

ProxiPhage: accurate viral genome reconstruction and host assignment with proximity-ligation sequencing

Gherman Uritskiy¹, Maximillian Press², Andrew Wiser¹, Jonas Grove¹, Ben Auch¹, Steve Eacker¹, Shawn Sullivan¹, Derek M. Bickhart³, Timothy P. L. Smith⁴, Mathew Sullivan^{5,6,7}, and Ivan Liachko¹✉

¹Phase Genomics, Seattle, WA 98109, USA

²Inscripta, Boulder, CO 80301, USA

³USDA Dairy Forage Research Center, Madison, WI 53593, USA

⁴USDA Meat Animal Research Center, Clay Center, NE 68933, USA

⁵Department of Microbiology, Ohio State University, Columbus, OH, 43210, USA

⁶Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, OH, 43210, USA

⁷Center of Microbiome Science, Ohio State University, Columbus, OH, 43210, USA

Viruses play crucial roles in the microbial composition and nutrient turnover of microbiomes, and yet they remain relatively understudied in their native environments. Despite many advancements in high-throughput whole-genome sequencing (WGS), sequence assembly, and annotation of viruses, the reconstruction of full-length viral genomes directly from metagenomic sequencing is possible primarily for the most abundant phage using long-read sequencing technologies. Additionally, the prediction of their cellular hosts remains difficult from conventional metagenomic sequencing alone. To address these gaps in the field and to accelerate the study of viruses directly in their native microbiomes, we developed ProxiPhage – a end-to-end bioinformatics platform for viral genome reconstruction and host attribution from metagenomic data using proximity-ligated library sequencing (i.e., Hi-C). We demonstrate the capabilities of the platform by recovering and characterizing the meta-virome of a variety of microbiomes, including a complex sheep fecal microbiome that has also been deeply sequenced with long-read HiFi reads, allowing for the assessment and benchmarking of the new methods. ProxiPhage can extract near-complete viral genomes even from highly fragmented short-read assemblies and can reliably predict their cellular hosts with minimal false positives. To our knowledge, ProxiPhage is the first software for performing these tasks. Being significantly cheaper than long-read sequencing of comparable depth, the incorporation of proximity-ligation sequencing in microbiome research using ProxiPhage shows promise to greatly accelerate future advancements in the field.

virus | phage | meta-virome | hi-c | metagenome | binning | microbiome | proximity ligation | virus host | metagenome-assembled genome

Correspondence: ivan@phagegenomics.com

Introduction

In the past two decades, the study of microbiome composition and function has risen to the forefront of both medical and basic research (1, 2). In host-associated microbiomes, metagenomic whole-genome sequencing (WGS) has been widely deployed to show that microbiota composition and metabolic function have great effects on the health of their host (1, 3, 4). Indeed, the gut microbiome alone has been linked with a wide range of human diseases and disorders (5, 6). Similarly, microbiomes found in water reservoirs (7), soil (8), and waste systems (9) were also found to play critical

roles in modulating the chemistry of their respective environments. However, while such research primarily focused on prokaryotic microorganisms, viruses have also been shown to have major effects on microbiome dynamics (10, 11). Playing the roles of predators in their respective ecosystems, their lytic activity makes them a critical component of population control and nutrient turnover (12), while their lysogenic activity has been demonstrated to play a role in antimicrobial resistance (AMR) gene transfer (13). Their contributions to microbiome composition dynamics make viruses critical to study in both medical and basic research.

High-throughput metagenomic WGS of microbial communities allowed the study of viruses directly in their native environments, and advancements in viral sequence annotation led to a rapid expansion of viral sequence databases (14, 15). However, the assembly of complete viral genomes from metagenomes remained a major challenge due to their fast mutation rates and subsequent high heterogeneity in their sequences (16, 17). Long-read sequencing technologies from both Oxford Nanopore and Pacific Biosciences allow for the extraction of near-complete viral sequences (18, 19), however, such sequencing is prohibitively expensive compared to shotgun sequencing of comparable depth and is typically only applied to the most abundant viruses (20). In prokaryotic genomes, this same challenge was overcome with the emergence of metagenomic binning software, which can predict groups of contigs belonging to the same genome. Coupled with methods to estimate the accuracy of such groupings with prokaryotic universal single-copy marker genes (21), such software has allowed the recovery of metagenome-assembled genomes (MAGs) from complex and diverse microbial communities (22–24).

Despite the widespread use and acceptance of metagenomic binning for prokaryotic genome recovery, similar advances have not been made for reconstructing viral metagenome-assembled genomes (vMAGs). One of the main challenges has been the inability to assess the completion and contamination of vMAGs from single-copy universal marker genes, as there is no such set known to exist for viruses (25). Several attempts have been made to bin viral contigs of select large phage genomes using conventional approaches (17, 26), however, such studies are rare due to the challenges of resolv-

ing closely related viral strains. Metagenomic binning approaches utilizing proximity-ligated library sequencing (Hi-C, 3C, and other derivatives of chromosome conformation capture) show particular promise to reconstruct vMAGs. Several studies reported being able to use the proximity-ligation signal to bin viral contigs together with their cellular host genome, with a high likelihood of these viral contigs belonging to the same original genome(27, 28). Marbouy et. al remarked on using a custom application of the Louvain algorithm to reconstruct several possible vMAGs (29). However, these approaches rely on the incorrect assumptions that each prokaryote may only host a single virus and that every virus may infect only one host (30), and were custom methods tailored to specific samples. To our knowledge, there is no currently available software designed for genome-resolved binning of vMAGs from fragmented metagenomic assemblies. The recent development of CheckV – software that can assess the completeness of viral sequences by comparing them to a large database of known viruses (31) enables such a software to be built and benchmarked.

Knowing the cellular hosts of viruses is critical for understanding their role in the microbiome, however, this information is lost during conventional shotgun or long-read sequencing, except viruses integrated into the host genomes (prophages) (32). In the past, most virus-host association studies focused on probe- and emulsion-based interaction capture (33) and CRISPR array spacer alignment (34), however, proximity-ligated library sequencing allows for the most robust and high-throughput approach, as internalized viral DNA can be ligated with the host chromosome DNA during cross-linking (34). One way of utilizing such data is to directly group viral contigs with their host genome during metagenomic binning with the proximity-ligation signal. This approach has been widely documented to produce robust virus-host associations (35, 36), however, this relies on the assumption that every virus can only infect one host, and is largely limited to prophages because more transient infections are likely to have weaker linkage signals (28). The second approach is to directly compare all viral sequences with all possible host MAGs to look for pairs with high proximity-ligation linkage signals, which allows for higher sensitivity and the identification of multiple host interactions from much weaker proximity ligation signal (37, 38). However, false-positive associations from poor library quality and read mis-alignments commonly found in proximity-ligation sequencing require robust linkage strength normalization and noise subtraction. To our knowledge, there is no software to perform such analysis, which has only been achieved on a sample-by-sample basis with custom methods. To overcome these gaps in the field and to accelerate high-throughput virus discovery and characterization, we developed ProxiPhage – a comprehensive end-to-end analysis platform for vMAG reconstruction and host prediction using proximity-ligated sequencing data.

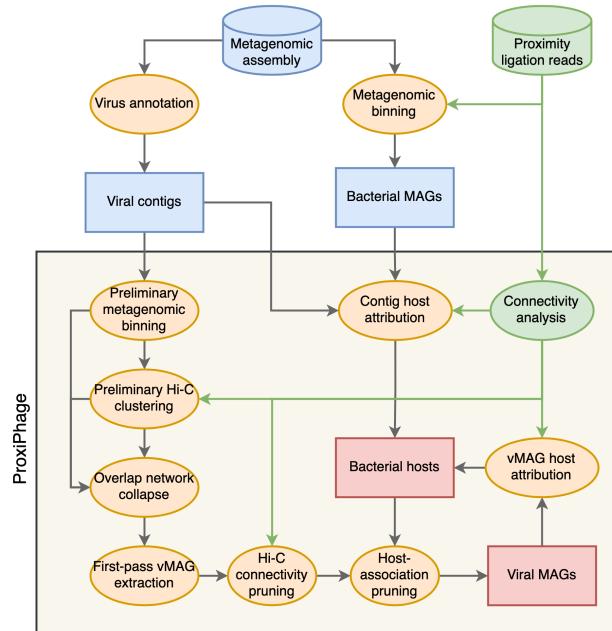


Fig. 1. Computational pipeline. Flowchart showing the outline of data processing in ProxiPhage. Cylinders represent main input data, rectangles represent sequence data, and ellipses represent methods or software. Green represents Hi-C data and uses of the Hi-C data in the pipeline, orange represents the main components of the platform, blue represents input metagenomic data, and red represents the main ProxiPhage outputs.

Results

Viral MAG extraction improved genome completion.

Viral contigs identified with VirSorter2 (39) in the short-read assembly were grouped based on likely membership to original viral genomes with the viral binning function of ProxiPhage (Fig. 1; see Methods for details). In short, clustering contigs based on Hi-C read linkages allowed for the extraction of groups of contigs that were likely in the same cell, phage envelope, or were otherwise in proximity with each other at the time of sampling. On the other hand, binning the contigs using more conventional metagenomic metrics such as tetranucleotide frequency profiles and mean read coverages allowed for the grouping of viral contigs that are likely of similar phylogeny and abundance. Intersecting and resolving these two cluster sets allows the reconstruction of viral metagenome-assembled genomes (vMAGs) even when multiple viruses are present in the same host.

In the sheep gut metagenome, ProxiPhage was able to place 791 viral contigs into 315 vMAGs (Table 1). The sizes of the viral MAGs ranged from 2 - 10 contigs or 11 - 197 Kb. While the majority of the vMAGs contained a small number of contigs, the clustering notably increased the average length of the viral genomes from 18 Kb to 45 Kb and the final sequence N50 from 23 Kb to 58 Kb (Fig. S1). The genome completion of the ProxiPhage vMAGs was compared to that of the original viral contigs using CheckV (31), which uses an extensive viral lineage and protein database to estimate the completion of a given phage sequence. We found that the ProxiPhage vMAGs had drastically improved CheckV completion met-

Category	Statistic	Sheep gut	Human stool	Cow rumen	Wastewater
General statistics	Hi-C reads (millions)	100	100	100	95.3
	WGS reads (million)	100	100	100	100
	Assembly size (Mb)	873	479	531	822
	Assembly N50	4138	5922	3871	2511
	Number of contigs	274,316	118,092	179,930	371,131
Viral contigs	Prokaryotic MAGs	365	155	178	238
	Number of viral contigs	2341	1609	1298	1013
	Unfiltered virus-host links	30,811	92,438	49,317	6561
	Filtered virus-host links	2497	1602	1158	1024
	Viruses with hosts	1559	1262	1003	705
Viral MAGs	Percent viruses with hosts	66.60%	78.43%	77.27%	69.60%
	Contigs in vMAGs	791	714	599	340
	Number of vMAGs	315	205	148	105
	Unfiltered vMAG-host links	11,015	18,692	10,469	1538
	Filtered vMAG-host links	373	215	146	126
	vMAGs with hosts	218	181	122	87
	Percent vMAGs with hosts	69.21%	88.29%	82.43%	82.86%

Table 1. Metagenomic data, assembly, binning and viral host attribution statistics from the four samples processed in this study.

rics (Fig. 2A). In particular, the number of near-complete viral genomes (>90% completion) improved from 9 to 73 after binning (Fig. 2C). For 276 out of the 351 (88%) of the vMAGs with a reliable reference, the completion improvements were also assessed using the alignments to the reference viral genomes from the long-read assembly (see next section).

HiFi long-read assembly allows for vMAG validation.

Due to the lack of a reliable universal marker gene set for viruses, the false-positive rates (or contamination) of the viral contig clustering needed to be evaluated with long-read technologies. The same sheep gut microbiome sample that was used for the short-read whole-genome sequencing (WGS) assembly was also deeply sequenced with long-read HiFi reads (38). The resulting long-read assembly had much greater contiguity than the short-read assembly – N50 of 279,621 bp and 4,138 bp, respectively. This allowed for the validation and accuracy assessment of vMAGs generated from the short-read assembly. To serve as a reliable reference for both completion and contamination estimation, phage sequences were annotated and excised from the long-read contigs using VirSorter2 (see Methods). The resulting excised viral genomes from the HiFi assembly were assumed to be 100% complete and 0% contaminated for the purposes of validating vMAGs from the short-read assembly. The short-read and long-read phage sequences were aligned to each other and the consistency of the vMAGs with the reference phages was assessed to estimate the percent completion and percent contamination of each vMAG that was present in the long-read assembly. To evaluate the validity of viral genome assessment using a long-read assembly, the estimated completion percentages of each phage contig and vMAG were compared to that estimated with CheckV. We found that the completion estimates from both methods produced similar values

for the vast majority of viral sequences (Fig. S2), with the majority of the completion estimates being within 10% of each other, and more than 81% being within 20% of each other. However, there were some sequences the completion of which was over-estimated by either CheckV or long-read viral references, particularly in short viral sequences.

Viral MAGs are supported by the long-read assembly.

Using the reference long-read virus sequences to estimate the completion of the vMAGs generated from the short-read assembly confirmed that the clustering method significantly improved viral genome completion (Fig. 2B), and increased the number of near-complete viral genomes from just 1 to 58 (Fig. 2C). The long-read references also allowed for the evaluation of vMAG contamination – erroneous groupings of contigs that originated from different phages. We found that most of the contigs from any given vMAG aligned to a single long-read viral genome reference, confirming that the contigs originated from the same phage in the sample (Fig. 2B). However, there were still several instances of contigs from the same vMAG aligning to different reference phages. In the network visualization of select clusters (Fig. S3), possible contamination can be seen at vMAGs identifiers 23, 266, 140, 20, and 50. In total, we found that just 17 of the 315 vMAGs (6%) had notable contamination of >10% (Table 2). The majority of these false positives were found to be closely related prophage sequences incorporated into the sample bacterial genome and thus could not be separated by the ProxiPhage algorithm.

Novel linkage signal normalization for evaluating virus-host interactions.

ProxiPhage also features a novel approach for using proximity-ligation sequencing to infer the likely prokaryotic MAG hosts for viral contigs and vMAGs. The host finder

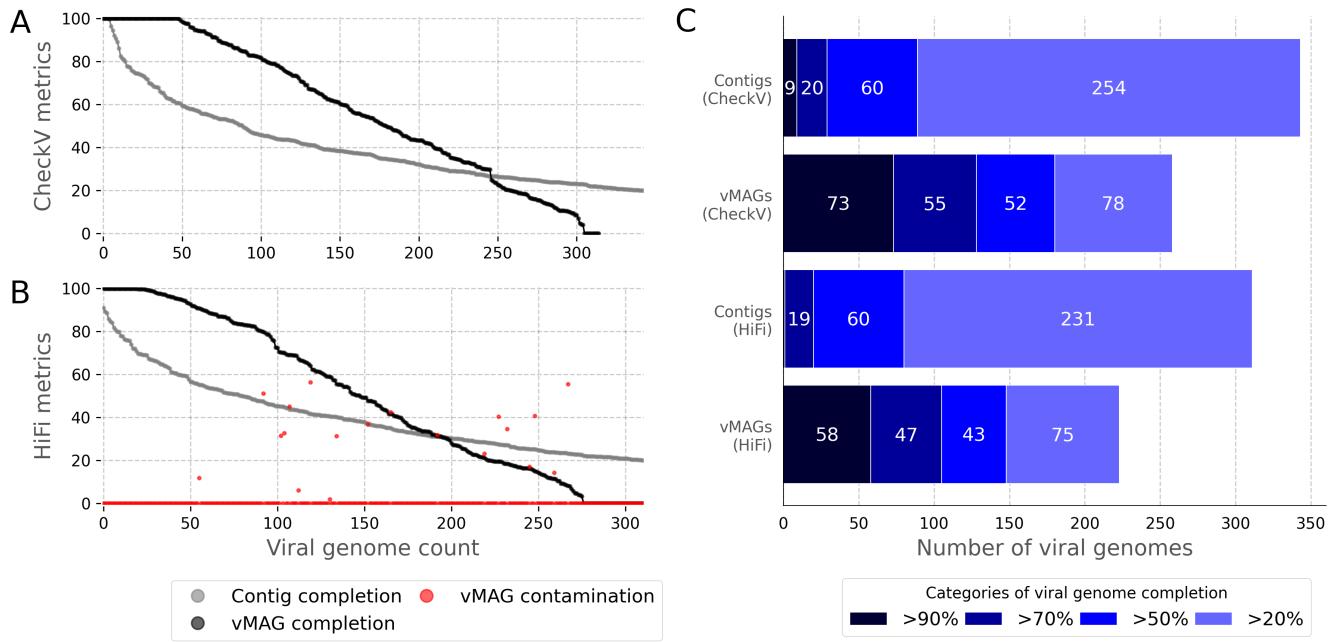


Fig. 2. Viral MAG validation. The completion and contamination of unbinned viral contigs and binned vMAGs from a sheep gut metagenome estimated with CheckV (A) and with viral references extracted from long-read HiFi assembly of the same sample (B), and a bar plot showing the number of high-completion viral genomes in the contigs and vMAGs (C).

Quality	Completion	Contamination	Count	Percent
Near-complete	>90%	<10%	57	20.65%
Moderate completion	>50%	<10%	141	51.09%
Low completion	>0%	<10%	259	93.84%
Over-contaminated	>0%	>10%	17	6.16%
Testable	>0%	NA	276	100.00%
Total	NA	NA	315	NA

Table 2. The number of vMAGs from a sheep gut metagenome falling into broad quality categories after evaluation with reference viral sequences from a long-read HiFi assembly.

independently evaluates each possible virus-host pair with at least 2 proximity-ligated read pairs linking them to estimate the average copy count of the virus genome per prokaryotic genome (see Methods; Formula 1). The density (links per Kb²) of the virus-host links is also compared normalized to the average expected virus per cell copy count and is compared to the average intra-genome Hi-C connectivity of the prokaryotic host (Formula 2) to evaluate the likelihood that the given MAG is the correct host for the virus. An important feature of these normalization methods is that they can be reliably used to threshold linkage data generated from a variety of Hi-C sequencing depths. The Hi-C sequences were rarefied *in-silico* to produce libraries ranging from 100 thousand to 100 million reads, and the resulting subsets were used to compute linkage metrics from the same virus-host pairs. Standardizing the linkage strengths by calculating the estimated viral copy count (Fig. 3A) and normalized connectivity ratio (Fig. 3B) revealed that these estimates do not significantly change with reduced Hi-C library size. In addition to this, we observed a concentration of both these values around the value of 1 – the theoretically expected value for both metrics (Fig. 3A-B; see next section). Taken together,

this suggests that these metrics can be used to reliably assess virus-host linkages regardless of the library size.

Unsupervised virus-host linkage thresholding.

Evaluating the linkage strength based on the advanced normalization metrics allows for a more robust separation of true positives from false positives because the expected values are known. For the copy count metric, a value of 1 suggests that on average every cell in the population has 1 copy of the virus and thus can be likely a true virus-host linkage. Likewise, a normalized linkage ratio of 1 means that the virus was connected to the host genome with the same signal strength as if it were part of the host genome. Note that both these metrics only represent the average state sampled microbiome, and in reality, the number of viral copies per cell may vary across the population.

The threshold boundary chosen for the minimum copy count metric has a major impact on the number of interactions captured and the false-positive rate of the classifier. To overcome this, ProxiPhage automatically assesses the likely false-positive rates and false-negative rates at each threshold and

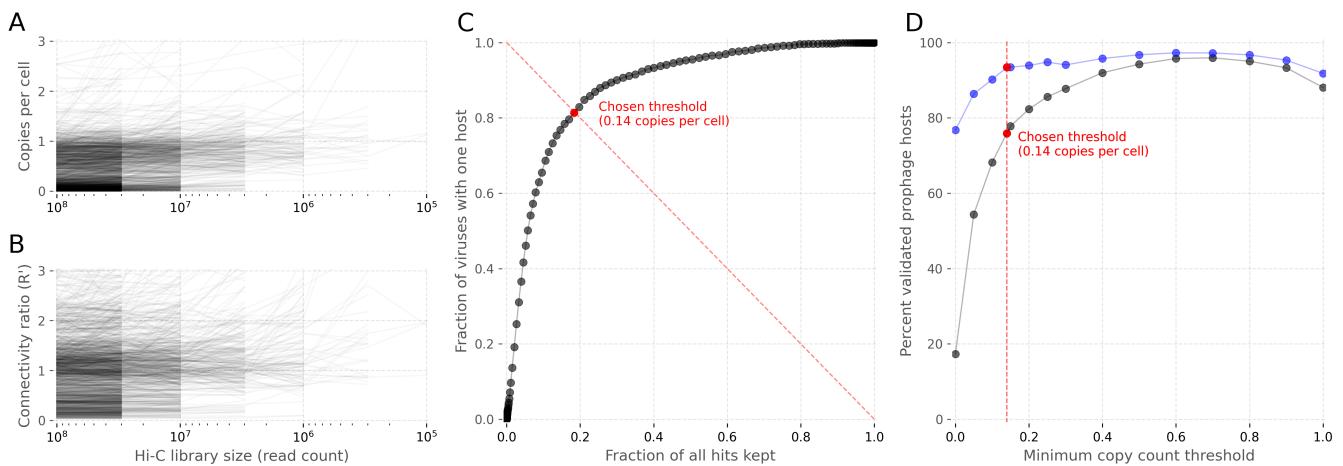


Fig. 3. Host finding thresholding. Effect of Hi-C read rarefaction (x-axis) on A) viral copy count per cell estimates and B) the normalized linkage ratio R' . Each line represents one virus-host association tracked across multiple rarefied Hi-C libraries. C) Receiver operating characteristic curve showing the decline in the number of bacteriophage host associations (x-axis) and in the number of phages with at least one host (y-axis) as the threshold of the minimum average copy count of each phage genome in its host is raised. D) The percent of prophage host links that were validated with HiFi long-read sequencing at different viral copy count thresholds. The red lines show the chosen threshold of 0.14 viral copies per cell.

selects the optimal value based on the results (see Methods). In short, ProxiPhage constructs a receiver operating characteristic (ROC) curve and chooses a threshold that minimizes the fraction of kept the virus-host links while maximizing the number of viruses that still have one host (Fig. 3A). During testing, we observed that the area under the ROC curve (AUC) of this analysis and the chosen threshold can vary significantly depending on the quality of the proximity ligated library and the complexity of the sampled community. For the sheep fecal microbiome, the area under the curve (AUC) was relatively low – 0.88, and ProxiPhage selected a minimum copy count threshold value of 0.14 viral copies per cell.

To evaluate the accuracy of the automated thresholding and to validate this host-finding method, the long-read assembly was used to estimate the false-positive rates in the resulting associations. HiFi long-reads do not carry any inter-molecule information, so this validation was limited to prophages – viral sequences integrated into the host genome. The host sequence flanking prophages on the long-read contigs was compared to the sequence of the host MAG(s) that the same prophage was linked to in the short-read assembly to determine if the host association was correct (see Methods). As expected, the true-positive virus-host assignment rate improved as the minimum copy count threshold was increased, peaking at 96% support (Fig. 3B, black line). The automatically detected optimal threshold was placed at the point in the curve where further increasing the threshold started having diminishing returns and resulted in 731 prophage-host links with a true-positive rate of 74%. However, some of the false positives from this method could also be explained by the prophages being present in several hosts (but present only once in the long-read assembly). When the validation was rerun with prophages that were only assigned a single host (Fig. 3B, blue line), the automatically chosen threshold left 516 prophage-host links with a true-positive rate of 93%.

ProxiPhage host attribution is sensitive and specific.

The annotated 2341 viral contigs were cross-referenced with 365 prokaryotic MAGs extracted from the same assembly using the ProxiPhage host attribution algorithm, resulting in a total of 30,811 possible virus-host pairs with at least 1 Hi-C link (Fig. S4A). Applying multiple automated filtered steps to the data (see Methods) filtered out the vast majority of these links, leaving just 2497 links that were predicted to be true positives (Fig. S4B, Fig. 6A). In total, the algorithm was able to identify a prokaryotic host to 1559 (67%) of the viral contigs. Of these, 516 were found to be prophages with a single host (see Methods), with a true host attribution rate of 93%, as validated with the long-read reference assembly. While the majority of the viral contigs were assigned just one host, there were several viral contigs that were linked to multiple prokaryotic MAGs, suggesting possible promiscuous viruses (Fig. S4, vertical lines). Likewise, some prokaryotic MAGs were linked with many viral contigs, which could be fragments of the same viruses or indicate co-infection (Fig. S4, horizontal lines). Reassuringly, most contigs that were clustered together into a vMAG were assigned identical or similar prokaryotic hosts, except where coverage dropout caused likely false negatives (Fig. 5). It should be noted, however, that some of the weaker host associations still appear to be clear outliers in their respective vMAG clusters, indicating the presence of residual false-positive associations even after thresholding.

The host attribution algorithm in ProxiPhage works more reliably on complete vMAGs since the algorithm has more sequence length to use for estimating connectivity likelihood. When using the vMAGs for host assignment, applying the automatic filtering on the unfiltered connectivity matrix (Fig. S5A) results in a cleaner matrix, with relatively few predicted promiscuous phages or host co-infections (Fig. S5B). Interestingly, the algorithm predicted the optimal copy count value to be 0.14 – the same as in the original viral contigs.

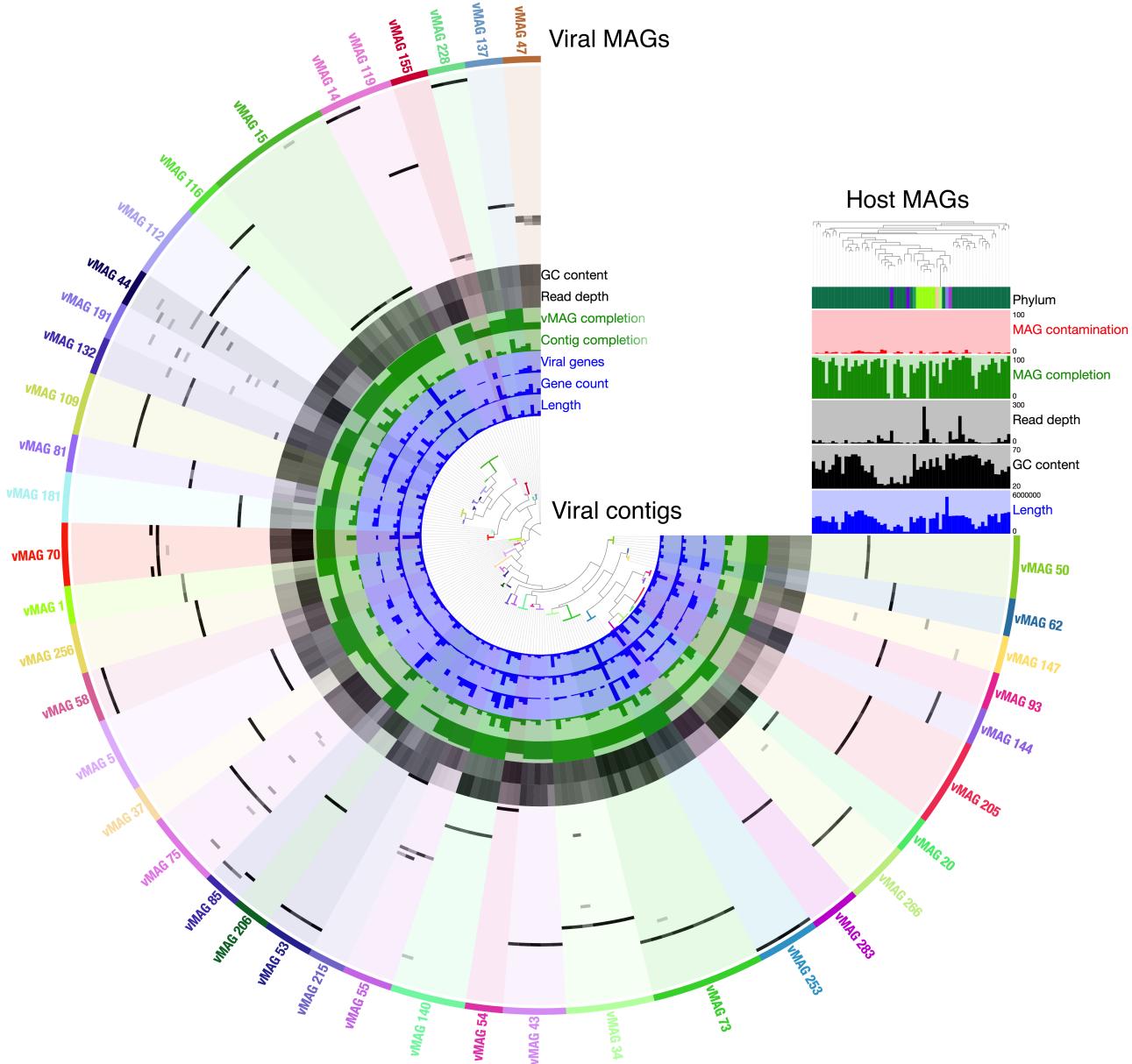


Fig. 4. Viral MAGs and their hosts. Viral metagenome assembled genomes (vMAGs, outer ring) derived from a sheep gut metagenome, with their associated contigs in the same color. Darker bars indicate higher estimated copy count. Additional circular layers (viral) and bar plots (prokaryotic host MAGs, upper-right) represent characteristics of viral contigs or host MAGs respectively, including gene content, length, GC%, and estimates of completion based on alignment to a long-read metagenomic assembly. Radial grey bars indicate a viral-host association, with the intensity encoding estimated viral copy count per cell. Only vMAGs with at least 3 contigs and a host found for every contig are shown to fit this visualization.

This reduced the possible 11,015 vMAG-host links with at least 1 Hi-C link to just 373 links after thresholding, which still associated a total of 219 vMAGs (70%) with a prokaryotic MAG host.

ProxiPhage can resolve viral genomes and their hosts from a variety of sample types.

The sheep gut microbiome used for benchmarking the accuracy of viral binning and host finding is the only currently available sample that has been sequenced to such a high depth with long-read sequencing (38). However, to validate

the broad applicability of ProxiPhage to a variety of sample types, including medically and environmentally relevant microbiomes, we processed three more samples – a fecal sample from a healthy human donor, a cow rumen sample, and a sample from a wastewater treatment plant (see Methods). Using the same computational analysis pipeline as the sheep fecal sample, we assembled and annotated a total of 1609, 1298, and 1013 viral contigs from these samples, respectively (Table 1; Fig. S1). These viral contigs were then binned into 205, 148, and 104 vMAGs in each sample, and their completion was compared to that of the original viral contigs with

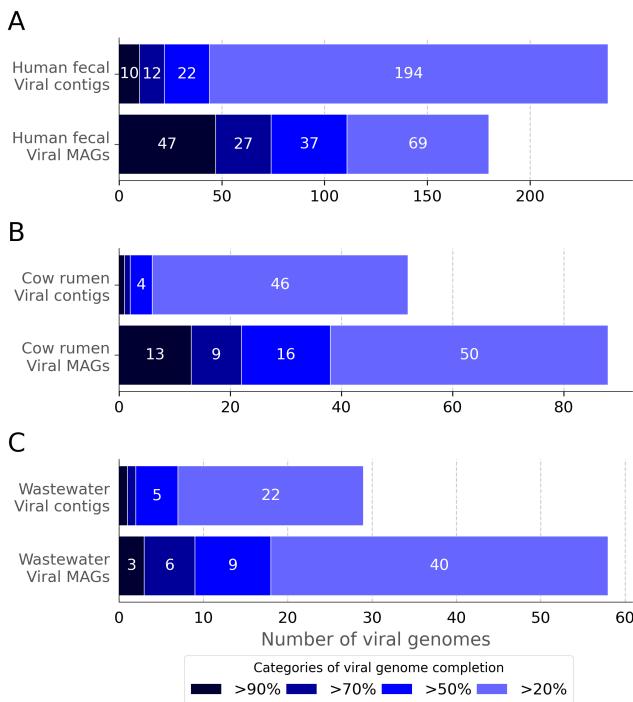


Fig. 5. Viral MAG extraction in additional samples. Viral genome completion (estimated with CheckV) of the original and binned viral contigs from additional benchmarking metagenomic samples extracted from A) human stool, B) cow rumen, and C) wastewater. Bar plots show the number of genomes at different completion cut-offs in the original viral contigs and vMAGs, as estimated with CheckV.

CheckV (Fig. 5). We found that in every sample, ProxiPhage significantly increased the completeness of the resulting viral genomes. The impact of viral binning was particularly notable in the human fecal and cow rumen samples, where many of the vMAGs had 10 - 27 contigs.

The metagenomic assemblies from the human fecal, cow rumen, and wastewater samples were also binned into prokaryotic MAGs to test the host attribution pipeline. The viruses from these three samples were referenced against the MAGs of their respective communities to identify their likely cellular host genomes. The application of automated minimum copy count filtering on the three samples resulted in significantly different thresholds being chosen – 0.03, 0.03, and 0.12 copies per cell, respectively. For the human fecal and cow rumen sample, the AUC of the ROC analysis was 0.97 and 0.98 viral copies per cell, respectively, suggesting a very high signal-to-noise ratio in the Hi-C linkage data, while the AUC for the wastewater sample was lower, at 0.86. After final thresholding, ProxiPhage kept 1602, 1158, and 1024 virus-host links, and a total of 1262 (78%), 1002 (77%), and 705 (69%) viruses were assigned at least one host in the human fecal, cow rumen, and wastewater samples, respectively (Fig. 6; Table 1). Similar to the results from the sheep sample, most of the viral contigs from the same vMAGs were assigned to the same host(s), confirming the accuracy of both groupings (Fig. 6, top color bar).

ProxiPhage is a commercial end-to-end service.

Viral sequence annotation, binning of vMAGs, and prokary-

otic MAG host association are all streamlined and automated in ProxiPhage as part of the umbrella ProxiMeta platform service, available at <https://proximeta.phasegenomics.com>. The platform is based on a modular cloud service based on Amazon's AWS service. It is a complete end-to-end service, at a minimum requiring the metagenomic assembly as Hi-C reads from the user, and returns viral annotation, vMAG clustering, and host attribution results. Optional analysis of plasmid sequences and other mobile elements is also available using the same workflow.

Discussion

ProxiPhage is the first automated software capable of accurate extraction of near-complete viral genomes from highly fragmented metagenomic assemblies by using proximity-ligation sequencing such as Hi-C. This approach is significantly cheaper and more scalable to large studies compared to long-read sequencing of comparable depth, while also enabling the association of the viruses with their hosts (20). Viral binning has been previously indirectly achieved with custom analysis of select samples by placing multiple viral contigs together with their host genome sequences (27–29), however, this approach can fail in events of promiscuous phage infections (30), which have been observed in all four microbiomes investigated in this study. By separating the viral contigs with both proximity-ligation signal as well as conventional binning methods, ProxiPhage allows for viral genome-resolved de-convolution even in events of phage co-infection, promiscuous phages, and relatively low Hi-C coverage. The underlying assumption behind the method is that the clusters from the former method represent contigs that were present in the same cell or envelope, while clusters from the latter approach represent groups of contigs with similar phylogeny and abundance. Incorporating both these signals when extracting the vMAGs allows ProxiPhage to separate viruses that may have been co-infecting cells. Since the majority of virus discovery efforts focus on extracting them directly from metagenomic assemblies (40) and often result in recovering short contigs from fragmented genomes (17, 41), ProxiPhage has the potential to greatly accelerate the discovery of near-complete viral genomes.

In addition to improving viral genome completion during virus discovery, proximity ligation signal has the added benefit of capturing interactions of the viruses with their respective hosts. In both environmental and medical applications, capturing this information is crucial for understanding the impact and contribution of viruses on the functioning of their respective communities (42). The high throughput of proximity ligation sequencing makes it the most scalable approach for use in sample-wide virus association and characterization studies (34, 37). Several studies have been able to predict the prokaryotic MAG hosts of viruses using this data type, however, these studies relied on careful investigation and custom thresholding to produce their results. ProxiPhage is not only the first unsupervised software for such analysis,

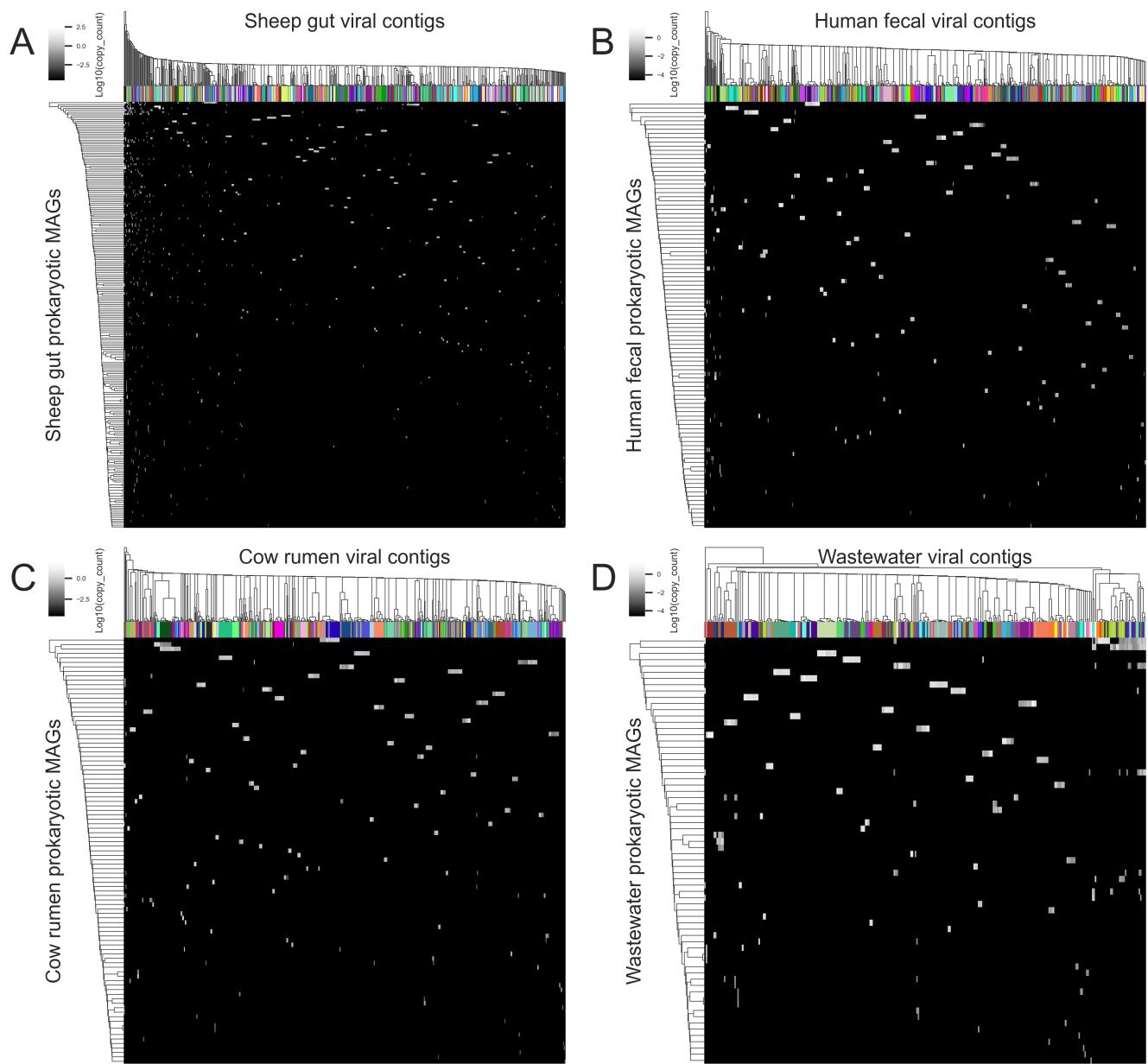


Fig. 6. Viral host assignment in additional samples. Prokaryotic hosts identified for viral contigs with ProxiPhage from additional benchmarking microbiome samples extracted from A) sheep gut, B) human stool, C) cow rumen, and D) wastewater. The color map represents the log of the estimated average copy count of each phage genome in its host. Rows and columns are clustered according to linkage similarity with seaborn clustermap. The colors above the phage contigs label the viral MAG that they belong to. Only viral contigs from viral MAGs are shown.

but it also overcomes many challenges of proximity linkage data standardization and thresholding, making it appropriate for use on a variety of sequencing depths, library qualities, and community compositions. As demonstrated in the four metagenomes processed in this study, this allows ProxiPhage to yield robust virus-host associations for the majority of viruses, making it appropriate for large-scale viral infection screening studies. In addition to this, the provided viral copy count per cell estimates can provide useful information about the nature of the virus-host interaction, such as predicting lytic or lysogenic infections.

Neither the binning nor the host attribution features in ProxiPhage have any sequence-specific methods, suggesting that

it may also be applied for the study of other mobile genomic elements, such as plasmid sequences. However, this functionality was not shown or benchmarked in this study, primarily because the sheep gut metagenome used for benchmarking was found to have very few plasmid sequences. The challenges of plasmid genome binning and host association are, in principle, very similar to those of prophage analysis (36), and plasmid binning and host attribution are very difficult to achieve without proximity-ligation data (43, 44). Similar to viruses, plasmid sequence reconstruction (45) and host association (36, 46) have been previously achieved using proximity-ligation sequencing, however, ProxiPhage is the first unsupervised software capable of this function. The possibility of high throughput plasmidome de-convolution could

have a great impact on microbial resistome characterization in a variety of research and medical fields (44).

The automated vMAG extraction host attribution features of ProxiPhage make it the first software capable of such analysis, and its accuracy and high throughput have potential to aid the study of viruses and their cellular hosts in many basic research and medical fields. In environmental and host-associated metagenomic studies, this platform can accelerate the discovery and characterization of novel viral clades, and the automated host predictions could improve our understanding of the role of phages in the composition dynamics and nutrient cycling of their respective communities (47). In clinical settings, our platform can be applied for screening the potential safety and efficacy of fecal matter transplants (FMR) and phage therapies, and for predicting the effects of such treatments on specific patients (48, 49). Finally, while ProxiPhage is a commercial service, the methods outlined in this work could serve as a stepping stone for future academic software for similar analysis of viral data.

Methods

Read and assembly pre-processing.

The DNA extracted from sheep gut was processed and sequenced on as described in Bickhart et. al (38). To better represent common sequencing depth and to have the results comparable with other samples in this study, 100 million reads were randomly selected from the sequence pool and assembled with MegaHit (50) v1.2.9 using default parameters. The resulting assembly had 274,316 contigs at least 1 Kb in length and contained a total sequence length of 873.4 Mb with an N50 of 4138 bp (Table 1). A proximity ligation library was also prepared from the same sample using the ProxiMeta bacterial Hi-C kit from Phase Genomics, and 100 million reads were also randomly selected from the resulting 2x150 bp reads. These Hi-C reads were then aligned to the metagenomic assembly with Bwa (51) v0.7.17 and compressed with Samtools (52) v1.10. The alignment was then scanned to count the total number of long-range Hi-C interactions between contigs (different contigs or on the same contig but at least 10 Kb apart). Similarly, the number of close-range interactions was also counted and saved (at most 10 Kb apart on the same contig). Only alignments that were non-redundant, full-length, and with a maximum of 1 mismatch were considered in these tallies. A total of 14,597,959 long-range and 44,233,451 close range Hi-C interactions were recorded. Finally, a size-selected SMRTbell (53) library (9-14 Kb final fragment length) was prepared from the same sample for ultra-deep sequencing on the Sequel, yielding a total of 255 Gb of long-read HiFi data, as described in Bickhart et. al (38). These long reads were then assembled with metaFlye (54) to produce 60,050 contigs with a total of 3.43 Gb of sequence and an N50 of 279,621 bp. This HiFi assembly was used as a reference to evaluate binning and host association methods in this paper.

Viral sequence annotation.

Long contigs (>5 Kb) in both the short-read shotgun and long-read HiFi assemblies were annotated with VirSorter (39) v2.2.2 with default parameters to find likely viral sequences. For the short-read assembly, original unmodified contigs with at least 50% viral gene content were saved for subsequent downstream analysis. For the Hi-Fi long-read assembly, predicted viral genomes were excised with VirSorter2 and used as complete viral genome references. Using these methods, 2341 (N50 15,919 bp) and 8054 (N50 66,873 bp) viral sequences were annotated in the short-read and long-read assemblies, respectively.

ProxiPhage viral binning.

The ProxiPhage viral binning features a combination of proximity ligation signal clustering and conventional metagenomic binning approaches to overcome the limitations of either method. First, the viral sequences are binned with both methods to produce two preliminary sets of vMAG. By default, ProxiPhage uses Bin3C (35) v0.1.1 and MetaBAT (22) v2.15 with customized parameters and thresholds, however, other clustering software may be used. A non-redundant set overlap network is constructed from these two contig groupings such that nodes represent contig clusters from either of the two preliminary vMAG sets, and edges contain contigs that overlap between bin sets. These overlaps are then scanned and resolved through a proprietary greedy network collapse algorithm featured throughout the ProxiMeta (36) platform to produce a single set of vMAGs that is more accurate and complete than either of the original inputs. Each vMAG cluster is then additionally scanned and adjusted to ensure a minimum strength of Hi-C linkages between its contained contigs, which further reduces possible contamination. Finally, ProxiPhage viral binning also allows for additional pruning of vMAG clusters such that contigs from each cluster have similar or identical predicted prokaryotic hosts, although this feature was not utilized in this study since host commonality was one of the metrics used for vMAG accuracy assessment.

Viral MAG validation and benchmarking.

The viral genome completion of both viral contigs and vMAGs was estimated with CheckV (31) v0.7.0 with default parameters. Note that CheckV does not natively support the investigation of contig clusters, so to run CheckV on the vMAGs the sequences from each vMAG needed to be concatenated into a single sequence with 200 bp “N” spacers. Also note that the “contamination” metric produced by CheckV refers to the bacterial content of the sequences, and not binning false positives as it does CheckM (21). The quality of the viral contigs and vMAGs was also assessed by comparing the sequences to the reference phage sequences excised from the long-read HiFi assembly (38). The short-read and long-read phage sequences were aligned to each other using Blast (55) v2.11.0, and high-quality alignments (>95% percent identity, >100bp length) were saved. The reference alignment network was constructed from these alignments

using Cytoscape (56) v3.7.1. The best reference for each vMAG was determined as the reference to which the greatest percentage its sequence aligned, with a minimum of 1 Kb. The completion of the vMAG was estimated as the percentage of the reference that aligned to the vMAG (Formula 1), and its contamination was estimated as the percentage of the vMAG sequence that did not align to the reference (Formula 2). Contigs or contig segments that did not reliably align to any of the references were not counted in the contamination calculation to account for some of the short-read assembly sequences not being present in the long-read assembly.

Genome completion of a query viral sequence $\omega(q)$ calculated using a long-read reference assembly from the total length of the reference that aligned to the query $A(r)$ and the length of best reference genome $L(r)$:

$$\omega(q) = \frac{A(r)}{L(r)} \quad (1)$$

Genome contamination $\chi(q)$ of a query viral sequence calculated using a long-read reference assembly from the total length of the query $L(q)$, the length of the query that aligned to the best reference $A(q)$ and the length of the query that was not found to align to any reference sequence $L(u)$:

$$\chi(q) = \frac{L(q) - \sum A(q) - \sum L(u)}{L(q)} \quad (2)$$

Prokaryotic MAG extraction.

The metagenomic assembly was de-convoluted with the ProxiMeta (36) platform to extract draft prokaryotic genomes to be used for finding the likely hosts of the viral sequences. The completion and contamination of the MAGs were estimated with CheckM (21) v1.1.3. In total, 365 MAGs were formed, representing 34% of the total assembly sequence. Of these, 151 MAGs were of moderate quality (>50% completion and <10% contamination), and none of the 365 MAGs were over-contaminated (contamination >10%). All the resulting clusters were used for viral host attribution with ProxiPhage.

ProxiPhage host attribution.

The long-range Hi-C linkage data was scanned to identify viral contigs and prokaryotic MAGs with a Hi-C link between them. A combination of the Hi-C link count, viral read depth, and MAG read depth were then used to estimate the average copy count of each virus in each MAG (Formula 3). The density of Hi-C links per Kb² of sequence between the virus and the MAG was then compared to the connectivity of the MAG to itself and normalized to the estimated copy count to compute the normalized connectivity ratio (Formula 4). This value assesses the strength of the virus-host linkage in the context of what would be expected if the virus was found inside the cell, with a value of 1 being ideal. Virus-host linkages were then filtered to keep only connections with at least 2 Hi-C read links between the virus and host MAG, a connectivity ratio of 0.1, and intra-MAG connectivity of 10 links

to remove false positives. For the final threshold value, a receiver operating characteristic (ROC) curve is used to determine the optimal copy count cut-off value. The optimal cut-off was determined from the ROC curve as the value that produces the point to the top left of the plot, or the cut-off that removed the maximum number of virus-host links while still finding at least one host for the maximum number of viruses. Each virus is also evaluated for the fraction of host MAGs that it still had connections with to identify “sticky” sequences with a likely high proportion of false positives. These were corrected by removing linkages with an average copy count less than 80% of the highest copy count value for the given viral sequence. The above host assignment workflow also works identically for vMAG clusters, but with linkage and length values from different contigs being added together.

Average viral copy counts per cell C calculated from the virus abundance V , prokaryotic host abundance H , Hi-C links between the virus and host L , and total Hi-C links of the virus and all possible hosts $L(v)$:

$$C = \frac{V}{H} \frac{L}{\sum L(v)} \quad (3)$$

Normalized connectivity ratio R' calculated from the Hi-C connectivity density between the virus and host D_{VH} and of the host genome to itself D_H , and normalized to the virus abundance V , prokaryotic host abundance H , Hi-C links between the virus and host L , and total Hi-C links of the virus and all possible hosts $L(v)$:

$$R' = \frac{D_{VH}}{D_H} \frac{H \sum L(v)}{VL} \quad (4)$$

Prophage host validation.

The accuracy of prophage virus-host links found with the host attribution software in ProxiPhage was evaluated with the reference long-read HiFi assembly. The short-read viral sequences were aligned to the full HiFi assembly using Blast (55) v2.11.0, and high-quality alignments (> 95% percent identity, > 100 bp length) were saved. The best reference contig for each virus was defined as the contig to which the greatest percentage its sequence aligned to, with a minimum of 50%. If the HiFi contig still had more than 200 Kb of sequence that did not align to the viruses, the virus was declared a prophage and thus used for subsequent analysis. This unaligned bacterial sequence was compared to the sequence of the host MAG from the short-read assembly to which the virus was linked. If at least 10 Kb of the host sequence aligned to the host MAG, the virus-host link was considered correct. Only good quality MAGs (>50% completion, <10% contamination according to CheckM) were included in this analysis.

Anvi'o analysis.

A custom selection was taken from the full vMAG set for detailed visualization with Anvi'o 7 (57). This selection contained vMAGs that had at least 3 contigs and had at least

one host assigned to each of its contigs. The phylogenetic tree was constructed from the vMAG sequences using VICTOR (58), and the resulting Newick tree was then manually modified to replace the vMAG leaves with the contigs contained in each of the vMAGs. For the host MAGs, their taxonomies at the phylum level were estimated manually using a combination Kraken2 (59) v2.1.1 (using default options and the full standard database) and metaWRAP (60) v1.3.2 blobology and *classifybins* modules using default options. The phylogenetic tree of the host MAGs was then constructed with Anvi'o 7 (57) using a concatenated alignment of all common ribosomal genes. The estimated average read depth for both viral contigs and host MAGs were estimated with the *jgi_summarizebam_contigdepths* script in MetaBAT (22) v2.15.

Human fecal microbiome processing.

The human fecal sample analyzed in this study was collected from a healthy individual that consented to DNA sequencing. Libraries were prepared using the Phase Genomics ProxiMeta Hi-C kit version 3 following the manufacturer's protocol. In brief, approximately 250 mg of sample was homogenized in phospho-buffered saline (PBS) and pelleted by spinning at 17,000 x g for 1 min. The pellet was resuspended in 1 ml of crosslink solution and incubated at room temperature for 20 min at room temperature with rotational mixing. Crosslinking was terminated by the addition of 100 µl of Quench solution and incubation for 15 min at room temperature with mixing. After pelleting sample at 17,000 x g for 5 minutes, pellets were washed once in chromatin rinse buffer (CRB) and then resuspended in 700 µl of Phase Genomics Lysis buffer 1 and 250 µl of Lysis beads. The sample was placed in a Turbomix disruptor (Scientific Industries) and mixed at maximum speed for 20 minutes. The lysate was spun down briefly and the lysate transferred to a new microcentrifuge tube and chromatin pelleted by spinning at 17,000 x g for 5 minutes. The pellet was then washed with CRB and resuspended in 100 µl of Phase Genomics Lysis Buffer 2 and incubated at 65°C for 15 min. Chromatin was then bound to Recovery beads, washed with CRB, and then fragmented/ends filled in with biotinylated nucleotides at 37°C for 1 h. Beads were washed and resuspended in 100 µl of Proximity Ligation buffer and 5 µl of Proximity Ligation enzyme and incubated at 25°C for 4h. Reverse crosslinks enzyme was added at the sample was heated to 65°C for 1h to release DNA from crosslinked chromatin. DNA was purified using Recovery beads and biotinylated ligation junctions capture using streptavidin beads. Bead-bound DNA was used to generate a dual unique-indexed Illumina-compatible library using Phase Genomics library prep reagents. DNA for Shotgun WGS libraries were prepared using a ZymoBiotics DNA miniprep kit (Zymo Research). Libraries were prepared using Nextera XT (Illumina) following the manufacturer's protocol and 50 ng of input DNA.

Additional sample processing.

Additional benchmarking data was generated from the short-

read WGS and Hi-C sequencing data from the cow rumen metagenome (61), the wastewater benchmark metagenome (44), and the human fecal metagenome (see human fecal microbiome processing). All downstream sequence analyses, including rarefaction to 100 million reads, metagenomic assembly, viral annotation, viral binning, and host attribution were identical to the analysis performed on the main sheep metagenome (Table 1).

Data Availability

The raw data from the sheep fecal microbiome whole-genome shotgun and Hi-C sequencing is publicly available from NCBI Bioproject PRJNA595610. The human fecal microbiome data is available at [upload PGFecal2 to SRA]. The cow rumen microbiome data is available from Bioproject PRJEB21624, sample SAMEA104567052. The wastewater microbiome data is available from Bioproject PRJNA506462. The four short-read assemblies used in this study are available at https://bitbucket.org/phasegenomics/proxiphage_paper/src/main/assemblies.

The sheep fecal microbiome long-read HiFi assembly used as a reference in this study is available at DOI: <https://doi.org/10.5281/zenodo.4729049>.

All other data, intermediate files, and analysis scripts are publicly available from https://bitbucket.org/phasegenomics/proxiphage_paper.

The ProxiPhage viral analysis platform is available through the ProxiMeta service platform at <https://proximeta.phasegenomics.com>.

Acknowledgements

We thank the early users of ProxiPhage for their patience and feedback during testing and for their suggestions for the platform's features. We also thank the engineering and management teams of Phase Genomics for their support, help, and code review during ProxiPhage development, and the laboratory staff for constructing sequencing libraries. Finally, we thank the anonymous donor of the human fecal sample used in this study for their contribution.

Author Contributions

GU conceived, developed, and benchmarked ProxiPhage, processed the results, and wrote this manuscript; MP developed the first prototype of the viral host attribution algorithm; AW and JG deployed and maintained the ProxiPhage cloud service; DB and TS provided benchmarking sequence data, assemblies, and taxonomic annotations; SS and IL directed ProxiPhage development; BA, SE, and MS provided guidance and insight for the project; IL and GU conceived the project.

Competing Interests

GU, SE, AW, JG, BA, SS, and IL are employees of Phase Genomics. MP is an employee of Inscripta. All other authors have no competing interests.

Bibliography

1. C. M. Cullen, K. K. Aneja, S. Beyhan, C. E. Cho, S. Woloszynek, M. Convertino, S. J. McCoy, Y. Zhang, M. Z. Anderson, D. Alvarez-Ponce, E. Smirnova, L. Karstens, P. C. Dorrestein, H. Li, A. Sen Gupta, K. Cheung, J. G. Powers, Z. Zhao, and G. L. Rosen. Emerging priorities for microbiome research. *Front Microbiol*, 11:136, 2020. ISSN 1664-302X (Print) 1664-302X (Linking). doi: 10.3389/fmicb.2020.00136.
2. Y. Fan and O. Pedersen. Gut microbiota in human metabolic health and disease. *Nat Rev Microbiol*, 19(1):55–71, 2021. ISSN 1740-1534 (Electronic) 1740-1526 (Linking). doi: 10.1038/s41579-020-0433-9.
3. J. Jovel, J. Patterson, W. Wang, N. Hotte, S. O'Keefe, T. Mitchel, T. Perry, D. Kao, A. L. Mason, K. L. Madsen, and G. K. Wong. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front Microbiol*, 7:459, 2016. ISSN 1664-302X (Print) 1664-302X (Linking). doi: 10.3389/fmicb.2016.00459.
4. W. L. Wang, S. Y. Xu, Z. G. Ren, L. Tao, J. W. Jiang, and S. S. Zheng. Application of metagenomics in the human gut microbiome. *World J Gastroenterol*, 21(3):803–14, 2015. ISSN 2219-2840 (Electronic) 1007-9327 (Linking). doi: 10.3748/wjg.v21.i3.803.
5. M. Clapp, N. Aurora, L. Herrera, M. Bhakti, E. Wilen, and S. Wakefield. Gut microbiota's effect on mental health: The gut-brain axis. *Clin Pract*, 7(4):987, 2017. ISSN 2039-7275 (Print) 2039-7275 (Linking). doi: 10.4081/cp.2017.00459.
6. J. Durack and S. V. Lynch. The gut microbiome: Relationships with disease and opportunities for therapy. *J Exp Med*, 216(1):20–40, 2019. ISSN 1540-9538 (Electronic) 0022-1007 (Linking). doi: 10.1084/jem.20180448.
7. A. Bruno, A. Sandionigi, M. Bernasconi, A. Panio, M. Labra, and M. Casiraghi. Changes in the drinking water microbiome: Effects of water treatments along the flow of two drinking water treatment plants in a urbanized area, milan (italy). *Front Microbiol*, 9:2557, 2018. ISSN 1664-302X (Print) 1664-302X (Linking). doi: 10.3389/fmicb.2018.02557.
8. B. K. Singh, P. Trivedi, E. Egidi, C. A. Macdonald, and M. Delgado-Baquerizo. Crop microbiome and sustainable agriculture. *Nat Rev Microbiol*, 18(11):601–602, 2020. ISSN 1740-1534 (Electronic) 1740-1526 (Linking). doi: 10.1038/s41579-020-00446-y.
9. L. Miao and Z. Liu. Microbiome analysis and -omics studies of microbial denitrification processes in wastewater treatment: recent advances. *Sci China Life Sci*, 61(7):753–761, 2018. ISSN 1869-1889 (Electronic) 1674-7305 (Linking). doi: 10.1007/s11427-017-9228-2.
10. S. Divya Ganeshan and Z. Hosseinioudost. Phage therapy with a focus on the human microbiota. *Antibiotics (Basel)*, 8(3), 2019. ISSN 2079-6382 (Print) 2079-6382 (Linking). doi: 10.3390/antibiotics8030131.
11. T. D. S. Sutton and C. Hill. Gut bacteriophage: Current understanding and challenges. *Front Endocrinol (Lausanne)*, 10:784, 2019. ISSN 1664-2392 (Print) 1664-2392 (Linking). doi: 10.3389/fendo.2019.00784.
12. C. M. Robinson and J. K. Pfeiffer. Viruses and the microbiota. *Annu Rev Virol*, 1:55–69, 2014. ISSN 2327-056X (Print) 2327-056X (Linking). doi: 10.1146/annurev-virology-031413-085550.
13. K. Moon, J. H. Jeon, I. Kang, K. S. Park, K. Lee, C. J. Cha, S. H. Lee, and J. C. Cho. Freshwater viral metagenome reveals novel and functional phage-borne antibiotic resistance genes. *Microbiome*, 8(1):75, 2020. ISSN 2049-2618 (Electronic) 2049-2618 (Linking). doi: 10.1186/s40168-020-00863-4.
14. J. R. Brum and M. B. Sullivan. Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat Rev Microbiol*, 13(3):147–59, 2015. ISSN 1740-1534 (Electronic) 1740-1526 (Linking). doi: 10.1038/nrmicro3404.
15. L. F. Camarillo-Guerrero, A. Almeida, G. Rangel-Pineros, R. D. Finn, and T. D. Lawley. Massive expansion of human gut bacteriophage diversity. *Cell*, 184(4):1098–1109 e9, 2021. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi: 10.1016/j.cell.2021.01.029.
16. R. Rose, B. Constantines, A. Tapinos, D. L. Robertson, and M. Prosperi. Challenges in the analysis of viral metagenomes. *Virus Evol*, 2(2):vew022, 2016. ISSN 2057-1577 (Print) 2057-1577 (Linking). doi: 10.1093/ve/vew022.
17. S. L. Smits, R. Bodewes, A. Ruiz-Gonzalez, W. Baumgartner, M. P. Koopmans, A. D. Osterhaus, and A. C. Schurich. Assembly of viral genomes from metagenomes. *Front Microbiol*, 5:714, 2014. ISSN 1664-302X (Print) 1664-302X (Linking). doi: 10.3389/fmicb.2014.00714.
18. V. Somerville, S. Lutz, M. Schmid, D. Frei, A. Moser, S. Irmeler, J. E. Frey, and C. H. Ahrens. Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. *BMC Microbiol*, 19(1):143, 2019. ISSN 1471-2180 (Electronic) 1471-2180 (Linking). doi: 10.1186/s12866-019-1500-0.
19. J. Klumpp, D. E. Fouts, and S. Sozhamannan. Next generation sequencing technologies and the changing landscape of phage genomics. *Bacteriophage*, 2(3):190–199, 2012. ISSN 2159-7073 (Print) 2159-7073 (Linking). doi: 10.4161/bact.22111.
20. W. De Coster, M. H. Weissensteiner, and F. J. Sedlazeck. Towards population-scale long-read sequencing. *Nat Rev Genet*, 2021. ISSN 1471-0064 (Electronic) 1471-0056 (Linking). doi: 10.1038/s41576-021-00367-3.
21. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson. Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*, 25(7):1043–55, 2015. ISSN 1549-5469 (Electronic) 1088-9051 (Linking). doi: 10.1101/gr.186072.114.
22. D. D. Kang, F. Li, E. Kirton, A. Thomas, R. Egan, H. An, and Z. Wang. Metabat 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7:e7359, 2019. ISSN 2167-8359 (Print) 2167-8359 (Linking). doi: 10.7717/peerj.7359.
23. N. Sangwan, F. Xia, and J. A. Gilbert. Recovering complete and draft population genomes from metagenome datasets. *Microbiome*, 4:8, 2016. ISSN 2049-2618 (Electronic) 2049-2618 (Linking). doi: 10.1186/s40168-016-0154-5.
24. Y. Yue, H. Huang, Z. Qi, H. M. Dou, X. Y. Liu, T. F. Han, Y. Chen, X. J. Song, Y. H. Zhang, and J. Tu. Evaluating metagenomics tools for genome binning with real metagenomic datasets and cami datasets. *BMC Bioinformatics*, 21(1):334, 2020. ISSN 1471-2105 (Electronic) 1471-2105 (Linking). doi: 10.1186/s12859-020-03667-3.
25. S. Roux, E. M. Adriaenssens, B. E. Dutill, E. V. Koonin, A. M. Kropinski, M. Krupovic, J. H. Kuhn, R. Lavigne, J. R. Brister, A. Varsani, C. Amid, R. K. Aziz, S. R. Bordenstein, P. Bork, M. Breitbart, G. R. Cochrane, R. A. Daly, C. Desnues, M. B. Duhaime, J. B. Emerson, F. Enault, J. A. Fuhrman, P. Hingamp, P. Hugenholtz, B. L. Hurwitz, N. N. Ivanova, J. M. Labonte, K. B. Lee, R. R. Malmstrom, M. Martinez-Garcia, I. K. Mizrachi, H. Ogata, D. Paez-Espino, M. A. Petit, C. Putonti, T. Rattee, A. Reyes, F. Rodriguez-Valera, K. Rosario, L. Schriml, F. Schulz, G. F. Stewart, M. B. Sullivan, S. Sunagawa, C. A. Suttle, B. Temperton, S. G. Tringe, R. V. Thurber, N. S. Webster, K. L. Whiteson, S. W. Wilhelm, K. E. Wommack, T. Woyke, K. C. Wrighton, P. Yilmaz, T. Yoshida, M. J. Young, N. Yutin, L. Z. Allen, N. C. Kyrides, and E. A. Eloe-Fadrosh. Minimum information about an uncultivated virus genome (miuvig). *Nat Biotechnol*, 37(1):29–37, 2019. ISSN 1546-1696 (Electronic) 1087-0156 (Linking). doi: 10.1038/nbt.4306.
26. F. Schulz, J. Andreani, R. Francis, H. Boudjemaa, J. Y. Bou Khalil, J. Lee, B. La Scola, and T. Woyke. Advantages and limits of metagenomic assembly and binning of a giant virus. *mSystems*, 5(3), 2020. ISSN 2379-5077 (Print) 2379-5077 (Linking). doi: 10.1128/mSystems.00048-20.
27. L. Baudry, T. Foutel-Rodier, A. Thierry, R. Koszul, and M. Marbouty. Metator: A computational pipeline to recover high-quality metagenomic bins from mammalian gut proximity-ligation (meta3c) libraries. *Front Genet*, 10:753, 2019. ISSN 1664-8021 (Print) 1664-8021 (Linking). doi: 10.3389/fgene.2019.00753.
28. M. Marbouty, L. Baudry, A. Cournac, and R. Koszul. Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay. *Sci Adv*, 3(2):e1602105, 2017. ISSN 2375-2548 (Electronic) 2375-2548 (Linking). doi: 10.1126/sciadv.1602105.
29. M. Marbouty, A. Thierry, G. A. Millot, and R. Koszul. Metahic phage-bacteria infection network reveals active cycling phages of the healthy human gut. *Elife*, 10, 2021. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/elife.60608.
30. Daniel Cazares, Adrian Cazares, Wendy Figueroa, Gabriel Guarneros, Robert A. Edwards, and Pablo Vinuesa. Dynamics of infection in a novel group of promiscuous phages and hosts of multiple bacterial genera retrieved from river communities. *bioRxiv*, page 2020.08.07.242396, 2020. doi: 10.1101/2020.08.07.242396.
31. S. Nayfach, A. P. Camargo, F. Schulz, E. Eloe-Fadrosh, S. Roux, and N. C. Kyrides. Checkv assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol*, 39(5):578–585, 2021. ISSN 1546-1696 (Electronic) 1087-0156 (Linking). doi: 10.1038/s41587-020-00774-7.
32. K. Yahara, M. Suzuki, A. Hirabayashi, W. Suda, M. Hattori, Y. Suzuki, and Y. Okazaki. Long-read metagenomics using promethion uncovers oral bacteriophages and their interaction with host bacteria. *Nat Commun*, 12(1):27, 2021. ISSN 2041-1723 (Electronic) 2041-1723 (Linking). doi: 10.1038/s41467-020-20199-9.
33. E. G. Sakowski, K. Arora-Williams, F. Tian, A. A. Zayed, O. Zablocki, M. B. Sullivan, and S. P. Preheim. Interaction dynamics and virus-host range for estuarine actinophages captured by epipcr. *Nat Microbiol*, 6(5):630–642, 2021. ISSN 2058-5276 (Electronic) 2058-5276 (Linking). doi: 10.1038/s41564-021-00873-4.
34. R. A. Edwards, K. McNair, K. Faust, J. Raes, and B. E. Dutill. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol Rev*, 40(2):258–72, 2016. ISSN 1574-6976 (Electronic) 0168-6445 (Linking). doi: 10.1093/femsre/fuv048.
35. M. Z. DeMaere and A. E. Darling. bin3c: exploiting hi-c sequencing data to accurately resolve metagenome-assembled genomes. *Genome Biol*, 20(1):46, 2019. ISSN 1474-760X (Electronic) 1474-7596 (Linking). doi: 10.1186/s13059-019-1643-1.
36. Maximilian O. Press, Andrew H. Wiser, Zev N. Kronberg, Kyle W. Langford, Mingun Shakya, Chien-Chi Lo, Kathryn A. Mueller, Shawn T. Sullivan, Patrick S. G. Chain, and Ivan Liachko. Hi-c deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. *bioRxiv*, page 198713, 2017. doi: 10.1101/198713.
37. D. M. Bickhart, M. Watson, S. Koren, K. Panke-Buisse, L. M. Cersosimo, M. O. Press, C. P. Van Tassell, J. A. S. Van Kessel, B. J. Haley, S. W. Kim, C. Heiner, G. Suen, K. Bakshy, I. Liachko, S. T. Sullivan, P. R. Myer, J. Ghurye, M. Pop, P. J. Weimer, A. M. Phillippe, and T. P. L. Smith. Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation. *Genome Biol*, 20(1):153, 2019. ISSN 1474-760X (Electronic) 1474-7596 (Linking). doi: 10.1186/s13059-019-1760-x.
38. Derek M. Bickhart, Mikhail Kolmogorov, Elizabeth Tseng, Daniel M. Portik, Anton Korobeynikov, Ivan Tolstogovan, Gherman Uritskiy, Ivan Liachko, Shawn T. Sullivan, Sung Bong Shin, Alvah Zorea, Victoria Pascal Andreu, Kevin Panke-Buisse, Marnix H. Medema, Itzik Mizrachi, Pavel A. Pevzner, and Timothy P. L. Smith. Generation of lineage-resolved complete metagenome-assembled genomes by precision phasing. *bioRxiv*, page 2021.05.04.442591, 2021. doi: 10.1101/2021.05.04.442591.
39. J. Guo, B. Bolduc, A. A. Zayed, A. Varsani, G. Dominguez-Huerta, T. O. Delmont, A. A. Pratama, M. C. Gazitua, D. Vik, M. B. Sullivan, and S. Roux. Virsorter2: a multi-classifier, expert-guided approach to detect diverse dna and rna viruses. *Microbiome*, 9(1):37, 2021. ISSN 2049-2618 (Electronic) 2049-2618 (Linking). doi: 10.1186/s40168-020-00990-y.
40. J. L. Mokili, F. Rohwer, and B. E. Dutill. Metagenomics and future perspectives in virus discovery. *Curr Opin Virol*, 2(1):63–77, 2012. ISSN 1879-6265 (Electronic) 1879-6257 (Linking). doi: 10.1016/j.coviro.2011.12.004.
41. D. Antipov, M. Raiko, A. Lapidus, and P. A. Pevzner. Metaviral spades: assembly of viruses from metagenomic data. *Bioinformatics*, 36(14):4126–4129, 2020. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: 10.1093/bioinformatics/btaa490.

42. M. Martinez-Garcia, F. Santos, M. Moreno-Paz, V. Parro, and J. Anton. Unveiling viral-host interactions within the 'microbial dark matter'. *Nat Commun*, 5:4542, 2014. ISSN 2041-1723 (Electronic) 2041-1723 (Linking). doi: 10.1038/ncomms5542.
43. F. Maguire, B. Jia, K. L. Gray, W. Y. V. Lau, R. G. Beiko, and F. S. L. Brinkman. Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic islands. *Microb Genom*, 6(10), 2020. ISSN 2057-5858 (Electronic) 2057-5858 (Linking). doi: 10.1099/mgen.0.000436.
44. T. Stalder, M. O. Press, S. Sullivan, I. Liachko, and E. M. Top. Linking the resistome and plasmidome to the microbiome. *ISME J*, 13(10):2437–2446, 2019. ISSN 1751-7370 (Electronic) 1751-7362 (Linking). doi: 10.1038/s41396-019-0446-4.
45. C. W. Beitel, L. Froenicke, J. M. Lang, I. F. Korf, R. W. Michelmore, J. A. Eisen, and A. E. Darling. Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ*, 2:e415, 2014. ISSN 2167-8359 (Print) 2167-8359 (Linking). doi: 10.7717/peerj.415.
46. M. Marbouty, A. Cournac, J. F. Flot, H. Marie-Nelly, J. Mozziconacci, and R. Koszul. Metagenomic chromosome conformation capture (meta3c) unveils the diversity of chromosome organization in microorganisms. *eLife*, 3:e03318, 2014. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/eLife.03318.
47. S. Benler, N. Yutin, D. Antipov, M. Rayko, S. Shmakov, A. B. Gussow, P. Pevzner, and E. V. Koonin. Thousands of previously unknown phages discovered in whole-community human gut metagenomes. *Microbiome*, 9(1):78, 2021. ISSN 2049-2618 (Electronic) 2049-2618 (Linking). doi: 10.1186/s40168-021-01017-w.
48. B. O. Anonye. Commentary: Bacteriophage transfer during faecal microbiota transplantation in clostridium difficile infection is associated with treatment outcome. *Front Cell Infect Microbiol*, 8:104, 2018. ISSN 2235-2988 (Electronic) 2235-2988 (Linking). doi: 10.3389/fcimb.2018.00104.
49. D. M. Lin, B. Koskella, N. L. Ritz, D. Lin, A. Carroll-Portillo, and H. C. Lin. Transplanting fecal virus-like particles reduces high-fat diet-induced small intestinal bacterial overgrowth in mice. *Front Cell Infect Microbiol*, 9:348, 2019. ISSN 2235-2988 (Electronic) 2235-2988 (Linking). doi: 10.3389/fcimb.2019.00348.
50. D. Li, R. Luo, C. M. Liu, C. M. Leung, H. F. Ting, K. Sadakane, H. Yamashita, and T. W. Lam. Megahit v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, 102:3–11, 2016. ISSN 1095-9130 (Electronic) 1046-2023 (Linking). doi: 10.1016/j.ymeth.2016.02.020.
51. H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–60, 2009. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: 10.1093/bioinformatics/btp324.
52. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and Subgroup Genome Project Data Processing. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9, 2009. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: 10.1093/bioinformatics/btp352.
53. N. Kong, W. Ng, K. Thao, R. Agullo, A. Weis, K. S. Kim, J. Korlach, L. Hickey, L. Kelly, S. Lappin, and B. C. Weimer. Automation of pacbio smrtbell ngs library preparation for bacterial genome sequencing. *Stand Genomic Sci*, 12:27, 2017. ISSN 1944-3277 (Print) 1944-3277 (Linking). doi: 10.1186/s40793-017-0239-1.
54. M. Kolmogorov, D. M. Bickhart, B. Behsaz, A. Gurevich, M. Rayko, S. B. Shin, K. Kuhn, J. Yuan, E. Pollevikov, T. P. L. Smith, and P. A. Pevzner. metaflye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods*, 17(11):1103–1110, 2020. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). doi: 10.1038/s41592-020-00971-x.
55. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. Blast+: architecture and applications. *BMC Bioinformatics*, 10:421, 2009. ISSN 1471-2105 (Electronic) 1471-2105 (Linking). doi: 10.1186/1471-2105-10-421.
56. M. E. Smoot, K. Ono, J. Ruscheinski, P. L. Wang, and T. Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–2, 2011. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: 10.1093/bioinformatics/btq675.
57. A. M. Eren, E. Kiefl, A. Shaiber, I. Veseli, S. E. Miller, M. S. Schechter, I. Fink, J. N. Pan, M. Yousef, E. C. Fogarty, F. Trigodet, A. R. Watson, O. C. Esen, R. M. Moore, Q. Clayssen, M. D. Lee, V. Kivenson, E. D. Graham, B. D. Merrill, A. Karkman, D. Blankenberg, J. M. Eppley, A. Sjodin, J. J. Scott, X. Vazquez-Campos, L. J. McKay, E. A. McDaniel, S. L. R. Stevens, R. E. Anderson, J. Fuessel, A. Fernandez-Guerra, L. Maignien, T. O. Delmont, and A. D. Willis. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol*, 6(1):3–6, 2021. ISSN 2058-5276 (Electronic) 2058-5276 (Linking). doi: 10.1038/s41564-020-00834-3.
58. J. P. Meier-Kolthoff and M. Goker. Victor: genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics*, 33(21):3396–3404, 2017. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: 10.1093/bioinformatics/btx440.
59. D. E. Wood, J. Lu, and B. Langmead. Improved metagenomic analysis with kraken 2. *Genome Biol*, 20(1):257, 2019. ISSN 1474-760X (Electronic) 1474-7596 (Linking). doi: 10.1186/s13059-019-1891-0.
60. G. V. Uritskiy, J. DiRuggiero, and J. Taylor. Metawrap-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*, 6(1):158, 2018. ISSN 2049-2618 (Electronic) 2049-2618 (Linking). doi: 10.1186/s40168-018-0541-1.
61. R. D. Stewart, M. D. Auffret, A. Warr, A. H. Wiser, M. O. Press, K. W. Langford, I. Liachko, T. J. Snelling, R. J. Dewhurst, A. W. Walker, R. Roehe, and M. Watson. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat Commun*, 9(1):870, 2018. ISSN 2041-1723 (Electronic) 2041-1723 (Linking). doi: 10.1038/s41467-018-03317-6.

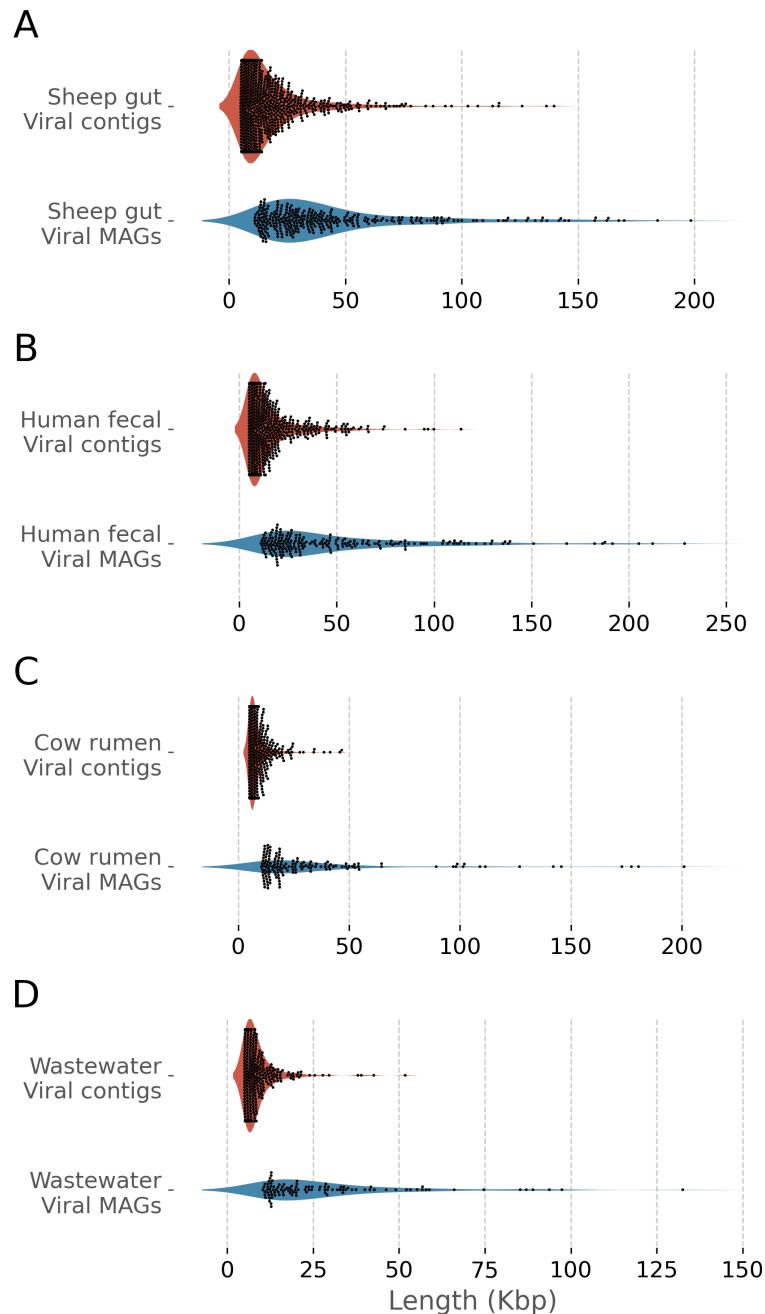


Fig. S1. Viral MAG lengths. The length distribution of viral sequences before (red) and after (blue) binning with ProxiPhage in metagenomic samples extracted from A) sheep gut, B) human stool, C) cow rumen, and D) wastewater.

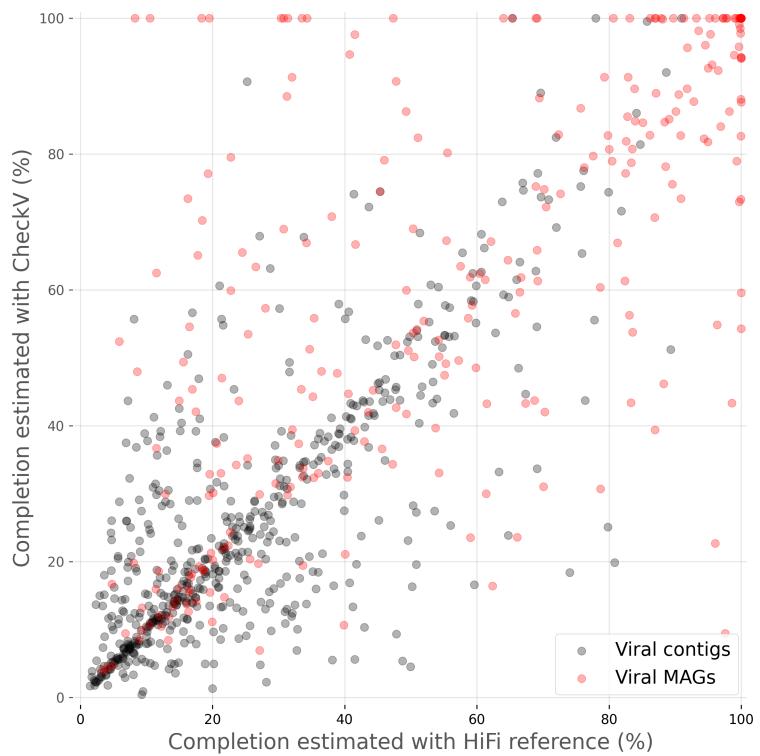


Fig. S2. Completion estimation comparison. Scatter plot of completion percentages of viral contigs and vMAGs from a sheep gut metagenome, estimated with CheckV (y-axis) and with reference excised phages from a long-read HiFi assembly (x-axis).

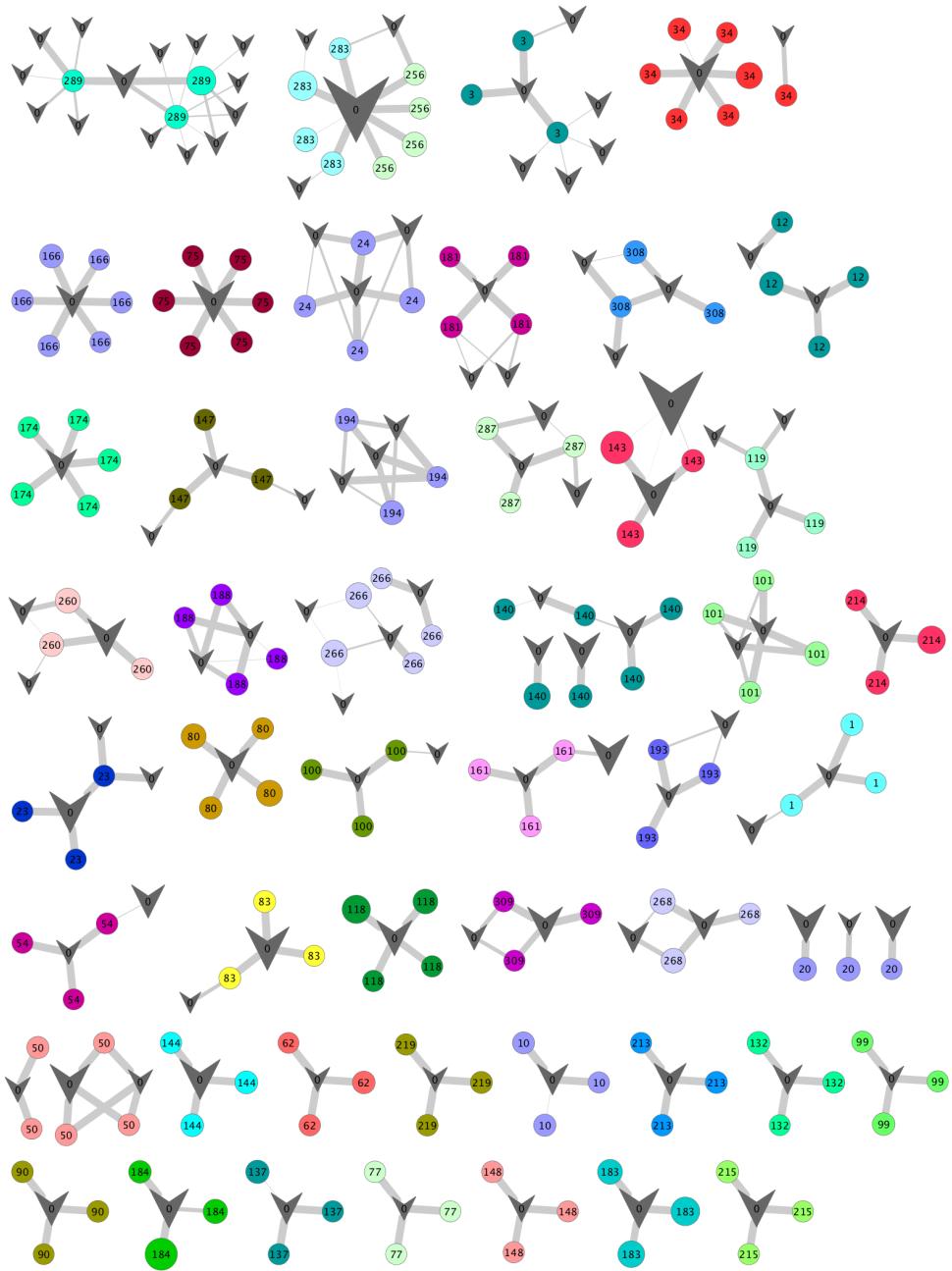


Fig. S3. Viral MAG validation network. Network showing long-read validation of vMAG contig clusters from a sheep gut metagenome. Viral contigs (circles, labeled with vMAG identifiers) are randomly colored according to the vMAG they belong to and linked to reference long-read viral genomes that they aligned to (grey check marks). The node size represents the viral sequence length, and the edge weight represents the percent of the short-read viral contig that aligned to the long-read reference. A random subset of 50 vMAGs was chosen for this visualization from a pool of vMAGs with at least 3 contigs and with at least one reference found for each contig.

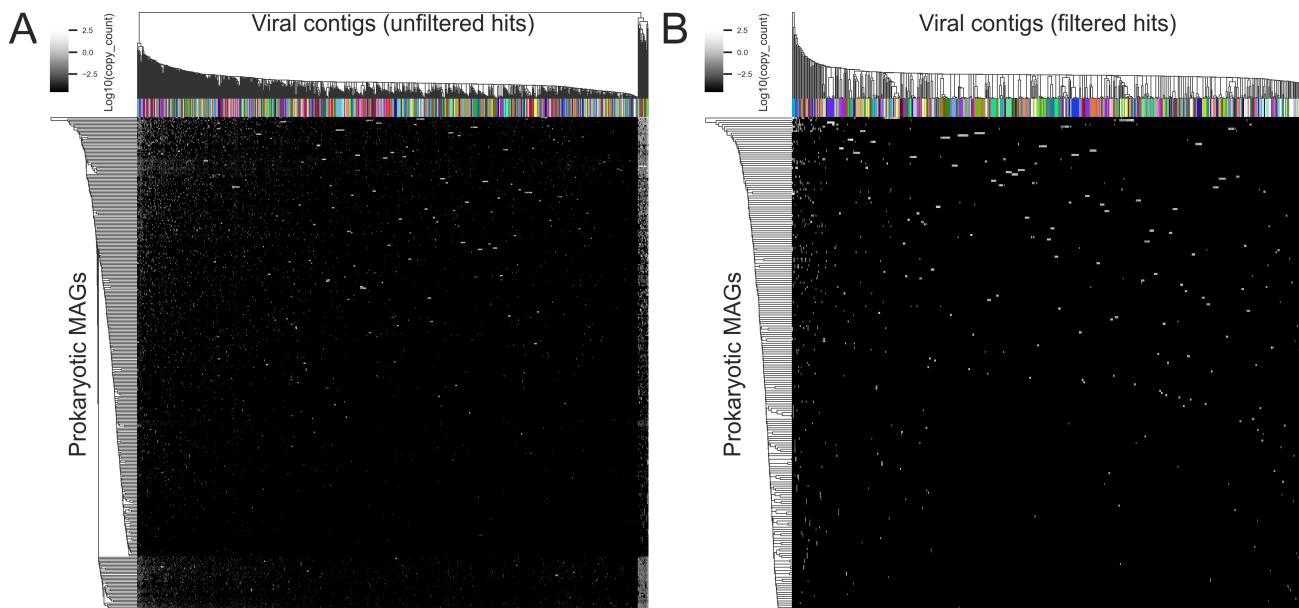


Fig. S4. Viral contig host predictions. Prokaryotic hosts identified for viral contigs from a sheep gut metagenome with ProxiPhage before (A) and after (B) thresholding. The color map represents the log of the estimated average copy count of each phage genome in its host. The colors above the phage contigs label the viral MAG that they belong to. Rows and columns are clustered according to linkage similarity with seaborn clustermap. Only viral contigs from viral MAGs are shown.

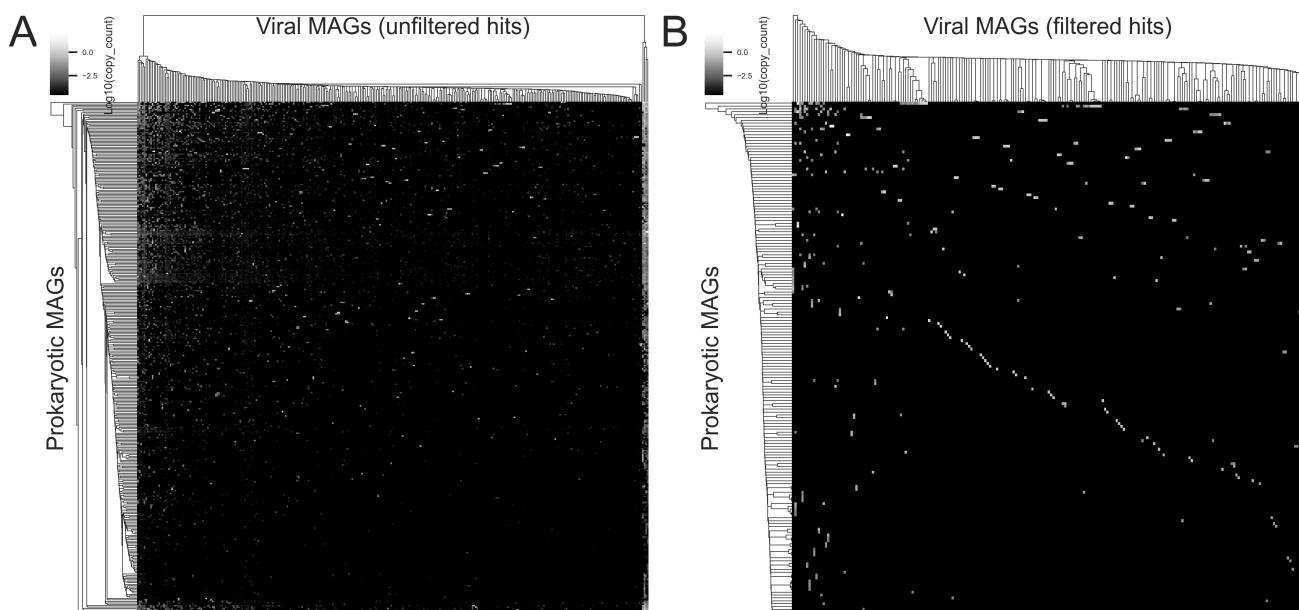


Fig. S5. Viral MAG host predictions. Prokaryotic hosts identified for viral MAGs from a sheep gut metagenome with ProxiPhage before (A) and after (B) thresholding. The color map represents the log of the estimated average copy count of each phage genome in its host. Rows and columns are clustered according to linkage similarity with seaborn clustermap.