A NETWORK-BASED INTEGRATED FRAMEWORK FOR PREDICTING VIRUS-HOST INTERACTIONS WITH APPLICATIONS

WEILI WANG[1,†], JIE REN[1,†,*], NATHAN A. AHLGREN[2], JED A. FUHRMAN[3], JONATHAN BRAUN[4] & FENGZHU SUN[1,5,*]

[1]*Molecular and Computational Biology Program, University of Southern California, Los Angeles, CA, USA;* [2]*Biology Department, Clark University, Worcester, MA, USA;* [3]*Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA;* [4]*Pathology and Laboratory Medicine, UCLA David Geffen School of Medicine, Los Angeles, CA, USA;* [5]*Centre for Computational Systems Biology, School of Mathematical Sciences, Fudan University, Shanghai, China.*
[†] *Equally contributed;* [*]*Corresponding authors (renj@usc.edu, fsun@usc.edu)*

Abstract. High-throughput metagenomic sequencing technologies have profoundly increased the ability to identify viral genomic sequences without isolation. With the discovery of new viruses, one of the most fundamental challenges is to predict their hosts. A pair of virus and host is referred as interacting if the virus infects the host. In this study, we develop a network-based integrated framework for predicting virus-host interactions based on various networks for virus-virus similarity, virus-host similarity, and known virus-host interactions. We integrate these networks systematically using a Markov random field model, in which the similarity is defined based on both alignment-based similarity measures including CRISPR matches and sequence homology, and alignment-free similarity measures including s_2^* and WIsH, to predict virus-host interactions. We evaluate our method on a benchmark dataset of 1,075 viruses and their hosts. Host prediction accuracy was 61% and 88% at the genus and phylum levels, respectively, representing a large improvement in host prediction accuracy over previous approaches. The model also markedly improves prediction accuracy for the hosts of metagenomically assembled short viral contigs. Requiring a minimum prediction score for making predictions further improves accuracy without much cost of the recall rate. We use the method to predict the hosts of new viruses discovered from single-cell amplified genomes and crAssphage, a ubiquitous virus assembled from human gut metagenomes. The predicted hosts are mostly consistent with the putative hosts, and other plausible hosts including *Akkermansia muciniphila* are suggested. Since *Akkermansia muciniphila* was shown to be associated with the effectiveness of cancer immunotherapy treatment and metabolic disorders, our study suggests that crAssphage may play important roles in these diseases. With the flexibility of incorporating other additional features, the model has great potential to be updated along with the development of the biological understanding of virus-host interactions.

Introduction

Viruses are the most abundant and highly diverse biological entities on earth[1,2]. Viruses can not replicate themselves independently and rely on the host cells they infect to replicate. Viruses infect all domains of life including archaea, bacteria, and eukaryotes. For prokaryotic viruses (those that infect bacteria and archaea), much is already known about their diversity[3,4], functions[5–7], and their impact on microbial communities through virus-host dynamic interactions[8–11]. In particular, prokaryotic viruses can significantly impact human health[12–14] and the functioning of many ecosystems[15–17] such as marine and soil habitats. Therefore, characterizing virus-host interactions is a critical component to understanding how biological systems work.

Viruses are traditionally studied using culture-based isolation techniques. The host of a cultured virus is directly known because it was obtained by culture on a particular host cell. The isolation approach, however, is low throughput, and a relatively small number of viruses have been discovered in this way versus the extant viral diversity. In particular, it was estimated that approximately 75%-85% of viruses are uncharacterized[11,18]. Metagenomic shotgun sequencing, especially the metagenomic sequencing of virus-like particles (VLPs), has tremendously accelerated the speed of uncovering unknown viruses. In metagenomic shotgun sequencing, the genetic materials of different types of microbial organisms including bacteria, archaea, protists, viruses, etc. in the same environment, regardless of cultivability, are sequenced altogether, resulting in a set of short reads randomly sampled from the microbial genomes.

A large number of new viral genomes or sequence fragments have been discovered from metagenomic data[5,19]. An example is crAssphage, a highly abundant ubiqitous human gut virus, that was revealed from the cross-assembly of 12 faecal viral metagenomic samples[20]. CrAssphage probably plays an important role in human gut, but it is still not possible to culture crAssphage in the lab so that its biological functions have not been characterized and the hosts of crAssphage have not been experimentally confirmed. Single-cell amplified genomes (SAGs) have recently become useful for viral studies as well. Since SAGs often contain both host and viral sequences[21,22], this new approach can potentially be effective to study virus-host interactions, although the current technique is still not mature enough to generate high-throughput data.

Identifying the hosts of viruses is important for the understanding of the impact of viruses on the hosts' dynamics and the hosts' impact on the microbial community. Computational methods have been developed to infer the hosts of new viruses. Many bacteria and archaea possess CRISPR virus defense systems whereby the host incorporates some virus DNA fragments into its own genome forming inter-spaced short palindromic repeats (CRISPRs) spacers. Therefore, sharing of CRISPR regions is the most direct evidence supporting virus-host interactions[16,20]. Genome alignment matches between virus and host genomes due to integrated prophages or horizontal gene transfers are another piece of strong evidence used in predicting the host of a virus[5,16]. However, the above methods are limited by their low accessibility: CRISPRs are absent from approximately 60% of sequenced bacterial genomes[23]; many viruses infect hosts under lytic mode without integration to the host genome; and only few viruses share host genes. For these viruses, CRISPRs and alignment-based approaches are not feasible for predicting the virus-host interactions.

Several investigators have utilized the fact that viruses are similar to their hosts compared with other unrelated host species in terms of their genomic signatures (or k-mer usage)[11,24–26]. They predicted the host of a virus as the one closest to the viral genome based on some dissimilarity measures using k-mers. The methods in general have decent prediction accuracy, though the mechanism behind this phenomenon is not fully understood. One plausible explanation is that the virus tends to adopt the codon used by its hosts in order to utilize the translational machinery of its host to replicate itself[27,28]. The recently developed dissimilarity measure d_2^* that subtracts expected k-mer frequency from the observed frequency achieves the highest accuracy among all current genomic signature-based measures, including the commonly used Euclidean and Manhattan distances[25]. Similarly, Galiez *et al.*[26] predicted the host of a virus to be the one under whose Markov chain model the viral sequence has the highest likelihood, and the method has good accuracy for short viral fragments. The genomic signature-based measures are often referred as alignment-free sequence comparison measures. The high correlation between virus and host abundance profiles across different samples also serves as evidence for virus-host interaction[20], but it does not have as high accuracy as the above methods[24]. Edwards *et al.*[24] has recently provided a comprehensive evaluation of several different computational approaches listed above for the virus-host prediction problem.

In addition to the methods using features defined between a pair of virus and host genomes, some researchers have used virus-virus similarity networks to infer the host of a query virus[29,30]. The high similarity between viruses may indicate common or very close host relatedness. In fact, network-based prediction models, whereby, unknown entities are predicted based on the features of their neighbors in a network, have been successfully applied for many biological problems, including predicting protein functions using a protein-protein interaction network[31,32], inference of disease genes based on a gene-gene network[33–35], and predicting the target of new drug using drug-drug, drug-target and target-target similarity networks[36]. A few attempts were made to exploit the possibility of predicting the host of a virus based on virus-virus network information. Different principles, such as gene homology[37,38], protein family[19] and genome similarity[29,39,40], were used to defined the virus-virus relationship in networks. Villarroel *et al.* proposed HostPhinder[29], a method to predict the host of a virus by searching for the virus that shares the most k-mers in a database with known hosts. Zhang *et al.*[30] identified the important k-mer features of viruses infecting the same host genera, and built a classifier to predict if a new virus belongs to the same group of viruses. One drawback of the network-based approach is that the performance can diminish if the query virus is highly divergent from the known viruses in the current network.

Though various methods have been proposed for predicting virus-host interactions, the highest accuracy is only 48% at the genus level using a single type of information such as virus-host alignment matches. With the exponentially increasing number of viruses uncovered, there is a huge demand for a tool that is able to accurately and rapidly predict the hosts of viruses, incorporating all types of virus-host and virus-virus features. In this paper, we develop a network-based integrated framework for predicting virus-host interactions based on multiple types of information: virus-virus similarity, virus-host alignment-free similarity, and virus-host shared CRISPR spacers and alignment-based matches. To the best of our knowledge, it is the first time that multiple types of features are effectively integrated into a stochastic network to complement each other and enhance the prediction accuracy of virus-host interaction. This powerful integrated framework markedly improves the accuracies in predicting virus-host interactions for complete viral genomes from 48% to 61% at the genus level, and yields 88% accuracy at the phylum level, the highest among all the existing methods. The prediction framework also has decent accuracy for shorter viral contigs even as short as 10 kb. We apply the framework to predict the hosts of 19 marine viruses discovered from SAGs in a previous study[21]. More than 94% of Proteobacteria viruses are predicted correctly, and the overall accuracy is 60% at the phylum level. The host of crAssphage is also inferred using the integrated framework. The predicted hosts include the putative host genus *Bacteroides*, and other potential hosts such as *Akkermansia muciniphila* are suggested based on our prediction results. Since *Akkermansia muciniphila* was shown to be associated with the effectiveness of cancer immunotherapy treatment[41] and metabolic disorders[42], our study suggests that crAssphage may play important roles in these diseases. Thus, the newly developed network-based integrated framework will greatly facilitate the characterization of the millions of unknown viruses from the big metagenomic sequencing projects.

Results

A novel network-based integrated framework for predicting virus-host interaction. We collected from NCBI a set of known virus-host interaction pairs, S_+, and a set of random virus-host pairs that most likely do not interact, S_-, as the data for this study. Our objective is to develop a machine learning approach to predict the probability for a query virus-host pair (v, b) to interact, denoted as $P(I(v, b) = 1)$, where $I(v, b)$ denotes the interaction status of a virus v and a host b with value 1 indicating interaction and 0 otherwise. In order to achieve the best performance, we comprehensively considered various factors that contribute to the interaction of a virus-host pair (v, b). First, the similarity among viruses indicates the relatedness of their hosts. If a virus is genetically close to viruses infecting a host, this virus is highly likely to infect the same host[29,30]. On the other hand, if a virus infects a host, the virus should be less similar to the viruses not infecting the host. Secondly, the similarity among hosts indicates the possibility of infection by the same virus[43,44]. If a new host belongs to the same taxon of the known host of the virus, then the host is likely to be infected by the virus. Third, the similarity between virus-host pairs in terms of genomic signatures reflects the likelihood of interaction. If the virus genome is similar to the host genome in terms of the alignment-free k-mer pattern usage, the pair is predicted to have a high chance of being interacting. Finally, the existence of virus-host shared CRISPR spacers and

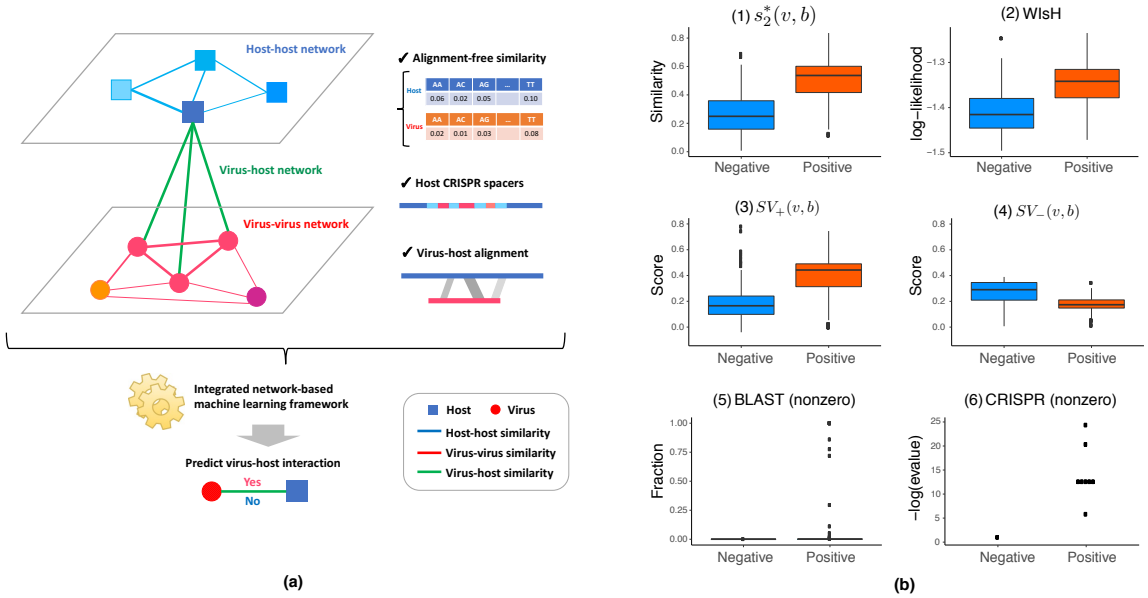

Figure 1. **(a)** Outline of the framework. A novel two-layer network is constructed for representing virus-virus, host-host, and virus-host similarities. Viruses (red circles) are connected (red edges) based on sequence similarity. Similarly, hosts (blue squares) are connected (blue edges) based on sequence similarity. The thick edges indicate high degrees of similarity. The interaction between a pair of virus and host (green edges) can be predicted using multiple types of features: 1) the similarity between the virus and the other viruses infecting the host; 2) the similarity between the host and the other hosts infected by the virus; 3) the alignment-free sequence similarity between the virus and the host based on k-mer frequencies; 4) the existence of shared CRISPR spacer between the virus and the host; 5) the alignment-based matches between the virus and the host. Finally, a network-based machine learning model is used to integrate all different types of features and to predict the likelihood of the interaction of a virus-host pair. **(b)** Distributions of the different feature values in the 352 interacting virus-host pairs (positive set) and the same number of non-interacting virus-host pairs (negative set) in the training data: 1) Boxplots of similarity defined by $s_2^*(v,b)$. The horizontal bar displays the median, boxes display the first and third quartiles, whiskers depict minimum and maximum values, points beyond the end of the whiskers are outliers. 2) Boxplots of the log-likelihood scores given by WIsH; 3) Boxplots of $SV_+(v,b)$ scores. 4) Boxplots of the $SV_-(v,b)$ scores; 5) Boxplots of BLAST scores. There are 108 nonzero values in the positive set, and 79% of them are less than 0.01. 6) Dotplots of the CRISPR scores. Only 8 virus-host pairs in the postive set and 1 pair in the negative set share CRISPR spacers.

the alignment-based matches is strong evidence of the interaction. The virus-virus similarity, host-host similarity, and virus-host similarity form a novel two-layer network connecting viruses and hosts. Therefore, based on the features derived from the two-layer network, we developed an integrated network-based Markov random field (MRF) approach that systemetically and comprehensively integrates various types of features to predict interacting virus-host pairs. The outline of the framework is demonstrated in Fig. 1a.

The feature scores are significantly different between positive and negative virus-host pairs. We carefully designed multiple types of features that contribute to the prediction of virus-host interactions. To assess the discriminatory power of each feature, we compared the distributions of the feature values between the virus-host interacting pairs and the non-interacting pairs. A set of 352 known interacting pairs were used as the positive set, and a set of the same number of randomly selected virus-host pairs was randomly selected and used as the negative set. See Materials and Methods for details of the data collection and the simulation of negative pairs. We used one sided t-test to test if the feature values in the positive set is significantly higher/lower than the ones in the negative set.

First, the alignment-free method $s_2^*(v,b)$ was used to measure the similarity between a virus-host pair. The similarity measure s_2^* is essentially the d_2^* in our previous work[25] but is expressed as a similarity score rather than dissimilarity. The measure s_2^* has the advantage over other classic measures because of its precise correction of background noise, and has shown superior prediction accuracy in the problem of predicting virus-host interaction[25]. See Materials and Methods section for the definition of $s_2^*(v,b)$. The s_2^* score had significantly higher values (p-value $< 2.2e-16$ for one sided t-test) for positive virus-host pairs than the negative pairs (Fig. 1b(1)). The mean s_2^* similarity score between positive pairs was 0.50 while the mean s_2^* similarity between negative pairs was 0.26.

The WIsH score, proposed by Galiez *et al.*[26], is another alignment-free simialrity measure that computes the log-likelihood of a viral genome coming from a host Markov chain model. We computed the WIsH scores for both positive and negative pairs, and it showed that the WIsH score has significantly different means (p-value $1e-10$) between the positive pairs and the negative pairs (Fig. 1b(2)). In fact, we observed that WIsH and s_2^* were highly correlated (Pearson correlation coefficient $\rho = 0.83$). We predicted a virus-host pair as interacting if one of the similarity measures, s_2^* or WIsH, is above a threshold and, by changing the threshold, the corresponding receiver operating characteristic curve (ROC) was plotted. The area under the receiver operating characteristic curve (AUROC), which measures the discriminative ability between positive and negative pairs, was 0.88 for s_2^* and 0.83 for WIsh (see the ROC curves in Supplementary Fig.B.1). Though the distinguishing power using WIsH was less than that of s_2^* using complete genomes, WIsH was previously shown to be more effective than s_2^* when predicting hosts of partical viral genomes[26]. Therefore, we decided to use s_2^* to measure virus-host alignment-free similarity when the length of viral sequence is close to the size of complete genome, and to use WIsH to measure the virus-host similarity for contigs shorter than 10 kb.

Second, for a given virus-host pair (v,b), we defined the similarity between virus v and other viruses infecting the same host b, denoted as $SV_+(v,b)$, and likewise, the similarity between virus v and other viruses not infecting the host b, denoted as $SV_-(v,b)$. See Materials and Methods for the details of the definitions. We hypothesized that, for a true interacting virus-host pair (v,b), other viruses that infect the same host b should exhibit high similarity to the virus v, resulting a high $SV_+(v,b)$, and at the same time, other viruses not infecting the host b should have low similarity to the virus v, resulting a low $SV_-(v,b)$. For a non-interacting virus-host pair, the above trend of $SV_+(v,b)$ and $SV_-(v,b)$ should be opposite. Consistent with our hypothesis, $SV_+(v,b)$ scores were significantly higher for positive virus-host pairs than negative pairs, and vice versa for $SV_-(v,b)$ scores (both p-values <1e-5, Fig. 1b(3-4)).

Third, we defined the CRISPR score and alignment-based matching score. The CRIPSR score is defined as the highest aligned scores between the host spacers and the viral genome, and the alignment-based matching score is defined as the fraction of virus genome mapped to the host genome using `blastn`. Thus for simplicity, we refer the alignment-based matching score to as the BLAST score. See the Material and Methods section for the details. Both CRISPR and BLAST scores were significantly higher for the true interacting virus-host pairs than the non-interacting pairs with p-values 0.004 and 8e-5 for one sided t-test, respectively. Fig. 1b(5,6) shows the limited availability of CRISPR and BLAST features though. For the alignment-based BLAST score, only 108 (31%) out of 352 positive virus-host pairs had nonzero values, and 79% of the scores were less than 0.01. The CRISPR scores were even more sparse, with only eight nonzero values (2.3%) out of 352 in the positive set, which can explain why its p-value was not as significant as the other measures.

The integrated approach markedly increases the prediction accuracy of the hosts of viruses. We integrated the multiple types of features proposed previously to predict virus-host interactions using a general framework of MRF, where the nodes were virus-host pairs (VHP) and edges were the similarity between the VHPs. See Materials and Methods for the details. We investigated the prediction accuracies of the newly developed integrated models in equations (4) and (5), and we compared the accuracies with those using the individual features. The model in Eq. (4) incorporates the network features like the virus-virus similarity SV_+ and SV_- and the virus-host similarity s_2^*, with the CRISPR score, and the model in Eq. (5) combines all types of features involved in the network, as well as the CRISPR and BLAST scores. For each of the integrated models, we learned the parameters using the 352 positive and negative virus-host pairs, and then tested the trained model on the remaining 1,075 viruses for which their true hosts are known against 31,986 candidate hosts. The estimated coefficients and the corresponding p-values of the features are shown in Table 1. All the coefficients had the expected signs that were consistent with the observations in Fig. 1b, and the p-values for the coefficients were all < 0.05. Note

that though the p-value for S_{CRISPR} was 0.02, it was not as small as the p-values for the other features due to its limited availability.

We assessed the prediction accuracies of the trained models using an independent set of 1,075 viruses at different taxonomic levels, including genus, family, order, class, and phylum. The details are described in Materials and Methods. Compared to the single feature prediction methods using s_2^* and CRISPR score alone, the prediction accuracy of the integrated approach combining network features and CRISPR score was markedly increased (Fig. 2a). At the genus level, prediction accuracy was 34.33% and 33.58% when using s_2^* and CRISPR, respectively, while using network similarity features together with CRISPR score (Eq.(4)) increased prediction accuracy to 48.3%, or a 1.4-fold increase.

Alignment-based BLAST scores alone had relatively high prediction accuracy itself. Incorporating BLAST scores into the network model by Eq. (5) yielded an even higher increase in host prediction accuracy of 60.4%, a 12% improvement. For the higher levels of taxonomy like family, order, class and phylum, the network-based integrated framework also achieved large improvement over the single-based predictions, yielding 74%, 75%, 84%, and 88% prediction accuracy, respectively. Thus, the model with the highest prediction accuracy including all types of features by Eq. (5) was used in the subsequent host prediction applications. These results also demonstrate the great potential of our integrated framework: when any other meaningful features contributing the virus-host interaction become available in the future, the model can be further extended to provide more accuate predicitons.

Model		s_2^*	S_{WIsH}	SV_+	SV_-	S_{CRISPR}	S_{blastn}
Complete genomes using Eq. (5)	Coefficients	13.54	-	3.40	-29.30	0.19	2791.1
	p-value	<2e-16	-	6e-05	<2e-16	0.02	0.001
Short contigs using Eq. (6)	Coefficients	-	21.70	5.40	-17.60	0.33	4107.7
	p-value	-	1e-10	1e-11	<2e-16	0.001	8e-05

Table 1. The estimated coefficients and the corresponding p-values of the features for predicting hosts of complete viral genomes using the network-based integrated model in Eq. (5), and for predicting hosts of short viral contigs using the model in Eq. (6). Since different negative training sets yielded slightly different estimated coefficients of the features, we show one example here.

The integrated approach markedly improves the host prediction accuracy on short viral sequences. The viral contigs assembled from the metagenomic data are often partial viral genomes. We developed the integrated model in Eq. (6) using WIsH scores instead of s_2^* for measuring the alignment-free similarity between the virus and the host. We evaluated the accuracy of the model for predicting the hosts of viral contigs at various lengths, and investigated the effect of viral sequence length on the prediction accuracy. To evaluate the performance of host prediction for short viral contigs, we randomly sub-sampled fragments of different lengths (5 kb, 10 kb, and 20 kb) from each of the 1,075 viral genomes. For a given viral genome and a fixed contig length, we randomly chose a segment of the fixed length uniformly from the genome. If the fixed length was longer than the size of the complete genome, we took the entire genome. This procedure was repeated 30 times for each contig length. We then computed all the features of the contigs in the same manner as for the complete viral genome analyses, with the only difference of replacing the s_2^* similarity with the WIsH scores[26]. The model was trained with the same set of 352 virus-host positive and negative pairs using the same scheme as before by replacing s_2^* with the WIsH score. Similar to the model for predicting complete viral genomes, the estimated coefficients is shown in Table 1. With the trained model, we predicted the hosts for all the sub-sampled contigs.

The average prediction accuracies for each contig length are shown in Fig. 2b. Our model (solid lines) achieved a large improvement compared to the results by WIsH alone (dashed lines). For example, when the contig length was 20 kb, the prediction accuracy using our model was about 13-18% higher than that of WIsH at the genus, family and order levels. As expected, the prediction accuracy of our model (solid lines) decreased with decreasing contig length. For instance, at the genus level, the accuracy decreased from 61% for the complete genome, to 52% for 20 kb, 50% for 10 kb, and 44% for 5 kb. The standard error of the accuracy also increased as the contig length decreased. Note that the prediction accuracy of WIsH was not highly affected by the contig length. Given the results, we provide our framework with two models for host prediction: one for complete or nearly complete viral genomes using the model in Eq. (5), and one for short viral contigs using the model in Eq. (6).

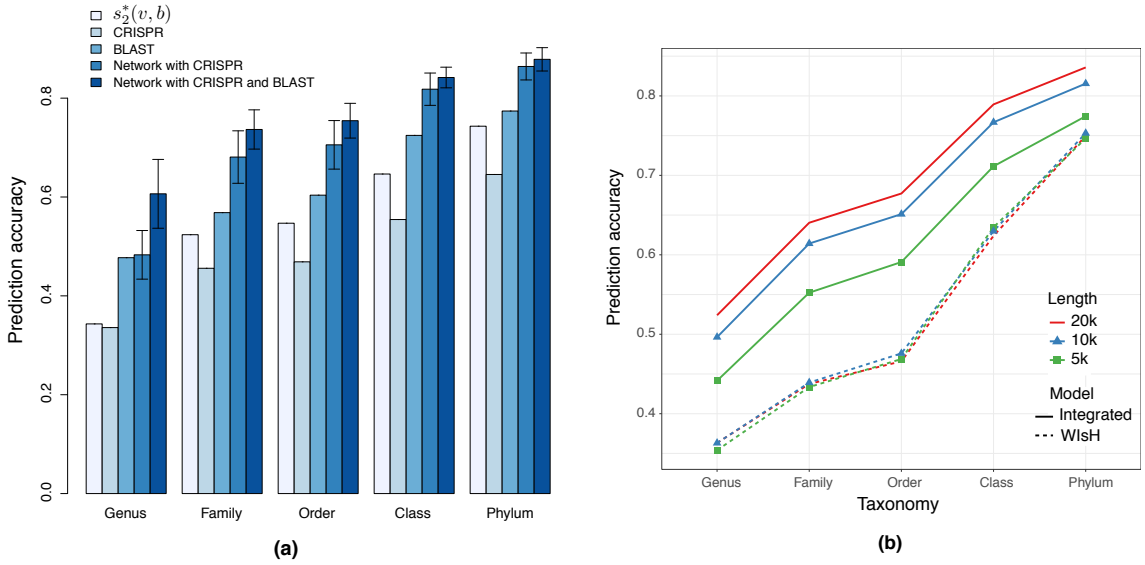

Figure 2. **(a)** Comparison of prediction accuracies of the different approaches on a test set of 1,075 viruses against 31,986 candidate hosts. The results are binned by taxonomic level. For each bin, the first three bars show the prediction accuracies using the individual features of $s_2^*(v,b)$, CRISPR score, or alignment-based similarity score, respectively. The fourth and fifth bars show the prediction accuracies using the integrated models proposed in Eq. (4) and (5), trained based on the 352 positive virus-host pairs and the same number of randomly sampled negative pairs. The model in Eq. (4) incorporates the network-based features $SV_+(v,b)$ and $SV_-(v,b)$, alignment-free virus-host similarity $s_2^*(v,b)$, and the CRISPR match scores. The model in Eq. (5) includes all features in Eq. (4) plus the alignment-based (BLAST) similarity feature. The error bars in the last two bars indicate the 95% confidence interval of the accuracies, based on 100 different training sets. **(b)** Comparison of prediction accuracies based on simulated viral fragments of size 5 kb (green line), 10 kb (blue line), and 20 kb (red line), respectively. The accuracy is compared between the integrated model in Eq. (6) (solid line) and WIsH (dash line) at different taxonomic levels. The standard errors for the prediction accuracy at genus level of 20 kb, 10 kb, and 5 kb are 0.001, 0.001 and 0.002 for WIsH and 0.002, 0.002, 0.014 for the integrated model, respectively.

Thresholding on the probability score further improves accuracy. In many situations, investigators are interested in making sure the predicted hosts are correct, i.e. the predictions have high precision. Therefore, we investigated how the accuracy changes by thresholding on the predicted probability of interaction $P(I(v,b)=1)$. In the above analysis, we predicted the host of every virus as the one with the highest score. However, sometimes even the highest score was relatively low. For example, as shown in Fig.3a, the highest prediction score among the 31,986 hosts for some viruses in the complete genome test set was as low as 0.4. Low scores may occur, for example, when the true host is not in the database of possible hosts. In order to improve the prediction accuracy, we can set a threshold for making predictions and not make a prediction if the highest score is below the threshold. For instance, when a threshold was set at 0.8, there was a markedly improvement of prediction accuracy at all taxonomic levels. Specifically at the genus level, accuracy was improved by 7%, from 61% to 68%. As a tradeoff, the proportion of viruses whose hosts can be predicted among all the viruses (the recall rate) decreased. However, this tradeoff turned out to have minimal impact on the prediction result. In fact, when the threshold was set at 0.8, the accuracy improved to 68% at the genus level, while the recall rate only went down to 86%. The model only missed 3% of the viruses for which it should have predicted correctly but chose not to predict. In practice, this small sacrifice can provide much more confidence for prediction in practice.

Prediction accuracy varies for different virus families. Viruses from three major families, siphoviridae, myoviridae, and podoviridae, are highly represented in our evaluation set. Previous host predictions with s_2^* showed notable differences in prediction accuracy among these families [25]. Therefore, we examined

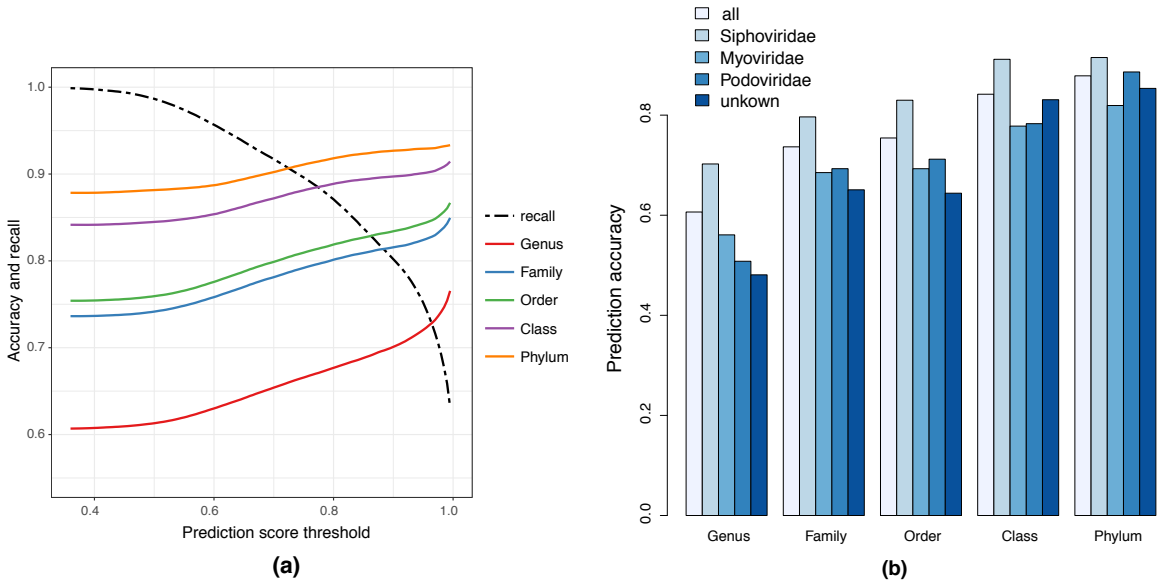

Figure 3. Improvement in host prediction by thresholding on the prediction score and differences in prediction accuracy across families. **(a)** By applying a threshold, predictions were made only when the prediction score is above the threshold. The proportion of viruses that can be predicted (recall rate) decreases as the prediction accuracy at all levels increases. **(b)** Prediction accuracies for different virus families, siphoviruses, myoviruses, and podoviruses (no thresholding). Viruses for which their virus families were not listed in the Genbank files are groups together as 'unknown'.

prediction accuracies using our udpated models (Fig. 3b). We found that the siphoviridae in our dataset had generally higher prediction accuracy than other types of viruses, achieving 70% accuracy compared with the average accuracy of 61% for all types of viruses, consistent with previous results using s_2^* scores alone at the genus level. This observation may be partially explained by the fact that siphoviruses and podoviridae typically have relatively narrow host ranges and myoviruses often have the broadest host ranges[45–47]. The prediction accuracies for the different virus families with various thresholds on the prediction score are discussed in Supplementary Fig.B.2.

Host prediction for viruses in single-cell amplified genomes (SAGs) validates the high accuracy of the integrated approach. Single-cell genomics has emerged as an important complement to metagenomic approaches in the study of microbes[48]. The ability of SAGs to recover both viral and host DNA sequences from actively infected cells helps shed light on *in situ* virus-host interactions. Labonte *et al.*[21] computationally identified complete or near-complete viral genomes from 20 out of 58 previously published microbial SAGs collected from surface ocean[49]. Detailed information on the SAGs can be found in Labonte *et al.*[21] and its supplementary materials. In this study, we employed our integrative method to predict the hosts of the 20 viral genomes and compared the host prediction results against the host taxonomy originally inferred from the SAG sequences. Similar to the scheme as mentioned before, we trained the model based on 352 known viruses, and calculated the alignment-free similarity scores, BLAST and CRISPR scores between the 20 viral genomes and 31,986 hosts. Eighteen out of the 20 SAG's hosts were present in the 31,986 host database. For fair evaluation, we removed all the contigs that were identical to the viral genomes in a preliminary check using `blastn`.

Since the 20 viruses represented nearly complete genomes, we used the model in Eq. (5) for host prediction. The prediction results for 18 of the viruses are shown in Supplementary Table B.1. We did not include the prediction result for AAA164-O14 due to its low prediction scores with the highest prediction score of 0.77. We also excluded AAA164-N20 from prediction analysis because the virus from this SAG was highly similar to a eukaryotic virus and probably was the result of non-specific attachment of a viral particle to the host cell[21]. We compared the predicted taxa with the reported host taxa and evaluated the prediction accuracy using the same criterion as in the Materials and Methods section.

Our model achieved 60.2% overall accuracy at the phylum level for the 18 SAG-derived viruses. The predictions were 94.4% correct for Proteobacteria and 56.3% for Verrucomicrobia, two phyla that make up the majority of the 18 viral genomes (14/18). Our model missed one of the SAGs in the phylum of Marinimicrobia (AAA160-C11), and one in each of the phyla Bacteroidetes (AAA160-P02) and Thaumarchaeota (AAA160-J20). It is worth noticing that the three phyla: Marinimicrobia, Thaumarchaeota, and Verrucomicrobia are all under-represented groups in the 31,986 candidate hosts we used in this study. In fact, for the three phyla, there are merely 21, 53, and 37 hosts among the 31,986 possible hosts, with a chance of random selection of only 0.066%, 0.18%, and 0.097%, respectively. None of the three phyla is present in our training data, which also makes the prediction challenging. In addition, since the host taxonomic groups of the viruses were inferred computationally based on the host sequences, it is possible that the host can be falsely inferred for some viruses when the coverage of their hosts was not enough. For example, it was shown that only <1% and 5% of the host genomes for AAA160-J20 and AAA164-P11 were recovered. Though the estimated risks of contamination with free viral particles or attachment of viruses to non-specific host cells was low, there was a chance that some of the viruses recovered from SAGs did not represent the interactions. All the factors can potentially affect the prediction result in this study.

Our integrated approach supports previous finding for the host of crAssphage and also suggests another potential host *Akkermansia muciniphila*. In a recent study, Dutilh *et al.*[20] discovered a highly abundant bacteriophage, crAssphage, in human faecal samples using viral metagenomic sequencing data. Though crAssphage is ubiquitous in human gut samples and comprises up to 90% of the sequencing reads in faecal viral metagenome[20], little is known about the biological significance and the hosts of crAssphage, due to the difficulty of culturing crAssphage and the high divergence between crAssphage and known viruses. Different methods have been used to predict the hosts of crAssphage. Dutilh *et al.*[20] predicted its host as the phylum Bacteroidetes using the co-occurrence profiling between crAssphage and 404 potential human gut bacteria hosts across 151 human gut metagenomes from the Human Microbiome Project. Ahlgren *et al.*[25] compared the alignment-free similarity between crAssphage and the potential hosts, and predicted the hosts as those with highest similarity to crAssphage. The genera, *Bacteroides*, *Coprobacillus* and *Fusobacterium*, were detected as having significantly high similarity to crAssphage.

Our integrated approach supports the previous finding that members of the phylum Bacteroidetes are potential hosts of crAssphage. Among the top 100 ranked hosts predicted for crAssphage, three hosts in the phylum of Bacteroidetes were ranked 11th, the 55th and the 66th: *Prevotella oryzae* DSM 17970, *Prevotella paludivivens* DSM 17968, and *Prevotella bryantii* B14 respectively with overall prediction scores of 0.966, 0.951, and 0.950, respectively. Each had the same CRISPR score of 2.797. In addition, the strain *Porphyromonas bennonis* DSM 23058, in the phylum of Bacteroidetes, also has a high CRISPR score of 8.282 and ranked at the 53th with an overall prediction score of 0.951. A different *Prevotella* strain, *Prevotella intermedia* 17, was previously suggested as a host of crAssphage by Dutilh *et al.*[20] when searching CRISPR spacers among a set of 3,177 complete bacterial genomes against crAssphage.

In order to yield better prediction, we restricted the host set to the 360 Human Microbiome Project (HMP) host genomes previously used in Ahlgren *et al.*[25]. Since many human gut bacteria are uncharacterized at the strain level and the 360 host genomes are only general representation, we investigated host predictions at higher taxonomic levels, such as the genus and the phylum levels, to identify which genus or phylum is more likely to be the host of crAssphage.

We used Wilcoxon rank-sum test to determine if hosts of a particular taxon have significantly higher prediction scores than all other taxa. To achieve a reasonable statistical test power, we only tested the taxa that have at least 20 members in the 360 HMP host set. Thus, we performed the test on 4 phyla (Bacteroidetes, Firmicutes, Proteobacteria and Actinobacteria), and 3 genera (*Bacteroides, Lactobacillus and Escherichia*). The corresponding p-values for the taxa are shown in Fig. 4(a,b). After Bonferroni correction, only the genus *Bacteroides* (p-value 3.34e-08) and the phylum Bacteroidetes (p-value 5.28e-08) were significant.

We also investigated the effect size for different taxa to evaluate the difference between the prediction scores of hosts in a hypothetical taxon and that in the controlled unrelated taxon. The phlyum Bacteroidetes and the genus *Bacteroides* had the largest effect sizes, 0.64 and 0.81, respectively, compared with other taxa (Fig. 4). The effect size > 0.5 was considered as a medium to large effect, suggesting that the crAssphage and the hosts in Bacteroidetes and *Bacteroides* have strong interactions. In addition,

among the 360 HMP hosts, we observed that two host genera in the phylum of Bacterioidetes, *Prevotella* and *Parabacteroides*, had CRISPR spacers that are similar to crAssphage with the scores of 2.80 and 0.60, respectively. All the above facts support the claim in the previous studies[20,50] that the phylum of Bacteroidetes could be the candidate host of crAssphage.

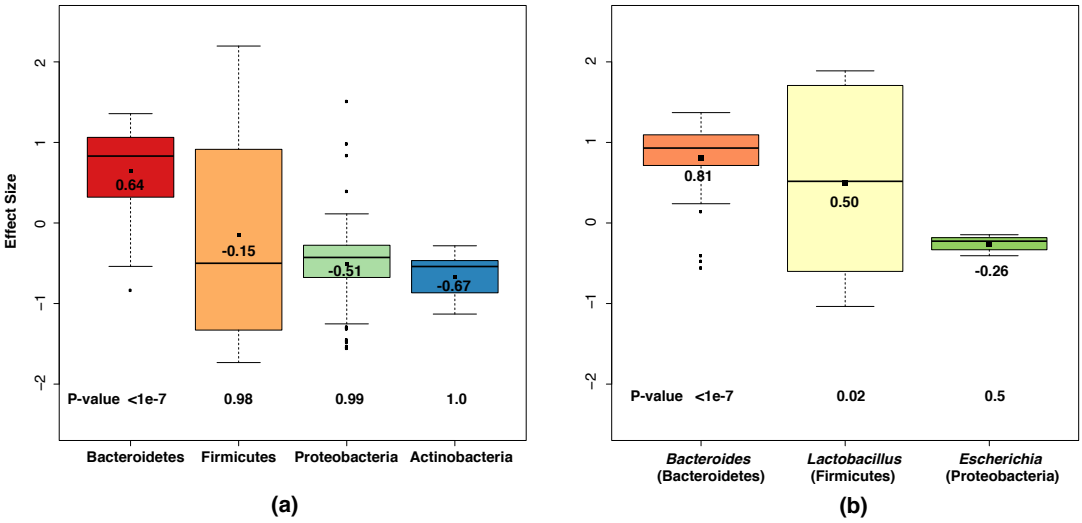

Figure 4. The p-values of the rank-sum tests and the effect sizes for the **(a)** four phyla (Bacteroidetes, Firmicutes, Proteobacteria and Actinobacteria) and **(b)** three genera (*Bacteroides, Lactobacillus and Escherichia*) studied in prediction of the host of crAssphage, for evaluating whether hosts from a particular taxonomic class have significantly higher prediction scores than the other hosts. The horizontal bars display the median, boxes display the first and third quartiles, whiskers depict minimum and maximum values, and the squares and the numbers below them display the effect sizes.

Our prediction model also indicated *Akkermansia muciniphila* from the phylum Verrucomicrobia as a potential host for crAssphage. Using our new model for complete genomes in Eq. (5), *Akkermansia muciniphila* had the highest ranked score of 1 out all of possible 31,986 hosts. Carefully examination of the individual measures contributing to this result explain why it was so highly ranked. Among the 31,986 hosts, only *Akkermansia muciniphila* yielded a positive BLAST alignment score. Its score of 5.1% indicated that this strain has high similarity (97.2% identity on average) with 5.1% of the crAssphage genome, or about 5,000 bp. These regions of similarity ranged from 40 to 400 bp and were scattered across the *Akkermansia muciniphila* genome

The alignment-free similarity score between crAssphage and *Akkermansia muciniphila* was $s_2^* = 0.339$, which was higher than the average value of s_2^* between crAssphage and all other host bacterial genomes. In comparison, the *Bacteroides* strains had s_2^* scores in the range of 0.272 to 0.495. Although two CRISPRs were found along the *Akkermansia muciniphila* genome[51], the associated spacers could not be mapped to crAssphage. Since no viruses are known to infect *Akkermansia muciniphila*, its $SV_+(v, b)$ score was 0.

To support this prediction further, we investigated the co-occurrence pattern between crAssphage and *Akkermansia muciniphila* in human gut metagenomic samples. Dutilh *et al.*[20] previously calculated the Spearman correlation between the abundance profiles of crAssphage and 404 bacteria from the HMP1 metagenome set, but their host collection did not contain *Akkermansia muciniphila*. We repeated this analysis by now including *Akkermansia muciniphila*, and excluded metagenomes that did not have any reads that mapped to crAssphage. By Spearman's rank correlation, *Akkermansia muciniphila* was the most highly correlated with a coefficient of 0.29 (p-value = 0.003). In comparison, all other strains including many *Bacteroides* strains has coefficients of < 0.25. Thus, the co-occurrence patterns between crAssphage and *Akkermansia muciniphila* also strongly support it as a potential host.

Discussion

The dynamics between virus and prokaryotic host interactions play important roles in human health and ecosystems. Millions of unknown viruses have been identified using the high-throughput metagenomic sequencing technology, but little is known about their biological functions and the prokaryotic hosts with which they interact. We developed a network-based integrated framework for predicting the hosts of prokaryotic viruses. The new method provides a huge improvement on the prediction accuracy compared with previous methods by integrating multiple measures for informing host prediction. Based on the evaluation of the methods using a large benchmark dataset containing 1,075 viruses and 31,986 hosts, the method achieves 61% and 88% prediction accuracy at the genus and the phylum levels, respectively, yielding 27% and 14% improvements at the genus and the phylum levels comparing to the highest accuracy achieved by previous methods.

The novel two-layer network of virus-virus, host-host, and virus-host genomic similarity lays the foundation for this method. The employment of the two-layer networks is inspired by the underlying biological phenomena. First, it is observed that genetically similar viruses tend to infect closely related hosts [43,44]. So the host of a new virus can be inferred based on the similarity to related viruses with known hosts, if they exist. Similarly, since the similar hosts can be infected by the same type of viruses, new hosts of a virus can be inferred based on the similarity between the new hosts and the known hosts. Second, because viruses depend on the cellular machinery of their host to replicate, viruses often share highly similar patterns in codon usage or short nucleotide words with their hosts. Then the host of a new virus can be indicated using nucleotide word similarity between the virus and the candidate hosts [11,24,25]. Thus, the two-layer network model is a natural formulation of biological system stated above. Despite the fact that the viruses in our current database only have one reported host for each virus such that host-host network connections cannot be incorporated into the prediction model, the novel two-layer network can be fully utilized in the future as the multiple hosts of viruses are revealed.

Multiple types of features, including virus-host shared CRISPR spacers and virus-host BLAST matches, combined with the network-based features, were incorporated to obtain an integrated framework for host prediction. The CRISPR and BLAST features are based on the biological phenomenon that some virus and their hosts share a portion of their genomes due to CRISPR defense system, horizontal gene transfer, or prophage integration. Though these features have been investigated individually in previous studies [24–26,52], it is the first time that multiple types of features have been integrated into a unified framework for virus-host prediction. Not surprisingly, the results show that the integrated method combining all features achieves the highest prediction accuracy among all models tested that use different combinations of the various features.

We also markedly improved the host prediction accuracies on shorter viral fragments at all taxonomical levels as compared to WIsH [26], a recently released software developed for this particular purpose. Metagenomically assembled viral genomes are most often partial and incomplete. The CRISPR and BLAST features depend highly on the overlap of particular regions between viral and host genomes, and thus can have limited utility in host prediction for metagenomic viral contigs. In addition, the alignment-free similarity score s_2^* requires a certain size of sequence to obtain high statistical power. Therefore, it makes sense that the prediction accuracy is reduced for short viral sequences. Our method is able to obtain 49%, 46% and 41% prediction accuracies at the genus level for 20 kb, 10 kb and 5 kb sequence lengths, respectively. The prediction accuracies for 20 kb, 10 kb, and 5 kb contigs are all above 75% at the phylum level. In practice, we recommend using at least 10 kb for host prediction at the genus level and at least 5 kb for prediction at the phylum level.

Setting a minimum threshold for making predictions displayed a notable improvement on accuracy and we show that this improvement comes at a minimal cost in recall. We also investigated the host prediction accuracy for different groups of viruses. Specifically, our observations indicate that viruses in the siphoviridae group have higher prediction accuracy than others, since siphoviruses tend to have a narrower range of target hosts [46,47]. Likewise, restricting the possible hosts from all available prokaryotic genomes to a focused set of relevant microbes can help improve prediction accuracy, as in the case of crAssphage for host prediction using HMP genomes only.

We used our new model to predict the hosts of viral genomes identified from SAGs to evaluate the method in real metagenomic data. The true hosts of the SAGs were previously inferred at the phylum level. Five of the six Proteobacteria viruses and five of the eight Verrucomicrobia viruses were predicted correctly at the phylum level. The low prediction accuracy for Verrucomicriobia viruses is probably

impacted by the under-representation of this phylum in our host database, and the fact that no viruses in our virus database infect this host. The latter fact highlights the utility and power of our model's multi-layered approach.

We also employed our model to predict host of crAssphage, an abundant and ubiquitous human gut virus. Our results support that a member of the phylum Bacteroidetes is the probable host of crAssphage, as was previously suggested[20,50]. Our model interestingly also suggests *Akkermansia muciniphila*, a member of the phylum Verrucomicrobia, as a new possible host of crAssphage since it was the most highly ranked strain in our model results. Its high BLAST score in particular explains its high prediction ranking. It is particularly compelling that only *Akkermansia muciniphila* yielded a positive BLAST score among ∼32,000 possible hosts. Recombination between viruses and hosts can occur during infection, so these results are suggestive of interaction between *Akkermansia muciniphila* and crAssphage. It is somewhat surprising however that no significant nucleotide similarity was detected between crAssphage and Bacteroidetes strains. Perhaps the Bacteroidetes strains in our analysis have not experienced recently recombination with crAssphage that could be detected by nucleotide searches. It was observed previously, though, that some crAssphage genes show similarity to Bacteroidetes genes at the protein level[20].

We similarly found a strong co-abundance correlation between *Akkermansia muciniphila* and crAssphage in human gut metagenomes and in fact this correlation was stronger than to all other hosts tested including *Bacteroides* strains. Co-abundance approaches were used to previously suggest *Bacteroides* strains as a probable host of crAssphage but did not include *Akkermansia muciniphila*[20]. Results from co-abundance approaches for inferring virus-host interactions must be interpreted cautiously since virus-host interactions may not always yield positive or negative associations dependent on the complexity of virus lifestyles (e.g. lytic vs. lysogenic) and when during an infection cycle samples were taken[53]. Likewise non-specific hosts and viruses can exhibit spurious correlations due to the computational bias in terms of the compositional data where the abundance vector is constrained to a constant sum, or the biological bias where for example those hosts share a similar niche to the virus true hosts. Nonetheless, the combined results of significant co-abundance along with our model results provide compelling evidence to further investigate *Akkermansia muciniphila* as a potential host of crAssphage.

Akkermansia muciniphila has been found to be negatively associated with inflammation, obesity and type 2 diabetes[42,54,55], and a very recent study revealed that *Akkermansia muciniphila* helps to improve the efficacy of the immunotherapy[41]. If *Akkermansia muciniphila* proves to be infected by the ubiquitous crAssphage, it could have implications for human health and our understanding of the complex interactions among intestinal tissues, prokaryotic cells, and their viruses in the human gut. Note that, though we collected as many host genomes as possible from NCBI, the current host database still cannot represent the extant host diversity. This means that the predicted hosts cannot be confirmed purely based on computational methods. Further experimental work is needed to validate as true hosts of crAssphage.

A major advantage of our network-based integrated framework is that the framework can be easily extended to incorporate more meaningful features that can better inform virus-host interactions in the future. Virus-host co-abundance profiles have been shown to provide some evidence of virus-host interactions[56], but Edwards *et al.*[24] suggested that its performance on host prediction was relatively poor compared to other measures such as CRISPR and sequence homology. Our preliminary analysis of incorporating such co-abundance data as a feature, however, showed reduced prediction accuracy (see Supplementary Fig. B.3), so we did not report on this in detail here. If however future refinement of co-abundance data can demonstrate improved utility in host prediction, it could then be included in our tool's framework.

Sequence-based and alignment-based measures such as CRISPR and BLAST scores generally suffer from the limited availability, but can provide solid evidence for virus-host interactions when a strong signal is present. On the other hand, alignment-free s_2^* similarity can be computed for any virus-host pairs, but may not always perform as strongly as CRISPR and BLAST evidence. We compared the prediction accuracies for s_2^* score and BLAST score when the hosts belonging to the true host genus of the viruses are removed from the candidates. The result showed that when the specific hosts were removed, the prediction accuracy for BLAST at the class level decreased largely to 0.20, while the accuracy for s_2^* was 0.32 (Supplementary Fig. B.4). It is shown that alignment-based method depends heavily on the existence of the true host in the database, and it can perform much worse than the alignment-free based method for predicting hosts of new viruses when true host genus are not in the host candidate set.

Therefore, the integrated framework combining both alignment and alignment-free based features helps to complement the two methods and improve the overall prediction accuracy.

We suspect that the lower host prediction accuracy found when using our method on SAG viruses is influenced largely by the absence of related hosts in our training set. This in fact is why the SAG method has been employed, to expand the catalog of genomes for difficult-to-culture microbes. Tentatively, the viral genomes in the SAGs application can be added to our training data, to promote the success rate of detecting hosts in phyla Verrucomicrobia, Marinimicrobia and Thaumarchaeota in future predictions. This problem of low representation of relevant genomes in the host database will be ameliorated over time as more genomes are sequenced. It is expected that with an increased dataset of hosts and virus-host interactions for training our models, the prediction accuracy of our method will increase further.

In summary, our novel network-based integrated approach provides a flexible framework with the potential to be updated along with the development of new computational theories and biological understanding in virus-host interactions. Despite the stated limitations, our results demonstrate the strong prediction ability of the model, and this approach will be valuable for identifying the putative hosts of newly discovered viral genomes.

Materials and Methods

Data sets. We used the same data as in[25] that includes 1,427 RefSeq viral genomes and 31,986 prokaryotic (bacterial and archaeal) genomes downloaded from NCBI on 5/8/2015. The viral data set includes all viruses at that time for which the taxonomy of its host was reported at least at the genus level. The hosts of the viruses where the viruses were originally isolated were collected based on the key words 'isolate_host=' or 'host=' within each Genbank file. Furthermore, for a subset of 352 viral genomes, their hosts were reported at either strain, subspecies, or serovar, and only a single host genome was reported in the NCBI genome database for that particular strain, subspecies or serovar. The accession numbers of the viruses and their hosts can be found in the supplementary material of Ahlgren *et al.*[25].

We used the developed computational methods to predict the host of crAssphage[20], a newly discovered virus from the human gut samples that is abundant and ubiquitous in human gut. The NCBI accession number for crAssphage is NC_024711.1. We also applied our methods to predict the host of viral sequences in 20 SAGs sampled from the Gulf of Marine[21] to see if our method can correctly identify the host genomes, assuming those viruses specifically infect those associated host cells without data contamination. The sequences of the 20 viruses were provided by Labonte *et al.*[21] in the supplemental material.

A Markov random field (MRF) approach for virus-host interactions. We formulate the virus-host interactions using a Markov random field (MRF) model[31,57]. Given a set of viruses $\{v_1, v_2, ..., v_n\}$ and a set of hosts $\{b_1, b_2, ..., b_m\}$, we define the set of virus-host pairs (VHP) and their interaction statuses,

$$\mathcal{K} = \{\kappa_{ij} = I(v_i, b_j), i = 1, 2, .., n, j = 1, 2, .., m\},$$

where $I(v, b) = 1$ if v infects b and $I(v, b) = 0$ otherwise. We construct a VHP network where nodes are VHPs and edge weights are the pairwise similarity between two VHPs.

The likelihood of an assignment $\mathcal{K}$ of the VHP network is proportional to the likelihood of the assignments of the VHP nodes and the likelihood of the edges given the assignment. Let the probability of having interaction between a pair of virus and host be π. The probability of an assignment of $\mathcal{K}$ is

$$\prod_{i,j} \pi^{\kappa_{ij}} (1-\pi)^{1-\kappa_{ij}} = \left(\frac{\pi}{1-\pi}\right)^{F_1} (1-\pi)^F = \lambda \exp\left(\beta F_1\right),$$

where $F_1 = \sum \kappa_{ij}$, $F = ||\mathcal{K}||$ is the size of $\mathcal{K}$, $\beta = \log(\frac{\pi}{1-\pi})$, and $\lambda = (1-\pi)^F$.

Now consider the relationship between two VHPs in the network. The probability of two similar VHPs having the same 0-1 status is higher than the probability of having different 0-1 assignments. Let $S_{ij,i'j'}$ be the similarity between two VHPs (v_i, b_j) and $(v_{i'}, b_{j'})$. We model the the probability of $\mathcal{K}$ conditional on the similarity network as proportional to

$$\exp(\gamma_1 F_{01} + \gamma_2 F_{11} + F_{00}),$$

where (γ_1, γ_2) are parameters indicating the penalty/award of having the 0-1 and 1-1 assignment for a pair of neighbors. The $F_{cc'}$ is defined as the total similarities between $(c, c'), c, c' = 0, 1$ pairs of VHPs,

namely

$$F_{01} = \sum_{(i,j)\neq(i',j')\in\mathcal{K}} (1-\kappa_{ij})\kappa'_{i'j'}S_{ij,i'j'} + (1-\kappa'_{i'j'})\kappa_{ij}S_{ij,i'j'},$$
$$F_{11} = \sum_{(i,j)\neq(i',j')\in\mathcal{K}} \kappa_{ij}\kappa'_{i'j'}S_{ij,i'j'},$$
$$F_{00} = \sum_{(i,j)\neq(i',j')\in\mathcal{K}} (1-\kappa_{ij})(1-\kappa'_{i'j'})S_{ij,i'j'}.$$

Therefore, the probability of the assignment conditional on the similarity network is,

$$\Pr(\mathcal{K}|\theta) = \frac{1}{Z(\theta)}\exp(U(\mathcal{K})) = \frac{1}{Z(\theta)}\exp\left(\beta F_1 + \gamma_1 F_{01} + \gamma_2 F_{11} + F_{00}\right)$$

where $\theta = (\beta, \gamma_1, \gamma_2)$ are the parameters, and $Z(\theta)$ is the normalizing factor.

With this distribution function, for any $\kappa_{ij} \in \mathcal{K}$, we can calculate

$$\frac{\Pr(\kappa_{ij}=1|\mathcal{K}_{[-ij]})}{Pr(\kappa_{ij}=0|\mathcal{K}_{[-ij]})} = \exp\left(\beta + (\gamma_2-\gamma_1)m_1^{ij} + (\gamma_1-1)m_0^{ij}\right)$$

where $\mathcal{K}_{[-ij]} = \mathcal{K}\setminus\kappa_{ij}$, $m_1^{ij} = \sum_{\kappa'_{i'j'}\in\mathcal{K}_{[-ij]},\kappa'_{i'j'}=1} S_{ij,i'j'}$, $m_0^{ij} = \sum_{\kappa'_{i'j'}\in\mathcal{K}_{[-ij]},\kappa'_{i'j'}=0} S_{ij,i'j'}$. Then the log-odds of the probability $\Pr(\kappa_{ij}=1|\mathcal{K}_{[-ij]},\theta)$ is,

$$\text{logit}\big(\Pr(\kappa_{ij}=1|\mathcal{K}_{[-ij]},\theta)\big) = \beta + (\gamma_2-\gamma_1)m_1^{ij} + (\gamma_1-1)m_0^{ij}.$$

Denote $\gamma_+ = \gamma_2 - \gamma_1$ and $\gamma_- = \gamma_1 - 1$. Then we have,

$$\text{logit}\big(\Pr(\kappa_{ij}=1|\mathcal{K}_{[-ij]},\theta)\big) = \beta + \gamma_+ m_1^{ij} + \gamma_- m_0^{ij}.$$

The similarity between two VHPs and the generalized probability model for a VHP to interact. The MRF network model is constructed based on the similarity between two VHPs $S_{ij,i'j'}$. Various similarity measures between VHPs can be defined. In this study, we define the similarity between two VHPs as the similarity between the two viruses and the similarity between the two hosts, namely, $S_{ij,i'j'} = S_{v_i,v_{i'}} + S_{b_j,b_{j'}}$. To measure the similarity between two genomic sequences, we previously developed novel measures d_2^* and d_2^S for alignment-free sequence comparison based on k-mers as genomic signatures[58–61], and showed that the dissimilarity measure d_2^* and d_2^S have high correlation with alignment-based distance measures[62]. Since viruses are highly diverse and alignments of highly divergent sequences are not reliable, alignment-free measures are more suitable for sequence comparison than the alignment-based methods. Furthermore, Ahlgren *et al.*[25] showed that d_2^* outperformed d_2^S for the comparison of virus and bacterial sequences for the purpose of virus-host interaction prediction. Therefore, here we choose to use d_2^* and transform it to s_2^* to measure the similarity between two sequences.

For each sequence, we represent it by the normalized k-mer frequency vector ($\tilde{f}_{\mathbf{w}}, \mathbf{w}\in\mathcal{A}^k$), where $\mathcal{A}$ is the set of alphabet $\{A, C, G, T\}$, k is the length of word, and

$$\tilde{f}_{\mathbf{w}} = (N_{\mathbf{w}} - E_{\mathbf{w}})/\sqrt{E_w},$$

with $N_{\mathbf{w}}$ and $E_{\mathbf{w}}$ being the observed and expected numbers of occurrences of word $\mathbf{w}$ in the sequence. The expected count is calculated under a certain Markov chain model. Since it was shown in[25] that $k = 6$ and second order Markov chain performed well in virus-host interaction prediction, we choose $k = 6$ and second order Markov chain in this study. The similarity between two sequences, s_2^*, is defined as the un-centered correlation between their corresponding normalized frequency vectors. That is,

$$s_2^* = 1 - 2d_2^* = \sum_{\mathbf{w}\in\mathcal{A}^k} \bar{f}_{\mathbf{w}}^{(1)}\bar{f}_{\mathbf{w}}^{(2)}$$

where $d_2^*(v,b)$ is the dissimilarity measure used in the previous studies, and $\bar{f}_{\mathbf{w}} = \tilde{f}_{\mathbf{w}}/||f||$ with $||f||$ being the Euclid norm of the feature vector $f = \left(\tilde{f}_{\mathbf{w}}, \mathbf{w}\in\mathcal{A}^k\right)$. Thus, we define the similarity

$$S_{ij,i'j'} = s_2^*(v_i, v_{i'})I(b_j = b_{j'}) + s_2^*(b_j, b_{j'})I(v_i = v_{i'}).$$

Plug $S_{ij,i'j'}$ into the logit function,

$$\text{logit}\big(\Pr(\kappa_{ij}=1|\mathcal{K}_{[-ij]},\theta)\big) = \beta + \gamma_+ SV_+^{ij} + \gamma_- SB_+^{ij} + \gamma_- SV_-^{ij} + \gamma_- SB_-^{ij}. \quad (1)$$

$$SV_+^{ij} = \sum_{I(v',b_j)=1, v'\neq v_i} s_2^*(v', v_i),$$
$$SB_+^{ij} = \sum_{I(v_i,b')=1, b'\neq b_j} s_2^*(b', b_j),$$
$$SV_-^{ij} = \sum_{I(v',b_j)=0, v'\neq v_i} s_2^*(v', v_i),$$
$$SB_-^{ij} = \sum_{I(v_i,b')=0, b'\neq b_j} s_2^*(b', b_j).$$

The above formulation takes into account both the similarity network between viruses, and the similarity network between hosts. In our dataset, however, each virus has only one reported host. So when we train the model using the current dataset, both SB_+^{ij} and SB_-^{ij} are set to zero. Then the model reduces to,

$$\text{logit}\big(\Pr(\kappa_{ij} = 1|\mathcal{K}_{[-ij]}, \theta)\big) = \beta + \gamma_+ SV_+^{ij} + \gamma_- SV_-^{ij}.$$

Though the terms SB_+^{ij} and SB_-^{ij} cannot be used given the current dataset, as more virus-host pairs are collected in the training data, the host-host similarity network will contribute to the prediction model and the two-layer MRF network will be fully utilized based on Eq. (1).

The assumption that any VHP has the same probability π for interaction is not realistic. Different pairs of virus and host have different features that determines the probability of interaction. For example, the probability can be associated with the similarity between the virus and the host. Thus, the probability π is modeled specifically to each individual pair (v_i, b_j),

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha + \beta s_2^*(v_i, b_j). \tag{2}$$

Then the logit model with the generalized probability can be written as,

$$\text{logit}\left(\Pr(\kappa_{ij} = 1|\kappa_{[-ij]}, \theta)\right) = \alpha + \beta s_2^*(v_i, b_j) + \gamma_+ SV_+^{ij} + \gamma_- SV_-^{ij}.$$

Therefore, the network-based MRF for predicting virus-host interaction is finally written as a logistic regression model where the predictors are the features of virus-virus similarity and virus-host similarity,

$$\text{logit}(\Pr(I(v,b) = 1)) = \alpha + \beta s_2^*(v,b) + \gamma_+ SV_+(v,b) + \gamma_- SV_-(v,b), \tag{3}$$

where α is a constant, $(\beta, \gamma_+, \gamma_-)$ measure the contributions of the features $s_2^*(v,b)$, $SV_+(v,b)$, and $SV_-(v,b)$, respectively. We expect that β and γ_+ to be positive and γ_- to be negative. However, we do not make these assumptions and let the data inform us the values of these parameters. To learn the parameters, we trained the model in a smaller training dataset, and predicted virus-host interactions in the network of all viruses and hosts. Since the scales of $SV_+(v,b)$ and $SV_-(v,b)$ are proportional to the scale of the dataset, in practice we used the normalized variables, that is,

$$SV_+(v,b) = \frac{1}{||H_b||}\sum_{v'\in H_b} s_2^*(v,v'), \quad SV_-(v,b) = \frac{1}{||H_b^c||}\sum_{v'\in H_b^c} s_2^*(v,v'),$$

where $H_b = \{v'|I(v',b) = 1, v' \neq v\}$, $H_b^c = \{v'|I(v',b) = 0, v' \neq v\}$, and $||\cdot||$ is the size of the set. When H_b or H_b^c is an empty set, the value of $SV_+(v,b)$ or $SV_-(v,b)$ is set to zero.

To achieve the best performance, in addition to the similarity score s_2^*, we integrate other types of features, including the CRISPR score and the alignment score between the virus v_i and host b_j into the framework.

Sharing of CRISPR spacers between the virus and the host. CRISPR systems play an important role as an adaptive and heritable immune system of prokaryotes. They help the host fight against the invasion of specific viruses by inserting small fragments of viral genomes (typically 21-72bp) as spacers into a CRISPR locus. The spacers are transcribed and are used as a guide by a Cas complex to target the degradation of the corresponding viral DNA[63].

Given a host genome, the CRISPR locus can be computationally located and thus the spacers can be extracted. In our study, we used the CRISPR Recognition Tool (`CRT`)[64] to find spacers. The spacers in a host genome (if available) were aligned to a viral genome by `blastn`[65] and alignment with E-value less than 1 were recorded. This threshold was chosen the same as the one used in a previous study[24]. For each pair of virus and host, we define the score $S_{CRISPR}(v,b)$ as the largest value of $-\log$(E-value). If

there is no match between a virus and host, a score of zero is assigned. The detailed parameters for the implementation of the procedures are given in the supplementary materials.

With the CRISPR information, we modify the assumption of π_{ij} in Eq. (2) to

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha + \beta s_2^*(v_i, b_j) + \eta S_{CRISPR}(v_i, G_{b_j}),$$

and our logistic regression model in Eq. (3) to

$$\begin{aligned}\text{logit}(\Pr(I(v,b)=1)) =& \alpha + \beta s_2^*(v,b) + \gamma_+ SV_+(v,b) + \gamma_- SV_-(v,b) \\ &+ \eta S_{CRISPR}(v, G_b),\end{aligned} \tag{4}$$

(Network with CRISPR)

where G_b is the set of hosts that belong to the same genus as host b, and

$$S_{CRISPR}(v, G_b) = \max_{b' \in G_b} S_{CRISPR}(v, b').$$

Due to the limited availability of CRISPR information in the training data, as shown in Fig.1b, we group hosts by genus for the CRISPR feature.

The fraction of virus genome aligning to the host genome. Viruses and their hosts frequently exchange genetic materials and viruses play important roles in horizontal gene transfer. Therefore, similar regions in virus and host genomes can provide a strong evidence for linking a virus to its potential host. On the one hand, phages, especially those temperate phages, are able to integrate their own genomes to the hosts. On the other hand, phages can obtain genetic material from their hosts. If a genetic element brings an evolutionary advantage to the virus, the borrowed genetic segment will be preserved in the viral genome[24]. One example is cyanophages, phages that infect cyanobacteria. Many cyanophages have acquired and express host photosystem genes that are thought to bolster photosynthetic energy during infection.[66].

Similar to the method in[24], we used `blastn` to find similarities between each pair of virus and host genomes. For each pair of virus and host, their similarity, $S_{blastn}(v, b)$, is defined as the fraction of the virus genome that can be mapped to the host genome. Only matches with percent identity higher than 90% are used for prediction. Note that different parts of the virus genome can be matched to different positions on the host genome and all contribute to the coverage percentage. We used the same parameter setting as in[24] for our analysis. The detailed parameter settings can be found in the supplementary material.

Finally, with the CRISPR feature and the alignment-based similarity, we have the following model:

$$\begin{aligned}\text{logit}(\Pr(I(v,b)=1)) =& \alpha + \beta s_2^*(v,b) + \gamma_+ SV_+(v,b) + \gamma_- SV_-(v,b) \\ &+ \eta S_{CRISPR}(v, G_b) + \delta S_{blastn}(v,b).\end{aligned} \tag{5}$$

(Network with CRISPR and BLAST)

Incorporation of WIsH score for predicting host of virus contigs. In many metagenomic studies, the whole genome of a virus may not be available. Instead, only parts of the virus genome referred as contigs that were assembled from shotgun reads are known. Several algorithms such as VirFinder and VirSorter etc.[67–72] can be used to decide if the contigs come from virus genomes. Our objective is to predict the hosts for full virus genomes as well as viral contigs.

Galiez *et al.*[26] recently developed a program, WIsH, to predict the hosts of viral contigs and showed that WIsH outperforms d_2^* for predicting the hosts of viral contigs as short as 5kb. WIsH trains a homogeneous Markov chain model for each bacteria genome, and calculates the likelihood of a viral contig based on each Markov chain model. Instead of using $s_2^*(v, b)$ as a feature, we hereby replace it with the log-likelihood of viral contig v fitting to the Markov chain model of bacteria b, $S_{WIsH}(v, b)$. Then the model for predicting the host b of viral contig v becomes,

$$\begin{aligned}\text{logit}(\Pr(I(v,b)=1)) =& \alpha + \beta S_{WIsH}(v,b) + \gamma_+ SV_+(v,b) + \gamma_- SV_-(v,b) \\ &+ \eta S_{CRISPR}(v, G_b) + \delta S_{blastn}(v,b).\end{aligned} \tag{6}$$

(Network using WIsH scores with CRISPR and BLAST)

Note that both $SV_+(v,b)$ and $SV_-(v,b)$ are still computed by s_2^*, since WIsH is not able to depict the similarities between viral contigs.

Model training and evaluation. Among the 1,427 viruses, we use the set of 352 viruses whose exact host genome sequences were known and the set of their corresponding 71 hosts as the positive training set. We randomly select 352 pairs of virus-host within the 352 viruses and 71 hosts as negative training data. To alleviate potential false negative interactions, we require that the the selected hosts for each virus is not in the same genus level as the true host. We then learned the model based on the training data for the various models.

It is possible that the selected 352 non-interacting pairs may contain some positive-yet-unknown interaction pairs, which may influence the training and test results. We recognized this possibility while assuming the fraction of such pairs is relatively low since the virus-host interaction is specific so that the overall fraction of virus-host interacting pairs among all the pairs is very small. The additional requirement that the host in a virus-host pair of a different genus level further mitigated this potential problem.

The trained models are then used to predict the hosts of the remaining 1,075 viruses against 31,986 candidate prokaryotic hosts. For each virus, we estimate its probability of infecting any bacteria, and the bacteria with the highest probability was predicted as its host. For a taxonomic group $\mathcal{L}$ at an upper taxonomic level containing a set of hosts, we define the prediction score between v and $\mathcal{L}$ as the maximum probability between v and all hosts in $\mathcal{L}$, that is

$$P(I(v,\mathcal{L})=1)=\max_{b\in\mathcal{L}} P(I(v,b)=1).$$

We predict the host group of the virus v by the one having the highest prediction score $P(I(v,\mathcal{L})=1)$. In case of ties, we first checked the number of hosts having the highest probability in each group and chose the one with the largest number of hosts having the highest probability. Further, if there are more than one taxon with the same number of bacteria having the highest probability, all of taxa are reported.

We then compared the predicted host taxonomic groups with the true taxonomic group of every virus at several taxonomic levels: genus, family, order, class, and phylum. At a particular taxonomic level $\mathcal{L}$, let $\mathcal{T}_v$ be the set of predicted groups and $C_{\mathcal{L}}(v)=I(h_v,\mathcal{T}_v)/||\mathcal{T}_v||$, where $I(h_v,\mathcal{T}_v)=1$ if the true host of v, h_v, belongs to the set of the predicted host groups $\mathcal{T}_v$, and $I(h_v,\mathcal{T}_v)=0$, otherwise. The prediction accuracy for a certain taxonomic level is defined as

$$Acc_{\mathcal{L}}=\frac{1}{||\mathcal{V}||}\sum_{v\in\mathcal{V}} C_{\mathcal{L}}(v),$$

where $\mathcal{V}$ is the set of viruses for prediction.

Statistical measures for determining the significant host taxon. We employ a few statistical methods to determine the host taxon at a high level, such as the genus and phylum levels, that most likely associate with the virus. Determine the host at a high level is especially helpful in practice, when the true host strain is not but its close relatives are in the host database. Wilcoxon rank-sum test is used to test for each taxon if the prediction score $P(I(v,b)=1)$ for the hosts in this taxon has higher mean than the hosts in other taxa. To achieve a reasonable statistical test power, we only test the taxa that include more than 20 host strains. Bonferroni correction is applied to adjust the p-values derived from multiple tests.

The effect size is used to measure the difference in the mean scores between two groups. The effect size is not affected by sample size, enabling us to compare across different taxa even if they include different numbers of hosts[73,74]. The effect size is defined as the standardized difference between the mean of prediction scores $P(I(v,b)=1)$ of the hosts in a particular taxon and the mean of the scores of the hosts in other taxa, normalized by the standard deviation of the scores in other taxa. We compute the Z-score as $Z_{(v,b)}=\frac{P(I(v,b)=1)-\bar{P}(I(v,\cdot)=1)}{\sigma(P(I(v,\cdot)=1))}$, where $\bar{P}(I(v,\cdot)=1)$ is the mean of the predicted scores between v and all the hosts in other taxa, and $\sigma(P(I(v,\cdot)=1))$ is the standard deviation of the predicted scores between v and the hosts in other taxa pairs. Then the effect size for a taxon is the mean of $Z_{(v,b)}$ for all hosts in the taxon. The effect size >0.5 is considered as medium to large effect[74].

Data and code availability. All the relevant data and code are available from the authors upon request.

Acknowledgements. We thank Dr. Michael S. Waterman and Dr. J. Cesar Ignacio Espinoza at University of Southern California (USC) for useful discussions, and Dr. Yang Lu at USC for providing the CAFE software. Dr. Yang Lu and Ms. Mengge Zhang at USC participated in the discussions and some of the calculations at the beginning of the project. This work was supported by NIH grant R01GM120624 to JF, FS, and NA; NSF grant DMS-1518001 to FS and JF; and grant GBMF3779 from the Gordon and Betty Moore Foundation Marine Microbiology Initiative to JF. JR was partially supported by a USC Provost Fellowship.

Author contributions. J.R. and F.S. conceived the study and designed the method. W.W. and J.R. developed the algorithms and implemented the data analysis. W.W., J.R., F.S., and N.A. wrote the manuscript. N.A. J.F. and J.B. helped in discussing and interpreting the results, and revised the manuscript critically. All authors read and approved the final manuscript.

References

[1] Breitbart, M. & Rohwer, F. Here a virus, there a virus, everywhere the same virus? *Trends in microbiology* **13**, 278–284 (2005).

[2] Breitbart, M. *et al.* Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences* **99**, 14250–14255 (2002).

[3] Fierer, N. *et al.* Metagenomic and small-subunit rrna analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Applied and environmental microbiology* **73**, 7059–7066 (2007).

[4] Hurwitz, B. L. & Sullivan, M. B. The pacific ocean virome (pov): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PloS one* **8**, e57355 (2013).

[5] Waller, A. S. *et al.* Classification and quantification of bacteriophage taxa in human gut metagenomes. *The ISME journal* **8**, 1391–1402 (2014).

[6] Brum, J. R. *et al.* Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).

[7] Minot, S. *et al.* The human gut virome: inter-individual variation and dynamic response to diet. *Genome research* **21**, 1616–1625 (2011).

[8] Mirzaei, M. K. & Maurice, C. F. Menage a trois in the human gut: interactions between host, bacteria and phages. *Nature Reviews Microbiology* **15**, 397–408 (2017).

[9] Hannigan, G. D., Duhaime, M. B., Koutra, D. & Schloss, P. D. Biogeography & environmental conditions shape phage & bacteria interaction networks across the human microbiome. *bioRxiv* 144642 (2017).

[10] Gómez, P. & Buckling, A. Bacteria-phage antagonistic coevolution in soil. *Science* **332**, 106–109 (2011).

[11] Roux, S., Hallam, S. J., Woyke, T. & Sullivan, M. B. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *Elife* **4**, e08490 (2015).

[12] Norman, J. M. *et al.* Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160**, 447–460 (2015).

[13] Reyes, A. *et al.* Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *Proceedings of the National Academy of Sciences* **112**, 11941–11946 (2015).

[14] Mills, S. *et al.* Movers and shakers: influence of bacteriophages in shaping the mammalian gut microbiota. *Gut microbes* **4**, 4–16 (2013).

[15] Srinivasiah, S. *et al.* Phages across the biosphere: contrasts of viruses in soil and aquatic environments. *Research in Microbiology* **159**, 349–357 (2008).

[16] Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).

[17] Rohwer, F., Prangishvili, D. & Lindell, D. Roles of viruses in the environment. *Environmental microbiology* **11**, 2771–2774 (2009).

[18] Cann, A. J., Fandrich, S. E. & Heaphy, S. Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. *Virus genes* **30**, 151–156 (2005).

[19] Paez-Espino, D. *et al.* Uncovering earths virome. *Nature* **536**, 425–430 (2016).

[20] Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature communications* **5** (2014).

[21] Labonté, J. M. *et al.* Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *The ISME journal* **9**, 2386 (2015).

[22] Lan, F., Demaree, B., Ahmed, N. & Abate, A. R. Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nature biotechnology* (2017).

[23] Grissa, I., Vergnaud, G. & Pourcel, C. The crisprdb database and tools to display crisprs and to generate dictionaries of spacers and repeats. *BMC bioinformatics* **8**, 172 (2007).

[24] Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage–host relationships. *FEMS microbiology reviews* **40**, 258–272 (2016).

[25] Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A. & Sun, F. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Research* **45**, 39–53 (2017).

[26] Galiez, C., Siebert, M., Enault, F., Vincent, J. & Söding, J. Wish: who is the host? predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* **33**, 3113–3114 (2017).

[27] Carbone, A. Codon bias is a major factor explaining phage evolution in translationally biased hosts. *Journal of Molecular Evolution* **66**, 210–223 (2008).

[28] Gouy, M. & Gautier, C. Codon usage in bacteria: correlation with gene expressivity. *Nucleic acids research* **10**, 7055–7074 (1982).

[29] Villarroel, J. *et al.* Hostphinder: a phage host prediction tool. *Viruses* **8**, 116 (2016).

[30] Zhang, M. *et al.* Prediction of virus-host infectious association by supervised learning methods. *BMC bioinformatics* **18**, 60 (2017).

[31] Deng, M., Zhang, K., Mehta, S., Chen, T. & Sun, F. Prediction of protein function using protein–protein interaction data. *Journal of Computational Biology* **10**, 947–960 (2003).

[32] Jansen, R. *et al.* A bayesian networks approach for predicting protein-protein interactions from genomic data. *science* **302**, 449–453 (2003).

[33] Jiang, R., Gan, M. & He, P. Constructing a gene semantic similarity network for the inference of disease genes. *BMC systems biology* **5**, S2 (2011).

[34] Freudenberg, J. & Propping, P. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics* **18**, S110–S115 (2002).

[35] Zhang, W., Sun, F. & Jiang, R. Integrating multiple protein-protein interaction networks to prioritize disease genes: a bayesian regression approach. *BMC bioinformatics* **12**, S11 (2011).

[36] Cheng, F. *et al.* Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS computational biology* **8**, e1002503 (2012).

[37] Shapiro, J. W. & Putonti, C. Gene networks provide a high-resolution view of bacteriophage ecology. *bioRxiv* 148668 (2017).

[38] Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. Reticulate representation of evolutionary and functional relationships between phage genomes. *Molecular biology and evolution* **25**, 762–777 (2008).

[39] Wu, G. A., Jun, S.-R., Sims, G. E. & Kim, S.-H. Whole-proteome phylogeny of large dsdna virus families by an alignment-free method. *Proceedings of the National Academy of Sciences* **106**, 12826–12831 (2009).

[40] Zhang, Q., Jun, S.-R., Leuze, M., Ussery, D. & Nookaew, I. Viral phylogenomics using an alignment-free method: A three-step approach to determine optimal length of k-mer. *Scientific reports* **7**, 40712 (2017).

[41] Routy, B. *et al.* Gut microbiome influences efficacy of pd-1–based immunotherapy against epithelial tumors. *Science* (2017). URL `http://science.sciencemag.org/content/early/2017/11/01/science.aan3706`. `http://science.sciencemag.org/content/early/2017/11/01/science.aan3706.full.pdf`.

[42] Derrien, M., Belzer, C. & de Vos, W. M. Akkermansia muciniphila and its role in regulating host functions. *Microbial pathogenesis* **106**, 171–181 (2017).

[43] Flores, C. O., Meyer, J. R., Valverde, S., Farr, L. & Weitz, J. S. Statistical structure of host–phage interactions. *Proceedings of the National Academy of Sciences* **108**, E288–E297 (2011).

[44] Flores, C. O., Valverde, S. & Weitz, J. S. Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages. *The ISME journal* **7**, 520 (2013).

[45] Sullivan, M. B., Waterbury, J. B. & Chisholm, S. W. Cyanophages infecting the oceanic cyanobacterium prochlorococcus. *Nature* **424**, 1047–1051 (2003).

[46] Wichels, A. *et al.* Bacteriophage diversity in the north sea. *Applied and Environmental Microbiology* **64**, 4128–4133 (1998).

[47] Chibani-Chennoufi, S., Bruttin, A., Dillmann, M.-L. & Brüssow, H. Phage-host interaction: an ecological perspective. *Journal of bacteriology* **186**, 3677–3686 (2004).

[48] Stepanauskas, R. Single cell genomics: an individual look at microbes. *Current opinion in microbiology* **15**, 613–620 (2012).

[49] Swan, B. K. *et al.* Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proceedings of the National Academy of Sciences* **110**, 11463–11468 (2013).

[50] Yutin, N. *et al.* Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nature microbiology* **3**, 38 (2018).

[51] van Passel, M. W. *et al.* The genome of akkermansia muciniphila, a dedicated intestinal mucin degrader, and its use in exploring intestinal metagenomes. *PloS one* **6**, e16876 (2011).

[52] Shapiro, J. W. & Putonti, C. Gene networks provide a high-resolution view of bacteriophage ecology. *bioRxiv* (2017). URL `https://www.biorxiv.org/content/early/2017/06/11/148668`. `https://www.biorxiv.org/content/early/2017/06/11/148668.full.pdf`.

[53] Weitz, J. S., Beckett, S. J., Brum, J. R., Cael, B. & Dushoff, J. Lysis, lysogeny and virus–microbe ratios. *Nature* **549**, E1 (2017).

[54] Dao, M. C. *et al.* Akkermansia muciniphila and improved metabolic health during a dietary intervention in obesity: relationship with gut microbiome richness and ecology. *Gut* gutjnl–2014 (2015).

[55] Caesar, R., Tremaroli, V., Kovatcheva-Datchary, P., Cani, P. D. & Bäckhed, F. Crosstalk between gut microbiota and dietary lipids aggravates wat inflammation through tlr signaling. *Cell metabolism* **22**, 658–668 (2015).

[56] Stern, A., Mick, E., Tirosh, I., Sagy, O. & Sorek, R. Crispr targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome research* **22**, 1985–1994 (2012).

[57] Li, S. Z. Markov random field models in computer vision. In *European conference on computer vision*, 361–370 (Springer, 1994).

[58] Song, K. *et al.* Alignment-free sequence comparison based on next-generation sequencing reads. *Journal of Computational Biology* **20**, 64–79 (2013).

[59] Song, K. *et al.* New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Briefings in Bioinformatics* **15**, 343–353 (2014).

[60] Wan, L., Reinert, G., Sun, F. & Waterman, M. S. Alignment-free sequence comparison (ii): theoretical power of comparison statistics. *Journal of Computational Biology* **17**, 1467–1490 (2010).

[61] Reinert, G., Chew, D., Sun, F. & Waterman, M. S. Alignment-free sequence comparison (i): statistics and power. *Journal of Computational Biology* **16**, 1615–1634 (2009).

[62] Ren, J. *et al.* Inference of markovian properties of molecular sequences from ngs data and applications to comparative genomics. *Bioinformatics* **32**, 993–1000 (2015).

[63] Horvath, P. & Barrangou, R. Crispr/cas, the immune system of bacteria and archaea. *Science* **327**, 167–170 (2010).

[64] Bland, C. *et al.* Crispr recognition tool (crt): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC bioinformatics* **8**, 209 (2007).

[65] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403–410 (1990).

[66] Sharon, I. *et al.* Comparative metagenomics of microbial traits within oceanic viral communities. *The ISME journal* **5**, 1178 (2011).

[67] Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. Virfinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).

[68] Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. Virsorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).

[69] Roux, S. *et al.* Metavir: a web server dedicated to virome analysis. *Bioinformatics* **27**, 3074–3075 (2011).

[70] Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome research* **26**, 1721–1729 (2016).

[71] Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using diamond. *Nature methods* **12**, 59–60 (2015).

[72] Truong, D. T. *et al.* Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nature methods* **12**, 902–903 (2015).

[73] Sullivan, G. M. & Feinn, R. Using effect sizeor why the p value is not enough. *Journal of graduate medical education* **4**, 279–282 (2012).

[74] Coe, R. It's the effect size, stupid: What effect size is and why it is important (2002).

[75] Consortium, H. M. P. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).

[76] Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).

Appendix A. Supplementary methods

Sharing of CRISPRs between the virus and the host. We used `CRT1.2-CLI`[64] to find CRISPRs in all bacterial genomes, with parameters `-minNR 3 -minRL 20 -maxRL 50 -minSL 20 -maxSL 60 -searchWL 7`. All recognized CRISPRs were merged to one file to construct a BLAST database using `makeblastdb` (`BLAST 2.6.0`). We then searched all viral genomes against the database using `blastn` with parameters `-evalue 1 -gapopen 10 -penalty -1 -gapextend 2 -word_size 7 perc_identity 90 -dust no -task blastn-short`.

Fraction of virus genome mapped to the host genome. A BLAST database was created for all bacterial genomes by `makeblastdb` (`BLAST 2.6.0`). We then searched all viral genomes against the database by `blastn` with parameters `-word_size 11 -evalue 0.01 -reward 1 -penalty -2 -gapopen 0 -gapextend 0 perc_identity 90`.

Incorporation of the WIsH score for predicting hosts of virus contigs. WIsH[26] scores were computed using `WIsH 1.0` with the default parameters.

Consideration of co-abundance profile. In order to investigate if co-abundance can help with prediction of virus-host interaction, we tried to incorporate this feature to the model in a smaller data set to evaluate its contribution. The data set included a subset of 2,695 prokaryotic reference genomes and 1,403 viruses. A total of 148 stool metagenomic samples from the Human Microbiome Project (HMP)[75] and 103 metagenomes from Tara Ocean (filter size 0.22 to 3 μm)[76] were collected. We used `centrifuge` (`centrifuge-1.0.3-beta`) to compute the abundance of virus and bacteria genomes in each of the metagenome, resulting in an abundance profile of 251-dimensional vector for every virus and bacteria genome. The co-abundance feature $S_{co-abundance}(v,b)$ is defined by the Pearson correlation between the abundance profiles for the pair of virus and bacteria. We then modified the integrated model to

$$\text{logit}(P\{I(v,b)=1\}) = \alpha + \beta s_2^*(v,b) + \gamma_+ SV_+(v,b) + \gamma_- SV_-(v,b) + \delta S_{co-abundance}(v,b). \tag{A.1}$$

We compared the performance of this model with that of the model in Eq. (3). Both models were trained based on a subset of 308 viruses and 50 hosts, including 308 pairs of true interacting pairs and 308 randomly chosen negative pairs. After the models were trained, we predicted the hosts of 1,095 viruses. The results are shown in Fig. B.3. The co-abundance feature itself had weak prediction ability and adding it to the model did not help prediction. Therefore, we did not consider it as a feature in the final model presented in the main text.

Appendix B. Supplementary tables and figures

SAG name	True phylum	Predicted phylum
AAA160-P02	Bacteroidetes	Firmicutes
AAA160-C11	Marinimicrobia	Firmicutes
AAA160-I06	Marinimicrobia	**Marinimicrobia**
AAA076-E06	Proteobacteria	**Proteobacteria**
AAA160-D02	Proteobacteria	**Proteobacteria**
AAA160-J14	Proteobacteria	**Proteobacteria**
AAA160-J18	Proteobacteria	**Proteobacteria**
AAA168-P09	Proteobacteria	**Proteobacteria** [a]
AAA300-J04	Proteobacteria	**Proteobacteria**
AAA160-J20	Thaumarchaeota	Verrucomicrobia
AAA164-A08	Verrucomicrobia	**Verrucomicrobia**
AAA164-A21	Verrucomicrobia	Proteobacteria
AAA164-B23	Verrucomicrobia	**Verrucomicrobia**
AAA164-I21	Verrucomicrobia	**Verrucomicrobia**
AAA164-L15	Verrucomicrobia	Actinobacteria
AAA164-M04	Verrucomicrobia	**Verrucomicrobia**
AAA164-P11	Verrucomicrobia	Proteobacteria
AAA168-E21	Verrucomicrobia	**Verrucomicrobia** [b]

[a] Two hosts from Verrucomicrobia and Firmicutes also have the highest score.
[b] Both Proteobacteria and Verrucomicrobia have the same number of hosts with the highest prediction score.

Table B.1. Host predictions for viral genomes revealed in SAGs and the comparison between the predicted hosts and the reported hosts at the phylum level in Labonte *et al.*[21]. The names in bold suggest correct predictions.

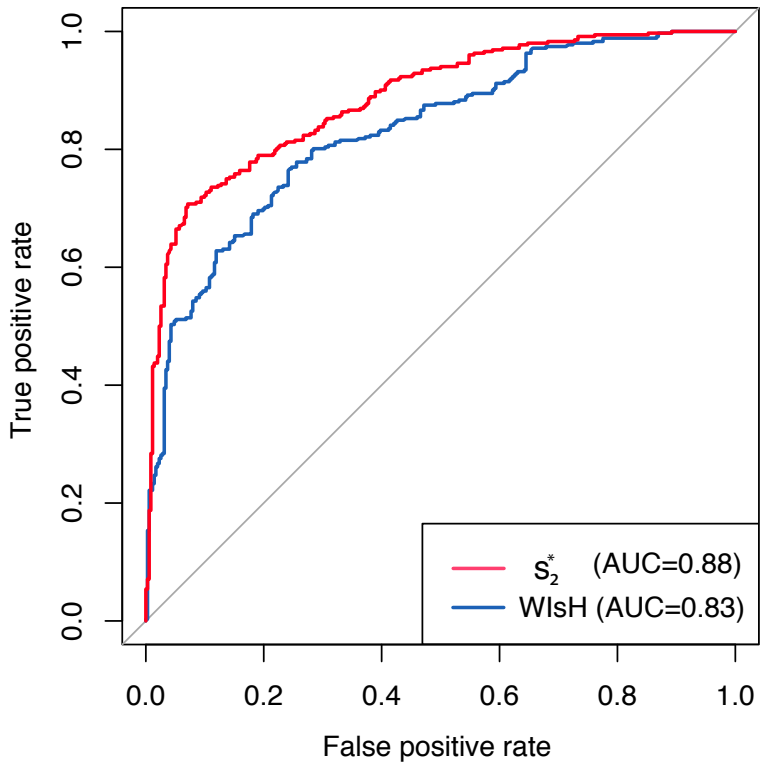

Figure B.1. ROC curves for predicting virus-host interactions using s_2^* and WIsH based on 352 virus-host interactions and randomly chosen negative virus-host complete genome pairs.

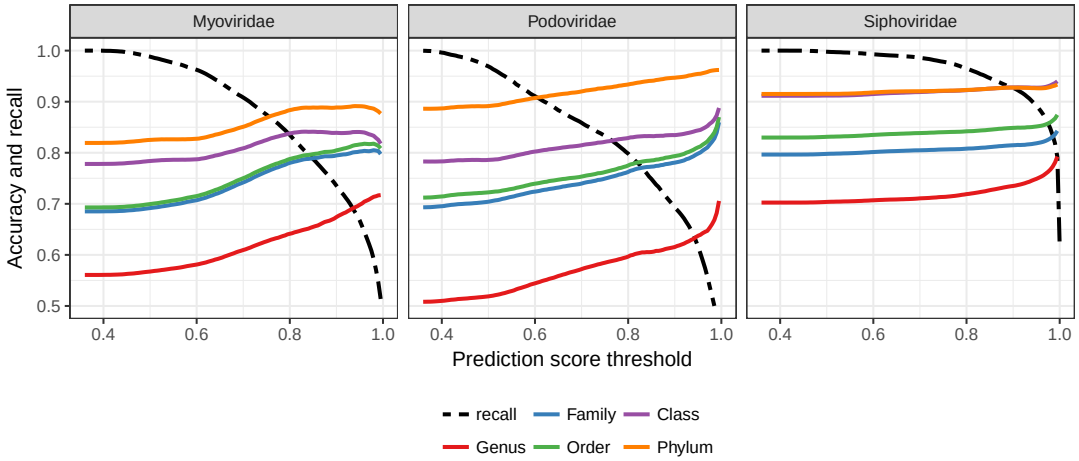

Figure B.2. Prediction improvement by thresholding on the prediction score when predicting on three viral families.

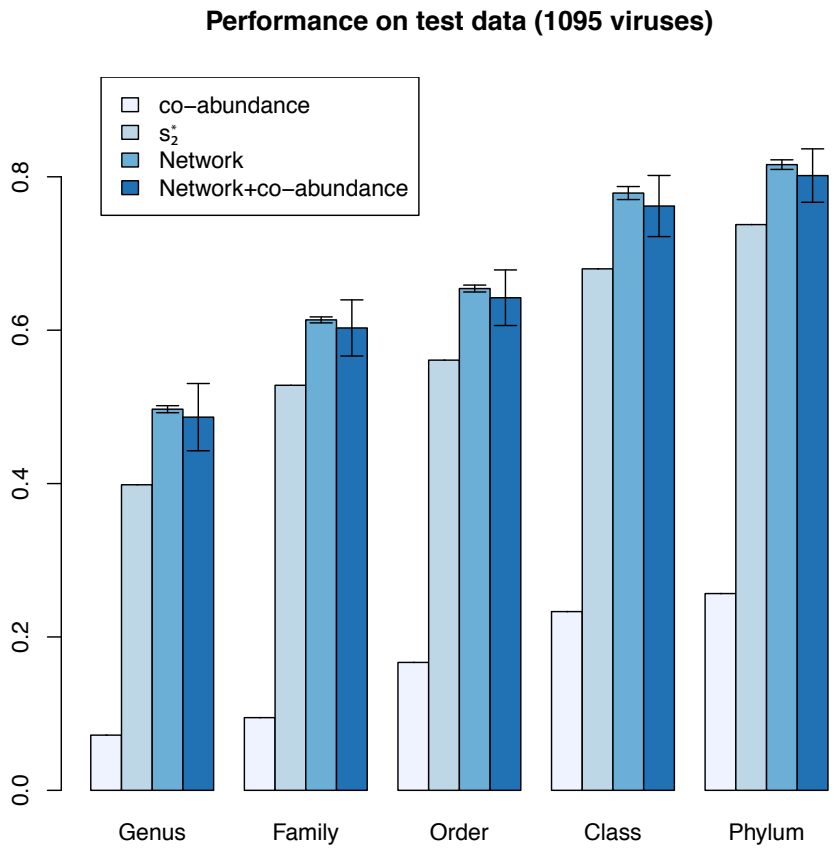

Figure B.3. The prediction accuracies of virus-host interactions for different methods using the individual feature, co-abundance, s_2^*, and the integrated network-based model in Eq. (3), and the integrated network-based model combined with the co-abundance feature in Eq. (A.1), respectively, from left to right. The results are binned by taxonomic level. Error bars show the 95% confidence intervals of the accuracies, based on 100 different (randomly selected) negative training sets.

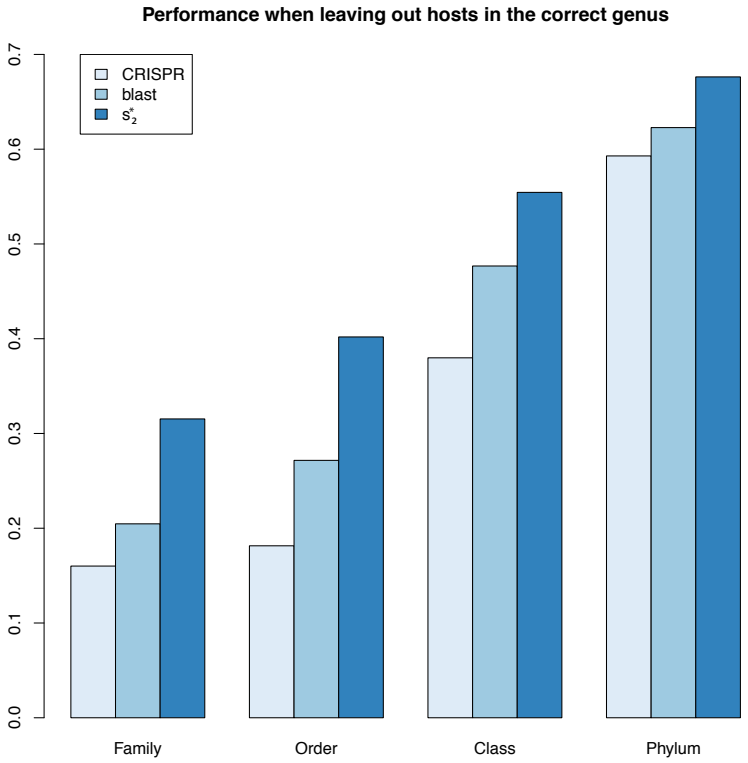

Figure B.4. The prediction accuracies of CRISPR, BLAST and s_2^*, respectively, when the hosts in the true genus level are excluded. Predictions were made by excluding all the true hosts in the genus level and were evaluated at higher taxonomic levels. Average prediction accuracies for the set of 1,075 viruses are shown. The performance for alignment-free measure s_2^* was least susceptible to this situation where the true host(s) was missing from the candidates.