VIRUS EVOLUTION

Virus Evolution, 2024, **10(1)**, 1–11
DOI: https://doi.org/10.1093/ve/vead083
Advance Access Publication 20 December 2023
Research Article

Computational host range prediction—The good, the bad, and the ugly

Abigail A. Howell,[†] Cyril J. Versoza,[†] and Susanne P. Pfeifer[*,‡]

Center for Evolution and Medicine, School of Life Sciences, Arizona State University, Tempe, AZ 85281, USA
[†]These authors contributed equally to the project.
[‡]https://orcid.org/0000-0003-1378-2913
*Corresponding author: E-mail: susanne@spfeiferlab.org

Abstract

The rapid emergence and spread of antimicrobial resistance across the globe have 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 sixteen broad-spectrum (polyvalent) bacteriophages with experimentally validated host ranges, we here benchmark the performance of eleven 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 per cent) could be reached at the strain-level, albeit at low levels of precision (<40 per cent). 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.

Keywords: bacteriophage; virus; host prediction; bioinformatics; genomics.

1. Introduction

Due to the rise of antimicrobial resistance—projected to lead to an estimated 10 million deaths per year (Furfaro, Payne, and Chang 2018) and an economic loss of up to US$100 trillion by 2050 across the globe (according to projections resulting from a high-burden-of-resistance model, which considered drug resistance to *Escherichia coli*, *Klebsiella pneumoniae*, and *Staphylococcus aureus* infections as well as HIV, malaria, and tuberculosis; O'Neill 2016)—bacteriophages (i.e. viruses that infect, and replicate within, bacteria) are now being routinely used in a wide variety of fields as an 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, McLean, and Palombo 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, Woolston, and Sulakvelidze 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, Seviour, and Tillett 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, Alavidze, and Morris 2001; Nobrega et al. 2015), bacteriophage display to discover antibodies, peptides, or proteins that bind to, e.g. cancer cells (Pande, Szewczyk, and Grover 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, Dedrick, and Schooley 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 regard to growth media, temperature, pH, and UV light—which may not fully capture the complexity of natural environments, in particular the organs frequently targeted by bacteriophage therapy (human lungs and gastrointestinal system). Moreover, culturing bacteriophages and performing host assays remain laborious, time-consuming, and expensive processes, thus limiting their potential for scalable high-throughput screening (Wade 2002; Edwards and Rohwer 2005; Coutinho, Edwards, and Rodríguez-Valera 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 two categories: (1) 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 and (2) alignment-free methods based on sequence composition such as oligonucleotide or *k*-mer (i.e. nucleotide sequences of length *k*) frequencies that may result, e.g. 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, to predict bacteriophage host ranges. In addition to methods based on single features, machine-/deep-learning-based methods trained on experimentally validated datasets of bacteriophage-host interactions have been used 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). More recently, such machine learning frameworks have also been utilized to predict the host taxonomy of uncultivated viruses infecting archaea and bacteria from high-throughput metagenomics data (e.g. iPHoP [Roux et al. 2023]).

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; Shang and Sun 2021, 2022; Amgarten et al. 2022; 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 thirteen polyvalent *Gordonia* bacteriophages with experimentally validated host ranges, we here assess the performance of eleven computational host range prediction tools and discuss important factors to consider when implementing these computational methods.

2. Materials and methods

2.1 Experimental data

Computational host range prediction tools were evaluated using three polyvalent *E. coli* bacteriophages—HY01 (Lee et al. 2016), KFS-EC3 (Kim, Adeyemi, and Park 2021), and SFP10 (Park et al. 2012)—as well as thirteen polyvalent *Gordonia* bacteriophages—GTE2 (Petrovski, Seviour, and Tillett 2011a), GTE7 (Petrovski, Seviour, and Tillett 2011b), GTE5 and GRU1 (Petrovski, Tillett, and Seviour 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, the host range of bacteriophage HY01 was previously determined by cultivating bacterial strains at 37°C in four different growth media: Lysogeny Broth, Tryptic Soy Broth, Brain Heart Infusion, and de Man-Rogosa-Sharpe (for additional details, see Lee et al. 2016). Similarly, the host range of bacteriophages KFS-EC3 and SFP10 was established by cultivating bacterial strains at 37°C in Tryptic Soy Broth and Luria-Bertani Broth, respectively (as described by Kim, Adeyemi, and Park 2021 and Park et al. 2012, respectively). For *Gordonia* bacteriophages GMA2–GMA7, GRU1, GRU3, GTE2, and GTE6–GTE8, bacterial strains were grown on PYCa liquid media at 30°C (for experimental details, see Petrovski, Seviour, and Tillett 2011a,b; Petrovski, Tillett, and Seviour 2012; Dyson et al. 2015). All bacteriophage-host pairs were tested using a drop spot assay with agar plates of their respective media.

Genome assemblies for all bacteriophages were downloaded from NCBI (using the accession numbers provided in Supplementary Tables S1 and S2). Publicly available 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.

2.1.1 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).

2.1.2 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, Jaron, and Schatz 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.

2.2 Computational host range prediction

Computational host range prediction tools can be divided into two groups: (1) confirmatory methods that utilize a set of bacterial genomes provided by the user to infer the likelihood of a bacteriophage-host interaction and (2) 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, Barylski, and Karlowski 2021), PHIST v.1.1 (Zielezinski, Deorowicz, and Gudyś 2022), Prokaryotic virus Host Predictor (PHP) v.1.0 (Lu et al. 2021), VirHostMatcher (VHM) 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 (VHMN) v.1.0 (Wang et al. 2020), and VPF-Class v.1.0 (Pons et al. 2021). For the confirmatory tools (Phirbo, PHIST, PHP, VHM, and WIsH), performance was evaluated in terms of sensitivity ($TP/(TP+FN)$, with TP being the number of true positives and FN the number of false negatives), specificity ($TN/(TN+FP)$, with TN being the number of true negatives and FP the number of false positives), accuracy ($(TP+TN)/(TP+TN+FP+FN)$), and precision ($TP/(TP+FP)$) based on the experimentally validated host and non-host bacterial strains for which genome assemblies were available (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 study 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 species 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 bacteriophages known to infect host species included in this study. In contrast, exploratory tools predict bacteriophage-host interactions based on inbuilt databases either at the species-level (CHERRY and VHMN) or at the 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).

2.3 Comparative genomic analyses

Pairwise average nucleotide identities (ANIs) between (1) the three *E. coli* bacteriophages HY01, KFS-EC3, and SFP10, as well as the thirteen *Gordonia* bacteriophages GMA2-7, GRU1, GRU3, GTE2, and GTE5-8 (Supplementary Fig. S1) and (2) the experimentally validated host and non-host genomes as well as genomes of closely related bacterial strains included in the exploratory tool databases (Supplementary Figs. 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 Fig. S4).

3. Results and discussion

The performance of eleven computational host prediction tools—CHERRY (Shang and Sun 2022), HostG (Shang and Sun 2021), Phirbo (Zielezinski, Barylski, and Karlowski 2021), PHIST (Zielezinski, Deorowicz, and Gudyś 2022), PHP (Lu et al. 2021), RaFAH (Coutinho et al. 2021), vHULK (Amgarten et al. 2022), VHM (Ahlgren et al. 2017), VHMN (Wang et al. 2020), VPF-Class (Pons et al. 2021), and WIsH (Galiez et al. 2017)—was evaluated using three polyvalent *E. coli* bacteriophages and thirteen polyvalent *Gordonia* bacteriophages for which host ranges were previously experimentally validated (for details, see Supplementary Tables S1 and S2).

3.1 Confirmatory tools

The five confirmatory tools—Phirbo (Zielezinski, Barylski, and Karlowski 2021), PHIST (Zielezinski, Deorowicz, and Gudyś 2022), PHP (Lu et al. 2021), VHM (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 sixteen 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 a 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 per cent–67.8 per cent (Supplementary Table S4). Highly conserved single-copy orthologous actinobacteria genes (BUSCOs) demonstrated that these *Gordonia* assemblies are nearly complete, containing between 98.0 per cent (*Gordonia rubripertincta* DSM 43197) and 99.4 per cent (*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 per cent) (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, VHM's background-subtracting d_2^* similarity measure yielded a much lower sensitivity (12.9 per cent); only WIsH's stringent 8-mer approach exhibited a lower recall (0.0 per cent), identifying none of the genuine host strains of the sixteen polyvalent bacteriophages. At the same time, the usage of longer *k*-mers also increased specificity, from 55.3 per cent in PHP to 83.5 per cent and 90.6 per cent in WIsH and VHM, respectively. Notably, none of the predictions of VHM and WIsH passed the recommended tool-specific thresholds for any of the *E. coli* and *Gordonia* bacteriophages, respectively (Fig. 1). More generally, fewer results were observed for *Gordonia* bacteriophages, with PHP and VHM 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 ANIs between the bacteriophages, see Supplementary Fig. S1), likely due to the fact that *E. coli* is a more widely studied model organism than *Gordonia*.

In contrast to PHP and VHM, 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 species 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 per cent) and accuracy (31.6 per cent) 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 per cent, specificity: 43.8 per cent, precision: 0.0 per cent, and accuracy: 36.8 per cent), 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 per cent sensitivity and 88.2 per cent 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 per cent 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 fourteen bacteriophages were obligatorily lytic, thus leaving no genetic trace in

Table 1. Performance of computational host range prediction tools. Performance of the confirmatory tools Phirbo, PHP, VHM, and WIsH as well as the species-level exploratory tools CHERRY and VHMN and the genus-level exploratory tools HostG, RaFAH, 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 were 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, Adeyemi, and Park 2021). Additional details about predicted bacteriophage-host interactions that passed recommended tool-specific thresholds are provided in Supplementary Tables S5, S6, and S8).

		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))[a]: 1442)	**77.4%**	55.3%	**38.7%**	61.2%
		VHM (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[b]: 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[c]: 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%

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

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 Fig. 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 et al. (2016) who utilized a set of bacteriophages with known isolation hosts to demonstrate that alignment-free methods (such as PHP, VHM, 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 per cent (PHP and WIsH) to 69.8 per cent (Phirbo and VHM)—similar to the level of accuracy previously observed for these tools (~20 per cent–60 per cent prediction accuracy at the genus-level for alignment-based methods [Edwards et al. 2016; Ahlgren et al. 2017; Zielezinski, Barylski, and Karlowski 2021] and ~30 per cent–70 per cent 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 per cent for WIsH (which did not identify any true positives) to 33.3 per cent, 37.5 per cent, and 38.7 per cent for VHM, 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 Figs. S1–S3). Notably, most genuine hosts were only identified by a single tool, PHP, with a limited number identified by multiple tools (Fig. 2).

3.2 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), several of 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), VHMN (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: (1) 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 (2) VHMN—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 VHM'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 VHMN in terms of sensitivity (47.6 per cent vs 10.0 per cent), precision (60.6 per cent vs 28.6 per cent), and accuracy (93.6 per cent vs 91.7 per cent) at a similar level of specificity (97.4 per cent vs 98.1 per cent) (Table 1, and see Supplementary Tables S5 and S8).

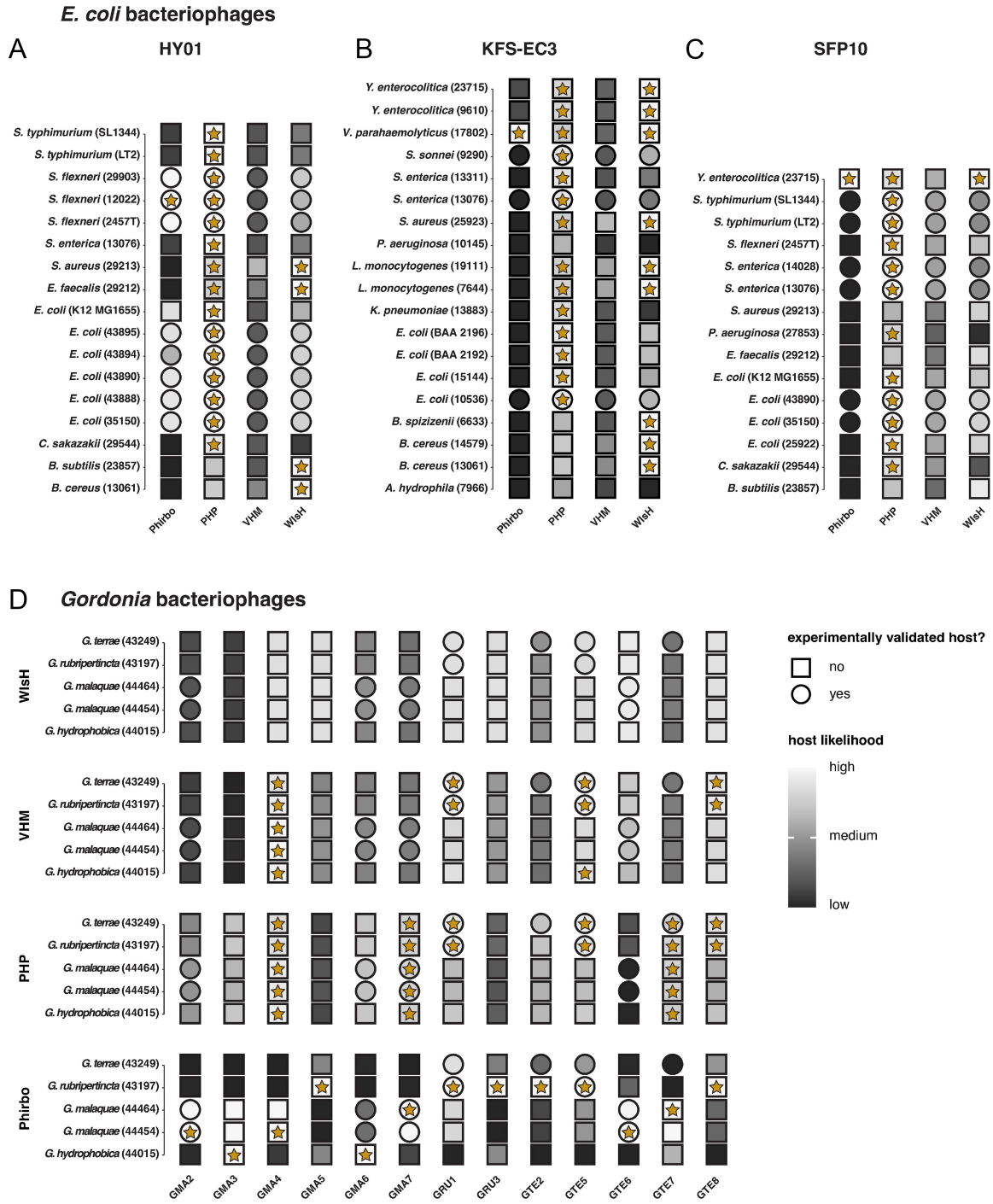

Figure 1. Computational host predictions for three *E. coli* bacteriophages—(A) HY01, (B) KFS-EC3, and (C) SFP10—and (D) thirteen *Gordonia* bacteriophages—GMA2-7, GRU1, GRU3, GTE2, and GTE5-8—for a set of experimentally validated host and non-host strains (Supplementary Tables S1 and S2) obtained using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

The remaining four exploratory tools predict hosts at the genus-level: (1) 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, (2) 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, (3) 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 database (Grazziotin, Koonin, and Kristensen 2017) to infer the host genus, and (4) 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 from the IMG/VR database (Paez-Espino et al. 2016). At the genus-level, RaFAH

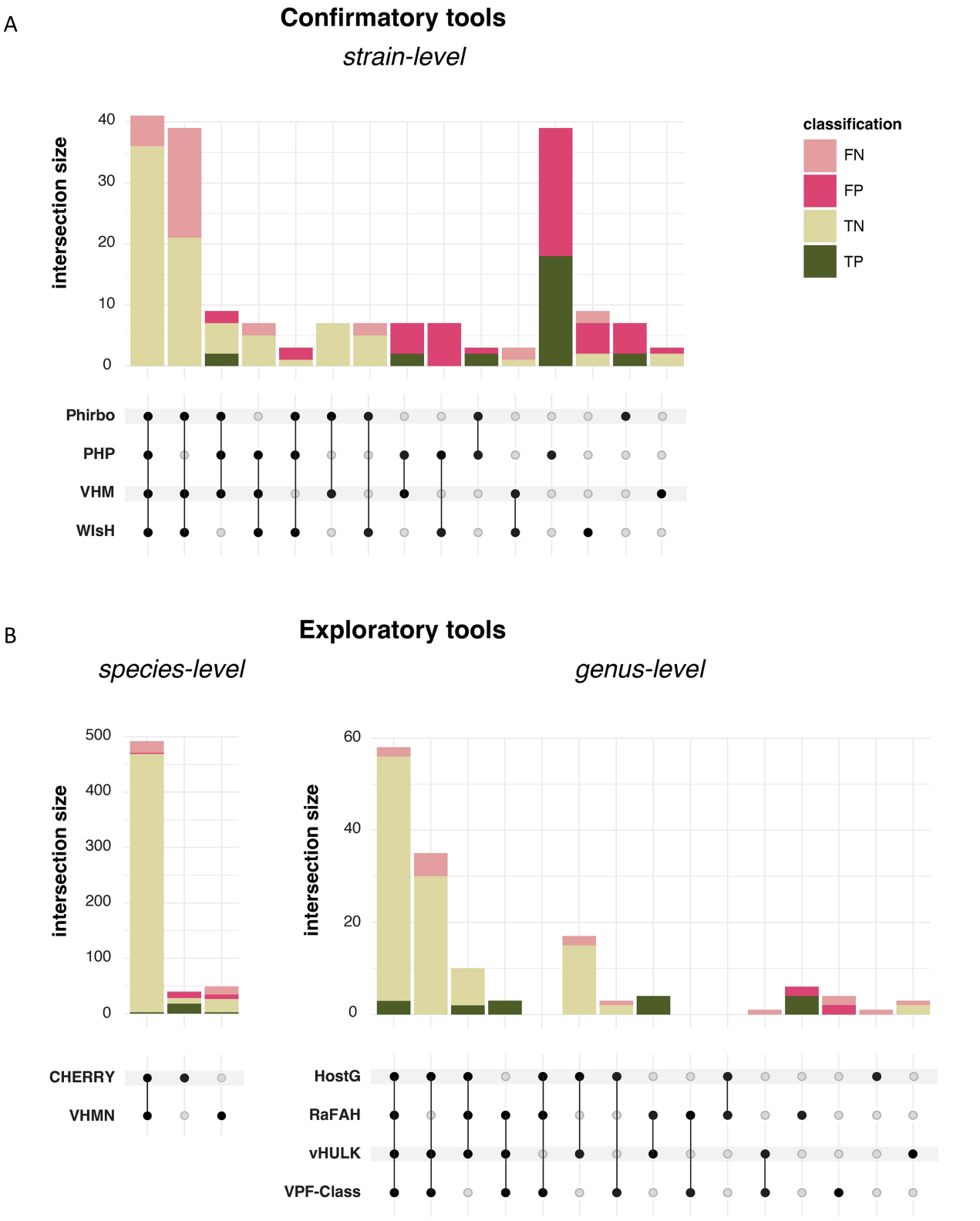

Figure 2. Performance of eleven computational host range prediction tools based on experimentally validated bacteriophage-host interactions. Each column in the upset plot corresponds to an intersection set of true positives (TP; shown in green), true negatives (TN; olive), false positives (FP; pink), and false negatives (FN; rose) between sets of host range prediction tools. Rows beneath each barplot correspond to the tools, with full circles connected by black lines displaying the sets that are being compared in a particular column. (A) The confirmatory tools Phirbo, PHP, VHM, and WIsH utilize a set of provided bacterial genomes to infer the likelihood of strain-specific bacteriophage-host interactions. (B) Exploratory tools predict bacteriophage-host interactions based on an internal database of putative host genomes either at the species-level (CHERRY and VHMN) or genus-level (HostG, RaFAH, vHULK, and VPF-Class).

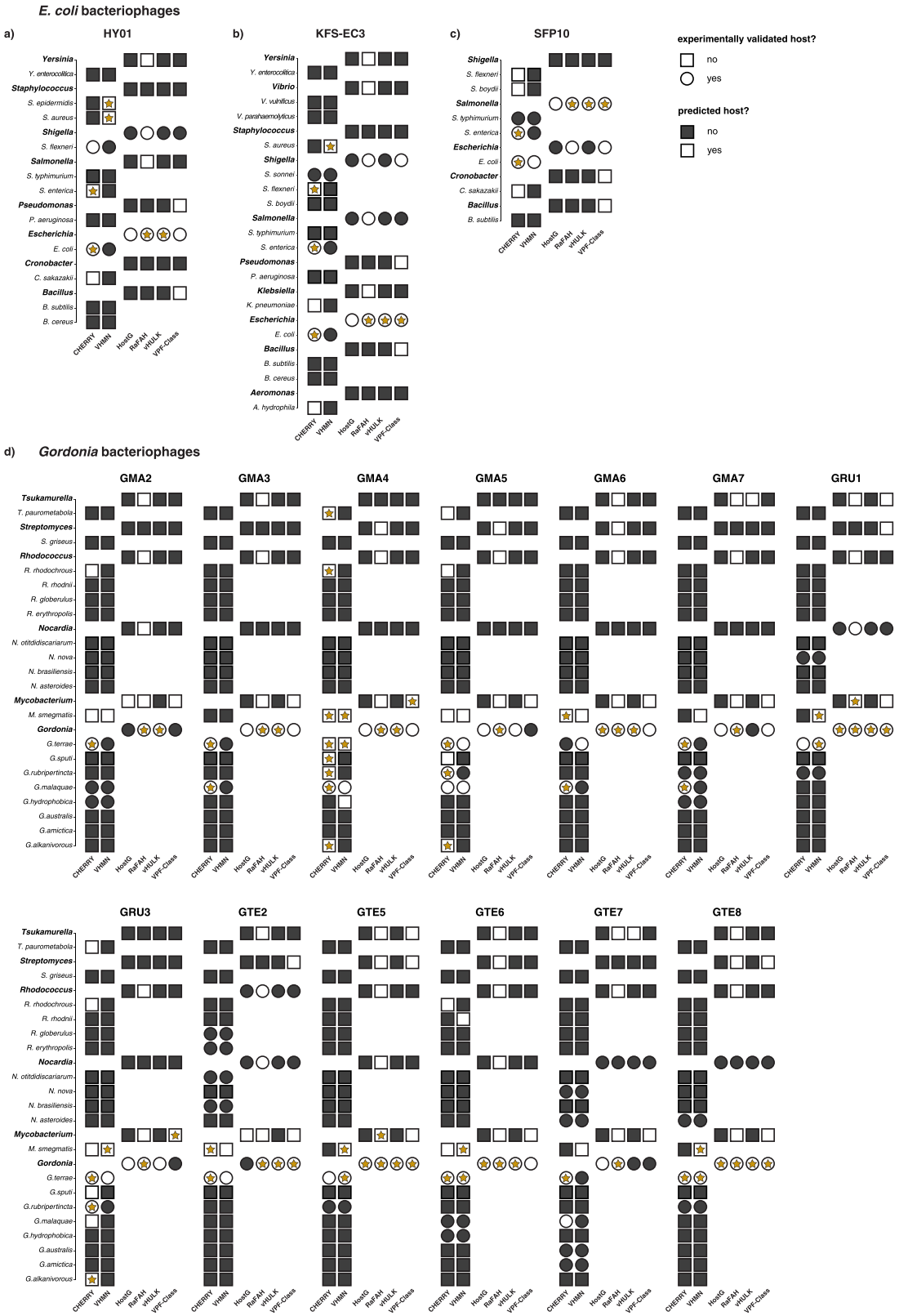

Figure 3. Computational host predictions for three *E. coli* bacteriophages—(A) HY01, (B) KFS-EC3, and (C). SFP10—and (D) thirteen *Gordonia* bacteriophages—GMA2-7, GRU1, GRU3, GTE2, and GTE5-8—for a set of experimentally validated host and non-host strains (Supplementary Tables S1 and S2 as well as Tables 1 in Park et al. 2012; Dyson et al. 2015; Lee et al. 2016; Kim, Adeyemi, and Park 2021) obtained using the species-level exploratory tools CHERRY and VHMN as well as the genus-level exploratory tools HostG, RaFAH, 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.

exhibited the highest recall (88.9 per cent) and accuracy (95.1 per cent) (Table 1)—higher than the ∼60 per cent genus-level accuracy previously reported (see Fig. 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 thirteen *Gordonia* bacteriophages (Fig. 3). In comparison, HostG, vHULK, and VPF-Class showed a sensitivity ranging from 31.3 per cent (HostG) to 52.2 per cent (vHULK) and an accuracy ranging from 87.6 per cent (VPF-Class)—similar to the 86.4 per cent genus-level accuracy reported by the developers (see Table 5 in Pons et al. 2021)—to 91.7 per cent (vHULK). However, RaFAH's sensitivity came at a cost of a slightly worse specificity (RaFAH: 96.9 per cent; VPF-Class: 97.7 per cent; HostG: 100.0 per cent; vHULK: 100.0 per cent). Moreover, both HostG and vHULK were more precise (100 per cent each) than RaFAH (88.9 per cent) and VPF-Class (75.0 per cent). Similar to the confirmatory tools, few genuine hosts were identified by multiple species-level exploratory tools (Fig. 2).

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

4. 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 eleven 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 of high specificity in order to aid in the prioritization of experimental efforts to identify the most suitable bacteriophage(s) for any given application. Finally, it is important to acknowledge that the study presented here is based on host susceptibility data for polyvalent bacteriophages that were obtained at a single experimental condition. As host susceptibility may depend on the laboratory conditions, future work will need to focus on the comprehensive characterization of bacteriophage growth rates and fitness

assays under a variety of experimental conditions (particularly with regard to growth media and incubation temperatures) to further improve *in silico* host range prediction.

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 with the accession numbers provided in Supplementary Tables S1 and S2 (bacteriophage and *E. coli* assemblies) and under BioProject PRJNA1021557 (*de novo* assemblies of *Gordonia* strains). Analysis scripts are available at https://github.com/PfeiferLab/host_range_prediction.

Supplementary data

Supplementary data is available at *VEVOLU* Journal online.

Funding

This work was supported by a National Science Foundation CAREER award to SPP [grant number DEB-2045343].

Acknowledgements

DNA isolation was performed at the Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures GmbH, library preparation and sequencing was performed at the Cold Spring Harbor Laboratory Genome Center, and computations were performed at Arizona State University's High Performance Computing facility.

Conflict of interest: None declared.

References

Ahlgren, N. A. et al. (2017) 'Alignment-free d_2^* Oligonucleotide Frequency Dissimilarity Measure Improves Prediction of Hosts from Metagenomically-derived Viral Sequences', *Nucleic Acids Research*, 45: 39–53.

Amgarten, D. et al. (2022) 'vHULK, a New Tool for Bacteriophage Host Prediction Based on Annotated Genomic Features and Neural Networks', *PHAGE (New Rochelle, N.Y.)*, 3: 204–12.

Arndt, D. et al. (2016) 'PHASTER: A Better, Faster Version of the PHAST Phage Search Tool', *Nucleic Acids Research*, 44: W16–W21.

Baláž, A. et al. (2023) 'PHERI – Phage Host ExploRation Pipeline', *Microorganisms*, 11: 1398.

Carbone, A. (2008) 'Codon Bias Is a Major Factor Explaining Phage Evolution in Translationally Biased Hosts', *Journal of Molecular Evolution*, 66: 210–23.

——— (2022) 'CHERRY: A Computational Method for Accurate Prediction of Virus–prokaryotic Interactions Using a Graph Encoder–decoder Model', *Briefings in Bioinformatics*, 23: bbac182.

Coclet, C., and Roux, S. (2021) 'Global Overview and Major Challenges of Host Prediction Methods for Uncultivated Phages', *Current Opinion in Virology*, 49: 117–26.

Coutinho, F. H. et al. (2021) 'RaFAH: Host Prediction for Viruses of Bacteria and Archaea Based on Protein Content', *Patterns (NY)*, 2: 100274.

Coutinho, F. H., Edwards, R. A., and Rodríguez-Valera, F. (2019) 'Charting the Diversity of Uncultured Viruses of Archaea and Bacteria', *BMC Biology*, 17: 1–16.

Dedrick, R. M. et al. (2021) 'Potent Antibody-mediated Neutralization Limits Bacteriophage Treatment of a Pulmonary *Mycobacterium Abscessus* Infection', *Nature Medicine*, 27: 1357–61.

Dedrick, R. M. et al. (2023) 'Phage Therapy of *Mycobacterium* Infections: Compassionate Use of Phages in 20 Patients with Drug-Resistant Mycobacterial Disease', *Clinical Infectious Diseases*, 76: 103–12.

Dyson, Z. A. et al. (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. et al. (2016) 'Computational Approaches to Predict Bacteriophage-host Relationships', *FEMS Microbiology Reviews.*, 40: 258–72.

Edwards, R. A., and Rohwer, F. (2005) 'Viral Metagenomics', *Nature Reviews, Microbiology*, 3: 504–10.

Eren, A. M. et al. (2015) 'Anvi'o: An Advanced Analysis and Visualization Platform for 'Omics Data', *PeerJ*, 3: e1319.

Furfaro, L. L., Payne, M. S., and Chang, B. J. (2018) 'Bacteriophage Therapy: Clinical Trials and Regulatory Hurdles', *Frontiers in Cellular & Infection Microbiology*, 8: 376.

Galiez, C. et al. (2017) 'WIsH: Who Is the Host? Predicting Prokaryotic Hosts from Metagenomic Phage Contigs', *Bioinformatics*, 33: 3113–4.

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 Research*, 45: D491–D498.

Harada, L. K. et al. (2018) 'Biotechnological Applications of Bacteriophages: State of the Art', *Microbiological Research*, 212-213: 38–58.

Hatfull, G. F., Dedrick, R. M., and Schooley, R. T. (2022) 'Phage Therapy for Antibiotic-resistant Bacterial Infections', *Annual Review of Medicine*, 73: 197–211.

Inglis, L. K., and Edwards, R. A. (2022) 'How Metagenomics Has Transformed Our Understanding of Bacteriophages in Microbiome Research', *Microorganisms*, 10: 1671.

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.

Kolmogorov, M. et al. (2019) 'Assembly of Long, Error-prone Reads Using Repeat Graphs', *Nature Biotechnology*, 37: 540–6.

Koskella, B., and Brockhurst, M. A. (2014) 'Bacteria-phage Coevolution as a Driver of Ecological and Evolutionary Processes in Microbial Communities', *FEMS Microbiology Reviews.*, 38: 916–31.

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', *Biological Control*, 165: 104817.

Lee, H. et al. (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', *Journal of Open Source Software*, 4: 1236.

Little, J. S. et al. (2022) 'Bacteriophage Treatment of Disseminated Cutaneous *Mycobacterium Chelonae* Infection', *Nature Communications*, 13: 2313.

López-Cuevas, O. et al. (2021) 'Bacteriophage Applications for Fresh Produce Food Safety', *International Journal of Environmental Health Research*, 31: 687–702.

Lu, C. et al. (2021) 'Prokaryotic Virus Host Predictor: A Gaussian Model for Host Prediction of Prokaryotic Viruses in Metagenomics', *BMC Biology*, 19: 5.

Makarova, K. S. et al. (2020) 'Evolutionary Classification of CRISPR-Cas Systems: A Burst of Class 2 and Derived Variants', *Nature Reviews, Microbiology*, 18: 67–83.

Manni, M. et al. (2021) 'BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes', *Molecular Biology and Evolution*, 38: 4647–54.

Marçais, G., and Kingsford, C. (2011) 'A Fast, Lock-free Approach for Efficient Parallel Counting of Occurrences of *K*-mers', *Bioinformatics*, 27: 764–70.

Monk, A. B. et al. (2010) 'Bacteriophage Applications: Where are We Now?', *Letters in Applied Microbiology*, 51: 363–9.

Moye, Z. D., Woolston, J., and Sulakvelidze, A. (2018) 'Bacteriophage Applications for Food Production and Processing', *Viruses*, 10: 205.

Nick, J. A. et al. (2022) 'Host and Pathogen Response to Bacteriophage Engineered against *Mycobacterium Abscessus* Lung Infection', *Cell*, 185: 1860–74.

Nobrega, F. L. et al. (2015) 'Revisiting Phage Therapy: New Applications for Old Resources', *Trends in Microbiology*, 23: 185–91.

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', *Journal of Microbiology & Biotechnology*, 27: 2075–88.

Omidfar, K., and Daneshpour, M. (2015) 'Advances in Phage Display Technology for Drug Discovery', *Expert Opinion on Drug Discovery*, 10: 651–69.

O'Neill, J. (2016) 'Tackling Drug-resistant Infections Globally: Final Report and Recommendations', Government of the United Kingdom and Wellcome Trust: London, UK.

Paez-Espino, D. et al. (2016) 'Uncovering Earth's Virome', *Nature*, 536: 425–30.

Pande, J., Szewczyk, M. M., and Grover, A. K. (2010) 'Phage Display: Concept, Innovations, Applications and Future', *Biotechnology Advances*, 28: 849–58.

Park, M. et al. (2012) 'Characterization and Comparative Genomic Analysis of a Novel Bacteriophage, SFP10, Simultaneously Inhibiting Both *Salmonella Enterica* and *Escherichia Coli* O157:H7', *Applied and Environmental Microbiology*, 78: 58–69.

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', *Applied and Environmental Microbiology*, 77: 3923–9.

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', *Applied and Environmental Microbiology*, 78: 42–7.

Pons, J. C. et al. (2021) 'VPF-Class: Taxonomic Assignment and Host Prediction of Uncultivated Viruses Based on Viral Protein Families', *Bioinformatics*, 37: 1805–13.

——— (2011b) 'Prevention of *Gordonia* and *Nocardia* Stabilized Foam Formation by Using Bacteriophage GTE7', *Applied and Environmental Microbiology*, 77: 7864–7.

Ranallo-Benavidez, T. R., Jaron, K. S., and Schatz, M. C. (2020) 'GenomeScope 2.0 And Smudgeplot for Reference-free Profiling of Polyploid Genomes', *Nature Communications*, 11: 1432.

Roux, S. et al. (2023) 'iPHoP: An Integrated Machine Learning Framework to Maximize Host Prediction for Metagenome-derived Viruses of Archaea and Bacteria', *PLoS Biology.*, 21: e3002083.

Shang, J., and Sun, Y. (2021) 'Predicting the Hosts of Prokaryotic Viruses Using GCN-based Semi-supervised Learning', *BMC Biology*, 19: 250.

Soffer, N. et al. (2017) 'Bacteriophage Preparation Lytic for *Shigella* Significantly Reduces *Shigella Sonnei* Contamination in Various Foods', *PLoS One*, 12: e0175256.

Sulakvelidze, A., Alavidze, Z., and Morris, J. G., Jr (2001) 'Bacteriophage Therapy', *Antimicrobial Agents and Chemotherapy*, 45: 649–59.

Vaks, L., and Benhar, I. (2011) '*In Vivo* Characteristics of Targeted Drug-carrying Filamentous Bacteriophage Nanomedicines', *Journal of Nanobiotechnology*, 9: 58.

Versoza, C. J., and Pfeifer, S. P. (2022) 'Computational Prediction of Bacteriophage Host Ranges', *Microorganisms*, 10: 149.

Vurture, G. W. et al. (2017) 'GenomeScope: Fast Reference-free Genome Profiling from Short Reads', *Bioinformatics*, 33: 2202–4.

Wade, W. (2002) 'Unculturable Bacteria – the Uncharacterized Organisms that Cause Oral Infections', *Journal of the Royal Society of Medicine*, 95: 81–3.

Wang, W. et al. (2020) 'A Network-based Integrated Framework for Predicting Virus–prokaryote Interactions', *NAR Genomics and Bioinformatics*, 2: lqaa044.

Zhang, X. et al. (2019) 'SalmoFresh™ Effectiveness in Controlling *Salmonella* on Romaine Lettuce, Mung Bean Sprouts and Seeds', *International Journal of Food Microbiology*, 305: 108250.

Zielezinski, A., Barylski, J., and Karlowski, W. M. (2021) 'Taxonomy-aware, Sequence Similarity Ranking Reliably Predicts Phage–host Relationships', *BMC Biology*, 19: 223.

Zielezinski, A., Deorowicz, S., and Gudyś, A. (2022) 'PHIST: Fast and Accurate Prediction of Prokaryotic Hosts from Metagenomic Viral Sequences', *Bioinformatics*, 38: 1447–9.

Virus Evolution, 2024, **10(1)**, 1–11
DOI: https://doi.org/10.1093/ve/vead083
Advance Access Publication 20 December 2023
Research Article