*microorganisms*

MDPI

*Review*

# Computational Prediction of Bacteriophage Host Ranges

**Cyril J. Versoza [1] and Susanne P. Pfeifer [2,*]**

[1] Center for Evolution and Medicine, School of Life Sciences, Arizona State University, Tempe, AZ 85281, USA; cversoza@asu.edu
[2] Center for Mechanisms of Evolution, School of Life Sciences, Arizona State University, Tempe, AZ 85281, USA
* Correspondence: susanne.pfeifer@asu.edu

**Abstract:** Increased antibiotic resistance has prompted the development of bacteriophage agents for a multitude of applications in agriculture, biotechnology, and medicine. A key factor in the choice of agents for these applications is the host range of a bacteriophage, i.e., the bacterial genera, species, and strains a bacteriophage is able to infect. Although experimental explorations of host ranges remain the gold standard, such investigations are inherently limited to a small number of viruses and bacteria amendable to cultivation. Here, we review recently developed bioinformatic tools that offer a promising and high-throughput alternative by computationally predicting the putative host ranges of bacteriophages, including those challenging to grow in laboratory environments.

**Keywords:** bacteriophages; bacterial hosts; bioinformatic tools; host ranges



**Citation:** Versoza, C.J.; Pfeifer, S.P. Computational Prediction of Bacteriophage Host Ranges. *Microorganisms* **2022**, *10*, 149. https://doi.org/10.3390/microorganisms10010149

Academic Editor: Igor V. Babkin

Received: 17 December 2021
Accepted: 11 January 2022
Published: 12 January 2022



## 1. Introduction

There are approximately $10^{31}$ viruses on earth [1]—more than stars in the observable universe. The vast majority of this diverse virosphere consists of bacteriophages, i.e., viruses that infect and prey on bacteria. Independently discovered by Frederick William Twort and Félix d'Herelle in the early 1900s [2,3], these abundant biological entities have since been routinely used for a multitude of purposes—ranging from diagnostics [4], to drug design and discovery [5,6], to vaccine development [7], to agriculture [8], to food preservation and safety [9], and to wastewater treatment [10].

In order to leverage the bactericidal effects of bacteriophages for these applications, bacterial host ranges (i.e., collections of bacterial species and strains that support the life cycle of the bacteriophage) need to be established. Several experimental techniques allow for the study of bacteriophage–host relationships (such as spot, plaque, and liquid assays, viral tagging, microfluidic PCR, phageFISH, and single-cell genomics [11]). However, they are often time- and labor-intensive, costly, and can be scientifically challenging (e.g., due to inconclusive or absent signs of infection [12]). These approaches are also inherently limited in scope due to both the bacterial cultures used in the experiments—with a limited number of microbial hosts [13,14] and viruses [15,16] being amendable to cultivation—as well as the conditions under which they are performed in the laboratory (such as growth media and temperature [17]).

Recent advances in sequencing technologies have enabled the discovery and identification of bacteriophages and their hosts from environmental (rather than cultivated) samples, thus providing an important avenue to comprehensively study the natural viral diversity [18,19]. In concert with these technical advances, many bioinformatic approaches have been developed to computationally predict putative bacteriophage host ranges at large scale, based on genomic features shared between bacteriophages and their bacterial hosts through their co-evolution over time. Although predictive by their nature, these tools can highlight the most promising candidates for subsequent experimental work to validate the bacteriophage's ability to identify and adsorb to the host, as well as to characterize infection cycles, bacteriophage–host interactions, and lysis efficacy.

In this review, we provide an overview of several available computational host prediction methods, discuss similarities and differences in their design, and provide key considerations when choosing between different approaches.

## 2. Methods to Computationally Predict Bacteriophage Host Ranges

Bioinformatic approaches to computationally predict putative bacteriophage host ranges can be broadly classified into three categories: (i) alignment-based methods based on sequence homology and sequence similarity, (ii) alignment-free methods based on sequence composition and genomic features, and (iii) machine-learning-based methods.

### *2.1. Alignment-Based Methods*

Many factors can impact bacteriophage host specificity. Temperate bacteriophages can integrate their own genomes into that of their bacterial hosts as lysogenic prophages [20]. This process often alters the phenotype of the host, which can lead to an increased fitness (e.g., by providing antibiotic resistance, increasing virulence, producing toxins, or preventing further (super)infections; see review by Touchon and colleagues [21]). At the same time, many bacterial hosts guard themselves against virulent bacteriophages and other invaders by employing a variety of restriction-modification (RM) and clustered regularly interspaced short palindromic repeats (CRISPRs)/Cas (CRISPR-associated protein) strategies [22,23]. In the latter case, a stretch of nucleotides from the invasive genetic material is incorporated into a CRISPR spacer array upon infection (adaptation), and this new spacer is used as a guide to create site-specific cleavages, ultimately leading to the degradation of the invading bacteriophage (immunity) [24]. In both scenarios, the host genome is ultimately altered by, or due to, the invading bacteriophage.

Alignment-based methods rely on these host–virus shared sequences to computationally predict host ranges from sequence homology (i.e., the common evolutionary ancestry between sequences) and sequence similarity. Many alignment-based methods—including the most prominent example, the Basic Local Alignment Search Tool (BLAST [25])—are straightforward to use, for example by comparing a user-provided viral sequence with those of putative bacterial hosts publicly available in well-maintained (reference) databases. Consequently, the inference of virus–host relationships through alignment-based methods is limited by the comprehensiveness and completeness of the used databases. On the one hand, sequences of bacteriophages that infect a single host not yet present in a database might yield no results; on the other hand, sequences of bacteriophages that exhibit a broad host range might yield multiple results, often ranked by some user-defined criteria (e.g., the overall length of similar sequence) to improve manual/visual dissemination. However, such rankings can also introduce challenges: (i) rankings may change depending on the criteria and thresholds used, (ii) the highest ranked result may not be the most prevalent host (or, in fact, it may not be a host at all), (iii) mosaic bacteriophage genomes may point towards several (equally well supported) related hosts, and (iv) comparable results may arise between distantly related viruses and bacterial species due to spurious alignments or other artifacts.

To circumvent some of these issues, Zielezinski and colleagues [26] developed a computational tool, Phirbo, that exploits the full range of BLAST results. Phirbo works under the assumption that the similarity between a pair of bacteriophage and host sequences is proportional to the overlap between their independent BLAST searches against the same dataset. Specifically, Phirbo generates two ranked lists from two independent BLAST searches—one using a bacteriophage-reference dataset and one using a host-reference dataset—and compares them using the Ranked-Biased Overlap metric [27], a procedure that has been shown to improve precision compared to several other state-of-the-art host prediction tools [26].

### *2.2. Alignment-Free Methods*

Viral and host sequences may lack sequence homology, making them less well-suited for alignment-based methods. In these cases, alignment-free methods offer a promising alternative to infer bacteriophage–host relationships by studying the similarity in patterns of sequence composition, such as codon usage or oligonucleotide (short nucleotide fragment) frequency [11]. Such similarities in patterns of sequence composition are expected from first principles. For example, viruses frequently corrupt the translational machinery of their hosts to synthesize their own viral proteins [28], and this synthesis is generally more efficient if the codon usage patterns of the virus matches that of its host [29,30]. Taking advantage of this relationship, Crane, Versoza and colleagues [31] determined the codon usage bias of 129 mycobacteriophages across 14 putative mycobacterial hosts using COUSIN [32] to obtain important insights into putative mycobacterial host ranges in nature. Bacteriophage genomes can also acquire molecular characteristics of their hosts due to exposure to similar genome-wide mutational pressures, a process referred to as 'genome amelioration' [33–35]. By matching the nucleotide composition of their hosts, bacteriophages are able to avoid host RM systems that recognize specific tetranucleotides [36].

Alignment-free, sequence composition-dependent tools can be categorized by whether the genome-wide signature of a viral sequence is compared to (i) a database of potential hosts (virus–host similarity), or (ii) a database of viruses with known hosts (virus-virus similarity). Examples of the first category include VirHostMatcher [37] which calculates virus–host similarity by comparing oligonucleotide frequencies between the viral sequence and those of potential hosts, and WIsH [38] which calculates virus–host similarity in terms of differences in frequencies of oligonucleotides of a specified length (so-called '*k*-mers'). In contrast, HostPhinder [39], an example of the second category, uses virus-virus similarity measures, assuming that similar oligonucleotide usage between viruses indicates shared or closely related hosts.

### *2.3. Machine-Learning Methods*

In addition to alignment-based and alignment-free methods, machine-learning (ML) approaches have found a home in bacteriophage research in general [40] and in the prediction of bacteriophage–host interactions specifically [41]. In order to infer virus–host relationships, ML approaches utilize 'features', i.e., measurable properties of the object being analyzed such as the nucleotide and amino acid content of the viral genome, amino acid properties, and protein domains (see [42] for a comparison of feature representations). For example, both the Host Taxon Predictor (HTP) [43] and the Prokaryotic virus Host Predictor (PHP) [44] tools use nucleotide features to predict bacteriophage–host interactions, with HTP representing the bacteriophage sequence using absolute and relative frequencies of oligonucleotides as well as nucleic acid types, and PHP using a Gaussian model to predict hosts based on the oligonucleotide frequency differences between viral and host genome sequences. In contrast, PredPHI (Predicting Phage–Host Interactions [45]) identifies putative bacteriophage hosts using a mix of amino acid frequency, chemical composition, and molecular weight as feature representations. Similarly, VirHostMatcher-Net [46] integrates multiple features, including virus-virus similarity, virus–host alignment-free similarity, virus–host alignment-based similarity, and virus–host CRISPR-based similarity, to predict virus–host interactions. BacteriophageHostPrediction [41] uses more than 200 features—ranging from genomic sequences (such as nucleotide and codon frequencies and GC-content), to protein sequences (such as amino acid frequency), to protein secondary structure (such as $\alpha$-helix and $\beta$-sheet frequencies), and to physicochemical properties (such as molecular weight and isoelectric point)–to represent receptor-binding proteins which play a crucial role in determining host specificity by recognizing receptors on the surface of the bacterial host [47]. At a higher level of sequence representation, PHERI [48] infers bacterial hosts from bacteriophage sequences through annotated protein sequence clusters.

## 3. Bacteriophage–Host Databases

Experimental evidence through bacteriophage isolation and cultivation remains, whenever possible, the gold standard in determining bacteriophage host ranges. However, experimental validation is often time- and labor-intensive. For example, nearly half a decade passed between the initial prediction and concrete experimental evidence that crAssphage–a highly abundant bacteriophage in the human gut microbiome–can infect bacteria of the genus *Bacteroides* [49,50]. As a consequence, information regarding bacteriophage–host relationships remains sparse, with information deposited in the well-established National Center for Biotechnology Information (NCBI) RefSeq and GenBank databases often being either restricted to the genus and/or species level or limited to a handful of samples [51]. The recently developed Viral Host Range database (VHRdb [52])–a web-based tool that integrates host range data as an analysis tool and search engine–aims to collect additional data by allowing researchers to directly share their experimental findings with the scientific community (at the time of writing, 16,715 interactions between 760 viruses and 1923 hosts have been recorded). Given the need of validated training datasets, bacteriophage–host databases such as VHRdb are expected to play a significant role in the development of future ML methods.

## 4. Method Choice: Key Considerations

### *4.1. Prediction Accuracy*

Apart from their underlying algorithms, bacteriophage–host prediction tools also differ in their prediction accuracy, i.e., the percentage of bacteriophages for which the taxonomy of their predicted and known hosts agree [46]. Prediction accuracy can be reported at different taxonomic levels—ranging from the family, genus, and species levels down to the phylum and domain levels. It is thus important to consider which taxonomic levels were measured when selecting the most appropriate tool for any analysis. Methodological differences (such as the type of data included in the benchmarking process) can further contribute to differences in prediction accuracy between tools. Hence, comparisons should ideally be performed using a uniform benchmarking dataset. Using such uniform benchmarking data, Zielezinski and colleagues [26] performed a comparison between a variety of alignment-based, alignment-free, and ML-based host-range prediction tools, demonstrating that tools based on sequence homology generally have a higher predictive accuracy than those reliant on sequence composition similarity (see their Tables 1 and 2).

A challenge faced by researchers working with environmental samples is the non-uniform abundance of microbial species present in a metagenomic sample. As sequencing technologies are optimized for moderate- to high-coverage individual samples, metagenomic samples often result in different read coverage profiles across different genomes [53]. Due to these differences, contigs (a gapless stretch of nucleotide sequence generated by overlapping sequencing reads [54]) obtained from metagenomic samples are frequently short, resulting in genome assemblies that are fragmented and/or incomplete [55]. This is a non-negligible factor in the prediction accuracy of most tools, with short viral contigs (<10 kb) generally experiencing a significant drop in prediction accuracy [26,37,44]. A notable exception in this regard is the tool WIsH, which matches VirHostMatcher's full-length genome prediction accuracy with merely 3 kb of nucleotide sequence, thus establishing itself as an alignment-free alternative for samples containing short viral contigs.

### *4.2. Usability*

Operating system restrictions can be an important aspect in the choice of a suitable bacteriophage–host prediction tool. In order to facilitate both automation and reproducibility, the majority of prediction tools rely on the command line interface (CLI) embedded within UNIX-based operating systems (such as Linux and macOS) (see Table 1). Consequently, users of other operating systems (such as Windows and Chrome OS) will need to either purchase a dedicated machine or install the necessary operating system on an available machine, for example via dual boot or a virtual machine. Windows users can also

leverage the Windows Subsystem for Linux (WSL) to allow native Linux programs to run on Windows.

**Table 1.** Computational methods in predicting bacteriophage host ranges.

| | Prediction Tool | Input | Output | User Interface | Key Considerations | Reference |
|---|---|---|---|---|---|---|
| **Alignment-based** | Phirbo | two ranked lists (phage and host genomes) | phage–host predictions | CLI (Python) | Linux and macOS multi-threading support | [26] |
| **Alignment-free** | HostPhinder | phage FASTA file | predicted hosts | web-based | not limited to any OS | [39] |
| | VirHostMatcher | phage FASTA file host FASTA file taxonomy text file | dissimilarity index phage–host pairs | CLI (Python) | Linux, macOS, Windows | [37] |
| | WIsH | phage FASTA file host FASTA file | predicted hosts | CLI (C++) | Linux and macOS multi-threading support | [38] |
| **Machine-learning-based** | Bacteriophage-HostPrediction | phage FASTA file | predicted hosts | CLI (Python) | Linux and macOS | [41] |
| | Host Taxon Predictor (HTP) | phage FASTA file | Predicted host lineages | CLI (Python) | Linux and macOS | [43] |
| | Prokaryotic virus Host Predictor (PHP) | phage FASTA file | predicted hosts | CLI (Python); web-based | Linux and macOS user-defined training models | [44] |
| | PredPHI | protein sequences (phage–host pairs) | phage-host predictions | CLI (Python) | Linux and macOS | [45] |
| | PHERI | phage FASTA file | predicted hosts predicted shared genes protein sequence clusters | CLI (Python) | Linux and macOS | [48] |
| | VirHostMatcher-Net | phage FASTA file | predicted hosts | CLI (Python) | Linux and macOS multi-threading support | [46] |

CLI, command line interface; OS, operating system; FASTA file, text file representing nucleotide or amino acid sequences.

Web-based prediction tools (such as HostPhinder and PHP) offer a valuable alternative. Apart from being user-friendly and intuitive, web-based tools avoid the inconvenience of installation and potential dependency issues, as their only requirement is a compatible browser. However, a major drawback of web-based tools is their cap on input data. For example, while the web-based version of PHP is limited to <100 viruses, the stand-alone version can analyze datasets that are orders of magnitude larger [44]. An additional advantage of many phage–host prediction CLI tools (including Phirbo, WIsH, and VirHostMatcher-Net) is multi-threading which increases the speed of the analyses.

## 5. Conclusions

Due to their bactericidal effects, bacteriophages are now routinely used for a multitude of biotechnological and clinical purposes, including personalized phage therapy to treat multi-drug resistant infections [56]. Although large-scale bacteriophage banks (such as the Phage Directory [57]) offer a broad range of bacteriophages to the scientific community, the host range that a bacteriophage can infect must be known in order to effectively guide the usage of bacteriophages in these disciplines. Traditional methods to experimentally characterize host ranges–phage isolation and cultivation–remain the gold standard. However, they are time-intensive and thus ill-suited for large-scale analyses. Recently developed computational prediction tools offer a promising alternative, allowing researchers to narrow down the sheer quantity of potential hosts to a limited set that can feasibly (and more cost-efficiently) be tested in a laboratory setting. As tools employ different strategies to predict bacteriophage–host relationships–each with their own advantages and disadvan-

tages, the use of multiple, complementary prediction tools can help to select the most promising candidates, especially for bacteriophages with large host ranges. For example, if time and computational resources permit, a three-way combination of alignment-based, alignment-free, and ML approaches may be used to select those that have been predicted by all three strategies for experimental validation as well as characterization of infection cycles and bacteriophage–host interactions. Although a vast diversity of bacteriophages and bacterial hosts remain to be discovered, advances in genomic databases, machine learning, and high-performance computing have begun to pave the way towards even more sophisticated and accurate computational methods in the near future.

**Author Contributions:** C.J.V. and S.P.P. wrote this review. All authors have read and agreed to the published version of the manuscript.

**Funding:** S.P.P. is supported by a National Science Foundation CAREER grant (DEB-2045343).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Rohwer, F. Global Phage Diversity. *Cell* **2003**, *113*, 141. [CrossRef]
2. Twort, F. An investigation on the nature of ultra-microscopic viruses. *Lancet* **1915**, *186*, 1241–1243. [CrossRef]
3. d'Herelle, F. Sur un microbe invisible antagoniste des bacilles dysentériques. *C. R. Acad. Sci. Paris* **1917**, *165*, 373–375.
4. Schofield, D.; Sharp, N.J.; Westwater, C. Phage-based platforms for the clinical detection of human bacterial pathogens. *Bacteriophage* **2012**, *2*, 105–121. [CrossRef] [PubMed]
5. Molek, P.; Strukelj, B.; Bratkovic, T. Peptide Phage Display as a Tool for Drug Discovery: Targeting Membrane Receptors. *Molecules* **2011**, *16*, 857–887. [CrossRef]
6. Nixon, A.E.; Sexton, D.J.; Ladner, R.C. Drugs derived from phage display: From candidate identification to clinical practice. *mAbs* **2014**, *6*, 73–85. [CrossRef] [PubMed]
7. Bao, Q.; Li, X.; Han, G.; Zhu, Y.; Mao, C.; Yang, M. Phage-based vaccines. *Adv. Drug Deliv. Rev.* **2019**, *145*, 40–56. [CrossRef]
8. Buttimer, C.; McAuliffe, O.; Ross, R.P.; Hill, C.; O'Mahony, J.; Coffey, A. Bacteriophages and Bacterial Plant Diseases. *Front. Microbiol.* **2017**, *8*, 34. [CrossRef]
9. Fenton, M.; McAuliffe, O.; O'Mahony, J.; Coffey, A. Recombinant bacteriophage lysins as antibacterials. *Bioeng. Bugs* **2010**, *1*, 9–16. [CrossRef]
10. Jassim, S.A.A.; Limoges, R.G.; El-Cheikh, H. Bacteriophage biocontrol in wastewater treatment. *World J. Microbiol. Biotechnol.* **2016**, *32*, 1–10. [CrossRef]
11. Edwards, R.; McNair, K.; Faust, K.; Raes, J.; Dutilh, B.E. Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol. Rev.* **2016**, *40*, 258–272. [CrossRef]
12. Hanna, L.F.; Matthews, T.D.; Dinsdale, E.A.; Hasty, D.; Edwards, R.A. Characterization of the ELPhiS Prophage from Salmonella enterica Serovar Enteritidis Strain LK5. *Appl. Environ. Microbiol.* **2012**, *78*, 1785–1793. [CrossRef] [PubMed]
13. Wade, W. Unculturable bacteria—the uncharacterized organisms that cause oral infections. *J. R. Soc. Med.* **2002**, *95*, 81–83.
14. Edwards, R.A.; Rohwer, F. Viral metagenomics. *Nat. Rev. Microbiol.* **2005**, *3*, 504–510. [CrossRef] [PubMed]
15. Breitbart, M.; Salamon, P.; Andresen, B.; Mahaffy, J.M.; Segall, A.M.; Mead, D.; Azam, F.; Rohwer, F. Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 14250–14255. [CrossRef]
16. Coutinho, F.H.; Edwards, R.A.; Rodríguez-Valera, F. Charting the diversity of uncultured viruses of Archaea and Bacteria. *BMC Biol.* **2019**, *17*, 1–16. [CrossRef]
17. Clokie, M.R.J.; Kropinski, A. (Eds.) *Bacteriophages: Methods and Protocols. Volume 1: Isolation, Characterization, and Interactions*, 1st ed.; Humana Press: Totowa, NJ, USA, 2009.
18. Zhang, T.; Breitbart, M.; Lee, W.H.; Run, J.-Q.; Wei, C.L.; Soh, S.W.L.; Hibberd, M.; Liu, E.T.; Rohwer, F.; Ruan, Y. RNA Viral Community in Human Feces: Prevalence of Plant Pathogenic Viruses. *PLoS Biol.* **2005**, *4*, e3. [CrossRef] [PubMed]
19. Reyes, A.; Haynes, M.; Hanson, N.; Angly, F.E.; Heath, A.C.; Rohwer, F.; Gordon, J.I. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **2010**, *466*, 334–338. [CrossRef] [PubMed]
20. Monteiro, R.; Pires, D.P.; Costa, A.R.; Azeredo, J. Phage Therapy: Going Temperate? *Trends Microbiol.* **2019**, *27*, 368–378. [CrossRef] [PubMed]
21. Touchon, M.; Sousa, J.A.M.D.; Rocha, E.P. Embracing the enemy: The diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. *Curr. Opin. Microbiol.* **2017**, *38*, 66–73. [CrossRef]

22. Barrangou, R.; Fremaux, C.; Deveau, H.; Richards, M.; Boyaval, P.; Moineau, S.; Romero, D.A.; Horvath, P. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* **2007**, *315*, 1709–1712. [CrossRef]
23. Jiang, F.; Doudna, J.A. CRISPR–Cas9 structures and mechanisms. *Annu. Rev. Biophys.* **2017**, *46*, 505–529. [CrossRef]
24. Marraffini, L.A. CRISPR-Cas immunity in prokaryotes. *Nature* **2015**, *526*, 55–61. [CrossRef]
25. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
26. Zielezinski, A.; Barylski, J.; Karlowski, W.M. Taxonomy-aware, sequence similarity ranking reliably predicts phage–host relationships. *BMC Biol.* **2021**, *19*, 1–14. [CrossRef]
27. Webber, W.; Moffat, A.; Zobel, J. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.* **2010**, *28*, 1–38. [CrossRef]
28. Rampersad, S.; Tennant, P. Chapter 3—Replication and expression strategies of viruses. In *Viruses*; Tennant, P., Fermin, G., Foster, J.E., Eds.; Academic Press: Cambridge, MA, USA, 2018; pp. 55–82.
29. Kunisawa, T.; Kanaya, S.; Kutter, E. Comparison of Synonymous Codon Distribution Patterns of Bacteriophage and Host Genomes. *DNA Res.* **1998**, *5*, 319–326. [CrossRef]
30. Lucks, J.B.; Nelson, D.R.; Kudla, G.R.; Plotkin, J.B. Genome Landscapes and Bacteriophage Codon Usage. *PLoS Comput. Biol.* **2008**, *4*, e1000001. [CrossRef] [PubMed]
31. Crane, A.; Versoza, C.J.; Hua, T.; Kapoor, R.; Lloyd, L.; Mehta, R.; Menolascino, J.; Morais, A.; Munig, S.; Patel, Z.; et al. Phylogenetic relationships and codon usage bias amongst cluster K mycobacteriophages. *G3 Genes Genomes Genet.* **2021**, *11*, 291. [CrossRef]
32. Bourret, J.; Alizon, S.; Bravo, I.G. COUSIN (COdon Usage Similarity INdex): A Normalized Measure of Codon Usage Preferences. *Genome Biol. Evol.* **2019**, *11*, 3523–3528. [CrossRef] [PubMed]
33. Lawrence, J.G.; Ochman, H. Amelioration of Bacterial Genomes: Rates of Change and Exchange. *J. Mol. Evol.* **1997**, *44*, 383–397. [CrossRef]
34. Pride, D.T.; Wassenaar, T.M.; Ghose, C.; Blaser, M.J. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. *BMC Genom.* **2006**, *7*, 8. [CrossRef] [PubMed]
35. Marri, P.R.; Golding, G.B. Gene amelioration demonstrated: The journey of nascent genes in bacteria. *Genome* **2008**, *51*, 164–168. [CrossRef]
36. Pride, D.T.; Meinersmann, R.J.; Wassenaar, T.; Blaser, M.J. Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases. *Genome Res.* **2003**, *13*, 145–158. [CrossRef]
37. Ahlgren, N.A.; Ren, J.; Lu, Y.Y.; Fuhrman, J.; Sun, F. Alignment-free d2* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomical-ly-derived viral sequences. *Nucleic Acids Res.* **2017**, *45*, 39–53. [CrossRef]
38. Galiez, C.; Siebert, M.; Enault, F.; Vincent, J.; Söding, J. WIsH: Who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* **2017**, *33*, 3113–3114. [CrossRef]
39. Villarroel, J.; Kleinheinz, K.A.; Jurtz, V.I.; Zschach, H.; Lund, O.; Nielsen, M.; Larsen, M.V. Host Phinder: A Phage Host Prediction Tool. *Viruses* **2016**, *8*, 116. [CrossRef] [PubMed]
40. Nami, Y.; Imeni, N.; Panahi, B. Application of machine learning in bacteriophage research. *BMC Microbiol.* **2021**, *21*, 1–8. [CrossRef] [PubMed]
41. Boeckaerts, D.; Stock, M.; Criel, B.; Gerstmans, H.; De Baets, B.; Briers, Y. Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. *Sci. Rep.* **2021**, *11*, 1–14. [CrossRef]
42. Young, F.; Rogers, S.; Robertson, D.L. Predicting host taxonomic information from viral genomes: A comparison of feature representations. *PLoS Comput. Biol.* **2020**, *16*, e1007894. [CrossRef] [PubMed]
43. Gałan, W.; Bak, M.; Jakubowska, M. Host Taxon Predictor—A Tool for Predicting Taxon of the Host of a Newly Discovered Virus. *Sci. Rep.* **2019**, *9*, 3436. [CrossRef]
44. Lu, C.; Zhang, Z.; Cai, Z.; Zhu, Z.; Qiu, Y.; Wu, A.; Jiang, T.; Zheng, H.; Peng, Y. Prokaryotic virus host predictor: A Gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC Biol.* **2021**, *19*, 1–11. [CrossRef] [PubMed]
45. Li, M.; Wang, Y.; Li, F.; Zhao, Y.; Liu, M.; Zhang, S.; Bin, Y.; Smith, A.I.; Webb, G.I.; Li, J.; et al. A Deep Learning-Based Method for Identification of Bacteriophage-Host Interaction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *18*, 1801–1810. [CrossRef] [PubMed]
46. Wang, W.; Ren, J.; Tang, K.; Dart, E.; Ignacio-Espinoza, J.C.; Fuhrman, J.A.; Braun, J.; Sun, F.; Ahlgren, N.A. A network-based integrated framework for predicting virus–prokaryote interactions. *NAR Genom. Bioinform.* **2020**, *2*, lqaa044. [CrossRef]
47. Dams, D.; Brøndsted, L.; Drulis-Kawa, Z.; Briers, Y. Engineering of receptor-binding proteins in bacteriophages and phage tail-like bacteriocins. *Biochem. Soc. Trans.* **2019**, *47*, 449–460. [CrossRef]
48. Baláž, A.; Kajsík, M.; Budiš, J.; Szemeš, T.; Turňa, J. PHERI-Phage Host Exploration Pipeline. *bioRxiv* **2020**. [CrossRef]
49. Dutilh, B.E.; Cassman, N.; McNair, K.; Sanchez, S.E.; Silva, G.G.Z.; Boling, L.; Barr, J.; Speth, D.; Seguritan, V.; Aziz, R.; et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **2014**, *5*, 4498. [CrossRef]
50. Shkoporov, A.N.; Khokhlova, E.V.; Fitzgerald, C.B.; Stockdale, S.R.; Draper, L.A.; Ross, R.P.; Hill, C. ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nat. Commun.* **2018**, *9*, 1–8. [CrossRef] [PubMed]

51. Sayers, E.W.; Beck, J.; Bolton, E.E.; Bourexis, D.; Brister, J.R.; Canese, K.; Comeau, D.C.; Funk, K.; Kim, S.; Klimke, W.; et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2021**, *49*, D10–D17. [CrossRef]
52. Lamy-Besnier, Q.; Brancotte, B.; Brancotte, H.M.; Ménager, L.D. Viral Host Range database, an online tool for recording, analyzing and disseminating virus–host interactions. *Bioinformatics* **2021**, *37*, 2798. [CrossRef]
53. Lapidus, A.L.; Korobeynikov, A.I. Metagenomic data assembly–the way of decoding unknown microorganisms. *Front. Microbiol.* **2021**, *12*, 653. [CrossRef] [PubMed]
54. Staden, R. A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Res.* **1980**, *8*, 3673–3694. [CrossRef]
55. Wooley, J.C.; Godzik, A.; Friedberg, I. A Primer on Metagenomics. *PLoS Comput. Biol.* **2010**, *6*, e1000667. [CrossRef] [PubMed]
56. Pirnay, J.-P. Phage Therapy in the Year 2035. *Front. Microbiol.* **2020**, *11*, 1171. [CrossRef]
57. Sacher, J.; Zheng, J.; McCallin, S. Sourcing phages for compassionate use. *Microbiol. Aust.* **2019**, *40*, 24. [CrossRef]