Discovery and Characterization of Novel Bacteriophage from Yellowstone National Park

Jonathan Deaton, Feiqiao Brian Yu, Stephen Quake
Stanford University Department of Bioengineering

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 microbes. 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.

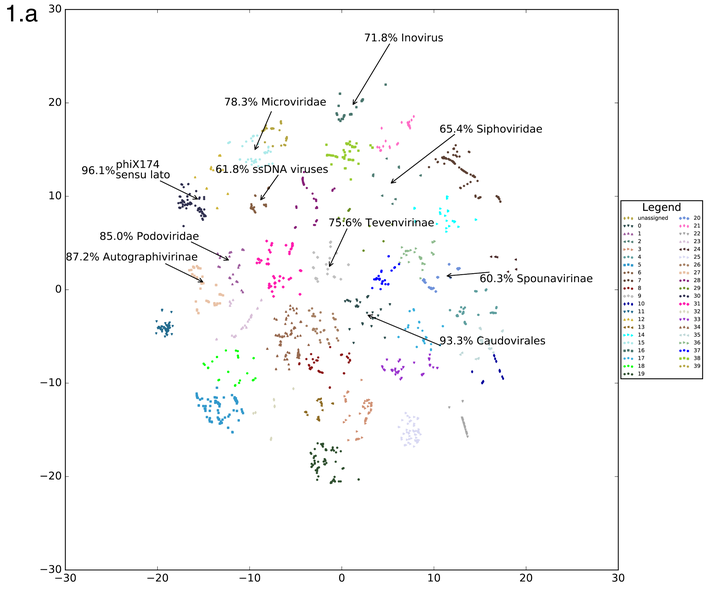
K-mer Analysis

k-mer ($k = 4$)

.....GTACTGATCGAGTACGTCA.....

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 occurrences 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.

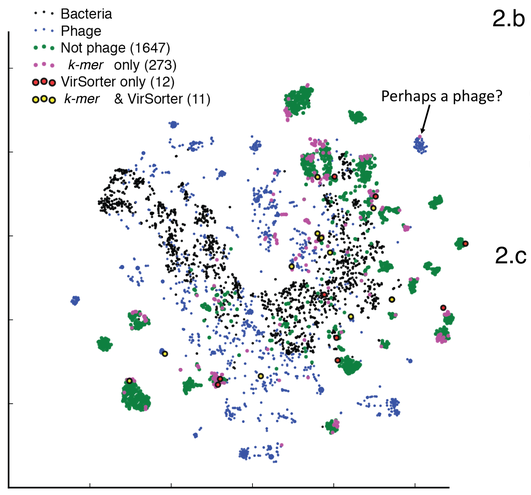
1.a



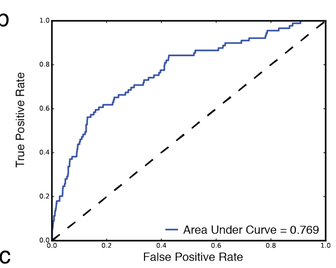
Results

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 those microbes.

2.a



2.b



2.c

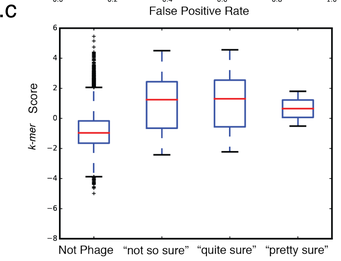
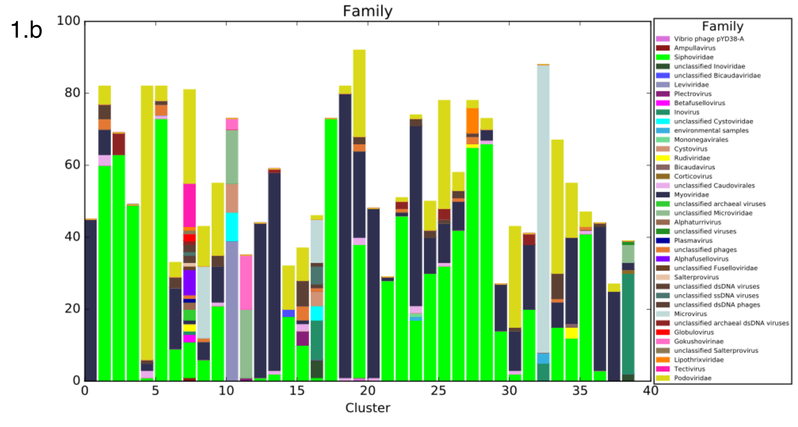


Figure 2.a is a two-dimensional t-SNE scatterplot of tetramer frequency vectors from reference phage, reference bacteria, and metagenomic contigs from Yellowstone National Park. This scatterplot shows that many phage predicted by *k*-mer analysis and VirSorter lie in proximity to clusters of known phage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

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 of 2015. VirSroter version 1.0.3 was used in phage genome identification, and JGI's Integrated Microbial Genomes (IMG) annotation pipeline was used to annotate genes on putative phage contigs.

1.b

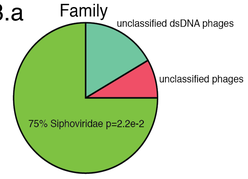


Novel Phages

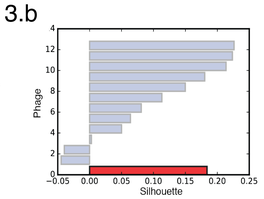
We identified 106 genomic fragments and complete genomes of novel phages through the use of VirSorter, the IMG annotation pipeline, and our own *k*-mer based analysis pipeline. Given that all the novel phage genomes were found in hot springs, they code for phages that are both tolerant of thermal environments and likely to infect thermophilic hosts. We also used *k*-mer based clustering to predict the taxa of several phages.

In order to consider a *k*-mer based taxonomic classification for a phage, that phage must have been assigned to a cluster enriched for a single taxon of known phages, and have a cluster silhouette value within one standard deviation of those from the known phage in the cluster. Some of these predictions were supported by presence of genes. For instance, contig 1753 (Figure 3a-c) contains a tail protein and was assigned to an enriched cluster of *Siphoviridae*. These two characteristics are evidence that this phage might be classified as *Siphoviridae.*

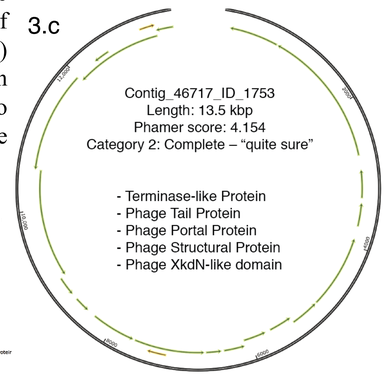
3.a



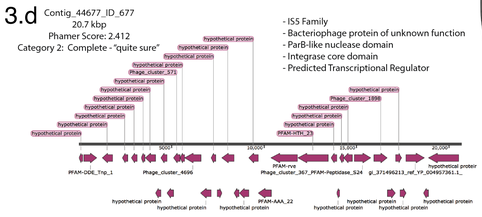
3.b



3.c



3.d



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

1. Dr. A. Edwards, K. McNair, K. Fraust, J. Raes and B. E. Dutilh, "Computational approaches to predict bacteriophage–host relationships," FEMS Microbiology Reviews, 2015.
2. J. C. Wooley, A. Godzik and I. Friedberg, "A Primer on Metagenomics," PloS Computational Biology, vol. 6, no. 2, 26 2 2010.
3. R. A. Edwards and F. Rohwer, "Viral Metagenomics," Nature Reviews Microbiology, pp. 504-510, 2005.
4. B. L. Hurwitz, J. M. U'Ren and K. Youens-Clark, "Computational prospecting the great viral unknown," FEMS Microbiology Letters, 2016.
5. S. Roux, F. Enault, B. L. Hurwitz and M. B. Sullivan, "VirSorter: Mining viral signal from microbial genomic data.," PeerJ, 2015.
6. V. Trifonov and R. Rabadan, "Frequency Analysis Techniques for Identification of Viral Genetic Data," mBio, pp. 156-10, 2010.
7. 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.
8. 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.
9. 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.
10. D. Papamichail, S. S. Skiena, D. Van Der Lelie and S. R. Mccorkle, "Bacteria Population Assay Via k-mer Analysis," 2004.
11. R. Ounit, S. Wanamaker, T. J. Close and S. Lonardi, "CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mer," BMC Genomics, 2015.
12. 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.
13. L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," Journal of Machine Learning Reserach, 2008.
14. R. A. Edwards, K. McNair, K. Fraust, J. Raes and B. Dutilh, "Computational approaches to predict bacteriophage–host relationships," FEMS Microbiology Reviews, 2015.
15. "Bacteriophages." *ZeptoMetrix*. ZeptoMetrix, n.d. Web. <http://www.zeptometrix.com/store/bacteriophage/>.
16. Nordstrom, Kirk D. "Bijah Spring Details - Yellowstone National Park." *Montana State University*. Montana State University, 28 July 2000. Web. 1 Oct. 2016. <http://www.rcn.montana.edu/Features/Detail.aspx?id=6695>.

Stanford | Bioengineering

Jonathan Deaton | Quake Lab | jdeaton@stanford.edu