PROJECT TITLE: Harvesting Viral Genomes from Sequences of Complex Communities

PARTICIPANT NAMES: Andrea Garretto and Ally Miley

PROJECT DESCRIPTION:

Introduction to Bacteriophages. The most abundant organisms on Earth are viruses, most notably bacteriophages or viruses that infect bacteria (Clokie et al. 2011). Viruses are infectious agents consisting of DNA or RNA surrounded by a protein coat. When a bacteriophage (phage) infects a host cell, they interject their own genetic material into the bacterium and use the bacterium to replicate their own DNA (or RNA). Phages can then rupture the cell wall, releasing new progeny to go on and infect other susceptible host cells (the lytic life cycle), or integrate their DNA within the host (the lysogenic life cycle). By entering the lysogenic cycle, a phage uses the host to replicate its own viral DNA whenever the host replicates. The integrated phage (also referred to as a prophage) lies dormant until environmental cues induce the phage to separate its DNA from the host, reassemble, and lyse out of the bacterial host cell, therefore re-entering the lysogenic life cycle. The consequence of lyses has profound effects on structuring of bacterial communities (Clokie et al. 2011; Jacquet et al. 2010). Phages mediate host mortality on a global scale (Berdjeb et al. 2011) and drive bacterial genetic diversity (Winget et al. 2011) in nature. Given their importance, phages have been explored within multiple diverse environments on Earth, e.g. polar icecaps (Yu et al. 2015), soil (Srinivasiah et al. 2015), and even Lake Michigan (Watkins et al. 2015). However, the amount of identified phages is significantly limited compared to the great abundance of the organisms. The number of bacterial classifications far outweighs viral due to the difficult nature of phage isolation. This obstacle restricts the amount of viral information available in genetic databases which hinders further investigation. Additionally, viral data that is available tends to be riddled with inaccuracies.

Introduction to Viral Genomics: DNA sequencing of viral genomes can be conducted for: 1) an isolated virus; 2) a viral culture; or 3) a viral community. The distinction here is that in the first case, DNA is representative of *only* the viral species, whereas the second case may include DNA from the virus' host species. Ideally we would be able to isolate the DNA of a single viral species such that all sequences generated were representative of this one, individual virus. However, this is often not the case; viruses cannot replicate without their host. Frequently host DNA "contaminates" viral genome sequencing efforts (Hatzopoulos et al. 2016). Moreover, studies which focus on a viral isolate are limited in scope. The vast majority of viruses cannot be cultivated in the lab. Thus sequencing of complex communities which include many different viral species, e.g. the entire viral population within the human gut, provide a unique opportunity to characterize viruses never before seen. Computationally speaking, however, analysis of this latter case is far more challenging than the first two. This is the focus of our proposal. Some of the tools developed for the sequencing of single viral species samples, however, can be utilized here as can tools created for complex bacterial communities (when adaptations are made specific for viral analyses). Figure 2 illustrates the process – from biological sample through analysis – for all three types of viral sequencing projects.

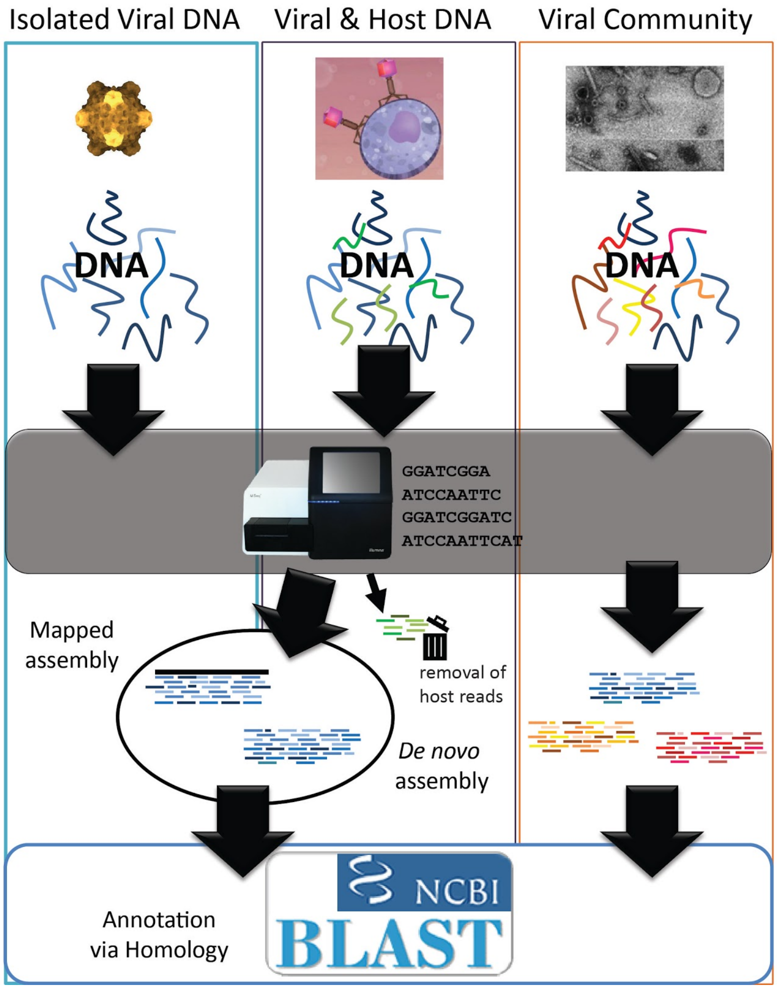

Figure 2. Workflow of viral sequencing projects.

Assembly. Sequencing reads can be assembled into longer sequences called contigs (perhaps whole gene sequences or complete genomes) via one of two approaches. The first approach is via mapping. Mapping strategies (tools such as Bowtie (XXX)) take the raw sequencing reads and compares them to a reference genome sequence. Thus a complete genome sequence can be recreated by the use of a reference genome. This approach works well when the complete genome of the virus is known, or a near-neighbor (close relative) is known. When there is no reference genome, as is the case for the vast majority of viral species on earth, *de novo* sequence assembly must be conducted. There are a large number of *de novo* assemblers used, some more popular than others. Here we will discuss two which we have used and found success with when assembling viral genomes.

- ***Velvet:*** One of the more popular, albeit older, de Bruijn graph assemblers used to process NGS *de novo* data is Velvet (Vázquez-Castellanos et al.). Generally, *k*-mers (substring of length *k*) are used to reconstruct a string of nucleotides. This process of reconstructing a string of nucleotides is called fragment assembly, which leads to the use of de Bruijn graphs in the next steps of assembly (refer to Pevzner et al. for more information on de Bruijn graphs). Velvet can handle relatively short, paired reads of bacterial genomes with N50 contig lengths of up to 50 kb. One of the more novel features of Velvet at the time it was created was its ability to focus on topological features: tips, bulges, and erroneous errors. Focusing on these features removes the need to rely on differences between expected coverage of true sequences, and coverage of random errors, which permits errors to be distributed randomly in the reads. Using velvet, tips, bulges, and erroneous errors are removed consecutively (Zerbino and Birney).

- ***SPAdes:*** SPAdes is a recently developed assembler. It too uses de Bruijn graph, however, it uses bi-kmers. *k*-mers are used in SPAdes is for the initial building of de Bruijn graphs, after which they are dismissed. There are four stages in total for SPAdes assembly detailed by Bankevich *et al*. Stage one is assembly graph construction, which is a de Bruijn graph deconstruction, similar to bulge or bubble removal in Velvet. The SPAdes version of the de Bruijn graph deconstruction employs new bulge removal algorithms, finds and deletes chimeric reads, combines biread information into histograms identifying their distance, and permits backtracking on graph operations. Next follows *k*-bimer adjustment, in which the distances between k-mers are estimated using analysis of the distance histograms, and the paths in the assembly graph. Then the paired assembly graph, inspired by the original PDBG (paired de Bruijn graphs) approach, is constructed. The final stage involves the construction of DNA sequences of contigs, and mapping reads to contigs through backtracking simplifications of the graph.

Annotation. The sequences generated by any one of the various assemblers listed are predominantly annotated by BLAST (Basic Local Alignment Search Tool). BLAST programs are used for going through protein and DNA databases and searching for sequence similarities. BLAST works by taking sequences of nucleotides or proteins of a given species, and compares them against any other species specified. By looking at the similarities as well as differences in the sequences, commonalities in function and ancestry between varying life forms can be identified. When using BLAST, it is important to keep in mind that an increased threshold parameter will result in an increase in speed, but also an increase in the probability of missing weaker similarities (Altschul et al.). Thus for annotation and analysis of viromic data, the contigs created by the assembler are compared to known sequences via BLAST; the contigs are queried against sequence data repositories identifying similarities representative of the taxa (species) and/or functionality present within the sequenced isolate/community. This methodology, however, is completely dependent upon the data available -- which as we've previously mentioned is scant for viruses. Nevertheless, this is the only real approach available.

Hypothesis: We believe that complete viral genomes can be harvested from available sequence data of complex communities. As detailed here, this is computationally challenging given the size of available data and the unknown characteristics of the harvested genomes. We propose to develop a pipeline to facilitate this process.

Proposed Methods:

We propose to develop a pipeline that will incorporate many of the existing tools available as well as integrate new functionality tailored to the analysis of viral genomes. A key feature of this tool will be the minimized human interaction and the fact that it will encompass the entire process, from raw sequence data through analysis. Presently there are not many, in fact there are just two - Metavir and VIROME, viral genome tools available, particularly for complex communities. One tool, Metavir, takes in contigs and analyzes each by BLASTing it against a viral genome database. Ascertaining if complete genomes are present is not straight-forward using this tool. VIROME follows a similar approach, although it uses a different database.

The proposed pipeline will integrate SPAdes and Velvet, utilizing parameters of each which will be fine-tuned for viral genome construction. The output of this assembly process will be piped into gene prediction. Predicted genes will then be compared to publicly available annotated gene sequences as well as other viral data sets. While the former will be analogous to BLAST searches (and we will in fact use BLAST), this latter comparison will be a new development. As the vast majority of viral sequence data shows no sequence homology to viral strains characterized within the laboratory, we anticipate that most predicted coding regions will have no BLAST hits. However, we can begin to scratch the surface of this "unknown" by asking, "Is this gene likely viral?" Thus the predicted genes will be compared to existing viral datasets. This may also identify uncharacterized genes of interest -- genes that are frequently found in viral communities, but for which we do not know what their function is.

Mentorship and Plan of Work: The three members of the team will meet weekly for the express purpose of discussing progress on the proposed project. During these meetings, progress as well as challenges will be discussed and plans for the coming week's efforts will be determined. In addition to these meetings, ad hoc meetings with either the entire team or one student and the mentor will be frequent. Andrea and Ally will each keep a research notebook documenting their work. They will also create and maintain an online blog detailing their research progress and links to relevant literature and/or online resources.

References:

1. Berdjeb L, Pollet T, Domaizon I, Jacquet S. 2011. Effect of grazers and viruses on bacterial community structure and production in two contrasting trophic lakes. BMC Microbiol 11, 88.
2. Clokie MRJ, Millard AD, Letarov AV, Heaphy S. 2011. Phages in nature. Bacteriophage 1, 31-45.
3. Jacquet S, Miki T, Noble R, Peduzzi P, Wilhelm S. 2010. Viruses in aquatic ecosystems: important advancements of the last 20 years and prospects for the future in the field of microbial oceanography and limnology. Advances in Oceanography and Limnology 1, 97-141.
4. Srinivasiah S, Lovett J, Ghosh D, Roy K, Fuhrmann JJ, Radosevich M, Wommack KE. 2015. Dynamics of autochthonous soil viral communities parallels dynamics of host communities under nutrient stimulation. FEMS Microbiol Ecol 91, pii: fiv063.
5. Watkins SC, Kuehnle N, Ruggeri CA, Malki K, Bruder K, Elayyan J, Damisch K, Vahora N, O'Malley P, Ruggles-Sage B, Romer Z, Putonti C. 2015. Assessment of a metaviromic dataset generated from nearshore Lake Michigan. Marine Fresh Res, In Press.
6. Winget DM, Helton RR, Williamson KE, Bench SR, Williamson SJ, Wommack KE. 2011. Repeating patterns of virioplankton production within an estuarine ecosystem. Proc Natl Acad Sci USA 108, 11506-11511.
7. Yu Z-C, Chen XL, Shen QT, Zhao DL, Tang BL, Su HN, Wu ZY, Qin QL, Xie BB, Zhang XY, Yu Y, Zhou BC, Chen B, Zhang YZ. 2015. Filamentous phages prevalent in *Pseudoalteromonas* spp. confer properties advantageous to host survival in Arctic sea ice. ISME J 9, 871-881.

STUDENT ACTIVITIES & RESPONSIBILITIES:

Each student will be responsible for their:

- Contribution to the survey of existing computational tools
- Engagement in the design of software tool
- Attendance at weekly 1-hour group meetings. Each student will present her progress once a month at one of these meetings. Other weeks will be dedicated to discussion of progress in a more informal setting and/or discussion of relevant literature (akin to a Journal Club).
- Participation in Loyola Undergraduate Research Opportunities Program (LUROP) workshops. LUROP hosts a number of events each year, e.g. searching for Graduate Schools, graduate school applications, workshops for preparing research posters, workshops for writing abstracts. The applicants will attend applicable sessions.
- Maintaining a lab notebook including all details of their research
- Maintaining a weekly blog accounting their research progress.

Each student will be required to devote 10-15 hours per week on the project. Before starting our research project, the students will attend Loyola's Responsible Conduct in Research and Scholarship training. A critical component of research is presentation of one's research. In addition to the presenting to the group, each student will also participate in area undergraduate symposia. In the Spring semester, Andrea and Ally will present their research at the Chicago Area Undergraduate Research Symposium (CAURS) as well as the Loyola University Research Symposium. In May 2017, Andrea and Ally will present to the scientific community at the Great Lakes Bioinformatics (GLBio) Conference; this will also give them the opportunity to interact with graduate students, post-docs, and faculty in the field.

Team Members Include:

Andrea Garretto**:** Andrea began working in the Putonti lab in Summer 2015. Since joining the lab, she has worked at the bench with several different phages. She has also assisted in the assembly and annotation of the genomes of phage isolates. She has worked through the software platform Geneious which provides a user-friendly GUI to many of the assembly and annotation tools to be examined here. In addition to her work in the Putonti lab, during the Spring 2016 semester, she also worked under the mentorship of Dr. Qunfeng Dong at Loyola's Health Sciences Campus. There she worked in collaboration with another Bioinformatics undergraduate student investigating tools for bacterial genomics. Andrea is a Bioinformatics major also minoring in Biostatistics.

Ally Miley: Ally began working in the Putonti lab in January 2016. Her work in the laboratory has thus far been primarily molecular. She has mastered the skills of DNA extraction and PCR and preparation of DNA for sequencing. In parallel to her work in in the Putonti lab, during the Spring 2016 semester, she was also enrolled in a course Bioinformatics Survey. As part of this course, Ally was introduced to many publicly available Bioinformatics software and data repositories, including the National Center for Biotechnology Information (NCBI) GenBank. Ally is a double major in Bioinformatics and Computer Science.

FACULTY ACTIVITIES & RESPONSIBILITIES:

Dr. Putonti is a computational biologist with experimental expertise with phages. Her lab has previously isolated, sequenced, and characterized phages from the environment. In addition, a manuscript from her lab is currently under review in which 9 phage genomes were harvested from publicly available complex data sets. This prior work is a key proof-of-concept requiring several months of work. The proposed pipeline will facilitate such discoveries in a fraction of the time. Dr. Putonti will assist the team in identifying key tools as well as programming; she has been a C++ programmer for decades and Python programmer for several years. In addition to providing computational support, Dr. Putonti is well versed in bacteriophage genomics and genetics. Dr. Putonti has a space dedicated for computational biology research in the Life Science Building at Loyola University Chicago's main campus. This room has a number of machines with varying operating systems and architectures. Andrea and Ally will have onsite and remote access to this space and its resources.

HOW THE PROJECT CONTRIBUTES/ADVANCES THE MISSIONS OF THE CREU PROGRAM:

IMPACT ON THE GOAL OF CREU: This proposal speaks to the goals of the CREU program in several ways. First, the entire team is female. The students will get to work closely with women whose professions rely on the power of computing. Loyola has an active Women in Science and Mathematics club. Although not a requirement, the students will be encouraged to attend activities sponsored by this group. Dr. Putonti was a co-PI on a previously funded NSF project, "Collaborative Research: BPC-A: Improving Metropolitan Participation to Accelerate Computing Throughput and Success", and as a result has connections to several area high schools. In the Spring 2015 semester, the students will visit an area high school to discuss research in computer science. Second, many of the university workshops (see link in Student Activities and Responsibilities Section) available to the students are geared towards preparing and exposing undergraduate students to graduate programs. Last, the program is interdisciplinary in nature. With the recent explosion of biological data being produced, bioinformaticians and computational biologists are in high demand. This is an extremely fast growing field. The recently published "Bioinformatics Market by Sector, Segment & Application – Global Forecasts to 2017" (by Research And Markets) estimates that the bioinformatics market is poised to expand 20.9% between 2012 and 2017. The report goes on to say that this growth is subject to several factors, most notably a possible "dearth of skilled personnel". Many of these jobs require post-baccalaureate training. The CREU sponsored experience will enrich Andrea and Ally's formal training, providing them with skills that will be of great value to future employers as well as graduate programs.