Project 1 Proposal: Genome Assembly of *Aliivibrio fischeri*

**Background**

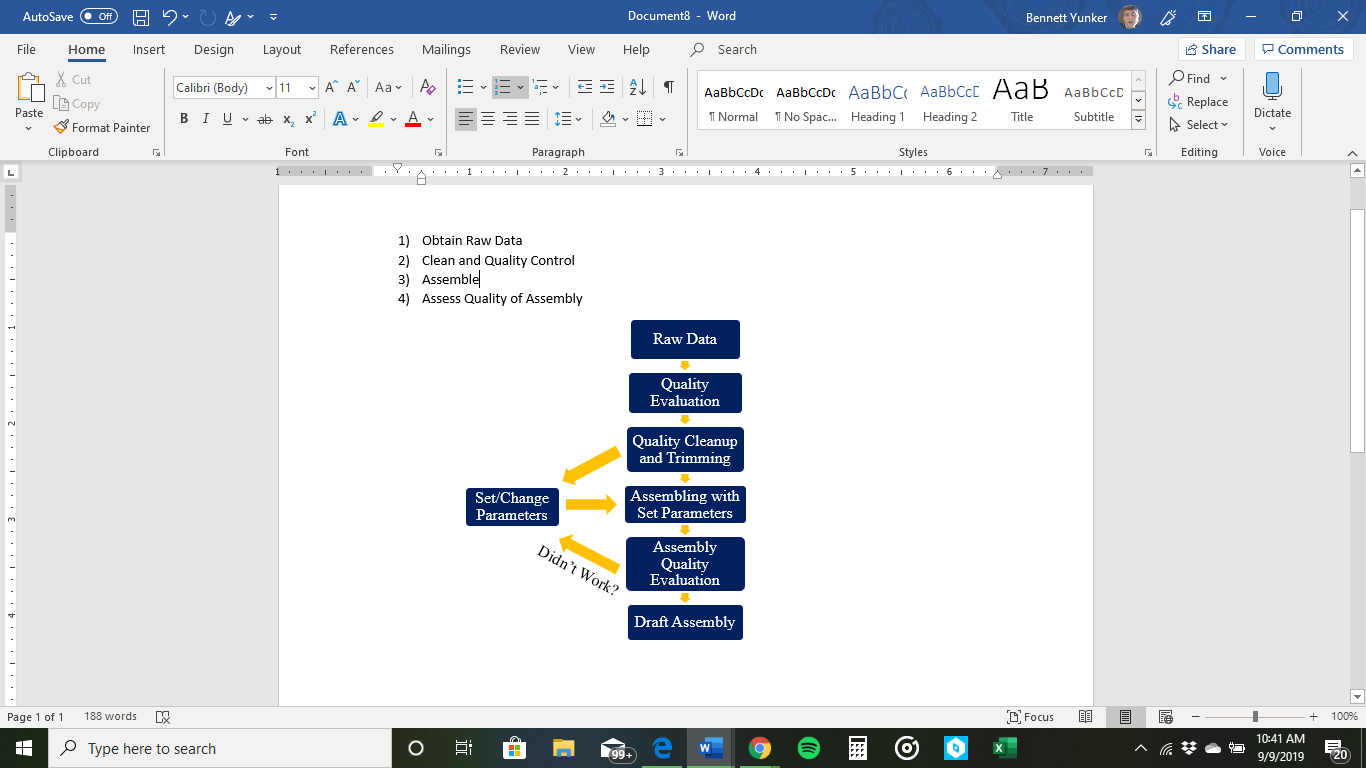
*Aliivibrio fischeri* (also referred to as *Vibrio fischeri*) is the bacterial symbiont of the Hawaiian bobtail squid *Euprymna scolopes*. These bacteria colonize a light organ within *E. scolopes* that then emits *A. fischeri*’s bioluminescence while the squid feeds nocturnally. The luminescence breaks up the squid’s profile against the night sky, protecting it from predation. The colonization by *A. fischeri* requires the squid to select against other environmental bacteria. Once *A. fischeri* has established itself in the crypts of the light organ, the bacteria induce morphological changes in its host to create a mature light organ (Nyholm and McFall-Ngai, 2004). Elucidating the regulatory networks involved in each of these processes may shed light on the genetic mechanisms by which symbioses evolve and are maintained. While the signaling pathway during early selection and colonization has begun to be characterized, the regulatory network underlying the morphological response is largely unknown.

Data provided through genome assembly may lead to information on how *A. fischeri* and its host have co-evolved as well as shed light on genes relating to symbiosis and the crosstalk between organisms. Symbiosis is an emerging research frontier; we are only beginning to understand the ways in which microbes influence our own gene expression. The ease of laboratory culture and non-obligate nature of the *Aliivibrio-Euprymna* symbiosis establishes this system as a promising model for studying various aspects of symbiosis in other bacterial-animal systems. The availability of a high-quality genome may also benefit the medical field, as the genome of *A. fisheri* can be compared to that of its pathogenic relatives including *Vibrio cholerae, Vibrio parahaemolyticus*, and *Vibrio vulnificus (*Ruby et al, 2005*).*

The *A. fischeri* genome had been sequenced and assembled in 2005 by Ruby *et al* and then resequenced in 2008 to resolve errors (Mandel et al, 2008). In 2019, a *A. fischeri* genome was published utilizing Illumina Next Generation Sequencing and was assembled using the SPAdes assembler. However, a higher quality and more complete assembly may be achieved through using a different assembly tool, such as ABySS. We aim to take a large, high coverage set of reads from the 2019 *A. fischeri* Illumina sequencing data set and assemble them *de novo* into a genome using the ABySS genome assembler and the SPRUCE computer cluster at the WVU High Performance Computing (HPC) center.

**Methods**

The raw Illumina Sequencing data was obtained from the NCBI Sequence Read Archive under accession SRR8647324. The data consists of 250bp paired-end reads with a total size of 529.2M bases. Based on previously published genome assemblies, the complete *A. fischeri* genome consists of two circular chromosomes with a combined total length of 4.2Mbp (Ruby et al, 2005). The raw data utilized for this assembly should provide approximately 125x coverage for the expected genome size, allowing for flexibility during initial quality control processing of sequence data and the potential to discard a higher proportion of reads in case of poor read quality.



*Figure 1: Assembly Pipeline*. General analysis flow will depend on various outputs from the data inspection, quality control, and assembly output steps.The above flowchart represents the expected pipeline approach.

Quality control will be conducted with the QUAST quality assembly tool (Gurevich et al, 2013). Before assembly, our data will be inspected for read quality, sequence degradation, and contaminants that may be represented as highly common k-mers. Trimming approach will depend on QC output. Leading and trailing base trimming will be used to excise poor quality regions from the beginning or ends of sequences, respectively. Alternatively, if initial inspection reveals poor base call quality intermittently throughout reads, sliding window trimming may be utilized instead to excise low quality regions throughout the sequence. After trimming functions are completed, orphaned reads will be partitioned and short fragments will be thrown out if below a specified minimum read length. The coverage of this data set will allow for more rigorous trimming and read length parameters will still maintaining the read depth needed to assemble the genome.

For *de novo* assembly, we will be using ABySS 2.0 (Jackman et al, 2017). ABySS (Assembly By Short Sequences) is a two-stage assembly algorithm consisting of generating k-mers from sequence reads and extending contigs using pair-mate information (Simpson et al, 2009). ABySS was initially chosen based on a review paper by Khan (2018). For a prokaryotic, paired-end Illumina NGS data set, Khan states that Velvet and ABySS provide the best results in terms of memory usage, assembly time, and efficiency. ABySS assemblies provided a higher N50 contig length than other assembly methods, suggesting increased accuracy. While Velvet was superior in sequencing time and memory usage, the small size of the data set will create only a marginal, acceptable increase in assembly time in ABySS. As the *A. fischeri* genome has been assembled previously, we prioritized accuracy over efficiency. Institutional support through WVU HPC/Spruce will support more demanding assembly methods. In addition to increased accuracy, ABySS assemblies had a 66.3% genome fraction while Velvet results had a 57.1% genome fraction, suggesting less misaligned bases occurred in a genome while assembling with ABySS.

ABySS was also compared to SPAdes in a GAGE-B assessment for bacterial genome assembly by Magoc et al (2012). This study compared multiple open source assemblers against five prokaryotic data sets at both 100bp and 250bp read lengths. While the summarized data suggests SPAdes generally outperforms ABySS, the analysis using 250bp MiSeq data showed that ABySS did outperform SPAdes in one out of five genome assemblies as estimated by N50 values (Magoc et al, 2012). The authors also supported Khan’s assertion that ABySS produced one of the lowest error rates across multiple prokaryotic data sets and both read lengths. Although SPAdes may be the optimal assembler for another *Vibrio* species- *V. cholerae*- the data had lower coverage and a higher %GC content than our selected data. Under more optimal conditions and a different read environment, this may not be the case for *A. fischeri*. To test this, we are proceeding with ABySS assembly.

During our pipeline analysis, we will run multiple assemblies at various *k*-values appropriate for our final average read length and evaluate the output quality of each using QUAST’s adjusted N50 values before finding an optimal *k* for final assembly. After assembling the genome, we will run a final analysis of the contigs to evaluate quality through N*A*50 and genome fraction percentages. Gaps in scaffolding can be filled with additional software such as GapFiller (Boetzer and Pirovano, 2012).

*Division of Labor*

Naomi Williamson and Ellie Spahr prepared the proposal document, with contributions from Mason Tatro and Ben Yunker for identifying an assembler and methods section. Mason and Ben will lead the analysis portion and Naomi and Ellie will lead the final paper, with everyone contributing to the other sections.

**Expected Outcomes**

We expect that our *de novo* assembly with ABySS will provide a highly accurate construction of the *A. fischeri* genome. The final assembly is expected to be 4.2Mbps with two circular chromosomes (2.9Mbp and 1.3Mpbs, respectively) (Ruby et al, 2005). Due to potential repeat-rich regions, we may end up with more than two scaffolds, as this is a common shortcoming of ABySS but a common element within genoms. Additional challenges for ABySS involve accommodating orphan reads, gaps, and sequence misalignments. While we plan to partition orphaned reads during initial data processing, we must consider gaps and misalignments after genome assembly. Software to fill gaps includes GapFiller, which is an automated strategy that uses paired reads to close gaps in scaffolds (Boetzer and Pirovano, 2012). QUAST compares generated data to existing, related genomes to identify misassembled contig lengths. (Gurevich et al, 2013) By employing these strategies, we hope to compensate for the anticipated assembly challenges.

**References**

Boetzer, M., & Pirovano, W, Toward almost closed genomes with GapFiller. *Genome biology*, *13*(6), R56. (2012).

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G., QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075. (2013)

Jackman, S. D., Vandervalk, B. P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S. A., ... & Birol, I., ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome research*, *27*(5), 768-777. (2017)

Khan AR, Pervez MT, Babar ME, Naveed N, Shoaib M. A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective. Evolutionary Bioinformatics.

Nyholm, S. V., & McFall-Ngai, M., The winnowing: establishing the squid–Vibrio symbiosis. *Nature Reviews Microbiology*, *2*(8), 632. (2004).

Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., ... & Salzberg, S. L., GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics*, *29*(14), 1718-1725. (2013).

Mandel, M. J., Stabb, E. V., & Ruby, E. G., Comparative genomics-based investigation of resequencing targets in Vibrio fischeri: focus on point miscalls and artefactual expansions. *BMC genomics*, *9*(1), 138. (2008).

Ruby, E. G., Urbanowski, M., Campbell, J., Dunn, A., Faini, M., Gunsalus, R., ... & Schaefer, A., Complete genome sequence of Vibrio fischeri: a symbiotic bacterium with pathogenic congeners. *Proceedings of the National Academy of Sciences*, *102*(8), 3004-3009. (2005).

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I., ABySS: a parallel assembler for short read sequence data. *Genome research*, *19*(6), 1117-1123. (2009).