Accumulation of Single Nucleotide Polymorphism (SNP) Mutations in *Escherichia coli* Grown Under Food Production Conditions and Their Importance in Outbreak Strain Epidemiology

By

James Austin Markell

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Science

in

Biology

Carleton University

Ottawa, Ontario

Abstract

Food borne illness outbreak investigations require accurate genotyping to identify outbreak clusters and link them to isolates recovered from possible sources of infection. Traditional genotyping methods such as PFGE lack the resolution to differentiate highly similar but epidemiologically unrelated isolates. Single Nucleotide Polymorphism (SNP) analysis of whole genome sequences shows promise in providing the required level of resolution, but lacks field relevant data to aid interpretation of results. Here, three *E. coli* strains of serotypes commonly associated with foodborne illness outbreaks were used to inoculate lettuce growing under field conditions and recovered weekly for SNP analysis. This pilot study indicates that the number of SNP differences accumulated in these strains while growing under field relevant conditions is very low, but that variability exists between strains and further study is warranted. These results will aid interpretation of SNP analyses during food outbreak investigations and help support resulting regulatory decisions.

Key words

Single Nucleotide Polymorphisms (SNPs), foodborne illness outbreak investigation, genotyping, genomic based comparison, Enterohemorrhagic *E. coli* (EHEC), field relevant

Acknowledgements

A great many people helped make this thesis a reality. First and foremost, I would like to thank Dr. Burton Blais for inviting me into his lab as an undergraduate, and helping me find my way through the academic maze to the place I find myself today. Along with Burton, Dr. Dominic Lambert and Dr. Catherine Carrillo have been instrumental to my success. Their guidance, advice, help and encouragement pushed me along when I needed it most and made this thesis a reality. Dr. Alex Wong and Dr. Rees Kassen are owed enormous thanks, both for their expert advice and for their enthusiasm for the project. Dr. Pascal Delaquis and his team in B.C. were indispensable and heavily involved in making this work field-relevant. Dr. Ken Dewer and Jessica Wasserscheid went above and beyond the call of duty and provided much needed reference assembly help. Dr. Adam Koziol, Mike Knowles and Jackson Eyres gave me my first taste of bioinformatics, and helped me during the times when it was too much for me to handle on my own. To Paul Manninger, Martine Gauthier, and the entire lab group at OLC I owe a huge amount of appreciation. The lab has been my home for a while, and I will miss it when I'm gone.

This work was funded by the Canadian Food Inspection Agency (CFIA) research and technological development budget.

Statement of Contributions

In this thesis, I carried out all wet-lab bacterial isolations, pre-sequencing growth, PCRs, sequencing data trimming and error correction, SNP analyses, and interpretation of results.

Lettuce inoculation, growth, and harvesting was carried out at Dr. Pascal Delaquis' laboratory, sequencing was performed by Paul Manninger at CFIA OLC, and reference strains were sequenced and assembled by Dr. Ken Dewer and his team at Genome Quebec (McGill University). Mike Knowles provided the unique identifier PCR primer pipeline and primers.

Table of Contents

List of Tables

List of Figures

I. Introduction

1. Impact of Foodborne illness and Enterohemorrhagic *E. coli*

It is estimated that there are over 4 million cases of foodborne illness each year in Canada. Of those 4 million, an estimated 845 resulting in hospitalization can be attributed to pathogenic *Escherichia coli* (Thomas et al. 2013; Thomas et al. 2008; Thomas et al. 2015). In addition to health and safety concerns, the burden on the Canadian economy from foodborne illness is estimated at $3.7 billion per year (Majowicz et al. 2004).

One of the most dangerous and potentially deadly types of foodborne illness is caused by enterohemorrhagic *E. coli* (EHEC), strains of *E. coli* that are a subgroup of Shiga-like toxin (*stx*) producing *E. coli* (STEC) associated with severe clinical illness in humans. Shiga-like toxins, also referred to as verotoxins, can cause hemorrhagic colitis (HC) and haemolytic uraemic syndrome (HUS) (Proulx et al. 2001). Symptoms include bloody diarrhea, colitis, hemolytic anemia, acute kidney failure, and potentially death (Corrigan et al. 2001; Levine et al. 1987; Nataro & Kaper 1998). These pathogenic strains of *E. coli* contain several other virulence factors such as intimin (*eae*) that allow them to colonize areas of the intestines of their host that are normally populated by harmless commensal strains (Donnenberg et al. 1993). Fibrillae and outer membrane proteins such as intimin can facilitate adhesion to mucosal cells, and indirectly induce cytoskeletal restructuring, cytokine activity, and, in EHEC strains in particular, internalization of the bacterial cell into host cells. The toxins secreted by these *E. coli* can interfere with host cell function and lead to cell death (kaper et al 2004). EHEC strains are thought to have evolved from enteropathogenic *E. coli* (EPEC) strains following

phage insertion of the characteristic *stx* genes, which are generally either stx_1 or stx_2 or their derivatives (Reid et al 2000).

2. Traditional *E. coli* Serotyping

Traditionally, *E. coli* have been classified based on their serotype, the structure of the surfaced-exposed lipopolysaccharide (LPS) and flagellar antigens (O and H antigens, respectively). These strain classifications are often used to identify foodborne illness outbreak strains of *E. coli*, despite lacking a direct connection to pathogenicity (Orskov et al. 1977). Strains that share a serogroup with known pathogenic strains are most likely closely related to them and therefore also assumed to have similar virulence characteristics, often referred to as "guilt by association". In 2003 Karmali et al proposed a seropathotype scheme placing different serotypes into seropathotype groups A to E based on their apparent link to disease severity in humans (Karmali et al. 2003). This scheme was widely used, but does not account for strains that show an O-antigen that is linked to human disease but are not themselves pathogenic, novel serogroups, or mutation of a strain such as the outbreak of *E. coli* O104:H4 in Germany in 2011 (Mellman et al. 2011). In a move similar to Karmali's seropathotype scheme, on September 2001 the U.S. Department of Agriculture's Food Safety and Inspection Service (FSIS) announced its intent to test for seven main serogroups most closely related to severe illness in humans in raw ground beef (USDA-FSIS, 2011).

These serogroups, O157, O26, O45, O103, O111, O121, and O145, are commonly referred to as the seven priority serogroups and are generally the serogroups that primary testing methods used by major governing bodies target when testing for

foodborne EHEC strains (Gill et al. 2012; Huszczynski et al. 2013; Blais et al. 2014; ISO, 2012; USDA-FSIS, 2014). These classifications are very useful for screening and regulatory testing, but are not specific enough to be of benefit during an outbreak investigation. Due to the fairly broad diversity within serotype strains of *E. coli*, matching two isolates based on serotype alone is not enough to confirm an epidemiological link. Traditional typing methods based on phenotypic features such as biotyping, serotyping, or phage typing (PT), can also be used as indicators of relation to virulent strains, but do not independently provide sufficient detail to support decisive regulatory action in an outbreak situation, especially given possible economic and health and safety impact of these actions. Likewise, plasmid analysis has been shown to be very useful on occasion (Horby et al. 2003), but most often cannot alone provide sufficient discrimination as plasmids can easily be lost or acquired by pathogenic bacteria.

3. Current Genotyping Methods

To ensure sufficient discrimination, most regulatory bodies use more discriminatory comparative methods. Pulse field gel electrophoresis (PFGE), one of the most common strategies for determining the relatedness of different isolates, uses genome restriction enzyme digest followed by analysis of fragment lengths via gel electrophoresis (Schwartz & Cantor, 1984). It is currently widely considered the gold standard for genotyping isolates for cluster determination and outbreak identification.

While generally successful, this method remains time consuming, and is not entirely reliable. It is possible for epidemiologically unrelated isolates to display the same digest pattern; for example, non-related strains of *Salmonella enteritidis* can occasionally

display identical PFGE patterns (Orson et al. 2007). It has also been shown that closely related strains can exhibit different PFGE profiles such as in the case of mixed success in the discrimination of *Campylobacter jejuni* isolates and certain *Listeria monocytogenes* strains (Gilmour et al. 2010; Taboada et al. 2013). PFGE simply does not have the discriminatory power to differentiate isolates with very similar genomes. In *E. coli* genomes rearrangements and insertions/deletions (indels) are far more common than single base pair mutations, which further confuses the matter as epidemiologically related strains can display different PFGE profiles based on a single genome rearrangement (Kudva, 2002). This makes it difficult to make regulatory decisions based on isolates with very similar profiles, and reduces the overall effectiveness of these approaches in comparative outbreak genomics.

Other molecular typing methods have been developed to remediate these issues and improve discriminatory power for epidemiological investigations. Polymerase Chain Reaction (PCR) based methods, such as random amplified polymorphic DNA (RAPD), repetitive sequence (rep-PCR), restriction fragment length polymorphism (PCR-RFLP) and amplified fragment length polymorphism (AFLP) are only a few of the many methods developed that look at selected regions of pathogen isolates' genomes. These methods have all been successfully used to differentiate isolates in outbreak investigations (Foley et al. 2007). Likewise, alternative approaches targeting only the sequences of highly conserved housekeeping genes such as ribotyping or multi-locus sequence typing (MLST), or repeat regions such as the variable number of tandem repeats (VNTRs) or multi-locus VNTR analysis (MLVA) methods (Miya et al. 2012) have also been developed to distinguish isolates. The latter two methods are based on

sequence databases that must first be populated via DNA sequencing experiments. All of the aforementioned typing methods rely on genetic differences dispersed over the entire genome, accumulated over time (point mutations, inversion, deletions, and insertions) to provide information on the relatedness of isolates. Unfortunately, they can lack the resolution required to identify small but significant changes because they use only a limited fraction of their targeted genome (Lukjanceko et al. 2010).

4. Single Nucleotide Polymorphism (SNP) Mutations

The majority of the methods discussed in the above section compare bacterial strains based on differences in their genomes in the specific DNA regions being studied; for example only tandem repeats or only conserved ribosomal housekeeping sequences. These genomic differences can have a multitude of sources, but in *E. coli* they most commonly arise from insertions, deletions, and genomic rearrangements (Kudva et al. 2002). Single nucleotide polymorphisms (SNPs) are single base pair mutations within a genome. Single nucleotide variations between genomes are referred to as single nucleotide variants (SNVs), and as such the terms SNV and SNP are often used synonymously. In *E. coli* these single base variants arise spontaneously from random mutation that can be caused by many different factors. Unlike insertions and deletions that can cause large shifts in reading frames and multiple amino acid residue changes, SNPs only affect a single residue at a time. Analysis of these SNPs should be less effected by single genomic change events, such as a phage insertion changing a PFGE profile.

SNP mutations can be 1) located in non-coding regions and thus cause no phenotypic change, 2) synonymous and cause no change to the resulting residue due to the redundant nature of DNA to amino acid sequence translation, 3) non-synonymous and cause little to no conformational changes in the final protein structure, for example a change between small uncharged residues, or 4) non-synonymous and cause changes that have large effects on protein structure, such as a mutation causing a change from a large hydrophobic residue to a proline residue (Bryant et al. 2012). Synonymous and non-synonymous but lower impact mutations (i.e. changes between residues with very similar sidechains outside of the active site of the folded protein) are less likely to be selected against, and are thus more likely to be retained by bacterial populations. Some synonymous mutations may increase the efficiency of translation and may be selected for, especially given the high codon bias seen in bacteria (Ochman 2003). Non-synonymous mutations that are beneficial are often strongly selected for, while non-beneficial mutations are usually selected against. Thus accumulation of SNPs in different bacterial strains is based on both the rate of spontaneous mutation and the effects of natural selective pressures on fixing or removing the SNPs from the population.

While the study of phenotypic changes resulting from SNP mutations is a very interesting field that can be useful in evolutionary and phylogenetic studies, from a food regulatory perspective SNP variation between strains has the potential to be a very powerful tool for direct genetic epidemiological comparison of outbreak isolates. SNPs occur less frequently than other genomic changes, and can be less complicated to identify and quantify. Where a single large genomic event such as a phage DNA insertion might cause a large change in PFGE profiles, it might have little impact on SNP analysis.

5. SNP-Based Genotyping

In contrast to traditional methods such as PFGE and MLST, whole genome SNP analysis looks at the entire genetic material of the organism on a base-to-base level, which allows for a very high level of discrimination between samples. A difference in SNP profiles could mean the difference between calling strains with identical PFGE profiles epidemiologically related or unrelated. The fact that SNPs are less frequent than the indels that cause most variation in PFGE profiles can also help determine if strains with very similar but slightly different PFGE profiles are related over an epidemiological timeline (Kudva et al. 2002). The greater discriminatory power also allows for a finer gradient of differences; i.e. genetic distance resulting in only a small change in PFGE profile might correspond to many SNP differences over a whole genome.

A 2013 study using SNP analysis to compare *Salmonella enteritidis* isolates showing identical PFGE profiles showed between 100 and 600 SNP differences between isolates (Allard et al. 2013). In 2009 and 2010, an outbreak of Salmonellosis was traced back to the source of contamination using SNP analysis of isolates obtained during the investigation after conventional typing methods such as PFGE had failed to differentiate the highly clonal isolates (Lienau et al. 2008). This highlights that proper investigation of foodborne outbreaks require accurate and timely clustering of isolates thought to be related, and that SNP analysis can provide that level of resolution where other methods fail.

The rapid evolution of next-generation sequencing and bioinformatics techniques continues to bridge the gap in analysis time and cost between WGS-based comparison

techniques and protocols such as PFGE, making the higher discriminatory power of WGS-based SNP calling an ever more attainable goal during a major outbreak. Direct genetic comparison of outbreak associated bacterial isolates can determine outbreak scope and sources of contamination. This can help tracking the spread and progression of particular strains over a relatively short time span.

6. The Benefits of Whole-Genome-Sequence-Based Typing

Traditional wet lab genotyping techniques such as PFGE have limited portability, as the results garnered are useful only for the specific technique used to gather them. Whole genome sequences can be used for analyses other than SNP analysis, and can in fact also be used to perform in silico typing such as MLST and PFGE in a fraction of the time of wet lab techniques, effectively replacing them once a sequence has been obtained.

Retaining outbreak isolate sequences can allow for the creation of large databases for analysis by future methods. While traditional methods only gather and retain method-specific information such as PFGE profile, WGS gathers all available genetic data in an easy to access format. Applying emerging WGS analysis techniques to the data could yield interesting discoveries that would otherwise be missed if only PFGE or MLST profiles were available from past isolates. The ability to test new techniques on a large historical set of data can facilitate the validation of emerging WGS analysis methods without waiting for new data; applying a new method to these WGS databases can be accomplished rapidly and immediately provides historical context for the newly emerging method dating back to the establishment of the database. These factors are driving a strong push by regulatory bodies to seriously invest in WGS strategies. The

Public Health Agency of Canada's (PHAC) primary means of monitoring emerging foodborne pathogens is PulseNet. While PulseNet currently uses PFGE extensively it has already begun to rely on whole genome sequence typing as well (Hunter et al. 2005; CDC, 2016; Dunn, 2016).

Genomic analysis using whole genome sequencing can provide unrivaled DNA fingerprinting capability and offer tremendous potential for food safety applications. Analyses such as these were used during the 2011 *E. coli* O104:H4 European outbreak to characterize the strain involved. The strain in question lacked the traditional *eae* intimin markers used to detect EHEC bacteria of this type and was not a serotype traditionally closely associated with human pathogenicity. This made it difficult to screen for isolates and determine the food vehicle of infection. Comparative genomic analyses were later used to compare the strain's genome to those of other known pathogenic and non-pathogenic strains of the same serotype. These studies identified newly acquired virulence genes in the strain and provided a better understanding of this organism's pathogenic capabilities over time (Mellmann et al. 2011). Had this information been available to public health risk assessors from the outset of the outbreak high resolution SNP typing may have facilitated accurate identification of the causative strain and helped identify related isolates, mitigating public health risks. The use of a SNP-based genomic comparison approach in particular would have allowed for high-fidelity identification of the strain in epidemiological samples tested during the outbreak independently of serotype and pathogenic gene profile.

Genomic analyses have been used to study genetic diversity, genome plasticity and niche adaptation in important foodborne pathogens for a number of years (Den

Bakker et al. 2008; Deng et al. 2010; Orsi et al. 2011; Grim et al. 2011; Kim et al. 2013). Still, application of these studies in a food safety context is in its infancy and the interpretation of genomic data from foodborne pathogens to support regulatory interventions remains challenging given the lack of definitive reference studies (Gilmour et al. 2013). Trace back investigations are essential for discovering the root causes of outbreaks and aiding in preventing them in the future. Despite this, they are notoriously difficult due to the inherently fluid nature of bacterial genomes. Bacteria found in crop fields, chicken farms or cattle processing facilities change over time to better survive under the selective pressures unique to these food production environments. The intrinsic differences in the rate at which DNA accumulates changes both within and between foodborne pathogen strains and species remains a little studied field with important implications in regulatory sciences.

SNP-based genotyping could also be applied to other areas where traditional genotyping lacks the required resolution. For example if re-occurring contamination is identified in a production facility, SNP analysis might reveal whether the closely related isolates arise from an influx of new contaminants, or a persistent contaminant strain within the facility. Coupled with analysis of isolates recovered from primary sources that supply the facility, the source of the contamination might even be identified. This information would be invaluable for correcting the issue.

7. Current Challenges in Whole-Genome-Sequence-based Genotyping

With the advent of accessible and rapid sequencing technologies and the heavy investments regulatory bodies are making into whole genome sequencing techniques, the

main issue being faced by WGS-based strategies is interpretation of results. With a nearly exponential increase in available data, accurate analysis and interpretation of this data remains a challenge. While the high level of discrimination WGS techniques such as SNP-based comparisons provide can be powerful, it is difficult to extrapolate results of these tests to real-world applications. For instance it can be difficult to determine the number of SNP difference required to exclude an isolate from an identified cluster of related isolates. Fast evolving pathogens (such as *E. coli*) may develop many SNP mutations between strains that are epidemiologically linked, while highly clonal species may not develop any at all. This makes it exceedingly difficult to accurately attribute an outbreak strain to a source without a reliable reference of basal mutation rate, especially over the time course of a foodborne illness outbreak.

There is a distinct lack of studies focussed on determining the basal rate of SNP accumulation in outbreak bacteria in field-relevant conditions. Health risk assessment decisions are based on information gained from analyses supported by a body of work composed of many studies. The lack of available studies of this type can make it difficult to base decisions on SNP analysis. PFGE and MLST methods have large databases of data and a multitude of studies validating their methodologies and providing references literature for interpreting data. Historical data built up over time further validates the assumptions made by these methods and lends them the reputation of reliability required for regulators to trust their results. Replicating this volume of work for the purposes of SNP analysis data interpretation will allow more accurate and confident health risk assessments built off of SNP data in much the same way PFGE is currently relied upon. PulseNet is currently working on a retroactive study of 1000 genomes to evaluate PFGE

and whole genome sequence analysis methods concurrently and help develop interpretation guidelines, which is a step in the right direction. As any decisions made by regulators need to stand up to the scrutiny of litigation, developing supporting research of this type is of high importance.

8. Bioinformatics tools

SNP calling relies on high quality, high fidelity raw reads. The publicly available trimmomatic bioinformatics tool trims the leading and lagging ends of reads based on a user supplied quality cut off. This is a key feature of the trimming parameter as the illumina MiSeq sequencing platform used for this experiment has a well-documented error bias towards the ends of its raw reads. Trimmomatic also checks for any MiSeq tag sequences that may have been missed, and trims the read based on a sliding average quality score when the average quality score of 4 consecutive bases drops below a user supplied threshold (Bolger et al. 2015). BBduk, an alternative bioinformatics tool similar to trimmomatic, works in much the same way (Bushnell 2016). Both programs also remove reads that are below a certain length threshold after trimming has been completed to improve read mapping fidelity.

There are many SNP calling platforms currently available, and more are being developed at a surprising pace. Although most of these platforms and pipelines are intended for eukaryotic organism genomics, there are still several available for bacterial genomic study. Several different programs and pipelines for SNP determination are used by public health agencies worldwide. Currently two primary methods of SNP calling are

available. Reference assembly free, where reads are directly compared, and reference assembly based, where reads are aligned to a reference assembly.

kSNP is an assembly reference free SNP calling program developed by Gardner *et al* (Gardner et al. 2015). kSNP breaks reads down into smaller kmers of identical lengths using a sliding window to ensure all data is included (i.e. bases 1 to 51 of a read are kmer number 1, bases 2 to 52 of the same read are kmer number 2, etc). It then compares all available kmers, looking for those that differ by only the center base of the sequences. In this way, it can determine SNPs between all available sequences without the use of a reference for comparison (Gardner et al. 2015). There are obvious benefits to this approach. A poor reference in reference based SNP calling can introduce false positive SNPs and similar errors, an issue avoided by not requiring a reference. Allowing for a matrix of SNP differences with comparisons between each genome in the analysis as opposed to comparing all genomes only to the reference improves the chances of locating SNPs located in genomic regions not covered by a reference genome.

Despite these obvious benefits, kSNP suffers from a number of drawbacks. There is no reference assembly to help identify DNA that is repeated in different areas of the genome. This can lead to comparison of reads that belong in different regions of the genomes if their sequences are similar enough. To address this kSNP removes kmers within a single genome that match except for the central base prior to comparison with other genomes in an analysis. This can remove sections of the genome from the analysis if there are any single base pair read errors. If repeated DNA regions are of poor quality, repeats may be removed from different areas of the genome in different samples. This may lead to comparison of kmers obtained from different areas of the genome, which

may appear as SNPs if they differ even if the genomes are actually identical, as kSNP has no reference to demonstrate that they are located in different areas of the genome. This may lead to a number of false positive SNP results. Another drawback is that SNPs that are located within one kmer length of each other will not be found by kSNP, as it removes kmers that are different in positions other than the center position.

While kSNP is assembly reference free, most SNP calling programs involve mapping of reads to a reference genome. In North America, the United States Food and Drug Administration's (FDA) Center for Food Safety and Applied Nutrition (CFSAN) has developed a reference based SNP analysis pipeline, referred to as the CFSAN SNP Pipeline (Davis et al. 2015). This program determines SNP positions based on read alignments to a reference, generating a matrix of SNP positions between all samples in an analysis run. Similarly, collaborators from the Public Health Agency of Canada (PHAC) British Colombia Center for Disease Control (BCCDC) and various universities have created an open sourced web platform based SNP analysis tool. This Single Nucleotide Variant PHYLogenomic (SNVPhyl) pipeline uses the Integrated Rapid Infectious Disease Analysis (IRIDA) web platform to perform all alignments and SNP calls (Petkau et al. 2016).

Both of these programs align raw reads to a reference genome and identify SNPs between the aligned raw reads and the reference genome. These pipelines do not analyse other mutation events such as indels, as these genome changes may be more strongly affected by single large genomic changes such as phage insertion or recombination, complicating analysis (Canchaya et al. 2003).

Reference based analysis avoids many of the issues kSNP's reference free approach experiences, however introducing a reference is not without its draw backs. As no quality scoring data is available in reference assembly based SNP analysis, these programs assume the reference genome is accurate; if any single base errors are present they will appear as SNPs when compared to the query genomes (Davios et al. 2015; Petkau et al. 2016). Running the reference genome against the raw reads used to create it and correcting for base pairs that appear as SNPs can mitigate this issue. Areas that are difficult to accurately sequence and assemble into the reference genome will also be troublesome for SNP calling program's alignment software. If the SNP calling program can't find an alignment of sufficient quality the base miscall will not be caught by this corrective run.

On a related note if alignments between query reads and the reference are not of sufficient quality those regions will be omitted from the analysis. This again stresses the importance of high quality sequencing data, and proper error correction and trimming of that data. A reliable and closed reference assembly can also help alleviate this issue. One point of interest regarding incomplete reference assemblies is that the gaps between contigs are the areas that both assembly programs and the SNP calling alignment programs are likely to have difficulty with. So, while these areas are less likely to be represented in a genome assembly, it is unlikely that high quality SNP calls would be obtained from these regions regardless.

9. Association of Field Lettuce and Foodborne Illness

One common route of human infection by EHEC strains is through pre-harvest infection of lettuce (Lynch et al 2009). Infection of lettuce intended for distribution and human consumption can be caused by agricultural irrigation systems drawing from contaminated watersheds, application of contaminated manure or compost on the lettuce fields, or direct and indirect contamination from the feces of infected wildlife and livestock (Nicholson et al. 2005). Flooding caused by heavy rain levels also appears to play a role in mobilizing *E. coli*.

While EHEC is traditionally associated with bovine meat products, it has been increasingly linked to lettuce, melon, radish, sprouts, spinach, and drinking water (Warriner et al. 2009; Allen et al. 2013; Lynch et al. 2009). In addition to a rising number of Canadian produce outbreaks, 22 documented produce outbreaks occurred from 1995 to 2006 in the United States, predominantly linked to contamination of California grown spinach and lettuce (Ravel, 2009; Cooley et al. 2007). Recent studies have shown that *E. coli* pathogen strains can migrate into the tissue of fresh produce grown in field, both on leaf surfaces and through contaminated soil in the root system. This indicates that washing lettuce prior to sale or consumption simply isn't enough to remove the presence of undesirable *E. coli* strains, which is in fact backed by recent studies into the effectiveness of surface washing lettuce isolates with chlorinated water (Delaquis et al. 2002). Other studies have proven the long-term survival of *E. coli* O157:H7 in manure, indicating that even with the precautions currently practiced by the agricultural industry produce-borne pathogenic *E. coli* is not an issue that will be alleviated in the near future (Solomon et al. 2002).

10. Anatomy of a Foodborne Illness Outbreak Investigation

Outbreak investigations typically follow the same rough methodology, regardless of what regulatory body is carrying out the investigation. There are several points during these investigations that rely on accurate typing methods, where the high resolution of SNP methods could be of use. Initially health risk monitoring programs will recognise a cluster of similar illnesses. Once the possibility of an issue has been established a cluster of related events will be identified to draw a circle for epidemiological study. Typing methods are used to determine which individual cases should be included or excluded from the identified cluster based on how closely related the isolates from those cases are, usually reported in the form of a phylogenetic tree. After characterizing the cluster, epidemiological investigations are used to drill down and find a common source for the outbreak (WHO, 2008). It is extremely important to accurately identify the cluster to facilitate these epidemiological investigations. Having a narrow selection of clearly linked cases helps to identify possible sources faster and reduces false leads. For instance if questionnaires about recent food products consumed are used as an epidemiological tool, the results of these questionnaires will be easier to interpret if the cases were accurately linked, as all patients affected by the outbreak will share a common source of infection. If a patient affected with a strain that is unrelated to the cluster is incorrectly included in it, it could potentially complicate the investigation as at least one questionnaire response will likely not include the causative agent of infection. SNP typing and other WGS methods could provide the resolution to obtain far more accurate clusters.

The epidemiological investigation will usually lead to specific sources such as a particular food commodity, and eventually testing of said food commodities will lead to a production plant, and ideally specific production lots. Once isolates from a potential source are obtained, they must again be linked to the cluster to ensure that the correct source of contamination has been identified. Clearly linking an isolate identified as the source of the illness to clinical isolates during an outbreak investigation is a crucial step prior to any regulatory action. The goal of any food safety investigation is to accurately identify the source of a foodborne illness and to take regulatory action to mitigate public health risk and financial impact to industry. As such, accurately identifying the source of the illnesses in a timely manner is of paramount importance. As discussed above, typing methods currently in use can encounter difficulties in resolving closely related or clonal strains due to a variety of drawbacks inherent to them. WGS analysis methods are already being used to characterize clinical and outbreak strains with greater resolution (Roetzer et al. 2013), and although they do not yet have the body of evidence required to be reliable within the scope of a foodborne illness outbreak they show remarkable promise once decision makers have a better understanding of how to analyse the data they provide.

11. The Importance of Field Relevant Parameters

The basal rate of SNP mutation events has been estimated in many studies, traditionally by measuring the rate of reversion to wild-type for lacZ mutant *E. coli* strains grown on lactose-minimal media (Cupples et al. 1989; Bryant et al. 2012). The error rates of *E. coli* DNA polymerases, incorporating for proofreading functions, are estimated to be roughly 5×10^{-10} per base per replication (Fijalkowska et al. 2012), while the actual error rate of *E. coli* DNA replication has been determined experimentally by

whole genome sequencing to be as low as 1 X 10^{-3} per genome per generation (Lee et al. 2012). Note the distinction here between the rate of SNP mutation events, and that of the rate of SNP accumulation. The mutation rate is a random process unaffected by the forces of natural selection and effects such as genetic drift. It also doesn't account for population load or mutational pressures from the environment of the population. It has been shown that rates calculated in mutation rate studies do not match those seen in direct genomic comparisons of isolates of related strains (Bryant et al. 2012).

A 2012 review of laboratory based mutation rate studies showed less than 4 SNPs per 100 generations in 21 strains spanning 7 studies, only 4 of which showed more than 1 SNP per 100 generations (Dettman et al. 2012). There are numerous studies of this type, but they do not account for the effects of field conditions. A 2011 study looking at 14 uropathogenic *E. coli* clones isolated from a single source of persistent contamination over three years found only 20 SNP changes between the genomes (Reeves et al. 2001). This study investigates SNP accumulation rate as opposed to SNP mutation rate, but still lacks food safety relevancy as it is dealing with human and canine host growth conditions as opposed to food production relevant conditions which may exhibit very different selective conditions.

Studies of SNP mutation rates commonly use generic apathogenic strains. It has been shown that pathogenic strains of bacteria commonly acquire mutations at a greater basal rate than apathogenic strains (LeClerc et al. 1996; Wirth et al. 2006, Matic & Radman, 1997). Thus, studies looking at the basal rate of mutation in strains only distantly related to pathogenic strains may underestimate the rate when extrapolated to common outbreak strains.

Strains grown under food production relevant conditions, such as in field grown lettuce, are subject to many stresses and pressures not encountered in a laboratory environment that may affect the rate of SNP mutation and accumulation. UV sunlight has a direct mutagenic effect on DNA, which increases the rate of mutation. This is a well-studied phenomenon, related to the high error rate of DNA repair mechanisms relative to transcription mechanisms (Bates et al. 1989). Selective pressures such as scarcity of resources, defences of the host lettuce plant, and stiff competition with regards to the lettuce's native biome may also influence the rate of both SNP mutation and accumulation. Mutation rates may directly increase due to the SOS response to these cell stressors, which up regulates error prone DNA repair processes (Janion 2008). These environmental factors also exert selective pressures, causing an increase or decrease in fixation of SNPs in the population depending on their effect on the fitness of the cells. Low population sizes may also affect fixation rates due to factors such as genetic drift.

Determining the rate of SNP accumulation divorced from these factors may have little bearing on the number of SNPs you would expect a population of pathogenic bacteria to accumulate while growing under field-relevant conditions. While it is not feasible to study all of these environmental effects on SNP accumulation in a laboratory setting, it is possible to study the overall effect of these influences by studying the rate of SNP accumulation in a bacterial population grown under field relevant conditions.

12. Intention of This Study

The lack of field-relevant data supporting SNP analysis remains a gap in the development of SNP based regulatory strategies. One of the most common questions

asked when evaluating whole genome sequence based typing data is that of how many SNPs you expect to find if isolates were from a common food source.

To address that question, lettuce was inoculated with strains of *E. coli* closely related to pathogenic strains (i.e. sharing a serotype with one of the priority seven serotypes as discussed above) and grown under typical field conditions. The lettuce was harvested in one week intervals, and the inoculated populations were recovered via a selective enrichment process based on current regulatory standard practices and sequenced for SNP analysis.

The intent of this study is to provide data on the basal rate of SNP accumulation in *E. coli* populations grown in real world conditions. This information can be used to help health risk assessors interpret SNP data during outbreak investigations, lending credibility to any decisions made based on such analyses. The hypothesis is that very few SNPs will accumulate over the timespan.

II. Materials and Methods

1. Microorganisms and Culture Conditions

Three *E. coli* strains lacking *stx1* and stx2 genes were used for the experiments described herein. The *E. coli* O157:H7 strain is a nalidixic acid resistant derivative of the ATCC 700728 (NCTC 12900) parent strain, OLC culture collection number 811 (Bezanson et al. 2012). The O103:H2 and O111:NM strains used were obtained from the CFIA's culture stock at the Ottawa Carling Laboratory (OLC), and were originally isolated by the Laboratory for Foodborne Zoonosis (LFZ) Guelph and Ottawa Lab Carling (OLC), respectively. They are OLC culture collection numbers 969 and 682 respectively. Stock cultures were kept at -80ºC in tryptic soy broth (TSB, Oxoid, Ltd., Basingstoke, UK) containing 10% glycerol and control working cultures were kept on nutrient agar at 4ºC (NA, Oxoid, Ltd., Basingstoke, UK). Colonies were isolated from stock by 24-hour incubation of tryptic soy agar plates (Sigma-Aldrich, St. Louis, MO), followed by a 24 hour incubation at 37ºC with agitation. Isolates recovered from the lettuce were stored at -80ºC in modified TSB (mTSB, Oxoid, Ltd., Basingstoke, UK) containing 15% glycerol.

2. Preparation of Field Plots

Romaine lettuce seedlings (Cos/Romaine cv. Parris Island; *Lactuca sativa* var. *longifolia*) were grown in a greenhouse for 3 weeks and hardened in an outdoor frame for 3-4 days (i.e., acclimated to outdoor elements). They were then planted in field soil and grown for 2-3 weeks or until the plants were roughly 10-15 cm high prior to inoculation. Plants were planted in 6 double planted rows separated by 150 cm and 76 cm high plastic

fencing to reduce lateral transfer by dust or rain splash. Two uninoculated guard rows were included at either end to prevent outside contamination. 120 plants were planted per row, which were then thinned to 30 plants per row (60 plants per plot) once established (see Figure 1). All growth was completed at Pacific Agri-Foods Research Center in Summerland, BC.

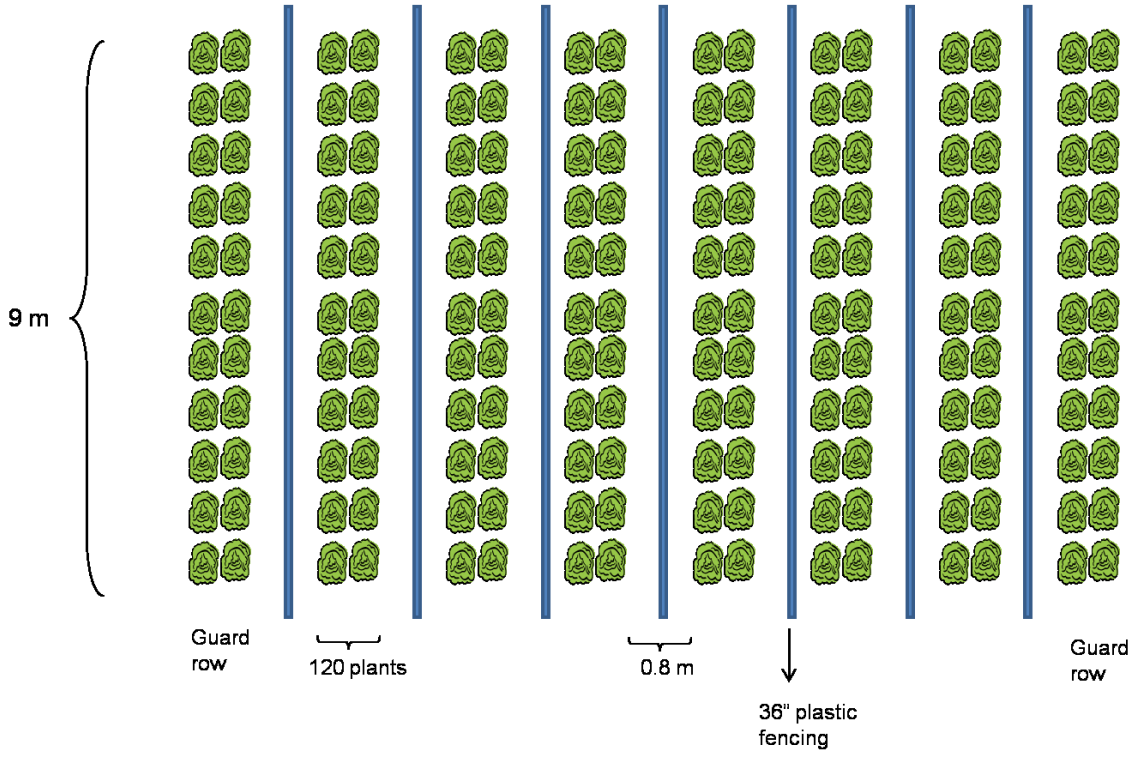

Figure 1: Planting scheme for field lettuce prior to inoculation with *E. coli*

3. Lettuce Inoculation and Growth

Lettuce plants were inoculated after 3-4 weeks of field growth, when at a height of roughly 10-15 cm. Aerated overnight TSB cultures incubated at 37ºC of each of the strains were diluted 10-fold twice and 100 fold once with sterile distilled water to obtain a cell density of roughly 10^6 CFU/ml. This inoculum was then evenly applied to all plants

and exposed soil in a given row by dispensing from a watering can held roughly 30 cm above the plants while walking down the row, resulting in roughly 2L per row (16 ml per plant). Plants were watered as necessary, up to 8 times per day for 30 minutes. 1.0 g/L 20:20:20 (N:P:K) fertilizer was applied once weekly, and manual weeding was performed by analysts wearing disposable boot covers.

4. Recovery of Lettuce Isolates in Year One

Three heads of lettuce for each strain were removed at random from each row immediately following inoculation and after 1, 2, 3 and 4 weeks of growth by cutting the plant stalk 2.5-3 cm above ground level with a sterile knife. Analysts wore disposable boot coverings and gloves during sample collection.

A representative tissue sample of 25 g was transferred into sterile sample bags via sterile tongs. Samples were taken at random from interior and exterior leaves at leaf tip and stalk levels. 225 ml of m-TSB was added to each sample bag, which were then mechanically agitated for 2 minutes and incubated overnight at 42ºC. Four hours into the incubation 2.5 ml of 1% vancomycin/cefsulodin in sterile water was added to each sample bag. After incubation 100 µl of each sample was added to 9.9 ml of 0.1% sterile peptone (Bacto Peptone, BD Biosciences, New Jersey, U.S.A). This was then serially diluted 10-fold to create 10^{-3} to 10^{-6} diluted samples. 100 µl of each dilution was then spread plated to a plate and incubated at 37ºC overnight. Rainbow agar was used for the O103 and O111 strains, while cefixime/tellurite sorbitol MacConkey agar (ct-SMAC) with 0.025 mg/ml nalidixic acid was used for the nalidixic acid resistant *E. coli* O157 strain. Two isolated colonies matching control plate colony morphology for each sample

were then used to inoculate 1 ml of mTSB. After an overnight incubation 10 µl was put aside for a unique identifier PCR and the remainder was stored at 4ºC.

5. Recovery of Lettuce Isolates in Year Two

The experiment was repeated the following year using the same inoculation, growth, and sample collection parameters. 225 g of sample tissue was used with 450 ml of mTSB instead of 25 g / 225 ml mTSB. Final concentrations of 10 µg/ml and 3 µg/ml vancomycin/cefsulodin were obtained by adding 1 mg/ml stock solutions after 4 hours of incubation, as opposed to the 0.1% final concentration used in year one. The enriched mTSB broth was then transferred directly to MAC plates containing vancomycin and cefsulodin (MAC-VC) using a sterile loop. After overnight incubation at 42ºC characteristic colonies were transferred to tryptone soy agar (TSA) plates, and then to 1 ml of mTSB containing 15% glycerol and frozen at -80ºC. The stored samples were used to streak SMAC-CT plates to confirm purity. Three colonies from each plate were then used to inoculate 1.8 ml of mTSB, which was used for DNA extraction as described below.

6. SigSeekr

All publicly available non-redundant *E. coli* genomes were retrieved from GenBank and combined with locally generated whole-genome sequences to create a genomic sequence database. The SigSeekr pipeline (http://github.com/OLC-LOC-Bioinformatics/SigSeekr) (Knowles et al. 2016) was then used to identify DNA sequences unique to the query strains using BLASTn 2.2.29+ (Camacho et al. 2008) as follows: Query sequences with matches to the genomic sequence database reporting

initial E-values ≤ 1.0 and ≥ 90% identity were removed and substituted with degenerate bases (N), and a string of least common sequences linked by the degenerate bases (which were ignored in subsequent BLASTn database searches, reducing redundancies and improving speed) were generated. Repeated unique sequences were then eliminated from the string of query sequences using a fuzzy matching algorithm to prevent a duplication of signature sequences in the output. Sequences below 200 base pairs were eliminated from the string to ensure suitability for PCR amplification. When no sequences were returned in the string, the process was recursively iterated using lower E-values until at least one was found. The resulting string, containing the least common sequence(s), was then used to query the entire pan-genomic database (including target strain) for quality control. Once a unique identifier sequence was obtained primers were designed to amplify the region using Primer-BLAST (Ye et al. 2012) and obtained from IDT (Integrated DNA Technologies, Coralville, Iowa, USA). Primer sequences are available in **Table 1**. Note that no unique identifier PCR primer set was identified for the O157 strain, which was characterized primarily by its uninhibited growth on ct-SMAC-nal plates.

Table 1: Unique identifier primers used for confirmation of isolate identity

Strain	Forward (5'->3')	Reverse (5'->3')	Product Length
O111:NM	AGGCACCCAGACACGTTAAA	CACCATGCTGTGCTGTATGC	463
O103:H2	GCTGGCCTGAACACCTGTAT	GCAAGCTTCTCTTGGGGGAA	473

Bacterial lysate was obtained for each O103 and O111 year-one sample by adding 10 µl of the incubated mTSB to 10 µl of 2% v/v triton X-100 (Sigma-Aldrich) and heating at 100ºC for 10 minutes. 5 µl of lysate was then added to 45 µl PCR mixture containing 2.5 units HotStar *Taq* and 1 x HotStar PCR buffer (Qiagen Inc., Mississauga, ON, Canada), 2.5 mM $MgCl_2$, 200 µM of each dNTP and 0.2 µM of that strain's unique identifier PCR forward and reverse primers. Note that 5 µl of DNA extract from each year two samples was used directly in the place of lysate for the procedure. The PCR was carried out in a BioRad Touch thermal cycler (Bio-Rad Laboratories, Inc) using the following conditions: 15 min at 94 °C, followed by 35 cycles of: 30 s at 94 °C, 30 s at 55 °C, and 1 min 30 s at 72 °C. An additional 2 min at 72 °C followed the last cycle. PCR products were analyzed by electrophoresis using either the FlashGel (Lonza, Rockland, ME, USA) or the QIAxcel (Qiagen Inc.) systems, following the manufacturers' instructions. Results were considered positive if a band at the corresponding probe sequence size was detectable.

7. DNA Isolation

Samples that tested positive for the unique identifier PCR test were used to inoculate 1.8 ml of mTSB. This was then incubated for 4.5 hours at 37ºC. The enriched broth was then concentrated by spinning down 1 ml of the broth at 13,000 RPM for 2 min and removing 600 µl of the supernatant. After resuspension the samples were DNA extracted via the Maxwell 16 Cell SEV DNA Purification Kit (Promega, Madison, WI, USA). The remaining 800 ml of broth was diluted to 30% glycerol and frozen at -80ºC.

8. Sequencing of Field Isolates

DNA was quantified using the Quant-iT High-Sensitivity DNA Assay Kit (Life Technologies Inc., Burlington, ON, Canada). Sequencing libraries were constructed from 1 ng of genomic DNA using the Nextera XT DNA Sample Preparation Kit (Illumina, Inc., San Diego, CA, USA) and the Nextera XT Index Kit (Illumina, Inc.). Paired-end sequencing was performed on the Illumina MiSeq Platform (Illumina, Inc.) using the 600 cycle MiSeq Reagent Kit v3.

Sequencing reads obtained from the recovered isolates were assembled using SPAdes 3.7.1 (Bankevich et al. 2012), and ribosomal multilocus sequence typing (rMLST) was conducted on the assemblies using a BLAST-based custom python script (https://github.com/OLC-Bioinformatics/MLST) and databases downloaded from http://pubmlst.org/ (Jolley et al. 2012; Wirth et al. 2006).

9. DNA Isolation and Illumina MiSeq Sequencing of Initial Inoculum Strains

Initial inoculum strains were used to inoculate 1.8 ml of mTSB. This was then incubated for 4.5 hours at 37ºC. The enriched broth was then concentrated by spinning down 1 ml of the broth at 13,000 RPM for 2 minutes and removing 600µl of the supernatant. After resuspension the samples were DNA extracted via the Maxwell 16 Cell LEV DNA Purification Kit (Promega, Madison, WI, USA). The remaining 800 ml of broth was diluted to 30% glycerol and frozen at -80ºC. The DNA was quantified using the Quant-iT High-Sensitivity DNA Assay Kit (Life Technologies Inc., Burlington, ON, Canada). Sequencing libraries were constructed in duplicate from 1 ng of genomic DNA using the Nextera XT DNA Sample Preparation Kit (Illumina, Inc., San Diego, CA,

USA) and the Nextera XT Index Kit (Illumina, Inc.). Paired-end sequencing was performed on the Illumina MiSeq Platform using the 600 cycle MiSeq Reagent Kit v3 (Illumina, Inc.).

10. DNA isolation, Sequencing, and Assembly of Reference Strains

DNA libraries were prepared following the Pacific Biosciences "20 kb Template Preparation Using BluePippin Size-Selection System" protocol. 7.5 µg of high molecular weight genomic DNA (final volume of 100 µl) was sheared using the Covaris g-TUBES (Covaris Inc., Woburn, Massachusetts, USA) at 4200 RPM for 60 seconds on each side, in an Eppendorf centrifuge 5424 (Eppendorf, Hamburg, Germany). Sheared DNA was size selected on a BluePippin system (Sage Science Inc., Beverly, MA, USA) using a cut-off range of 7 kb to 50 kb. The DNA damage repair, end repair and SMRT cell ligation steps were performed as described in the template preparation protocol with the SMRT cell Template Prep Kit 1.0 reagents (Pacific Biosciences, Menlo Park, CA, USA). Sequencing primers were annealed at a final concentration of 0.8333 nM and the P6 polymerase was bound at 0.500 nM. The libraries were sequenced on a PacBio RSII instrument at a loading concentration (on-plate) of 80pM using the MagBead OneCellPerWell loading protocol, DNA sequencing kit 4.0, SMRT cells v3 and 4 hour movies.

Preliminary contig assemblies were done with smrtanalysis version 2.3.0.140936.p2, incorporating BLASR for long read correction, Celera for assembly, and Quiver for read realignment and final basecalling. In-house scripts were used for chromosomes and associated plasmids edge trimming, gap closure, and confirmation of

circularity. The chromosomal sequences were then re-oriented to begin with the dnaA origin of replication gene. Additional in-house scripts, including BLAT and BWA, were used to align the Illumina MiSeq datasets (obtained as described above) to confirm basecalling and correct homopolymer lengths when needed.

Reference assemblies were compared to the raw read files used to create them using SNVPHYL (version 1.0.1 Paired_End) with the recommended default settings, including a minimum coverage cut-off of 10, a minimum mean mapping of 30 and a SNV abundance ratio of 0.75. The resulting SNP in the O103:H2 reference file and the 9 SNPs in the O111:NM file were then manually corrected. No SNPs identified in any field isolates contained SNPs in these positions.

11. Trimming and Error Correction

Sequencing errors in illumina MiSeq reads were corrected using Quake (version 0.3) with a k-mer size of 33 (Kelley et al. 2010). Trimmomatic was used to trim reads prior to SNP analysis with a minimum length cut off of 50, lead and lag trim of 25, and a minimum window quality trim of 30 (Bolger et al. 2014). Reads were also trimmed in parallel with bbduk using a q value of 15 with right and left trimming, a minimum cut off length of 50, and the illumina clip option to remove any leftover illumina adapter reads (Bushnell 2016).

12. SNP Analysis

Reads trimmed using bbduk and trimmomatic in parallel (as discussed above) were analysed against the reference assemblies outlined above. SNVPhyl (version 1.0.1 Paired_End) was used with the recommended default settings, including a minimum

coverage cut-off of 10, a minimum mean mapping of 30 and a SNV abundance ratio of 0.75, to compare all isolates of each strain to their reference concurrently (Petkau et al 2016). CFSAN's SNP Pipeline was used with default settings to compare all isolates of each strain to their reference concurrently (Davis et al 2015).

Resulting identified SNPs were quality checked by aligning the raw read data that contained the SNP to the reference using SMALT mapping and viewed using Tablet to ensure they contained 100% read consensus (Ponstingl 2014; Milne et al. 2013).

13. Sanger Sequencing Confirmation of Select SNP positions

A selection of identified SNPs were confirmed via Sanger sequencing. SNPs confirmed in this way were Sanger sequenced for both the isolate in which the SNP was identified and an unrelated isolate that did not contain the SNP. Forward and reverse PCR primers were designed to amplify a roughly 250 to 300 base pair region incorporating each SNP position using Primer3 (Untergasser et al. 2012; Koressaar & Remm, 2007). A second set of primers were designed as sequencing primers for Sanger sequencing nested within the amplified PCR products, at least 60 base pairs up stream of the SNP positions. Where primers of sufficient quality could not be produced within the amplified DNA, the forward or reverse primer from the initial amplification was used in place. All primers were obtained from IDT (Integrated DNA Technologies, Coralville, Iowa, USA). Amplification was performed using the following PCR parameters:

2.5 µl of DNA extract was added to 27.5 µl PCR mixture containing 1x TopTaq (Qiagen Inc. Alameda, CA, USA) and 1.0 µM of that strain's unique identifier PCR forward and reverse primers. The PCR was carried out in a BioRad Touch thermal cycler

(Bio-Rad Laboratories Inc. Berkely, CA, USA) using the following conditions: 15 min at 94 °C, followed by 35 cycles of: 30 s at 94 °C, 30 s at 58 °C, and 1 min at 68 °C. An additional 2 min at 68 °C followed the last cycle. PCR products were confirmed via electrophoresis using the QIAxcel system (Qiagen Inc. Alameda, CA, USA), following the manufacturers' instructions. Resultant amplicons were sequenced at Genome Quebec Innovation Center via a 3730xl DNA Analyzer (Applied Biosystems, CAL, USA).

14. SNP Comparison of Strains to Strains of the Same Serotype

SNVPhyl (version 1.0.1 Paired_End) was used with the recommended default settings, including a minimum coverage cut-off of 10, a minimum mean mapping of 30, a SNV abundance ratio of 0.75 and GTR+ γ model as default and tree support values estimated using PhyML's approximate likelihood ratio test, to compare each reference assembly to a selection of strains showing the same O-antigen as the reference (Petkau et al. 2016). These genomes were obtained from the CFIA Ottawa Lab Carling (OLC)'s culture collection and represent all O103 and O111 sequenced strains available in the collection, as well as 20% of available sequenced O157 strains.

III. Results

1. Recovery of isolates from field lettuce

Three lettuce heads were harvested immediately following inoculation for each strain, and again after every seven day period. The recovery method was applied to each sample, but was not successful in all cases, especially in the second year.

In an attempt to improve the performance of the recovery protocol, it was repeated over two separate summers with the recovery performed at separate laboratories by different personnel. This mitigated the possibility of laboratory technician error affecting the yield of the recovery efforts. The year one recovery protocol was performed in Ottawa, Ontario, while the year two protocol was performed in Burnaby, British Colombia. The second year of the experiment used a slightly different protocol that had proven to be reliable in the past incorporating a larger tissue sample and direct primary enrichment broth plating (Bezanson et al. 2012). This protocol did not appear to improve the yield of the recovery. Both methods used the CFIA's Compendium of Analytical Methods MFLP-52 methodology as a baseline, with slight modifications to account for the nalidixic acid resistant feature of the O157 strain (Government of Canada, H.C. 1999). A representative tissue sample of the lettuce was mechanically agitated to release possible internally contained bacterium, and then incubated at 42^{o}C in mTSB with added cefsulodin and vancomycin to inhibit non-target growth, primarily gram positives. *E. coli* is more tolerant of these higher temperatures and antibiotics than some bacteria native to the lettuce (Kalchayanand et al. 2013; Tillman et al. 2012; Windham et al. 2013).

Figure 2 shows the results of the recovery procedures. Isolates were recovered from 3 heads of lettuce for each strain at time zero (immediately after inoculation). Isolates from three heads of lettuce harvested after one week of growth were recovered for the O103 and O111 strains, while no week-one isolates were recovered for the O157 strain. In the first year three O103 isolates were recovered, along with two each of the O157 and O111 strains. In the second year three isolates each of the O157 and O103 strains were recovered. No O111 strain isolates were recovered past week one in the second year, and nothing was isolated past week 2 in the second year. In the first year three O103, two O111 and one O157 isolates were recovered. Nothing was recovered from any lettuce samples gathered after 3 weeks of field growth.

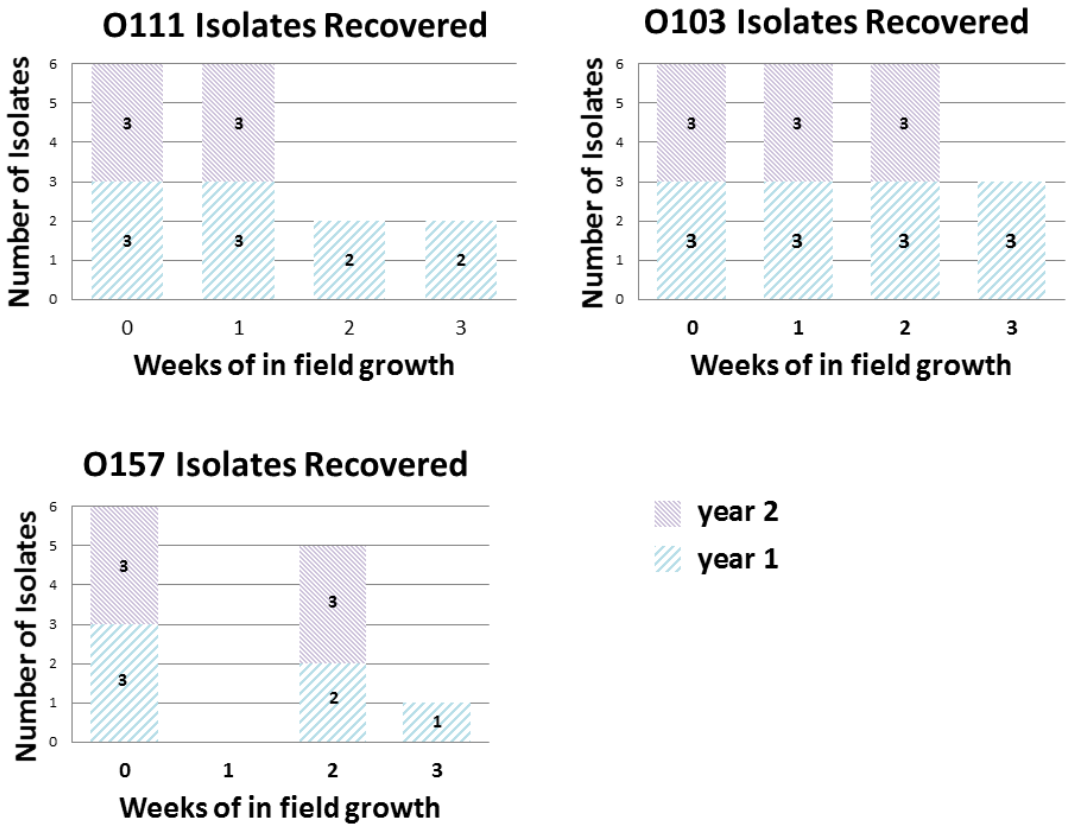

Figure 2: Recovery of inoculated isolates from field lettuce, repeated over two years

2. Confirmation of Correct Isolation

Isolates obtained from the recovery procedure were confirmed to be descendants of the initial inoculum by two primary methods. Sequences of DNA unique to the O111 and O103 strains were obtained by algorithmically comparing the whole genomes of the initial inoculum strains to NCBI's entire strain database. Primers were then designed to amplify these unique regions, and a simple PCR amplification assay was used to ensure the isolated colonies contained the unique DNA sequence. *In silico* MLST typing was performed on all sequenced isolates to ensure they shared a MLST profile with the initial inoculum strains. Both methods confirmed that each isolate was a match to the initial inoculum.

Ct-SMAC-nal plates selecting for the nalidixic acid resistant O157:H7 strain using a low concentration of nalidixic acid were used in place of the differential rainbow agar plates used for the O111:NM and O103:H2 strain isolates. While no unique identifier PCR was identified by SigSeekr for the O157:H7 strain, the selective nalidixic plates isolated the strain from background growth sufficiently to ensure that the correct strain was recovered. The strains grown on differential media were identified by morphology compared to control cultures, and confirmed via the methods discussed above. While the O157:H7 selective Nal plates had pure and homologous colony morphologies, the rainbow agar differential plates used to isolate the other strains showed a wide diversity of morphologies. While the colonies matching the control cultures were easy to identify, this further illustrates the diversity of bacteria native to field environments such as the lettuce used for this experiment.

3.0 SNP Analysis

All SNPs identified by SNVPHYL were also identified by the CFSAN SNP pipeline, and all SNPs were identified in raw reads trimmed with both the strict Trimmomatic and more lenient BBDUK trimming protocols. None of the sequences surrounding any SNP position were located in more than one position in the reference genome (as ascertained via BLAST), and all SNPs had perfect read consensus and a depth of coverage of at least 5 reads. All SNPs identified were located at least 1000 base pairs apart, and appeared to be spread randomly throughout the genome with no identifiable pattern or specific genomic features associated with their locations.

3.1 SNP Analysis of O157:H7 and O103:H2 Strains

No SNPs were identified in any recovered O157 isolates relative to the parent inoculum. A single SNP was identified in an O103:H2 isolate recovered from a lettuce sample taken after two weeks of growth. This SNP was located at base pair position 3510540 relative to the reference genome, located in a gene coding for an uncharacterized host specific protein. It is a transversion change from a cytosine to a guanine and within context of the open reading frame it changes the third position base, resulting in a silent mutation.

3.2 SNP Analysis of O111 Strains

Tables 2 and 3 show the SNPs identified in each of the recovered O111 isolates as well as any characterized genes and changes to the codon they are found in, if any. The number of SNPs per isolate ranged from one to twelve, with the average and the mode being five per isolate. Of the 93 SNPs, 15 were located in non-coding or unidentified

regions, and 10 were synonymous mutations (i.e. no change to residue identity). Of the non-synonymous mutations 11 were changes between two residues with small non-polar sidechains, 20 were changes between residues containing a positively charged side chain to a polar uncharged side chain or a polar uncharged to a positively charged side chain, 9 were negatively charged side chains to small uncharged side chains, 7 were a small side chain or polar side chain to a proline, 1 was a change to a stop codon, and 3 were changes from large uncharged residues to a cysteine. These changes ranged from nearly synonymous (i.e. non-polar to non-polar small sidechains) to possibly causing distinct phenotypical changes in the resulting protein (i.e. small side chain to a proline). Three of the unique identified SNPs were transversion mutations, while the remaining SNPs were transitions.

In addition to the SNPs identified in each isolate, there were 8 SNPs identified in 6 of the year one isolates, and 9 SNPs identified in the other 4 year one isolates. 5 of the 8 SNPs identified in the six year one isolates as discussed above were also identified in every year two isolate, along with 6 additional SNPs. **Table 4** shows these SNPs and their associated open reading frame information were applicable.

Sanger sequencing was used to confirm a selection of SNP results. Both the isolate containing the SNP and an isolate not containing the SNP were PCR amplified for a roughly 250 bp region surrounding the SNP position, and then Sanger sequenced using a nested sequencing primer where possible. SNPs confirmed in this manner are indicated in **Tables 2, 3, and 4**.

Table 2: SNPs identified in each isolate recovered from field lettuce during the first year of the experiment (year 1) and associated open reading frame information.

SNPs marked with an asterisk were confirmed via Sanger sequencing

Week	Isolate	Position on reference genome	SNP	Change to open reading frame codon	Gene associated with open reading frame
Immediately followinginoculation	Isolate 1	1484916	A -> C	Lys to Thr	Ethanolamine operon regulatory protein
		*2137681	A -> C	Leu to Stop	hypothetical protein
		2346088	T -> G	Leu to Pro	Flagellar hook-associated protein FlgK
		*2845774	T -> G	Lys to Thr	COG1649 predicted glycoside hydrolase
		*2909261	T -> G	Silent	Putative transport protein
	Isolate 2	*978489	T -> G	Ile to Met	2-acylglycerophosphoethanolamine acyltransferase (EC 2.3.1.40) / Acyl-[acyl-carrier-protein] synthetase (EC 6.2.1.20)
		*2066460	A -> C	Glu to Ala	L,D-transpeptidase YcbB
		*2968038	T -> G	Cys to Gly	PTS system, maltose and glucose-specific IIC component (EC 2.7.1.69) / IIB component (EC 2.7.1.69)
		*4913013	T -> G	Leu to Trp	Excinuclease ABC subunit A
		*5113518	A -> C	Lys to Gln	formate dehydrogenase formation protein FdhE
	Isolate 3	1144062	T -> G	Glu to Ala	6-phospho-beta-glucosidase ascB (EC 3.2.1.86)
		2722963	A -> C	Non-Coding	Non-Coding
		4934733	A -> C	Non-Coding	Non-Coding
		5186237	A -> C	Silent	Cof protein, HD superfamily hydrolase
		2734044	A -> C	Ile to Leu	Phenylacetate-CoA oxygenase, PaaG subunit
Week 1	Isolate 1	1576546	A -> C	Lys to Thr	Fimbriae usher protein StfC
		1964115	A -> C	Asp to Ala	Soluble aldose sugar dehydrogenase, PQQ-dependent (EC 1.1.5.-)
		4486546	T -> G	Phe to Cys	Phosphoglycerol transferase I (EC 2.7.8.20)
		5034577	T -> C	Non-Coding	Non-Coding
		5060112	A -> C	Phe to Leu	Cystathionine gamma-synthase (EC 2.5.1.48)
	Isolate 2	397215	A -> C	Lys to Thr	Methyl-directed repair DNA adenine methylase (EC 2.1.1.72)
		1003253	A -> C	Leu to Arg	N-acetylglutamate synthase (EC 2.3.1.1)
		1095444	A -> C	Met to Leu	CRISPR-associated helicase Cas3
		1654548	T -> G	Asn to His	Acetate kinase (EC 2.7.2.1)
		2145818	T -> G	Silent	Prophage Clp protease-like protein
		2276137	T -> G	Ser to Ala	Citrate:6-N-acetyl-6-N-hydroxy-L-lysine ligase, alpha subunit (EC 6.3.2.27), aerobactin biosynthesis protein IucA @ Siderophore synthetase superfamily, group A @ Siderophore synthetase large component, acetyltransferase
		1914010	A -> C	Non-Coding	Non-Coding
		2893421	T -> G	Asn to His	FIG00638480: hypothetical protein
		3265404	T -> G	Non-Coding	Non-Coding
		4562027	T -> G	Non-Coding	Non-Coding
		4751022	A -> C	Thr to Pro	Mg(2+) transport ATPase, P-type (EC 3.6.3.2)
		4873722	A -> C	Gln to His	Transcriptional regulator of D-allose utilization, RpiR family
	Isolate 3	515061	T -> G	Lys to Gln	Outer membrane stress sensor protease DegS
		1554275	A -> C	Ile to Leu	Membrane fusion component of tripartite multidrug resistance system
		1895676	T -> G	Non-Coding	Non-Coding
		2413427	T -> G	Asp to Glu	Efa1-Lymphostatin-like protein
		2661551	A -> C	Thr to Pro	FIG00644273: hypothetical protein
		2774601	A -> C	Thr to Pro	Tellurite resistance protein TehA
		2882401	T -> G	Non-Coding	Non-Coding
		3957782	T -> G	Met to Arg	Putative inner membrane protein
		4491824	T -> G	Glu to Gly	Methyl-accepting chemotaxis protein I (serine chemoreceptor protein)
		4771235	A -> C	Ser to Ala	FIG00638157: hypothetical protein
		252002	G -> A	Non-Coding	Non-Coding

Week 2	Isolate 1	535254	T -> G	Asn to His	Glutamate synthase [NADPH] large chain (EC 1.4.1.13)
		723237	T -> G	Ile to Leu	Modulator of drug activity B
		1026353	T -> G	Asp to Ala	ClpB protein
		1787975	T -> G	Silent	Fructose-specific phosphocarrier protein HPr (EC 2.7.1.69) / PTS system, fructose-specific IIA component (EC 2.7.1.69)
		2869603	A -> C	Ile to Met	chaperone FimC
		3610345	A -> C	Non-Coding	Non-Coding
	Isolate 2	1574831	T -> G	Phe to Leu	Major fimbrial subunit StfA
		2553086	T -> G	Asp to Glu	Respiratory nitrate reductase alpha chain (EC 1.7.99.4)
		2748153	A -> C	Glu to Ala	FIG00638289: hypothetical protein
		4200230	T -> G	Silent	IcmF-related protein
		5125221	A -> C	Gln to His	Aldolase YihT
Week 3	Isolate 1	*168699	A -> C	silent	Xylose isomerase (EC 5.3.1.5)
		*2783439	A -> C	Asn to His	FIG00638106: hypothetical protein
		*2897529	A -> C	Asn to His	LysR family transcriptional regulator YneJ
		2953659	T -> G	Lys to Gln	Fumarate hydratase class II (EC 4.2.1.2)
		*3827309	T -> G	Leu to Trp	DNA-binding heavy metal response regulator
	Isolate 2	2734044	A -> C	Ile to Leu	Phenylacetate-CoA oxygenase, PaaG subunit

Table 3: SNPs identified in each isolate recovered from field lettuce during a repeat of the experiment (year 2) and associated open reading frame information. No isolates were recovered past one week of in field growth

SNPs marked with an asterisk were confirmed via Sanger sequencing

Week	Isolate	Position on reference genome	SNP	Change to open reading frame codon	Gene associated with open reading frame
Immediately following inoculation	Isolate 1	1926400	A -> C	Lys to Gln	Ribosomal RNA large subunit methyltransferase F (EC 2.1.1.51)
		*4380035	A -> C	Thr to Pro	RNA polymerase associated protein RapA (EC 3.6.1.-)
		43277	T -> G	Silent	Putative transport protein
		332430	T -> G	Ile to Met	Putative DNA processing chain A
		*3708013	T -> G	Thr to Pro	Phosphoglucomutase (EC 5.4.2.2)
		*5107319	T -> G	Silent	FIG00638858: hypothetical protein
	Isolate 2	22506	T -> G	Non-Coding	Non-Coding
		3538661	T -> G	Lys to Asn	Hydroxymethylpyrimidine phosphate kinase ThiD (EC 2.7.4.7)
		4462237	A -> C	Glu to Ala	Lipoate-protein ligase A
		4948748	T -> G	Asn to His	YjbF outer membrane lipoprotein
	Isolate 3	742758	T -> G	Phe to Cyc	Biopolymer transport protein ExbD/TolR
		2261123	T -> G	Ile to Leu	putative membrane protein
		2586239	A -> C	Phe to Val	Indole-3-glycerol phosphate synthase (EC 4.1.1.48) / Phosphoribosylanthranilate isomerase (EC 5.3.1.24)
		4929262	A -> C	Glu to Ala	Glycerol-3-phosphate acyltransferase (EC 2.3.1.15)
		5076833	A -> C	Lys to Asn	Glycerol uptake facilitator protein
Week 1	Isolate 1	1220116	A -> C	Asn to His	2-keto-3-deoxy-D-arabino-heptulosonate-7- phosphate synthase I alpha (EC 2.5.1.54)
		2634395	T -> G	Phe to Cys	Phage shock protein B
	Isolate 2	860147	T -> G	Non-Coding	Non-Coding
		1692029	A -> C	Leu to Arg	Polymyxin resistance protein PmrL, sucrose-6 phosphate hydrolase
		3539981	T -> G	Non-Coding	Non-Coding
	Isolate 3	1185319	T -> G	Asp to Ala	FIG00638140: hypothetical protein
		1216423	T -> G	Silent	tRNA (Guanine37-N1) -methyltransferase (EC 2.1.1.31)
		1282737	A -> C	Leu to Val	Uracil-DNA glycosylase, family 1
		1419046	T -> G	Leu to Arg	Putative cytochrome C-type biogenesis protein
		1678295	T -> G	Phe to Val	NADH-ubiquinone oxidoreductase chain N (EC 1.6.5.3)
		2414492	T -> G	Asp to Glu	Efa1-Lymphostatin-like protein
		2502541	T -> G	Val to Gly	Error-prone, lesion bypass DNA polymerase V (UmuC)
		2861265	T -> G	Lys to Asn	Putative formate dehydrogenase oxidoreductase protein
		3237108	T -> G	Non-Coding	Non-Coding
		3379570	A -> C	His to Pro	5-Hydroxyisourate Hydrolase (HIUase) (EC 3.5.2.17)
		4730502	A -> C	Silent	Inorganic pyrophosphatase (EC 3.6.1.1)
		5304083	T -> G	Silent	Xanthine/uracil/thiamine/ascorbate permease family protein

Table 4: SNPs found in more than one O111:NM isolate recovered from field lettuce and associated open reading frame information

SNPs marked with an asterisk were confirmed via Sanger sequencing

	Position on reference genome	SNP	Change to open reading frame codon	Gene associated with open reading frame
SNPs found in all of the following isolates recovered in year one: -immediately after innoculation (isolate 2) -after one week of field growth (isolates 1 and 3) -after 2 weeks (isolates 1 and 2) and -after three weeks of growth (isolate 1)	*5199550	T -> G	Glu to Ala	ATP-dependent DNA helicase UvrD/PcrA
	*3145694	A -> C	Silent	Choline-glycine betaine transporter
	*2093343	T -> A	Silent	Flavodoxin reductases (ferredoxin-NADPH reductases) family 1
	*4835381	A -> C	Glu to Ala	Fumarate respiration sensor kinase protein DcuS
	*2677081	A -> C	Gln to His	Phage major capsid protein
	2931335	T -> G	Phe to Cys	Anaerobic dimethyl sulfoxide reductase chain C (EC 1.8.5.3)
	20338	T -> G	Leu to Val	PTS system, maltose and glucose-specific IIC component (EC 2.7.1.69) / IIB component (EC 2.7.1.69)
	595595	T -> G	Glu to Ala	LppC putative lipoprotein
SNPs found in all of the following isolates recovered in year one: -immediately after innoculation (isolates 1 and 3) -after one week of field growth (isolate 2) and -after three weeks of growth (isolate 2)	4044685	T -> G	Asn to His	Putative flagellin structural protein
	146353	T -> G	Gln to Pro	Transcriptional regulator, LysR family
	3171049	T -> G	Non-Coding	Non-Coding
	775019	A -> C	Ile to Leu	Type III secretion inner membrane protein (YscR,SpaR,HrcR,EscR,homologous to flagellar export components)
	958863	A -> C	Val to Gly	IncI1 plasmid conjugative transfer putative membrane protein PilT
	2171203	A -> C	Lys to Gln	Uptake hydrogenase large subunit (EC 1.12.99.6)
	2179193	A -> C	Glu to Ala	Phosphoanhydride phosphohydrolase (EC 3.1.3.2) (pH 2.5 acid phosphatase) (AP) / 4- phytase (EC 3.1.3.26)
	3182874	A -> C	Leu to Phe	Ribosomal RNA small subunit methyltransferase F (EC 2.1.1.)
	3369799	A -> C	Ile to Met	Putative inner membrane protein
SNPs found in all Year 2 samples	4835381	A -> C	Glu to Ala	Fumarate respiration sensor kinase protein DcuS
	2677081	A -> C	Gln to His	Phage major capsid protein
	2931335	T -> G	Phe to Cys	Anaerobic dimethyl sulfoxide reductase chain C (EC 1.8.5.3)
	*20338	T -> G	Leu to Val	PTS system, maltose and glucose-specific IIC component (EC 2.7.1.69) / IIB component (EC 2.7.1.69)
	*595595	T -> G	Glu to Ala	LppC putative lipoprotein
	*190811	T -> G	Asn to Lys	Phosphoethanolamine transferase specific for the outer Kdo residue of lipopolysaccharide
	*1594062	T -> G	Lys to Thr	Cell division protein
	*1966405	T -> G	STOP to Gln	D-alanyl-D-alanine carboxypeptidase (EC 3.4.16.4)
	*3215999	T -> G	Tyr to Asp	Phage-related protein
	*3671401	T -> G	Silent	FIG00639598: hypothetical protein
	*4261064	T -> G	Ser to Arg	HtrA protease/chaperone protein

4. SNP Comparison of Assembled Reference Genomes to Other Strains

Figures 3, 4, and 5 show the maximum likelihood trees produced by comparing each strain used in the experiment to a selection of strains of the same serotype. The reference assemblies used in this study were used as the reference assembly strain for these SNP analyses.

The O103 strains showed between 9914 and 12310 SNPs between themselves and the reference, with the majority of strains clustering together with between 0 and 225 SNP differences between strains within the cluster. There was a similar number of SNP differences between the cluster, the two outlier O103 strains, and the reference. The O111 strains showed 235 to 251 SNP differences between themselves and the reference strain, with two outliers and two distinct clusters each consisting of strains within 20 SNP differences of each other. The number of SNP differences between the clusters, the outliers and the reference are consistent with the number of SNPs between the reference and each individual strain. The O157:H7 strains showed between 342 and 1547 SNP differences from the reference, with clusters showing a similar distance between them and a distance of less than 100 SNPs defining each cluster.

This comparison shows that the strains used for this study, although lacking the *stx* and *eae* virulence genes, are still representative of their respective serotypes.

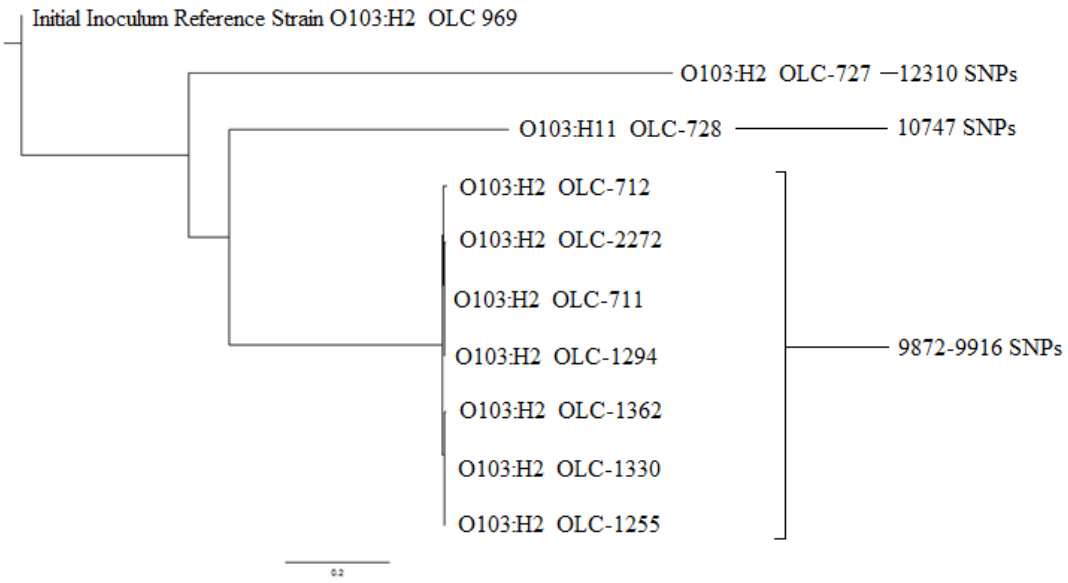

Figure 3: Maximum likelihood phylogenetic tree of *E. coli* O103:H2 strain used to inoculate field lettuce compared to a selection of O103 strains from the CFIA Ottawa Laboratory Carling (OLC) culture collection using SNVPhyl (V1.0.1 Paired_end) with default parameters, including a minimum coverage cut-off of 10, a minimum mean mapping of 30, a SNV abundance ratio of 0.75 and GTR+ γ model as default and tree support values estimated using PhyML's approximate likelihood ratio test (Petkau et al. 2016).

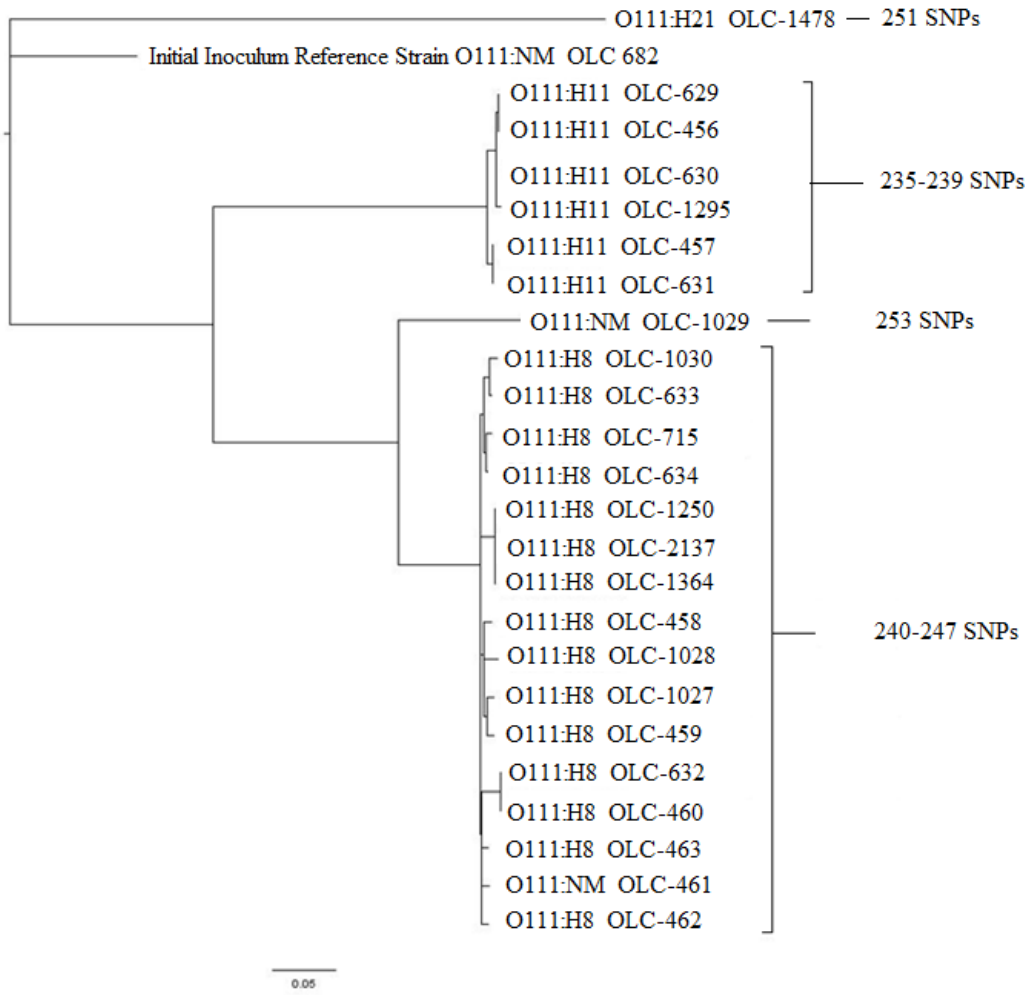

Figure 4: Maximum likelihood phylogenetic tree of an *E. coli* O111:NM strain used to inoculate field lettuce compared to a selection of O111 strains from the CFIA Ottawa Laboratory Carling (OLC) culture collection using SNVPhyl (V1.0.1 Paired_end) with default parameters, including a minimum coverage cut-off of 10, a minimum mean mapping of 30, a SNV abundance ratio of 0.75 and GTR+ γ model as default and tree support values estimated using PhyML's approximate likelihood ratio test (Petkau et al. 2016).

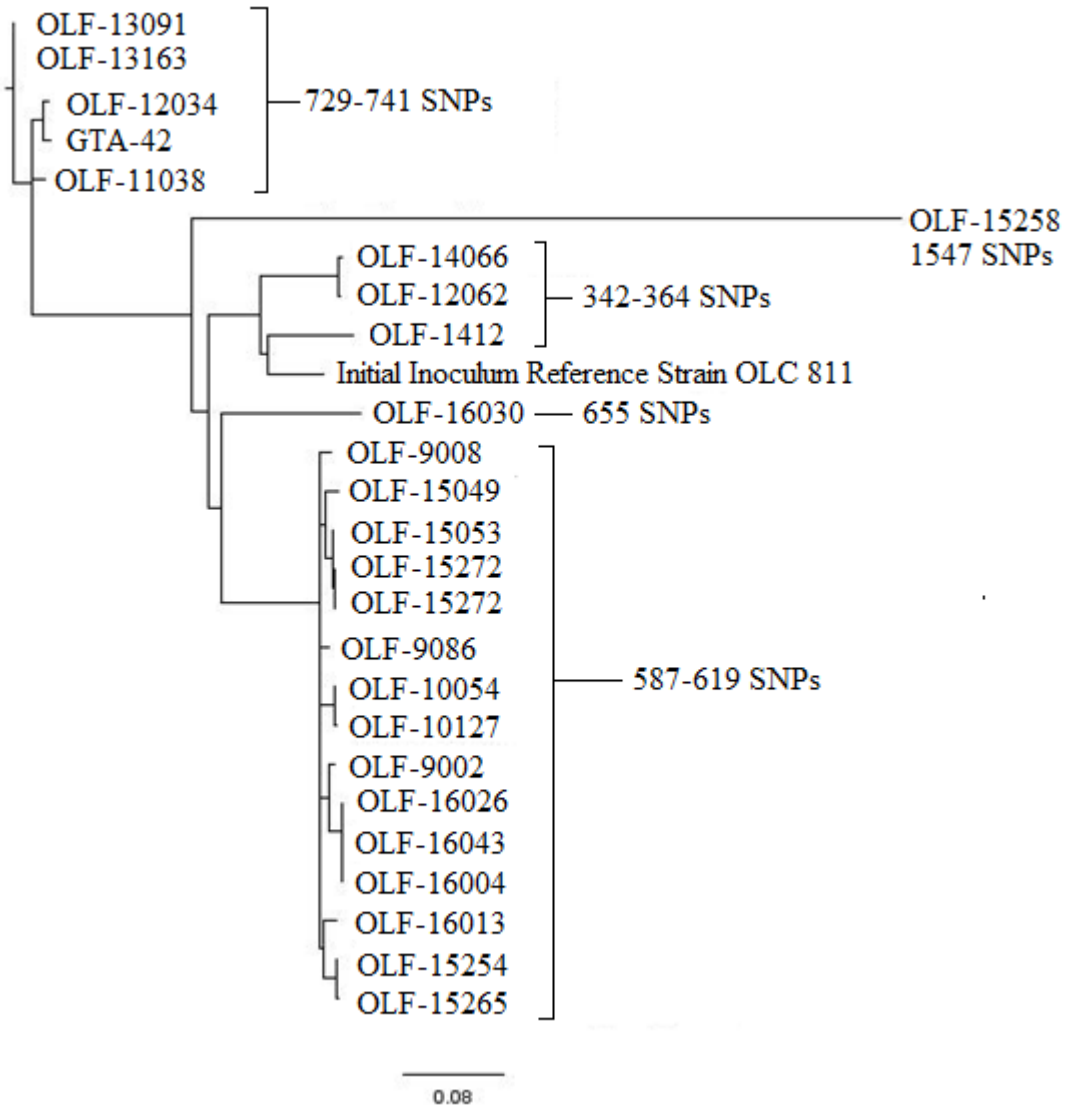

Figure 5: Maximum likelihood phylogenetic tree of an *E. coli* O157:H7 strain used to inoculate field lettuce compared to a selection of O157:H7 strains from the CFIA Ottawa Laboratory Carling (OLC) and Ottawa Laboratory Fallowfield (OLF) culture collections using SNVPhyl (V1.0.1 Paired_end) with default parameters, including a minimum coverage cut-off of 10, a minimum mean mapping of 30, a SNV abundance ratio of 0.75 and GTR+ γ model as default and tree support values estimated using PhyML's approximate likelihood ratio test (Petkau et al. 2016).

IV. Discussion

1. Recovery of Isolates

The selective enrichment protocol used to recover isolates from the lettuce plants in this experiment was derived from the MFLP-52, a method from the CFIA's compendium of analytical methods (Government of Canada, H.C. 1999). This is a standard method used by the CFIA to isolate EHEC from food sources, and as such accurately reflects the effects of isolation methods used during an actual outbreak scenario. This protocol saw mixed success. No isolates were recovered past three weeks of field growth, and recovery within the three weeks was unreliable (see **Figure 2**). There are several possible causes for this. The strains may have failed to colonize the lettuce initially, or may have died out under the stresses associated with the environment. These particular strains may utilize the roots or earth around the plants as their main reservoir, which were not included in the samples harvested from the field (Ibekwe et al. 2004). It is also possible that they were heavily out competed to the point where they were not recoverable, which given the high level of colony morphology diversity of non-target colonies on the differential plates is entirely plausible. Initial inoculum levels may also influence *E. coli* survivability in field lettuce (McKellar et al. 2014).

The soil and root system was not included in the analysis as industrial lettuce production procedures do not directly sample the soil and the experiment was designed to mimic a lettuce production scenario (Delaquis et al. 2002). Future experiments may look at SNP accumulation in bacteria located in the soil surrounding the plants, especially given it has been shown to be a reservoir for *E. coli* (Ibekwe et al. 2004). It is possible

that splash back from contaminated soil would explain why no O157:H7 isolates were recovered after one week, but several were recovered after two and three weeks of growth as shown in **Figure 2**. It is also possible that the bacteria died off to an unrecoverable level before building a population back up. Further experiments would need to be conducted to confirm either of these possible explanations.

Difficulty recovering inoculated bacteria from field lettuce has been well documented (Bezanson et al. 2012; Schuenzel et al. 2002; Johannessen et al. 2005). It has been shown that O157:H7 cells decrease over 1000-fold almost immediately after inoculation (Moyne et al. 2013). This may also help explain the dearth of studies of this type; lab environment studies are far more reliable, at the cost of accurate representation of the growing conditions.

2. Effects of Selective Enrichment on SNP Accumulation

It is possible that the selective nature of the recovery protocol used for the experiment had an effect on SNP accumulation in the recovered strains. The selective pressure applied during the enrichment may have selected for or against SNPs that were beneficial or deleterious in the enrichment broth environment (Liamkaew et al. 2012). This would effectively fix SNPs that would not remain in the population under field conditions, or remove SNPs that would otherwise be fixed. It is also possible that the growth conditions during the recovery protocol allowed for larger population sizes and faster generation times than the in-field growth conditions. This may have resulted in more opportunity for SNP mutation and accumulation during the recovery protocol than during the in-field growth. . It is also possible that selecting a single isolated colony prior

to sequencing acted as a genetic bottleneck, effectively fixing SNPs in that isolate are not representative of the population as a whole. This would help explain why a similar number of SNPs were identified in the O111:NM isolates obtained immediately after inoculation and those recovered over the three week growth period.

Any typing performed to characterize an outbreak or attempt to locate a source for an outbreak would require isolation of the causative strain by enrichment methods very similar to those used in this study. Thus, any effect on SNP accumulation caused by the recovery protocol here mimics the effects expected during an outbreak investigation, and increases the direct applicability of conclusions drawn from this experiment to regulatory decisions.

3. The Importance of Mimicking Real World Conditions

Due to its common association with EHEC outbreaks lettuce production was chosen for this study as a preliminary look at SNP accumulation rates under food production relevant conditions (Warriner et al. 2009; Lynch et al. 2009; Nicholson et al. 2005; Cooley et al. 2007). While any food production relevant environment associated with foodborne illness outbreaks might have been chosen as a representative model, field lettuce in particular was chosen due to the mutagenic effects of UV sunlight, stiff selective pressure from the harsh environment and high levels of resource competition between the inoculated strains and native bacteria found in the field-grown lettuce environment (Bezanson et al. 2012). Field lettuce represents one of the harshest environments a wild population of bacteria can inhabit, within the context of food production. In addition to using a field relevant environment for growth the method used

to inoculate the field lettuce isolates was chosen to replicate a contamination event, such as irrigation or flooding of a field with a contaminated water source (Lynch et al. 2009).

It has been shown that factors such as starvation conditions, competition for resources, availability of carbon sources and general bacterial stress induce a significantly greater rate of mutation in *E. coli* and other bacteria (Jacobs et al. 1997; Shapiro et al. 1997; Hughes et al. 1997; Shapiro et al. 1984). Adaptive mutation may also play a part in changes to mutation rates in stressed bacteria, and has been demonstrated in *E. coli* growing under stressed conditions (Foster et al. 1990; Foster 2004; Bihan et al. 2015). *rpoS* in particular, a gene regulating error-prone repair DNA polymerase IV, is upregulated in bacteria under population pressure and can lead to an increase in mutation rate (Layton & Foster 2003). The temperature variations, competition, and resource scarcity present in field lettuce will all effect the doubling time of the bacteria, which in turn will affect SNP accumulation in the population over time as a slower doubling time implies less generations over time, and therefore less opportunity to develop SNPs via replication errors. Lower population size and turn over would have similar effects in addition to introducing genetic bottleneck and genetic drift effects which can also play an important role in mutation fixation rates. Genetic drift can fix potentially deleterious genes where they might be removed in a larger population due to selective forces (Kuo et al. 2009).

These environmental factors can also increase selective pressures for fixing beneficial mutations and removing deleterious ones, further affecting mutation accumulation rates. It has been shown that bacteria are biased towards using more efficient codons for highly transcribed genes. As such, even SNPs causing silent

mutations may be under positive selective pressure as they may improve the efficiency of protein synthesis (Ochman, 2003).

Accounting for all of these factors affecting SNP accumulation rates and their complex interactions in a laboratory setting is not feasible. Thus, the best way to accurately reflect their influences is to simply carry out the experiment under field relevant conditions, eliminating the need to artificially replicate the effects of the environment.

In addition to choosing growing conditions relevant to food production, strains bearing close relation to outbreak associated strains were chosen as model organisms. It has been shown that pathogenic bacteria have higher mutation rates than their non-pathogenic counterparts (LeClerc et al. 1996; Falush et al. 2006). This likely increases adaptation to changing host defenses, but also complicates reference studies of this type, as mutation rates can vary greatly between strains and species. To ensure an accurate representation of expected mutation rate, bacteria as similar as possible to actual pathogenic strains should be used to model mutation rates. To this end bacteria sharing serotypes commonly associated with human pathogenicity were used.

The O103:H2 and O111:NM strains are wild type isolates originally isolated by the CFIA and PHAC respectively, while the O157:H7 strain is a naladixic acid derivative of American Type Culture Collection (ATCC) 700728 (also known as NCTC 12900). ATCC 700728 was originally isolated by Australia's Federal Public Health organization (FPH). The serotypes chosen all belong to the seven priority EHEC serotypes described by FSIS (USDA-FSIS, 2011). O157:H7 in particular is closely associated with severe

human illness, and is often tested for independently of other serotypes by food regulation governing bodies (Karmali et al. 2003). The naladixic acid resistant derivative was used instead of the wild type to aid in the selective enrichment recovery of the bacteria inoculum from the lettuce (Bezanson et al. 2012).

Unfortunately, nalidixic acid resistance was not available for the other strains, and inducing the resistance would potentially introduce more genetic difference between the strains used and the wild type isolates, contrary to the purpose of the experiment. All three strains were missing the *eae* and *stx* genes that code for the intimin and Shiga-like toxins that make the strains enterohemorrhagic, but are otherwise genetically as similar as possible to the pathogenic strains sharing their serotypes. This ensures an accurate representation of mutation rates for these strains to mimic what could be expected in an outbreak scenario. **Figures 3, 4, and 5** show the maximum likelihood trees of each strain compared to a representative selection of strains showing the same O-antigen. The reference strains show a similar number of SNP differences to each queried strain as the strains all show between themselves. This shows that although the strains used were apathogenic, they are still representative of these serotypes. Note that while using the actual pathogenic strains would be ideal, the ethical ramifications of inoculating field lettuce with verotoxigenic EHEC of potential human pathogenicity prevents it from being practical.

4. Reference Assemblies

As most SNP analysis programs are reference based they require an accurate assembled genome to ensure reliable results. To create an appropriate reference for this

study two different sequencing technologies were used; Pacific Biosciences Single Molecule Real-Time (SMRT) sequencing (commonly referred to as PacBio sequencing) and Illumina MiSeq paired end sequencing. Sequencing of the DNA was performed in duplicate via both methods to ensure adequate coverage. Illumina MiSeq paired end sequencing was also used to obtain high quality reads for all isolates recovered from lettuce during the course of the experiment.

Illumina MiSeq is relatively fast and inexpensive when compared to other available technologies. The platform delivers a high depth of coverage for the majority of the genome, and delivers one of the lowest read error rates available. However MiSeq sequencing can introduce systematic errors, especially downstream of and within repeat and low complexity regions. Short read lengths make resolution of these low complexity and highly repetitive regions very difficult, which can result in large gaps in the genome of assemblies built using only MiSeq data. Proper trimming of raw reads is also crucial due to MiSeq's lower fidelity near the start and end of reads (Quail et al. 2012; Laehnemann et al. 2016).

PacBio SMRT sequencing is more costly than MiSeq in both time and resources, but has significantly longer read lengths. While it is far more error prone than illumina MiSeq, its errors are randomly spread throughout the genome and thus easier to correct for with a higher depth of coverage. Its longer read lengths allow it to bridge gaps left in MiSeq assemblies, and are especially helpful in resolving low complexity and highly repetitive regions. (Quail et al. 2012).

High quality reference assemblies were produced using the PacBio reads as assembled by smrtanalysis incorporating BLASR for long read error correction, Celera for assembly and Quiver for read realignment and final base calling. The PacBio sequence data provided an excellent scaffold, covering regions difficult to assemble using short read based sequencing technology like MiSeq. Reference strains were incubated for four hours and DNA extraction was performed while the samples were still in log exponential growth phase. The variation in depth of coverage across each genome was correlated with proximity to the origin of replication, allowing the contigs in the assemblies to be rearranged based on depth of coverage (i.e. proximity to the origin of replication). This vastly improved scaffolding. Further assembly efforts using this ordered contig scaffold provided closed reference genomes of all three strains. The Illumina MiSeq reads were then used for error correction, providing high depth of coverage and high fidelity reads to correct for PacBio's error prone reads and to confirm homopolymer lengths where required.

The resulting reference genome assemblies showed coverages of 401x, 499x, and 469x for the O157:H7, O111:NM and O103:H2 strains respectively. Genomes were 5.54, 5.47, and 5.21 million base pairs, with GC contents of 50.55%, 50.66%, and 50.55%.

One of the limitations of mixing both long and short read sequencing technologies for reference creation and only short read sequencing for the isolate SNP analysis is that difficult to resolve regions of the genome that are assembled using long read technology reads may invite incorrect alignments by SNP calling programs that are only aligning short reads. This is a difficult issue to address, as long read sequence technology would need to be coupled to short reads to ensure fidelity in field isolates, and currently no SNP

analysis programs are optimized for this sort of analysis. In the context of outbreak isolate genotyping it is not necessary to correct for this issue as it should be systematic, affecting all isolates equally relative to the reference. Thus, it should not influence the analysis. It does serve to again underline the idea that SNP analysis results, much like all epidemiological and genotyping results, must be interpreted in context and in conjunction with all other available information

While making a very high quality reference such as those used for this experiment may not be feasible during an outbreak analysis, a reasonable compromise could be made. Excluding regions of low complexity or high repetition from the analysis will decrease the likelihood of false positive results. Using a MiSeq data only assembly may effectively accomplish this already, as these regions are mostly omitted from MiSeq only assemblies already as inter-contig gaps, where the assembly programs fail to resolve the genome sequence. As these omitted regions are usually difficult for alignment software, it is unlikely that any high-quality SNPs would be found there in any event. The SNVPhyl pipeline has already taken steps to address this, as it allows users to incorporate a masking file specifically to address this issue (Petkau et al. 2016). While a masking file was not included in this experiment, one incorporating known phage and repetitive sequences could be made to reduce the likelihood of false SNP calls.

5. O157:H7 and O103:H2 SNPs

The O157 and O103 strains showed almost no SNP accumulation over the duration of this experiment. A single SNP was identified in an O103:H2 isolate recovered after two weeks of in-field growth. No other SNPs were identified in any of the O157:H7

or O103:H2 isolates recovered from the lettuce. This implies that the basal rate of accumulation of SNPs in these specific strains under these conditions is extremely low over the timeline of an outbreak scenario. The SNP identified in the O103:H2 isolate was in a region coding for an uncharacterized host specific phage related protein, and coded for a silent mutation. It has been noted in past studies that even *E. coli* O157:H7 isolates that show distinct PFGE patterns may have extremely similar genomes, and show little variation (Noller et al. 2003). Along with the results of this study, the implication is that a number of SNP differences between closely related O157:H7 or O103:H2 isolates is a strong indication that the isolates are not epidemiologically related.

6. O111:NM SNPs

Tables 2 and 3 show the SNPs identified in the O111 strain isolates recovered during the experiment. Far more SNPs were identified in this strain relative to the other two, but overall very few SNPs per isolate were identified. Both the mode and average number of SNPs per isolate was 5.

In addition to the SNPs identified in each isolate, there were 8 SNPs identified in 6 of the year one isolates, and 9 SNPs identified in the other 4 year one isolates. These SNPs divide the year one isolates into two distinct genotypes, henceforth referred to as genotypes 1 and 2, respectively. Five of the eight SNPs that define genotype 1 in year one were also identified in every year two isolate, along with 6 additional SNPs. This third genotype is assumed to have arisen from the same base genotype as the year one genotype 1. Although it is possible that these SNPs accumulated separately in the two populations due to positive selective pressure, it is unlikely (Fong et al. 2005). One of the

SNPs occurred in a non-coding region, and several SNPs coded for synonymous or near synonymous changes. The most likely explanation for these SNPs is that a mixed culture was used for the initial inoculation in year one, giving rise to genotypes 1 and 2. In the second year either a pure culture consisting of only the genotype giving rise to genotype 1 in year one was used to inoculate, or the genotype 2 strain was outcompeted in year two.

The total number of SNPs in each isolate appears to be unassociated with the amount of time the recovered isolate spent growing in field conditions. There are a few possible explanations for this. The variation in SNP accumulation over this short of a time period might be too high to surmise a pattern given the limited number of samples recovered during this study. The accumulation of SNPs in this strain might happen primarily during the pre-inoculation growth of the bacteria or during the recovery of the strains from the field and not during the period the strains spent growing in the field. As discussed above, it is also possible that SNPs that did develop during the in-field growth were selected against during the selective enrichment isolation process, and thus only SNPs that were not deleterious to strains growing in the isolation media were identified in the isolated strains (Liamkaew et al. 2012). Growth during the recovery protocol may have allowed for more opportunity for SNP mutation and accumulation, as it may have allowed for higher population sizes and more generations compared to the in-field growth.

While it appears that no clear relationship between the time spent growing in field and the number of accumulated SNPs can be established over this timeline and sample size, the timeline used for this study is similar to that which would be expected during a

food safety investigation (WHO, 2001). As such the number of SNPs identified in the isolates in this experiment is representative of what would be expected in epidemiologically related isolates of this strain over the course of a food safety investigation. Assuming future studies support the conclusion that the rate of SNP accumulation over a foodborne illness relevant timeline is not as strongly effected by the time spent in field compared to the basal rate of SNP mutation and accumulation inherent to the strain, it may be that further studies of this nature looking at other strains of *E. coli* will only need to be performed over a single week of growth. This would greatly reduce the financial and time cost associated with this type of research.

There are a few different explanations for the greater accumulation of SNPs in the O111:NM strain than the other two strains. It is possible that the O111:NM strain contains mutator alleles such as *rpoS* that cause an increase in mutation rate or an increased SOS response (Turrientes et al. 2013; Janion 2008). This may be supported by the high transversion to transition rate observed in the SNPs identified in the O111:NM strain. While it has been shown that transition mutations are far more common than transversion mutations, this ratio appears to be somewhat higher than expected (Lee et al. 2012). Changes to transversion/transition rates have been associated with changes to DNA repair associated *mut* genes such as *mutS* (Zhao and Winkler 2000). A change in one of these genes could explain the increased number of SNPs found in this strain compared to the other 2 strains.

It is also possible that these strains are at different points on the fitness landscape, and that the O111:NM strain is less adapted to the growth conditions than the other two strains. A greater level of mutation accumulation would be expected in this case as more

as yet unrealized beneficial mutations may be available to an strain further from a fitness peak relative to an adapted strain, increasing the accumulation of SNPs due to selective pressure (Galhardo et al. 2012). An un-adapted strain may also be under greater cell stress than an adapted one, increasing the rate of SNP mutation (Janion 2008).

Further research could possibly test these theories. Comparing the relative rates of synonymous to non-synonymous mutation may shed identify if a strain is near the bottom or top of its fitness landscape, as non-synonymous mutations are more common than synonymous mutations in cells that less adapted to their surroundings and under selective pressures (Kryazhimskiy & Plotkin 2008; Ochman et al. 2003). The use of fluctuation assays could ascertain if the O111:NM strain shows a higher rate of mutation relative to the other strains (Ochman et al. 2003; Foster 2007).

7. The Importance of Context and a Call for Further Research

This pilot study indicates that it would not be expected to find more than an average of 5 SNP differences between epidemiologically related *E. coli* isolates. The limitation of this study, however, is the relatively low number of replicates and the fact that only three different strains of bacteria were used. It has been shown that the rate of mutation of bacteria can vary greatly, as in the case of so called super mutator variants. Upregulation of genes such as *rpoS* and *mut*S in particular have been shown to greatly increase mutation rate (Turrientes et al. 2013; Shaver & Sniegowski 2003). Contamination events may also include mixed cultures of strains showing a great diversity of SNP mutations. In these cases a much greater number of SNPs might be

expected between epidemiological related isolates during a foodborne illness outbreak than indicated by this study.

Even within this study's limited selection of strains, substantially more SNPs were identified in the O111:NM strain than the other two strains, indicating that it has a higher basal rate of SNP accumulation under these conditions. While the total number of SNPs isolated from the recovered O111:NM isolates remained low, the difference between the numbers of SNPs identified in this strain compared to the other two strains highlights the fact that basal rates of SNP accumulation vary even between the fairly closely related strains used in this study. This has important implications for future SNP based applications.

The existence of these factors highlights two major points. One is that more research of this type would be greatly beneficial. A large body of supporting evidence would lend a much higher credibility to the results of SNP analyses such as these, especially in cases where they provide greater resolution than and contradict traditional method results such as PFGE and MLST. Further research would also identify strains that are particularly susceptible to outlier rates of SNP accumulation. As of this date, this body of evidence is not available. As more data becomes available, a weight of evidence will lend support to regulatory decisions based on SNP analysis; this study is but one of many that must be completed to obtain a broad view of what to expect between epidemiologically related isolates. Further research may focus on using a greater variety of strains and sample matrices.

The second point is one that is already very familiar to regulators and epidemiologists. The results of methods such as SNP analysis must be interpreted within the context of the investigation in which they are utilized. If the strain suspected to be the causative agent is highly clonal and two isolates show many SNP differences, it is likely that they are unrelated. Finding the same number of SNP differences between isolates that have been shown to be fast mutating variants or of a genotype known to have highly variable genomes would indicate they are epidemiologically unrelated. SNP data needs to be analysed holistically with all other available typing and epidemiological data to ensure it is utilised accurately and effectively. The finer resolution offered by SNP analysis simply provides another tool to help accurately interpret outbreak investigative data.

While there is pressure from regulatory bodies to identify a set statistical number of SNP differences to expect between epidemiologically related isolates, it is becoming more apparent as these tools are seeing an increase in use during outbreak scenarios that this number does not exist (Leekitcharoenphon et al. 2014). The number of SNPs to expect between isolates depend on the context in which they are isolated and the nature of the strain you are investigating.

With these facts in mind, this study has shown that for these strains under these conditions a low number of SNP differences can be expected between isolates related over an epidemiological timeline, as outlined in the hypothesis. It has also shown that food production relevant conditions and foodborne illness outbreak investigation isolation procedures do not appear to have a drastic effect on the rate of SNP accumulation.

8. An Example Illustrating the Application of This Research

In a hypothetical situation, 50 O157:H7 EHEC clinical isolates have been identified over a two-week period. 49 of the strains show identical PFGE profiles, while one shows a very similar profile. SNP analysis shows that of the 50 strains, 30 have no SNP differences compared to an isolate identified in a potential food contamination source. 7 have 1 SNP difference, 3 have 20-30 SNP differences, and the remaining 10 have over 150 SNP differences (including the one displaying a different PFGE profile). It would be exceedingly to accurately cluster these isolates without higher discrimination, which SNP analysis provides. When creating clusters incorporating the SNP analysis, how should it be interpreted? According to the results of this study, the possible source contamination is almost definitely not related to the >150 SNP isolates, very unlikely to be related to the >20 SNP isolates, but quite possibly related to the 1 or less SNP isolates. The investigators can now look at epidemiological information from only 37 sources instead of 50, greatly facilitating and simplifying their task. While this example may seem contrived, situations such as this are common in identification of foodborne illness outbreaks (Lienau et al. 2008; Roetzer et al. 2013; Kudva et al. 2002; Allard et al. 2013)

V. Conclusions:

When conducting studies of this type it is important that the parameters used are as close to the real-world conditions as feasible. This ensures that the results of the study are accurate and transferable to applications outside of the lab. While the extra effort and cost associated with studies of this type can be onerous, it is the duty of food regulators to ensure their methods and references are reliable and accurate to ensure public health and safety and reduce financial impact to industry where ever possible. That being said, the preliminary results garnered by this study indicate that very few SNPs accrue in these serotypes under field relevant conditions. If further studies confirm these results and indicate that the effects of food production relevant selective pressures do not greatly affect SNP accumulation, it may be that further studies can be performed in lab at a much lower cost and difficulty. This would greatly facilitate the gathering of accurate data to map the basal rate of SNP formation in various strains and species of foodborne illness related pathogens. The variation in SNP accumulation rates observed between the three strains used in this experiment serves to highlight the fact that more studies of this type would be highly beneficial.

Based on the low SNP accumulation rates observed during this experiment, it appears that these serotypes would not be expected to display many SNPs between epidemiologically related isolates. This is a key finding, as interpreting SNP analysis data remains challenging when the number of SNPs between isolates is below 100. This study provides data to support decisions based on SNP differences numbering in this range. As with all typing methods, any guidelines elucidated by these studies as to the basal rate of

SNP accumulation would have to be used in concert with common sense and alongside other typing methods.

VI. References:

Allard, M.W., Luo, Y., Strain, E., Pettengill, J., Timme, R., Wang, C., Li, C., Keys, C.E., Zheng, J., Stones, R., et al. (2013). On the Evolutionary History, Population Genetics and Diversity among Isolates of Salmonella Enteritidis PFGE Pattern JEGX01.0004. PLOS ONE *8*, e55254.

den Bakker, H.C., Didelot, X., Fortes, E.D., Nightingale, K.K., and Wiedmann, M. (2008). Lineage specific recombination rates and microevolution in Listeria monocytogenes. BMC Evol Biol *8*, 277.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol *19*, 455–477.

Bates, H., Randall, S.K., Rayssiguier, C., Bridges, B.A., Goodman, M.F., and Radman, M. (1989). Spontaneous and UV-induced mutations in Escherichia coli K-12 strains with altered or absent DNA polymerase I. J Bacteriol *171*, 2480–2484.

Bezanson, G., Delaquis, P., Bach, S., McKellar, R., Topp, E., Gill, A., Blais, B., and Gilmour, M. (2012). Comparative examination of Escherichia coli O157:H7 survival on romaine lettuce and in soil at two independent experimental sites. J Food Prot *75*, 480–487.

Blais, B., Deschênes, M., Huszczynski, G., and Gauthier, M. (2014). Enterohemorrhagic Escherichia coli colony check assay for the identification of serogroups O26, O45, O103, O111, O121, O145, and O157 colonies isolated on plating media. J Food Prot *77*, 1212–1218.

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

Bryant, J., Chewapreecha, C., and Bentley, S.D. (2012). Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences. Future Microbiol *7*, 1283–1296.

Bushnell B. (2016). BBTools User Guide. Available at http://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brüssow, H. (2003). Prophage Genomics. Microbiol Mol Biol Rev *67*, 238–276.

Center for Disease Control and Prevention (2016). PulseNet Methods: Whole Genome Sequencing (WGS). Available at https://www.cdc.gov/pulsenet/pathogens/wgs.html

Cooley, M., Carychao, D., Crawford-Miksza, L., Jay, M.T., Myers, C., Rose, C., Keys, C., Farrar, J., and Mandrell, R.E. (2007). Incidence and Tracking of Escherichia coli O157:H7 in a Major Produce Production Region in California. PLOS ONE *2*, e1159.

Corrigan, J.J., and Boineau, F.G. (2001). Hemolytic-uremic syndrome. Pediatr Rev *22*, 365–369.

Cupples, C.G., Cabrera, M., Cruz, C., and Miller, J.H. (1990). A Set of Lacz Mutations in Escherichia Coli That Allow Rapid Detection of Specific Frameshift Mutations. Genetics *125*, 275–280.

Davis, S., Pettengill, J.B., Luo, Y., Payne, J., Shpuntoff, A., Rand, H., and Strain, E. (2015). CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. J Comput Sci *1*, e20.

Del Fabbro, C., Scalabrin, S., Morgante, M., and Giorgi, F.M. (2013). An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. PLoS One *8*.

Delaquis, S., Stewart, S., Cazaux, S., and Toivonen, P. (2002). Survival and growth of Listeria monocytogenes and Escherichia coli O157:H7 in ready-to-eat iceberg lettuce washed in warm chlorinated water. J Food Prot *65*, 459–464.

Deng, X., Phillippy, A.M., Li, Z., Salzberg, S.L., and Zhang, W. (2010). Probing the pan-genome of Listeria monocytogenes: new insights into intraspecific niche expansion and genomic diversification. BMC Genomics *11*, 500.

Dettman, J.R., Rodrigue, N., Melnyk, A.H., Wong, A., Bailey, S.F., and Kassen, R. (2012). Evolutionary insight from whole-genome sequencing of experimentally evolved microbes. Mol Ecology *21*, 2058–2077.

Donnenberg, M.S., Tzipori, S., McKee, M.L., O'Brien, A.D., Alroy, J., and Kaper, J.B. (1993). The role of the eae gene of enterohemorrhagic Escherichia coli in intimate attachment in vitro and in a porcine model. J Clin Invest *92*, 1418–1424.

Dunn, J.R. (2016). Whole-Genome Sequencing: Opportunities and Challenges for Public Health, Food-borne Outbreak Investigations, and the Global Food Supply. J Infect Dis *213*, 499–501.

Fijalkowska, I.J., Schaaper, R.M., and Jonczyk, P. (2012). DNA replication fidelity in Escherichia coli: a multi-DNA polymerase affair. FEMS Microbiol Rev *36*, 1105–1121.

Foley, I., Marsh, P., Wellington, E.M., Smith, A.W., and Brown, M.R. (1999). General stress response master regulator *rpoS* is expressed in human infection: a possible role in chronicity. J. Antimicrob Chemother *43*, 164–165.

Foley, S.L., Zhao, S., and Walker, R.D. (2007). Comparison of molecular typing methods for the differentiation of Salmonella foodborne pathogens. Foodborne Pathog Dis *4*, 253–276.

Fong, S.S., Joyce, A.R., and Palsson, B.Ø. (2005). Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. Genome Res *15*, 1365–1372.

Foster, P.L. (1993). Adaptive mutation: the uses of adversity. Annu Rev Microbiol *47*, 467–504.

Foster, P.L. (2000). Adaptive Mutation in Escherichia coli. Cold Spring Harb Symp Quant Biol *65*, 21–29.

Foster, P.L. (2006). Methods for Determining Spontaneous Mutation Rates. Methods Enzymol *409*, 195–213.

Galhardo, R.S., Hastings, P.J., and Rosenberg, S.M. (2007). Mutation as a Stress Response and the Regulation of Evolvability. Crit Rev Biochem Mol Biol *42*, 399–435

Gardner, S.N., Slezak, T., and Hall, B.G. (2015). kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics *31*, 2877–2878.

Gill, A., Martinez-Perez, A., McIlwham, S., and Blais, B. (2012). Development of a method for the detection of verotoxin-producing Escherichia coli in food. J Food Prot *75*, 827–837.

Gilmour, M.W., Graham, M., Van Domselaar, G., Tyler, S., Kent, H., Trout-Yakel, K.M., Larios, O., Allen, V., Lee, B., and Nadon, C. (2010). High-throughput genome sequencing of two Listeria monocytogenes clinical isolates during a large foodborne outbreak. BMC Genomics *11*, 120.

Gilmour, M.W., Graham, M., Reimer, A., and Domselaar, G.V. (2013). Public Health Genomics and the New Molecular Epidemiology of Bacterial Pathogens. PHG *16*, 25–30.

Government of Canada, H.C. (1999). The Compendium of Analytical Methods: Microbiol Methods - Health Canada.

Grim, C.J., Kotewicz, M.L., Power, K.A., Gopinath, G., Franco, A.A., Jarvis, K.G., Yan, Q.Q., Jackson, S.A., Sathyamoorthy, V., Hu, L., et al. (2013). Pan-genome analysis of the emerging foodborne pathogen Cronobacterspp. suggests a species-level bidirectional divergence driven by niche adaptation. BMC Genomics *14*, 366.

Horby, P.W., O'Brien, S.J., Adak, G.K., Graham, C., Hawker, J.I., Hunter, P., Lane, C., Lawson, A.J., Mitchell, R.T., Reacher, M.H., et al. (2003). A national outbreak of multi-resistant Salmonella enterica serovar Typhimurium definitive phage type (DT) 104 associated with consumption of lettuce. Epidemiol Infect *130*, 169–178.

Hughes, D., and Andersson, D.I. (1997). Carbon starvation of Salmonella typhimurium does not cause a general increase of mutation rates. J Bacteriol *179*, 6688–6691.

Hunter, S.B., Vauterin, P., Lambert-Fair, M.A., Van Duyne, M.S., Kubota, K., Graves, L., Wrigley, D., Barrett, T., and Ribot, E. (2005). Establishment of a Universal Size Standard Strain for Use with the PulseNet Standardized Pulsed-Field Gel Electrophoresis Protocols: Converting the National Databases to the New Size Standard. J Clin Microbiol *43*, 1045–1050.

Huszczynski, G., Gauthier, M., Mohajer, S., Gill, A., and Blais, B. (2013). Method for the detection of priority Shiga toxin-producing Escherichia coli in beef trim. J Food Prot *76*, 1689–1696.

Ibekwe, A.M., Watt, P.M., Shouse, P.J., and Grieve, C.M. (2004). Fate of Escherichia coli O157:H7 in irrigation water on soils and plants as validated by culture method and real-time PCR. Can J Microbiol *50*, 1007–1014.

ISO (2012). ISO/TS 13136:2012 Microbiology of Food and Animal Feed–Real-Time Polymerase Chain Reaction (PCR)-Based Method for the Detection of Food-Borne Pathogens–Horizontal Method for the Detection of Shiga Toxin-Producing Escherichia coli (STEC) and the determination of O157 O111 O26 O103 and O145 Serogroups. Available at: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=53328

Jacobs, K.L., and Grogan, D.W. (1997). Rates of spontaneous mutation in an archaeon from geothermal environments. J Bacteriol *179*, 3298–3303.

Janion, C. (2008). Inducible SOS Response System of DNA Repair and Mutagenesis in Escherichia coli. Int J Biol Sci *4*, 338–344.

Johannessen, G.S., Bengtsson, G.B., Heier, B.T., Bredholt, S., Wasteson, Y., and Rørvik, L.M. (2005). Potential uptake of Escherichia coli O157:H7 from organic manure into crisphead lettuce. Appl Environ Microbiol *71*, 2221–2225.

Jolley, K.A., Bliss, C.M., Bennett, J.S., Bratcher, H.B., Brehony, C., Colles, F.M., Wimalarathna, H., Harrison, O.B., Sheppard, S.K., Cody, A.J., et al. (2012). Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiol (Reading, Engl.) *158*, 1005–1015.

Kalchayanand, N., Arthur, T.M., Bosilevac, J.M., Wells, J.E., and Wheeler, T.L. (2013). Chromogenic agar medium for detection and isolation of Escherichia coli serogroups O26, O45, O103, O111, O121, and O145 from fresh beef and cattle feces. J Food Prot *76*, 192–199.

Kaper, J.B., Nataro, J.P., and Mobley, H.L. (2004). Pathogenic Escherichia coli. Nat. Rev Microbiol *2*, 123–140.

Karmali, M.A., Mascarenhas, M., Shen, S., Ziebell, K., Johnson, S., Reid-Smith, R., Isaac-Renton, J., Clark, C., Rahn, K., and Kaper, J.B. (2003). Association of genomic O island 122 of Escherichia coli EDL 933 with verocytotoxin-producing Escherichia coli

seropathotypes that are linked to epidemic and/or serious disease. J Clin Microbiol *41*, 4930–4940.

Kelley, D.R., Schatz, M.C., and Salzberg, S.L. (2010). Quake: quality-aware detection and correction of sequencing errors. Genome Biol *11*, R116.

Kim, K.-W., Burt, D.W., Kim, H., and Cho, S. (2013). Identification of Differentially Evolved Genes: An Alternative Approach to Detection of Accelerated Molecular Evolution from Genome-Wide Comparative Data. Evol Bioinformatics *2013*, 285–299.

Knowles, M., Stinson, S., Lambert, D., Carrillo, C., Koziol, A., Gauthier, M., and Blais, B. (2016). Genomic Tools for Customized Recovery and Detection of Foodborne Shiga Toxigenic Escherichia coli. J Food Prot *79*, 2066–2077.

Koressaar, T., and Remm, M. (2007). Enhancements and modifications of primer design program Primer3. Bioinformatics *23*, 1289–1291.

Kozyreva, V.K., Crandall, J., Sabol, A., Poe, A., Zhang, P., Concepción-Acevedo, J., Schroeder, M.N., Wagner, D., Higa, J., Trees, E., et al. (2016). Laboratory Investigation of *Salmonella enterica* serovar Poona Outbreak in California: Comparison of Pulsed-Field Gel Electrophoresis (PFGE) and Whole Genome Sequencing (WGS) Results. PLoS Curr.

Kryazhimskiy, S., and Plotkin, J.B. (2008). The Population Genetics of dN/dS. PLOS Genetics *4*, e1000304.

Kuban, W., Jonczyk, P., Gawel, D., Malanowska, K., Schaaper, R.M., and Fijalkowska, I.J. (2004). Role of Escherichia coli DNA Polymerase IV in In Vivo Replication Fidelity. J Bacteriol *186*, 4802–4807.

Kudva, I.T., Evans, P.S., Perna, N.T., Barrett, T.J., Ausubel, F.M., Blattner, F.R., and Calderwood, S.B. (2002). Strains of Escherichia coli O157:H7 Differ Primarily by Insertions or Deletions, Not Single-Nucleotide Polymorphisms. J Bacteriol *184*, 1873–1879.

Kuo, C.-H., Moran, N.A., and Ochman, H. (2009). The consequences of genetic drift for bacterial genome complexity. Genome Res *19*, 1450–1454.

Laehnemann, D., Borkhardt, A., and McHardy, A.C. (2016). Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. Brief Bioinformatics *17*, 154–179.

Layton, J.C., and Foster, P.L. (2003). Error-prone DNA polymerase IV is controlled by the stress-response sigma factor, *RpoS*, in Escherichia coli. Mol Microbiol *50*, 549–561.

Le Bihan, G., Jubelin, G., Garneau, P., Bernalier-Donadille, A., Martin, C., Beaudry, F., and Harel, J. (2015). Transcriptome analysis of Escherichia coli O157:H7 grown in vitro in the sterile-filtrated cecal content of human gut microbiota associated rats reveals an adaptive expression of metabolic and virulence genes. Microbes Infect *17*, 23–33.

LeClerc, J.E., Li, B., Payne, W.L., and Cebula, T.A. (1996). High mutation frequencies among Escherichia coli and Salmonella pathogens. Science *274*, 1208–1211.

Lee, H., Popodi, E., Tang, H., and Foster, P.L. (2012). Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. PNAS *109*, E2774–E2783.

Leekitcharoenphon, P., Nielsen, E.M., Kaas, R.S., Lund, O., and Aarestrup, F.M. (2014). Evaluation of Whole Genome Sequencing for Outbreak Detection of Salmonella enterica. PLoS One *9*.

Levine, M.M. (1987). Escherichia coli that cause diarrhea: enterotoxigenic, enteropathogenic, enteroinvasive, enterohemorrhagic, and enteroadherent. J Infect Dis *155*, 377–389.

Liamkaew, R., Kosonpisita, S., Supanivatina, P., Saeteawa, N., and Thipayarat, A. (2012). Effect of selective enrichment media on selectivity and isolation of Listeria from non-Listeria strains in suspended cell culture. Procedia Engineering *32*, 119–125.

Lienau, E.K., Strain, E., Wang, C., Zheng, J., Ottesen, A.R., Keys, C.E., Hammack, T.S., Musser, S.M., Brown, E.W., Allard, M.W., et al. (2011). Identification of a Salmonellosis Outbreak by Means of Molecular Sequencing. New Eng J Med *364*, 981–982.

Lukjancenko, O., Wassenaar, T.M., and Ussery, D.W. (2010). Comparison of 61 sequenced Escherichia coli genomes. Microb Ecol *60*, 708–720.

Lynch, M.F., Tauxe, R.V., and Hedberg, C.W. (2009). The growing burden of foodborne outbreaks due to contaminated fresh produce: risks and opportunities. Epidemiol Infect *137*, 307–315.

Majowicz, S.E., Doré, K., Flint, J.A., Edge, V.L., Read, S., Buffett, M.C., McEwen, S., McNab, W.B., Stacey, D., Sockett, P., et al. (2004). Magnitude and distribution of acute, self-reported gastrointestinal illness in a Canadian community. Epidemiol Infect *132*, 607–617.

Matic, I., Radman, M., Taddei, F., Picard, B., Doit, C., Bingen, E., Denamur, E., and Elion, J. (1997). Highly Variable Mutation Rates in Commensal and Pathogenic Escherichia coli. Science *277*, 1833–1834.

McKellar, R.C., Peréz-Rodríguez, F., Harris, L.J., Moyne, A.-L., Blais, B., Topp, E., Bezanson, G., Bach, S., and Delaquis, P. (2014). Evaluation of different approaches for modeling Escherichia coli O157:H7 survival on field lettuce. Int J Food Microbiol *184*, 74–85.

Mellmann, A., Harmsen, D., Cummings, C.A., Zentz, E.B., Leopold, S.R., Rico, A., Prior, K., Szczepanowski, R., Ji, Y., Zhang, W., et al. (2011). Prospective Genomic Characterization of the German Enterohemorrhagic Escherichia coli O104:H4 Outbreak by Rapid Next Generation Sequencing Technology. PLOS ONE *6*, e22751.

Milne, I., Stephen, G., Bayer, M., Cock, P.J.A., Pritchard, L., Cardle, L., Shaw, P.D., and Marshall, D. (2013). Using Tablet for visual exploration of second-generation sequencing data. Brief Bioinform *14*, 193–202.

Miya, S., Takahashi, H., Kamimura, C., Nakagawa, M., Kuda, T., and Kimura, B. (2012). Highly discriminatory typing method for Listeria monocytogenes using polymorphic tandem repeat regions. J Microbiol Meth *90*, 285–291.

Moyne, A.-L., Harris, L.J., and Marco, M.L. (2013). Assessments of Total and Viable Escherichia coli O157:H7 on Field and Laboratory Grown Lettuce. PLOS ONE *8*, e70643.

Nataro, J.P., and Kaper, J.B. (1998). Diarrheagenic Escherichia coli. Clin Microbiol Rev *11*, 142–201.

Nicholson, F.A., Groves, S.J., and Chambers, B.J. (2005). Pathogen survival during livestock manure storage and following land application. Bioresour Technol *96*, 135–143.

Noller, A.C., McEllistrem, M.C., Stine, O.C., J. Glenn Morris, J., Boxrud, D.J., Dixon, B., and Harrison, L.H. (2003). Multilocus Sequence Typing Reveals a Lack of Diversity among Escherichia coli O157:H7 Isolates That Are Distinct by Pulsed-Field Gel Electrophoresis. J Clin Microbiol *41*, 675–679.

Ochman, H. (2003). Neutral mutations and neutral substitutions in bacterial genomes. Mol Biol Evol *20*, 2091–2096.

Orsi, R.H., Bakker, H.C. den, and Wiedmann, M. (2011). Listeria monocytogenes lineages: Genomics, evolution, ecology, and phenotypic characteristics. International J Med Microbiol *301*, 79–96.

Orskov, I., Orskov, F., Jann, B., and Jann, K. (1977). Serology, chemistry, and genetics of O and K antigens of Escherichia coli. Bacteriol Rev *41*, 667–710.

Petkau, A., Mabon, P., Sieffert, C., Knox, N., Cabral, J., Iskander, M., Iskander, M., Weedmark, K., Zaheer, R., Katz, L.S., et al. (2016). SNVPhyl: A Single Nucleotide Variant Phylogenomics pipeline for microbial genomic epidemiology. bioRxiv 092940.

Proulx, F., Seidman, E.G., and Karpman, D. (2001). Pathogenesis of Shiga Toxin-Associated Hemolytic Uremic Syndrome. Pediatr Res *50*, 163–171.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics *13*, 341.

Reeves, P.R., Liu, B., Zhou, Z., Li, D., Guo, D., Ren, Y., Clabots, C., Lan, R., Johnson, J.R., and Wang, L. (2011). Rates of Mutation and Host Transmission for an Escherichia coli Clone over 3 Years. PLOS ONE *6*, e26907.

Reid, S.D., Herbelin, C.J., Bumbaugh, A.C., Selander, R.K., and Whittam, T.S. (2000). Parallel evolution of virulence in pathogenic Escherichia coli. Nature *406*, 64–67.

Roetzer, A., Diel, R., Kohl, T.A., Rückert, C., Nübel, U., Blom, J., Wirth, T., Jaenicke, S., Schuback, S., Rüsch-Gerdes, S., et al. (2013). Whole Genome Sequencing versus Traditional Genotyping for Investigation of a Mycobacterium tuberculosis Outbreak: A Longitudinal Molecular Epidemiological Study. PLOS Medicine *10*, e1001387.

Schuenzel, K.M., and Harrison, M.A. (2002). Microbial antagonists of foodborne pathogens on fresh, minimally processed vegetables. J Food Prot *65*, 1909–1915.

Schwartz, D.C., and Cantor, C.R. (1984). Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell *37*, 67–75.

Shapiro, J.A. (1984). Observations on the formation of clones containing araB-lacZ cistron fusions. Mol Gen Genet *194*, 79–90.

Shapiro, J.A. (1997). Genome organization, natural genetic engineering and adaptive mutation. Trends Genet *13*, 98–104.

Solomon, E.B., Yaron, S., and Matthews, K.R. (2002). Transmission of Escherichia coli O157:H7 from contaminated manure and irrigation water to lettuce plant tissue and its subsequent internalization. Appl Environ Microbiol *68*, 397–400.

Taboada, E.N., Clark, C.G., Sproston, E.L., and Carrillo, C.D. (2013). Current methods for molecular typing of Campylobacter species. J Microbiol Methods *95*, 24–31.

Thomas, M.K., Majowicz, S.E., Pollari, F., and Sockett, P.N. (2008). Burden of acute gastrointestinal illness in Canada, 1999-2007: interim summary of NSAGI activities. Can Commun Dis Rep *34*, 8–15.

Thomas, M.K., Murray, R., Flockhart, L., Pintar, K., Pollari, F., Fazil, A., Nesbitt, A., and Marshall, B. (2013). Estimates of the burden of foodborne illness in Canada for 30 specified pathogens and unspecified agents, circa 2006. Foodborne Pathog Dis *10*, 639–648.

Thomas, M.K., Murray, R., Flockhart, L., Pintar, K., Fazil, A., Nesbitt, A., Marshall, B., Tataryn, J., and Pollari, F. (2015). Estimates of Foodborne Illness–Related Hospitalizations and Deaths in Canada for 30 Specified Pathogens and Unspecified Agents. Foodborne Pathog Dis *12*, 820–827.

Tillman, G.E., Wasilenko, J.L., Simmons, M., Lauze, T.A., Minicozzi, J., Oakley, B.B., Narang, N., Fratamico, P., and Cray, A.C. (2012). Isolation of Shiga toxin-producing Escherichia coli serogroups O26, O45, O103, O111, O121, and O145 from ground beef using modified rainbow agar and post-immunomagnetic separation acid treatment. J Food Prot *75*, 1548–1554.

Treangen, T.J., and Salzberg, S.L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet *13*, 36–46.

Turrientes, M.-C., Baquero, F., Levin, B.R., Martínez, J.-L., Ripoll, A., González-Alba, J.-M., Tobes, R., Manrique, M., Baquero, M.-R., Rodríguez-Domínguez, M.-J., et al. (2013a). Normal Mutation Rate Variants Arise in a Mutator (Mut S) Escherichia coli Population. PLOS ONE *8*, e72963.

Turrientes, M.-C., Baquero, F., Levin, B.R., Martínez, J.-L., Ripoll, A., González-Alba, J.-M., Tobes, R., Manrique, M., Baquero, M.-R., Rodríguez-Domínguez, M.-J., et al. (2013b). Normal Mutation Rate Variants Arise in a Mutator (Mut S) Escherichia coli Population. PLOS ONE *8*, e72963.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., and Rozen, S.G. (2012). Primer3—new capabilities and interfaces. Nucleic Acids Res *40*, e115.

U.S. Department of Agriculture, Food Safety Inspection Service, Federal Register. (2011). Shiga toxin–producing Escherichia coli in certain raw beef. Docket no. FSIS-2010-0023. Available at: www.fsis.usda.gov/OPPDE/rdad/FRPubs/2010-0023.htm

USDA-FSIS (2014). Detection and Isolation of non-O157 Shiga Toxin-Producing Escherichia coli *(STEC)* from Meat Products and Carcass and Environmental Sponges. Available at: http://www.fsis.usda.gov/wps/wcm/connect/7ffc02b5-3d33-4a79-b50c-81f208893204/MLG-5B.pdf?MOD=AJPERES

Warriner, K., Huber, A., Namvar, A., Fan, W., and Dunfield, K. (2009). Chapter 4 Recent Advances in the Microbial Safety of Fresh Fruits and Vegetables. B-A in F and N Research, ed (Academic Press), pp. 155–208.

Windham, W.R., Yoon, S.-C., Ladely, S.R., Haley, J.A., Heitschmidt, J.W., Lawrence, K.C., Park, B., Narrang, N., and Cray, W.C. (2013). Detection by hyperspectral imaging of shiga toxin-producing Escherichia coli serogroups O26, O45, O103, O111, O121, and O145 on rainbow agar. J Food Prot *76*, 1129–1136.

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L.H., Karch, H., Reeves, P.R., Maiden, M.C.J., Ochman, H., et al. (2006). Sex and virulence in Escherichia coli: an evolutionary perspective. Mol Microbiol *60*, 1136–1151.

World Health Organisation. (2008). Foodborne Disease Outbreaks – Guidelines for Investigations and Control. Available at http://www.who.int/foodsafety/publications/foodborne_disease/outbreak_guidelines.pdf

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T.L. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics *13*, 134.

Zhao, J., and Winkler, M.E. (2000). Reduction of GC --> TA transversion mutation by overexpression of *MutS* in Escherichia coli K-12. J Bacteriol *182*, 5025–5028.