bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Seeker: Alignment-free identification of bacteriophage genomes by deep learning

Noam Auslander[1,*], Ayal B. Gussow[1,*#], Sean Benler[1], Yuri I. Wolf[1], Eugene V. Koonin[1,#]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

* These authors contributed equally
[#]for correspondence: koonin@ncbi.nlm.nih.gov, ayal.gussow@nih.gov

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Summary

Advances in metagenomics enable massive discovery of diverse, distinct microbes and viruses. Bacteriophages, the most abundant biological entity on Earth, evolve rapidly, and therefore, detection of unknown bacteriophages in sequence datasets is a challenge. The existing methods rely on sequence similarity to known bacteriophage sequences, impeding the identification and characterization of distinct bacteriophage families. We present Seeker, a deep-learning tool for reference-free identification of phage sequences. Seeker allows rapid detection of phages in sequence datasets and clean differentiation of phage sequences from bacterial ones, even for phages with little sequence similarity to established phage families. We comprehensively validate Seeker's ability to identify unknown phages and employ Seeker to detect unknown phages, some of which are highly divergent from known phage families. We provide a web portal (seeker.pythonanywhere.com) and a user-friendly python package (https://github.com/gussow/seeker) allowing researchers to easily apply Seeker in metagenomic studies, for the detection of diverse unknown bacteriophages.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

Bacteriophages, viruses that infect bacteria (phages, for short), are ubiquitous and abundant in every type of biome, and their interactions with microbial communities heavily influence microbial ecology, impact biogeochemical cycling in various ecosystems, and to a large extent, shape the evolution of cellular organisms (Busby et al., 2013; Edwards and Rohwer, 2005; Fuhrman, 1999; Gilbert et al., 2018; Reyes et al., 2012; Rodriguez-Valera et al., 2009; Rohwer and Thurber, 2009; Wommack and Colwell, 2000). Recently, the development of non-culture based, metagenomic sequencing has allowed researchers to detect numerous, diverse bacteriophages in sequence data from almost every environment, further demonstrating their broad impact on the functions of microbial communities, such as, for example, animal gut, soil, and ocean microbiomes. In particular, it has been recently shown that the human gut microbiota harbors abundant bacteriophages (Hurwitz et al., 2016) that profoundly influence human metabolism and immunity (Cani et al., 2009; Kernbauer et al., 2014; Norman et al., 2015), with clear therapeutic implications (Kumarasamy et al., 2010; Norman et al., 2015; Reyes et al., 2012) for diseases such as irritable-bowel syndrome and non-alcoholic fatty liver disease (Tripathi et al., 2018). Yet, our understanding of the viral diversity in the majority of microbial communities is limited, given that most of the microbes from such communities have not been cultivated, complicating virus discovery (Delwart, 2007).

Metagenomic studies using high throughput sequencing technology generate ample amounts of short read sequences from prokaryotic cells in microbial communities regardless of the cultivability of cells. Hence, multiple new viruses can be discovered from metagenomic sequencing data, substantially advancing our knowledge of the virus diversity in different types of communities (Simmonds et al., 2017). However, to characterize habitat-specific viromes, it is essential to efficiently extract viral sequences from complex mixtures of virus and host sequences. The existing tools for the identification of phages and prophages rely on sequence similarity (Akhter et al., 2012; Arndt et al., 2016; Fouts, 2006; Lima-Mendez et al., 2008; Roux et al., 2015; Zhou et al., 2011), gene prediction (Arndt et al., 2016; Roux et al., 2015; Zhou et al., 2011) or distribution of nucleotide *k*-mers and specific sequence signatures (Akhter et al., 2012; Ren et al., 2017). Due to this knowledge-based approach, the available methods are largely limited to the detection of viruses with sequences significantly similar to those of already known viruses. Because of the high evolution rate typical of virus genome sequences, distinct groups of viruses often have little in common with previously recognized viruses, impeding their identification by sequence similarity, even using the most sensitive of the available methods for protein sequence comparison. Therefore, identification of previously unknown major groups of bacteriophages remains an open challenge.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Here we introduce Seeker, an alignment-free method that leverages recent advances in deep learning to detect phages without comparing to reference genomes. Seeker employs Long Short-Term Memory (LSTM) models which, by contrast to other sequence learning methods, maintain a long memory of sequences, and thus can identify distant dependencies within sequences to distinguish phages from bacteria. Seeker is unbiased, reference-free, and is not based on pre-determined sequence features (i.e. genes, repeats, *k*-mers or sequence signatures), but rather, is trained to read through a complete DNA sequence, weighing the likelihood of it belonging to a phage genome. This makes Seeker a fitting choice for learning DNA sequence context including long term dependencies and subtle patterns, more powerfully than any method that explicitly extracts motifs or relies on direct sequence similarity. In addition, Seeker does not require substantive computational resources and can run up to 90 times faster than existing methods. To demonstrate the utility of Seeker on specific test cases, we used this method to identify previously unknown phages from human and sheep gut microbiomes as well as environmental metagenome data. Some of the detected phages are highly divergent from known phage families and at least one will likely become the founder of a distinct phage family or a higher taxon.

We provide a web portal (seeker.pythonanywhere.com) and a python package (https://github.com/gussow/seeker) for the application of Seeker and visualization of the results. This work demonstrates that deep learning tools can overcome the limitations of alignments and gene comparisons for the identification of new phages, and has the potential to facilitate other complex sequence prediction tasks.

Results

Seeker: a method to differentiate phages from bacteria

Seeker employs LSTM networks, a type of neural network that is structured to learn order dependence in prediction problems (Hochreiter and Schmidhuber, 1997). Conceptually, LSTM networks process DNA by looking at each position in the sequence and passing information from one step in the network to the next, thus allowing information to persist. This persistence allows LSTMs to learn subtle patterns in the data which are inaccessible to other existing machine learning methods and contribute to the classification of the input sequences.

To train our model to differentiate phage from bacterial sequence, we required a set of sequences from each category. To this end, we first curated high-confidence data of positive (phages, n=2,232) and negative (bacteria, n=75) samples (the relatively small subset of the available bacterial genomes was used to make the positive and negative

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

datasets comparable in size; see Methods for details, Table S1). To train the models, the genomes in the dataset were segmented into fragments of 1 kilobase pairs (Kbp) which were then converted into vectors to be used as the input for the LSTM models (Methods, Fig. 1). To maximally speed up the convergence of the training process, the input was ordered by training difficulty (Bengio et al., 2009) (see Methods). Following training the models on the high-confidence set, we expanded the training to include a more diversified set of bacteria and phage strains (n=13,443 phages, n=1,269 bacteria; see Methods for details, Table S2). At this point, a test set was set aside from these data for method assessment, and the remaining genomes were divided into training and validation sets (Fig. 1). We assessed the method against the test set and found that the classification scores assigned by the two models are strongly correlated (Pearson's ρ=0.95), and can distinguish viral from bacterial sequences with high confidence (Python model AUC = 0.91, Matlab model AUC = 0.92, Fig. 1, Table S3).

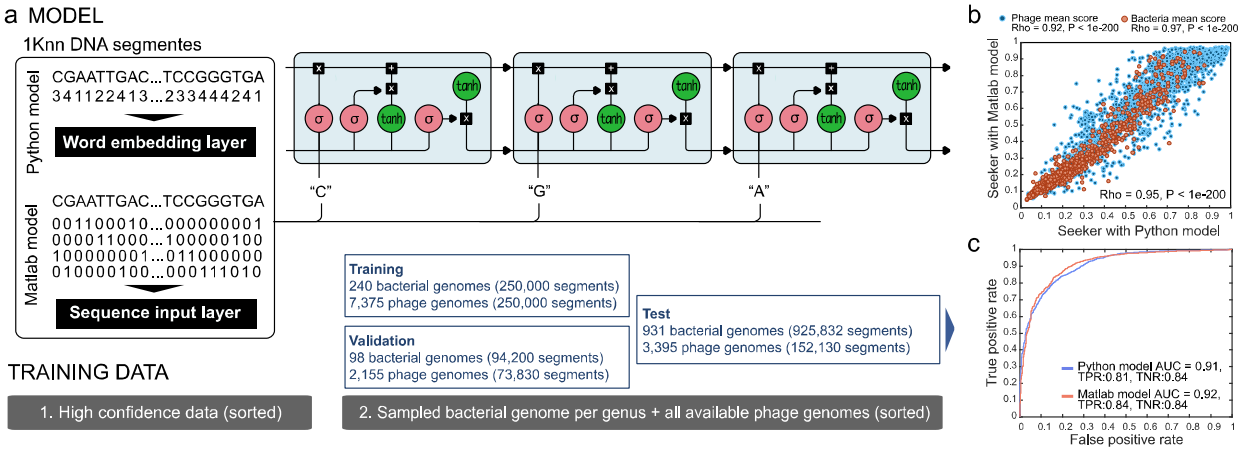

Figure 1. Seeker: a machine learning method for phage genome identification. (a) Models and training pipeline. **(b)** Scatter plot showing the test scores of the model trained with an embedding layer (Python model, x-axis) vs that of the model using a sequence input layer (Matlab model, y-axis). **(c)** ROC (Receiver Operating Characteristic) classification curves predicting phage vs. bacterial genome on the left-out test set.

We next compared Seeker to the existing approaches for phage genome identification focusing on the two most popular methods for metagenomic virus discovery, VirSorter (Roux et al., 2015) and VirFinder (Ren et al., 2017). VirSorter determines whether a DNA sequence is viral by dividing it into genes and applying BLAST to detect similarity to known viral proteins. VirFinder works by searching for specific *k*-mer signatures that were frequently observed in known viral sequences. To produce an unbiased comparison of these methods, we required a test set that was not seen during the

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

training of any of the methods. Thus, we obtained environmental sequences from 6 phage families that were added to NCBI after 2018, and therefore were excluded from the training datasets used to train any of the methods (Table S4). Applied to these phage sequences, Seeker performed substantially better than either VirSorter or VirFinder (Seeker True Positive rate (TPR) = 0.90, VirSorter TPR = 0.37, VirFinder TPR = 0.67; Fig2a, b). In addition, to obtain a test set with even less similarity to the sequences used for training, we tested all three methods on shotgun-sequencing datasets from the NCBI (n=419 phages, n=1042 bacteria, see Methods, Table S5). Once again, we found that Seeker substantially outperformed the other two methods (TPR = 0.96, Fig. 2c). Together, these results demonstrate Seeker's superior ability to detect phages that were not seen during training. In addition, Seeker is much faster than the existing approaches, and its runtime is linear with respect to the input length (Fig 2d).

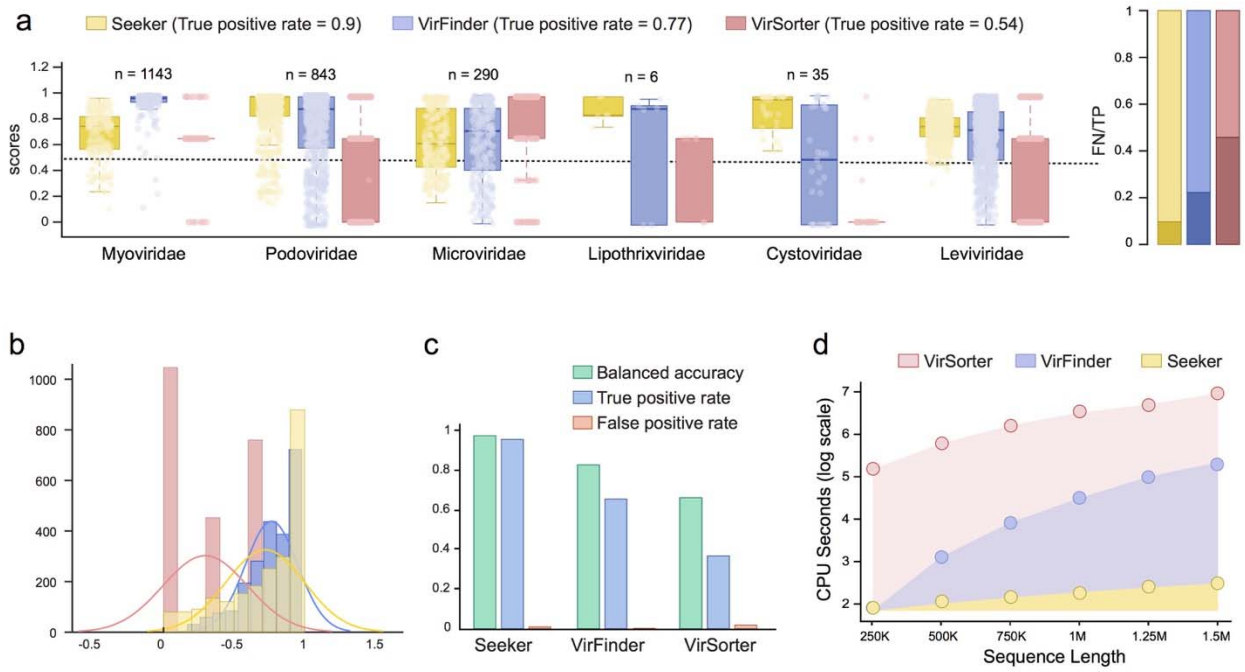

Figure 2. Comparison of the performance of Seeker to those of VirFinder and VirSorter.

(a) Left panel**:** boxplots showing Seeker (yellow), VirFinder (Blue) and VirSorter (red) scores assigned to sequences from 6 bacteriophage families used for testing. The scores are overlaid as dot plots (VirSorter scores are converted from confidence levels to 0-1). Right Panel: bar plots showing the True Positive (lighter shade) vs. False Negative (darker shade) rates of each approach.

(b) Histogram showing the distribution of scores in all sequences in the 6 families (color code similar to panel (a)).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(c) Bar plots showing the balanced accuracy (green), true positive rate (blue) and false positive rate (red) of Seeker, VirFinder and VirSorter on the shotgun sequencing test set.
(d) The CPU seconds (y-axis) of Seeker (yellow), VirFinder (blue) and VirSorter (red) for varying input size (x-axis).

Using Seeker for phage discovery

Encouraged by the results of Seeker testing against diverse sets of known phages, we used this method to search metagenomic sequence datasets for previously undetected phage genomes. We filtered four metagenomic sequencing projects for circular contigs with a high Seeker score, for a total of 367 candidate phage genomes (Table S6). Each candidate was then searched for the protein sequences of three phage markers, i.e. protein-coding genes that are represented in all known tailed phages (Grazziotin et al., 2017), namely, terminase (large subunit), capsid and portal proteins (see Methods for details). We found that, for 311 of the candidates (85%), we were able to detect at least one of these markers (Fig. 3a, Table S6), most often, the terminase, the most conserved of the three protein markers, sequence-wise. The remaining candidates are either not phages and therefore false positives, or contain extremely divergent forms of these markers. In the majority of the candidates where the markers were detected, the sequences of the marker proteins are substantially dissimilar from their closest known homologs, with less than 50% identity (Fig. 3b,c), further indicating the novelty of these phages detected by Seeker.

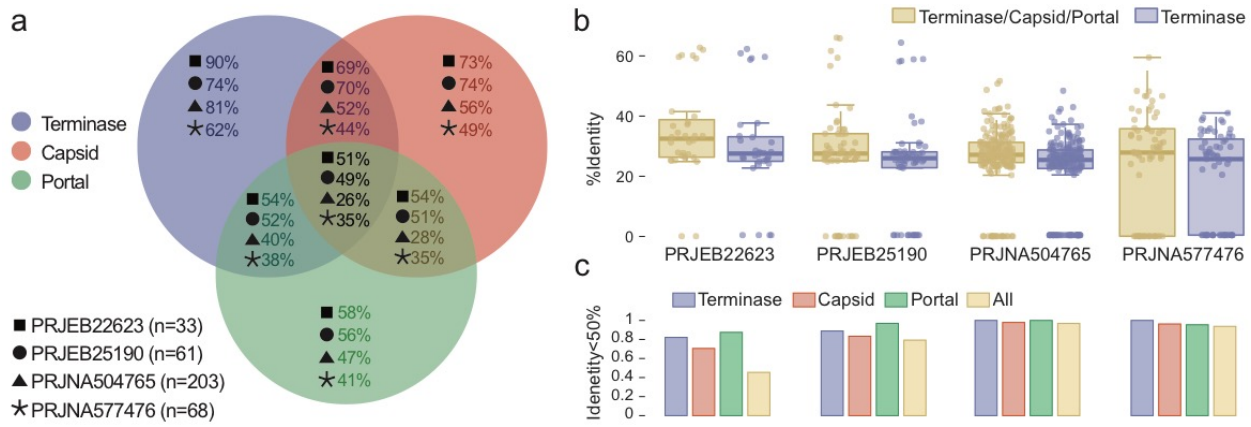

Figure 3. Terminase, capsid and portal proteins in candidate novel bacteriophages.
(a) Ven diagram indicating, for each metagenomic project analyzed (marked with different shapes), the percentage of Seeker identified circular contigs with combinations of the three phage markers (color-coded).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(b) Box plots showing the maximum percent identity of all three markers and their closest known homologs, and that of the terminase separately, for each project.
(c) The proportion of proteins with less than 50% identity to their closest known homologs for each of the three markers, and for all of the three markers combined (color-coded), displayed by project.

We explored in detail 5 of the unknown phages discovered by Seeker in this set, with an explicit focus on the phages that bore the least sequence similarity to known phages (see Methods for details), starting with 2 of the phages detected in gut metagenomes. The first of these (*OLNE01000568.1*), which we refer to as Flitwick, was detected in a human gut metagenome. Flitwick has a 33,716bp circular genome, with 25 predicted genes, and uses an alternative genetic code, with readthrough of amber stop codons. This could, in part, explain why this phage has not been previously identified. We annotated 13 of Flitwick's genes (52%, Fig. 4a, Supplementary Data 1-3), in particular, several encoding structural proteins including the major capsid protein and the large terminase subunit. We additionally detected four tRNA genes in the phage genome one of which is predicted to be the suppressor of the amber stop codon. The position of Flitwick in the phylogenetic tree of the large terminase subunit shows that this is a distinct member of the *Siphoviridae* family (Fig. 4b).

The second phage (*ODAI012083904.1*), which we refer to as Regulus, was detected in sheep rumen metagenome (Methods). Regulus has a 432,079bp circular genome and is thus a previously unknown "jumbo" phage (Yuan and Gao, 2017), with 554 predicted genes, of which we were able to annotate 127 (23%). Regulus also uses an amber-readthrough genetic code. In the phylogenetic tree of the large terminase subunits, Regulus forms a distinct branch in the Myoviridae family (Fig. 4d).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

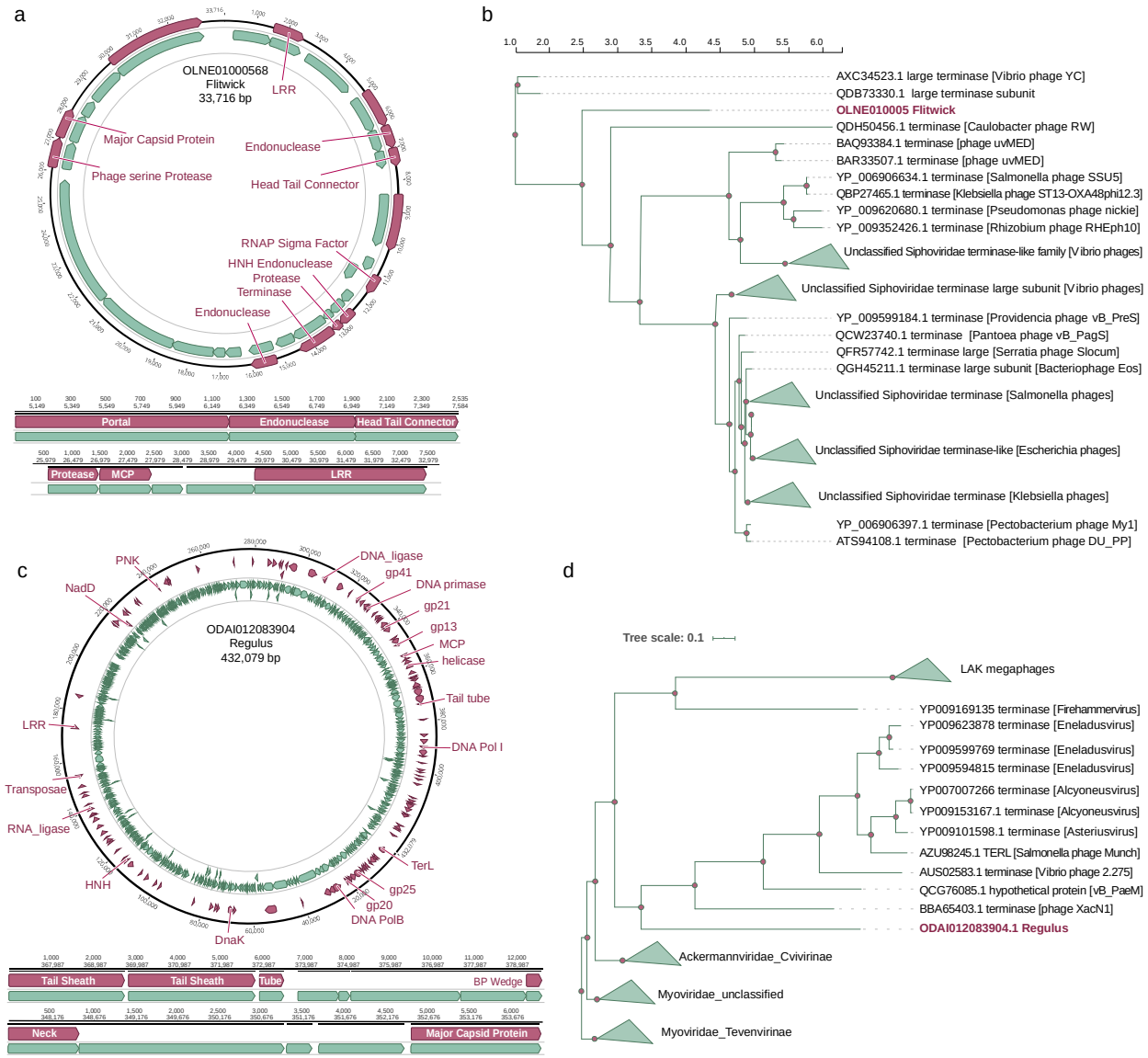

Figure 4. Novel bacteriophages identified by Seeker in gut metagenomes.

(a) Annotated gene map of the Flitwick phage.

(b) Phylogenetic tree of large terminase subunit for the relatives of the Flitwick phage.

(c) Annotated gene map of the Regulus phage.

(d) Phylogenetic tree of large terminase subunits for the relatives of the Regulus phage.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Identification of these phages with Seeker illustrates its ability to detect phage sequences that are distantly related to phages that were seen during training, and additionally, demonstrates that Seeker does not depend on the genetic code used by a phage.

We next explored in detail 3 of the environmental metagenome phages detected by Seeker, all of which are divergent from any known phage family. The first of these (*SDBT011083.1*), which we named Ignotus, has a 46,652bp circular genome with 88 predicted genes, of which 17 (19%) could be annotated (Fig 5a). The predicted terminase and capsid protein are too divergent to be reliably aligned with the other phage terminase or capsid proteins (although recognized at a statistically significant level), and therefore, we were unable to reconstruct a phylogenetic tree (Supplementary Data 1-3). Thus, Ignotus will, probably, become the founder of a distinct phage family or a higher taxon.

The second phage in this set (*WNFG01000004.1*), named Alastor, has a 164,887bp circular genome with 223 predicted genes, of which 68 (31%) could be annotated (Fig. 5b, Supplementary Data 1-3). The major capsid and portal proteins of this phage are moderately similar to proteins in a subset of the phages in the family Herelleviridae, but its terminase is highly divergent from known terminases (Fig. 5d). The last phage we analyzed from this set (*WNGI01000014.1*), named Wulfric, has a 103,078bp circular genome with 133 predicted genes, of which 43 (32%) could be annotated (Fig. 5c, Supplementary Data 1-3). Phylogenetic analysis of the large terminase subunits shows that Wulfric is a distinct member of the family Podoviridae (Fig. 5e).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

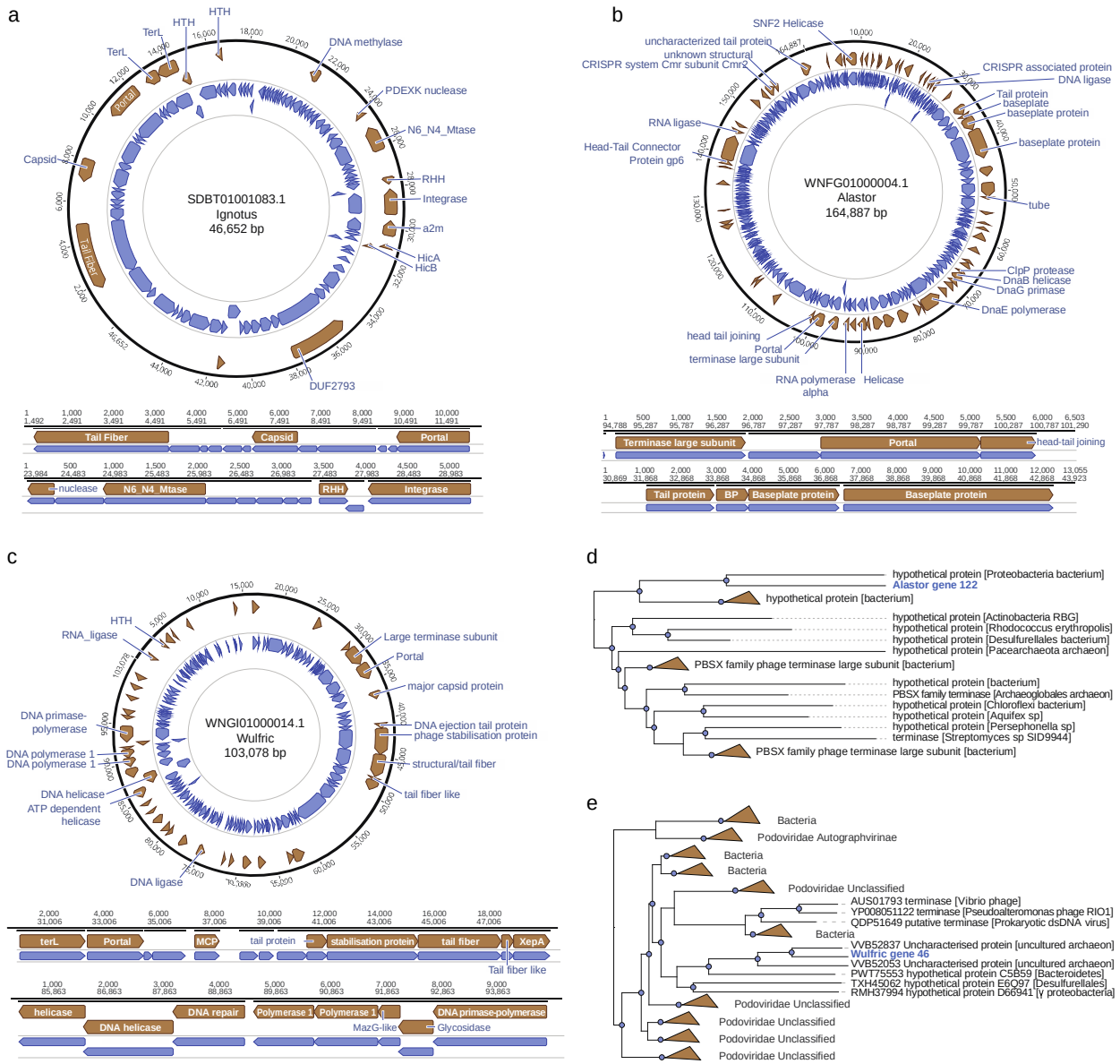

Figure 5. Novel bacteriophages identified by Seeker in environmental metagenomes.

(a)-(c) Annotated gene maps of the phages Ignotus**,** Alastor and Wulfric, respectively.

(d)-(e) Phylogenetic trees constructed from the large terminase subunits of the phages Alastor and Wulfric, respectively.

Discussion

Bacteriophages play a vital role in nearly every ecosystem and, through their presence in microbiomes, directly impact human health. Metagenomic sequencing has brought about a new era of bacteriophage discovery, where the crucial hurdle is the ability to extract viral sequences and discover unknown bacteriophages from a large pool of metagenomic sequences. Existing methods, which detect phage sequences based on

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

direct similarity to the phages present in the current databases, are often inadequate for detection of phages distantly related to the known ones, and are slow when applied to long sequences and large datasets.

Recent advances in deep learning have demonstrated the enormous power these approaches can wield in detecting otherwise opaque patterns and trends in complex datasets (Eraslan et al., 2019). Here, we utilize LSTM neural networks that, to our knowledge, have not been previously employed to detect the origin of a DNA sequence, to enable reference-free detection of viral sequences from large-scale sequencing data. A comparison of Seeker with two most widely used methods for phage identification in sequence databases, VirSorter and VirFinder, demonstrated a substantially better performance of Seeker on diverse sets of phage sequences, combined with a much higher speed. Furthermore, in contrast to the existing approaches, Seeker maintains an equally high level of performance when applied to viral sequences with little similarity to those seen during its training. This feature is exemplified by the previously unknown phages detected with Seeker, some of which are highly divergent from known phage families, and at least one of which will likely become the founder of a distinct phage family or a higher taxon.

Seeker is freely and publicly available, as a webtool (seeker.pythonanywhere.com), a command-line tool and a Python package (https://github.com/gussow/seeker). Researchers can easily utilize any of these avenues to process their metagenomic datasets and rapidly discover previously unknown phages. Given its ability to detect a wide diversity of bacteriophages, we expect that widespread application of Seeker leads to the discovery of numerous phages, some of which would represent distinct families or even higher taxa in the forthcoming new phage taxonomy (Adriaenssens et al., 2020). This work demonstrates that LSTM neural networks can learn long-term dependencies within DNA sequences and thus can efficiently tackle tasks that are not easily amenable to standard techniques based on explicit sequence similarity. Future studies are warranted to evaluate the approach developed here for other sequence categories.

STAR Methods

Long Short-Term Memory (LSTM) machines

Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Networks that take a sequence as input for various prediction tasks (Hochreiter and Schmidhuber, 1997). Each LSTM network unit defines a nucleotide n_i in a sequence, and is composed of the following components:

1. $f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f)$
2. $i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i)$

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

3. $so_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o)$
4. $c_t = f_t \odot c_{t-1} + i_t \odot \sigma_c(W_f x_t + U_f h_{t-1} + b_f)$
5. $h_t = f_t \odot \sigma_c(c_t)$

where the initial values are $c_0 = 0$ and $h_0 = 0$. $\odot$ denotes the Hadamard product. X_t are the input vectors to the LSTM unit. f_t, i_t and o_t are the activation vectors for the forget gate, input gate and output gate, respectively. h_t is the output vector of the LSTM unit, and c_t is cell state vector. W and U are the weight matrices and b are the bias matrices that are learned during training. σ are the non-linear functions, where σ_g is a sigmoid function and σ_c is the $tanh$ function.

All LSTM networks used in this work are sequence-to-label LSTMs with 5 hidden layers, with a softmax and classification layer and were trained using Adam optimizer (Kingma and Ba, 2015), where the maximal epoch for training is set to 100. The mini batch size used for each training iteration was set to 27, with a standard gradient-clipping threshold set to 1.

All DNA sequence data was consecutively segmented into non-overlapping sequences of length 1 Kbp and used as input for two different types of layers, which were trained to input data into the LSTM layers:

a. **Python Keras word embedding layer**, for which the DNA sequence input was transformed into integers ("A" = 1, "T" = 2, "C" = 3 and "G" = 4), with vocabulary size 5 and input length defined to 1,000,
b. **Matlab sequence input layer,** using channel-wise normalization for zero-center normalization, for which the DNA sequence input was transformed using one-hot encoding ("A" = 1000, "T" = 0100, "C" = 0010 and "G" = 0001). This model was converted into a Keras model once the training was completed.

<u>Data Curation</u>

a. Data used for training step 1. The positive set was obtained from RefSeq(O'Leary et al., 2016) and consisted of 80% of all RefSeq phages (n=2,232 phages, Table S1). For the negative set, we curated a high-confidence, non-redundant set from the reference bacteria set (n=75 bacteria, from the *ncbi-genome-download* project). From the latter, all instances of known phage and prophage sequences were removed using exact match of at least 100 nucleotides with any phage in our positive set or in phage sequences obtained from the PHASTER database (Arndt et al., 2016). The phages were randomly subset from a total of 2,750 RefSeq phages to balance the positive and negative sets, yielding n=80,000 phage and bacteria fragments.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

b. Data used for training step 2. To expand the positive set, we obtained an additional larger set consisting of all annotated complete genome phages found in an exhaustive search of online databases (https://www.ncbi.nlm.nih.gov/, https://www.ebi.ac.uk/genomes/phage.html).
For bacterial genomes, we randomly sampled a single representative per bacterial genus (Tables S2, S3) and downloaded the genomes from NCBI (n=1,269 bacteria; 13,443 phages). 240 bacterial and 7,375 phage genomes were used for training (yielding 250,000 phage and bacteria fragments), 98 bacterial and 2,155 phage genomes were used as validation set, and 931 bacterial and 3,931 phage genomes were left out for testing and never included in training

Sorting the training data by difficulty

a. Sorting data from training step 1. We aim to sort our initial high-confidence data by training difficulty (from easy to hard), to speed up the convergence of the training process and hence reduce the risk of overfitting (Bengio et al., 2009). To this end, we approximate the difficulty of a training sample (phage or bacterial) by the average ROC AUC obtained with LSTMs trained on its genome. We hence trained LSTMs from randomly chosen combinations of phage and bacterial genomes in the high-confidence training, such that each phage or bacterial genome was used to train 5 models, where in each iteration, the performance was evaluated on the rest of the training set. Then, each phage or bacterial sequence in the set was assigned a score indicating the average performance of a network trained using it; the training data was sorted by these scores, and given as input to the LSTMs for the first training phase.
b. Sorting data from training step 2. For the second step of training, each sample was assigned a value indicating the average performance of the LSTM networks generated from training step 1 on all its 1 Kbp segments. The step 2 training data was ordered by this objective and given as input to the LSTMs for the second training phase.

Comparison to VirFinder and VirSorter

To compare the performance of Seeker for phage detection to those of VirFinder and VirSorter, we first obtained sequences that were submitted to NCBI after 2018 (hence were not available for training VirFinder and VirSorter) and were not included in training Seeker either. We examined different families of phages, and included families with more than 5 phage sequences of length>750bp submitted after 2018, yielding 2270 genomes from 6 families. We applied Seeker, VirFinder and VirSorter to these genomes

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

and calculated the detection (true positive) rate. Seeker and VirFinder output continuous scores between 0 and 1, so we used a cutoff of 0.5 to determine whether a sequence comes from a phage. VirSorter outputs phages predicted at different confidence levels (1-3), so we considered all confidence levels as a positive prediction.

Second, to obtain a dataset with less similarity to the sequences used for training by these methods, we obtained shotgun-sequencing datasets from NCBI (all annotated shotgun phage sequences n=419, and 1042 unclassified shotgun bacterial sequences added to NCBI after 2017, to reduce class imbalance in this set). We applied Seeker, VirFinder and VirSorter to these genomes and calculated true and false positive rates and balanced accuracies, using similar thresholds.

To compare the runtime of Seeker to those of VirFinder and VirSorter, we downloaded a bacterial genome from NCBI (*NC_011750.1*) and created six segments from its nucleotide sequence, starting from the first base and continuing in steps of 250,000, so that the first segment was 250 Kbp, the second one was 500 Kbp, and so forth, until the sixth one (1,500 Kbp). Each segment was used as input to each of the three methods and recorded the number of CPU seconds for each run.

Identification of unknown bacteriophages

To identify unknown bacteriophages, we applied Seeker to unclassified metagenomic data from 4 projects. We searched for circular sequences (those with a direct overlap > 15 bp at the genome termini) of length >30kb, assigned with high Seeker scores (top 10% of each database and larger than 0.7), yielding 367 contigs in total (33, 61, 203 and 68 from PRJEB22623, PRJEB25190, PRJNA504765 and PRJNA577476, respectively, Table S6). To quantify the proportion of unknown phages within these datasets, we ran six frame translation on each contig, ignoring stop codons, and PSI-blasted (Altschul et al., 1997) the resulting proteins against CDD (Lu et al., 2020) and PVOG (Grazziotin et al., 2017) with E-value cutoff of 0.1. in addition, we applied BlastX to each contig, with E-value cutoff of 1E-4. From these, hits to terminase, capsid and portal proteins were retained, where 311 contigs (85%) had a hit to at least one of these three. For each of identified protein, the maximum percent of identity was obtained using BlastX against NR (Table S6).

The resulting sequences were then filtered to include only those with less than 1% overlap with existing phage sequences (using blastn, query coverage less than 1%). The protein sequences of these candidates were predicted using Prodigal (Hyatt et al., 2010) (v2.6.3) with the parameter set for metagenome mode (-p meta). The protein sequences of these candidates were compared to the phage subset of the NR protein database (accessed 12/2020) using BlastP. We filtered for candidates in which fewer

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

than 50% of the predicted proteins had BlastP hits to the proteins in this database and less than 33% of the proteins had hits to a single phage family (with E-value < 1e-6). The candidates that met these criteria were taken to represent 'unknown' phages.

Characterization of unknown bacteriophages

Each predicted protein sequence of the candidate phages was used as a query for psi-blast (Altschul et al., 1997) against the NR database (accessed 12/2019) to construct a multiple sequence alignment (MSA). The resulting MSA was used as a query against the NCBI CDD database (accessed 12/2019 (Lu et al., 2020)) with an E-value cutoff of < 0.1. Additional annotations were generated with hhblits (Remmert et al., 2012), using the MSAs constructed above as queries to search the PDB database clustered to 70% maximum pairwise sequence identity (downloaded from http://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs/, accessed 12/2019). The presence of tRNAs on selected contigs was assessed for the amber-readthrough genomes using tRNA-scan-SE (Lowe and Eddy, 1996) (v2.0) with a bitscore cutoff of 35.

For the amber-readthrough phages, the initial prediction of protein-coding genes was performed using the standard genetic code. In both cases, the following was observed: (a) the homolog of the large terminase subunit was small and contained a TAG stop codon but successfully aligned to known terminases when the stop codon was ignored; (b) when translating with an amber-readthrough genetic code, the size of most of the genes substantially increased, enabling us to annotate genes for which otherwise no homologs were detected. Given these lines of evidence, these phages are assumed to be using an amber-readthrough genetic code, as previously documented for several gut metagenomes (Ivanova et al., 2014).

Phylogenetic trees construction

For each candidate phage and its relatives, a phylogenetic tree was constructed based on the predicted terminase large subunit protein sequence. To create an alignment, terminase sequence was run against NR using PSI-Blast, and sequences with e-value < 0.01 were retrieved. Sequences were then clustered at 70% identity using mmclust (Steinegger and Söding, 2017). The resulting sequences were aligned using muscle(Edgar, 2004), and then filtered to include those with less than 50% gaps. The resulting alignment was used to construct a tree using FastTree, with default parameters (Price et al., 2009).

Acknowledgements

The authors thank Kira S. Makarova for helpful discussions. This research was supported by the Intramural Research Program of the National Library of Medicine at the NIH. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

Adriaenssens, E.M., Sullivan, M.B., Knezevic, P., van Zyl, L.J., Sarkar, B.L., Dutilh, B.E., Alfenas-Zerbini, P., Łobocka, M., Tong, Y., Brister, J.R., et al. (2020). Taxonomy of prokaryotic viruses: 2018-2019 update from the ICTV Bacterial and Archaeal Viruses Subcommittee. Arch. Virol.

Akhter, S., Aziz, R.K., and Edwards, R.A. (2012). PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. Nucleic Acids Res.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res.

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

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In ACM International Conference Proceeding Series, p.

Busby, B., Kristensen, D.M., and Koonin, E. V. (2013). Contribution of phage-derived genomic islands to the virulence of facultative bacterial pathogens. Environ. Microbiol.

Cani, P.D., Possemiers, S., Van De Wiele, T., Guiot, Y., Everard, A., Rottier, O., Geurts, L., Naslain, D., Neyrinck, A., Lambert, D.M., et al. (2009). Changes in gut microbiota control inflammation in obese mice through a mechanism involving GLP-2-driven improvement of gut permeability. Gut.

Delwart, E.L. (2007). Viral metagenomics. Rev. Med. Virol.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Edwards, R.A., and Rohwer, F. (2005). Viral metagenomics. Nat. Rev. Microbiol.

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F.J. (2019). Deep learning: new computational modelling techniques for genomics. Nat. Rev. Genet. *20*, 389–403.

Fouts, D.E. (2006). Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res.

Fuhrman, J.A. (1999). Marine viruses and their biogeochemical and ecological effects. Nature.

Gilbert, J.A., Blaser, M.J., Caporaso, J.G., Jansson, J.K., Lynch, S. V., and Knight, R. (2018). Current understanding of the human microbiome. Nat. Med.

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

Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. Neural Comput. *9*, 1735–1780.

Hurwitz, B.L., U'Ren, J.M., and Youens-Clark, K. (2016). Computational prospecting the great viral unknown. FEMS Microbiol. Lett.

Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics.

Ivanova, N.N., Schwientek, P., Tripp, H.J., Rinke, C., Pati, A., Huntemann, M., Visel, A., Woyke, T., Kyrpides, N.C., and Rubin, E.M. (2014). Stop codon reassignments in the

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

wild. Science (80-.).
Kernbauer, E., Ding, Y., and Cadwell, K. (2014). An enteric virus can replace the beneficial function of commensal bacteria. Nature.
Kingma, D.P., and Ba, J.L. (2015). Adam: a Method for Stochastic Optimization. Int. Conf. Learn. Represent. 2015 1–15.
Kumarasamy, K.K., Toleman, M.A., Walsh, T.R., Bagaria, J., Butt, F., Balakrishnan, R., Chaudhary, U., Doumith, M., Giske, C.G., Irfan, S., et al. (2010). Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: A molecular, biological, and epidemiological study. Lancet Infect. Dis.
Lima-Mendez, G., Van Helden, J., Toussaint, A., and Leplae, R. (2008). Prophinder: A computational tool for prophage prediction in prokaryotic genomes. Bioinformatics.
Lowe, T.M., and Eddy, S.R. (1996). TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res.
Lu, S., Wang, J., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Marchler, G.H., Song, J.S., et al. (2020). CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. *48*, D265–D268.
Norman, J.M., Handley, S.A., Baldridge, M.T., Droit, L., Liu, C.Y., Keller, B.C., Kambal, A., Monaco, C.L., Zhao, G., Fleshner, P., et al. (2015). Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell.
O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res.
Price, M.N., Dehal, P.S., and Arkin, A.P. (2009). Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. *26*, 1641–1650.
Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods.
Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., and Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome.
Reyes, A., Semenkovich, N.P., Whiteson, K., Rohwer, F., and Gordon, J.I. (2012). Going viral: Next-generation sequencing applied to phage populations in the human gut. Nat. Rev. Microbiol.
Rodriguez-Valera, F., Martin-Cuadrado, A.B., Rodriguez-Brito, B., Pašić, L., Thingstad, T.F., Rohwer, F., and Mira, A. (2009). Explaining microbial population genomics through phage predation. Nat. Rev. Microbiol.
Rohwer, F., and Thurber, R.V. (2009). Viruses manipulate the marine environment. Nature.
Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015). VirSorter: Mining viral signal from microbial genomic data. PeerJ.
Simmonds, P., Adams, M.J., Benkő, M., Breitbart, M., Brister, J.R., Carstens, E.B., Davison, A.J., Delwart, E., Gorbalenya, A.E., Harrach, B., et al. (2017). Consensus statement: Virus taxonomy in the age of metagenomics. Nat. Rev. Microbiol. *15*, 161–168.
Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025783; this version posted April 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Tripathi, A., Debelius, J., Brenner, D.A., Karin, M., Loomba, R., Schnabl, B., and Knight, R. (2018). The gut-liver axis and the intersection with the microbiome. Nat. Rev. Gastroenterol. Hepatol.

Wommack, K.E., and Colwell, R.R. (2000). Virioplankton: Viruses in Aquatic Ecosystems. Microbiol. Mol. Biol. Rev.

Yuan, Y., and Gao, M. (2017). Jumbo Bacteriophages: An Overview. Front. Microbiol. *8*, 403.

Zhou, Y., Liang, Y., Lynch, K.H., Dennis, J.J., and Wishart, D.S. (2011). PHAST: A Fast Phage Search Tool. Nucleic Acids Res.