**IDENTIFICATION OF HIGHLY CONSERVED BACILLUS ORFS OF UNKNOWN FUNCTION**

Bacteriophages (Phages) are the most numerous [1] and highly diverse biological entities in the biosphere. The genetic diversity of these biological entities gives rise to numerous novel genes. A recent comparative genomic study done in the Pope lab analyzed 657 genomes [2]. Of these genomes, 69,633 ORFs were identified and grouped into 5,205 phams, of which 1,613 (31%) were orphams. This suggests bacteriophages are not only highly diverse, they also contain an exuberant amount of unexplored genetic information. While much of their genetic information is unknown, many phages such as Lamda [3] and T7 [4] have been studied extensively. These studies can be used as models for further exploration of other phage genomes. In this present study, we propose to investigate the function of highly conserved proteins in *Bacillus* phages by overexpression of *Bacillus* bacteria. The data generated by this study will establish a foundation for future functional analysis. Discovering the function of these highly conserved unknown proteins paves the way for a better understanding of viral – host interactions.

We will use an overexpression study by Jeroen Wagemans [5] and colleagues completed in 2014 as a model for our investigation of protein function. This study identified 26 proteins of unknown function found to be translated during the early bacterial infection process of *Pseudomonas* phages. Each of these proteins were cloned into the entry vector pUC18-mini-Tn7T-Lac, which is E. *coli* and P. *aeruginosa* compatible. Then transformed into the P. *aeruginosa* PAO1 bacteria with plasmid pTNS2 in order to facilitate the integration of the ORF into its hosts genome. The result was a single ORF in each host genome to be overexpressed. Cells were grown in various serial dilutions on media with and without IPTG present as a transcription inducing agent. Phenotypes at various stages of cell growth were observed. Of the 26 proteins overexpressed, 6 of them (Gp 7, 8, 14, 15, 18, and 30) were found to have a phenotypic impact on host bacterial growth. These 6 proteins were then selected for yeast two-hybrid assays for more detailed analysis of protein function. This experiment was repeated in both E. *coli* MG1655 and P. *aeruginosa* PA14 to verify the accuracy of results in P. *aeruginosa* PAO1 since the experimental phage does not infect the other host bacteria on its own.

The VCU SEA PHAGES program has over the years built a diverse library of sequenced and annotated *Bacillus* phage genomes. The *Bacillus* genus consists of the ATC family (B. anthracis, B. thuringiensis, B. cereus), which are all closely related by sequence. These bacteria are rod shaped sporulating gram-positive bacteria [6]. At VCU the B. thuringiensis phages are studied since its host is not a human pathogen but is still closely related to B. anthracis and B. cereus, which are human pathogens. Phages that infect these bacteria have the potential to be used therapeutically to treat their infectious host in humans. Due to the growing problem of bacterial resistance to antibiotics many scientists are looking elsewhere for alternatives, one such alternative being phage therapy. However, more must be known about these phages in order to be used to combat bacterial infections in humans safely. Gaining a better understanding of phage genomes, protein function and their protein – protein interactions with host bacteria could have a major impact on the food industry, human health and our quality of life. We propose to investigate the function of unknown proteins by establishing overexpression assays with an entry and gateway expression vector system to screen for phenotypes so that we can identify proteins for further functional analysis.

**Preliminary Data**

Detailed analysis of 83 *Bacillus* phage genomes using various bioinformatics tools allowed us to identify a set of highly conserved genes of unknown function. These biological entities can be categorized into clusters based on sequence similarity and shared gene content. Analysis of genome sequence similarity, generated by dot plot (**Fig.1**), categorized each of the 83 genomes into 13 clusters. Dot plot analysis is used to organize sequences with 50% or more similarity into clusters. Lines with darker shading indicate a higher level of similarity; cluster E contains 11 genomes with dark lines to show a comparison of sequence similarity within the closely related cluster. Cluster A, also containing 11 genomes, shows well defined dark lines indicating a higher level of conservation for sequence similarity compared to cluster E. While cluster A exhibits a conserved sequence similarity, the genome on the outer edge of the cluster shows a faint line. Despite the genome having over 50% similarity in the cluster, its sequence is less conserved when compared to the other genomes. SplitsTree imaging was done to verify results observed from Dot Plot analysis (**Fig.2**). SplitsTree software organizes genomes into clusters, much like Dot Plot, however, this program compares genomes by shared protein content instead of sequence similarity. Interestingly enough, both Dot Plot and SplitsTree yielded similar results. Lattice structures within the tree represent areas of protein diversity, while areas with thicker and longer lines indicate conserved proteins between genomes. Cluster A exhibits a long, darkened line with little lattice networking suggesting a group of conserved proteins shared among genomes. Cluster E, however, shows a long line with lattice networking throughout, indicating protein diversity with in the cluster. Although the tree organizes majority of the genomes into distinct clusters, the center of the tree shows an area where a large number of genomes share proteins in common. Given the diverse nature of phages, it can be speculated that proteins present in numerous genomes and clusters are essential to their viral life cycle, thus making highly conserved proteins of unknown function valuable candidates to further research.

The comparative genomic tool Phamerator [6] organizes proteins by sequence similarity into “phamilies” using clustalw and blastp scores. As of August 2015, Phamerator organized 14,922 phage proteins into 3,638 phamilies from a total of 83 *Bacillus* phage genomes. Of the 83 phage genomes, 70 genomes were myovirus and siphovirus morphology and 13 were podovirus. Examination of the 3,638 phams showed a significant number of phams that are highly conserved among closely related phages. For the purposes of this study we identified the top 50 most highly conserved proteins for detailed functional annotation using HHpred and blastp. Within the group of 50 highly conserved genes were 20 genes with no known function. The 50 conserve proteins of interest were located on the tree by deleting them from the data set and regenerating the plot **(Fig.3)**. When the plot was regenerated the center ‘trunk’ of the tree and much of the lattice networking among clusters was absent. Of these 50 conserved proteins our 20 unknown ORFs are found in 5 of the 7 myovirus clusters and in 28 to 43 phage genomes**(Fig.4)**. Establishing an assay to help identify the function of these 20 genes is the overall goal for this study.

The Bacillus phage Phrodo is a myoviridae with a double stranded DNA genome and a contractile tail **(Fig.5)** that was isolated by the SEA PHAGES program at VCU in 2014 using the host bacteria *Bacillus thrungunensis.* Phrodo is our experimental phage of choice because its DNA and genetic information is readily available and a vast majority of our genes of interest are found within Phrodo’s genome. These genes of interest are also found in majority of the 13 clusters, including cluster E where Phrodo is located. While cluster E is fairly diverse in comparison to the highly conserved cluster A, it is interesting that majority of the top 55 highly conserved proteins (including unknowns) are found in Phrodo’s genome and numerous other genomes found in cluster E. The diversity of this clusters makes it an interesting choice to focus on given its representation of the natural diversity of phage genomes. This observation could further support the speculation that proteins present in numerous genomes and clusters are essential to their viral life cycle.

Of the 50 highly conserved genes 5-7 genes will be selected as controls. Criteria for selecting controls will be based on its presence in Phrodo, ability to undergo PCR with appropriate primers, and having a well studied protein- protein interaction or host interaction by the scientific community. A list has been formulated for controls based on the criteria aforementioned**(Table 1)**. In addition to a list of controls, 15-20 unknown genes will be selected for overexpression given similar criteria. Both lists are our likely targets for this overexpression assay.

|  |
| --- |
|  Table 1. *Bacillus* phage proteins of interest  |
| phams | Frequency | Predicted function | Notes  |
| 1152 | 39 | RNA polymerase sigma factor |  |
| 3118 | 50 | Tail assembly chaperone |  |
| 269 | 40 | ssDNA binding |  |
| 1732 | 39 | Tape measure / tail fiber? |  |
| 393 | 41 | Thymidylate synthase |  |
| 106 | 28 | Ftsk / SpoIIIE |  |
|  | 28-42 | Unknown | Hypothetical proteins |

**Methods**

To begin the process of establishing overexpression assays in *Bacillus* bacteria, the control and functionally unknown genes must undergo polymerase chain reaction (PCR) and be cloned into appropriate expression vectors. Wells containing our *Bacillus* phage genes of interest are not readily available, so PCR is necessary in order to conduct this study. Phrodo DNA will be used for PCR due to its accessibility with all target genes being present.

The first step to PCR is primer design of each gene to be overexpressed. Primers are currently being designed for control genes and will be ordered from New England BioLabs. Guides and online tools for primer design used in this study can be found at New England BioLabs and Integrated DNA technology websites. Requirements for primer design include a section of the beginning and end of each gene (20-40bp in length) running 5’-3’ with a GC content of ~50% and a melting point of ~55oC. Requirements may be flexible given the naturally low GC content of *Bacillus* phages*.* In total, the primer design process of all genes for overexpression will ~2 weeks plus a few additional days for shipping.

Once primers have been designed and delivered the PCR process can begin. For PCR, phage DNA will be amplified using LongAmp *Taq* DNA polymerase from the New England BioLabs. This experiment will complete two-step PCR using methods from the Heidelberg European Molecular Biology Laboratory (EMBL). Step one of DNA amplification will involve combining 12a *ttB*1and12a *ttB*2 sites to the forward and reverse ends of each primer sequence with buffer and DNA polymerase and running 10 cycles of Denaturing, Annealing and Extension after two minutes of denaturing prior to beginning the cycles. Step two combines step one PCR product with 12a *ttB*1and12a *ttB*2 adapter primers with buffer and DNA polymerase, then denaturing and running 5+ cycles. The product of step two will be purified using a DNA clean-up kit. When genes have been successfully amplified they will be cloned and transferred to entry vectors in *E. coli* bacteria.

Overexpression assays will be used to gain an understanding of phage-host interactions by expressing various known and unknown proteins, as well as establishing thorough functional investigation of highly conserved phage proteins of unknown function. While T7 and Lambda have been studied in detail, little is known about other species of bacterial viruses. The *Bacillus* phages and their host provide an excellent unexplored model for further functional analysis of protein interactions in other species of bacteria and viruses. This study will begin the process of overexpressing unknown proteins in *Bacillus* phages using gateway clones containing our experimental proteins from PCR.

After the PCR product has been purified, a BP Clonase reaction will be performed to prepare entry vectors for transformation into chemically competent cells. PCR product is combined with TE buffer, pDEST14 entry vector and BP Clonase II enzyme. The following mixture is incubated over night at 25oC. The next morning protein Kinase K is added and incubated at 37oC briefly to stop the Clonase reaction. When incubation is complete the Clonase product is transformed into chemically competent E. *coli* cells. Chemically competent cells for E. *coli* have already been prepared and stored at -80oC in preparation for this experiment. Once cells containing entry vectors have completed multiple rounds of cloning a mini prep will be done using a Nucleic Acid and Protein Purification kit from Macherey-Nagel to isolate the plasmid DNA containing our experimental ORF. The isolated plasmids are then transferred to an expression vector by the LR Clonase II reaction. During this reaction the pDEST14 entry vector is combined with the pDG148 GW expression vector along with TE buffer and LR Clonase II enzyme. The mixture is then incubated overnight at 25oC and the following morning briefly incubated at 37oC after the addition of protein Kinase K to stop the Clonase reaction. The expression vector pDG148 GW will be used for this experiment because it is inducible with IPTG and the vector is compatible with both E. *coli* and *Bacillus* cells. The final step in the overexpression assay is transforming the Clonase product into chemically competent *Bacillus* bacteria. For overexpression to occur the bacteria will be plated on media containing IPTG to ensure ORFs are overexpressed in their host.

**Discussion**

This experiment will use an expression and entry vector system to overexpress *Bacillus* phage proteins in *Bacillus* bacteria. Expression and entry vectors will be used so that ORFs will be located in a plasmid and ready for use in future experiments, such as yeast two-hybrid assays. We have chosen pDONR/ Zeo **(Fig.6)** as our entry vector since it contains a phage T7 promoter and is compatible with E. *coli* bacteria. Majority of the prep work for this study will be done in E. coli bacteria because they are easy to grow and well studied. Chemically competent E. coli cells have already been prepared for transformations with plasmid DNA. Vector pDG148 GW **(Fig.7)** will be used as our expression vector since it is compatible with both E. *coli* and *Bacillus* bacteria. This vector contains the Pspac promoter for *Bacillus* bacteria upstream from the experimental ORF. This promoter allows up to induce and control the rate of ORF transcription in bacteria with the use of IPTG. IPTG will be used at various concentrations in media to observe any effects the rate of ORF transcription may have on the cell. This vector is actively being requested from the creator in preparation for this experiment and chemically competent Bacillus cells will be made prior to beginning the overexpression assay.

The purpose of this experiment is to establish assays to screen for any phage-host interactions that express phenotypes. Proteins that express interesting phenotypes will be selected for further functional analysis. Experiments planned for the future on unknown proteins that express phenotypes include performing Yeast 2 Hybrid and knockout assays. For Yeast 2 Hybrid and knockout assays majority of the preparation will already be complete. This experiment provides amplified Phrodo DNA from PCR, entry and expression vectors and competent *E.* *coli* and *Bacillus* cells.

This experiment will be conducted in the Uetz Lab room 333 in the Trani Life Science building. Lab work is expected to begin the week finals end and will continue throughout the summer. Hours spent in the lab will be ~5 hours depending on the wait time for individual experiments to incubate. Assays will be run as many times as necessary to verify phenotype results are correctly observed. When overexpression is complete other experiments will follow given the preliminary data generated from this experiment. The mid term project is planned to be turned in by July 14, approximately 9 weeks (**Fig. 8**) into the lab work.



 **Figure 8.** Visualization of work plan to identify highly conserved *Bacillus* phage proteins of

 unknown function for PCR and overexpression so that we may select candidates with interesting

 observations for future functional analysis. This plan is designed to reach two aims using the

 chronological plan of action shown above.

**References**

[ 1 ]Jakutytė, Lina et al. “Bacteriophage Infection in Rod-Shaped Gram-Positive Bacteria: Evidence for a Preferential Polar Route for Phage SPP1 Entry in Bacillus Subtilis.” *Journal of Bacteriology* 193.18 (2011): 4893–4903. *PMC*. Web. 23 Feb. 2016.

[2] Wommack, K. Eric, and Colwell, Rita R. "Virioplankton: Viruses in Aquatic Ecosystems." *Microbiology and Molecular Biology Reviews*64.1 (2000): 69.

[3] Pope, Welkin H, Charles A Bowman, Daniel A Russell, Deborah Jacobs-Sera, David J Asai, Steven G Cresawn, William R Jacobs, Roger W Hendrix, Jeffrey G Lawrence, and Graham F Hatfull. "Whole Genome Comparison of a Large Collection of Mycobacteriophages Reveals a Continuum of Phage Genetic Diversity." *ELife* 4 (2015): E06416.

[4] Maynard, Nathaniel D., Elsa W. Birch, Jayodita C. Sanghvi, Lu Chen, Miriam V. Gutschow, Markus W. Covert, and Ivan Matic. "A Forward-Genetic Screen and Dynamic Analysis of Lambda Phage Host-Dependencies Reveals an Extensive Interaction Network and a New Anti-Viral Strategy (Host Genetic Requirements for Lambda Infection)."*PLoS Genetics* 6.7 (2010): E1001017.

[5] Qimron, Udi, Boriana Marintcheva, Stanley Tabor, and Charles C. Richardson. "Genomewide Screens for Escherichia Coli Genes Affecting Growth of T7 Bacteriophage." *Proceedings of the National Academy of Sciences of the United States of America* 103.50 (2006): 19039-9044.

[6]

[7] Cresawn, Steven G., Matt Bogel, Nathan Day, Deborah Jacobs-Sera, Roger W. Hendrix, and Graham F. Hatfull. "Phamerator: A Bioinformatic Tool for Comparative Bacteriophage Genomics." BMC Bioinformatics 12.1 (2011): 395.