**PROJECT 2 Proposal**

**Group: nkob**

**Members: Heather Baldwin, Ida Holásková, Denzel Middleton**

**Date Oct 22, 2019**

**Genome architecture analysis of *Listeria monocytogenes***

**BACKGROUND**

*Listeria monocytogenes* *(L. monocytogenes)* is a pathogenic gram-positive bacterium that is known as one of the most virulent foodborne pathogens (*1*). Listeriosis in humans and ruminants is the root of abortion, septicemia, gastroenteritis, and central nervous system (CNS) infections (*2*, *3*). *L. monocytogenes* strains are grouped into four distinct phylogenetic lineages, I, II, III and IV (*4*, *5*). In general, the *L. monocytogenes* strains of lineage I are associated with CNS infections in humans and animals, while strains belonging to lineage II typically clusters food and environmental strains (*6*).

Intention of this experiment is to map sequenced reads from *L. monocytogenes* (LMNC088) lineage II strain to *L. monocytogenes* (JF5861) lineage I reference strain in order to identify single nucleotide polymorphisms (SNP) in genomic structure. The significance of this study is to identify genetic variability between lineage I and lineage II strains of *L. monocytogenes.* The purpose isto identify possible genomic characteristics in an environmental lineage II strain that may differentiate it from the human-virulent lineage I strain using comparative genomic tools. The innovative segment of this experiment is the use of two lineages of listeria; a listeria strain with a smaller genome due to lost genes will be mapped to a listeria strain with more genes, that also happens to be more virulent.

**METHODS**

**Data Availability and Retrieval**

In order to address the genetic separation of *L. monocytogenes*, whole genome sequencing (WGS) was conducted on isolates of *L. monocytogenes*. Genome mapping will be performed *in silico* by utilizing a Illumina HiSeq sequencing dataset for *L. monocytogenes* (Accession: [ERX2192466](https://www.ncbi.nlm.nih.gov/sra/ERX2192466%5Baccn%5D)) available from the NCBI Sequence Read Archive (SRA) in fastq.gz file format. Paired end sequencing was conducted on a *L. monocytogenes* isolate using Illumina HiSeq and resulted in a data file consisting of 2.5M spots, 755.8M bases, and 336.3Mb of downloads. The *L. monocytogenes* dataset is paired-end sequencing which is preferred over single-end sequences due to higher quality and accuracy-alignable sequence data (*7*).

The *L. monocytogenes* reads ([ERX2192466](https://www.ncbi.nlm.nih.gov/sra/ERX2192466%5Baccn%5D)) will be mapped to the reference genome, *L. monocytogenes* (Accession: JF5861 complete genome) which is available from NCBI’s Assembly platform in the .fna file format. This is an assembled, complete genome with full genome representation, submitted by University of Bern (Switzerland). The reference genome consists of one chromosome and total ungapped length of 2,913,696 bp. The preliminary coverage calculation for lineage II raw data revealed 260x genome coverage (2,512,294\*151\*2)/2,913,696) with respect to the reference genome (lineage I strain), which should be sufficient depth for mapping and to determine unbiased statistical estimates of genomic variability.

**Experimental Software Pipeline**

The workflow of raw data processing will be carried out via the High-Performance Computing center (Spruce Knob), available at West Virginia University. The experimental software pipeline is demonstrated in Figure 1.

*Fastqc (v0.11.7)* and *Trimmomatic (v0.38)*

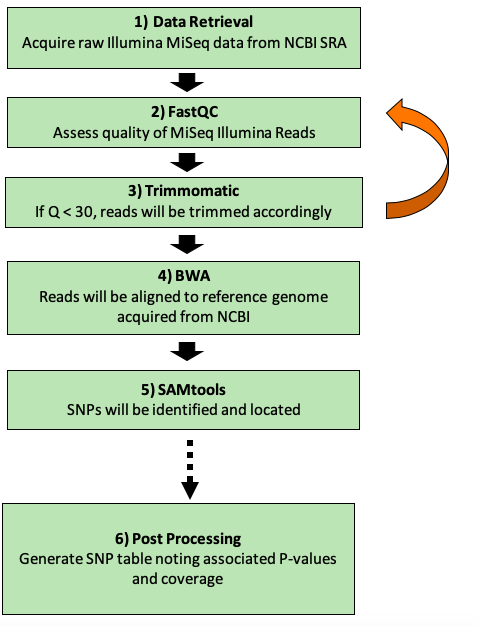
To evaluate the quality of the *L. monocytogenes* dataset, FastQC will be used to assess the quality of the raw data as FastQC is both available on Spruce and is a widely accepted application for quality control (*8*). The resulting FastQC file will report the per base sequence quality, sequence content, N content, sequence duplication levels, overrepresented sequences, and adapter content of the dataset. Trimmomatic will be used to trim the dataset as it is specifically designed for paired-end data and available on Spruce (*9*). The preliminary Fastqc report exhibited that reads decline in quality (Q < 30) near the 3’ end. Therefore, Trimmomatic will be utilized to trim the 3’ end of reads to 140 bp initially and if necessary, further trimming parameters will be explored.

*BWA (v0.7.17)*

Following quality control and trimming, Burrows-Wheeler Aligner (BWA) will be used to map the WGS reads of *L. monocytogenes* (LMNC088) lineage II strain to the *L. monocytogenes* (JF5861) lineage I reference strain. The BWA was designed to map low-divergent sequences to reference genomes of large sizes allowing mismatches as well as gaps. It is based on backward search with Burrows-Wheeler Transform (BWT) and supports paired-end mapping as well as generates mapping quality. Specifically, BWA-MEM will be used out of the three algorithms available from BWA. Being the latest BWA algorithm, BWA-MEM is the best suited for Illumina sequence reads larger than 100 bp and is considered to have improved accuracy (*10*). The BWA parameters will be set to the default settings unless it is deemed necessary to manipulate the criterion.

*SAMtools (v0.1.19)*

As the output of BWA is Sequence Alignment/Map format (SAM), SAMtools will be utilized to generate alignments in a per-position format which will enable the calling of variants. SAMtools is a library and software package which facilitates the manipulation of SAM alignment files. The location and incidence of SNPs (*11*). The utilization of BWA along with SAMtools is superior to human identification of nucleotide variances as it eliminates the risk of human error as well as bias. The variant call format (vcf) file generated by SAMtools will then be used to generate a table of necessary SNP information (coverage, P-value). Preliminary threshold parameters for variant calling will consist of a point coverage minimum of 50x and a p-value = 0.05. The p-value of 0.05, which is biologically relevant, will be used initially due to the goal of obtaining all SNPs that may affect phenotype.



**Figure 1.** An overview of the genome mapping protocol.

**Division of Labor**

Data retrieval from NCBI was a group effort in which all members assisted in identifying appropriate datasets (reference and raw reads). HB will write code for quality control using FastQC and Trimmomatic. DM will write code for sequence alignment/mapping using BWA. IH will write code for variant calling using SAMtools. Generation of the final report will be conducted by all nkob members with equal contribution to the writing of the manuscript. Issues throughout the process of implementation will be resolved by the use of software manuals or codes found online to assist in the completion of experimentation.

**EXPECTED OUTCOME**

A genome of lineage II strain of *L. monocytogenes* is expected to be mapped to a reference genome of a lineage I (more virulent strain) *L. monocytogenes*. After obtaining a map, the variants shall be determined, and their positions noted. Summary of the variants, such as how many SNPs were located in the examined genome will be based on the output of SAMtools and built-in statistical probability. Table of outcome with the frequencies of particular types of variants will be generated.

The preliminary quality control on raw reads of *L. monocytogenes (Strain LMNC088, Lineage II)*, using fastqc revealed 260x genome depth coverage with respect to the reference genome *(Strain JF5861, Lineage I)*. However, we noticed the untrimmed right reads being of a lower quality, which is often observed with Illumina sequencing, thus we will use trimming and that will reduce the coverage. This alternative approach will utilize estimated coverage 241x after trimming to 140 bp [(2,512,294\*140\*2)/2,913,696], which should still provide sufficient depth for mapping and variant calling. In the event of low-quality data, NCBI has additional matches to whole genome sequencing of *L. monocytogenes* isolates, for which the dataset can be changed to use an alternate raw data set to carry out the experiment. In the event of preliminary threshold parameters in SAMtools (p-value=0.05, coverage 50x) revealing an excessive number of significant variants, an alternate approach could be to modify the main two parameters (lower the p-value and increase the coverage) to narrow the output to the strongest candidates of variants calls. If time permits, the Integrative Genomics Viewer (IGV) application will be utilized to take a closer look at the alignment to note the occurrence of missing genes as well as gene/SNP association (*12*).

**REFERENCES**

1. C. Zhang, M. Zhang, J. Ju, J. Nietfeldt, J. Wise, P. M. Terry, M. Olson, S. D. Kachman, M. Wiedmann, M. Samadpour, A. K. Benson, Genome Diversification in Phylogenetic Lineages I and II of Listeria monocytogenes: Identification of Segments Unique to Lineage II Populations. *J Bacteriol*. **185**, 5573–5584 (2003).

2. J. C. Low, W. Donachie, A review of Listeria monocytogenes and listeriosis. *The Veterinary Journal*. **153**, 9–29 (1997).

3. A. Oevermann, A. Zurbriggen, M. Vandevelde, Rhombencephalitis Caused by Listeria monocytogenes in Humans and Ruminants: A Zoonosis on the Rise? *Interdiscip Perspect Infect Dis*. **2010** (2010), doi:10.1155/2010/632513.

4. T. J. Ward, T. F. Ducey, T. Usgaard, K. A. Dunn, J. P. Bielawski, Multilocus Genotyping Assays for Single Nucleotide Polymorphism-Based Subtyping of Listeria monocytogenes Isolates. *Appl Environ Microbiol*. **74**, 7629–7642 (2008).

5. H. C. den Bakker, B. N. Bundrant, E. D. Fortes, R. H. Orsi, M. Wiedmann, A Population Genetics-Based and Phylogenetic Approach to Understanding the Evolution of Virulence in the Genus Listeria. *Appl Environ Microbiol*. **76**, 6085–6100 (2010).

6. L. Balandyté, I. Brodard, J. Frey, A. Oevermann, C. Abril, Ruminant Rhombencephalitis-Associated Listeria monocytogenes Alleles Linked to a Multilocus Variable-Number Tandem-Repeat Analysis Complex ▿. *Appl Environ Microbiol*. **77**, 8325–8335 (2011).

7. T. Nakazato, T. Ohta, H. Bono, Experimental Design-Based Functional Mining and Characterization of High-Throughput Sequencing Data in the Sequence Read Archive. *PLoS ONE*. **8**, e77910 (2013).

8. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data, (available at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

9. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**, 2114–2120 (2014).

10. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. **25**, 1754–1760 (2009).

11. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

12. J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J. P. Mesirov, Integrative Genomics Viewer. *Nat Biotechnol*. **29**, 24–26 (2011).