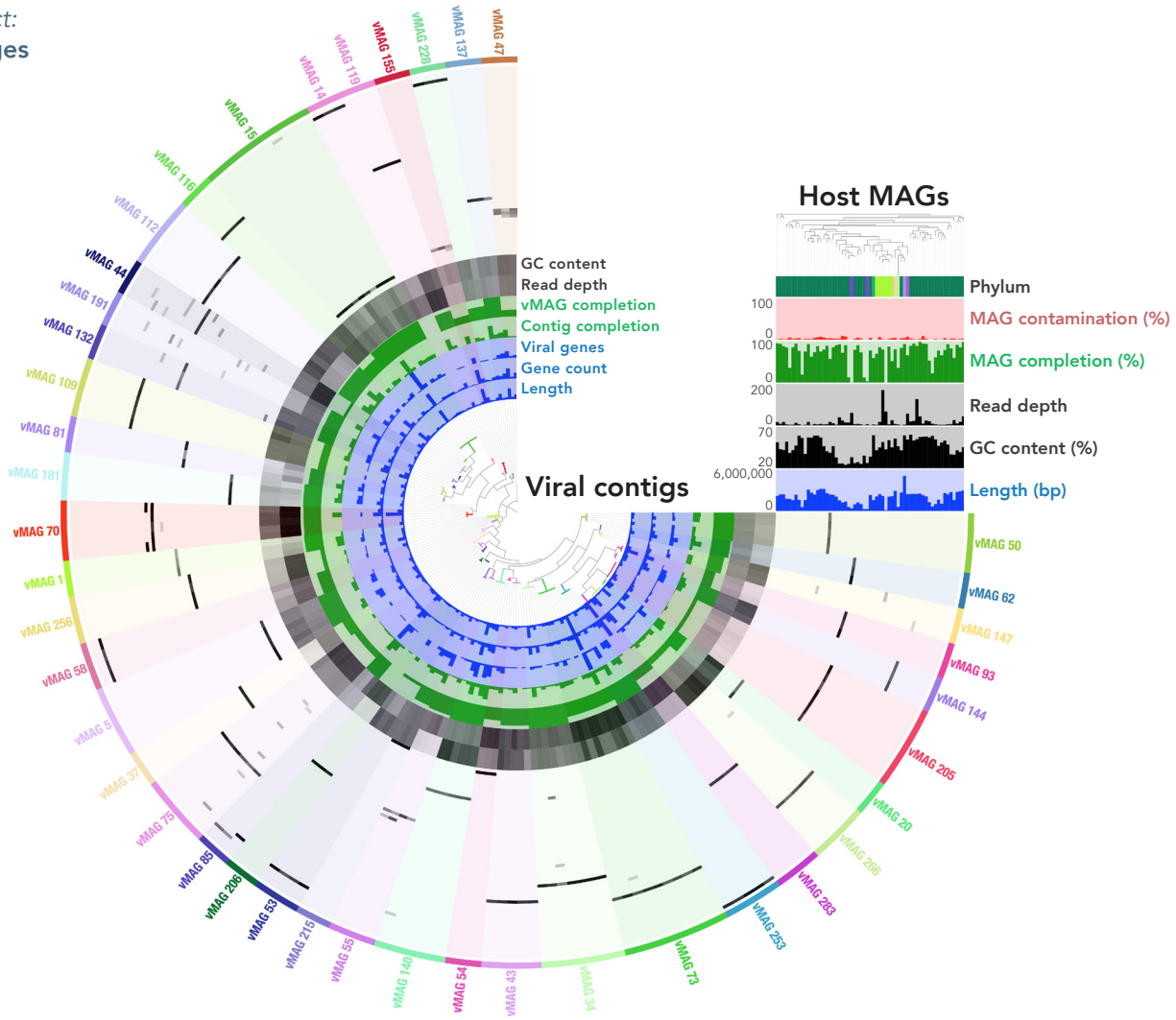
PHASE GENOMICS

Application Note

Accurate viral genome reconstruction and host attribution with proximity-guided metagenomics

The ProxiMeta™ Platform employs proximity ligation technology to significantly improve *de novo* binning and host attribution of bacterial and viral genomes recovered from complex metagenomic samples.

Graphical Abstract:
Connecting Phages with their Hosts



Circular plot of viral metagenome-assembled genomes (vMAGs, outer ring), constructed with the ProxiMeta Platform from short-read shotgun and proximity ligation (Hi-C) sequencing data. All 42 vMAGs depicted here consist of at least 3 contigs of ≥5 kb each, and every contig has a physical host connectivity signal. Contigs associated with each vMAG are shown in the same color. Bars within a contig represent physical links to a given microbial MAG reconstructed from the same sample, indicating a viral-host association. Darker bars represent higher estimated copy counts. Bars aligned across a vMAG designate connections of multiple contigs to the same host. Additional circular layers (viral) and bar plots (hosts, upper right) represent characteristics of individual contigs or MAGs, respectively. These include length, GC content and estimated completeness based on alignment to a long-read metagenomic assembly.[1]

Introduction

Viruses are ubiquitous and are the most abundant biological entities in the biosphere. Yet, they represent the largest unexplored genetic information space on earth. Viruses infect bacteria, archaea, and eukaryotes. As important vectors of horizontal gene transfer, they shape the evolution and population dynamics of their microbial hosts, as well as the natural and man-made ecosystems in which they occur.[2,3]

Metagenomics has become a driving force in the study of environmental viromes, contributing vastly to our knowledge of viral diversity and enabling functional characterization. Increasing sequencing read depth and decreasing cost have enabled routine recovery of metagenomes from human and animal microbiomes, and environmental samples. However, metagenome deconvolution and assembly pipelines have to rely on *a priori* knowledge, statistical assumptions, and binning algorithms, as shotgun sequencing data lacks genomic contiguity information. These analysis methods cannot tell with certainty which sequences originated from which cell in a complex microbial community. This leaves metagenome-assembled genomes (MAGs) incomplete and contaminated. In addition, conventional binning approaches are unable to accurately associate mobile genetic elements, such as bacterial plasmids, viruses or bacteriophages, or integrons and transposons, with their hosts. Inversely, this means that *bona fide* hosts for the viruses in metagenomic samples remain largely unidentified, leaving huge gaps in our understanding of the biological and ecological roles of these viruses.

The Phase Genomics ProxiMeta™ Metagenome Deconvolution Platform[4] employs proximity ligation (Hi-C)[5] technology to capture physical interactions between sequences within the same cell. The ProxiMeta analysis pipeline tool augments metagenomic binning with this additional layer of linkage information to reconstruct more, high-quality bacterial and viral genomes, and enable specific and sensitive host attribution of DNA viruses (phages).

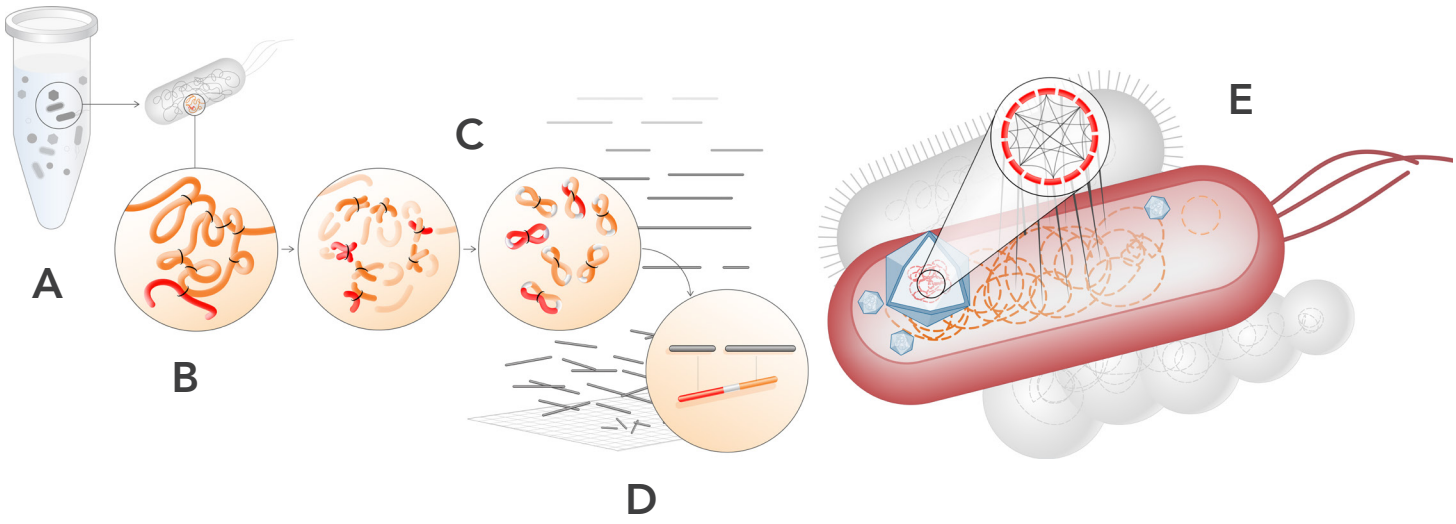


Figure 1. Overview of proximity-guided metagenome deconvolution and host attribution. Metagenomic samples are comprised of complex populations of archaea, bacteria and fungi (**A**). In addition to chromosomal DNA, cells contain genetic material such as plasmids, viruses or bacteriophages, transposons, integrons, and other mobile genetic elements. The first step in the generation of a proximity ligation library is crosslinking (**B**), which captures the physical interactions between DNA fragments in every individual cell. Digestion and ligation create chimeric junctions (**C**) that are sequenced and analyzed in combination with short- or long-read shotgun assemblies (**D**). The proximity ligation data provides an additional layer of information that is used to reconstruct more, high-quality bacterial and viral genomes than traditional binning approaches. In addition, the physical linkage information enables accurate viral-host attribution.

Library Preparation and Sequencing

To demonstrate the advantages of the ProxiMeta™ Platform for viral genome reconstruction and host attribution, several libraries were constructed from DNA extracted from an animal fecal sample:

- For shotgun sequencing on the Illumina® platform, a library was prepared with the TruSeq® DNA PCR-Free kit (Illumina). A total of 512 million reads (2 x 150 bp) were generated on an NextSeq™ 500 instrument, and were randomly downsampled to 100 million reads. A 837-Mb assembly was generated with MEGAHIT[6] using default parameters.
- A proximity ligation library was prepared with the ProxiMeta Hi-C Kit (Phase Genomics). Sequencing (2 x 150 bp) was performed on an Illumina HiSeq® 2000 instrument. Data were randomly downsampled to 100 million reads for analysis with the ProxiMeta pipeline.
- For ultra-high coverage long-read sequencing on the PacBio® platform, a size-selected HiFi SMRTbell® library (9 – 14 kb final fragment length) was prepared and sequenced using a combination of the Sequel® I and II systems. A total of 46 SMRT® Cells yielded 255 Gb of circular consensus sequencing (CCS) reads, which were filtered and assembled with metaFlye,[7] yielding a 3.4-Gb assembly.

The short-read and proximity ligation sequencing data were used for the construction of bacterial and viral metagenome-assembled genomes (MAGs) and host-attribution with the ProxiMeta analysis pipeline. The long-read data were used to validate the outputs from the ProxiMeta pipeline.

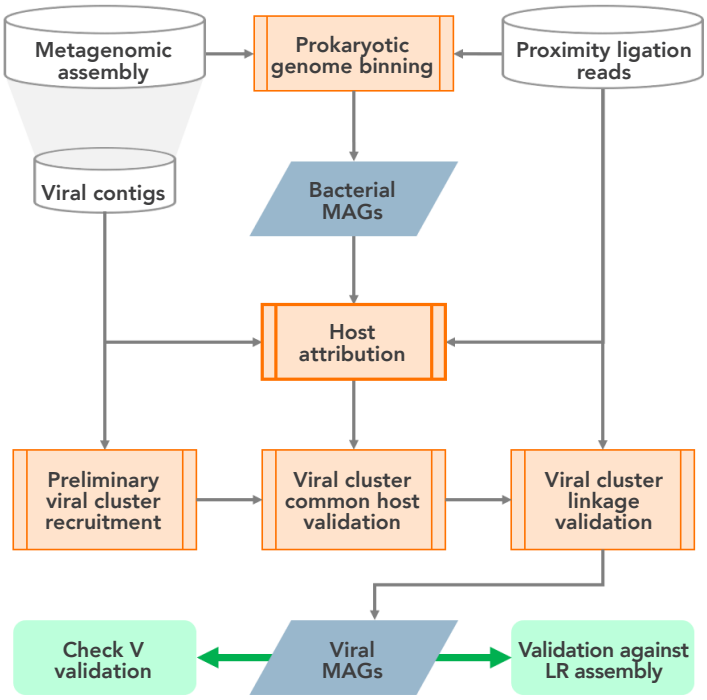


Figure 2. Overview of the ProxiMeta analysis pipeline for viral genome reconstruction and host attribution. The process starts with two key inputs: (i) annotated viral contigs from a metagenomic assembly derived from shotgun sequencing data, and (ii) proximity ligation reads from a Hi-C library prep. The ProxiMeta pipeline uses these inputs to produce high-quality metagenome-assembled genomes (MAGs) for both the bacteria and DNA viruses in the sample. The physical linkage data captured in proximity ligation reads are used to improve the number and quality of MAGs, and to accurately assign viruses to bacterial host cells.

In this study, the quality and accuracy of reconstructed viral MAGs were validated using CheckV,[9] or alignment to reference genomes derived from a long-read (LR) assembly for the same sample. These validation steps are not part of the standard ProxiMeta workflow.

Data Analysis and Results

Viral binning with the ProxiMeta pipeline significantly increases the completeness and quality of reconstructed viral genomes

Viral contigs derived from the shotgun assembly were annotated with VirSorter2[8] and clustered into putative viral MAGs (vMAGs) using the viral binning feature of the ProxiMeta pipeline (Figure 2). In the initial grouping stage, the pipeline uses classical metagenomic binning techniques (tetranucleotide frequency and coverage depth similarities) in combination with proximity ligation data to generate preliminary viral clusters. The two groupings are compared using an overlap network, and collapsed into a final, improved viral bin set. These bins are processed through several progressively more stringent quality filters to remove false positive associations. The two primary filters are (i) designed to detect physical links shared between contigs in the same cluster, and (ii) verify that contigs from the same cluster have the same bacterial hosts. Together, these approaches allow the pipeline to separate out viral genomes that may be infecting multiple bacterial hosts and thus have physical links that are grouping them into a viral pan-genome.

A total of 1,163 vMAGs were generated from the animal fecal sample. Individual vMAGs were comprised of 2 – 23 contigs (average of 3 contigs per vMAG), and ranged between 2.2 kb and 256 kb in size (20 kb on average).

To validate the ProxiMeta™ pipeline, the quality of reconstructed viral genomes was assessed using two different approaches:

- First, genome completeness was assessed by comparing vMAGs to unbinned viral contigs using CheckV. This algorithm uses an extensive viral lineage and protein database to estimate the completeness of viral genomes.[8] Results are summarized in Figure 3A.
- To obtain a more robust vMAG quality benchmark, clusters were also validated against a PacBio® HiFi assembly for the same sample. In short, VirSorter2 was used to annotate viral contigs in the long-read assembly. Complete prophage genomes were excised from the long-read contigs, and used as reference genomes. Viral MAGs and unbinned viral contigs from the short-read assembly were aligned to these reference genomes using BLAST.[10]

  A custom algorithm was subsequently used to evaluate vMAG quality. To this end, high-quality alignments (>95% identity, >100 bp) were used to find the best long-read viral reference for each vMAG (i.e., the reference to which the greatest fraction of the viral MAG sequence aligned). After applying stringent parameters (80% alignment of at least one contig), reliable references were obtained for 505 of the 1,163 vMAGs. Those alignments were used to calculate completeness (percentage of the reference virus genome aligned to the vMAG), and contamination (percentage of the vMAG that did not align to the reference). Results are summarized in Figure 3B and Figure 4.

Both validation strategies confirmed a significant gain in high-quality reconstructed viral genomes with the ProxiMeta pipeline, as compared to the unbinned assembly. The CheckV analysis showed a 10-fold increase (from 9 to 91) in the number of near-complete viral genomes (>95% completion, <10%), whereas the benefit was even higher (from 3 to 74) using the long-read approach.

Contamination rates calculated from reference genome alignments further allowed us to evaluate the accuracy (false positive rate) of clustering with the ProxiMeta pipeline. As shown in Figure 4, only 21 of the 505 testable vMAGs returned a contamination score >10%. This represented just 4% of all clusters.

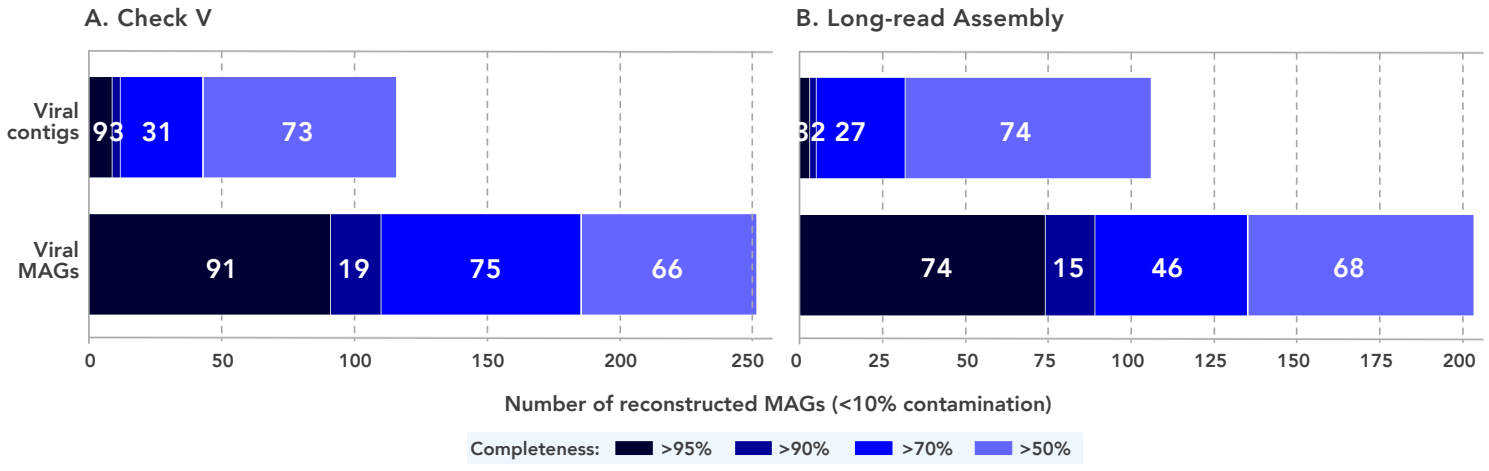
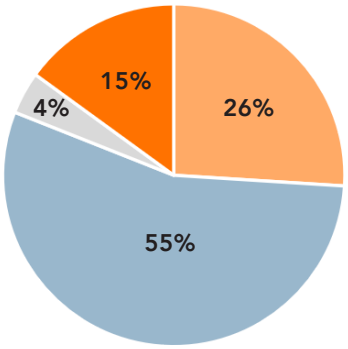


Figure 3. Number and completeness of high-quality (>50% complete <10% contaminated) reconstructed viral genomes, before and after binning with the ProxiMeta pipeline. Completeness and contamination of individual contigs (from the short-read assembly) and Proximeta-reconstructed vMAGs were estimated with (**A**) viral marker genes using CheckV, or (**B**) alignment to reference phage genomes excised from the long-read assembly. Both approaches confirmed that proximity ligation-based binning produces significantly more high-quality viral genomes than the unbinned assembly.



Quality	Completeness	Contamination	Count	%
Near-complete	>95%	<5%	75	15
Moderate	>50%	<10%	129	26
Partial	>0%	<10%	280	55
Contaminated	>0%	>10%	21	4

Total vMAGs: 1,163 Untestable: 658 (57%) Testable: 505 (43%)

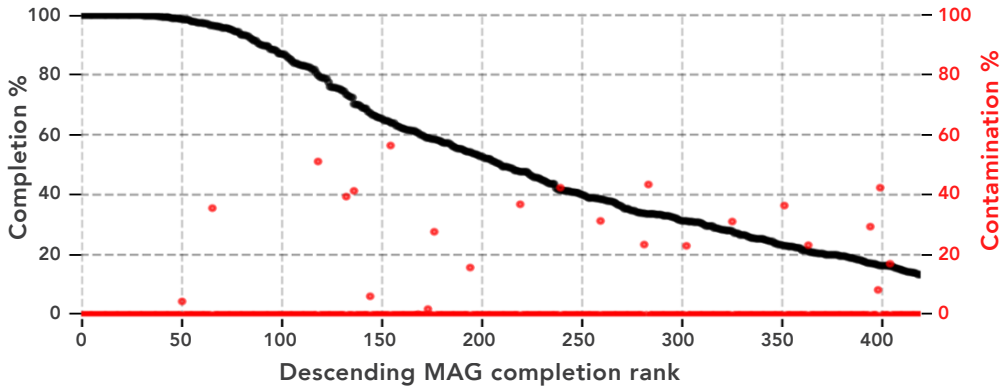


Figure 4. Quality and accuracy of viral MAGs reconstructed with the ProxiMeta pipeline. Completeness and contamination levels of testable viral MAGs were estimated by aligning each to a reference phage genome, assembled from HiFi sequencing on the PacBio® platform. Only 4% of the 505 testable MAGs were regarded as contaminated, confirming that the pipeline reconstructs viral genomes with a high degree of accuracy.

The ProxiMeta™ pipeline's host attribution algorithms dynamically assign bacterial hosts to viruses

The pipeline's host attribution function is designed to find reliable physical links between bacterial and viral genomes (MAGs) in order to predict the likely bacterial host(s) for each virus. The physical links between a virus and a putative host are used to estimate the average copy count of the virus genome per prokaryotic cell, using the following formula:

$$C = \frac{V}{H}\frac{L}{\Sigma L(v)}$$

where the average viral copy count per cell (C) is calculated from virus abundance (V), prokaryotic host abundance (H), physical links between the virus and host (L), and total links between the virus and all possible hosts.

It is important to consider that physical links only reflect the **average** connectivity within a given microbial community and not the median. In other words, a copy count of 1 may mean that all cells have 1 copy of a phage, but it could also mean that 1% of the cells have 100 copies.

Next, the estimated copy count is used to statistically evaluate each virus-host association in the context of how connected the bacterial genome is to itself. This is done to determine if the connectivity density is similar to what would be expected by random chance if this was the correct host. For this evaluation, the following formula is used:

$$R' = \frac{D_{VH}}{C\,D_H}$$

where the normalized connectivity ratio (R') is calculated from the connectivity density (links per kb^2) between the virus and host (D_{VH}) and of the host genome to itself (D_H), and normalized to the average virus copy count per cell (C).

Finally, a receiver operating characteristic (ROC) curve is constructed to determine the optimal cut-off value for the minimum copy count (among other internal thresholds) for the particular metagenomic sample (Figure 5). This allows the ProxiMeta pipeline to reliably remove false-positive virus-host associations from the data.

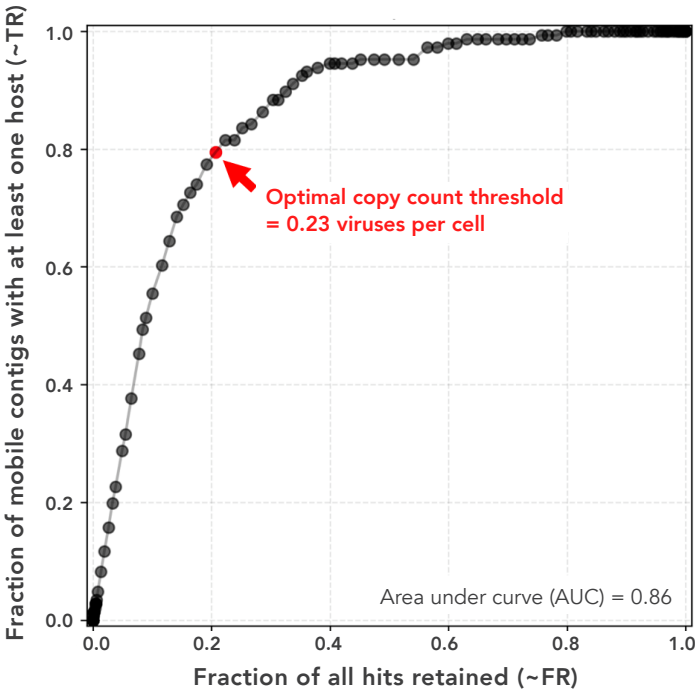


Figure 5. Receiver operating characteristics (ROC) curve showing the decline in the number of bacteria-phage host associations (x-axis) and in the number of phages with at least one host (y-axis) as the threshold of the minimum average copy count of each phage genome in its host is raised. The area under curve was 0.86 and the optimal copy count threshold was determined to be 0.23 viruses per cell.

Using this dynamic host attribution algorithm, prokaryotic MAG hosts of at least 3 kb in length were identified for 3,483 out of 5,568 viral contigs (63%). The final host connectivity matrix for a subset of these binned viral contigs are shown in Figure 6. In addition to identifying the exact taxonomy of the hosts of most viruses in a given sample, the ProxiMeta™ pipeline is also capable of revealing putative co-infection events (horizontal lines in Figure 6) and promiscuous phages infecting multiple hosts (vertical lines in Figure 6). These data can be used to further curate or validate virus-host attributions, as contigs belonging to the same cluster are expected to have the same hosts, unless coverage dropout causes false-negatives (color bar in Figure 6).

Host attribution with the ProxiMeta pipeline is both sensitive and specific

The host attribution function of the ProxiMeta pipeline evaluates each virus-host linkage in the context of the connectivity that would be expected by random chance if it were a real connection. This allows the algorithm to assess the accuracy of even very low-coverage sequences, and assign a host to a virus from as little as two proximity ligation reads. Of the 8,944 virus-hosts associations made for the animal fecal sample, 57% (5,114 associations) were based on fewer than five physical read connections—which is the standard cut-off used in other host-attribution pipelines.

Given the high sensitivity, it was critical to evaluate the specificity (false-positive rate) of host assignment. To this end the long-read assembly was revisited. Unlike proximity ligation reads, long reads do not carry inter-molecule association information and cannot be used to verify most virus-host pairs. However, this is possible when focusing on prophages, which were abundant in this data set. The sequence flanking the prophage in the long-read contig was compared against the sequence of the predicted host MAG from the short-read assembly to determine whether it was the same host. Of the 498 prophages with an identified host, at least 444 (89%) were validated using the long-read assembly. Note that it is possible that the remaining 54 prophage associations are still accurate, but that the flanking bacterial sequences from the long-read contig were not included in the host MAG during genome binning.

Conclusion

In this study we used an animal fecal sample to illustrate the key features and benefits of the ProxiMeta Platform. Hundreds of high-quality viral genomes (<10% contamination) were reconstructed from a modest amount of short shotgun plus proximity ligation reads (100 million each), with push-button convenience—a feat that is otherwise only possible with ultra-deep, high-fidelity, long-read sequencing and custom scripts. In addition, prokaryotic MAG hosts of at least 3 kb in length were identified for 63% (3,483 out of 5,568) of the viral contigs. The completeness and quality of reconstructed viral MAGs were validated using a well-known algorithm (CheckV), as well as alignment to reference genomes derived from a PacBio® HiFi assembly for the same sample. The long-read assembly was also used to validate the high specificity and sensitivity of viral host attribution. Evidence of viral co-infection of the same host, as well as promiscuous phages (infecting different hosts) was uncovered.

The ProxiMeta Platform is the only commercially available technology designed to apply proximity ligation data to the deconvolution of complex metagenomic data sets. The ProxiMeta analysis pipeline utilizes this genomic contiguity information in an innovative workflow to enable significant improvements in the quantity, completeness and quality of reconstructed viral genomes. In addition, the pipeline employs a sophisticated statistical approach to dynamically assign bacterial hosts to viruses, making it the first platform for reliable viral host attribution in metagenomic samples.

As the "perfect predators" of microbial communities,[11] phages and DNA viruses impact the genomic plasticity of their hosts and ecosystems, with functional impacts all the way up the proverbial food chain. As such, the ProxiMeta Platform is a powerful tool for emerging applications, such as viral therapy and fecal microbiota transplants.

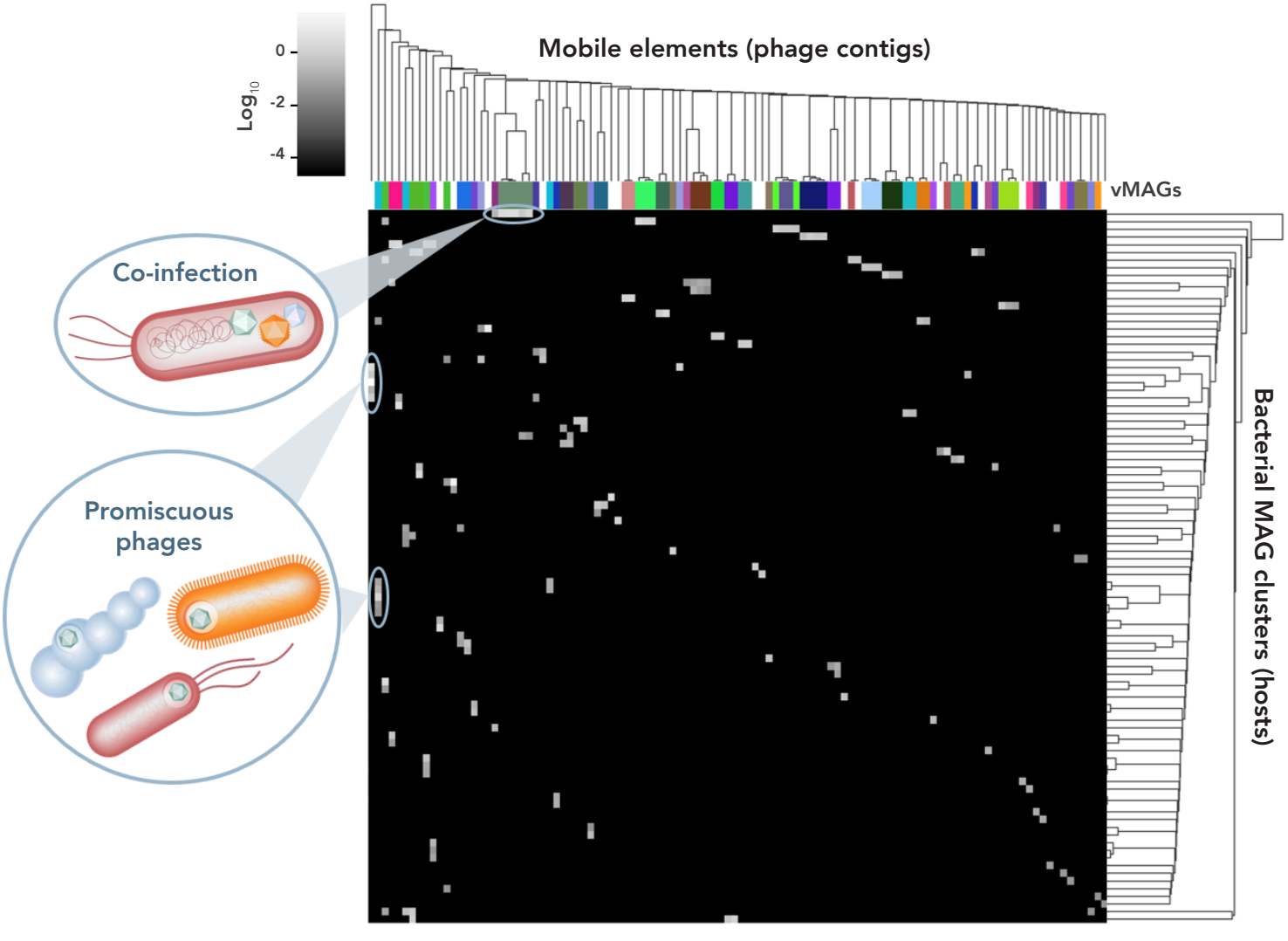


Figure 6. Bacterial hosts identified for viral contigs with the ProxiMeta pipeline. The color map represents the log of the estimated average copy count of each phage genome in its host. Colors above phage contigs designate the vMAG that they belong to. Only viral contigs form near-complete vMAGs are shown. Call-outs emphasize examples of putative co-infection events (horizontal lines) and promiscuous phages infecting multiple hosts (vertical lines).

References

1. Uritskiy G, et al. Accurate viral genome reconstruction and host assignment with proximity-ligation sequencing. *bioRxiv* 2021.06.14.448389. doi: 10.1101/2021.06.14.448389.
2. Rohwer F, et al. Roles of viruses in the environment. *Environ Microbiol* 2009; 11:2771-2774. doi: 10.1111/j.1462-2920.2009.02101.x.
3. Dávila-Ramos S, et al. A review on viral metagenomics in extreme environments. *Front Microbiol* 2019; 10:2403. doi: 10.3389/fmicb.2019.02403.
4. Capture a complete picture of complex microbial communities, including the moving parts. Phase Genomics Application Note 2020. phasegenomics.com/wp-content/uploads/2020/09/ProxiMeta-Application-Note_Aug-2020.pdf.
5. Burton JN, et al. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3* (Bethesda, Md.) 2014; 4(7):1339–1346. doi: 10.1534/g3.114.011825.
6. Li D, et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015; 31(10):1674-1676. doi: 10.1093/bioinformatics/btv033.
7. Kolmogorov M, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* 2020; 17:1103-1110. doi: 10.1038/s41592-020-00971-x.
8. Guo J, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* 2021; 9:37. doi: 10.1186/s40168-020-00990-y.
9. Nayfach S, et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* 2021; 39:578–585. doi: 10.1038/s41587-020-00774-7.
10. https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=References.
11. Strathdee S and Patterson T. The perfect predator: a scientist's race to save her husband from a deadly superbug. New York: Hachette Books, 2019. ISBN: 9780316418089.

Learn more about the ProxiMeta™ Platform at phasegenomics.com/products/proximeta/

PHASE GENOMICS

info@phasegenomics.com
www.phasegenomics.com

Phone: 1-833-742-7436
Twitter: @PhaseGenomics