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

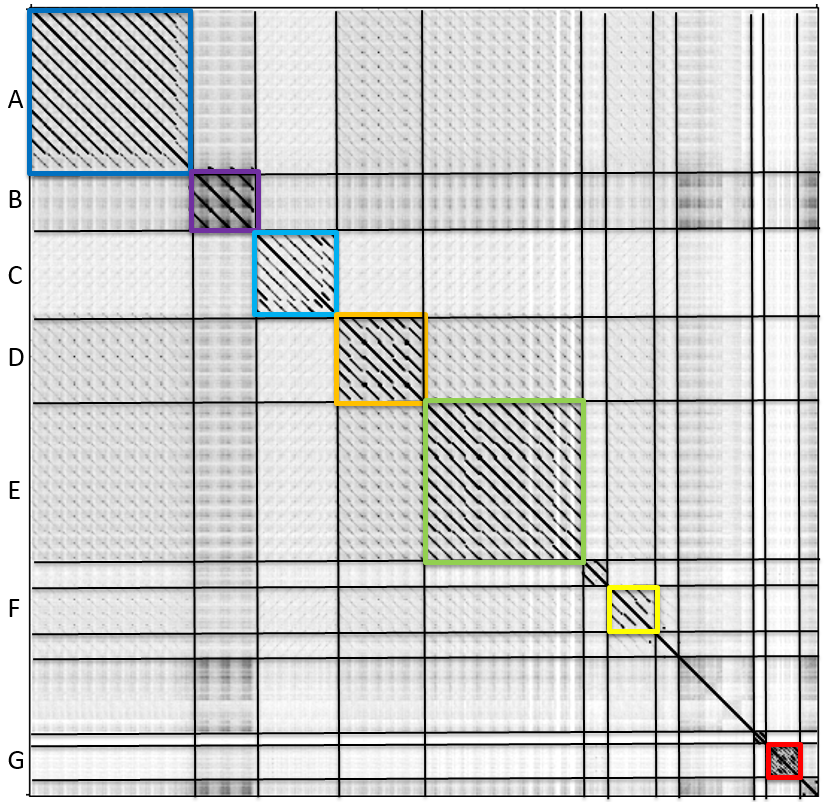
Bacteriophages (Phages) are the most numerous [1] and highly diverse biological entities in the biosphere. The genetic diversity of these biological entities gives rise to numerous novel genes. A recent comparative genomic study done in the Pope lab analyzed 657 genomes [2]. Of these genomes, 69,633 ORFs were identified and grouped into 5,205 phams, of which 1,613 (31%) were orphams. This suggests bacteriophages are not only highly diverse, but also contain an exuberant amount of unexplored genetic information. While much of their genetic information is unknown, many phages such as Lambda [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 in *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.

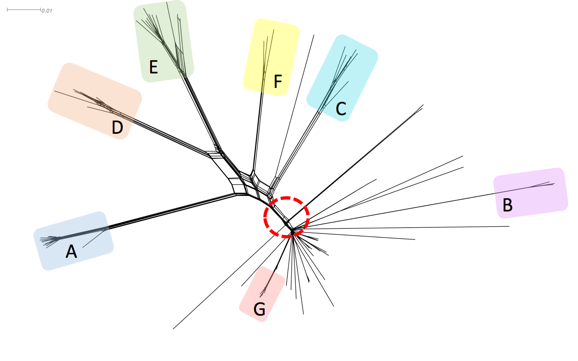
We will use an overexpression study by Jeroen Wagemans [5] and colleagues completed in 2014 as a model for our investigation of protein function. This study identified 26 proteins of unknown function found to be translated during the early bacterial infection process of *Pseudomonas* phages. Each of 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 (protein 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.

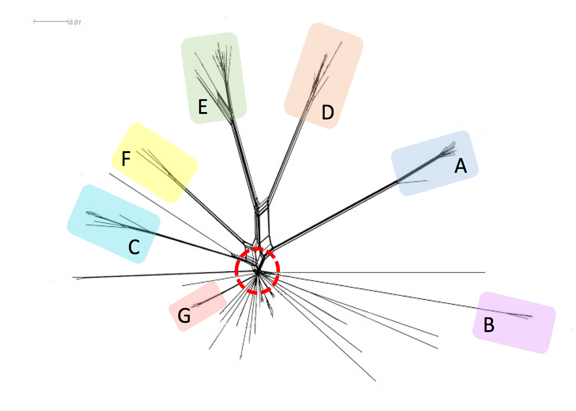
**Preliminary Data**,

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. Dot plot works by comparing each genome to itself and all other genomes. Lines with darker shading indicate a higher level of similarity. The thick dark line diagonally across the plot represents genomes compared to themselves. Cluster A containing 11 genomes, shows well defined darkened lines indicating a high level of conservation for sequence similarity, with the exception of one genome on the outer edge. Cluster E, also containing 11 genomes, has slightly darkened lines suggesting a lower level of sequence similarity compared to cluster A.

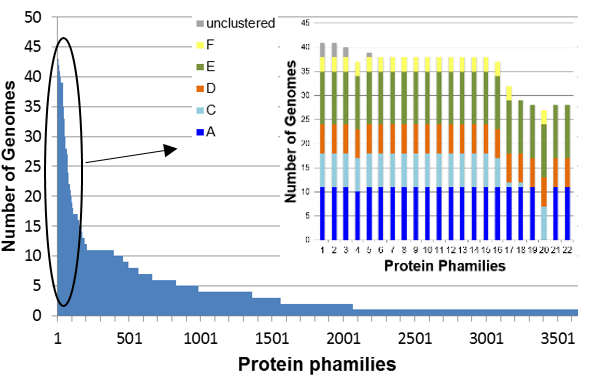
**Figure 1**. Dot plot image showing the genome sequence similarity of 70 myoviridae *Bacillus* phages. Group by sequence similarity reveals 7 distinct clusters.

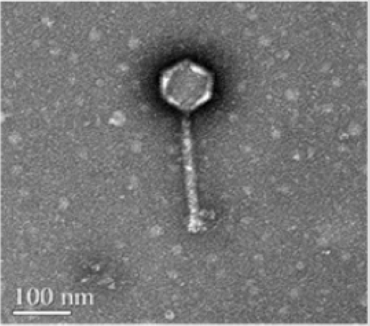
SplitsTree imaging was done to verify results observed from Dot Plot analysis (**Fig.2**). SplitsTree software organizes genomes into clusters, much like Dot Plot, however, this program compares genomes by shared protein content instead of sequence similarity. Interestingly enough, both Dot Plot and 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 within the cluster. Although the tree organizes majority of the genomes into distinct clusters with proteins uniquely shared within each cluster, the center of the tree shows an area where a large number of genomes share proteins in common across multiple clusters. 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.

**Figure 2.** SplitsTree image of 83 bacillus phage genomes grouped into clusters based on protein content similarity.

The comparative genomic tool Phamerator [7] 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 55 most highly conserved proteins for detailed functional annotation using HHpred and blastp. Within the group of 55 highly conserved genes were 22 genes with no known function. The 55 conserved proteins of interest were located on the tree by deleting them from the SplitsTree 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 55 conserved proteins our 22 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 22 genes is the overall goal for this study.

**Figure 3.** SplitsTree image of 83 bacillus phage genomes with the top 55 most highly conserved proteins absent.

**Figure 4.** Image shows the top 22 unknown proteins and the frequency of their presence in each cluster.

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 thuringiensis.* 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 still 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.

**Figure 5.** TEM image of phage Phrodo.

Of the 55 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 | DNA Binding |
| 3181 | 50 | Tail assembly chaperone | Binds to tape measure to ensure phage tail matures to the correct length [8] |
| 269 | 40 | ssDNA binding | DNA Binding |
| 1732 | 39 | Tape measure / tail fiber? | Structural |
| 393 | 41 | Thymidylate synthase | Nucleotide metabolism |
| 106 | 28 | Ftsk / SpoIIIE | *Bacillus* phage sporulation |
| \* | 28-42 | Unknown | Hypothetical proteins |

\*unknown phams: 1957, 1471, 3201, 2846, 3118, 58, 735, 1164, 1205, 1224, 2799, 2853, 2879, 1480, 385, 3625, 100, 162, 793, 1011

**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. Well plates containing our *Bacillus* phage genes of interest are not readily available, so PCR is necessary in order to conduct this study. Phrodo will be used for PCR as a DNA template.

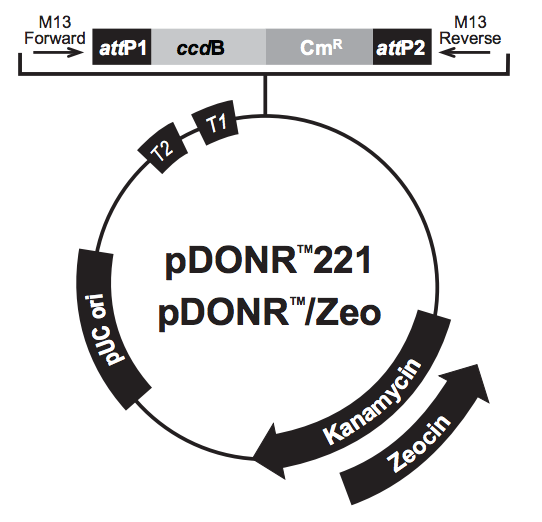
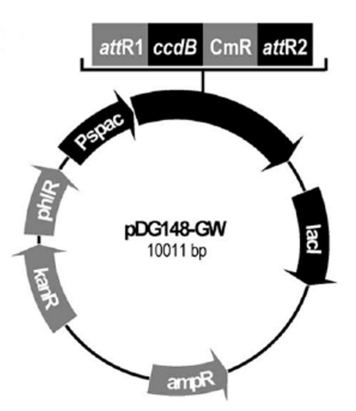
The first step to PCR is primer design of each gene to be overexpressed. Primers are currently being designed for control genes and will be ordered from New England BioLabs. Guides and online tools for primer design used in this study can be found at New England BioLabs and Integrated DNA technology 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*.*

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 12 bp long *attB*1and *attB*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 12 bp long *attB*1and *attB*2 adapter primers with buffer and DNA polymerase, then denaturing and running 5+ cycles. The adapter primers are added as an extension to ensure the ORF is not cut short during reactions involving ligase. 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.

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, pDONR / Zeo 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. The BP clonase reaction mimics phage-host relations. Naturally phage DNA has *att*P sites and the host has *att*B sites, when these sites combine the result is *att*L sites [9]. For BP clonase reactions the ORF product from PCR has *att*P sites and the empty plasmid contains *att*B sites, when the sites are combined *att*L sites form. In both situations the result is *att*L sites flanking the ORF and inserted into either a plasmid or bacterial genome.

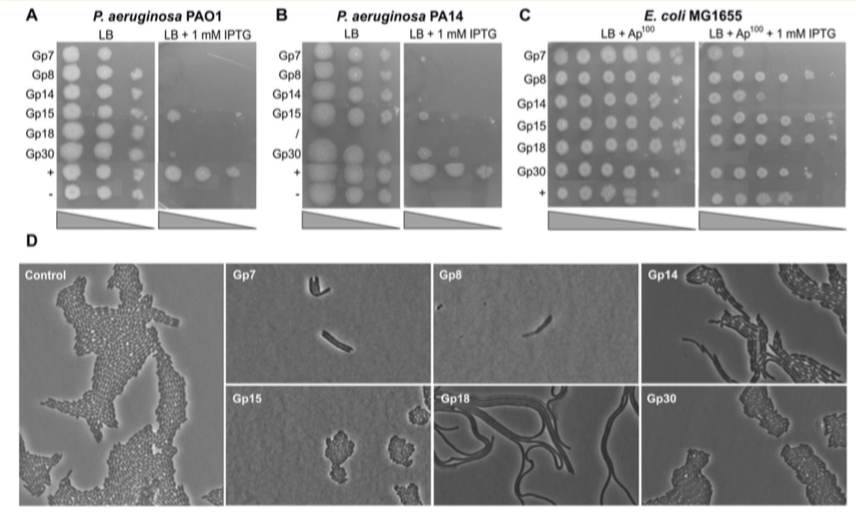
Once cells containing entry vectors have completed transformation into E. *coli* a mini prep will be done using a Nucleic Acid and Protein Purification kit to isolate the plasmid DNA containing our experimental ORF. The isolated plasmids are then transferred to an expression vector by the LR Clonase II reaction. During this reaction the pDONR / Zeo 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. LR clonase, similar to BP clonase, mimics phage – host interactions. When the LR clonase reaction occurs the ORF containing *att*L sites isolated from an entry vector is combined with the *att*R sites in the expression vector. Combining the *att*L and *att*R sites results in an expression clone with an ORF flanked by *att*B sites. When this process of combining *att*L and *att*R sites occurs naturally between phage and host the result is an *att*L site at the beginning of the ORF and an *att*R site at the end. Once LR clonase is complete the product is transformed into chemically competent *Bacillus* bacteria. For overexpression to occur the bacteria will be plated on media containing IPTG to induce ORF translation.

**Discussion**

Wagemans and colleagues used a shuttle vector system to overexpress their experimental proteins, just as we intend to do. However, their system results in a single copy of an ORF integrated in the host bacteria. This experiment will use an expression and entry vector system to overexpress *Bacillus* phage proteins in *Bacillus* bacteria. The result is a cell containing plasmids with an ORF instead of a single ORF integrated in the cells genome. We chose to use expression and entry vectors so we can control the rate of transcription with confidence that the bacterial genome is not interfering with the gene expression. By beginning with entry vectors we will have a vector containing our ORF that is easily transferable to other vectors for future experiments. We have chosen pDONR/ Zeo **(Fig.6)** as our entry vector since it contains a phage T1/T2 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 [10]. This vector is actively being requested from the creator in preparation for this experiment. It is compatible with both E. *coli* and *Bacillus* bacteria. Chemically competent *Bacillus* will be made prior to beginning gene overexpressions. This vector contains the Pspac promoter for *Bacillus* bacteria upstream from the experimental ORF. The promoter allows us to induce and control the rate of ORF transcription in bacteria with the use of IPTG. Media will contain various concentrations of IPTG to observe any effects the rate of ORF transcription may have on the cell. IPTG works by binding to promoters upstream from experimental ORFs, this allows RNA polymerase to bind to DNA and begin transcription. This process is modeled after the Lac promoter in E. *coli.*

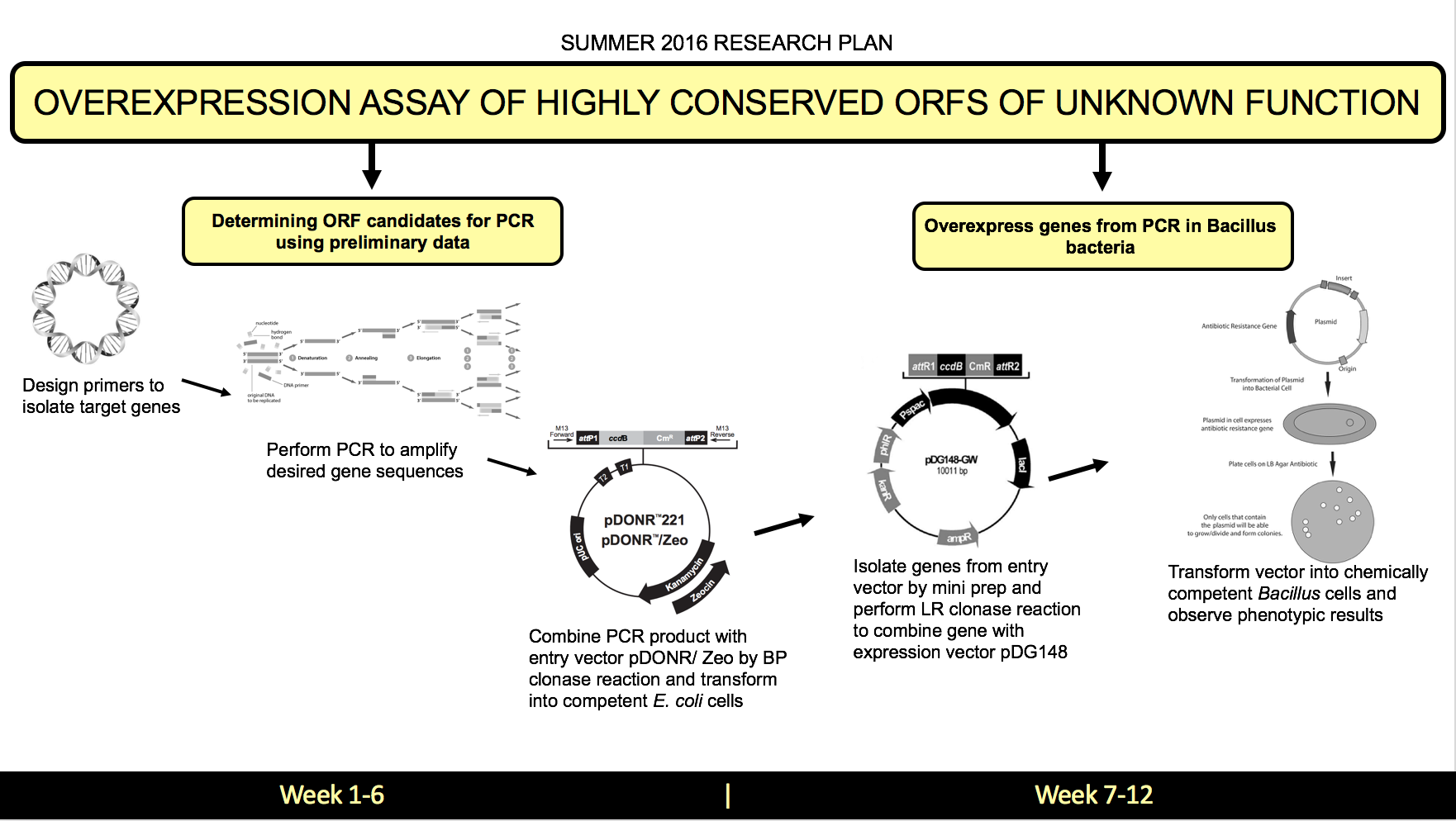
**Figure 6.** Visualization of pDONR/ Zeo vector and all the genes present.

**Figure 7.** pDG148 GW vector showing all genes it contains. ccdB and CmR is where the ORF is inserted.

The study our experiment is modeled after provides an excellent example for results we should expect from an overexpression assay. Of the 26 proteins they chose to over express, 6 of them (protein 7, 8, 14, 15, 18, 30) resulted in either cell growth inhibition or interesting phenotypes (**Fig. 8**). Wagemans and colleagues observed cells that could not properly divide after one correct division when protein 7 was overexpressed. The resulting cells were elongated with two nuclei. Protein 8, when over expressed, was observed to cause cells to begin cell division correctly but promptly stop mid division, resulting in attached daughter cells that burst several hours after division has halted. Both protein 14 and 18 induced in cells were observed to exhibit a long filamentous phenotype. While 4 of the 6 proteins cause morphological effects in cells, proteins 15 and 30 affect cells by inhibiting their growth. In our study we will screen for similar results when inducing proteins for overexpression. We aim to observe cell death, inhibited cell growth or abnormal cell division.

**Figure 8.** Image extracted from Wagemans et al. showing the effects 6 of their 26 proteins had on their host cells when overexpressed. Our study will screen for similar results.

The purpose of this experiment is to establish assays to screen for any phage protein-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 Yeast 2 Hybrid assays. For Yeast 2 Hybrid assays majority of the preparation will already be complete. This experiment provides amplified Phrodo DNA from PCR, entry and expression vectors containing experimental ORFs and competent *E.* *coli* and *Bacillus* cells.



**Figure 9.** 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] Wommack, K. Eric, and Colwell, Rita R. "Virioplankton: Viruses in Aquatic Ecosystems." *Microbiology and Molecular Biology Reviews*64.1 (2000): 69.

[2] 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

[3] 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.

[4] 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.

[5] Wagemans, Jeroen, Bob G. Blasdel, An Van Den Bossche, Birgit Uytterhoeven, Jeroen De Smet, Jan Paeshuyse, William Cenens, Abram Aertsen, Peter Uetz, Anne-Sophie Delattre, Pieter-Jan Ceyssens, and Rob Lavigne. "Functional Elucidation of Antibacterial Phage ORFans Targeting P Seudomonas Aeruginosa." *Cell Microbiol Cellular Microbiology* 16.12 (2014): 1822-835. Web.

[6]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.

[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.

[8] Xu, Hendrix, and Duda. "Chaperone–Protein Interactions That Mediate Assembly of the Bacteriophage Lambda Tail to the Correct Length."*Journal of Molecular Biology* 426.5 (2014): 1004-018. Web.

[9] "BxSeqTools Ultimate Molecular Cloning Guides - Gateway Cloning."*BxSeqTools Ultimate Molecular Cloning Guides - Gateway Cloning*. BioInfoRx, 2016. Web. 08 May 2016. <http://bioinforx.com/lims1/bxseqtools/ultimate-molecular-cloning-guides/help\_cloning\_gw.php>.

[10] Dobrijevic, Dragana, Gaetana Di Liberto, Kosei Tanaka, Tomas De Wouters, Rozenn Dervyn, Samira Boudebbouze, Johan Binesse, Hervé M. Blottière, Alexandre Jamet, Emmanuelle Maguin, and Maarten Van De Guchte. "High-Throughput System for the Presentation of Secreted and Surface-Exposed Proteins from Gram-Positive Bacteria in Functional Metagenomics Studies." *PLoS ONE* 8.6 (2013): n. pag.