Fall 12-1-2012

Determination of Potential Azotobacter vinelandii Bacteriophage Gene Sequences: A Bioinformatics Approach

Anish Reddy
Case Western Reserve University

Sarah Bagby
Case Western Reserve University, scb126@case.edu

Determination of Potential *Azotobacter vinelandii* Phage Gene Sequences: A Bioinformatics Approach

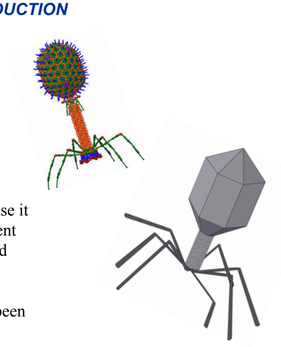
Case Western Reserve University
Anish Reddy, Department of Biology
Dr. Sarah Bagby, Department of Biology

ABSTRACT

Microbial communities play a key role in shaping many diverse ecosystems through their biogeochemical contributions. These communities comprise not only bacteria and archaea but also their viruses, whose reproduction profoundly affects host cell biology. While many bacterial species have been well characterized, the challenges of isolation and sequencing have hampered the study of environmental viruses such as bacteriophages. The model organism *Azotobacter vinelandii* is a common nitrogen-fixing soil bacterium. To our knowledge, no *Azotobacter* phages currently exist in culture. However, modern bioinformatic and database approaches can be used to identify the metagenomic sequences that may derive from bacteriophages that infect *Azotobacter*. We have developed a pipeline that scans metagenomic samples from the IMG JGI database to identify phage sequences whose abundance co-varies with *Azotobacter vinelandii* abundance. The viral sequences identified can be grouped into clusters, and the resulting clusters can then be analyzed for auxiliary metabolic genes (AMGs). These candidate sequences may then be used to guide phage isolation strategies and predict phage ecological impact.

INTRODUCTION

Azotobacter vinelandii, a bacterial species commonly found in many soils around the world, is a model organism to understand nitrogen fixation, respiration, hydrogen production and assimilation, and enzyme kinetics. Unusually, it fixes nitrogen aerobically, even though the enzymes that perform nitrogen fixation are damaged by molecular oxygen if left unprotected. Because it encodes three distinct nitrogenases, dependent respectively on molybdenum, vanadium, and iron cofactors, it is able to perform nitrogen fixation under a range of trace-metal conditions [3]. Due to these factors, it has been an important model system in research regarding the biogeochemical nitrogen cycle.

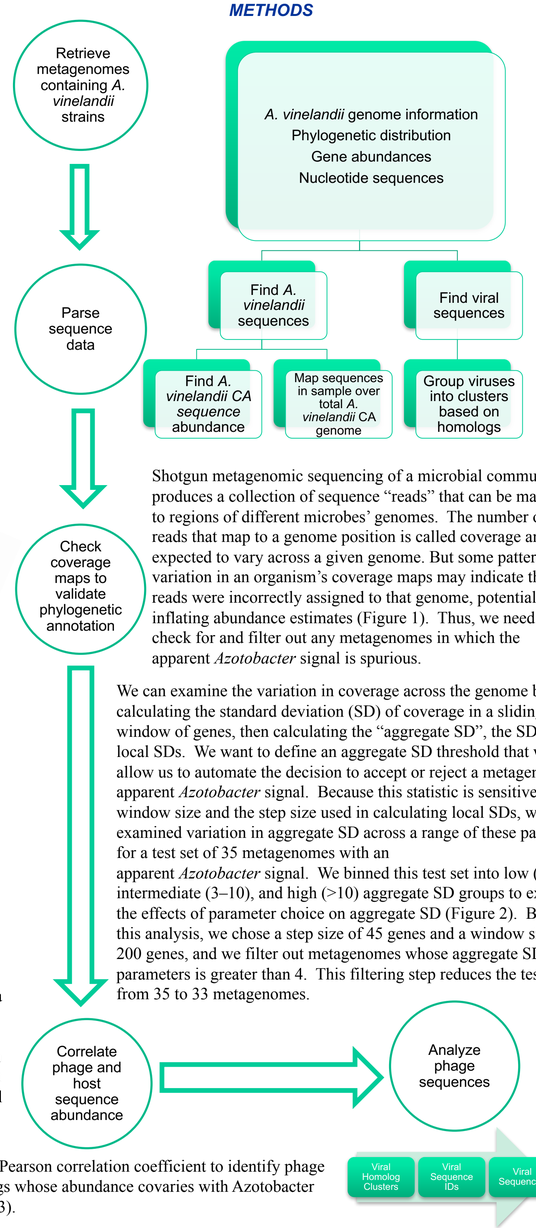


Despite the large body of research investigating *Azotobacter vinelandii*'s environmental role and impact, one important area has been largely ignored—the interactions that this organism has with bacteriophages. We hypothesize that these interactions' effects on *Azotobacter* physiology and community metabolism significantly shape the microbes' role in global biogeochemical cycles.

Currently, little is known about soil viruses and how they affect microbial populations, through either microbial mortality rates or altered metabolic activity [4,5]. This gap is particularly acute for *Azotobacter*. But we expect viruses to be a key player in soil microbial ecology, based on results in other habitats, which show that large populations of microbes are generally infected with (largely uncharacterized) phages at any given time [4], and that these phages act as a form of population control and can drastically impact metabolic activity, such as the effects of many phages on carbon-cycling bacteria and the associated impact on the carbon cycle [5].

Similarly, bacteriophage infection of *Azotobacter vinelandii* could have a large effect on the biogeochemical nitrogen cycle and on any models of this cycle. Thus, it is important to characterize *Azotobacter* bacteriophage and their impact on host physiology. This will give us a better understanding of the process of nitrogen fixation in nature and the overall nitrogen cycle, with potential applications as wide-ranging as industrial production and environmental policy.

METHODS



Shotgun metagenomic sequencing of a microbial community produces a collection of sequence "reads" that can be mapped to regions of different microbes' genomes. The number of reads that map to a genome position is called coverage and is expected to vary across a given genome. But some patterns of variation in an organism's coverage maps may indicate that reads were incorrectly assigned to that genome, potentially inflating abundance estimates (Figure 1). Thus, we need to check for and filter out any metagenomes in which the apparent *Azotobacter* signal is spurious.

We can examine the variation in coverage across the genome by calculating the standard deviation (SD) of coverage in a sliding window of genes, then calculating the "aggregate SD", the SD of those local SDs. We want to define an aggregate SD threshold that would allow us to automate the decision to accept or reject a metagenome's apparent *Azotobacter* signal. Because this statistic is sensitive to the window size and the step size used in calculating local SDs, we examined variation in aggregate SD across a range of these parameters for a test set of 35 metagenomes with an apparent *Azotobacter* signal. We binned this test set into low (<3), intermediate (3–10), and high (>10) aggregate SD groups to examine the effects of parameter choice on aggregate SD (Figure 2). Based on this analysis, we chose a step size of 45 genes and a window size of 200 genes, and we filter out metagenomes whose aggregate SD at these parameters is greater than 4. This filtering step reduces the test set from 35 to 33 metagenomes.

Use the Pearson correlation coefficient to identify phage homologs whose abundance covaries with Azotobacter (Figure 3).

RESULTS

The parameters for the pipeline run in this trial were: 50%+ match to be considered an *A. vinelandii* CA strain; 30%+ match to be considered for an initial viral homolog cluster; window size of 200 steps and rolling increment of 45 steps; threshold of 4 or under for a "valid" sample; threshold of r = 0.6 + and p < .01 for correlation analysis.

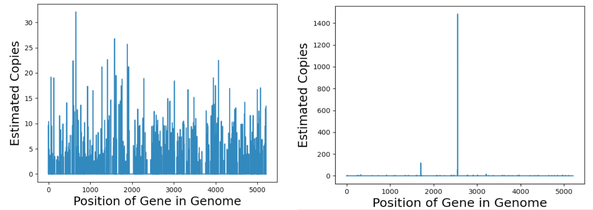


Figure #1: Coverage maps for *Azotobacter vinelandii* in two different metagenomes. The map at left is consistent with the presence of *Azotobacter* in the sampled community; the map at right strongly suggests spurious annotation. Note the difference in y-axis scales.

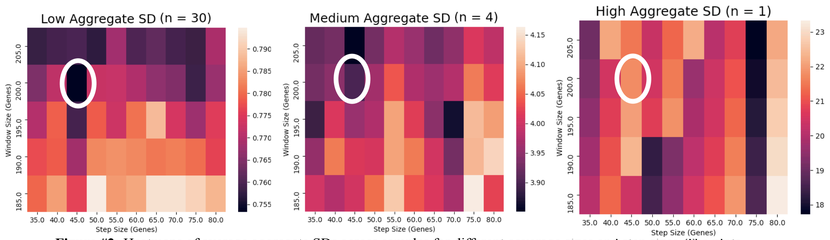


Figure #2: Heatmaps of average aggregate SDs across samples for different coverage sizes and step sizes. The plots represent data in the first bin (left), second bin (middle), and third bin (right). The circle represents the set parameters that was chosen to run the validation step with (coverage size of 200 and step size of 45). Note the difference in color scales.

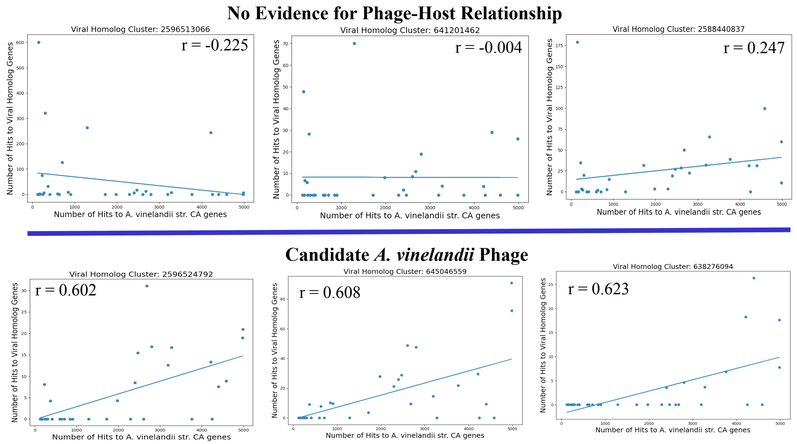


Figure #3: Metagenomic correlation analysis for six representative phage homolog clusters. As a proxy for species abundance in each metagenome, we used the total number of hits to the *A. vinelandii* str. CA genome or to the viral homolog cluster. Of the 34331 phage homologs assessed, six were strongly correlated (r > 0.6) with *Azotobacter* abundance. Note the difference in y-axis scales.

DISCUSSION

33 of the initial 35 datasets passed the validation step. Then, through this pipeline, 6 initial viral homolog clusters were found. These 6 homolog clusters represented 221 viral gene sequences.

It is important to remember that the results of this correlation analysis will change depending on the parameters used, such as the percent similarity to be considered an *A. vinelandii CA* gene sequence and the similarity to be considered for an initial viral homolog cluster. Thus, any outputs from this pipeline should always be analyzed in the context of what the parameter values were.

Additionally, this model is easily expandable to other bacterial strains, as the only change would be the reference genome that the gene counts are getting mapped to.

FUTURE DIRECTIONS

- Automate downloading of data from IMG database.
- Download more samples for greater accuracy.
- Expand model to work with other bacterial strains.
- Explore alternative summary statistics for the aggregate SD filter.
- Identify examples of different coverage pathologies and test aggregate SD filter further.
- Cluster analysis of viral gene sequences.
- Analyze viral clusters for AMGs.
- Analyze viral clusters for other genes of interest, such as large phage genes.

ACKNOWLEDGEMENTS

I would like to acknowledge CWRU SOURCE. the Bruce Rakay Summer Research Fellowship, and Dr. Sarah Bagby for providing me support to carry out this research. The Bagby lab's *Azotobacter* work is part of the DOE-funded VirSoil project.

REFERENCES

[1] Noar, J.D.., Bruno-Bárcena, J.M. (2018). Azotobacter vinelandii: the source of 100 years of discoveries and many more to come. *Microbiology*, 164(4), 421-436.
[2] Lipman, J.G. (1903). Experiments on the transformation and fixation of nitrogen by bacteria. *New Jersey State Agric Exp Sta Ann Rep*, 24, 217–285.
[3] Premakumar, R., Lemos, E., Bishop, P. (1984). Evidence for two dinitrogenase reductases under regulatory control bymolybdenum in Azotobacter vinelandii. *Biochim Biophys Acta*, 797, 64–70.
[4] Howard-Varona, C., Lindback, M.M., Bastien, G.E. et al. (2020). Phage-specific metabolic reprogramming of virocells. *ISME* J.
[5] Trubl, G., Jang, H. B., Roux, S., Emerson, J. B., Solonenko, N., Vik, D. R., et al. (2018). Soil Viruses Are Underexplored Players in Ecosystem Carbon Processing. *mSystems*, 3(5), e00076-18. doi:10.1128/mSystems.00076-18.