WHOLE GENOME DNA SEQUENCE ANALYSIS OF *SALMONELLA ENTERICA* SUBSPECIES ISOLATED FROM ENVIRONMENTAL SOIL AND FECAL SAMPLES IN WESTERN NORTH CAROLINA

A thesis presented to the faculty of the Graduate School of Western Carolina University in partial fulfillment of the requirements for the degree of Master of Science in Biology

By

David Austin Russell

Director: Dr. Mark Wilson
Associate Professor and Director, Forensic Science Program
Department of Chemistry and Physics

Committee Members: Dr. Indrani Bose, Biology
Dr. Seán O'Connell, Biology

September 2015

ACKNOWLEDGEMENTS

I would like to thank my committee and director for their assistance and encouragement. In particular I would like to thank Dr. Wilson for his role as my mentor.

I also extend special thanks to the following people, without whom this thesis would not have been possible: Patricia Melton and Sofia Lilly, Britannia Bintz my academic advisor, from the FDA, Mark Allard, Ruth Timme, Rebecca Bell, Erik Burrows, and David Melka, and from Illumina®, Cydne Holt, Joe Valero, Tom Richardson, and Kathryn Stephens.

Lastly, I offer my warmest regards to family and friends, especially my parents, for their continued support along the way.

TABLE OF CONTENTS

LISTS OF TABLES

LIST OF FIGURES

ABSTRACT

WHOLE GENOME DNA SEQUENCE ANALYSIS OF *SALMONELLA ENTERICA* SUBSPECIES ISOLATED FROM ENVIRONMENTAL SOIL AND FECAL SAMPLES IN WESTERN NORTH CAROLINA

David Austin Russell, M.S.

Western Carolina University (September, 2015)

Director: Dr. Mark R. Wilson

Foodborne bacterial pathogens like *Salmonella* genera remain of interest to regulatory agencies like the FDA and CDC. As a foodborne pathogen, capable of causing serious illness in both human and non-human animals, the CDC has listed *Salmonella* spp. as potential bioterrorism agents. From a forensic perspective, accurate and rapid identification of *Salmonella* subspecies is essential for successful investigation of foodborne outbreaks or suspected biocrimes. Massively parallel sequencing (MPS) provides investigators with a streamlined, cost-effective method to rapidly sequence the whole bacterial genome. To study the genetic variation of naturally occurring *Salmonella* spp., environmental samples were collected from areas around freshwater lakes, rivers and ponds in the Piedmont and mountains of western North Carolina. Nineteen *Salmonella* isolates were sequenced using the Illumina MiSeq producing high quality sequence data that were submitted to NCBI in an effort to build a comprehensive database containing whole genome sequences of bacterial pathogens. Distance–based phylogenetic trees were created using the sequence information. This method was shown to be susceptible to the quality of the given sequence data. kSNP, a SNP analysis program to create phylogenetic trees, was shown to produce trees of similar quality without the influences of sequence quality as found in distance-

based trees. Ultimately, the databases generated from MPS data can serve as a repository of phylogenetic information and population data to most effectively answer questions germane to bacterial forensics, such as identifying the source of a foodborne outbreak

CHAPTER ONE: INTRODUCTION

When a pathogenic organism threatens public and animal safety, or in instances of biocrimes, investigators are called to: 1) determine the source and routes of transmission, 2) identify the type of infectious agent, 3) characterize the nature of the outbreak or event, and 4) identify the person(s) responsible. From a forensic perspective, the ability to rapidly and accurately detect subtle differences between highly clonal bacterial populations of *Salmonella* spp. can assist investigators in elucidating the source of a threat.

Foodborne illnesses are a major cause for concern to public health in the United States. Approximately 42,000 cases of *Salmonella* infections (Salmonellosis) are reported to the CDC each year, although it is estimated that over one million people in the U.S. acquire foodborne infection of non-typhoidal *Salmonella* annually[1]. Discriminative methods for sub-culturing, identification, and sequencing of suspect *Salmonella* species for purposes of providing accurate trace back is paramount in instances of environmental contamination, biocrimes, and foodborne outbreaks.

Infection with *Salmonella enterica* subspecies can result in foodborne illness. The symptoms of *S. enterica* infection are diarrhea, fever, and abdominal cramps, which often develop 12 to 72 hours after infection. The illness usually lasts 4 to 7 days, and is self-limiting. However, in extreme cases it can enter the bloodstream and cause death[2,3]. The CDC has listed *Salmonella* as potential bioterrorism agents due to the severity of pathogenicity as well as the ubiquitous nature of the genus. *Salmonella enterica* is often found in foods including but not limited to meats, cheeses and nuts. As an enteric organism it is found in human and other mammalian animal digestive tracts as well as in reptiles[4].

Fresh fruits and vegetables have been largely identified as a primary source for the introduction of *Salmonella spp.* into the food supply. Additionally, *Salmonella* is often detected in sewage, freshwater, groundwater, and the soil[5-7]. Traditionally, health organizations, like the FDA and CDC, which monitor the outbreaks, have been able to characterize the isolates responsible. However, these methods are time consuming, laborious, and require specialized equipment[8]. In rural agricultural areas where fresh produce is grown and livestock are reared, the use of animal waste and human bio solids (treated sewage sludge) are a common farming practice employed as a practical way to fertilize the soil[5,9-11]. Animals are carriers of *Salmonella* and the use of their feces as fertilizer is one mode of introduction into the soil. Animal waste is used in greater amounts in agricultural areas as compared to human bio solids, thus it is the most common source of *Salmonella* contamination [5,12,13].

Salmonella can be introduced into the food supply from produce grown in contaminated soil. One proposed method is the adhesion of *Salmonella* to plant surfaces from soil splashing during rain events[5,9,13]. Produce allowed to germinate in contaminated soil also becomes susceptible to the internal colonization of *Salmonella*. Fresh produce contamination can also be attributed to the use of contaminated irrigation water[9,10,14,15]. For example, in a foodborne outbreak in 2005, *Salmonella* Newport, a serotype of *Salmonella enterica* subspecies *enterica* isolated from tomatoes, was traced back to the use of contaminated irrigation ponds[16]

There have been several foodborne outbreaks of *Salmonella* spp. in the U.S. originating from fresh produce[14,17-19]. One of the earlier applications of massively parallel sequencing (MPS) in this area came in 2009 when MPS was used for the molecular tracking of an outbreak of *Salmonella* Montevideo. This serovar was associated with contaminated red and black peppers used in spiced-meat production that affected 300 people in 44 states[20]. The observed PFGE

patterns of this particular outbreak-associated isolate; obtained from contaminated spiced-meats and clinical samples, appeared indistinguishable. A MPS approach was implemented to more fully resolve 35 genomes of *Salmonella enterica* subtype Montevideo collected from clinical samples, as well as geographically disparate food sources collected during previous outbreaks of this serovar. Genome sequencing data clearly revealed that there were subtle differences between a particular clinical isolate from California and the outbreak strain. Out of the entire genome, which is approximately 4.9 Mbp in length, there were only 56 single nucleotide polymorphisms (SNP) differences and a 100kb insertion of a bacteriophage[20], thus highlighting how advantageous GS is as a tool for distinguishing between closely related bacterial pathogens and identifying minor genetic differences between them.

1.1 Background for this study

Massively parallel sequencing, also called Next Generation Sequencing (NGS), has become a valuable tool in bacterial epidemiology and molecular microbiology research. The power of MPS allows researchers to generate sequencing data from bacterial genomes at an extraordinary level of resolution; extending to the ability to detect single nucleotide changes within entire genomes[21]. This level of resolution becomes important to health organizations like the CDC and FDA that monitor and identify outbreaks of foodborne bacterial pathogens.

Pulsed-field gel electrophoresis (PFGE) is the primary subtyping method used by PulseNet, a national laboratory network that studies foodborne disease organisms like *Salmonella*, to produce unique DNA fingerprints for pathogenic bacteria[22] It is a non-sequence-based typing method that utilizes restriction enzymes to cut bacterial DNA at specific locations and separate the resulting fragments by agarose gel electrophoresis. Unlike conventional gel electrophoresis, the polarity of the electrical current is alternated at predetermined time intervals,

which enable better separation of larger DNA fragments, generating a pattern or DNA fingerprint. Historically, pulsed-field gel electrophoresis (PFGE) has been regarded by clinical laboratories as the "gold standard" for determining the molecular relatedness of bacterial isolates, yeast, and fungi[23-26].

Once obtained, DNA fingerprints can be searched against other patterns in the PulseNet database to determine if the samples could have originated from a common source[8,22,24-26]. However, when comparing patterns of highly clonal and closely related bacterial populations, the ability to distinguish between these populations is limited when using the PFGE technique alone.

Sanger sequencing is a DNA sequencing approach developed By Frederick Sanger in 1975. The use of Sanger sequencing has been the method of choice to sequence both genomic DNA and mitochondrial (mtDNA) DNA. This method greatly improved on previous sequencing techniques developed by others during the same time and even his own 'plus minus' technique that he developed a few years prior. The classical chain terminator method involved the use of single-stranded DNA (ssDNA) template, DNA polymerase, standard nucleotide triphosphates (dNTP's), and radiolabeled dideoxynucleotide triphosphates (ddNTP's). This sequencing method required four individual reactions that contained many copies of fragmented ssDNA template, DNA polymerase, dNTP's and one of the four of the radioactively tagged ddNTP's (ddATP, ddCTP, ddGTP, & ddTTP). These ddNTP's lack a 3' hydroxyl (OH) group required to form the phosphodiester bond between the two nucleotides in the growing chain catalyzed by a DNA polymerase. This dideoxy-characteristic terminates DNA extension of the growing chain. Following DNA extension from the bound primers, the DNA is denatured and the terminated fragments are separated by gel-electrophoresis. Each lane of the gel should contain the

extension-termination product of only one nucleotide. The bands were visualized using autoradiography and the DNA sequence was determined by reading the x-ray film or gel image.

The foundations of the Sanger chain-terminating dideoxy DNA sequencing method used in early genome sequencing gave rise to automated DNA sequencing using fluorescent dyes to detect electrophoretically resolved DNA fragments within a instrument designed for this purpose. The use of fluorescent dyes and automated instruments eliminated the need for radioisotopes and toxic chemicals and has greatly simplified DNA sequencing, so much that the human genome project was completed using these technologies; making this Nobel Prize winning technique the most preferred method in the past 30 years[27]. After the completion of the Human Genome project and the development of the automated capillary electrophoresis instruments, researchers desired more powerful sequencing technologies that could obtain higher throughput and, importantly, were economical.

The entire human genome was sequenced using Sanger/capillary-based sequencing with an output of 3Gb (1X coverage) at a cost of $3 billion over 13 years. MPS offers a different approach to the limited scalability of traditional Sanger sequencing through the use of micro rectors or by attaching target DNA to be sequenced to a solid surface. These techniques are extremely high-throughput, allowing for millions of sequencing reactions to occur in parallel[28]. It is this massively parallel sequencing that has set NGS apart from conventional Sanger/capillary-based sequencing. MPS produces thousands to millions of reads per sample, increasing the depth of coverage (the average number of times each base is read in a sequencing run) to levels that are orders of magnitude higher than traditional Sanger sequencing. This allows for rapid sequencing of large stretches of DNA that may span the entire genome. The human genome can now be sequenced using MPS in a single run at a cost of approximately $15,000 with 30Gb of output

(10-fold increase on coverage)[28]. Due to the scalability of MPS technology, now several small bacterial or viral genomes can be sequenced simultaneously within a single run, while maintaining higher output and remaining cost-effective.

DNA sequencing data is being deposited into public databases at a faster rate than in the recent past, including thorough and complete metadata describing the isolates (i.e. information regarding the source, strain, serovar, location, etc.). Clinical databases pertaining to foodborne pathogens are typically created to study a specific disease outbreak (e.g., *S. enterica* Montevideo) currently threatening public health. However, these data sets are often not accessible to the public and provide little in the way of studying the natural population distribution of *Salmonella* species occurring within the environment. These databases can be extremely helpful in assisting investigators and public health officials during outbreak events. A complete database, containing whole genome sequencing data coupled with metadata of both clinical and environmental isolates, can provide known geographical distributions and past outbreak associations of a particular *Salmonella* serovar.

In an effort to build a comprehensive database containing whole genome sequences of bacterial pathogens, the FDA and other public health officials have coordinated an international network of laboratories to sequence pathogens collected from foodborne outbreaks, contaminated food products and environmental sources. These genomic sequences are archived in a public reference database called GenomeTrakr®[29]. Bioinformatic support and analysis for this open-access database is provided by the National Center for Biotechnology Information (NCBI). The GenomeTrakr® database currently contains over 10,500 *Salmonella* spp. isolates. The network is currently sequencing on average 800 isolates each month and as the database grows, so too will its strength as an investigative tool[29,30]. While this database is enabling

focused investigations into the root source of an outbreak, it will have the additional benefit of supporting researchers in understanding the conditions that lead up to an environmental contamination of agricultural products.

1.2 Objectives

Western Carolina University (WCU) has partnered with the U.S. Food and Drug Administration to utilize MPS methods for the characterization of naturally occurring *Salmonella* spp. The goals of the project were split among three phases, which are outlined in Figure 2, beginning with developing a sequencing strategy utilizing the Illumina® MiSeq®, a bench top MPS platform, to sequence the entire genomes of six bacterial isolates previously characterized as *Salmonella enterica* subsp. *enterica* serovar Enteritidis. These isolates were provided by the Center for Food Safety and Applied Nutrition (CFSAN) of the FDA for sequencing on the Illumina® MiSeq® instrument. This phase of the project was conducted to assess the effectiveness and efficiency of the library preparation and sequencing protocols for bacterial genomes on a bench top sequencer. This effort helped to assist the FDA with development of bioinformatic tools and pipelines that would help to shuttle data from sequencing platforms to private and public databases for storage, making these data available to investigators during an outbreak event. To facilitate the assistance to the FDA's data analysis pipeline construction, WCU's Forensic Science Program became a contributing laboratory in the GenomeTrakr® network.

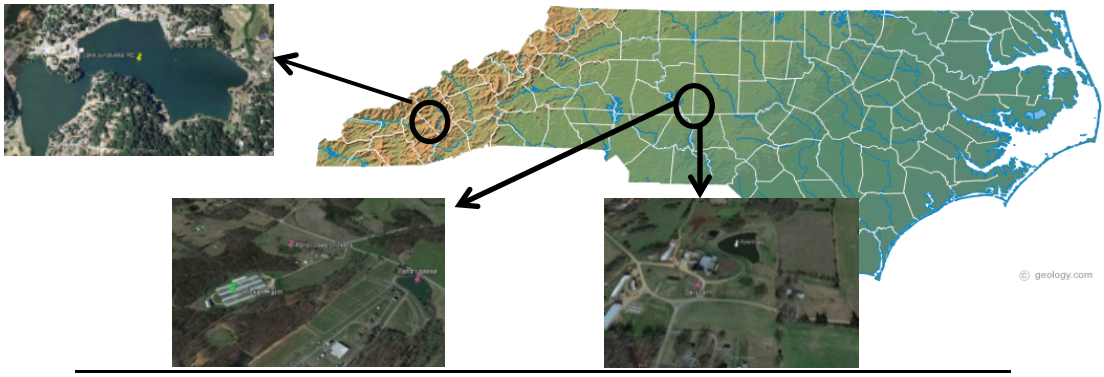
The second phase of this project assessed the microbiological techniques that are used by the FDA to select, differentiate, isolate and identify *Salmonella* spp. from different environmental sources. These techniques were used to design a protocol to isolate *Salmonella* spp. from environmental soil and fecal samples. The protocol was initially tested using bacterial

control strains purchased from American Type Culture Collection (ATTC.). The FDA suggested the use of control strains, because three of the six strains produce atypical colony morphology and unusual phenotypic reactions on the selective and deferential media being used. All six strains were used to evaluate the modified FDA protocol including familiarity with the colony morphology and biochemical responses of both non-S*almonella* species (e.g., *Enterococcus faecalis* and *Escherichia coli*) and S*almonella* species (*S. enterica, S. diarizonae, S. abortusequi*) in and on the different microbiological media.

The method by which the bacterial DNA was extracted was also evaluated using these bacterial strains (see methods for complete list). The DNA extraction method chosen was a kit-based extraction procedure that was also evaluated using test bacterial strains. The procedure included both hands on and automated tasks and was designed for forensically relevant samples such as buccal swabs, blood, gum and other materials. This kit, however, is not optimized for the use of bacterial samples. The protocol was amended to allow for addition of enzymes and incubation steps to ensure that the cells were lysed before DNA purification.

Once preliminary evaluation of the methods and protocols for isolation of bacterial isolates, DNA extraction, and genome sequencing were completed, the project progressed into the final phase, which involved obtaining environmental samples.

To maximize the detection of genetic variation of naturally occurring *Salmonella* spp., environmental samples were collected from areas around freshwater lakes, rivers and ponds in the Piedmont and mountains of western North Carolina. The sampling locations were in close proximity to waterfowl habitats, agricultural farmland, and animal rearing facilities. Sites such as these provide the most probable locations for collection of environmental *Salmonella* spp. (Figure 1).



Piedmont North Carolina		Western North Carolina	
DFP Pond	Diary Farm	Lake Junaluska	Parker Farm
5 Soil, 5 Fecal	5 Soil, 5 Fecal	5 Soil, 5 Fecal	5 Soil, 5 Fecal

Figure 1. Sampling sites and sample types. Listed are four sampling sites in the Piedmont and western North Carolina mountains and the types of samples that were collected from each site to culture naturally occurring *Salmonella* spp

Following collection, the samples were sub-cultured to enrichment, selective, and differential media, in serial steps, to select for probable *Salmonella* isolates. Enrichment media allowed for only some bacteria present within the chosen sample to become cultured after being removed from the environment, selective media uses defined chemicals (Table 1), which allow for the growth of specific microorganisms while inhibiting the growth of others; this includes selection of gram-negative bacteria, and inhibiting the growth of gram-positive bacteria, and differential media allowed for preliminary identification of a specific genus or species from pure cultures[31]. The use of all three different media types allowed for the isolation of *Salmonella.*

Table 1. Common reagents in differential and selective media. These reagents select for the isolation of *Salmonella* species based on their biochemical reactions.

Ingredient	Purpose
Bismuth Sulfite	Inhibitor of Gram-positive bacteria and other coliforms (selective media)
Brilliant green	Inhibitor of Gram-positive bacteria and other coliforms (selective media)
Bromocresol purple	pH indicator of fermentation reaction (both)
Ferrous sulfate	Reaction with hydrogen sulfide (H_2S) produce black precipitate; Indicator of H_2S production (differential media)
Sodium Thiosulfate	Reduced to hydrogen sulfide, indicator of H_2S production (differential media)
Phenol red	pH indicator of fermentation, decarboxylation, and deamination reactions (both)
L-lysine	Used to detect enzymes (lysine decarboxylase and lysine deaminase) (both)
XLT-4 Supplement	Inhibits the growth of non-*Salmonella* organisms (differential media)

Serological and biochemical tests were used to confirm that the bacterial isolates, obtained from the environment and then cultured in the laboratory were in fact *Salmonella.* Subtyping of identified *Salmonella* spp. was conducted using pulsed-field gel electrophoresis (PFGE). The DNA of the *Salmonella* isolates was extracted for subsequent sequencing on the Illumina® MiSeq™ to further characterize the whole genome. Using genome-sequencing data, phylogenetic trees were produced showing the genetic relationships of environmental *Salmonella spp.* populations in select locations in North Carolina.

1.3 Illumina® genome sequencing

The Illumina® MiSeq® is a MPS platform that is leading the sequencing industry and is capable of a range of applications from targeted amplicon re-sequencing to small genome sequencing, which are performed using its unique and highly accurate sequencing chemistry[32,33]. The Illumina® MiSeq® massively parallel sequencer is a sequencing-by-synthesis instrument that employs on-instrument cluster generation on the solid surface (flow cell) using bridge amplification and reversible-terminator chemistry for the detection of single base incorporation events (see methods for more details). Until recently (2012) the majority of MPS instruments have been geared towards large-scale applications, disregarding the needs of smaller laboratories working at a much smaller scale[33]. Illumina® has recently accommodated the needs of smaller

laboratories with a slightly lower throughput instrument. The MiSeq® is smaller than earlier versions, improving on reaction, run costs and the turn around time needed to obtain quality data. Its smaller footprint makes it suitable and economical for smaller research and clinical laboratories.

One initial and critical step in MPS is constructing DNA libraries. The DNA is prepared in a way so it is compatible with the sequencing system being used. This generally involves several core steps to prepare DNA for MPS analysis which include: fragmenting the target DNA, ensuring the DNA is double stranded, adding adaptors to the ends of the DNA fragments, and verifying the concentration of the final library product for sequencing. Along with ensuring adequate library products, the size of the inserts is equally as important. There are different approaches to fragmenting the input DNA: physical, chemical and enzymatic,[34] each capable of producing various ranges in DNA fragment size. The fragmentation method is important for library preparation to allow for the appropriate read length your wanting to achieve during sequencing.

For this project, DNA libraries were prepared using Illumina® Nextera® XT DNA library preparation kit. This kit utilizes enzymatic fragmentation. A transposase enzyme simultaneously fragments and then tags (Tagmentation) the dsDNA with adaptor sequences[35]. The incorporated adaptors serve as priming sites during limited cycle PCR that adds index sequences and sequencing adaptors to both ends of the tagmented DNA, thus enabling dual indexed sequencing of pooled libraries. The index sequences act as "barcodes," which allow for a high degree of multiplexing; which is sequencing of many different samples simultaneously[36]. Following sequencing, data from each multiplexed sample can be separated using the indices. This library preparation method includes steps to eliminate very small library fragments from the population

and to normalize the quantity of each library to ensure equal representation when libraries are pooled for sequencing.

1.4 Sequencing analysis

One of the pioneering approaches for sequence analysis involves creating sequence alignments. This is one way of arranging DNA sequences to identify areas of sequence overlap. Following sequencing, the raw sequencing reads are put together into larger segments using a similar genome as a reference or assembled *de novo* (without a reference). During sequence alignment, gaps are often inserted to allow for the sequences to align. For sequences that share a common ancestor, gaps in the alignment could indicate indels (insertion or deletion mutations) of nucleotides or genes. Mismatches in the aligned sequences could indicate point mutations.

Alignment–based methods can be performed by either pairwise comparisons or through multiple sequence alignment (MSA)[37,38]. Pairwise alignment compares each sequence to every other sequence until all comparisons have been made. MSA requires more complex methods to align multiple sequences utilizing an iterative method to repeatedly re-align sequences as more sequences are added to the growing MSA. The use of a reference genome in MSA does introduce another level of complexity requiring prior knowledge of the sequences under comparison.

Phylogenetic trees are used to represent evolutionary relationships and a history of organisms under study, in other words, a phylogeny. Phylogenetic trees are often inferred from DNA sequences and other data. Traditionally trees were constructed using two distinct kinds of methods, character and distance, both of which used sequence alignments as input. A distance matrix is created most simply by counting the number of dis-similarities between nucleotide sequences from pairwise comparison of each sample in the study. Clustering algorithms, like

neighbor joining or Unweighted Pair Group Method with Arithmetic mean (UPGMA), use a distance matrix to construct the tree.

Character-based methods, which include maximum likelihood (ML) and parsimony, use sequence alignments to construct and refine phylogenetic trees by searching all possible tree combinations to find the tree that best fits the observed sequence data[39]. ML tree estimations use substitution models to assign probability to specific mutations and rates of mutations that occur within each character state. Parsimony tree estimations simplify ML trees by requiring some assumptions to be made to apply a value or "cost" associated to specific evolutionary changes: such as nucleotide substitutions and insertions/deletions. Algorithms search through the information space containing all possible trees to find the tree with the smallest cumulative cost.

Character-based methods do have an advantage over distance methods in that they can reliably place character changes on a tree by introducing hierarchal weights to these changes. With enough supporting data SNP changes can define particular species groups. One major disadvantage of using character-based estimations is that the exhaustive searching of trees can be computationally extensive and time consuming and only effective for very small datasets. Distance based methods are fast and computationally efficient, because they do not take into consideration the assumption and weighted probabilities that are used in ML calculations. Additionally, gene absence and presence can be highly variable in diverse taxa[40] reducing the resolution leading to potential bias in the inferred phylogeny[41].

Alignment-free methods take a different approach to building matrices for character and distance trees. Some organisms under study may not have been sequenced before, or in the case of metagenomics (the simultaneous sequencing of entire populations) the identity of the organisms being sequenced may not be known. In the case of newly sequenced genomes, there

may only be partial genomes available or there may only be a few subspecies that have been extensively studied for use as comparison. This lack of appropriate sequences effectively eliminates the use of sequence alignments as a way to construct phylogenetic trees.

Alignment-free methods appear to eliminate some of the limitations of MSA by being computationally less intensive, fast and comparatively accurate[40,42,43]. This particular approach is based on extracting DNA subsequences of defined length called *words* or *k-mers*. The method consists of collecting sets of k-mers, from each sequence, and comparing those sets of k-mers pairwise to calculate pairwise distances. A distance tree is then built by comparing the number of shared k-mers. The more similar sequences are to each other, the smaller the pairwise distance[44] and the closer they will cluster on a distance-based tree. Currently NCBI has a growing tree of *Salmonella* genome sequence data that is built using this method. The k-mer distance tree method eliminates many of the issues that MSA imposes on sequence alignment of large and complex bacterial genomes.

Newer methods that identify single nucleotide polymorphisms (SNPs) from genome sequencing data result in a higher phylogenetic resolution for determining evolutionary histories of bacterial populations; even from multiple strains within the same clonal lineage[45,46].

Two SNP-based programs were used for data analysis. These were designed to identify SNPs and create phylogenetic trees based on genome sequencing data. The first is a web-server program that contains the tools for automatic SNP analysis and tree construction based on SNP data[47]. The web server, called snpTree, is freely accessible at http://www.cbs.dtu.dk/services/snpTree-1.0/. The server was created to handle GS data from assembled and unassembled raw-sequences. snpTree requires the use of a reference genome that can be uploaded by the user, or chosen from a list that contains over 2,000 complete genomes

collected from the NCBI genome database. The built-in toolbox uses currently available programs for mapping, genotyping and SNP calling (e.g. Burrows-Wheeler Aligner[48] (BWA) and SAMtools[49]). The design of snpTree allows users to modify only a few of the settings with the goal of being user friendly for users with limited bioinformatic knowledge and experience[46].

The second SNP analysis program, kSNP v2, requires more bioinformatic knowledge. It is freely available at http://sourceforge.net/projects/ksnp/. It is also designed to identify SNPs across the entire genome. It offers many more options (arguments) that can be defined by the user, and because of the many applications of SNP analysis, there are numerous output files that can be used for downstream analysis. kSNP is also capable of handling assembled genomes and contigs as well as raw sequences[50,51]. Unlike the snpTree server, which requires fastq input files, the kSNP program only allows input of one fasta file. This limitation requires the user to convert the raw reads from fastq to fasta, merge reads, and concatenate merged files.

In-house analysis of *Salmonella* sequencing data was performed using each program followed by evaluating the possible affects of upstream sequencing quality on the inferred phylogenetic trees. K-mer distance trees are more susceptible to poorer-quality sequencing data because of the way the trees are built (similar subsequences is indicative of smaller evolutionary distance), but SNP analysis tends not to be affected by the sequencing quality. Because kSNP utilizes k-mers to identify SNP loci; it is of interest to see if quality will impact the analyses.

The resulting whole genome sequence data, combined with other whole genome sequences from *Salmonella* isolates, were used to produce a phylogenetic tree showing the locations of *Salmonella* spp. occurring naturally in North Carolina. In addition to phylogenetic analysis, sequence data was uploaded to the Sequence Read Archive (SRA) of the National

Center for Biotechnology Information (NCBI) as part of WCU's involvement and support in the GenomeTrakr® network.

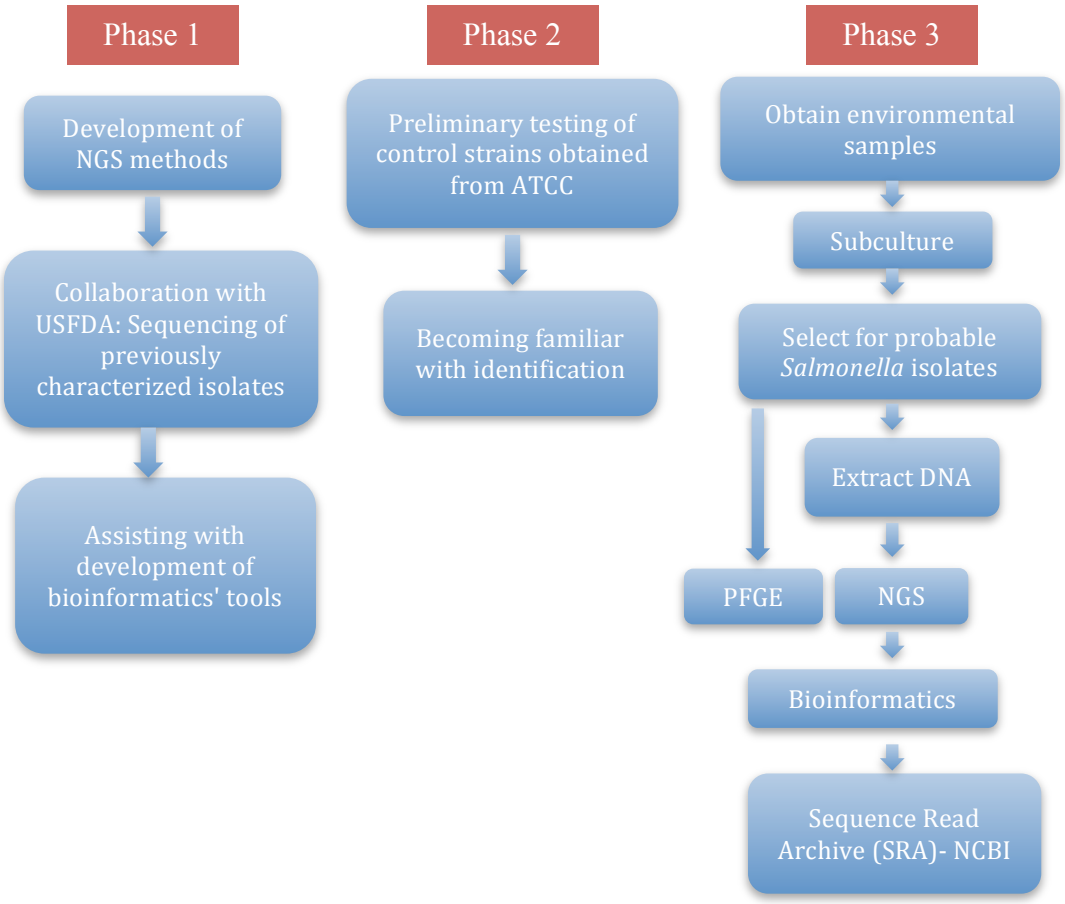


Figure 2. **WCU *Salmonella* project workflow**. This workflow summarizes each stage and accomplished goals throughout the project.

CHAPTER TWO: MATERIALS AND METHODS

Specific methods were utilized to isolate *Salmonella* from environmental samples. The identities of the bacterial isolates were confirmed as *Salmonella* using microbiological, biochemical and serological techniques. Genomic DNA was purified from confirmed *Salmonella* isolates and characterization of *Salmonella* was performed using next-generation sequencing and SNP analysis programs.

2.1 Bacterial controls

The ATCC™ strains (*Enterococcus faecalis* (Andrewes and Horder) Schleifer and KilpperBalz (ATCC® 1943™); *Escherichia coli* (Migula) Castellani and Chalmers (ATCC® 10536™); *Salmonella enterica* subsp. *enterica* (*ex* Kauffmann and Edwards) Le Minor and Popoff serovar Newport (ATCC® 6962™); *Salmonella enterica* subsp. *diarizonae* (Le Minor et al.) Le Minor and Popoff (ATCC® 29934™); *Salmonella enterica* subsp. *diarizonae* (Le Minor et al.) Le Minor and Popoff (ATCC® 12325™); *Salmonella enterica* subsp. *enterica* (*ex* Kauffmann and Edwards) Le Minor and Popoff serovar Abortusequi (ATCC® 9842™) were purchased as recommended controls during the isolation technique used by the FDA. The lyophilized cells were rehydrated in the appropriate liquid media and maintained in culture.

The isolation techniques that were used in this project were part of the Bacteriological Analytical Manual (BAM); which is a compilation of the FDA-preferred assays for the testing of foods and cosmetics[52]. Chapter 5 details various ways to isolate *Salmonella* spp. from various food and cosmetic items. These methods use traditional microbiological techniques to recover, grow and isolate bacterial colonies on and in different culture media. The cellular morphology

and, more importantly, the phenotypic response of the organism to the particular media will give some indication of the particular bacterium present.

2.2 Sample collection and isolation

Samples that were collected from each location were weighed using a digital scale (Fisher Scientific) on site before placing into Whirl-pak® bags (Nasco). Masses of 100±4g soil/sediment samples and 25±3g fecal samples were placed in Whirl-pak® bags and placed in a styrofoam cooler with ice packs to keep the samples cool. A total volume of 225ml of Modified Buffered Peptone Water (MBPW), supplemented with Acriflavine [10mg/L], Cefsulodin [10mg/L], and vancomycin [8mg/L], was added to each sample within 72 hours after collection. The bags containing MBPW were agitated for 1-2 minutes by hand to suspend samples in the pre-enrichment media and incubated for 24±3 hours at 35°C allowing for recovery of bacteria present in the sample.

After incubation the samples were removed from 35°C. One milliliter of each sample was transferred to 10ml of tetrothionate (TT) broth supplemented with novobiocin at a concentration of 20mg/L. Additionally; 0.1ml of each sample was transferred to 10ml of Rappaport-Vasidillias (RV) broth. Both pre-enrichments were placed at 42±0.5°C for 24 hours. Following the 24-hour incubation, a 10µl loopful was removed from both TT and RV and streaked to each of the three selective plates, bismuth sulfate (BS) agar, Hektoen enteric (HE) agar, and XTL-4 agar.

The plates were incubated at 35°C for approximately 24 hours. The following day a well-isolated colony, demonstrating the appropriate characteristics of *Salmonella* on HE[a], BS[b], and

[a] Blue -green colonies with or without black centers. Many cultures of *Salmonella* may produce colonies with large, glossy black centers

[b] Brown, gray, or black colonies; sometimes they have a metallic sheen. Surrounding medium is usually brown at first, but may turn black in time with increased incubation, producing the so-called halo effect

XLT-4[c] was removed and re-streaked to a new plate of HE and XLT-4 to confirm and further isolate any suspected *Salmonella* colonies. Again, a well-isolated colony appearing as *Salmonella* and producing the appropriate phenotypic response in the plate media was selected with a sterile needle and stabbed to the butt of a tube containing triple sugar iron (TSI) and streaked along its slant. The same needle was used to stab the butt of the tube containing lysine iron agar (LIA) twice and then streaked along its slant. The tubes that appear as *Salmonella* spp. were retained, and if both slants and isolation streaks exhibited characteristics of *Salmonella* the samples were carried onto preliminary identification using the EnteroPluri-test.

2.3 *Salmonella* identification

The EnteroPluri-test is a single use device for the identification of *Enterobacteriaceae* and other gram-negative bacteria[53]. The test is divided into 12 sections; each compartment contains a unique culture media used for identification based on the pattern of response. The device (Table 2) allows for simultaneous inoculation of all 12 media compartments and detection of 15 biochemical reactions. This commercially available product was supplemented into the biochemical testing workflow for the presumptive identification of *Salmonella*. One or two colonies were selected from only the TSI slant and inoculated to the EnteroPluri tube and incubated for 24 hours at 35°C. The EnteroPluri tubes that were indicative of possible *Salmonella* spp. were investigated further using *Salmonella* antisera.

[c] Pink colonies with or without black centers (H_2S negative). Many cultures of *Salmonella* may produce colonies with large, glossy black centers or may appear as almost completely black colonies (H_2S positive)

Table 2. EnteroPluri test code sheet. The table below represents the 15-biochemical reactions that are involved with this compartmentalized testing system. A coding system allows for generic identification of Enterobacteriaceae.

	Group 1			Group2			Group 3			Group 4			Group 5		
Test	Glucose	Gas	Lysine	Ornithine	H_2S	Indole	Adonitol	Lactose	Arabinose	Sorbitol	VP	Dulcitol	PA	Urea	Citrate
Positivity Code	4	2	1	4	2	1	4	2	1	4	2	1	4	2	1
Result	+	+	+	+	+	-	-	-	+	+	-	+	-	-	+
Code sum		**7**			**6**			**1**			**5**			**1**	
Identification	*Salmonella spp.* subsp. *Cholerasuis*									*Escherichia coli*					

Salmonella spp. contains specific cell surface antigens. Serological testing was performed on presumptive positive *Salmonella* isolates using polyvalent flagellar (H) and polyvalent somatic (O) antisera[54] targeting these particular antigens. The antiserum, obtained from PRO-LAB diagnostics, contains specific *Salmonella* antibodies which, in the presence of homologous antigens (i.e., *Salmonella cells*), cause bacterial aggregation[54]. Positive cultures were indicated by agglutination (granular clumping) in the test mixture, (liquid culture plus antiserum), and not in the control (liquid culture plus saline).

The colonies that were used for the EnteroPluri-test were also used to prepare liquid cultures in brain heart infusion (BHI) broth for same-day analysis. On a clean glass slide and using a wax pencil, 4 boxes approximately $4cm^2$ were drawn in pairs and are referred to as the test chambers (Figure 3).

A 3mm loop of prepared 0.85% saline solution was placed in both control (CTRL) boxes along with a loopful of liquid culture. A loopful of liquid culture was added to each of the antisera chambers along with the appropriate antisera. The glass slide was rocked back and forth for approximately 1 minute. The glass slide was then observed with a Nikon Eclipse TS1000 under 40X magnification. Observation of distinct agglutination (granular clumping) within 60

seconds, and a lack of agglutination in the saline control were indicative of a positive result for *Salmonella*.

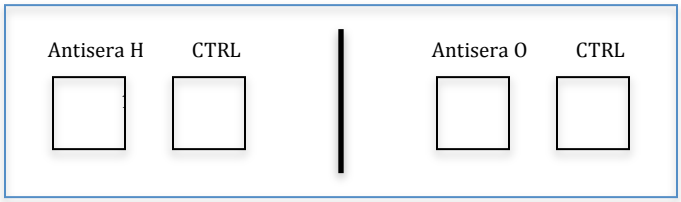


Figure 3. Antisera test chamber to identify *Salmonella* spp.

2.4 DNA extraction and quantification

Extraction of bacterial DNA was performed using the Qiagen® EZ1® Advanced XL using the EZ1® DNA Investigator® kit (Qiagen). The EZ1® Advanced XL is a fully automated instrument that can simultaneously purify DNA from up to 14 samples[55]

Isolates that were indicated as possible *Salmonella* species via the EnteroPluri Test, and confirmed by the serological test were inoculated into 5 ml of nutrient broth (Difco) and incubated for 24 hours at 37°C. Two hundred microliters of the overnight culture was removed and centrifuged for 5 min at 8500 rpm. The supernatant was removed and the pellet was re-suspended in 180 µl of Buffer G2 (Qiagen) and 20µl of proteinase K.

The re-suspended pellet of gram-negative bacteria was incubated at 56°C for 15 minutes. After the initial preparation step, samples were loaded onto the EZ1 Advanced XL along with a reagent cartridge provided with the kit. Once on the instrument, following the "tip-dance" protocol, the samples were subjected to further lysis and subsequent DNA purification with the use of magnetic beads coated with silica. In the presence of chaotropic salts the DNA binds to the silica beads and with the assistance of the magnet; the DNA was removed from the lysate and

washed of any residual cellular material and salts. After a series of wash steps the DNA was eluted into 100µl of TE buffer.

Purified bacterial DNA was quantified using Invitrogen's ™ Qubit® 2.0 fluorometer. The double stranded DNA (dsDNA) high-sensitivity (HS) assays used for quantification are highly selective for dsDNA over RNA that may be present; containing fluorescent dyes that are specific to targets of interest and only emit a fluorescent signal when the dye is bound to the dsDNA[56]. This assay requires the preparation of a working solution consisting of a 1:200 dilution of HS reagent: dsDNA HS buffer. The standards were prepared by combining 190µl of the working solution and 10µl of the Qubit™ standard. For each of the samples to be quantified, 195µl of the working solution was combined with 5µl of extracted DNA. Samples were vortexed for 2-3 seconds and then allowed to incubate at room temperature for 2 minutes before loading in the instrument.

2.5 Pulsed field gel electrophoresis (PFGE)

2.5.1 Culture growth

Samples were removed from -20°C storage and packaged in dry ice for transport to FDA in Maryland. In 2-ml cryo-vials, which remained frozen until their removal from dry ice and allowed to thaw. Once the 25% glycerol and culture suspensions were thawed, a 3 ml loop of each sample was inoculated to half of a petri plate containing trypticase soy Agar with 5% defibrinated sheep blood (TSA-SB). Cultures were incubated at 37°C for 14-18 hours.

2.5.2 Making plugs

Growth from the agar plates was removed using a sterile cotton swab and suspended in 2 ml of cell suspension buffer (CSB)[100mM Tris:100 mM EDTA, pH 8.0]. The concentration of the cell suspension was measured using a microscan turbidity meter (Dade Behring) and adjusted

to values that range from 0.40-0.45. Prepared TE buffer (10mM Tris:1 mM EDTA, pH 8.0) was used to make 1% SeaKem Gold agarose for PFGE plugs. Prepared melted agarose and TE were placed in a water bath (54-55°C) to equilibrate for 15 minutes or until use.

2.5.3 Casting plugs

Four hundred microliters of the cell suspensions were transferred to 1.5 microcentrifuge tubes to which 20µl of proteinase K (20mg/ml stock) was added. 400µl of melted 1% SeaKem Gold agarose was added to the 400µl cell suspensions and mixed by pipetting up and down 3-5 times. The temperature of the agarose was maintained by keeping the flask of agarose and cell suspensions in the water bath (54-55°C) during this procedure. Once the cell suspension and agarose were mixed, part of the mixture was transferred to the appropriate wells of a plug mold and allowed to solidify at room temperature for 10-15 minutes. Two plugs were cast for each sample.

2.5.4 Lysis of cells in plugs

To prepare the master mix, 5ml of prepared Cell Lysis Buffer (CLB)[50mM Tris: 50 mM EDTA, pH 8.0 +1% Sarcosyl] per tube and 25µl of proteinase K stock solution (20mg/ml) per tube were added to an appropriately sized flask and mixed well. Five milliliters of the master mix was added to each of the labeled 50ml polypropylene screw-cap tubes. Excess agarose from the plug molds were removed with a razorblade and discarded. The two plugs for each sample were removed with a spatula and placed into its corresponding tube making sure the plugs were submerged in the proteinase K/lysis buffer. Fifty milliliter tubes were incubated in a 54-55°C shaker water bath for 1.5-2 hours.

2.5.5 Washing of plugs after lysis

After incubation, the tubes were removed from the water bath and the lysis buffer was poured off using screened caps to avoid losing the plugs. Ten to fifteen milliliters of pre-heated (54-55°C) sterile ultrapure water (CLRW) was added to each tube. The tubes were returned to the shaker water bath for 10-15 minutes. After incubation, the water was poured off and the water wash step was repeated. 10-15ml of pre-heated sterile TE buffer (10mM Tris:1mM EDTA, pH 8.0) was added to each tube after the final water wash. All tubes were placed in a shaker water bath for 10-15 minutes. After incubation, the TE buffer was decanted and the TE wash step was repeated three more times. After the final wash step, the TE buffer was removed and replaced with 5-10 ml of sterile TE. The sample tubes were capped and placed at 4°C overnight.

2.5.6 Restriction digestion of DNA in agarose plugs

Restriction buffer was prepared by diluting 10X restriction buffer 1:10 with sterile ultrapure water (CLRW) [CLRW: 180µl/Plug slice + 10X restriction buffer: 20µl/plug slice = total vol. 200µl]. A single plug was removed from each tube and placed on a sterile disposable petri dish. A 2.0-2.5 mm wide slice was taken from each test sample and the *Salmonella* ser. Braenderup H9812 size standards with a razor blade. The cut slices were placed into corresponding 1.5-ml microcentrifuge tubes containing 200µl of the diluted restriction buffer (1X). Sample and control plug slices were incubated at room temperature for 10-15 minutes. Following incubation the restriction buffer was removed from the tubes and discarded. Two hundred microliters of prepared restriction enzyme master mix [CLRW: 173µl/plug slice + 10X restriction buffer: 20 µl/plug slice + BSA (10mg/ml): 2 µl/plug slice + Xbal (10U/µl): 5 µl/plug slice = total vol.:200µl] was added to each sample and control tube. Samples were incubated in a 37°C heat block for 1.5-2hours.

2.5.7 Casting agarose gel

1% SeaKem Gold Agarose was prepared in 0.5X Tris-Borate EDTA Buffer (TBE), melted and placed in a water bath (55-60°C) to equilibrate for 15 min or until use. Digested plug slices were removed from the 37°C heating block. *Salmonella* ser. Braenderup H9812 size standards were loaded onto the bottom of the comb teeth. The sample plug slices were loaded on the remaining teeth; keeping note of their location. The slices were allowed to air dry while positioned on the comb for 3-5 minutes. The comb, with teeth 10mm wide, was placed in the gel mold positioned atop a leveling platform, and the cooled (55-60°C) agarose was poured and allowed to solidify for 30-45 minutes before the comb was removed. Two liters of freshly prepared 0.5X TBE was added to the chamber and the pumps were calibrated to a flow rate of 1 liter/minute and the cooling module set to 14°C 30 minutes prior to running the gel. Once removed from the mold, the gel was secured into the electrophoresis chamber and run under the following conditions on CHEF mapper: Voltage Gradient: 6 V/cm; Included angle: 120°; Ramping: linear; Initial switch time: 2.16s; Final switch time 63.8 s; Runtime: 19h; Initial milliamps: 120ma; and Temperature: 14°C.

2.5.8 Staining and documenting gel

After completion of the run, the gel was removed from the chamber and stained for 20-30 minutes with ethidium bromide (EtBr) by diluting 40µl of EtBr stock solution (10mg/ml) with 400ml of ultrapure water (CLRW) in a covered container. Gels were de-stained with 500µl of reagent grade water for approximately 30 minutes. The de-stained gel was imaged according to the directions provided with the imaging equipment. Files formatted as tif files were analyzed using BioNumerics software program.

2.6 Illumina® Nextera® XT DNA library preparation

Genomic libraries were prepared using Illumina® Nextera® XT DNA library preparation kit; revision C (part #15031942). Initially, genomic DNA was randomly fragmented producing lengths of approximately 300bps and simultaneously tagged with adaptors. During a limited cycle amplification reaction, oligonucleotides complementary to sequencing primers and indices were added. The products were cleaned using 90 µl of Agencourt® AMPure® XP beads (Beckman Coulter). Following normalization, the libraries were pooled and then diluted. The protocol was amended to account for over-clustering in previous sequencing runs. The recommended dilution factor of 25X or 24µl of DNA input was increased to 28.5X or 21µl of DNA input, effectively decreasing the amount of DNA input into the diluted library by 12.5% (Increasing or decreasing the input volume of the pooled library into the diluted library by 10-20% can resolve over-clustering and under-clustering issues, respectively). PhiX control was used in all runs on the Illumina® MiSeq®, and was spiked in at 1% by volume into the diluted amplicon library (DAL).

2.7 Illumina® MiSeq® sequencing chemistry

Fragmented DNA molecules with ligated adapter oligos at either end were able to hybridize to one of two complementary oligos that are attached to the surface of the flow cell. The flow cell is the solid substrate that anchors the DNA in place during sequencing[32]. Once the adaptor region of the DNA fragment binds to the oligo on the flow cell, a polymerase creates a complementary sequence of the hybridized DNA fragment. The resulting dsDNA is denatured and the template strand is removed, leaving its complement covalently bound to the flow cell (Figure 4). Each strand was then clonally amplified through bridge amplification. The bound DNA molecule folds over to hybridize the free adaptor region to the second oligo on the flow cell. The polymerase generates the complementary strand forming a double stranded DNA

bridge. The resulting double-stranded DNA molecule is denatured leaving two single stranded DNA molecules (forward and reverse) covalently bound to the flow cell. This process is repeated and occurs simultaneously across the flow cell resulting in the formation of clusters. Each cluster is a clonal population of a single DNA fragment[57]. After amplification the reverse strands are removed from each cluster leaving only the forward strand to be sequenced. The 3' ends are blocked to prevent unwanted binding[58,59].

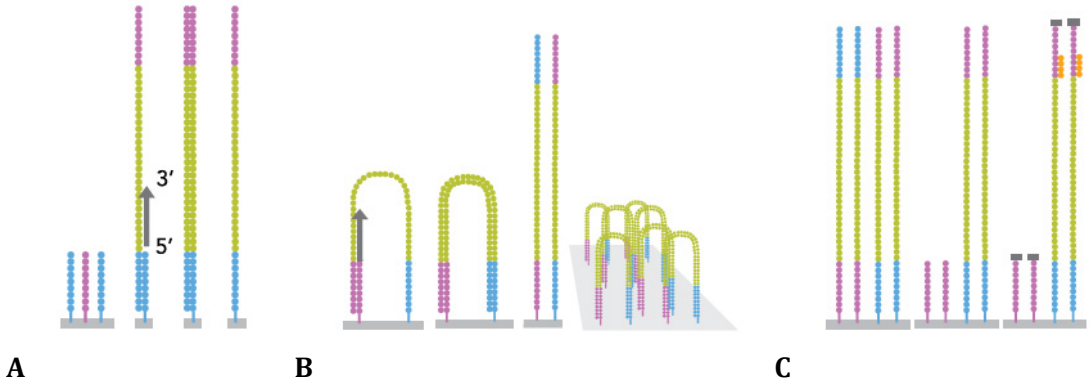


Figure 4. Illumina® MiSeq® cluster generation by bridge amplification. DNA fragments are clonally amplified through bridge amplification to form clusters. A) During Nextera XT Library preparation, DNA fragments are equipped with adaptor sequences that are able to bind to two types of oligos on the flow cell surface. Once bound, a DNA polymerase creates the complementary strand. The template molecule is denatured and washed away. B) The adaptor region of the bound DNA molecule binds to the other oligo on the flow cell. A DNA polymerase creates a complementary strand, forming a double-stranded DNA bridge, which is then denatured. The process occurs over and over simultaneously across the surface of the flow cell, generating millions of clusters. This results in the clonal amplification of all the DNA fragments. C) Following clonal amplification each cluster contains both the forward and reverse strand. The reverse strands are cleaved and washed away, leaving only the forward strands. The 3' ends are blocked to prevent unwanted primer binding (https://www.illumina.com/technology/next-generation-sequencing/sequencing-technology.html)

Once the on-instrument cluster generation is complete, sequencing can take place. All four of the reversible-terminator dNTP's are flowed across the flow cell simultaneously, with

only one modified nucleotide incorporated at a time during each cycle[60,61]. This natural competition decreases incorporation bias. Single base incorporation is ensured due to the fact that the ddNTP's contain a blocking group on the 3' end (Figure 5)[32,57,61]. Following incorporation, the flow cell is washed of any remaining un-incorporated nucleotides. An LED light source then excites the incorporated fluorescently labeled ddNTP, and the resulting light emission is recorded by a CCD camera[57]. Each nucleotide has a fluorophore, with a characteristic emission wavelength. After the signal has been recorded, the fluorophore is cleaved, and the terminator is removed exposing the 3' hydroxyl group for the next base incorporation[58]. This process continues until the read length is reached. DNA fragments can then undergo a paired-end-turn-around.

Paired-end sequencing allows for both the forward and reverse strands to be sequenced, creating dual interrogation of bases form both directions of the same template, increasing the accuracy of base calls. For this to occur, the sequencing product is removed, the indices are sequenced, and the opposite strand of the DNA template is de-protected. This allows the adaptor region on the DNA fragment to hybridize to a complementary oligonucleotide on the flow cell forming the DNA bridge. The polymerase creates the complementary strand, and the dsDNA bridge is then denatured. The forward strand, which was used as a template, is removed and washed away, leaving the reverse strand available for sequencing.

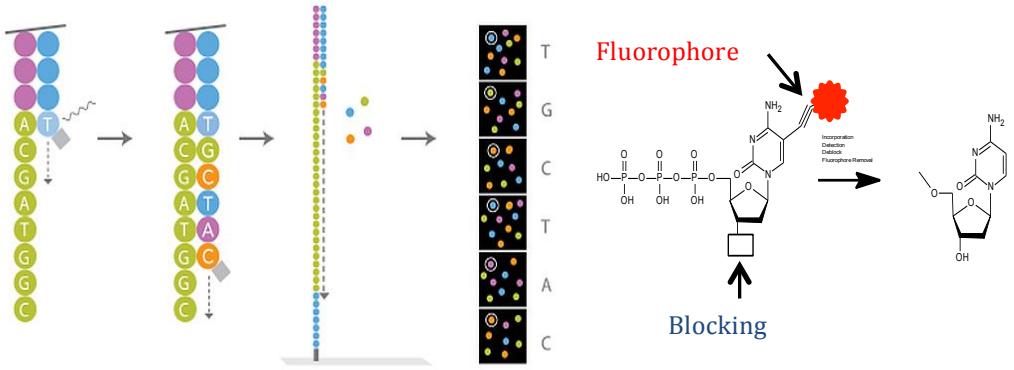


Figure 5. Illumina® MiSeq® sequencing chemistry. Only one nucleotide is incorporated at a time. After incorporation the fluorophore is excited and its resulting emission is captured on CCD. Once the image is captured the blocking group and fluorophore are cleaved away.[32]

Sequencing files were stored as raw reads on the instrument and within the cloud computing and storage environment for Illumina's next-generation sequencing platforms, called BaseSpace. This cloud platform allowed for real-time monitoring of data and viewing of quality performance metrics. Three key metrics were monitored during and after the completion of the run to assess the overall quality; the data quality score (Q-Score); the cluster density (k/mm^2); and the clusters passing (%) filter. The performance parameters used for the MiSeq v2 reagent kit 2x150 bp were as follows: >80% if bases with a Q-score of Q30 (averaged across the entire run), cluster density of 500-1000k/mm^2 and >80% of cluster passing filter. These key parameters, as well as other performance features, were monitored on the instrument or remotely through BaseSpace.

2.8 SNP analysis programs

snpTree is a online server that was used to analyze SNPs and produce SNP trees, utilizing an in-house toolbox containing available programs for next-generation sequencing analysis. SnpTree can analyze assembled genomes and raw sequencing reads.

Raw reads, downloaded from the MiSeq® instrument, were uploaded to the snpTree server for SNP analysis. Prior to the sequence reads being mapped to a user uploaded reference genome, or a reference chosen from the thousands of references collected from NCBIs Genome database, the fastq files were filtered and trimmed using the following criteria: (i) reads with N's, or ambiguous bases, were removed, (ii) reads that match at a minimum of 25nt of a primer or adaptor sequence on the 5' end were trimmed, (iii) the bases at the 3' end of the reads were trimmed until the Q-score was ≥ 20 (A Q-score is a logarithmic measure of estimated error for that base, for instance, a Q-score of 20 (Q20) means there is a probability of 0.01% that the base call was incorrect), (iv) The minimum average quality of the reads used was Q20 and the minimum size of reads was 20bp. Once filtered and trimmed, the raw reads were aligned to the chosen reference using BWA[48]. SAM tools were used for SNP calling and filtering. This is the software package for aligning DNA sequences in a generic alignment format (SAM/BAM format)[49]. The user-selected parameters for filtering SNPs were minimum coverage (number of times it was observed) and minimum distance between SNPs (the number of bp between a SNP). Both were set to 20. These parameters remained the same for both data sets analyzed. The VCF files containing the called SNPs were aligned and passed to Fastree[62] a program that created a maximum likelihood tree. Several output files were produced and made available after the server has processed the job such as (i) images of the tree, made from the identified SNPs, in PNG and SVG format (ii) SNP files in Newick format that were used to visualize the SNP trees in phylogenetic programs like FigTree, (iii) SNP annotation files which include an overview of all identified SNP positions, amino acid changes between reference and query genomes, SNPs differences between genomes, and the genomes that contain each particular SNP.

kSNP[51] v2.1 is a SNP analysis program that is designed to run in the terminal heads for Apple OS and Linux CPUs. kSNP was run on a Linux 64-bit CPU server at Western Carolina University. The kSNP package contains all the necessary programs as well as third party programs that were used to identify SNPs and estimate phylogenetic trees. Before analysis began, the raw fastq files generated on the MiSeq® were processed in two ways: (i) the files were filtered and trimmed according to the criteria specified by the snpTree server, (ii) fastq files were converted to fasta and then merged and concatenated to produce a single fasta file to be used as the input file for kSNP. Filtering and trimming of the raw fastq files was done using the tools on Galaxy's public space at http://usegalaxy.org. Galaxy is an open source web based platform for data intensive research.

The raw sequencing data was processed using the following workflow: (i) *FastQC* was used to generate a quality control report to assess data quality from high-throughput sequencers (this tool was used before and after the quality trimming and filtering), (ii) *FASTQ Groomer*[63] was used to convert from Sanger & Illumina 1.8+ quality score format to the recommended Sanger Quality score format (which Galaxy tools are designed to work with), (iii) *FASTQ Quality Trimmer*[63] was used to trim the 3' ends by using a sliding window (A window size of 1 stepped from the 3' to 5' direction 1bp until the aggregation or total quality score of the base(s) within the window met a minimum quality score of ≥ 20), (iv) finally, *Filter FASTQ*[63] was used to filter reads on a minimum size of 20bp and an average quality score of ≥ 20. This workflow produced fastq files that have their 3' tails trimmed using a quality score of ≥ 20, a minimum average quality of reads ≥ 20, and read lengths that are ≥ 20bp .

kSNP requires a single .fasta input file. Fasta files filtered and trimmed using Galaxy tools were converted from fastq to fasta, essentially removing the imbedded quality scores. The

kSNP package contains tools for fasta file preparation for use in the kSNP program. *merge_fasta_reads* was used to concatenate the multiple unassembled (raw) reads in a .fasta file into a single sequence under one fasta header. *cat* was used to then combine the merged files created into a single .fasta file which was used as the input file. Raw sequences require additional file(s) to be generated to accompany the input file. *genome_names* was used to extract the genome names of each of the genomes in the input file and generate a list containing those names. This allows kSNP to process the unassembled genomes through an extra step to remove k-mers that only occur less than a specified number of times. This has the effect of avoiding the consideration of sequencing errors as SNPs. *Kchooser* was used to identify the optimum value of k (the best k-mer size) in which kSNP is likely to identify SNPs from the given dataset.

kSNP used the input file and enumerated all of the k-mers in each genome using jellyfish[64]. This program removes, from raw genomes, k-mers that only occur once, and k-mers that would result in allele conflict. kSNP compared all k-mers across all genomes to identify SNP loci[d], meaning it identifies k-mers in which there are allelic differences between at least two genomes. SNP alleles were enumerated for each genome by comparing the k-mer list for that genome to the list of SNP loci. SNP matrices were generated from core SNPs (SNPs that are shared in all genomes). Trees were built using maximum likelihood with FastTree[62].

2.9 Independent sequence assembly and analysis

SNP analysis was performed by the New York State Department of Health. *Salmonella* Enteritidis strain P125109 was used as a reference genome to map the sequence reads and find positions with single nucleotide polymorphisms (SNPs). The raw reads were mapped over the reference genome using BWA-MEM Version: 0.7.5a-r405 with default parameters. The reads were sorted and duplicate reads were removed using Picard-tools Version 1.27. Read mapping

[d] A SNP locus is defined by the k-mer sequence surrounding the central base, which is the SNP allele.

statistics were extracted with Samtools flagstat Version: 0.1.19- 4428cd and final coverage statistics were retrieved using genomeCoverageBed from the Bedtools package v2.17.0.

A final read pileup was generated using Samtools mpileup and the variant call file (VCF) was produced with BCFtools Version: 0.1.19-44428cd, ignoring indels. Each individual genome position in the VCF file (variant and wt positions) was assessed to determine the exact nucleotide state in the sequenced genome and to create a high quality consensus sequence.

To identify a SNP, a genome position was required to have at least 20x depth of coverage of high quality mapped reads with 95% of the reads in agreement, as determined by the DP4 field in the VCF file. Positions that failed these requirements, or positions that mapped over phage-associated islands and/or repeat regions, were marked as unknown state (N's) in the consensus sequence. Genomic coordinates corresponding to phage sequences and repetitive elements in the reference genome were determined using Phast and Mummer, respectively.

The SNP alignment was created by comparing all of the resulting consensus sequences and retrieving positions where at least one of the sequences experienced a nucleotide change compared to the reference genome. The maximum likelihood phylogenetic tree was calculated with PhyML using a K80 (K2P) model, no gamma and the SPR tree search algorithm. A SNP heatmap was calculated in R version 3.1.2 with the Package 'gplots' using the ratio of [# of SNP differences / total number of non-'N' positions] between any pairwise consensus sequence comparisons. The numbers within each cell in the heatmap correspond to the number of SNP differences between each sample. Pascal Lapierre, Ph.D, a research scientist at the NY department of Health, Wadsworth Center conducted the analysis. The samples were run through a pipeline that he built for an internal *Salmonella* surveillance project.

CHAPTER THREE: RESULTS

3.1 Sample collection and identification

Using traditional microbiological techniques to recover, grow and isolate bacterial colonies, 25 isolates from the Piedmont region and 21 isolates from the mountain region were presumptively identified using the EnteroPluri-test (lyoflochem). Identifications were made using the EnteroPluri codebook.

In addition to the EnteroPluri test all 25 isolates from the Piedmont and 21 from the mountains were evaluated using polyvalent 'O' (somatic) and monovalent 'H' (flagella) *Salmonella* antisera. ATCC controls were evaluated along with the samples to ensure proper performance of the antisera. Table 3 shows that five isolates from the Piedmont and 14 isolates from the mountains tested positive for *Salmonella* spp. according to the manufacturer's guidelines. All isolates from both locations were analyzed with the EnteroPluri tubes. Some of the identifications produced two different genera (i.e. *Salmonella* spp. ad *Escherichia coli*), in which case the isolates were evaluated with a more specific test utilizing *Salmonella* antisera. Isolates that were confirmed as *Salmonella* based on the serological test, in conjunction with all other tests performed, including the EnteroPluri and the phenotypic responses to the media (Table 3).

Table 3. EnteroPluri and serological test results. Samples that were positive for both the poly H and O antisera were deemed positive *Salmonella* species. The EnteroPluri tests support these results.

Sample ID	Sample Name	Sample Type	Entero code	Poly H	Poly O
PED_002	CTD.02.TT.XLT4.XLT4.TSI	SOIL	76151	+	+
PED_003	CTD.02.TT.BS.XLT4.TSI	SOIL	66151	+	+
PED_011	CTD.04.TT.XLT4.HE.TSI	SOIL	76100	+	+
PED_019	CTD.03.TT.HE.XLT4.TSI	SOIL	62100	+	+
PED_021	CTD.02.TT.HE.HE.TSI	SOIL	76151	+	+
WNC_001	PFR.01.RV.HE.HE.TSI	SOIL	66151	+	+
WNC_003	LJL.03.TT.BS	SOIL	66151	+	+
WNC_004	LJL.02.TT.HE.HE.TSI	FECAL	66151	+	+
WNC_005	LJL.03.TT.XLT4.XLT4.TSI	SOIL	66151	+	+
WNC_006	PFR.02.TT.XLT4.HE.TSI	SOIL	76151	+	+
WNC_008	PFR.03.RV.XLT.HE.TSI	SOIL	76151	+	+
WNC_012	PFR.02.RV.HE.XLT.TSI	SOIL	76151	+	+
WNC_013	PFR.05.TT.XLT	SOIL	76151	+	+
WNC_014	PFR.01.TT.XLT.HE.TSI	SOIL	76151	+	+
WNC_015	PFR.03.TT.XLT.HE.TSI	SOIL	-	+	+
WNC_016	PFR.01.RV.XLT4.XLT4.TSI	SOIL	-	+	+
WNC_017	PFR.05.TT.HE.HE.TSI	SOIL	66151	+	+
WNC_018	PFR.02.RV.XLT.HE.TSI	SOIL	76151	+	+
WNC_019	PFR.01.TT.BS	UKN	76151	+	+

3.2 Quantification

DNA from 19 bacterial isolates was extracted using the DNA Investigator kit, which is a kit designed for forensic samples, but was modified to allow for bacterial DNA extraction (as described in the methods section). Preliminary data (not shown) supports the modification and that the kit was able to extract bacterial DNA at least as well as the QIamp DNA mini kit. DNA was quantified in duplicate using the Qubit HS dsDNA kits. Throughout the duration of the project, double stranded bacterial DNA concentrations, obtained from environmental *Salmonella* isolates remained higher than the recommended input amount for library preparation using Illumina's Nextera XT library preparation kit (Table 4)

Table 4. DNA quantifications. Qubit HS dsDNA quantifications taken throughout the project. The last measurement was taken in August 2014. The dilution volume is the volume of the extract to be added to TE buffer to obtain 0.2ng of DNA (total volume: 500µl) for library preparation. Each concentration is an average of duplicate readings.

Sample ID	Quantification Date (mm/yy)			
	Sep-13 Concentration [ng/µl]	Mar-14 Concentration [ng/µl]	Aug-14 Concentration [ng/µl]	Dilution Volume [µl]
PED_002	12.15	8.68	9.22	10.85
PED_003	9.23	6.92	7.22	13.85
PED_011	13.30	10.2	10.75	9.3
PED_019	13.40	10.65	11.95	8.37
PED_021	12.00	8.68	9.06	11.04
PED_rb	undetc.	undetc.	undetc.	-
WNC_001	15	11.2	12.95	7.72
WNC_003	13.1	8.7	39.54	10.48
WNC_004	13.3	9.32	10.2	9.8
WNC_005	12.8	8.62	9.3	10.75
WNC_006	14.9	10.5	11.4	8.77
WNC_008	14.75	10.55	11.95	8.37
WNC_012	14.8	10.45	13.5	7.41
WNC_013	17.3	11.2	13.4	7.46
WNC_014	17.4	11.5	14.75	6.78
WNC_015	17.65	10.6	13.95	7.17
WNC_016	16.2	12.35	14.4	6.94
WNC_017	17.45	12.05	14.45	6.92
WNC_018	16.65	10.85	14.5	6.9
WNC_019	14.95	11.5	12.85	7.78
WNC_rb	undetc.	undetc.	undetc.	-
WNC_rb2	undetc.	undetc.	undetc.	-

Pulsed-field gel electrophoresis (PFGE) was conducted using the Standard Operating Procedure (SOP) for PulseNet PFGE of *Salmonella* serotypes; as described in methods section. The procedure was conducted at the FDA's CFSAN lab in College Park, MD. PFGE gels we photographed and analyzed using Bionumerics software. The resulting patterns, normalized to a standard, were searched in PulseNet™. Initially, the patterns from 15 of the 19 isolates were searched for in the PulseNet™ database returning: 10) *S.* Newport, 2) *S.* Bareilly, 2) *S.* Typhimurium. One isolate (WNC_008) was only identified by its antigenic formula: I 4, [5], 12:b:-. Due to poor quality results, four isolates were reanalyzed. Only sample DFP011 produced a different pattern after a second PFGE run.

Sample ID	Sample type	PFGE ID	PULSE-Net Pattern	Serotype
WNC_012	Soil	LJL12	JJPX010119	Newport
WNC_013	Soil	LJL13	JJPX010120	Newport
WNC_014	Soil	LJL14	JJPX010121	Newport
WNC_015	Soil	LJL15	JJPX010122	Newport
WNC_016	Soil	LJL16	JJPX010123	Newport
WNC_017	Soil	LJL17	JJPX010124	Newport
WNC_018	Soil	LJL18	JJPX010125	Newport
WNC_019	Soil	LJL19	JJPX010126	Newport
WNC_006	Soil	LJL6	JJPX010127	Newport
WNC_001	Soil	LJL1	JJPX010128	Newport
PED_002	Soil	DFP2	JAPX01.0975	Bareilly
PED_003	Soil	DFP3	JAPX01.0975	Bareilly
PED_021	Soil	DFP21	JAPX01.0975	Bareilly
WNC_008	Soil	LJL8	I 4, [5], 12:b:-	Paratyphi B
PED_011	Soil	DFP11	JAPX01.0975	Bareilly
PED_019	Soil	DFP19	Unique Pattern	-
WNC_004	Fecal	LJL4	Unconfirmed	Typhimurium
WNC_005	Soil	LJL5	JPXX01.1298	Typhimurium
WNC_003	Soil	LJL3	JPXX01.1298	Typhimurium

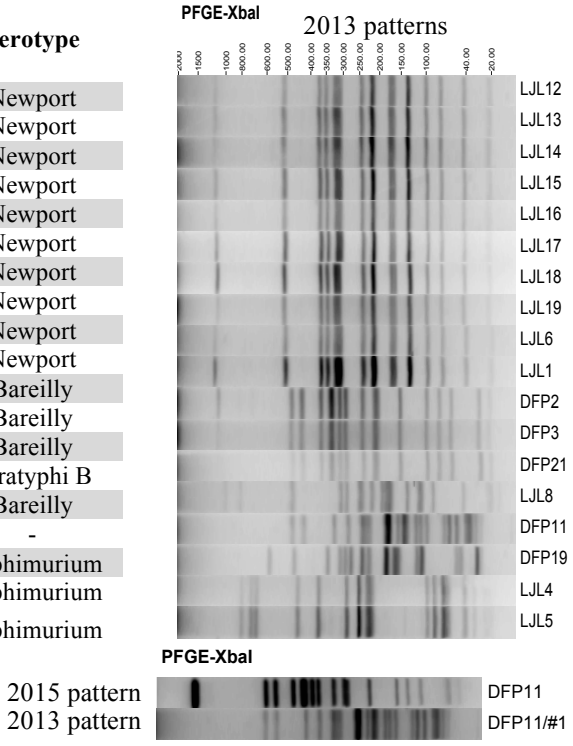


Figure 6. PFGE patterns and associated serotypes. PFGE on 19 *Salmonella* isolates resulted in 18 PFGE patterns. The Pattern for sample WNC_003 was not completely digested by the restriction enzyme Xbal. Re-performance of PFGE on samples WNC_003 (pattern not shown) and PED_011 yielded patterns sufficient for PulseNet search.

3.3 Genome sequencing

Genome sequencing (GS) data from 19 *Salmonella* isolates were obtained in four runs (6-7 genomes/run). One run (RUN A [Table 5]) failed based on the overall lower than expected run quality, outlined in the instructions from the v2 2x151 kit being used, resulting in two genomes being unidentified and present in very low amounts. The nature of the issue was due to possibly over clustering of the flow cell, but attributed to ssDNA being at -20˚C for an extended period during Nextera XT Library Preparation. New libraries were generated for the genomes affected by this run. All genomes in that particular batch were re-sequenced (RUN D [Table 5]). This re-sequencing step produced data that of the expected quality, resulting in an entire representation of all collected and identified S*almonella* species. There were two data sets created, Alpha (**α**)

and Delta (**δ**), for use in the SNP analysis. One data set, Alpha (**α**), (Runs A, B & C) contained sequencing data from run A (Table 5) that was deemed poor quality based on the real-time run statistics and performance expectations. The second data set, Delta (**δ**), (Runs B, C & D) included the same sequencing data, with the exception that only the genomes that were batched in the single run that failed (Run A) were replaced with new sequencing data of the same genomes that were sequenced in a fourth (Run D) MiSeq® run.

Table 5. Environmental *Salmonella* sequencing lists by runs. Each run column contains a list of the isolates that were sequenced. The same isolates were sequenced in run A and D.

Run A 09/10/2014	Run B 09/17/2014	Run C 09/25/2014	Run D 12/09/2014
PED-002	WNC-012	WNC-018	PED-002
PED-003	WNC-013	WNC-019	PED-003
PED-021	WNC-014	WNC-006	PED-021
PED-011	WNC-015	WNC-001	PED-011
PED-019	WNC-016	WNC-008	PED-019
WNC-004	WNC-017	WNC-003	WNC-004
WNC-005	RB (EXTR)	RB (EXTR)	WNC-005
RB (EXTR)			RB (EXTR)

3.4 NCBI *Salmonella* tree

In order to observe any differences in clustering related to the overall quality of the sequencing data both datasets (α and δ) were submitted to the Sequence Read Archive (SRA) at NCBI. Isolates from each dataset were placed on the pathogen tree for *Salmonella*. Only 14 isolates, from the alpha (α) dataset, were placed on the tree during the initial submission. Using NCBI's Genome Workbench isolates were searched using PRJNA260089 to locate their placement on the tree. Figure 7 shows the clustering of 9 isolates with PFGE patterns indicative of *Salmonella enterica* subsp *enterica* serovar Newport among other isolates with similar identifications. The large branch length for isolate PED002 (Figure 8) is attributed to the very

low sequencing coverage of this particular isolate (Table 6). It should be noted that of the five isolates not placed on the tree, two are missing due to the absence of sequencing product (PED003 & WNC005) in the original sequencing run, two are missing (PED019 & PED011) due to being mixtures and the final isolate (WNC006) is missing for technical reasons; stemming from identical sample ID's created prior to sequencing. Isolate WNC006 has since been placed on the tree and clusters as expected.

Sequencing data from isolates arising from run D (within the δ-dataset) replaced the isolates previously placed on the tree (run A). The δ-dataset contained high-quality sequencing reads. Isolates from this dataset were placed on the tree resulting in clustering of all 19 isolates collected from the four sampling sites in North Carolina. These isolates that are currently on the tree are clustering together and represent the four different serotype associations seen in Figure 7.

Figure 7. NCBI pathogen tree of *Salmonella.* This figure shows the placement of the Newport serotype isolates collected in Western North Carolina (NC).

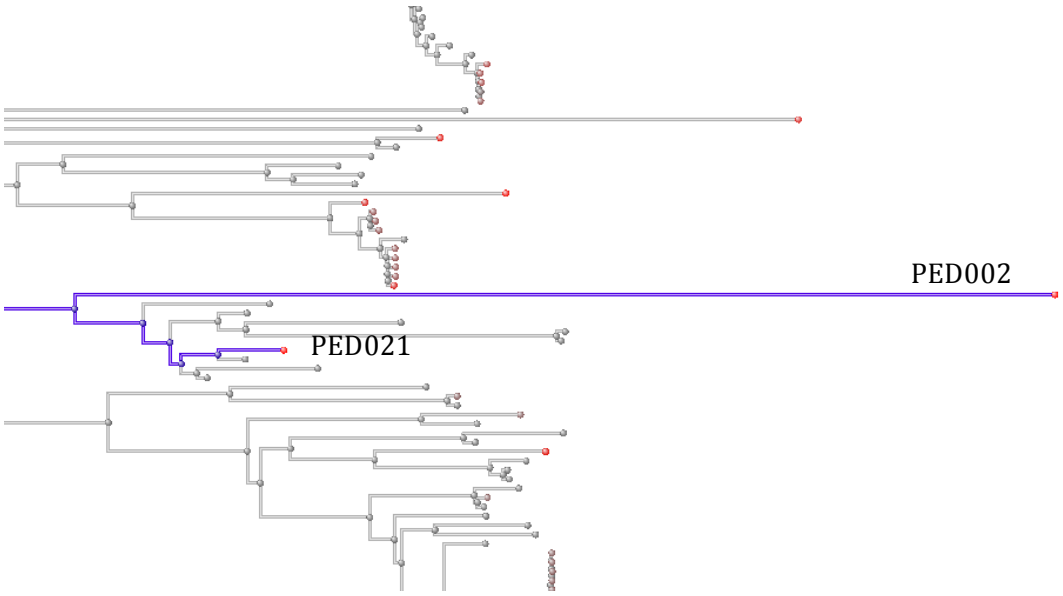


Figure 8. Large distance due to low quality. This figure illustrates the separation of two isolates sharing similar PFGE pattern due solely to low sequencing quality.

3.5 Independent SNP analysis

The Wadsworth Center (http://www.wadsworth.org/) in Albany, New York performed SNP analysis on the 19 *Salmonella* isolates sequenced on the Illumina® MiSeq®. Two datasets were generated that contained i) a run with low quality reads (**α**) and ii) a separate sequencing run of the same isolates in which the sequencing results were of higher quality (**δ**).

To observe whether or not sequencing quality has an effect on the SNP analysis results, two phylogenetic trees were produced. Figure 9 shows a SNP tree that represents the **α** data set (Runs A, B & C [Table 5]). The tree produced shows many genomes clustering in the tree as expected according to the PFGE results. Figure 6 shows the associated serovars based on PFGE patterns and their similarity with patterns in the PulseNet™ database. Based on the identified serovars, the SNP trees reflect the expected clustering of the corresponding genomes. Included in both trees, in addition to the *Salmonella* genomes collected from the environment, are genome references of the various serovars that were identified via PFGE, as well as closely related genomes.

Figure 9 shows two outlying genomes, PED-011 & PED-019. Upon further analysis, using Kraken[65]; a fast and accurate program for assigning taxonomic labels for metagenomic DNA sequences[65], these samples appeared not to be pure samples, but were a mixture of many bacterial species. PFGE patterns identified PED-002 as potentially a *S.* Bareilly serovar, however the NY-SNP tree suggested that PED-002 was more closely related to a *S.* typhi serovar. The average coverage for all isolates in Run A (Table 6) with the exception of PED-021 was below10x. Isolates PED-003 & WNC-005 are not included in this tree because the indices for these two libraries were not found, an indication that very little sequencing product was present.

Figure 10 is the second NY-SNP tree constructed with the **δ**-data set (Runs B, C, & D [Table 5]). In this tree, isolates PED-011 & PED-019 were not included. Analysis of these isolates, using Kraken, revealed the same mixed results; therefore they were excluded from the SNP analysis. The tree in Figure 7 has grouped the PED-002, PED-003, and PED-021 samples together. These represent one of the sampling sites in the Piedmont of North Carolina. The identity of these isolates was believed to be *S.* Bareilly, but the tree shows a placement more closely related to *S*. Newport serovar. The PFGE patterns were not definitive, but the banding pattern for these isolates, as seen in Figure 4, appear to be identical, which is depicted in the SNP tree. In this data set, isolate WNC-005 was successfully sequenced and grouped with WNC-003 and WNC-004. These environmental isolates were believed to be possibly a *S.* typhimurium serovar (according to PFGE patterns), which is what the SNP tree indicates. The node that groups WNC-003, 4 & 5 together is common in all samples that were collected from Lake Junaluska, in the Western North Carolina mountains, that contained *Salmonella* spp. these isolates were obtained from soil (WNC-003 & WNC-005) and fecal matter (WNC-004) collected within approximately a $4ft^2$ area.

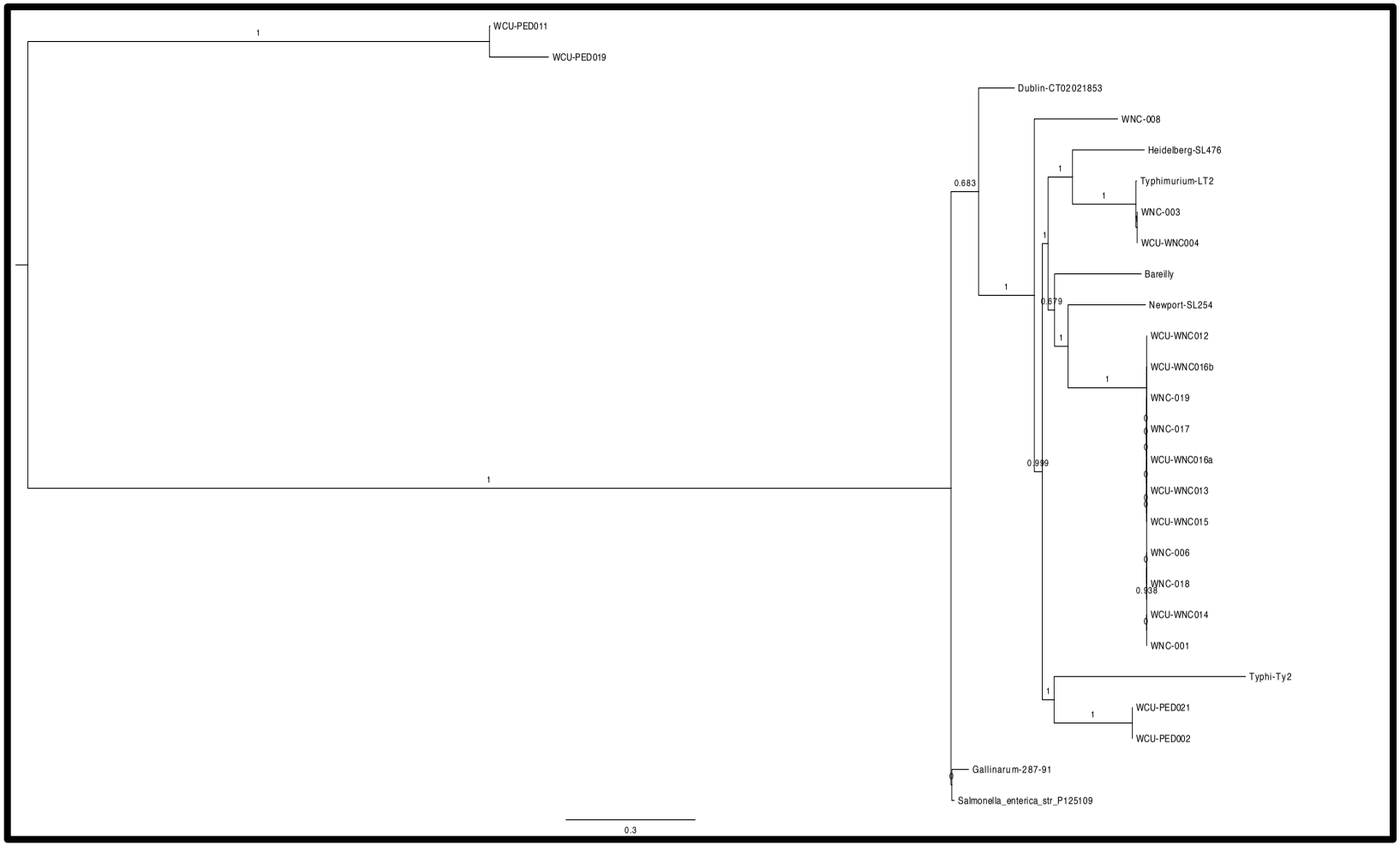


Figure 9. Alpha (α) data SNP tree. These results were obtained from runs A, B, & C. and isolates are clustering together as expected with their associated serovars that were included in the tree construction.

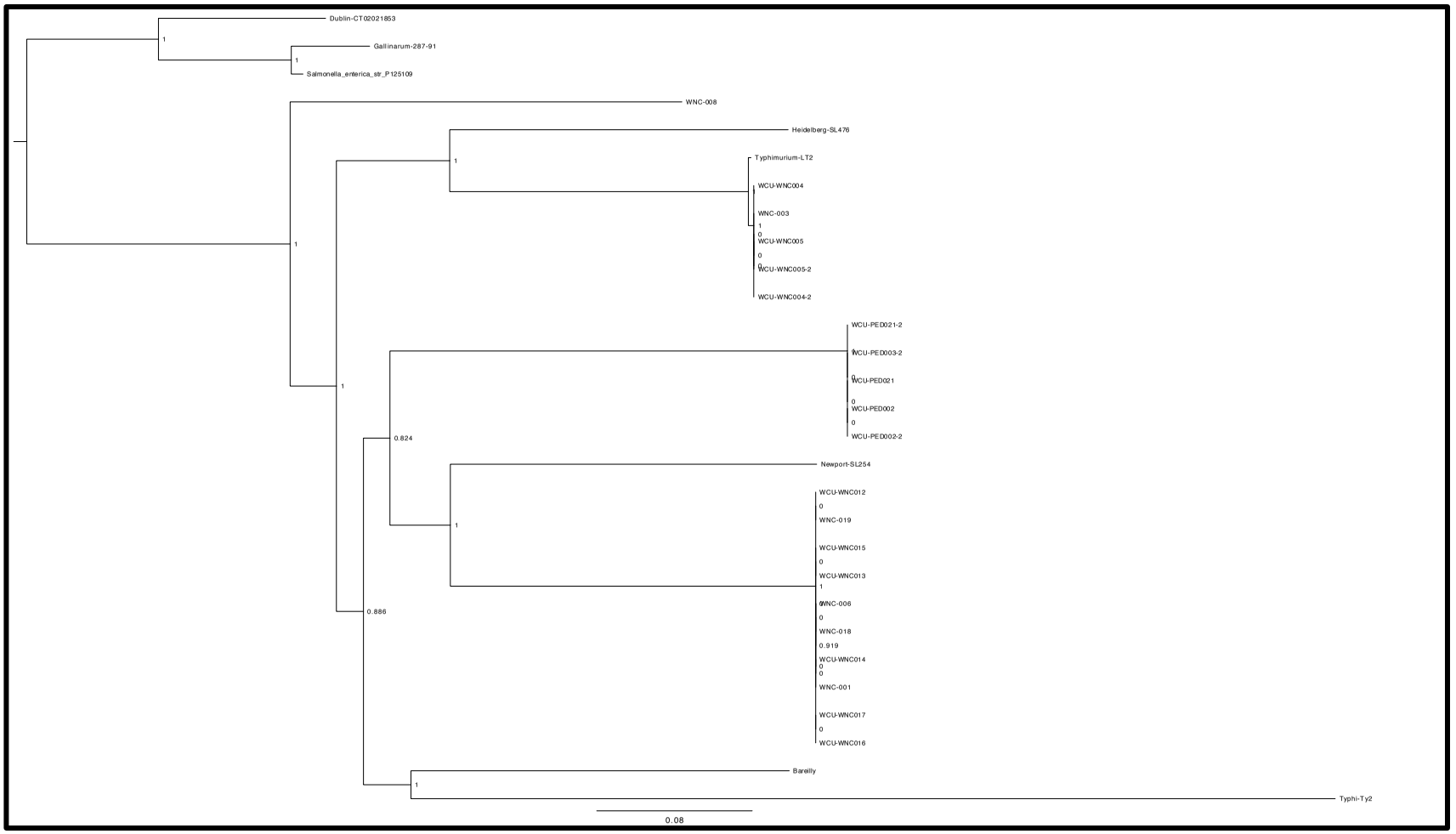


Figure 10. Delta (δ) data SNP tree. These results were obtained from Runs B, C, & D. This high quality data included the isolates PED003 and WNC005 where are clustering as expected. The -2 at the end of the isolate name signifies the isolate that were used to build this tree. The sequencing data used to build the tree in Figure 5, were included in this tree to observe and change in topology between datasets

Table 6. NY SNP tree run statistics. Shown are the percentage of reads that mapped to the reference (*S. Enteritidis* strain P125109), the average depths of coverage of the mapped reads, and the percent of the genome covered.

Sample	% Reads Mapped	% Correctly paired	% Coverage	Avg. Depth
Bareilly	94.02%	80.65%	94.18%	70.3
Dublin-CT02021853	95.56%	83.70%	98.22%	73.29
Gallinarum-287-91	98.06%	86.07%	97.04%	72.26
Heidelberg-SL476	90.33%	77.26%	93.78%	70
Newport-SL254	91.94%	78.75%	94.09%	70.29
Typhi-Ty2	89.18%	73.46%	90.38%	67.58
Typhimurium-LT2	92.08%	78.61%	94.63%	70.78
WCU-PED002-2	90.37%	89.61%	94.25%	51.02
WCU-PED002-1	83.30%	76.77%	92.11%	5.66
WCU-PED003-2	89.42%	88.03%	94.30%	118.51
WCU-PED021-2	89.99%	88.70%	94.26%	57.12
WCU-PED021-1	81.45%	75.92%	94.24%	26.68
WCU-WNC004-2	89.46%	88.02%	95.27%	124.47
WCU-WNC004-1	79.61%	72.65%	95.18%	20.7
WCU-WNC005-2	89.90%	88.55%	95.22%	38.86
WCU-PED011*	17.82%	11.48%	39.50%	4.16
WCU-PED019*	20.78%	16.76%	40.52%	4.78
WCU-WNC012	89.84%	89.07%	95.40%	31.12
WCU-WNC013	88.18%	86.78%	95.46%	106.38
WCU-WNC014	87.83%	86.72%	95.47%	145.24
WCU-WNC015	88.08%	86.82%	95.44%	95.82
WCU-WNC016	88.72%	87.34%	95.45%	126.55
WCU-WNC017	88.29%	87.00%	95.44%	101.36
WNC-001	89.41%	88.57%	95.39%	28.62
WNC-003	89.92%	88.24%	95.26%	56.58
WNC-006	89.39%	87.98%	95.45%	117.18
WNC-008	89.77%	88.29%	95.56%	59.35
WNC-018	88.42%	86.72%	95.44%	93.97
WNC-019	88.91%	88.16%	95.42%	43.95

3.6 kSNP® trees

Trees produced using the α-dataset (Figures 11 & 12) show clustering of environmental isolates using core SNPs (SNPs that are shared among all genomes being analyzed). This tree contains 17 of the 19 isolates sequenced within the α-dataset. Isolates PED003 & WNC005 did not generate any sequencing reads during Run 1 (see Table 6. A minimum k-mer count (MKC) of 10 (default) was applied; clustering isolates by sampling site. The isolates also clustered consistently with their PFGE patterns (Figure 6). PED011 & PED019 in both trees (Figures 11 & 12) were mixed bacterial cultures and the branch lengths for these isolates are proportional to the number of SNP differences. The MKC was increased to 20, as shown in Figure 8, to reduce the number of SNPs being called for analysis due to sequencing error. This reduced the number of core SNPs from 1197 to 8. All 8 SNP differences occurred within the PED011 & PED019 isolates for this particular tree, suggesting that the remaining isolates are similar.

Trees produced using the δ-dataset (Figures 13 & 14) also show the clustering of the environmental isolates using core SNP. These trees were products of high quality sequencing data using a MKC of 20. Figure 13 shows the clustering of all 19 of the environmental isolates collected. A total of 10,901 SNPs were identified as core SNPs generating four distinct clades with fully resolved clusters. Figure 14 shows the clustering of isolates along with four complete genomes from Genbank. The assembled genomes were included to show the relationship that the environmental isolates have with each other and a common relative; based on PFGE serovar associations. The addition of the assembled genomes reduced the number of core SNPs (726) that were used to generate the tree.

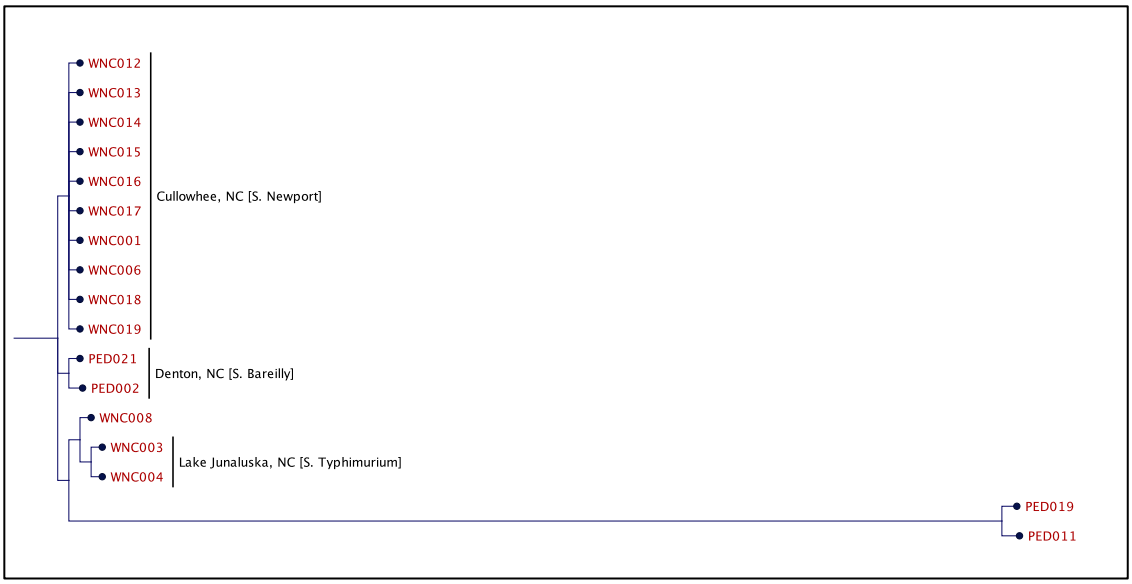


Figure 11. α-kSNP tree 1. Phylogenetic tree created from the α-dataset using core SNP loci. Minimum k-mer count (MKC) of 10 (default) was used. kSNP identified 1171 core SNPs to build the tree with 5 distinct and fully resolved clusters. Isolates cluster as expected from PFGE results (Figure 6) and sub-trees are labeled based on sampling site. WNC008 was collected from the Cullowhee, NC site and was identified as being a different serovar of *Salmonella* from isolates collected at Lake Junaluska and Cullowhee sites.

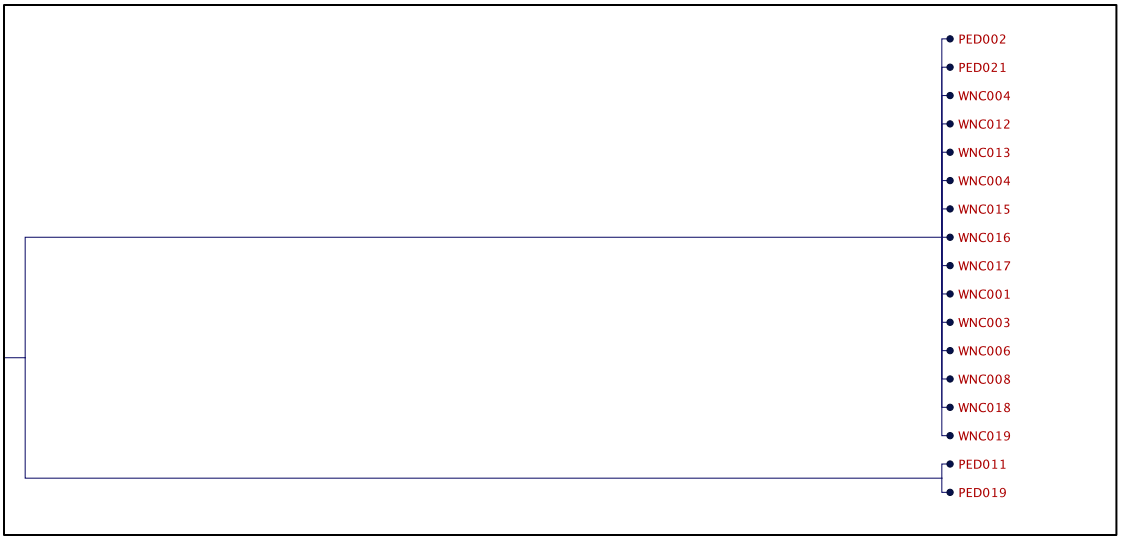


Figure 12. α-kSNP tree 2. Phylogenetic tree produced with kSNP from α-dataset using only core SNP loci. MKC of 20 was used. kSNP identified only 8 core SNPs that define the 2 sub-trees. All SNP differences were identified to be within the isolates PED011 & PED019. MKC of 20 decreased the total number of core loci that are attributed to the low quality genomes present in the α-dataset.

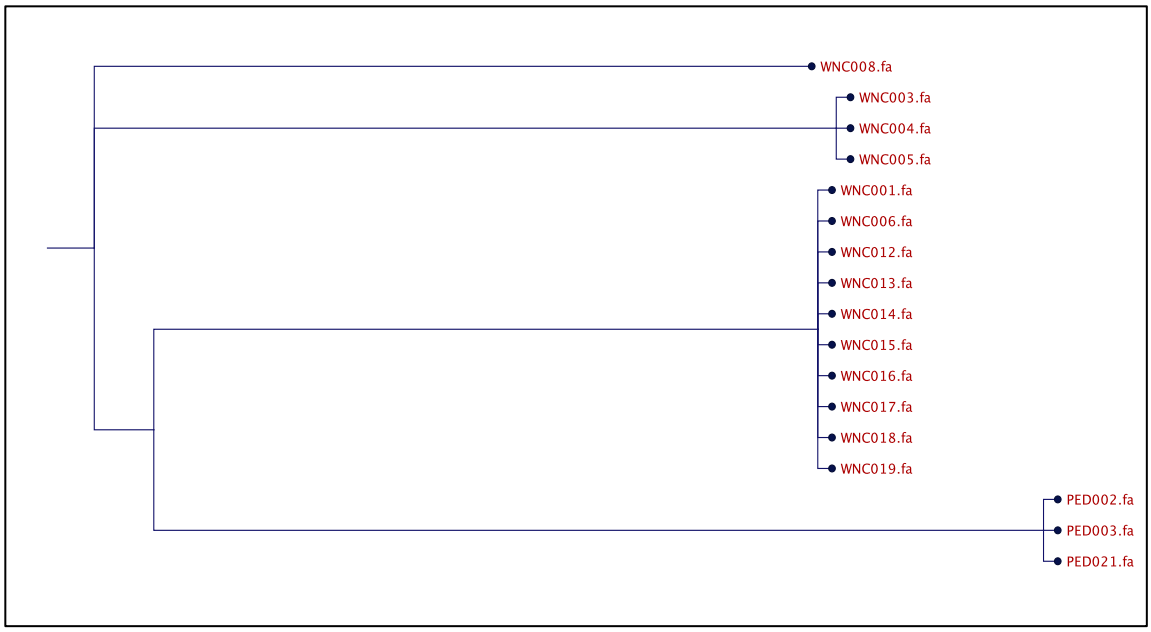


Figure 13. δ-kSNP tree. kSNP phylogenetic tree estimated using the δ-dataset. This tree used a MKC of 20. Isolates PED011 & PED019 were removed from this analysis because of being identified as mixed bacterial species. kSNP identified 10,901 SNPs as being core. kSNP was able to fully resolve isolates within δ-dataset and all isolates clustered as expected (according to PFGE and independent SNP analysis)- Sequence reads in the δ-dataset was representative of all 19 *Salmonella* isolates collected.

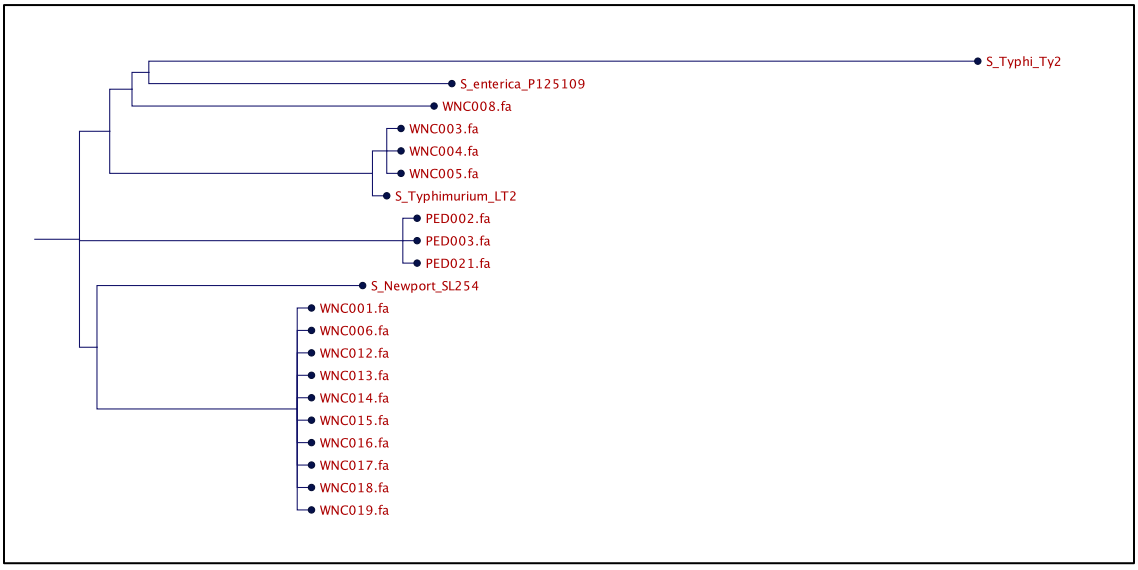


Figure 14. δ-kSNP tree assembled. Phylogenetic tree produced using δ-dataset with the addition of 4 complete genomes downloaded from Genbank that were used in the independent SNP analysis. kSNP clustered the environmental isolates as expected, and the overall formation is consistent with the tree in Figure 7.

CHAPTER FOUR: DISCUSSION

During outbreak events or investigations, the speed and accuracy of analysis is crucial. Some analysis tools utilized to analyze GS data of foodborne pathogens provide information quickly, but the accuracy of these analyses have been shown to be influenced by the quality and quantity of the sequencing data generated. Algorithms that create distance trees, like the reference tree of *Salmonella* used within the Genome Trackr® network; are affected by the total amount of sequence information provided during sequencing. Although the quality and amount of sequencing data is a limitation, these trees provide investigators with fast and reliable information regarding the relationships of the samples in question, which is important during multistate outbreak investigations.

Distance-based metrics rely on the overall distances between isolates in a pairwise manner to cluster isolates together. Distance measurements, that utilize short subsequences [k-mers], are created by counting the total number of k-mers that are the same between each isolate. The matrix, used to construct the tree, is calculated from the total sequence data available from each isolate under comparison[66].The more similar any two are to one another, the smaller the calculated distance between them, and thus the closer they will cluster. However, clustering of isolates on a distance tree can be influenced by any parameter that can change the distance metric (e.g., total amount of sequencing data that passes filter). Subtle differences in the total sequencing output that is less than the total size of the genome would create only a partial genome comparison. When comparing a complete genome to a partial genome, even if originating from the same isolate, slight differences were expected in the clustering results. This is due to the different number of k-mers being used for comparison in the partial genome and therefore it is seen as a genetic difference using this method.

In a recent mock outbreak exercise, conducted between the FDA's Center for Food Safety and Applied Nutrition (CFSAN) and other Public Health lab partners, 112 clinical and environmental isolates from 11 previous outbreaks were prepared and divided among seven state and federal labs for a single-blind analysis using whole genome sequencing. The outcome of this exercise showed that the Genome Trackr® system was performing as expected overall, however there were some unexpected anomalies that were seen within the clustering.

The investigators noticed that these outbreak isolates, although epidemiologically linked and obtained from earlier outbreaks, were clustering differently before and during the exercise. The results also showed that one isolate, that was inadvertently sequenced twice six months apart, showed a slight change in position of clustering on the tree; possibly attributed to the total amount of genome sequence obtained which could be related to the overall quality.

Similar anomalies were identified with the sequencing data used in this project. One isolate in particular, PED002, clustered away from isolate PED021 with a long branch length even though the two isolates shared very similar PFGE patterns. It was believed to be associated with low quality sequencing data in the original run. This was confirmed upon submission of the second dataset containing higher quality sequencing data, resulting in these two isolates, along with PED003, clustering closer together with other isolates of similar serotype as expected from associated PFGE patterns and independent SNP analysis.

4.1 SNP analysis

Next generation sequencing data analysis can be time consuming, and the time sensitive nature of forensic investigations is relevant to the successful resolution of a case. Also, the rapid analysis of any information has its trade offs; fast analysis tends to trade off some accuracy. The enormous amount of sequencing data that can be generated with higher-throughput sequencers

has led to the need for more bioinformatic programs and software support in to effectively analyze the large datasets. It stands to reason that the advancements in these programs had to be equally paced with the rapidly advancing sequencing technologies. Better bioinformatics support can also help research laboratories remain productive whilst the computer analysis is being conducted. Many analyses (e.g., SNP analysis) are created to perform a specific task and contain, usually in a web server, a pipeline of analysis programs. Most online servers are tailored for researchers without much experience in bioinformatics to allow for a more hands-off approach. These servers, however, offer the user limited control over how the data will be processed. Normally the sequencing data is uploaded, a few adjustable parameter-values are selected, and the data is analyzed. This allows the researcher to continue conducting experiments and allows the computer to do all the work. Nevertheless, those that are more bioinformatically inclined may prefer analysis in which the user has more control over how the sequencing data will be treated. Desktop software programs allow the user to optimize the analysis pipeline in order to generate high quality data specific to the research being conducted.

snpTree® allows for raw sequence data to be uploaded to the web-based server and then be analyzed. This toolbox, which contains all of the programs necessary for SNP analysis, was able to infer a phylogeny of the *Salmonella* isolates collected. The results from this program were compared to previous PFGE analysis and an independent SNP analysis done by the New York Department of Health.

Using SnpTree, the results obtained were inconsistent and unable to produce trees containing all of the isolates contained within the two datasets (Alpha and Delta). For both datasets, various trees returned missing isolates. Several attempts were made to produce SNP trees that contained all of the isolates within a particular dataset (α or δ), but the resulting trees

were inconsistently missing isolates. The sequencing reads for some of the isolates that were omitted were of poor quality, which is explained by the fact that the server filters and trims the raw sequencing data prior to analysis. What is not explainable was the fact that in other cases, the trees produced were missing an isolate that was a product of a high quality-sequencing run. Apparently the manner in which the files are uploaded contributed to the sporadic omission of some of the data. Prior to the sequencing files being uploaded, they were identified as paired-end (one file for forward reads, and one file for reverse reads). If the server encountered an issue with the uploading of an unpaired file, it would omit that particular isolate. What is interesting is the datasets that were being uploaded contained the same files and were never modified, yet the trees returned by snpTree did not consistently omit the same isolate in each of the attempts. Further investigation is warranted in this area.

Analysis time for snpTree was on average approximately 16-20 hours and snpTree offered a similar adjustable parameter as kSNP in that the minimum coverage for a SNP call could be adjusted. In snpTree, raw reads were filtered and trimmed (see Methods) before analysis, and were mapped to a reference. This could account for the extended analysis time. The same reference used for the independent analysis was also used in snpTree. The original phylogenetic estimation generated (excluding the fact that on occasions there was missing data), were not clustering the isolates as expected in either the α- and δ-datasets. Due to the factors mentioned above, no comparisons were made between kSNP and snpTree outputs.

Analysis using kSNP was done locally using a Linux server housed at Western Carolina University. Its ability to infer the independent SNP analysis and the clustering of isolates also determined phylogeny identified from PFGE patterns. Sequencing reads, filtered and trimmed (see Methods) were analyzed within 2 hours; a fraction of the time needed for snpTree. The

phylogenetic trees produced consistently clustered isolates as expected according to PFGE results. Quantity and quality of sequencing data were shown to have a minor affect on the clustering of isolates of similar PFGE patterns when the two datasets were compared.

4.2 kSNP trees

When a default minimum k-mer count (MKC) of 10 was applied to these samples, it did result in the expected clustering of the isolates within the dataset. However, when a more stringent MKC was applied, the number of core SNPs that were used to build the α-kSNP trees dropped from 1,171 to 8. This was attributed to the extremely low coverage across most of the genomes, and the fact that PED011 and PED019 were mixtures. Independent analyses of the α-dataset showed that PED002, 11, & 19 had an average depth of approximately 5.66, 4.16 & 4.78 respectively. In both α-kSNP trees, large branch lengths separated the nodes due to a large number of SNP differences (relative to the total number of core SNPs). The total number of SNPs identified in the α-dataset using the kSNPs default MKC value of 10x was 1,195,794. The total number of SNPs identified with kSNP MKC of 20x was only 36,588. As expected, increasing the MKC decreased the total number of SNPs and core SNPs identified due to the lower quality and average coverage. The MKC parameter is used in unassembled raw reads to reduce the number of SNPs called that maybe due to a potential sequencing error. Default MKC of 10 is used with the assumption of 100x coverage and using a lower MKC of 5 is recommended when dealing with low coverage (i.e. <25x). Independent analysis showed that the sequencing data from Runs B & C (Table 5) had an average coverage that ranged from 28x-145x; 7 of the 12 isolates had and average depth of >80x. The increase of MKC between the α-trees illustrated the idea that very poor data can influence phylogenetic trees when using a pipeline intended for high quality sequence data. kSNP was still able to produce a phylogenetic

tree using a lower MKC, in which more SNPs are included in the analysis. The sequencing reads in run A (Table 5) would generally not be acceptable or used in any data analysis having such low coverage and quality. This run was included as part of the α-dataset to evaluate whether or not sequencing quality/quantity has an impact on phylogenetic tree estimation using SNP analysis, as it does on distance-based trees.

By comparison, the δ-dataset was created to represent 'High-Quality' data. The only portion of this dataset that changed was the NGS data from Run D (Table 5) created in part to evaluate the affects of quality/quantity of unassembled–raw reads on distance based and SNP based phylogenetic tree estimations. The δ-dataset was used to evaluate a computer-based package designed for SNP analysis without the use of a reference genome. Prior knowledge to the identity of the environmental isolates came from PFGE patterns and independent analysis performed by the New York Department of Health. The branch lengths within the different clades suggest that the isolate within that sub-tree are very similar; based on the core SNPS used to produce the tree. The inclusion of the 4 completed genomes from Genbank into the δ-dataset was in an attempt to tease out any other SNP loci that might not have been included in the original tree (Figure 13) and highlight any differences among the environmental isolates within a particular region. The addition of the assembled genome reduced the number of core SNPs (726) that were used to generate the tree.

Using raw reads without the need for a reference would reduce the potential of bias by only considering three variables: the programs that perform the analysis, and the quality and quantity of the data being analyzed. Programs requiring a reference would introduce more complex issues such as the choice of a reference genome that could affect the results. This would require some prior knowledge of the isolate, such as the PFGE pattern. PFGE is an important

tool, but as mentioned above, it poses some considerable resolution issues[17,66,67]. Also, in addition to its limited resolving power; PFGE does take a considerable amount of time to generate results. The process from sample collection to DNA extraction is 7 calendar days and PFGE analysis would increase that timeline by an additional 5 days.

The questions being considered with the δ-dataset are: i) can this SNP analysis program be equally as accurate in its ability to identify SNPs and cluster isolates without a reference, ii) can this tool be wielded by someone with minimal knowledge in bioinformatics and iii) during a foodborne outbreak could use of high quality whole-genome sequence data in conjunction with available analysis tools, like the kSNP, circumvent the use of PFGE as being a pivotal identification step? If the answer to any of these questions is yes, then this would limit the use of PFGE to a supplemental tool, placing emphasis rather on the phylogenetic analysis during active investigations or prosecutions arising from the deliberate use of pathogens as weapons.

In summary, this research has shown to be able to isolate and identify environmental *Salmonella* using the optimized methodologies currently being utilized by the FDA. New techniques are being investigated to potentially reduce the number of samples being processed during an outbreak. The FDA is currently developing a real-time PCR assay for *invA* (a subunit of a protein that makes up the type three secretion system (TTSS) in S*almonella* species) that can pre-screen samples after the enrichment step reducing the workload on investigators.

Although this research was not designed around the time optimization of the entire process from sample collection to readable data, each aspect of the workflow was investigated. The implementation of commercially available kits was used to identify bacterial colonies as S*almonella*. Automation during DNA extraction greatly reduced the hands on time that would have been required, if performed manually.

Most importantly, high-throughput MPS has shown to be vital tool in the workflow. When dealing with a multistate outbreak the ability to accurately identify the infectious agent can be crucial. NGS has demonstrated its ability to sequence multiple *Salmonella* genomes per run. Genome sequencing is not a new concept, rather an evolving one. The advancement in NGS technology has improved the genetic resolution as well as provided the ability to look deeper into the genomes to identify unique differences.

High quality sequencing data generated by the MiSeq® was uploaded to NCBI where phylogenetic trees were created; representing the relationship or similarities between the environmental *Salmonella* populations in North Carolina. The way in which these phylogenetic trees are created was shown to be influenced by the quality of the sequencing data. The distance method is a fast and relatively accurate technique to provide a quick explanation of what is being observed. However, due to these influences in the outcome of the data generated, other methods that take on a different approach, utilizing SNPs, was investigated.

SNP analysis traditionally involved the use of a reference genome becoming computationally extensive rather quickly. The approach taken here utilized a reference free method. Phylogenic trees were created utilizing the kSNP program and were less influenced by the quality of the sequencing data given.

As databases grow in size containing high quality sequencing information coupled with phylogenetic tree analysis of foodborne pathogens MPS may, in the near future, become the new gold standard for epidemiological investigations.

WORKS CITED

1. Library C. Salmonella Fast Facts - CNN.com. 2015.
2. Porwollik S, Boyd EF, Choy C, Cheng P, Florea L, Proctor E, McClelland M. Characterization of Salmonella enterica Subspecies I Genovars by Use of Microarrays. Journal of Bacteriology 2004;186(17):5883-5898.
3. Prevention CfDCa. Diagnosis and Treatment - Salmonella. <http://www.cdc.gov/salmonella/general/diagnosis.html%3E%3Cfiles/267/CDC - Diagnosis and Treatment - Salmonella.html>.
4. Prevention CfDCa. Salmonella. <http://www.cdc.gov/salmonella/%3E.
5. Jacobsen CS, Bech TB. Soil survival of Salmonella and transfer to freshwater and fresh produce. Food Research International 2012;45(2):557-566.
6. Hubálek Z. An annotated checklist of pathogenic microorganisms associated with migratory birds. Journal of Wildlife Diseases 2004;40(4):639-659.
7. Baudart J, Lemarchand K, Brisabois A, Lebaron P. Diversity of Salmonella Strains Isolated from the Aquatic Environment as Determined by Serotyping and Amplification of the Ribosomal DNA Spacer Regions. Applied and Environmental Microbiology 2000;66(4):1544-1552.
8. Sibley CD, Peirano G, Church DL. Molecular methods for pathogen and microbial community detection and characterization: Current and potential application in diagnostic microbiology. Infection, Genetics and Evolution 2012;12(3):505-521.
9. Islam M, Morgan J, Doyle MP, Phatak SC, Millner P, Jiang X. Fate of Salmonella enterica Serovar Typhimurium on Carrots and Radishes Grown in Fields Treated with Contaminated Manure Composts or Irrigation Water. Applied and Environmental Microbiology 2004;70(4):2497-2502.
10. Barak JD, Liang AS. Role of Soil, Crop Debris, and a Plant Pathogen in Salmonella enterica Contamination of Tomato Plants. PLoS ONE 2008;3(2).
11. US EPA O. Introduction. 2014.
12. Gerba CP, Smith JE. Sources of pathogenic microorganisms and their fate during land application of wastes. Journal of Environmental Quality 2005;34(1):42-48.
13. Winfield MD, Groisman EA. Role of Nonhost Environments in the Lifestyles of Salmonella and Escherichia coli. Applied and Environmental Microbiology 2003;69(7):3687-3694.
14. Zheng J, Allard S, Reynolds S, Millner P, Arce G, Blodgett RJ, Brown EW. Colonization and Internalization of Salmonella enterica in Tomato Plants. Applied and Environmental Microbiology 2013;79(8):2494-2502.
15. Hintz LD, Boyer RR, Ponder MA, Williams RC, Rideout SL. Recovery of Salmonella enterica Newport Introduced through Irrigation Water from Tomato (Lycopersicum esculentum) Fruit, Roots, Stems, and Leaves. HortScience 2010;45(4):675-678.
16. Greene SK, Daly ER, Talbot EA, Demma LJ, Holzbauer S, Patel NJ, Hill TA, Walderhaug MO, Hoekstra RM, Lynch MF and others. Recurrent multistate outbreak of Salmonella Newport associated with tomatoes from contaminated fields, 2005. Epidemiology and Infection 2008;136(2):157-165.
17. Allard MW, Luo Y, Strain E, Li C, Keys CE, Son I, Stones R, Musser SM, Brown EW. High resolution clustering of Salmonella enterica serovar Montevideo strains using a next-generation sequencing approach. BMC Genomics 2012;13(1).
18. Maki DG. Coming to Grips with Foodborne Infection — Peanut Butter, Peppers, and Nationwide Salmonella Outbreaks. New England Journal of Medicine 2009;360(10):949-953.

19. Cao G, Meng J, Strain E, Stones R, Pettengill J, Zhao S, McDermott P, Brown E, Allard M. Phylogenetics and differentiation of Salmonella Newport lineages by whole genome sequencing. PloS one 2013;8(2).
20. Lienau EK, Strain E, Wang C, Zheng J, Ottesen AR, Keys CE, Hammack TS, Musser SM, Brown EW, Allard MW and others. Identification of a salmonellosis outbreak by means of molecular sequencing. The New England Journal Of Medicine 2011;364(10):981-982.
21. Parkhill J, Wren BW. Bacterial epidemiology and biology - lessons from genome sequencing. Genome Biology 2011;12(10).
22. Genetic Testing. Food Safety News, Food Safety News: Bill Marler.
23. Beadle J, Wright M, McNeely L, Bennett JW. Electrophoretic karyotype analysis in fungi. Advances in applied microbiology 2003;53:243-270.
24. Goering RV. Pulsed field gel electrophoresis: A review of application and interpretation in the molecular epidemiology of infectious disease. Infection, Genetics and Evolution 2010;10(7):866-875.
25. Olive DM, Bean P. Principles and Applications of Methods for DNA-Based Typing of Microbial Organisms. Journal of Clinical Microbiology 1999;37(6):1661-1669.
26. Prevention CfDCa. 2013 7/1/2013. Pulsed-field Gel Electrophoresis (PFGE). <http://www.cdc.gov/pulsenet/pathogens/pfge.html%3E. 7/1/2013.
27. Schuster SC. Next-generation sequencing transforms today's biology. Nature Methods 2007;5(1):16-18.
28. Reis-Filho JS. Next-generation sequencing. Breast Cancer Research 2009;11(Suppl 3).
29. Nutrition CfFSaA. 2015 Whole Genome Sequencing (WGS) Program. Center for Food Safety and Applied Nutrition <http://www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS/ - Investigators>.
30. Nutrition CfFSaA. 2015 Whole Genome Sequencing (WGS) Program - GenomeTrakr Fast Facts. Center for Food Safety and Applied Nutrition <http://www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS/ucm403550.htm%3E.
31. Zimbro MJ, Power DA. Difco & BBL manual: manual of microbiological culture media. Sparks, MD: Becton Dickinson and Co.; 2009.
32. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR and others. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;456(7218):53-59.
33. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 2012;13(1).
34. Steven R. Head HKK, Sarah A. LaMere, Thomas Whisenant, Filip Van Nieuwerburgh, Daniel R. Salomon, and Phillip Ordoukhanian. BioTechniques - Library construction for next-generation sequencing: Overviews and challenges. BioTechniques 2015;56(2):16.
35. Syed F, Grunenwald H, Caruccio N. Next-generation sequencing library preparation: simultaneous fragmentation and tagging using in vitro transposition. Nature Methods 2009;6(11).
36. Illumina. Nextera® XT DNA Sample Preparation Guide.
37. Mullan L. Pairwise sequence alignment—it's all about us! 2006.
38. Batzoglou S. The many faces of sequence alignment. 2005.
39. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nature Reviews Genetics 2012;13(5):303-314.
40. Chan CX, Ragan MA. Next-generation phylogenomics. Biology Direct 2013;8(1).
41. Allard MW, Luo Y, Strain E, Pettengill J, Timme R, Wang C, Li C, Keys CE, Zheng J, Stones R and others. On the Evolutionary History, Population Genetics and Diversity among Isolates of Salmonella Enteritidis PFGE Pattern JEGX01.0004. PLoS ONE 2013;8(1).

42. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, Segre JA. Tracking a Hospital Outbreak of Carbapenem-Resistant Klebsiella pneumoniae with Whole-Genome Sequencing. Science Translational Medicine 2012;4(148):148ra116-148ra116.
43. Blaisdell BE. Average values of a dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a computer-generated model system. J Mol Evol 1989;29(6):538-47.
44. Beres SB, Carroll RK, Shea PR, Sitkiewicz I, Martinez-Gutierrez JC, Low DE, McGeer A, Willey BM, Green K, Tyrrell GJ and others. Molecular complexity of successive bacterial epidemics deconvoluted by comparative pathogenomics. Proc Natl Acad Sci U S A. Volume 107. United States2010. p 4371-6.
45. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill F-X, Goodhead I, Rance R, Baker S, Maskell DJ, Wain J and others. High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. Nature Genetics 2008;40(8):987-993.
46. Leekitcharoenphon P, Kaas RS, Thomsen MCF, Friis C, Rasmussen S, Aarestrup FM. snpTree - a web-server to identify and construct SNP trees from whole genome sequence data. BMC Genomics 2012;13(Suppl 7).
47. Leekitcharoenphon P, Nielsen EM, Kaas RS, Lund O, Aarestrup FM. Evaluation of Whole Genome Sequencing for Outbreak Detection of Salmonella enterica. PLoS ONE 2014;9(2).
48. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. Volume 25. England2009. p 1754-60.
49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25(16):2078-2079.
50. Gardner SN, Hall BG. When Whole-Genome Alignments Just Won't Work: kSNP v2 Software for Alignment-Free SNP Discovery and Phylogenetics of Hundreds of Microbial Genomes. PLoS ONE 2013;8(12).
51. Gardner SN, Slezak T. Scalable SNP Analyses of 100+ Bacterial or Viral Genomes. Journal of Forensic Research 2010;01(03).
52. Andrews WH, Jacobson A, Hammack T. US Food & Drug Administration.
53. EnteroPluri-Test. Lioflichem.
54. Salmonella Antisera. Pro-Lab Diagnostics.
55. EZ1® Advanced XL User Manual. Qiagen; 2009.
56. Qubit® 2.0 Fluorometer User Manual. Life Technologies; 2010.
57. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. Comparison of Next-Generation Sequencing Systems. Journal of Biomedicine and Biotechnology 2012;2012:1-11.
58. Metzker ML. Sequencing technologies — the next generation. Nature Reviews Genetics 2010;11(1):31-46.
59. Shendure J, Ji H. Next-generation DNA sequencing. Nature Biotechnology 2008;26(10):1135-1145.
60. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet 2010;11(1):31-46.
61. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics 2011;12(6):443-451.
62. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE 2010;5(3).
63. Blankenberg D, Gordon A, Kuster GV, Coraor N, Taylor J, Nekrutenko A. Manipulation of FASTQ data with Galaxy. Bioinformatics 2010;26(14):1783-1785.
64. Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 2011;27(6):764-770.
65. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology 2014;15(3).
66. Wilson M. Associate Professor & Director, Forensic Science Program. 2015.

67. Gilmour MW, Graham M, Domselaar GV, Tyler S, Kent H, Trout-Yakel KM, Larios O, Allen V, Lee B, Nadon C. High-throughput genome sequencing of two Listeria monocytogenes clinical isolates during a large foodborne outbreak. BMC Genomics 2010;11(1).