Prokaryotic Genomes- Salmonella

Andrea Soles

Biol 425- Modern Genetics

**Introduction:**

PulseNet is the national molecular subtyping network for foodborne bacterial disease surveillance. It was created by the Centers for Disease Control and Prevention and other state labs to aid in the subtyping of bacterial foodborne pathogens for epidemiologic purposes (Swaminathan et al., 2001). PulseNet started off by typing one pathogen, *Escherichia coli* O157:H7, and now they have three more foodborne pathogens (Salmonella, Listeria monocytogenes, and Shigella) being subtyped and studied (Swaminathan et al., 2001).

Salmonella is a rod-shaped Gram-negative bacterium that belongs to the Enterobacteriaceae family. More commonly, it is known as a bacterium that causes a foodborne illness. Salmonella can be found in animal intestines and is released from the body by feces (Mayo Clinic, 2018). There are multiple foods that people can eat that could contain Salmonella. Raw meat and seafood, raw eggs, and contaminated fruits and vegetables are the main food sources of Salmonella (Mayo Clinic, 2018). People infected with Salmonella usually get a fever, abdominal pain, and diarrhea within 12 to 72 hours after consumption of Salmonella in their food (CDC, 2018). These symptoms last up to around 4 to 7 days and can be treated with antibiotics (CDC, 2018). However, if not treated right away, hospitalization and possibly death can occur in severe cases (CDC, 2018).

Salmonella is an important bacterium to study because it is a good model organism that can be a representative for other pathogens related to food borne illnesses. A model organism is widely studied and its genetic contents are well documented and studied. More importantly, the knowledge of the host-pathogen interaction pathway is essential to understand (Garai et al., 2012). Since Salmonella can infect a wide variety of hosts causing multiple diseases, it can be useful in studying for PulseNet. The purpose of this study was to aid PulseNet by subtyping a Salmonella isolate, SRR7506697 (DG-10 for short), that was collected and sequenced from a group in Virginia. The gene annotation and assembly methods were done with the help of the website, Galaxy Trakr. The results from this study will be used for comparisons and other studies for PulseNet in the future.

**Assembly:**

First, Trimmomatic was used to essentially clean up DG-10. Trimmomatic is a tool that performs various tasks useful for trimming of ilumina paired-end and single ended data (Bolger & Lohse, 2014). The sliding widow option was used to clean up the data. Sliding window uses the average quality to determine what stays or what is trimmed off if it does not belong within the threshold (Bolger & Lohse, 2014). For this sequence, 4 was the number of bases to average across and 20 was the average quality required. The total amount of sequences in the DG-10 forward and DG-10 reverse was 497,167. After the program Trimmomatic ran, the DG-10 forward and reverse paired ends sequences went down to a total of 485,644 sequences. Trimmomatic cut out 11,523 sequences. The overall percent change after Trimmomatic was performed was 2.37%.

The assembly program SPAdes was done of three separate data sets: Raw Data, Raw Data with Read Error Correction, and Trimmed Data. SPAdes is a genome assembler for regular and single-cell projects (Bankevich et al., 2012). The parameters used were to run the K-mer values: 21, 33, 55, 77, 99, and 127. Then the program, QUAST, was performed on each of the assembled sequences. QUAST is a program that gives a quality assessment for genomic assemblies (Alexey et al., 2013). QUAST gave assembly values, the number of contigs, the total length, GC% content, N50, and other values pertaining to contigs. Those values are shown in Table 1. The number of contigs is the same for the two raw data sets but increases by three with the trimmed data. The total length is decreased a little with read error correction and decreased a lot with the trimmed data. The N50 values are the same between the three. Keeping those values in mind, the best assembly to continue on analysis with is the Raw Data because it has the highest total length and the N50 number is not changed when the sequenced is cleaned up. A De Brujin graph was then created with the computer software Bandage and the results of the Raw Data are shown below in Figure 2.

**Analyses**:

First for the analysis process of DG-10, ABRicate was performed on the raw data. This program allows for a mass screening of contigs, a continuous sequence resulting from the reassembly of small DNA fragments, and identifies the presence of antibiotic resistance genes (Seemann, 2016). Table 2 includes a list of few antibiotic resistance genes found during the screening. The table includes their percent coverage and percent identity. For example, (Tet)TetR was found during the screening with 100% coverage and identity. The expression of Tetracycline, antibiotic, resistance genes is regulated by Tet repressor proteins. TetR has an important role in helping researchers understand antibiotic resistance mechanisms better and has the potential to help with finding a way to treat antibiotic resistance bacteria, like DG-10.

Three additional analyses were done to describe and characterize this strain of Salmonella. The first one was done using the tool KmerResistance from the Center of Genomic Epidemiology. This tool maps the co-occurrence of k-mers between the whole genome sequence and a database of resistance genes (Clausen et al., 2016). K-mers are all the possible subsequences (of a certain length) from a read obtained through DNA sequencing. In Table 3, the results for this tool generated a more detailed name and description of DG-10. DG-10 is *Salmonella* *enterica* Serotype Typhimurium DT 104, which is known to have multi-drug resistance (Meunier et al., 2002). Table 3 also includes more specific genes that modified themselves to help with resistance mechanisms. For example, laa\_1\_NC\_003197 is associated with the cell-wall hydrolase involved in septum cleavage in cell division (UniProt, 2018).

The second analysis done using the tool MLST. MLST stands for Multi Locus Sequence Typing. It is a program that can characterize isolates of bacterial species using sequences of internal fragments called housekeeping genes (Larsen et al., 2012). Each one of those genes, the different sequences in the bacteria are assigned as distinct alleles and the alleles at the seven loci make up the sequence type (Larsen et al., 2012). In Table 4, the seven alleles or housekeeping genes are listed. They are very specific to DG-10 and help with differentiating it from other strains.

The third and final analysis done of DG-10 was the program, PlasmidFinder. This program can identify plasmids in total or partial sequenced isolates of bacteria (Carattoli et al., 2014). This tool is noted to be accurate however few false positives have occurred. The plasmid identified was IncA/C2 which is a *Klebsiella pneumoniae* strain Kp7 plasmid. IncA/C2 was also identified as an antimicrobial resistance gene in Table 2.

**Discussion**:

The major finding of this project was that a complete name for DG-10 was obtained. *Salmonella* *enterica* Serotype Typhimurium DT 104 is known across the world for causing a foodborne illness that is hard to treat due to its antibiotic resistance properties. By subtyping the Salmonella isolate, SRR7506697, found in Virginia for PulseNet, valuable information was given to them to aid them in possibly creating a new drug that can by-pass the antimicrobial and antibiotic resistant genes. If someone were to ingest this bacterium on their food, this new drug can help get rid of their foodborne illness without all of the trial and error of taking different antibiotics that are not effective in killing this particular strain of Salmonella.

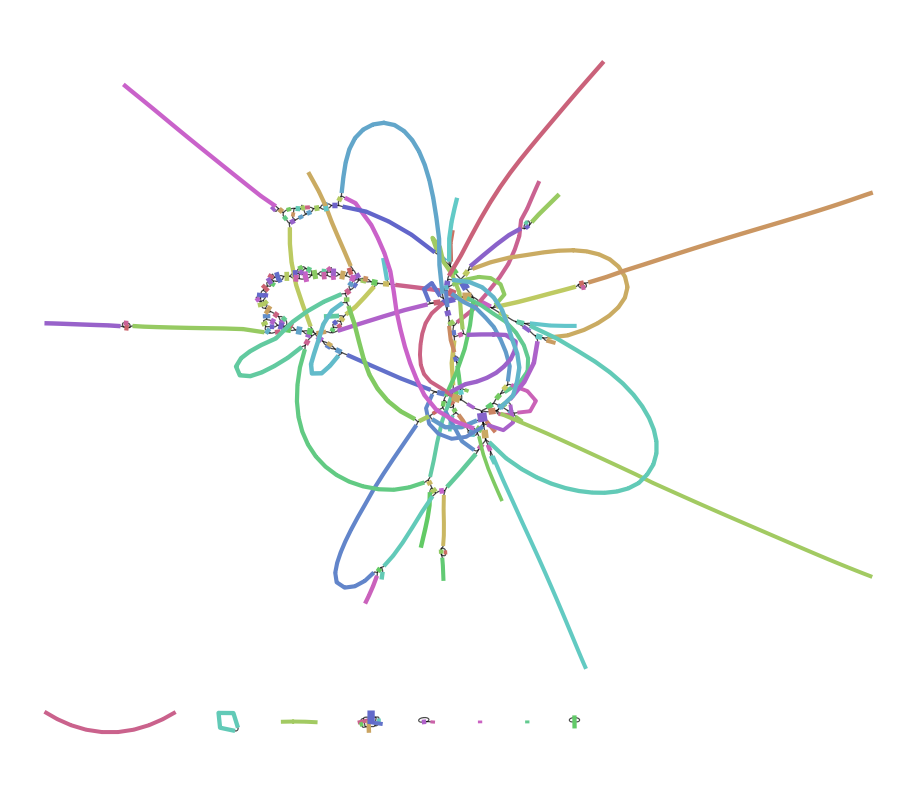
**Figures and Tables:**



**Figure 1. Per Base Sequence Quality Before and After Trimmomatic of DG-10**. a.) The per base sequence quality of the forward and b.) the per base sequence quality of the reverse both show on their graphs that the quality is a very poor towards the end of the sequences. c.) The per base sequence quality of forward paired ends and d.) the per base sequence quality of the reverse paired ends have been trimmed up with Trimmomatic. The poor-quality sequences have been removed and box and whiskers plots are not in the red anymore.

**Table 1. QUAST Analysis Comparison Table.** The values for GC% and N50 are the exact same while the other values are not significantly different from each other for assembly, number of contigs, and total length.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Assembly | # contigs | Total Length | GC% | N50 |
| Raw Data | 94 | 65 | 4973186 | 52.05 | 164260 |
| Raw Data with Read Error Correction | 95 | 65 | 4971305 | 52.05 | 164260 |
| Trimmed Data | 99 | 68 | 4971593 | 52.05 | 164260 |



**Figure 2. De Bruijn Graph of the Raw Data Sequence.** The raw data was generated into a De Bruijn graph to show the overlapping of k-mer values in this assembled sequence.

**Table 2. Antimicrobial Resistance Gene Detection Results.** This table includes some examples of antimicrobial resistance genes found using ABRicate. None of them had gaps in the alignment and the majority of them have 100 percent coverage and identity in the Salmonella sequence.

|  |  |  |  |
| --- | --- | --- | --- |
| Gene | %Coverage | %Identity | Gaps |
| (Bla)AMPH\_Ecoli | 100 | 80.05 | 0 |
| (Bla)Penicillin\_Binding\_Protein\_Ecoli | 100 | 83.23 | 0 |
| (Tet)TetR | 100 | 100 | 0 |
| (AGly)Aac6-laa | 100 | 100 | 0 |
| lncA/C2\_1 | 100 | 100 | 0 |

**Table 3. K-mer Resistance Results.** A more detailed name and description of the Salmonella strain is given below. Also, the genes that modified themselves to help with resistance mechanisms are listed below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Template | Score | Expected | Template Length | P Value | Template Coverage |
| CP014965.1 Salmonella enterica subsp. enterica serovar Typhimurium str. CDC 2010K-1587, complete genome aac(6’)- | 5661702 | 85 | 154762 | 1.0e-26 | 99.94 |
| laa\_1\_NC\_003197 | 19379 | 7 | 438 | 1.0e-26 | 100.00 |
| sul2\_2\_AY034138 | 15101 | 14 | 816 | 1.0e-26 | 100.00 |
| tet(A)\_4\_AJ517790 | 26367 | 16 | 1200 | 1.0e-26 | 99.58 |

**Table 4. MLST Results.** Seven alleles were identified with the MLST tool. The alignment and allele length are provided below. All of the alleles have zero gaps and 100 percent identity and coverage in the sequence.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Locus | Identity | Coverage | Alignment Length | Allele Length | Gaps | Allele |
| aroC | 100 | 100 | 501 | 501 | 0 | aroC\_10 |
| dnaN | 100 | 100 | 501 | 501 | 0 | dnaN\_7 |
| hemD | 100 | 100 | 432 | 432 | 0 | hemD\_12 |
| hisD | 100 | 100 | 501 | 501 | 0 | hisD\_9 |
| purE | 100 | 100 | 399 | 399 | 0 | purE\_5 |
| sucA | 100 | 100 | 501 | 501 | 0 | sucA\_9 |
| thrA | 100 | 100 | 501 | 501 | 0 | thrA\_2 |

**Works Cited:**

Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, Glenn Tesler. QUAST: quality assessment tool for genomce assemblies. Bioinformatics. 2013. 29(8):1072-1075.

Bankevich, Anton and Nurk, Sergey and Antipov, Dmitry and Gurevich, Alexey A. and Dvorkin, Mikhail and Kulikov, Alexander S. and Lesin, Valery M. and Nikolenko, Sergey I. and Pham, Son and Prjibelski, Andrey D. and et al. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology.19(5):455-477*.*[[doi:10.1089/cmb.2012.0021](http://dx.doi.org/10.1089/cmb.2012.0021)][[Link](http://dx.doi.org/10.1089/cmb.2012.0021)]

Bolger, Anthony M. and Lohse, Marc and Usadel, Bjoern. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30(15):2114-2120.[[doi:10.1093/bioinformatics/btu170](http://dx.doi.org/10.1093/bioinformatics/btu170)][[Link](http://dx.doi.org/10.1093/bioinformatics/btu170)]

Carattoli A, Zankari E, Garcia-Fernandez A, Voldby M, Lund O, Villa L, Aarestrup FM, Hasman H. 2014. PlasmidFinder and pMLST: in silico detection and typing of plasmids. Antimicrob. Agents Chemother.

Clausen P, Zankari E, Aarestrup F, Lund O. Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. Journal of Antimicrobial Chemotherapy. 2016. 71(9):2484-2488.

Garai P, Gnanadhas D, Chakravotty D. 2012. *Salmonella enterica* serovars Typhimurium and Typhi as model organisms. Virulence. 3(4):377-388.

Meunier D, Boyd D, Mulvey M, Baucheron S, Mammina C, Nastasi A, Chaslus-Dancla E, Cloeckaert A. 2002. *Salmonella enterica* Serotype Typhimurium DT 104 Antibiotic Resistance Genomic Island I in Serotype Paratyphi B. CDC. 8(4).

Larsen MV, Consentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbek L, Sicheritz-Ponten T, Ussery DW, Aarestrup FM, Lund O. 2012. Multilocus Sequence Typing of Total Genome Sequenced Bacteria. J. Clin. Micobiol. 50(4):1355-1361.

Salmonella infection. Mayo Clinic. [accessed 2018 Nov 12];

<https://www.mayoclinic.org/diseases-conditions/salmonella/symptoms-causes/syc-20355329>

Salmonella. Centers for Disease Control and Prevention. [accessed 2018 Nov 12];

<https://www.cdc.gov/salmonella/general/index.html>

Seemann, Torsten (2016). ABRicate: mass screening of contigs for antiobiotic resistance genes.

Swaminathan B, Barrett T, Hunter S, Tauxe R. 2001. PulseNet: The Molecular Subtyping Network for Foodborne Bacterial Disease Surveillance, United States. Emerging Infectious Diseases. 7(3):382-389.

UniProtKB-P26366 (AMIB\_SALTY). UniProt. [accessed 2018 Nov 29]; <https://www.uniprot.org/uniprot/P26366>

https://en.wikipedia.org/wiki/TetR