**Genome Architecture Analysis of L*isteria monocytogenes***

**INTRODUCTION**

*Listeria monocytogenes (LM*) 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*). *Listeria monocytogenes* strains are grouped into four distinct phylogenetic lineages, I, II, III and IV (*4*, *5*). In general, the *LM* 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*, *7*). The goal of this experiment was to map sequenced reads from *LM* (LMNC088)lineage II strain to *LM* (JF5861) lineage I reference strain in order to identify the number of single nucleotide polymorphisms (SNP) and insertions/deletions (indels) in genomic structure. The significance of this study is to identify genetic variability between lineage I and lineage II strains of *LM*. The purpose is to 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.

**RESPONSE TO REVIEWERS**

Authors appreciate the comments and helpful suggestions from reviewers. We took the time to consider them to make our project more meaningful, clear and innovative. The p-value included in the decision making on the variant call has been re-considered. In the variant calling in SAMtools, the probability based on Bayesian model is implemented. The depth of each variant is an important factor in the determination of error. The p-value is calculated for each individual variant to assist with the decision by providing an estimate of likelihood of the call being called incorrectly. The null hypothesis (Ho) is the variant call is incorrect. Alternative hypothesis the particular variant call is correct. Small p-value represents the probability we (or software) call the variant (reject the Ho), if it is truly an error. Smaller the p-value, the smaller chance the variant was called incorrectly. We initially planned on using the cut-off p-value 0.05, but in that case, some corresponding quality scores (QS) are below 20, so we adjusted our decision to p-value 0.01. This way, keeping the cut-off point 0.01, all the variants with QS less than 20 will be excluded. The mathematical relationship of QS and p-value is QS = -10\*log(prob(call)), where prob(call) is the probability that the variant call is incorrect. Therefore, one really needs just one or the other for the decision. The correction in the language of p-value being “biologically relevant” was applied in the proposal.The biological significance of the variant would be determined in future steps, such as amino acid codon analysis and annotation of the variants, for example, is this variant found possibly related to special proteins called internalins, which *LM* use for attachment to the host, colonization and cell-to-cell spread. In terms of the innovative portion of our project, the main objective was to identify genomic variants amongst two strains of *LM,* specifically identifying SNPs and indels. The objective was not to focus on specific genes associated with virulence.

**RESULTS**

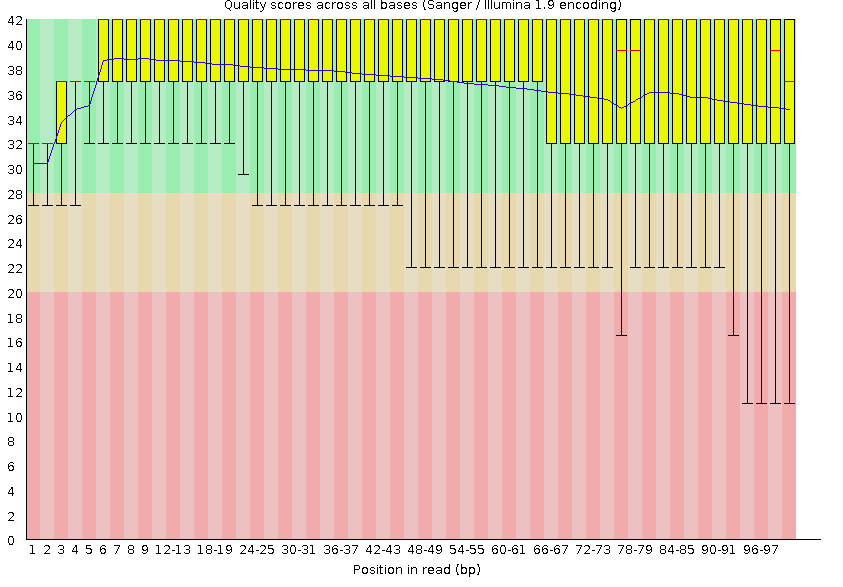
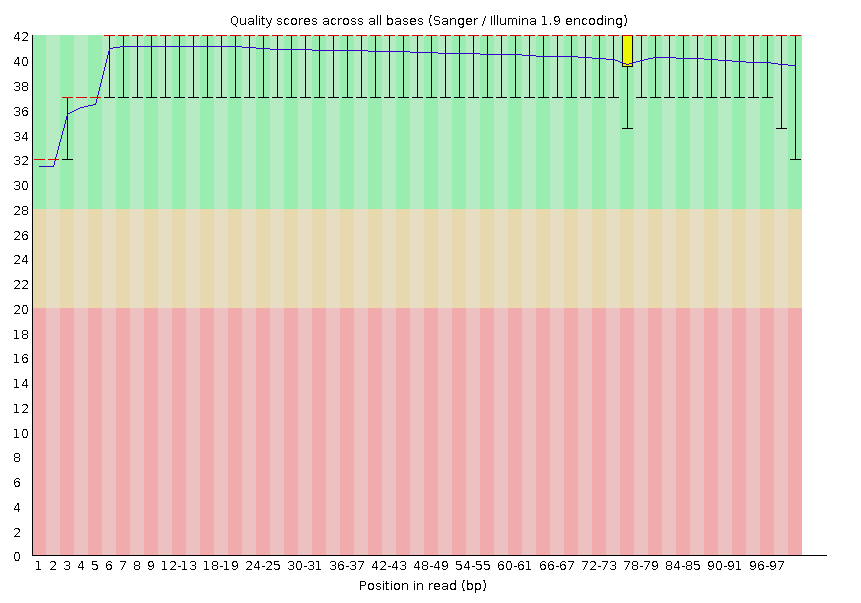
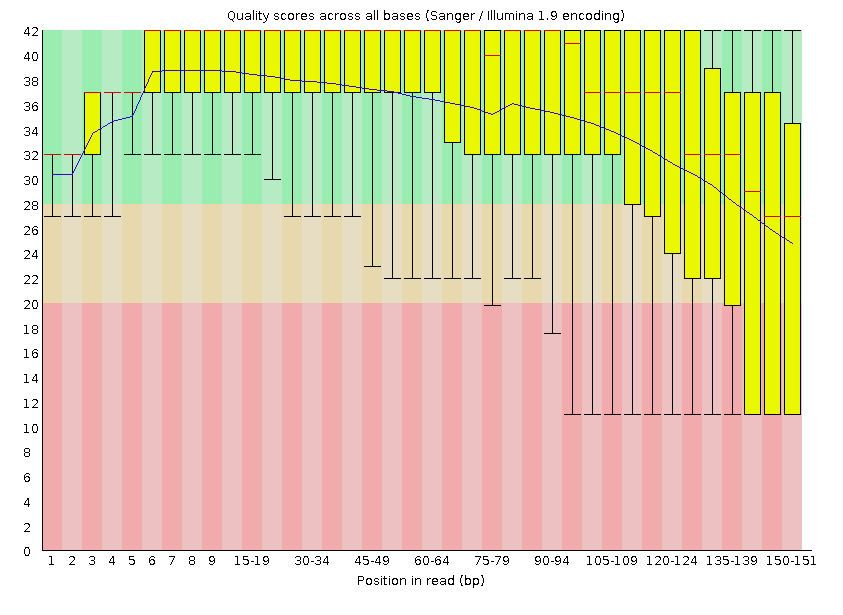
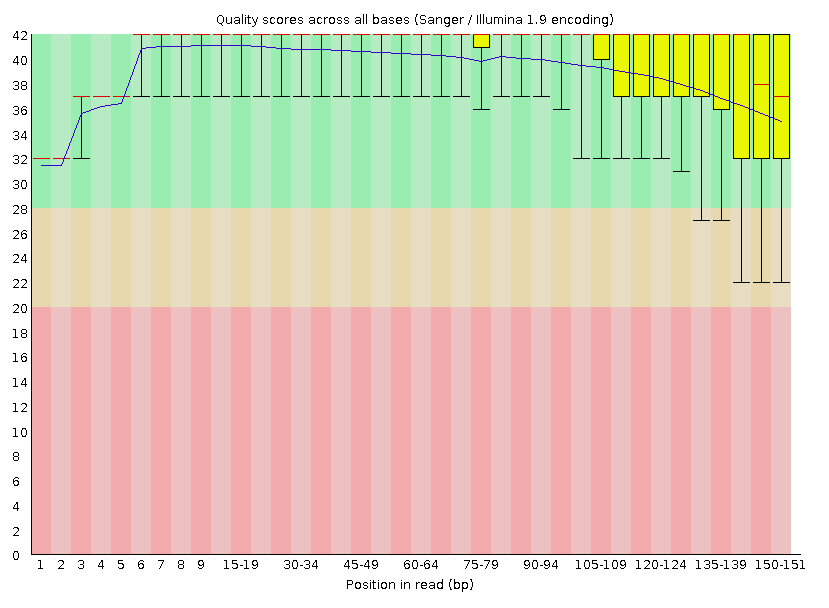
***Data Retrieval***

In order to address the genetic separation of *L. monocytogenes*, whole genome sequencing (WGS) was conducted on isolates of *L. monocytogenes*. Genome mapping was 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* reads ([ERX2192466](https://www.ncbi.nlm.nih.gov/sra/ERX2192466%5Baccn%5D)) were mapped to the reference genome, *L. monocytogenes* (Accession: JF5861 complete genome) which is available from NCBI’s Assembly platform in the .fna/.gff 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.

***Quality Control Processing (****Fastqc v0.11.7 and Trimmomatic v0.38)*

The mean 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 was estimated to be sufficient depth for mapping and to determine unbiased statistical estimates of genomic variability. The preliminary Fastqc report of the L. monocytogenes lineage II dataset exhibited that reads decline in quality (Q < 30) near the 3’ end. **Figure 1A** depicts read 1 base quality scores, the quality score reduced at about position 135, as shown by the variability of >135 bp position appearing in the yellow zone (mid/low quality reads). **Figure 1B** illustrates read 2 base quality scores having poorer quality than read 1, with the base quality scores declining at about position 100, with the mean base quality scores > 100 bp position appearing in the red zone (low quality reads). Trimmomatic was utilized to trim the 3’ end of both reads to 100 bp in order to reduce the chance of calling rare SNPs and noise. This resulted in the genomic coverage of 172x (2,510,242\*100\*2)/2,913,696). **Figure 1C** depicts the base quality score results for read 1 after trimming to 100 bp, all reads are found in the green zone (high quality reads). **Figure 1D** illustrates read 2 base quality scores after trimming to 100 bp, this demonstrates the most desired quality scores with fewer variability of quality scores at each position found in the red zone.

******

**D**

**C**

**B**

**A**

**Figure 1.** (A-D) Per sequence quality scores of reads from Read 1 (A and C) and Read 2 (B and D) obtained via fastqc. The raw (untrimmed; A and B) sequences contained reads with low Q scores (Q < 20) towards the 3’ end, therefore, trimmomatic was utilized to trim reads to 100 bp and remove reads shorter than 100 bp from the data set. This resulted in a data set with Q scores greater than 30 for both reads (C and D).

***BWA (v0.7.17), SAMtools (v0.1.19), & bcftools (v0.9)***

Burrows-Wheeler Aligner (BWA) was used to align the WGS reads of *L. monocytogenes* (LMNC088) lineage II strain to the *L. monocytogenes* (JF5861) lineage I reference strain. The untrimmed data set was investigated to determine the number of variants in comparison to the trimmed dataset with a max depth of 250. When comparing variant calls between trimmed and untrimmed datasets, indels were the only variants reported while the untrimmed data generated 470 variants and the data trimmed to 100 bp resulted in 492 variants, as shown in Table 1. Further the Q score (QS) for both data sets ranged from 5-228 with a median QS of 107.5 and 106 for the untrimmed and trimmed to 100 data sets, respectively. The per base depth of the untrimmed and trimmed to 100 data sets ranged from 1-249 and 1-259, respectively. In addition, to investigate the impact of increasing the maximum depth (-d) for the bcftools mpileup command, the -d value was changed from 250 to 10,000. This resulted in the identification of 488 indels with a median QS of 106. Collectively, the maximum depth of 250 was used to move forward in the project as this is the default setting in bcftools and increasing the maximum depth to 10,000 did not have a notable effect on the number of variants called variant type, % CDS variants, QS range, median QS, nor the per base depth range (**Table 1**).

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Untrimmed** | **Trimmed 100, d250** | **Trimmed 100, d10,000** |
| No.Variants | 470 | 492 | 488 |
| Variant Type | indels | indels | indels |
| % CDS Variants | 24 | 22 | 22 |
| QS Range | 5-228 | 5-228 | 3.7-228 |
| Median QS | 107.5 | 106 | 106 |
| Per Base Depth (range) | 1-249 | 1-247 | 1-222 |

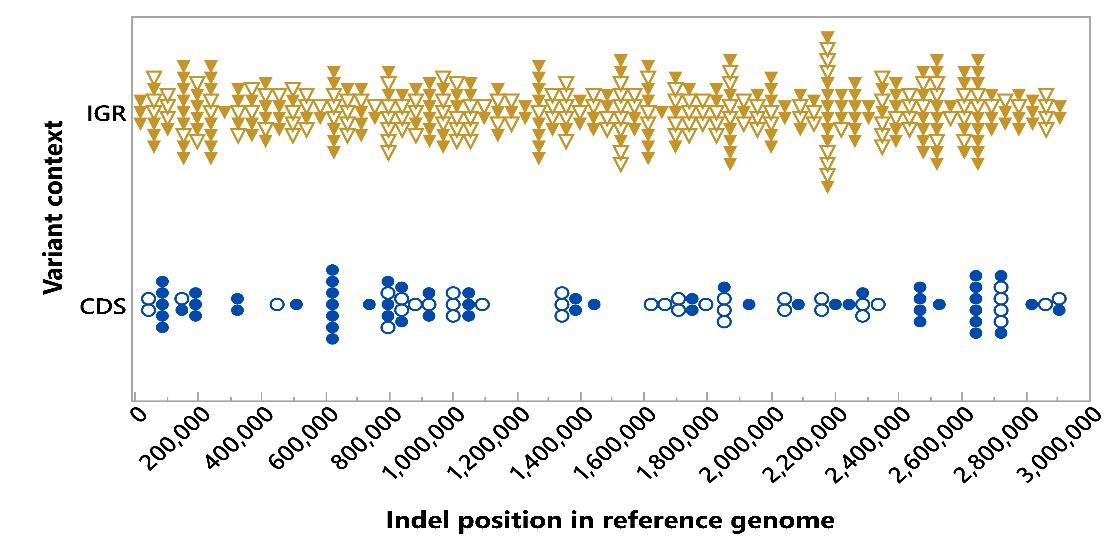
**Table 1.** The results generated from mapping *LM* (LMNC088; lineage II) to the *LM* reference genome (F5861; lineage I) prior to sorting the variants statistically. The trimmed reads, reads trimmed to 100 excluding reads shorter than 100 bp with the maximum depth set at 250, and reads trimmed to 100 excluding reads shorter than 100 bp with the maximum depth set at 10,000 resulted in the identification of exclusively indels. Increasing the maximum depth (-d) did not have an effect on increasing the number of variants identified, nor the QS of the variants.

**Post Processing**

There were 492 variants listed in our initial output. Examination of the variant’s types revealed that all of the 492 were indels. The QS for all 492 found indels varies from 5 to 228 (mean 115.3, SD 68.25). The QS was calculated as -10\*log(prob(call)), where prob(call) is the probability that the variant call is incorrect. The corresponding p-values for all the quality scores < 20 are greater than 0.01 with max p-value found of 0.31 corresponding to the lowest quality score 5. However, the remaining variants with relatively high p-value and low-quality score (<20) had variant depth above 100 (n=31). This first post processing step filtered out unacceptable quality scores, leaving 460 variants with QS>20 (p<0.01). Variant depth (per base) of all 492 indels varied from 1 to 247 and even after filtering out the low-quality calls, remaining 460 calls still varied in per base depth from 1 to 247. The distribution of the variant depth was left skewed indicating 13 outliers with a low depth (detected by JMP (v. Pro14.0.0) statistical software package. Thus, our next post processing step was to filter away the calls with undesirable variant depth, less than 106. This excluded 13 calls plus one additional row without variant context information, leaving 446 indels. Of the indels identified, 253 were insertions while 193 were deletions. The indels represented 352 intergenomic regions (79%, IGR) and 94 were identified as coding regions CDS (21%) (**Figure 2**).

|  |  |
| --- | --- |
| **Trimmed to 100, max depth 250** | |
| No. of Variants | 446 |
| Variant Type | indels |
| Insertions/Deletions | 253/193 |
| % CDS Variants | 21 |
| CDS Insertions/Deletions | 56/38 |
| QS Range | 20-228 |
| Median QS | 115 |
| Per Base Depth (range) | 106-247 |

**Table 2.** Results obtained post-sorting the bcftools data for reads trimmed to 100 bp and excluding reads < 100 bp with maximum depth set at 250 for bcftools mpileup. All variants included in the table have an associated p-value of ≦ 0.01.



**Figure 2.** WGS of *LM* of Lineage II (LMNC088) was aligned with *LM* sequence of lineage I (JF5861, reference). After quality and per base depth filtering, 446 Indels, as represented by the markers, were recognized and their approximate position is identified in the plot. Gold triangles represent intergenomic region (IGR, 79%) and blue dots represent the protein coding sequences (CDS, 21%). Empty markers (triangles and circles) indicate deletions, while full markers indicate insertions.

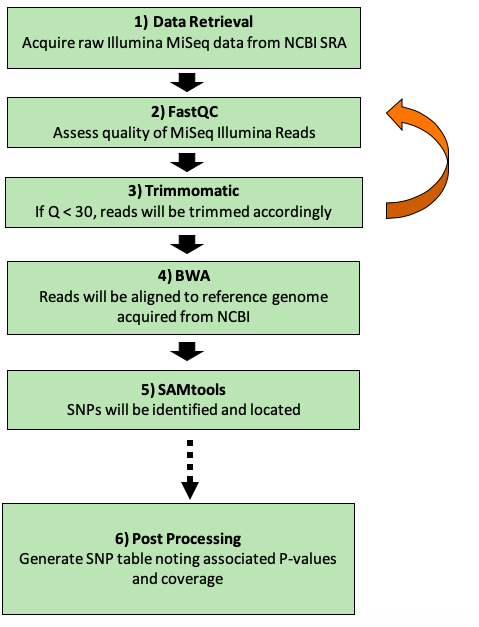
**DISCUSSION**

The objective of this experiment was to align sequenced reads from *LM* (LMNC088)lineage II strain to *LM* (JF5861) lineage I reference strain in order to identify genomic variability in genomic structure. Due to the virulent nature of the reference genome, we were interested in identifying the number of SNPs and/or indels in the non-virulent lineage II strain of *LM.* Through the use of BWA, SAMtools, and bcftools, we identified 446 genomic variants in *LM* lineage II in comparison to the *LM* lineage I reference genome. All variants were indels with 253 insertions and 193 deletions.

The pipeline implemented throughout this module varied little in software use and more in command use. The process of quality control with the Fastqc program revealed variation within our quality scores across all bases. Consistently complications arose with read 2 having too many bases in the less preferred low-quality area as compared to read 1. This could be a result of Illumina's sequencing-by-synthesis method, essentially, the more cycles run, the noisier the signal or even a product of improper library construction. Trimommatic command line tool, utilizing the paired end mode, enabled us to trim the Illumina (FASTQ) data after we detected the presence of low-quality bases. A vital lesson learned by nkob was to read the manual for the particular version of the software and visit blogs of people who use the software, or blogs of developers for ideas and command changes. For example, the SAMtools v. 0.1.19 has some required options or arguments in the view and sort commands, while they may be set as default in the newer versions. Another lesson learned throughout the process was in times of error, running each step individually will allow the error to be determined at the precise step. For example, when conducting BWA, SAMtools, & bcftools we kept receiving an error code after the run, we would copy the error code sent and use the less command to identify the error. After further confusion of error identification we ran each step separately and the modification to the provided script was not to include .bam after file.sorted, the samtools would do it automatically. Failure to do this generated .bam.bam file and the pipeline was interrupted before the next step. Initially, the egrep command was implemented to sort the vcf file to generate a file with variants only. However, after experiencing difficulty with the egrep command, it was decided to move forward with the addition of the -v (variants only) parameter. We utilized the command -v (variants only) in bcftools call, to directly generate the list of variants which omitted the need for further sorting. The final lesson learned was to keep the scratch folder organized. Normally, when new files happen to have the same name as the existing files in the scratch folder, they would be automatically replaced with the new content. This fact was taken advantage of, to save time, especially when trying multiple (about 10) variations of the code. However, there is an exception to every rule. If one realizes later that some useful results were actually generated despite the error message from Spruce, if the files are replaced every time, and only the newest copy was kept the old content is gone. Keeping the things properly labelled helped us to utilize the older results and made it easier to share the input/output files among the group members.

Future direction of these experiments can include the biological relevance in amino acid (aa) changes observed in genomic structure. The three kinds of aa indels discovered with the highest frequency were histidine (n=12), followed by leucine (8) and serine (n=7). Future detailed analysis considering the region of the genome may determine if these could be regions responsible for the differential virulence lineage of the two compared genomes. In our case, the deletions in CDS region may be the most relevant, as those may lead to discovery of proteins involved in bacterial adherence to host and virulence. Overall, the maximum depth set at the default parameter of 250 resulted in the greatest number of variants called. A total of 446 indels were called after the variants were sorted by p-value with a strict cutoff at p-value 0.01. Keeping the cut-off point at 0.01, variants with QS less than 20 were excluded. In addition, we have learned the important steps necessary to implement variant calling when comparing genomic variations in genomic structures between lineage I and lineage II strains of *LM*.

**PIPELINE CODE**



**Figure 3.** An overview of the protocol for detection of genomic variance.

***Fastqc***

**fastq**  ListII\_R1.fastq ListII\_R2.fastq

#fastqc was utilized to address the quality of the raw paired end dataset selected

***Trimmomatic***

**trimmomatic** PE -threads 6 ListII\_R1.fastq ListII\_R2.fastq -baseout ListII\_crop100.fastq CROP:100

#the paired end reads were cropped to 100 bp by trimming the 3’ end of the reads

**fastqc** ListII\_crop100\_1P.fastq ListII\_crop100\_2P.fastq

#the quality was assessed again by using fastqc

**trimmomatic** PE -threads 6 ListII\_crop100\_1P.fastq ListII\_crop100\_2P.fastq -baseout ListII\_crop100\_minlen100 fastq MINLEN:100

#reads less than 100 bp were excluded from the dataset

**fastqc** ListII\_crop100\_minlen100\_1P.fastq ListII\_crop100\_minlen100\_2P.fastq

#the quality was assessed again by using fastqc

***BWA, Samtools, and generation of variant summary table***

**bwa index** LM\_L1\_JF5861\_ref.fna

#The index command of BWA was utilized to index the reference genome. First, the index of the reference genome was created by utilizing the index command of BWA.

**bwa mem** -t 12 -c 50000 -P -B 3 LM\_L1\_JF5861\_ref.fna ListII\_crop100\_minlen100\_1P.fastq ListII\_crop100\_minlen100\_2P.fastq > LM\_L2\_trim100.sam

#The MEM mapping algorithm of BWA was used and the coverage (-c) cutoff was set at 50000. Additionally, the data was flagged as being paired end (-P) and the mismatch penalty (-B) was set to 3. This generated a .sam file.

**samtools view** -Sb -o LM\_L2\_trim100.bam LM\_L2\_trim100.sam

#The generated .sam file was converted to a .bam file.

**samtools sort** LM\_L2\_trim100.bam LM\_L2\_trim100.sorted

#The sam file was converted to an ordered bam file which is the required input for calling variants.

**samtools index** LM\_L2\_trim100.sorted.bam LM\_L2\_trim100.sorted.bai

#The sorted bam file had to be indexed by creating the binary index.

**bcftools mpileup** --threads 12 -6 -A -d 250 -f LM\_L1\_JF5861\_ref.fna LM\_L2\_trim100.sorted.bam -O b -o LM\_L2\_trim100.bcf

#Variant calling was executed in two steps with the first being the bcftools mpileup command which resulted in the creation of the binary file coding for every position in our reference genome. For this command, the maximal per sample read depth (-d) was modified as two variations of -d (250 and 10,000) were used for the trimmed data while the -d parameter for the untrimmed data was set at the default (250).

**bcftools call** -o LM\_L2\_trim100.vcf -v --threads 12 -A -m LM\_L2\_trim100.bcf

#Using the bcftools call command, variants were called. A .vcf file which contained only variants, due to the addition of the -v, was generated.

./vcf2table.pl LM\_L1\_JF5861\_ref.gff -g LM\_L1\_JF5861\_ref.fna < LM\_L2\_trim100.vcf > LM\_L2\_trim100.variant\_table.txt

# The filtered vcf file was combined with the reference genome (fna) and genomic feature (gff) files to identify the biological context of the variants.

**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. L. Aguilar-Bultet, P. Nicholson, L. Rychener, M. Dreyer, B. Gözel, F. C. Origgi, A. Oevermann, J. Frey, L. Falquet, Genetic Separation of Listeria monocytogenes Causing Central Nervous System Infections in Animals. *Front Cell Infect Microbiol*. **8** (2018), doi:10.3389/fcimb.2018.00020.