Title

Global overview and major challenges of host prediction methods for uncultivated phages



Authors

Coclet, Clément

Roux, Simon

Publication Date

2021-08-01

DOI

10.1016/j.coviro.2021.05.003

Peer reviewed

# Global Overview And Major Challenges Of Host Prediction Methods For Uncultivated Phages

Clément Coclet[1], Simon Roux[1*]

[1] DOE Joint Genome Institute, Berkeley, CA, USA
[*] Correspondence to: sroux@lbl.gov

## Highlights

- Global phage diversity is now primarily explored through cultivation-independent metagenomics
- Metagenome-derived phages are typically linked to their host(s) via *in silico* predictions
- Multiple alignment-dependent and alignment-free methods for host predictions have been proposed
- Recent integrative approaches combining several methods into a single prediction seem most promising
- Eventually, complementary *in silico* predictions and *in vitro* assays will enable the reconstruction of entire phage-host networks

## Abstract

Bacterial communities play critical roles across all of Earth's biomes, affecting human health and global ecosystem functioning. They do so under strong constraints exerted by viruses, i.e., bacteriophages or "phages". Phages can reshape bacterial communities' structure, influence long-term evolution of bacterial populations, and alter host cell metabolism during infection. Metagenomics approaches, i.e., shotgun sequencing of environmental DNA or RNA, recently enabled large-scale exploration of phage genomic diversity, yielding several millions of phage genomes now to be further analyzed and characterized. One major challenge however is the lack of direct host information for these phages. Several methods and tools have been proposed to bioinformatically predict the potential host(s) of uncultivated phages based only on genome sequence information. Here we review these different approaches and highlight their distinct strengths and limitations. We also outline complementary experimental assays which are being proposed to validate and refine these bioinformatic predictions.

## Introduction

Most bacterial cells can be infected by bacteriophages ("phages"), that constitute the overwhelming majority of the global virosphere. Laboratory cultivation of a virus-host pair was until recently the primary approach used to explore viral diversity and establish viral taxonomy, and it remains the only way to comprehensively characterize a virus infection cycle and molecular interactions with its host cell [1,2]. However, culturing phages can be experimentally challenging, especially since phages may require specific conditions to grow [3], most bacteria have not yet been cultured leading to a limitation in host availability [4], and some phages, e.g. the ones establishing a lysogenic (i.e., latent) infection cycle, may be challenging to observe and detect [5,6].

Recent development in next generation sequencing (NGS) techniques has uncoupled virus discovery from virus isolation. With the advent of metagenomics, genetic material from microbes and viruses can be sequenced directly from a sample regardless of cultivability, which tremendously accelerated the discovery of novel viruses [7–10]. Metagenomic studies revealed that environmental phage sequences were frequently unrelated to isolate genomes, highlighting the huge diversity of the virosphere, while also providing unprecedented insights into the potential roles, dynamics, and host interactions of these uncultivated phages [11–13]. These analyses come with a fundamental disadvantage however: while cultured viruses are inherently associated to one of their host(s), i.e., the one on which the virus was isolated from, this is not the case for phages identified only from a sequence assembled from a metagenome. Hence, the vast majority of metagenome-derived phages are not associated to a specific host, and "who infects whom?" remains largely unknown [10,13].

In the absence of well-established high-throughput experimental methods to establish phage-host associations for uncultivated phages, researchers have to rely on bioinformatic predictions to associate metagenome-derived phages with their potential host(s). These predictions are typically based on molecular signals (features) of coevolution and/or arms race between phages and their host(s), including exact matches to reference viral or host genomes, matches to host-encoded CRISPR spacers, and sequence composition analyses. Most of these tools however are challenged by the uneven representation of host (i.e., bacterial) diversity in public genome databases, the fact uncultivated phage genomes are often fragmented and partial, and the nuances of phage-host relationships which can not always be easily translated as a binary "infect" vs "does not infect" framework. Here, we review the range of computational approaches currently available for sequence-based prediction of phage-host pairs, and highlight some recent developments in complementary experimental methods.

## Host-phage prediction approaches by bioinformatics tools

Bioinformatic tools for host-phage predictions require as input a phage genome and, for some approaches, a set of candidate host genomes. Some methods based on co-abundance of phages and their hosts also require a set of metagenomes, often from a time series, from which phage and host coverage can be tracked and correlated [14,15]. Correlation results can be however challenging to interpret and sometimes misleading, especially in complex communities for which phage-host correlation patterns vary with sample frequency and across infection types and dynamics [16,17]. In this review, we focus instead on methods based only on the phage genome sequence and, if required, candidate host genomes.

### *I. Alignment-dependent approaches*

Prediction methods based on sequence alignment can be broadly classified in two main categories: the ones based on nucleotide similarity between a query virus and host genomes, and the ones based on similarity between a query virus and known marker genes, i.e., genes encoded exclusively by viruses infecting a specific host taxon (Figure 1, Table 1).

#### I.1 Approaches based on nucleotide similarity to host genomes

RefSeq genomes of Bacteria and Archaea are typically used as the host reference database to identify nucleotide sequence similarity between the input virus and candidate hosts. These regions of sequence similarity can correspond to integrated proviruses, host-encoded CRISPR spacers, auxiliary metabolic genes (AMGs), and/or shared tRNAs, which reflect different arms race and/or co-evolution processes [17] (Table 1). Large (i.e., several kilobases or more) regions of sequence similarity typically result from the integration of a prophage closely related to the query phage in the host genome. This prophage may be still intact or partially degraded [18]. Short regions of sequence similarity resulting from genes horizontally transferred (e.g., AMGs) or insertion sites (e.g. tRNAs) are also typically associated with past successful infections and adaptation of phages to their bacterial host [19]. Conversely, similarity between the query phage and a bacterial CRISPR spacer reflects instead a successful defense of this bacteria against a closely related phage: in this case the specific bacteria is likely able to resist to this phage, and the high level of sequence similarity between phage genomes and CRISPR spacer stems from the ongoing arms race between phages and hosts [20].

Blastn (nucleotide-nucleotide) and blastx (nucleotide-protein) searches [21] are most frequently used to identify regions of sequence similarity, with predicted host(s) identified from hits above defined cutoffs on e-value, bit score, match length, and/or number of mismatches. Importantly, robust identification of short matches such as CRISPR spacers or tRNA require either the use of custom sequence similarity search strategies [22], or some adjustments to the blast search including using the blastn-short task, turning off the dust filtering, and applying stringent filtering criteria allowing only up to one or two mismatches across the whole sequence [17]. When using stringent criteria, alignment-dependent predictions often display a high accuracy, i.e., high percentage of correct phage-host pairs, but a low recall, i.e., low percentage of phage-host pairs predicted compared to the total number of input phages [17]. The resolution (i.e., taxonomic rank) of the final host prediction depends on the type and score of the matche(s), while the prediction accuracy can be even improved by considering multiple hits for each query, e.g. through a “lowest common ancestor” approach applied to a defined number of best-scoring matches [17,22,23].

In addition to prokaryotic genomes from the RefSeq database, it can be beneficial to complement the host reference database with genomes from uncultivated microbes, obtained via single-cell sorting (SAGs) or metagenome assembly and binning (MAGs) [24]. SAGs are generated from microbial cells individually sorted, amplified, and sequenced [25]. These genomes can thus be used for both host prediction and viral discovery, as viral genomes can be assembled *de novo* as part of a SAG generated from an infected cell [26–29]. However, most SAGs are incomplete, fragmented, and because of the amplification step, can also be prone to cross-contamination which should always be considered when leveraging these genomes for host prediction [26]. Meanwhile, MAGs are genomes composed of one or more metagenomic contigs, grouped together based on sequence composition and/or gene content features suggesting these contigs belong to the same genome [24]. As for SAGs, contamination, i.e., unrelated contigs being wrongly gathered in the same MAG, is a potential issue when using these genomes for host prediction, especially when contigs are mostly or entirely viral, in which case their grouping in a MAG is frequently erroneous. Global analysis across biomes and taxa suggested that considering only contigs with no viral region predicted or a viral region representing < 2/3rd of the contig length for host prediction, strongly limited the prediction errors linked to MAG contamination [30]. When properly curated and filtered however, both SAGs and MAGs can be very helpful for host prediction as they increase the diversity of the host reference database and can be derived from the same ecosystem, sampling location, or even sample as the query virus, i.e., they are more likely to represent (one of the) the “true” host(s) [30–32].

I.2 Approaches based on viral marker genes

In contrast to methods leveraging sequence similarity between virus and candidate host genomes, another group of tools relies instead on a comparison between a query phage and a set of pre-defined phage marker genes (Table 1). In Viral Host UnveiLing Kit (vHULK), phage predicted protein sequences are affiliated to the

Prokaryotic Virus Orthologous Group (pVOGs) database [33], and the pVOG list of reach query genome is used as input for two deep neural networks which provide a prediction of host species and genus, along with a measure for prediction confidence (i.e., entropy value) [34]. VPF-Class compared phage predicted proteins against a subset of Viral Protein Families (VPFs, [35]) and derive host prediction and confidence scores from the list of VPFs detected on each query genome, first at the host domain level and then to the family and genus levels, based on the distribution of these VPFs in reference phage genomes [36]. Finally, in the tool Random Forest Assignment of Hosts (RaFAH), predicted proteins are compared to a custom database of HMM profiles obtained from isolated phages and uncultivated phages with a high-confidence host prediction. The list of HMM profiles identified for each query virus is then used as input to a Random Forest Classifier which provides a prediction score (between 0 and 1) for each possible host from phylum to genus [37].

Overall, marker-gene-based approaches display a high accuracy and, on the test sets used in their respective benchmarks, medium to high recall (Table 1). These estimations are however biased by the relatively low number of "entirely distinct" phages available in reference databases: even when attempting to control for this database bias, the training and test sets used to evaluate these tools are often more similar to each other than newly assembled phages from environmental metagenomes would be to reference phage genomes. The performance of these approaches will thus likely be lower (especially in terms of recall) when processing real datasets with entirely novel phages. On the other hand, the performance of these tools should increase in the future as more phages from diverse hosts and environments are added to the genome databases.

*II. Alignment-free approaches*

The second major group of phage-host prediction method does not rely on sequence or profile alignment, but instead establishes putative phage-host pairs based on sequence composition feature(s) of the whole phage genome or of phage predicted proteins (Figure 1). Similarities in sequence composition between phage and host genomes are primarily due to adaptation of the phage genome to the host replication, transcription, and translation machinery [38]. Overall, these methods have the potential to provide host prediction for a broader diversity of phages as they do not strictly require the presence of a closely related phage or host in the databases, however they tend to display a lower accuracy than alignment-based approaches (Table 1).

II.1 Sequence composition features used for alignment-free host prediction

The most common sequence composition feature used for host prediction is the comparison of overall k-mer (DNA sequences of length k) composition of the query genome against a reference database. While HostPhinder predicts phage-host interactions by comparing 16-mers composition between query phage and known phage genome sequences [39], most tools instead compare k-mer frequencies between the query phage and a host reference genome database, based on the assumption that virus and host genomes often display similar sequence composition bias and k-mer frequencies patterns (Table 1). Prokaryotic virus host Predictor (PHP) [40] and VirHostMatcher (VHM) [41] respectively compare 4-mer and 6-mer frequency vectors between the query phage and a database of reference host genomes. "Who Is the Host?" (WIsH) do a similar k-mer-based comparison, by training a homogeneous Markov model of order 8 on each host genome, and computing the likelihood of a query contig under each of the trained Markov models [42]. This Markov model approach was developed to better handle short phage contigs, for which estimated k-mer frequencies become very noisy. Finally, ILMF-VH uses a "hybrid" approach in which the query phage is compared via a 6-mer distance to a database of known phages, which are themselves connected to a host taxon network built from host-host 6-mer distances and similarity in "interaction profiles", i.e., similarity in the list of phages associated with each host [43].

As an alternative to nucleotide k-mer frequencies, other alignment-free host prediction methods have been proposed based on the properties of predicted protein sequences in phage and host genomes (Table 1). These features typically include the frequency of each amino-acid in the predicted protein sequences, the abundance of

selected chemical elements in the proteins (e.g., Carbon, Hydrogen, Nitrogen, Oxygen, and/or Sulfur) and the molecular weight of the protein. While markedly different from nucleotide k-mer frequencies evaluated from whole genomes, these protein sequence features similarly enable host prediction without requiring a sequence alignment step [44–46].

II.2 The "prediction engines": from features to host prediction

Alignment-free measure of similarity between genomes, either between two virus genomes or between a virus and a host genome, are fast to compute and have the potential to identify similarity in genome composition even in the absence of clearly detectable regions of sequence similarity. The results obtained can however be challenging to interpret. Specifically, once a set of similarity values or scores is obtained for a given query phage against a database of host genomes, the next critical step is to aggregate these in a single host prediction and evaluate its confidence.

The first tools leveraging alignment-free methods, including HostPhinder and VHM, both proposed empirical cutoffs to filter the results obtained [39,41]. VHM benchmarks also suggested that results were improved when considering a consensus approach, e.g., of the 30 most similar hosts [41]. WIsH computes a p-value for each phage-host pair based on the Gaussian null distribution pre-computed for each host from a diverse set of phages, while also observing that a consensus approach of the most similar hosts improved overall accuracy [42].

In contrast, the most recent tools leverage advances in machine learning to identify reliable predictions, including Gaussian models [40], neighborhood regularized logistic matrix factorization [43], and deep convolution neural network [45]. A critical and challenging aspect of these techniques is the establishment of robust and balanced training and test sets which should ideally represent a diverse range of viruses, hosts, and virus-host interactions, to avoid over-estimating the performance of these tools, i.e. to prevent overfitting. While automated methods such as K-means clustering of positive and negative pairs have been proposed to help balance the training and test sets, these are still ultimately strongly dependent and biased by the limited number and diversity of phage-host pairs currently available in databases.

II.3 Performance of sequence composition-based host prediction tools

Overall, alignment-free tools tend to offer a higher recall but lower precision than alignment-based tools (Table 1). Across the different tools, accuracy values of 30%-70% have been reported [17,40–43,45] but there is currently no systematic benchmark of these different tools on a common set of genomes as was done in ref. [17] for alignment-based approaches, so that a direct comparison of these tools' performance is challenging to establish. When processing uncultivated viral genomes from various datasets and ecosystems, alignment-based and alignment-free approaches typically provided host predictions for distinct sets of viruses, with alignment-free tools often displaying a higher recall, i.e. larger number of predictions [31,32,47]. This is consistent with the expectation that alignment-free tools are able to detect overall similarity in genome composition between two genomes without alignable regions, with the potential to predict hosts for entirely novel phages (Table 1). Finally, as for alignment-based approach, the use of uncultivated microbial genomes (SAGs and/or MAGs) in the host reference database helped increase the number of host predictions obtained [31].

*III. Integrative approaches*

Since the different methods available for predicting phage-host interactions each display specific challenges and limitations, new tools have been developed that integrate multiple approaches to maximize both the recall and accuracy of phage-host prediction. Specifically, VirHostMatcher-Net [48] and PHISDetector (for Phage-Host Interaction Signals) [49] integrate multiple alignment-free and alignment-based features and score the overall probability of individual phage-host pairs using machine-learning models. Both tools integrate features based on: (i) the alignment-free sequence similarity between phage and host based on k-mer frequencies; (ii) the

existence of shared CRISPR spacers between phage and host; and (iii) alignment-based sequence matches between phage and host. VirHostMatcher-Net, a two-layer network prediction framework, also leverages virus-virus and host-host similarities, while PHISDetector integrates additional features from putative prophage regions in bacterial genomes and protein-protein interactions. Both tools reported increased recall and accuracy compared to single-approach predictions, and such integrative approaches thus appear to be a promising way to improve host prediction by combining methods with distinct strengths and limitations (Table 1) [48,49]. Nevertheless, these integrative approaches are still very new and will need to be further evaluated across a broad range of phages, hosts, and ecosystems, to better understand their potential pitfalls.

*IV. Example case of host prediction result and interpretation*

In 2014, Dutilh et al. identified a new phage genome called crAssphage (for Cross-Assembly phage), from a combined assembly of human fecal microbiome metagenomes [8]. Using large-scale read mapping, this phage was found to be highly abundant and ubiquitous in human gut microbiomes, and further metagenome analyses revealed that crAssphage was one representative of a relatively broad crAss-like family that can reach high abundance in diverse habitats, both animal-associated and environmental [50,51]. In the absence of a cultivated representative, crAss-like phages had to be linked to putative host(s) via *in silico* prediction.

Using alignment-based methods, bacteria from the Bacteroidetes phylum, in particular within the *Bacteroides*, *Prevotella*, and *Porphyromonas* genera, were consistently identified as the most likely hosts of crAss-like phages [8,50,52]. Specifically, several genes encoded by crAss-like phages could be tentatively linked to Bacteroidetes, including a DNA primase which appears to be monophyletic in the crAss-like family and forms a strongly supported clade with the primases of Bacteroidetes, as well as predicted proteins including a "BACON" (Bacteroidetes-associated carbohydrate-binding) domain [50,52]. Meanwhile, matches to CRISPR spacers also pointed towards Bacteroidetes hosts. The original description of the crAssphage genome identified 2 hits to CRISPR spacers in a *Prevotella* and a *Bacteroides* genome, however the alignments included several ($\geq$ 3) mismatches so that these CRISPR matches by themselves do not provide a high-confidence host prediction [8]. A high-confidence prediction could be obtained for another crAss-like phage however, based on a perfect match to a CRISPR spacer in a *Porphyromonas* genome [50]. Finally, a co-occurrence analysis based on read mapping also suggested that crAssphage infected host(s) in the Bacteroidetes phylum [8].

Several alignment-free methods have also been applied to the crAssphage genome, but without providing a confident and consistent host prediction. Using VHM with a focused set of 360 Human Microbiome Project (HMP) host genomes found in the human gut, the most similar strain to crAssphage was a member of the genus *Coprobacillus* in the bacterial phylum Firmicutes [41]. Conversely, WIsH did predict Bacteroidetes hosts for several crAss-like phages [52]. However in both cases, these predictions were obtained with low confidence scores, indicating that they remain uncertain.

Taken individually, most of these results do not provide a robust and high-confidence host prediction by themselves. However, by comparing and integrating the results of multiple independent approaches, crAss-like phages can be confidently linked to several genera within the *Bacteroidetes* phylum. This host prediction was further confirmed by the first two cultivated representatives of the crAss-like family, which were isolated on Bacteroidetes hosts, specifically in the *Bacteroides* genus [53,54]. Hence, while the exact host(s) of the original crAssphage is still not determined, this example illustrates how integrating multiple *in silico* host predictions can help in identifying the most likely host(s) of entirely novel phages known only from metagenome assemblies.

*V. Experimental approaches to predict phage-host relationships*

As a complement to bioinformatic predictions, different experimental approaches have been developed and used for linking novel phages to their host(s). The most sensitive tools to identify which phage infects which bacterium are spot and plaque assays, in which cultures of potential hosts are challenged by phages, and

infections are identified based on host cell clearing [2]. These are however usually limited in throughput and biased towards high-virulence phages, since low-virulence or lysogenic phages may not kill enough host cells to be readily detected. Complementarily, new methods are available to quickly identify from a complex community which phages can bind to and potentially infect a cultivated microbial host. These include viral tagging (VT), which relies on Fluorescence-activated Cell Sorting (FACS) to sort microbial cells to which fluorescently labeled viral particles have bound [55,56], and AdsorpSeq, which leverages the differential migration rates of phage particles bound to host receptors compared to free phage particles in agarose gel electrophoresis [57]. Importantly, in both cases, the approach detects successful binding between virus particle and host cell but does not guarantee that a full infection cycle occurred or would even be possible.

Alternatively, cultivation-independent *in vitro* assays have been developed that typically rely on the co-detection of a host and a viral marker in a host cell, either from fluorescent markers or by physically linking the two markers. Among fluorescent-based methods, PhageFISH, an epifluorescence microscopy-based method, is able to detect both replicating and encapsidated (intracellular and extracellular) phage DNA [58]. Microfluidic digital multiplex PCR [59] and droplet digital PCR (ddPCR) [60] have also been used to co-localize host and viral marker genes. In these approaches, environmental microbial cells are loaded onto a digital PCR array panel or encapsulated in microdroplets such that the majority of chambers/droplets are either empty or contained a single bacterium, and a PCR with fluorescent reporters is used to assess the co-occurrence of a viral marker gene and a host marker gene. The main limitation of these approaches however resides in the probe design, which need to be specific to enable reliable identification viruses and hosts, and is not amenable to large and/or diverse groups.

Finally, new methods have been recently developed that attempt to co-localize viral and host genomes (or marker genes) using physical linkage instead of fluorescence. For instance, emulsion paired isolation-concatenation PCR (epicPCR) proposes to first encapsulate individual cells into microdroplets, then perform a fusion PCR step which joins a viral marker gene and a host marker gene. Fused amplicons, each originating from a single droplet, can then be sequenced to identify the specific virus and host that were co-localized. This method was recently used to link uncultivated phages to their host in a tidal estuary environment [61]. Although the lack of a universal viral marker gene remains an issue for epicPCR, PCR primers can be designed for common viral genes that typically target a broader diversity than fluorescent probes [62]. Alternatively viral and host DNA can also be linked via high-throughput chromosomal confirmation capture (Hi-C) [63,64]. A single Hi-C assay typically produces millions of 'contacts', where each contact reflects two sequence fragments, including potentially phage and host fragments from an ongoing infection. Recently, Ignacio-Espinoza et al. [65] further proposed the XRM-Seq (Ribosome cross-linking and sequencing) method, an adapted Hi-C method to ligate host's rRNA to viral transcripts, allowing to directly link viral gene expression to specific hosts. For all co-localization methods however, the virus-host pairs reflect the presence of the phage in its host cell, but cannot confirm active infections or differentiate between lytic or lysogenic infections.

**Conclusions**

We are only starting to explore the immense diversity of viruses, especially bacteriophages, in nature. Emerging computational and experimental approaches improving our ability to link viruses and hosts without the need to cultivate all possible virus-host pairs will undoubtedly contribute to a better understanding of global viral ecology. With the growing number of computational approaches available, combining multiple host prediction tools appears to be a reliable strategy to optimize recall and accuracy of phage-host pair detection [48,49]. Critically, large-scale data sets of verified virus-host links are necessary for a more holistic and predictive evaluation of virus-host interactions in nature. Meanwhile, conceptual models describing virus-host relationships will also need to be enriched based on experimental studies of host range and virion production to better reflect the broad range of existing virus-host dynamics. Specifically, phage-host linking methods will need to move

from qualitative ("who infects whom") to quantitative ("which proportion of a host population is infected by whom") to better understand viral roles and impacts on natural communities [19].

In parallel to host-phage prediction approaches, tools aiming at inferring the specific host of novel eukaryotic viruses are also being developed and refined. Host Taxon Predictor (HTP) [66] and VIrus Deep learning HOst Prediction (VIDHOP) [67] are tools based on simple sequence features (nucleotide k-mer frequencies) that perform host prediction at the ~ domain level (i.e., eukaryote vs prokaryote) or at the species level within specific eukaryotic groups (e.g., animals). However, these methods are limited in the number of different hosts they can predict and mainly focus on in human virus species or disease-related viruses. There is currently a lack of tools that can predict hosts of viruses infecting unicellular eukaryotes (amoebas, protozoans, and microalgae) or of giant viruses, which are now more readily assembled and detected from environmental sequencing datasets [9]. Ultimately, our ability to identify putative hosts of uncultivated viruses infecting bacteria, archaea, and eukaryotic hosts is critical to robustly analyze increasingly complex environmental sequencing datasets and to design targeted virus isolation assays.

## Figures & Tables

| Method category | Tool | Date | Reference database used | Features | Main Strengths / Advantages | Main Challenges / Limitations |
|---|---|---|---|---|---|---|
| Alignment-dependent | Blastn [21] | NA | Host genomes, genes, and CRISPR spacers | CRISPR spacers, AMG, tRNA | High specificity, up to the strain level [17,22,30]. Potential to identify host(s) of phages unrelated to a phage reference genome. | Strong dependency on host genome database, often leading to a low recall. This can be partially alleviated by using uncultivated microbial genomes (SAGs and MAGs) from relevant samples/ecosystems. For CRISPR spacer matches, can only be applied to hosts which encode a CRISPR system, which proportion vary by environment [68] |
| | SpacePHARER [22] | 2020 | Host CRISPR spacers | CRISPR spacers | | |
| | RaFAH [37] | 2020 | Phage marker genes / HMM profiles | Custom HMM profiles | High accuracy for phages related to known references, does not depend on a host reference database. | Only applicable to phages sharing at least 1 marker gene with a known phage reference. |
| | vHULK [34] | 2020 | Phage marker genes / HMM profiles | Protein families (pVOGs) [33] | | |
| | VPF-Class [36] | 2021 | Phage marker genes / HMM profiles | Protein families (VPF) [36] | | |
| Alignment-free | HostPhinder [39] | 2016 | Phage genomes | Nucleotide 16-mer frequencies | Independent of gene prediction and host reference database. | Only applicable to phages similar to at least 1 phage reference genome |
| | ILMF-VH [43] | 2019 | Phage genomes | Nucleotide 6-mer frequencies | | |
| | VirHostMatcher [41] | 2017 | Host genomes | Nucleotide 6-mer frequencies | Independent of gene prediction and phage reference database, i.e. able to identify hosts for entirely novel phages without any related reference. Typically high recall, especially when the host database includes uncultivated microbes (SAGs / MAGs) from relevant samples/ecosystems. | High rate of false-positives compared to other methods, often requires additional statistical test and/or consensus approaches to post-process the results. For "Boeckaerts's method", the approach was only tested on phages infecting *E. coli*, *S. enterica*, and *C. difficile*. |
| | WIsH [42] | 2017 | Host genomes | Nucleotide 8-mer frequencies | | |
| | PHP [40] | 2020 | Host genomes | Nucleotide 4-mer frequencies | | |
| | PredPhi [45] | 2020 | Host proteomes | Properties of protein sequences | | |
| | Leite's methods [44] | 2018 | Host proteomes | Properties of protein sequences | | |
| | Boeckaerts's method [46] | 2021 | Subset of host proteome (RBP) | Receptor-binding proteins | | |
| Integrative | PHISDetector [49] | 2020 | Phage and Host genomes | Multiple features | Improved accuracy and sensitivity by considering and integrating multiple signals. | Integrative approaches are still relatively new and need to be further evaluated. Integrative methods may also require a longer compute time since they need results from several individual prediction approaches. |
| | VirHostMatcher Net [48] | 2020 | Host genomes | Multiple features | | |

**Table 1. Overview of bioinformatic tools and approaches used for phage-host relationship predictions.** Because there are no systematic benchmark of these different tools on a single set of diverse phage genomes, a direct and quantitative comparison of the tools performance is challenging to establish. Hence, we opted to outline "qualitative" broad strengths and limitations of each tool category, to help users identify which tool(s) may be most appropriate for their research question.

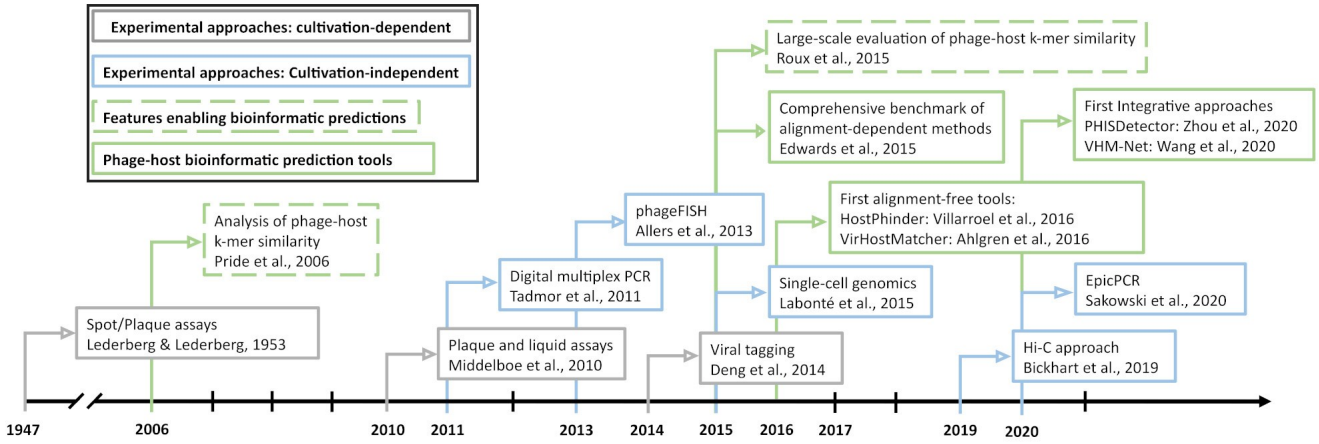


**Figure 1. Overall timeline of bioinformatic and experimental approaches used to identify phage-host pairs.** Key advances or new approaches for both *in vitro* (blue) and *in silico* (green) predictions are indicated, and correspond to the following references [17,29,38,39,41,48,55,58,59,61,64,69–71].

**Acknowledgements**

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. This work was specifically supported by the U.S. Department of Energy, Office of Science, Biological and Environmental Research, Early Career Research Program awarded under UC-DOE Prime Contract DE-AC02-05CH11231.

**References**

1. Luong T, Salabarria AC, Edwards RA, Roach DR: **Standardized bacteriophage purification for personalized phage therapy**. *Nat Protoc* 2020, **15**:2867–2890.

2. Hyman P: **Phages for phage therapy: Isolation, characterization, and host range breadth**. *Pharmaceuticals* 2019, **12**.

3. Raya RR, Hébert EM: **Bacteriophages : methods and protocols**. *Methods Mol Biol* 2009, **501**:23–32.

4. Steen AD, Crits-Christoph A, Carini P, DeAngelis KM, Fierer N, Lloyd KG, Cameron Thrash J: **High proportions of bacteria and archaea across most biomes remain uncultured**. *ISME J* 2019, **13**:3126–3130.

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

6. Feiner R, Argov T, Rabinovich L, Sigal N, Borovok I, Herskovits AA: **A new perspective on lysogeny: Prophages as active regulatory switches of bacteria**. *Nat Rev Microbiol* 2015, **13**:641–650.

7. Páez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC: **Uncovering Earth's virome**. *Nature* 2016, **536**:425–430.

8. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, Barr JJ, Speth DR, Seguritan V, Aziz RK, et al.: **A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes**. *Nat Commun* 2014, **5**:4498.

9. Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, McMahon KD, Konstantinidis KT, Eloe-Fadrosh EA, Kyrpides NC, et al.: **Giant virus diversity and host interactions through global metagenomics**. *Nature* 2020, **578**.

10. Roux S, Páez-Espino D, Chen I-MA, Palaniappan K, Ratner A, Chu K, Reddy TBK, Nayfach S, Schulz F, Call L, et al.: **IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses**. *Nucleic Acids Res* 2020, doi:10.1093/nar/gkaa946.

11. Dutilh BE, Reyes A, Hall RJ, Whiteson KL: **Virus Discovery by Metagenomics: The (Im)possibilities**. *Front Microbiol* 2017, **8**:5–7.

12. Dion MB, Oechslin F, Moineau S: **Phage diversity, genomics and phylogeny**. *Nat Rev Microbiol* 2020, **18**:125–138.

13. Roux S, Adriaenssens EM, Dutilh BE, Koonin E V., Kropinski AM, Krupovic M, Kuhn JH, Lavigne R, Brister JR, Varsani A, et al.: **Minimum information about an uncultivated virus genome (MIUVIG)**. *Nat Biotechnol* 2019, **37**:29–37.

14. Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, Brussaard CPDD, Dutilh BE, Thompson FL: **Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans**. *Nat Commun* 2017, **8**:1–12.

15. Arkhipova K, Skvortsov T, Quinn JP, McGrath JW, Allen CC, Dutilh BE, McElarney Y, Kulakov LA: **Temporal dynamics of uncultured viruses: a new dimension in viral diversity**. *ISME J* 2017, **12**:199–211.

16. Coenen AR, Weitz JS: **Limitations of Correlation-Based Inference in Complex Virus-Microbe Communities**. *mSystems* 2018, **3**:7–9.

17. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE: **Computational approaches to predict bacteriophage-host relationships**. *FEMS Microbiol Rev* 2016, **40**:258–272.

18. Canchaya C, Fournous G, Brüssow H: **The impact of prophages on bacterial chromosomes.** *Mol Microbiol* 2004, **53**:9–18.

19. Breitbart M, Bonnain C, Malki K, Sawaya NA: **Phage puppet masters of the marine microbial realm**. *Nat Microbiol* 2018, **3**:754–766.

20. Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, Brouns SJJ, Charpentier E, Cheng D, Haft DH, Horvath P, et al.: **Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants**. *Nat Rev Microbiol* 2020, **18**:67–83.

21. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications**. *BMC Bioinformatics* 2009, **10**:421.

22. Zhang R, Mirdita M, Karin EL, Norroy C, Galiez C, Soeding J: **SpacePHARER: Sensitive identification of phages from CRISPR spacers in prokaryotic hosts**. *bioRxiv* 2020, doi:10.1101/2020.05.15.090266.

23. Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R: **Expanding the marine virosphere using metagenomics.** *PLoS Genet* 2013, **9**:e1003987.

24. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, et al.: **Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea**. *Nat Biotechnol* 2017, **35**:725–731.

25. Pachiadaki MG, Brown JJM, Brown JJM, Bezuidt O, Berube PM, Biller SJ, Poulton NJ, Burkart MD, La Clair JJ, Chisholm SW, et al.: **Charting the Complexity of the Marine Microbiome through Single-Cell Genomics**. *Cell* 2019, **179**:1623-1635.e11.

26. Munson-Mcgee JH, Peng S, Dewerff S, Stepanauskas R, Whitaker RJ, Weitz JS, Young MJ: **A virus or more in (nearly) every cell: Ubiquitous networks of virus-host interactions in extreme environments**. *ISME J* 2018, **12**:1706–1714.

27. Jarett JK, Džunková M, Schulz F, Roux S, Paez-Espino D, Eloe-Fadrosh E, Jungbluth SP, Ivanova N, Spear JR, Carr SA, et al.: **Insights into the dynamics between viruses and their hosts in a hot spring microbial mat**. *ISME J* 2020, doi:10.1038/s41396-020-0705-4.

28. Roux S, Hawley AK, Torres Beltran M, Scofield M, Schwientek P, Stepanauskas R, Woyke T, Hallam SJ, Sullivan MB: **Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta- genomics.** *Elife* 2014, **3**:e03125.

29. Labonté JM, Swan BK, Poulos BT, Luo H, Koren S, Hallam SJ, Sullivan MB, Woyke T, Wommack EK, Stepanauskas R: **Single cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton**. *ISME J* 2015, **9**:2386–99.

30. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, Wu D, Paez-Espino D, Chen IM, Huntemann M, et al.: **A genomic catalog of Earth's microbiomes**. *Nat Biotechnol* 2020, doi:10.1038/s41587-020-0718-6.

31. Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang H Bin, Singleton CM, Solden LM, Naas AE, Boyd JA, et al.: **Host-linked soil viral ecology along a permafrost thaw gradient**. *Nat Microbiol* 2018, **3**:870–880.

32. Dalcin Martins P, Danczak RE, Roux S, Frank J, Borton MA, Wolfe RA, Burris MN, Wilkins MJ: **Viral and metabolic controls on high rates of microbial sulfur and carbon cycling in wetland ecosystems**. *Microbiome* 2018, **6**:1–17.

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

34. Amgarten D, Vázquez Iha BK, Piroupo CM, da Silva AM, Setubal JC: **vHULK, A new tool for bacteriophage host prediction based on annotated genomic features and deep neural networks**. *bioRxiv* 2020, doi:10.1101/2020.12.06.413476.

35. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC: **Uncovering Earth's virome**. *Nature* 2016, **536**:425–430.

36. Pons JC, Paez-Espino D, Riera G, Ivanova N, Kyrpides NC, Llabrés M: **VPF-Class : taxonomic assignment and host prediction of uncultivated viruses based on viral protein families**. *Bioinformatics* 2021, doi:10.1093/bioinformatics/btab026.

37. Coutinho FH, Zaragoza-Solas A, López-Pérez M, Barylski J, Zielezinski A, Dutilh BE, Edwards RA, Rodriguez-Valera F, Diego S: **RaFAH: A superior method for virus-host prediction**. *bioRxiv* 2020, doi:10.1101/2020.09.25.313155.

38. Roux S, Hallam SJ, Woyke T, Sullivan MB: **Viral dark matter and virus-host interactions resolved from publicly available microbial genomes.** *Elife* 2015, **4**:e08490.

39. Villarroel J, Kleinheinz KA, Jurtz VI, Zschach H, Lund O, Nielsen M, Larsen MV: **HostPhinder: A phage host prediction tool**. *Viruses* 2016, **8**:116.

40. 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.

41. Ahlgren N, Ren J, Lu YY, Fuhrman JA, Sun F: **Alignment-free $d_2^*$ oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences**. *Nucleic Acids Res* 2016, **45**:39–53.

42. Galiez C, Siebert M, Enault F, Vincent J, Söding J, Enault F, Vincent J, Söding J: **WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs**. *Bioinformatics* 2017, **33**:3113–14.

43. Liu D, Ma Y, Jiang X, He T: **Predicting virus-host association by Kernelized logistic matrix factorization and similarity network fusion**. *BMC Bioinformatics* 2019, **20**:1–10.

44. Leite DMC, Brochet X, Resch G, Que YA, Neves A, Peña-Reyes C: **Computational prediction of inter-species relationships through omics data analysis and machine learning**. *BMC Bioinformatics* 2018, **19**.

45. Li M, Wang Y, Li F, Zhao Y, Liu M, Zhang S, Bin Y, Smith AI, Webb G, Li J, et al.: **A Deep Learning-Based Method for Identification of Bacteriophage-Host Interaction**. *IEEE/ACM Trans Comput Biol Bioinforma* 2020, doi:10.1109/tcbb.2020.3017386.

46. 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.

47. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, Poulos BT, Solonenko N, Lara E, Poulain J, et al.: **Ecogenomics and potential biogeochemical impacts of uncultivated globally abundant ocean viruses**. *Nature* 2016, **537**:689–93.

48. Wang W, Ren J, Tang K, Dart E, Ignacio-Espinoza JC, Fuhrman JAA, Braun J, Sun F, Ahlgren NA: **A network-based integrated framework for predicting virus–prokaryote interactions**. *NAR Genomics Bioinforma* 2020, **2**:1–19.

49. Zhang F, Zhou F, Gan R, Ren C, Jia Y, Yu L, Huang Z: **PHISDetector: A web tool to detect diverse in silico phage-host interaction signals**. *bioRxiv* 2019, doi:10.1101/661074.

50. Yutin N, Makarova KS, Gussow AB, Krupovic M, Segall A, Edwards RA, Koonin E V.: **Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut**. *Nat Microbiol* 2018, **3**:38–46.

51. Yutin N, Benler S, Shmakov SA, Wolf YI, Tolstoy I, Rayko M, Antipov D, Pevzner PA, Koonin E V.: **Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features**. *Nat Commun* 2021, **12**:1–11.

52. De Jonge PA, Von Meijenfeldt FAB, Van Rooijen LE, Brouns SJJ, Dutilh BE: **Evolution of BACON domain tandem repeats in crassphage and novel gut bacteriophage lineages**. *Viruses* 2019, **11**:1–16.

53. Shkoporov AN, Khokhlova E V., Fitzgerald CB, Stockdale SR, Draper LA, Ross RP, Hill C: **ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis**. *Nat Commun* 2018, **9**:1–8.

54. Guerin E, Shkoporov AN, Stockdale SR, Comas JC, Khokhlova E V., Clooney AG, Daly KM, Draper LA, Stephens N, Scholz D, et al.: **Isolation and characterisation of ΦcrAss002, a crAss-like phage from the human gut that infects Bacteroides xylanisolvens**. *Microbiome* 2021, **9**:1–21.

55. Deng L, Ignacio-Espinoza JCJC, Gregory ACA, Poulos BT, Weitz JS, Hugenholtz P, Sullivan MB: **Viral tagging reveals discrete populations in Synechococcus viral genome sequence space**. *Nature* 2014, **513**:242–45.

56. Džunková M, Low SJ, Daly JN, Deng L, Rinke C, Hugenholtz P: **Defining the human gut host–phage network through single-cell viral tagging**. *Nat Microbiol* 2019, **4**:2192–2203.

57. de Jonge PA, von Meijenfeldt FAB, Costa AR, Nobrega FL, Brouns SJJ, Dutilh BE: **Adsorption Sequencing as a Rapid Method to Link Environmental Bacteriophages to Hosts**. *iScience* 2020, **23**:101439.

58. Allers E, Moraru C, Duhaime MB, Beneze E, Solonenko N, Canosa JB, Amann R, Sullivan MB: **Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses**. *Environ Microbiol* 2013, **15**:2306–18.

59. Tadmor AD, Ottesen EA, Leadbetter JR, Phillips R, Gallant JA, Cramer P, Weintraub H, Rich A, Maas S, Platas AA, et al.: **Probing Individual Environmental Bacteria for Viruses by Using Microfluidic Digital PCR**. *Science (80- )* 2011, **333**:58–62.

60. Morella NM, Yang SC, Hernandez CA, Koskella B: **Rapid quantification of bacteriophages and their bacterial hosts in vitro and in vivo using droplet digital PCR**. *J Virol Methods* 2018, **259**:18–24.

61. Sakowski EG, Arora-Williams K, Tian F, Zayed AA, Zablocki O, Sullivan MB, Preheim SP: **Interaction dynamics and virus–host range for estuarine actinophages captured by epicPCR**. *Nat Microbiol* 2021, doi:10.1038/s41564-021-00873-4.

62. Adriaenssens EM, Cowan DA: **Using signature genes as tools to assess environmental viral ecology and diversity**. *Appl Environ Microbiol* 2014, **80**:4470–4480.

63. Marbouty M, Thierry A, Millot GA, Koszul R: **MetaHiC phage-bacteria infection network reveals active cycling phages of the healthy human gut**. *Elife* 2021, **10**:1–51.

64. Bickhart DM, Watson M, Koren S, Panke-Buisse K, Cersosimo LM, Press MO, Van Tassell CP, Van Kessel JAS, Haley BJ, Kim SW, et al.: **Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation**. *Genome Biol* 2019, **20**:1–18.

65. Ignacio-Espinoza JC, Laperriere SM, Yeh YC, Weissman J, Hou S, Long AM, Fuhrman JA: **Ribosome-linked mRNA-rRNA chimeras reveal active novel virus host associations**. *bioRxiv* 2020, doi:10.1101/2020.10.30.332502.

66. Gałan W, Bąk M, Jakubowska M, Ga W: **Host Taxon Predictor - A Tool for Predicting Taxon of the Host of a Newly Discovered Virus**. *Sci Rep* 2019, **9**:1–13.

67. Mock F, Viehweger A, Barth E, Marz M: **VIDHOP, viral host prediction with Deep Learning**. *bioRxiv* 2019, doi:10.1101/575571.

68. Burstein D, Sun CL, Brown CT, Sharon I, Anantharaman K, Probst AJ, Thomas BC, Banfield JF: **Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems.** *Nat Commun* 2016, **7**:10613.

69. Lederberg EM, Lederberg J: **Studies on lysogenization in Escherichia coli.** *Genetics* 1953, **38**:51–64.

70. Pride DT, Wassenaar TM, Ghose C, Blaser MJ: **Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses.** *BMC Genomics* 2006, **7**:1–13.

71. Zhang F, Zhou F, Gan R, Ren C, Jia Y, Yu L, Huang Z: **PHISDetector: a tool to detect diverse in silico phage-host interaction signals for virome studies**. *bioRxiv* 2020, doi:10.1101/661074.