HostPhinder: A Phage Host Prediction Tool

Villarroel, Julia; Kleinheinz, Kortine Annina; Jurtz, Vanessa Isabell; Zschach, Henrike; Lund, Ole; Nielsen, Morten; Larsen, Mette Voldby

Published in:
Viruses

Link to article, DOI:
10.3390/v8050116

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

viruses

MDPI

Article

HostPhinder: A Phage Host Prediction Tool

Julia Villarroel [1,*], Kortine Annina Kleinheinz [1], Vanessa Isabell Jurtz [1], Henrike Zschach [1], Ole Lund [1], Morten Nielsen [1,2] and Mette Voldby Larsen [1,*]

[1] Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark; kortinekleinheinz@gmx.de (K.A.K.); vanessa@cbs.dtu.dk (V.I.J.); henrike@cbs.dtu.dk (H.Z.); lund@cbs.dtu.dk (O.L.); mniel@cbs.dtu.dk (M.N.)

[2] Instituto de Investigaciones Biotecnológicas, Universidad de San Martín, CP(1650) San Martín, Prov. de Buenos Aires, Argentina

* Correspondence: juliavi@cbs.dtu.dk (J.V.); metteb@cbs.dtu.dk (M.V.L.); Tel.: +45-4525-2425 (M.V.L.); Fax: +45-4593-1585 (M.V.L.)

Academic Editor: Rob Lavigne
Received: 23 December 2015; Accepted: 19 April 2016; Published: 4 May 2016

Abstract: The current dramatic increase of antibiotic resistant bacteria has revitalised the interest in bacteriophages as alternative antibacterial treatment. Meanwhile, the development of bioinformatics methods for analysing genomic data places high-throughput approaches for phage characterization within reach. Here, we present HostPhinder, a tool aimed at predicting the bacterial host of phages by examining the phage genome sequence. Using a reference database of 2196 phages with known hosts, HostPhinder predicts the host species of a query phage as the host of the most genomically similar reference phages. As a measure of genomic similarity the number of co-occurring k-mers (DNA sequences of length k) is used. Using an independent evaluation set, HostPhinder was able to correctly predict host genus and species for 81% and 74% of the phages respectively, giving predictions for more phages than BLAST and significantly outperforming BLAST on phages for which both had predictions. HostPhinder predictions on phage draft genomes from the INTESTI phage cocktail corresponded well with the advertised targets of the cocktail. Our study indicates that for most phages genomic similarity correlates well with related bacterial hosts. HostPhinder is available as an interactive web service [1] and as a stand alone download from the Docker registry [2].

Keywords: "host specificity"; prediction; genome; k-mers

1. Introduction

In 2012, the World Health Organization (WHO) announced the beginning of the end of the antibiotic era, and the possible return to a time when even trivial bacterial infections could turn out to be fatal [3]. Since then, the problem of antimicrobial resistance has continued to grow and in the foreword to the WHO report "Antimicrobial resistance: global report on surveillance 2014" it is stated that "A post-antibiotic era-in which common infections and minor injuries can kill-far from being an apocalyptic fantasy, is instead a very real possibility for the 21st century" [4]. As emphasized by WHO there is an urgent need for treatment alternatives, one such being bacteriophages (phages). The idea of using phages for the treatment of bacterial infections dates back to 1919, when French-Canadian microbiologist Félix d'Herelle used them for treating a patient with severe bacillary dysentery [5]. For a number of historical reasons, phage therapy never became general practice in the West, although it has been used extensively in countries from the former Eastern bloc [6–9]. Several recent studies from the West have also demonstrated the effectiveness of phages as antibacterial treatment [10–13], and more countries are currently revisiting phage therapy [14,15]. Phages have furthermore been suggested for use in the agriculture and food industries [16,17]. Examples include their use for reducing *Campylobacter jejuni* colonisation of broiler chickens [18] and the growth of *E. coli* in milk [19].

For a phage to successfully infect a bacterial host, the phage must adsorb to the bacterial surface through recognition of specific host receptors, e.g., proteins, LPS, or cell wall polysaccharides. Phage adsorption to an appropriate surface receptor is, however, only the first step required for successful infection. Several host defence mechanisms must also be overcome: Restriction-Modification (RM) systems have been shown to be present in more than 90% of sequenced bacterial genomes [20]. These systems include restriction enzymes that degrade incoming phage DNA with appropriate target sequences. Some bacteria contain Clustered Regular Interspaced Short Palindromic Repeats (CRISPR) loci, which together with the CRISPR-associated (cas) genes encode an adaptive anti-phage immune system [21]. Phage abortive systems (Abi systems) allow infected bacteria to commit "altruistic suicide" thereby preventing the spread of the phage within the bacterial community [22]. Other factors such as successful gene transcription and translation based on amino acid or tRNA availability further limit the host range [23]. Bacteria and phages have from the outset of their coexistence been engaged in a vehement arms race leading to intricate coevolutionary processes, and for each of the defence mechanisms mentioned above, examples exist of phages that have evolved to circumvent them [24,25]. The arms race has contributed to bacterial as well as phage diversity [26] and entails that phage host determination is influenced by multiple genes and genome features distributed across the phage genome. Although examples exist of phages that have extended their host range based on only a few mutations [27], the extended host range is typically limited to different strains of the same species. Apart from polyvalent enterobacteria phages, which are able to infect members of phylogenetically linked genera within the *Enterobacteriaceae* family, e.g., *Escherichia*, *Shigella*, and *Klebsiella* [28,29], most phages have been found to be specific to a particular genus [30]. This has been indicated by studies examining proteins, not entire proteomes [31], as has the "Phage Proteomic Tree", which is based on completely sequenced phage genomes [32], and analysis of genome type for Mycobacteriophages and host preference [33].

In this study, we extend the observation that genetically similar phages often share the same bacterial host species and hypothesize that it should be possible to predict the host species of a phage by searching for the most genetically similar phages in a database of reference phages with known hosts. In the developed method, called HostPhinder, genetic similarity is defined as the number of co-occurring k-mers between the query phage and phages in the reference database. K-mers are stretches of DNA with a length of k, and their use as a measure of genetic relatedness dates back to Woese and Fox and their groundbreaking paper from 1977, which uncovered Archaea as a separate branch in the tree of life [34]. Woese and Fox limited their analysis to k-mers (they used the term oligonucleotides) in 16S (18S) ribosomal RNA, but since phages do not have 16S rRNA genes or any other genes which are common to all phages [32], and because high-throughput sequencing methods have made the entire genome of phages easily available, HostPhinder examines the complete genome. Further, for bacteria we have previously shown that the co-occurrence of k-mers across the entire genome performs superior to other whole-genome or single locus based approaches for inferring genetic relatedness [35]. The splitting of entire phage genomes into overlapping k-mers may furthermore be an advantage in relation to the highly mosaic phage genome structure [36,37].

We believe that a method enabling prediction of the bacterial hosts of phages will be useful for several reasons. Firstly, phages have for many years been used to treat bacterial infections in countries belonging to the former Eastern bloc. The Eliava Institute in Tbilisi, Georgia has in particular been dominant in this regard and produce cocktails containing a mixture of phages for a range of bacterial infections. One of the steps towards adopting phage therapy in the West, is likely to be a full characterization of the content of these cocktails, which due to the way they are manufactured is not known [38]. Further, the current approach to exploration of many ecological niches is done by untargeted sequencing of samples isolated directly from the environment, so called metagenomics. This enables identification of phage and bacterial sequences without knowledge of the link between them, and importantly also enables identification of bacteria, and hence phages, that cannot be cultured. HostPhinder could help establish the link between phages and bacteria, which might be an important

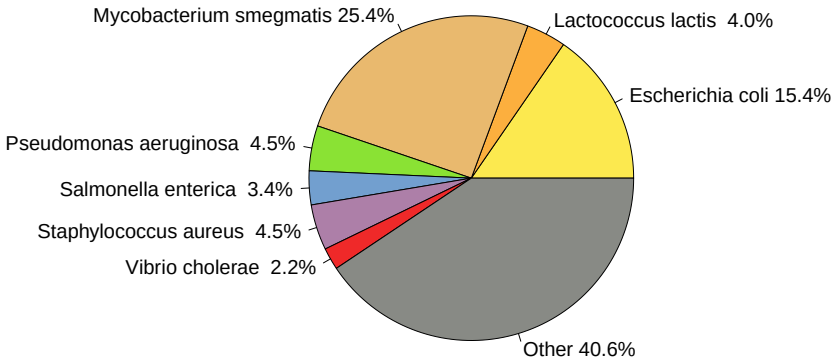
step towards understanding, e.g., the microbiome of the human gut, and possibly associations between the microbiome and clinical parameters of the human host [39].

2. Materials and Methods

2.1. Whole Genome Phage Sequences from Public Databases

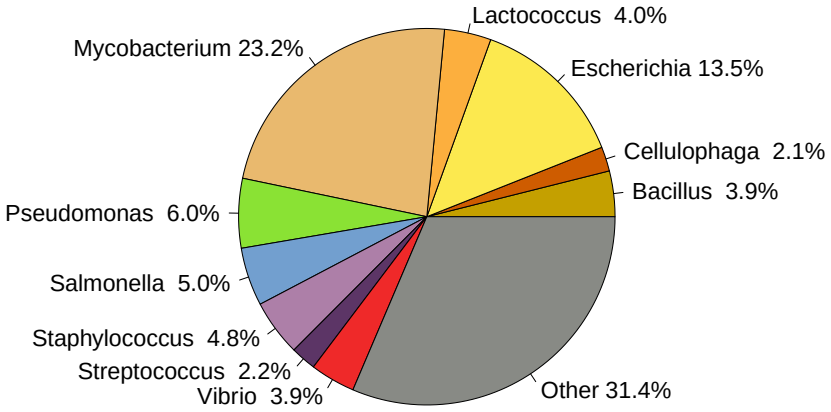
A set of public phage Whole Genome Sequences (WGS) was collected in August 2014: First, lists of phage WGS IDs were obtained from Phages.ids–VBI mirrors page [40], the NCBI viral Genome Resource [41], the EMBL EBI phage genomes list [42], and the phagesdb databases for Mycobacteriophages [43], Arthrobacter [44], Bacillus [45], and Streptomyces [46]. The resulting unique list of IDs was uploaded to the Batch Entrez service of NCBI to retrieve the corresponding WGS. Furthermore genome sequences were downloaded from the PhAnToMe genomes database and from NCBI searching for "(phage [Title]) AND complete genome".

Only entries indicating "complete genome" in the DEFINITION field of the GeneBank file and which host taxonomy was specified at least at the genus level were included. Entries annotated as "prophage" in the DEFINITION were removed. Hosts annotated as *Salmonella Typhimurium* were re-annotated as *Salmonella enterica* according to current nomenclature [47]. Finally, only the genus was taken into account for hosts with species specified as "sp." followed by an alphanumeric code; for example *Synechococcus* sp. *WH7803* was re-annotated as *Synechococcus*. 2196 phages had annotated host genus, here called $\text{phages}_{\text{genus}}$ dataset, and of these, 1871 had annotated species as well, $\text{phages}_{\text{species}}$. A total of 209 different host species and 129 different genera were represented among the phages (this data is available in HostPhinder's repository [48]). Figure 1 shows the distribution of hosts in the dataset.



(**a**) Distribution of host species.

Figure 1. *Cont.*



(**b**) Distribution of host genera.

Figure 1. Hosts represented in the database. Species (**a**) and genera (**b**) representations are displayed in the same genera-colour code.

2.2. Data Partitioning and Clustering

In this study, a 4-fold cross validation setup was used to assess the ability of the host prediction method to generalize to previously unseen data. Five data partitions were made, and one partition, $\text{phage}_{\text{eval}}$ was left aside during the entire process of parameter optimization. Once the parameters were optimized, the prediction accuracy was evaluated on this $\text{phage}_{\text{eval}}$ set, using the entire $\text{phage}_{\text{train,test}}$ set as reference database (Supplementary Materials Figure S1). In this setup, the performance of the evaluation set is hence completely unbiased towards the model parameter optimizations.

A reliable, *i.e.*, not overfitted, evaluation can only be made if phage genomes in the training-test and evaluation sets are not too similar to each other. Indeed, if a phage genome in the training set is almost identical to a genome in the evaluation set, it would be a simple task for HostPhinder to predict its host, leading to an overestimation of the method's ability to generalize to previously unseen data. To avoid such a bias we clustered the genomes according to 16-mer similarity by means of a Hobohm 1 approach [49]. The Hobohm approach consists in the formation of a final list of representative phage genomes, here called seeds. After the first sequence in a randomly sorted list enters the seed list and forms a seed, the following sequences are each checked for similarity (number of overlapping 16-mers) to each seed in the final list. Only if significantly different to the seed sequences, the new sequence will be included in the seed list. Otherwise, it will be linked to the most similar seed as member of the same cluster. The similarity between two genomes was measured in terms of frac_q (see Equation (4) in section "K-mer-based resemblance measures") using a threshold $\text{frac}_q > 0.7$. This threshold was chosen because the resulting clustering was most similar (93%) to the clustering obtained with a BLAST-Hobohm1 approach, where the similarity threshold was set to >90% genomewide ID (data not shown). The k-mer-Hobohm1 analysis resulted in 293 clusters with at least 2 sequences and 1121 singlets. The total number of seeds was hence 1414 containing 1 to 97 sequences. To separate the clustered phages in train-test and evaluation sets, the 1414 seeds were sorted by host alphabetical order, and secondly by size and alternately distributed between 5 partitions. This assured an equal host and genome size representation among partitions. Finally remaining members of each cluster were integrated into the partition of their respective seed. Sequences within the same cluster shared the host; therefore the unbiased host distribution was maintained also after integrating members of the clusters in each partition (see Supplementary Materials Figure S2). Subsets of each of these partitions were made, which comprised all phages that contained information about the species of the host, overall constituting the $\text{phages}_{\text{species}}$ dataset. The host and size distribution between partitions remained conserved (see Supplementary Materials Figures S2–S4). As stated above, one partition was next left aside for final evaluation, $\text{phages}_{\text{eval}}$, and the remaining 4 formed the train-test set,

$\text{phages}_{\text{train-test}}$. The final $\text{phages}_{\text{train-test,genus}}$ set contained 1818 phages (115 genera and 190 species), the $\text{phages}_{\text{eval,genus}}$ set contained 378 phages (72 genera, 96 species), while the $\text{phages}_{\text{train-test,species}}$ set consisted of 1546 phages and the $\text{phages}_{\text{eval,species}}$ set consisted of 325 phages (data available in HostPhinder's repository [48]).

2.3. K-mer-Based Resemblance Measures

Under the assumption that phages infecting the same bacterial host share genomic features, the host of a query phage should be predictable by searching for the most genomically similar phages in a reference database of phages with annotated hosts. The reference database was build from phage genome sequences and their reverse complements by splitting both into k-mers and sliding a window of length k along the sequences with step-size 1.

Query sequences were likewise split into k-mers, and for each reference sequence i having at least one k-mer in common with the query, a score, S_i, was defined as the number of identical unique k-mers between query and template. This score was subsequently used to determine the expectation value E_i:

$$E_i = N_{\text{Hits}} \frac{l_{u,i}}{L_{u,\text{tot}}} \tag{1}$$

where N_{Hits} is the sum of scores over all references, $l_{u,i}$, is the total number of unique k-mers found in the reference sequence i and in its reverse complement and $L_{u,\text{tot}}$ is the sum of unique k-mers over all references in the database. This expectation value was used to obtain a z-score:

$$z_i = \frac{S_i - E_i}{\sqrt{S_i + E_i + \eta}} \tag{2}$$

with $\eta = 0.001$ being a pseudocount used to avoid division by zero. Using SciPy, a two-sided p-value was generated from the z-score. All p-values were corrected using the Bonferroni method [50] by multiplying each p-value by the number of reference phages in the database:

$$p_{\text{corr}} = p_i * N_{\text{ref}} \tag{3}$$

where N_{ref} is the number of reference sequences in the database. HostPhinder outputs only significant hits, *i.e.*, $p_{\text{corr}} < 0.05$. Additionally, the values $\text{frac}_{q,i}$ and $\text{frac}_{d,i}$ were estimated. They represent the ratio of the score and the number of unique k-mers in query and reference sequences respectively:

$$\text{frac}_{q,i} = \frac{S_i}{q_{u,i} + \eta} \tag{4}$$

where $q_{u,i}$ is the number of unique query k-mers and $\eta = 0.001$ avoids division by zero. The value of $\text{frac}_{q,i}$, falling between 0 and 1, gives a direct indication of how much of the query sequence matched to the reference phage.

$$\text{frac}_{d,i} = \frac{S_i}{l_{u,i} + \eta} \tag{5}$$

where $l_{u,i}$ is the number of unique k-mers in the reference sequence and in its complement. Therefore, $\text{frac}_{d,i}$ falls between 0.5 and 1 if query and reference are identical, depending on the number of additional unique k-mers found in the reversed complement. The two measures are hence not directly comparable. Finally the coverage was determined as a measure of how much of the reference sequence is covered by the total number of k-mers in the query that match the reference:

$$\text{coverage}_i = \frac{2q_{\text{matched},i}}{l_{u,i} + \eta} \tag{6}$$

where $q_{\text{matched},i}$ is the total number of k-mers in the query that were matched to reference i, and l_i is the total number of k-mers in the reference. Both of these values include identical k-mers and do not only

count unique k-mers. The factor 2 is included to account for the additionally used reverse complement sequence of the reference to obtain l_i. The coverage can be larger than 1 if the query contains k-mers that could be matched multiple times.

2.4. Determining the Measure and Selection Criteria for Final Prediction

As described above, 5 measures were calculated for the similarity of a query phage to each of the phages in the reference database: score, z-score, frac_q, frac_d, and coverage. The optimal measure was determined in a simple 4 fold cross-validation setup. Here in turn, 3 of the 4 data sets were used as reference database for predicting the host for each query phage in the left out test set (see Supplementary Materials Figure S1, left). The host was inferred from the host of the reference phage with the highest value of similarity measure. This was repeated 4 times so that all 4 partitions were used as test set, and an overall performance for the given measure was calculated by concatenating the predictions of the 4 test sets. For each measure the average and interval of confidence was assessed through 100 bootstrap resamplings with replacement for each test set and calculating the overall accuracy. On a pairwise comparison based on 1000 bootstrap resamplings, coverage outperformed the other measures and was therefore chosen for further analysis. A number of different selection criteria can be used for the final prediction of the host of a query phage. We tested and compared the efficacy of 4 selection criteria that are each described in detail below.

2.4.1. Criterion 1: Host of Best-Matching Reference Phage

The host of the reference phage with the highest coverage value was selected as predicted host. This is the selection criterion used above to define the optimal similarity measure.

2.4.2. Criterion 2: Majority Host among Top-10 Reference Phages

As predicted host, the most abundant host among the hosts of the top 10 reference phages with the highest coverage values was selected. In case of a tie, the most abundant host with the highest coverage, was selected.

In cases where the coverage of non-top reference phages is far below the coverage of the top reference phage, it might not be advantageous to consider them in the selection criterion. To accommodate this, two additional criteria, criteria 3 and 4, were developed.

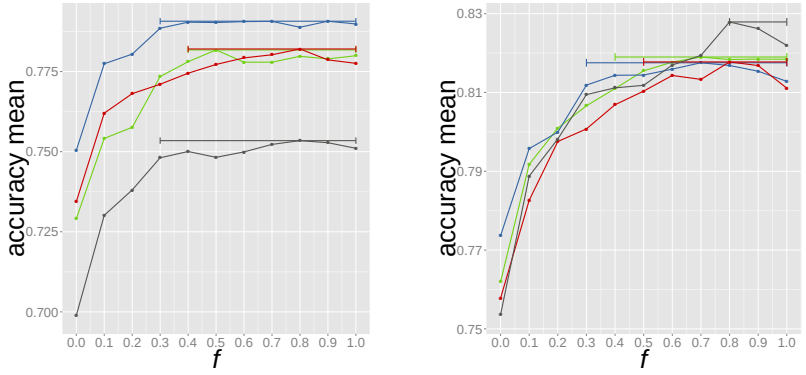
2.4.3. Criterion 3: Majority Host among Reference Phages above Coverage Threshold

As predicted host, the most abundant host among the phages with a coverage value above a given threshold was selected. The threshold was defined as a fraction of the highest coverage:

$$\text{coverage}_{\text{threshold}} = f\,\text{coverage}_1 \tag{7}$$

where f (fraction) is a number in the range 0.0–1.0. Note that $f = 0.0$ means considering all significant predictions, whilst $f = 1.0$ corresponds to selecting the host of the reference phage with the highest coverage (criterion 1). The optimal value of f was determined through a nested 3 fold cross-validation to avoid biased estimates of performances that would result from using the same cross validation used to select the optimal criterion. Here in turn, 3 data partitions were used as tripartite train-test set in a procedure called inner cross-validation. Within the tripartite set, 2 partitions were sequentially used as reference database for predicting the host for the left out test set using Equation (7) for a given value of f. This was repeated 3 times within each tripartite set so that all 3 partitions were used as test set and an overall performance for the given f value was calculated (see Supplementary Materials Figure S5). For each f value the average accuracy was assessed through 100 bootstrap resamplings with replacement for each inner cross validation loop. The same procedure was repeated 4 times so that each tripartite combination was analysed leading to 4 estimates of the optimal f value. The accuracy *vs.* f values curves are shown in Figure 2 for prediction of species and genus. The horizontal bars span

f values that yield at least 99% of the highest accuracy in the relative tripartite combination. Given these performance curves, an f value of 0.8 was chosen within the highest performance range, Figure 2.



(**a**) Species. (**b**) Genus.

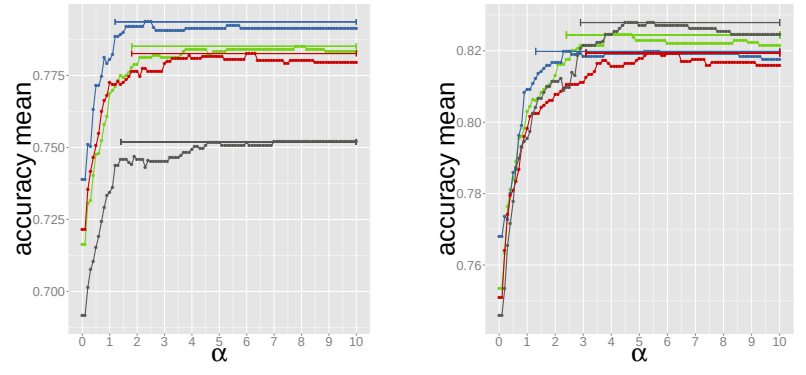
Figure 2. Accuracy *vs.* f values obtained from the 4 loops of inner cross validation. Each dot represents the averaged accuracy for species (a) and genus (b) prediction over 100 bootstrap resamplings. The bars cover the range of f values for which the accuracy is 99% the highest accuracy in the specific tripartite set.

2.4.4. Criterion 4: Summing up Normalized Coverage Values of Phages with Same Host

In the scoring method, coverage values of all significant reference phages were normalised by division by the highest coverage, coverage_1, and raised to the power of an arbitrary number, $\alpha > 0$.

$$\text{score}_i = \left(\frac{\text{coverage}_i}{\text{coverage}_1} \right)^{\alpha} \tag{8}$$

Next, scores of hits with the same host were summed up and the host was predicted as the one with the highest score. The higher the value of α, the higher the score of the first hit, the closer this method is to criterion 1. Values of α in the range 0.0–10.0 were tested. As for the criterion 3, the optimal α was determined through a nested 3 fold cross-validation setup (see Supplementary Materials Figure S5) and led to the selection of $\alpha = 6.0$ within the range that yielded the highest accuracy in the 4 tripartite train-test sets (see Figure 3).



(**a**) Species. (**b**) Genus.

Figure 3. Accuracy *vs.* α values for prediction of species (a) and genus (b) in each tripartite set. Each dot represents the averaged accuracy over 100 bootstrap resamplings. The bars cover the range of α values for which the accuracy is 99% the highest accuracy.

2.5. Programming Language and Speed of Execution

The algorithm was written in Python and Bash.

On an Intel(R) Xeon(R) CPU E5-4610 v2 @ 2.30GHz computer, using 2 cores and 10 GB RAM, HostPhinder average running time is of 61.1662 s for host species prediction and 109.622 s for genus prediction. The longer runtime for genus prediction is due the larger database used for genus predictions. These values were calculated on the evaluation set.

2.6. BLAST Evaluation

The accuracy of the HostPhinder k-mer based approach was compared to the state-of-the-art tool in bioinformatics, BLAST [51]. BLAST performance was assessed on the $\text{phages}_{\text{eval}}$ set using the $\text{phages}_{\text{train-test}}$ set to create a local nucleotide BLAST database. The host associated to the hit with the lowest E-value and secondarily highest bit score was returned as prediction.

2.7. Establishing an Evaluation Set of Predicted Prophages

The PhiSpy prophage prediction tool [52] was used to predict prophages in 2679 complete bacterial genomes collected from NCBI [53]. PhiSpy was run once on each genome resulting in a total of 7559 predicted bacterial prophages in 2074 genomes. Of these, 2796 were from bacterial species that were also included in the HostPhinder reference database. In the following, these predicted prophages will be referred to as the $\text{prophages}_{\text{species}}$ set. A total of 4639 predicted prophages were from genera that were included in the reference database of HostPhinder. They will be referred to as the $\text{prophages}_{\text{genus}}$ set.

Furthermore 261 manually verified prophages were downloaded from PhiSpy and phage_finder directories from Phantome [54] and HostPhinder prediction was tested on them.

2.8. Host Prediction of INTESTI Bacteriophage Cocktail

The Georgian George Eliava Institute of Bacteriophages, Microbiology and Virology has developed phage cocktails (mixtures of phages) since the 1950s. One of these, the INTESTI bacteriophage cocktail, claims to contain sterile filtrates of phage lysates effective against *Staphylococcus*, *Enterococcus*, *Proteus*, *Shigella*, *Salmonella*, *Escherichia coli*, and *Pseudomonas aeruginosa* for the treatment of intestinal bacterial infections. The cocktail was sequenced directly on an Illumina MiSeq platform and de novo assembled to contigs, which were further grouped into 19 draft genomes each hypothesized to represent close to complete phage genomes, and 4 smaller groups hypothesized to represent fragments of phage genomes previously described [38]. The host genus and species of each of these 23 groups was predicted by the final HostPhinder method using the 4th criterion with $\alpha = 6.0$.

3. Results

In this study, we developed and benchmarked HostPhinder, a bioinformatics tool for predicting the bacterial host species of phages. The method is based on the assumption that genetically similar phages are likely to share bacterial hosts. For performing the predictions, HostPhinder relies on a reference database in which WGS data from phages with annotated hosts have been split into k-mers. The genomes of the query phages for which the hosts should be predicted are likewise split into k-mers, and the number of co-occurring k-mers between the query phage and the phages in the reference database is used as a measure of genetic similarity.

3.1. Developing and Benchmarking the HostPhinder Method

Initial analysis on a small dataset indicated that k-mers of length 15–20 nt led to comparable predictive performances. In contrast, shorter k-mers were too unspecific and led to a lower final accuracy, while longer k-mers were too specific and led to more query phages for which no predictions at all could be made (data not shown). Based on these results and a previous study that showed

16-mers to be optimal, when using a k-mer based approach for bacterial species identification [35], 16 was chosen as the k-mer length in the following.

In the initial testing of the basic genetic similarity assumption of HostPhinder, 5 measures were evaluated for estimating the similarity of the query phage to the reference phages as described in Materials and Methods. For each measure, the query host was inferred from the host of the reference hit with the highest similarity. Table 1 shows the performance of each similarity measure in this initial testing.

Table 1. Overall performance of different similarity measures on $\text{phages}_{\text{train-test}}$.

	Score	**z**	**frac$_q$**	**frac$_d$**	**Coverage**
Species (%)	77.03 ± 0.112	77.81 ± 0.111	77.24 ± 0.111	78.43 ± 0.111	$\mathbf{78.76 \pm 0.108}$
Genus (%)	81.43 ± 0.096	82.02 ± 0.094	81.78 ± 0.094	$\mathbf{83.07 \pm 0.09}$	82.84 ± 0.092

The measures' accuracies in predicting the query phage host species of the training-test set were pairwise compared by 1000 bootstrap resamplings with replacement. Coverage performed significantly better than other measures (p-value < 0.05), apart from frac$_d$, which in turn did not significantly outperformed coverage. Since coverage showed the highest performance in predicting the host species, it was chosen as the measure used when further optimizing HostPhinder prediction at the species level. Next, the performance of 4 scoring methods for host selection was compared (see Material and Methods for criteria description and parameter optimization). For each selection criterion only significant hits were considered ($p_{\text{corr}} < 0.05$) and the number of queries with predictions was constant for all criteria allowing a direct comparison of criteria efficacy. Using the model parameters determined above, the 4 criteria were compared in terms of overall accuracy in a 4 fold cross-validation system. In turn, 3 of the 4 partitions were used as reference database for predicting the host for the left out test set using each criterion. This was repeated 4 times so that all 4 partitions were used as test set, and an overall performance for the given criterion was calculated. For each criterion the average and interval of confidence was assessed through 100 bootstrap resamplings with replacement for each test set and calculating the overall accuracy. Table 2 shows the overall accuracy on $\text{phages}_{\text{train-test,genus}}$ and $\text{phages}_{\text{train-test,species}}$ sets for each criterion on genus and species level, respectively. Bacterial host genera and species were not predicted for 5.8% $\text{phages}_{\text{train-test,genus}}$ and 5.6% $\text{phages}_{\text{train-test,species}}$ phages respectively.

Table 2. Average and mean standard error of the overall HostPhinder performance over 100 $\text{phages}_{\text{train-test}}$ set resamplings with replacement.

Method	**Criterion 1 (First Host)**	**Criterion 2 (Majority Host among Top-10)**	**Criterion 3 (Coverage Threshold, $f = 0.8$)**	**Criterion 4 (Summing up Normalized Coverage Values, $\alpha = 6.0$)**
Accuracy, Species (%)	78.76 ± 0.108	74.79 ± 0.102	79.1 ± 0.104	$\mathbf{79.13 \pm 0.105}$
Accuracy, Genus (%)	82.84 ± 0.092	80.41 ± 0.099	83.61 ± 0.092	$\mathbf{83.72 \pm 0.092}$

Criterion 4 with $\alpha = 6.0$ had the highest predictive value, with an accuracy of 79% and 84% for species and genus respectively, even though it only significantly outperforms criterion 2.

Some hosts are substantially more frequent than others in the data set. This could potentially lead to a bias in the prediction, and a subsequent sub-optimal predictive performance. To investigate this, modified versions of criteria 2–4 were tested, where the sequences in the reference database were

clustered according to Hobohm 1 algorithm [49], and only the highest scoring element within one cluster was used in the prediction schema. This did not, however, improve the performance.

Based on the above benchmarking procedures, the final method called HostPhinder was developed. The reference database was generated by splitting all phage genomes in the entire phage set into 16-mers using a step-size of 1. After searching through the database, HostPhinder examines the coverage measure and creates a hits list, *i.e.*, phages significantly similar to the query. The final host species and genus is given according to criterion 4 with an $\alpha = 6.0$. HostPhinder is freely available as a web server [1] and as a Docker image [2].

3.2. Evaluating HostPhinder's Performance on Complete and Partial Genomes

HostPhinder was evaluated on the $\text{phages}_{\text{eval,genus}}$ and $\text{phages}_{\text{eval,species}}$ sets containing phages from public databases. HostPhinder was able to correctly predict the bacterial host species and genera of 74.24% $\pm$ 0.270% and 81.39% $\pm$ 0.206% of the phages respectively. In the evaluation set, 4.0% (3.44%) of the phages could not be matched to any phage in the database when predicting on species (genus) level. We speculated that the accuracy of the HostPhinder method is depending on the coverage value of its prediction. That is, the higher the coverage value, the higher the accuracy. To quantify if this is indeed the case, we show in Figure 4 the accuracy on the evaluation set at different intervals of the coverage value. No hit appeared to have range 0.8 < coverage $\leq$ 0.9 for species. For species as well as genus level, it can be seen that predictions based on a coverage value below 0.1 are only correct for 47% (species) and 63% (genus) of the phages. At the other end of the scale, predictions based on a coverage value above 0.7 (species) and 0.8 (genus) are correct in all instances.

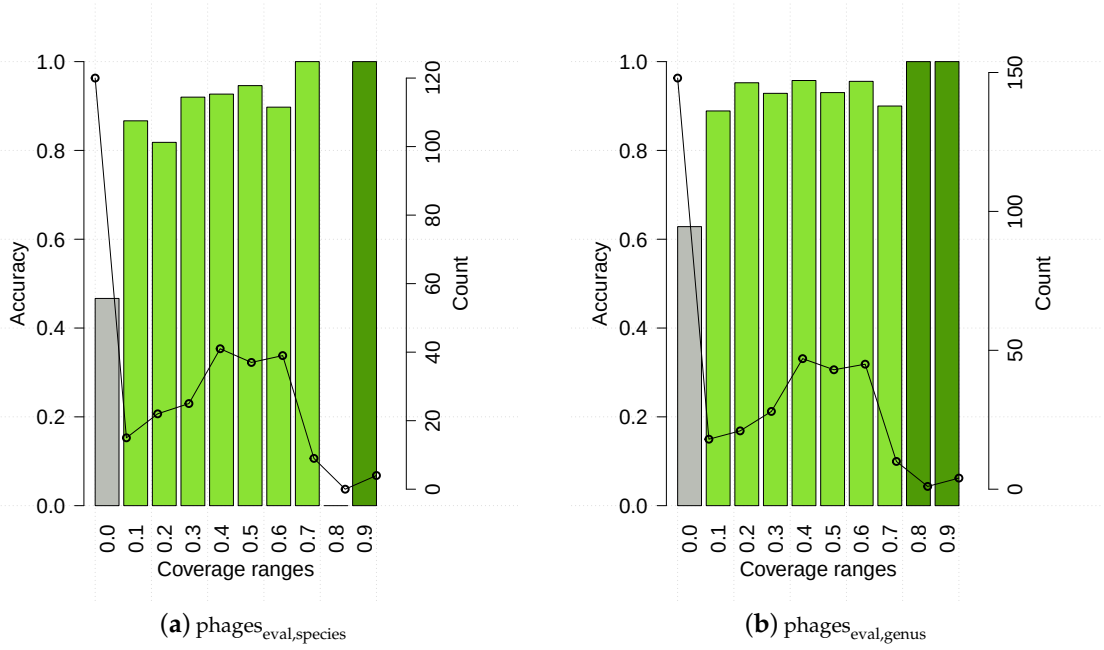
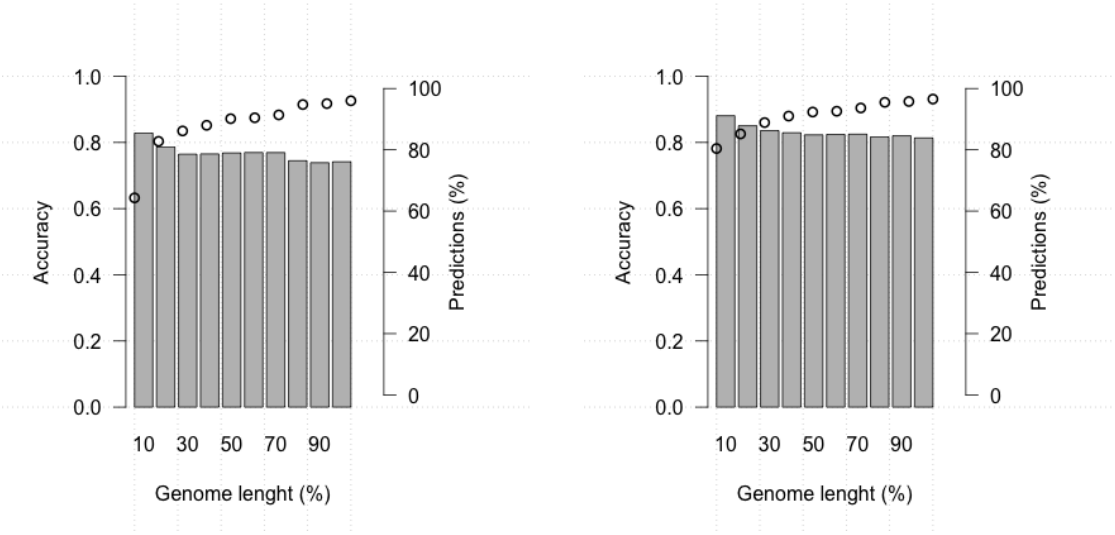


Figure 4. HostPhinder's accuracy (bar) and prediction counts (line) on $\text{phages}_{\text{eval}}$ at different coverage ranges. The values displayed on the x axis are the lower limit of that range. With exception of the last bin which includes all entries with coverage >0.9, all ranges are right-closed with upper limit $x + 0.1$. Poorly reliable results are in grey, while reliable and highly reliable results are in green and dark green respectively. Results on HostPhinder's web server [1] are displayed using the same colour code.

Assembly of metagenomic samples often do not results in entire phage genomes. To assess how the completeness of a phage genome affects HostPhinder performance, we ran the tool on the evaluation set where each genome was gradually reduced by 10%, 20%, ... ,90% of its total length.

Figure 5 shows the accuracy and the number of predictions for each percentage of genome length. HostPhinder maintained the prediction accuracy but made gradually fewer predictions as the fraction of genome given as query is decreased.



(**a**) $\text{phages}_{\text{eval,species}}$ (**b**) $\text{phages}_{\text{eval,genus}}$

Figure 5. HostPhinder's accuracy (bar) and percentages of predictions (dots) on $\text{phages}_{\text{eval}}$ at different percentages of genome length from 10% to 100% of total genome length.

Generally, HostPhinder returned predictions at 10% genome length for those genomes which prediction at complete genome length had a higher coverage. The average coverage for predictions made at complete genome length but not at 10% genome length was 0.023, while the average coverage for commonly predicted was 0.36.

We next examined if HostPhinder always correctly predicted particular host species or genera (Table 3). Only hosts occurring at least 3 times in the $\text{phages}_{\text{eval}}$ set are listed. All phages in the $\text{phages}_{\text{eval}}$ set that target these hosts listed in Table 3 were correctly predicted. Additionally, none of these hosts were erroneously predicted as targets of other phages.

Table 3. List of host species (left) and genera (right), which HostPhinder predicts correctly.

Species	Representation in $\text{phages}_{\text{train-test,species}}$	Genus	Representation in $\text{phages}_{\text{train-test,genus}}$
Enterococcus faecalis	15	*Acinetobacter*	16
Listeria monocytogenes	21	*Listeria*	26
Propionibacterium acnes	21	*Propionibacterium*	24
Vibrio cholerae	35	*Streptococcus*	39
		Streptomyces	11
		Thermus	5

HostPhinder also worked effectively for predicting the host of phages, which according to the initial clustering were of different types; in fact in the HostPhinder dataset there are 14 different types of *Enterococcus faecalis* phages, 13 types of *Listeria monocytogenes* phages and 21 types of *Vibrio cholerae* phages and all phages known to infect these host have been correctly predicted, see Table 3.

Figures 6 and 7 show right and wrong predictions for species and genera respectively. To ease comprehension of the plots, hosts were grouped by phyla, which are displayed on the left side of the figures. Rows are alternatively shaded and column names are enhanced with the same

colour of the phylum of belonging. The heatmaps are read from right to left and then downwards; expressely, the phage related to the host identified by the row name, on the right, was predicted (red intensity of the cell) to infect the host identified by the column name in the lower part of the figure. As an example, *Alteromonas macleodii* phages, the row encompassed in a blue horizontal box in Figure 6, occurred four times in the $\text{phages}_{\text{eval,species}}$ set, as indicated by the number within parenthesis beside the host name, and all of them were wrongly predicted to be *S. aureus* phages (vertical blue box) as indicated by the intense red colour of the square in the intersection between the two blue boxes; of note, there were 69 *S. aureus* phages in the $\text{phages}_{\text{train-test,species}}$ data set and no *Alteromonas macleodii* phages.

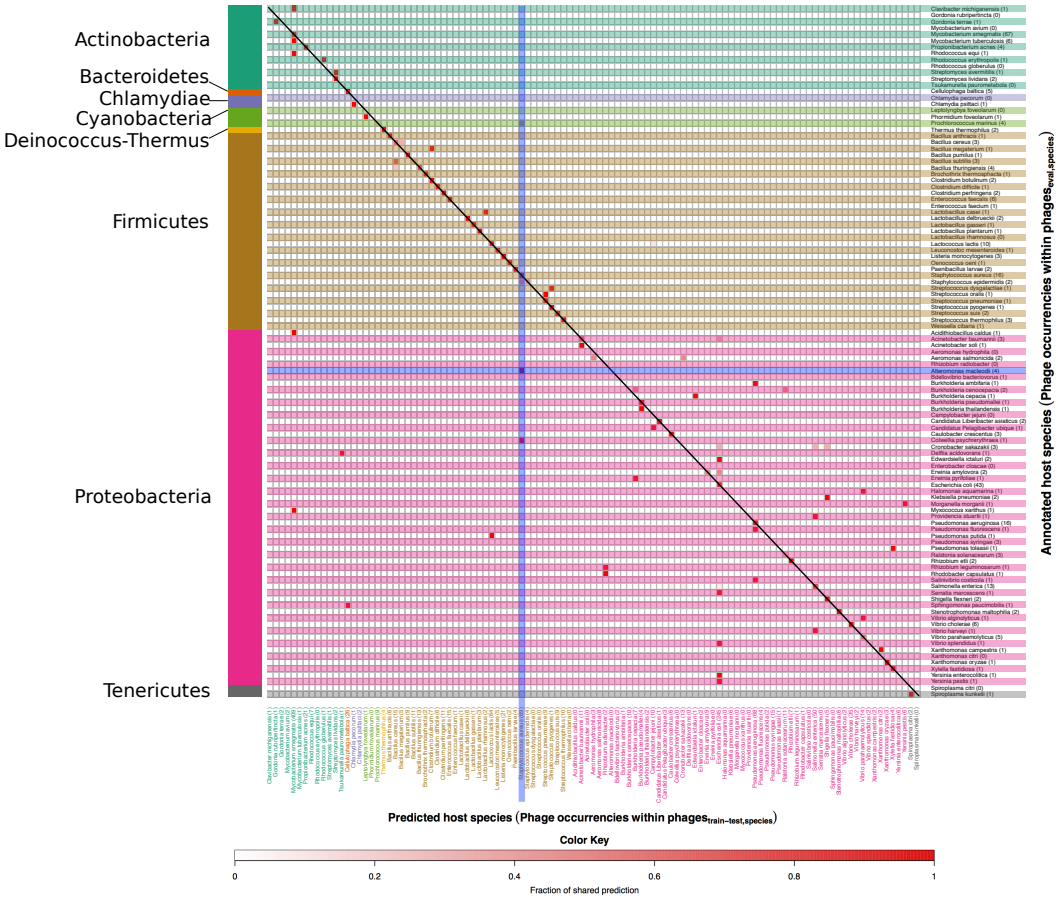


Figure 6. Heatmap of annotated *vs.* predicted host species in the $\text{phages}_{\text{eval,species}}$ set. In this figure correct as well as mispredicted host species can be seen. Annotated host species are listed along the *y* axis, while predicted ones are on the *x* axis. The number after each species on the *y* axis and the *x* axis also indicate the occurrences of phages in the $\text{phage}_{\text{eval,species}}$ and in the $\text{phages}_{\text{train-test}}$ respectively. Host species are grouped according to the respective phylum, which are indicated on the left side of the figure. The colour scale indicates the fraction of phages predicted as targeting a particular host and goes from white, no phages, to red, 100% of the phages. Accordingly, the colour itself is not an indicator of correctness of the prediction, and red colours along the diagonal represent correct predictions.

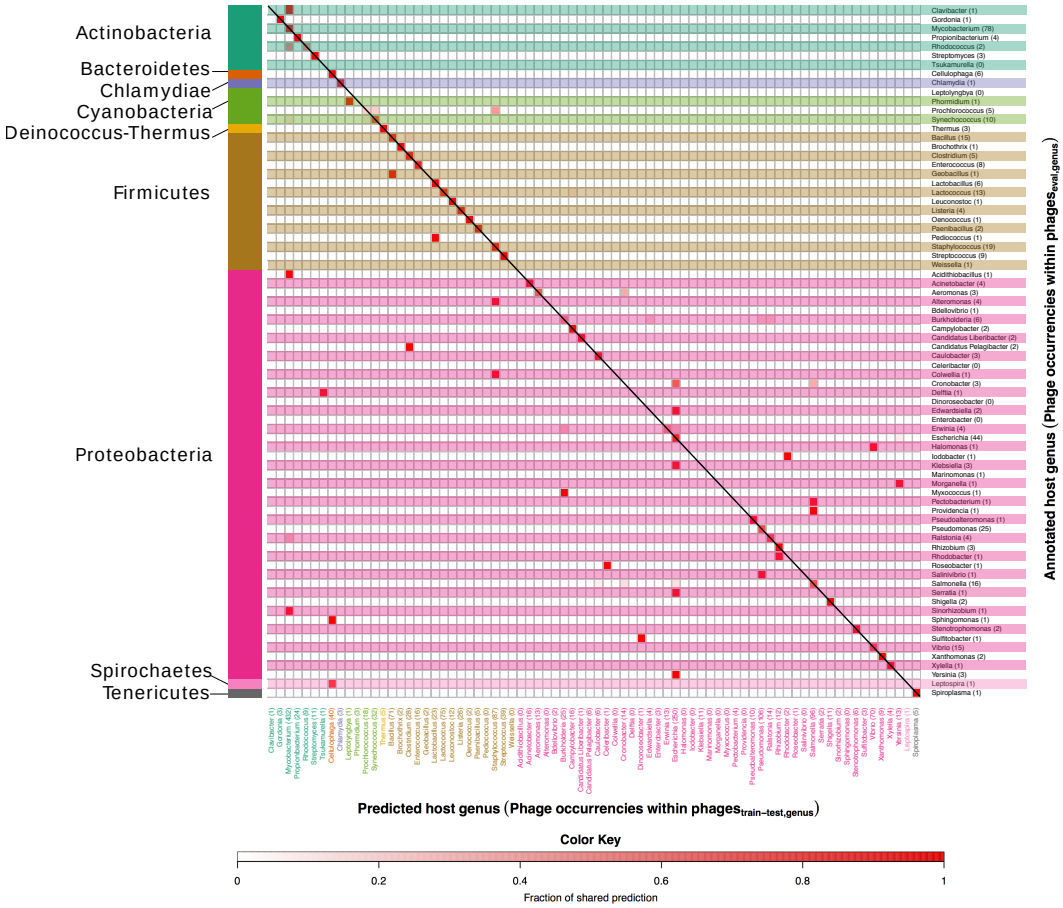


Figure 7. Heatmap of annotated *vs.* predicted host genera in the $\text{phages}_{\text{eval,genera}}$ set. In this figure correct as well as mispredicted host genera can be seen. Annotated host genera are listed along the *y* axis, while predicted ones are on the *x* axis. The number after each genus on the *y* axis and the *x* axis indicate the number of occurrences of phages in the $\text{phage}_{\text{eval,genus}}$ and $\text{phages}_{\text{train-test,genus}}$ respectively. Host genera are grouped according to the respective phylum, which are indicated on the left side of the figure. The colour scale indicates the fraction of phages predicted as targeting a particular host and goes from white, no phages, to intense red, 100% of the phages. Accordingly, the colour is in itself not an indicator of correctness of the prediction, and red colours along the diagonal represent correct predictions.

At species level, phages with mispredicted hosts are often predicted to target a host of the same genus as the annotated host (see small deviations from the diagonal in Figure 6). As examples, the 3 phages annotated to target *Bacillus subtilis* are predicted to target either *B. subtilis* or *Bacillus cereus*. For some phages the mispredicted host is, however, of an entirely different genus, e.g., the phage annotated to target *Yersinia enterocolitica* and the phage annotated to target *Yersinia pestis* are both predicted to target *E. coli*. For species as well as genera there is a tendency that phages with mispredicted hosts are predicted to target the most frequent hosts in the $\text{phages}_{\text{train-test}}$ set, e.g., *E. coli* and *Mycobacterium smegmatis* on species level and *Escherichia* and *Mycobacterium* on genus level. What is important to note is that inaccurate predictions were finding related hosts. For example, imprecise predictions of phages infecting *Proteobacteria* (the ones within the brown region) were still falling within the phylum of *Proteobacteria*. This indicates a relatedness in terms of genome sequence among phages infecting different hosts belonging to the same phylum.

3.3. Comparing HostPhinder to BLAST

Next, the HostPhinder performance on $phages_{eval}$ was compared to BLAST. Table 4 summarises the results.

Table 4. HostPhinder and BLAST performance comparison on the $phages_{eval}$ set.

	BLAST	**HostPhinder**
No. of predictions, training on $phages_{train\text{-}test,genus}$	90%	97%
No. of predictions, training on $phages_{train\text{-}test,species}$	91%	96%
Accuracy on common predictions (GENERA) (%)	84.66 ± 0.188	85.13 ± 0.176
Accuracy on common predictions (SPECIES) (%)	76.92 ± 0.252	78.69 ± 0.237

HostPhinder was able to make host predictions for more phages than the BLAST-based method. For the phages that both methods were able to make a prediction for, HostPhinder outperformed BLAST on both genus and species level. The observed better performance of HostPhinder on species level is significant ($p < 0.05$). HostPhinder correctly predicted 25% among 24 (genera) and 10% among 20 (species) predictions not covered by BLAST. Moreover when inferring the host genus of a phage for which HostPhinder gave no prediction, BLAST match to the most closely related phage resulted in the wrong prediction.

3.4. HostPhinder's Performance on Predicted Prophages and Establishment of Confidence Threshold

To further evaluate the performance of HostPhinder and to establish a confidence threshold for the predictive value, we examined if HostPhinder was able to identify the bacterial hosts of predicted prophages on the premise that prophages are phages that have at one point infected the host that they are currently found in. The predicted prophages provide a dataset diverse enough to define a reliability threshold that can be generalized and applied to previously unseen data. For this purpose, we predicted prophages in 2679 bacterial genomes using PhiSpy [52]. Without any threshold value set, HostPhinder was able to correctly predict approximately 45% and 47% of the species and genus respectively. The accuracy was calculated over the number of phages that HostPhinder was able to make a prediction for.

As for $phages_{eval}$, the results on PhiSpy predicted prophages were binned into coverage ranges (Figure 8, upper panels). The accuracy pattern for prophages generally resembled the one for the evaluation set, *i.e.*, it had low accuracy for coverage ≤ 1, and 100% accuracy above a certain threshold, which in this case is 0.8 for species. There is an unexpected drop in accuracy for coverage values >0.9 (genus), which a bootstrap analysis proved non significant ($p > 0.05$). To further cofirm the thresholds, we ran HostPhinder on 261 manually verified prophages, downloaded from PhAnToMe.org, which resulted in 63.57 % $\pm$ 0.356 % and 78.69 % $\pm$ 0.262 % prediction accuracy of species and genus respectively. Accuracy distribution for this dataset among different coverage ranges can be seen in Figure 8, lower panels. Based on observations $phages_{eval}$ and on prophages, HostPhinder considers trustable results with coverage value higher than 0.1, and it applies a conservative threshold of 0.8 to distinguish highly trustable results.

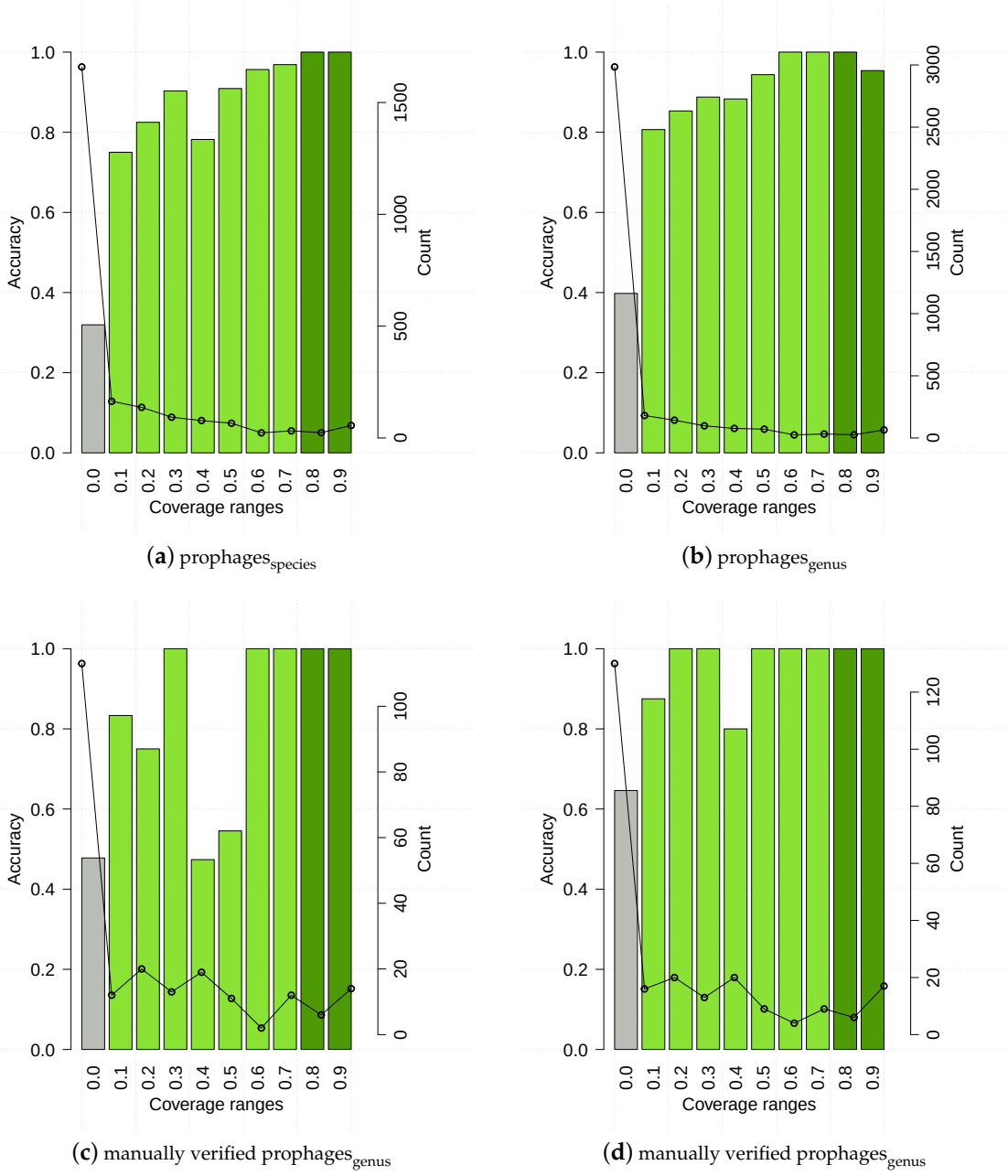


Figure 8. HostPhinder's accuracy (bar) and prediction counts (line) on prophages predicted by PhySpy, upper panels, and manually verified prophages, lower panels, at different coverage ranges. The values displayed on the x axis are the lower limit of that range. With exception of the last bin which includes all entries with coverage >0.9, all ranges are right-closed with upper limit $x + 0.1$. Poorly reliable results are in grey, while reliable and highly reliable results are in green and dark green respectively.

3.5. Host Analysis of Phages from Therapeutic Phage Cocktail from the Georgian George Eliava Institute

In a recent study, we examined the content of an INTESTI bacteriophage cocktail from the Georgian George Eliava Institute. According to the packing, the cocktail is effective against *Staphylococcus*, *Enterococcus*, *Proteus*, *Shigella*, *Salmonella*, *Escherichia coli*, and *Pseudomonas aeruginosa* infections [38]. A total of 19 phage draft genomes were identified that were hypothesized to represent close to complete phage genomes. An additional set of four sequences represented fragments of phage

genomes. Here, we used HostPhinder in an attempt to predict host genera and species of these phage draft genomes and fragments. Table 5 provides an overview.

Table 5. Overview of the results of HostPhinder predicting the hosts of 19 phage draft genomes (name starts with a "D" and *Proteus*) and 4 phage genome fragments (name starts with an "F") from the INTESTI phage cocktail.

Draft ID	Genus	Species	Coverage
D1	*Staphylococcus*	*Staphylococcus aureus*	1.000
D13	*Salmonella*	*Salmonella enterica*	0.840
D5	*Escherichia*	*Escherichia coli*	0.740
D3	*Pseudomonas*	*Pseudomonas aeruginosa*	0.690
D11	*Salmonella*	*Salmonella enterica*	0.610
D10	*Escherichia*	*Escherichia coli*	0.500
D14	*Escherichia*	*Escherichia coli*	0.490
D17	*Escherichia*	*Escherichia coli*	0.460
D9	*Escherichia*	*Escherichia coli*	0.450
D15	*Escherichia*	*Escherichia coli*	0.450
D16	*Sodalis*	*Sodalis glossinidius*	0.430
D4	*Escherichia*	*Escherichia coli*	0.420
D18	*Salmonella*	*Salmonella enterica*	0.380
D8	*Escherichia*	*Shigella flexneri*	0.300
F1	*Pseudomonas*	*Pseudomonas aeruginosa*	0.270
D2	*Escherichia*	*Escherichia coli*	0.250
D7	*Enterococcus*	*Enterococcus faecalis*	0.230
D12	*Enterococcus*	*Enterococcus faecium*	0.180
F2	*Salmonella*	*Salmonella enterica*	0.078
F4	*Escherichia*	*Escherichia coli*	0.014
D6	*Enterococcus*	*Enterococcus faecalis*	0.011
F3	*Escherichia*	*Escherichia coli*	0.011
Proteus	*Escherichia*	*Escherichia coli*	0.003

For six of the seven bacterial targets of the cocktail, HostPhinder predicted at least one phage targeting this type of bacteria. The only bacterium that was not predicted among the hosts was *Proteus*. Instead, the phage that was experimentally found to infect *Proteus* [38], was predicted as an *E. coli* phage with a coverage of 0.0026. This is not surprising, as the HostPhinder database contains no examples of *Proteus* phages. A *Sodalis glossinidius* was predicted, not corresponding to any of the anticipated targets. This bacterium is an endosymbiont of the tsetse fly [50] and its prediction was based on a coverage value of 0.43, where predictions with coverages above 0.2 have approximately 80% chance of being correct (see Figures 4 and 8). The predicted hosts of the 4 phage fragments were generally based on a lower coverage than the 19 phage draft genomes, indicating that these predictions are less certain.

4. Discussion

In the present study, we developed a fast and simple method for prediction of phage hosts. Other studies have previously focused on the identification of phage-host pairs. Experimental methods examining phage-host interactions include mining viral signals from SAG (single amplified genomes) datasets; microfluidic digital PCR and phageFISH [55]. Recently, M. Martínez-García *et al.* combined single-cell genomics and microarrays technology to assign viruses to hosts depending on hybridization

allowing for discovery of new virus-host pairs directly on a metagenomic samples without requiring cultivation or relying on genomic information [56]. In another study, Roux *et al.* developed a bioinformatics tool VirSorter [57], which was able to identify more than 12,000 virus-host linkages from publicly available bacterial and archeal genomes. In their study they analysed the virus-host adaptation in compositions in terms of mono- di- tri- tetra-nucleotide frequency and codon usage [58] showing the strongest signal of adaptation to host genome given by tetranucleotide frequency (TNF). A further classification method for phage host prediction, MGTAXA was developed by Williamson *et al.* in their metagenomic study of the marine microbe in the Indian Ocean [59]. MGTAXA links viral sequences to the highest scoring host taxonomic model based on polynucleotide genome composition similarity between phage and bacterial genomes. The software is not conveniently available anymore (as of December 2015) and we therefore could not compare its performance to HostPhinder's. Finally, a recent publication by Edwards *et al.* reviewed the predictive power of several computational tools for predicting the host of a given phage based on genome information [60]. The authors highlighted the importance of such tools for the characterization of uncoltured virus from metagenomes, and found that homology-based approaches had the strongest signals for predicting phage-host interactions.

HostPhinder bases its predictions on co-occurring k-mers between the query phage genome and the genomes of reference phages with known hosts. Kmer-based approaches have recently been implemented for genome assembly [61], fast classification [62,63] and annotation [64] of metagenomes. Considering the highly mosaic structure of phage genomes, one of the advantages of using k-mers for phage host predictions is that the exact order of genetic elements does not influence the outcome, only their presence or absence.

On an independent evaluation set, HostPhinder was found to perform well, when predicting the hosts of phages currently found in public databases. A remarkable 74% accuracy for the host species and 81% for the host genus were obtained. Some hosts were consistently easier to predict than others. This was for example the case for *P. acnes*, where the host of all annotated *P. acnes* phages in the evaluation set were correctly predicted, while no non-*P. acnes* phages were erroneously predicted as such. The observation is in concordance with previous studies showing that *P. acnes* phages constitute a homogenous group, sharing 85% nucleotide sequence and having similar genome length [65,66]. Furthermore the examined *P. acnes* phages were not able to infect other members of the *Propionibacterium* genus [65,67]. For many of the mispredicted hosts of HostPhinder, the genus of the annotated and predicted host was the same, which might be considered concurrent with the ability of some phages to infect more than one species within a genus. Examples of such broad host range phages are *Salmonella* Phage Felix O1 [68], Mycobacteriophage D29 [69] and *Yersinia* Phage PY100 [70]. It is hence possible that the mispredicted phages are polyvalent, *i.e.*, capable of infecting more than one bacterial species. Alternatively they may represent actual misprediction by HostPhinder caused by closely related phages targeting different host species. In some cases, the host predicted by HostPhinder did not even belong to the same genus as the annotated host, e.g., the three *Yersinia* phages were all predicted to infect *Escherichia* with coverage values that indicate a reliable result, namely 0.57, 0.6 and 0.13. Indeed the genome sequence of the *Y. pestis* phage phiA1122 has been found to be closely related to coliphage T7, sharing 89% nucleotide identity [71]. Despite this high nucleotide identity, PhiA1122 is not able to infect *E. coli*, and has even been used by the Center for Disease Control and Prevention of the United States as a diagnostic agent to identify *Y. pestis* [72].

When applying HostPhinder to phage draft genomes and fragments from the INTESTI phage cocktail, the predicted hosts corresponded well with the advertised targets of the cocktail. One phage draft genome was, however, predicted to target *Sodalis glossinidius*, an endosymbiont of the tsetse fly. Excluding the remote possibility that phages targeting this bacterium has been added to the cocktail, it is likely that the HostPhinder prediction is incorrect or that the phage is able to infect *S. glossinidius* as well as one of the targets of the cocktail. A study by Ho-Won and Kyoung-Ho Kim has shown close relation in comparative genomic and phylogenetic analyses between EP23, a phage that infects

E. coli and *Shigella sonnei* and, SO-1, which infects *S. glossinidius* [73]. It was, however, not examined if the phages were able to cross-infect the hosts.

Many phages have a very narrow host range and only target specific strains within a particular species. This feature has been used extensively previously, when typing, e.g., *S. enterica* [74] and *S. aureus* [75]. HostPhinder is not able to perform predictions beyond species level, partly due to the hosts of most phages in the public databases not being annotated beyond this. Further, to perform predictions down to specific strains of bacteria more factors than the mere genome resemblance would likely have to be taken into account, e.g., by examining the receptor binding proteins, identifying the number of restriction sites in the phage genomes or analysing the CRISPR regions of the host genome.

Another limitation to the performance of HostPhinder is the accuracy of the breadth of annotated host(s) of the references phages. Most of the reference phages had only one annotated host, although many examples exist of phages that are able to infect closely or even distantly related bacteria [76–78]. Further, the performance of HostPhinder depends on the size and completeness of the underlying database. As an example, at the time of compiling the database for this study, no *Proteus* phage genomes were available in public databases. Hence it is inherently impossible for the HostPhinder method to predict any query phage as a *Proteus* phage. Indeed, HostPhinder predicted an experimentally identified *Proteus* phage from the INTESTI phage cocktail as an *E. coli* phage, albeit based on a coverage value of 0.003 indicating that the prediction was not reliable. Carson *et al.* demonstrated the capability of a coli-proteus phage isolated from a Russian cocktail of equally eradicating *E. coli* and *Proteus mirabilis* biofilms [79], evincing the potential of some phages to infect both species. As more phage genomes become available, we will update HostPhinder database to ensure its continued high performance.

Despite the limitations in HostPhinder, we envision that the tool will be useful for narrowing down the list of potential hosts. With the growing availability of metagenome samples, new approaches are necessary to firstly identify phages and secondly, determine their host. Thanks to its capability of promptly identifying potential phage-host interactions, the HostPhinder tool has potential applications in ecology, human gut microbiocenosis studies, and other viral metagenomics analyses, where there is need to shed light on the nature of phages.

The current of HostPhinder is very simple, only taking into account genomic information about the phage. Further development of the tool will expand this, taking the genome of the host into account, which we expect will enable us to make predictions beyond host species level.

5. Conclusions

The current antibiotics resistance crisis warrants new ways to combat bacterial infections. For decades, phage therapy has been used for this purpose in countries belonging to the former Eastern Bloc, and to ensure transfer of the technology to the West, it is important to establish a pool of well-characterized phages. The presented HostPhinder method provides the phage community with an easy-to-use tool for predicting the host genus and species of query phages, usable when searching for phages with appropriate host specificity and for correlating phages and hosts in ecological and metagenomic studies. HostPhinder is freely available as a web server [1] and as a Docker image [2].

Acknowledgments: This work was supported by the Center for Genomic Epidemiology at the Technical University of Denmark and funded by grant 09-067103/DSF from the Danish Council for Strategic Research, grant nr. 14-3056 from Oticon Fonden and grant 14-70-0955 from Otto Moensted Fonden.

Author Contributions: Mette Voldby Larsen conceived the method; Ole Lund wrote the script to read the k-mers; Kortine Annina Kleinheinz developed the preliminary version of the method; Morten Nielsen designed the method optimization; Julia Villarroel downloaded whole genome sequence data, performed method optimization, analysed the data and finalized the method; Henrike Zschach predicted the prophages; Vanessa Isabell Jurtz designed the Hobohm experiments and set up HostPhinder web server; Julia Villarroel built HostPhinder Docker image; Julia Villarroel, Mette Voldby Larsen and Morten Nielsen wrote the paper. All authors contributed in reviewing the paper.

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

References

1. HostPhinder web service. Available online: http://cge.cbs.dtu.dk/services/HostPhinder (accessed on 1 April 2016).
2. HostPhinder Docker image. Available online: https://registry.hub.docker.com/u/julvi/hostphinder (accessed on 1 April 2016).
3. Kapi, A. The evolving threat of antimicrobial resistance: Options for action. *Indian J. Med. Res.* **2014**, *139*, 182–183.
4. WHO. *Antimicrobial Resistance: Global Report on Surveillance*; World Health Organization: Geneva, Switzerland, 2014.
5. Harper, D.; Anderson, J.; Enright, M. Phage therapy: Delivering on the promise. *Ther. Deliv.* **2011**, *2*, 935–947.
6. Kutateladze, M.; Adamia, R. Bacteriophages as potential new therapeutics to replace or supplement antibiotics. *Trends Biotechnol.* **2010**, *28*, 591–595.
7. Kutateladze, M.; Adamia, R. Phage therapy experience at the Eliava Institute. *Méd. Mal. Infect.* **2008**, *38*, 426–430.
8. Miedzybrodzki, R.; Borysowski, J.; Weber-Dabrowska, B.; Fortuna, W.; Letkiewicz, S.; Szufnarowski, K.; Pawelczyk, Z.; Rogóz, P.; Klak, M.; Wojtasik, E.; *et al*. Charpter 3—Clinical aspects of phage therapy. *Adv. Virus Res.* **2012**, *83*, 73–121.
9. Weber-Dąbrowska, B.; Mulczyk, M.; Górski, A. Bacteriophage therapy of bacterial infections: An update of our institute's experience. In *Inflammation*; Springer: Netherlands, 2001; pp. 201–209.
10. Biswas, B.; Adhya, S.; Washart, P.; Paul, B.; Trostel, A.N.; Powell, B.; Carlton, R.; Merril, C.R. Bacteriophage therapy rescues mice bacteremic from a clinical isolate of vancomycin-resistant *Enterococcus faecium*. *Infect. Immun.* **2002**, *70*, 204–210.
11. Capparelli, R.; Parlato, M.; Borriello, G.; Salvatore, P.; Iannelli, D. Experimental phage therapy against *Staphylococcus aureus* in mice. *Antimicrob. Agents Chemother.* **2007**, *51*, 2765–2773.
12. Smith, H.W.; Huggins, M. Successful treatment of experimental *Escherichia coli* infections in mice using phage: Its general superiority over antibiotics. *J. Gen. Microbiol.* **1982**, *128*, 307–318.
13. Wright, A.; Hawkins, C.; Änggård, E.; Harper, D. A controlled clinical trial of a therapeutic bacteriophage preparation in chronic otitis due to antibiotic-resistant *Pseudomonas aeruginosa*; A preliminary report of efficacy. *Clin. Otolaryngol.* **2009**, *34*, 349–357.
14. Matsuzaki, S.; Uchiyama, J.; Takemura-Uchiyama, I.; Daibata, M. Perspective: The age of the phage. *Nature* **2014**, *509*, doi:10.1038/509S9a.
15. Reardon, S. Phage therapy gets revitalized. *Nature* **2014**, *510*, doi:10.1038/510015a.
16. Sulakvelidze, A. Using lytic bacteriophages to eliminate or significantly reduce contamination of food by foodborne bacterial pathogens. *J. Sci. Food Agric.* **2013**, *93*, 3137–3146.
17. Guenther, S.; Huwyler, D.; Richard, S.; Loessner, M.J. Virulent bacteriophage for efficient biocontrol of *Listeria monocytogenes* in ready-to-eat foods. *Appl. Environ. Microbiol.* **2009**, *75*, 93–100.
18. Carrillo, C.L.; Atterbury, R.; El-Shibiny, A.; Connerton, P.; Dillon, E.; Scott, A.; Connerton, I. Bacteriophage therapy to reduce *Campylobacter jejuni* colonization of broiler chickens. *Appl. Environ. Microbiol.* **2005**, *71*, 6554–6563.
19. McLean, S.K.; Dunn, L.A.; Palombo, E.A. Phage inhibition of *Escherichia coli* in ultrahigh-temperature-treated and raw milk. *Foodborne Pathog. Dis.* **2013**, *10*, 956–962.
20. Stern, A.; Sorek, R. The phage-host arms race: Shaping the evolution of microbes. *Bioessays* **2011**, *33*, 43–51.
21. Deveau, H.; Garneau, J.E.; Moineau, S. CRISPR/Cas system and its role in phage-bacteria interactions. *Annu. Rev. Microbiol.* **2010**, *64*, 475–493.
22. Fineran, P.C.; Blower, T.R.; Foulds, I.J.; Humphreys, D.P.; Lilley, K.S.; Salmond, G.P. The phage abortive infection system, ToxIN, functions as a protein-RNA toxin-antitoxin pair. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 894–899.

23. Carbone, A. Codon bias is a major factor explaining phage evolution in translationally biased hosts. *J. Mol. Evol.* **2008**, *66*, 210–223.
24. Blower, T.R.; Evans, T.J.; Przybilski, R.; Fineran, P.C.; Salmond, G.P. Viral evasion of a bacterial suicide system by RNA-based molecular mimicry enables infectious altruism. *PLoS Genet.* **2012**, *8*, e1003023.
25. Labrie, S.J.; Samson, J.E.; Moineau, S. Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.* **2010**, *8*, 317–327.
26. Weitz, J.S.; Hartman, H.; Levin, S.A. Coevolutionary arms races between bacteria and bacteriophage. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 9535–9540.
27. Duffy, S.; Turner, P.E.; Burch, C.L. Pleiotropic costs of niche expansion in the RNA bacteriophage Φ6. *Genetics* **2006**, *172*, 751–757.
28. Amarillas, L.; Cháidez-Quiroz, C.; Sañudo-Barajas, A.; León-Félix, J. Complete genome sequence of a polyvalent bacteriophage, phiKP26, active on *Salmonella* and *Escherichia coli*. *Arch. Virol.* **2013**, *158*, 2395–2398.
29. Loessner, M.J.; Neugirg, E.; Zink, R.; Scherer, S. Isolation, classification and molecular characterization of bacteriophages for Enterobacter species. *J. Gen. Microbiol.* **1993**, *139*, 2627–2633.
30. Koskella, B.; Meaden, S. Understanding bacteriophage specificity in natural microbial communities. *Viruses* **2013**, *5*, 806–823.
31. Casjens, S.R. Diversity among the tailed-bacteriophages that infect the Enterobacteriaceae. *Res. Microbiol.* **2008**, *159*, 340–348.
32. Rohwer, F.; Edwards, R. The Phage Proteomic Tree: A genome-based taxonomy for phage. *J. Bacteriol.* **2002**, *184*, 4529–4535.
33. Jacobs-Sera, D.; Marinelli, L.J.; Bowman, C.; Broussard, G.W.; Bustamante, C.G.; Boyle, M.M.; Petrova, Z.O.; Dedrick, R.M.; Pope, W.H.; Advancing, S.E.A.P.H.; *et al*. On the nature of mycobacteriophage diversity and host preference. *Virology* **2012**, *434*, 187–201.
34. Woese, C.R.; Fox, G.E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 5088–5090.
35. Larsen, M.V.; Cosentino, S.; Lukjancenko, O.; Saputra, D.; Rasmussen, S.; Hasman, H.; Sicheritz-Pontén, T.; Aarestrup, F.M.; Ussery, D.W.; Lund, O. Benchmarking of methods for genomic taxonomy. *J. Clin. Microbiol.* **2014**, *52*, 1529–1539.
36. Hendrix, R.W. Bacteriophage genomics. *Curr. Opin. Microbiol.* **2003**, *6*, 506–511.
37. Lawrence, J.G.; Hatfull, G.F.; Hendrix, R.W. Imbroglios of viral taxonomy: Genetic exchange and failings of phenetic approaches. *J. Bacteriol.* **2002**, *184*, 4891–4905.
38. Zschach, H.; Joensen, K.G.; Lindhard, B.; Lund, O.; Goderdzishvili, M.; Chkonia, I.; Jgenti, G.; Kvatadze, N.; Alavidze, Z.; Kutter, E.M.; *et al*. What can we learn from a metagenomic analysis of a Georgian bacteriophage cocktail? *Viruses* **2015**, *7*, 6570–6589.
39. Nielsen, H.B.; Almeida, M.; Juncker, A.S.; Rasmussen, S.; Li, J.; Sunagawa, S.; Plichta, D.R.; Gautier, L.; Pedersen, A.G.; le Chatelier, E.; *et al*. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **2014**, *32*, 822–828.
40. Phages.ids - VBI mirrors page. Available online: http://mirrors.vbi.vt.edu/mirrors/ftp.ncbi.nih.gov/genomes/IDS/Phages.ids (accessed on 1 April 2016).
41. NCBI viral Genome Resource. Available online: http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi (accessed on 1 April 2016).
42. EMBL EBI phage genomes list. Available online: http://www.ebi.ac.uk/genomes/phage.html (accessed on 1 April 2016).
43. phagesdb for Mycobacteriophages. Available online: http://phagesdb.org/ (accessed on 1 April 2016).
44. phagesdb for Arthrobacter. Available online: http://arthrobacter.phagesdb.org/ (accessed on 1 April 2016).
45. phagesdb for Bacillus. Available online: http://bacillus.phagesdb.org/ (accessed on 1 April 2016).
46. phagesdb for Streptomyces. Available online: http://streptomyces.phagesdb.org/ (accessed on 1 April 2016).
47. Euzéby, J.P. List of Bacterial Names with Standing in Nomenclature: A folder available on the Internet. *Int. J. Syst. Bacteriol.* **1997**, *47*, 590–592.
48. HostPhinder Github repository. Available online: https://github.com/julvi/HostPhinder (accessed on 1 April 2016).
49. Hobohm, U.; Scharf, M.; Schneider, R.; Sander, C. Selection of representative protein data sets. *Protein Sci.* **1992**, *1*, 409–417.

50. Bonferroni, C.E. *Teoria Statistica Delle Classi e Calcolo Delle Probabilita*; Libreria Internazionale Seeber: Firenze, Italy, 1936.
51. 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.
52. Akhter, S.; Aziz, R.K.; Edwards, R.A. PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. *Nucleic Acids Res.* **2012**, *40*, doi:10.1093/nar/gks406.
53. NCBI complete bacterial genomes. Available online: ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/ (accessed on 1 April 2016).
54. Phantome manually verified prophages, dating 14 March 2012. Available online: http://www.phantome.org/Downloads/Prophages/PhiSpy/Manually_Verified/ (accessed on 1 April 2016).
55. Dang, V.T.; Sullivan, M.B. Emerging methods to study bacteriophage infection at the single-cell level. *Front. Microbiol.* **2014**, *5*, doi:10.3389/fmicb.2014.00724.
56. Martínez-García, M.; Santos, F.; Moreno-Paz, M.; Parro, V.; Antón, J. Unveiling viral-host interactions within the 'microbial dark matter'. *Nat. Commun.* **2014**, *5*, doi:10.1038/ncomms5542.
57. Roux, S.; Enault, F.; Hurwitz, B.L.; Sullivan, M.B. VirSorter: Mining viral signal from microbial genomic data. *PeerJ* **2015**, *3*, doi:10.7717/peerj.985.
58. Roux, S.; Hallam, S.J.; Woyke, T.; Sullivan, M.B. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *eLife* **2015**, *4*, doi:10.7554/eLife.08490.
59. Williamson, S.J.; Allen, L.Z.; Lorenzi, H.A.; Fadrosh, D.W.; Brami, D.; Thiagarajan, M.; McCrow, J.P.; Tovchigrechko, A.; Yooseph, S.; Venter, J.C. Metagenomic exploration of viruses throughout the Indian Ocean. *PLoS ONE* **2012**, *7*, e42047.
60. Edwards, R.A.; McNair, K.; Faust, K.; Raes, J.; Dutilh, B.E. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol. Rev.* **2016**, *40*, 258–272.
61. Marçais, G.; Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **2011**, *27*, 764–770.
62. Kawulok, J.; Deorowicz, S. CoMeta: Classification of metagenomes using k-mers. *PLoS ONE* **2015**, *10*, e0121453.
63. Wood, D.E.; Salzberg, S.L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **2014**, *15*, doi:10.1186/gb-2014-15-3-r46.
64. Edwards, R.A.; Olson, R.; Disz, T.; Pusch, G.D.; Vonstein, V.; Stevens, R.; Overbeek, R. Real Time Metagenomics: Using k-mers to annotate metagenomes. *Bioinformatics* **2012**, *28*, 3316–3317.
65. Marinelli, L.J.; Fitz-Gibbon, S.; Hayes, C.; Bowman, C.; Inkeles, M.; Loncaric, A.; Russell, D.A.; Jacobs-Sera, D.; Cokus, S.; Pellegrini, M.; *et al*. *Propionibacterium acnes* bacteriophages display limited genetic diversity and broad killing activity against bacterial skin isolates. *MBio* **2012**, *3*, doi:10.1128/mBio.00279-12 .
66. Liu, J.; Yan, R.; Zhong, Q.; Ngo, S.; Bangayan, N.J.; Nguyen, L.; Lui, T.; Liu, M.; Erfe, M.C.; Craft, N.; *et al*. The diversity and host interactions of *Propionibacterium acnes* bacteriophages on human skin. *ISME J.* **2015**, *9*, 2078–2093.
67. Farrar, M.D.; Howson, K.M.; Bojar, R.A.; West, D.; Towler, J.C.; Parry, J.; Pelton, K.; Holland, K.T. Genome sequence and analysis of a *Propionibacterium acnes* bacteriophage. *J. Bacteriol.* **2007**, *189*, 4161–4167.
68. Kuhn, J.; Suissa, M.; Chiswell, D.; Azriel, A.; Berman, B.; Shahar, D.; Reznick, S.; Sharf, R.; Wyse, J.; Bar-On, T.; *et al*. A bacteriophage reagent for Salmonella: Molecular studies on Felix 01. *Int. J. Food Microbiol.* **2002**, *74*, 217–227.
69. Ford, M.E.; Sarkis, G.J.; Belanger, A.E.; Hendrix, R.W.; Hatfull, G.F. Genome structure of mycobacteriophage D29: Implications for phage evolution. *J. Mol. Biol.* **1998**, *279*, 143–164.
70. Schwudke, D.; Ergin, A.; Michael, K.; Volkmar, S.; Appel, B.; Knabner, D.; Konietzny, A.; Strauch, E. Broad-host-range Yersinia phage PY100: Genome sequence, proteome analysis of virions, and DNA packaging strategy. *J. Bacteriol.* **2008**, *190*, 332–342.
71. Garcia, E.; Elliott, J.M.; Ramanculov, E.; Chain, P.S.; Chu, M.C.; Molineux, I.J. The genome sequence of Yersinia pestis bacteriophage φA1122 reveals an intimate history with the coliphage T3 and T7 genomes. *J. Bacteriol.* **2003**, *185*, 5248–5262.
72. Zhao, X.; Cui, Y.; Yan, Y.; Du, Z.; Tan, Y.; Yang, H.; Bi, Y.; Zhang, P.; Zhou, L.; Zhou, D.; *et al*. Outer membrane proteins Ail and OmpF of Yersinia pestis are involved in the adsorption of T7-related bacteriophage Yep-phi. *J. Virol.* **2013**, *87*, 12260–12269.

73. Chang, H.W.; Kim, K.H. Comparative genomic analysis of bacteriophage EP23 infecting *Shigella sonnei* and *Escherichia coli*. *J. Microbiol.* **2011**, *49*, 927–934.
74. De Lappe, N.; Doran, G.; O'Connor, J.; O'Hare, C.; Cormican, M. Characterization of bacteriophages used in the *Salmonella enterica* serovar Enteritidis phage-typing scheme. *J. Med. Microbiol.* **2009**, *58*, 86–93.
75. Hood, A. Phage typing of Staphylococcus aureus. *J. Hyg.* **1953**, *51*, 1–15.
76. Bielke, L.; Higgins, S.; Donoghue, A.; Donoghue, D.; Hargis, B. Salmonella host range of bacteriophages that infect multiple genera. *Poult. Sci.* **2007**, *86*, 2536–2540.
77. Jensen, E.C.; Schrader, H.S.; Rieland, B.; Thompson, T.L.; Lee, K.W.; Nickerson, K.W.; Kokjohn, T.A. Prevalence of broad-host-range lytic bacteriophages of *Sphaerotilus natans*, *Escherichia coli*, and *Pseudomonas aeruginosa*. *Appl. Environ. Microbiol.* **1998**, *64*, 575–580.
78. Olsen, R.H.; Siak, J.S.; Gray, R.H. Characteristics of PRD1, a plasmid-dependent broad host range DNA bacteriophage. *J. Virol.* **1974**, *14*, 689–699.
79. Carson, L.; Gorman, S.P.; Gilmore, B.F. The use of lytic bacteriophages in the prevention and eradication of biofilms of *Proteus mirabilis* and *Escherichia coli*. *FEMS Immunol. Med. Microbiol.* **2010**, *59*, 447–455.