The Design and Implementation of Genomics Tools for Enhanced Detection and Characterization of Verotoxigenic *Escherichia coli* Implicated in Foodborne Illness

by

Michael Knowles

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
For the M.Sc. degree in
Bioinformatics

Department of Biology
Faculty of Science
University of Ottawa

Abstract

Motivation Control strains of bacterial pathogens are commonly processed in parallel with test samples in food microbiology laboratories assuring the satisfactory performance of materials used in the analytical procedure. Before positive findings can be reported for risk management purposes, analysts must have a means of verifying that pathogenic bacteria recovered from test samples are not due to inadvertent contamination with the control strain. In addition to detection of laboratory control strains, genomic antimicrobial resistance (AMR) prediction tools may be very useful in supporting food-borne illness outbreak investigations through their application in the analysis of bacterial genomes from causative strains. The AMR marker profile from an outbreak strain of interest could be compared with related bacteria to determine the presence of unique resistance markers enabling customization of selective enrichment media facilitating their recovery from samples during food safety investigations. Food safety investigations focusing on both control strain and priority bacterial pathogen detection, characterization and selective recovery would benefit from a practical means of differentiating between phylogenetically similar isolates.

Results We describe the implementation of a novel algorithm for the identification of DNA signature sequences enabling differentiation of phenotypically related bacterial strains within the same species. This tool, which we have named SigSeekr, was applied in the analysis of whole genome sequence data from select *Escherchia coli* and *Salmonella enterica* strains, enabled the development of primers for their specific detection and differentiation from closely related bacteria using polymerase chain reaction (PCR) techniques. We applied this tool enabling the identification of a common positive control strain using a simple PCR procedure. We combined SigSeekr with selective recovery (using AMR) and identification tools for the efficient detection of verotoxin-producing *Escherichia coli* (VTEC) outbreak strains within the timeframe of an active investigation.

Conclusion We used SigSeekr in a simple PCR procedure to differentiate three *E. coli* and one *Salmonella enterica* between phylogenetically strains including control strains. We also discuss the relative merits of different AMR marker prediction tools employing bioinformatics alignment algorithms along with comprehensive AMR gene databases such as CARD (e.g., Antimicrobial Resistance Marker Identifier [ARMI]), or publicly available tools which use more limited acquired AMR gene databases (*e.g.*, ResFinder), and describe how such tools can be utilized to good effect in a typical outbreak investigation.

Acknowledgements

This work could have only been accomplished with the help and support of my supervisors Dr. Burton Blais and Dr. Douglas Johnson. And the whole team at the Canadian Food Inspection Agency including Dr. Adam Koziol, Dr. Catherine Carrillo, and Dr. Dominic Lambert for bioinformatics support; Paul Manninger and Ashley Cooper for sequencing bacterial strains used in this study; and Mylène Deschênes, George Huszczynski, Austin Markell, and Rachel Hitayezu for technical assistance with PCR analyses. My advisory committee made up of my supervisors, Dr. Dominic Lambert, Dr. Alex Wong, and Dr. Rees Kassen were essential in offering alternative research perspectives.

Funding The Genome Research and Development Initiative, an interdepartmental Food and Water Safety project consortium (comprised of researchers from Agriculture and Agri-food Canada, the Canadian Food Inspection Agency, Environment Canada, Health Canada, the National Research Council of Canada, and the Public Health Agency of Canada).

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Current State of Foodborne Pathogen Detection

Current pathogen detection methods implemented by the Canadian Food Inspection Agency (CFIA) and other food safety regulatory agencies detect 7 verotoxin-producing *Escherichia coli* (VTEC) priority strains defined by serotype O-antigen markers (O26, O45, O103, O111, O121, O145 and O157) with partially-inclusive virulence detection methods (Eppinger et al. 2011; Blais, Gauthier, et al. 2012; Huszczynski et al. 2013). The current VTEC detection procedure is associated with key virulence gene markers (*stx*1, *stx*2, and *eae*) and inclusion with the O-serogroup family as targets for multiplex polymerase chain reaction (PCR) in a cloth-based hybridization array system (CHAS) (Blais, Gauthier, et al. 2012). *E. coli* O157:H7 had a rapid emergence from an unknown strain in 1982 to become the dominant hemorrhagic *E. coli* serotype in North America, responsible for the majority of foodborne illness outbreaks (Eppinger et al. 2011). However, the history of foodborne disease outbreak is rife with examples of causative strains with unanticipated characteristics (e.g., the 2011 German outbreak in which the aetiologic agent belonged to serogroup O104 and lacked the definitive virulence marker *eae*) (Eppinger et al. 2011). These characteristics can make developing detection methods to account for all contingencies difficult, and this is further complicated by variability among non-O157 VTEC strains with respect to resistance of commonly used selective agents (e.g., tellurite, novobicin) in the enrichment process. Therefore, the current state of food safety regulation and trace-back analysis may not allow for the detection of all possible virulent *E. coli*.

Traditional techniques for the detection of pathogenic bacteria in foods rely on a

multi-step process involving pre-enrichment in a selective broth, followed by plating to obtain colony isolates, which are then purified and subjected to a battery of biochemical and serological tests to confirm their identity. The process of definitively identifying bacterial colonies on primary isolation plates can take up to one week to complete because of the requirement for growth and expression of phenotypic characteristics specific to the organism. In some cases (e.g., detection of VTEC of public health concern), phenotypic methods are entirely impractical as a means of identification. Ultimately, these techniques are limited in terms of the type of information (e.g., risk profiling) that can be garnered from an isolate to underscore risk management decisions.

VTEC infections can result in serious medical conditions including bloody diarrhea, hemolytic-uremic syndrome (HUS), kidney failure, microangiopathic hemolytic anemia, and can occasionally be fatal. There are no biochemical features by which most so-called priority VTEC strains can be differentiated from commensal *E. coli* or other VTEC which are not a public health concern. The VTEC method utilized by the Canadian Food Inspection Agency (Blais, Gauthier, et al. 2012; Huszczynski et al. 2013) features a PCR procedure (EHEC-7 CHAS) for the identification of colony isolates on the basis of these defining gene markers within one work day (Gill et al. 2012). "Positive" primary isolates are shipped thereafter to a specialized typing laboratory for further analysis by multiple-locus variable number tandem repeat analysis (MLVA) and pulsed-field gel electrophoresis (PFGE), a process that requires several days and incurs delays in the resolution of outbreak investigations.

1.2 Next-generation sequencing (NGS) and Assembly

The state of the art in NGS technology is nearing the point where clinical isolates of bacteria implicated in foodborne disease outbreaks will be routinely sequenced in reference laboratories at an early stage during such events. Assembly of bacterial raw sequence data from NGS equipment is performed with *de novo* or reference guided assembly (Eppinger et al. 2011). Initially, *de novo* assembly was designed to build contiguous sequences (contigs) from DNA short read sequencing data produced by next generation sequencing. This is predominantly used when studying novel data from organisms for which a reference genome is unavailable for guided assembly. Additionally, unlike reference mapping algorithms – which discard those NGS data not found in a reference genome – it is possible to identify genome insertions and deletions, such as pathogenic islands (Eppinger et al. 2011).

Leading-edge genomic technologies open new possibilities for comprehensive analyses of microbial isolates recovered from food and clinical samples. NGS technologies can now render a bacterial genome much faster and at a significantly lower cost than previously possible (Metzker 2010). The value of rapid benchtop sequencing in the investigation of foodborne disease outbreaks is becoming increasingly accepted (Leekitcharoenphon et al. 2014; den Bakker et al. 2014; Evans et al. 2014; Trees et al. 2014; Schmid et al. 2014; Franz et al. 2014; Joensen et al. 2014; Eyre et al. 2012). Implementing NGS capacity in analytical laboratories supporting food inspection programs would generate high-resolution strain characterization enabling unambiguous identification of pathogens, facilitate detection of relevant genetic markers underpinning the development of risk profiles, eliminate delays associated with shipping isolates to typing facilities, and provide a one-method-fits-all solution for the identification of food pathogens.

1.2.1 *De novo* Genome Assembly using de Bruijn Graphs

Genomics and bioinformatics is heavily reliant on a particular field of math known as graph theory. Our software (discussed later in Chapters 2 and 4) for both SigSeekr and ARMI rely on genomes that have been *de novo* assembled with de Bruijn graphs. In graph theory, the graph is made up of the fundamental parts: nodes, and edges similar to leaves and branches in phylogeny trees. The original graph, devised by Euler, was adapted by de Bruijn to find a cyclic sequence of letters taken from a given alphabet for which every possible word of a certain length k appears as a string of consecutive characters in the cyclic sequence exactly once. The de Bruijn graph was originally intended to solve the superstring problem. This problem involved finding the shortest circular superstring containing all possible substrings of length k over a given alphabet (Compeau et al. 2011). Figure 1.1 shows the construction of graph B using an alphabet composed of the digits 0 and 1 for which every $(k-1)$-mer is assigned to a node; connect one $(k-1)$-mer by a directed edge to a second $(k-1)$-mer if there is some k-mer whose prefix is the former and whose suffix is the latter. Edges of the de Bruijn graph represent all possible k-mer, and thus an Eulerian cycle in B represents a shortest (cyclic) superstring that contains each k-mer exactly once. By checking that the indegree and outdegree of every node in B equals the size of the alphabet, we can verify that B contains an Eulerian cycle. In turn, we can construct an Eulerian cycle using Euler's algorithm, therefore solving the superstring problem (Compeau et al. 2011).

Fig. 1.2a illustrates how graphs are used in *de novo* genome assembly; using the simple

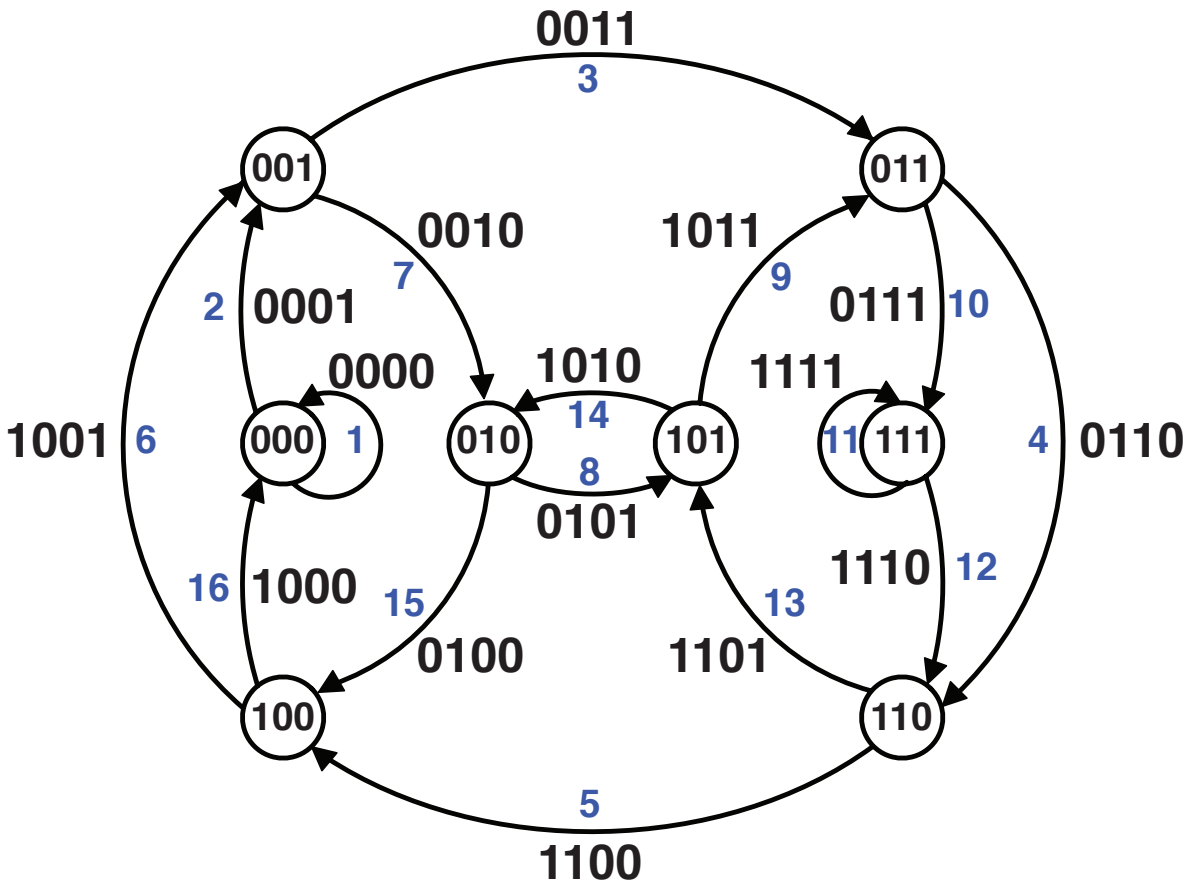

Figure 1.1: The de Bruijn graph B for $k = 4$ and a two-character alphabet composed of the digits 0 and 1. This graph has an Eulerian cycle because each node has indegree and outdegree equal to 2. Following the blue numbered edges in order from 1 to 16 traces an Eulerian cycle **0**000, **0**001, **0**011, **0**110, **1**100, **1**001, **0**010, **0**101, **1**011, **0**111, **1**111, **1**110, **1**101, **1**010, **0**100, **1**000. Recording the first character (in boldface) of each edge label spells the cyclic superstring **0000110010111101**. Adapted with permission from Compeau et al. (2011)

example with five very short reads (CGTGCAA, ATGGCGT, CAATGGC, GGCGTGC and TGCAATG) sequenced from a small circular genome, ATGGCGTGCA. Current NGS methods produce reads that vary in length, but the most popular technology generates 300-nucleotide reads (Metzker 2010). A graph in which each read is represented by a node and overlap between reads is represented by an arrow (called a 'directed edge') joining two reads was used as a straightforward method for assembling reads into longer contiguous sequences – and the one used for assembling the human genome in 2001 as well as for all other projects based on Sanger sequencing. For instance, two nodes representing reads may be connected with a directed edge if the reads overlap by at least five nucleotides. Fig. 1.2b (Compeau et al. 2011)

Traversal along the edges of the graph in Fig. 1.2c provides an aid for understanding in the case of genome assembly. The traversal path traces a series of overlapping reads, and thus represents a candidate assembly. Specifically, following the path **AT**GGCGT $\rightarrow$ **GG**CGTGC $\rightarrow$ **CG**TGCAA $\rightarrow$ **TG**CAATG $\rightarrow$ **CA**ATGGC $\rightarrow$ **AT**GGCGT, this induces a 'Hamiltonian cycle' in our graph. A Hamiltonian cycle is a traversal of a graph that travels to every node exactly once and ends at the starting node, meaning that each read will be included once in the assembly. The circular genome ATGGCGTGCA, which is computed by concatenating the first two nucleotides in each read in such a Hamiltonian cycle, contains all five reads and thus reconstructs the original genome (although we may have to 'wrap around' the genome, for example, to locate CAATGGC in **ATGGC**GTG**CA**) (Compeau et al. 2011).

Modern assemblers usually work with strings of a particular length k (k-mers), which are shorter than entire reads (*e.g.*a 100-nucleotide read may be divided into 46 overlapping 55-mers). The Hamiltonian cycle approach to make use of 3-mers by constructing a graph in Fig. 1.2. 1) From a set of reads, make a node for every k-mer appearing as a consecutive substring of one of these reads (e.g., in Fig. 1.2, ATG, TGG, GGC, GCG, CGT, GTG, TGC, GCA, CAA and AAT). 2) Given a k-mer, define its 'suffix' as the string formed by all its nucleotides except the first one and its 'prefix' as the string formed by all of its nucleotides except the last one. Connect one k-mer to another using a directed edge if the suffix of the former equals the prefix of the latter – that is, if the two k-mers completely overlap except for one nucleotide at each end ($k-1$-mer Fig. 1.2c). 3) The Hamiltonian cycle will be produced in a graph representation of a candidate genome evidenced by a minimal length because it visits each detected k-mer (node) exactly once (Compeau et al. 2011). Although this is trivial for a genome of 10bp a similar graph for a single run of an Illumina (San Diego, CA) sequencer that generates

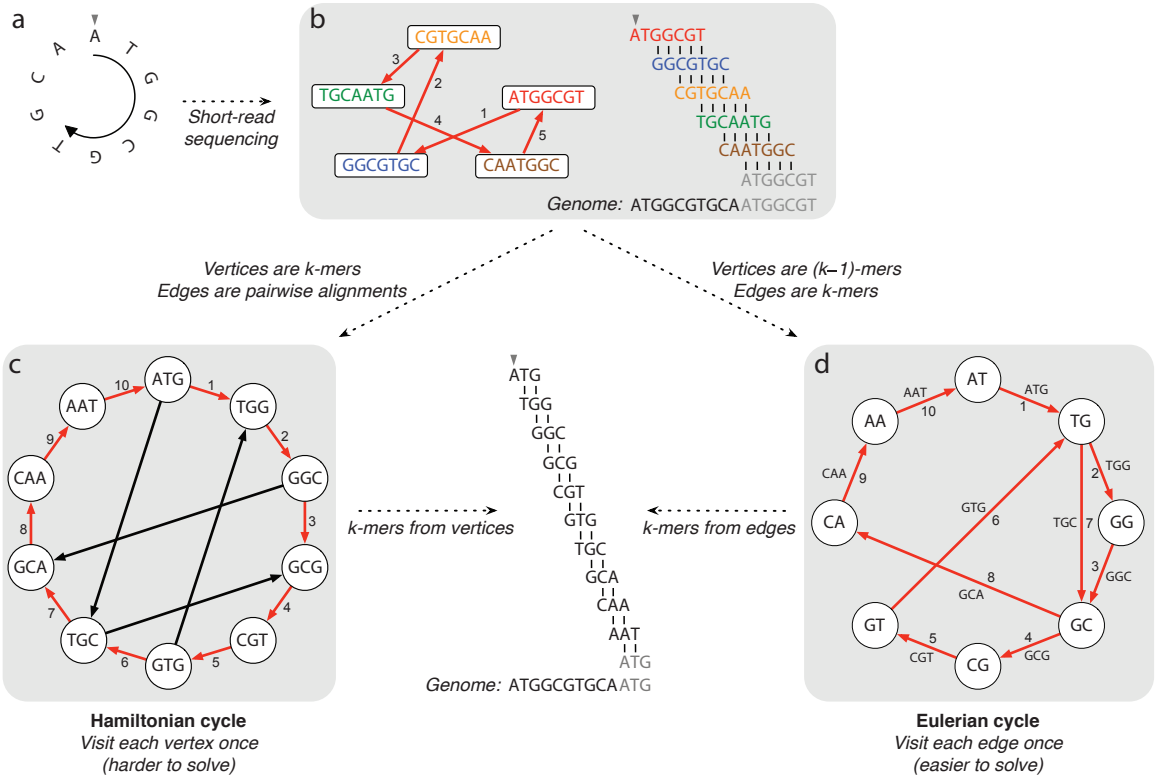

Figure 1.2: Genome assembly with graphs (a) An example small circular genome. (b) In traditional Sanger sequencing algorithms, reads were represented as nodes in a graph, and edges represented alignments between reads which can be represented by overlapping nodes. Reconstruction of the circular genome by combining alignments between successive reads by a Hamiltonian cycle following the edges in numerical order. At the end of the cycle, the sequence wraps around to the start of the genome. (c) A k-mer based approach can assemble the graph by first splitting reads into all possible k-mers. Reconstruction of the circular genome by forming an alignment in which each successive k-mer (from successive nodes) is shifted by one position is accomplished in a Hamiltonian cycle indicated the red edges. (d) Modern short-read assembly algorithms construct a de Bruijn graph by representing all k-mer prefixes and suffixes as nodes and then drawing edges that represent k-mers having a particular prefix and suffix. Adapted with permission from Compeau et al. (2011)

many reads. A million (10^6) reads will require a trillion (10^{12}) pairwise alignments. A billion (109) reads necessitate a quintillion (10^{12}) alignments.

As noted in the previous section, finding a cycle that visits all nodes of a graph exactly once (called the Hamiltonian cycle problem) is a difficult computational problem (*NP*-Hard). It is often simpler to find a cycle that visits all edges of a graph exactly once (Knuth 2008) because directed edges only allow one-way traversal (Eulerian cycle). This algorithmic contrast has motivated computer scientists to solve the problem of DNA short-read assembly with other approaches. Instead of assigning each k-mer contained in some read to a node; reducing the node size by assigning each such k-mer to an edge. This allows the construction of a de Bruijn graph in Fig. 1.2d: 1) Form a node for every distinct prefix or suffix of a k-mer, meaning that a given sequence of length $(k-1)$ (e.g., AT, TG, GG, GC, CG, GT, CA and AA) can appear only once as a node of the graph. Then, connect node x to node y with a directed edge if some k-mer (e.g., ATG) has prefix x (e.g., AT) and suffix y (e.g., TG), and label the edge with this k-mer (Compeau et al. 2011).

1.2.2 SPAdes

The SPAdes assembler was designed with the purpose of assembling next-generation sequencing data retrieved from single-cell sequencing. This makes SPAdes an effective assembly algorithm for preserving the data on DNA variability of foodborne bacterial pathogens (Eppinger et al. 2011). The algorithm is specially designed to process highly non-uniform read coverage, chimeric reads, and elevated levels of sequence errors. The processes described by Bankevich et al. (2012) can also easily be applied to multicell sequencing since it suffers from similar symptoms but to a lesser degree. In NGS data, chimeric reads can complicate the ensuing assembly by causing bifurcation of the de Bruijn graph and giving rise to artificial contiguous sequences. SPAdes benefits from an error correction procedure adapted from Quake, thus reducing chimeric reads and sequencing errors (Kelley et al. 2010).

SPAdes utilizes multisized de Bruijn graph which allows employing different values of k. The use of smaller values of k in low-coverage regions minimize fragmentation, and larger values of k in high coverage regions to decrease repeat collapsing. This process is effectively detects and removes bulge/bubble and chimeric reads. SPAdes employs the basic concept of paired de Bruijn graphs. However, paired de Bruijn works well on paired-end reads with fixed insert size. Therefore SPAdes estimates 'distances' instead

of using insert sizes i.

$$d = i - L \tag{1.1}$$

Distance d of a paired-end read is defined as the difference of i and the read length L. The distances are estimated by utilizing k-bimer adjustment approach. A k-bimer consisting of k-mers α and β together with the estimated distance between them in a genome $P(\alpha|\beta, d)$. This approach breaks the paired-end reads into pairs of k-mers which are transformed to define pairs of edges (biedges) in the de Bruijn graphs. These sets of biedges are involved in the estimation of distances between edges paths between k-mers α and β. By clustering, the optimal distance estimate is chosen from each cluster (stage 2, above). To construct paired de Bruijn graph, the rectangle graphs are employed in SPAdes (stage 3). Rectangle graphs approach was first introduced in 2012 to construct paired de Bruijn graphs with doubtful distances (Bankevich et al. 2012).

1.2.3 Alternatives to NGS

While NGS is a state of the art process, it is still expensive for routine testing and the average run of an Illumina MiSeq requires three days to complete (Metzker 2010). Recent work at the Canadian Food Inspection Agency (CFIA) has yielded a method for rapid detection of a foodborne pathogen of public health concern using an Illumina MiSeq and performing on-the-fly analysis while the sequencing reactions are still occurring (Lambert et al. 2015). Alternatively, polymerase chain reaction (PCR) is a process of amplifying a DNA fragment with oligonucleotides to a quantity that is detectable using agarose gel electrophoresis. PCR requires only a few hours to complete with significantly lower cost than the NGS method employed by Lambert et al. (2015). In this manner, a sample could be detected and characterized using specialized primers that are unique to the sample within minimal time and cost, enabling same day reporting of results and realizing the full advantage of a rapid colony identification method.

1.3 DNA Alignment Algorithms

In a foodborne illness outbreak investigation scenario, genomics and bioinformatics are rapid and cost effective *in silico* strategies for strain-specific target discovery. Alignment tools, such as Basic Local Alignment Tool (BLAST) and bowtie2 offer the ability to take genomic sequence and align them against a reference genomic sequence. In the brief

timeframe of foodborne illness outbreak investigation scenarios characterized in Fig. 4.3, BLASTn is a quick and efficient way to find genes of public health concern including the Shiga-toxin genes, *stx*1 or *stx*2, the intimin-coding gene, *eae*, and markers for the specific serogroups of concern (e.g., O26, O45, O103, O111, O121, O145 and O157) (Blais, Gauthier, et al. 2012; Huszczynski et al. 2013). Corollary to this, BLAST requires assembled genomes (described earlier), however, for genomic data that is unassembled a different algorithm is required because the large number short reads produced by modern sequencers for a single sample is too inefficient with BLASTn. Short read alignment software is becoming the new norm for rapid foodborne pathogen detection (Lambert et al. 2015). Alignment of 'raw' next-generation sequencing (NGS) data using a Burrows-Wheeler transform and FM index is both rapid and accurate with bowtie2 (Langmead and Salzberg 2012). ARMI (Chapter 4) is currently being tested with bowtie2. As an alternative to gene alignment, modern DNA alignment algorithms should make it possible to eliminate DNA sequences in a sample using a database of non-target sequences. The software SigSeekr was developed to create signature species-specific DNA targets through pan-genomic analysis. We use a database of whole genome assemblies of the same species as non-target sequences in an attempt to remove common sequences from a target genome. The BLAST algorithm uses a set of statistical functions for alignment while bowtie2 uses a compression algorithm (Burrows-Wheeler transform) to reduce an array (FM index) which is used to align to a target.

1.3.1 Basic Local Sequence Alignment Tool (BLAST)

The workflow of SigSeekr (Chapter 2) relies on NCBI BLASTn. It performs genome searches by first masking low-complexity regions using an nucleotide-triplet frequency assessment algorithm (DUST). BLAST creates a k-mer word list of DNA segments using a siding window 11 bases long by default. A neighbourhood score is determined and is used with a predefined threshold to limit mismatches. High scoring pairs (HSPs) are assembled into a tree then iterated over to extend the HSPs in the database sequence until an HSP falls below a predetermined cutoff score. A visual representation of HSP graphs is displayed in Figure 1.3. An expectation E value is calculated to assess the statistical significance of the HSP matches using the Gumbel extreme value distribution (EVD). In accordance with the Gumbel EVD, the probability P of observing a score S equal to or greater than x is given by the equation

$$p\left(S \geq x\right) = 1 - \exp\left(-e^{-\lambda(x-\mu)}\right) \tag{1.2}$$

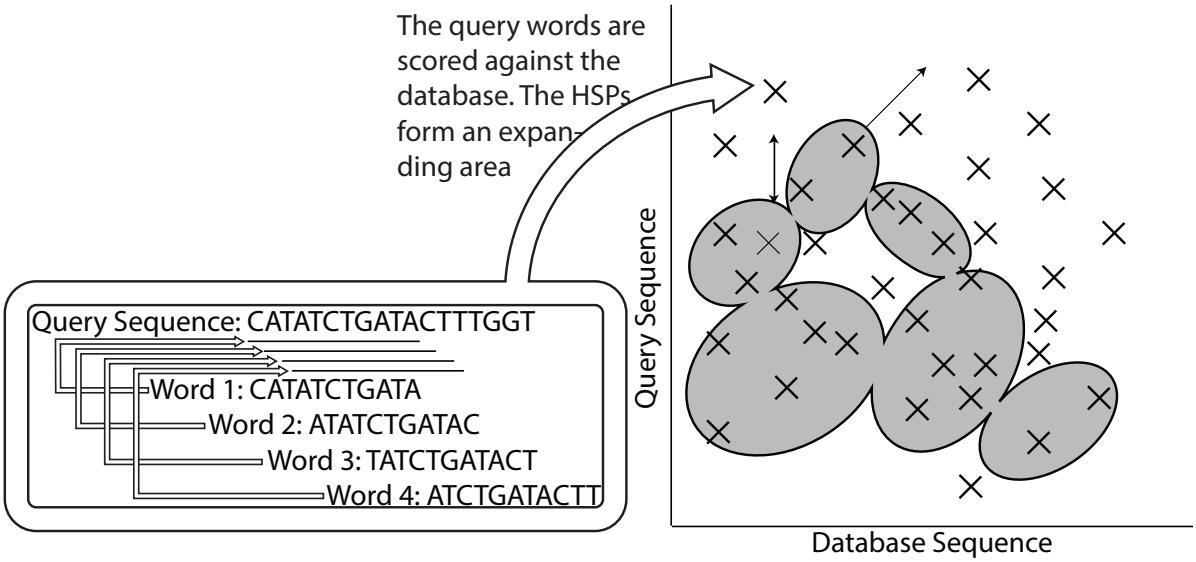

Figure 1.3: The construction of an array of k-mers with a sliding window of 11mer and the query words are shown to be represented as a high-scoring pair (HSP) in the database. The HSPs are expanded outward based on optimal scores.

where

$$\mu = {}^{[\log(Kmn)]}\!/_{\lambda} \tag{1.3}$$

Where S refers to the score obtained from a random sequence alignment with the database (brut score) and x is the observed original score of the alignment. Thus K and λ are estimated by fitting the distribution to an ungapped local alignment. m and n are the effective lengths of the query and database sequences, respectively. Therefore, the E-value can be obtained from the normalized S (bitscore) in the following equation:

$$E = mn2^{-S'} \tag{1.4}$$

where

$$S' = {}^{(\lambda S - \ln K)}\!/_{\ln 2} \tag{1.5}$$

The E-value of a BLAST database search calculates probability of match using predefined cutoff values, ungapped local alignment scores and parameters of the substitution matrix. In a BLAST database search, the E-value describes the expected number of hits that would occur by random chance alone; short and relatively dissimilar BLAST matches have high E-values, and long BLAST matches with high sequence identity have low E-values (Altschul et al. 1990). Therefore, elimination of HSPs from a genome can be manipulated precisely with the E-value of BLAST to remove matching results from the target sequence. This ability may make it possible to eliminate common sequences from an assembled genome when compared to a large non-target genome database.

1.3.2 Burrows-Wheeler transform and FM index for alignment

The Burrows-Wheeler transform (BWT), developed by Burrows and Wheeler (1994), is the reversible permutation of the characters of a string. This concept was originally used for data compression. Figure 1.4a describes the BWT of string T = abaaba\$. The '\$' symbol is used on the end of T because it is the first character in lexicographically order. T is then shifted one character so that the last character of the previous string becomes the first character of the next string. The resulting strings are sorted lexicographic into a Burrows-Wheeler matrix (BWM). The last column of the sorted matrix becomes the Burrows-Wheeler transform (Burrows and Wheeler 1994). Therefore the characters of BWT(T) are sorted by their 'right-context' of every character in T this tends to put like characters with one another. This lends additional structure to BWT(T), tending to make it more compressible. As seen in Figure 1.4b BWMs bear resemblance to suffix

arrays, the others are the same because the '$' allows the string to complete therefore any follow characters have no weight on sorting. This allows quick construction of a BWT using a suffix array method using the following definition of BWT(i) (Li and Durbin 2009):

$$BWT[i] = \begin{cases} T[SA[i]-1] & \text{if } SA[i] > 0 \\ \$ & \text{if } SA[i] = 0 \end{cases} \tag{1.6}$$

The benefit of the BWT is two fold, first it allows for complete permutation of the original string and second, the BWT benefits from a low memory footprint because of its adaptation to compression. Subsequently, with the addition of a suffix array, the BWT can be used to align exact matches of strings (*e.g.*, a genome) Ferragina and Manzini (2000) developed the FM index using the first column of the BWM(T) to provide an opportunity to efficiently query a prefix trie a for exact matches of a given string P. A prefix trie is a hierarchical data structure commonly used in computer science to organize data for efficient re*trie*val. The prefix trie, used in example of BWM(T), is a representation of all possible backward search paths of the FM index. BWM(T) in Figure 1.4 can also be represented as a prefix trie (Li and Durbin 2009). The position of P can be further elucidated with a tally matrix that lists the occurrence of characters in BWT(i) – at a given interval to reduce memory usage. The FM index lookup benefits from an $\mathcal{O}(|P|)$ efficiency for a given string (Ferragina and Manzini 2000).

The FM index can be adapted to query for inexact matches of a string as described by Li and Durbin (2009). The algorithm uses a similar backward search to sample distinct substrings W in a genome. The process is bounded by the array D where $D(i)$ is the lower bound of the number of differences by the $W[0, i]$. The better the array D is estimated, the smaller space and the more efficient the algorithm is. The naïve bound of the array is achieved by setting $D(i) = 0$ for all i, but the resulting algorithm is inefficient since it is clearly exponential in the number of different combinations obtained (Li and Durbin 2009). The algorithm described by Li and Durbin (2009) is guaranteed to find all differences z but there are modifications to adapt to genomic data. First, penalties for introducing mismatches, gaps, and gap extensions. Second, a heap-like data structure is employed to keep partial alignments rather than recursion and prioritized on the alignments of these partial alignments. The reverse complement is processed concurrently with the algorithm. Third, an iterative strategy is adopted to reduce suboptimal hits, if the top interval is unique and has z difference, only search of hit with up to $z + 1$ differences. Bowtie 2 developed by (Langmead and Salzberg 2012) uses a similar algorithm

but it incorporates a dynamic programming approach to which is easily single-instruction multiple-data (SIMD) parallelized (*e.g.*, the type of processing commonly employed in modern computer graphics processing units (GPUs)). To better understand the role of the array D, we use the example of searching W = bba in T = abaaba\$. If we set $D(i) = 0$ and i to disallow gaps, the compression of a BWT applied to 'raw' NGS data increases the rate at which it may be processed. This rate can be reduced to a worst-case complexity when using a FM index to a perform a search of the 'raw' genome against a target nucleotide sequence (*e.g.*, a gene). This efficiency is known as big 'O' notation depicted by $\mathcal{O}$. With the BWT and FM index (and prefix trie) employed by bowtie2 (Langmead and Salzberg 2012), a efficiency of $\mathcal{O}(n \log n)$. This complexity makes it useful when processing large amounts of data retrieved from modern NGS sequencers. The efficiency can make on-the-fly anaylsis applications like those performed by Lambert et al. (2015) possible. We can also employ applications like bowtie2 for analysis of antimicrobial resistance (AMR) in a NGS sample.

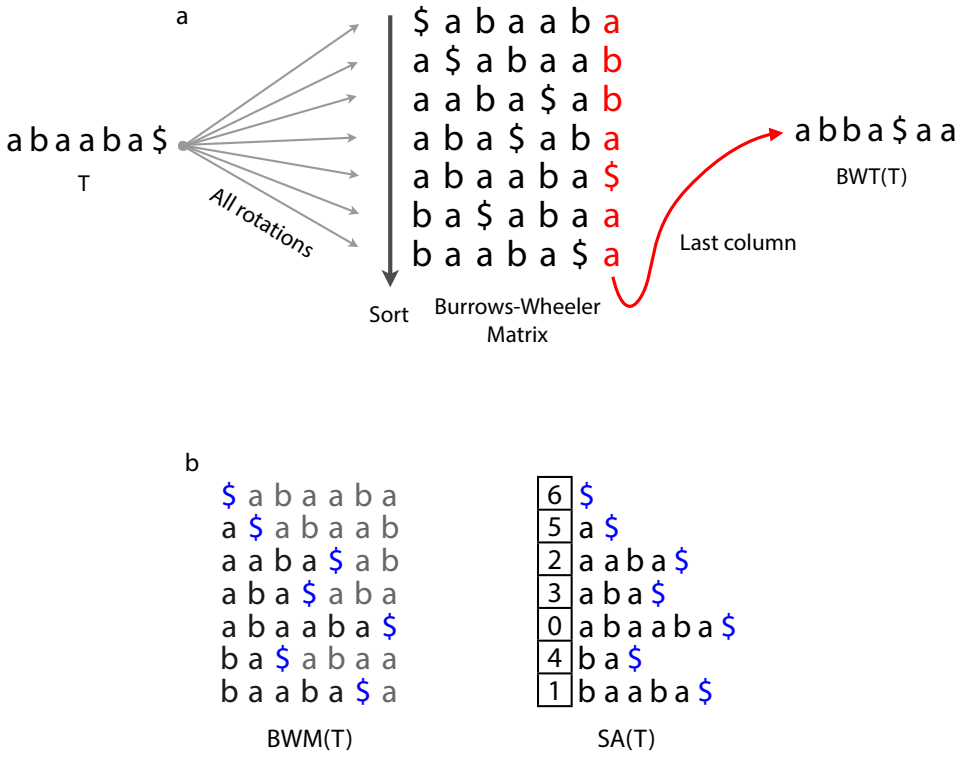

Figure 1.4: Overview of Burrows-Wheeler transform algorithm (a) The construction of a Burrows-Wheeler transform from the the string 'abaaba$' ($T$). The '$' symbol is used on the end of the string because it is the first character in lexicographic order. The string is then shifted one character so that the last character of the previous string becomes the first character of the next string. The resulting strings are sorted lexicographically into a Burrows-Wheeler matrix. The last column of the sorted matrix becomes the Burrows-Wheeler transform. (b) Resemblance of Burrows-Wheeler matrix (BWM) to suffix array (SA) for string T.

1.4 Antimicrobial Resistance

In recent years, multidrug resistance in clinical pathogens has been rising, whereas the process for new antibiotic development and approval continues to dwindle. As a result, we face the prospect of returning to a 'pre-antibiotic era' where an increasing number of infections can no longer be treated effectively with our current arsenal of drugs. It is clear that dissemination of antimicrobial resistance (AMR) extends beyond clinical samples to include agricultural and environment microbial communities (Gibson et al. 2015). Recent studies have found AMR genes have been found in foodborne pathogens the investigation into the the full impact of AR bacteria on human health is still unknown (Gibson et al. 2015; McDermott et al. 2002). Antimicrobials are often extensively used in raising of animals in developed nations and as a result of exposure to this selective pressure the bacteria in food animals favour immunity to antimicrobials while the wildtype bacteria tend to be susceptible to such agents (McDermott et al. 2002). Since AMR genes are often found on mobile genetic elements (*i.e.*, transposons, plasmids, bacteriophages, and integrons) in bacteria, there is a possibility of transfer of genetic material among bacterial pathogens in matrices such as food and animal feed (McDermott et al. 2002; Piddock 1996). This immunity to antimicrobial agents has been observed since the beginnings of therapies in both the human and veterinary fields and is a fundamental caveat to fighting infectious disease (McDermott et al. 2002).Three scenarios have been proposed by which the use of antimicrobials in food animals could have potential human health implications: 1) antimicrobial-resistant zoonotic bacterial pathogens are selected, and food is contaminated during slaughter and/or preparation. After consumption of the food, these pathogens cause an infection that requires antibiotic treatment and therapy is compromised; 2) antimicrobial-resistant bacteria non-pathogenic to humans are selected in the animal. When the contaminated food is ingested, the bacteria transfer resistance determinants to other bacteria in the human gut, commensal and potential pathogens; and 3) antibiotics remain as residues in food products, which allows the selection of antibiotic-resistant bacteria after the food is consumed (McDermott et al. 2002; Piddock 1996). In these circumstances, it is vital to for surveillance of AMR abundance in food and feed, and subsequent studies on AMRs in resulting clinical samples. Surveillance is needed to better understand which practices in food production might be leading to the AMR prevalence, and to understand the modes of transmission of genetic material between foodborne bacteria.

The occurrence of bacteria with AMR characteristics (e.g., resistance to therapeutic

antibiotics) in foods is widely regarded as a serious public health threat. Outbreaks of foodborne illness caused by AMR in bacteria are becoming more common, for example, a recent outbreak of salmonellosis in the US linked to chicken contaminated with a multi-drug resistant strain of *Salmonella* Heidelberg (CDC 2013). In North America antibiotics are regularly used to supplement livestock feeds because of their growth-promoting properties. There is some evidence indicating that this practice has given rise to the acquisition of AMR traits in microbiota associated with feeds, with the potential for transfer of AMR factors to bacterial pathogens associated with the human gut, ultimately impacting on the efficacy of therapeutic antibiotic use in the treatment of infectious diseases (Barton 2000). The occurrence of AMR bacteria in animal manures (cattle, chicken and pigs) has been documented, with noted resistance to β-lactams, aminoglycosides, chloramphenicol, tetracyclines, among others (Mathew et al. 2007). Once the gut microbiota of food animals have acquired AMR traits, drug-resistant bacteria will accumulate in manure that is spread on fields and enters waterways, allowing these bacteria to spread far and wide and ultimately back up the food chain to humans.

The widespread occurrence of AMR in foodborne pathogenic bacteria does allow the use of antimicrobial enriched media for selective recovery of said pathogen. Leading-edge genomics approaches provide new possibilities for comprehensive analyses of foodborne bacteria fostering the development of highly targeted tools enabling the efficient detection of outbreak strains. Next-generation sequencing technologies can now render a bacterial genome much faster (possibly within a single working day) and at a significantly lower cost (about one hundred dollars) than previously possible, making it feasible to sequence foodborne isolates within the time frame of a food safety event (Lambert et al. 2015). Current outbreak investigation routines typically involve whole genome sequencing (WGS) of clinical isolates by public health laboratories, making it possible for partner agencies (e.g., food inspection laboratories) to rapidly acquire WGS data which may offer important clues about the particular characteristics of the causative organism to aid in its detection in food samples. For example, it may be possible to determine a priori the presence of antibiotic resistance markers and strain-specific DNA sequences enabling the application of customized selective recovery and identification methods (e.g., polymerase chain reaction (PCR)) enhancing their detection against a background of non-target bacteria.

A number of tools are currently available to predict AMR from bacterial WGS data [e.g., ResFinder (Kleinheinz et al. 2014; Zankari et al. 2012), SEAR (Rowe et al. 2015), Resistance Gene Identifier (McArthur et al. 2013), and an in-house tool developed by the

present investigators named Antimicrobial Resistance Marker Identifier (ARMI)]. These AMR marker prediction tools (APTs) rely on curated international AMR gene databases such as CARD (McArthur et al. 2013), ARDB (B. Liu and Pop 2009), and ARG-ANNOT (Gupta et al. 2014). WGS-based methods for prediction of AMR phenotype have been shown to be highly accurate (Tyson et al. 2015; Zankari et al. 2012). Here we propose the systematic application of such tools (e.g., the ARMI tool developed in our laboratory, and ResFinder) for the identification of antibiotic resistance in a locally sequenced bacterial strain, such as a clinical VTEC isolate, to enable the design of selective media for use in its specific recovery in food samples during food safety investigations. Combining APTs with a previously described a bioinformatics pipeline (SigSeekr) for the WGS-based design of specific PCR primers (Knowles et al. 2015) targeting a strain of interest, it should be possible to deploy customized detection methods to rapidly identify foods implicated in a food safety incident. The feasibility of such an approach was investigated with a model laboratory verotoxin-producing *Escherichia coli* (VTEC) strain subjected to analysis using APTs and SigSeekr tools developed in our laboratory, and the detection of this strain in ground beef by customized selective enrichment and PCR techniques predicated on this bioinformatics platform.

A benefit of AMR prevalence in bacteria allows for selective enrichment. The two databases the Comprehensive Antibiotic Resistance Database (CARD) (McArthur et al. 2013) and the Antibiotic Resistance Genes Database (ARDB) (B. Liu and Pop 2009) serve to catalogue and annotate genes conferring AMR in bacteria. These databases are meticulously curated to reduce errors, this makes these databases a useful tool for designing of custom enrichment broths for food safety traceback investigations.

1.5 Preamble

1.5.1 Hypothesis

Ultimately, this work has been built on the premise of improving conventional laboratory procedures and traceback analyses used in food regulatory science at the Canadian Food Inspection Agency (CFIA). We want to show that it is possible build bioinformatics software to predict signature DNA sequences in a target verotoxin-producing *Escherichia coli* (VTEC) and that the predicted signature sequence can differentiate our target from phenotypically similar bacteria at a particular point in time. To prove this we have designed an experiment in Chapter 2 where we use the signature sequence of a target VTEC

in PCR primers and perform a PCR against a panel of phenotypically similar bacteria and bacterial different species. Confirmation of this hypothesis would require very few of the non-target bacteria within the panel to produce a PCR amplicon and the target VTEC to produce an amplicon in isolated testing. If we are able to accurately predict the presence or absence of a specific VTEC strain within a sample of phenotypically similar bacteria, it will be possible to use this method to improve routine laboratory procedures at the CFIA. We want to prove it is also possible to build bioinformatics software to predict the antimicrobial resistance (AMR) of a VTEC strain at a particular point in time. In Chapter 4 we design an experiment to test the AMR predictions using media made with the particular antibiotic of which resistance is predicted. We will also contrast the results with other similar software.

The corollary of signature sequences and signature AMRs would allow an experimenter to differentiate target VTEC and non-target VTEC in a food matrix (*i.e.*, ground beef) that was allowed incubate in a broth of predicted antimicrobials. The experiment designed in Chapter 4 involves acquiring the signature sequence and signature AMR predictions of a target VTEC. The target VTEC is a added to a mixture of food matrix, growth media, and predicted antimicrobials then allowed to incubate. Our expectation of the experiment is a reduction in the variety of bacterial populations when compared to a food matrix incubated in similar conditions without antimicrobials. To test the expected recovery of the target VTEC, we will use its signature sequences in PCR on each isolated microbial colony in the sample and compare results to a set of controls. We want to prove there is a higher number of colonies that produce a PCR amplicon the food matrix incubated in antimicrobials than a similar food matrix grown without antimicrobials.

1.5.2 Objectives

Currently, clinical isolates of bacteria implicated in foodborne illness are routinely sequenced at an early stage during outbreaks. The objectives are three parts building on top of each other. (1) We exploited reductive whole genome screening to inform the design of strain-specific PCR probes and primers. These primers are proven capable to provide conclusive results on presence of a strain in a panel of closely-related organisms. A comprehensive literature/public database search was conducted to catalogue AMR sequences known to exist among VTEC and related bacteria (e.g., ARDB, CARD, and PATRIC) (B. Liu and Pop 2009; McArthur et al. 2013; Wattam et al. 2014). (2) Followed

by an analysis to identify AMRs in WGS data from a small selection of VTEC strains addressed in the FWS GRDI project (different priority serotypes, as well as multiple strains within a given serotype). For comparative purposes, publicly available WGS information for commensal *E. coli* (and possibly other food microflora, if available) will likewise be analyzed to determine the "background" AMR load, which may need to be discounted in devising customized selective media for enhanced recovery of outbreak strains. (3) We will attempt to recreate a scenario for the integration of genomics technology in an outbreak investigation in Fig. 4.3.

Chapter 2

SigSeekr: Short Signature Sequence Identifier

Preface

Copyright

Contributions of Collaborators

The following chapter is written as a manuscript in preparation for submission by Michael Knowles, Adam G. Koziol, Burton W. Blais, and Dominic Lambert titled "SigSeekr: Short Signature Sequence Identifier" in the *Bioinformatics* as an Applications Note. Typically, this is a two-page report on software developed for bioinformatics purposes. I, Michael Knowles, assembled all experimental materials and performed all of the experiments and developed algorithms in the following manuscript with the exception of the following:

- Collection of *Escherchia coli* and *Salmonella enterica* isolates (B. W. Blais, and D. Lambert)
- Sequencing of bacteria (A. G. Koziol, A. Cooper, M. Deschênes, P. Manniger)
- PCR experiments detailed in Table 2.4 (M. Deschênes)

Burton W. Blais, Dominic Lambert and I, Michael Knowles, wrote the original draft of this paper and all revised versions.

2.1 Abstract

2.1.1 Motivation

Food safety investigations focusing on detection and characterization of priority bacterial pathogens, such as *Salmonella enterica* and *Escherichia coli*, would benefit from a practical means of differentiating between phylogenetically similar isolates.

2.1.2 Results

We describe the implementation of a novel python-based tool for the identification of DNA signature sequences enabling differentiation of phenotypically related bacterial strains within the same species. This tool, which we have named SigSeekr, was applied in the analysis of whole genome sequence data from select *Escherchia coli* and *Salmonella enterica* strains, enabled the development of primers for their specific detection and differentiation from closely related bacteria using polymerase chain reaction (PCR) techniques.

2.1.3 Availability

SigSeekr is available online at https://github.com/OLC-LOC-Bioinformatics/SigSeekr

2.2 Introduction

The rapid identification of a specific pathogenic bacterial strain responsible for a foodborne illness outbreak is a key element in traceback investigations. Traditional culture, phenotypic and molecular typing tests can identify specific outbreak strains, but they are time-consuming to perform, costly, and often lack the discriminatory power to distinguish an outbreak strain from other closely related bacteria which may be present in investigative samples. Polymerase chain reaction (PCR) methods can be very useful in expediting the process of identifying target bacteria in enrichment cultures of food samples, but current approaches cannot distinguish bacteria at the strain level. The

availability of large whole-genome sequence databases and powerful bioinformatics tools has the potential to support the development of rapid and cost-effective strategies for the elaboration of PCR tests targeting specific bacterial strains using pan-genomic analyses (PGA)(C.-C. Ho et al. 2012). Here we present SigSeekr: a multithreaded parallel pan-genomic analysis software using an iterative search process for the identification of strain-specific sequences. This software facilitates the design of customized PCR tests targeting specific bacterial strains, such as a strain implicated in a foodborne illness outbreak, or even to determine if a positive test result might due to inadvertent contamination of the sample by control bacterial strains (Blais, Martinez-Perez, Gauthier, et al. 2008) commonly used in the laboratory environment (Knowles et al. 2015). Performed alongside standard detection procedures, such PCR tests can significantly reduce the time required to identify tainted food commodities implicated in and the scope of a foodborne illness outbreak (Knowles et al. 2015). Other key features of SigSeeker include optimized memory use, a ribosomal multilocus sequence typing function (Jolley et al. 2012), and strain-level differentiation and recursion of Basic Local Alignment Tool nucleotide (BLASTn) searches using progressively reduced E-values.

2.3 Methods

2.3.1 Pan-genome sequence database

All publicly available *Escherichia coli* (1600) and *Salmonella* (1200) GenBank sequences (i.e., unannotated genomic contigs, closed assembled pseudomolecules (chromosome-like DNA fragment), and plasmids were retrieved from NCBI GenBank using the Biopython application programming interface v1.64 (Cock et al. 2009) Entrez module (NCBI, June 2015) and were combined with the locally generated sequences to form the database used by SigSeekr. To ensure that all genome entries were represented only once in the database, an optional *in silico* ribosomal multilocus sequence typing assay (rMLST), previously shown to be sufficiently discriminatory for differentiation of isolates at the strain level (Jolley et al. 2012), was used to identify possibly redundant entries (e.g. multiple runs of the same genome entered using slightly different names). These entries were removed from the SigSeekr database to ensure that sequences unique to the target strain were not mistakenly eliminated. The applicable entries are provided in the standard output data. The rMLST subroutine was omitted in cases where the target strain had a similar sequence type to non-redundant genome sequences.

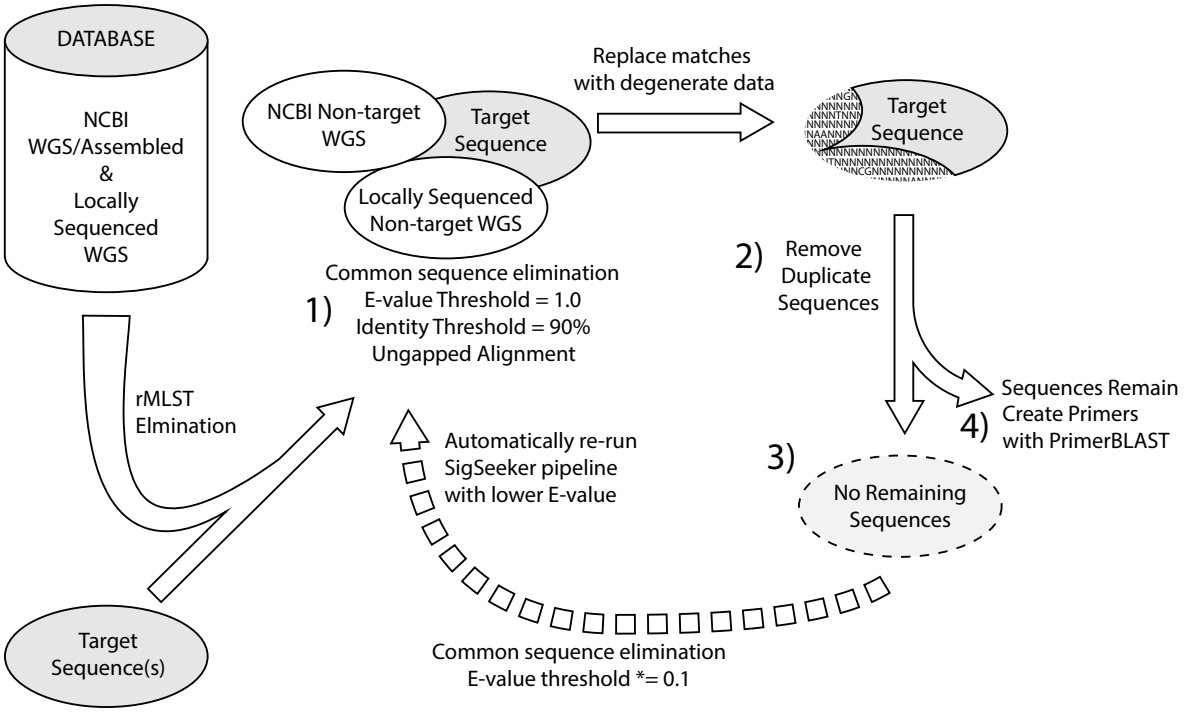

Figure 2.1: Visual representation of SigSeekr elements of database building with ribosomal multilocus sequence typing assay (rMLST) and common sequence elimination using Basic Local Alignment Tool nucleotide (BLASTn). The default parameters are E-value of 1 to start (subsquently reduced by an order of magnitude) with a percent identity at 90% and a target length at 100 throughout. The output is any unmatched sequences of the search (these constitute a unique sequence) then Primer-BLAST software is applied to design polymerase chain reaction (PCR) primers.

2.3.2 Algorithm design and pipeline optimization

The workflow of SigSeekr relies on BLAST. In a BLAST database search the E-value describes the expected number of hits that would occur by random chance alone; short and relatively dissimilar BLAST alignments have high E-values, whereas long BLAST alignments with high sequence identity have low E-values (Altschul et al. 1990). Therefore, lowering the E-value threshold of a BLAST database search will reduce the number of alignments returned (thereby reducing the amount of sequence data eliminated from the target genome).

The signature sequence identification segment of SigSeeker is an implementation of substring matching using BLAST with the caveat that the substring (i.e., unique sequence) is unknown and the noise (i.e., the genome) must be removed to discover it. We accomplished this in parallel with the use of multiple threads of BLASTn and memory locks during processing of the BLAST output to achieve parallelism without overwriting SigSeekr data in memory.

The SigSeekr pipeline, illustrated in Figure 2.1, uses BLASTn 2.2.29+ (Altschul et al. 1990) to eliminate target strain (query) sequences from the pool of potential signature sequences as follows: 1) query sequences with matches to the pan-genomic sequence database reporting initial E-values ≤ 1.0 and $\geq 90\%$ identity are 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) is generated; 2) repeated unique sequences were eliminated from the string of query sequences using a fuzzy matching algorithm to prevent a duplication of signature sequences in the output and sequences below 200 base pairs are eliminated from the string to ensure suitability for PCR amplification; 3) when no sequences are returned in the string, the process is recursively iterated using lower E-values until at least one is found. 4) The resulting string, containing the least common sequence(s), can then be used to query the entire pan-genomic database (including target strain) as a quality control.

The software was implemented in Python v2.7.3 with the use of Biopython application programming interface v1.64 (Cock et al. 2009) BlastCommandLine, SearchIO and SeqIO and optimized to prevent excessive I/O to the hard disk and keep the majority of information in memory. The memory usage experienced a maximum of 2GB with the *E. coli* database. An average run time of 15 minutes was found using a Xeon E3-1650v2 (6 cores at 3.5 GHz) workstation and 64GB of RAM.

2.4 Evaluation

A SigSeekr validation study was previously described for the development of a specific polymerase chain reaction (PCR) method targeting the lab control strain *E. coli* O157:H7 (ATCC 35150 NCBI: JYIO00000000.1) (Knowles et al. 2015). To demonstrate the broader applicability of SigSeekr to other bacterial strains, signature sequences were identified in three other shiga-toxigenic *E. coli* and one *S. enterica* target strains to design specific PCR methods targeting each strain Tables 2.1–2.4 . Primer pairs were designed using Primer-BLAST (Ye et al. 2012), and the specificity of each PCR was verified with panels of closely related bacteria. For strains *E. coli* O157:H7 (ATCC 35150), *E. coli* O157:H7 (NCTC 12900) and *E. coli* O111:NM (98-8338), it was possible to design primers producing 325, 191 and 183 bp amplicons, respectively. The specificity of these primers was verified against a panel of 82 phenotypically related *E. coli* strains, including 5 serogroup O111 strains and 51 serogroup O157 strains. In each case, only the target strain produced an amplicon with its specific primer pair, demonstrating the high degree of specificity of the primers for their target DNA sequences in the intended strain, and the absence of these unique sequences from other strains (Table 2.5). For *S. enterica* Mishmarhaemek (NCBI: JZUU00000000.1 (Cooper et al. 2015)) a combination of two signature sequences was required to distinguish it from other phenotypically similar salmonellae in the reference *Salmonella* database, culminating in a multiplex PCR using 2 sets of primer pairs producing 554 and 150 bp amplicons with the target strain, but not with 99 non-target *S. enterica* or 30 non-*Salmonellae* strains (Table 2.6). In Table 2.6, 0/129 and 1/129 non-target strains produced the 554 bp and 150 bp amplicon respectively. Primer-BLAST (Ye et al. 2012) predicted one other *S. enterica* in the GenBank database would produce the 554 bp amplicon, this strain was not available for *in vitro* evaluation. These results indicate that multiple sequences identified in a target strain can be combined to achieve a unique PCR signature providing a high level of specificity for the identification of the strain when a single unique signature sequence cannot be found.

2.5 Conclusion

We have shown that SigSeekr can be used to identify DNA sequences found uniquely in a given target strain or a group of closely related strains. Such unique sequences can be used individually or combination for greater specificity, to design primers for single

or multiplex polymerase chain reaction (PCR) methods enabling specific detection of a strain of interest.

2.6 Acknowledgement

The authors thank Paul Manninger, Ashley Cooper and Catherine Carrillo, Mylène Deschênes, Martine Gauthier, George Huszczynski, Austin Markell, and Rachel Hitayezu for technical assistance with polymerase chain reaction (PCR) analyses and sequencing bacterial strains used in this study.

2.7 Supplemental Information

Table 2.1: Sequences of oligonucleotide primers used within this work

Strain	Code[a]	Sequence ($5' \rightarrow 3'$)	Amplicon size (bp)
S. enterica Mishmarhaemek	1602-150 F	ATCGATGGTGCCTTCGGC	150
	1602-150 R	AAAAGCGGGCAAAACAAAAGG	
	1602-554 F	GGAAACCAACCTCAGTATGT	554
	1602-554 R	CATACCCGCTTTCATCACTA	
E. coli ATCC35150	795-325-F	TCAAAGCCACTCTGTAGGGAA	325
	795-325-R	AGCTGCAGATTACGATCCCC	
E. coli NCTC12900	811-191-F	CATATCTGATACTTTGGTTCCCACT	191
	811-191-R	TAAGCGGCTGCTGACCT	
E. coli 98-8338	714-183-F	GTCGCGAGACATTACGACAG	183
	714-183-R	CCGGGGCGGTTCAAGAG	

[a]Designed for this study

Table 2.2: polymerase chain reaction (PCR) master mix preparation

Reagents	Volume (µL)	Final concentration
DNase Free Water	17.875	
10 x PCR Buffer [a] (15 mM $MgCl_2$)	2.5	1 x (1.5mM $MgCl_2$)
dNTPs mix (10 mM ea.)	0.5	0.2mM ea.
Primers F/R[b] (5 µM ea.)	1	0.2 µM ea.
Qiagen HotStarTaq (5 U/µL)	0.125	0.025 U/µL
Template[c]	2.5	
Final volume	25	

[a] This reagent is included with the Qiagen HotStar Taq polymerase Kit.

[b] To reduce the number of pipetting steps when preparing the PCR master mix the forward and reverse primer pair for a given target may be mixed in equimolar amounts as follows: Mix 25 µl of each the 100 µM forward and reverse primer stocks for a given primer pair (F+R). In this mix the primers are now at a concentration of 50 µM. Store the 50µl aliquots of F+R mix at -20^{o}C. When preparing the PCR master mix, dilute the 50 µM aliquot of F+R primer mix 1:10 in 1X TE, to a working concentration of 5 µM (e.g. 5 µl F+R primer mix added to 45 µl 1X TE).

[c] DNA template is prepared by suspending a colony in 100 µL of 1% TritonX-100. Heat 10 min at 100°C.

Table 2.3: *E. coli* PCR cycling program

Cycle	Temperature	Time
1. Initial denaturation	95°C	10 min
2. 40 Cycle of:		
Denaturation	95°C	45 sec
Annealing	53°C	45 sec
Elongation	72°C	45 sec
3. Final elongation	72°C	10 min
4. Final cooling	10°C	Hold

Table 2.4: *S. enterica* PCR cycling program

Cycle	Temperature	Time
1. Initial denaturation	94°C	15 min
2. 35 Cycle of:		
Denaturation	94°C	30 sec
Annealing	55°C	30 sec
Elongation	72°C	90 sec
3. Final elongation	72°C	10 min
4. Final cooling	10°C	Hold

Table 2.5: Bacterial strains used in the evaluation of the *E. coli* PCR[a]

Organism	Culture collection no.[b]	Source[c]	PCR reactivity of *E.coli* strains		
			ATCC 35150 (325 bp)	*NCTC 12900 (191 bp)*	*98-8338 (183 bp)*
E. coli O157:H7	OLC-469	S. Read	-	-	-
E. coli O157:H7	OLC-470	S. Read	-	-	-
E. coli O157:H7	OLC-471	S. Read	-	-	-
E. coli O157:H7	OLC-472	S. Read	-	-	-
E. coli O157:H7	OLC-473	S. Read	-	-	-
E. coli O157:H7	OLC-474	S. Read	-	-	-
E. coli O157:H7	OLC-475	S. Read	-	-	-
E. coli O157:H7	OLC-476	S. Read	-	-	-
E. coli O157:H7	OLC-477	S. Read	-	-	-
E. coli O157:H7	OLC-478	S. Read	-	-	-
E. coli O157:H7	OLC-479	S. Read	-	-	-
E. coli O157:H7	OLC-480	S. Read	-	-	-
E. coli O157:H7	OLC-481	S. Read	-	-	-
E. coli O157:H7	OLC-482	S. Read	-	-	-
E. coli O157:H7	OLC-483	S. Read	-	-	-
E. coli O157:H7	OLC-484	S. Read	-	-	-
E. coli O157:H7	OLC-718	M. Gilmour	-	-	-
***E. coli* O157:H7 NalR**	**OLC-795**	**OLC**	+	-	-
E. coli O157:H7	OLC-796	OLC	-	-	-
***E. coli* O157:H7 (ATCC 35150)**	**OLC-797**	ATCC 35150	+	-	-
E. coli O157:H7	OLC-803	SHY	-	-	-
E. coli O157:H7	OLC-804	SHY	-	-	-
E. coli O157:H7	OLC-805	SHY	-	-	-
E. coli O157:H7	OLC-806	SHY	-	-	-

Organism	Culture collection no.[b]	Source[c]	PCR reactivity of *E.coli* strains		
			ATCC 35150 (325 bp)	*NCTC 12900 (191 bp)*	*98-8338 (183 bp)*
E. coli O157:H7	OLC-807	BUR	-	-	-
E. coli O157:H7	OLC-808	BUR	-	-	-
E. coli O157:H7	OLC-809	G. Bezanson	-	-	-
E. coli O157:H7	OLC-810	G. Bezanson	-	-	-
***E. coli* O157:H7 (NCTC 12900)**	**OLC-811**	OLC	-	+	-
E. coli O157:H7	OLC-996	F. Scheutz	-	-	-
E. coli O157:H7	OLC-1042	F. Scheutz	-	-	-
E. coli O157:H7	GTA-EHEC6	Beef trim	-	-	-
E. coli O157:H7	GTA-EHEC7	Ground beef	-	-	-
E. coli O157:H7	GTA-EHEC8	Beef trim	-	-	-
E. coli O157:H7	GTA-EHEC9	Beef trim	-	-	-
E. coli O157:H7	GTA-EHEC10	Beef burger	-	-	-
E. coli O157:H7	GTA-EHEC13	Beef trim	-	-	-
E. coli O157:H7	GTA-EHEC40	Beef trim	-	-	-
E. coli O157:H7	GTA-EHEC42	Veal liver	-	-	-
E. coli O157:H7	GTA-EHEC46	Veal liver	-	-	-
E. coli O157:H7	GTA-EHEC47	Walnuts	-	-	-
E. coli O157:H7	GTA-EHEC50	Frozen processed raw beef patties	-	-	-
E. coli O157:H7	GTA-EHEC52	Frozen processed raw beef patties	-	-	-

Organism	Culture collection no.[b]	Source[c]	PCR reactivity of *E.coli* strains		
			ATCC 35150 (325 bp)	*NCTC 12900 (191 bp)*	*98-8338 (183 bp)*
E. coli O157:H7	GTA-EHEC55	Frozen processed raw beef patties	-	-	-
E. coli O157:H7	GTA-EHEC68	Ground beef	-	-	-
E. coli O157:H7	GTA-EHEC70	Frozen spiced beef burger	-	-	-
E. coli O157:H7	GTA-EHEC73	Frozen spiced beef burger	-	-	-
E. coli O157:H7	GTA-EHEC75	Frozen spiced beef burger	-	-	-
E. coli O157:H7	GTA-EHEC81	Frozen spiced beef burger	-	-	-
E. coli O157:H7	GTA-EHEC178	Veal cheek meat	-	-	-
E. coli O157:H7	GTA-EHEC179	Beef burger	-	-	-
E. coli O26:H11	OLC-464	S. Read	-	-	-
E. coli O26:H11	OLC-465	S. Read	-	-	-
E. coli O26:H11	OLC-466	S. Read	-	-	-
E. coli O26:H11	OLC-725	M. Gilmour	-	-	-
E. coli O26:H11	OLC-731	M. Gilmour	-	-	-
E. coli O45:H2	OLC-716	M. Gilmour	-	-	-
E. coli O103:H2	OLC-679	M. Gilmour	-	-	-
E. coli O103:H25	OLC-680	M. Gilmour	-	-	-
E. coli O103:H2	OLC-711	M. Gilmour	-	-	-
E. coli O103:H2	OLC-712	M. Gilmour	-	-	-
E. coli O103:H25	OLC-727	M. Gilmour	-	-	-
E. coli O103:H11	OLC-728	M. Gilmour	-	-	-

Organism	Culture collection no.[b]	Source[c]	PCR reactivity of *E.coli* strains *ATCC 35150 (325 bp)*	*NCTC 12900 (191 bp)*	*98-8338 (183 bp)*
E. coli O111:H11	OLC-455	S. Read	-	-	-
E. coli O111:H11	OLC-457	S. Read	-	-	-
E. coli O111:H11	OLC-629	R. Johnson	-	-	-
***E. coli* O111:NM (strain 98-8338)**	**OLC-714**	M. Gilmour	-	-	+
E. coli O111:NM	OLC-715	M. Gilmour	-	-	-
E. coli O121:NM	OLC-677	M. Gilmour	-	-	-
E. coli O121:H19	OLC-678	M. Gilmour	-	-	-
E. coli O121:H19	OLC-710	M. Gilmour	-	-	-
E. coli O121:H19	OLC-789	OLC	-	-	-
E. coli O121:NM	OLC-791	OLC	-	-	-
E. coli O121:NM	OLC-792	OLC	-	-	-
E. coli O121:H7	OLC-982	H. Tabor	-	-	-
E. coli O145:NM	OLC-675	M. Gilmour	-	-	-
E. coli O145:NM	OLC-676	M. Gilmour	-	-	-
E. coli O145	OLC-983	H. Tabor	-	-	-
E. coli O5:NM	OLC-467	S. Read	-	-	-
E. coli O5:NM	OLC-468	S. Read	-	-	-
E. coli O119:H25	OLC-667	OLC	-	-	-
E. coli O76:H19	OLC-669	OLC	-	-	-
E. coli O115:H18	OLC-708	M. Gilmour	-	-	-
E. coli O5:NM	OLC-709	M. Gilmour	-	-	-
E. coli O55:H7	OLC-717	M. Gilmour	-	-	-
E. coli O91:H21	OLC-720	M. Gilmour	-	-	-
E. coli O117:H7	OLC-723	M. Gilmour	-	-	-
E. coli O113:H21	OLC-726	M. Gilmour	-	-	-
E. coli O6:H34	OLC-729	M. Gilmour	-	-	-
E. coli O146:H21	OLC-730	M. Gilmour	-	-	-
E. coli O177:NM	OLC-732	M. Gilmour	-	-	-

Organism	Culture collection no.[b]	Source[c]	PCR reactivity of *E.coli* strains		
			ATCC 35150 (325 bp)	*NCTC 12900 (191 bp)*	*98-8338 (183 bp)*
E. coli O85:H1	OLC-733	M. Gilmour	-	-	-
E. coli O48:H21	OLC-994	F. Scheutz	-	-	-
E. coli O174:H21	OLC-995	F. Scheutz	-	-	-
E. coli O118:H12	OLC-997	F. Scheutz	-	-	-
E. coli O73:H18	OLC-998	F. Scheutz	-	-	-
E. coli O2:H25	OLC-999	F. Scheutz	-	-	-
E. coli O8:K85ab:Hrough	OLC-1000	F. Scheutz	-	-	-
E. coli O128ac:H2	OLC-1001	F. Scheutz	-	-	-
E. coli O174:H8	OLC-1002	F. Scheutz	-	-	-
E. coli O139:K12:H1	OLC-1003	F. Scheutz	-	-	-
Shigella sonnei	OLC-24	ATCC 29930	-	-	-
Klebsiella pneumonia	OLC-27	ATCC 13883	-	-	-
Proteus vulgaris	OLC-28	ATCC 13315	-	-	-
Serratia marcescens	OLC-30	ATCC 13880	-	-	-
Proteus aeruginosa	OLC-32	ATCC 10145	-	-	-
Citrobacter freundii	OLC-36	ATCC 8090	-	-	-
Salmonella enterica ser. Enteritidis	OLC-47	ATCC 13076	-	-	-
S. enterica ser. Typhimurium	OLC-59	ATCC 14028	-	-	-
Bacillus cereus	OLC-146	ATCC 14579	-	-	-
Enterobacter faecalis	OLC-147	ATCC 19433	-	-	-
B. subtilis	OLC-161	ATCC 6051	-	-	-
Staphylococcus epidermidis	OLC-244	ATCC 12228	-	-	-

Organism	Culture collection no.[b]	Source[c]	PCR reactivity of *E.coli* strains		
			ATCC 35150 (325 bp)	*NCTC 12900 (191 bp)*	*98-8338 (183 bp)*

[a]Bacterial lysates were subjected to the 795 PCR procedure as described in Methods, and PCR products were analyzed using the Flashgel electrophoresis system. PCR results for the 325-bp and IAC amplicons, respectively, are indicated as + (band present) or – (band absent)

[b]Culture collection reference number. OLC, Ottawa Laboratory-Carling culture collection; GTA, Greater Toronto Area Laboratory culture collection. OLC-795 and the parent strain from which it was derived are indicated in bold.

[c]Strains were obtained from various culture collections as follows: M. Gilmour and H. Tabor, National Microbiology Laboratory, Public Health Agency of Canada; R. Johnson and S. Read, Laboratory for Foodborne Zoonoses, Public Health Agency of Canada; G. Bezanson, Agriculture and Agri-Food Canada; St-Hyacinthe Laboratory (SHY), Canadian Food Inspection Agency; Burnaby Laboratory (BUR), Canadian Food Inspection Agency; F. Scheutz, Unit of Foodborne Infections, Statens Serum Institut.

Table 2.6: Bacterial strains used in evaluation of *Samonella enterica* subsp. enterica serovar Mishmarhaemek multiplex PCR[a]

Organism	Culture collection no.[b]	Source[c]	PCR reactivity of primer pairs		
			554 bp	*150 bp*	*Both*
S. typhimurium	OLC-33	ATCC 19585	-	-	-
S. Montevideo	OLC-34	ATCC 8387	-	-	-
S. salamae	OLC-37	ATCC 43972	-	-	-
S. arizonae	OLC-38	ATCC 13314	-	-	-
S. diarizonae	OLC-39	ATCC 43973	-	-	-
S. houtenae	OLC-40	ATCC 43974	-	-	-
S. bongori	OLC-41	ATTC 43975	-	-	-
S. indica	OLC-42	ATCC 43976	-	-	-
S. agona	OLC-43	CFIA Ottawa	-	-	-
S. braenderup	OLC-44	CFIA Ottawa	-	-	-
S. blockley	OLC-45	CFIA Ottawa	-	-	-
S. haardt	OLC-46	CFIA Ottawa	-	-	-
S. enteritidis	OLC-47	ATCC 13076	-	-	-
S. fresno	OLC-48	CFIA Ottawa	-	-	-
S. london	OLC-49	CFIA Ottawa	-	-	-
S. newington	OLC-50	ATCC 29628	-	-	-
S. thomasville	OLC-51	CFIA Ottawa	-	-	-
S. senflenberg	OLC-52	ATCC 8400	-	-	-
S. rubislaw	OLC-53	ATCC 10717	-	-	-
S. poona	OLC-54	CFIA Ottawa	-	-	-
S. worthington	OLC-55	ATCC 9607	-	-	-
S. cerro	OLC-56	ATCC - 10723	-	-	-
S. alachua	OLC-57	CFIA Ottawa	-	-	-
S. johannesburg	OLC-58	CFIA Ottawa	-	-	-
S. typhimurium	OLC-59	ATCC 14028	-	-	-
S. paratyphi A	OLC-60	ATCC 9150	-	-	-

Organism	Culture collection no.[b]	Source[c]	PCR reactivity of primer pairs		
			554 bp	*150 bp*	*Both*
S. bleadon	OLC-61	CFIA Ottawa	-	-	-
S. minnesota	OLC-62	ATCC 9700	-	-	-
S. westhampton	OLC-94	CFIA St-Hyacinthe	-	-	-
S. abony	OLC-95	CFIA St-Hyacinthe	-	-	-
S. enteriditis	OLC-98	CFIA Ottawa	-	-	-
S. vom	OLC-210	CFIA Ottawa	-	-	-
S. reading	OLC-236	CFIA Ottawa	-	-	-
S. ealing	OLC-359	Dr.Poppe, PHAC-LFZ	-	-	-
S. arizonae	OLC-421	Dr.Poppe, PHAC-LFZ	-	-	-
S. urbana	OLC-495	CFIA Ottawa	-	-	-
S. iverness	OLC-496	CFIA Ottawa	-	-	-
S. kahla	OLC-497	CFIA Ottawa	-	-	-
S. bredeney	OLC-499	CFIA Ottawa	-	-	-
S. infantis	OLC-500	CFIA Ottawa	-	-	-
S. hadan	OLC-501	CFIA Ottawa	-	-	-
S. heidelberg	OLC-502	CFIA Ottawa	-	-	-
S. kentucky	OLC-504	CFIA Ottawa	-	-	-
S. ohio	OLC-507	CFIA Ottawa	-	-	-
S. oranienburg	OLC-508	CFIA Ottawa	-	-	-
S. blukwa	OLC-509	CFIA Ottawa	-	-	-
S. binza	OLC-510	CFIA Ottawa	-	-	-
S. Zwickau	OLC-512	CFIA Ottawa	-	-	-
S. enteritidis	OLC-554	CFIA Ottawa	-	-	-
S. enteritidis	OLC-555	CFIA Ottawa	-	-	-
S. typhimurium	OLC-595	Dr.Poppe, PHAC-LFZ	-	-	-

Organism	Culture collection no.[b]	Source[c]	PCR reactivity of primer pairs		
			554 bp	*150 bp*	*Both*
S. typhimurium	OLC-649	Dr.Devenish, CFIA OLF	-	-	-
S. kiambu	OLC-693	Dr.Devenish, CFIA OLF	-	-	-
S. 1:Rough-O:1:26	OLC-701	Dr.Devenish, CFIA OLF	-	-	-
S. gallinarum	OLC-753	ATCC 9184	-	-	-
S. malawi	OLC-767	CFIA Calgary	-	-	-
S. maregrosso	OLC-768	CFIA Calgary	-	-	-
S. berta	OLC-799	ATCC 8392	-	-	-
S. Indica	OLC-801	ATCC 43976	-	-	-
S. rissen	OLC-1146	CFIA Ottawa	-	-	-
S. sandiego	OLC-1147	CFIA Ottawa	-	-	-
S. thompson	OLC-1148	CFIA Ottawa	-	-	-
S. pomona	OLC-1150	CFIA Ottawa	-	-	-
S. tennessee	OLC-1153	CFIA Ottawa	-	-	-
S. anatum	OLC-1156	CFIA Ottawa	-	-	-
S. mbandaka	OLC-1157	CFIA Ottawa	-	-	-
S. schwarzengrund	OLC-1161	CFIA Ottawa	-	-	-
S. newport	OLC-1164	CFIA Ottawa	-	-	-
S. typhimurium	OLC-1165	CFIA Ottawa	-	-	-
S. anatum	OLC-1168	CFIA Ottawa	-	-	-
S. I:6,7,14:-:1,5	OLC-1172	CFIA Ottawa	-	-	-
S. mishmar haemek	**OLC-1602**	CFIA Ottawa	+	+	+
S. I:rough -0:-:-	OLC-1694	CFIA Ottawa	-	-	-
S. stanley	OLC-1709	CFIA Ottawa	-	-	-
S. havana	OLC-1712	CFIA Ottawa	-	-	-
S. livingstone	OLC-1714	CFIA Ottawa	-	-	-
S. meleagridis	OLC-1717	CFIA Ottawa	-	-	-
S. cubana	OLC-1718	CFIA Ottawa	-	-	-

Organism	Culture collection no.[b]	Source[c]	PCR reactivity of primer pairs		
			554 bp	*150 bp*	*Both*
S. orion	OLC-1728	CFIA Ottawa	-	-	-
S. muenster	OLC-1729	CFIA Ottawa	-	-	-
S. 1:4, 12:b:-	OLC-1754	CFIA Ottawa	-	-	-
S. kaimbu	OLC-1757	CFIA Ottawa	-	-	-
S. amsterdam	OLC-1763	CFIA Ottawa	-	-	-
S. hadar	OLC-1829	CFIA Ottawa	-	-	-
S. 1:4, 12:-:1, 7	OLC-1849	CFIA Ottawa	-	-	-
S. adelaide	OLC-1856	CFIA Ottawa	-	-	-
S. 1:10:eh:-	OLC-1891	CFIA Ottawa	-	-	-
S. l:8,20:-:z6	OLC-1900	CFIA Ottawa	-	-	-
S. hartford	OLC-1943	CFIA Ottawa	-	-	-
S. putten	OLC-1945	CFIA Ottawa	-	-	-
S. I:10:-:1,5	OLC-1947	CFIA Ottawa	-	-	-
S. 1:43:24,223:-	OLC-1950	CFIA Ottawa	-	+	-
S. l:4, 12:d:-	OLC-1951	CFIA Ottawa	-	-	-
S. llandoff	OLC-2000	CFIA Ottawa	-	-	-
S. 1:8,20:210:-	OLC-2040	CFIA Ottawa	-	-	-
S. 1:6,7:b:-	OLC-2044	CFIA Ottawa	-	-	-
S. I:6	OLC-2057	CFIA GTA	-	-	-
S. Saintpaul	OLC-2058	CFIA GTA	-	-	-
S. Freetown	OLC-2065	CFIA GTA	-	-	-
S. II:48:z10	OLC-2068	CFIA GTA	-	-	-
Listeria innocua	OLC-4	ATCC 33090	-	-	-
Listeria ivanovii	OLC-5	ATCC19119	-	-	-
Listeria seeligeri	OLC-11	CFIA Ottawa	-	-	-
Listeria welshimeri	OLC-12	CFIA Ottawa	-	-	-
Listeria grayi	OLC-13	CFIA Ottawa	-	-	-
Listeria murrayi	OLC-14	CFIA Ottawa	-	-	-
Listeria monocytogenes	OLC-15	CFIA Ottawa	-	-	-

Organism	Culture collection no.[b]	Source[c]	PCR reactivity of primer pairs		
			554 bp	*150 bp*	*Both*
Shigella sonnei	OLC-24	ATCC 29930	-	-	-
Klebsiella pneumoniae	OLC-27	ATCC 13883	-	-	-
Proteus vulgaris	OLC-28	ATCC 13315	-	-	-
Serratia marcescens	OLC-30	ATCC 13880	-	-	-
Pseudomonas aeruginosa	OLC-32	ATCC 10145	-	-	-
Staphylococcus aureus	OLC-35	ATCC 12600	-	-	-
Citrobacter freundii	OLC-36	ATCC 8090	-	-	-
Yersinia enterocolitica	OLC-106	M.Skurnik	-	-	-
Bacillus cereus	OLC-146	ATCC 14579	-	-	-
Enterococcus Farcalis	OLC-147	ATCC 19433	-	-	-
Bacillus subtilis	OLC-161	ATCC 6051	-	-	-
Staphylococcus epidermidis	OLC-244	ATCC 12228	-	-	-
Escherichia coli O111:H11	OLC-455	Dr. Read, PHAC-LFZ	-	-	-
Escherichia coli O26:H11	OLC-464	Dr. Read, PHAC-LFZ	-	-	-
Hafnia alvei	OLC-517	CFIA Ottawa	-	-	-
Morganella morganii	OLC-641	CFIA Ottawa	-	-	-
Escherichia coli O145:NM	OLC-675	Dr Gilmour, PHAC-NML	-	-	-
Escherichia coli O103:H2	OLC-679	Dr Gilmour, PHAC-NML	-	-	-
Escherichia coli O121:H19	OLC-710	Dr Gilmour, PHAC-NML	-	-	-
Escherichia coli O45:H2	OLC-716	Dr.Gilmour, PHAC-NML	-	-	-
Escherichia coli O157:H7	OLC-797	CFIA Ottawa	-	-	-

Organism	Culture collection no.[b]	Source[c]	PCR reactivity of primer pairs		
			554 bp	*150 bp*	*Both*
Escherichia coli	OLC-1676	CFIA-GTA	-	-	-
Escherichia coli	OLC-1678	CFIA-GTA	-	-	-

[a]Colony from Nutrient or BHI Agar suspended in 100uL of 1% Tritonx-100 follow by heating at 100°C for 10 min. The extracts were cooled to room temperature and used immediately in the Salmonella mishmar procedure.

[b]Ottawa Laboratory Carling culture collection reference.

[c]Strains were obtained from various culture collections: ATCC; Ottawa Laboratory – Carling and Fallowfield, St-Hyacinthe Laboratory, Calgary Laboratory; GTA Laboratory (Canadian Food Inspection Agency); Dr.Poppe and Dr. Read (Laboratory for Foodborne Zoonoses, Public Health Agency of Canada); M.Skurnik (University of Helsinki-Finland); Dr Gilmour (National Microbiology Laboratory, Public Health Agency of Canada).

[d]Reactions are scored qualitatively as follows: Positive (presence of a band at 150bp and 554 bp on a gel); Negative (no band)

Chapter 3

Polymerase chain reaction for the specific detection of an *Escherichia coli* O157:H7 laboratory control strain

Preface

Copyright

Contributions of Collaborators

The following chapter is written as a manuscript published by Michael Knowles, Dominic Lambert, George Huszczynski, Martine Gauthier, and Burton W. Blais titled "Polymerase chain reaction for the specific detection of an *Escherichia coli* O157:H7 laboratory control strain" in the *Journal of Food Protection* with the consent to republish this work granted by the International Association for Food Protection (IAFP). I, Michael Knowles, assembled all experimental materials and performed all of the experiments and developed algorithms in the following manuscript with the exception of the following:

- Collection of *Escherchia coli* and *Salmonella enterica* isolates (B. W. Blais, and D.

Lambert)

- Sequencing of bacteria (M. Deschênes, and P. Manniger)
- Independent testing of experimental results (G. Huszczynski and M. Gauthier)

Coauthors Dominic Lambert, George Huszczynski, Martine Gauthier, Burton W. Blais and I, Michael Knowles, wrote the original draft of this paper and all revised versions.

3.1 Abstract

Control strains of bacterial pathogens such as *Escherichia coli* O157:H7 are commonly processed in parallel with test samples in food microbiology laboratories as a quality control measure assuring the satisfactory performance of materials used in the analytical procedure. Before positive findings can be reported for risk management purposes, analysts must have a means of verifying that pathogenic bacteria (e.g., *E. coli* O157:H7) recovered from test samples are not due to inadvertent contamination with the control strain routinely handled in the laboratory environment. Here we report on the application of an in-house bioinformatic pipeline for the identification of unique genomic signature sequences in the development of specific oligonucleotide primers enabling the identification of a common positive control strain, *E. coli* O157:H7 (ATCC 35150), using a simple polymerase chain reaction (PCR) procedure.

3.2 Introduction

Canadian Food Inspection Agency (CFIA) laboratories conducting analyses of food inspection samples to determine the presence of the priority food pathogen *Escherichia coli* O157:H7 employ a validated polymerase chain reaction (PCR) technique (MFLP-22) to identify colony isolates recovered on plating media (Blais, Martinez-Perez, Huszczynski, et al. 2013). This technique features the Cloth-based Hybridization Array System (CHAS) platform in which PCR products resulting from the amplification of key target genes by multiplex PCR are detected by hybridization with an array of immobilized capture probes on a polyester cloth support (Martinez-Perez and Blais 2010). The entire procedure can be completed in less than six hours, that is, on the same day that colonies are observed on primary isolation plates. This is a marked advantage over traditional biochemical techniques, which require two days or more to complete, especially in

the context of foodborne illness outbreak investigations, where timeliness in producing evidence supporting risk management decisions is paramount.

Food microbiology testing laboratories may analyze hundreds of food samples for the presence of different pathogens every week. A common quality control measure employed in these laboratories is the parallel processing of a positive control bacterial strain with each set of samples. When a sample tests positive for the presence of a specific pathogen, laboratories must ensure that the positive result is not due to contamination of the test sample by the control strain being handled in the same environment. Previously, a nalidixic acid-resistant (Nal^R) mutant (strain OLC-795) derived from the common *E. coli* O157:H7 control strain (ATCC 35150) was developed to serve as distinguishable control strain for routine use in CFIA food microbiology testing laboratories (Blais, Martinez-Perez, Gauthier, et al. 2008). This particular strain is distinguishable because of its ability to grow vigorously on plating media containing nalidixic acid, distinguishing it from other *E. coli* O157:H7 isolates which generally lack this property. Thus, before a positive result for *E. coli* O157:H7 can be reported for risk assessment purposes it is necessary to demonstrate the inability of the food isolate to grow on plating media containing nalidixic acid, a process requiring a minimum of 22 h. This diminishes the principal advantage of rapid colony identification techniques such as MFLP-22, since the results of analysis cannot be reported at the time of identification, but instead must be deferred until growth on nalidixic acid has been assessed.

The process of determining identity with the laboratory control strain would be expedited if the same colony lysate used for the CHAS procedure in MFLP-22 could also be tested in parallel by a simple PCR method targeting sequences unique to the control strain. In this manner, contamination of the sample by the laboratory control strain could be discounted within the same time frame required to complete MFLP-22, enabling same day reporting of results and realizing the full advantage of the rapid colony identification method.

We have developed a bioinformatic pipeline for the identification of unique genomic signature sequences in bacterial strains, which we have named SigSeekr. In this approach, whole-genome sequence data for the subject strain is analyzed by reductive Basic Local Alignment Tool (BLAST) (Altschul et al. 1990; C.-C. Ho et al. 2012) against extensive databases of in-house and public genome sequences of closely related organisms in order to identify candidate strain-specific sequences from which primers can be derived. Essentially, BLASTn reports an E-value, describing the expected number of "hits" that would occur by random chance alone; short and relatively dissimilar BLAST matches have high

E-values, and long BLAST matches with high sequence identity have low E-values. The E-values generated during a BLASTn search can therefore be used to identify sequences homologous to those found in the subject strain, and remove them from the dataset. This procedure was applied in the analysis of the positive control strain regularly employed at CFIA, OLC-795, in order to design PCR primers that are highly specific for this strain for deployment to food microbiology testing laboratories. Because the parent strain from which OLC-795 was derived is commonly available from the American Type Culture Collection (ATCC 35150), we also demonstrate the suitability of this PCR procedure for the identification of that strain, making the approach described herein broadly applicable to all food microbiology testing laboratories.

3.3 Materials and Methods

3.3.1 Bacterial growth and genomic DNA extraction.

Bacterial strains (Table 3.1) were grown on nutrient agar (Difco, Becton, Dickinson & Co) overnight (14-16 h) at 37°C, and genomic DNA was extracted from single colonies using the Maxwell 16 Cell LEV DNA Purification Kit (Promega, Madison, WI). Genomic DNA was quantified using the Quant-iT High-Sensitivity DNA Assay Kit (Life Technologies Inc., Burlington, ON).

3.3.2 Whole-genome sequencing and assembly.

Two hundred and fifty *E.coli* strains from the Ottawa Laboratory (Carling) culture collection including *E. coli* O157:H7 ATCC 35150 (*viz.* OLC-797) and its nalidixic acid-resistant mutant derivative (OLC-795 NCBI: JYIO00000000) were sequenced. Sequencing libraries were constructed from 1 ng of genomic DNA using the Nextera XT DNA Sample Preparation Kit (Illumina, Inc., San Diego, CA) and the Nextera XT Index Kit (Illumina, Inc., San Diego, CA). Paired-end sequencing of sixteen multiplexed samples was performed on the Illumina MiSeq Platform (Illumina, Inc., San Diego, CA) using a 300 cycle MiSeq Reagent Kit v2 and 1746204 paired-end reads were generated. Sequencing errors in reads were corrected using Quake (version 0.3) with a k-mer size of 15 (Kelley et al. 2010). Sequencing reads were assembled using SPAdes 3.1.1 (Bankevich et al. 2012) with a k-mer range (21, 33, 55, 77, 99, 127) and error correction parameters. Assembly of the OLC-795 strain sequence resulted in 143 contigs with an average of

52.93-fold genome coverage. The combined length of the draft genome size is 5.32 Mbp with a G+C content of 50.28%.

3.3.3 *E. coli* pan-genome sequence database

All publicly available *E. coli* genome entries (i.e., 1600 unannotated genomic contigs, closed assembled pseudomolecules (chromosome-like DNA fragment), and plasmids were retrieved from NCBI GenBank using the Biopython application programming interface (Cock et al. 2009) Entrez module (NCBI 2013). GenBank sequences were combined with the locally generated sequences to form the database used by SigSeekr. Note that gene annotation was not required for present purposes, since the *in silico* subtraction conducted by SigSeekr is applicable to the analysis of genome fragments. To ensure that all genome entries were represented only once in the database, an *in silico* ribosomal multilocus sequence typing assay (rMLST), previously shown to be sufficiently discriminatory for differentiation of isolates at the strain level (Jolley et al. 2012), was used to identify possible identical entries. Entries harbouring identical ribosomal sequence types (rST) were further compared using BLAST2 (Altschul et al. 1990) and redundant entries (e.g. multiple runs of the same genome entered using slightly different names) were removed from the analysis.

3.3.4 Unique sequence identification and primer design.

The SigSeekr pipeline uses each contig individually to query the pan-genomic sequence database using BLASTn 2.2.29+ (Altschul et al. 1990) to eliminate control strain (query) sequences from the pool of potential signature sequences as follows: 1) query sequences with matches reporting initial E-values ≤ 1.0 and $\geq 90\%$ identity are removed and substituted with degenerate bases, generating a string of least common sequences linked by the degenerate bases (with the degenerate bases being ignored in subsequent BLASTn database searches, reducing redundancies and improving speed); 2) unique repeat sequences are eliminated from the string of query sequences using a second BLASTn search and a cut-off E-value of 1×10^{-40} to prevent a duplication of signature sequences in the output; 3) sequences below 200 base pairs are eliminated from the string to ensure suitability for polymerase chain reaction (PCR) amplification; 4) when no sequences are returned in the string, the process is recursively iterated using lower E-values until at least one is found. The resulting string, containing the least common sequence(s), can then be used to query the entire pan-genomic database (including the target strain) as a

quality control. When used to identify a sequence unique to OLC-795, the pipeline identified a 930-nucleotide fragment (NCBI: JYIO00000000 Contig 0017 [2776:3101]). The unique sequence was also found in its parent strain ATCC 35150. SigSeekr is available at (http://github.com/OLC-LOC-Bioinformatics/SigSeekr). Primers specific to the OLC-795 strain unique sequence were designed (Table 3.2) using Primer-BLAST (Ye et al. 2012) and obtained from a commercial supplier (Coralville, IA, USA).

3.3.5 Polymerase chain reaction procedure (795 PCR)

Bacterial colony lysates were prepared by suspension of a bacterial colony in 200 µl of 1% (v/v) Triton X-100 (Sigma-Aldrich Co., St. Louis, MO, USA), followed by heating at 100°C for 10 min. For the PCR, 5 µl of bacterial lysate was added to 45 µl of multiplex PCR mixture containing 2.5 units HotStar *Taq* and 1 x HotStar PCR buffer (Qiagen Inc., Mississauga, ON, Canada), 2.5 mM $MgCl_2$, 400 µM of each dNTP, 0.4 µM each of the 795-F and 795-R primers, and 0.1 µM each of the IAC-1 and IAC-2 primers. The IAC primers served as an internal amplification control (Leggate and Blais 2006). The PCR was carried out in a Mastercycler gradient thermal cycler (Eppendorf, Hamburg, Germany) 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) systems (following the manufacturers' instructions).

3.4 Results and Discussion

The 795 polymerase chain reaction (PCR) procedure is designed to identify false-positive results due to in-lab contamination of investigative food samples with the laboratory control strain. The procedure is intended to be used in conjunction with a rapid colony identification method such as the *E. coli* O157:H7 CHAS (Blais and Martinez-Perez 2011), which has been validated to support its use as a regulatory test method (MFLP-22) (Blais, Martinez-Perez, Huszczynski, et al. 2013) for the timely identification of presumptive colonies recovered on primary plating media using standard enrichment techniques (Anonymous 2014). Hence, the present PCR system was designed for use with a basic thermocycler to be compatible with the CHAS. Basic PCR procedures are robust and offer great flexibility for primer design and are more accommodating of

different amplicon sizes. To determine whether a colony of interest on a primary isolation plate might be the control strain (OLC-795), the colony is lysed and subjected to a PCR procedure using primers designed to amplify a 325-bp sequence unique to the control strain (plus an internal amplification control (Leggate and Blais 2006)), as determined using a novel signature sequence-finding genome analysis pipeline (SigSeekr) developed for this purpose. Production of the 325-bp PCR product may be determined by either gel or capillary electrophoresis (Fig. 1), according to user preference.

The inclusivity and exclusivity characteristics of the 795 PCR procedure were verified using a panel of 113 different bacterial strains, featuring 51 *E. coli* O157:H7 (including OLC-795 and the ATCC 35150 strain from which it was derived), 50 non-O157 *E. coli* and 12 other Gram-negative and Gram-positive bacteria (Table 3.1). The 325-bp amplicon anticipated for strain OLC-795 (and its ATCC 35150 parent) was only observed with the latter two strains (Table 3.1), and not with any of the 99 other *E. coli* (including 49 different *E. coli* O157:H7 strains), nor with the non-*E. coli* bacteria. The IAC amplicon was evident in every PCR product, confirming the validity of the negative PCR results. These results demonstrate the high degree of specificity of the 795 primers for their intended target DNA sequences in the laboratory control strain OLC-795 (and its ATCC 35150 parent), and the absence of these unique sequences from other *E. coli* O157:H7 strains. The generation of the 325-bp amplicon with both OLC-795 and ATCC 35150 is consistent with the results of a comparative SNP analysis of the two strains, revealing only one base difference between the genomes located in the *gyrA* gene (*E. coli* OLC-797 (JXUS00000000)) and presumably conferring the Nal^R trait (not shown). The recent trend toward an increase in the prevalence of genetic determinants for anti-microbial resistance in *E. coli* isolated from foods and food production environments (Bonyadian et al. 2014; Jahne et al. 2014; Nawaz et al. 2015) may ultimately make the Nal^R trait less reliable as a primary means of distinguishing the control strain OLC-795 from natural foodborne isolates (Blais, Martinez-Perez, Gauthier, et al. 2008), underscoring the need for alternative approaches such as the 795 PCR method.

The present PCR analysis could be completed within 4 h, which matches the time-frame required to complete rapid colony identification procedures such as that described in MFLP-22 (Blais, Martinez-Perez, Huszczynski, et al. 2013). Thus, in a foodborne illness outbreak investigation, where timeliness in the identification of food sample contaminated with *E. coli* O157:H7 is paramount, use of the 795 PCR in conjunction with a rapid colony identification method will enable definitive identification of the target pathogen on the same day that colonies are obtained on primary isolation plates. This is

a marked advantage over the previous approach (Blais, Martinez-Perez, Gauthier, et al. 2008) in which results could not be reported to underscore risk management actions before completion of an overnight growth period to determine the nalidixic acid resistance status of an isolate.

The approach described presently was made possible by the application of next-generation sequencing technology, which now makes it feasible to elucidate whole bacterial genome sequences in a standard microbiology laboratory setting. The availability of powerful bioinformatic tools (e.g., SigSeekr) for the treatment of complex genomic datasets enables the rapid comparative analysis of a genome (e.g., strain OLC-795) to identify unique primer sequences specific to the strain of interest. There are many potentially useful applications of such an approach. For example, the state of the art in next-generation sequencing technology has reached the point where clinical isolates implicated in foodborne disease outbreaks are routinely sequenced in public health laboratories at an early stage during such events. With the application of bioinformatic tools such as SigSeekr, it should be possible to identify outbreak strain-specific marker sequences for the development of custom methods (e.g., PCR tests) that can be rapidly deployed to food testing labs conducting analyses in support of outbreak investigations. The SigSeekr approach described presently was designed to enable the identification of unique sequences within any bacterial strain of interest, provided that suitable genome databases are available for comparative analyses. Work is currently underway to develop specific PCR methods to distinguish laboratory control strains of the other major food pathogens routinely analysed at CFIA, including *Salmonella* and *Listeria monocytogenes*.

These results demonstrate that the 795 PCR is highly specific for the control strain (and the parent strain from which it was derived) currently in use at CFIA food microbiology testing laboratories. Therefore, this method is suitable for use in conjunction with MFLP-22, or any other rapid colony identification method, for same-day confirmation of isolates on primary plating media. It is suggested that the same colony lysate preparation used as the source of template DNA for MFLP-22 can also be used for the 795 PCR procedure, which can be performed in parallel. For laboratories not in possession of OLC-795 the readily obtainable ATCC 35150 strain may be used as a lab control strain in conjunction with this PCR procedure.

3.5 Acknowledgments

The authors thank Paul Manninger and Catherine Carrillo for sequencing bacterial strains used in this study, and Adam Koziol for assistance with bioinformatics analyses. We also acknowledge the federal GRDI interdepartmental Food and Water Safety project consortium (comprised of researchers from Agriculture and Agri-food Canada, the Canadian Food Inspection Agency, Environment Canada, Health Canada, the National Research Council of Canada, and the Public Health Agency of Canada) for useful discussions and contributions.

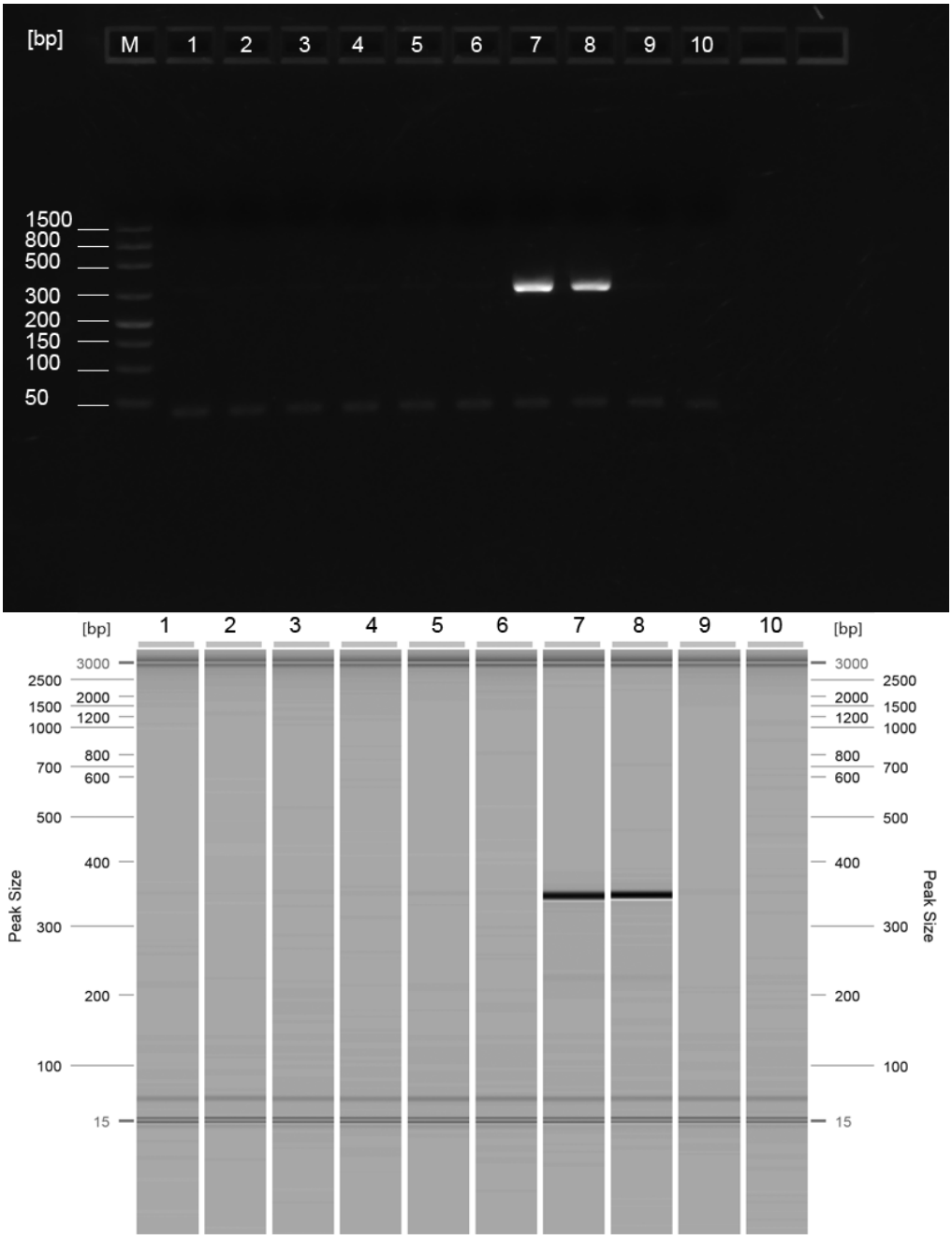

Figure 3.1: Analysis of PCR products from selected *E. coli* O157:H7 strains by electrophoresis. The 795 PCR products obtained with eight strains selected from Table 3.1 were visualized by rapid gel slab (upper panel) and capillary (lower panel) electrophoresis systems: M, molecular size ladder; 1, OLC-803; 2, OLC-804; 3, OLC-805; 4, OLC-806; 5, OLC-807; 6, OLC-808; 7, OLC-795; 8, ATCC 35150; 9, Lysis buffer blank; 10, reagent blank. The 325-bp and 40 bp IAC amplicons are indicated by block arrows.

3.6 Supplemental Information

Table 3.1: Bacterial strains used in the evaluation of the 795 polymerase chain reaction (PCR)[a]

Organism	Culture collection no.[b]	Source[c]	PCR reactivity	
			795	*IAC*
E. coli O157:H7	OLC-469	S. Read	-	+
E. coli O157:H7	OLC-470	S. Read	-	+
E. coli O157:H7	OLC-471	S. Read	-	+
E. coli O157:H7	OLC-472	S. Read	-	+
E. coli O157:H7	OLC-473	S. Read	-	+
E. coli O157:H7	OLC-474	S. Read	-	+
E. coli O157:H7	OLC-475	S. Read	-	+
E. coli O157:H7	OLC-476	S. Read	-	+
E. coli O157:H7	OLC-477	S. Read	-	+
E. coli O157:H7	OLC-478	S. Read	-	+
E. coli O157:H7	OLC-479	S. Read	-	+
E. coli O157:H7	OLC-480	S. Read	-	+
E. coli O157:H7	OLC-481	S. Read	-	+
E. coli O157:H7	OLC-482	S. Read	-	+
E. coli O157:H7	OLC-483	S. Read	-	+
E. coli O157:H7	OLC-484	S. Read	-	+
E. coli O157:H7	OLC-718	M. Gilmour	-	+
***E. coli* O157:H7 NalR**	**OLC-795**	**OLC**	+	+
E. coli O157:H7	OLC-796	OLC	-	+
***E. coli* O157:H7 (ATCC 35150)**	**OLC-797**	ATCC 35150	+	+
E. coli O157:H7	OLC-803	SHY	-	+
E. coli O157:H7	OLC-804	SHY	-	+

Organism	Culture collection no.[b]	Source[c]	PCR reactivity	
			795	*IAC*
E. coli O157:H7	OLC-805	SHY	-	+
E. coli O157:H7	OLC-806	SHY	-	+
E. coli O157:H7	OLC-807	BUR	-	+
E. coli O157:H7	OLC-808	BUR	-	+
E. coli O157:H7	OLC-809	G. Bezanson	-	+
E. coli O157:H7	OLC-810	G. Bezanson	-	+
E. coli O157:H7	OLC-811	OLC	-	+
E. coli O157:H7	OLC-996	F. Scheutz	-	+
E. coli O157:H7	OLC-1042	F. Scheutz	-	+
E. coli O157:H7	GTA-EHEC6	Beef trim	-	+
E. coli O157:H7	GTA-EHEC7	Ground beef	-	+
E. coli O157:H7	GTA-EHEC8	Beef trim	-	+
E. coli O157:H7	GTA-EHEC9	Beef trim	-	+
E. coli O157:H7	GTA-EHEC10	Beef burger	-	+
E. coli O157:H7	GTA-EHEC13	Beef trim	-	+
E. coli O157:H7	GTA-EHEC40	Beef trim	-	+
E. coli O157:H7	GTA-EHEC42	Veal liver	-	+
E. coli O157:H7	GTA-EHEC46	Veal liver	-	+
E. coli O157:H7	GTA-EHEC47	Walnuts	-	+
E. coli O157:H7	GTA-EHEC50	Frozen processed raw beef patties	-	+
E. coli O157:H7	GTA-EHEC52	Frozen processed raw beef patties	-	+
E. coli O157:H7	GTA-EHEC55	Frozen processed raw beef patties	-	+
E. coli O157:H7	GTA-EHEC68	Ground beef	-	+
E. coli O157:H7	GTA-EHEC70	Frozen spiced beef burger	-	+
E. coli O157:H7	GTA-EHEC73	Frozen spiced beef burger	-	+

Organism	Culture collection no.[b]	Source[c]	PCR reactivity	
			795	*IAC*
E. coli O157:H7	GTA-EHEC75	Frozen spiced beef burger	-	+
E. coli O157:H7	GTA-EHEC81	Frozen spiced beef burger	-	+
E. coli O157:H7	GTA-EHEC178	Veal cheek meat	-	+
E. coli O157:H7	GTA-EHEC179	Beef burger	-	+
E. coli O26:H11	OLC-464	S. Read	-	+
E. coli O26:H11	OLC-465	S. Read	-	+
E. coli O26:H11	OLC-466	S. Read	-	+
E. coli O26:H11	OLC-725	M. Gilmour	-	+
E. coli O26:H11	OLC-731	M. Gilmour	-	+
E. coli O45:H2	OLC-716	M. Gilmour	-	+
E. coli O103:H2	OLC-679	M. Gilmour	-	+
E. coli O103:H25	OLC-680	M. Gilmour	-	+
E. coli O103:H2	OLC-711	M. Gilmour	-	+
E. coli O103:H2	OLC-712	M. Gilmour	-	+
E. coli O103:H25	OLC-727	M. Gilmour	-	+
E. coli O103:H11	OLC-728	M. Gilmour	-	+
E. coli O111:H11	OLC-455	S. Read	-	+
E. coli O111:H11	OLC-457	S. Read	-	+
E. coli O111:H11	OLC-629	R. Johnson	-	+
E. coli O111:NM	OLC-714	M. Gilmour	-	+
E. coli O111:NM	OLC-715	M. Gilmour	-	+
E. coli O121:NM	OLC-677	M. Gilmour	-	+
E. coli O121:H19	OLC-678	M. Gilmour	-	+
E. coli O121:H19	OLC-710	M. Gilmour	-	+
E. coli O121:H19	OLC-789	OLC	-	+
E. coli O121:NM	OLC-791	OLC	-	+
E. coli O121:NM	OLC-792	OLC	-	+
E. coli O121:H7	OLC-982	H. Tabor	-	+

Organism	Culture collection no.[b]	Source[c]	PCR reactivity	
			795	*IAC*
E. coli O145:NM	OLC-675	M. Gilmour	-	+
E. coli O145:NM	OLC-676	M. Gilmour	-	+
E. coli O145	OLC-983	H. Tabor	-	+
E. coli O5:NM	OLC-467	S. Read	-	+
E. coli O5:NM	OLC-468	S. Read	-	+
E. coli O119:H25	OLC-667	OLC	-	+
E. coli O76:H19	OLC-669	OLC	-	+
E. coli O115:H18	OLC-708	M. Gilmour	-	+
E. coli O5:NM	OLC-709	M. Gilmour	-	+
E. coli O55:H7	OLC-717	M. Gilmour	-	+
E. coli O91:H21	OLC-720	M. Gilmour	-	+
E. coli O117:H7	OLC-723	M. Gilmour	-	+
E. coli O113:H21	OLC-726	M. Gilmour	-	+
E. coli O6:H34	OLC-729	M. Gilmour	-	+
E. coli O146:H21	OLC-730	M. Gilmour	-	+
E. coli O177:NM	OLC-732	M. Gilmour	-	+
E. coli O85:H1	OLC-733	M. Gilmour	-	+
E. coli O48:H21	OLC-994	F. Scheutz	-	+
E. coli O174:H21	OLC-995	F. Scheutz	-	+
E. coli O118:H12	OLC-997	F. Scheutz	-	+
E. coli O73:H18	OLC-998	F. Scheutz	-	+
E. coli O2:H25	OLC-999	F. Scheutz	-	+
E. coli O8:K85ab:Hrough	OLC-1000	F. Scheutz	-	+
E. coli O128ac:H2	OLC-1001	F. Scheutz	-	+
E. coli O174:H8	OLC-1002	F. Scheutz	-	+
E. coli O139:K12:H1	OLC-1003	F. Scheutz	-	+
Shigella sonnei	OLC-24	ATCC 29930	-	+
Klebsiella pneumonia	OLC-27	ATCC 13883	-	+
Proteus vulgaris	OLC-28	ATCC 13315	-	+

Organism	Culture collection no.[b]	Source[c]	PCR reactivity	
			795	*IAC*
Serratia marcescens	OLC-30	ATCC 13880	-	+
Proteus aeruginosa	OLC-32	ATCC 10145	-	+
Citrobacter freundii	OLC-36	ATCC 8090	-	+
Salmonella enterica ser. Enteritidis	OLC-47	ATCC 13076	-	+
S. enterica ser. Typhimurium	OLC-59	ATCC 14028	-	+
Bacillus cereus	OLC-146	ATCC 14579	-	+
Enterobacter faecalis	OLC-147	ATCC 19433	-	+
B. subtilis	OLC-161	ATCC 6051	-	+
Staphylococcus epidermidis	OLC-244	ATCC 12228	-	+

[a]Bacterial lysates were subjected to the 795 PCR procedure as described in Methods, and PCR products were analyzed using the Flashgel electrophoresis system. PCR results for the 325-bp and IAC amplicons, respectively, are indicated as + (band present) or – (band absent).

[b]Culture collection reference number. OLC, Ottawa Laboratory-Carling culture collection; GTA, Greater Toronto Area Laboratory culture collection. OLC-795 and the parent strain from which it was derived are indicated in bold.

[c]Strains were obtained from various culture collections as follows: M. Gilmour and H. Tabor, National Microbiology Laboratory, Public Health Agency of Canada; R. Johnson and S. Read, Laboratory for Foodborne Zoonoses, Public Health Agency of Canada; G. Bezanson, Agriculture and Agri-Food Canada; St-Hyacinthe Laboratory (SHY), Canadian Food Inspection Agency; Burnaby Laboratory (BUR), Canadian Food Inspection Agency; F. Scheutz, Unit of Foodborne Infections, Statens Serum Institut.

Table 3.2: Oligonucleotide primer sequences for use in the 795 PCR procedure [a]

Primer	Sequence (5'→3')
795-F	TCA AAG CCA CTC TGT AGG GAA
795-R	AGC TGC AGA TTA CGA TCC CC
IAC-1	CAT AAT ATC ACT CGC GTC CGT TGA AGC TTA
IAC-2	GAC GAA ATC GTA AGC TTC AA

[a]Oligonucleotide primer sequences for use in the 795 PCR procedure

Chapter 4

Genomic tools for Customized Recovery and Detection of Foodborne Shiga-Toxigenic *Escherichia coli*

Preface

Copyright

Contributions of Collaborators

The following chapter is written as a manuscript submitted by Michael Knowles, Sara Stinson, Dominic Lambert, Catherine Carrillo, Adam Koziol, Martine Gauthier, and Burton W. Blais titled "Genomic tools for Customized Recovery and Detection of Foodborne Shiga-Toxigenic *Escherichia coli*" in the *Journal of Food Protection* with the consent to republish this work granted by the International Association for Food Protection (IAFP). I, Michael Knowles, assembled all experimental materials and performed all of the experiments and developed algorithms in the following manuscript with the exception of the following:

- Collection of *Escherchia coli* and *Salmonella enterica* isolates (B. W. Blais, and D.

Lambert)

- Sequencing of bacteria (C. Carrillo, and P. Manniger)
- SNP anaylsis in Fig. 4.1 (A. Koziol, and C. Carrillo)
- PCR analysis of *E. coli* O111 strains detailed in Fig. 4.1 (M. Gauthier)
- Laboratory evaluations detailed in the text, Tables 4.3 and 4.4 (S. Stinson)

Coauthors Sara Stinson, Dominic Lambert, Catherine Carrillo, Adam Koziol, Martine Gauthier, Burton Blais and I, Michael Knowles, wrote the original draft of this paper and all revised versions.

4.1 Abstract

Genomic antimicrobial resistance (AMR) prediction tools may be very useful in supporting foodborne illness outbreak investigations through their application in the analysis of bacterial genomes from causative strains. The AMR marker profile from an outbreak strain of interest could be compared with related bacteria to determine the presence of unique resistance markers enabling customization of selective enrichment media facilitating their recovery from samples during food safety investigations. Different possibilities for such analyses include in-house pipelines employing bioinformatics alignment algorithms along with comprehensive AMR gene databases such as CARD (e.g., Antimicrobial Resistance Marker Identifier [ARMI]), or publicly available tools which use more limited acquired AMR gene databases (e.g., ResFinder). Combined with a previously reported pipeline (SigSeekr) designed to identify specific DNA sequences associated with a particular strain, or closely related sub-types, for their rapid identification by polymerase chain reaction (PCR), it should be possible to deploy custom recovery and identification tools for the efficient detection of Shiga-toxigenic *Escherichia coli* (STEC) outbreak strains within the timeframe of an active investigation. Using a laboratory STEC strain as a model, trimethoprim resistance identified by both ARMI and ResFinder was used as a basis for its selective recovery against a background of commensal *E. coli* bacteria in ground beef samples. Enrichment in modified trypticase soy broth containing trimethoprim greatly enhanced the recovery of low numbers of model strain cells inoculated in ground beef samples, as verified by the enumeration of colonies on plating media using a strain-specific PCR method to determine the recovery efficiency for the target strain. We discuss the relative merits of different AMR marker prediction tools for this purpose, and

describe how such tools can be utilized to good effect in a typical outbreak investigation scenario.

4.2 Introduction

Shiga-toxigenic *Escherichia coli* (STEC) are an important cause of food- and water-borne gastroenteritis, both sporadically and in outbreaks, potentially producing serious illness in infected persons (Tarr et al. 2005). The investigation of foodborne illness outbreaks involves the coordinated actions of many regulatory authorities, such as public health laboratories which identify the causative organism in clinical specimens (a key step in the identification of an outbreak event), risk assessment and food inspection professionals who must all work together toward the timely identification of contamination sources in order to minimize public exposure (e.g., by removing contaminated food products from the marketplace). The food testing laboratory has an important role in supporting food safety investigations through the identification of implicated production lots to determine the source and scope of a contamination incident. For this purpose, it is essential that food testing protocols provide the most informative test results regarding the salient characteristics of food bacterial isolates linking them to the case in hand within the shortest timeframe possible.

Despite recent efforts of food safety regulatory agencies to implement test methods targeting specific sub-groups of priority STEC (Blais, Gauthier, et al. 2012; Huszczynski et al. 2013; USDA 2011) it has been difficult to precisely define all of the members from this broad family that are of public health concern, and there have been outbreaks of foodborne disease where the causative strain had unanticipated characteristics (e.g., the 2011 German outbreak in which the etiologic agent belonged to serogroup O104, not typically included among the "big seven" priority serogroups, and lacked the definitive virulence marker *eae*) (Rasko et al. 2011). Problems associated with the detection of such a adaptable family of emerging pathogens are many-fold. The recovery of target pathogens during food safety investigations is often hampered by the overwhelming presence of competing endogenous microbiota in food commodities such as ground beef and sprouts. Most strategies for the recovery and detection of traditional pathogens in foods have exploited well-characterized properties such as resistance to selective agents (e.g., antibiotics or toxic compounds) which can be incorporated in enrichment media. However, the STEC are a heterogeneous family with considerable variability in resistance to antibiotics and other selective agents (Gill et al. 2012), making it difficult to develop

universal selective enrichment conditions enabling consistent recovery of all or even a predefined sub-set of types. Furthermore, the rapid identification of food samples containing an STEC strain implicated in a public health event can be hampered by false positive results due to the relatively high prevalence of non-pathogenic STEC in certain types of commodities such as beef (Bosilevac and Koohmaraie 2011).

Leading-edge genomics approaches provide new possibilities for comprehensive analyses of foodborne bacteria fostering the development of highly targeted tools enabling the efficient detection of outbreak strains. Next-generation sequencing technologies can now render a bacterial genome much faster (possibly within a single working day) and at a significantly lower cost (about one hundred dollars) than previously possible, making it feasible to sequence foodborne isolates within the time frame of a food safety event (Lambert et al. 2015). Current outbreak investigation routines typically involve whole genome sequencing (WGS) of clinical isolates by public health laboratories, making it possible for partner agencies (e.g., food inspection laboratories) to rapidly acquire WGS data which may offer important clues about the particular characteristics of the causative organism to aid in its detection in food samples. For example, it may be possible to determine a priori the presence of antibiotic resistance markers and strain-specific DNA sequences enabling the application of customized selective recovery and identification methods (e.g., polymerase chain reaction (PCR)) enhancing their detection against a background of non-target bacteria.

A number of tools are currently available to predict antimicrobial resistance (AMR) from bacterial WGS data [e.g., ResFinder (Kleinheinz et al. 2014; Zankari et al. 2012), SEAR (Rowe et al. 2015), Resistance Gene Identifier (McArthur et al. 2013), and an in-house tool developed by the present investigators named Antimicrobial Resistance Marker Identifier (ARMI)]. These AMR marker prediction tools (APTs) rely on curated international AMR gene databases such as Comprehensive Antibiotic Resistance Database (CARD) (McArthur et al. 2013), Antibiotic Resistance Genes Database (ARDB) (B. Liu and Pop 2009), and Antibiotic Resistance Gene-ANNOTation (ARG-ANNOT) (Gupta et al. 2014). WGS-based methods for prediction of AMR phenotype have been shown to be highly accurate (Tyson et al. 2015; Zankari et al. 2012). Here we propose the systematic application of such tools (e.g., the ARMI tool developed in our laboratory, and ResFinder) for the identification of antibiotic resistance in a locally sequenced bacterial strain, such as a clinical VTEC isolate, to enable the design of selective media for use in its specific recovery in food samples during food safety investigations. Combining APTs with a previously described a bioinformatics pipeline (SigSeekr) for the WGS-based de-

sign of specific PCR primers (Knowles et al. 2015) targeting a strain of interest, it should be possible to deploy customized detection methods to rapidly identify foods implicated in a food safety incident. The feasibility of such an approach was investigated with a model laboratory STEC strain subjected to analysis using APTs and SigSeekr tools developed in our laboratory, and the detection of this strain in ground beef by customized selective enrichment and PCR techniques predicated on this bioinformatics platform.

4.3 Materials and Methods

4.3.1 Chemicals and Reagents

Antimicrobial susceptibility disks were obtained from Oxoid, Thermo Fisher Scientific Inc. (Ottawa, Ontario), and included apramycin at 15µg (CT0545B); chloramphenicol at 10µg, 30µg and 50µg (CT0012B, CT0013B, CT0014B); erythromycin at 5µg, 10µg, 15µg and 30µg (CT0066B, CT0019B, CT0020, CT0021B); gentamicin at 10µg, 30µg, 120µg, and 200µg (CT0024B, CT0072B, CT0794B, CT0695B); kanamycin at 5µg and 30µg (CT0025B, CT0026B); neomycin at 10µg and 30µg (CT0032B, CT0033B); penicillin G at 1 unit, 1.5 units, 2 units, 5 units, and 10 units (CT0152B, CT0042B, CT0088B, CT0124B, CT0043B); spectinomycin at 10µg, 25µg, and 100µg (CT0046B, CT0411B, CT0823B); (10) streptomycin at 10µg, 25µg, and 100µg (CT0047B, CT0048B, CT1897B); sulphonamide at 300µg (CT0059B); and trimethoprim at 1.25µg, 2.5µg, and 5µg (CT0057B, CT0070B, and CT0076B).

Antibiotics were obtained from Sigma-Aldrich (St. Louis, Mo), and included apramycin (A2024), chloramphenicol (C1919), erythromycin (E1300000), florfenicol (F1427), gentamicin (G1264), hygromycin B (H3274), kanamycin A (K1377), neomycin (N0401000), penicillin G (P3032), spectinomycin (no. PHR1441), streptomycin (S6501), sulfamethoxazole (S7507), and trimethoprim (T7883).

4.3.2 Bacterial strains

A variety of different STEC strains from the Ottawa Laboratory Carling (OLC) (Canadian Food Inspection Agency) culture collection (Table 4.1) were used for antimicrobial resistance (AMR) marker prediction analysis. An additional 27 strains of *E. coli* serogroup O111 (Fig. 4.1) from the OLC culture collection were used to validate the 714 polymerase chain reaction (PCR) (see below). Bacteria were routinely grown on nutrient

agar (NA) (Oxoid Ltd., Basingstoke, Hampshire, England) for 16-20 h at 37°C.

4.3.3 Broth method for determination of AMR phenotypes

Antibiotic stock solutions were prepared at a concentration of 10 mg/mL in deionized distilled water or dimethyl sulfoxide, depending on solubility, and then filter sterilized by passage through 0.22µm filters. Antibiotics were diluted in modified tryptone soya broth (mTSB) (Oxoid Ltd., Basingstoke, Hampshire, England) to final concentrations of 10 μg/mL and 100 μg/mL. Bacteria grown in mTSB were inoculated (ca. 10^4 CFU) into 3 mL of antibiotic-broth solution and incubated at 37°C for 16-20h. Growth was assessed by measuring turbidity (A_{580}) using a DensiCHECK™ Plus instrument (BioMérieux Inc., Montreal, Quebec) and recorded in McFarland turbidity units. Typically, growth manifested as turbidity producing McFarland values above 4.0, whereas the absence of visible growth (no turdbidity) corresponded to McFarland values below 0.10. A strain was considered resistant to an antibiotic if the colonies produced significant growth (McFarland value > 1.0) in its presence. Each determination was carried out in duplicate.

4.3.4 Disc diffusion method for determination of AMR phenotypes

Bacteria were suspended in peptone to a density of 0.5 McFarland turbidity units (approximately 1-2$\times 10^8$ CFU/mL). Suspensions were spread on Mueller-Hinton Agar (MHA) (Oxoid, Thermo Fisher Scientific Inc., Ottawa, Ontario, CM0337B). Six antibiotic discs were applied to the MHA plates in duplicate using a disk dispenser (Oxoid, Thermo Fisher Scientific Inc., Ottawa, Ontario). MHA plates were then inverted and incubated at 37°C for 16-20 h. The zone edge was determined to be at the point of complete inhibition and the diameters of the zones of inhibition were measured to the nearest millimeter. STEC strains were determined to be phenotypically susceptible or resistant to a given antibiotic based on minimum inhibitory concentration (MIC) and breakpoints for zones of inhibition. Zones of inhibition were compared to published standards for *Enterobacteriaceae* from the European Committee on Antimicrobial Susceptibility Testing (EUCAST) and the Clinical and Laboratory Standards Institute (EUCAST 2016; CLSI 2014).

4.3.5 Whole–genome sequencing and assembly

Bacterial strains were cultured on nutrient agar (Difco, Becton, Dickinson & Co) overnight (14-16 hrs) at 37°C or in BHI broth for 4 to 7 h at 37°C, and genomic DNA was extracted using the Maxwell® 16 Cell LEV DNA Purification kit (Promega, Madison, WI). DNA was quantified using the Quant-iT™ High-Sensitivity DNA Assay Kit (Life Technologies Inc., Burlington, ON). Sequencing libraries were constructed from 1 ng of genomic DNA using the Nextera XT DNA Sample Preparation Kit (Illumina, Inc., San Diego, CA) and the Nextera XT Index Kit (Illumina, Inc.). Paired-end sequencing was performed on the Illumina MiSeq Platform (Illumina, Inc.) using 600 cycle MiSeq Reagent Kit v3. Sequencing errors in reads were corrected using Quake (version 0.3) with a k-mer size of 15 (Kelley et al. 2010). Sequencing reads were assembled using SPAdes 3.7.1 (Bankevich et al. 2012) with a k-mer range (21, 33, 55, 77, 99, 127) and error correction parameters. Contigs shorter than 1000 bp were removed from the assemblies. Ribosomal multilocus sequence typing (rMLST) was conducted on the assembled genomes using a BLAST-based custom python script (https://github.com/adamkoziol/MLST) and databases downloaded from http://pubmlst.org/ (Jolley et al. 2012; Wirth et al. 2006)

4.3.6 Evolutionary analyses

Phylogenetic relationships among the E. coli isolates was inferred by analysis of high quality SNPs in genome assemblies using kSNP version 3.0 (Gardner et al. 2015), using a k-mer size of 51. Maximum parsimony trees were constructed in MEGA7 using core SNPs (Kumar et al. 2016)

4.3.7 ARMI predictions

To assess the relative abundance of AMR genes in *E. coli* species, all publicly available *Escherichia coli* (1600) GenBank sequences (i.e., unannotated genomic contigs, closed assembled pseudomolecules (chromosome-like DNA fragment), and plasmids were retrieved from NCBI GenBank using the Biopython application programming interface v1.64 (Cock *et al.*, 2009) Entrez module (NCBI, June 2015) and were combined with the locally generated sequences to form the database used by ARMI. The Comprehensive Antibiotic Resistance Database (CARD), Antibiotic Resistance Genes Database (ARDB) (B. Liu and Pop 2009), and Antibiotic Resistance Gene-ANNOTation (ARG-ANNOT) (Gupta et al.

2014) databases were retrieved from their respective source to build BLAST and Bowtie2 databases for ARMI (CARD databases and ARDB are retrieved automatically as part of the installation process located at github.com/OLC-Bioinformatics/GeneSeekr). The ARMI analyses of the acquired genomes were performed against the AMR databases. A cut-off E-value of 10^{-7} and sequence identity of 85% were used in accordance with recent findings by Yang *et al.* (Yang et al. 2016).

4.3.8 ResFinder Predictions

Genome assemblies were uploaded to the ResFinder web server, version 2.1 (Kleinheinz et al. 2014). A sequence identity threshold of $\geq 98.00\%$ and a match length threshold of $\geq 60.00\%$ was selected for a search using genes in the ResFinder database against input genome contiguous sequences. The default format of the input genome was selected as assembled genome/contigs (Kleinheinz et al. 2014). The ResFinder output consisted of a list of resistance genes/accession numbers, alignments of the target sequences to these genes, organized under subheadings indicating resistance to antibiotic classes. Further elucidation of the antimicrobial compounds to which each gene conferred resistance required both scanning the journal article linked to the NCBI gene accession numbers, searching for the associated gene accession number and gene name in ARDB (Table 4.2).

4.3.9 Unique sequence identification and primer design

The SigSeekr pipeline uses BLASTn 2.2.29+ (Altschul et al. 1990) to eliminate target strain (query) sequences from a pool of potential signature sequences as follows: (1) query sequences with matches to the pan-genomic sequence database reporting initial E-values ≤ 1.0 and $\geq 90\%$ identity are 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) is generated; (2) repeated unique sequences are eliminated from the string of query sequences using a fuzzy matching algorithm to prevent a duplication of signature sequences in the output and sequences below 200 base pairs are eliminated from the string to ensure suitability for PCR amplification; (3) when no sequences are returned in the string, the process is recursively iterated using lower E-values until at least one is found; and (4) the resulting string, containing the least common sequence(s), is then used to query the entire pan-genomic database (including target strain) as a quality control.

The software was implemented in Python v2.7.3 with the use of Biopython application programming interface v1.64 (Cock et al. 2009) BlastCommandLine, SearchIO and SeqIO and optimized to prevent excessive I/O to the hard disk and keep the majority of information in memory. The memory usage experienced a maximum of 2GB with the *E. coli* database. An average run time of 15 minutes was found using a Xeon E3-1650v2 (6 cores at 3.5 GHz) workstation and 64GB of RAM.

When used to determine a sequence unique to strain OLC-714, the pipeline identified a 183-nucleotide fragment (http://github.com/OLC-LOC-Bioinformatics/SigSeekr). Primers specific to the OLC-714 strain unique sequence (Table 2.1) were designed using Primer-BLAST (Ye et al. 2012), and then obtained from a commercial supplier (Coralville, IA, USA).

4.3.10 Polymerase chain reaction procedure (714 PCR)

Bacterial colony lysates were prepared by suspension of a bacterial colony in 200 µl of 1% (v/v) Triton X-100 (Sigma-Aldrich Co., St. Louis, MO, USA), followed by heating at 100°C for 10 min. Enrichment broth culture lysates (see below) were prepared by combining 100 µl of culture with 100 µl of 2% (v/v) Triton X-100, followed by heating at 100°C for 10 min. For the PCR, 5 µl of bacterial lysate was added to 45 µl of multiplex PCR mixture containing 2.5 units HotStar *Taq* and 1 x HotStar PCR buffer (Qiagen Inc., Mississauga, ON, Canada), 2.5 mM $MgCl_2$, 400 µM of each dNTP, 0.4 µM each of the 714-F and 714-R primers, and 0.1 µM each of the IAC-1 and IAC-2 primers (Leggate and Blais 2006). The IAC primers served as an internal amplification control (Leggate and Blais 2006). The PCR was carried out in a Mastercycler gradient thermal cycler (Eppendorf, Hamburg, Germany) 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) systems (following the manufacturers' instructions).

4.3.11 Ground beef recovery studies.

Fresh ground beef was obtained from a local retailer. Total aerobic colony counts (ACC) were performed by mechanically digesting 25 g of ground beef in 225 ml of Tryptone Soya Broth (TSB) (Oxoid Ltd.), and plating serial dilutions on NA plates which were incubated for 16-20 h at 37°C. The ACC value for the ground beef was 2.3×10^4 cfu/g.

For enrichment studies, 25 g of ground beef were stomached in 225 ml of mTSB, and 10 ml portions with and without 100 µg/ml of trimethoprim (Sigma-Aldrich Co., St. Louis, MO) were inoculated with 0 or 7 cfu of *E. coli* O111:NM (OLC-714), and then incubated for 16-20 h at 37°C. Enrichment cultures were tested by the 714 PCR (above), and portions were plated on Rainbow Agar® (Biolog, Hayward, CA, USA) to obtain isolated colonies (30-150 colonies per plate). All isolated colonies observed on each plate were subjected to the 714-PCR, and the recovery efficiency was calculated by determining the proportion of total colonies assayed that were positive.

4.4 Results and Discussion

The ability to identify antimicrobial resistance (AMR) markers in a bacterial genome is largely predicated on the nature of the reference gene database which the prediction algorithm is directed to search. The ARMI analysis relies on the Comprehensive Antibiotic Resistance Database (CARD) (McArthur et al. 2013) which is an actively maintained, comprehensive resource incorporating acquired resistance genes. In addition, this database contains resistance genes that have not yet been associated with pathogenic bacteria. The comprehensive nature of CARD offers annotation of genes which may confer resistance to antimicrobial agents that are not commonly used in clinical settings, therefore the selective assay designed using CARD may have both greater specificity with target bacteria and the potential to significantly reduce time for recovery. In contrast, ResFinder has been developed and evaluated using data from organisms of clinical importance, and focusses on classes of antimicrobial agents that are of clinical importance (Zankari et al. 2012). For the purpose of the development of customized enrichment broths, the increased scope of the CARD database was advantageous. The extensive curation of this database was also extremely useful, as ARMI was able to parse data to rapidly identify target antibiotics (Table 4.1). The ARMI tool designed in our laboratory matches bacterial strain genomic data with one of the most comprehensive AMR gene databases currently available, CARD, and its output is a listing of specific antibiotics to which the organism is putatively resistant. The primary output of ResFinder is in the form of antibiotic classes, though specific genes are identified and their function (i.e., target antibiotic) can be determined by cross-referencing the corresponding entries in the National Center for Biotechnology Information (NCBI) repository and inferring resistance.

4.4.1 Comparative evaluation of AMR marker prediction tools (APTs)

The AMR marker profiles generated by these two APTs were compared for a panel of 18 different *E. coli* (mostly STEC) strains. A total of 306 and 96 AMR predictions were made by ARMI and ResFinder, respectively (Table 4.1). In most cases ARMI determined all of the antibiotic resistance traits obtained with ResFinder. ARMI also detected additional resistance genes not covered by ResFinder (which is focused on "acquired" AMR determinants). For some antibiotic classes, ARMI gave a higher number of determinations because of its ability to delineate specific AMR markers (e.g., SFA, SDZ, SDD, SDX, SMZ SMX, SSZ and SIX) within a broader class (SUL) as determined by ResFinder. The primary output of ResFinder is in the form of gene designations indicating broad antibiotic classes (e.g., aminoglycosides).

One of the main goals of these studies was to develop a practical, standardized approach to exploit the predicted AMR traits of bacteria (e.g., an outbreak strain) for customization of selective enrichment media. Therefore, we studied the predictive value of the two APTs under typical broth enrichment conditions in which the antibiotics could be used to recover a target strain from a food sample. For this purpose, we selected a sub-set of 10 different antibiotics (Table 4.3) to determine the correlation between AMR marker prediction and the resistance phenotype. Since AMR phenotypes are defined on the basis of minimal inhibitory concentrations (MICs), which vary for different antibiotics and classes of bacteria, we sought to standardize the conditions for their use by focusing on antibiotic concentrations within range of their respective resistance breakpoints (CLSI 2014; EUCAST 2016). The 18 STEC strains (Table 4.1) were inoculated in broth containing 10 and 100 µg/ml of each antibiotic for which resistance was predicted using the two APTs. For the panel of antibiotics examined, ARMI gave an insignificantly higher number of total predictions with a positive predictive value (positive predictive value (PPV) of 92%, compared to a PPV of 90% obtained with ResFinder in Table 4.3. In order to achieve this level of performance with ARMI it was necessary to modify the search database to eliminate problematic AMR markers. For example, an earlier version of ARMI in which markers for erythromycin were included resulted in a significantly diminished PPV by the broth method because 16 predictions were made for this antibiotic, but none were substantiated by evidence of growth with either level of this antibiotic (not shown). One possibility for this high rate of "false predictions" could have been due to the presence in CARD of *mdtEF-TolC*, a common component of efflux pumps associated

with a variety export functions not necessarily related to antibiotic resistance (Deininger et al. 2011). Elimination of this marker from the search field significantly improved the PPV (Table 4.3) primarily because no erythromycin predictions were made. Likewise, equivocal results can sometimes occur with ResFinder due to ambiguities originating with the source of the predicted AMR. For example, while kanamycin resistance may be indicated by the occurrence of *aph*(3')-IIa in OLC 1334 (Tables 2.1 and 3.2), Najibi et al. (2012) found that either kanamycin or neomycin resistance may be associated with *E. coli* strains bearing this marker. This finding underscores the importance of carefully searching the database to ensure a high PPV for the APT, which is necessary for the present purpose of accurately determining an antibiotic which may be used for selective recovery of a strain of interest. While this may limit the number of antibiotics which can be considered for present purposes, it is more important to ensure the reliability of the APT in identifying an AMR trait. To ensure that the observed predictive values were not skewed by the choice of broth enrichment method to determine resistance phenotype, the same panel of strains was also tested for resistance to a sub-set of antibiotics using a standard disc diffusion approach (CLSI 2014; EUCAST 2016). The number of antibiotics tested (total of 7) was limited in this case by the commercial availability of discs. The disc method elicited PPVs of 92% for both APTs (Table 4.4), mirroring the performance observed with the broth method (Table 4.3), and providing corroborating evidence underscoring the validity of the latter approach for determining resistance characteristics.

4.4.2 Customized selective enrichment and detection of a model STEC strain in ground beef

The applicability of APTs in informing the customization of target strain-specific selective enrichment media was examined by studying the recovery of a model target strain from a food sample containing a high level of background bacteria (e.g., ground beef). For this purpose, an *E. coli* O111:NM strain (OLC 714) was selected as a model, along with the use of trimethoprim for its selective recovery from inoculated ground beef, since resistance to this antibiotic predicted for this strain (Table 4.1) occurred with a moderate degree of rarity (4 occurrences as determined by ResFinder and ARMI) among the panel of 18 *E. coli* strains examined. Recovery of the model strain was assessed by assaying enrichment cultures using a PCR method (714-polymerase chain reaction (PCR)) featuring primers (Table 2.1) targeting DNA sequences associated with the strain of interest identified by

SigSeekr. The narrow reactivity of this PCR method for OLC-714 (and closely related bacteria [i.e., some other *E. coli* O111 strains]) was confirmed through the observation that these primers only produced the expected amplicon with the target strain and several very closely related *E. coli* O111 strains of the same sequence type (ST-16), but not with other *E. coli* O111 strains (including some ST-16 strains, several ST-21, and one each of ST-40 and ST-149) (Fig. 4.1), nor with a broader panel of *E. coli* strains (Table 4.1).

The efficacy of trimethoprim as a strain-specific selective agent aiding in the recovery of *E. coli* O111:NM (OLC-714) from ground beef samples containing a high level of background bacteria was examined using the 714 PCR for detection of the target organism in broth enrichment cultures as well as to identify colonies subsequently recovered on plating media. The use of trimethoprim was necessary to enable detection of strain OLC-714 directly in the enrichment broth culture of an inoculated sample (Fig. 4.2 inset, lane 2), presumably because in the absence of trimethoprim the background microbiota inhibited the growth of the target strain (Fig. 4.2, inset, lane 3). No PCR product was produced with the uninoculated samples regardless of whether or not trimethoprim was present (Fig. 4.2 inset, lanes 4 and 5), nor was there any evidence of interference with PCR performance from any of the enrichment or sample preparation components (Fig. 4.2 inset, lanes 6-11), supporting the conclusion that trimethoprim was solely responsible for enabling successful detection of the target organism in the enrichment sample.

The impact of trimethoprim on recovery efficiency for the target strain was studied by plating portions of duplicate ground beef enrichment samples on Rainbow Agar® and determining the proportion of recovered colonies which bore identity to the subject strain using the 714-PCR. Large numbers of colonies were recovered from samples devoid of the target strain enriched in the absence of trimethoprim, but none of these colonies tested positive with the 714-PCR (Fig. 4.2, panel A), suggesting that they were the product of enrichment of the ground beef background microbiota. Large numbers of colonies were also recovered in the absence of trimethoprim from samples inoculated with the target strain, but none of these were positive in the 714-PCR (Fig. 4.2, panel B), indicating that the recovery of OLC 714 was inhibited by an overwhelming level of background microbiota (which were present in greater than 3000-fold excess relative to the inoculated strain). Very few colonies were recovered from uninoculated samples enriched in the presence of trimethoprim (Fig. 4.2, panel C), none of which were reactive in the 714-PCR, demonstrating the effectiveness of trimethoprim in suppressing the growth of the background bacteria. Finally, when inoculated ground beef samples were enriched in the presence of trimethoprim almost all of the colonies recovered on plat-

ing media produced positive results with the 714-PCR (Fig. 4.2, panel D), suggesting that the background microbiota were effectively suppressed, allowing the unhampered growth of the target strain. These results clearly demonstrate the beneficial impact of incorporating trimethoprim in the enrichment broth.

While the present experiments are limited in that only one antibiotic resistance scenario (trimethoprim) was studied in detail, it may be reasonably inferred from the results presented in Table 4.3, which demonstrate successful cultivation of various STEC strains in enrichment broth incorporating different antibiotics, that other antibiotic resistance traits should be amenable to the concept of customization of selective enrichment tailored to a strain of interest. More work will be needed to validate APTs and better define which AMR predictions and antibiotic growth conditions will work reliably in a food enrichment scenario such as the present example. Work is underway in our laboratory to improve the APTs and develop a catalogue of antibiotics (and the conditions for their use) from which a selection could be made on an *ad hoc* basis to support an active food safety investigation.

A model for the general deployment of these *ad hoc* strain-specific tools in support of outbreak investigations is illustrated in Fig. 4.3. Outbreak events associated with consumption of contaminated food products are initially recognized when public health laboratories recover matching clinical isolates from several affected individuals. The genomes of clinical isolates are rapidly sequenced for high resolution typing analyses to establish phylogenetic relationships among different isolates. Once an outbreak is recognized the sequence is rapidly disseminated to partner investigative agencies such as the food inspection laboratory. This can be achieved through participation in global data basing initiatives such as Genome Trakr (Genome Trakr Network 2015), or if the information is of a confidential nature, through a secure network link. Patient-specific metadata associated with the genome sequence are unnecessary and may be omitted to alleviate privacy concerns. The food testing laboratory applies APT and SigSeekr analyses to the clinical WGS data and determines whether a validated AMR marker can be identified to aid in the selective recovery of the outbreak strain, along with specific PCR primers enabling its detection (e.g., in enrichment broth cultures or colony isolates). Once the strain has been identified in food products targeted regulatory action can be taken to minimize public exposure to the specified hazard, such as recalling specific production lots from the marketplace. This approach is the very epitome of the risk-based food safety paradigm.

An essential requirement is that the AMR prediction pipeline be validated (i.e., geno-

type accurately predicts phenotype), and that the conditions (i.e., concentrations) for use of different antibiotics as selective enrichment agents can be standardized. We are currently developing a catalogue of AMR markers predicted by the ARMI tool for which we can demonstrate the corresponding resistance phenotypes among an extended panel of STEC strains. The effective concentrations for each validated antibiotic trait in selective enrichment are also being characterized. The process of verifying the exclusivity characteristics of *ad hoc* PCR primers can be expedited by using 96-well microtiter plates pre-formulated with genomic DNA preparations from different bacterial strains. We have found that wells with dried genomic DNA extracts can be stored in a vacuum-sealed package at 4°C for up to one year, and therefore, it is possible to prepare ready-for-use plates ahead of time for rapid *ad hoc* primer verification prior to their deployment. Inclusivity verification would of course not be possible unless the laboratory had the actual outbreak strain in hand, which may not be feasible during an active outbreak investigation. However, it should be possible to synthesize an oligonucleotide corresponding to the entire amplicon sequence encompassed by the primers for use as a positive control which can be incurred in a portion of enrichment broth culture or used on a standalone basis with each sample set.

The similar performance characteristics obtained with the two APTs was achieved through careful curation of the AMR gene databases to avoid ambiguous search results (i.e., poorly defined AMR genes or broad class determinations). This is an important consideration in achieving accurate results for the present purpose of informing the composition of strain-specific selective enrichment media. Based on our results, either ARMI or ResFinder can be used, though the latter requires considerable effort to precisely identify candidate antibiotics which can be incorporated as selective agents in enrichment media.

To safeguard against the possibility of false negative results due to bioinformatics errors, portions of test samples can also be processed in parallel using standard methodology (Huszczynski et al. 2013). In the current approach the process of identifying a resistance trait that is more or less unique to the target strain is carried out manually (i.e., comparing the AMR profile of a subject strain to those of a set of closely related bacteria). We are currently working on the further refinement of the AMR prediction tool providing for the identification of rare traits through an automated process for comparison of the subject AMR profile with large databases of closely related bacteria, including typical commensals associated with typical food commodities.

4.5 Acknowledgements

The authors thank Paul Manninger and Mylène Deschênes for providing technical laboratory support. We also acknowledge the Government of Canada Genomic Research and Development Initiative for providing partial funding for this work, as well as the GRDI Interdepartmental Food and Water Safety project consortium (comprised of researchers from Agriculture and Agri-food Canada, the Canadian Food Inspection Agency, Environment Canada, Health Canada, the National Research Council of Canada, and the Public Health Agency of Canada) for useful discussions and contributions.

4.6 Supplemental Information

Table 4.1: AMR predictions for *E. coli* strains[a]

Organism	Culture collection no.	APT	Inferred antibiotic resistance[b]
E. coli O111:H11	OLC-455	ARMI	AMK, BAC, BUT, CHL, CST, GENB, ISP, KAN, LVDA, LVDB, NEO, PRM, PMB, RSM, SPT, STR
		ResFinder	BLA, CHL, GENB, KAN, LVD, NEO, PRM, RSM, SPT, STR, SUL, TET
E. coli O26:H11	OLC-637	ARMI	BAC, CST, MAF, PMB, STR, SFA, SDZ, SDD, SDX, SMZ, SMX, SSZ, SIX, SUL, TET
		ResFinder	STR, SUL, TET
E. coli O26:H11	OLC-639	ARMI	BAC, CST, MAF, PMB, STR, SFA, SDZ, SDD, SDX, SMZ, SMX, SSZ, SIX, SUL, TET
		ResFinder	STR, SUL, TET

Organism	Culture collection no.	APT	Inferred antibiotic resistance[b]
E. coli O111:NM	OLC-714	ARMI	AMK, AM, ABK, ASTM, BAC, BLA, BUT, CHL, CST, DAP, DBK, GENB, GENC, HYG, ISP, KAN, LVDA, LVDB, MAF, NEO, NET, PRM, PMB, RSM, SIS, SPT, STR, SFA, SDZ, SDD, SDX, SMZ, SMX, SSZ, SIX, SUL, TOB, TMP
		ResFinder	BLA, CHL, DBK, GENB, GENC, KAN, LVD, NEO, NET, PRM, RSM, SIS, SPT, STR, SUL, TET, TOB, TMP
E. coli O145:NM	OLC-719	ARMI	BAC, CST, MAF, PMB, STR, SFA, SDZ, SDD, SDX, SMZ, SMX, SSZ, SIX, SUL, TET
		ResFinder	STR, SUL, TET
E. coli O103:H2	OLC-969	ARMI	BAC, BLA, CST, DAP, MAF, PMB, STR, SFA, SDZ, SDD, SDX, SMZ, SMX, SSZ, SIX, SUL, TMP
		ResFinder	BLA, STR, SUL, TET, TMP
E. coli O111:H8	OLC-1064	ARMI	BAC, CST, PMB, STR
		ResFinder	BLA, STR
E. coli O157:H7	OLC-1068	ARMI	BAC, BLA, CST, DAP, MAF, PMB, STR, SFA, SDZ, SDD, SDX, SMZ, SMX, SSZ, SIX, SUL, TMP
		ResFinder	BLA, STR, SUL, TET, TMP.

Organism	Culture collection no.	APT	Inferred antibiotic resistance[b]
E. coli (not typed)	OLC-1105	ARMI	BAC, BLA, CST, DAP, MAF, PMB, STR, SFA, SDZ, SDD, SDX, SMZ, SMX, SSZ, SIX, SUL, TMP
		ResFinder	BLA, STR, SUL, TET, TMP
E. coli O157:H7	OLC-1112	ARMI	AMK, AM, ABK, ASTM, BAC, BUT, CST, DBK, GENB, GENC, HYG, ISP, KAN, LVDA, LVDB, MAF, NEO, NET, PRM, PMB, RSM, SIS, SPT, STR, SFA, SDZ, SDD, SDX, SMZ, SMX, SSZ, SIX, SUL, TOB
		ResFinder	GENB, GENC, KAN, SIS, SPT, STR, SUL, TET
E. coli O157:H7	OLC-1128	ARMI	BAC, CST, MAF, PMB, SPT, STR, SFA, SDZ, SDD, SDX, SMZ, SMX, SSZ, SIX, SUL
		ResFinder	SPT, STR, SUL, TET
E. coli O103:H2	OLC-1330	ARMI	BAC, BLA, CST, MAF, PMB, STR, SFA, SDZ, SDD, SDX, SMZ, SMX, SSZ, SIX, SUL
		ResFinder	BLA, STR, SUL
E. coli O2:H25	OLC-1334	ARMI	CST, MAF, PMB, SPT, STR, SFA, SDZ, SDD, SDX, SMZ, SMX, SSZ, SIX, SUL
		ResFinder	SPT, STR, SUL, TET
E. coli O2:H25	OLC-1337	ARMI	BAC, CHL, CST, FFC, MAF, PMB, STR, SFA, SDZ, SDD, SDX, SMZ, SMX, SSZ, SIX, SUL

Organism	Culture collection no.	APT	Inferred antibiotic resistance[b]
		ResFinder	BLA, CHL, FFC, STR, SUL, TET
E. coli O157:H7	OLC-1356	ARMI	BAC, CST, MAF, PMB, SPT, STR, SFA, SDZ, SDD, SDX, SMZ, SMX, SSZ, SIX, SUL
		ResFinder	SPT, STR, SUL, TET
E. coli O157:H7	OLC-1357	ARMI	BAC, CST, MAF, PMB, SPT, STR, SFA, SDZ, SDD, SDX, SMZ, SMX, SSZ, SIX, SUL
		ResFinder	SPT, STR, SUL, TET
E. coli O157:H7	OLC-1358	ARMI	BAC, CST, MAF, PMB, SPT, STR, SFA, SDZ, SDD, SDX, SMZ, SMX, SSZ, SIX, SUL
		ResFinder	SPT, STR, SUL, TET
E. coli (not typed)	OLC-1558	ARMI	BAC, CST, MAF, PMB, STR, SFA, SDZ, SDD, SDX, SMZ, SMX, SSZ, SIX, SUL, TET
		ResFinder	STR, SUL, TET

[a] A variety of *E. coli* strains from the culture collection of the Ottawa Laboratory Carling (Canadian Food Inspection Agency) were subjected to whole genome sequencing and antimicrobial resistance (AMR) marker analysis using the ARMI and ResFinder tools.

[b]Amikacin (AMK); apramycin (AM); arbekacin (ABK); astromicin (ASTM); bacitracin (BAC); beta-lactam (BLA); Butirosin (BUT); chloramphenicol (CHL); cloxacillin (CLX); colistin (CST); diaminopyrimidine (DAP); dibekacin (DBK); erythromycin (ERY); Florfenicol (FFC); Geneticin (G418); gentamicin B (GENB); gentamicin C (GENC); hygromycin B (HYG); isepamicin (ISP); kanamycin A (KAN); lividomycin (LVD); lividomycin A (LVDA); lividomycin B (LVDB); mafenide (MAF); nalidixic acid (NAL); neomycin (NEO); netilmicin (NET); norfloxacin (NOR); novobiocin (NVB); oxacillin (OXA); paromomycin (PRM); penicillin (PEN); polymyxin (PMB); ribostamycin (RSM); sisomicin (SIS); spectinomycin (SPT); streptomycin (STR); sulfacetamide (SFA); sulfadiazine (SDZ); sulfadimidine (SDD); sulfadoxine (SDX); sulfamethizole (SMZ); sulfamethoxazole (SMX); sulfasalazine (SSZ); sulfisox-

Table 4.2: Determination of antimicrobial resistance (AMR) with Resfinder[a]

OLC no.	Resistance Gene	AMR Class	Accession Number	AMR	Reference
455	*aadA24*	Aminoglycoside	AM711129	STR SPT	Rodríguez et al. 2008
	aph(3')-Ic	Aminoglycoside	X62115	KAN NEO	Lee et al. 1990
	strA	Aminoglycoside	M96392	STR	Lee et al. 1990
	strB	Aminoglycoside	M96392	STR	Lee et al. 1990
	catA1	Phenicol	V00622	CHL	Alton and Vapnek 1979
	sul1	Sulphonamide	CP002151	SUL	Szczepanowski et al. 2011
	tet(A)	Tetracycline	AJ517790	TET	Sørum et al. 2003
637	*strB*	Aminoglycoside	M96392	STR	Lee et al. 1990
	strA	Aminoglycoside	M96392	STR	Lee et al. 1990
	sul2	Sulphonamide	GQ421466	SUL	Kuo et al. 2010
	tet(B)	Tetracycline	AP000342	TET	Giufré et al. 2015
639	*strB*	Aminoglycoside	M96392	STR	Lee et al. 1990
	strA	Aminoglycoside	M96392	STR	Lee et al. 1990
	sul2	Sulphonamide	GQ421466	SUL	Kuo et al. 2010
	tet(B)	Tetracycline	AP000342	TET	Giufré et al. 2015
714	*aac(3)-IId*	Aminoglycoside	EU022314	GEN	P.-L. Ho et al. 2010
	strB	Aminoglycoside	M96392	STR	Lee et al. 1990
	aph(3')-Ia	Aminoglycoside	V00359	KAN NEO	Oka et al. 1981
	aadA5	Aminoglycoside	AF137361	STR SPT	Sandvang 1999
	strA	Aminoglycoside	AF321551	STR	Sundin 2002
	$bla_{\text{TEM-1B}}$	Beta-lactam	JF910132	PEN	Kubasova et al. 2014

OLC no.	Resistance Gene	Antibiotic Class	Accession Number	AMR	Reference
	catA1	Phenicol	V00622	CHL	Alton and Vapnek 1979
	sul2	Sulphonamide	HQ840942	SUL	Cain and Hall 2012
	sul1	Sulphonamide	CP002151	SUL	Szczepanowski et al. 2011
	tet(A)	Tetracycline	AJ517790	TET	Sørum et al. 2003
	dfrA17	Trimethoprim	FJ460238	TMP	Wiesner et al. 2009
719	*strB*	Aminoglycoside	M96392	STR	Lee et al. 1990
	strA	Aminoglycoside	M96392	STR	Lee et al. 1990
	sul2	Sulphonamide	GQ421466	SUL	Kuo et al. 2010
	tet(B)	Tetracycline	AP000342	TET	Giufré et al. 2015
969	*strA*	Aminoglycoside	AF321551	STR	Sundin 2002
	strB	Aminoglycoside	M96392	STR	Lee et al. 1990
	$bla_{\text{TEM-IB}}$	Beta-lactam	JF910132	PEN	Kubasova et al. 2014
	sul2	Sulphonamide	GQ421466	SUL	Kuo et al. 2010
	tet(A)	Tetracycline	AJ517790	TET	Sørum et al. 2003
	dfrA8	Trimethoprim	U10186	TMP	Sundström, Jansson, et al. 1995
1064	*strA*	Aminoglycoside	M96392	STR	Lee et al. 1990
	strB	Aminoglycoside	M96392	STR	Lee et al. 1990
1068	*strB*	Aminoglycoside	M96392	STR	Lee et al. 1990
	strA	Aminoglycoside	AF321551	STR	Sundin 2002
	$bla_{\text{TEM-1B}}$	Beta-lactam	JF910132	PEN	Kubasova et al. 2014
	$bla_{\text{CTX-M-15}}$	Beta-lactam	DQ302097	PEN	W. Liu et al. 2009
	sul1	Sulphonamide	CP002151	SUL	Szczepanowski et al. 2011
	sul2	Sulphonamide	HQ840942	SUL	Cain and Hall 2012
	tet(A)	Tetracycline	AJ517790	TET	Sørum et al. 2003

OLC no.	Resistance Gene	Antibiotic Class	Accession Number	AMR	Reference
	dfrA7	Trimethoprim	JF806498	TMP	Labar et al. 2012
1105	*strB*	Aminoglycoside	M96392	STR	Lee et al. 1990
	strA	Aminoglycoside	AF321551	STR	Sundin 2002
	$bla_{\text{TEM-1A}}$	Beta-lactam	HM749966	PEN	Bailey et al. 2011
	sul2	Sulphonamide	HQ840942	SUL	Cain and Hall 2012
	tet(A)	Tetracycline	AJ517790	TET	Sørum et al. 2003
	dfrA5	Trimethoprim	X12868	TMP	Sundström, Rådström, et al. 1988
1112	aadA1	Aminoglycoside	JQ414041	STR SPT	Kor et al. 2013
	aac(3)-VIa	Aminoglycoside	M88012	GEN	Rather et al. 1993
	sul1	Sulphonamide	CP002151	SUL	Szczepanowski et al. 2011
	tet(A)	Tetracycline	AJ517790	TET	Sørum et al. 2003
1128	*aadA1*	Aminoglycoside	JN815078	STR SPT	Najibi et al. 2012
	sul1	Sulphonamide	CP002151	SUL	Szczepanowski et al. 2011
	tet(A)	Tetracycline	AJ517790	TET	Sørum et al. 2003
1330	*strB*	Aminoglycoside	M96392	STR	Lee et al. 1990
	strA	Aminoglycoside	AF321551	STR	Sundin 2002
	$bla_{\text{TEM-1C}}$	Beta-lactam	FJ560503	PEN	Zou et al. 2011
	sul2	Sulphonamide	GQ421466	SUL	Kuo et al. 2010
1334	*aph(6)-Ic*	Aminoglycoside	X01702	STR	Mazodier et al. 1985
	aph(3')-IIa	Aminoglycoside	X57709	KAN NEO	Blázquez et al. 1991
	aadA1	Aminoglycoside	JN815078	STR SPT	Najibi et al. 2012

OLC no.	Resistance Gene	Antibiotic Class	Accession Number	AMR	Reference
	sul1	Sulphonamide	AY224185	SUL	Dubois et al. 2003
	tet(A)	Tetracycline	AJ517790	TET	Sørum et al. 2003
1337	*strB*	Aminoglycoside	M96392	STR	Lee et al. 1990
	strA	Aminoglycoside	AF321551	STR	Sundin 2002
	floR	Phenicol	AF118107	CHL FFC	Arcangioli et al. 1999
	sul2	Sulphonamide	GQ421466	SUL	Kuo et al. 2010
	tet(A)	Tetracycline	AJ517790	TET	Sørum et al. 2003
1356	*aadA1*	Aminoglycoside	JN815078	STR SPT	Najibi et al. 2012
	sul1	Sulphonamide	CP002151	SUL	Szczepanowski et al. 2011
	tet(A)	Tetracycline	AJ517790	TET	Sørum et al. 2003
1357	*aadA1*	Aminoglycoside	JN815078	STR SPT	Najibi et al. 2012
	sul1	Sulphonamide	AY224185	SUL	Dubois et al. 2003
	tet(A)	Tetracycline	AJ517790	TET	Sørum et al. 2003
1358	*aadA1*	Aminoglycoside	JN815078	STR SPT	Najibi et al. 2012
	sul1	Sulphonamide	CP002151	SUL	Szczepanowski et al. 2011
	tet(A)	Tetracycline	AJ517790	TET	Sørum et al. 2003
1558	*strB*	Aminoglycoside	M96392	STR	Lee et al. 1990
	strA	Aminoglycoside	M96392	STR	Lee et al. 1990
	sul2	Sulphonamide	GQ421466	SUL	Kuo et al. 2010
	tet(B)	Tetracycline	AP000342	TET	Giufré et al. 2015

[a]ResFinder v2.1 (Accessed March 31, 2016).

Table 4.3: AMR predictive accuracy for STEC strains by broth method[a]

Antibiotic	ARMI	ResFinder
Chloramphenicol	3 / 3	3 / 3
Florfenicol	1 / 1	1 / 1
Gentamicin C	2 / 2	2 / 2
Kanamycin	3 / 3	2 / 3
Neomycin	2 / 3	2 / 3
Penicillin G	5 / 5	5 / 5
Spectinomycin	8 / 8	8 / 8
Streptomycin	15 / 18	15 / 18
Sulfamethoxazole	15 / 16	16 / 17
Trimethoprim	4 / 4	4 / 4
Total	58 / 63	58 / 64

Table 4.4: AMR predictive accuracy for STEC strains by disc diffusion method[a]

Antibiotic	ARMI	ResFinder
Apramycin	2 / 2	0 / 0
Chloramphenicol	3 / 3	3 / 3
Kanamycin	2 / 3	2 / 3
Neomycin	2 / 3	2 / 3
Spectinomycin	8 / 8	8 / 8
Sulfamethoxazole	15 / 16	16 / 17
Trimethoprim	4 / 4	4 / 4
Total	36 / 39	35 / 38

[a] A panel of 18 *E. coli* strains (Table 4.1) was subjected to growth in broth culture in the presence of 10 and 100 µg/ml of each antibiotic, and growth was measured turbidimetrically and subjected to AMR determination using a commercial disc diffusion method. Results for Table 4.3 are reported as total number of cultures for all bacteria showing growth (resistance) per number of predictions for a particular antibiotic. Note that for each instance where resistance was noted, growth occurred at both antibiotic concentrations (with the exception of chloramphenicol, for which growth was only observed at 10 µg/ml). ARMI positive predictive value = 92.%; ResFinder positive predictive value = 91%. Results inTable 4.4 are reported as total number of positive AMR observations per number of predictions for a particular antibiotic. ARMI positive predictive value = 92%; ResFinder positive predictive value = 92%.

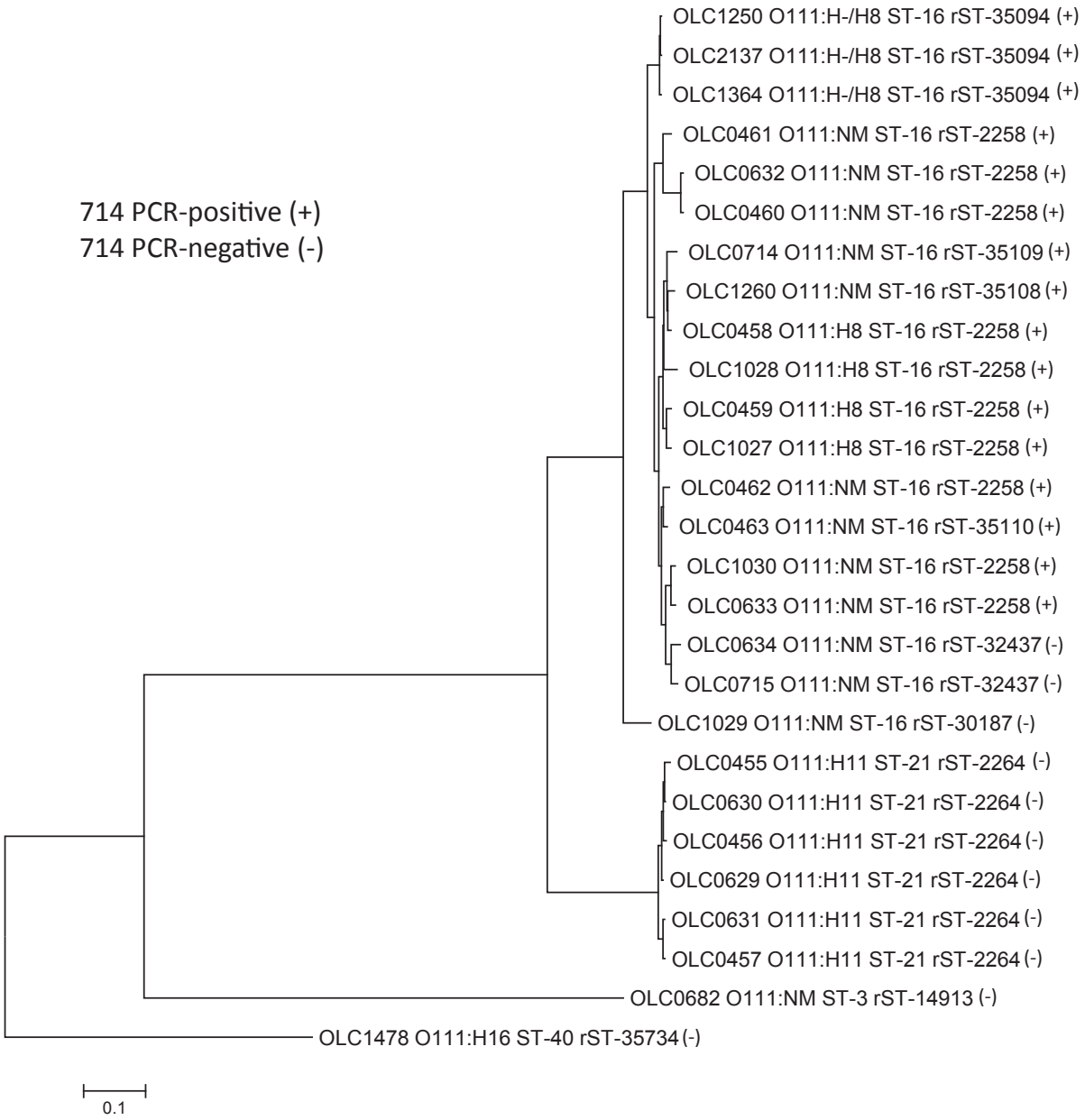

Figure 4.1: Reactivity of 714 – PCR with different *E. coli* serogroup O111 strains. The evolutionary history of the strains used in this study was inferred from core SNPs identified wth kSNP v3 using the Maximum Parsimony method. Serotype, multilocus sequence type (MLST) and ribosomal sequence type (rST) of isolates are indicated.

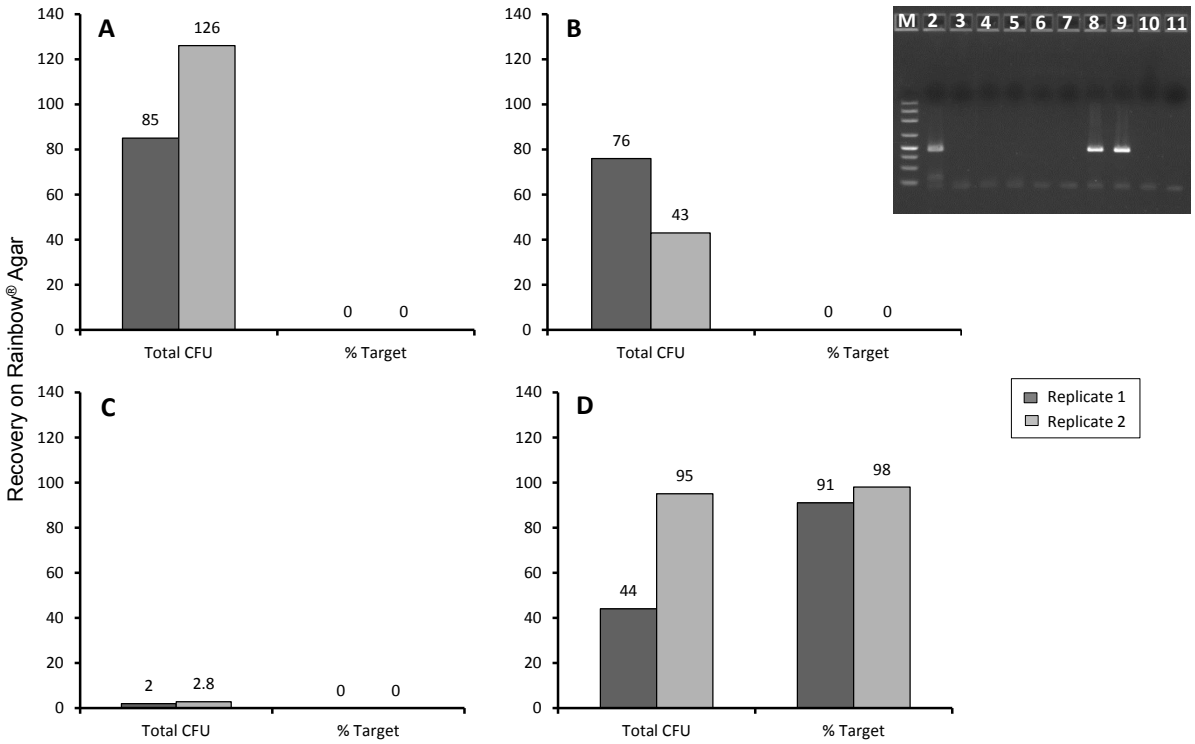

Figure 4.2: Recovery of *E. coli* O111:NM strain OLC-714 from ground beef. A, uninoculated ground beef enriched in the absence of trimethoprim; B, inoculated ground beef enriched in the absence of trimethoprim; C, uninoculated ground beef enriched in the presence of trimethoprim; D, inoculated ground beef enriched in the presence of trimethoprim. (Inset: analysis of enrichment cultures by the 714 polymerase chain reaction (PCR): M, 50-1500 bp DNA Marker; 2, inoculated ground beef enriched with trimethoprim; 3, inoculated ground beef enriched without trimethoprim; 4, un-inoculated ground beef enriched with trimethoprim; 5, un-inoculated ground beef enriched without trimethoprim; 6, un-inoculated enrichment medium without trimethoprim; 7, un-inoculated enrichment medium with trimethoprim; 8, inoculated enrichment medium without trimethoprim; 9, inoculated enrichment medium with trimethoprim; 10, PCR control: 1% Triton X-100; 11, PCR control: PCR reaction mixture without template DNA).

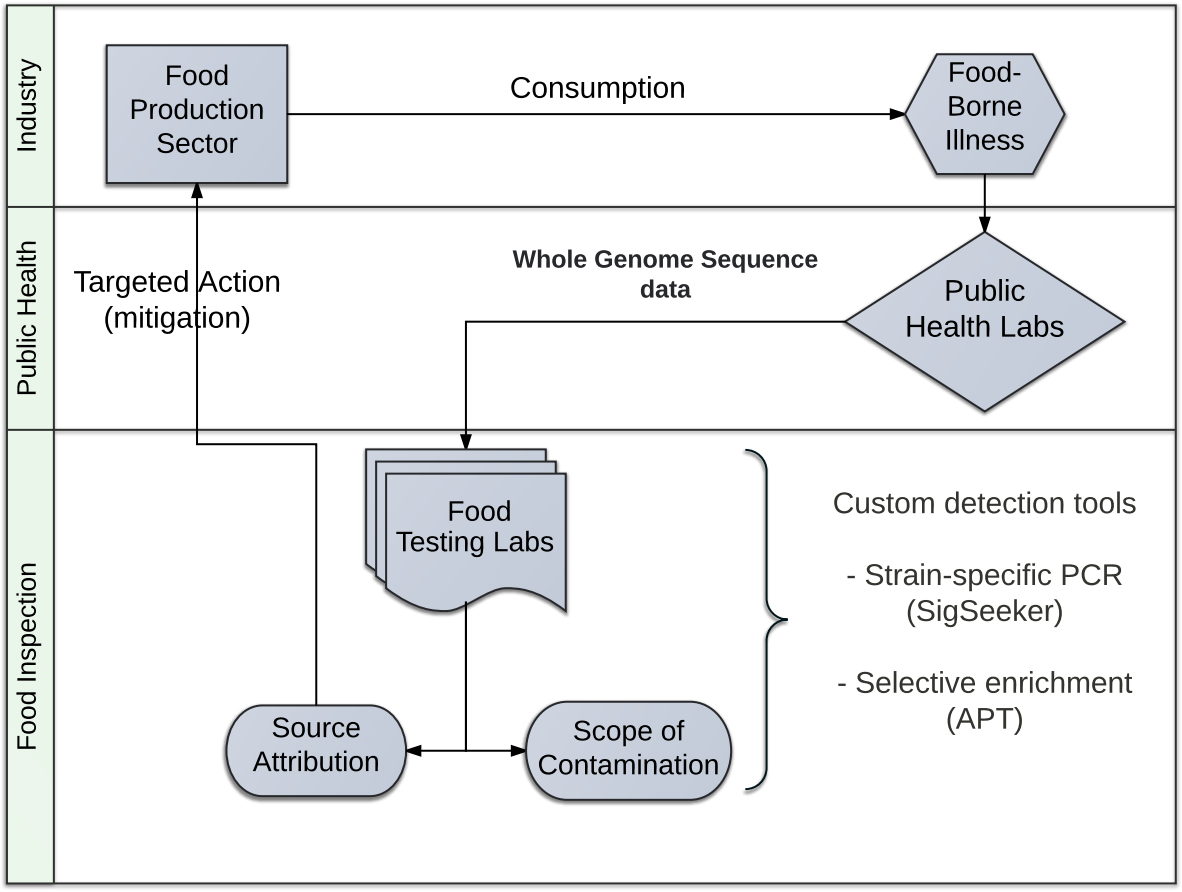

Figure 4.3: Scenario for the integration of genomics technology in an outbreak investigation.

Chapter 5

Discussion

5.1 Bioinformatics in Unique Bacterial Strain Detection

Chapter 2 described an avenue to discover sequences of DNA within *de novo* assembled *Escherichia coli* and *Salmonella* strains which are suitable to design primers using well-established tools like Primer-BLAST (Ye et al. 2012). We evaluated this tool which we named SigSeekr, to find 100% specificity and sensitivity to four target strains. SigSeekr was able to fulfill our objectives of identifying unique DNA sequences in a bacterial target.

In a recent fast-track communication, the ssGeneFinder webserver was able to find PCR targets for an *E. coli* strain O104:H4 using a database of nine *E. coli* genomes (C.-C. Ho et al. 2012). Although this method was successful for clinical isolates from human stool samples where the bacterial environment is well characterized, we found no results when comparing *E. coli* O157:H7 to a database of *E. coli* O157:H7/NM with the ssGeneFinder method of *in silico* subtraction. This prompted us to devise a new strategy while maintaining the concept of *in silico* subtraction with Basic Local Alignment Tool nucleotide (BLASTn). When the strategy was changed to involve a variable E-value of the sequence we were able to find unique strain-specific sequences with masking of genomes being manipulated precisely with E-value variation of a BLAST database search. Multiple iterations of BLAST *in silico* subtraction also proved unacceptable for our purposes because the E-value fell below a threshold that produced unique results. The calculation of the E-value in a BLAST database search includes the length and identity of the matching pair as parameters (Altschul et al. 1990). This can prove useful

when employing a BLAST database search for finding relatively distant similarities in organisms (Altschul et al. 1990). However, since our target size was relatively small, 200 nucleotides, and a low E-value could return large sequences with relatively low percent identity, we included an identity cutoff of 90%. Therefore, sequences from genomes in the database matching those in our target sequence would only be eliminated if they were above the E-value and below the percent identity of database search.

The PCR described in Chapter 2 and Table 2.5 depicts a delineation of 3 *E. coli* against the rest of the 100 bacterial isolates in the panel of 91 phenotypically related *E. coli* strains, including 5 serogroup O111 strains and 51 serogroup O157 strains and 12 non-*E. coli*. This is indicative that the database of 1600 genomes collected from GenBank has an inclusivity to encompass variations in *E. coli* strains not present in the database. The SigSeekr pipeline is capable of generating unique sequences that are 100% exclusive on a panel of phylogenetically close bacteria. The identification of both the parent and progeny from a bioinformatics designed probe show 100% inclusivity. The exclusion of the parental genome sequence for OLC 797, OLC 795, in Table 3.1 from the genomic database confirms that inclusivity upon the accumulation of a point mutation is not altered. Mutation occurs naturally in populations and therefore, its added variable of randomness does not significantly alter effective probe generation. In fact this mitigates some issues faced in trace back analysis of foods to divergent populations of the same strain in distinct regions (Eppinger et al. 2011).

We expanded our scope to evaluate SigSeeker with *Salmonella enterica* subsp. enterica Serovar Mishmarhaemek, a strain commonly used a positive control in *Salmonella enterica* routine food testing, but with a combination of two signature sequences was required to distinguish it from other phenotypically similar salmonellae in the reference *Salmonella* database, culminating in a multiplex polymerase chain reaction (PCR) using 2 sets of primer pairs producing 554 and 150 bp amplicons with the target strain, but not with 99 non-target *S. enterica* or 30 non-*Salmonellae* strains (Table 2.6). This evaluation exemplifies that SigSeekr is adaptable to find multiple sequences that are unique to a target strain. In fact, one of the two primer sets designed by SigSeekr appeared to produce a 554 bp amplicon with only *Salmonella enterica* subsp. enterica Serovar Mishmarhaemek (Table 2.6). However, a *Salmonella enterica* strain exists in the GenBank database that produces this amplicon. This is indicative that the database of 120 genomes collected from GenBank has an inclusivity to encompass variations in *Salmonellae* strains not present in the database. From isolation of the bacteria to PCR probe generation our method uses genomics and bioinformatics to inform the design of strain-specific PCR

primers for precise detection of ad hoc outbreak VTEC strains in foods. Our method is novel in the field of pan-genomic analysis because of our multi-step iterative algorithm to conclusively differentiate bacterial strains from close relatives. For instance, PGAP (Zhao et al. 2012) and a recent study on *Propionibacterium acnes* (Tomida et al. 2013) pan-genomic analyses use BLAST for genome alignment and focus on single nucleotide polymorphisms (SNPs) for differentiation. A previous *Streptococcus agalactiae* study (Tettelin et al. 2005) and Panseq (Laing et al. 2010) pan-genomic analyses use whole genome alignment software. These pan-genomic methods focus on detecting diversity of amongst the pan-genome population. SNP analysis allows researchers to determine phylogenetic distances relations quickly with a desktop computer (Zhao et al. 2012; Tomida et al. 2013) but their inherent short length makes it unsuitable for engineering unique PCR probes. Alternatively, whole genome alignment affords the possibility of unique sequence discovery for multiple genome comparison (Edwards and Holt 2013). However, whole genome alignment software is cumbersome and slow compared to BLAST heuristics implemented in our *in silico* subtraction algorithms (Edwards and Holt 2013; Altschul et al. 1990). This resulted in our SigSeekr pipeline producing quick and precise unique sequence results an aptitude for ad hoc PCR probe creation through the use of genomics and bioinformatics to inform the design. Other methods of pan-genomic analysis have been studied and are either inefficient or ineffective at producing significantly unique PCR probes for a single *E. coli* isolate. This includes genotyping methods like MLST but this requires equipment and training and lacks the same specificity of pan-genomic informed methods (Laing et al. 2010). Moreover, many methods are focused on determining homologous sequences (Zhao et al. 2012; Tomida et al. 2013; Tettelin et al. 2005; Laing et al. 2010).

Our SigSeekr pipeline with an E-value of 10^{-55} and a cutoff identity of 90%, and was found to return nucleotide sequences unique to a locally sequenced *Escherichia coli* and *Salmonella enterica* subsp. enterica Serovar Mishmarhaemek (Cooper et al. 2015). This sequence was manually confirmed against NCBI Online nucleotide BLAST chromosome database, returning no hit. Based on the unique sequence, Primer-BLAST created two primer pairs for each strain (Table 2.1). These primers were used in binary classification PCR assays with other isolates (Tables 2.5 and 2.6) These results demonstrate the effectiveness of *in silico* subtraction approach described here.

Probes generated by our SigSeekr pipeline are simple and expedient in implementation to frontline testing laboratories which are often directly responsible for traceback analyses. The *ad hoc* nature of our platform will allow for contingencies where only

the sequence data of the isolate and not the physical isolate itself are available. This platform would simply require further validation in a variety of scenarios often used in verotoxin-producing *Escherichia coli* (VTEC) detection protocols.

The SigSeekr tool was successfully used to supplement traditional laboratory techniques using PCR to detect a laboratory control strain in Chapter 3. The experiments described in Chapter 3 shortened traditional techniques by 20 hours. In some cases (e.g., detection of VTEC of public health concern), phenotypic methods are entirely impractical as a means of identification. Ultimately, these techniques are limited in terms of the type of information (e.g., risk profiling) that can be garnered from an isolate to underscore risk management decisions. This application exemplifies the possibility to design primers for PCR that can unique identify strains of VTEC or *Salmonella* with unanticipated virulence characteristics.

5.2 Combining Genomic Tools for Selective Recovery of Bacteria

Genomics approaches provide novel possibilities for comprehensive analysis of foodborne bacteria such as verotoxin-producing *Escherichia coli* (VTEC), fostering the development of tools that would support the rapid detection of outbreak strains. Next-generation sequencing (NGS) technologies can now render a bacterial genome much faster and at a significantly lower cost than previously possible (Lambert et al. 2015). It may be possible to determine the presence of antibiotic resistance markers through the use of bioinformatics tools that would enable the application of a customized selective recovery media, enhancing detection against the competing background of non-target bacteria. In order to utilize these genomic practices, AMR predictive tools must first be validated to determine how accurately they are able to predict phenotypic resistance.

A number of tools are currently available to predict AMR genes from bacterial WGS data, however only two tools exist that include curated international AMR gene databases that are relatively complete and regularly updated. The first tool is ARMI, which is based on the gene databases ARDB (B. Liu and Pop 2009), CARD (McArthur et al. 2013), and ARG-ANNOT (Gupta et al. 2014). The second tool is ResFinder (Kleinheinz et al. 2014), which is based on ARDB and thorough literature searches. It is important to note that ARDB has not been regularly updated; however the ResFinder tool has been consistently maintained. The ability to identify AMR markers in a bacterial genome is largely

predicated on the nature of the reference gene database, which the prediction algorithm is directed to search. For the purpose of the development of customized enrichment broths, the increased scope of the CARD database was advantageous. The extensive curation of this database was also extremely useful, as ARMI was able to parse data to rapidly identify target antibiotics (Table 4.1). The ARMI tool designed in our laboratory matches bacterial strain genomic data with one of the most comprehensive antimicrobial resistance (AMR) gene databases currently available, CARD, and its output is a listing of specific antibiotics to which the organism is putatively resistant. The primary output of ResFinder is in the form of antibiotic classes, though specific genes are identified and their function (i.e., target antibiotic) can be determined by cross-referencing the corresponding entries in the National Center for Biotechnology Information (NCBI) repository and inferring resistance.

The main object of this research was to integrate genomics tools for the selective recovery and identification of an *E. coli* by having only a genomic sequence. Chapters 2 and 3 describe the identification methods using genomics to find unique sequences then perform a polymerase chain reaction (PCR) assay. To integrate a selective element, we performed a comprehensive literature/public database search to catalogue AMR sequences known to exist among VTEC and related bacteria (e.g., ARDB, CARD, and PATRIC) (B. Liu and Pop 2009; McArthur et al. 2013; Wattam et al. 2014). Followed by an analysis to identify AMRs in of WGS data from a small selection of VTEC strains from our bacterial database (different priority serotypes, as well as multiple strains within a given serotype). A model for the general deployment of these *ad hoc* strain-specific tools in support of outbreak investigations is illustrated in Fig. 4.3. Outbreak events associated with consumption of contaminated food products are initially recognized when public health laboratories recover matching clinical isolates from several affected individuals. The genomes of clinical isolates are rapidly sequenced for high resolution typing analyses to establish phylogenetic relationships among different isolates. Once an outbreak is recognized the sequence is rapidly disseminated to partner investigative agencies such as the food inspection laboratory.

The applicability of AMR marker prediction tools (APTs) in informing the customization of target strain-specific selective enrichment media was examined by studying the recovery of a model target strain from a food sample containing a high level of background bacteria (e.g., ground beef). For this purpose, an *E. coli* O111:NM strain (OLC 714) was selected as a model, along with the use of trimethoprim for its selective recovery from inoculated ground beef, since resistance to this antibiotic predicted for this strain

(Table 4.1) occurred with a moderate degree of rarity (4 occurrences as determined by ResFinder and ARMI) among the panel of 18 *E. coli* strains examined. Recovery of the model strain was assessed by assaying enrichment cultures using a PCR method (714-PCR) featuring primers (Table 2.1) targeting DNA sequences associated with the strain of interest identified by SigSeekr. The narrow reactivity of this PCR method for OLC-714 (and closely related bacteria [i.e., some other *E. coli* O111 strains]) was confirmed through the observation that these primers only produced the expected amplicon with the target strain and several very closely related *E. coli* O111 strains of the same sequence type (ST-16), but not with other *E. coli* O111 strains (Fig. 4.1), nor with a broader panel of *E. coli* strains (Table 4.1).

The efficacy of trimethoprim as a strain-specific selective agent aiding in the recovery of *E. coli* O111:NM (OLC-714) from ground beef samples containing a high level of background bacteria was examined using the 714 PCR for detection of the target organism in broth enrichment cultures as well as to identify colonies subsequently recovered on plating media. The use of trimethoprim was necessary to enable detection of strain OLC-714 directly in the enrichment broth culture of an inoculated sample (Fig. 4.2 inset, lane 2), presumably because in the absence of trimethoprim the background microbiota inhibited the growth of the target strain (Fig. 4.2, inset, lane 3). No PCR product was produced with the uninoculated samples regardless of whether or not trimethoprim was present (Fig. 4.2 inset, lanes 4 and 5), nor was there any evidence of interference with PCR performance from any of the enrichment or sample preparation components (Fig. 4.2 inset, lanes 6-11), supporting the conclusion that trimethoprim was solely responsible for enabling successful detection of the target organism in the enrichment sample.

The impact of trimethoprim on recovery efficiency for the target strain was studied by plating portions of duplicate ground beef enrichment samples on Rainbow Agar® and determining the proportion of recovered colonies which bore identity to the subject strain using the 714-PCR. Large numbers of colonies were recovered from samples devoid of the target strain enriched in the absence of trimethoprim, but none of these colonies tested positive with the 714-PCR (Fig. 4.2, panel A), suggesting that they were the product of enrichment of the ground beef background microbiota. Large numbers of colonies were also recovered in the absence of trimethoprim from samples inoculated with the target strain, but none of these were positive in the 714-PCR (Fig. 4.2, panel B), indicating that the recovery of OLC 714 was inhibited by an overwhelming level of background microbiota (which were present in greater than 3000-fold excess relative to the inoculated strain). Very few colonies were recovered from uninoculated samples

enriched in the presence of trimethoprim (Fig. 4.2, panel C), none of which were reactive in the 714-PCR, demonstrating the effectiveness of trimethoprim in suppressing the growth of the background bacteria. Finally, when inoculated ground beef samples were enriched in the presence of trimethoprim almost all of the colonies recovered on plating media produced positive results with the 714-PCR (Fig. 4.2, panel D), suggesting that the background microbiota were effectively suppressed, allowing the unhampered growth of the target strain. These results clearly demonstrate the beneficial impact of incorporating trimethoprim in the enrichment broth.

For the panel of antibiotics examined, ARMI gave an slightly higher number of total predictions with a positive predictive value (positive predictive value (PPV) of 92%, compared to a PPV of 90% obtained with ResFinder in Table 4.3. In order to achieve this level of performance with ARMI it was necessary to modify the search database to eliminate problematic AMR markers.

While the present experiments are limited in that only one antibiotic resistance scenario (trimethoprim) was studied in detail, it may be reasonably inferred from the results presented in Table 4.3, which demonstrate successful cultivation of various VTEC strains in enrichment broth incorporating different antibiotics, that other antibiotic resistance traits should be amenable to the concept of customization of selective enrichment tailored to a strain of interest. More work will be needed to validate APTs and better define which AMR predictions and antibiotic growth conditions will work reliable in a food enrichment scenario such as the present example. Work is underway in our laboratory to improve the APTs and develop a catalogue of antibiotics (and the conditions for their use) from which a selection could be made on an *ad hoc* basis to support an active food safety investigation.

5.3 Obstacles in Research

During the work completed in Chapter 4 (Genomic tools for Customized Recovery and Detection of Foodborne Shiga-Toxigenic *Escherichia coli*) the primers designed for our target strain (OLC 714) were not fully exclusive to our target strain. Several verotoxin-producing *Escherichia coli* (VTEC) O111 strains were further tested using kSNP (Gardner et al. 2015) to determine their genetic distances in Fig. 4.1. We determined that these additional 27 strains of *E. coli* serogroup O111 from the OLC culture collection were related closely enough to produce the 183 bp amplicon in Table 2.1 in the PCR assay in Table 2.3. Therefore, with just the PCR assay alone several of the non-target

strains would give a positive result. However, none of these non-target O111 VTEC strain carried the modified dihydrofolate reductase gene (*dfr*A17 found in OLC 714) capable of withstanding prolonged incubation in trimethoprim (100 g/L). We were able to show that the two assays using ARMI and SigSeekr combined gave a more conclusive result. The research with ARMI and SigSeekr was conducted for short-term purposes in mind. We were aware at the onset of experimentation methods like SigSeekr and ARMI were not suitable for long-term investigation due to the rate of DNA mutation within VTEC generations. Therefore SigSeekr and ARMI predictions could not reliably used if the bacteria are exposed to different environments and made to adapt. The SigSeekr and ARMI approaches are equivalent to a DNA snapshot at a particular point in time which is completely suited in foodbourne illness trackback investigations.

5.4 Future Work

Further work has been completed after submission of Chapter 2 for publication that allows SigSeekr to find regions that containing unique variation only at the end of a suitably large (200 bp or greater) nucleotide sequence. We enhanced the functionality of regular expression that allowed us to overlap matching results so we could find common nucleotide sequences flanked by small (approx. 30mer) unique regions on either end. This was accomplished through the use of a model that treated each 30mer unique region separately (with lookaheads and lookbehinds) then concatenated the sequence by converting the degenerate bases back to genomic data so that the common element of sequence was attached to a unique sequence on either end to make an optimal amplicon. Then we expanded the functionality of regular expressions (which are not normally able to overlap matches) to a sliding window model where the matches sequentially use the ending unique sequence as a beginning for the next result.

A conclusive demonstration of the link between the occurrence of antimicrobial resistance (AMR) bacteria in foods and food production practices such as supplementation of animal feeds with antibiotics and the presence of metal contaminants in fertilizers would inform the development of best practices (and possibly new regulatory policies) safeguarding public health from avoidable AMR risks. There is a need for systematic studies to provide information leading to a better understanding of the role of food manufacturing practices and impinging factors on the emergence of AMR bacteria. The CFIA is ideally positioned to undertake studies of linearly connected pre- and post-harvest samples in the Canadian food production system owing to the structure of current sampling

programs targeting fertilizers, animal feeds and animal-derived foods.

While traditional approaches of measuring the resistome involve the isolation and cultivation of live bacteria in the presence of different antibiotics, more recent metagenomic approaches in which DNA extracted from the entire microbiome of a sample (e.g., cattle manure, meats, etc.) is queried for the presence of AMR gene markers can yield a more comprehensive picture of the true potential for transfer of resistance from such a reservoir (Wichmann et al. 2014). Though metagenomic analysis in the context of food safety is known (Bergholz et al. 2014), current approaches are more suited to an academic line of enquiry and are not designed for implementation as food inspection tools supporting ready assessment of the AMR burden associated with different samples collected throughout the food production continuum. To be suitable as an inspection tool, an analytical procedure needs to be simple to carry out and the analytical targets (AMR markers) and the type of sample in which they are queried must be clearly defined. Ideally, the analytical procedure (which includes the steps of sample preparation, whole genome sequencing and bioinformatics analysis) should be simple, fit-for-purpose and provide a user-friendly report for broad interpretation. The sample type and the interval of sampling should be selected on the basis of maximizing opportunities for observation of AMR gene flow among readily recoverable microbial populations or subpopulations that are representative of the most relevant vectors trafficking AMR traits throughout the food production continuum and into human populations. Such "sentinel" bacteria should constitute major intersections where microbial hazard vectors, environmental reservoirs and food manufacturing inputs converge. It is hereby posited that for production systems based on meat animals such as cattle, poultry, swine, etc., the *Enterobacteriacea* constitute a reasonable candidate sub-population for this purpose.

Future work for ARMI has already commenced, the new version allows the input of 'raw' whole-genome sequence (WGS) data. The input WGS is aligned to the AMR gene by ARMI with bowtie2 (Langmead and Salzberg 2012) using a Burrow-Wheeler alignment algorithm. Bowtie2 allows alteration of Eq. (1.6) to allow for 'fuzzy' alignments along a prefix trie. The raw ARMI is still in a testing phase, ultimately the tool is capable of detecting background AMR (resistome) in a WGS metagenome sample of commensal bacteria in a food commodity. This will require a benchmark to compare with tools like Resfinder (Kleinheinz et al. 2014), and SRST2 (Inouye et al. 2014) which already process 'raw' data. Metagenomic samples such as those from fertilizers, animal feeds and animal-derived food, would be ideal candidates for pilot surveillance studies on the detection and characterization of AMRs using ARMI as an AMR marker prediction tool.

Appendix A

Glossary of Terms

de novo Latin idiom to begin anew. In genomics, assembling short reads to create full-length (sometimes novel) sequences. iv, 2, 3

lookahead Matches at a position where the pattern inside the lookahead can be matched. Matches only the position. It does not consume any characters or expand the match.. 92

lookbehind Matches at a position if the pattern inside the lookbehind can be matched ending at that position.. 92

sensitivity (also called the true positive rate, or the recall in some fields) measures the proportion of positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition).. 85

specificity (also called the true negative rate) measures the proportion of negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition).. 85

Appendix B

Glossary of Acronyms

AMR antimicrobial resistance. ii, 13, 15–18, 58, 60, 61, 63, 64, 66–68, 70, 71, 75, 81, 89, 91–93

APT AMR marker prediction tool. 17, 60, 61, 67, 68, 70, 71, 89, 91, 93

ARDB Antibiotic Resistance Genes Database. 60, 63, 64

ARG-ANNOT Antibiotic Resistance Gene-ANNOTation. 60, 63

BLAST Basic Local Alignment Tool. 8, 9, 11, 22–24, 43, 44, 85–87

CARD The Comprehensive Antibiotic Resistance Database. 60, 63, 66, 67

CFIA Canadian Food Inspection Agency. 1, 8, 17, 18

NGS next-generation sequencing. iv, 2, 3, 5, 7–9, 13, 88

PCR The polymerase chain reaction is a process used in molecular biology to amplify a single copy or a few copies of a piece of DNA across several orders of magnitude, generating thousands to millions of copies of a particular DNA sequence.. ii, vii, 1, 2, 8, 16, 17, 20–40, 42–48, 51–56, 58, 60, 61, 64–66, 68–71, 83, 86, 88–91

PPV $= \frac{\text{number of true positives}}{\text{number of true positives+number of false positives}} = \frac{\text{number of true positives}}{\text{number of positive calls}}$. 67, 68, 91

VTEC verotoxin-producing *Escherichia coli*. ii, 1, 2, 17, 18, 88, 89, 91, 92

WGS whole-genome sequence. 93

Appendix C

References

Alton, N. K. and D. Vapnek (1979). "Nucleotide sequence analysis of the chloramphenicol resistance transposon Tn9." *Nature*, 282 (5741): pp. 864–869.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (1990). "Basic local alignment search tool." *Journal of Molecular Biology*, 215 (3): pp. 403–410.

Anonymous (2014). "Isolation of *Escherichia coli* O157:H7/NM from foods and environmental surface samples (MFHPB-10)." In: *Compendium of Analytical Methods*. Vol. 2. Accessed March 20, 2015. Available at: http://www.hc-sc.gc.ca/fn-an/res-rech/analy-meth/microbio/volume2-eng.php.

Arcangioli, M. A., S. Leroy-Sétrin, J. L. Martel, and E. Chaslus-Dancla (1999). "A new chloramphenicol and florfenicol resistance gene flanked by two integron structures in Salmonella typhimurium DT104." *FEMS Microbiol Lett*, 174 (2): pp. 327–332.

Bailey, J. K., J. L. Pinyon, S. Anantham, and R. M. Hall (2011). "Distribution of the blaTEM gene and blaTEM-containing transposons in commensal Escherichia coli." *J Antimicrob Chemother*, 66 (4): pp. 745–751.

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner (2012). "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing." *Journal of Computational Biology*, 19 (5): pp. 455–477. ISSN: 1557-8666.

Barton, M. D. (2000). "Antibiotic use in animal feed and its impact on human health." *Nutrition research reviews*, 13: pp. 279–99. ISSN: 0954-4224 (Linking).

Bergholz, T. M., A. I. Moreno Switt, and M. Wiedmann (2014). "Omics approaches in food safety: fulfilling the promise?" *Trends in Microbiology*, 22 (5): pp. 275–281. ISSN: 0966-842X.

Blais, B. W., M. Gauthier, M. Deschênes, and G. Huszczynski (2012). "Polyester cloth-based hybridization array system for identification of enterohemorrhagic *Escherichia coli* serogroups O26, O45, O103, O111, O121, O145, and O157." *Journal of Food Protection*, 75 (9): pp. 1691–1697.

Blais, B. W. and A. Martinez-Perez (2011). "A Simple PCR-Based Macroarray System for Detection of Multiple Gene Markers in the Identification of Priority Enterohemorrhagic *Escherichia coli*." *Journal of Food Protection*, 74 (3): pp. 365–372.

Blais, B. W., A. Martinez-Perez, M. Gauthier, R. Allain, F. Pagotto, and K. Tyler (2008). "Development of Unique Bacterial Strains for Use as Positive Controls in the Food Microbiology Testing Laboratory." *Journal of Food Protection*, 71 (11): pp. 2301–2306.

Blais, B. W., A. Martinez-Perez, G. Huszczynski, S. White, and B. Gingras (2013). "Characterization of verotoxigenic*Escherichia coli*O157:H7 colonies by polymerase chain reaction (PCR) and cloth-based hybridization array system (CHAS) (MFLP-22)." In: *Compendium of Analytical Methods*. Vol. 3. Accessed March 20, 2015. Available at: http://www.hc-sc.gc.ca/fn-an/res-rech/analy-meth/microbio/volume3-eng.php.

Blázquez, J., J. Davies, and F. Moreno (1991). "Mutations in the aphA-2 gene of transposon Tn5 mapping within the regions highly conserved in aminoglycoside-phosphotransferases strongly reduce aminoglycoside resistance." *Mol Microbiol*, 5 (6): pp. 1511–1518.

Bonyadian, M., H. Moshtaghi, and M. Akhavan Taheri (2014). "Molecular characterization and antibiotic resistance of enterotoxigenic and entero-aggregative *Escherichia coli* isolated from raw milk and unpasteurized cheeses." *Veterinary Research Forum : an International Quarterly Journal*, 5: pp. 29–34. ISSN: 2322-3618 (Electronic).

Bosilevac, J. M. and M. Koohmaraie (2011). "Prevalence and Characterization of Non-O157 Shiga Toxin-Producing *Escherichia coli* Isolates from Commercial Ground Beef in the United States." *Applied and Environmental Microbiology*, 77 (6): pp. 2103–2112.

Burrows, M. and D. Wheeler (1994). *A Block-sorting Lossless Data Compression Algorithm*. A block-sorting lossless data compression algorithm no. 124. Digital, Systems Research Center.

Cain, A. K. and R. M. Hall (2012). "Evolution of a multiple antibiotic resistance region in IncHI1 plasmids: reshaping resistance regions in situ." *J Antimicrob Chemother*, 67 (12): pp. 2848–2853.

CDC (2013). *National Antimicrobial Resistance Monitoring System (NARMS) for Enteric Bacteria: Human Isolates Final Report*. Atlanta, Georgia: U.S. Department of Health and Human Services, CDC.

Clinical and Laboratory Standards Institute (2014). *Performance Standards for Antimicrobial Susceptibility Testing; Twenty-Fourth Informational Supplement. CLSI document M100-S24*. 950 West Valley Road, Suite 2500, Wayne, Pennsylvania 19087 USA: Clinical and Laboratory Standards Institute. ISBN: 1-56238-897-5.

Cock, P. J. A., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon (2009). "Biopython: freely available Python tools for computational molecular biology and bioinformatics." *Bioinformatics*, 25 (11): pp. 1422–1423.

Compeau, P. E. C., P. A. Pevzner, and G. Tesler (2011). "How to apply de Bruijn graphs to genome assembly." *Nature Biotechnology*, 29 (11): pp. 987–991. ISSN: 1087-0156.

Cooper, A., D. Lambert, A. G. Koziol, K. Seyer, and C. D. Carrillo (2015). "Draft Genome Sequence of *Salmonella enterica* subsp. enterica Serovar Mishmarhaemek Isolated from Bovine Feces." *Genome Announcements*, 3 (5).

Deininger, K. N. W., A. Horikawa, R. D. Kitko, R. Tatsumi, J. L. Rosner, M. Wachi, and J. L. Slonczewski (2011). "A Requirement of TolC and MDR Efflux Pumps for Acid Adaptation and GadAB Induction in *Escherichia coli*." *PLoS ONE*, 6 (4): pp. 1–7.

den Bakker, H. C., M. W. Allard, D. Bopp, E. W. Brown, J. Fontana, Z. Iqbal, A. Kinney, R. Limberger, K. A. Musser, M. Shudt, E. Strain, M. Wiedmann, and W. J. Wolfgang (2014). "Rapid whole-genome sequencing for surveillance of *Salmonella enterica* serovar enteritidis." *Emerging Infectious Diseases*, 20 (8): pp. 1306–1314.

Dubois, V., C. Arpin, C. Quentin, J. Texier-Maugein, L. Poirel, and P. Nordmann (2003). "Decreased susceptibility to cefepime in a clinical strain of Escherichia coli related to plasmid- and integron-encoded OXA-30 beta-lactamase." *Antimicrob Agents Chemother*, 47 (7): pp. 2380–2381.

Edwards, D. and K. Holt (2013). "Beginner's guide to comparative bacterial genome analysis using next-generation sequence data." *Microbial Informatics and Experimentation*, 3 (1): p. 2.

Eppinger, M., M. K. Mammel, J. E. Leclerc, J. Ravel, and T. A. Cebula (2011). "Genomic anatomy of *Escherichia coli* O157:H7 outbreaks." *Proceedings of the National Academy of Sciences of the United States of America*, 108 (50): pp. 20142–20147.

European Committee for Antimicrobial Susceptibility Testing (2016). *Breakpoint tables for interpretation of MICs and zone diameters. Version 6.0., 1-92.* URL: `http://www.eucast.org` (visited on 04/01/2015).

Evans, P. S., Y. Luo, T. Muruvanda, S. Ayers, B. Hiatt, M. Hoffman, S. Zhao, M. W. Allard, and E. W. Brown (2014). "Complete Genome Sequences of *Salmonella enterica* Serovar Heidelberg Strains Associated with a Multistate Food-Borne Illness Investigation." *Genome Announc*, 2 (3).

Eyre, D. W., T. Golubchik, N. C. Gordon, R. Bowden, P. Piazza, E. M. Batty, C. L. C. Ip, D. J. Wilson, X. Didelot, L. O'Connor, R. Lay, D. Buck, A. M. Kearns, A. Shaw, J. Paul, M. H. Wilcox, P. J. Donnelly, T. E. A. Peto, A. S. Walker, and D. W. Crook (2012). "A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance." *BMJ Open*, 2 (3).

Ferragina, P. and G. Manzini (2000). "Opportunistic data structures with applications." In: *Proceedings: 41st Annual Symposium on Foundations of Computer Science*, pp. 390–398.

Franz, E., P. Delaquis, S. Morabito, L. Beutin, K. Gobius, D. A. Rasko, J. Bono, N. French, J. Osek, B. Lindstedt, M. Muniesa, S. Manning, J. LeJeune, T. Callaway, S. Beatson, M. Eppinger, T. Dallman, K. J. Forbes, H. Aarts, D. L. Pearl, V. P. J. Gannon, C. R. Laing, and N. J. C. Strachan (2014). "Exploiting the explosion of information associated with whole genome sequencing to tackle Shiga toxin-producing *Escherichia coli* (STEC) in global food production systems." *International Journal of Food Microbiology*, 187: pp. 57–72.

Gardner, S. N., T. Slezak, and B. G. Hall (2015). "kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome." *Bioinformatics*, 31 (17): pp. 2877–2878.

Genome Trakr Network (2015). *Joining and Using the Genome Trakr Network.* URL: `http://www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS/ucm363134.htm` (visited on 04/01/2015).

Gibson, M. K., K. J. Forsberg, and G. Dantas (2015). "Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology." *ISME Journal*, 9 (1): pp. 207–216. ISSN: 1751-7362.

Gill, A., A. Martinez-Perez, S. McIlwham, and B. Blais (2012). "Development of a method for the detection of verotoxin-producing *Escherichia coli* in food." *Journal of Food Protection*, 75 (5): pp. 827–837.

Giufré, M., M. Accogli, C. Graziani, L. Busani, and M. Cerquetti (2015). "Whole-Genome Sequences of Multidrug-Resistant Escherichia coli Strains Sharing the Same Sequence Type (ST410) and Isolated from Human and Avian Sources in Italy." *Genome Announc*, 3 (4).

Gupta, S. K., B. R. Padmanabhan, S. M. Diene, R. Lopez-Rojas, M. Kempf, L. Landraud, and J.-M. Rolain (2014). "ARG-ANNOT, a New Bioinformatic Tool To Discover Antibiotic Resistance Genes in Bacterial Genomes." *Antimicrobial Agents and Chemotherapy*, 58 (1): pp. 212–220.

Ho, C.-C., A. K. L. Wu, C. W. S. Tse, K.-Y. Yuen, S. K. P. Lau, and P. C. Y. Woo (2012). "Automated Pangenomic Analysis in Target Selection for PCR Detection and Identification of Bacteria by Use of ssGeneFinder Webserver and Its Application to *Salmonella enterica* Serovar Typhi." *J. Clin. Microbiol.* 50 (6): pp. 1905–1911.

Ho, P.-L., R. C. Wong, S. W. Lo, K.-H. Chow, S. S. Wong, and T.-L. Que (2010). "Genetic identity of aminoglycoside-resistance genes in Escherichia coli isolates from human and animal sources." *J Med Microbiol*, 59 (Pt 6): pp. 702–707.

Huszczynski, G., M. Gauthier, S. Mohajer, A. Gill, and B. Blais (2013). "Method for the detection of priority Shiga toxin-producing *Escherichia coli* in beef trim." *Journal of Food Protection*, 76 (10): pp. 1689–1696.

Inouye, M., H. Dashnow, L.-A. Raven, M. Schultz, B. Pope, T. Tomita, J. Zobel, and K. Holt (2014). "SRST2: Rapid genomic surveillance for public health and hospital microbiology labs." *Genome Medicine*, 6 (11): p. 90. ISSN: 1756-994X.

Jahne, M. A., S. W. Rogers, I. P. Ramler, E. Holder, and G. Hayes (2014). "Hierarchal clustering yields insight into multidrug-resistant bacteria isolated from a cattle feedlot wastewater treatment system." *Environmental Monitoring and Assessment*, 187 (1), 4168. ISSN: 0167-6369.

Joensen, K. G., F. Scheutz, O. Lund, H. Hasman, R. S. Kaas, E. M. Nielsen, and F. M. Aarestrup (2014). "Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*." *Journal of Clinical Microbiology*, 52 (5): pp. 1501–1510.

Jolley, K. A., C. M. Bliss, J. S. Bennett, H. B. Bratcher, C. Brehony, F. M. Colles, H. Wimalarathna, O. B. Harrison, S. K. Sheppard, A. J. Cody, and M. C. J. Maiden

(2012). "Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain." *Microbiology*, 158 (4): pp. 1005–1015.

Kelley, D., M. Schatz, and S. Salzberg (2010). "Quake: quality-aware detection and correction of sequencing errors." *Genome Biology*, 11 (11): R116. ISSN: 1465-6906.

Kleinheinz, K. A., K. G. Joensen, and M. V. Larsen (2014). "Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and E. coli virulence genes in bacteriophage and prophage nucleotide sequences." *Bacteriophage*, 4 (2). PMID: 24575358: e27943.

Knowles, M., D. Lambert, G. Huszczynski, M. Gauthier, and B. W. Blais (2015). "PCR for the Specific Detection of an *Escherichia coli* O157:H7 Laboratory Control Strain." *Journal of Food Protection*, 78 (9): pp. 1738–1744.

Knuth, D. E. (2008). "The Art of Computer Programming." In: 1st ed. Vol. 4. Boston: Addison–Wesley Professional. Chap. Introduction to Combinatorial Algorithms and Boolean Functions, pp. 1–20. ISBN: 0321534964, 9780321534965.

Kor, S.-B., Q.-C. Choo, and C.-H. Chew (2013). "New integron gene arrays from multiresistant clinical isolates of members of the Enterobacteriaceae and Pseudomonas aeruginosa from hospitals in Malaysia." *J Med Microbiol*, 62 (Pt 3): pp. 412–420.

Kubasova, T., J. Matiasovicova, I. Rychlik, and H. Juricova (2014). "Complete sequence of multidrug resistance p9134 plasmid and its variants including natural recombinant with the virulence plasmid of Salmonella serovar Typhimurium." *Plasmid*, 76C: pp. 8–14.

Kumar, S., G. Stecher, and K. Tamura (2016). "MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets." *Molecular Biology and Evolution*.

Kuo, S.-C., C.-P. Fung, Y.-T. Lee, C.-P. Chen, and T.-L. Chen (2010). "Bacteremia due to Acinetobacter genomic species 10." *J Clin Microbiol*, 48 (2): pp. 586–590.

Labar, A. S., J. S. Millman, E. Ruebush, J. A. Opintan, R. A. Bishar, A. O. Aboderin, M. J. Newman, A. Lamikanra, and I. N. Okeke (2012). "Regional dissemination of a trimethoprim-resistance gene cassette via a successful transposable element." *PLoS One*, 7 (5): e38142.

Laing, C., C. Buchanan, E. Taboada, Y. Zhang, A. Kropinski, A. Villegas, J. Thomas, and V. Gannon (2010). "Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions." *BMC Bioinformatics*, 11 (1): p. 461.

Lambert, D., C. D. Carrillo, A. G. Koziol, P. Manninger, and B. W. Blais (2015). "GeneSippr: A Rapid Whole-Genome Approach for the Identification and Characterization

of Foodborne Pathogens such as Priority Shiga Toxigenic *Escherichia coli*." *PLoS ONE*, 10 (4): pp. 1–19.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." *Nature Methods*, 9 (4): pp. 357–359. ISSN: 1548-7091.

Lee, K. Y., J. D. Hopkins, and M. Syvanen (1990). "Direct involvement of IS26 in an antibiotic resistance operon." *J Bacteriol*, 172 (6): pp. 3229–3236.

Leekitcharoenphon, P., E. M. Nielsen, R. S. Kaas, O. Lund, and F. M. Aarestrup (2014). "Evaluation of whole genome sequencing for outbreak detection of *Salmonella enterica*." *PLoS One*, 9 (2): e87991.

Leggate, J. and B. W. Blais (2006). "An Internal Amplification Control System Based on Primer-Dimer Formation for PCR Product Detection by DNA Hybridization." *Journal of Food Protection*, 69 (9): pp. 2280–2284.

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." *Bioinformatics*, 25 (14): pp. 1754–1760.

Liu, B. and M. Pop (2009). "ARDB— Antibiotic Resistance Genes Database." *Nucleic Acids Research*, 37 (suppl 1): pp. D443–D447.

Liu, W., L. Chen, H. Li, H. Duan, Y. Zhang, X. Liang, X. Li, M. Zou, L. Xu, and P. M. Hawkey (2009). "Novel CTX-M beta-lactamase genotype distribution and spread into multiple species of Enterobacteriaceae in Changsha, Southern China." *J Antimicrob Chemother*, 63 (5): pp. 895–900.

Martinez-Perez, A. and B. W. Blais (2010). "Cloth-based hybridization array system for the identification of textitEscherichia coli O157:H7." *Food Control*, 21 (10): pp. 1354–1359. ISSN: 0956-7135.

Mathew, A. G., R. Cissell, and S. Liamthong (2007). "Antibiotic Resistance in Bacteria Associated with Food Animals: A United States Perspective of Livestock Production." *Foodborne Pathogens and Disease*, 4 (2): pp. 115–133. ISSN: 1535-3141.

Mazodier, P., P. Cossart, E. Giraud, and F. Gasser (1985). "Completion of the nucleotide sequence of the central region of Tn5 confirms the presence of three resistance genes." *Nucleic Acids Res*, 13 (1): pp. 195–205.

McArthur, A. G., N. Waglechner, F. Nizam, A. Yan, M. A. Azad, A. J. Baylay, K. Bhullar, M. J. Canova, G. De Pascale, L. Ejim, L. Kalan, A. M. King, K. Koteva, M. Morar, M. R. Mulvey, J. S. O'Brien, A. C. Pawlowski, L. J. V. Piddock, P. Spanogiannopoulos, A. D. Sutherland, I. Tang, P. L. Taylor, M. Thaker, W. Wang, M. Yan, T. Yu, and G. D. Wright (2013). "The Comprehensive Antibiotic Resistance Database." *Antimicrobial Agents and Chemotherapy*, 57 (7): pp. 3348–3357.

McDermott, P. F., S. Zhao, D. D. Wagner, S. Simjee, R. D. Walker, and D. G. White (2002). "The food safety perspective of antibiotic resistance." *Anim Biotechnol*, 13 (1): pp. 71–84.

Metzker, M. L. (2010). "Sequencing technologies – the next generation." *Nature Reviews. Genetics*, 11 (1): pp. 31–46. ISSN: 1471-0056.

Najibi, S., B. Bakhshi, S. Fallahzad, M. R. Pourshafie, M. Katouli, M. Sattari, M. Alebouyeh, and M. Tajbakhsh (2012). "Distribution of class 1 integrons among enteropathogenic Escherichia coli." *Can J Microbiol*, 58 (5): pp. 637–643.

Nawaz, M., K. Sung, O. Kweon, S. Khan, S. Nawaz, and R. Steele (2015). "Characterisation of novel mutations involved in quinolone resistance in *Escherichia coli* isolated from imported shrimp." *International Journal of Antimicrobial Agents*, 45 (5): pp. 471–476. ISSN: 0924-8579.

NCBI Resource Coordinators (2013). "Database resources of the National Center for Biotechnology Information." *Nucleic Acids Research*, 41 (D1): pp. D8–D20.

Oka, A., H. Sugisaki, and M. Takanami (1981). "Nucleotide sequence of the kanamycin resistance transposon Tn903." *J Mol Biol*, 147 (2): pp. 217–226.

Piddock, L. J. V. (1996). "Does the use of antimicrobial agents in veterinary medicine and animal husbandry select antibiotic-resistant bacteria that infect man and compromise antimicrobial chemotherapy?" *Journal of Antimicrobial Chemotherapy*, 38 (1): pp. 1–3.

Rasko, D. A., D. R. Webster, J. W. Sahl, A. Bashir, N. Boisen, F. Scheutz, E. E. Paxinos, R. Sebra, C.-S. Chin, D. Iliopoulos, A. Klammer, P. Peluso, L. Lee, A. O. Kislyuk, J. Bullard, A. Kasarskis, S. Wang, J. Eid, D. Rank, J. C. Redman, S. R. Steyert, J. Frimodt-Møller, C. Struve, A. M. Petersen, K. A. Krogfelt, J. P. Nataro, E. E. Schadt, and M. K. Waldor (2011). "Origins of the *E. coli* Strain Causing an Outbreak of Hemolytic–Uremic Syndrome in Germany." *New England Journal of Medicine*, 365 (8). PMID: 21793740: pp. 709–717.

Rather, P. N., P. A. Mann, R. Mierzwa, R. S. Hare, G. H. Miller, and K. J. Shaw (1993). "Analysis of the aac(3)-VIa gene encoding a novel 3-N-acetyltransferase." *Antimicrob Agents Chemother*, 37 (10): pp. 2074–2079.

Rodríguez, I., M. R. Rodicio, S. Herrera-León, A. Echeita, and M. C. Mendoza (2008). "Class 1 integrons in multidrug-resistant non-typhoidal Salmonella enterica isolated in Spain between 2002 and 2004." *Int J Antimicrob Agents*, 32 (2): pp. 158–164.

Rowe, W., K. S. Baker, D. Verner-Jeffreys, C. Baker-Austin, J. J. Ryan, D. Maskell, and G. Pearce (2015). "Search Engine for Antimicrobial Resistance: A Cloud Compatible

Pipeline and Web Interface for Rapidly Detecting Antimicrobial Resistance Genes Directly from Sequence Data." *PLoS ONE*, 10 (7): pp. 1–12.

Sandvang, D. (1999). "Novel streptomycin and spectinomycin resistance gene as a gene cassette within a class 1 integron isolated from Escherichia coli." *Antimicrob Agents Chemother*, 43 (12): pp. 3036–3038.

Schmid, D., F. Allerberger, S. Huhulescu, A. Pietzka, C. Amar, S. Kleta, R. Prager, K. Preußel, E. Aichinger, and A. Mellmann (2014). "Whole genome sequencing as a tool to investigate a cluster of seven cases of listeriosis in Austria and Germany, 2011-2013." *Clin Microbiol Infect*, 20 (5): pp. 431–436.

Sørum, H., T. M. L'Abée-Lund, A. Solberg, and A. Wold (2003). "Integron-containing IncU R plasmids pRAS1 and pAr-32 from the fish pathogen Aeromonas salmonicida." *Antimicrob Agents Chemother*, 47 (4): pp. 1285–1290.

Sundin, G. W. (2002). "Distinct recent lineages of the strA- strB streptomycin-resistance genes in clinical and environmental bacteria." *Curr Microbiol*, 45 (1): pp. 63–69.

Sundström, L., C. Jansson, K. Bremer, E. Heikkilä, B. Olsson-Liljequist, and O. Sköld (1995). "A new dhfrVIII trimethoprim-resistance gene, flanked by IS26, whose product is remote from other dihydrofolate reductases in parsimony analysis." *Gene*, 154 (1): pp. 7–14.

Sundström, L., P. Rådström, G. Swedberg, and O. Sköld (1988). "Site-specific recombination promotes linkage between trimethoprim- and sulfonamide resistance genes. Sequence characterization of dhfrV and sulI and a recombination active locus of Tn21." *Mol Gen Genet*, 213 (2-3): pp. 191–201.

Szczepanowski, R., F. Eikmeyer, J. Harfmann, J. Blom, L. M. Rogers, E. M. Top, and A. Schlüter (2011). "Sequencing and comparative analysis of IncP-1α antibiotic resistance plasmids reveal a highly conserved backbone and differences within accessory regions." *J Biotechnol*, 155 (1): pp. 95–103.

Tarr, P. I., C. A. Gordon, and W. L. Chandler (2005). "Shiga-toxin-producing *Escherichia coli* and haemolytic uraemic syndrome." *The Lancet*, 365 (9464): pp. 1073–1086. ISSN: 0140-6736.

Tettelin, H. et al. (2005). "Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome"." *Proc. Natl. Acad. Sci. U. S. A.* 102 (39): pp. 13950–13955.

Tomida, S., L. Nguyen, B.-H. Chiu, J. Liu, E. Sodergren, G. M. Weinstock, and H. Li (2013). "Pan-Genome and Comparative Genome Analyses of *Propionibacterium acnes*

Reveal Its Genomic Diversity in the Healthy and Diseased Human Skin Microbiome." *mBio*, 4 (3).

Trees, E., N. Strockbine, S. Changayil, S. Ranganathan, K. Zhao, R. Weil, D. MacCannell, A. Sabol, A. Schmidtke, H. Martin, D. Stripling, E. M. Ribot, and P. Gerner-Smidt (2014). "Genome Sequences of 228 Shiga Toxin-Producing *Escherichia coli* Isolates and 12 Isolates Representing Other Diarrheagenic *E.coli* Pathotypes." *Genome Announcements*, 2 (4).

Tyson, G. H., P. F. McDermott, C. Li, Y. Chen, D. A. Tadesse, S. Mukherjee, S. Bodeis-Jones, C. Kabera, S. A. Gaines, G. H. Loneragan, T. S. Edrington, M. Torrence, D. M. Harhay, and S. Zhao (2015). "WGS accurately predicts antimicrobial resistance in *Escherichia coli*." *Journal of Antimicrobial Chemotherapy*, 70 (10): pp. 2763–2769.

United States Department of Agriculture Food Safety Inspection Service Federal Register (2011). *Shiga Toxin-Producing Escherichia coli in Certain Raw Beef. Docket no. FSIS-2010-0023*. URL: `http://www.fsis.usda.gov/OPPDE/rdad/FRPubs/2010-0023.htm` (visited on 04/01/2015).

Wattam, A. R., D. Abraham, O. Dalay, T. L. Disz, T. Driscoll, J. L. Gabbard, J. J. Gillespie, R. Gough, D. Hix, R. Kenyon, D. Machi, C. Mao, E. K. Nordberg, R. Olson, R. Overbeek, G. D. Pusch, M. Shukla, J. Schulman, R. L. Stevens, D. E. Sullivan, V. Vonstein, A. Warren, R. Will, M. J. C. Wilson, H. S. Yoo, C. Zhang, Y. Zhang, and B. W. Sobral (2014). "PATRIC, the bacterial bioinformatics database and analysis resource." *Nucleic Acids Research*, 42 (Database issue): pp. D581–D591.

Wichmann, F., N. Udikovic-Kolic, S. Andrew, and J. Handelsman (2014). "Diverse Antibiotic Resistance Genes in Dairy Cow Manure." *mBio*, 5 (2).

Wiesner, M., M. B. Zaidi, E. Calva, M. Fernández-Mora, J. J. Calva, and C. Silva (2009). "Association of virulence plasmid and antibiotic resistance determinants with chromosomal multilocus genotypes in Mexican Salmonella enterica serovar Typhimurium strains." *BMC Microbiol*, 9: p. 131.

Wirth, T., D. Falush, R. Lan, F. Colles, P. Mensa, L. H. Wieler, H. Karch, P. R. Reeves, M. C. J. Maiden, H. Ochman, and M. Achtman (2006). "Sex and virulence in *Escherichia coli*: an evolutionary perspective." *Molecular Microbiology*, 60 (5): pp. 1136–1151.

Yang, Y., X. Jiang, B. Cai, L. Ma, B. Li, A. Zhang, J. R. Cole, J. M. Tiedje, and T. Zhang (2016). "ARGs-OAP: Online Analysis Pipeline for Anti-biotic Resistance Genes Detection from Meta-genomic Data Using an Integrated Structured ARG-database." *Bioinformatics*.

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

Zankari, E., H. Hasman, S. Cosentino, M. Vestergaard, S. Rasmussen, O. Lund, F. M. Aarestrup, and M. V. Larsen (2012). "Identification of acquired antimicrobial resistance genes." *Journal of Antimicrobial Chemotherapy*, 67 (11): pp. 2640–2644.

Zhao, Y., J. Wu, J. Yang, S. Sun, J. Xiao, and J. Yu (2012). "PGAP: pan-genomes analysis pipeline." *Bioinformatics*, 28 (3): pp. 416–418.

Zou, L.-K., H.-N. Wang, B. Zeng, A.-Y. Zhang, J.-N. Li, X.-T. Li, G.-B. Tian, K. Wei, Y.-S. Zhou, C.-W. Xu, and Z.-R. Yang (2011). "Phenotypic and genotypic characterization of β-lactam resistance in Klebsiella pneumoniae isolated from swine." *Vet Microbiol*, 149 (1-2): pp. 139–146.