Megan Curry

Introduction paragraph

Genetics 425

11/12/18

Unknown *Salmonella* Isolates and Responding Data

**Introduction:**

PulseNet is national laboratory network program that directly works with studying and tracking foodborne pathogens. The program uses pulsed-field gel electrophoresis and whole genome sequencing to preform DNA fingerprinting of the bacteria causing the illness and detects thousands of outbreaks, whether it is local or multistate. Matching the fingerprints helps investigators find where the outbreak started and is currently happening, this allows investigators to quickly prevent further illnesses. The program has been running for twenty years now and is always improving and updating. PulseNet is estimated to prevent 270,000 illnesses every year. The three most common foodborne illnesses that are being tracked include: *Salmonella, E.Coli,* and *Listeria.* The program is tracking these illnesses because they are the most frequent and lethal to humans. It is very important to manage and track these illnesses because of how fast and easily they spread (Centers, 2018).

 Currently, the Genetics 425 class is sequencing *Salmonella* isolates. *Salmonella enterica* is the listed genus and it has seven species and subspecies. There are 2,463 serotypes within the subspecies (Brenner, 2000). Since there are so many serotypes and outbreaks, it is critical and important to use *Salmonella* as a model organism. With *Salmonella* it is always changing, and scientists are finding new ways to sequence it. Many studies have already been done on the genetic makeup of *Salmonella.* It will be a good model organism because it can be easily modified, tracked, and it is similar to other foodborne illnesses.

 Lastly, it is important to track these outbreaks because they can be so lethal. *Salmonella* spreads in different ways. At a buffet meal in a restaurant *Salmonella* spread from microwaving chicken and incorrectly cooking it (Evans, 1995). Another outbreak took place in a nursing home that caused patients to be hospitalized for a while (Choi, 1990). A third study was done on a large outbreak during a medical conference in Wales (Glynn, 1992). The newest and biggest issue arising is *Salmonella* creating a battle for the poultry and turkey industries. It is generating a large stress and decline in the production of poultry and turkey. With foodborne illnesses they are easily spread; meaning that one contamination of a farm can be spread all over the country. These outbreaks can cause issues such as arthritis and newer illnesses (Rohekar, 2008). With further analysis and tracking, *Salmonella* numbers and outbreaks can be potentially decreased and more understood. This is the main purpose of this study and the sequencing that is currently taking place.

**Methods**

*Salmonella* isolates were distributed, and each student ran analyses of one isolate with the use of the Galaxy Tracker program. Each isolate contained the bacteria of an unknown *Salmonella* strand. With the use of Galaxy students ran the isolate through Trimmomatic to produce trimmed, cleaned, and concise contigs. Contigs are defined as a contiguous sequence of DNA that are created from assembling the overlapping sequenced fragments of a specific chromosome (Medical, 2016). The contigs were then ran through a FastQC system to report the sequence differences in the forward and reverse reads for the original data and the trimmed data. FastQC also provided a report for the sequence quality and the sequence length on the forward and reverse reads for both data sets. The contigs were then ran through Quast to give a Spades report on the contigs specific information. This report provided us with information on the number of contigs greater than zero and one thousand, the GC percent, N50 values, and many more important analyses. After the Quast report was ran a full Spades report was done. The report was downloaded and opened in a program called Bandage. The program showed a graphic of the original and trimmed contig data, and represented bad and good contig lengths in the loops. Once all the data was collected further analyses were done with different programs. MLST is a program that was used to find proteins in the strand. Each protein listed gave different information on their makeup, percentage values, and created the ideology of what the unknown isolate was made of. Two other programs were ran as well, but crashed and didn’t provide any further analysis.

**Results:**

After running FastQC and Trimmomatic on the DG-17 sequence for the forward and reverse reads, very comparative statistical data was presented. Figure 1 shows the report of the sequence lengths. Parts A and B have identical sequence lengths and are 27,133 bases shorter than parts C and D (also identical) because Trimmomatic was ran. With Trimmomatic it cleans the data and removes unneeded bases leading to the smaller sequence lengths in parts A and B. Figure 2 shows the sequence quality across all bases. Part A, B, and D have good data quality, and part C has poor data quality. This can be analyzed through the yellow boxes presented in each part of the graphs. The larger and more common the boxes are the poorer the data is. Part A and B show less boxes and smaller boxes because the sequences were ran through Trimmomatic. Trimmomatic should have removed the poor data and it did. Part D shows good quality, this is a random and an irregular occurrence because it should show just as poor data as part C does. Part C and D started with the original sequences, so they should both show poor quality, but this didn’t happen. Figure 3, presents the distribution of the sequence lengths. Part A and B have random spikes across the graphs because Trimmomatic can go haywire on the distribution. This happens because it cuts out the bad data and makes the sequences smaller in certain areas. This shows the changes in size and the inconsistency of the graphs. Part C and D have a start line going across the graphs because the sequences are all the same length. Once all of the data was collected, it was easier to see the difference that an error correcting program creates.

After running Quast and Spades on the original and trimmed contigs an even larger comparative statistical data was presented. Figure 4 shows the Quast report of the contigs. Part A shows the original data on contigs 4 and 3. Part B shows the trimmed data on contigs 12 and 11. Comparatively part B’s contig lengths and N50 values are larger than part A’s because part B is the clean data. With this being said, the data has less issues or base misreads which results in longer contig lengths. Figure 5 shows the Quast GC content of the contigs. Part A shows an almost perfect peak for the original data, and part B also shows an almost perfect peak for the trimmed data. Both graphs relay that there was no contamination and that the data was from one organism. This makes sense because *salmonella* data was only used so it should be one organism, but it also shows that good data was used, and no contamination was processed. The final figure, figure 6, presents the Spades report for the assembly graph. Part A shows the original contig data. There are shorter colored loops which represents short contigs and thus results in bad data. Part B shows the trimmed contig data. It has larger loops meaning the contig lengths are larger and thus results in better data.

MLST was ran on the trimmed data to find protein/gene matches. There was only one protein/gene that showed 100% coverage and 100% identity. The protein was PrgI and is found in the *Salmonella typhimurium* strain which could reveal that that is the identity of our unknown strand (UniProt, 2018).

**Discussion**

The *salmonella* isolates that were ran and analyzed produced data that was very helpful. There was no expectation to receive good data, or have it work as well as it did. There were only minor inconveniences that deterred the data. Programs such as Kmer resistance and Prokka didn’t not run correctly and left no data to analyze. Possibilities of this failure could include: system downtime, running errors, incorrect inputs of data, and system overload. Further analysis and data programming can be done on the *salmonella* strand to find more data pertaining to the strand. If the strand is *Salmonella typhimurium,* it will be a very serviceable strand to further research with because it is a suitable PCR target with unique gene sequences (Rahn, 1992).

**Figures:**



A

B

C



D

**Figure 1. FastQC Report of Sequence Lengths.** (A) Statistics after trimmomatic on DG-17 forward reads for the R1 paired sequence. (B) Statistics after trimmomatic statistics on DG-17 reverse reads for R2 paired sequence. (C) FastQC statistics before trimmomatic on DG-17 forward sequence. (D) FastQC statictics before trimmomatic on DG-17 reverse sequence.



B

A



D

C

**Figure 2. Sequence Quality.** (A) Quality scores across all bases after trimmomatic on DG-17 forward reads for the R1 paired sequence. (B) Quality scores across all bases after trimmomatic statistics on DG-17 reverse reads for R2 paired sequence. (C) FastQC quality scores across all bases before trimmomatic on DG-17 forward sequence. (D) FastQC quality scores across all bases before trimmomatic on DG-17 reverse sequence.



C

D

B

A

**Figure 3. Sequence Length.** (A) Distribution of sequence lengths over all sequences after trimmomatic on DG-17 forward reads for the R1 paired sequence. (B) Distribution of sequence lenghts over all sequences after trimmomatic statistics on DG-17 reverse reads for R2 paired sequence. (C) FastQC distribution of sequence lengths over all sequences before trimmomatic on DG-17 forward sequence. (D) FastQC quality distribution of sequence lenghts over all sequences before trimmomatic on DG-17 reverse sequence.



B

A



**Figure 4. Quast report.** (A) Spades on data 4 and 3 contigs (original untrimmed data). (B) Spades on data 12 and 11 contigs (trimmed data).



A



B

**Figure 5. Quast GC content graph.** (A) GC content percent for Spades on data 4 and 3 contigs. (B) GC content percent for Spades on data 12 and 11 contigs.



A

B



**Figure 6. Spades report.** (A) Assembly graph for data 4 and 3 contigs. (B) Assembly graph for data 12 and 11 contigs.

Bibliography:

Choi, Yoshikawa, Bridge, Schlaifer, Osterweil, Reid, and Norman. 1990 May. *Salmonella* outbreak in a nursing home. *NCBI.* 38(5):531-534.

F.W. Brenner, R.G. Villar, F.J. Angulo, R. Tauxe, B. Swaminathan. 2000 Jul. *Salmonella* nomenclature. *Journal of Clinical Microbiology.* 38(7):2465-2467.

Judith R. Glynn, Stephen R. Palmer. 1992 Dec. Incubation period, severity of disease, and infecting dose: evidence from a*Salmonella* outbreak. *American Journal of Epidemiology.* 136(11):1369-1377.

K. Rahn, S.A. De Grandis, R.C. Clarke, S.A. McEwen, J.E. Galan, C. Ginocchio, R. Curtiss III, C.L. Gyles. 1992 Jan. Amplification of an invA gene sequence of *Salmonella* typhimurium by polymerase chain reaction as a specific method of detection of *Salmonella*. *Molecular and Cellular Probes.* 6(4):271-279.

M.R. Evans, S.M. Parry, C. D. Ribeiro. 1995 Oct. *Salmonella* outbreak from microwave cooked food. *Epidemiology and Infection.* 115(2):227-230.

Sherry Rohekar, Florence W.L. Tsui, Hing Wo Tsui, Nancy XI, Reena Riarh, Rose Bilotta, Robert D. Inman. 2008 Aug. Symptomatic acute reactive arthritis after an outbreak of *Salmonella. The Journal of Rheumatology.* 35(8):1599-1602.

2016. Medical definition of contig. *MedicineNet.*

2018. Centers for disease control and prevention. *PulseNet.*

2018. UniProtKB - P41784 (PRGI\_SALTY). *UniProt.*