. doi: 10.1073/pnas.1635160100.

Zárate S, Pond SLK, Shapshak P, Frost SDW. 2007. Comparative study of methods for detecting sequence compartmentalization in human immunodeficiency virus type 1. *J Virol.* 81(12):6643–6651.

Zhang,X., Niu,Y.D., Nan,Y., Stanford,K., Holley,R., McAllister,T. and Narváez-Bravo,C. (2019) SalmoFreshÔ effectiveness in controlling Salmonella on romaine lettuce, mung bean sprouts and seeds. *Int J Food Microbiol*, 305, 108250.

Zheng K, et al..(2014). Epidermal growth factor receptor-PI3K signalling controls cofilin activity to facilitate herpes simplex virus 1 entry into neuronal cells. *mBio* 15(1):e00958-13.

Zielezinski A, Barylski J, Karlowski WM (2021). Taxonomy-aware, sequencing similarity ranking reliably predicts phage-host relationships. *BMC Biol;*19(1):223.

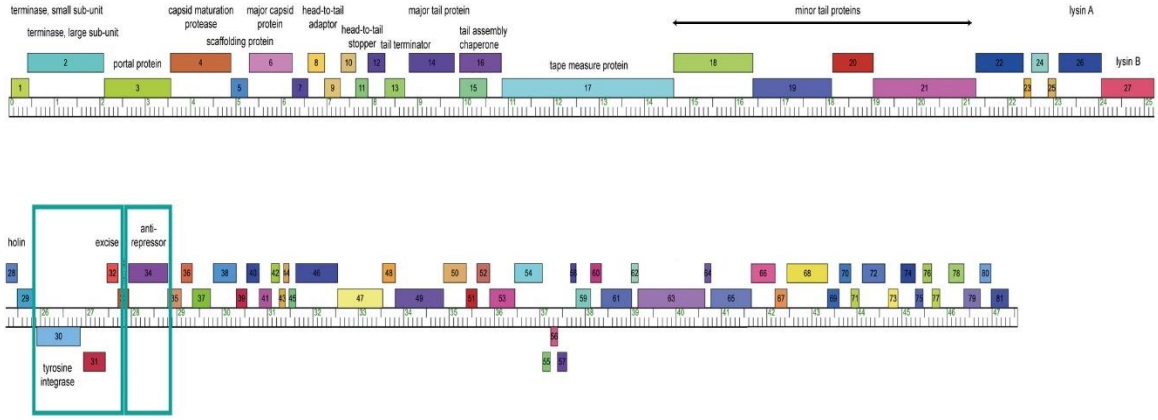
Zielezinski,A., Deorowicz,S. and Gudyś,A. (2022) PHIST: fast and accurate prediction of prokaryotic hosts from metagenomic viral sequences. *Bioinformatics*, 38, 1447–1449.

Zimmermann L., Stephens A., Nam S.-Z., Rau D., Kübler J., Lozajic M., Gabler F., Söding J., Lupas A.N., Alva V. (2018). A completely reimplemented MPI Bioinformatics Toolkit with a new HHpred server at its core. *J. Mol. Biol*. 430:2237–2243. doi: 10.1016/j.jmb.2017.12.007.

Zwart MP, Elena SF. (2015). Matters of size: genetic bottlenecks in virus infection and their potential impact on evolution. *Annu Rev Virol*. 2(1):161–179.

APPENDIX A

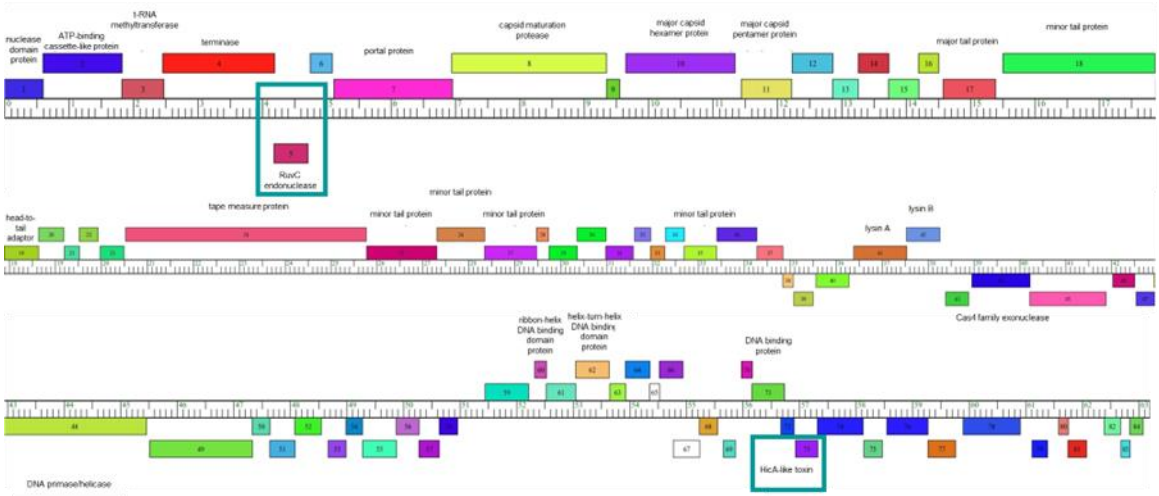
GENOME OF THE CLUSTER P MYCOBACTERIOPHAGE PHEGASUS.



Appendix A. Genome of the cluster P mycobacteriophage Phegasus. Protein-coding genes on the forward or reverse strands with their putative functional assignments (if available) are displayed above or below the ruler, respectively. The integration-dependent immunity system (genes 30 to 32 and 34) is indicated by teal-colored boxes. ssDNA, single-stranded DNA.

APPENDIX B

GENOME OF THE CLUSTER DR BACTERIOPHAGE BIGGITYBASS.



Appendix B. Genome of the cluster DR bacteriophage BiggityBass. Protein-coding genes on the forward or reverse strands with their putative functional assignments (if available) are displayed above or below the ruler, respectively. The RuvC-like resolvase (gene 5) and the *hicA*-like toxin (gene 73) are indicated by teal-colored boxes. ssDNA, single-stranded DNA.

APPENDIX C

PERMISSION FROM CO-AUTHORS

The chapter titled “Phylogenomic analyses and host range prediction of cluster P mycobacteriophages” was published in 2022 in *G3*. The paper had 20 contributing authors. Abigail A. Howell was a co first author. The original publication can be found at: https://academic.oup.com/g3journal/article/12/11/jkac244/6696222. The corresponding author, Pfeifer, S.P. has consented for the publication to be included in this dissertation by Abigail A. Howell.

The chapter titled “Comparative Genomics of Closely-Related Gordonia Cluster DR Bacteriophages” was was published in 2022 in *Viruses*. The paper had 20 contributing authors. Abigail A. Howell was a co first author. The original publication can be found at: https://academic.oup.com/g3journal/article/12/11/jkac244/6696222. The corresponding author, Pfeifer, S.P. has consented for the publication to be included in this dissertation by Abigail A. Howell.

The chapter titled “Developing an Appropriate Evolutionary Baseline Model for the Study of Human Cytomegalovirus” was published in 2023 in *Genome Biology and Evolution.* The paper had 6 contributing authors. Abigail A. Howell was the first author. The corresponding author, Pfeifer, S.P. has consented for the publication to be included in this dissertation by Abigail A. Howell.