Introduction
Bacteriophages (phages) are ancient biological entities and the most abundant living things on the planet, at an estimated 10^31 particles. Phages are viruses that infect microorganisms such as bacteria and archaea, and play important roles in microbial communities such as lateral gene transfer and gene duplication. Despite the abundance of phage populations around the world, we understand little of their genetic diversity, owing to difficulties in culturing phages with no known culturable host. Also, many phages exist in small populations, making them difficult to study with microscopy or classic laboratory methods. High-throughput DNA sequencing has allowed researchers to bypass these problems and study elusive phage species by sequencing environmental samples. This practice, called metagenomics, is responsible for the recent explosion of discovered phage genomes. Studying phages in this manner requires the ability to computationally analyze DNA sequences to determine which represent genomic fragments of phages, and which are more likely from other microorganisms. In this study, we used used existing computational tools, and created a new k-mer based analysis tool, to identify and classify novel phage DNA sequences. We applied these tools to environmental samples taken from hot springs in Yellowstone National Park.

K-mer Analysis

**k-mer (k = 4)**

\[ \text{GTACTGATCGTACGTA} \ldots \]

k-mer are short DNA sequences of length k. Because there are 4 base pairs of DNA, for a given value of k, there are 4^k possible k-mers. Long DNA sequences may be compared on the basis of k-mer frequencies by counting the occurrence of each of the 4^k possible k-mers and normalizing. To analyze unidentified DNA sequences found in environmental samples, we created an analysis pipeline that compares tetramer (k = 4) frequencies of newly discovered sequences to those of previously discovered phage genomes.

Materials & Methods

Samples were collected from hot springs in Yellowstone National Park and prepared using Fluidigm’s C1 automated sample preparation system. Libraries were created using Illumina’s Nextera library preparation protocol and sequenced on Illumina sequencing platforms. Reference phage and bacteria genomes used in k-mer were taken from NCBI in October 2015. VirSorter version 1.0.3 was used in phage genome identification, and JGI’s Integrated Microbial Genomes (IMG) annotation pipeline was used to annotate genomes on putative phage contigs.

Results

Our analysis of the 2255 phage genomes available in NCBI revealed that when clustered on the basis of tetramer frequency, many clusters of phages are enriched with a single viral taxon. (Figure 1) When we compared the performance of our k-mer frequency analysis tool to that of VirSorter, an automated phage identification tool, we learned that tetramer frequencies have predictive power in phage identification, but have limited positive predictive value. By creating two-dimensional embeddings of k-mer frequencies with t-SNE, we observed that many contigs form tight clusters, some of which contain DNA sequences identified as phage. (Figure 2) These clusters are hypothesized to be collections of fragments from single microbial genomes, and the phages located within these clusters are hypothesized to infect these microbes.

![Figure 2a](image1.png)

Figure 2a is a two-dimensional t-SNE scatterplot of tetramer frequency vectors from reference phage, reference bacteria, and metagenomic contigs from Yellowstone National Park. This scatterplot shows that many phage predicted by k-mer analysis and VirSorter lie in proximity to clusters of known phage genomes. Figures 2b and 2c were generated by unidentified sequences and show that k-mer frequencies have predictive power, but have a weak positive predictive value.

![Figure 2b](image2.png)

![Figure 2c](image3.png)

Future Direction

We intend to continue this work by identifying hosts for each viral contig, and further characterizing the taxa of each novel phage. Additionally, we would like to improve the performance of our k-mer based analysis pipeline by adding the ability to examine other features like the presence of viral genes and structures. Finally, given that k-mer based phage identification has weak positive predictive value, and therefore should not be used alone, we would like to integrate this tool into preexisting phage identification tools in order to improve their predictive performance.

References


