PhaMers identifies novel bacteriophage sequences from thermophilic hot springs

By
Jonathan Deaton

Under the supervision of Dr. Stephen Quake, Bioengineering Department, Stanford University

February, 2017

PhaMers identifies novel bacteriophage sequences from thermophilic hot springs

By
Jonathan Deaton
February 23, 2017

Approved: _____
Stephen Quake, Ph.D.
February 23, 2017
Research Advisor from the Department of Bioengineering

Approved: _____
Russ Altman, Ph.D., MD.
February 23, 2017
Faculty Reader from the Department of Bioengineering

Approved: _____
Karl Deisseroth, Ph.D., MD.
February 23, 2017
Chair for Undergraduate Education in the Department of Bioengineering

Preface and acknowledgements

I would like to thank Professor Stephen Quake for graciously overseeing my work in his lab for the past two years, and for giving me the initial concept for this project. Professor Quake generously provided the resources and environment that facilitated this research, enabled my growth as a scientist, and helped my overall learning as a student. In addition to being an advisor and a role model as a scientist, he has been an excellent source of knowledge and guidance in project direction. I am incredibly grateful to have had the opportunity to work in the Quake Lab, as it has been one of the most enriching experiences of my time as an undergraduate at Stanford.

I am incredibly grateful to Brian Yu for his continued help and advising throughout my undergraduate career. Brian oversaw my first experiences in laboratory science and has generously provided me with nearly all of my laboratory skills. Not only has Brian aided me with technical challenges of my projects, but he has also gifted me the indispensable skill of thinking like a research scientist. A student cannot be luckier than I have been in having such a patient and knowledgeable teacher as Brian.

I would also like to thank my faculty reader, Professor Russ Altman, for agreeing to read and advise my thesis. I also am grateful for the inspiration and bio-computational toolset given to me in his introductory bioinformatics class. Both of these were very necessary for my completion of this project.

I would like to acknowledge Paul Blainey for sample collection, and the members of the Stanford Stem Cell Institute Sequencing Facility: Norma Neff, Jennifer Okamoto, Gary Mantalas, and Ben Passarelli. I acknowledge the generous financial support from the DOE JGI's Emerging Technologies Opportunities Program (ETOP), Templeton Foundation, Stanford Graduate Fellowship, NSF GRFP, and the Stanford Bioengineering REU program.

Finally, I would like to thank my parents, Angelica and David Deaton, for the support that they have given me throughout my entire life. I owe all of my accomplishments and successes to their expert parenting and guidance.

Table of Contents

List of Figures

Abstract

Discovering novel phage sequences from metagenomic data is often challenging. This study presents PhaMers (Phage *k*-Mers), a phage identification tool that uses supervised learning to identify metagenomic sequences (contigs) as phage or non-phage on the basis of tetranucleotide frequencies. PhaMers compares the tetranucleotide frequencies of metagenomic contigs to those of phage and bacteria reference genomes from online databases. Using PhaMers, we identified 103 novel phage sequences in hot spring samples from Yellowstone National Park. We applied a microfluidic-based mini-metagenomic approach [1] to sequence environmental samples and produce metagenomic sequence datasets. We analyzed assembled contigs using PhaMers and VirSorter [2], a publicly available phage identification and annotation pipeline. We present the performance of PhaMers in identifying genomic fragments of phages and its ability to predict phage taxonomic classification. We also present putative hosts and taxa for some novel phage sequences. PhaMers is available for public use at https://github.com/jondeaton/PhaMers.

Introduction

Bacteriophages (phages) are viruses that infect microorganisms such as bacteria and archaea. Phages play important roles in microbial communities such as gene duplication and lateral gene transfer, and at an estimated 10^{31} viral particles, are also the most abundant biological entities on the planet [3]. Despite this abundance, our understanding of the genomic diversity of phages is limited to thousands of partial and completed genomic sequences. Before Next Generation Sequencing made high-throughput metagenomics possible, our understanding of phage genomics was limited to phages that could be cultured in plaques [4]. With these powerful tools in hand, our understanding of phage genetic diversity expands and is limited only by the throughput of metagenomic sequencing and our ability to computationally characterize phage sequences [5]. The frequently small size of phage genomes (many are smaller than 5,000 base pairs) (Fig. 1d) and nonexistence of any pervasive genetic marker (e.g., rDNA used in bacterial genome identification) complicates the identification of phage genomic fragments [6]. Common approaches to phage identification involve the use of Hidden Markov Models (HMM) to produce gene annotations, followed by a search for coding regions that are homologous to known viral genes [7]. The viral genes that are typically searched for (called hallmark or marker genes) may include the terminase large and small subunits, major capsid, coat, tail, and portal proteins [2]. The rationale behind this approach is grounded in the central dogma of molecular biology, and it often performs well for identifying species and annotating novel sequences in metagenomic samples. However, this approach is weak in that is requires accurate multiple sequence alignments, and may fail at identifying phages with no known genetic homologues. A less common approach involves considering the frequencies of short oligonucleotides of length k (k-mers) [8]. Frequencies of k-mers vary among species and can be used to predict phylogeny and taxonomy.

In this thesis, we present the development and application of PhaMers, a novel phage identification algorithm that uses tetranucleotide (4-mer) frequencies and supervised learning to identify viral contigs in metagenomic sequencing data. PhaMers also integrates and presents results from VirSorter, an automated phage identification pipeline [2]. Both algorithms were applied to metagenomic sequences (contigs) generated from the sequencing of environmental samples taken from three locations in Yellowstone National Park (YNP), resulting in the

identification and characterization of 103 novel bacteriophages. Additionally, gene predictions made by VirSorter were verified using the Integrated Microbial Genomes & Microbiomes (IMG) annotation pipeline [9].

PhaMers applies supervised machine learning algorithms in the tetranucleotide frequency feature space to predict whether a DNA sequence is likely that of phage. PhaMers compares the tetranucleotide frequencies of unidentified metagenomic contigs to those of sequences from established datasets of phage and bacterial genomes from RefSeq and GenBank, respectively. PhaMers performed well when tested on these datasets with K-fold cross-validation, showing 89% sensitivity, 98.7% specificity, and 98.5% positive predictive value (PPV). By comparing putative phages identified by VirSorter and PhaMers in samples from YNP, we have shown that PhaMers remains predictive when applied to metagenomics samples. This being the case, we used PhaMers' scoring algorithm to increase confidence of viral sequence prediction.

In addition, we explored the relationship between tetranucleotide frequencies and phage taxonomy, both in a reference dataset of known phages and in the novel phages identified from YNP. We demonstrate visualization of metagenomic sequencing data by embedding tetranucleotide frequency vectors in two dimensions using t-Distributed Stochastic Neighbor Embedding (t-SNE), a manifold learning dimensionality reduction algorithm well suited for visualizing high dimensionality data [10].

This thesis begins with a background section that reviews the microbiology of phages, our motivation and methodologies for studying them, the field of metagenomics, and computational tools used in this research. Materials and methods follows with a detailed overview of how samples were collected, prepared, sequenced, and analyzed. The results section begins with results from computational analyses of our reference dataset of phages, followed by statistical test results quantifying the predictive performance of PhaMers. Next are the results from the analysis of metagenomic contigs from YNP. In the discussion section, we review the genomic attributes of several novel phages, and the advantages and disadvantages of PhaMers. This thesis concludes with a discussion of future research that could make use of and extend this work.

Background

Bacteriophage Overview

Viruses are small biologic entities typically between 20 and 200 nm in length that contain a genetic code stored in single- or double-stranded RNA or DNA surrounded by a coat of proteins. Viruses infect all manner of life (i.e., prokaryotes and eukaryotes); however, those that infect bacteria or archaea are referred to as bacteriophages (phages).

A phage's occupation of a host cell is divided into the lytic cycle and the lysogenic cycle. The lytic cycle includes the virus binding to the host, inserting its genetic material, replicating in the host, and then destroying the host through lysis. The lytic cycle begins with the phage's binding to the plasma membrane or cell wall at specific binding sites. In the case of tailed phages like *Caudovirales*, this binding occurs at the end of the phage's tail. These interactions are typically mediated by complex baseplate, tail spike, or tail fiber proteins [11]. The phage then penetrates the plasma membrane, inserting its DNA or RNA genome into the cytoplasm of the host. The genome may be in the form of a circular plasmid or a linear strand.

At this point, a phage may either continue with the lytic cycle or enter the lysogenic cycle. Phages that bypass the lysogenic cycle, such as *bacteriophage T4* [12], are said to be virulent, whereas those that participate in the lysogenic cycle are said to be temperate [13]. Research has shown that approximately 60% of sequenced bacterial genomes contain temperate virus regions [2] [14] [15]. The lysogenic cycle begins with the integration of the phage genome into the genome of the host, a process that is mediated by phage integrase proteins. In the case of temperate RNA viruses (retroviruses), this integration is proceeded by the reverse transcription of the phage's genome from RNA into DNA through the action of reverse transcriptase. When the phage's genome is part of the bacterial genome, it is referred to as a prophage. The phage may then stay integrated in the genome of the host for an indefinite amount of time, replicating by taking advantage of the host's genomic replication mechanism. In a process known as induction, a phage can transition back into the lytic cycle, replicating itself with the host's protein expression machinery, and recapitulating itself in its full protein-coated form in the cytoplasm. For lytic phages, the lytic cycle concludes with the lysis of the host membrane. Because phages never produce proteins with secretory signals, they must create lesions in the

host's membrane and protein coat through the action of proteins like holin and murein hydrolases such as lysozymes, glucosaminidases, amidases, and endopeptidases [16] [17]. For non-lytic phages, the host is left intact while phage particles "bud out" and go on to infect other hosts.

The structure of a phage particle is relatively simple: it contains a nucleotide genome surrounded by a protein coat. The proteins that polymerize to form the coat surrounding the genome are often called capsid proteins, and they are chaperoned into place during phage assembly by scaffolding proteins. Because the head of the phage is often small and the genome can be very big, the phage must pack its genome into the head using tremendous force generated by the terminase protein. The terminase protein is frequently encoded in two genes called the terminase large and small subunits. In the case of tailed phages, the tail is constructed of tail proteins that provide structure; a portal protein through which the phage's genome passes while entering the host; and some spike, baseplate, or tail fiber protein at the distal end of the phage's tail. Temperate phages include integrase proteins that are responsible for pasting the genome of the phage into the genome of the host. Phages also frequently encode tRNA molecules, which can be used to stop or alter the functioning of the host's ribosome machinery.

We classify phages (and viruses in general) taxonomically in a fashion similar to how we classify plants and animals at the levels of kingdom, phylum, class, etc. The International Committee on Taxonomy of Viruses (ICTV) publishes a taxonomic classification standard for all known viruses that divides them into taxa at the levels of Baltimore classification, order, family, sub-family, and species. Though it has been proposed [18], the ICTV taxonomic classification of phages does not contain any sequence-based classification scheme such as the 16S based classification used for microorganisms. Instead, classification is based on physical attributes of the viral particle. The Baltimore classification divides viruses into taxa based on the chemical form of their genome (e.g., single-stranded RNA, double-stranded DNA). Lower taxonomic classifications are delineated by other physical features such as tail morphology, head geometry, or method of replication inside the host. The majority of phages with known genomic sequences for are double-stranded DNA phages (dsDNA), and of those, the vast majority belong to the order *Caudovirales. Caudovirales* are, like many taxa, sub-classified on the basis of tail morphology. *Caudovirales* contain the families of *Podoviridae*, *Myoviridae*, and *Siphoviridae*.

Podoviridae are phages that have short non-contractile tails, whereas *Myoviridae* and *Siphoviridae* have long tails, which are contractile for *Myoviridae* and non-contractile for *Siphoviridae* [19].

Phages play unique roles in microbial communities. In aquatic environments, there are typically on the order of ten billion phages per liter [20]. Phages are critical in nutrient turnover [21], such as in the transfer of carbon from bacterial biomass into the pool of organic matter supporting microbial communities. Phages that infect cyanobacteria (cyanophages) have been shown to encode genes that regulate the photosynthetic systems of cyanobacteria [22] [23]. Many phages contain genes that modulate the metabolisms of their hosts, which has been hypothesized to be a mechanism by which microbial communities regulate metabolite flux [24].

Motivation

Microorganisms like bacteria, archaea, and phages, are ubiquitous entities in the biosphere, playing key roles through their interactions with each other, their environments, and multicellular organisms. Understanding microorganisms is a key step to understanding our own biologic functions and those of the environment. It is also a crucial step in the development of enzymatic biotechnologies.

This study focuses on the discovery and characterization of novel phages. This pursuit offers many prospects to the scientific community and society, which include the potential discovery of new antibiotics [25], novel cancer therapies [26], and phage therapy to treat bacterial infections [27] [28] [29]. Another possibility is the identification of novel genetic elements that can be used for research or by the public. One example is integrases and transposases, which are often present in prophage genomes. Integrases and transposases are used extensively in protocols for DNA sequencing and molecular biology. Though it was not found in a phage, the Taq Polymerase enzyme (now used ubiquitously in molecular biology for PCR) was originally found in the genome of *Termus Aquaticus,* a bacteria discovered in hot springs very near to those studied in this work [30].

One of the first approaches that scientists take in the study of microorganisms is to characterize their genetic composition. One reason this approach is so common is because of the speed and accuracy with which modern DNA sequencers can determine an organism's genomic sequence. Another reason is that that the genomic sequence of a microorganism significantly describes its characteristics. This information, combined with knowledge of the microorganism's environment and an understanding of the genetic underpinnings of biologic function, can reveal nearly everything about that microorganism. One major challenge in studying microorganisms gnomically is that we lack a true understanding of how to determine biologic function from arbitrary genetic code. This is in part due to the very narrow understanding of the genetic diversity on earth. It is estimated that there are 5.3×10^{31} megabases (Mb) in the biosphere [31], only a tiny fraction of which have actually been sequenced. For much of the genetic material that has been sequenced, we lack an understanding of how that genetic code manifests itself in biologic functions. Given that protein engineering is a relatively young field, protein engineers generally begin their construction of novel peptides from the starting point of functioning proteins found in nature. Metagenomics is one main asset available to us for discovering novel proteins that might be already useful or potentially useful once engineered.

Metagenomics

Metagenomics is the study of genetic material from environmental samples. This field is dedicated to expanding our knowledge of the genetic diversity of the biosphere. The beginning of metagenomics was enabled by the advent of Sanger sequencing in the 1970s. Sanger sequencing is a low-throughput method of DNA sequencing that allowed for the sequencing of the genomes of the first two phages: ϕX174 [32] and bacteriophage MS2 [33]. The study of bacteriophages using metagenomic approaches is central to expanding our understanding of their genetic diversity, as less than one percent of viral hosts are culturable. Phages were first described in the early 1900s [34]; however, viral metagenomics did not begin until 2002 with the sequencing of uncultured marine viruses [35]. Since then, and with the progression of DNA sequencing technologies, we have been able to sequence thousands of genomes of bacteria, archaea, and viruses.

With these powerful tools at our disposal, the challenge of phage discovery shifts towards the computational challenge of identifying phage genomes when presented with many unannotated sequences from the environment.

There are two traditional approaches to metagenomic studies: bulk metagenomics and single-cell metagenomics. Bulk metagenomics involves the sequencing of genetic material in environmental samples without first separating microorganisms at the single-cell level. This approach has the advantage of producing a high-level view of the genetic diversity of an environment, yet often fails to capture the characteristics of microorganisms that exist in low abundance. Single-cell metagenomics offers a solution to this problem through the sorting and sequencing of individual cells, which can yield the characterization of microorganisms at very low abundances but is limited by the throughput of single-cell sequencing systems.

Mini-Metagenomics is a recent method of sample preparation intermediate to single-cell and bulk metagenomics. Mini-metagenomics offers the single-cell resolution of single-cell metagenomics but with increased throughput that can be comparable to bulk metagenomic methods. Instead of capturing a single cell in a chamber (as is done in single-cell metagenomics), mini-metagenomics aims to capture between five and ten bacterial cells per chamber to sequence together. We used a mini-metagenomics protocol because of the advantages that it offers in viral metagenomics. One advantage is the ability to sequence phages that exist in low abundance with increased throughput. Mini-metagenomics may have an advantage in terms of identifying phage sequences, because if a cell is infected, when that cell is captured in a chamber, the effective concentration of the phage genome is increased. Another advantage is that identifying the phage's host based on presence patterns of phages and bacteria in the chambers of the microfluidic device may be possible.

Computational Approaches

There are many existing algorithms and software tools for identifying phages in metagenomic sequencing data. Some of these include Prophinder [36], PhiSpy [37], Phage_Finder [38], Prophage Finder [39], and PHAST [40]. All of these tools use sequence alignment algorithms to search for viral gene enrichment. In order to functionally annotate a DNA sequence, each

putative coding region in that sequence must be compared to a database with sets of aligned coding regions from other organisms. Creating these sets of aligned genes, known as multiple sequence alignments (MSA), requires a wealth of annotated DNA sequences. Fortunately, sequence-driven microbiology has existed for long enough to have made such data available and computationally useful. Despite this, incorrect annotations lead to errors in future annotations. One of the biggest challenges with this approach is that sequences with no genetic homologues can remain completely unclassified.

Of the algorithms listed, PhiSpy is unique in that it combines gene annotation with a search for "phage words." These phage words are 12-base pair sequences of DNA that appear only in the genomes of the prophages of their reference dataset. Approaches like this are uncommon in phage identification because of their inferior accuracy and that they reveal little insight into the function of genes in the phage's genome. This approach has the advantage of quantifying similarities among sequences when no high-similarity alignments exist. Another advantage of this approach is reduced computational expense as compared to sequence alignment methods [41] [42].

Because of these advantages, oligonucleotide content analysis is used extensively in computational biology. Past efforts have shown that tetranucleotide frequencies can be used to bin metagenomic contigs into groups that represent bacteria genomes and to assign taxonomy to these bacteria [43]. Similarly, *k*-mer frequencies in 16S sequences can be used to predict bacterial taxonomy [44] [45]. Oligonucleotide analysis has also been applied to virology. For instance, it has been shown that *k*-mer frequencies of a phage's genome are predictive of the phage's host. [4] [46] [47]. Other research has shown that tetranucleotide frequencies can serve as a distinguishing pattern of phage genomes, likely due to the specificity of viral hosts and the evolutionary pressure that viral genomes experience [47] [48] [49] [50].

VirSorter is a phage prediction algorithm that was published in 2015. We decided to use VirSorter in conjunction with PhaMers in this work because it builds upon previous algorithms (Prophinder), and has demonstrated superior performance to all of the algorithms listed (with the exception of Prophage Finder, which was not compared). VirSorter was useful to us because it

uses sequence alignments methods that produce gene annotations that could be visualized, externally validated, and used to understand phage morphology and function.

Machine Learning

In this work we made use of several methodologies in the field of machine learning (ML), which we will review briefly before beginning. ML is a field focused on how we might instruct machines to make decisions. In the case of this thesis, the relevant decision to be made is whether a particular sequence of DNA is that of a phage. ML is largely divided into two categories: supervised learning, in which a dataset of pre-classified data is available, and unsupervised learning, in which no such dataset is available. PhaMers would be classified as a supervised learning algorithm because it makes use of reference datasets of DNA sequences each of which is known with certainty to be either a phage or bacteria. In the study of our datasets and in the PhaMers scoring algorithm, we used the unsupervised techniques of clustering, specifically the *k*-means and DBSCAN algorithms. These algorithms assign subsets of data points to a series of "clusters" based on a metric of dissimilarity between points. The result is a classification of data such that points within a cluster are said to be more similar to each other than to points in other clusters. The *k*-means algorithm accomplishes this by randomly setting the coordinates of *k* (user-specified) cluster centroids, iteratively moving them by assigning points to the closest centroid, and setting new centroids as the mean of the points assigned to each cluster. The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm [51] assigns clusters by looking for groups of points all of which have at a sufficient number of points (minPts, user-specified) within some distance (epsilon, user-specified). Since some points may not exist in such a group, not all points may be assigned to a cluster by DBSCAN (unlike *k*-means). Finally, we used t-SNE (t-Distributed Stochastic Neighbor Embedding), a dimensionality reduction algorithm, to visualize tetranucleotide frequency points in $4^4 = 256$ dimensions. The t-SNE algorithm, unlike common dimensionality reduction algorithms such as PCA (Principal Component Analysis), does not try to preserve distances between the high and low dimensionality points. Rather, t-SNE aims to preserve probabilities that two points are considered to be "neighbors." By using a different probability distribution in the high and low dimensionality space (Gaussian and Cauchy, respectively), t-SNE moves adjacent points

together, while pushing dissimilar points apart. The result is non-deterministic, two- or three-dimensional embedding of data points that facilitates visualization and clustering.

Materials and Methods

Reference database generation

In order to construct and test supervised learning algorithms, we assembled a dataset of DNA sequences that were known to represent phages and another dataset of sequences that were known to be from bacteria. The reference dataset of genomic phage sequences was assembled using the phages available as of October of 2015 on RefSeq. A complete list of all viral accession numbers, made available on NCBI, was downloaded and used to find accession numbers for all viruses that infect bacteria or archaea. This set of accession numbers was used to access and compile a set of all phage sequences in fasta format, which was subsequently used for tetranucleotide frequency analysis. The reference dataset of bacterial genomic sequences was generated from genomic assemblies available on GenBank. Bacterial species were selected at random from the sub-directories at ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/, and the latest genomic assembly fasta files were used for analysis in PhaMers. Because we were also interested in the taxonomic classification of phages, we used BioPython and Entrez [52] to automate the retrieval of each phage's taxonomic classification from the NCBI database.

We wrote code in Python that uses a sliding window approach to count and store the occurrences of each of the 256 tetranucleotide sequences in each of the 2255 phage and 2255 bacterial genomic sequences. Because bacteria fasta files contained multiple sequences, we summed the tetranucleotide frequencies of all sequences in each bacteria fasta file.

PhaMers Development

Code and algorithms used by PhaMers were tested in MATLAB and implemented in the Python 2.7 programming language. Python 2.7 was used to write scripts for parsing VirSorter and IMG output files and for integration with PhaMers data. All PhaMers scripts are available at https://github.com/jondeaton/PhaMers. The Python library Matplotlib was used for plot generation. We automated the visualization of contig gene annotations with the

dna_features_viewer 0.1.0 (pypi.python.org/pypi/dna_features_viewer/0.1.0) Python package. We also used SnapGene to visualize long contigs.

PhaMers classification of known phage and bacterial genomes

While developing and verifying PhaMers' predictive discrimination between phage and non-phage genomic sequences, we tested the discriminatory power of supervised learning algorithms on our reference dataset. To do so, we normalized each tetranucleotide count vector by the total number of tetranucleotides counted to produce tetranucleotide frequency vectors, thereby discrediting differences in vectors due to sequence length. We then performed 20-fold cross-validation on different scoring algorithms, wherein we divided both the phage and bacterial datasets into subdivisions and scored each subdivision with the remaining nineteen subdivisions as training data.

We tested the following supervised learning algorithms: Support Vector Machine, Kernel Density Estimation, K-Nearest Neighbors (KNN), and Nearest Centroid, as well as linear combinations of results from each (Fig. 5a). KNN was chosen as the primary classifier because it performed with the lowest false positive rate while maintaining >90% sensitivity. We varied the parameter specifying the number of neighbors used in KNN classification (K) from 3 to 20 during cross-validation. Increasing K yielded marginally decreased performance, hence informing our choice of K=3. To add additional information into the final PhaMers score, we took the initial classification by KNN to be -1 (non-phage) or 1 (phage) and added to it a parameter between -1 and 1 that quantified the proximity of a point to phage clusters and distance away from bacterial clusters (Nearest Centroid).

Sample Collection

Prior to testing PhaMers on metagenomic datasets, environmental samples from three different hot springs in Yellowstone National Park were collected, stored, prepared, sequenced, and assembled. Sample #1 was collected from sediments of the Bijah Spring in the Mammoth Norris Corridor area. Sample #2 was collected from sediment near Mound Spring in the Lower Geyser Basin region. Sample #3 was collected from a spring in Mammoth Geyser Basin. During

collection, samples were placed in 2 mL tubes and soaked in 50% ethanol. Samples were transferred to -80˚C for long-term storage upon return.

Prior to sequencing, environmental samples were processed using the microfluidic-based mini-metagenomic protocol based on the Fluidigm C1 Auto Prep IFC (Integrated Fluidic Circuit). Each sample was thawed on ice and subsequently vortexed briefly to re-suspend cells and not large particles and debris. 1 mL of sample from the top of the tube was transferred to a new 1.5 mL tube and centrifuged at 5000 × g for 10 min to pellet cells. Supernatant was removed and cells were re-suspended in 1% NaCl. Following resuspension, we performed microscopy to quantify cell concentration. We diluted samples in 1% NaCl or PBS to a concentration of ~2×10^6 cell/mL so that each chamber of the Fluidigm C1 microfluidic IFC would contain approximately 10 cells.

To reduce the chance of MDA amplification of DNA contaminants, we treated the C1 microfluidic chip, all tubes, and buffers with ultraviolet irradiation (Strategene) for 30 min. We did not treat reagents containing enzymes, oligonucleotides, or dNTPs. After treatment, the C1 IFC was primed with a standard protocol (Fluidigm). The diluted samples containing bacterial cells were loaded into the chip using a modified protocol that excludes the washing step, as the capture sites in the C1 chip are too large to capture and hold single bacterial cells. With this modification, each capture chamber acted as a volume of liquid throughout which approximately 5 to 10 bacterial cells were suspended, which would subsequently would be prepared for sequencing as a single sample.

Following cell loading, whole genome amplification was performed with MDA in the 96 reaction chambers of the C1 chip. To facilitate the lysis of gram-positive bacteria, we added a lysozyme (Epicenter) digestion at 37˚C for 30 minutes prior to alkaline denaturation of DNA. Denaturation was performed at 65˚C for 10 minutes. We then performed neutralization followed MDA (Qiagen REPLI-g single cell kit) for 2 h 45 min at 30 °C. We adjusted all reagent concentrations to match the 384 well plate-based protocol developed by the single-cell group at DOE's Joint Genome Institute but adapted for volume of the Fluidigm C1 IFC.

Amplified genomic DNA from each of the 96 sub-sample chambers was harvested from the C1 chip into a 96-well plate. We used a high-sensitivity large fragment analysis kit (AATI) to quantify the oligonucleotide concentrations of each sub-sample. Quantification results were used to dilute each sample to 0.1–0.3 ng/μL so as to be in range for the Nextera XT library preparation pipeline. Nextera XT V2 libraries (Illumina) with dual sequencing indices were prepared, pooled, and purified with 0.75 volumes of AMPure beads (Agencourt). Each library pool was sequenced on an Illumina NextSeq (Illumina) with 2x150 bp sequencing runs.

Sequencing reads were filtered with Trimmomatic V0.30 in paired end mode with options "ILLUMINACLIP:adapters.fa:3:30:10:3:TRUE SLIDINGWINDOW:10:25 MAXINFO:120:0.3 LEADING:30 TRAILING:30 MINLEN:30" to remove possible occurrences of Nextera indices and low quality bases. Filtered reads from each sub-sample were clustered using DNACLUST [53], with k=5 and a similarity threshold of 0.98, to remove reads from highly covered regions and thereby reduce computational expense of genomic assembly. We performed read assembly using SPAdes V3.5.0 [54] with the single-cell and careful flags asserted. To assemble genomic fragments of the same species found in different chambers, corrected reads from each sub-sample were extracted combined, and re-assembled via SPAdes V3.5.0 with *k*-mer values of 33, 55, 77, and 99.

PhaMers classification of metagenomic contigs

PhaMers' scoring algorithm was used to score assembled metagenomic contigs longer than 5 kilobase-pairs (kbp). PhaMers uses BioPython to parse fasta formatted files of assembled contigs, and tabulates tetranucleotide frequencies before scoring. Results are saved for subsequent analysis in a human-readable format. To assign putative taxonomic classifications to phages, we used *k*-means (*k*=86) to cluster the tetranucleotide frequencies of each putative phage with those of the phages from the reference data set, and examined the taxonomic composition of the phages in the cluster that each contig was assigned to. We considered a contig to have been assigned to a taxonomically enriched cluster if phages from a single taxon compose a proportion of the cluster greater than 50% and is statistically significantly greater than the proportion that that taxa represents in the entire dataset. We considered it evidence of a putative phage's taxonomic classification if a phage was assigned to an enriched cluster and if its silhouette was

within one standard deviation of the mean of the cluster silhouettes of the reference phages in the cluster.

VirSorter analysis of metagenomic contigs

Samples were analyzed using the VirSorter 1.0.3 phage identification pipeline available through the iPlant Discovery Environment on the iPlant collaborative website, made available by CyVerse. (https://de.iplantcollaborative.org/de/) We set VirSorter to use all bacterial and archaeal virus genomes in Refseq as of January 2014 for the analysis of all metagenomic datasets.

Annotation of putative phage contigs

To create functional gene annotations of putative phage sequences, we uploaded metagenomic datasets to JGI's Integrated Microbial Genomes Expert Review online database (IMG/ER). Annotation was performed via IMG/ER [55]. Structural annotations were performed to identify CRISPRs (pilercr) [56], tRNA (tRNAscan) [57], and rRNA (hmmsearch). Protein coding genes were identified with a four ab initio gene prediction tools: GeneMark [58], Prodigal [59], MetaGeneAnnotator [60], and FragGeneScan [61]. Functional annotation was achieved by associating protein coding genes with COGs, Pfams, KO terms, and EC numbers. A phylogenetic lineage was assigned to each contig based on gene assignment.

Results

Taxonomic Analysis of Reference Dataset

Our preliminary results come from studying the relationship between tetranucleotide frequencies and phage taxonomy using the reference dataset of known phage sequences. In this pursuit, we reduced the dimensionality of our reference dataset of phage tetranucleotide frequency vectors from 256 to two using t-SNE (Fig. 3a). We then clustered the reduced dimensionality tetranucleotide frequency vectors using DBSCAN and quantified the prevalence of each taxa in each cluster. Given that DBSCAN labels some points as noise, only about 60% of the phages in the dataset were assigned to a cluster.

Through altering the algorithmic parameters under which phages are assigned to clusters, we

revealed similarities among taxa. The algorithmic parameters implicated in cluster assignment include perplexity values used in t-SNE, minimum number of points per cluster and epsilon values used in DBSCAN, preset number of clusters (*k*) in *k*-means clustering, choice of clustering algorithm (*k*-means or DBSCAN), and whether clusters were assigned before or after dimensionality reduction with t-SNE. In general, strict cluster assignment, such as using DBSCAN on reduced dimensionality data with a low epsilon value (epsilon = 1.5), resulted in the enrichment of many clusters with a single taxon at the Baltimore, Order, and Family levels (Fig. 3b-d). More liberal cluster assignment, such as with *k*-means on raw tetranucleotide frequency vectors, resulted in fewer enriched clusters and greater homogeneity in the distributions of taxa across clusters (Fig. 4a-d). These clusterings revealed similarities in tetranucleotide frequencies between phages of different taxa. For instance, *k*-means clustering (*k* = 40) raw tetranucleotide frequencies resulted in the assignment of the dsDNA *Cellulophaga phage phiSM*, the dsDNA *Lactococcus phage 936 sensu lato*, and *Skunalikevirus* to the same cluster (cluster 10). This result indicates that phages of these taxa have similar genomic tetranucleotide frequencies. Many clusters of phages are enriched for a single phage taxon, which indicates that an unidentified phage could be classified taxonomically by examining the taxa of the phages in the cluster that it was assigned to.

Single-stranded RNA (ssRNA) viruses primarily occupy two clusters, owing to their composition of the genera *Allolevirus* and *Levivirus*. These genera are generally classified as *Enterobacteriophge MS2* and *Enterobacteriophage Qß*, respectively [62]. Using *k*-means (*k* = 40) clustering on tetranucleotide frequencies resulted in the assignment of both taxa to cluster 8 (Fig. 4b-d). Generally, relaxing clustering parameters resulted in phages of these taxa occupying the same cluster, whereas more stringent clustering resulted in assignments to distinct clusters. This result indicates that *Allolevirus* and *Levivirus* each have signature tetranucleotide frequencies yet are more similar to each other than to most other phages. Single-stranded DNA phages are another taxon which tends to forms enriched clusters. The biggest of these clusters is formed due to the abundance of the well-characterized 5.4 kbp *Enterobacteria phage phiX174 sensu lato*. The second largest of these clusters is enriched with the 5.5 kbp *Enterobacteria phage G4 sensu lato*.

By studying the tetranucleotide frequencies of known phages, we have shown that tetranucleotide frequencies can predict a phage's taxonomy. Enrichment is predominantly caused by the abundance of phage genomes with high similarity in the dataset of phages from RefSeq. In this dataset, there are many sequences from the same species in duplicate, but only for a handful of base pairs. For certain taxa there are only one or a few sequences. For instance, this dataset contains only 24 double-stranded RNA (dsRNA) phages, which only contain the species of *Pseudomonas phage Φ8* and *Φ6*, both tending to occupy in the same cluster. Similarly, the small number of ssRNA and ssDNA phages generate few clusters generally free of phages from other taxa. These clusters likely form due to the high global sequence similarity of phages with these taxa, rather than an over-arching similarity in tetranucleotide frequencies. For these reasons, the utility of tetranucleotide frequencies in predicting the taxa of phages is limited by the comprehensiveness of our reference datasets of phage genomes.

PhaMers Algorithmic Performance

In addition to differentiating phage sequences from each other, tetranucleotide frequencies can differentiate genomic fragments of phages from those of bacteria with high accuracy. This is useful because the abundance of bacterial genomic fragments in metagenomic datasets often hampers discovery of novel viral genomes. We sought to understand how tetranucleotide frequencies may best distinguish phage sequences from bacterial sequences. To do so, we developed a machine-learning algorithm that compares tetranucleotide frequencies of metagenomic contigs to those of phages from RefSeq and bacteria from GenBank, used as ground truth. We observed that a nearest neighbors approach (KNN) yielded a sensitivity above 90% with a false positive rate below 1% when tested with 20-fold cross-validation (Fig. 5a). We chose a linear combination of KNN and a cluster centroid proximity metric to use for phage identification in PhaMers. The PhaMers score S_P for a tetranucleotide frequency vector $\boldsymbol{v}$ is given by

$$S_P(v) = KNN_3(v) + C_P(v)$$

where the *KNN* ($K = 3$) score is given by

$$KNN_3(v) = \begin{cases} 1, & N > K/2 \\ -1, & N < K/2 \end{cases}$$

and N is the number of tetranucleotide frequency vectors of the K closest to $\boldsymbol{v}$ that are phages, and the cluster proximity metric C_p is given by

$$C_P(v) = \tanh \frac{R_b - R_p}{R_b + R_p}$$

with R_b and R_p given by the Euclidian distance from the nearest cluster centroid of bacteria and phages, respectively, as assigned by *k*-means clustering. PhaMers gives scores that are positive for predicted phages and negative otherwise (Fig. 5b). We chose this combination because the cluster proximity metric performed well on its own and quantifies relative distances to large groups of reference data, whereas KNN classifies based on more local data points. This combination adds additional information the score, and increases the area beneath the Receiver Operating Characteristic (ROC) curve. Phages that were classified correctly are generally those which are well-studied and heavily represented in the dataset. PhaMers performed well when tested on these datasets with 20-fold cross-validation, yielding 91.8% sensitivity, 99.3% specificity, and 99.2% positive predictive value (PPV).

PhaMers easily classifies phages of the taxa of *Enterobacteria phage T4*, T4-like and T7-like viruses, *Propionibacterium phage*, and *Lactococcus phage ASCC.* The phages which received negative scores (misclassified) were often from the taxa *unclassified Siphoviridae*, *unclassified Podoviridae*, or *unclassified Myoviridae*. These phages were misclassified because there are few phages with genetic similarity to them in the dataset. To discount the effect of overrepresented phages on PhaMers' performance estimation, we repeated cross-validation with a reduced dataset of phages that contains only one phage of each taxon. Tetranucleotide frequencies distinguished this set phages from bacteria with only marginally decreased accuracy (Fig. 6a-b).

We also assessed how sequence length affects PhaMers' performance. We performed 20-fold cross-validation with random subsequences of the genomes in the reference dataset. We observed that the area under the curve (AUC) of the ROC curve dropped most significantly as

sequences were cut to lengths smaller than 5 kbp (Fig. 5c). This result informed our choice of 5 kbp as the sequence length requirement for metagenomic contigs to be scored with PhaMers.

Analysis of Metagenomic Datasets

We analyzed our three sets of assembled metagenomic contigs from YNP with the VirSorter phage prediction pipeline and scored all contigs longer than 5 kbp with PhaMers. VirSorter identified putative phages in 30 contigs from Bijah Road Side and 108 from Mound Spring. Six of the phages from Mound Spring dataset were identified as potentially prophage. Of the 5594 contigs longer than 5 kbp, 845 received positive PhaMers scores. Of the 23 and 89 contigs in the Bijah Spring and Mound Spring datasets identified as potentially viral by VirSorter, 10 and 52 were given positive PhaMers scores, respectively. In the Bijah Spring and Mound Spring datasets, 257 and 526 contigs, respectively, received positive PhaMers scores which were not labeled as phages by VirSorter. In these datasets, 1663 and 3036 contigs, respectively, were not labeled phages by VirSorter or PhaMers. These results suggest that PhaMers and VirSorter are in large part consistent with predictions made for the majority of the contigs. In addition, we noted that contigs identified as phages by VirSorter tended to receive higher scores than those which were not (Fig. 7c, 8). Interestingly, we noticed that the average PhaMers score sometimes decreased with increasing confidence of classification by VirSorter. That many contigs received positive PhaMers scores that were not labeled as phages by VirSorter indicates that PhaMers may identify phages that VirSorter does not.

Algorithmic Results Comparison

Although PhaMers performed well when cross-validated with reference datasets, we needed to explore how much its predictive power extended to metagenomic datasets. We did this by comparing its predictions with those made by VirSorter. Under the assumption that all the putative phages classified with high confidence (categories 1, 2, 4, or 5) by VirSorter are correct classifications but with some missed phages, let X_{TN} be the number of phages that both PhaMers and VirSorter missed, and X_{FP} be the number of phages missed by VirSorter, but that PhaMers identified. The actual number of true positives (*TP*), true negatives (*TN*), false positives (*FP*), and false negatives (*FN*) in PhaMers' predictions are then given by

$$TP = TP_c + X_{FP}$$
$$TN = TN_c - X_{TN}$$
$$FP = FP_c - X_{FP}$$
$$FN = FN_c + X_{TN}$$

where TP_c, TN_c, FP_c, and FN_c are the true positives, true negatives, false positives, and false negatives calculated as though VirSorter predictions are truth. With these equations, PhaMers' accuracy (ACC), may be estimated if it may be assumed that $|X_{TN} - X_{FP}|$ is small as compared to $TP_c + TN_c$, and is given by:

$$ACC = \frac{TP_c + TN_c + X_{FP} - X_{TN}}{TP_c + FP_c + TN_c + FN_c}$$

When comparing PhaMers to VirSorter with this analysis, PhaMers performed with ~85% accuracy on both metagenomic datasets. This result indicated that PhaMers' predictive power extends beyond classification of the reference datasets and can make accurate predictions on metagenomic datasets. To further validate this hypothesis, we created an ROC curve from PhaMers' predictions based on VirSorter predictions (Fig 7b). Although VirSorter identifications are not ground truth, the specificity and sensitivity values in this ROC curve are estimates that provide a sense of the predictive performance of PhaMers. This ROC curve further confirms that PhaMers maintains predictive power on metagenomic datasets.

Identification of Novel Phages

Our threshold for considering a contig to count as a putative phage was a high-confidence annotation by VirSorter (categories 1, 2, 4, or 5) or a low confidence annotation by VirSorter (categories 3 and 6) with a positive PhaMers score. Given these criteria, 19 contigs from Bijah Road Side, and 83 contigs from Mound Spring are predicted to be phages. We also analyzed a single contig from a metagenomic dataset generated from samples taken from Mammoth Geyser Basin (Sample #3). This single contig was identified as a phage through the use of PhaMers and protein sequence alignment search (BLASTP), but not VirSorter.

To deepen our analysis of these novel phages, we used t-SNE to visualize the tetranucleotide frequencies of all contigs in each dataset with those of the phages and bacteria in the PhaMers reference dataset (Fig. 7a, 9-10). This visualization reveals that contigs tend to form tight clusters, which are hypothesized to be fragments of the same or similar bacterial genomes. We observe that many putative phage contigs predicted by VirSorter and PhaMers lie in proximity to clusters of known phage genomes in these embeddings.

For the Bijah Spring and Mound Spring datasets, we also used the Integrated Microbial Genomes and Microbiomes (IMG) gene annotation pipeline to make gene predictions for all contigs. These gene annotations validated our identifications of viral contigs by confirming the presence of viral marker genes with an independent method. Many gene annotations concurred with those given by VirSorter, though for many contigs, IMG identified additional genes which VirSorter did not. This is likely due to the differences in genomic datasets used for gene annotation by IMG and VirSorter. These additional gene annotations and their associated phylogenies from IMG further characterized these novel phage sequences.

Taxonomic Predictions for Novel Phages

To predict phage taxonomy from tetranucleotides, we used *k*-means clustering to cluster known phage tetranucleotide frequency vectors with those of novel phages. We looked for novel phages that were assigned to clusters enriched with reference phages of a single taxon. Clusters were labeled as enriched for a taxon if phages of that taxon constituted more than half of the cluster and their prevalence was significantly greater than the proportion that the taxon represents in the reference dataset (Fig. 1a-c). If a contig met these criteria and had a positive silhouette value within a standard deviation of the mean cluster silhouette value, we took this as evidence for the taxonomy of the phage. Of the 24 contigs (five from Bijah Spring and 18 from Mound Spring, one from Mammoth Geyser Basin) that met these criteria, most were assigned to clusters enriched by phages of the taxa *Siphoviridae, Podoviridae,* or *Myoviridae* (Fig. 12-17).

Discussion

Discussion of Novel Phages

Several of the predicted phage sequences are particularly interesting due to their predicted taxonomy and proteomic phylogeny. A 13,504 base pair contig (Contig 1753) from the Mound Spring dataset was predicted to be a phage by VirSorter (category 2, "quite sure") and received a score of 1.14 by PhaMers (Fig. 11a). When analyzed for tetranucleotide similarity to known phages, this contig was assigned to a cluster of known phages enriched with *Siphoviridae* (Fig. 11b-c). Given this analysis and that the contig contains predicted phage tail and portal proteins (annotated by both VirSorter and IMG), we predict that this contig is a genomic fragment of a phage that would be classified as *Siphoviridae*. Phylogenetic predictions of coding regions on this contig made by IMG revealed that 15 of the 19 putative proteins have greatest homology to genes from *Clostridiales*. These findings suggest that Contig 1753 is fragment or complete genome of a *Siphoviridae* that infects *Clostridiales*, or that many of the phages with proteomic similarity to this contig in the IMG database come from the genomes of *Clostridiales*.

Another contig (Contig 677) of length 20,664 base pairs is interesting due to its enrichment of genes homologous to those of the thermophilic bacteria *Hydrogenobacter* (Fig. 11e). This contig was identified as a phage by VirSorter (category 2, "quite sure") and was given a score of 0.97 by PhaMers. This contig contains 36 putative coding regions, 33 of which were predicted by IMG to have greatest homology to *Hydrogenobacter* of the phylum *Aquificae.* This gene enrichment and that this contig was found in a sample taken from a hot geyser spring suggest that Contig 677 is a genomic fragment from a phage that infects thermophilic bacteria. The enrichment for genes homologous to those of *Hydrogenobacter* is either an indication that *Hydrogenobacter* is this phage's natural host or that there is an overrepresentation of *Hydrogenobacter* genes in the IMG database similar to those in this contig. Putative coding regions on this contig include transposase, integrase, a phage protein of unknown function, and predicted transcriptional regulator. Many of the putative coding regions on this contig are unidentified, indicating that many of this phage's genomic functions remain unknown.

One of the longest contigs (Contig 15) that was identified as putative phage by VirSorter is 136.9 kbp long (Fig. 11d). This contig was classified as a category 2 ("pretty sure") putative phage by VirSorter due to its enrichment for viral hallmark genes. These genes include terminase-like coding regions, phage baseplate J, and baseplate assembly genes. Other viral proteins on this

contig include a major capsid, T4-like capsid assembly protein, and viral prohead core protein protease, which are involved in the formation of the mature head coating (capsid) [63]. Additionally, this phage contains many proteins that interact with DNA, including 5' to 3' endonuclease, Superfamily II DNA/RNA helicase, DNA Polymerase Elongation Subunit, Recombination Endonuclease VII, and a homing endonuclease. This contig also contains 10 tRNA genes. Interestingly, this contig did not receive a positive PhaMers score. We hypothesize that this is due to the abundance of tRNA and DNA binding proteins such as DNA polymerases which although often seen in phages, are more characteristic of bacterial genomes.

Finally, in the sample taken from Mammoth Geyser Basin, we characterized a single novel viral contig of length 35,211 bp (Fig. 12). This contig was not identified as phage by VirSorter but was given a score of 0.90 by PhaMers, and an NCBI BLAST nucleotide and protein sequence similarity search revealed similarities to the thermophilic archaea phage genera *Sulfobus filamentous* and *Acidianus filamentous*. These phages are of the *Betalipothrixvirus* genus, of the family *Lipothrixviridae* [64]. Interestingly, this contig was assigned, on the basis of tetranucleotide frequencies, to a cluster enriched with phages of the species *Lactococcus phage 936 sensu lato,* which are *unclassified Siphoviridae.* Protein BLAST searches revealed a 596-amino acid putative protein with 73% identity to a Holiday junction branch migration helicase from *Acidianus filamentous virus* 9. This search also found a 563 amino acid putative protein with 72% identity to a helicase from *Acidianus filamentous virus* 9. In addition, a 1038-amino acid putative protein on this contig has 48% identity to the 1349-amino acid conserved hypothetical protein of *Acidianus filamentous virus 3*. This feature does not appear in the genome of *Sulfobus filamentous* [64]. This feature indicates that this phage has greater genetic similarity to *Acidianus filamentous* than to *Sulfobus filamentous*. Additionally, this contig has homologous tail and baseplate proteins to those of *bacteriophage T4*. These results suggest that this contig represents the genome of a thermophilic phage of the *Betalipothrixvirus* genus that should be classified as a new species of *Acidianus filamentous*.

PhaMers Advantages and Limitations

The advantages of PhaMers are that it requires little computation and scales easily with increased knowledge of phage genomic diversity. The utility of tetranucleotide frequencies (and therefore

PhaMers) for phage identification in metagenomic sequences is determined by the comprehensiveness of our reference datasets of genomic material from phages. As reference datasets of phages grow to include a more comprehensive representation of global phage diversity, tetranucleotide frequency analysis could be used for rapid identification and taxonomic classification of phages.

PhaMers has several limitations. First, PhaMers is limited in its ability to classify short contigs. This is because as contigs become short, tetranucleotide counts become sparse and contain little information about the contig. Also, given that PhaMers uses supervised learning, its results are sensitive to the choice of reference data used for scoring. Since the genomic datasets were chosen at random from GenBank in this study, optimization of the reference dataset could yield improved performance. Another limitation of PhaMers is that it does not provide any insight about the biologic functions of identified phages. For this reason, we suggest that PhaMers be used in conjunction with gene annotation tools.

One limitation of this work is that the accuracy with which tetranucleotide frequencies identify phages in metagenomic sequences is overestimated by cross-validation. This is due to the repetitive characteristics of the reference dataset of phages. Though we performed cross-validation on a reduced dataset to minimize these effects, the abundance of highly similar phage genomes still allows for easy classification of those sequences during K-fold cross-validation. This results in an overestimated sensitivity and specificity that may not extend to metagenomic datasets.

Conclusions

Phages are unique biologic entities that play important roles in microbial communities. Furthering our understanding of phages not only improves our general understanding of ecology, but also helps solve problems in bioengineering and medicine. Metagenomics is a primary methodology for studying phages which cannot be cultured. This work has not only used metagenomics to discover novel bacteriophages, but also produced a novel algorithm and computational tool for this purpose.

Our results have demonstrated the utility of the novel bio-computational tool PhaMers and the Mini-Metagenomics protocol in the identification and characterization of novel phages. Through this work, we have identified and characterized a total of 103 novel genomic fragments of phages. As a secondary result, we have demonstrated how tetranucleotide frequencies can be used to predict the taxonomy of phages. Another secondary result has been to show that t-SNE can be a useful visualization tool to compare entire metagenomic datasets with genomic reference datasets.

Future Work

Future work should investigate the metabolic impact that genes found on these novel phages might have on the microbial community in Yellowstone hot springs. Future research should also aim to characterize the structures and functions of hypothetical coding regions on viral contigs. Given that these phages exist in thermophilic environments, interesting results may come from comparing genes found on novel phages to their non-thermophilic homologues. This investigation may elucidate the amino acid changes required for successful adaptation to thermophilic environments.

With regard to PhaMers, future work should integrate PhaMers into gene annotation-based viral prediction algorithms like VirSorter. The incorporation of PhaMers and other phage identification tools into one comprehensive algorithm would produce a tool with performance superior to that of any that currently exist. We also suggest that more research be done to determine the optimal set of non-phage reference genomes that are used in PhaMers. Similarly, future work should expand and optimize PhaMers' set of reference phages. Since the beginning of our research, many new phage genomes have been identified. These could be incorporated into and improve PhaMers. In addition, the reference dataset of phages was compiled with phages from RefSeq; however, the Joint Genomics Institute has online metagenomic datasets, many with viral fragments that could also be included in the reference dataset of PhaMers.

PhaMers is written entirely in Python. While this choice was made to facilitate research and development, the computational performance of PhaMers could be improved considerably. First, PhaMers counts *k*-mers using a single-process function that performs orders of magnitude slower

than other *k*-mer counting software tools like Jellyfish [65]. PhaMers would be improved by importing and using the computationally efficient *k*-mer counting tools in Jellyfish. PhaMers' performance could also be improved by writing it in a more efficient programming language such as C++.

Finally, future research should identify the hosts of the novel phages identified here. This might be done by looking for CRISPR sequences found in bacterial contigs, analyzing structures of proteins that facilitate host-phage interaction, or through *k*-mer based clustering of metagenomic contigs. A phage located in a cluster of bacterial contigs from a single bacteria may infect that bacteria. As a final suggestion, future investigation should characterize the contigs labeled as phages by PhaMers, but which were not identified as viral by VirSorter. As with the case of the *Acidianus filamentous* phage, PhaMers may have identified many phages that VirSorter missed.

References

[1] F. B. Yu, P. Blainey, F. Schulz, T. Woyke and M. Horowitz, "Microfluidic-based mini-metagenomics enables discovery of novel microbial lineages from complex environmental samples," *Science,* 2016.

[2] S. Roux, F. Enault, B. L. Hurwitz and M. B. Sullivan, "VirSorter: Mining viral signal from microbial genomic data.," *PeerJ,* 2015.

[3] G. F. Hatfull, ""Bacteriophage Genomics"," *Current opinion in Microbiology,* pp. 447-453, 2008.

[4] R. A. Edwards, K. McNair, K. Fraust, J. Raes and B. E. Dutilh, "Computational approaches to predict bacteriophage–host relationships," *FEMS Microbiology Reviews,* 2015.

[5] J. C. Wooley, A. Godzik and I. Friedberg, "A Primer on Metagenomics," *PloS Computational Biology,* vol. 6, no. 2, 26 2 2010.

[6] R. A. Edwards and F. Rohwer, "Viral Metagenomics," *Nature Reviews Microbiology,* pp. 504-510, 2005.

[7] B. L. Hurwitz, J. M. U'Ren and K. Youens-Clark, "Computational prospecting the great viral unknown," *FEMS Microbiology Letters,* 2016.

[8] V. Trifonov and R. Rabadan, "Frequency Analysis Techniques for Identification of Viral Genetic Data," *mBio,* pp. 156-10, 2010.

[9] M. Victor M., C. I-Min A., P. Krishna, C. Ken, S. Ernest, P. Manoj, R. Anna, H. Jinghua, W. Tanja, H. Marcel, A. Iain, B. Konstantinos, V. Neha, M. Konstantinos, P. Amrita, N. N. Ivanova and N. C. Kyrpides, "IMG 4 version of the integrated microbial genomes comparative analysis system," *Nucleic Acids Research,* vol. 42, no. D1, 2013.

[10] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Reserach,* 2008.

[11] S. Chaturongakul and P. Ounjai, "Phage–host interplay: examples from tailed phages and Gram-negative bacterial pathogens," *Front Microbiol,* vol. 5, no. 442, 20 August 2014.

[12] C. K. Mathews, "Bacteriophage T4," *eLS. John Wiley & Sons Ltd, Chichester,* August 2015.

[13] G. Bertani, "Lysogenic Versus Lytic Cycle of Phage Multiplication," *Cold Spring Harb Symp Quant Biol,* vol. 18, pp. 65-70, 1953.

[14] S. Casjens, "Prophages and bacterial genomics: what have we learned so far?," *Molecular Microbiology,* vol. 49, pp. 277-300, 2003.

[15] C. Canchaya, G. Fournous and H. Brünssow, "The impact of prophages on bacterial chromosomes," *Molecular Microbiology,* vol. 53, pp. 9-18, 2004.

[16] P. Veiga-Crespo, J. Barros-Velázquez and T. G. Villa, "What can bacteriophages do for us?," *Communicating Current Research and Educational Topics and Trends in Applied Microbiology ,* 2007.

[17] R. Young, "Bacteriophage lysis: mechanism and regulation," *Microbiol Rev,* vol. 56, no. 3, p. 430–481, September 1992.

[18] F. Rohwer and R. Edwards, "The Phage Proteomic Tree: a Genome-Based Taxonomy for Phage," *J Bacteriology,* vol. 184, no. 16, pp. 4529-4535, August 2002.

[19] A. Fokine and M. G. Rossmann, "Molecular architecture of tailed double-stranded DNA phages," *Bacteriophage,* 21 Feb 2014.

[20] Ø. Bergh, K. Y. BØrsheim, G. Bratbak and M. Heldal, "High abundance of viruses found in aquatic environments," *Nature,* vol. 340, pp. 467-468, 10 August 1989.

[21] M. Middelboe, N. O. G. Jørgensen and N. Kroer, "Effects of Viruses on Nutrient Turnover and Growth Efficiency of Noninfected Marine Bacterioplankton," *Applied and Environmental Microbiology,* vol. 62, no. 6, pp. 1991-1997, June 1996.

[22] M. B. Sullivan, D. Lindell, J. A. Lee, L. R. Thompson, J. P. Bielawski and S. W. Chisholm, "Prevalence and Evolution of Core Photosystem II Genes in Marine Cyanobacterial Viruses and Their Hosts," *PloS Biol,* vol. 4, no. 8, August 2006.

[23] D. Lindell, J. D. Jaffe, Z. I. Johnson, G. M. Church and S. W. Chisholm, "Photosynthesis genes in marine viruses yield proteins during host infection," *Nature,* vol. 438, no. 3, November 2005.

[24] B. L. Hurwitz, S. J. Hallam and M. B. Sullivan, "Metabolic reprogramming by viruses in the sunlit and dark ocean," *Genome Biology,* vol. 14, 7 November 2013.

[25] A. R. M. Coates and Y. Hu, "Novel approaches to developing new antibiotics for bacterial infections," *Br J Pharmacol,* vol. 152, no. 8, pp. 1147-1154, December 2007.

[26] H. Bar, I. Yacoby and I. Benhar, "Killing cancer cells by targeted drug-carrying phage nanomedicines," *BMC Biotechnology,* vol. 8, no. 37, 3 April 2008.

[27] M. Skurnik, "Phage therapy: Facts and fiction," *International Journal of Medical Microbiology,* vol. 296, no. 1, pp. 5-14, 15 February 2006.

[28] K. Bush, P. Courvalin, G. Dantas, J. Davies, B. Eisenstein, P. Huovinen, G. Jacoby, R. Kishony, B. Kreiswirth, E. Kutter, S. Lerner, S. Levy, K. Lewis, O. Lomovskaya, J. Miller, S. Mobashery, L. Piddock, S. Projan, C. Thomas, A. Tomasz, P. Tulkens, T. Walsh, J. Watson, J. Witkowski, W. Witte, G. Wright, P. Yeh and H. Zgurskaya, "Tackling antibiotic resistance," *Nature Reviews Microbiology,* vol. 9, p. 894–896, 2011.

[29] F. Robrega, A. Costa, L. Kluskens and J. Azerdo, "Revisiting phage therapy: new applications for old resources," *Trends in Microbiology,* vol. 23, pp. 185-191, 2015.

[30] A. Chien, D. B. Edgar and J. M. Trela, "Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus," *J Bacteriol,* vol. 127, no. 3, pp. 1550-1557, September 1976.

[31] H. K. E. Landenmark, D. H. Forgan and C. S. Cockell, "An Estimate of the Total DNA in the Biosphere," *PLoS Biol,* vol. 13, no. 6, 11 June 2015.

[32] F. Sanger, A. R. Coulston, T. Friedmann, G. M. Air, B. G. Barrell, N. L. Brown, J. C. Fiddes, C. A. Hutchison III, P. M. Slocombe and M. Smith, "The nucleotide sequence of bacteriophage φX174," *Nature,* vol. 265, pp. 687-695, 24 February 1977.

[33] W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaertt, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van Den Berghe, G. Volckaert and M. Yserbaert, "Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene," *Nature,* vol. 260, pp. 500-507, 8 April 1976.

[34] L. Lond, L. Lond and M.C.R.S., "An Investigation on the Nature of Ultra- Microscopic Viruses," *The Lancet,* vol. 2, no. 4814, pp. 1241-1243, 4 December 1915.

[35] M. Breitbart, P. Salamon, B. Andersen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam and F. Rohwer, "Genomic analysis of uncultured marine viral communities," *PNAS,* vol. 99, no. 22, 16 October 2002.

[36] G. Lima-Mendez, J. V. Helden, A. Toussaint and R. Leplae, "Prophinder: a computational tool for prophage prediction in prokaryotic genomes," *Bioinformatics,* vol. 24, no. 6, 30 January 2008.

[37] S. Akhter, R. K. Aziz and R. A. Edwards, "PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies," *Nucl Acids Res,* vol. 40, no. 16, 14 May 2012.

[38] D. E. Fouts, "Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences," *Nucl Acids Res,* vol. 34, no. 20, 26 September 2006.

[39] B. D. Robert and M. Bose, "Prophage Finder: A Prophage Loci Prediction Tool for Prokaryotic Genome Sequences," *In Silico Biology,* vol. 6, no. 3, pp. 223-227, 2006.

[40] Y. Zhou, Y. Liang, K. H. Lynch, J. J. Dennis and D. S. Wishart, "PHAST: a fast phage search tool," *Nucl Acids Res,* vol. 39, 14 June 2011.

[41] R. Ounit, S. Wanamaker, T. J. Close and S. Lonardi, "CLARK: fast and accurate classification of metagenomic and genomic sequences usign discriminative k-mer," *BMC Genomics,* 2015.

[42] D. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments," *BioMed Central Genome Biology,* 2014.

[43] S. S. Mande, M. H. Mohammed and T. S. Ghosh, "Classification of metagenomic sequences: methods and challenges," *Briefings in Bioinformatics,* vol. 13, no. 6, p. 669681, 24 7 2012.

[44] N. Chaudhary, A. K. Sharma, P. Agarwal, A. Gupta and V. K. Sharma, "16S Classifier: A Tool for Fast and Accurate Taxonomic Classification of 16S rRNA Hypervariable Regions in Metagenomic Datasets," *PLoS ONS,* 2015.

[45] D. Papamichail, S. S. Skiena, D. Van Der Lelie and S. R. Mccorkle, "Bacteria Population Assay Via k-mer Analysis," 2004.

[46] J. Villarroel, K. A. Kleinheinz, V. I. Jurtz, H. Zschach, O. Lund, M. Nielsen and M. V. Larsen, "HostPhinder: A Phage Host Prediction Tool," *Viruses,* vol. 8, 2016.

[47] D. T. Pride, T. M. Wassenaar, C. Ghose and M. J. Blaser, "Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses," *BMC Genomics,* 2006.

[48] K. V. Srividhya, V. Alaguraj, G. Poornima, D. Kumar, G. P. Singh, L. Raghavenderan, A. V. S. K. Mohan Katta, P. Mehta and S. Krishnaswamy, "Identification of Prophages in Bacterial Genomes by Dinucleotide Relative Abundance Difference," *Plos ONE,* 21 November 2007.

[49] P. Deschavanne, M. S. DuBow and C. Regeard, "The use of genomic signature distance between bacteriophages and their hosts displays evolutionary relationships and phage growth cycle determination," *Virology Journal,* vol. 7, p. 163, 17 7 2010.

[50] D. T. Pride, R. J. Meinersmann, T. M. Wassenaar and M. J. Blaser, "Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases," *Genome Research,* vol. 13, pp. 145-158, 14 1 2003.

[51] M. Ester, H.-P. Kriegel and J. Sander, "A Density-Based Algorithm for Discovering Clusters," *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining,* 1996.

[52] NCBI Resource Coordinators, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Res.,* vol. 41, no. Database issue, 28 Dec 2013.

[53] M. Ghodsi, B. Lui and M. Pop, "DNACLUST: accurate and efficient clusteringof phylogenetic marker genes," *BMC Bioinformatics,* vol. 12, 2011.

[54] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. A. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev and P. A. Pevzner, "SPAdes: A New Genome Assembly Algorithmand Its Applications to Single-Cell Sequencing," *J Comput Biol,* vol. 19, no. 5, pp. 455-477, 19 May 2012.

[55] M. Huntemann, N. N. Ivanova, K. Mavromatis, J. H. Tripp, D. Paesespino, K. Tennessen, K. Palaniappan, E. Szeto, M. Pillay, I.-M. A. Chen, A. Pati, T. Nielsen, V. M. Markowitz and N. C. Kyrpides, "The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4)," *Stand Genomic Sci,* vol. 11, 24 Feb 2016.

[56] R. C. Edgar, "PILER-CR: Fast and accurate identification of CRISPR repeats," *BMC Bioinformatics,* vol. 8, no. 18, 20 January 2007.

[57] T. M. Lowe and S. R. Eddy, "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.," *Nucleic Acids Res.,* vol. 25, no. 25, pp. 955-964, 1 March 1997.

[58] J. Besmer and M. Borodovsky, "GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses John," *Nucleic Acids Research,* vol. 33, 20 April 2005.

[59] D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer and L. J. Hauser, "Prodigal: prokaryotic gene recognition and translation initiation site identification," *BMC Bioinformatics,* vol. 11, no. 119, 8 March 2010.

[60] H. Noguchi, T. Taniguchi and T. Itoh, "MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes," *DNA Res,* vol. 15, no. 6, pp. 387-396, 15 December 2008.

[61] M. Rho, H. Tang and Y. Ye, "FragGeneScan: predicting genes in short and error-prone reads," *Nucleic Acids Res.,* vol. 38, no. 20, p. 191, 30 August 2010.

[62] J. Manlioff and H.-W. Ackermann, "Taxonomy of bacterial viruses: establishment of tailed virus genera and the other Caudovirales," *Archives of Virology,* vol. 143, no. 10, pp. 2051-2063, Oct 1998.

[63] T. Dokland, "Scaffolding proteins and their role in viral assembly," *Cell Mol Life Sci,* vol. 56, no. 7-8, pp. 580-603, 15 Nov 1999.

[64] G. Vestergaard, R. Aramayo, T. Basta, M. Häring, X. Peng, K. Brügger, L. Chen, R. Rachel, N. Boisset, R. A. Garrett and D. Prangishvili, "Structure of the Acidianus Filamentous Virus 3 and Comparative Genomics of Related Archaeal Lipothrixviruses," *Journal of Virology,* vol. 82, no. 1, pp. 371-381, 17 Oct 2007.

[65] G. Marçais and C. Kingsford, "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers," *Bioinformatics,* vol. 27, no. 6, pp. 764-770, 7 January 2011.

[66] R. A. Edwards, K. McNair, K. Fraust, J. Raes and B. Dutilh, "Computational approaches to predict bacteriophage–host relationships," *FEMS Microiology Reviews,* 2015.

[67] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society,* vol. 36, no. 2, pp. 111-147, 1974.

[68] S. Sun, S. Gao, K. Kondabagil, Y. Xiang, M. G. Rossmann and V. B. Rao, "Structure and function of the small terminase component of the DNA packaging machine in T4-like bacteriophages," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 109, no. 3, pp. 817-822, 17 Jan 2012.

[69] P. G. Leiman, M. M. Shneider, V. A. Koxtyuchenko, P. R. Chipman, V. V. Mesyanzhinov and M. G. Rossmann, "Structure and Location of Gene Product 8 in the Bacteriophage T4 Baseplate," *Journal of Molecular Biology,* vol. 328, no. 4, pp. 821-833, 9 May 2003.

Figures

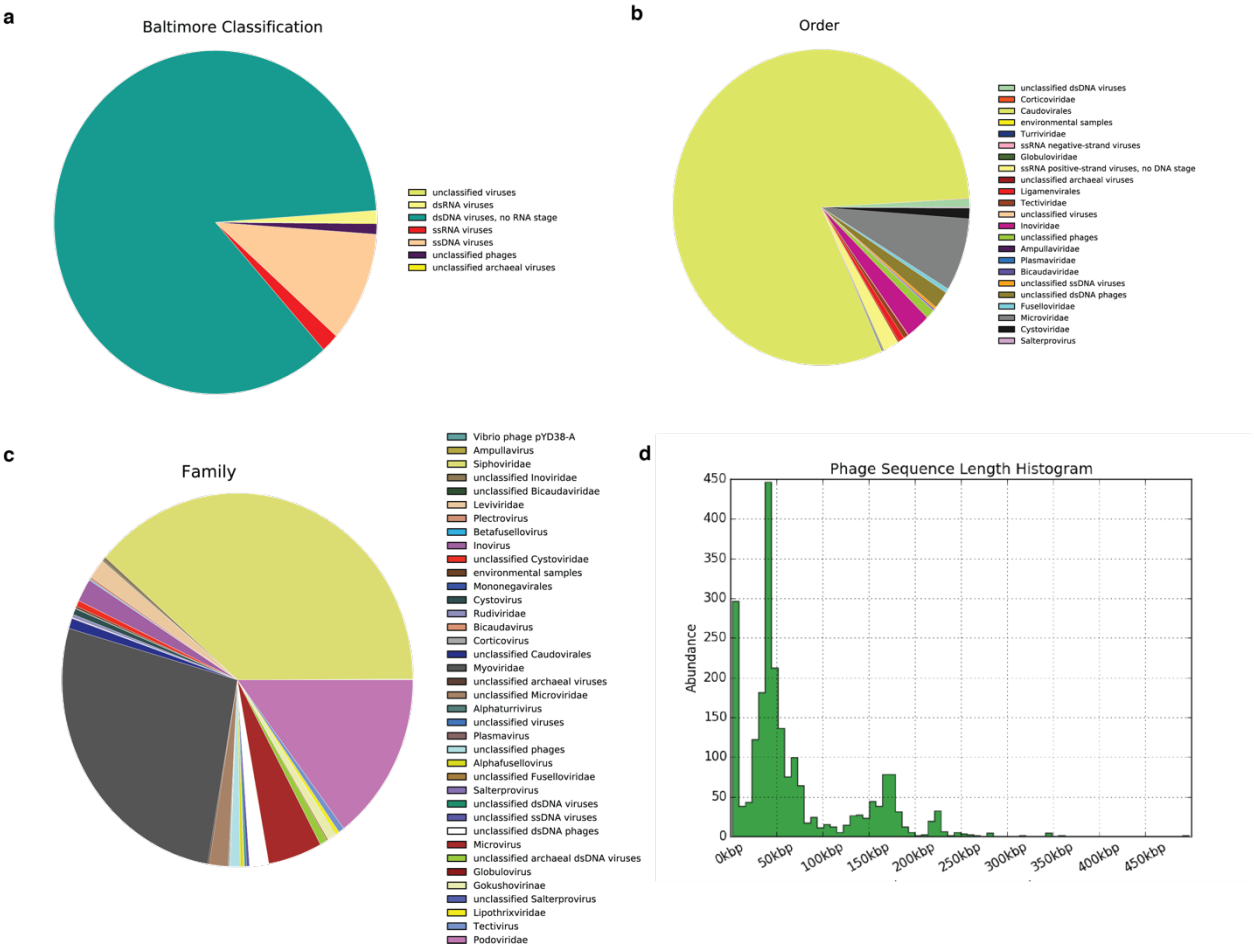


Figure 1: Phage reference dataset attributes

a, The proportion of bacteriophages with a given Baltimore classification in the phage reference dataset used by PhaMers. **b,** The proportion of bacteriophages in each order in the phage reference dataset. **c,** The proportion of bacteriophages in each family in the phage reference dataset. **d,** Histogram showing the distribution of genome size for the 2255 phages in the reference dataset.

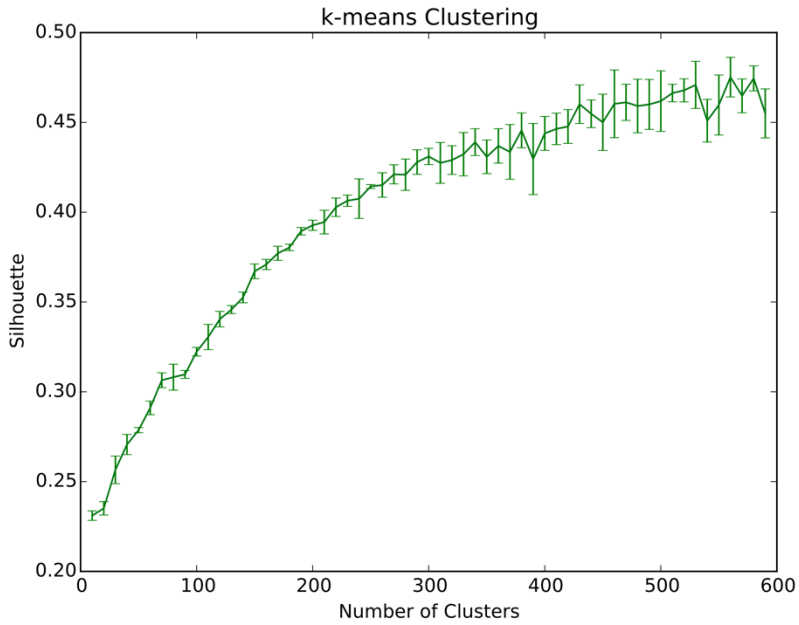


Figure 2: Phage tetranucleotide frequency clustering optimization

This plot shows the average cluster silhouette score for clusters made from tetranucleotide frequencies of phages from the reference dataset. Error bars show standard deviation of the average of all silhouette scores for all phages after clustering with *k*-means five times. This data was used to inform choice of *k* in learning algorithms.

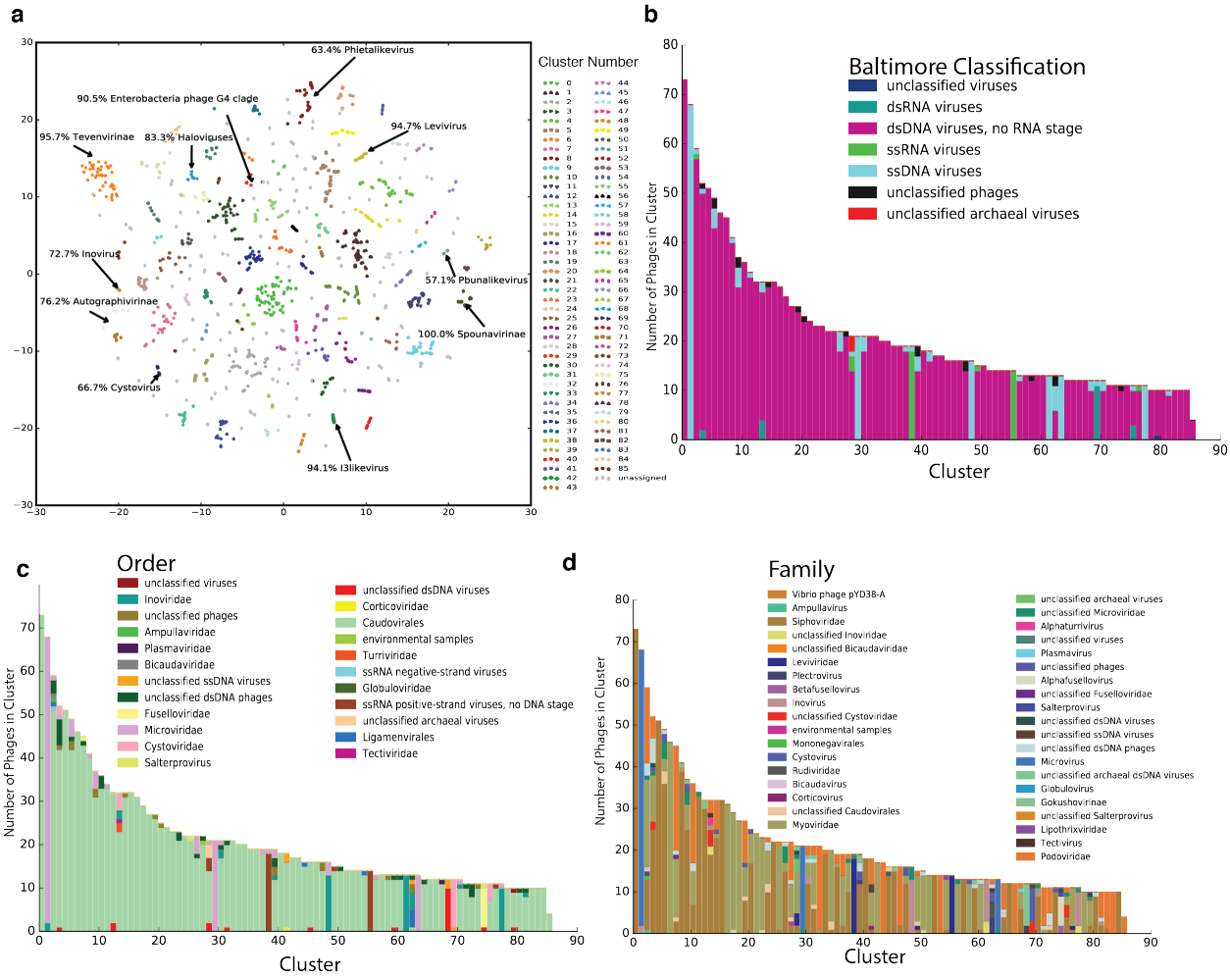


Figure 3: Phage tetranucleotide characteristics

a, t-SNE representation of tetranucleotide frequency vectors of 2255 phage genomic sequences from RefSeq. Clusters assigned with reduced dimensionality (2D) embedded tetranucleotide frequencies using DBSCAN (epsilon = 1.5, min points = 10). Some clusters enriched with phage of a single taxon are labeled with percentages denoting the proportion of phages in that cluster which belong to that enriched taxon. A cluster is considered enriched with a taxon if the proportion of phages belonging to that taxon in the cluster is greater than 50% and statistically significantly greater than that in the entire reference dataset, as tested using Pearson's chi-squared test. **b-d**, Compositions of taxa for phages assigned to the clusters shown in **a** at the Baltimore classification (**b**), Order (**c**), and Family (**d**) levels. Cluster numbers correspond with those in **a**.

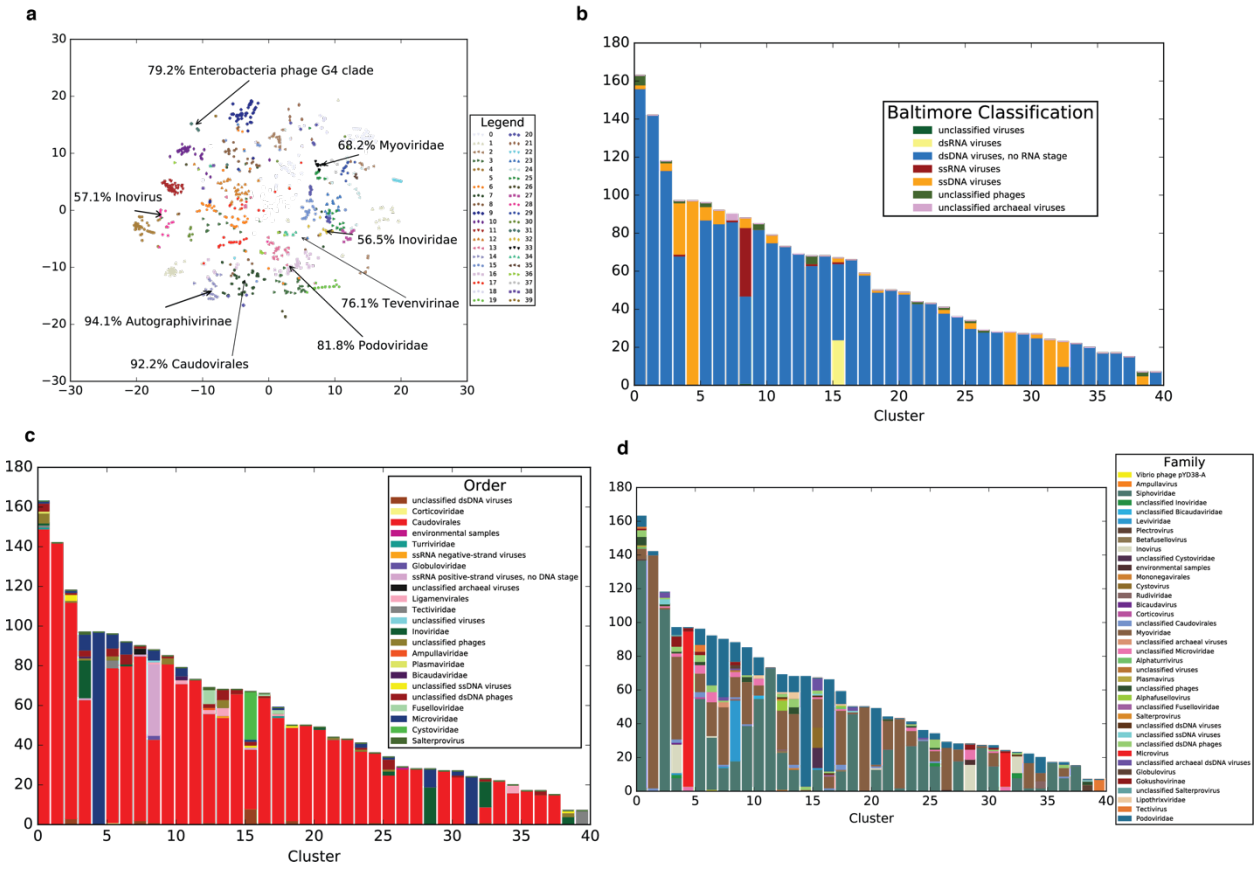


Figure 4: Phage tetranucleotide characteristics (alternate clustering)

a, t-SNE representation of tetranucleotide frequency vectors of 2255 phage genomic sequences from RefSeq. Clusters assigned with tetranucleotide frequencies using *k*-means (*k*=40). **b-d**, Compositions of taxa for phages assigned to the clusters shown in **a** at the Baltimore classification (**b**), Order (**c**), and Family (**d**) levels. Cluster numbers correspond with those in **a**.

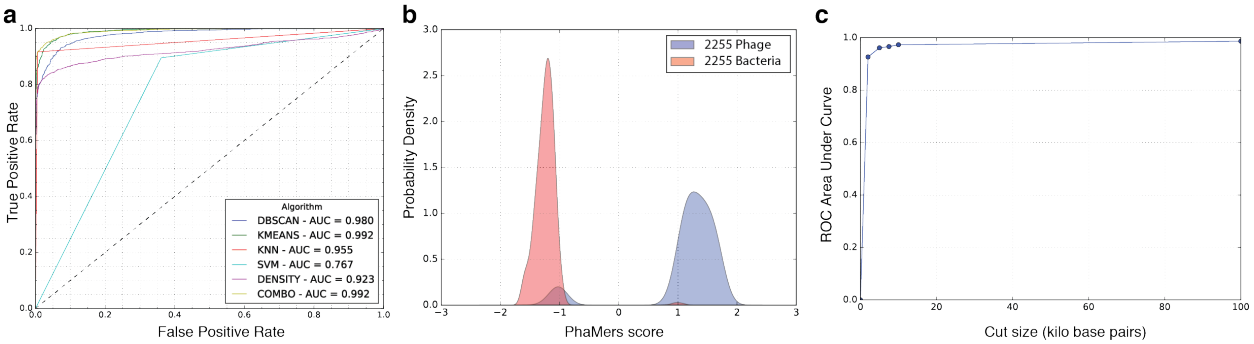


Figure 5: Algorithmic design and K-fold cross validation of PhaMers

a, Receiver Operator Characteristic (ROC) for supervised learning algorithms used on the reference datasets of phages and bacteria as tested by 20-fold cross validation. Feature vectors were tetranucleotide frequencies calculated from the 2255 phage and 2255 bacterial sequences in the reference dataset. "DBSCAN" and "KMEANS" scoring algorithms use the cluster proximity metric described in the main text, based on cluster assignments given by DBSCAN and *k*-means, respectively. "COMBO" is the PhaMers scoring algorithm described in the main text. "KNN" represents a K-Nearest-Neighbors algorithm, and "SVM" represents a Support Vector Machine approach. "DENSITY" stands for a custom algorithm that uses Kernel Density Estimation to approximate the probability density of phage and bacteria data points, and then gives a score as the log-ratio of the two probability densities. **b**, Distributions of PhaMers scores for 2255 phage and 2255 bacterial genomes. Scores were calculated with 20-fold cross validation. The small blue population at approximately score = -1 are the phages in the datasets which were misclassified as bacteria. There is also a smaller population of bacteria which were misclassified as phages shown at score = 1. **c**, Predictive performance (AUC) of PhaMers as a function of reference sequence length. The reference datasets of phage and bacterial genomes were cut randomly to sizes of 2.5 kbp, 5 kbp, 7.5 kbp, 10 kbp, and 100 kbp. Predictive performance, show on the y-axis, is given by the area beneath the ROC curve, which drops as sequences were cut to lengths shorter than 5 kbp.

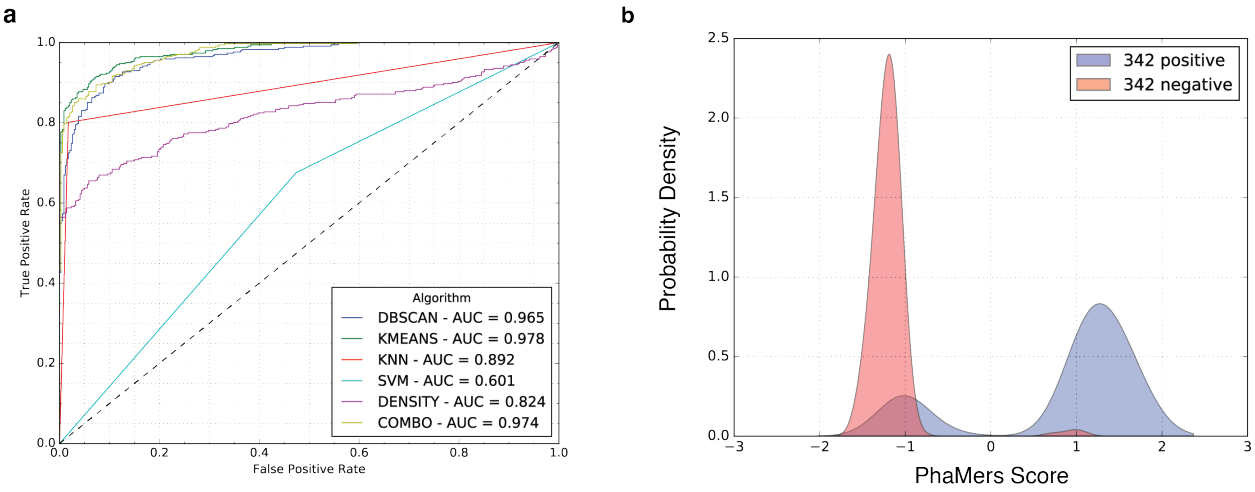


Figure 6: PhaMers performance on filtered phage dataset

a, ROC curves for 20-fold cross validation of supervised learning algorithms using the reduced dataset of phages. The dataset was constructed from the original dataset such that no two phages with the same taxonomic classification were included in the reduced dataset. This reduced dataset contains 342 phages of the 2255 phages in the original dataset. A random subset of 342 of the original bacterial genomes were selected as negative data points in this analysis. **b,** Distributions of PhaMers scores given to the 342 phages (blue) and 342 bacteria (red) during cross validation.

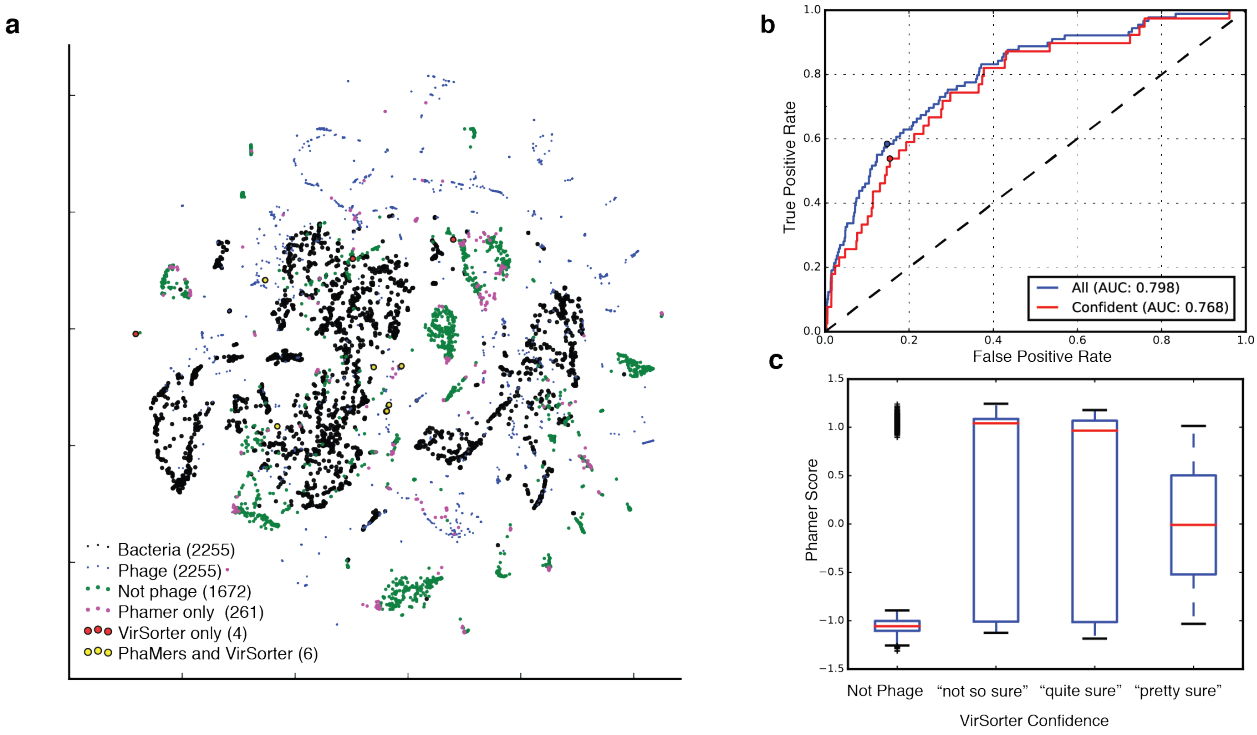


Figure 7: Comparison of PhaMers with VirSorter

a, t-Distributed Stochastic Neighbor Embedding (t-SNE) of tetranucleotide frequency vectors. Reference datasets of phage (blue) and bacteria (black) are shown in conjunction with contigs longer than 5 kbp from the Bijah Spring dataset. Points representing contigs are colored according to whether they were labeled as viral by VirSorter (Red), PhaMers (purple), both(yellow), or neither (green). **b,** ROC curves for predictions made by PhaMers on the Bijah Spring dataset as though VirSorter predictions are truth. The blue curve shows ROC curve if all VirSorter categories of putative phages are counted as phages, and the red curve shows the result if only confident (categories 1,2,4,5) classifications are considered phages. Dots on the curves show the resulting sensitivity and specificity values for a PhaMers score threshold of zero. **c**, Box plot showing distributions of PhaMers scores as a function of the confidence assigned to the predictions made by VirSorter. Contigs predicted to be viral by VirSorter tend to have a higher PhaMers score than those which are not predicted to be viral.

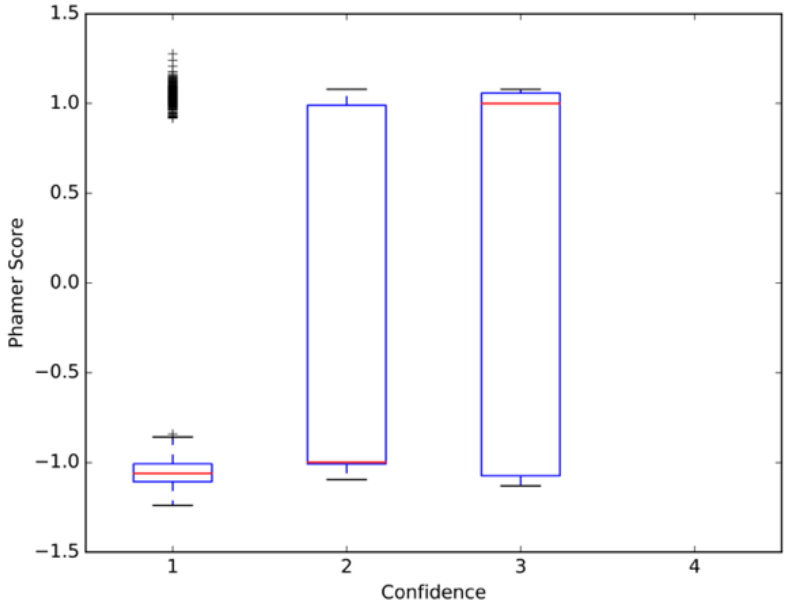


Figure 8: Boxplot of PhaMers scores vs. VirSorter Confidence for Bijah Spring

This boxplot shows the distributions of PhaMers scores given to contigs from Bijah Road Side as a function of the confidence with which they were predicted to be viral by VirSorter. On the x-axis, 1 indicates contigs that were not predicted to be viral by VirSorter. Confidence of 2, 3, and 4 indicate contigs that were predicted to be viral with low, medium, and high confidence, respectively. This plot shows that contigs classified as phages by VirSorter with greater confidence tend to be assigned higher PhaMers scores.

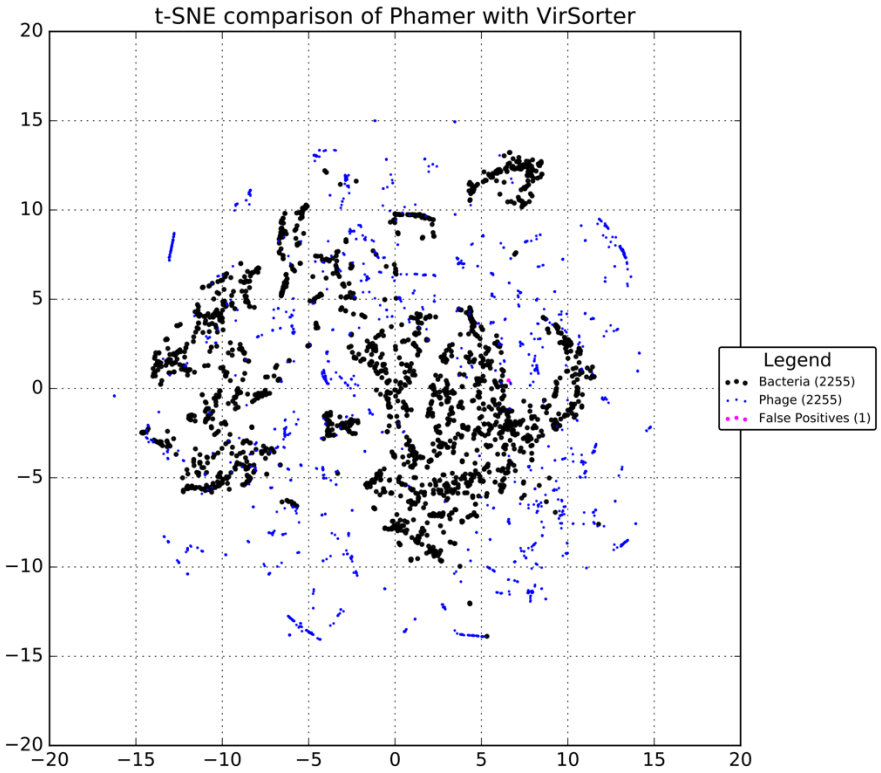


Figure 9: Mammoth Geyser Basin t-SNE of tetranucleotide frequencies

This plot shows a t-SNE two-dimensional embedding of tetranucleotide frequencies from the reference dataset of phages (blue) and bacteria (black), and from the single *Acidianus filamentous* contig (shown in purple). This plot shows that this phage's tetranucleotide signature is not easily indefinable as a phage, given that it lies in proximity to a large group of bacterial genomes.

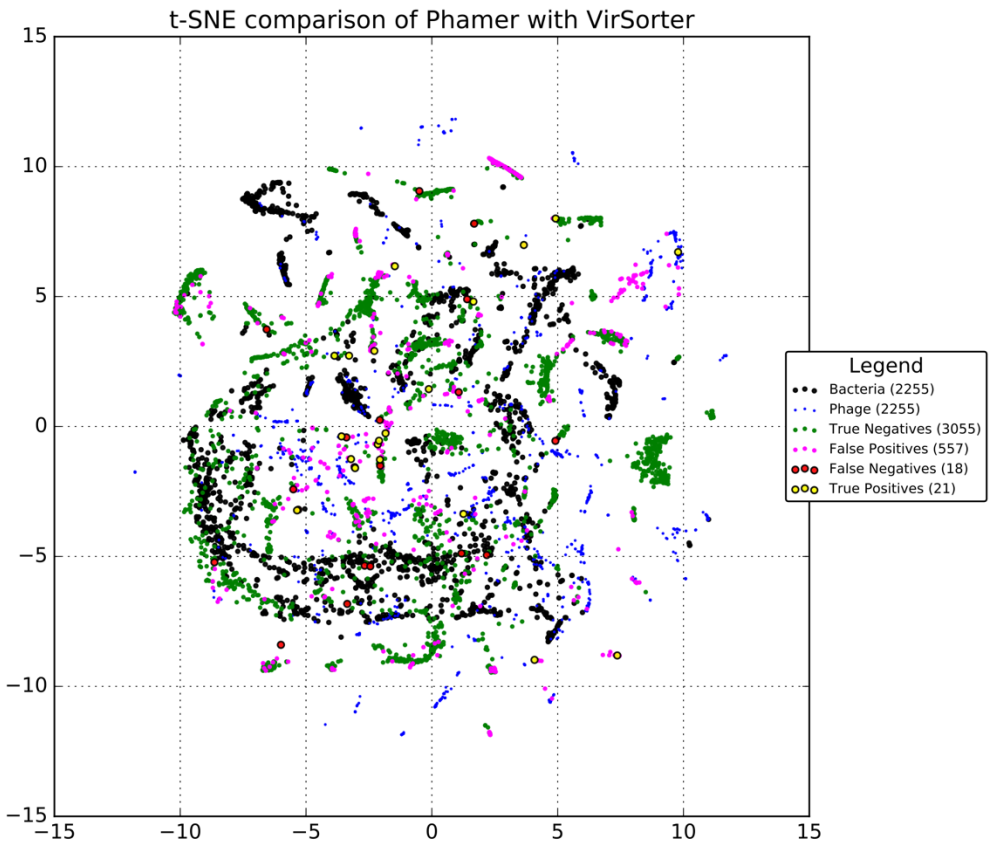


Figure 10: Lower Geyser Basin t-SNE of tetranucleotide frequencies

This plot shows a t-SNE embedding of tetranucleotide frequencies from the reference dataset of phages (blue) and bacteria (black), and those of the contigs from the Mound Spring sample. This plot shows that many of the phages which were identified by both VirSorter and PhaMers (yellow) lie in a region, at approximately (x, y) = (-2.5, -1), that lacks bacterial genomes, allowing easy identification of phages. This plot also shows phages which were identified despite being located within clusters of contigs that are likely mostly from a single bacterial genome, such as the one at (x, y) = (5, 7.5).

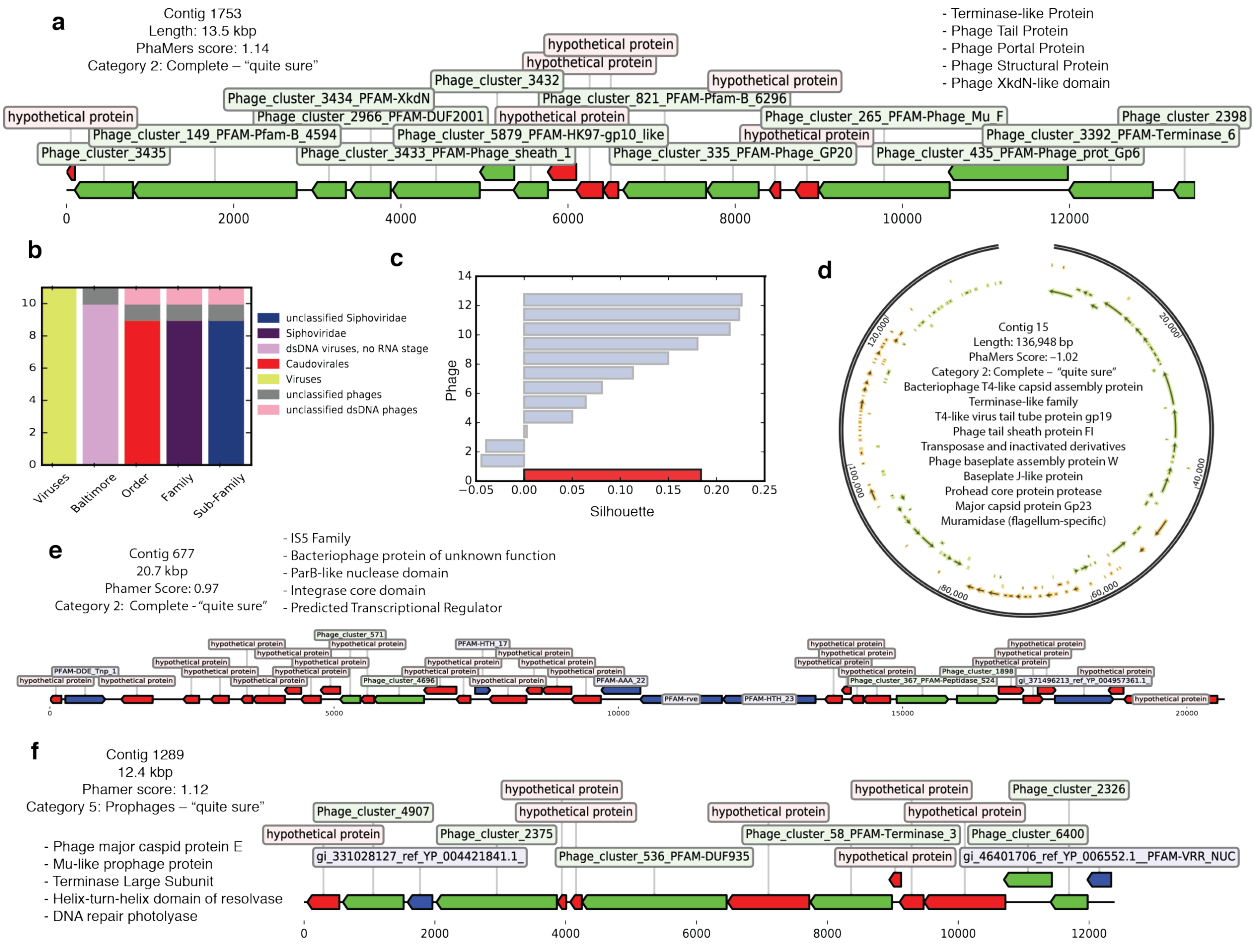
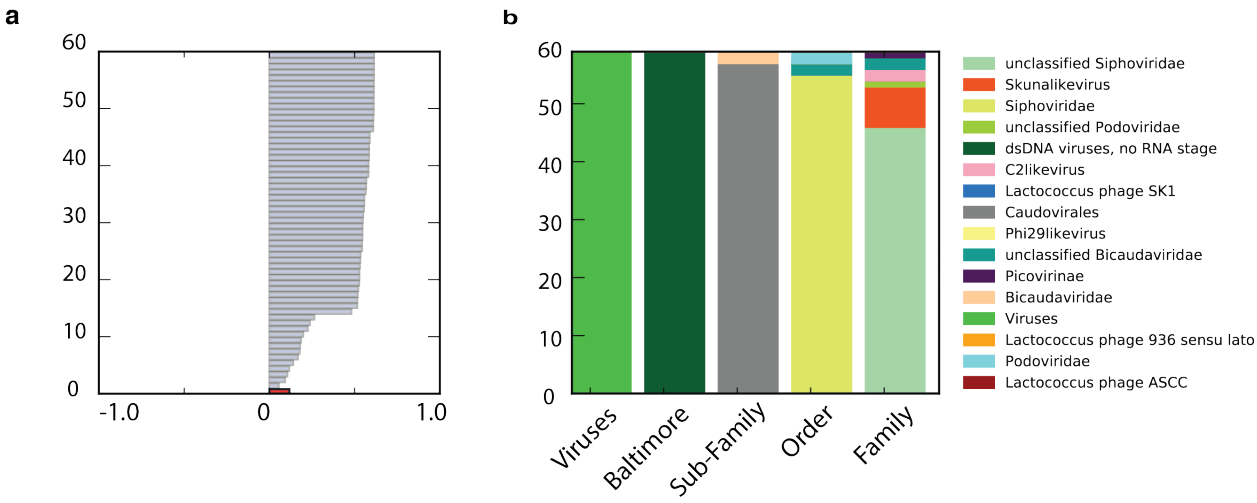


Figure 11: Diagrams of novel phage contigs

a, Diagram of 13.5 kbp contig (Contig 1753) from metagenomic sequencing of Mound Spring. Putative coding regions identified by VirSorter are shown as arrows. Some viral gene annotations given by Integrated Microbial Genomes and Microbiomes annotation pipeline (IMG) available on the Joint Genome Institute's (JGI) website are listed in upper right. **b**, Bar chart representing the composition of taxa for phages in the cluster which contig 1753 was assigned by *k*-means (*k*=86) on the basis of tetranucleotide frequencies. **c**, Cluster silhouette values for reference phages (blue) and contig 1752 (red) assigned by *k*-means clustering. **d**, Diagram of contig 15 with putative coding regions annotated by VirSorter shown by green and orange arrows. Viral gene annotations found by IMG are listed along with contig size, PhaMers score, and VirSorter confidence. This contig is displayed as circular for visualization, rather than to represent its structure. **e**, Diagram of contig 677 showing gene annotations given by VirSorter. Listed above are contig length, PhaMers score, VirSorter confidence and viral gene annotations given by IMG. **f**, Diagram of contig 1289 showing gene annotations given by VirSorter. Listed to the left are contig length, PhaMers score, VirSorter confidence and viral gene annotations given by IMG.



SubSample_IL0320-714-518_ODE_1_length_35211_cov_14002.5_ID_1
Category: No VirSorter Category
Phamer score: 0.8958932613
Cluster enriched: 57.6% Lactococcus phage 936 sensu lato p=9.6e-134 (Sub-Family)

Figure 12: Tetranucleotide characteristics of Acidianus filamentous contig

a, Cluster silhouette values for phages which were assigned to the same cluster (*k*-means clustering) as the *Acidianus Filamentous* putative phage contig on the basis of tetranucleotide frequencies. **b,** Taxonomic composition of that cluster of reference phages, with enrichment for *unclassified Siphoviridae*.

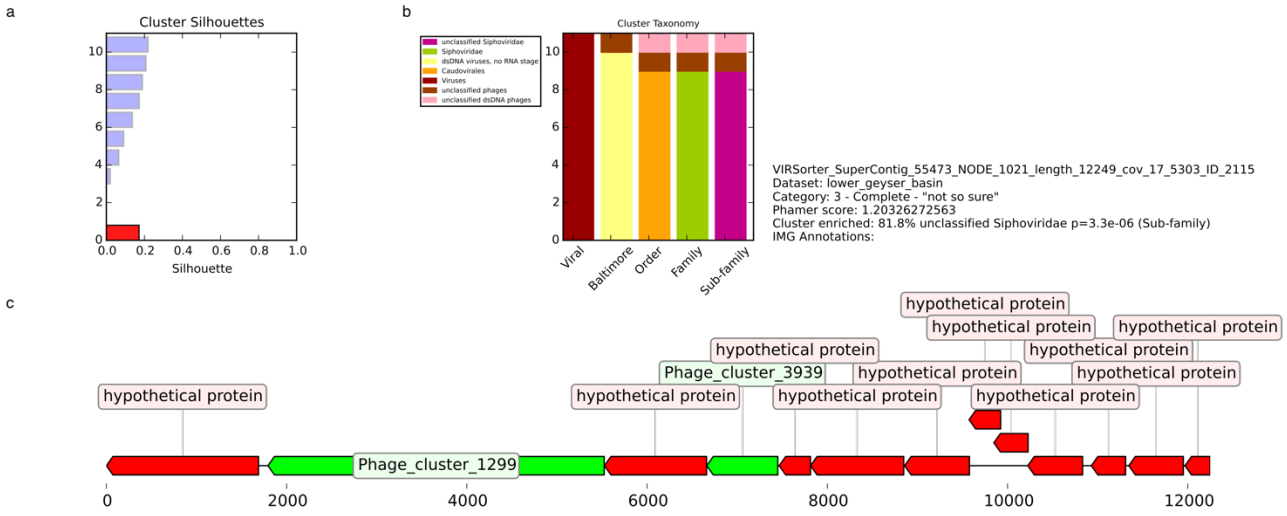


Figure 13: Diagram of contig 2115

a, Cluster silhouette values for phages which were assigned to the same cluster (*k*-means clustering) as the contig 2115 (Mound Spring**)** on the basis of tetranucleotide frequencies. **b,** Taxonomic composition of that cluster of reference phages, with enrichment for *unclassified Siphoviridae*. **c,** Genetic feature diagram of contig 2115 given by VirSorter.

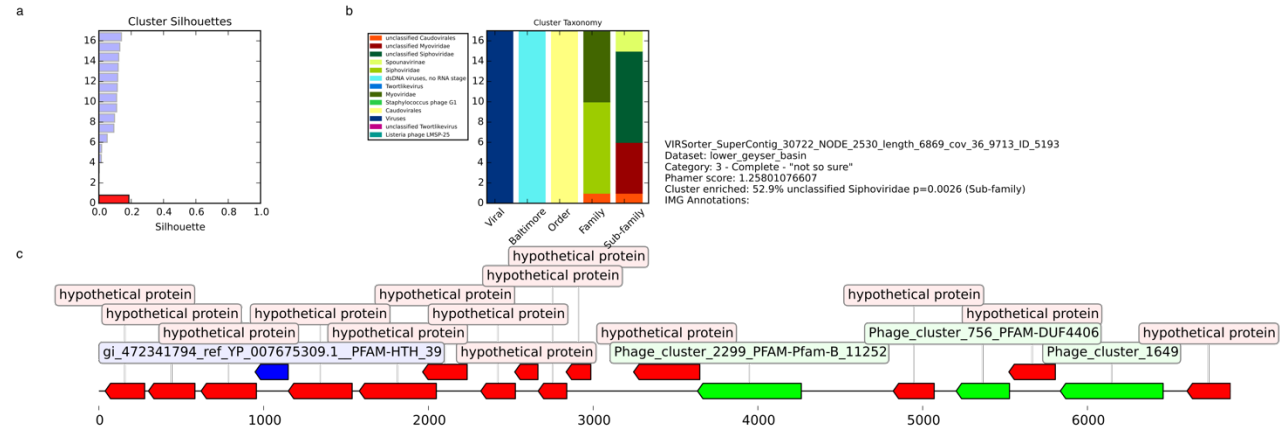


Figure 14: Diagram of contig 5193

a, Cluster silhouette values for phages which were assigned to the same cluster (*k*-means clustering) as the contig 5193 (Mound Spring**)** on the basis of tetranucleotide frequencies. **b,** Taxonomic composition of that cluster of reference phages, with enrichment for *unclassified Siphoviridae*. **c,** Genetic feature diagram of contig 5193 given by VirSorter.

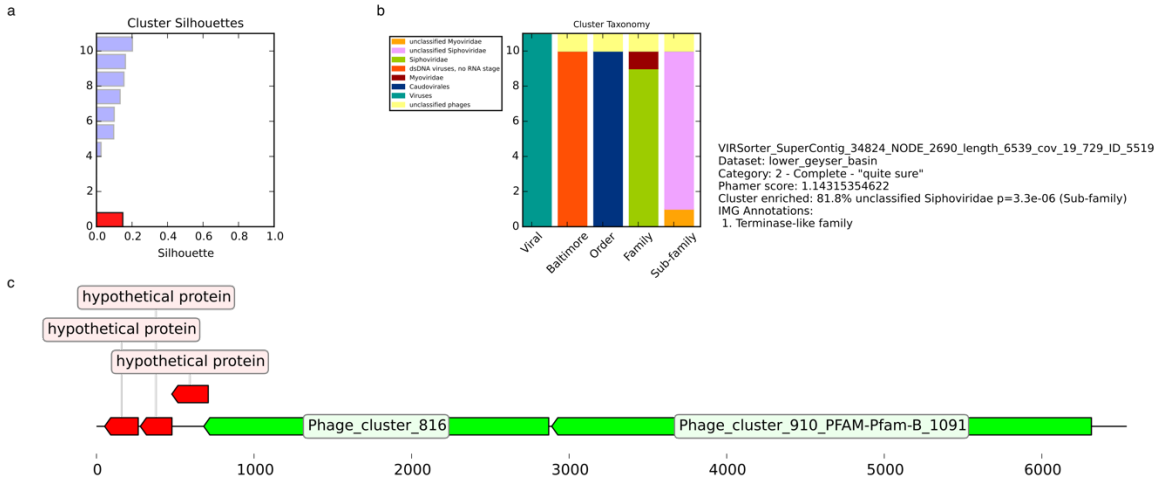


Figure 15: Diagram of contig 5519

a, Cluster silhouette values for phages which were assigned to the same cluster (*k*-means clustering) as the contig 5519 (Mound Spring**)** on the basis of tetranucleotide frequencies. **b,** Taxonomic composition of that cluster of reference phages, with enrichment for *unclassified Siphoviridae*. **c,** Genetic feature diagram of contig 5519 given by VirSorter.

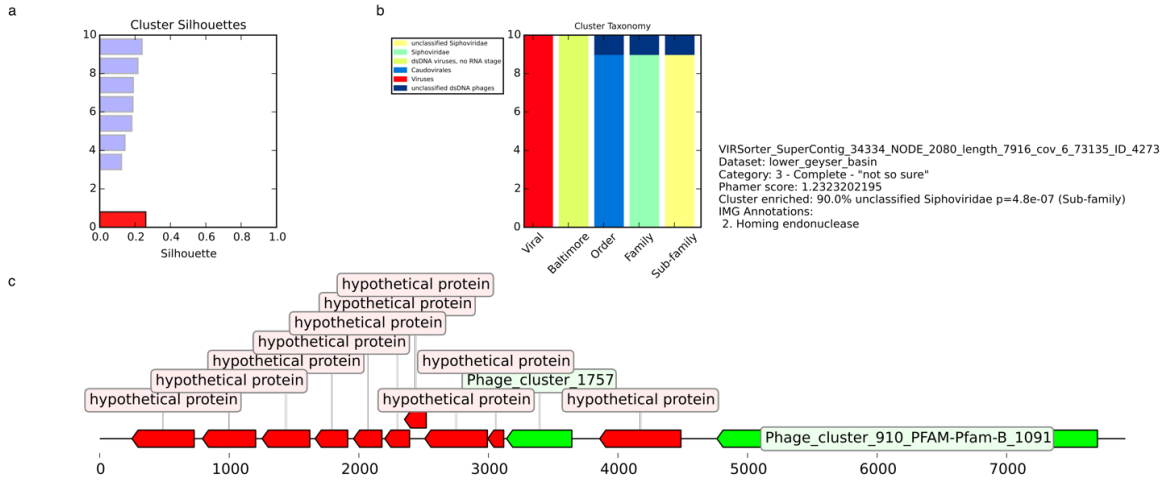


Figure 16: Diagram of contig 4273

a, Cluster silhouette values for phages which were assigned to the same cluster (*k*-means clustering) as the contig 4273 (Mound Spring**)** on the basis of tetranucleotide frequencies. **b,** Taxonomic composition of that cluster of reference phages, with enrichment for *unclassified Siphoviridae*. **c,** Genetic feature diagram of contig 4273 given by VirSorter.

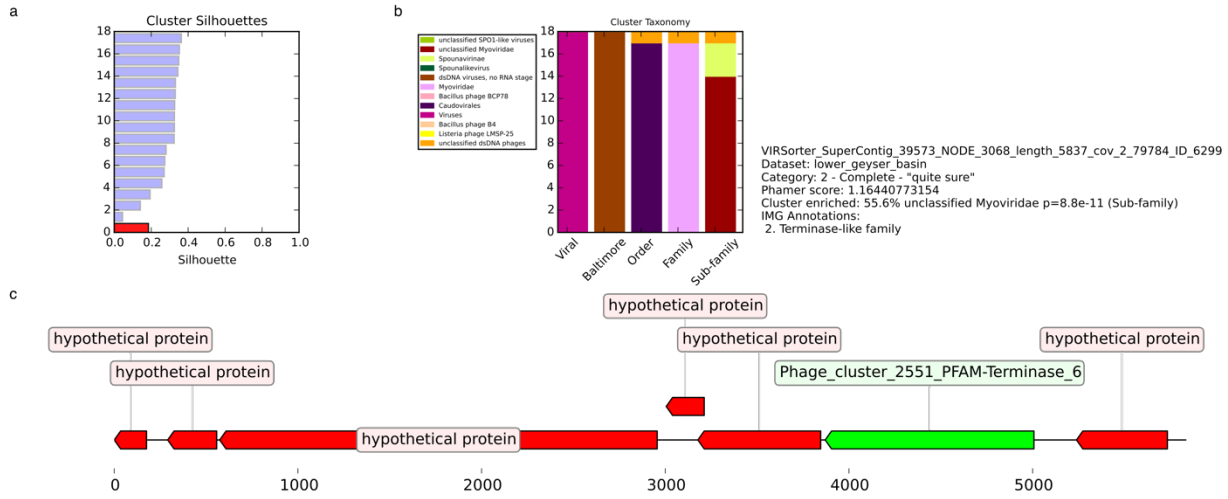


Figure 17: Diagram of contig 6299

a, Cluster silhouette values for phages which were assigned to the same cluster (*k*-means clustering) as the contig 6299 (Mound Spring**)** on the basis of tetranucleotide frequencies. **b,** Taxonomic composition of that cluster of reference phages, with enrichment for *unclassified Myoviridae*. **c,** Genetic feature diagram of contig 6299 given by VirSorter.