DEVELOPMENT OF A QUANTITATIVE MODEL FOR COMPARING THE GENOMIC AND EPIDEMIOLOGICAL SIGNAL OF FOODBORNE PATHOGENS: IMPROVING THE APPLICATION OF WHOLE-GENOME SEQUENCING TO INFECTIOUS DISEASE EPIDEMIOLOGY

BENJAMIN M HETMAN
Bachelor of Science (Hon), University of Victoria, 2012

A Thesis
Submitted to the School of Graduate Studies
of the University of Lethbridge
in Partial Fulfillment of the
Requirements for the Degree

MASTER OF SCIENCE

Department of Biological Sciences
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

DEVELOPMENT OF A QUANTITATIVE MODEL FOR COMPARING THE GENOMIC AND EPIDEMIOLOGICAL SIGNAL OF FOODBORNE PATHOGENS: IMPROVING THE APPLICATION OF WHOLE-GENOME SEQUENCING TO INFECTIOUS DISEASE EPIDEMIOLOGY

BENJAMIN M HETMAN

Date of Defence: April 27, 2016

Dr. James E Thomas Co-Supervisor	Professor	Ph.D.
Dr. Eduardo N Taboada Co-Supervisor	Adjunct Professor	Ph.D.
Dr. Victor PJ Gannon Thesis Examination Committee Member	Adjunct Professor	Ph.D.
Dr. G Douglas Inglis Thesis Examination Committee Member	Adjunct Professor	Ph.D.
Dr. Anthony Russell Chair, Thesis Examination Committee	Assistant Professor	Ph.D.

Abstract

Interpreting microbial whole genome sequencing data remains an ongoing challenge in the fields of public health and epidemiology. For this thesis, 274 isolates of the human bacterial pathogen *Campylobacter jejuni* were selected for sequencing on the basis of their genotype and sampling metadata. A novel core genome typing method revealed that the genomic signal of bacterial isolates is not always concordant with their underlying epidemiology. To systematically examine this relationship, I developed an analytical model for quantifying the epidemiological similarity of bacterial isolates based on their sampling metadata, allowing for direct comparison to their genomic similarities. Applying this model to my dataset highlighted certain genotypes that were present throughout several diverse ecologies in disproportionately high amounts. A competitive recovery experiment revealed that particular genotypes seen in high prevalence in national and international repositories display preferential growth under laboratory conditions, providing evidence for systematic bias in infectious disease surveillance systems.

Acknowledgments

This body of work is the result of a collaborative effort, and I would be remiss if I didn't acknowledge the input of the many people responsible for its completion. First, I would like to sincerely thank my co-supervisors. Ed: not everyone can email their supervisor at 2:00am and have a response by 3:00am, and I am truly grateful for all your input to both my professional and personal development these past years. Jim, you've been there to shield me from the beauracracy of the educational system and facilitate my time as a graduate student, for that I am thankful.

To my committee members, thank you for your insights and support throughout my degree. Your enthusiasm and encouragement helped to keep the light at the end of the tunnel visible.

To my *dysfunctional* lab family, past and present, thank you for simply making every single day *better*. Steven and Cody, you've been there every single step of the way, and I would have been lost without you. Ruth, you're the most pleasant office mate I could ask for, thank you for putting up with me. Peter, your constant push for newer and better has always inspired me to develop my own abilities. Dor, I can always count on you for a brew and an adventure. Cassandra, you brought me in to all of this, so in a good way, I blame you.

Finally, to all my family and friends, thank you for your patience, your understanding and your unwavering support as I continue thoughout my academic pursuits.

Contents

List of Tables

List of Figures

List of Abbreviations

AFLP	Amplified Fragment Length Polymorphism (Analysis)
AR	Adjusted Rand Coefficient
AW	Adjusted Wallace Coefficient
BLAST	Basic Local Alignment Search Tool
CC	Clonal Complex
CGF	Comparative Genomic Fingerprinting
cgMLST	Core-Genome Multi-Locus Sequence Typing
FS	Fisher Syndrome
GAS	Group A *Streptococcus*
GBS	Guillain-Barré Syndrome
HC	Hemolytic Colitis
HUS	Hemolytic Uremic Syndrome
MALDI-TOF/MS	Matrix Assisted Laser Desorption Ionization Time of Flight Mass Spectroscopy
MCGH	Microarray Comparative Genomic Hybridization
MEE	Multi-Enzyme Electrophoresis
MLST	Multi-Locus Sequence Typing
MLVA	Multi-Locus Variable Number of Tandem Repeat Analysis
MST	Minimum Spanning Tree
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
PFGE	Pulsed-Field Gel Electrophoresis
RAPD	Random Amplification of Polymorphic DNA (Analysis)
RFLP	Restriction Fragment Length Polymorphism (Analysis)
SID	Simpson's Index of Diversity
ST	Sequence Type
SVR	Short Variable Region
VNTR	Variable Number of Tandem Repeat (Analysis)
WGS	Whole Genome Sequencing

Chapter 1

Review of Current Literature

1.1 A Brief History of the Epidemiology of Infectious Disease

In 1847, Ignaz Philipp Semmelweis (1818-1865), a newly appointed surgeon in the Vienna General Hospital, sought to explain the high proportion of deaths occurring in the hospitals two maternity wards from puerperal, or *childbed* fever. For years leading up to Semmelweis' appointment as assistant in obstetrics, new mothers in the Vienna General Hospital suffered from post-delivery mortality rates as high as 18%. Early in his posting, Semmelweis noticed that deliveries by midwives resulted in drastically lower maternal death rate (2%) than those by medical students and physicians (13-18%), who routinely performed autopsies on cadavers prior to attending the maternity ward (Best, 2004; Wyklicky & Skopec, 1983).

Today, puerperal fever is known to be a condition resulting from postpartum infection of the female reproductive organs after exposure to contaminated medical personnel or equipment during childbirth; however, in the mid 19th century, theories on bacteriology and infectious disease were only in their infancy. Semmelweis hypothesized that the handling of corpses prior to delivery in the maternity ward was causing infection, as the midwives, with their lower rates of infection, were not involved with any autopsies. Semmelweis tested his theory by implementing a mandatory chlorinated hand and equipment-washing routine and watched as rates of postpartum mortality soon dropped to less than 1% (Best, 2004; Wyklicky & Skopec, 1983).

Epidemiology is the study of the causation and distribution of disease, and seeks to explain why individuals develop diseases at different times and with different susceptibilities (Tannock et al., 2013). The example of childbed fever is one of the first recorded case

studies in the epidemiology of infectious disease; Semmelweis determined the etiology of childbed fever by investigating why patients under treatment from different sources had different rates of sepsis, and used these discrepancies to formulate a hypothesis on how to reduce the rates of postpartum mortality to almost zero. It would take years before Semmelweis would be credited with hand-washing as a means for preventing infection, but basic hygiene is now regarded as one of the most potent interventions for public health available today (Freeman et al., 2014).

The first studies in epidemiology are credited to the Greek physician Hippocrates (*approx.* 460-370 BCE), who, in his works on medical literature, made an attempt to relate human disease to environmental factors rather than attribute illness to an internal imbalance of personal humours; he coined the terms epidemic (*on the people*) and endemic (*in the people*) which are still in use today. A disease epidemic is one that resurfaces throughout various populations and lasts a finite amount of time; it is not persistent in any given population. An endemic disease is one that persists and is typically confined to a given population. Pandemics occur rarely, but result when an epidemic is no longer confined to a finite population, and spans multiple nations or even continents (Duncan, 1988).

1.2 Modern Epidemiology of Infectious Disease

A disease outbreak caused by an infectious agent (epidemic) has the ability to advance to endemic (or pandemic) status if its secondary infection rate, or its ability to spread, is greater than the rate at which the population either recovers, or becomes resistant to the infection (Friedman & Kao, 2014). The goal of an epidemiologist is to uncover the drivers behind the infection rate and implement policy or change to limit the rate so that the disease eventually depletes itself. Thus, the main tools of an epidemiologist exist in data collection and analysis. In studying the causation and frequency of disease, data relating to the factors resulting disease, or to the overall epidemiology of a disease agent (i.e. infectious bacterium) can be distilled into three categorical components: (1) *source*, (2) *space*, (3)

time.

The source of an infectious disease agent can refer to several different factors, depending on the type of disease being considered. In the epidemiology of cancer, for example, source may be difficult to define, as cancer may develop over many years, and be influenced by a multitude of environmental and internal factors. The epidemiology of agents of infectious disease, which this review will be limited to, can be considered to be more straightforward in its definition of source drivers. The source of an infectious disease agent can be considered as the sink or reservoir that plays host to the agent of infection, which, when encountered by a person of sufficient susceptibility, causes disease in that person. These sources can be either environmental reservoirs, such as specific locations within rivers, lakes, beaches, soil; or animal hosts, namely wild or farm animals, insects, birds, fish and even other humans. The classification of source can be further dissected into layers of increased granularity by source material type: this can refer to the difference between live animal sampling (e.g. rectal swabs, blood samples), post-processing sampling (e.g. meat from a retail setting, grocery store milk, cheese) or drop sampling (soil and faecal sampling not taken directly from the animal sources). In many cases, sampling non-directly (i.e. environmental water or drop sampling) can lead to samples of mixed origin, thus presenting a significant challenge to discerning the true source of the disease agent.

Data representing the temporal and geospatial aspects of epidemiology are generally much more straightforward than that of source, but again can be divided into many layers of granularity. Temporal data, for example, may be considered as the date of sampling or the day when a patient started showing symptoms or was exposed to the infectious agent. When comparing yearly trends, temporal data may be considered on a much broader scale, where only the monthly or seasonal sampling time would be measured. Geospatial records, much like temporal data, can be focused at an almost infinite number of levels of detail, often depending on the dataset being analyzed. When investigating an outbreak of infectious disease, for example, a fine level of granularity may be used to pinpoint exactly where

patients may be contracting the infectious disease agent. Geospatial figures are also derived, however, on national and provincial scales, to measure the overall rates and trends of infection; thus only general geospatial data may need to be recovered.

When all three categories of data for tracking an infectious disease agent can be put to use, epidemiologists are able to employ mathematical and statistical methods to try and fit the data into a logical model, allowing for the projection of disease both geographically, and temporally. These models allow for both elucidating the origin of disease, while at the same time providing information as to the progression of disease, enabling enhanced intervention and prevention measures.

1.3 Molecular Epidemiology I: Phenotypic Classification of Pathogens

Molecular epidemiology is the adaptation of techniques from molecular biology for use in enhancing our understanding of the pathogenesis and spread of infectious disease agents. When performed in tandem with traditional epidemiological investigations, molecular epidemiology can improve intervention and prevention strategies for reducing the occurence of diseases. In the epidemiology of infectious disease agents, the application of molecular epidemiology takes two broad forms: (1) pathogen identification, and (2) pathogen fingerprinting or strain typing (Foxman & Riley, 2001). Molecular techniques for each of these applications can be classified as either phenotypic, where the technique relies on externally expressed characteristics of the microorganism; or genotypic classification, where techniques are employed that involve direct analysis of genetic elements either chromosomally or extra-chromosomally (Maslow et al., 1993).

1.3.1 Culture Methods

Culture-based identification methods are among the oldest and simplest means of characterizing pathogens from samples (Fleming, 1942). Used primarily in clinical settings, differential and selective culturing methods provide a means of identifying bacterial pathogens that are relatively fast, inexpensive, and do not require considerable technical expertise or

complex equipment to perform. At best, however, culturing methods only allow for the positive or negative identification of a bacterial species in a sample; they often do not provide adequate information to differentiate between different subspecies or strains, and the possibility remains for competitive growth of unwanted bacterial species (Corry et al., 1995). Apart from these limitations, however, obtaining pure cultures remains an important step in most molecular epidemiological analyses even today; thus we can expect to continue to a reliance on culture-based methods for the isolation and characterization of pathogens (Lagier et al., 2015a,b).

1.3.2 Biotyping

Biotyping is often integrated with the practice of culturing a microorganism, and is based on the differential expression of metabolic processes. Biotyping methods typically assess four main characteristics of the organism in question: (1) colony morphology; (2) chemical susceptibility and resistance; (3) environmental tolerances; and (4) biochemical reactions (Eberle & Kiess, 2012). Various biotyping procedures have historically been used for identifying species of the *Campylobacter* microbiota, a food-borne pathogen responsible for up to 1% of gastrointestinal illnesses (Wheeler et al., 1999). Published in 1980, Skirrow and Benjamin released a biotyping scheme suitable for routine microbiology that differentiated between four species and two biotypes of thermophilic campylobacters. The biotyping scheme consisted of growth at 25°C and 43°C, susceptibility testing to nalidixic acid, hippurate-hydrolysis, and hydrogen sulphide production in an iron-containing medium (Skirrow & Benjamin, 1980). This scheme was later improved upon by Lior in 1984, whose updated procedure extended the range of the Skirrow-Benjamin scheme to allow for recognizing four separate biotypes of *Campylobacter jejuni*, two biotypes of *Campylobacter coli*, and two biotypes of *Campylobacter lari* (Lior, 1984). A more extensive biotyping approach was employed by On and Holmes (1995), who tested 73 phenotypic characteristics of *Campylobacter* species, of which 67 proved effective for

differentiating 347 strains of *Campylobacteria* into 31 phenotypically related clusters. The taxa created by this extended phenotyping method were assessed to be reasonably accurate in dividing the species and subspecies of bacteria and thus the method showed promise for epidemiological use (On & Holmes, 1995).

1.3.3 Serotyping

Serotyping is a method by which bacterial strains are differentiated based upon differences in carbohydrates and proteins expressed on their cellular surface. Suspect bacterial culture is subjected to antibodies known to react to specific subtypes of the microorganism; if a reaction is seen (i.e. the culture contains the antigen specific for the antibody), the unknown strain is recorded as being the variety, or serotype, that is reactive to the antibody being used. A new serotype identification is achieved when no known anti-sera react to the organism being tested (Centers for Disease Control, 2014; Singleton & Sainsbury, 2007). Serotyping, like culture based methods, is one of the most long-standing procedures for classification of bacterial species; for *Salmonella* spp., serotyping schema have been in use for most of the past century, with the first publication of the *Kauffmann-White* serotyping scheme, based on the antigenic formula of the somatic (O-) and flagellar (H-) antigens, appearing in 1934 (Lancefield, 1933; *Salmonella*-Subcommittee of the International Society for Microbiology, 1934).

Serotyping for *Salmonella* spp. is still considered the *de facto* standard for nomenclature and classification of the organism, and is routinely used for epidemiological investigations and outbreak analyses today. However, reagents required for serotyping specific strains of pathogens are costly to develop and produce, and serotyping does not give a comprehensive understanding of the pathogenicity of the typed organism; thus more practical methods of typing bacteria are likely to replace serotyping in the near future (Nesbitt & Ravel, 2012; Sheth et al., 2011).

1.3.4 Multilocus Enzyme Electrophoresis

Multilocus enzyme electrophoresis (MEE) is a technique for characterizing bacterial species based on the electrophoretic mobility of several intracellular enzymes essential for the survival of the cell. The differential mobilities of the enzymes are determined by substitutions at the amino acid level, which are in turn defined by allelic changes at the gene locus. The location of specific enzymes on agarose gel is visualized post-electrophoresis by the addition of substrate which, when catalyzed by its respective enzyme, displays a colour indicating the specific reaction taking place. The banding patterns for each sample being subjected to MEE can then be compared to produce a population network derived from the allelic differences at the loci of several housekeeping genes (Stanley & Wilson, 2003).

Originally used for studying the population genetics of *Drosophila* and humans, MEE became an important tool for assessing population dynamics underlying microbial pathogens during the 1980s, mainly due to the efforts of Selander et al. and their work on *Escherichia coli* and *Shigella* spp. (Ochman et al., 1983; Selander & Levin, 1980). The work of Selander et al. demonstrated the clonal nature of *E. coli* populations, with genetic clones appearing in unrelated hosts, and low rates of recombination apparent among the isolates studied.

MEE offered advantages over earlier phenotypic-based molecular classification methods in that all isolates characterized were inherently typeable, as the housekeeping enzymes assessed by MEE are required by the organism for survival. Additionally, MEE offered enhanced resolution over typical serotyping results; serotypes are usually derived from two to three loci, whereas in the case of MEE, it was possible to resolve typing results from several enzymes, reducing the chances of isolates from various sources appearing identical in the typing results (Ochman et al., 1983).

Application of MEE to population genetics has helped to inform us about the epidemiology of pathogenic *E. coli* O157:H7 infections. Work done by Whittam et al. demonstrated that members belonging to the O157 serogroup were highly diverse genetically via

MEE; this finding helped refute the argument that O157:H7 was a recent descendant of the O157 serogroup and prompted further investigation into the origin of the pathogen (Whittam et al., 1988). MEE was used in further work to show that isolates of the O157:H7 strain that caused hemolytic colitis (HC) or hemolytic uremic syndrome (HUS) in patients from geographically diverse regions were almost identical in their electrophoretic type, yet they were also genetically distinct from other serotypes of the *E. coli* species that also caused HC or HUS. This finding suggested that *E. coli* O157:H7 are a group of recent derivatives from a single clone spread throughout North America (Whittam & Wilson, 1988).

While MEE proved useful to studying the molecular epidemiology of bacterial isolates, several factors limit its use in contemporary studies. MEE, like most gel-electrophoresis applications, requires ample time and skill to create and run starch gels with consistent results, which are less discriminatory than many other methods now available for less technical investment. Further, MEE was only able to provide results on a small subset of the enzymes available in a bacterium, thus, with methods now available for assessing the entire genome of an organism, MEE is no longer typically used as a technique in molecular epidemiology (Maslow et al., 1993; Stanley & Wilson, 2003; Boerlin, 1997).

1.3.5 Phage-Typing

Bacterial phage-typing functions on the basis that bacteriophage will lyse only bacterial cells that contain the specific antigen which allows adherence of the phage to the bacterial cell; the differentiation of which phages a bacterial strain is susceptible to allows for the classification of bacterial species subtypes (Anderson & Williams, 1956). In an early study by Craigie and Yen in 1938, it was discovered that the *Salmonella typhi* type II *Vi* bacteriophage becomes highly lytic for the specific strain on which it was last propogated (Craigie & Yen, 1938). This feature was exploited for use in epidemiological studies, as it was shown that epidemiologically related strains could be verified as being of the same origin by testing if their phage type matched using the *Vi* II phage after adaptation to a known

Salmonella strain; this adaptive method allows for a much higher level of differentiation than originally estimated using *Vi* bacteriophage types I-IV (Anderson & Williams, 1956; Craigie & Yen, 1938).

Since the mid-20th century, phage-typing has been employed for studying the epidemiology of several agents of infectious disease including *E. coli*, *Mycobacterium tuberculosis*, *Pseudomonas*, *Campylobacter*, *Listeria*, and *Salmonella* species (Haq et al., 2012). Recently, however, the efficacy of phage-typing for epidemiologic investigations has been brought into question. Two separate incidences involving phage-typing for epidemiological surveillance resulted in outbreaks of *Salmonella typhimurium* going unidentified in western Europe between 2003 and 2008. In both cases, different interpretations of phage lysis results by separate, but equally experienced and qualified laboratories resulted in the outbreak strain of *S. typhimurium* not being identified until further confirmation could be achieved by genetic-based molecular epidemiology (Baggesen et al., 2010). These incidences underscore the requirement for epidemiologic methods that are both portable between laboratories and unambiguous in the interpretation of results; with new technologies for molecular epidemiology having entered the field in recent years fitting both of these requirements, phage-typing has gradually ceased to be of use for most investigations in molecular epidemiology.

1.3.6 MALDI-TOF Typing

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF/MS) has recently been suggested as a rapid, cost-effective means of identifying and typing pathogenic bacteria in a clinical laboratory based on the identification of expressed cellular and extra-cellular proteins, and using computer software to match the expressed protein spectra with their associated strain, species and genus (Bright et al., 2002; Clark et al., 2013). One advantage of the MALDI-TOF/MS application for bacterial typing is the speed at which results can be derived. While many typing techniques require lengthy proto-

cols for culturing and processing cellular lysates, it has been shown that MALDI-TOF/MS analysis can be performed directly on bacterial culture, decreasing the time required for the procedure to under ten minutes per sample (Krishnamurthy & Ross, 1996). Furthermore, MALDI-TOF/MS has been shown to differentiate between pathogenic and non-pathogenic strains of bacteria by identification of strain-specific protein biomarkers (Clark et al., 2013; Krishnamurthy et al., 1996).

Though promising in clinical laboratories for rapid identification of bacteria, MALDI-TOF/MS has not yet seen widespread use for molecular typing and epidemiological investigations (Murray, 2010). While the MALDI-TOF/MS procedure can be carried out relatively quickly and inexpensively on a per-sample basis (e.g. approximately 10 minutes and $0.20), the high principal cost of purchasing the instrument (>$100,000) remains prohibitive (Stevenson et al., 2010; Vranakis et al., 2012). Additionally, varying results based on sample preparation being used, as well as only being able to identify some bacterial isolates to the genus or species level further limits the use of MALDI-TOF/MS for molecular epidemiologic investigations (Clark et al., 2013; Seng et al., 2009).

1.4 Molecular Epidemiology II : Genotypic Classification of Pathogens

1.4.1 Restriction (Amplified) Fragment Length Polymorphism

Restriction fragment length polymorphism (RFLP) analysis involves the amplification and subsequent cleavage of a target gene with restriction endonucleases to produce DNA fragments with variable lengths. For *C. jejuni*, an assay targeting the *fla*A flagellin gene was developed by Nachamkin et al. and published in 1993. The *fla*A gene was specifically targeted as it was found to contain high levels of sequence variability; thus, when cleaved with a restriction endonuclease, strain specific banding patterns were found when the DNA fragments were resolved via gel electrophoresis (Nachamkin et al., 1993).

An extension of RFLP, Amplified Fragment Length Polymorphism (AFLP) involves the use of two restriction endonucleases, a frequent cutter and a rare cutter, which cleave

the genomic DNA of the organism. The genomic fragments are selectively primed, tagged with fluorescent adapters and amplified via PCR, typically producing between 50 and 100 genomic fragments. The banding pattern created by resolving the fragments on polyacrylamide gel provide a unique genomic fingerprint for each strain of the organism (Vos et al., 1995). A major advantage of AFLP analysis is the non-selective nature of the restriction enzymes being used, and that they target restriction sites across the whole-genome; this allows for both portability and universality of the method, as no specific knowledge about the genome is required beforehand (Sabat et al., 2013; Savelkoul et al., 1999). Due to the increased discriminatory power of AFLP over other genomic methods, analysis by AFLP has proven to be of great benefit to investigations of the epidemiology of *Legionella pneumophila*, *Bacillus anthracis*, *Salmonella enteritidis* and *S. enterica*, *C. jejuni* and *C. coli*, *Helicobacter pylori*, and *Streptococcus pyogenes*; however the high cost of DNA extraction kits, enzyme, fluorescent tagging systems, and the time required for analysis all limit the extent to which AFLP can be employed in a typical laboratory (Sabat et al., 2013; Savelkoul et al., 1999).

1.4.2 Analysis by Variable Number of Tandem Repeats

Analysis of the variable number of tandem repeats (VNTR) has been widely used for studies in molecular ecology. Subdivision of species is performed on the basis of multiple short, repeating DNA sequences, or micro-satellites; the number of tandem repeat elements denotes an allelic variation inheritable by downstream lineages of the organism (Lunt et al., 1998). An extension of this method, multi-locus VNTR analysis (MLVA), uses the frequencies of micro-satellites at multiple loci of the genome in order to establish an allelic fingerprint useful in the molecular epidemiology of infectious disease (Sabat et al., 2013). MLVA is not a universal method, however, as analysis relies on pathogen species-specific primers for amplifying the VNTR regions; loci are amplified via PCR and the number of tandem repeats are calculated from the size of the resulting amplicons. Also, random

genetic elements such as insertions, deletions, or duplications can affect the size of the amplicons; therefore for exact reproducibility, sequencing of the amplicons is required (Sabat et al., 2013).

1.4.3 Random Amplification of Polymorphic DNA

Molecular typing by random amplification of polymorphic DNA (RAPD) analysis involves the use of short, arbitrary primers to randomly target unspecified genome sequence. Carried out under low annealing temperatures, the primers are able to hybridize even when multiple mismatches between the primer and target sequence occur. When the distance between two priming sites falls between 0.1-3kb, the connecting sequence is amplified via PCR and the resulting amplicons can be visualized via gel electrophoresis to form strain-specific banding arrays (Sabat et al., 2013). Genetic relationships among isolates of *E. coli* inferred by RAPD are in high agreement with those established via MEE, however, RAPD has been shown to provide higher discriminatory power. Further, compared to MEE or pulsed-field gel electrophoresis (PFGE), RAPD analysis is less costly and more efficient for typing multiple isolates (Wang et al., 1993). Though simple and inexpensive, the low stringency annealing conditions used in RAPD analysis produces fluctuating results with only slight changes to reagents, protocols or machines; thus both intra- and inter-laboratory reproducibility of the method are low (Sabat et al., 2013).

1.4.4 Pulsed Field Gel Electrophoresis

Originally developed for genomic analysis of yeast, pulsed-field gel electrophoresis (PFGE) allows for the resolution of high molecular weight molecules via gel electrophoresis; genomic DNA is subjected to restriction enzyme digestion, then loaded onto an agarose gel where perpendicular electrical fields are applied in alternating pulses resulting in a net forward momentum of the DNA molecules (Schwartz & Cantor, 1984). Due to its ability to resolve entire genomes and provide high levels of discriminatory power, PFGE has been considered the gold standard in molecular subtyping for the surveillance and epidemiology

of bacterial foodborne diseases. In 1996, the US Centers for Disease Control and Prevention launched a surveillance project for bacterial foodborne diseases entitled PulseNet, which currently monitors the occurrences of high-risk foodborne pathogens in the United States and generates PFGE profiles for collection in a national database; a similar network exists in Canada, facilitated by the National Microbiology Laboratory and is linked with PulseNet USA to provide comprehensive international molecular surveillance (Gerner-Smidt et al., 2006; MacDonald et al., 2004; Swaminathan et al., 2001).

The surveillance network provided by PulseNet framework has proven invaluable to epidemiological outbreak investigations in the USA and Canada; with routine molecular surveillance being performed at a national level, outbreaks can be identified that link together infections in multiple states or provinces. PFGE subtyping via the PulseNet network was successfully used to identify outbreaks of *E. coli* O157:H7 in the United States in 1997, where multiple isolates from infected persons were linked to a single outbreak from contaminated alfalfa seeds (Breuer et al., 2001). In 1999, the PulseNet protocol for PFGE was applied to isolates of *E. coli* O157:H7 to uncover an outbreak of gastrointestinal illness linked to the consumption of salami in British Columbia, Canada (MacDonald et al., 2004).

The fact that PFGE has remained a *gold-standard* in molecular epidemiological investigations for over 20 years attests to its usefulness as a molecular typing technique; the typing data is portable, reproducible, and provides high discriminatory power from analysis of the entire genome (Goering, 2010). However, several limitations exist which prevent PFGE from being considered as a continued long-term solution for epidemiologic investigations. While PFGE provides good discriminatory power, there is evidence that it may not provide enough discrimination to discern between highly clonal bacterial isolates. In a study by Champion et al. in 2002, the investigators found that some highly clonal lineages of *C. jejuni* were indistinguishable between outbreak and non-outbreak strains via analysis by PFGE (Champion et al., 2002).

Genomic instability has also been shown to limit the usefulness of PFGE for epidemio-

logic investigations. Isolates of clonal origin which undergo natural genomic rearrangement have been shown to possess diverse PFGE fingerprints while retaining identical serotypes, biotypes and phage types; these differences can confound investigations as to the origin of epidemiologically important isolates (Wassenaar et al., 1998). Finally, PFGE requires a significant amount of hands-on time to perform and even longer to allow for the resolution of large molecules (e.g. $>$12h), and involves specialized equipment not readily available for all public health laboratories; thus the need to shift to newer, quicker methods for molecular typing is increasingly paramount (Goering, 2010; Taboada et al., 2013).

1.4.5 Single Gene Sequencing and Multilocus Sequence Typing

Sequence analysis of the variations present in a single gene has been shown to provide discrimination similar to or better than serotyping, and has largely replaced serotyping as the standard method for assessing the molecular epidemiology of group A *Streptococcus* (GAS) (Beall et al., 1995; Sabat et al., 2013). The *emm* gene encodes for the highly variable M protein in GAS; a major virulence factor traditionally used in serotyping the pathogen. In a study by Beall et al. (1995), it was demonstrated that by amplifying and subsequently sequencing the variable region of the *emm* gene, sequence types could be derived that were of higher discriminatory power than the serological M-types. Single-gene typing systems do not, however, provide a reliable estimate of overall genetic similarity, as variability may be present throughout the rest of the genome; thus, single gene sequencing needs to be used in tandem with other, more comprehensive typing schemes to help define clonal relatedness (Beall et al., 1995).

Multi-locus sequence typing (MLST) extends the practice of single gene sequencing to multiple loci throughout the genome. MLST for *Campylobacter* species assesses the allelic differences at short DNA regions (*approx.* 300-500bp) within seven housekeeping genes in the *Campylobacter* genome; an allelic match at 7/7 loci is defined as an identical sequence type (ST) while a match at 4/7 loci defines a clonal complex (CC) (Dingle et al.,

2001). MLST has been used extensively for typing *Campylobacter* species and a public database currently exists online (http://pubmlst.org/campylobacter) that comprises typing data on over 32000 isolates, with over 7500 unique STs identified. MLST is recognized as an important tool for assessing the population structure of *Campylobacter* species as well as providing strong evidence for epidemiologic linkages; it has been used to link predominant STs to specific sources (French et al., 2005; Manning et al., 2003; Sheppard et al., 2009a,b) allowing for targeted intervention strategies aimed at reducing the burden of gastrointestinal illness (Sears et al., 2011).

MLST provides better discriminatory power than phenotypic methods and PFGE, while remaining highly portable and reproducible, however, in certain circumstances it has failed to provide enough resolution to adequately differentiate between closely related bacterial strains (Clark et al., 2005; Hall et al., 2010; Taboada et al., 2013). Several attempts have been made to mitigate this lack of resolution; separate typing methods have been used in tandem with traditional MLST approaches (i.e. use of the *fla*A/B SVR sequencing with MLST for typing *Campylobacter* spp.) (Clark et al., 2005), or else extended MLST schema have been developed to exploit additional loci within the genome for enhanced discriminatory power (i.e. eMLST, rMLST) (Dingle et al., 2008; Jolley et al., 2012).

1.4.6 DNA Microarray Approaches

Microarray-based comparative genomic hybridization (MCGH) involves the use of oligonucleotide or open-reading-frame (ORF) based hybridizing microarrays for simultaneously visualizing the presence or absence of up to thousands of target nucleotide sequences in a particular bacterial strain in a single experiment (Taboada et al., 2013). Further, dual fluorescent labelling in MCGH allows for direct comparison of genomic profiles between a control and test strain, enabling investigations into the comparative genomics of bacterial strains with varying pathogenic potential (Leonard et al., 2004; Taboada et al., 2007), as well as the roles that host-specific and environmental stressors may play on a pathogen

(Lucchini et al., 2001). MCGH has been shown to be accurate for clustering of epidemiologically linked isolates, providing superior resolution and discrimination when compared to methods such as serotyping and RADP (Leonard et al., 2003), while remaining highly concordant to MLST-derived relationships (Taboada et al., 2008).

Though MCGH provides excellent resolution for comprehensive genome analysis of bacterial isolates, several challenges exist that make its use in routine epidemiologic typing unlikely. MCGH was developed at a time when prohibitively expensive whole-genome sequencing was limited to analysis of only a handful of priority strains; thus, the costs of microarray development and reagents were well justified (Taboada et al., 2013). Today, however, it is inexpensive to sequence the entire genome of a bacterial isolate (e.g. <$100 CAD), thus analysis by MCGH is significantly less practical now than it was even a decade ago. Further, reproducibility of MCGH analyses between laboratories is questionable, as no standards exist to establish thresholding of presence / absence of genes based on the hybridization ratios. Finally, elements such as gene insertions, point mutations, and genetic rearrangements are often not detectable by ORF-based microarrays, thereby limiting their faculty to detect novel pathogens (Garaizar et al., 2006).

1.4.7 The Bacterial Pan-Genome

The bacterial pan-genome for describing the quintessential aggregate of genes contained within a bacterial species was proposed in 2005, after whole genome sequence analysis of eight Group B *Streptococci* revealed that the total number of genes known to be present within the species grew with the addition of each sequenced genome (Tettelin et al., 2005). Genes common to all strains in the comparison were designated as “core”; it is assumed that these genes are required for base function of the organism, thus they are necessarily present in all strains of the same species. Genes that were found to either be unique to a single strain, or found in up to all but one strain of the species are considered to be accessory gene content; these genes were proposed to confer features related to host and environmental

specificity, pathogenicity and virulence factors (Medini et al., 2005). As more genomes of a bacterial species are sequenced, the pan-genome grows, uncovering more genes that were previously not known to belong in the species. The core and accessory genomes change in size inversely proportional to one another; with the increase in numbers of sequenced strains, the true core genome becomes more refined to those genes essential to survival; each gene identified as non-core becomes redefined as an accessory gene. Thus, as more bacterial genomes are sequenced, the core genome of a bacterial species approaches the refined asymptote of true core genes, while the accessory genome continues to expand.

1.4.8 Typing based on Accessory Genome Content

Polymerase chain reaction (PCR) is a simple, cheap and effective means of identifying the presence or absence of one or more gene targets in a DNA sample; the discovery that a thermophilic DNA polymerase from *Thermus aquaticus* could be used for automating PCR cycles drastically facilitated molecular biology experiments and enabled the shift from phenotypic typing methods towards genotypic typing methods in modern epidemiology (Saiki et al., 1988).

Comparative genomic fingerprinting (CGF) is a recently developed micro-array or PCR-based assay that exploits the hypervariability of accessory genome content. Analysis by CGF involves determining the presence or absence of specified genes in the accessory genome; the dichotomous profile created by assessment across multiple loci determines the resulting CGF fingerprint, which is easily comparable across multiple laboratories. Several CGF assays have been developed for typing bacterial pathogens, including *Campylobacter* spp., *E. coli*, and *Streptococcus pneumonia* (Dagerhamn et al., 2008; Laing et al., 2008; Taboada et al., 2012). Molecular typing of *Campylobacter* spp. using CGF has been shown to achieve higher discriminatory power than MLST, while retaining high concordance to the groupings created by MLST analysis. As well, it has been suggested that CGF may be more suited to short-term epidemiologic investigations than MLST, due to its assessment of the

accessory genome content which is thought to comprise more content related to pathogenicity and virulence as compared to the core-genome (Taboada et al., 2012). Another major advantage of CGF is the low cost associated with performing the assay. Compared to gene-sequencing, MEE, or PFGE, reagents and equipment required for PCR amplification are easily acquired and inexpensive; finally, PCR amplification requires little technical expertise and is rapid to perform, thus making CGF typing an appealing technique for routine surveillance of bacterial pathogens (Clark et al., 2012; Taboada et al., 2013).

1.5 *C. jejuni* as a Model Organism

1.5.1 Impact on Public Health and Economic Burden of Illness

The genus *Campylobacter* comprises 33 known species of small, motile, Gram-negative, microaerophilic bacilli. The thermophilic campylobacters (notably *C. jejuni* and *C. coli*) are the leading cause of bacterial gastroenteritis worldwide, with *C. jejuni* (approximately 90% of campylobacteriosis cases) and *C. coli* causing infection in approximately 1% of industrialized populations annually (Allos & Blaser, 1995; Humphrey et al., 2007; Wheeler et al., 1999). Infection by *Campylobacter* species often goes unnoticed or unreported; symptoms of campylobacteriosis are typically non life-threatening, with abdominal pain, fever and self-limiting diarrhea presenting in the vast majority of cases (Young et al., 2007). More serious complications can occur following infection with *Campylobacter* involving intestinal, neurological or rheumatological disorders (Humphrey et al., 2007; Nachamkin et al., 1998). Fisher Syndrome (FS) or Guillan-Barré Syndrome (GBS) occurs in approximately 1 out of every 1000 prolonged infections with *C. jejuni*, and results in symptoms ranging from muscle weakness, respiratory failure, and mild to acute paralysis (Nachamkin, 2002).

Among the genus *Campylobacter*, thermophilic campylobacters constitute the principal threat to human health. A primary reservoir for *C. jejuni* is considered to be the intestinal tract of poultry livestock - with consumption of raw or under-cooked chicken implicated in a

large proportion of cases of campylobacteriosis. The avian gut provides optimal conditions for growth of campylobacters; a humid, microaerobic environment kept at a temperature of 42°C. Livestock, such as cattle, sheep and pigs as well as non-agricultural sources including migratory birds, domesticated animals, soil, manure, and untreated water are all considered to be major players in the transmission dynamics of *C. jejuni* (Soloman & Hoover, 1999).

Illness due to infection with *Campylobacter* species is thought to be largely sporadic; the occurrence of campylobacteriosis due to outbreaks is typically less than 2% of all cases (Silva et al., 2011). Because of the sporadic nature of the pathogen, outbreak analysis remains impractical for capturing the epidemiology of *Campylobacter* infections, thus investigations rely on data generated from hospitals and clinical diagnostic laboratories. In Canada, infection with *C. jejuni* accounts for approximately 145,000 annual cases of foodborne gastrointestinal illness, ranking *C. jejuni* first for infectious bacterial pathogens based on Canadian surveillance data (Thomas et al., 2013). At an estimated economic cost of almost $1100 CAD per case, gastroenteritis from infection with *Campylobacter* spp. represents an annual economic burden of over $150 million to the Canadian public (Majowicz et al., 2006). Epidemiologic investigations aimed at intervening in the transmission of *C. jejuni* therefore have the potential to alleviate significant costs to the Canadian economy, as well as decrease the overall rates of illness in the Canadian public. Thus, from both fiscal and health viewpoints, elucidating the epidemiology of *C. jejuni* in terms of transmission dynamics, population structure, and identification of principal reservoirs remains an important avenue for study.

1.5.2 Upward Trend of Overall *Campylobacter* Infections

The frequency of campylobacteriosis cases has increased over time since reporting of *Campylobacter* infections began in the 1980s (The European Food Safety Authority, 2015; Nichols et al., 2012). Two factors are possible for the rise in cases; changes in the structure of reporting the illness or a rise in the number and severity of risks involved in contracting

Campylobacter infection. Better identification and isolation procedures over the past 20 years could explain the overall rise in campylobacteriosis seen throughout the population; however, a similar long-term rise in cases has been seen as a systematic change across all diagnostic laboratories (Nichols et al., 2012). If disease rates were dependent solely on the *Campylobacter* isolation efficiency, then more resourced labs should be the ones to show increased rates, instead of an increased frequency across the general spectrum of reporting centres. Thus, it is reasonable to assume that the increase in cases of *Campylobacter* infection is largely due an increase in risk factors associated with the epidemiology of the pathogen.

1.5.3 Risk Factors for Campylobacteriosis

Age appears to be significantly correlated with incidences of campylobacteriosis; infections occur most in very young children (< 1 year of age) and between the ages of 20 to 50 years of age. There is a significant decline in cases for people over 60, and for those between 6 and 20 years of age. In all age categories, men report with more cases than women, with an approximate 15% increase in male : female ratios of cases (Friedman et al., 2004; Nichols et al., 2012). The reasons for the gender discrepancies as well as the two peaks in campylobacteriosis with regards to age are largely unknown, however, handling and consumption of poultry at a commercial establishment remains one of the highest risk factors for infection with *Campylobacter*; thus it is possible that young adult males consume more meals at restaurants and fast food establishments than do their female counterparts (Friedman et al., 2004).

Interestingly, in recent years there has been an increase in cases observed for people over 50 years of age. It is thought that three drivers may explain this rise among the older demographic. First, there has been an shift in overall population demographics: as a higher proportion of the population becomes over the age of 50, more cases of campylobacteriosis are logically seen in this demographic category (Nichols et al., 2012). Secondly, the

increased use of a class of drugs known as proton pump inhibitors that are used to combat acid-reflux in the stomach have been shown to have a positive correlation with infection by *Campylobacter* species (Tam et al., 2009). Finally, travel abroad is associated with higher rates of campylobacteriosis; as people tend to travel more in their retirement, they have a higher risk of contracting campylobacteriosis from foods prepared under questionable circumstances (Nichols et al., 2012; Neimann et al., 2003).

Seasonality trends have been observed in case frequencies of campylobacteriosis, particularly among young children. Campylobacteriosis cases rise in the spring and summer, and decline during the winter months, with consistent major peaks in June/July and a secondary, small peak shortly after the New Year (The European Food Safety Authority, 2015; Louis et al., 2005; Nichols et al., 2012; Deckert et al., 2014). The small peak seen in early January may involve increased consumption of poultry during the holiday season, while the drastic increase seen in the summer during warmer temperatures is correlated with several risk factors for infection. Increased cases in summer may be explained by changes in human behaviour during warmer periods of the year, changes in the prevalence of campylobacters in agricultural or environmental reservoirs, or a combination of both (Nylen et al., 2002). Consumption of barbequed foods, as well as consuming untreated water are both identified as risk factors for campylobacteriosis; thus it is reasonable to assume that recreational water use as well as picnic season may help to explain at least part of the summer increase in bacterial enteritis. Discrepancies in the rural and urban rates of infection during the summer, however, suggest that occupational hazards are also a major risk factor. Rural areas experience more seasonality with respect to rates of campylobacteriosis; warmer temperatures correspond to higher rates of infection suggesting a logical link to the increase in agricultural activities in the spring and summer months (Louis et al., 2005; Nichols et al., 2012).

Poultry is considered to be the single highest risk factor for the transmission of *C. jejuni* to humans; exposure to poultry raising (chicken farms, broiler houses), processing (slaugh-

terhouse, meat packing plant), and consumption (fast food, restaurant) all correspond to significantly higher rates of contracting campylobacteriosis (Tam et al., 2009; Friedman et al., 2004; Müllner et al., 2009). The contamination of broiler flocks with *C. jejuni* remains enigmatic; chicken flocks have been shown to be free of *Campylobacter* contamination at the hatchery and early rearing stages; however, once infection is introduced, contamination of a flock occurs within a matter of days (Herman et al., 2003; Shanker et al., 1990). Vertical transmission (i.e. transfer from chicken to egg to chick) is thought to occur rarely, if at all; thus horizontal transmission of *Campylobacter* to the flock is assumed to occur during rearing and slaughter, implicating environmental or animal vectors as sources of contamination (Callicott et al., 2006). Several external factors are purported to be responsible for contamination of chicken flocks: soil/water contamination, airborne transmission from nearby manure stacks, domestic animals, and contaminated flies or other insects may all contribute to the transmission of *Campylobacter* to flocks (Agunos et al., 2014; Nichols, 2005).

1.6 Molecular Epidemiology and Surveillance of *C. jejuni*

Traditional epidemiologic approaches have met with some success in elucidating the etiology and transmission dynamics of *C. jejuni*; however, the sporadic nature, combined with the high number of potential reservoirs makes it nearly impossible to pinpoint the sources of human infection with *C. jejuni* without confirmation at the molecular level. The application of molecular methodologies for the study of *Campylobacter* epidemiology has improved our understanding of the transmission pathways of the pathogen (Callicott et al., 2006; Kwan et al., 2008a), identified significant sources of human exposure useful in targeted public health interventions (Wilson et al., 2008; French & the Molecular Epidemiology and Veterinary Public Health Group, 2008; Sheppard et al., 2009b), and uncovered evolutionary mechanisms helpful in understanding the pathogenesis of *Campylobacter* species (Sheppard et al., 2008, 2013; Young et al., 2007).

To provide effective source attribution for *Campylobacter* species, several countries have developed routine sentinel surveillance initiatives to monitor *Campylobacter* prevalence from non-clinical settings. These sentinel surveillance programs employ molecular typing analysis that allows for the attribution of clinical isolates to potential sources from environmental, commercial and retail reservoirs. Due to the sporadic nature of *C. jejuni* infections and the organisms ability to survive in many animal and environmental sources, the results derived from routine molecular surveillance networks have proven to be essential for understanding the role of *Campylobacter* in public health (Sears et al., 2011; Müllner et al., 2009; French & the Molecular Epidemiology and Veterinary Public Health Group, 2008).

In Canada, a national surveillance network on transmissible enteric pathogens was implemented in 2005; FoodNet Canada (formerly C-Enternet) currently surveys three sentinel sites across the country. The prototype site was established in Waterloo, Ontario in 2005, followed by the establishment of a second site in Chilliwack, BC in 2010, and finally, a third site was founded in Calgary, AB in 2014 (Public Health Agency of Canada, 2006). These sentinel sites represent collaborations between local health units and provincial health laboratories and are facilitated by the Public Health Agency of Canada. Sample collection is routinely done at water, farm, and retail sources, while human samples are collected via passive surveillance systems at clinical laboratories. *Campylobacter* isolates collected by FoodNet Canada are subjected to molecular typing by CGF (Public Health Agency of Canada, 2010).

Results from molecular epidemiology performed on Canadian *Campylobacter* isolates have shown that poultry remains an important source of campylobacteriosis in humans, but the relative significance is likely dependent on demographic setting. Molecular typing data has demonstrated that while the majority of cases from urban settings share a similar CGF profile to poultry-derived isolates (i.e.$>$90% similarity), the genotypes of cases from rural settings are significantly less likely to associate with chicken-based clusters (Deckert et al.,

2014). These results suggest that rural inhabitants, especially children, are at higher risk for contracting campylobacteriosis from sources other than chicken, including direct exposure to cattle, contact with untreated water, or consumption of unpasteurized milk.

Molecular typing data from FoodNet Canada has also helped to implicate cattle as a significant factor in *Campylobacter* epidemiology, as well as provide evidence of widespread lineages of *Campylobacter*, which reside in an extensive array of animal and environmental reservoirs. Highly prevalent clusters of *Campylobacter* genotypes isolated from patients in areas of high agricultural production show strong associations to isolates derived from cattle sources, though isolation of campylobacters from retail beef samples is rare (Mutschall et al., 2013; Deckert et al., 2014). This finding corroborates the hypothesis that occupational hazards, where direct contact with livestock (including cattle) occurs, are likely a significant risk factor for contracting and spreading *Campylobacter* (French & the Molecular Epidemiology and Veterinary Public Health Group, 2008; Kapperud et al., 2003).

An Example of Molecular Surveillance of *C. jejuni*: New Zealand

Following its induction as a reportable illness in 1980, cases of campylobacteriosis in New Zealand rose steadily into the mid 2000s, with peak rates higher than 380 cases per 100,000 people; several investigations implicated poultry as a major risk factor for exposure to *C. jejuni* and an increase in poultry consumption in New Zealand during this time period supported these implications (Sears et al., 2011; Baker et al., 2006). To investigate the role of poultry in the epidemiology of *Campylobacter* infections in New Zealand and identify other important sources of exposure, a sentinel surveillance program was developed to work in collaboration with public health units across the country. In studies combining traditional epidemiology with molecular epidemiologic approaches, MLST was used to compare clinical data collected at public health units throughout the country with data from domestic animal, food, and environmental sources. Poultry was implicated as the primary contributor to human campylobacteriosis, followed by cattle, sheep and environmental sources

(French & the Molecular Epidemiology and Veterinary Public Health Group, 2008; Sears et al., 2011).

A three-year study using MLST data indicated that between 58-76% cases of human campylobacteriosis in New Zealand could be attributed to poultry sources (Müllner et al., 2009). New Zealand possesses a uniquely simple poultry supply, in that no raw poultry products are imported into the country, and domestic poultry production focuses almost exclusively on the local market (Baker et al., 2006; Müllner et al., 2009). This simple model of poultry production allowed for suppliers of poultry meat to be treated as individual sources, as opposed to grouping chicken as an individual source, and results from molecular subtyping were able to identify a single poultry supplier as contributing to over 60% of the annual cases of human *Campylobacter* infection (Müllner et al., 2009). Results from this and other studies also identified a rare subtype of human pathogenic *C. jejuni*, ST-474, which has appeared only sporadically in other countries, but appears widespread in New Zealand. The high prevalence of this subtype is thought to be a major contributing factor to the excessively high rates of campylobacteriosis in New Zealand (Müllner et al., 2009; French & the Molecular Epidemiology and Veterinary Public Health Group, 2008).

Both traditional and molecular epidemiologic investigations were leveraged to implement nation-wide interventions on the spread of *C. jejuni* from chicken meat to human beings. Interventions aimed at the food safety and poultry industry were implemented in New Zealand in 2006, and included both regulatory and voluntary procedures for the reduction of *C. jejuni*. Performance targets, rather than specific intervention strategies, were made on industry, and included improvements to both hygiene, and alterations to the process of chilling poultry meat prior to sale. Following the introduction of these intervention targets, a rapid decline was seen in the rates of campylobacteriosis in New Zealand; a 54% decrease representing a fall from the yearly average of over 350 cases per 100,000 to 161 cases per 100,000 occurred in only two years. The fiscal impact of such a decline in cases is estimated to have saved approximately $70 million (NZD) to the New Zealand economy

(Sears et al., 2011; Muellner et al., 2011).

The application of molecular subtyping to isolates of *C. jejuni* derived from human, domestic, and environmental sources helped to identify the relative importance of poultry consumption to campylobacteriosis in New Zealand, allowing for targeted approaches at reducing its spread. However, decline in the rates of gastroenteritis due to *Campylobacter* infection have since stalled, and the prevalence of campylobacteriosis continue to be much higher in New Zealand than the global average (Sears et al., 2011; Coker et al., 2002). While poultry is an important contributor to *Campylobacter* epidemiology, other routes of exposure remain to be implicated and only through continued routine surveillance by both public health units and sentinel laboratories will additional sources of *Campylobacter* infection be identified.

1.6.1 Challenges to Molecular Epidemiology of *C. jejuni*

Several genomic features make *C. jejuni* particularly complex for assessment via techniques used for molecular epidemiology. Many campylobacters are highly recombinogenic; they have been shown to readily exchange DNA with other bacteria both inter and intraspecially, and recombination is suggested to be the primary mechanism for evolutionary change of the organism (Wilson et al., 2009). Genome plasticity is a major challenge to most subtyping methods, as the underlying principle of molecular subtyping is that the generated fingerprint is an accurate proxy for the whole genome. If sections of the bacterial chromosome are frequently changing, then the assumption that two genomes are the same based upon typing a minor fraction of each genome is heavily flawed. Thus, subtyping approaches need to either (a) exploit the recombinogenic features of the *Campylobacter* genome to provide strong evidence for epidemiologic linkages; or (b) provide high enough resolution that instances of recombination are captured by the method and leveraged in the typing result.

MLST for the molecular subtyping of *C. jejuni* has been employed widely in re-

cent years by several investigations aimed at the population genetics and epidemiology of campylobacters circulating in the environment. The assay compares *Campylobacter* strains based on nucleotide sequence located within seven core genes in the *Campylobacter* genome: *aspA*, *glnA*, *gltA*, *glyA*, *pgm*, *tkt*, and *uncA* (Dingle et al., 2001). The strength in analysis by MLST is the reliance on allelic profiling, rather than single nucleotide analysis. High levels of recombination in the *Campylobacter* genome can obfuscate genomic similarity results derived by nucleotide-sequence based approaches; via MLST, any change to the nucleotide sequence of a gene is considered a single allelic difference thus rendering all recombination events and mutational point changes equivalent (Maiden et al., 1998; Sheppard et al., 2012).

Recombination is considered to be the principal driver behind evolution of the *Campylobacter* genome (Wilson et al., 2009). Using MLST, investigations have uncovered significant evidence for inter- and intra-species gene flow in the genomes of *C. jejuni* and *C. coli*, with several MLST sequence types identified as a hybrid allelic arrangement between the two species (Sheppard et al., 2008; Wilson et al., 2009). The high level of gene flow throughout the *Campylobacter* genome diminishes epidemiologic linkages gleaned from analysis with MLST; assessment at seven loci, which corresponds to only 0.5% of the *Campylobacter* genome, may not provide enough resolution to accurately discriminate between closely related isolates in short term epidemiological studies (Sails et al., 2003; Taboada et al., 2008). Whole genome assessment using MCGH has provided evidence that (a) MLST may not consistently capture overall genome similarity; and (b) significant genomic heterogeneity may exist among strains otherwise similar at the sequence-type level (Taboada et al., 2008). To produce higher resolution results, efforts have been made to develop MLST typing schemes based on sequencing a greater number of targets within the core genome of *C. jejuni*. Analysis by extended MLST (eMLST) and ribosomal MLST (rMLST) consists of sequencing 10 and 53 genes, respectively (Dingle et al., 2008; Jolley et al., 2012). While these methods provide higher resolution to studies comparing *Campy-*

lobacter genotypes, sequencing individual genes remains costly and time consuming, especially when considering the volume of isolates generated for routine surveillance purposes.

Analysis by MLST measures variation in the core genome of *C. jejuni* yet it has been proposed that features that may confer higher epidemiological relevance, such as host adaptation and survival under environmental stressors are imparted by accessory genome content. The accessory genome has been shown to contain regions of hypervariability; several of these regions are responsible for structures conferring pathogenicity such as the lipooligosaccharide layer, flagella, and surface polysaccharides (Parkhill et al., 2000). The ability to rapidly evolve these structures is considered to be a survival mechanism for bacterial pathogens, as it allows for the adaptation to hostile environments and host niches (Young et al., 2007; Bolton, 2015). Thus, while analysis of a small fraction of the core genome of *C. jejuni* is suitable for studies into evolutionary population genetics, short-term epidemiology may be better analyzed by methods capturing the variation in accessory genome content (Maiden et al., 1998).

1.7 A New Era of *Genomic Epidemiology*

Whole genome sequencing (WGS) for the analysis of bacterial pathogens has recently come to light as a highly informative, cost-effective method for providing the highest level of resolution for characterizing bacterial strains. Sequencing the whole genome provides the ultimate genetic map that can be used to compare bacterial strains using both their core and accessory gene content. Further, due to recent advances in next-generation sequencing technology, analysis of bacterial WGS has become less costly than most other molecular typing methods, including MLST, while still providing the information required for comparison with legacy datasets (Wetterstrand, 2015; Köser et al., 2012).

WGS analysis in pathogen outbreaks has already proven to be extraordinarily useful in recent years. An outbreak of tuberculosis in British Columbia, Canada was observed between May 2006 and December 2008. Using molecular typing methods including RFLP

and VNTR analysis, the outbreak was believed to be of a single, clonal origin; both RFLP and VNTR analyses identified only a single genotype of *Mycobacterium tuberculosis* in all cases. In a 2011 retrospective study by Gardy et al., outbreak isolates were subjected to analysis using a whole-genome SNP-based approach for measuring genetic similarity at increased resolution. In this approach combining WGS data with social network analysis, it was discovered that the rise in cases during the two-year period actually consisted of two separate, simultaneous outbreaks of *M. tuberculosis*, corresponding to a concurrent rise in the use of high-risk drugs (Gardy et al., 2011). The increased level of data granularity achieved by WGS analysis allowed for the resolution of these two outbreaks, in turn greatly improving the interpretation of the spread and causation of tuberculosis during this time.

WGS-based methods have also been applied to outbreaks involving foodborne bacterial pathogens. In 2011, an outbreak involving pathogenic *E. coli* was identified in Germany. Typical methods of molecular typing such as serotyping and PFGE identified the outbreak strain as subtype O104:H4, a strain seen previously, but rarely identified to cause extreme disease (Rasko & Webster, 2011). The O104:H4 strain implicated in the 2011 outbreak, however, displayed some of the highest rates of HUS ever seen, combined with abnormally high mortality rates. Rapid, real-time WGS analysis was performed alongside traditional molecular typing techniques; while other methods could not identify what caused the particularly high virulence demonstrated by the outbreak strain, WGS analysis provided evidence that the German O104:H4 was an enteroaggregative strain of *E. coli* that had recently acquired the potential to produce deadly Shiga-toxin. This strain therefore had the combined abilities to adhere to and colonize the gut at advanced levels, while producing Shiga-toxin in a highly localized fashion, resulting in the increased rates of HUS observed in the 2011 outbreak (Rasko & Webster, 2011; Grad et al., 2012; Rohde et al., 2011).

Recent reviews have suggested WGS for routine analysis in a public health setting, citing its value for both epidemiologic investigations and evolutionary population genetics (Didelot et al., 2012; Köser et al., 2012; Sabat et al., 2013; Taboada et al., 2013). While rou-

tine sequencing may have been prohibitively expensive as recently as the early 2000s, today, preparing and sequencing a bacterial genome can be accomplished for approximately $100 CAD per isolate, which rivals the cost of gold-standard typing methods such as MLST or PFGE (Wetterstrand, 2015). As discussed earlier in this review, routine pathogen surveillance has already shown to be of high value to epidemiologic and public health efforts, reducing the burden of illness from both fiscal and health standpoints. By replacing current molecular typing methods used in surveillance initiatives with WGS, we could effectively extend possible analyses to any level of DNA-based investigations; this could include high-resolution evolutionary studies, phenotype to genotype analyses, and studies concerning the carriage of mobile genetic elements, to name a few.

A significant challenge in performing WGS for routine surveillance of bacterial pathogens currently lies in the interpretation and management of the sheer volume of data generated; the *C. jejuni* isolate NCTC11168, sequenced in 2000, was found to contain 1,641,481 base pairs of DNA (Parkhill et al., 2000). Routine sequencing for surveillance of *C. jejuni* in Canada has the potential to generate hundreds to thousands of genome sequences each year, thus, bioinformatics tools are required for handling data of this magnitude, but also for calibrating the data for practical epidemiologic purposes. To this end, Kruczkiewicz et al. developed the Microbial *in-silico* typing tool (MIST), which leverages the Basic Local Alignment Search Tool (BLAST) to perform nucleotide sequence searches based on user-defined criteria (Camacho et al., 2009). Using this bioinformatic approach, results from traditional molecular typing assays, including MLST, VNTR, MLVA, and PCR-based assays can be approximated directly from the sequence data (Kruczkiewicz et al., 2013).

1.8 Objectives of the MSc. Thesis

The field of genomics has only recently advanced to the point where it is possible to use the complete genome sequence of many bacterial isolates to perform an epidemiologic

investigation; thus, nascent approaches to analyzing genomic data in an epidemiologic context are limited. In this thesis, I propose the use of *C. jejuni* as a model organism for investigating the potential benefits of using WGS for performing an epidemiologic investigation on a bacterial pathogen of high importance to public health. As discussed earlier in this review, *Campylobacter* species represent a significant burden to the public health of Canadians, as well as the economy. Infections from *Campylobacter* species are highly sporadic, and the *Campylobacter* genome is highly plastic, allowing the bacterium to successfully adapt and evolve in order to survive in a wide range of host environments. These features make elucidating the epidemiology of *C. jejuni* particularly difficult using molecular epidemiology techniques, thus the enhanced resolution and discriminatory power of WGS is perfectly suited to help assess the circulation of *C. jejuni* in Canada.

My first objective of this thesis was to perform WGS on a collection of Canadian *Campylobacter* isolates for use in an epidemiological study to determine the feasibility of elucidating the epidemiology of an important public health pathogen from genomic sequence data. Isolates were collected from a variety of clinical, environmental, and animal sources through the FoodNet Canada sentinel surveillance program and sampling initiatives local to southern Alberta. Information regarding the isolation date, location and source of sampling, the complete CGF profile and any ancillary molecular typing data (serotype, MLST profile) are stored in the Canadian CGF database comprising over 20,000 isolates. From this collection, I selected 298 isolates of *C. jejuni* according to their epidemiological relationships, as well as their genotypic relatedness from CGF profiling. In order to test the hypothesis that genomic relationships uncovered using WGS would better reflect the epidemiology of *C. jejuni* isolates compared to molecular typing methods, I subjected the sample of isolates to WGS and performed *in-silico* core-genome MLST to allow for the visualization of genomic relationships with their underlying epidemiology.

An inherent challenge in comparing the genomic relatedness between bacterial isolates with their epidemiologic relatedness, is the translation between qualitative and categorical

data types, respectively. My second thesis objective was to develop novel analytics for assessment of the relationship between the epidemiology and genomics of bacterial isolates. For Chapter 3, I hypothesized that by using basic epidemiological metadata (source, location, date) collected during the sampling of bacteria, I could develop a quantitative summary statistic encapsulating the epidemiologic relatedness between *C. jejuni* isolates. I further hypothesized that this summary statistic could then be compared directly to the genomic relatedness between isolates of *C. jejuni* to guide the interpretation of WGS data for informative and practical public health purposes.

An ongoing challenge in surveillance efforts is ensuring that isolates obtained through sampling of the many reservoirs of *Campylobacter* are representative of the population in circulation, which allows for assessment of the public health significance of the various subtypes in circulation. From the results of the epidemiological modelling in Chapter 3, genotypes of *C. jejuni* were identified that were genomically homogeneous, but corresponded to a wide range of reservoirs from very distinct ecologies. These results prompted an investigation of these genotypes in the Canadian CGF database, where it was found that a small collection of genotypes are represented in disproportionately high frequencies. Whether the sampling frequency of these isolates was elevated (a) due to their true frequency in the environment; or (b) because of an advantageous ability to out-compete other, less prominent genotypes under typical laboratory isolation conditions was unknown. Thus, to test the hypothesis that laboratory isolation conditions were biasing the recovery of isolates towards those from only a small subset of *Campylobacter* genotypes, I performed a controlled competitive growth experiment comparing the relative efficiencies of strains which appeared in low-prevalence in the CGF database with those that appeared in high-prevalence. The results from the competitive recovery experiments are presented in Chapter 4.

Chapter 2

Investigating the Genomic Epidemiology of Canadian *C. jejuni*: a pilot study

2.1 Preamble

Whole genome sequencing (WGS) of a bacterial organism is considered to provide the ultimate genetic map, containing all of the heritable traits that identify a bacterial isolate as both a member of a taxonomic species, as well as differentiating it as a uniquely identifiable strain (Hardison, 2003). With a recent decline in costs associated with WGS, routine analyses of bacterial isolates of public health concern using WGS approaches have become increasingly possible for diagnostic and public health laboratories across Canada and worldwide. The use of WGS based approaches for analyzing current and prospective pathogens has the potential to highlight relationships between pathogens derived at a clinical level, and bacteria circulating throughout the increasingly complex "farm to fork" continuum. Drawing upon these relationships allows for better informed intervention strategies and public health policies aimed at reducing the burden of illness from infectious bacterial pathogens (Köser et al., 2012; Larsen et al., 2012; Sabat et al., 2013; Taboada et al., 2013).

The ability to derive results from molecular typing techniques using *in-silico* methods is a useful exercise in connecting current WGS based investigations to datasets from years past. However, while WGS data allows for the estimation of molecular typing assays *in-silico*, this approach does not exploit WGS to its maximum potential. WGS data provides the ultimate level of resolution for comparing isolates; thus WGS analyses should reflect the high-resolution capabilities afforded by the technology. For the genomic analysis of 298 recently sequenced genomes of *C. jejuni*, I proposed the use of core-genome MLST (cgMLST), which captures the allelic variation found in the majority of genes in the *C. jejuni* core-genome. To test the hypothesis that a novel high resolution cgMLST

typing method maintains accurate genomic relationships between *Campylobacter* isolates while at the same time providing superior discriminatory power and resolution compared to molecular typing methods, results from analysis by cgMLST were compared to the *in-silico* molecular typing results derived from the WGS results of *C. jejuni* genomes using the comparing partitions framework. To test the hypothesis that genomic relationships observed via cgMLST exhibit higher concordance to the underlying epidemiologic relationships of *C. jejuni* isolates, metadata pertaining to source, time and geography were visually superimposed upon a genotype cluster of genomically related isolates.

2.2 Methods

2.2.1 Strain Selection for WGS

In order to make the best use of limited sequencing resources, a strain selection criteria was developed ensuring that as many pertinent questions as possible could be answered in the analysis of the final dataset of strains for WGS. Five guidelines to investigation were established and isolates of *C. jejuni* were selected such that the inter-strain relationships could be used to investigate at least one of the following criteria: *a*) population structure; *b*) epidemiological relationships; *c*) concurrent type-matched strains; *d*) temporal distribution; and *e*) source attribution. From genotype clusters in the CGF database, micro-clusters of two to five isolates were selected based on their partial or complete epidemiologic relatedness with respect to year, location, and source of sampling. For example, a micro-cluster of four isolates were be selected that all shared the same CGF fingerprint, the same animal source and province of isolation, but differed in the year of isolation such that after the first isolate was selected, each preceding isolate was sampled one year subsequent. The approach in this example would therefore allow us to examine the effect of temporal distribution on the genomic relatedness of the isolates.

A total of 139 isolates selected for sequencing were collected as part of the FoodNet Canada Enteric Disease Surveillance Network (formerly C-EnterNet) for the

years of 2005-2010. Sample collection procedures for these isolates can be found at http://www.phac-aspc.gc.ca/foodnetcanada/niedsp10-pnisme10/index-eng.php (Accessed 15 June, 2015). The remainder of isolates were sampled from a variety of local, provincial and federal health initiatives. For detailed sample information on the isolates used in this study, refer to Table 6.1. In total, 298 isolates were selected to be processed for WGS.

2.2.2 DNA Extraction and QC

Isolates selected for analysis were recovered from archival glycerol stocks (60% glycerol in PBS stored at -80°C). Stocks were streaked for isolation onto modified cefoperazone charcoal deoxycholate agar (mCCDA, Oxoid CM0739, with selective supplement SR0155E). Cultures were incubated for 24-48 hours in a tri-gas microaerobic environment (MAE, 10% CO_2, 5% O_2, 85% N_2) at 42°C. Single colonies were selected and spread to blood agar plates (BBL Blood Agar base, BD 211037, 5% sheep blood) and incubated overnight under MAE prior to harvesting biomass. Genomic DNA extractions were performed using the QIAgen genomic tip 20/G kit according to the manufacturers recommendations. Quantity and integrity of genomic DNA were assessed using the Quant-IT HS fluorometric assay (Life Technologies Q-33120) and via gel electrophoresis on 0.8% agarose, respectively; samples with poor DNA yield or with partially degraded DNA were re-extracted. CGF subtypes of all isolates were confirmed post-extraction as a quality control/assurance step by performing CGF analysis on the extracted genomic DNA used for whole genome sequencing (Taboada et al., 2012).

Paired End Tagged (PET) sequencing libraries were generated at the BC Cancer Agency Genome Sciences Centre (Vancouver, Canada) and WGS data was obtained using the Illumina HiSeq platform (100 bp PET reads). Eighty-three isolates were run per indexed sequencing lane (two lanes total) yielding, on average, 375-fold coverage per isolate. PET Libraries for the remaining isolates were prepared at the National Microbiology Laboratory (Winnipeg, Manitoba), and sequenced using the Illumina MiSeq sequencer, pooling

approximately 30 strains per run for coverage of approximately 80-100 fold per isolate.

Draft genome assemblies from both the Illumina HiSeq and MiSeq runs were generated *de-novo* using the St. Petersburg Academy genome assembler (SPAdes) (Bankevich et al., 2012) using a hash length of 55. Four of the 298 genomes sequenced did not pass assembly quality requirements and were thus removed from the dataset, yielding 294 draft genomes available for *in-silico* typing.

2.2.3 Comparison of Typing Methods and Visualization of Strain Relationships

In-silico typing results were generated for the Canadian *C. jejuni* dataset by subjecting the 294 sequenced draft genomes to BLAST analysis using the program Microbial *in-Silico* Typer (MIST) (Kruczkiewicz et al., 2013). MIST generates *in-silico* typing data by performing sequence homology searches on the draft genome assemblies based on a set of specifications input by the user. In the case of analysis by Sequence Typing, MIST searches for known alleles in the draft sequence data at each user-specified locus and records the allele number found in the draft genome; when a novel allele is found, a new allelic number is designated. *In-silico* typing assays were generated for MLST (Maiden et al., 1998) and rMLST (Jolley et al., 2012) from *Campylobacter* typing profiles provided at http://pubmlst.org/campylobacter and http://pubmlst.org/rmlst/ (Accessed 15 June, 2015).

In an effort to leverage the high resolution afforded by WGS analysis, a whole-genome MLST (wgMLST) approach for bacterial population analyses was employed similar to that described previously by Sheppard et al. (Sheppard et al., 2012), but focusing on the *Campylobacter* core genome only. Using the set of core allelic definitions for 1,343 loci provided at the *Campylobacter* pubMLST website (http://pubmlst.org/campylobacter/info/cgMLST.shtml Accessed 17 June, 2015) as homology queries, the collection of 294 draft genome assemblies were subjected to analysis using MIST, with the intent of establishing a subset of genomes with complete sequence data (i.e. no contig truncations) for all loci queried (Barker D, *personal communication*). The final core-genome MLST (cgMLST) assay com-

prised a total of 729 loci that exhibited no sequence truncations in 274 *Campylobacter* draft genome assemblies; this became the final dataset.

The comparing partitions framework provides metrics that allow for the direct comparison of the ability of molecular typing systems to differentiate between microbial isolates, while also measuring the level of agreement between the cluster membership of partitions created by each method (Carriço et al., 2006). This framework therefore allows for the evaluation of novel typing systems, such as cgMLST, against established methods that have been used extensively in the field of molecular epidemiology. Carillo et al. (2012) extended the concept of comparing typing methods against a *gold standard* by using single-nucleotide polymorphism (SNP) data at several levels of resolution, allowing for a comprehensive comparison of conventional typing methods against the "true" phylogeny of *C. jejuni* derived from whole-genome data (Carrillo et al., 2012). By testing the comparing partitions metrics against cgMLST at multiple levels of granularity (cluster thresholds), it becomes possible to investigate not only the effectiveness of cgMLST as a typing method for *C. jejuni*, but also estimate the relative strength and flexibility of the method compared to molecular typing methods.

Clustering results from each typing method were generated using the global optimal eBURST clustering algorithm implemented in the *Java*-based software, PHYLOViZ (Feil et al., 2004; Francisco et al., 2009, 2012). Full minimum spanning trees (MST) were calculated in order to derive the complete range of groupings available. As the PHYLOViZ software will automatically collapse isolates with identical typing data into a common node, a single dummy locus consisting of a unique identifier for each strain being tested was added to the data from each method. This extra locus allowed for the visualization of each individual isolate on the MST with a null effect on the clustering of isolates. Calculation of the Simpson's Index of Diversity was performed using the Comparing Partitions webserver, available at *www.comparingpartitions.info* (Accessed 15 June, 2015), and custom scripts were written in the R language for statistical computing to generate the Adjusted Rand and

Adjusted Wallace results (R Core Team, 2015). Source code for R scripts are available at https://github.com/hetmanb/Thesis

2.2.4 Analysis of the Genomic Epidemiology of *C. jejuni* Isolates

Sampling data recorded in the Canadian CGF database such as the location, date, and source matrix for each *C. jejuni* isolate was used in this study for the comparison of genomic data to the epidemiology of Canadian *C. jejuni*. Detailed information on the sampling data for each isolate used in this study are provided in Table 6.1. By superimposing epidemiologic attributes on the MST created using *Phyloviz*, similarities between the source, date and location of sampling were used to assess the concordance between the genomic and epidemiological relatedness of clustered isolates. In addition to typing profiles and curated epidemiologic metadata available in the Canadian CGF database, the *Campylobacter* pubMLST database was used to compare results from this study to an internationally relevant dataset comprising 32347 isolates at the time of writing. Information from the BIGSdb *Campylobacter* isolate database was used to obtain an epidemiologic break-down of the *in-silico* MLST sequence types (ST) in the current study (Jolley & Maiden, 2010).

2.3 Results and Discussion

2.3.1 Evaluation of cgMLST as an Effective Typing Tool

Figure 2.1 demonstrates the SID values derived at decreasing clustering thresholds for each of CGF, MLST, rMLST and cgMLST typing methods when used to assess the 274 *C. jejuni* genomes included in the dataset. Consistent with results seen elsewhere, CGF provides higher discriminatory power than MLST (Taboada et al., 2012), while rMLST displays enhanced discriminatory power compared to both CGF and MLST, yet does not meet the same level as that of cgMLST. In comparison to the other methods shown, cgMLST produces increased discriminatory power, with a maximum SID of 1.0, establishing 269 of a possible 274 discriminatory profiles (Table 6.2).

When applied to typing methods, Simpsons Index of Diversity (SID) measures the discriminatory power of a method by calculating the probability that any two typed strains sampled randomly from the population would be differentiable by the typing method being assessed (Grundmann et al., 2001; Hunter & Gaston, 1988; Simpson, 1949). The values calculated by the SID range from 0.0 to 1.0, with a value of 1.0 indicating that the method in question can discriminate between every strain tested; conversely, a value of 0.0 indicates that all strains in the population being tested would be ascribed an identical type assignment. To determine the discriminatory power of cgMLST compared to molecular typing methods, results from cgMLST, CGF, MLST, and rMLST were used to calculate SID metrics. Both CGF and MLST have been shown to have high concordance to one another with CGF providing superior discrimination (Taboada et al., 2012), while rMLST, targeting 52 ribosomal loci in the bacterial genome, was developed in part to provide increased resolution over the seven core genes used in traditional MLST analysis (Jolley et al., 2012).

A major advantage to using a high resolution scheme such as cgMLST is the flexibility afforded to performing genomic analyses. By assessing the cluster membership at decreasing thresholds of cgMLST (e.g. 100% threshold = all 729 cgMLST loci sequences must exactly match for two isolates to be considered identical; 90% threshold = 659 of the total 729 loci sequences need to match for two isolates to be considered equal) typing results generated by cgMLST can, in fine detail, be compared to those of other, legacy typing methods. By re-clustering the sample of 274 draft genomes using cgMLST at decreasing clustering thresholds, the cgMLST was shown to maintain the highest discriminatory power until a cluster threshold of 96% (equivalent to allowing 29 mismatches for clustering two isolates together), at which point it approximated the same maximum discriminatory power as rMLST.

The SID of cgMLST dropped to the maximum levels of CGF and MLST at 85% and 75% clustering thresholds, respectively, demonstrating the increased discriminatory power that cgMLST has over these methods. Further, as the clustering threshold for each typing

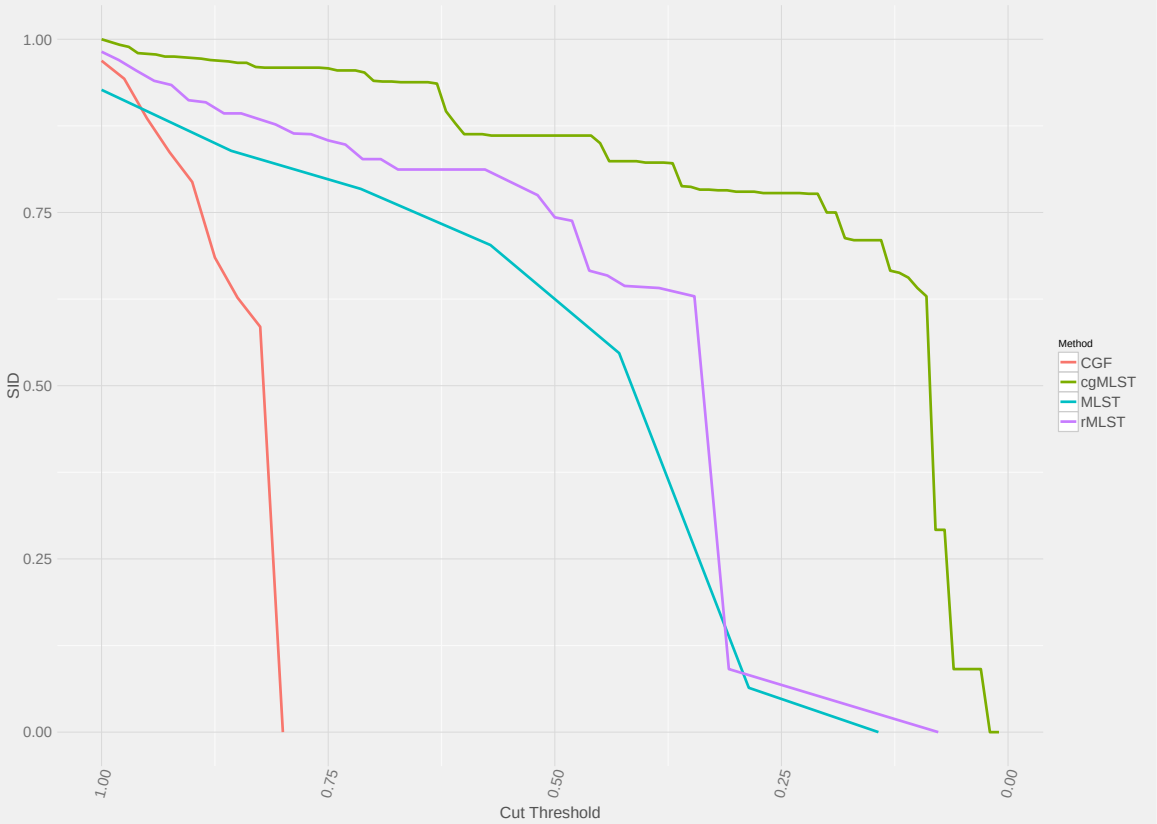

Figure 2.1: Simpsons Index of diversity for *in-silico* generated typing methods at indicated clustering thresholds. Clustering thresholds were generated using the goeBURST clustering algorithm, as supplied in the *Phyloviz* software package, and SID measurements were calculated using the Comparing Partitions online toolkit. See Table 6.2 for detailed results.

method was reduced, a sharp decline was observed in the SID of CGF, MLST, and rMLST at approximately 80%, 40% and 30%, respectively, while the SID of cgMLST remained robust to change until it approached a 10% clustering threshold, where it declined sharply (Figure 2.1). The robust discriminatory power of cgMLST thus allowed for testing the method across a wide range of clustering thresholds without sacrificing the ability of the method to discriminate among bacterial genomes.

The second metric proposed by the comparing partitions framework is the Adjusted Rand (AR) measurement, which is adopted from the Rand coefficient for partition agreement, but extends it to include the possibility that agreement between two partitions could occur by chance alone (Hubert & Arabie, 1985; Rand, 1971). By calculating the AR be-

tween two typing systems, it is possible to assess the bidirectional agreement between methods; that is, a measurement indicating the level of overall concordance between the membership of clusters created by method A (e.g. cgMLST) and those created by method B (e.g. CGF, MLST, rMLST). If pairs of isolates in method A are clustered together identically in method B, then the AR approaches a maximum of 1.0, as the cluster pairings in the two methods becomes more disparate, the AR decreases towards 0.

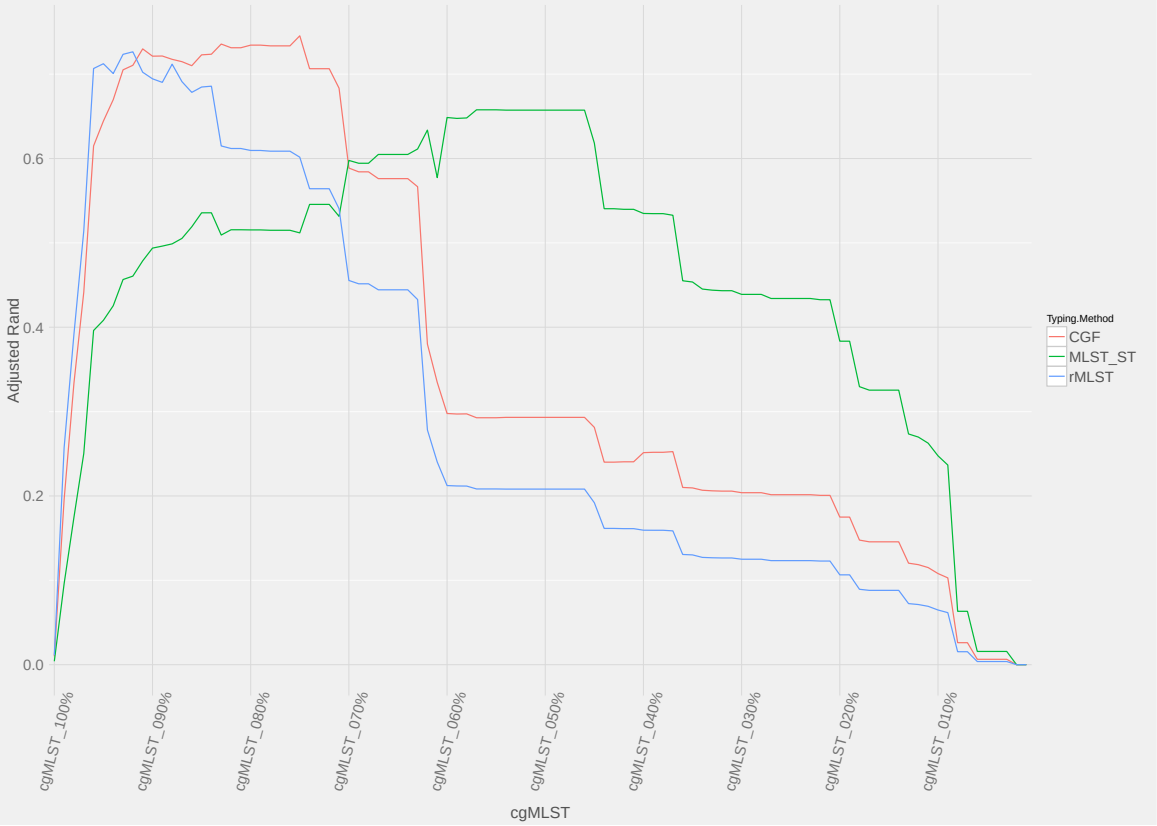

Figure 2.2: Adjusted Rand results for decreasing thresholds of cgMLST versus establish *C. jejuni* molecular typing methods. Cluster memberships for all thresholds were derived using the goeBURST algorithm from the software package *Phyloviz* and the Adjusted Rand statistic was calculated using a custom script written in the R language for statistical computing.

Thresholding cgMLST at 100% produced higher discriminatory power than any of the other typing methods being tested, thus the number of clusters produced by this method was far greater than that of CGF, MLST, and rMLST (Table 6.2). Large discrepancies in discriminatory power between typing methods negatively affects the AR, as the calculation

relies on pairs of isolates in the same clusters between two methods, however, by calculating the AR at incremental levels of decreasing granularity for cgMLST while keeping CGF, MLST and rMLST constant, it is possible to assess relative maximum of each of the methods being tested when compared against the cluster memberships of cgMLST (Figure 2.2).

As expected, at a 100% clustering threshold, an AR result approaching zero between cgMLST and each of the three legacy methods is observed due to the discrepancy in the number of clusters generated by each method (Figure 2.2). However, as the thresholding level for clustering cgMLST results decreases, a rise in the calculated AR is observed, with a maximum AR for CGF (AR = 0.745) achieved at 75% clustering of cgMLST, and maximum AR for MLST (AR = 0.658) and rMLST (AR = 0.723) at 57% and 93% clustering thresholds, respectively. These maximum values represent the point where cgMLST and each of the legacy typing methods best agree with one another, and give us confirmation that cgMLST maintains clusters generated by typing systems that have been used extensively for population genetics and epidemiologic studies.

As mentioned already, one challenge to using the AR to assess the congruence between typing methods is that the AR is negatively affected when one method being assessed has much higher discriminatory ability than the other method. Largely, this problem arises due to a lack of consideration of directionality in the AR calculation. The AR measures the overall congruence between two methods in a bidirectional fashion, thus it averages the fit between method A$\rightarrow$B and B$\rightarrow$A to result in a metric that is lessened when one method performs considerably worse than the other.

To help mitigate this issue of directionality, the Wallace coefficient was proposed, and later adjusted to include confidence intervals (Severiano et al., 2011; Wallace, 1983). The Adjusted Wallace coefficient (AW) provides two separate, directional measurements for the comparison of typing methods, namely the $AW_{A\rightarrow B}$, and the $AW_{B\rightarrow A}$. In calculating the AW for the congruence of cgMLST to other typing methods, a directional assessment of

whether cgMLST provides good fit to the partitions created by other typing methods can be accomplished, allowing for the identification of instances where a low AR result was due to (a) poor fit between both methods; or (b) the difference in discriminatory ability between methods devaluing the overall score.

Table 2.1: Adjusted Wallace values for different clustering thresholds of cgMLST as compared to the indicated typing tests. Results generated by *in-silico* typing of 274 isolates of *C. jejuni*. Directionality of results is indicated, with method A referring to cgMLST thresholds indicated in the first column, and method B pertaining to the typing methods indicated above. Values in parentheses correspond to Jackknife pseudo values for 95% confidence intervals.

	CGF (95% CI)		MLST (95% CI)		rMLST (95% CI)	
cgMLST	$AW_{A\rightarrow B}$	$AW_{B\rightarrow A}$	$AW_{A\rightarrow B}$	$AW_{B\rightarrow A}$	$AW_{A\rightarrow B}$	$AW_{B\rightarrow A}$
100%	1.00(1.00-1.00)	0.01(0.00-0.01)	1.00(1.00-1.00)	0.00(0.00-0.00)	0.66(0.00-1.00)	0.01(0.00-0.02)
99%	0.88(0.78-0.99)	0.11(0.06-0.16)	1.00(1.00-1.00)	0.05(0.03-0.07)	0.73(0.53-0.93)	0.16(0.09-0.22)
98%	0.87(0.77-0.96)	0.21(0.14-0.27)	0.99(0.97-1.00)	0.10(0.07-0.13)	0.68(0.53-0.83)	0.28(0.20-0.35)
97%	0.84(0.74-0.94)	0.30(0.23-0.37)	0.99(0.98-1.00)	0.14(0.10-0.19)	0.68(0.55-0.81)	0.41(0.31-0.51)
96%	0.79(0.68-0.91)	0.50(0.39-0.62)	0.97(0.94-1.00)	0.25(0.17-0.32)	0.68(0.55-0.81)	0.73(0.62-0.85)
95%	0.80(0.69-0.91)	0.54(0.43-0.65)	0.95(0.91-1.00)	0.26(0.18-0.34)	0.67(0.54-0.79)	0.76(0.65-0.88)
90%	0.77(0.67-0.87)	0.68(0.58-0.78)	0.94(0.89-0.99)	0.33(0.25-0.42)	0.58(0.46-0.69)	0.87(0.78-0.96)
85%	0.69(0.60-0.79)	0.76(0.66-0.85)	0.88(0.81-0.94)	0.39(0.29-0.48)	0.53(0.42-0.64)	0.98(0.95-1.00)
80%	0.64(0.54-0.75)	0.86(0.78-0.94)	0.74(0.64-0.84)	0.40(0.29-0.50)	0.44(0.34-0.54)	1.00(0.99-1.00)
75%	0.65(0.55-0.75)	0.88(0.81-0.95)	0.72(0.62-0.82)	0.40(0.29-0.50)	0.43(0.33-0.53)	1.00(0.99-1.00)
70%	0.44(0.35-0.54)	0.88(0.81-0.95)	0.67(0.58-0.76)	0.54(0.43-0.65)	0.29(0.22-0.37)	1.00(1.00-1.00)
65%	0.43(0.33-0.52)	0.88(0.81-0.95)	0.67(0.57-0.76)	0.55(0.44-0.66)	0.29(0.21-0.36)	1.00(1.00-1.00)
60%	0.18(0.13-0.23)	0.88(0.82-0.95)	0.49(0.39-0.58)	0.97(0.92-1.00)	0.12(0.08-0.16)	1.00(1.00-1.00)
50%	0.18(0.13-0.22)	0.89(0.82-0.95)	0.49(0.39-0.59)	1.00(1.00-1.00)	0.12(0.08-0.15)	1.00(1.00-1.00)
40%	0.14(0.11-0.18)	0.98(0.95-1.00)	0.37(0.27-0.46)	1.00(1.00-1.00)	0.09(0.06-0.11)	1.00(1.00-1.00)
30%	0.11(0.09-0.14)	1.00(1.00-1.00)	0.28(0.20-0.36)	1.00(1.00-1.00)	0.07(0.05-0.09)	1.00(1.00-1.00)

Examination of the AW coefficients derived from comparison of cgMLST and the other typing methods shows that in almost all cases, there is a significantly different result when considering the directionality of comparison (Table 2.1). Thus, the low AR metric observed between cgMLST clustered at 100% and CGF, MLST and rMLST typing results (Figure 2.2), were a result of the much lower discriminatory power of each typing method when compared to cgMLST. At a 100% clustering threshold, cgMLST perfectly predicts the partitions created via analysis by CGF and MLST, and predicts 66% of the cluster membership of rMLST, while still retaining higher discriminatory ability than any of the three contrasting methods. By decreasing the clustering threshold of cgMLST, the level at which cgMLST shows highest concordance with each of the other methods can be evaluated in both forward and reverse directions.

Results from the SID, AR and AW calculations performed here indicate that typing results generated by *in-silico* cgMLST provide much higher discriminatory power than other established typing methods tested, and retain the ability to cluster isolates into groupings that still agree with these accepted methods. The clustering threshold at which cgMLST is assessed can be reduced without deteriorating the ability of the method to remain highly discriminatory and produce clusters congruent with typical *Campylobacter* typing methods. This flexibility of the cgMLST method is critical to performing an investigation attempting to connect the genomic signal of *C. jejuni* isolates to their epidemiology.

Assessing the genomic epidemiology of *C. jejuni* isolates using cgMLST at a 100% clustering threshold is likely too highly discriminatory for such an analysis: 100% clustering similarity using cgMLST split the sample of 274 isolates into 269 separate clusters, removing potentially informative genomic connections among the isolates shown (Table 6.2). By reducing the clustering threshold that is used to assess the cgMLST results, an attempt can be made to establish clusters of isolates with high epidemiologic relevance. Care must be taken, however, to ensure that the selected clustering threshold results in large enough clusters such that the epidemiological relevance of genomically related isolates can

be assessed, without clustering isolates together with only moderate similarities.

2.3.2 Selecting Isolates based on Epidemiologic Relationships

Data from the Canadian CGF database were used to select isolates of *C. jejuni* for WGS based on their epidemiologic and genotypic relatedness. Isolates were first selected from CGF clusters sharing identical CGF fingerprints in an attempt to ensure a baseline level of genetic relatedness. The recorded year, province, and source of isolation were then used as epidemiologic factors to identify groups of isolates within the CGF clusters that demonstrated high epidemiologic similarity. In general, micro-clusters of three to five isolates were selected from each CGF cluster such that the genomic effects arising from changes in one or two epidemiologic factors could be assessed. Figure 2.3 demonstrates four examples of micro-clusters selected in this way, and shows the genomic differences between isolates as the number of different cgMLST loci between each of the isolates in each cluster.

The three isolates shown in Figure 2.3(A) correspond to *C. jejuni* sampled from Alberta, Canada in 2004, and share the 949.1.2 CGF fingerprint. Isolates CI-0168 and CI-0182 were sampled from cattle, while isolate CI-0392 was sampled from sewage near the same location and sampling date. Allelic differences among cgMLST profiles for each isolate appear to reflect the close epidemiologic relationship between the triad. Five loci differences between CI-0168 and CI-0182 represent a similarity of over 99% (724/729 loci) based on cgMLST. The genome of CI-0392, a *C. jejuni* isolate sampled from sewage, maintains a very strong similarity to the cattle isolates CI-0168 and CI-0182; only eight and 11 loci differences were observed between the CI-0392 genome and those of CI-0168 and CI-0182, respectively. The close membership of this grouping based on cgMLST profiles supports the selection criteria based on the epidemiology related to sampling conditions, as well as supports the groupings achieved by clustering with CGF.

Figure 2.3(B) includes three isolates of *C. jejuni* derived from river water in Ontario,

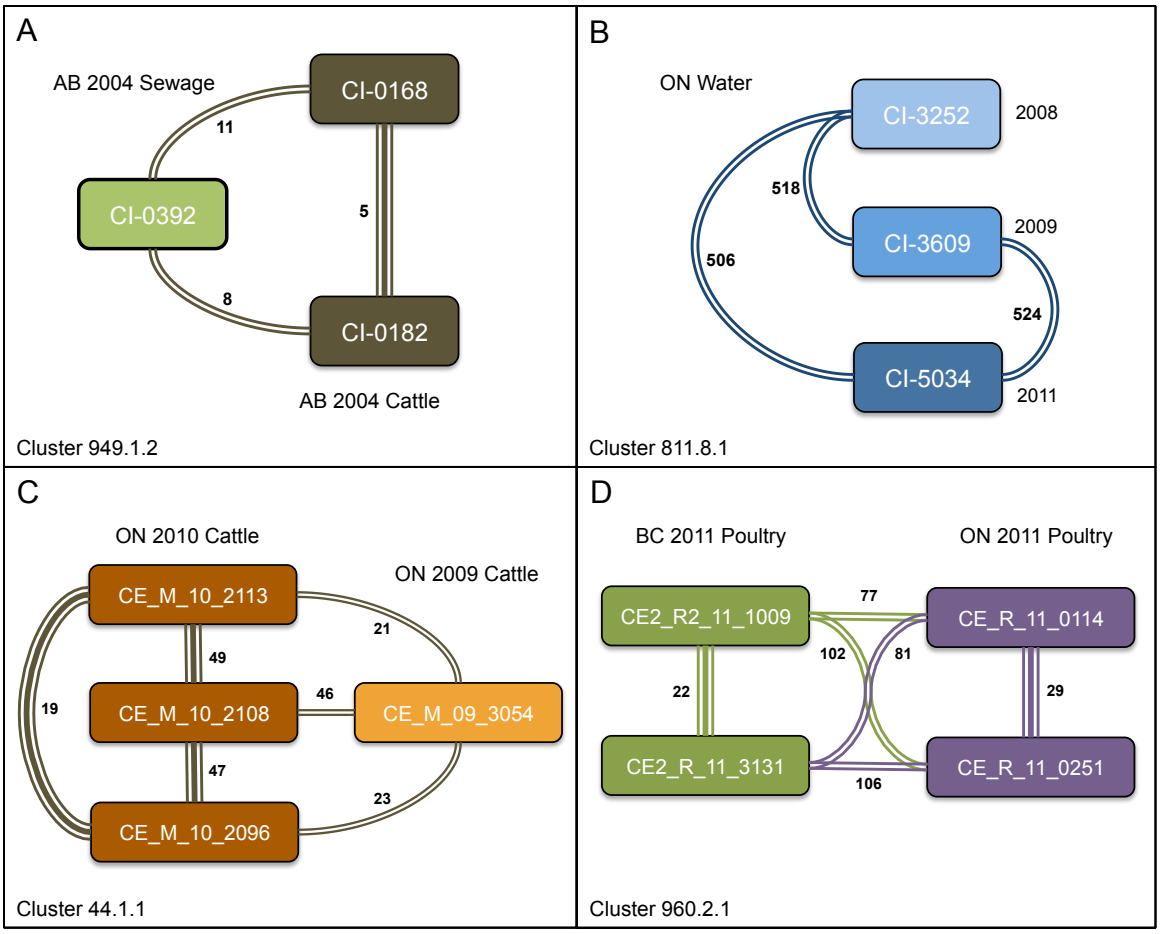

Figure 2.3: Epidemiologic relationships used to select isolates of *C. jejuni* for WGS. The number of shared epidemiologic attributes (e.g. Year, Province, Source) are indicated by the thickness of connecting lines between isolates. Numbers on connecting lines indicate the number of cgMLST allelic differences.

Canada, for the years 2008, 2009 and 2011 that all shared an identical CGF fingerprint. In selecting three isolates from the same sampling source and geography, the effect of temporal differences on the genetic closeness between isolates could be assessed; however, by comparing the cgMLST profiles of each of these isolates with their temporal data, a significant number of divergent cgMLST loci separating all three of the isolates was observed. Among these three isolates, few genomic effects were observed from a sampling difference of one year (2008-2009), two years (2009-2011) and three years (2008-2011). All three isolates showed high differentiation based on their cgMLST profiles, with distances of over 500 loci observed between all pairings. Based on these observations alone, it is impossible

to determine the magnitude of effect of temporal change on core genome difference. It may be the case that even a one year separation between samplings is too distant to draw connections between the core genomes of these *C. jejuni*, as there appeared to be little difference in the number of matching alleles between genomes sampled one year apart or three years apart. Alternatively, it is possible to have complex networks of influx leading to contamination of waterways from various agricultural and animal sources, thus while these isolates appeared to be resident to water, one or more may have been associated with a separate external source or geography.

The fact that isolates in Figure 2.3(B) did not reflect high similarity assessed via cgMLST, yet were selected from a CGF cluster of 100% similar indicates that accessory genome content may not provide a good estimate of core genome similarity. However, the CGF profile of these isolates (CGF cluster 811.8.1) represents only 16 positive loci among a possible 40 loci; thus it may be the case that extending the number of accessory loci used to compare these isolates via CGF would show more similar results to the analysis of the cgMLST profiles.

The isolates in Figure 2.3(C) were selected to assess differences in the core genome between *C. jejuni* sampled in the same year, versus one year apart within the lineage defined by the CGF cluster 44.1.1. All four isolates were sampled from similar sources (retail beef) and geography (Ontario), with three isolates sampled in 2010 and the fourth in 2009. Similar to the example shown in Figure 2.3(B), the cgMLST profiles of the genomes in Figure 2.3(C) did not show an obvious effect based on temporal differences. The number of differing loci between the cgMLST profiles of the genomes of isolates sampled in the same year showed almost no change compared to loci differences between genomes of isolates sampled one year apart; the genome of CE_M_10_2108 produced the largest number of diverging cgMLST loci when compared to neighbouring genomes, irrespective of the year of sampling.

The isolates in Figure 2.3(D) were chosen to investigate the effect of geogra-

phy on cgMLST profiles. Isolates CE2_R2_11_1009, CE2_R_11_3131, CE_R_11_0114 and CE_R_11_0251 were sampled during 2011 from retail poultry samples; however, CE2_R2_11_1009 and CE2_R_11_3131 were sampled in British Columbia while CE_R_11_0114 and CE_R_11_0251 were isolated in Ontario. Here, there is see a discernable difference in the cgMLST profiles with regards to location of sampling. The 22 cgMLST loci differences between the genomes of CE2_R2_11_1009 and CE2_R_11_3131 and 29 cgMLST loci differences between the genomes of CE_R_11_0114 and CE_R_11_0251 seem to indicate that isolates from physically closer sampling sites may posess more similar cgMLST profiles than those sampled from more distant locales. Isolates sampled from more distant locations appeared to exhibit less similarity between cgMLST profiles: a minimum of 77 cgMLST loci differences were observed between the genomes of isolates from British Columbia and those from Ontario (Figure 2.3(D)).

By leveraging the genotype data and epidemiologic metadata from the Canadian CGF database, clusters of epidemiologically linked isolates were selected to test the hypothesis that these relationships would coincide with strong evidence for association via cgMLST. Isolates shown in panels A and D from Figure 2.3 support the hypothesis that cgMLST similarity is a good indicator of geospatial and source relationships, and these examples help to reinforce the CGF method as a robust system for clustering isolates based on genomic and epidemiologic relatedness. Surprisingly, however, examples B and C demonstrated that predicting genomic similarity between isolates based on temporal relatedness can be challenging. The hypothesis that higher cgMLST similarity would be observed between isolates sampled at closer timepoints was not supported by this analysis; in both Figure 2.3(B and C), the year of sampling had essentially no effect on the number of discordant cgMLST loci between pairings.

2.3.3 Visualizing Genomic Relationships

Having examined the cgMLST distances between some micro-clusters of isolates selected from the Canadian CGF database for their genotypic and epidemiologic relationships, the ability of cgMLST to agnostically cluster isolates into epidemiologically related groupings was next tested. Because of the high discriminatory power of cgMLST it was necessary to select a clustering threshold with high enough discriminatory power that unrelated isolates would not be connected, while still providing sufficient apportionment to capture the epidemiologic relationships present. For the purposes of this investigation, cgMLST clustered both at the 90% and 80% thresholds was used. Thresholding cgMLST at 80% provides a discriminatory power between that of MLST and CGF and creates clusters large enough in size to investigate potential epidemiologic relationships while still establishing good AW metrics with the legacy typing methods (see Figure 2.1 and Table 6.2). By including the 90% thresholds in the analysis, more refined clusters within those established at the 80% level were visible, potentially identifying niche clusters of high specificity. This multi-faceted approach was used in an attempt to uncover stronger evidence of associated epidemiological relationships identified by the high resolution genomic clustering results.

Minimum spanning trees (MST) were generated in *Phyloviz* to visualize the genetic relatedness of the Canadian *C. jejuni* isolates via cgMLST-based analysis. The analysis was restricted to the largest cluster generated by the 80% clustering of cgMLST, yielding a group of 31 isolates of *C. jejuni* sampled from a variety of sources, timepoints and locations. Figure 2.4(A) depicts the MST of this cluster, with isolate names imposed on each of the nodes, and the number of mismatched cgMLST loci indicated by the number on the connecting branches between strains.

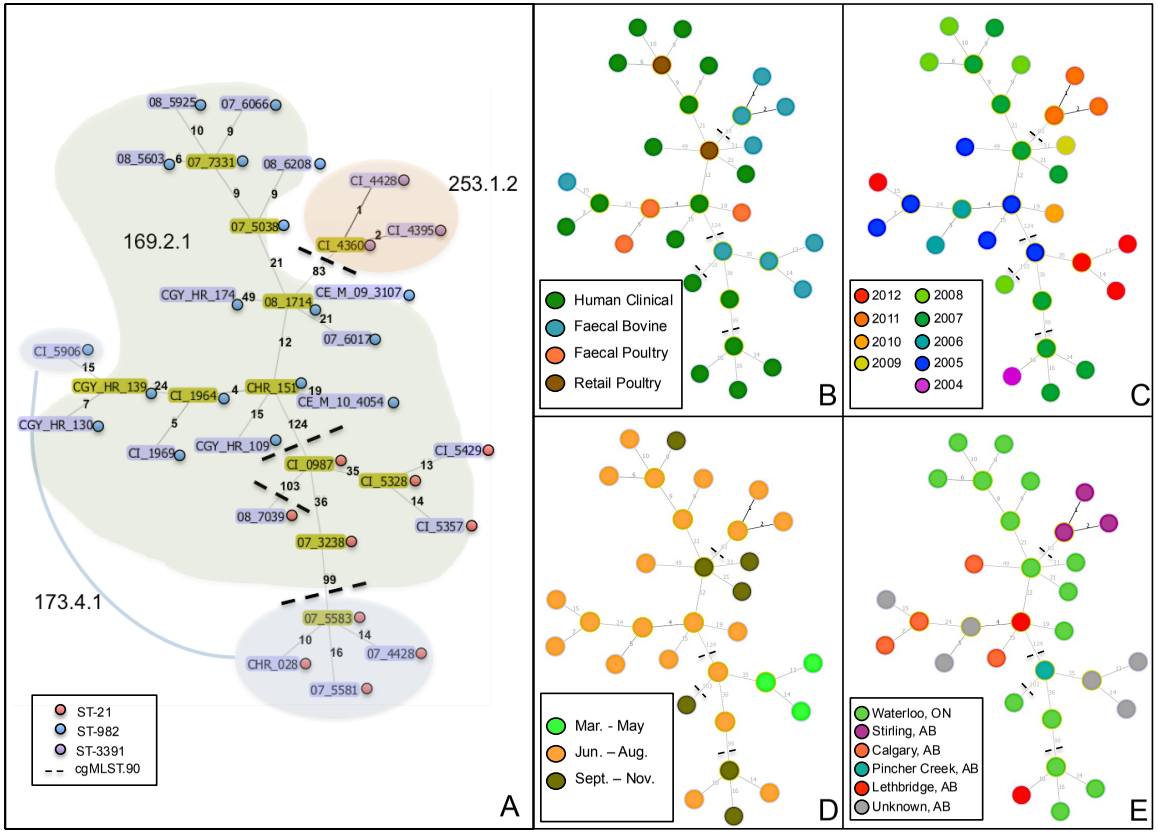

Figure 2.4: A) Minimum spanning tree of Canadian *C. jejuni* isolates generated via cgMLST at a threshold of 80%. Dotted lines indicate threshold cuts at 90%. Numbers on linkage branches indicate the total cgMLST loci between nodes with different allelic assignments. Relative MLST Sequence Types are indicated by red, blue and purple dots and referenced in the legend. CGF clusters are indicated by blue, orange and green halos. B-E) Epidemiological overlays for the MST in Part A, conforming to B) Source of isolation C) Year of sampling D) Season of sampling E) City or Province of isolation.

To compare the Canadian *C. jejuni* isolates here with international results from the *Campylobacter* pubMLST database, *in-silico* MLST was performed for each of the tested genomes which led to the discovery of three ST associated with the cluster in Figure 2.4(A): ST-21, ST-982 and ST-3391. Similarly, by cross-examining the isolates' respective CGF fingerprint data from the CGF database, three CGF fingerprints were found to be present: 169.1.2, 173.4.1 and 253.1.2. By comparing the cluster assignments generated by cgMLST_90% to MLST and CGF assignments, the higher discriminatory power of cgMLST allowed for an increase to the number of clusters generated in the dataset; five clusters were formed among the 31 genomes using cgMLST_90%, where the analysis by CGF and MLST only produced three clusters.

It is interesting to note that the two breakpoints located between CI-0987 to 08_7039 and 07_3238 to 07_5583 split the cluster of ST-21 associated isolates at branches indicating 103 and 99 loci differences, even though these isolates would otherwise be considered identical by typing via MLST. It has been suggested that "generalist" STs, including ST-21, may all possess phenotypes that confer advantages to *C. jejuni* with respect to survivability in the enviroment (Gripp et al., 2011). The observation that the single ST-21 cluster in Figure 2.4(A) was split into three clusters separated by a substantial number of allelic cgMLST mismatches seems to indicate that the ST-21 lineage actually constitues a number of sub-lineages that are not captured by the allelic variation at the seven loci used in the MLST assay. If each of these sub-lineages exhibited a distinct phenotype, then it may be the case that the "generalist" phenotype is an incorrect classification for isolates associated with the ST-21 designation. Rather, it is the sum of several distinct phenotypes that are incorrectly recognized as "generalist" behaviour. The number of ST-21 associated isolates in the current study is insufficient to test this hypothesis - however, it remains an avenue for future investigation.

Analysis by CGF clusters the genome of CI-5906 within the CGF fingerprint 173.4.1, indicating 40/40 matched loci from the accessory genome, even though an average of 170

core genome loci differences exist between the genomes of CI-5906 and the four other isolates from cluster 173.4.1 shown in Figure 2.4(A). The grouping of these isolates via CGF is a clear example where the similarity via accessory genome content differs substantially from the similarity observed via the core genome. The accessory genome is thought to contain genes responsible for rapid host adaptation (Medini et al., 2005); thus the similarity in the accessory genome content between these genomes may provide evidence as to the etiology of the human clinical isolates in this cluster.

2.3.4 Genomic Epidemiology of Canadian *C. jejuni*

Metadata pertaining to source, temporal, geospatial, and seasonal information was superimposed on the MST in Figure 2.4(A) to generate the four colour-coded MSTs seen in panels B-E. Figure 2.4(B) depicts the relationships of strains from Figure 2.4(A) with reference to the sources from which they were sampled. Overall, isolates from human clinical samples are dispersed throughout the 31-isolate cluster, with connections to faecal bovine, and retail and faecal poultry samples. This observation corroborates findings seen previously, where exposure to both poultry and cattle have been implicated as major risk factors for human campylobacteriosis (Nichols et al., 2012).

Based on comparing the the MST and sampling source metadata shown in Figures 2.4(A and B), genomes from the ST-21 group appear to be closely associated with cattle-derived isolates, while the ST-982 and ST-3391 groups exhibit a mixture of both poultry and cattle associations. This observation is supported by analysis of the CGF types 173.4.1, 169.1.2 and 253.1.2, which show predominent cattle associations in the CGF database, and associations to poultry, human and environmental sources as well. ST-21, a highly prevalent international sequence type, represents over 5% of the total isolates from the pubMLST *Campylobacter* database, a repository containing information on over 32,000 isolates. ST-21 has been sampled from a total of 22 various human, animal and environmental sources, though cattle and poultry remain predominant (6.75% and 4.09% respectively). When as-

sessed at the cgMLST_90% threshold, the ST-21 associated isolates in Figure 2.4(A) were split into two multi-isolate clusters, one associated with human clinical isolates; and another containing four faecal bovine isolates linked to a single isolate sampled from a clinical source (Figure 2.4(B)). The position of the human clinical isolate 07_3238 within a bovine-associated cluster suggests a possible cattle-associated etiology for the human clinical isolate; only 36 cgMLST loci differ between the genomes of 07_3238 and CI-0987, indicating a high degree of similarity between these two isolates sampled from human clinical and cattle sources.

Analysis of the *Campylobacter* pubMLST database revealed that isolates with the ST-982 sequence type designation have been isolated less frequently than ST-21, and show a high occurrence in North America compared to the UK and other European countries. An interesting feature of this analysis, is that based on data within pubMLST, ST-982 has not previously been reported in poultry samples, and may have been otherwise considered only cattle-associated. The present analysis clearly shows isolates of ST-982 associated with both retail and faecal poultry samples (Figures 2.4(A and B)), suggesting that this sequence type may possess a wider source-association than previously indicated. The three isolates associated with ST-3391, CI-4428, CI-4395 andCI-4360 were all sampled from bovine sources and possess remarkably similar genomes at the level of cgMLST. Only one instance of this ST has been recorded previously, from a UK cattle isolate in 2007. The associated CGF cluster: 253.1.2, contains over 50 isolates sampled from farm livestock sources, including both cattle and poultry. Thus, while ST-3391 may have been seen only rarely in european sampling data, the Canadian CGF data indicates that these strains are commonly found within a Canadian agricultural context.

The temporal distribution of isolates was visualized by the superimposition of the year of sampling in Figure 2.4(C). Isolates observed in Figure 2.4(A) were sampled from a range of nine years, spanning 2004 to 2012, providing evidence that these genomes may represent a persistent group of genotypes in Canada. This hypothesis is supported by sampling data

from the CGF database, as the three CGF clusters 173.4.1, 253.1.2 and 169.2.1 have been observed across a range of 9-15 years of sampling, and is also supported by the analysis of sampling source in Figure 2.4(B). By adapting to a diverse range of environmental and host niches, these *Campylobacter* strains may confer enhanced phenotypic features related to survival, thus enabling them to persist as a stable group of closely related genotypes throughout many years. The flexible nature of the *Campylobacter* genome is thought to allow it to rapidly adapt to new source environments; by acquiring genetic material via recombination with bacteria both inter and intra-specially, *C. jejuni* can potentially adapt to new source environments in much less time than selection by mutation would normally allow (Wilson et al., 2009). The genotypes observed here may therefore allow for increased adaptability, potentially promoting their success as closely related long-term *Campylobacter* genotypes (Croll & McDonald, 2012).

While strong concomitant temporal and genomic associations can be seen for some of the isolates assessed here (e.g. CI-4428, CI-4395 and CI-4360), other highly similar genomes are separated by several years in sampling date, (e.g. CHR_028 and 07_5583), suggesting that among the genotypes observed here, temporal similarity does not necessarily imply genomic relatedness. The corollary of this appears to also be true, as isolates having similar sampling dates may display large differences genomically, as is the case between *Campylobacter* isolates 07_3238 and 07_5583 (99 cgMLST loci differences), and CHR_151 and CI-0987 (124 cgMLST loci differences)(Figure 2.4(A and C)). While a strong connection may not be observed between the genomic and temporal relatedness of the genomes present in Figure 2.4(A), further investigation of clusters containing more highly restricted genotypes may be required to ascertain the full extent of temporal association on the strength of similarity between *C. jejuni* genomes.

The temporal data presented in Figure 2.4(C) was used to determine seasonal distributions of the *Campylobacter* isolates from 2004-2012 (Figure 2.4(D)). Isolates from this representative cluster were sampled from three seasons: Spring (March-May), Summer

(June-August) and Autumn (September-November). Isolation rates of *Campylobacter* from environmental samples and human campylobacteriosis cases are typically low in the winter season (December-February) and peak in early summer, consistent with warmer temperatures and increased outdoor recreational and agricultural activities. The seasonal peaks in human campylobacteriosis are most evident in children up to six years old, and increase among people who reside in rural areas; suggesting that the seasonal fluctuations in *Campylobacter* incidences may be more related to environmental circulation, rather than food consumption (Louis et al., 2005; Nichols et al., 2012).

Much like the temporal associations seen in Figure 2.4(C), there are incidences where seasonal similarities both coincide and contradict the underlying genomic relationships. The isolates CI_5429, CI_5328 and CI_5357 are closely related genomically, with only 13 cgMLST loci differing between CI_5429 and CI_5328, and were all sampled between the months of March and May, assigning them to the Spring sampling season. Examples of genomically similar isolates from different sampling seasons can also be seen, however, including isolates 07_5583 and CHR_028 that were sampled in the Fall and Summer seasons, and demonstrate only 10 cgMLST loci differences; and 07_7331 and 07_6066, also sampled in the Summer and Fall seasons, with only 9 cgMLST loci differences. A comprehensive analysis of only 31 isolates of *C. jejuni* is unlikely to uncover firm genomic trends in the context of seasonality, however, the observation that the majority of isolates were isolated from the summer season is consistent with trends seen elsewhere (Louis et al., 2005; Nichols et al., 2012).

Data representing the geographical regions of isolation was superimposed on the MST in Figure 2.4(E). As the majority of isolates sequenced for this study were sampled from Ontario and Alberta, these are the only provinces represented in the current cluster. Geographic metadata from the CGF database confirm that the genotypes present within the cluster exist across a wide geospatial range throughout Canada. Among the isolates present in Figure 2.4(E) there does not seem to be a strong indication that geographic sampling lo-

cation predicts genotypic similarity, as isolates sampled from nearby locations are shown to be separated by a large number of cgMLST loci (e.g. 99 loci difference separating 07_3238 and 07_5583); likewise isolates shown to be highly similar genotypically (e.g. 19 loci difference between CHR_151 and CE_M_10_4054) are seen to be sampled from widely separated locations across the country.

A potential factor affecting the reliability of assessing geographic location as an indicator of genomic similarity is the increased delocalization of the food industry in Canada. For example, consumption of chicken has been indicated as the number one risk factor for human campylobacteriosis (Friedman et al., 2004), but tracing the source of contaminated retail poultry can be difficult, as several stages lie in between the raising of commercial birds and their delivery to the supermarket; production stages may occur at centralized locations with distribution across the country. The human-derived *Campylobacter* isolates 07_5583 and CHR_028, are highly similar via cgMLST, although one case originates in Alberta, and the other in Ontario, a distance of several thousand kilometers. Without enhanced epidemiological information surrounding the potential exposure to *C. jejuni* for each of these cases, it is therefore difficult to ascertain the true geographic association between genotypes.

2.3.5 cgMLST for Uncovering Epidemiologic Clusters

By adjusting the threshold used to assess the high resolution cgMLST typing method, clusters of *C. jejuni* genomes were generated that reflected both broad (80%) genomic similarities, as well as provided information on more highly refined clusters, assessed at 90% cgMLST similarity. This approach allowed us to assess the cohesion of the largest 80% cluster produced from the sample dataset in regards to distinct epidemiologic parameters (Figure 2.4(B-E)). Based on an *ad-hoc* analysis using two cgMLST clustering thresholds, it appears that the ability for cgMLST to generate clusters of *Campylobacter* genomes with similar epidemiologic profiles may be limited. The 80% cluster presented in Figure 2.4 contains genomes derived from isolates pertaining to (a) several sampling sources; (b) a

wide geographic range; and (c) a temporal spread of almost a decade that represents three different sampling seasons. Applying a 90% clustering threshold split the larger cluster into much smaller groups that still demonstrated mixed epidemiologies, highlighting the challenge of performing *ad-hoc* analyses for genomic epidemiology.

Each of the categorical overlays in Figure 2.4(B-E) contained examples of close genomic relatedness between isolates that had little epidemiologic similarity; for example, isolates 07_5583 and CHR_028 represent a close genomic relationship, only differing at 10 of a possible 729 cgMLST loci; however, these isolates were sampled in distinct geographic locations and separated by three years in time of sampling. The temporal and geospatial data for this pair of isolates may appear to contradict the genomic similarity seen via cgMLST, yet this observation, combined with many of the examples listed in the source, temporal and geospatial analyses above, may actually provide evidence for the persistent nature of these genotypes in farming ecosystems. When considering the epidemiologic information provided for the isolates in Figure 2.4 as an aggregate of source, temporal and geospatial data, there appears to be a trend of these genotypes persisting in agriculturally impacted areas. Data from the CGF database indicates that these CGF types have persisted for up to 15 years of sampling, and have been predominantly associated with livestock sources from across Canada. The presence of these genotypes in distinct geographic regions and agriculturally associated sources may forecast an adaptation of these isolates to agricultural environments in general, accounting for their persistence in over a decade of sampling, and seasonality that largely coincides with farming activities.

Thus, while the cgMLST analysis here did not separate the genomes into disparate epidemiologic groupings, it did, more importantly, create connections between distinct ecologies that may provide insight to the transmission and survival dynamics of the *C. jejuni* in an agricultural ecosystem. The ability to form potentially informative connections between otherwise unrelated strains is, ultimately, far more useful than grouping isolates together by obvious nature, and allows for deeper investigation into potential sources of human illness.

2.4 Summary and Conclusion

While the use of bacterial WGS data for public health analyses has the potential to be a powerful tool for uncovering etiologies in the dissemination of infectious disease agents, significant challenges remain in the contextualization of the vast quantities of data output by WGS for use in epidemiologic analyses. Although the increasing availability of genomic data may encourage the use of the whole genome sequence for comparative genomic analyses, considerations such as data quality and robustness of comparisons need to be weighed. Generation of a closed, annotated genome represents a significant expenditure of effort, requiring either single-chromosome sequencing technologies that are prohibitively expensive for most public health laboratories, or multiple sequencing runs typically using Sanger-based sequencing to connect the contiguous sequences generated by draft shotgun sequencing approaches. Rather than invest the time and financial commitments to generating closed genomes for comparison, it is therefore more effective to leverage draft genome data, as this represents a more pragmatic approach for most laboratories.

The development of the cgMLST typing system used here was driven by a desire to leverage the high resolution and flexibility of draft WGS data in a means that is both robust and portable; the addition of new genomes to the analysis should not require the establishment of a novel typing scheme. By leveraging the majority of the *C. jejuni* core genome in the scheme, relationships observed between genomes are stable and indicative of true phylogenetic relatedness. Using the results from *in-silico* typing, cgMLST has been shown to not only outperform established typing systems in terms of discriminatory power, but maintain high concordance with groupings created by these methods, allowing direct comparisons with legacy results.

While the selection of isolates from the Canadian CGF database based on CGF profile and epidemiologic metadata proved to to be an effective way of comparing the genomic and epidemiologic relationships between strains of *C. jejuni*, there were instances where the genomic similarity assessed by cgMLST did not agree with the epidemiology. The cattle-

associated isolate, CE_M_10_2113, for example, exhibited 49 cgMLST loci mismatches with an isolate sampled under very close temporal, spatial and source conditions. When compared to the isolate CE_M_09_3054, sampled under similar source and geospatial conditions, but one year apart, only 21 cgMLST differences were observed (Figure 2.3(C)). By contrast, the temporal separation between the sampling dates of isolates CI-3252, CI-3609, and CI-5034 appeared to have little effect on differences observed in the cgMLST results of these three water-derived isolates (Figure 2.3(B)).

Performing these analyses in an *ad-hoc* manner requires a substantial amount of effort, and often, the results are not clear as to the strength of relationships present between the genomics and epidemiology of bacterial isolates. By assessing a collection of *C. jejuni* genomes at two clustering thresholds of cgMLST, isolate relationships could be assessed as part of a large inclusive cluster, as well as smaller, highly refined clusters. A more systematic approach, rather than *ad-hoc* analyses, assessing incremental cgMLST clustering thresholds from 100% to 1%, allows for a more comprehensive understanding of the relationship between the genomics and epidemiology of these isolates, and is explored further in Chapter 3.

Chapter 3

Quantitative Epidemiology for Assessing the Concordance between Epidemiologic and Genomic Similarities of WGS data from *C. jejuni*: Towards Improved Application of Genomic Epidemiology in Public Health

3.1 Preamble

Following the use of whole genome sequencing (WGS) of *C. jejuni* for a pilot study in Chapter 2, I established that WGS provides high-resolution genomic data that enables the simulation of various established typing methods (e.g. CGF, MLST) and varying levels of thresholding (e.g. cgMLST_90%) to facilitate assessment of genetic and epidemiologic relationships between isolates. The ability to adjust between thresholding levels of high-resolution typing data allows for assessments that can be highly discriminatory (e.g. cgMLST_100%), showing only the strongest genetic linkages, or highly inclusive (e.g. cgMLST_80%), producing large networks of isolates for studying inter-strain epidemiological relationships. For smaller datasets, e.g. approximately 30 strains, I used multiple thresholding levels to investigate the epidemiologic and genetic relatedness between members included in the same cluster in an *ad hoc* fashion.

In order to facilitate analyses of the genomic epidemiology of *C. jejuni* for use with larger datasets, a means of rapidly summarizing the general epidemiology of strains is required. To this end, I propose that a quantitative epidemiologic similarity metric measuring the magnitude of similarity between any two bacterial isolates can be developed based on fundamental epidemiologic metrics alone. These statistics include the source of isolation, the date on which the bacterial isolate was derived, and the location of sampling. By applying a standardized model incorporating these basic data, qualitative statistics can be transformed into measurable quantitative epidemiologic similarities, in turn allowing for

the clustering of isolates based solely on their epidemiologic metadata. Using the quantitative epidemiological summaries derived by this system, it will then becomes possible to assess the concordance of genomic strain relationships with relationships based on epidemiological data.

3.2 Methods

3.2.1 Strain Selection for Whole Genome Sequencing

All isolates used in this study were selected from the Canadian Comparative Genomic Fingerprint (CGF) database (comprising approximately 20,000 isolates) and pre-screened by comparison of their CGF fingerprints (Taboada et al., 2012) and relevant epidemiological sampling data. Guidelines to investigation were developed such that the relationships between strains of *C. jejuni* selected for sequencing could be used to investigate at least one of the following criterion: (1) population structure, (2) epidemiological relationships, (3) concurrent type-matched strains, (4) temporal distribution, (5) source attribution. On average, micro clusters of 3-4 strains were chosen based on their CGF fingerprints such that the groups satisfied at least one of the above guideline criteria. In total, 298 isolates of *C. jejuni* were selected for sequencing. A subset containing 139 of the total selected isolates were collected as part of the FoodNet Canada Enteric Disease Surveillance Network (formerly C-Enternet) for the years of 2005-2010, and supplemented with 159 isolates collected as part of local initiatives from Southern Alberta, British Columbia and New Brunswick, Canada. Sample collection procedures for isolates collected via FoodNet Canada can be found at http://www.phac-aspc.gc.ca/foodnetcanada/niedsp10-pnisme10/index-eng.php.

3.2.2 DNA Extraction and Sequencing

Isolates selected for analysis were recovered from archival glycerol stocks (60% glycerol in PBS stored at -80°C). Stocks were streaked for isolation onto modified cefoperazone charcoal deoxycholate agar (mCCDA, Oxoid CM0739, with selective supplement SR0155E). Cultures were incubated for 24-48 hours in a tri-gas microaerobic environment

(MAE, 10% CO_2, 5% O_2, 85% N_2) at 42°C. Single colonies were selected and spread to blood agar plates (BBL Blood Agar base, BD 211037, 5% sheep blood) and incubated overnight under MAE prior to harvesting biomass. Genomic DNA extractions were performed using the QIAgen genomic tip 20/G kit according to the manufacturers recommendations. Quantity and integrity of genomic DNA were assessed using the Quant-IT HS fluorometric assay (Life Technologies Q-33120) and via gel electrophoresis on 0.8% agarose, respectively; samples with poor DNA yield or with partially degraded DNA were re-extracted. CGF subtypes of all isolates were confirmed post-extraction as a quality control/assurance step by performing CGF analysis on the extracted genomic DNA used for whole genome sequencing (Taboada et al., 2012).

Paired End Tagged (PET) sequencing libraries were generated at the BC Cancer Agency Genome Sciences Centre (Vancouver, Canada) and WGS data was obtained using the Illumina HiSeq platform (100 bp PET reads). Eighty-three isolates were run per indexed sequencing lane (two lanes total) yielding, on average, 375-fold coverage per isolate. PET Libraries for the remaining isolates were prepared at the National Microbiology Laboratory (Winnipeg, Manitoba), and sequenced using the Illumina MiSeq sequencer, pooling approximately 30 strains per run for coverage of approximately 80-100 fold per isolate.

Draft genome assemblies from both the Illumina HiSeq and MiSeq runs were generated *de-novo* using the St. Petersburg Academy genome assembler (SPAdes) (Bankevich et al., 2012) using a hash length of 55. Four of the 298 genomes sequenced did not pass assembly quality requirements and were thus removed from the dataset, yielding 294 draft genomes available for *in-silico* typing.

3.2.3 *In-silico* Typing of Draft Genome Assemblies

In-silico cgMLST typing results were generated for the Canadian *C. jejuni* dataset as described in Chapter 2. Briefly, the collection of 294 draft genome assemblies were subjected to analysis using MIST (Kruczkiewicz et al., 2013), with the intent of establish-

ing a subset of genomes with complete sequence data (i.e. no contig truncations) for all loci queried (Barker D, *personal communication*). The final collection of cgMLST typed genomes comprised a total of 729 loci that exhibited no sequence truncations in 274 of our *Campylobacter* draft genomes; this became the final dataset.

3.2.4 Data Analysis

All calculations were performed using the R language for statistical computing (R Core Team, 2015). Geographical positioning system (GPS) coordinates were derived by entering sampling site information into Google Maps (available at https://www.google.com/maps) and recording the location data. Distances between GPS coordinates were calculated using the Haversine formula available in the R package fossil (Vavrek, 2015). Pairwise matrices of temporal and geospatial distances were generated in R using the base *dist* function and the euclidean distance calculation metric. A log10 correction was applied to both temporal and geospatial distances. Heatmap analyses were generated in R using the package Gplots (Warnes et al., 2015) and applying the single-linkage clustering function. Colour scales for graphical images were derived using the R package RColorBrewer (Neuwirth, 2015). Detailed colour-scales and frequency histograms for Figures 3.3-3.5 can be found in Appendix B. A repository containing scripts and R code used for analyses and figures are available online: https://github.com/hetmanb/thesis.git.

3.3 Results and Discussion

3.3.1 Development of the Model Framework

In the study of disease occurrence, a simple and popularly-accepted model is the *epidemiological triad*, which defines three factors whose interaction results in the causation of disease: host, agent, and environment (Dicker et al., 2006). Host refers to the human with the appropriate susceptibility to contract disease; the agent, in the study of infectious diseases, is the pathogen responsible for causing illness; the environment brings the host and agent together, allowing for the opportunity of exposure via a multitude of pathways. An

advantage of focusing on the surveillance of bacterial infectious disease for public health is that we can predetermined that the agent, as a human pathogen, when in contact with the host in sufficient numbers, will typically cause disease. Using this assumption, we can then ignore the factors relating to susceptibility of host and instead focus on the agent and how it circulates within the environment.

Where the interaction of agent and environment occurs, in the prelude to interacting with the target host, is what we may consider to be the *epidemiologic source* (Figure 3.1); this can relate to any number of items including water, soil and vegetation, insects, wild and domestic animals, food production from farm to fork, and even other human hosts acting as carriers of the disease agent. Sources can further be differentiated from one another by the addition of time and location data pertaining to the instance of sampling, in this way, we can define not only the vector of a human pathogen, but also pinpoint geographic spaces suspected in attributing disease to specific environmental and physical conditions, and narrow temporal ranges where disease occurrence was particularly high, indicative of potential outbreaks that were previously unaccounted for.

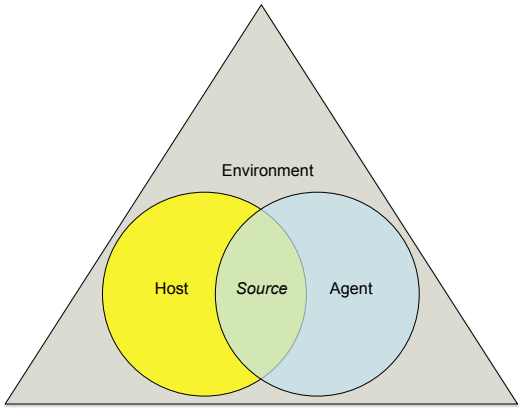

Figure 3.1: A depiction of the *Epidemiological Triad* adapted from (Dicker et al., 2006)

By identifying the source from which a bacterial isolate was sampled, the time or date of isolation, and the geography of the source sample, we arrive at three common, measurable metrics for assessing the epidemiologic similarity of any two bacterial isolates sampled in a surveillance setting. Further, a combination of these three metrics provides a unique statistic for describing the epidemiology of any sampled bacterial isolate; much like a sequence of genetic features can be merged to create a strain genotype in molecular epidemiology,

our aim here is to use a sequence of descriptive epidemiologic factors to describe the summarized epidemiology of an isolate, allowing for a quantitative epidemiologic strain-strain comparative summary. The "epi-type" of a bacterial isolate can thus be described by the following formula:

$$\varepsilon_{sym} = source + geospatial + temporal \tag{3.1}$$

Certain caveats, however, may be considered when assessing the bacterial epidemiologic type using the above equation; namely, the weight given to each factor used in the calculation should not necessarily be equal. Depending on the collection methods being used and the circumstances surrounding the sample collection, certain aspects of the data may be less reliable than others. For example, bacterial isolates sampled from the cloaca of migratory birds may have highly refined source and temporal factors included in their sampling records, however, measuring the geospatial component may be less reliable due to the potential distance travelled by migratory fowl in a very short time. In this instance it may prove more accurate to assign increased importance to the source and temporal variables of the calculation, while assigning less consequence to the geospatial variable. A further consideration may include knowledge about the agent being tested. A bacterial species known to be highly source-restricted may then require higher weight on the source variable compared to the geospatial and temporal counterparts, to account for increased importance when observing a change in the source. To account for these types of variation in the overall calculation of epidemiologic type, we propose a formula adapted from Equation 1 where the variables σ, γ, and τ are adjustable coefficients for re-assigning weights based on *a priori* considerations like those mentioned above:

$$\varepsilon_{sym} = \sigma(source) + \gamma(geospatial) + \tau(temporal) \tag{3.2}$$

3.3.2 Defining Components of the Model

While geospatial and temporal data are easily converted to a systematic numeric format (e.g. GPS coordinates, POSIX timestamp), source information is inherently more complex to quantify and to our knowledge, no system currently exists for measuring the likeliness of one source compared to another. Approaches at using the genetic likeliness of sources may provide a basis for similarity when assessing plant or animal hosts; however, when comparing water or soil samples, this method loses its effectiveness. Instead, since this model is aimed at measuring the epidemiologic similarities of bacterial isolates, we chose to employ categories commonly used in describing the epidemiology of enteric pathogens (Harding et al., 2014). To this end, sources were redefined as fitting to animal, human or environmental association, and then further differentiated based on subsequent epidemiologic attributes pertaining to each parental group. In essence, a line-list was created containing all the non-redundant sources in the dataset as the sample input, with descriptive epidemiologic attributes acting as the informative elements of the questionnaire. Each source exemplar was then scored independently across all epidemiologic categories with three possible outcomes for each attribute: (1) strong association; (0) little to no association; and (*) partial or potential association. In replacing each individual, descriptive source with a series of categorical epidemiologic scores, we effectively reduced all qualitative source descriptors into a consistent set of comparable, quantitative fingerprints. Once every source in the dataset was scored, it became possible to compare the outcome of the rubric in a pairwise manner; for combination of source pairs, the sequence of categorical scores were assessed in an allelic fashion, resulting in a match, mismatch, or partial match. At the end of each pairwise comparison, the score from matching epidemiologic attributes are summarized and assessed as a proportion of the total attributes examined. Thus, the source statistic from Equation 3.2 becomes:

$$\sigma(source) = \sigma\left(\sum_{i,j=1}^{n} (i+j)\frac{1}{n}\right) \tag{3.3}$$

where (i, j) are the scores from each source in the pairwise comparison starting with attribute 1 through to the number (n) attributes, $(i + j)$ is the value given to the resulting pairwise match, mismatch, or partial match from the comparison of the attribute across the two sources, and (n) is the total number of attributes being examined. Using this procedure, we are consequently able to assign a pairwise similarity to any two bacterial isolates based on their descriptive epidemiologic source attributes alone.

As mentioned above, temporal and geospatial data are readily comparable by measuring the sampling date information as a function of POSIX-time, defined as the number of seconds elapsed since January 01, 1970, and converting geographic location into a global positioning system (GPS) coordinate derived from the available sampling site geographic data. To calculate the relative temporal similarities of isolates in a dataset, the individual pairwise Euclidean distances are calculated based on the day of isolation of each isolate, then treated as a proportion of the largest distance in the dataset. A similar treatment is performed on the geospatial data, where Euclidean pairwise distances in km are calculated between each pairing of isolates, and compared to the largest distance in the dataset.

In order to account for diminishing returns on isolate similarities when geospatial or temporal distances are high (French et al., 2005), we chose to apply a logarithmic correction to the distribution of these data in the dataset. For example, when comparing the geospatial or temporal distance between two isolates, we propose that the closer they are to the same isolation place or date; the probability that they share a high similarity to one another becomes especially high. A distance of 3000 km separating isolates from Southern Alberta and Eastern Ontario likely bears the same dissimilarity as comparing isolates from Southern Alberta to New Brunswick, a distance approximating 4000 km. However, when comparing two isolates from the same sampling site to two isolates separated by a distance of 1000 km, the difference in geospatial distribution of 1000km should be treated with higher significance than the cross-country comparison. The logarithmic correction, therefore, is applied to shape the distribution of the resulting similarity values such that

they provide a greater significance to isolates of closer geographic distance. To illustrate the use of a logarithmic correction with regards to temporal data, we can apply a similar example as used with the geospatial reasoning: two isolates sampled at dates separated by six years likely provide no less epidemiologic strength to the investigation of similarity than two isolates separated by a three-year sampling period; in other words, both pairs of isolates separated by 6 years, and 3 years, respectively, are treated as approximately equal in dissimilarity. If we consider, however, a comparison of two isolates sampled on the same day versus two isolates sampled three years apart, the same three-year difference in sampling periods exists, but conceptually, we should expect the isolates from the single day of separation to be much more similar based on temporal information alone. Thus, applying a logarithmic correction to the distribution of temporal distances increases the weight of the similarity metric given to isolates that are separated by mere days and weeks, as opposed to years.

The formulae for calculating temporal and geospatial similarities are presented below.

$$\tau(temporal) = (\tau) log\left(\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}\right) \tag{3.4}$$

Where (x, y) represent the date of isolation of each pairing of isolates, in POSIX-time rounded to the nearest day.

$$\gamma(geospatial) = \gamma(log(dist_{ab})) \tag{3.5}$$

Where $(dist_{ab})$ is the physical distance, in km of each pairwise set of isolates in the dataset, calculated using the Haversine formula for deriving great-circle spherical distances from latitude and longitude coordinates (see Appendix B). Substituting Equations 3.3 to 3.5 into Equation 3.2 yields our final model for summarizing the basic epidemiologic similarity

between any two bacterial isolates, and is presented in Equation 3.6:

$$\varepsilon_{sym} = \sigma\left(\sum_{i,j=1}^{n}(i+j)\frac{1}{n}\right) + \tau\left(log\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}\right) + \gamma(log(dist_{ab})) \tag{3.6}$$

3.3.3 Applying Source Similarities to Isolates of *C. jejuni*

In this investigation, we have chosen to focus on the circulation of *C. jejuni* throughout Canadian rural, urban and clinical environments. As a high-priority foodborne pathogen, surveillance of *C. jejuni* has been largely focused on the environment surrounding the production and distribution of food: farm animals and their retail products, including poultry, pork and beef, soil and water samples from agricultural lands and watersheds, and even non-agricultural sources such as deer, migratory birds, and raccoons that have the potential to act as a reservoir or vector for *C. jejuni* throughout the Canadian ecosystem. The multitude of sources related to pathogen transmission and survival makes *C. jejuni* an excellent, if complex, model organism for an investigation aimed at identifying relationships based on descriptive epidemiological metrics alone.

Rather than try to assess the vast number of potential sources of *C. jejuni* from the environment to build a network of source similarities, we instead focused on establishing a foundation using only available source data from the Canadian CGF database. The database contains carefully curated metadata on over 20,000 *Campylobacter* isolates, and is managed by only a few curators, thus the granularity of the available source data has been kept largely consistent; this makes the process of identifying sources much more efficient than attempting to parse the same type of information from a database that contains data from many different contributors. The source data from the CGF database has been divided into three main categories, 1) Animal, 2) Environmental and 3) Human sources. These sources are then further refined into urban or rural associations (e.g. farm or retail location), food or non-food (e.g. farm or companion animal), sample medium (e.g. meat or faecal sample), and finally any specific information pertaining to the species of animal sampled (e.g.

chicken carcass, geese cloacal swab) or type of environmental location (e.g. recreational beach sand, irrigation canal). Table 3.1 lists the non-redundant source identifiers from the Canadian CGF database, assessed at four layers of granularity.

In order to develop a rubric to assign similarity scorings to the sources in Table 3.1, we began by employing a decision-tree schematic, starting with the level 1 source categories consisting of Animal, Human and Environmental source attributes. We attempted to define categories that were of epidemiologic relevance to the attribution of *C. jejuni* in the environment; namely, risk factors for *C. jejuni* often include consumption of poultry and contaminated food products, occupational hazards of working in either retail food handling or rural agricultural environments, and exposures to contaminated agricultural and recreational waters.

Figure 3.2 summarizes the core epidemiologic attributes and their subsequent characteristics that were used to develop an epidemiologic rubric in establishing source similarities.

After developing the line-list containing all non-redundant sources, each source was then scored against the epidemiologic attributes included in the rubric using three possible score outcomes defined earlier: (1) strong association, (0) little to no association and (*) partial association. Pairwise summary scores were then calculated using penalties that were set in relation to a full (1-1) pairing, which was established as 1.00 for the purposes of this study. For a match consisting of partial association versus little to no association (*-0), a correction factor was applied equivalent to that of 15% of a (1-1) match; for a match consisting of partial association versus strong association (*-1), a 35% correction factor was applied; for a (0-0) match (no association versus no association), a 90% correction factor (or 10% penalty) was applied in an effort to promote clustering based on the presence of similar attributes, rather than clustering based on the absence of the same attributes. Source-source pairwise scores were calculated based on each individual epidemiologic attribute in the line-list, and a summary similarity was derived for each pairing by the addition of each individual score, divided by the total possible maximum score.

Table 3.1: Non-redundant combinations of source categories populated from the Canadian CGF database.

Sample Type 1	Sample Type 2	Sample Type 3	Sample Type 4
Animal	Avian	Abattoir	Chicken
Animal	Avian	Faecal	Chicken
Animal	Avian	Retail	Chicken
Animal	Avian	Faecal	Duck
Animal	Avian	Faecal	Goose
Animal	Avian	Faecal	Pelican
Animal	Avian	Faecal	Seagull
Animal	Avian	Faecal	Sparrow
Animal	Avian	Faecal	Turkey
Animal	Avian	Retail	Turkey
Animal	Avian	Faecal	Wild-bird
Animal	Companion	Faecal	Cat
Animal	Companion	Faecal	Dog
Animal	Equine	Faecal	Donkey
Animal	Equine	Faecal	Horse
Animal	Miscellaneous	Faecal	Llama/Alpaca
Animal	Miscellaneous	Faecal	Peromyscus
Animal	Miscellaneous	Faecal	Raccoon
Animal	Miscellaneous	Faecal	Rattus
Animal	Miscellaneous	Faecal	Skunk
Animal	Miscellaneous	Faecal	Small-mammal
Animal	Porcine	Faecal	Pig
Animal	Porcine	Retail	Pig
Animal	Ruminant	Faecal	Buffalo
Animal	Ruminant	Abattoir	Cow
Animal	Ruminant	Faecal	Cow
Animal	Ruminant	Retail	Cow
Animal	Ruminant	Faecal	Deer
Animal	Ruminant	Faecal	Goat
Animal	Ruminant	Faecal	Sheep
Animal	Ruminant	Retail	Sheep
Environmental	Rural	Water	Lagoon
Environmental	Urban	Soil	Sand
Environmental	Urban	Water	Sewage
Environmental	Urban	Water	Water
Human	Clinical	Human	-

A heatmap depicting the similarities derived from the pairwise source analysis is shown in Figure 3.3, with darker colours indicating stronger relationships. Clusters of high similarity are outlined in black and annotated as clusters 1-6. From the histogram in Figure 3.3, a wide range of source similarities is seen to exist in the resulting pairwise matrix, with the minimum similarities between two sources equalling approximately 20%, and the highest non-self pairings approaching 90% (See Figure 7.1).

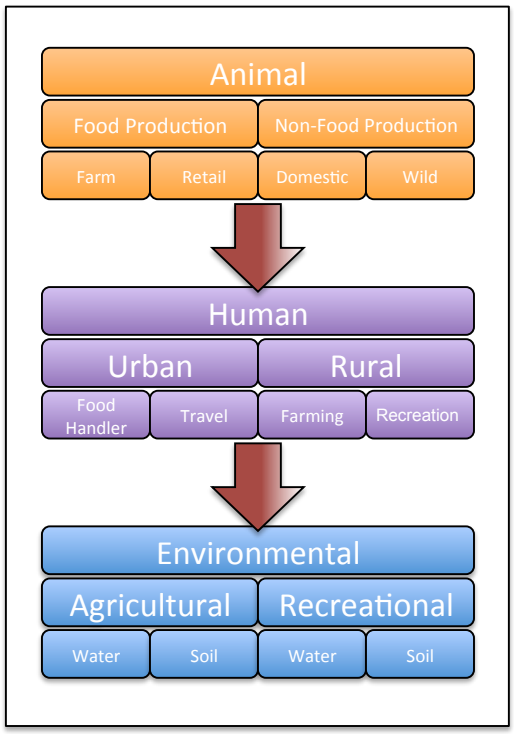

Figure 3.2: Core epidemiologic guidelines for scoring sources found in the Canadian CGF database. Each of the three major source attributes (Animal, Human, Environmental) are further broken down into minor categories.

In Cluster 1, human clinical sources are highlighted, and high similarities can be seen between Cluster 1 and Cluster 2; and Cluster 1 and Cluster 6, which represent sources from retail and farm origins, respectively. As people have increased contact with the retail food production system, as well as the agricultural industry, the strong associations seen between human and retail and farm clusters are expected. Furthermore, these pockets of high similarity are concordant with epidemiologic studies which indicate that occupations both in the agricultural and food-handling industries are significant risk factors for exposure to *C. jejuni*; as well, companion animals such as dogs and cats have been shown to be a potential vector of contamination to human beings (Friedman et al., 2004).

Cluster 4 is representative of animals associated with remote, non-agricultural environments. This cluster bears limited similarity to the clinical human source cluster, as interaction between people and these wilderness-associated animals occurs more rarely than

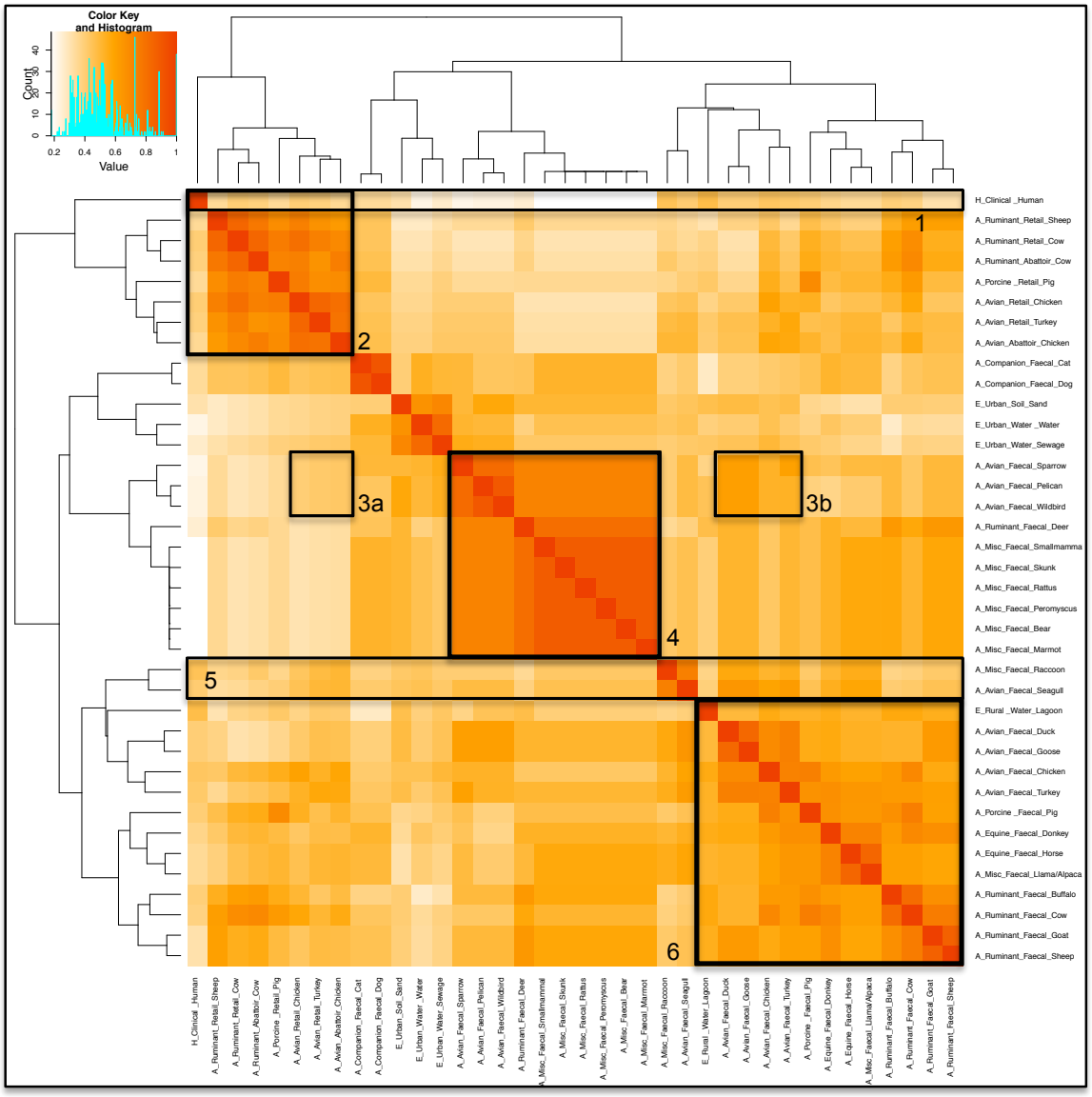

Figure 3.3: Graphical heat representation of the similarity of source metadata from non-redundant source metadata collection in the Canadian CGF database. Darker colouring indicates stronger association between subjects. A histogram portraying the colour key and the counts of isolates corresponding to the range of similarity scores is displayed top left. Clustering of the data was performed using the complete linkage algorithm.

between people and animal sources associated with domestic or agricultural environments. Clusters 3a and 3b highlight the similarity derived from shared avian characteristics, and demonstrate the epidemiological hierarchy achieved by the rubric. Though the three sources highlighted (Sparrow, Pelican, Wild Bird) all share common avian attributes with Turkey,

Chicken, Goose and Duck sources, the latter birds are clustered more strongly by their retail and agricultural related characteristics, which differ from their wilder counterparts. It follows that wild and water birds generally have limited contact with humans, and are thus separated from their agricultural cousins by their lower epidemiologic value to human-related environments.

Following the framework of epidemiologic attributes in the scoring rubric, the Raccoon and Seagull sources (Cluster 5) are highly similar to one another, even though biologically, they might be considered highly dissimilar. Raccoons are found widespread throughout rural and urban environments; they exhibit scavenger lifestyles and come into contact with many avian, land-dwelling and aquatic sources, providing an excellent vector for the potential transmission of *C. jejuni* (Lee et al., 2011). In an epidemiological context, seagulls exhibit similar behaviour as raccoons, surviving in a wide range of climates and ecosystems and interacting with a multitude of source vectors for *C. jejuni*. Again, the epidemiologic value of the potential interactions of these animals outweighs their biologic context, placing raccoon and seagull sources in a highly similar cluster of source similarity, even though biologically, they remain largely distant. Of note, the pervasive nature of scavenger type animals makes both seagulls and raccoons collectively unique in Figure 3.3, highlighted by their moderately strong similarity to all other sources shown, based on their high potential for association with the many different epidemiologic attributes being considered in this study.

3.3.4 Combining Source Similarities with Geospatial and Temporal Components

In order to derive the total epidemiologic similarities of strains from an investigation of the genomic epidemiology of Canadian *C. jejuni*, we subjected the sample metadata from 274 isolates of *C. jejuni* to the model in Equation 3.6. In an attempt to reflect the reliability of metadata obtained for the current dataset, we used coefficient values of 0.4, 0.4, and 0.2 for the values of σ, τ, and γ, respectively. To resolve these ratios of source and temporal

data compared with geospatial data, we first proposed an equal split between the three epidemiologic categories, with each data type comprising one-third of the final epidemiologic similarity. However, when assessed separately, the geospatial component of the dataset had several incomplete data points; e.g. while complete provincial data could always be assessed, city data was found lacking in 63 entries, resulting in 23% missing data. Rather than eliminate these samples from the analysis entirely, we assessed these locations as a generic provincial location based on Google Maps GPS data, (e.g. Ontario, Canada). Since the source and temporal components of the epidemiologic similarity calculation had complete data associated with each strain, we thus determined that they would each comprise equal strength in the calculation, while contributing more than the geospatial component.

Epidemiological groupings of 274 isolates of *C. jejuni* from across Canada are demonstrated in the heatmap shown in Figure 3.4 as a result of comparing each isolate on the basis of the epidemiologic metadata using the model described in Equation 3.6. A detailed version of the histogram presented in Figure 3.4 can be found in the supplementary Figure 7.2. Three major clades separate the isolates into their dominant epidemiologic categories; namely, (1) "Clinical", (2) "Animal" and (3) "Environmental" clusters. The sub-clusters defined by highly specific source, geospatial and temporal components accurately group isolates together that possess strong relationships at a minimum of two of the three possible categories used in the model. Even with a preference given to the source and temporal components of the model, the sub-clusters depicted in Figure 3.4 provide firm associations based on geospatial data. Of 16 sub-clusters listed in Table 3.2, only a single group (Cluster P) combines isolates sampled from more than a single province. If we examine the contents of Cluster P, we see that all isolates are derived from Environmental Waters, and are from a narrow temporal range spanning only two years (2006-2007); the strength of association based on source and temporal attributes using a coefficient of 0.4 for both σ and τ, causes them to cluster together despite their wide geospatial range.

An assessment of clusters that lie along the diagonal axis in Figure 3.4 provides strong

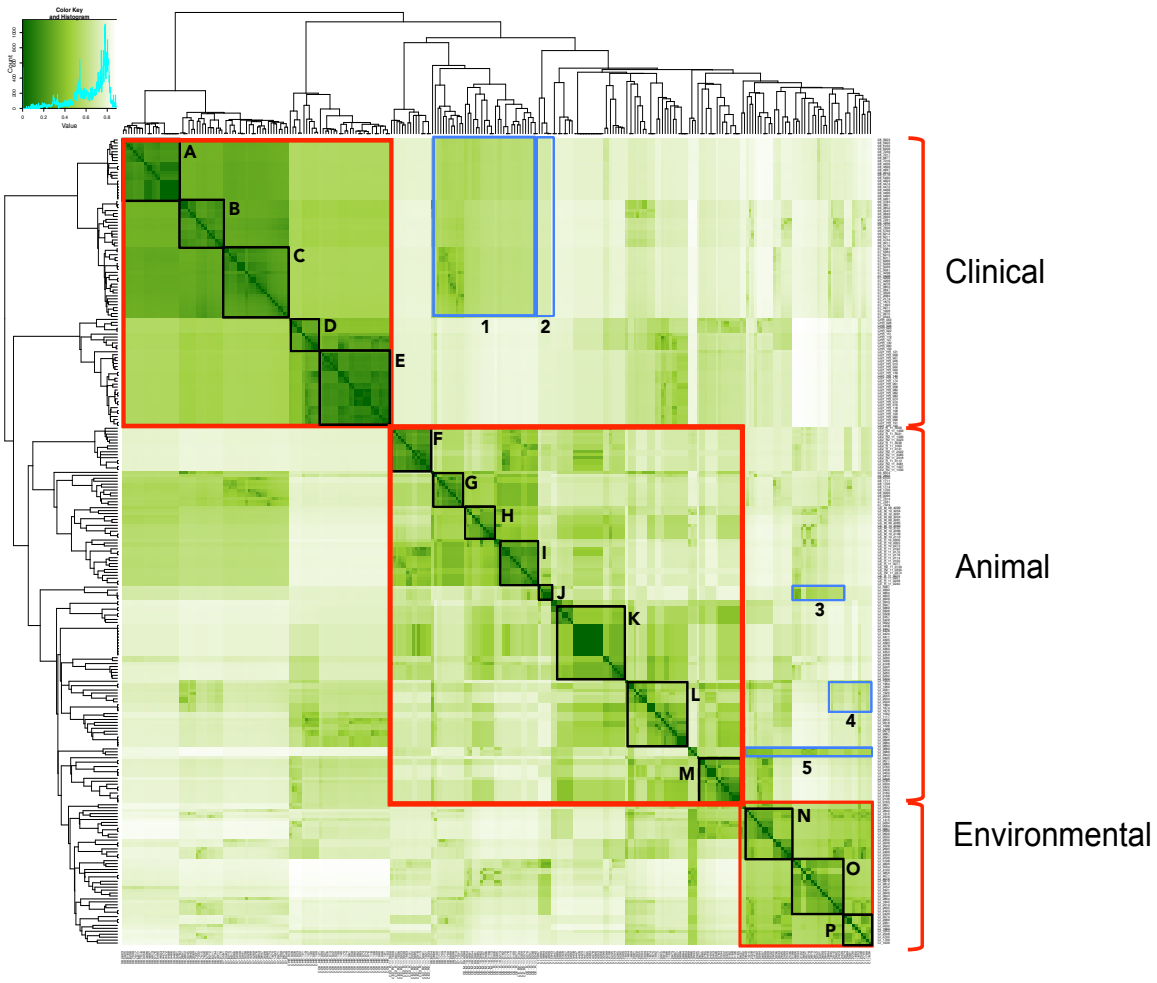

Figure 3.4: Graphical heat representation of the similarity of sequenced isolates of Canadian *C. jejuni* based on a summary of the basic epidemiological metadata calculated using Equation 3.6 and the source similarities represented in Figure 3.3. Source and temporal components each contributed 40% to the final comparison, while the geospatial component contributed only 20% due to less reliability of the data. Stronger colour indicates a higher degree of association. A histogram portraying the colour key and the counts of isolates corresponding to the range of similarity scores is displayed top left. Clustering of the data was performed using the complete linkage algorithm.

indication that the model expressed in Equation 3.6 generates high epidemiologic similarities when the isolates have similar epidemiologic attributes. However, the possibility exists that we could be over-fitting the data to ascribe to these small pockets of high similarity; for our model to be considered as a robust measure of epidemiologic similarity, there needs to be similarity generated not only at these primary sites of comparison, but also at secondary

Table 3.2: Summary of the minor clades annotated in the epidemiological heatmap presented in Figure 3.4. Letters A-P designate minor clusters indicated in Figure 3.4

	Source	Spatial	Temporal
A	Human Clinical	Waterloo, ON	2008
B	Human Clinical	Waterloo, ON	2006
C	Human Clinical	Waterloo, ON	2007
D	Human Clinical	Lethbridge, AB	2004-2005
E	Human Clinical	Calgary, AB	2005
F	Retail Poultry	Chilliwack, BC	2011
G	Retail Meats	Waterloo, ON	2005-2007
H	Faecal Poultry, Bovine	Waterloo, ON	2009-2010
I	Retail Poultry	Waterloo, ON	2010-2011
J	Faecal Raccoon	N/A, ON	2011-2012
K	Faecal Bovine	Lethbridge, Stirling, N/A, AB	2009-2012
L	Faecal Poultry, Bovine, Porcine	Various Locations in Southern AB	2005-2006
M	Faecal Poultry, Bovine	N/A, AB	2004
N	Environmental Water	N/A, Lethbridge, AB	2004-2007
O	Environmental Water	N/A, ON	2006-2011
P	Environmental Water	Sumas River, Salmon River, BC; Grand Falls, NB; Fort Macleod, AB	2006-2007

locations throughout the heatmap where isolates share only an incomplete portion of the same epidemiological characteristics. In other words, we should be able to identify clusters of heat that lie apart from the diagonal axis; these clusters of secondary heat outline epidemiologic comparisons that result in partial similarity generated between samples that share an incomplete set of common attributes used in the model.

In the heatmap presented in Figure 3.4, several clusters of secondary heat are outlined and designated as sub-clusters 1-5; these areas of moderate similarity are identified by regions of strong colour found apart from the diagonal axis that stand out from background levels of low-heat. Sub-cluster 1 shows a strong association between clinical isolates derived from the Waterloo region of Ontario and farm faecal and retail meat isolates sampled from the same region. Both of these groups share similar geography and temporal ranges, resulting in the increased associative heat. Sub-cluster 2 shows increased heat between Wa-

terloo clinical isolates with raccoon isolates sampled from the same region; this association underscores the method used for assessing the strength of source-source relationships. As discussed earlier, raccoons inhabit a wide range of both rural and urban niches, often coming into close contact with human food sources and waste; thus, this pocket of secondary heat is indicative of the urban niches shared by both raccoons and humans. By contrast, sub-cluster 3 depicts the interaction of raccoons with the non-urban environment, portraying moderate heat between raccoon isolates and environmental water sources located in the same geographic region.

The secondary heat depicted in sub-cluster 4 is associated with a strong temporal relationship; a group of faecal farm-animal derived strains sampled in the summer months (May-July) of 2006 show moderate similarity with environmental water-derived isolates sampled from a similar time period (2006-2007), emphasizing the epidemiological relationship established between agricultural isolates and irrigational waters. Finally, sub-cluster 5 highlights the strong source relationship between three water isolates that clustered primarily with animal isolates due to strong geospatial and temporal components of the epidemiologic similarity calculation. While these three samples are located in the midst of the major animal-based clade, by looking at the dendrogram assignment of Figure 3.4, we can see that isolates belonging to this sub-cluster are actually moderately dissimilar to the neighbouring animal isolates, and likely only assign to this parental clade due to incomplete geospatial information associated with the sampling data (Table 3.2).

3.3.5 Comparing Genomic and Epidemiologic Clustering Results

In an attempt to evaluate the overall genetic versus epidemiological clustering results, we subjected the total pairwise similarities from each of the cgMLST and epidemiological approaches to comparison via a rank-based hierarchical analysis. By transforming the pairwise similarity values from each method into a hierarchical rank based on the total number of comparisons, we effectively removed any bias resulting from a comparison of empirical

data across two largely different ranges of values. The resulting ranks were normalized and values from each pairwise evaluation were compared from each method (i.e. epidemiologic rank versus genomic rank). A graphical matrix of the results is shown in Figure 3.5, with isolate-pairs that were ranked similarly ($+/-1$ SD from 0) via both epidemiological similarity and cgMLST sequence similarity shown as white space. Isolate pairs highlighted in green signify pairs of isolates with similarity much higher via their epidemiological attributes compared to their genomic likeliness and conversely, isolate pairings in blue indicate pairs of isolates whose genomic similarity largely outweighed any similarity derived via their epidemiological scoring.

The histogram in Figure 3.5 (top-left), indicates that the majority of genomic to epidemiologic congruence lies within one SD of each other, suggesting at least a moderate degree of agreement between the two clustering methodologies; however, a noteworthy number of isolate clusters exist in the heatmap that pertain to groups that are significantly more similar based on their epidemiology or their genomic profiles alone (See Figure 7.3). These discongruent clusters may be more informative than clusters of high similarity between genomic and epidemiologic profiles, as they indicate isolate groups that are paired together either in their epidemiology or genomics under unexpected conditions. For example, clusters in green represent strains of *C. jejuni* that relate closely to one another based on their epidemiologic profiles, but are distantly related via their genomic signal. These clusters may be suggestive of certain sources, environments, or timespans that support the survival and circulation of many different genotypes of *C. jejuni* concomitantly. The circulation of genetic material throughout and between bacterial populations is considered to be a principal means of adaptation to microbial species, thus, by supporting the circulation of many bacterial genotypes in a single reservoir, the ecologic niches identified here may be promoting exchange between *C. jejuni* genomes, facilitating genetic exchange resulting in enhanced survival and spread of *C. jejuni* throughout other environments.

Clusters highlighted in blue in Figure 3.5 represent *C. jejuni* isolates from our dataset

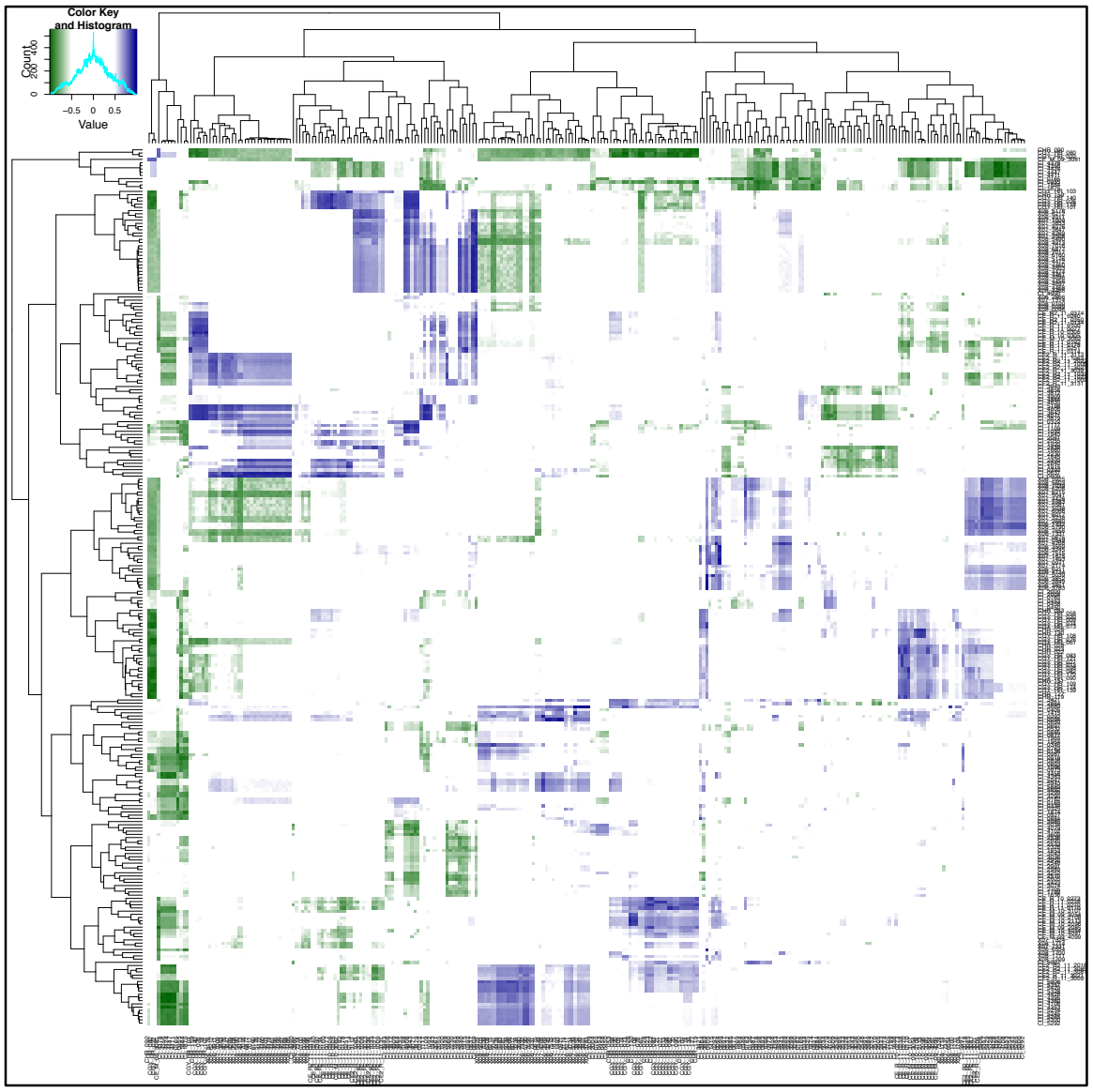

Figure 3.5: Graphical heat representation depicting the rank agreement between comparison similarities derived using genomic (cgMLST) clustering and epidemiologic clustering of 274 isolates of *C. jejuni*. Blue sections indicate stronger genomic similarities between isolate pairs, while those in green indicate stronger epidemiological relationships. White indicates a strong agreement ($+/-1$ SD) between similarities derived using both genomic and epidemiologic methods. A histogram portraying the colour key and the counts of isolates corresponding to the range of similarity scores is displayed top left. Data was clustered in R using the complete linkage algorithm.

that were sampled from disparate ecologic niches, yet share highly similar genomic profiles via cgMLST. The isolates present in these clusters appear in multiple animal and environmental sources from across Canada and persist throughout several years of sampling, and

may represent *C. jejuni* genomes that have enhanced ability for survival throughout a variety of environments. Certain MLST sequence types (ST) of *Campylobacter* have been implicated previously as generalist types (e.g. ST-45, ST-21) that could possess an enhanced ability for survival under a wide range of environmental and host conditions, conferring an increased risk for exposure to the human population (Gripp et al., 2011; Sheppard et al., 2014).

To test if the isolates identified by the clusters in blue corresponded to established generalist lineages, we performed *in-silico* typing using the MIST software package to derive the MLST sequence types (ST) for each isolate used in the study. Then, we selected both the left and right-tails (p = 0.05) from this distribution, pertaining to the isolate pairs that showed highest epidemiologic concordance, and genomic concordance, respectively. We then summarized the MLST ST present in each of the tails. Results from the analysis of frequency distribution of *C. jejuni* ST from each tail are shown in Figure 3.6. Data from the right-tail (blue) correspond to isolate pairs with high genomic and low epidemiologic similarity, while data shown in green represents isolate-pairs with high epidemiologic and low genomic similarity. In the right-tail distribution (shown in blue), ST-45 stands out as the dominant genotype (n = 29) from amongst 16 other ST (n = 39) present. The observation that a single ST from a lineage often associated with generalist behaviour comprises almost half of the ST found in the right tail of the distribution suggests that our rationale for the interpretation of Figure 3.5 is accurate.

To determine if the frequency of ST-45 observed in the right-tail distribution of Figure 3.6 was unique, or an artifact from the sample selection process for this study, we investigated the frequency of ST from the left-tail of the distribution as well. The green bars in Figure 3.6 indicate the frequency of ST among isolates in Figure 3.5 identified as being highly similar epidemiologically, and dissimilar genomically. A similar number of total isolates exist in both left and right tails (n = 65, 68 respectively), but no single ST appears to stand out among the left-tailed ST as observed among the right-tailed distribution; an

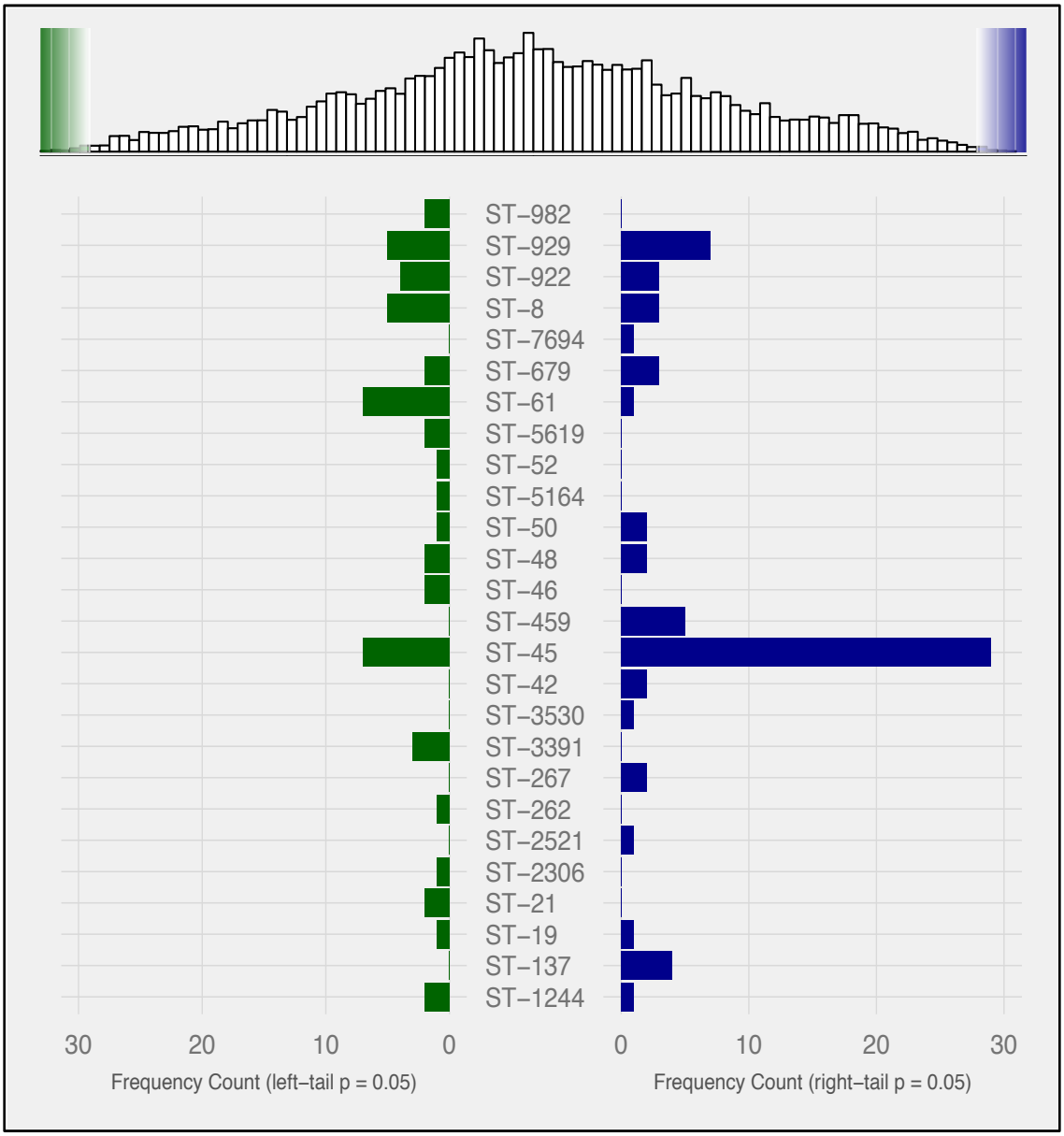

Figure 3.6: (Top) Population distribution of total comparison analysis (n = 75076) presented in Figure 3.5. Left and right tails (p = 0.05) are highlighted in green (n = 68) and blue (n = 65), indicating isolate pairs strongly related epidemiologically, and genomically, respectively. (Bottom) Frequency of ST located within left and right tails (p = 0.05) of the distribution pictured above.

explanation for this seems to follow our logic for examining Figure 3.5, and indeed, seems to be consistent with our hypothesis that we are identifying specific generalist genotypes found throughout a broad range of sampling environments, as well as those specific environmental sinks contributing to the survival of a wide range of *C. jejuni* genotypes.

Isolate clusters with high epidemiologic similarities and vastly disparate genomic signal would seem to be indicative of environmental and source sinks with conditions equally beneficial to both specialist and generalist genotypes of *C. jejuni*. It may be for this reason, that while we do observe the generalist ST-45 amongst the ST present in the left-tail, it appears in similar frequency to the other STs present. Among the right-tail STs, ST-45, a generalist genotype hypothesized to possess genomic features allowing it to survive in a great variety of conditions, dominates the distribution, occurring in much higher frequency than any of the neighbouring ST. While the population size, as well as the sample selection process for the genomes present in this study will inherently introduce ST bias into the proportions of ST present in the results, the significantly different distributions of ST present between the left and right tails of the data presented in Figure 3.5 seem to indicate that measuring the epidemiological similarities, as we have done here, is a useful tool in helping to identify both high-risk sources and environments that propagate a variety of *C. jejuni* subtypes, as well as those generalist subtypes that appear over a long time period in a wide variety of epidemiological sinks.

3.3.6 Assessing Congruence of Epidemiologic and Genomic Data

Assessment of genomic concordance of bacterial isolates with their underlying epidemiology has typically been performed in *ad hoc* analyses, with observations made as to the general epidemiologic characteristics among isolates on a cluster-by-cluster basis; this has been useful in identifying clusters that are host-restricted, widespread in occurrence, or that may be associated with especially pathogenic properties. However, by developing the means to quantify the basic epidemiologic similarity of a sample population of bacterial isolates, we have provided an avenue for direct comparison of the genomic signal of a sample bacterial population with its underlying epidemiology. Further, we can perform a statistical summary on the goodness of fit between the epidemiology of our sample, and genomic typing results, such as those derived from assessment with cgMLST.

Results from both the genomic and epidemiologic clustering methods can be measured at varying thresholds of resolution, allowing us to calibrate the level at which the concordance between the two methods is measured in fine detail. Here, we propose the use of the Adjusted Wallace Coefficient (AW) to measure the congruence between the cluster memberships of the 274 isolates of *C. jejuni* from our dataset based on the epidemiologic similarities derived using Equation 3.6, and genomic similarities generated by *in-silico* typing with cgMLST. The metric provided by the AW describes the probability that any two isolates clustered together in one method will also cluster together using the second method, and provides directionality to the result (i.e. how well method A fits method B, regardless of how well method B fits method A) (Severiano et al., 2011; Wallace, 1983). By measuring the overall fit of clusters of *C. jejuni* isolates with similar genomic profiles generated via cgMLST to clusters established based on quantifying and comparing the epidemiology of the isolates using our model described in Equation 3.6, we hope to assess the efficacy of the cgMLST method to group isolates together based on the underlying epidemiology from which the they were sampled. Figure 3.7 displays the unidirectional AW calculated at multiple threshold levels for the fit of cgMLST to epidemiologic clusters. Thresholds were compared by generating partition memberships at k number of fixed clusters for both cgMLST (k_{cgMLST} = 2-269 clusters) and epidemiologic clustering (k_{epi} = 4-213 clusters); relative threshold percentages of cgMLST are indicated by dotted lines on the figure as a visual aid.

From the results shown in Figure 3.7 it appears that cgMLST best assumes the partitioning of epidemiological clusters when assessed at a high level of cluster thresholding (i.e. $k_{cgMLST} > 200$), and when the epidemiologic-based partitioning is concurrently assessed at low cluster thresholding levels ($k_{epi} < 50$). However, at this wide disparity between method thresholds, there is a large difference in the number of clusters created by each method. As discussed in Chapter 2, when comparing results between highly discriminate methods versus methods with low discriminatory power, there is a high probability that the

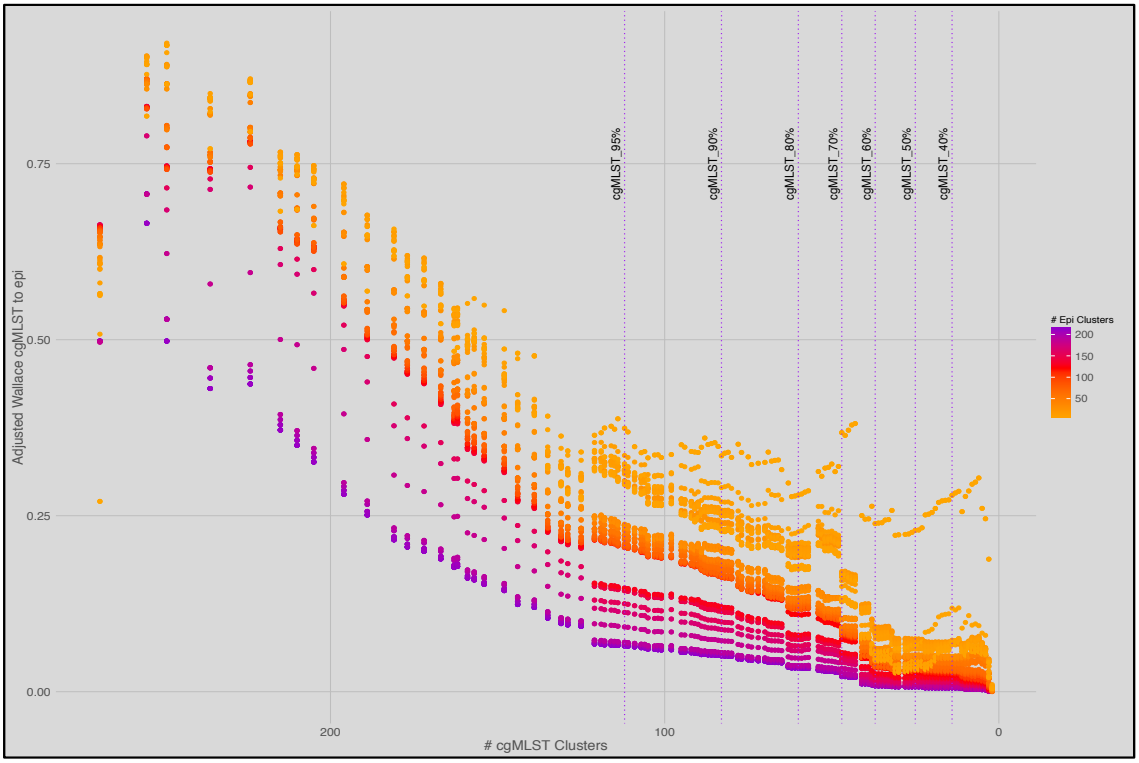

Figure 3.7: Adjusted Wallace (AW) scores for the comparison of clustering thresholds of cgMLST versus epidemiological clustering. AW results shown for directionality of cgMLST $\rightarrow$ epidemiological clustering only. Number of clusters (k) created at each level of thresholding indicated for cgMLST on the horizontal axis, and by the colour-scale indicated in the legend for epidemiologic clustering. Thresholds relative to percentile scores are indicated as vertical dotted lines on the figure.

few, high-similarity strain pairings created in the highly-discriminate analysis will also be formed in the large clusters created by the method at low discriminatory power, establishing agreement between the two methods. Further, as the discriminatory power of a method is increased, singleton clusters are produced, which, as they contain only single isolates, should not be included in the calculation of method agreement. Thus, in the case of fitting high-threshold cgMLST clustering to low-threshold epidemiologic clustering, it is likely that the overall fit of cgMLST to the epidemiologic partitions is forcefully inflated by the contrasting discriminatory powers of the cgMLST and epidemiologic clustering thresholds. The large vertical spread of the results in Figure 3.7 when $k_{cgMLST} > 200$ is therefore the result of many singleton clusters being produced at high levels of discriminatory power by

cgMLST. While high discriminatory power is often desirable, splitting a dataset into a collection of many small and single-member clusters does not allow for assessment of group relatedness, and artificially inflates the calculated AW, as singleton clusters are ignored in the calculation.

To test the influence of over-discrimination inflating the AW results seen in Figure 3.7, we performed an analysis by measuring the clusters generated by each method with respect to the average strength of similarity present between isolates in each cluster. To calculate this statistic, we measured the average pairwise similarity within each cluster by calculating the sum of all pairwise similarities between the isolates within each multi-isolate cluster, and divided by the total number of comparisons. We then multiplied this "intra-cluster cohesion" (ICC) by the number of isolates contained within the cluster, effectively assigning a weighted ICC to clusters based on the size of their isolate-membership. Finally, we computed the mean weighted ICC for all clusters generated at each specific clustering threshold by adding the weighted ICC from all clusters created at the desired threshold, and dividing by the number of isolates belonging to only multi-isolate clusters, producing a *Weighted Global Genomic Cohesion (no singletons)* ($WGGC\_ns$) and *Weighted Global Epidemiologic Cohesion (no singletons)* ($WGEC\_ns$).

To measure the effect of single-isolate clusters (resulting from selecting thresholds that are highly discriminatory), we changed the denominator in the mean weighted ICC calculation to equal the total number of isolates in the dataset including singletons, producing the *Weighted Global Genomic Cohesion (with singletons)* ($WGGC\_ws$) and *Weighted Global Epidemiologic Cohesion (with singletons)* ($WGEC\_ws$). In this way, we effectively penalize clustering thresholds that are overly discriminate, i.e. producing few multi-isolate clusters with high cohesion, and many singleton clusters. By comparing the mean weighted ICC calculated with the inclusion and exclusion of singleton clusters, we directly assess the effect that highly discriminate clustering thresholds have on inflating the congruence estimates provided by the AW statistic shown in Figure 3.7.

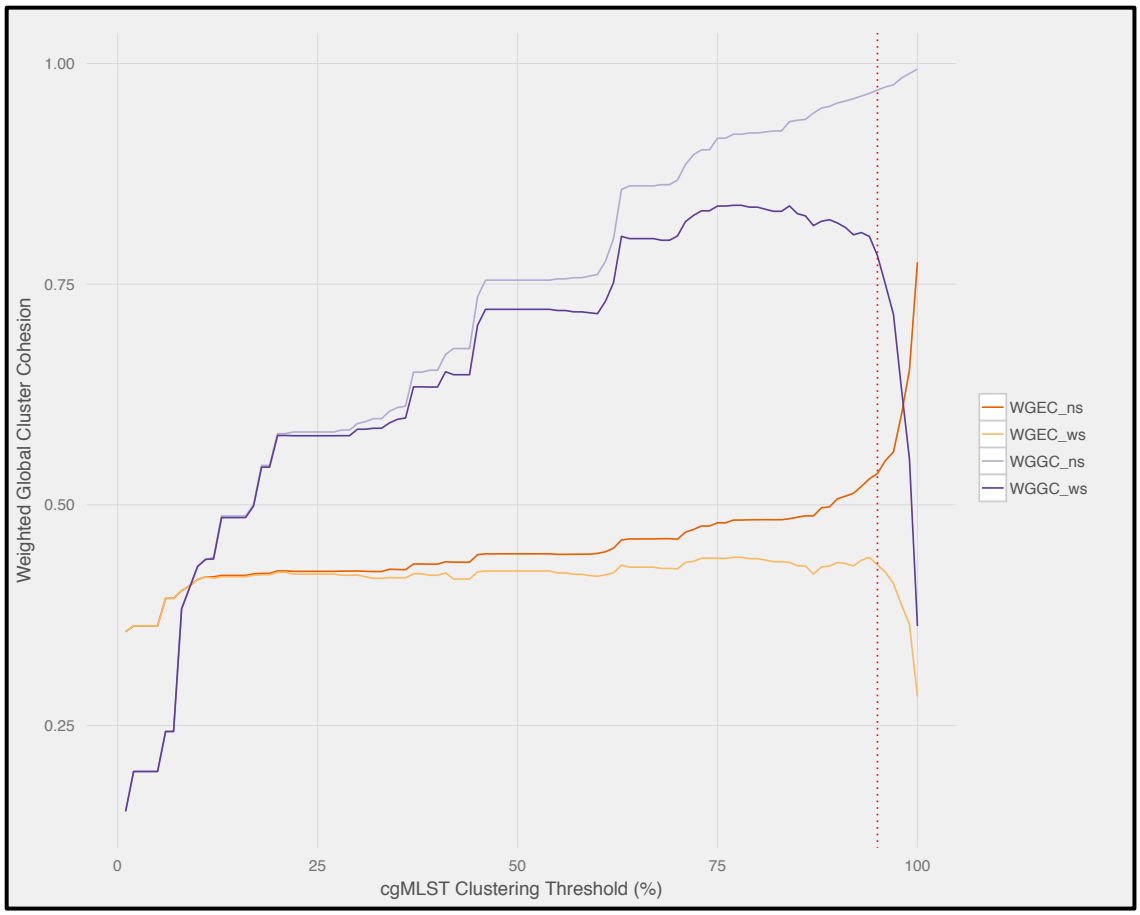

Figure 3.8: Weighted global intra-cluster cohesion for epidemiologic (WGEC) and genomic (WGGC) similarities of isolates within multi-isolate clusters. The exclusion of singletons (_ns) from the calculation is compared to the inclusion of singletons (_ws) for each increasing clustering threshold percent of cgMLST.

The results from the ICC analysis presented in Figure 3.8 support our hypothesis that the AW calculated for the fit of cgMLST to epidemiological clusters was inflated when $k_{cgMLST} > 200$. In both Figures 3.7 and 3.8, there is an extreme change in slope around the 95% cgMLST clustering threshold; by comparing the difference between the inclusion and exclusion of singleton clusters in our calculation of the ICC, we observed that a dramatic change in slope occurs at the point where an increased number of singleton clusters were generated. Thus, rather than selecting a threshold combination with the highest AW for the fit of cgMLST to the epidemiologic clustering, it may instead be more appropriate to choose from a range that represents a compromise between the optimum fit of cgMLST to

epi-clustering and preservation of the integrity of multi-isolate clusters. Thus, to illustrate the congruence of a single threshold of cgMLST and epidemiologic clustering, we have chosen a cgMLST threshold of 95% (k = 94 clusters), representing the highest clustering level of cgMLST before degeneration of multi-isolate clusters into singletons; paired with an epidemiologic clustering threshold of 55% (k = 15 clusters), a close approximation to the number of epidemiologic clusters established in Figure 3.4. A contrasting dendrogram-based comparison of the two clustering methods is presented in Figure 3.9 in the form of a *tanglegram*, with coloured lines drawn to emphasize the specific cluster membership of the individual isolates in each method.

In the tanglegram presented in Figure 3.9, several substructures exist indicating cohesive clusters that connect the two typing methods, indicated by parallel lines between dendrograms. These highly structured matches represent cases where the genotypes are highly specific to the epidemiology, i.e. endemic strains of *C. jejuni* with specific host, geospatial and temporal niches. Two large groups within the current dataset portray cohesive linkages between the epidemiologic and genotypic sub-clustering: a clade of environmental water samples (A), and one containing human clinical isolates (B). The genomic and epidemiologic structure present in these two groups may suggest epidemiologies with decreased genetic exchange; restricted exposure to genotypes of *C. jejuni* may limit the genotypes observed in epidemiologic clusters, or limited survival of *C. jejuni* genotypes within these ecological niches prevents their widespread occurence.

The environmental water derived isolates in Figure 3.9(A) are organized in a way that describes strong similarity between the genomic and epidemiologic clustering methods. While the connecting lines between clusters are not configured with total parallel structure, they are largely limited to a discrete section of each dendrogram, suggesting that relative to the rest of the genomes each dataset, they are considered a separate, cohesive group. The majority of water isolates in the cluster are genomically organized with long-branch terminal leaves, indicating a lack of genomic similarity among the members of the cluster.

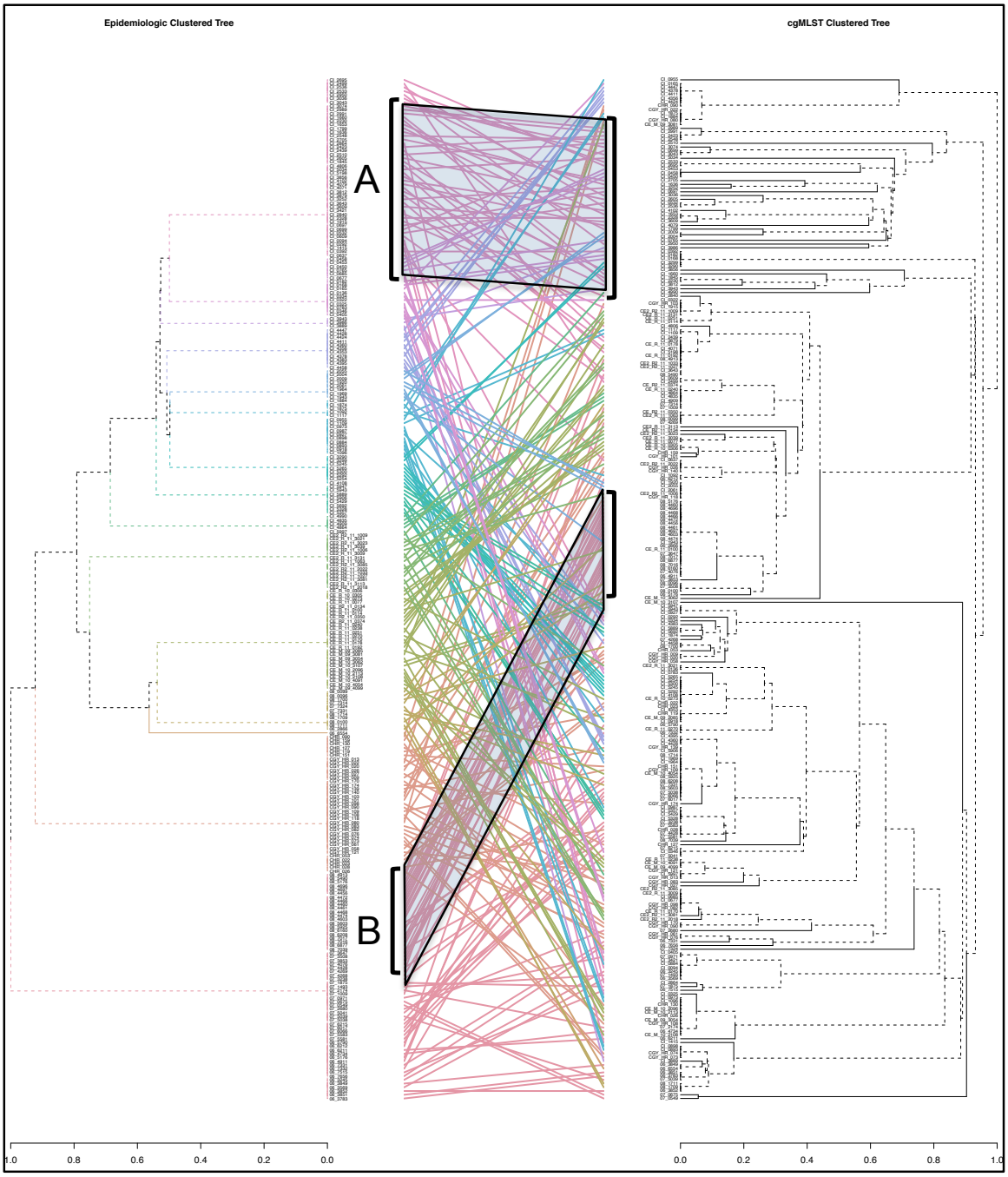

Figure 3.9: An opposing-dendrogram (*tanglegram*) analysis illustrating the degree of concordance between clustering of *C. jejuni* isolates using cgMLST at 95% and epidemiologic clustering at 55%. Connecting coloured lines demonstrate the relative positions of isolates on each dendrogram.

This observation suggests that these genomes may cluster together due to their dissimilarity with the rest of the dataset, as opposed to their inherent similarities toward one an-

other. *Campylobacter* infections follow seasonal trends indicative of a waterborne disease, however, there is still a significant knowledge gap in the epidemiology and genomics of waterborne campylobacters (Jones, 2001). Agricultural watersheds have many inputs, including waterfowl, faecal contaminants from nearby farming ecosystems and wildlife; the complexities present in the genotypic clustering of water isolates may reflect this complex collection of *Campylobacter* inputs into the environmental water reservoirs (Khan et al., 2014; Lee et al., 2011; Pitkänen, 2013).

The human clinical isolates in group B were derived from a set of *C. jejuni* identified from an outbreak of campylobacteriosis from Ontario in 2008 (Clark et al., 2012). These isolates were derived from a single summer camp during a short time span and were indistinguishable by various traditional molecular epidemiologic methods. When assessed by high resolution genotyping using CGF, they were shown to be highly related genotypically; the cgMLST data presented here confirms these observations. As these isolates are strongly linked epidemiologically (i.e. identified as a classical point-source outbreak) as well as genomically, they thus are connected by strong parallelism in connecting lines in the tanglegram analysis of Figure 3.9 (B).

While a global assessment of the organization of strains by the two methods in Figure 3.9 may appear to indicate poor concordance between the genomic clustering of cgMLST and our epidemiological clustering algorithm described in Equation 3.6, when assessed on a cluster-by-cluster basis, there does appear to be a high level of agreement in the membership of smaller clusters obtained using each method. The entanglement between connecting lines in Figure 3.9 is caused by (a) genomic clusters of isolates derived under several different sampling conditions; and (b) epidemiologic clusters that permit the survival of several genotypes of *C. jejuni*. Indeed, by dissecting the epidemiologic cluster from which the isolates highlighted in group (B) belong to, we can see that several genotypes exist pertaining to this epidemiologic grouping, which is not a surprising result. This type of entanglement suggests that it is possible for several different genotypes to circulate within identified epi-

demiologic clusters, and likewise, it is probable that highly similar genotypes may have several epidemiologic niches in which they can survive.

Though inconvenient for comparing the structure of these two methodologies, the flow of *Campylobacter* genotypes throughout a mixed assortment of epidemiologic sources likely reflects in part the success of *C. jejuni* as a human pathogen. In order to see a direct congruence between epidemiologic and genotypic clustering of isolates of *C. jejuni*, genotypes would necessarily be restricted to one epidemiologic source, and not permitted to spread from one source to another. Without being able to colonize or at least survive in multiple host types, there would be very little possibility of *C. jejuni* spreading to human hosts, thus limiting its potential as a human pathogen. Since *C. jejuni* is identified as the leading bacterial human pathogen in Canada and worldwide (Thomas et al., 2013), it follows that as a species, its pathogenic potential is not limited by an inability to survive and spread throughout various epidemiologic sources.

3.3.7 Application to Other Organisms

While we have applied our model for measuring epidemiologic similarities to a dataset comprised of isolates of *C. jejuni*, theoretically, a similar approach could be taken for assessing similarities of other bacterial species with high importance to public health. Fundamentally, the model presented in Equation 3.6 should be applicable to any pathogen that exists in various epidemiologic niches and can be transmitted from one reservoir to another. Importantly, however, the assessment of the source component for the model may change from one bacterial species to another, based on the critical epidemiologic attributes relevant to the organism in question. For the assessment of the source component of *C. jejuni*, we created a line list containing items commonly found in case records for reporting the occurrences of campylobacteriosis to public health units in Canada, and combined them with known risk factors found in current literature (Friedman et al., 2004; Newell et al., 2011; Nichols et al., 2012). In order to accurately reflect other bacterial organisms, a re-

view of epidemiologic factors could to be undertaken to ascertain the appropriate attributes for comparing sources in a quantitative fashion.

3.4 Summary and Conclusion

Here we have presented a simple model for quantitatively assessing the similarities of isolates of a human bacterial pathogen based on a comparison of their basic descriptive epidemiologic attributes. This assessment helps to provide not only a snapshot of the overall epidemiology of large datasets, but allows for the direct comparison to other means of comparing the isolates including their genomic profiles. By comparing the genomic information of strains back to their epidemiology in a directly quantitative manner, we are not only able to assess how closely a genomic method can approximate the underlying epidemiology of bacterial populations, but these comparisons also provide insight into the pathogenic potential of an organism, and the diversity of environments in which the pathogen can survive. Within a test dataset of *C. jejuni* genomes sampledfrom a wide variety of Canadian sources and temporal and geographical ranges, we have demonstrated that outbreaks of campylobacteriosis can be visually identified by their narrow epidemiologic and genomic variability. Background levels of incongruence between clustering methods based on epidemiologic and genomic similarities may reflect adaptive potential for *C. jejuni* to survive and spread throughout many different animal and environmental sources, which ultimately increases the likelihood of exposure to the human population, reflected in the abundant numbers of cases of campylobacteriosis seen in Canada each year.

While a complete and holistic assessment on the efficacy of cgMLST to group isolates in a way that is relevant both genomically and epidemiologically likely cannot be determined using a dataset of only 274 bacterial isolates, we can suggest that certain caveats exist when investigating the concordance between genomic and epidemiologic partitions. Namely, we must consider that the reported statistics referencing the epidemiology for each of the subjects in study are in fact only inferred from the circumstances surrounding the sampling and

isolation of each bacterial isolate, leaving us unaware of whether the recorded epidemiologic source of the isolate is in fact the originating reservoir, or merely a transitory vector contributing to the circulation of the pathogen from one reservoir to another. Furthermore, the very nature of a common pathogen suggests that it is able to traverse epidemiologic boundaries and contaminate disparate sources. If no cross-contamination in the environment existed for *C. jejuni*, for example, then studying the transmission dynamics of the pathogen would become a redundant exercise, and public health efforts at monitoring the flow of pathogen prevalence from non-human sources could be considered unnecessary. Since substantial evidence exists as to the transmission of *C. jejuni* from non-human to human sources, and from non-human to other non-human sources, we can assume that a certain level of transitory activity likely exists among the genotypes of isolates from epidemiological clusters generated in this study. Thus, when comparing genomic cluster membership to partitions created via epidemiologic association, we should in fact expect the transmission dynamics of the organism to result in some discordance between the two methods.

In order to assess the full extent to which cgMLST clusters bacterial isolates in agreement with their underlying epidemiology, a dataset comprising isolates from firmly established reservoirs and outbreak clusters needs to be assessed. The dataset used in this study, by contrast, was originally selected to fulfill a requirement for several different research aims, including an assessment of strains with similar comparative genomic fingerprints, but from vastly different epidemiologies, and vice versa. In selecting micro-clusters of isolates based on their genotypic and epidemiologic similarities, it is likely that we have biased the estimation of fit between cgMLST and epidemiologic clustering of *C. jejuni* isolates. In order to establish a complete summary of the extent that cgMLST clusters strains of *C. jejuni* in a way that produces concordance to their basic epidemiologic traits, a much larger dataset may be required. Even in light of this bias, however, we have established here that epidemiologic attributes can be used effectively to compute a quantitative summary of

the epidemiology of *C. jejuni* isolates using our analytical model, to be directly compared with high resolution genotyping results such as cgMLST. This comparison provides an avenue for optimally leveraging WGS data in studying the genomic epidemiology of bacterial pathogens, a critically important aspect of disease prevention and control for the future of public health.

Chapter 4

Genotypic Recovery of *C. jejuni* from Mixed Samples: Evidence for Bias in Laboratory Surveillance

4.1 Preamble

Molecular subtyping of isolates of *C. jejuni* obtained via sampling throughout the food chain and the environment has provided important insights on population structure and the distribution and prevalence of genotypes that pose an increased risk to human health (Sheppard et al., 2009b; Müllner et al., 2009; Wilson et al., 2008). Subtyping methods exist to help identify important *Campylobacter* genotypes that cause human disease, and to help elucidate the routes of transmission of this pathogen from its various sources to human hosts. An ongoing challenge in surveillance efforts is ensuring that bacterial isolates obtained through sampling of various reservoirs are representative of the population in circulation, which allows for the assessment of the epidemiological significance of subtypes identified. It is not known, however, to what extent laboratory isolation methods may affect the subtypes recovered from animal or environmental sampling. If certain subtypes are suited to growth under laboratory isolation conditions, this may affect downstream assessment of relative prevalence of *Campylobacter* subtypes and obscure our understanding of the population of campylobacters circulating in various reservoirs and ultimately the risks associated with them.

Many studies now exist that include molecular subtyping as a means of assessing the prevalence of *C. jejuni* subtypes from various important food and environmental sources. Often, in these studies a select number of common genotypes emerge as highly prominent subtypes - these few subtypes frequently dominate over half of the population of isolates derived from sampling, while less-common genotypes will only be isolated in minute pro-

portions (Gripp et al., 2011; Dingle et al., 2002; Sheppard et al., 2009b; Kwan et al., 2008a; Müllner et al., 2010). In comparing the highly prominent genotypes from across several studies, a common trend appears, with the same genotypes being routinely isolated with high frequency. In fact, when comparing the most isolated *Campylobacter* subtypes from studies investigating various sources, there is often an association with the most prominent subtypes observed in the global pubMLST database. The isolation frequency of these common genotypes may be explained by two different hypotheses; namely, either they truly exist in higher proportion throughout the sampled environments, or alternatively, these highly prominent campylobacters are amplified in the process of sampling and laboratory isolation.

In Chapter 3, I explored a method of quantifying the epidemiological metadata associated with strains of *C. jejuni* collected through routine surveillance initiatives in Canada. When I compared the clustering based on epidemiologic metadata to clusters derived from genotypic associations using cgMLST and visualized the results using a 2-channel heatmap (Figure 3.5), I found that several clusters existed that demonstrated significantly higher association via their epidemiology, as well as other clusters that were more highly related via their genotypic profiles. These clusters with low epidemiologic similarities and strong genotypic associations can be used to describe certain genotypes of *C. jejuni* that are able to persist throughout a range of environmental and host conditions, possibly including strains that are temporally persistent, widely dispersed geographically, and that exist in multiple host species: attributes that may be summarized as conferring enhanced survival advantages for particular *Campylobacter* subtypes.

To date, several studies have noted the potential for recovery bias in the isolation of *Campylobacter* from various sample matrices. The use of different isolation protocols and media, particularly pre-enrichment broth, has been shown to influence the frequency of detection and moreover, may affect the genotypic richness of observed subtypes in mixed-strain populations. While many studies have focused on differences in detection across

isolation methods, or the differential detection of thermophilic *Campylobacter* species, few have examined the effect isolation methods have on the observed diversity of subtypes within a given sample. Recently, Williams et al. and Ugarte-Ruiz et al. assessed the genotype diversity of *C. jejuni* isolated from poultry samples using a variety of isolation protocols and found method-specific differences in the genotypes and numbers of isolates observed, suggesting that enrichment methods, in particular, bias the genotypic population obtained (Williams et al., 2012; Ugarte-Ruiz et al., 2013, 2012). Although the dynamics of this bias are unknown, it could be that selective components within a method favour the growth of particular *Campylobacter* subtypes, or the corollary, could act as a stressor to hamper the growth of less well-suited genotypes. Furthermore, Williams et al. suggested the possibility that additional *C. jejuni* subtypes were likely present in samples but were not detected due to recovery bias (Williams et al., 2012).

When assessing subtype bias from naturally contaminated samples, there is by definition an inherent uncertainty in the outcome as it is impossible to know *a priori* the true population and proportions of strains in a given sample. To mitigate this uncertainty and to further investigate the effects of enrichment methods on the recovery of *C. jejuni* subtypes in multi-strain samples, I developed a controlled recovery experiment in which isolates of known *C. jejuni* CGF subtypes were co-cultured and re-isolated using parallel enriched and non-enriched isolation methods. Furthermore, to test the hypothesis that large clusters from MLST and CGF databases contained isolates favoured under laboratory growth conditions, competitive cohorts were selected that contained representative isolates from genotypic clusters that have historically been isolated in high and low amounts. Isolates for the controlled spike-in experiment were selected based on their CGF fingerprint such that each of four isolates were distinguishable from one another within each controlled cohort based upon their CGF subtype. The resulting frequencies of each subtype were then assessed post-recovery using the CGF method to ascertain the effects of enrichment on isolates from both large and small genotypic clusters.

4.2 Methods

4.2.1 Strain selection and Resuscitation from Archival Library

Three cohorts (A, B, and C) representing a multi-isolate sample containing four isolates of *C. jejuni* were selected for the controlled recovery experiments. Isolates for each cohort were chosen from the Canadian CGF Database such that each isolate possessed a unique CGF fingerprint differentiating it from other subtypes in the cohort based on CGF. Care was taken to ensure that the selected isolates represented a variety of sources, including environmental water, human clinical, and animal faecal samples (Table 4.1).

To explore the hypothesis that strains from highly predominant genotypes may have improved fitness in the context of laboratory isolation conditions compared to strains with less common genotypes, each cohort was designed to contain one strain from a highly prevalent genotype and three strains from less prevalent genotypes, based on their frequencies in the Canadian CGF database (representing approximately 20000 isolates from various Canadian sampling initiatives). Associated MLST sequence type data was also used to corroborate the classification of strains into highly prevalent or less prevalent genotypes wherever possible.

4.2.2 Microbiological Recovery Trials

Selected strains were resuscitated from glycerol stocks kept at -80°C by sub-culturing twice into brain heart infusion broth (BHIB) (Fisher Oxoid CM1135) and incubating for approximately 24 hours in a tri-gas microaerophilic (MAE) incubator at 42°C. Fresh glycerol stocks for use in the competitive recovery experiments were then made from the second round of growth in BHIB for approximately 20 hours incubation time and stored at -80°C for up to three months.

Competitive recovery experiments were performed in duplicate for each of three cohorts. Twenty-four hours prior to each laboratory trial, frozen glycerol stocks were thawed and 800 μl from each stock was pipetted into 20 mL BHIB and incubated in a tri-gas MAE

incubator at 42°C for 24 hours. The concentration of cells was then assessed using a spectrophotometer to measure the optical density at 600 nm (OD_{600}) and normalized quantities of broth culture for each of the four strains were co-inoculated into 10mL of 1x phosphate-buffered saline (PBS) to create the mixed strain sample.

Prior to pooling, an aliquot from each parental strain culture was taken for DNA extraction and CGF analysis as a pre-isolation control for genotype comparison post-recovery. Pooled samples were voretexed briefly, and serial dilutions (10^{-1} to 10^{-5}) of the pooled culture were prepared and inoculated in the manner described below for direct and enriched recoveries.

4.2.3 Direct Recovery

From each serial dilution of the pooled spike, 100 μl of liquid culture was spread-plated to each of five plates of Charcoal-Cefoperazone Deoxycholate Agar (CCDA)(Fisher Oxoid CM0739), which were then subjected to incubation in MAE at 42°C. Following approximately 24 hours incubation, plates from each dilution series were assessed for suitable growth (e.g. containing 30-300 well-formed colonies); five plates from a single dilution series were then selected for subsequent steps. Individual colonies were harvested and streaked for isolation on CCDA; 100 colonies in total were selected per trial from a maximum of three CCDA plates. Following 24 hour growth on CCDA, isolates were then sub-cultured to BBL blood agar (BD 211037) supplemented with 7% sheep blood for subsequent harvesting and DNA extraction.

4.2.4 Enriched Recovery

To test the effects of enrichment on the recovery of the *Campylobacter* cohorts, 100 μl from the each dilution of the control spike was transferred to 20 mL of Bolton Broth (BB) (Fisher Oxoid CM983) with modified BB supplement (Fisher Oxoid SR0208E) and incubated for 24 hours in a MAE tri-gas incubator at 42°C. Following enrichment, a ten-fold serial dilution starting with 1ml BB into 9 mL PBS was made, and 100 μl from dilutions

10^{-3} to 10^{-5} were spread to five CCDA plates per dilution to ensure representative growth and well separated colonies for selection. After a 24 hour growth period, the dilution series best representing 30-300 colonies per plate was chosen, and 100 colonies were selected from a maximum of three CCDA plates for subsequent sub-culturing to blood agar.

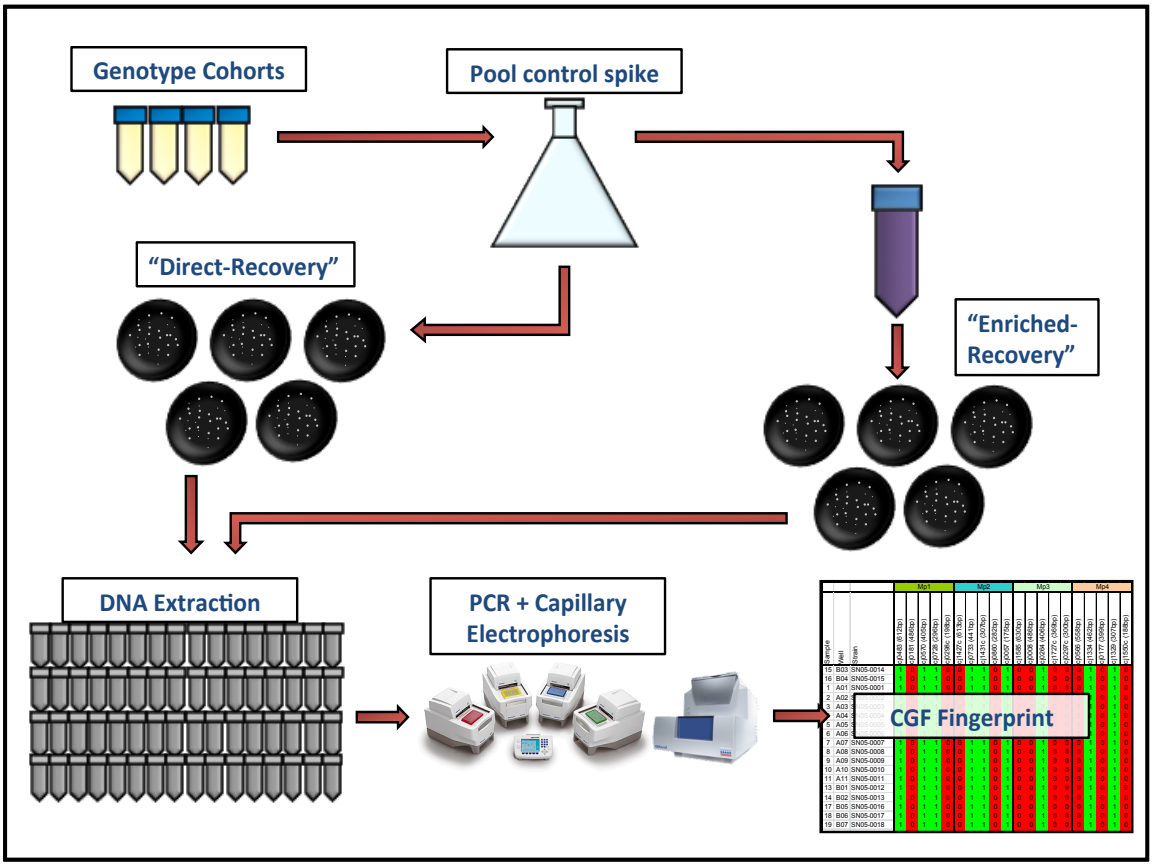

Figure 4.1: Schematic diagram of the microbiological recovery experiment.

4.2.5 DNA Extraction

For both direct and enriched recovery methods, DNA was extracted from 100 colonies using a modified protocol of the Epicentre Masterpure DNA Extraction kit (Epicentre MC85200). Briefly, cells were resuspended into 300 μl Cell and Tissue Lysis Solution (MTC096H) containing RNAaseA (1 μl of 5mg/mL) and proteinase K (5 μl of 50 mg/mL) and heated at 65°C for 30 to 60 minutes or until the lysates cleared. Samples were then cooled on ice and 175 μl of chilled MPC protein precipitation solution (MMP095H) was

added to each sample, after which they were vortexed and centrifuged to pellet the precipitate. Precipitation with 70% ethanol was used to clean and recover the DNA from the resulting supernatant. The DNA pellets were resuspended in buffer containing 1x Tris-EDTA and stored at -20°C or at 4°C short-term until PCR.

4.2.6 Verification of CGF fingerprints

Genotypes of each test strain were assessed using Comparative Genomic Fingerprinting (Taboada et al., 2012) prior to pooling, as well as after recovery (100 colonies each from direct and enriched recovery methods). To economize resources, each strain cohort was designed such that each fingerprint could be identified using a subset of 10 out of the 40 CGF loci. CGF types of recovered isolates were compared to those of the inoculum strains to determine the frequencies of recovery for each of the four strains within a cohort. Any isolates that could not be matched to one of the four parental strains (n = 21) were not included in downstream analyses.

4.2.7 Statistical Analyses

The effect of direct plating versus selective enrichment on the recovery of *Campylobacter* genotypes for each cohort was analyzed using IBM SPSS Version 21.

Pearsons Chi-Square test is useful in measuring if an observed distribution exists that is significantly different than expected, based on the sampling population. To assess the effect of isolation method on the distribution of recovered isolates from each cohort across two trials, a Chi-Square test was performed. The individual CGF types were used as independent variables and frequency of recovery for each strain was used as the dependent variable. The data was split using isolation method to assess the distribution of isolate recovery based on direct versus enriched isolation procedures.

To assess the relationship of parental CGF cluster size on the frequency of isolation from the microbiological trials, isolates from each cohort were given a rank corresponding to the size of the CGF cluster from which they originated (e.g. rank 1 to 4, with a rank of

1 corresponding to the cluster with the greatest relative number of isolates in the cohort). A mixed-design analysis of variance (ANOVA) was used to investigate the differences between multiple independent groups (Ranks), whilst also taking into account the effect of repeated measurements. An argument could be made that the trials represented independent observations, thus rendering a repeated-measures ANOVA inappropriate. However, as each strain used in this study was derived from a single vial of archival stock representing at least one clonal passage via laboratory isolation, for this analysis the trials were considered to be related, and thus the mixed design ANOVA was used to assess whether there was a significant change in the recovery of isolates depending on their assigned rank, and also to investigate if there was a significant interaction of genotype rank with the recovery method used. Outliers identified in the analysis were not removed, as they were deemed to represent true measurements of isolate recovery.

Trials 1 and 2 were used as the within-subject factors and Cluster Rank was input as the between subject factor for the analysis. Mauchlys test of Sphericity was attempted to test for equal variance between all possible pairs of groups. As there were only two levels of the repeated measures factor, however, Mauchly's test could not be computed, and we thus assumed that sphericity was violated. The Greenhouse Geisser correction measures the departure from perfect sphericity, and was used to correct for the degrees of freedom in the calculation of the F statistic to reduce the probability of a Type I statistical error.

To assess the probability of recovering all four isolates used in the spike broth mixture by enrichment and direct isolation methods, a binomial probability distribution was constructed using the frequencies from the genotype rank analysis and comparing the recovery of Rank 1 isolates with the sum of Ranks 2-4 as the two probabilities tested.

Figures were generated using the R statistical software program (R Core Team, 2015) and the package ggplot2 (Wickham & Winston, 2015).

4.3 Results

4.3.1 Database Analysis

As of June 2015, the Canadian CGF database contained 20080 isolates, 18925 of which belonged to multi-isolate clusters. To ensure a close comparison between the Canadian CGF and global pubMLST databases, we opted to perform CGF clustering at a threshold of 95% (i.e. 38/40 loci matching), as this reduced level of granularity has been shown to approximate the resolution achieved by MLST typing (Taboada et al., 2012). Within this subset of 18925 isolates, 1079 different multi-isolate genotypes were present at the 95% clustering level. This collection of 1079 genotypes was assessed for frequency bias, revealing a highly skewed distribution of genotypes. To compare these Canadian national distributions against a more geographically diverse database, we compared our findings to that of the publicly available *Campylobacter* MLST database, *pubMLST* (http://www.pubmlst.org/campylobacter, accessed June 15, 2015). At the time of access, the pubMLST database contained 32708 isolates sequence-typed by MLST, for a total of 2074 unique multi-isolate sequence-types (ST = 7/7 loci matches). Both the CGF and MLST databases portray similar skewness of genotype frequency; namely, there seems to exist few prominent genotype clusters which contribute in disproportionately high numbers to the overall population of *Campylobacter* isolates, while the remainder of the database population is constructed from many small clusters of isolates (n = $<$ 20 isolates). The log distributions of isolates by genotype are displayed in Figure 4.2 (*top*).

To test if the skewness of genotype distribution was related to an overabundance of single-source sampling, we assessed the source distributions of the ten most prevalent genotypes from each of the CGF and MLST databases. The ten most prevalent genotypes in the CGF database accounted for 20.45% (3870/18925) of the overall population of typed isolates; similarly, the ten most-frequently isolated ST in the MLST database accounted for 8487 isolates, equivalent to over 31% of the typed isolates from *pubMLST*. No evidence of host restriction appeared to exist across these high-frequency genotypes, with each geno-

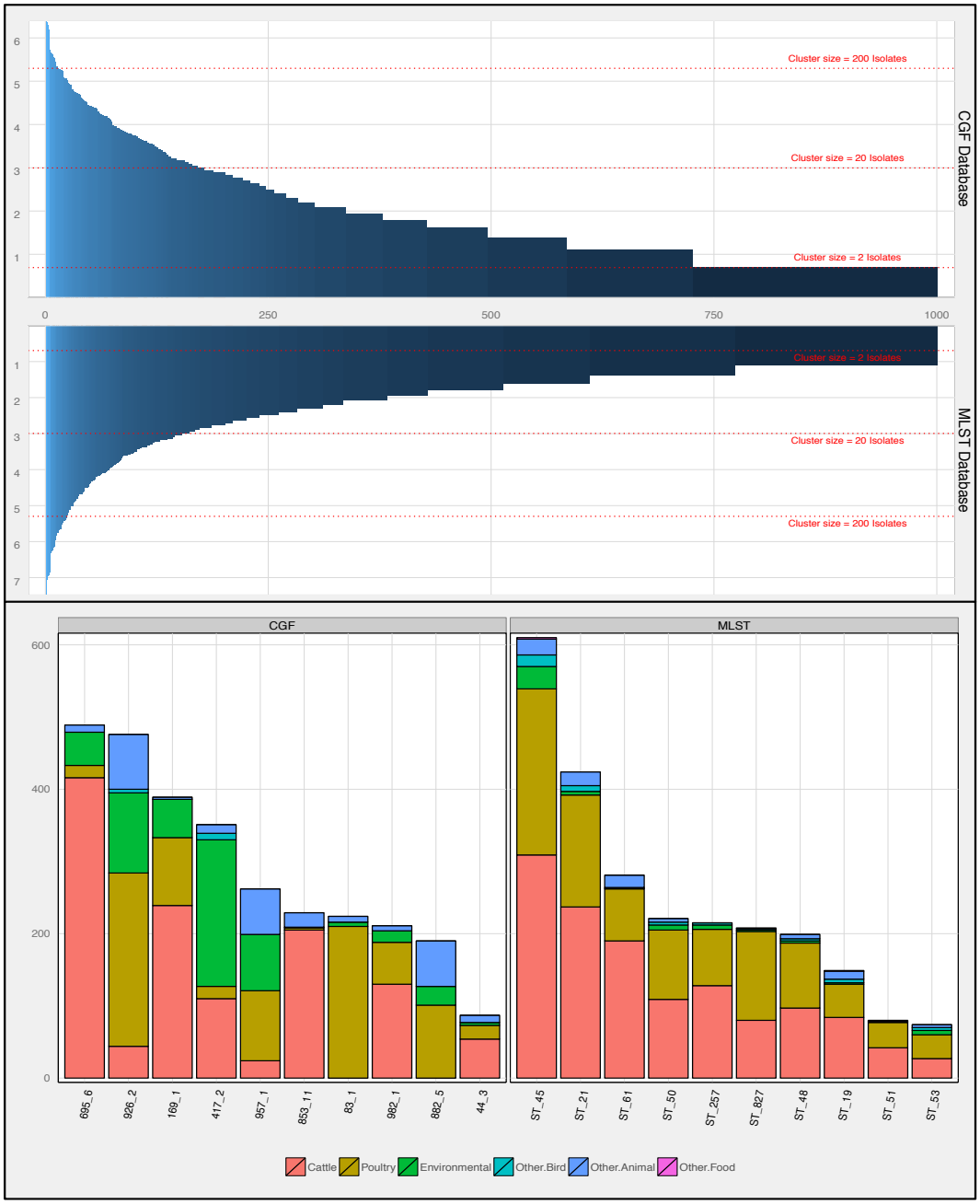

Figure 4.2: (*Top*) Log-scale frequency distribution of genotype clusters from the Canadian *Campylobacter* CGF database and the *Campylobacter* MLST database. Each bar along the horizontal axis represents a unique CGF 95% cluster (38/40 matching loci) or MLST Sequence Type (7/7 matching alleles). Only the top 1000 genotypes, ranked by frequency, are shown. (*Bottom*) Non-clinical source distribution of ten most prevalent clusters from CGF and MLST databases.

type being derived from multiple non-human sources (Figure 4.2(*bottom*)).

4.3.2 Recovery of Isolates from Laboratory Trials

Genotype cohort A consisted of four isolates derived from environmental water samples in southern Ontario, Canada collected as part of routine surveillance initiative in an area of agricultural activity. All four isolates used in this cohort shared similar source and isolation date and were originally isolated using typical laboratory isolation procedures including enrichment in BB. In both the direct and enriched recovery results, however, the isolate belonging to the largest genotype cluster, 957.1.1 was recovered in substantially greater amounts than the remaining isolates in the mixture, even though all four isolates were present in approximately equal proportions in the control spike. Following direct recovery, isolate CI-5178 amounted to 50% of the total number of recovered isolates. This observed proportion increased to 89% of the total number of isolates recovered when enrichment with BB supplemented with antibiotic was used prior to recovery. Additionally, isolates CI-5043 and CI-5039 were almost altogether absent after enrichment, representing only 1% of the recovered isolates, significantly less than 26% and 11% respectively, after direct recovery.

Cohort B consisted of four isolates from varying host and temporal sources, including chicken meat, water, and human clinical samples ranging from years 2004 to 2010. These isolates did not share similar origins of isolation, and yet the observed recovery of isolates with and without an enrichment stage is similar to that of the first cohort. Isolate CE_M_10_4053, belonging to the large genotype cluster 169.1.1, accounted for approximately 29% of the recovered isolates after direct plating; this number increased to over 47% after enrichment. Conversely, isolate 07_2680, belonging to the relatively less-frequent genotype cluster 83.7.1, accounted for 17% of the isolates recovered using direct plating; this number decreased by over half to 6.5% following recovery post-enrichment.

Cohort C contained a similar mixture of sources as cohort B, although different genotypes were chosen for the mixture. These genotypes produced results concordant with those from the previous cohorts. Isolate 07_1875, from cluster 735.5.1, increased in recov-

Table 4.1: *C. jejuni* isolates and counts from microbiological recovery trials. DR indicates direct recovery, ER indicates enrichment-based recovery method. Cluster Size indicates the number of isolates pertaining to the same cluster in the CGF database, measured at 100% cluster congruence (40/40 loci matches).

Strain	Cohort	Source	Year	DR 1	DR 2	ER 1	ER 2	Cluster Size
CI-5178	A	Water	2011	42	52	90	73	274
CI-4685	A	Water	2011	12	11	6	10	13
CI-5043	A	Water	2011	31	18	2	0	3
CI-5039	A	Water	2011	15	7	2	0	1
CE_M_10_4053	B	Chicken	2010	32	26	45	49	593
CI-2669	B	Water	2006	20	28	16	17	27
07_2680	B	Human	2007	18	16	8	5	38
CGY_HR_241	B	Human	2004	29	30	30	29	24
07_1875	C	Human	2007	34	18	52	53	217
CI-4820	C	Water	2011	23	46	21	29	51
CE_R_11_0073	C	Chicken	2011	13	12	0	1	71
CE_R_11_0249	C	Chicken	2011	28	22	26	16	67

ery frequency from 26.5% in the direct method to 53% in the enrichment method, while the least-recovered isolate, CE_R_11_0073 decreased from 12.7% in the direct method to 0.5% following enrichment. Total recoveries from each cohort are summarized in Figure 4.3(top).

4.3.3 Effect of Isolation Method on Recovery of Isolates

All instances except Cohort B - Trial 1 produced a significant difference in the distribution of recovered isolates based on the isolation method employed. When summarized, the overall distributions changed significantly as a function of method. Results from each test are presented in Table 4.2.

4.3.4 Effect of Cluster Rank on Recovery

No significant within-subject effects were found for **Trial**, indicating no significant change was observed in the mean recovery of ranked isolates between trials. However, there was a significant result obtained for the interaction of **Trial** by **Rank**, indicating that

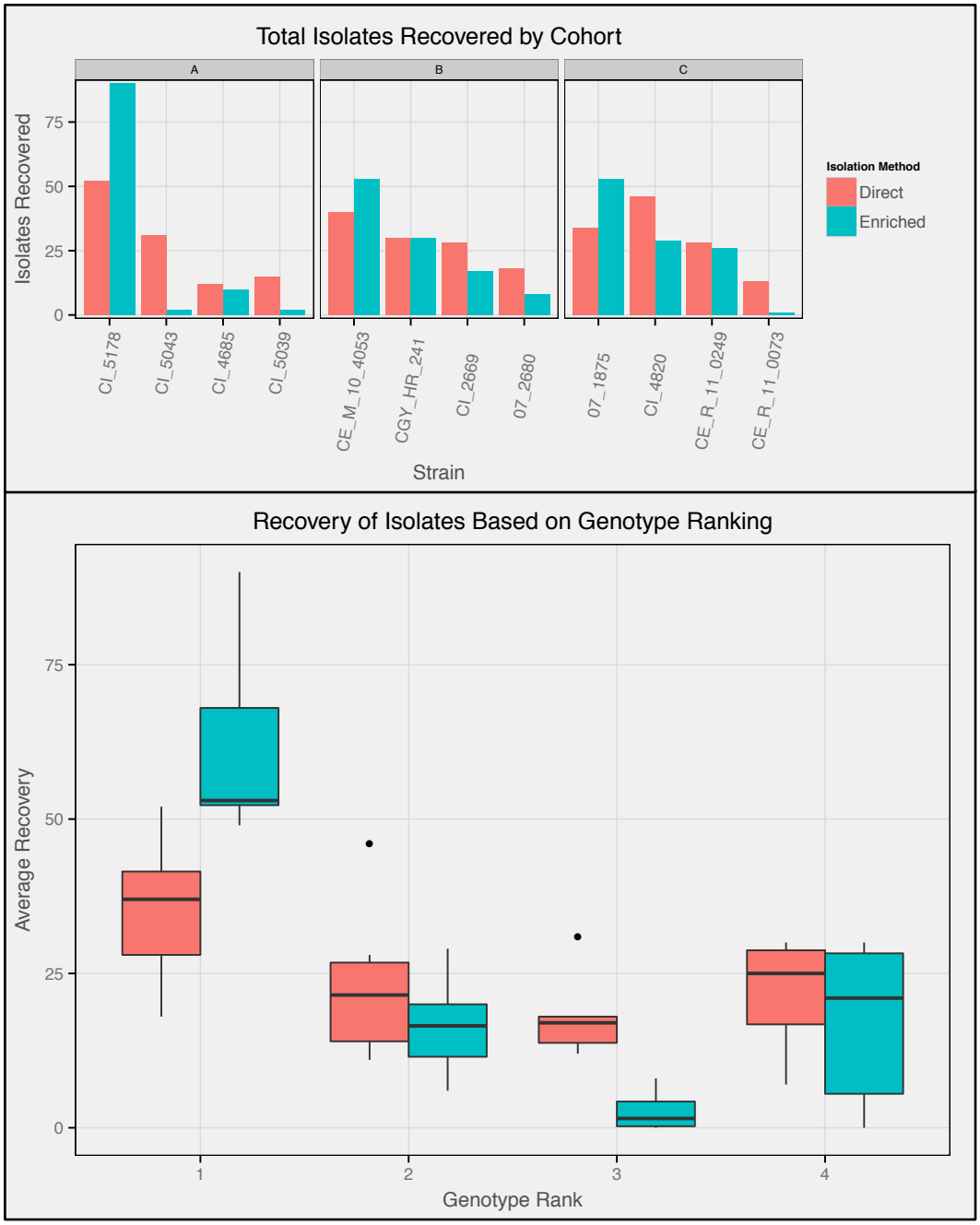

Figure 4.3: (Top) Summary of strain recoveries from microbiological recovery trials. Three cohorts were tested in duplicate. (Bottom) Recovery of genotypes based on genotype rank. Top, middle, and bottom horizontal lines on boxes indicate third, median and first quartile positions. Range of data is indicated by vertical lines extending from the boxes. Circles above boxes indicate statistical outlier observations.

the observed mean recoveries of isolates based on **Rank** changed from Trial 1 to Trial 2 ($F(3,20) = 3.116$, $p = 0.049$, partial $\eta^2 = 0.319$). There was a highly significant between-

Table 4.2: Pearson's Chi Squared test results for the distribution of strain recovery based on genotypic rank. An asterisk (*) indicates statistical significance at the 95% level of confidence.

	Trial 1	Trial 2
Cohort A	$\chi^2(3, n = 200) = 54.9, p < 0.001$*	χ^2 (3, n = 171) = 28.4, p < 0.001*
Cohort B	χ^2 (3, n = 214) = 6.1, p < 0.106	χ^2 (3, n = 200) = 15.5, p = 0.001*
Cohort C	χ^2 (3, n = 197) = 16.9, p = 0.001*	χ^2 (3, n = 197) = 31.4, p < 0.001*
Overall	**χ^2 (11, n = 611) = 77.9, p < 0.001***	**χ^2 (11, n = 568) = 75.4, p < 0.001***

subject effect found for **Rank** ($F(3,20) = 8.54$, $p = 0.001$, partial $\eta^2 = 0.562$) indicating a substantial difference between the mean recoveries of each of the isolates based on the ranking applied to each of their respective parental clusters.

Follow-up post-hoc analysis was performed using the Tukey Honest Significant Difference test to assess the individual differences in the recovery of isolates based on the assigned cluster rank. In each cohort, a significant difference was found between the mean recovery of Rank 1 isolates and the mean recovery of Ranks 2 (p = 0.006), 3 (p = 0.001) and 4 (p = 0.005), indicating that the isolates from the highest-frequency clusters were recovered in disproportionately higher amounts than those from the smaller clusters. No significant differences were found between the recoveries of ranks 2, 3 and 4. Results from the cluster rank analysis are summarized in Figure 4.3(bottom).

Table 4.3: Descriptive Statistics for Ranked Recoveries

		Trial 1			Trial 2		
Method	Cluster Rank	Mean	SD	N	Mean	SD	N
Direct	1	38.7	4.2	3	32.0	17.8	3
	2	18.3	5.7	3	28.3	17.5	3
	3	20.7	9.3	3	15.3	3.1	3
	4	24.0	7.8	3	19.7	11.7	3
Enriched	1	65.0	21.7	3	58.3	12.9	3
	2	14.3	7.6	3	18.7	9.6	3
	3	3.3	4.2	3	2.0	2.7	3
	4	19.3	15.1	3	15.0	14.5	3

4.3.5 Probability of Recovering Multiple Genotypes from Mixed Samples

Results from the binomial probability distribution are presented in (Figure 4.4). In order to achieve a 95% probability of recovering all four isolates from the spiked broth culture by direct isolation, a selection of 37 colonies was required, as opposed to a 84 colonies required when using an enrichment broth isolation procedure.

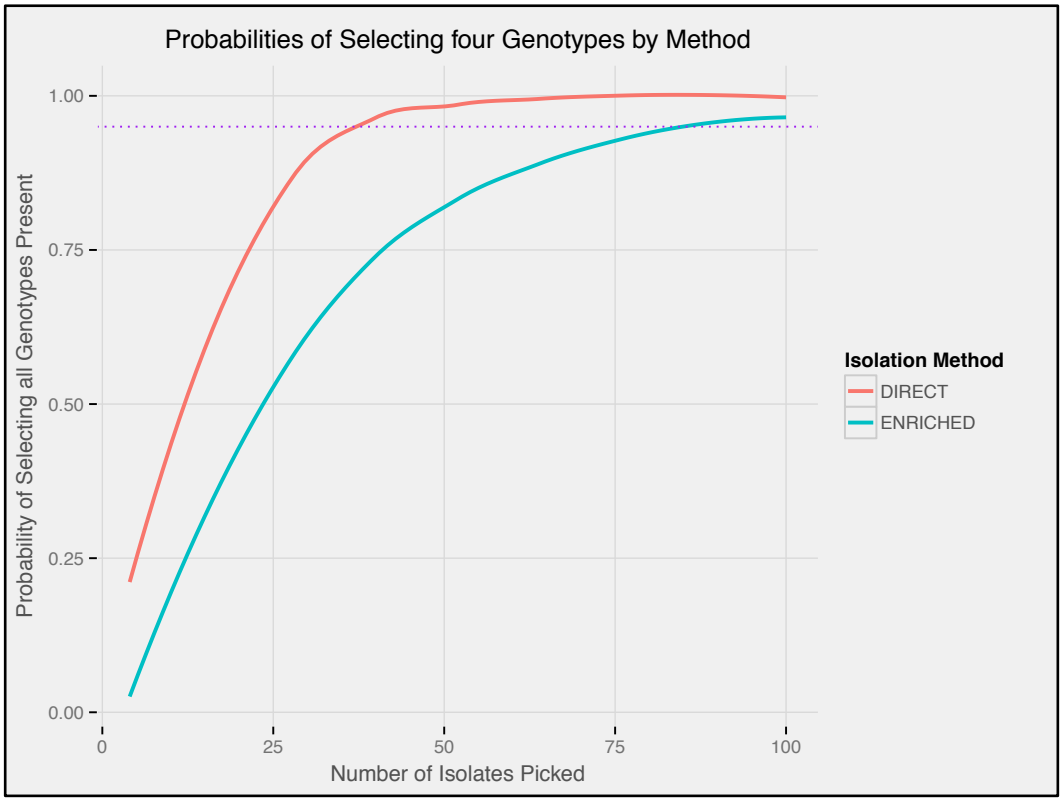

Figure 4.4: Probabilities of selecting all genotypes present in a 4-genotype mixed sample, based on isolation method.

4.4 Discussion

4.4.1 Enrichment Based Isolation methods

Enrichment media have been shown to be effective for the recovery of *Campylobacter* from a variety of samples including retail meat, animal faeces, and environmental water. The use of enrichment steps in the isolation of campylobacters was traditionally suggested for cases where low numbers of cells may be present (Hutchinson & Bolton, 1984), al-

though enrichment media are widely adopted in most modern-day standard isolation procedures. Numerous types and formulations of enrichment media have been developed, including Bolton, Preston, and Exeter broths (Corry et al., 1995), although comparative studies have yet to come to a consensus on which is preferred, leading to the ongoing development of novel enrichment media (Hayashi et al., 2013; Chon et al., 2013). Despite the lack of consensus, BB is one of the most widely used enrichment mediums, in part due to its being recommended by the International Standard Organization (ISO, 2006) and the US Food and Drug Administration (Hunt et al., 2001), and was chosen in this study for that reason.

Numerous studies comparing enrichment-based methods to direct plating methods have been conducted. In many cases sample enrichment was found to provide higher rates of *Campylobacter* recovery than direct plating (Habib et al., 2011), particularly when isolating from cattle faeces (Gharst et al., 2006; Atabay et al., 1998; Stanley et al., 1998; Garcia et al., 1985). Again, it is thought that enrichment enables the resuscitation and propagation from low numbers of organisms that may not have otherwise been detected through direct plating methods. By contrast, other studies, particularly those focused on isolation from poultry meat, have shown direct plating to provide higher rates of recovery than enrichment methods (Habib et al., 2008; Kiess et al., 2010; Musgrove et al., 2001). Here it was proposed that the large number of competing, non-*Campylobacter* organisms may have confounded detection when using enrichment-based methods, however, if this was the case, one would expect to find similar difficulties with isolation from cattle faeces, where campylobacters make up a very small proportion of the bacterial load. Nevertheless, similar hypotheses are often provided in defense of either outcome (direct or enriched) suggesting that additional research is required.

In this study we attempted to examine the effects of enrichment in comparison to direct plating on the recovery of *C. jejuni* subtypes from multi-strain samples in a controlled environment. To avoid confounding factors related to different sample matrices or background

organisms, the experiments were conducted using sterile broth spiked with known *C. jejuni* strains. If the null hypothesis for the recovery experiments were true, (i.e. no recovery bias exists isolation methods) one would expect to recover approximately equal proportions of each of the four genotypes within each cohort. In the direct plating trials, which were expected to most closely match the null hypothesis, only 1 of the 6 trials (Cohort B, Trial 1) showed no significant variation in the recovery of the four genotypes ($\chi^2(3, n = 214) = 6.13$, $p = 0.106$). The remaining direct plating trials from cohorts A, B and C showed statistically significant deviations in the number of expected isolates recovered from each genotype ($p = 0.022$), suggesting that some degree of bias exists among the recoveries of these isolates. The variations observed in the recovery of subtypes using the direct plating method may be due to method error (e.g. normalization of bacterial inoculum using A600), or inherent differences in competitive fitness and/or growth rate between the individual strains. Nevertheless, the direct recovery results serve to act as a baseline against which the enrichment recovery can be assessed.

All enrichment trials had statistically significant deviations from the null hypothesis. Moreover, the extent of recovery bias observed among the enrichment trials was significantly higher than that of the direct plating results suggesting that further bias occurs relative to that observed in the direct trials. This suggests that the enrichment process may favor the growth of certain genotypes either by directly encouraging the growth of specific *C. jejuni* subtypes, or by hampering the growth of others and thereby decreasing competition in the nutrient broth environment. Since all of the isolates used in this study were originally isolated through a BB enrichment procedure it seems less likely that components within the enrichment broth are directly suppressing the growth of less frequently isolated genotypes, but more likely that certain genotypes are being preferentially amplified. Also, because the sampling size remained constant across trials, if the growth of a single genotype was amplified, it follows that the remaining isolates would be recovered in lesser proportions regardless of the effect of direct suppression.

4.4.2 Genotype Frequency and Database Analysis

The abundance of a small number of highly dominant *Campylobacter* genotypes is well documented. Historically, the most predominant of these lineages have been characterized by multi-locus sequence typing (MSLT) as belonging to sequence types ST-45 and ST-21. These groups in particular are characterized as being poorly-host restricted, and are often referred to as "generalists" due to their widespread nature across multiple ecologic niches and apparent adaptability to diverse conditions (Gripp et al., 2011; Sheppard et al., 2011, 2013). It is possible that the genetic and phenotypic mosaicism underlying the backbone of these stable, dominant genotypes has provided a fitness advantage that makes them ecologically agile and well adapted to living in a wide variety of hosts and environments. Indeed, the implication of these types in a high proportion of human clinical cases worldwide is evidence that humans are another well-adapted-to niche.

The three highly-represented isolates used in this study came from water, poultry, and human sources; moreover, the genotype clusters to which each of the isolates belong are comprised of a broad range of sampling with respect to host, spatial and temporal sources. The CGF genotypes sampled: 169.1.2, 957.1.1, and 735.5.1 are not only among the ten most frequently occurring CGF fingerprints, but also correspond to the MLST sequence types ST-21, ST-45 and ST-42 respectively. STs 21 and 45 in particular are among the most frequently isolated genotypes in the MLST database. ST-42 is also characterized as a highly prevalent sequence type (ranked 12 in terms of frequency), and it has been suggested to be moderately host-restricted to cattle (Kwan et al., 2008a,b; Sheppard et al., 2014). However, when compared to the source distributions of ST-21 and ST-45 from the MLST database, there is little difference in the restrictive nature of the genotype, save that the predominant animal source is cattle, not poultry. Thus, ST-42 also appears to act as a "generalist", and we can hypothesize that it behaves in a similar fashion under laboratory conditions to the traditional generalists ST-21 and ST-45.

The results from the laboratory recovery experiments suggest a correlation between the

ability of a bacterial isolate to perform under typical laboratory isolation conditions and the size of the genotype cluster from which it came. In each of the repeated laboratory recovery trials, the genotype most frequently isolated based on historical sampling was able to out-compete the remaining isolates in all enrichment based isolations, even though all isolates were present in approximately equal quantities in the initial control spike. The significant bias afforded to dominant genotypes in laboratory isolation procedures would appear to indicate that enrichment media such as BB is yet another niche that dominant genotypes can exploit, at the expense of rarer subtypes. While few studies have examined the effect of isolation method on genotypic richness, this is the first, to our knowledge, to describe the occurrence of significant bias towards genotypes that are among the most highly prevalent worldwide. In cases of mixed-strain samples, competition with dominant genotypes is likely to lead to false negative findings and decreased genotypic richness.

4.5 Conclusions and Future Work

While it is not surprising that certain genotypes are well-adapted to a broad range of host niches and are thus isolated with high frequencies, the more significant implication is the absence of less frequently-isolated genotypes as a result of their being out-competed during isolation protocols. Routine surveillance is essential for determining the impact of *C. jejuni* on the health of the human population; by employing molecular subtyping protocols, high-risk genotypes can be identified for further study. However, when isolation methodologies bias recovery results by obscuring the presence of certain genotypes, we lose the ability to identify genotypes of high public health importance but with low adaptability to laboratory conditions. Likewise, in cases of mixed infections with clinical cases, laboratory recovery bias could be impacting our ability to determine the genotypes responsible for causing disease.

A variety of procedures have been suggested for improving the representativeness of recovery during isolation of *Campylobacter* spp. from samples. Suggestions include choos-

ing three to five colonies from agar plates following enrichment (Jørgensen et al., 2002), selecting all *Campylobacter*-like colonies present for subtyping (Heuer et al., 2001), performing isolation in the absence of selective antibiotics (Le Roux & Lastovica, 1998), and selecting all isolates portraying a unique morphology on an agar plate (Kramer et al., 2000). The international standard ISO 10272:2006 recommends up to 5 well formed colonies be selected for subtyping analysis (Habib et al., 2011; ISO, 2006). While it is impossible to know the subtype distribution in a sample *a priori*, since the dominant generalist genotypes appear to thrive throughout many host and environmental niches, the possibility of a sample containing an isolate from a generalist lineage would seem highly likely. The results generated by our study suggest that in a sample containing approximately equal proportions of different genotypes, there is little chance of recovering more than a single dominant genotype when only selecting up to five colonies for subtyping; this probability is further reduced when isolation procedures involving an enrichment stage are employed.

Our results here do not preclude the possibility that other non-generalist dominant genotypes known to be more host-restricted (for example ST-61, which is associated with cattle; and ST-353, which is associated with chicken) may also have a competitive advantage in an enrichment broth ecosystem. Further studies are required to determine if other frequenctly-isolated genotypes display similar competitive advantages in mixed strain samples or if certain low-frequency genotypes have a competitive advantage when grown in enrichment media. Furthermore, since the sample size remained consistent (100 randomly selected isolates per method per trial), it is impossible to conclude if non-dominant genotypes were being directly suppressed, or if the increased proportions of dominant genotypes were diluting the recovery of the non-dominant genotypes. Interestingly, the hierarchy of recovered strains found between direct and enriched methods was not necessarily conserved in each case, however, when assessed as a dichotomous outcome between the dominant versus non-dominant isolates, the overall trend of strains from higher-frequency clusters outperforming the rest of the mixture was consistent. To obtain a better understanding of the existing dy-

namics in these recovery experiments, in the future a more quantitative approach based on qPCR may be used to rapidly assess the overall relative proportions of isolates in a mixed broth culture.

Chapter 5

Summary and Conclusions

In the last century alone, studying the epidemiology of infectious disease has undergone significant transformations, augmenting the traditional, *shoe-leather* style of investigative epidemiology with new molecular techniques that help define the transmission and persistence of infectious disease agents.

In Chapter 1, I reviewed historic and contemporary techniques for the identification and characterization of bacterial pathogens and their application to epidemiological problems. These subtyping techniques can be condensed into two broad categories, namely, the phenotypic and genotypic classification of pathogens. Phenotypic techniques rely on classifying bacteria based on expressed cellular traits; culturing, biotyping and serotyping have been relied on for almost a century to help identify bacteria that pose an increased risk to humans. Advanced phenotypic techniques such as multi-enzyme electrophoresis, phage-typing and MALDI-TOF mass-spectrometry based typing were developed to provide increased levels of discrimination to more traditional phenotypic approaches, but still retain many of the same limitations as their predecessors. Phenotypic typing generally requires specialized expertise in the interpretation of results, which can fluctuate between users and laboratories. Furthermore, reagents and equipment costs can be quite high for many phenotypic typing methods, thus limiting the ability to type large numbers of bacterial isolates.

In contrast to phenotypic classification and typing, molecular typing approaches are generally robust to user interpretation, portable between different laboratories, and have a higher degree of discrimination to discern between bacterial species and subtypes. Genotyping techniques used in molecular epidemiology have progressed alongside the many advances in the field of molecular biology including the use of restriction enzymes, hy-

bridization, and PCR. An entire generation of molecular subtyping methods based on combinations of these various techniques ultimately lacks the resolution afforded by whole-genome sequencing, which is set to become the new standard in pathogen characterization used in molecular epidemiology.

Single and multi-gene DNA sequencing has been used for several decades now to differentiate bacterial subtypes, but only recently has sequencing of the entire bacterial genome become affordable enough to potentiate its use in epidemiological investigations. Several examples now exist where whole genome sequence (WGS) has proven invaluable for elucidating the transmission dynamics and microevolution of pathogens after or even during an outbreak. The successful implementation of WGS for epidemiologic investigations has now even prompted its increasing use in routine pathogen surveillance, however, several challenges remain in the interpretation of WGS for epidemiologic purposes.

In this thesis, I have chose to pursue a WGS-based investigation of *Campylobacter jejuni*. *C. jejuni* has been shown to survive and circulate in a multitude of animal and environmental sources and is among the highest contributors to bacterial gastrointestinal disease in Canada and worldwide. As a human pathogen, *C. jejuni* does not typically represent a life-threatening infectious disease, so resources aimed at elucidating its transmission and attribution are limited. Among several other risk factors for campylobacteriosis include preparation and consumption of poultry, exposure to rural environments, eating at commercial fast-food establishments, and domestic and international travel, and there appears to be an upward trend in cases of campylobacteriosis overall. Whether this incline in cases is due to an increase in the numbers of *C. jejuni* circulating in the environment, an increase in pathogenic potential of the organism, or just an improved ability to diagnose the pathogen, is yet unknown.

The aforementioned factors concerning the epidemiology of *C. jejuni* make it an ideal candidate for enhanced WGS-based epidemiologic investigations. As *C. jejuni* possesses a relatively small genome, approximately 1.6 Mb, and is readily found from a wide variety of

environmental, animal and clinical samples, sequencing the genome of *C. jejuni* is relatively inexpensive (<$100), providing far more information than previous typing methods are able to. To this end, I explored the use of whole-genome epidemiology of Canadian *C. jejuni* in this thesis, with three major topics investigated:

1. The use of WGS data in *ad hoc* epidemiological investigations
2. Quantifying basic epidemiologic attributes for comparison to WGS data
3. Sampling bias identified through whole-genome epidemiology

In Chapter 2, I sequenced 294 isolates of *C. jejuni* chosen from the Canadian CGF database. These isolates were selected from genomic clusters derived by analysis using CGF, a 40-gene PCR assay that determines the presence or absence of genes found within the accessory genome of *C. jejuni*. A critique could be made in my decision to use CGF for the basis of selecting *C. jejuni* isolates for a WGS-based core-genome MLST study, as a method based on the contents of the accessory genome may not accurately reflect the same relationships derived by an analysis of the core genome using WGS data. However, CGF has been shown to have extremely high correlation to MLST typing results, both in the literature (Taboada et al., 2012), as well as the results from Chapter 2, suggesting that a compatible phylogenetic signal can be achieved between accessory and core gene analyses. When I assessed the concordance of CGF typing results to cgMLST, a reasonably strong agreement between cluster memberships was produced by the two methods, with CGF predicting cgMLST results with up to 88.7% probability (Table 2.1). Thus, using CGF as an estimator of genetic relationships for a WGS based analysis was likely a sound choice, and it appears from the analysis in this thesis that the relationships construed by comparing accessory genome content does coincide, to a large extent, with relationships formed from comparing content from the core genome of *C. jejuni*.

Another note of concern for the selection of isolates in Chapter 2 was the criteria developed to maximize resources dedicated for whole-genome sequencing. Isolates were chosen

in pairs or triplets from CGF genotypes that were identified as being of interest due in part to their discordant epidemiologies - i.e. highly similar genotype clusters that contained isolates derived from a wide range of epidemiologies, whether it be temporally, geospatially, or host-associated. This selection process was originally developed to assess the extent to which CGF genotyping data is concordant with the underlying epidemiology of *C. jejuni* isolates. Due to this selection approach, my final sample population of 274 *C. jejuni* isolates can be considered a veritable mosaic of small genotype clusters from a variety of epidemiologic backgrounds, hence a thorough analysis of the efficacy of WGS-based typing for estimating the epidemiology of these bacterial strains is not fully possible using this dataset.

One of the limitations imposed by the diverse selection of genotypes from my sample dataset of 274 *C. jejuni* isolates is that the clusters derived from analysis with cgMLST at a threshold of 90% are limited in size. As seen in Figure 2.4, the largest cluster derived from the selection of *C. jejuni* contained only 31 isolates, from a total sample population of 274 isolates analyzed. By analyzing a larger dataset, including a larger number of isolates from CGF clusters that were highly related genotypically, I may have been able to produce larger cgMLST clusters for comparing the epidemiology of similar genotypes. Conversely, by selecting a sample of isolates from cohesive ecologic niches, it may have been possible to more comprehensively examine genomic similarities of isolates obtained under similar sampling conditions.

The confounding effect of having many small genomic clusters with varied epidemiology is readily seen in the tanglegram of Figure 3.9. Here, rather than seeing high levels of agreement between the epidemiology and the genomics on this set of *C. jejuni* isolates, there were many different genomic inputs to the various epidemiologic clusters, with a few examples of larger clusters of isolates that show high concordance between the genomic and epidemiological clustering, indicated by multiple parallel lines connecting the two dendrograms. Overall, the tanglegram in Figure 3.9 appears highly discordant, however, when

assessed more thoroughly, there are many instances where small clusters of two to five isolates cluster together both in the epidemiological dendrogram (*left*) as well as the cgMLST dendrogram (*right*); thus, when taking into consideration the sampling methodology used to select isolates from many small epidemiologic and genomic clusters, the results seen in Figure 3.9 appear consistent with a sampling approach that emphasized the selection of isolates from many small epidemiologic and genomic clusters.

Results from my quantitative model proposed in Chapter 3 generated logical groupings of strains of *C. jejuni* based only on basic epidemiological metadata. One of the major challenges in epidemiological investigations of cases of campylobacteriosis is the lack of in-depth follow-up due to constraints in time, resources and confidentiality of information. Thus, by providing a means to both quantitate easily-attained sampling data, and compare it to genomic data, I hope to benefit future endeavours in applied public health. By establishing a firm baseline between genomic profiles of bacterial isolates and their relative epidemiology, it may be possible to determine the probable exposure leading to infection for novel cases of campylobacteriosis based on WGS analysis of the clinical isolate. As the fields of public health and clinical microbiology continue to progress towards WGS analysis for all bacterial isolates, developing analytical methodologies that leverage the enhanced levels of information provided by these recent technologies is increasingly paramount.

In addition to the potential benefits of rapidly identifying the source of clinical bacterial illness, assessing whole populations of *C. jejuni* collected from surveillance could provide important evidence for establishing prevention strategies that stem the flow of *C. jejuni* upstream of human exposure. In Chapter 3, I discuss briefly the application of the epidemiological similarity model for identifying isolates belonging to either persistent genotypes with a high capacity for survival in varied environments, or conversely, ecological reservoirs that are able to support the continuation of many different genotypes of *C. jejuni*. Both of these situations pose an important risk to public health; persistent genotypes may lead to enhanced human exposure, leading to increased possibility of disease in the popula-

tion. By identifying these genotypes with increased potential for persistence, efforts can be made to screen for them in routine surveillance, providing early awareness to help combat potential outbreak scenarios. Epidemiologic sinks can collect a vast range of organisms, increasing the chances of human exposure to potentially pathogenic subtypes of bacteria; further, by permitting the circulation of many types of bacteria, they may promote an increased transmission of genetic material from pathogenic to otherwise benign subtypes of *C. jejuni*. With antimicrobial resistance currently considered a top priority worldwide, practicing stewardship by identifying and intervening at the level of environmental sinks may prove essential to the health of Canadians (Public Health Agency of Canada, 2016).

The investigation in Chapter 4 explores one of the situations identified by the comparisons made in Chapter 3, wherein certain genotypes of *C. jejuni* are found in disproportionate abundance among routine sampling spanning several years, hosts, and geographies. Through a set of controlled microbiological trials, I established that these genotypes, while pervasive and persistent in the environment, possess a level of fitness highly adapted to laboratory recovery conditions. This enhanced laboratory fitness allows for these genotypes to out-compete less laboratory-adapted strains in a sampling cohort, thus obscuring the true proportions of genotypes from mixed-strain environments. The significant potential for bias due to laboratory recovery methods in studies assessing genotype *richness* and diversity is cause for concern. With the results from Chapter 4, I hope to promote either modifications to isolation methods for *C. jejuni*, or pursue analytical corrections for measuring the dominance or persistence of certain genotypes from within mixed strain samples in order to improve our understanding of the true diversity of *Campylobacter* genotypes circulating in the environment and their impact on public health.

The work presented in this thesis serves to illustrate the use of WGS for practical pursuits in the context of the epidemiology of *C. jejuni*. I have assessed the use of a core-genome genotyping assay, *cgMLST* as a robust and high-resolution method for rapidly estimating strain-strain genomic relationships, developed an analytical model for quantitating

otherwise qualitative epidemiologic relationships, and shown how both genomic and epidemiologic information can be combined to identify isolates of increased public health risk. As the widespread use of WGS of bacterial isolates for public health continues to propagate alongside advances in molecular technologies, it will become increasingly important to be able to contextualize the quantitative genomic data with meaningful, practical epidemiology. It is my hope that the analyses provided in this thesis help to further the burgeoning field of *genomic epidemiology*.

Bibliography

Agunos, A., Waddell, L., Léger, D., & Taboada, E. (2014). A systematic review characterizing on-farm sources of *Campylobacter* spp. for broiler chickens. *PLoS ONE*, *9*.

Allos, B. & Blaser, M. (1995). *Campylobacter jejuni* and the expanding spectrum of related infections. *Clinical Infectious Diseases*, *20*, 1092–1099.

Anderson, E. S. & Williams, R. E. (1956). Bacteriophage typing of enteric pathogens and staphylococci and its use in epidemiology. *Journal of Clinical Pathology*, *9*, 94–127.

Atabay, H. I., Corry, J. E. L., & On, S. L. W. (1998). Identification of unusual *Campylobacter*-like isolates from poultry products as *Helicobacter pullorum*. *Journal of Applied Microbiology*, *84*, 1017–1024.

Baggesen, D. L., Sorensen, G., Nielsen, E. M., & Wegener, H. C. (2010). Phage typing of *Salmonella typhimurium* - is it still a useful tool for surveillance and outbreak investigation? *Euro surveillance : European Communicable Disease Bulletin*, *15*, 19471.

Baker, M., Wilson, N., Ikram, R., Chambers, S., Shoemack, P., & Cook, G. (2006). Regulation of chicken contamination urgently needed to control New Zealands serious campylobacteriosis epidemic. *The New Zealand Medical Journal*, *119*, 76–83.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*, 455–77.

Beall, B., Facklam, R., & Thompson, T. (1995). Sequencing *emm*-specific polymerase chain reaction products for routine and accurate typing of group A streptococci. *Journal of Clinical Microbiology*, *34*, 953–958.

Best, M. (2004). Ignaz Semmelweis and the birth of infection control. *Quality and Safety in Health Care*, *13*, 233–234.

Boerlin, P. (1997). Applications of multilocus enzyme electrophoresis in medical microbiology. *Journal of Microbiological Methods*, *28*, 221–231.

Bolton, D. J. (2015). *Campylobacter* virulence and survival factors. *Food Microbiology*, *48*, 99–108.

Breuer, T., Benkel, D. H., Shapiro, R. L., Hall, W. N., Winnett, M. M., Linn, M. J., Neimann, J., Barrett, T. J., Dietrich, S., Downes, F. P., et al. (2001). A multistate outbreak of *Escherichia coli* O157:H7 infections linked to alfalfa sprouts grown from contaminated seeds. *Emerging Infectious Diseases*, *7*, 977–982.

Bright, J. J., Claydon, M. A., Soufian, M., & Gordon, D. B. (2002). Rapid typing of bacteria using matrix-assisted laser desorption ionisation time-of-flight mass spectrometry and pattern recognition software. *Journal of Microbiological Methods*, *48*, 127–138.

Callicott, K. A., Frioriksdóttir, V., Reiersen, J., Lowman, R., Bisaillon, J. R., Gunnarsson, E., Berndtson, E., Hiett, K. L., Needleman, D. S., & Stern, N. J. (2006). Lack of evidence for vertical transmission of *Campylobacter* spp. in chickens. *Applied and Environmental Microbiology*, *72*, 5794–5798.

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

Carriço, J. A., Silva-Costa, C., Melo-Cristino, J., Pinto, F. R., De Lencastre, H., Almeida, J. S., & Ramirez, M. (2006). Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant *Streptococcus pyogenes*. *Journal of Clinical Microbiology*, *44*, 2524–2532.

Carrillo, C. D., Kruczkiewicz, P., Mutschall, S., Tudor, A., Clark, C., & Taboada, E. N. (2012). A Framework for Assessing the Concordance of Molecular Typing Methods and the True Strain Phylogeny of *Campylobacter jejuni* and C. *coli* Using Draft Genome Sequence Data. *Frontiers in Cellular and Infection Microbiology*, *2*, 1–12.

Centers for Disease Control (2014). Serotypes and the importance of serotyping *Salmonella*. Retrieved from http://www.cdc.gov/salmonella/reportspubs/salmonella-atlas/serotyping-importance.html. Accessed 15 Jan, 2015.

Champion, O. L., Best, E. L., & Frost, J. A. (2002). Comparison of pulsed-field gel electrophoresis and amplified fragment length polymorphism techniques for investigating outbreaks of enteritis due to campylobacters. *Journal of Clinical Microbiology*, *40*, 2263–2265.

Chon, J. W., Kim, H., Yim, J. H., Park, J. H., Kim, M. S., & Seo, K. H. (2013). Development of a selective enrichment broth supplemented with bacteriological charcoal and a high concentration of polymyxin B for the detection of *Campylobacter jejuni* and *Campylobacter coli* in chicken carcass rinses. *International Journal of Food Microbiology*, *162*, 308–310.

Clark, C. G., Bryden, L., Cuff, W. R., Johnson, P. L., Jamieson, F., Ciebin, B., & Wang, G. (2005). Use of the Oxford multilocus sequence typing protocol and sequencing of the flagellin short variable region to characterize isolates from a large outbreak of waterborne *Campylobacter* spp. strains in Walkerton, Ontario, Canada. *Journal of Clinical Microbiology*, *43*, 2080–2091.

Clark, C. G., Kruczkiewicz, P., Guan, C., McCorrister, S. J., Chong, P., Wylie, J., van Caeseele, P., Tabor, H. A., Snarr, P., Gilmour, M. W., et al. (2013). Evaluation of MALDI-TOF mass spectroscopy methods for determination of *Escherichia coli* pathotypes. *Journal of Microbiological Methods*, *94*, 180–191.

Clark, C. G., Taboada, E., Grant, C. C. R., Blakeston, C., Pollari, F., Marshall, B., Rahn, K., MacKinnon, J., Daignault, D., Pillai, D., et al. (2012). Comparison of molecular typing methods useful for detecting clusters of *Campylobacter jejuni* and C. *coli* isolates through routine surveillance. *Journal of Clinical Microbiology*, *50*, 798–809.

Coker, A. O., Isokpehi, R. D., Thomas, B. N., Amisu, K. O., & Larry Obi, C. (2002). Human campylobacteriosis in developing countries. *Emerging Infectious Diseases*, *8*, 237–243.

Corry, J., Post, D., Colin, P., & Laisney, M. (1995). Culture media for the isolation of campylobacters. *International Journal of Food Microbiology*, *26*, 43–76.

Craigie, J. & Yen, C. H. (1938). The demonstration of types of *B. typhosus* by means of preparation of type II *Vi* phage. *Canadian Journal of Public Health*, *29*, 448–463.

Croll, D. & McDonald, B. A. (2012). The accessory genome as a cradle for adaptive evolution in pathogens. *PLoS Pathogens*, *8*, 8–10.

Dagerhamn, J., Blomberg, C., Browall, S., Sjöström, K., Morfeldt, E., & Henriques-Normark, B. (2008). Determination of accessory gene patterns predicts the same relatedness among strains of *Streptococcus pneumoniae* as sequencing of housekeeping genes does and represents a novel approach in molecular epidemiology. *Journal of Clinical Microbiology*, *46*, 863–868.

Deckert, A. E., Taboada, E., Mutschall, S., Poljak, Z., Reid-Smith, R. J., Tamblyn, S., Morrell, L., Seliske, P., Jamieson, F. B., Irwin, R., et al. (2014). Molecular epidemiology of *Campylobacter jejuni* human and chicken isolates from two health units. *Foodborne Pathogens and Disease*, *11*, 150–5.

Dicker, R. C., Coronado, F., Koo, D., & Parrish, R. G. (2006). *Principles of Epidemiology in Public Health Practice : an Introduction to Applied Epidemiology and Biostatistics*. US Centers for Disease Control and Prevention (CDC), Atlanta, GA, 3rd edition.

Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. A., & Crook, D. W. (2012). Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews Genetics*, *13*, 601–612.

Dingle, K. E., Colles, F. M., Ure, R., Wagenaar, J. A., Duim, B., Bolton, F. J., Fox, A. J., Wareing, D. R. A., & Maiden, M. C. J. (2002). Molecular characterization of *Campylobacter jejuni* clones: a basis for epidemiologic investigation. *Emerging Infectious Diseases*, *8*, 949–955.

Dingle, K. E., Colles, F. M., Wareing, D. R. A., Ure, R., Fox, A. J., Bolton, F. E., Bootsma, H. J., Willems, R. J. L., Urwin, R., & Maiden, M. C. J. (2001). Multilocus sequence typing system for *Campylobacter jejuni*. *Journal of Clinical Microbiology*, *39*, 14–23.

Dingle, K. E., McCarthy, N. D., Cody, A. J., Peto, T. E. A., & Maiden, M. C. J. (2008). Extended sequence typing of *Campylobacter* spp., United Kingdom. *Emerging Infectious Diseases*, *14*, 1620–1622.

Duncan, D. F. (1988). Mankind's changing concepts of disease. In *Epidemiology: Basis for Disease Prevention and Health Promotion*, pp. 11–26. Macmillan, New York.

Eberle, K. N. & Kiess, A. S. (2012). Phenotypic and genotypic methods for typing *Campylobacter jejuni* and *Campylobacter coli* in poultry. *Poultry Science*, *91*, 255–64.

Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P., & Spratt, B. G. (2004). eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *Journal of Bacteriology*, *186*, 1518–1530.

Fleming, A. (1942). A simple method of using Penicillin, Tellurite, and Gentian Violet for differential culture. *British Medical Journal*, *1*, 547–548.

Foxman, B. & Riley, L. (2001). Molecular epidemiology: focus on infection. *American Journal of Epidemiology*, *153*, 1135–1141.

Francisco, A. P., Bugalho, M., Ramirez, M., & Carriço, J. A. (2009). Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. *BMC Bioinformatics*, *10*, 152.

Francisco, A. P., Vaz, C., Monteiro, P. T., Melo-Cristino, J., Ramirez, M., & Carriço, J. A. (2012). PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinformatics*, *13*, 87.

Freeman, M. C., Stocks, M. E., Cumming, O., Jeandron, A., Higgins, J. P. T., Wolf, J., Prüss-Ustün, A., Bonjour, S., Hunter, P. R., Fewtrell, L., et al. (2014). Hygiene and health: systematic review of handwashing practices worldwide and update of health effects. *Tropical Medicine & International Health*, *19*, 906–16.

French, N., Barrigas, M., Brown, P., Ribiero, P., Williams, N., Leatherbarrow, H., Birtles, R., Bolton, E., Fearnhead, P., & Fox, A. (2005). Spatial epidemiology and natural population structure of *Campylobacter jejuni* colonizing a farmland ecosystem. *Environmental Microbiology*, *7*, 1116–26.

French, N. & the Molecular Epidemiology and Veterinary Public Health Group (2008). Enhancing surveillance of potentially foodborne enteric diseases in New Zealand : human campylobacteriosis in the Manawatu. Technical report, Hopkirk Institute, Massey University.

Friedman, A. & Kao, C.-Y. (2014). Epidemiology of Infectious Disease. In A. Stevens & M. Mackey, editors, *Lecture Notes on Mathematical Modelling in the Life Sciences*, pp. 33–47. Springer International Publishing, Switzerland.

Friedman, C. R., Hoekstra, R. M., Samuel, M., Marcus, R., Bender, J., Shiferaw, B., Reddy, S., Ahuja, S. D., Helfrick, D. L., Hardnett, F., et al. (2004). Risk factors for sporadic *Campylobacter* infection in the United States: a case-control study in FoodNet sites. *Clinical Infectious Diseases*, *38*, S285–S296.

Garaizar, J., Rementeria, A., & Porwollik, S. (2006). DNA microarray technology: a new tool for the epidemiological typing of bacterial pathogens? *FEMS Immunology and Medical Microbiology*, *47*, 178–189.

Garcia, M. M., Lior, H., Stewart, R. B., Ruckerbauer, G. M., Trudel, J. R., & Skljarevski, A. (1985). Isolation, characterization, and serotyping of *Campylobacter jejuni* and *Campylobacter coli* from slaughter cattle. *Applied and Environmental Microbiology*, *49*, 667–672.

Gardy, J. L., Johnston, J. C., Ho Sui, S. J., Cook, V. J., Shah, L., Brodkin, E., Rempel, S., Moore, R., Zhao, Y., Holt, R., et al. (2011). Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *New England Journal of Medicine*, *364*, 730–739.

Gerner-Smidt, P., Hise, K., Kincaid, J., Hunter, S., Rolando, S., Hyytiä-Trees, E., Ribot, E. M., Swaminathan, B., & the PulseNet Taskforce (2006). PulseNet USA: a Five-Year Update. *Foodborne Pathogens and Disease*, *3*, 9–19.

Gharst, G., Hanson, D., & Kathariou, S. (2006). Effect of direct culture versus selective enrichment on the isolation of thermophilic *Campylobacter* from feces of mature cattle at harvest. *Journal of Food Protection*, *69*, 1024–1027.

Goering, R. V. (2010). Pulsed field gel electrophoresis: a review of application and interpretation in the molecular epidemiology of infectious disease. *Infection, Genetics and Evolution*, *10*, 866–875.

Grad, Y. H., Lipsitch, M., Feldgarden, M., Arachchi, H. M., Cerqueira, G. C., FitzGerald, M., Godfrey, P., Haas, B. J., Murphy, C. I., Russ, C., et al. (2012). Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. *Proceedings of the National Academy of Sciences*, *109*, 3065–3070.

Gripp, E., Hlahla, D., Didelot, X., Kops, F., Maurischat, S., Tedin, K., Alter, T., Ellerbroek, L., Schreiber, K., Schomburg, D., et al. (2011). Closely related *Campylobacter jejuni* strains from different sources reveal a generalist rather than a specialist lifestyle. *BMC Genomics*, *12*, 584.

Grundmann, H., Hori, S., & Tanner, G. (2001). Determining confidence intervals and the discriminatory abilities of typing methods for microorganisms. *Journal of Clinical Microbiology*, *39*, 4190–4192.

Habib, I., Sampers, I., Uyttendaele, M., Berkvens, D., & De Zutter, L. (2008). Performance characteristics and estimation of measurement uncertainty of three plating procedures for *Campylobacter* enumeration in chicken meat. *Food Microbiology*, *25*, 65–74.

Habib, I., Uyttendaele, M., & De Zutter, L. (2011). Evaluation of ISO 10272:2006 standard versus alternative enrichment and plating combinations for enumeration and detection of *Campylobacter* in chicken meat. *Food Microbiology*, *28*, 1117–1123.

Hall, B. G., Ehrlich, G. D., & Hu, F. Z. (2010). Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. *Microbiology*, *156*, 1060–1068.

Haq, I. U., Chaudhry, W. N., Akhtar, M. N., Andleeb, S., & Qadri, I. (2012). Bacteriophages and their implications on future biotechnology: a review. *Virology Journal*, *9*, 9.

Harding, S., Parmley, J., & Morrison, K. (2014). Using participatory epidemiology to assess factors contributing to common enteric pathogens in Ontario: results from a workshop held at the Ontario Veterinary College, University of Guelph, Ontario. *BMC Public Health*, *14*, 405.

Hardison, R. C. (2003). Comparative Genomics. *PLoS Biol*, *1*, e58.

Hayashi, M., Kubota-Hayashi, S., Natori, T., Mizuno, T., Miyata, M., Yoshida, S., Zhang, J., Kawamoto, K., Ohkusu, K., Makino, S., et al. (2013). Use of blood-free enrichment broth in the development of a rapid protocol to detect *Campylobacter* in twenty-five grams of chicken meat. *International Journal of Food Microbiology*, *163*, 41–46.

Herman, L., Heyndrickx, M., Grijspeerdt, K., Vandekerchove, D., Rollier, I., & De Zutter, L. (2003). Routes for *Campylobacter* contamination of poultry meat: epidemiological study from hatchery to slaughterhouse. *Epidemiology and Infection*, *131*, 1169–1180.

Heuer, O. E., Pedersen, K., Andersen, J. S., & Madsen, M. (2001). Prevalence and antimicrobial susceptibility of thermophilic *Campylobacter* in organic and conventional broiler flocks. *Letters in Applied Microbiology*, *33*, 269–274.

Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Humphrey, T., O'Brien, S., & Madsen, M. (2007). Campylobacters as zoonotic pathogens: a food production perspective. *International Journal of Food Microbiology*, *117*, 237–257.

Hunt, J., Abeyta, C., & Tran, T. (2001). Isolation of *Campylobacter* Species from Food and Water. Retrieved from http://www.fda.gov/Food/FoodScienceResearch/LaboratoryMethods/ucm072616.htm. Accessed 19 Jun, 2015.

Hunter, P. R. & Gaston, M. A. (1988). Numerical index of the discriminatory ability of typing systems : an application of Simpson's index of Diversity. *Journal of Clinical Microbiology*, *26*, 2465–2466.

Hutchinson, D. & Bolton, F. (1984). Improved blood free selective medium for isolating *Campylobacter jejuni* from faecal specimens. *Journal of Clinical Pathology*, *37*, 956–957.

ISO (2006). ISO 10272-1:2006. Microbiology of food and animal feeding stuffs. Horizontal method for detection and enumeration of *Campylobacter* spp. Part 1: Detection method. Technical report, International Standards Organization.

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

Jolley, K. A. & Maiden, M. C. J. (2010). BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, *11*, 595.

Jones, K. (2001). Campylobacters in water, sewage and the environment. *Journal of Applied Microbiology Symposium Supplement*, *90*, 68S–79S.

Jørgensen, F., Bailey, R., Williams, S., Henderson, P., Wareing, D. R. A., Bolton, F. J., Frost, J. A., Ward, L., & Humphrey, T. J. (2002). Prevalence and numbers of *Salmonella* and *Campylobacter* spp. on raw, whole chickens in relation to sampling methods. *International Journal of Food Microbiology*, *76*, 151–164.

Kapperud, G., Espeland, G., Wahl, E., Walde, A., Herikstad, H., Gustavsen, S., Tveits, I., Natas, O., Bevanger, L., & Digranes, A. (2003). Factors associated with increased and decreased risk of *Campylobacter* infection: a prospective case-control study in Norway. *American Journal of Epidemiology*, *158*, 234–242.

Khan, I. U. H., Gannon, V., Jokinen, C. C., Kent, R., Koning, W., Lapen, D. R., Medeiros, D., Miller, J., Neumann, N. F., Phillips, R., et al. (2014). A national investigation of the prevalence and diversity of thermophilic *Campylobacter* species in agricultural watersheds in Canada. *Water Research*, *61*, 243–252.

Kiess, A. S., Parker, H. M., & McDaniel, C. D. (2010). Evaluation of different selective media and culturing techniques for the quantification of *Campylobacter* spp. from broiler litter. *Poultry Science*, *89*, 1755–1762.

Köser, C. U., Ellington, M. J., Cartwright, E. J. P., Gillespie, S. H., Brown, N. M., Farrington, M., Holden, M. T. G., Dougan, G., Bentley, S. D., Parkhill, J., et al. (2012). Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathogens*, *8*.

Kramer, J. M., Frost, J. A., Bolton, F. J., & Wareing, D. R. (2000). *Campylobacter* contamination of raw meat and poultry at retail sale: identification of multiple types and comparison with isolates from human infection. *Journal of Food Protection*, *63*, 1654–1659.

Krishnamurthy, T. & Ross, P. L. (1996). Rapid identification of bacteria by direct matrix-assisted laser desorption/ionization mass spectrometric analysis of whole cells. *Rapid Communications in Mass Spectrometry*, *10*, 1992–1996.

Krishnamurthy, T., Ross, P. L., & Rajamani, U. (1996). Detection of pathogenic and non-pathogenic bacteria by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, *10*, 883–888.

Kruczkiewicz, P., Mutschall, S., Barker, D., Thomas, J., Van Domselaar, G., Gannon, V. P., Carrillo, C. D., & Eduardo, T. (2013). MIST: A Tool for Rapid in silico Generation of Molecular Data from Bacterial Genome Sequences. In *Proceedings of Bioinformatics 2013: 4th International Conference on Bioinformatics Models, Methods and Algorithms*, pp. 316–323. Barcelona, Spain.

Kwan, P. S. L., Barrigas, M., Bolton, F. J., French, N. P., Gowland, P., Kemp, R., Leatherbarrow, H., Upton, M., & Fox, A. J. (2008a). Molecular epidemiology of *Campylobacter jejuni* populations in dairy cattle, wildlife, and the environment in a farmland area. *Applied and Environmental Microbiology*, *74*, 5130–5138.

Kwan, P. S. L., Birtles, A., Bolton, F. J., French, N. P., Robinson, S. E., Newbold, L. S., Upton, M., & Fox, A. J. (2008b). Longitudinal study of the molecular epidemiology of *Campylobacter jejuni* in cattle on dairy farms. *Applied and Environmental Microbiology*, *74*, 3626–3633.

Lagier, J.-C., Edouard, S., Pagnier, I., Mediannikov, O., Drancourt, M., & Raoult, D. (2015a). Current and past strategies for bacterial culture in clinical microbiology. *Clinical Microbiology Reviews*, *28*, 208–236.

Lagier, J.-C., Hugon, P., Khelaifia, S., Fournier, P.-E., La Scola, B., & Raoult, D. (2015b). The rebirth of culture in microbiology through the example of culturomics to study human gut microbiota. *Clinical Microbiology Reviews*, *28*, 237–264.

Laing, C., Pegg, C., Yawney, D., Ziebell, K., Steele, M., Johnson, R., Thomas, J. E., Taboada, E. N., Zhang, Y., & Gannon, V. P. J. (2008). Rapid determination of *Escherichia coli* O157:H7 lineage types and molecular subtypes by using comparative genomic fingerprinting. *Applied and Environmental Microbiology*, *74*, 6606–6615.

Lancefield, R. C. (1933). A serological differentiation of human and other groups of hemolytic streptococci. *Journal of Experimental Medicine*, *57*, 571–595.

Larsen, M. V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R. L., Jelsbak, L., Sicheritz-Pontén, T., Ussery, D. W., Aarestrup, F. M., et al. (2012). Multilocus sequence typing of total-genome-sequenced bacteria. *Journal of Clinical Microbiology*, *50*, 1355–1361.

Le Roux, E. & Lastovica, A. (1998). The cape town protocol: how to isolate the most campylobacters for your dollar, pound, franc, yen, etc. In *Proceedings of the 9th International Workshop on Campylobacter, Helicobacter and Related Organisms*, pp. 31–33. Cape Town, South Africa: Institute of Child Health.

Lee, K., Iwata, T., Nakadai, A., Kato, T., Hayama, S., Taniguchi, T., & Hayashidani, H. (2011). Prevalence of *Salmonella*, *Yersinia* and *Campylobacter* spp. in feral raccoons (Procyon lotor) and masked palm civets (Paguma larvata) in Japan. *Zoonoses and Public Health*, *58*, 424–431.

Leonard, E. E., Takata, T., Blaser, M. J., Falkow, S., Tompkins, L. S., & Gaynor, E. C. (2003). Use of an open-reading frame-specific *Campylobacter jejuni* DNA microarray as a new genotyping tool for studying epidemiologically related isolates. *The Journal of Infectious Diseases*, *187*, 691–694.

Leonard, E. E., Tompkins, L. S., Falkow, S., & Nachamkin, I. (2004). Comparison of *Campylobacter jejuni* isolates implicated in Guillain-Barré syndrome and strains that cause enteritis by a DNA microarray. *Infection and Immunity*, *72*, 1199–1203.

Lior, H. (1984). New, extended biotyping scheme for *Campylobacter jejuni*, *Campylobacter coli*, and "*Campylobacter laridis*". *Journal of Clinical Microbiology*, *20*, 636–640.

Louis, V. R., Gillespie, I. A., O'Brien, S. J., Pussek-Cohen, E., Pearson, A. D., Colwell, R. R., Louis, R., Brien, S. J. O., & Russek-Cohen, E. (2005). Temperature-driven *Campylobacter* seasonality in England and Wales. *Applied and Environmental Microbiology*, *71*, 85–92.

Lucchini, S., Thompson, A., & Hinton, J. C. D. (2001). Microarrays for microbiologists. *Microbiology*, *147*, 1403–1414.

Lunt, D. H., Whipple, L. E., & Hyman, B. C. (1998). Mitochondrial DNA variable number tandem repeats (VNTRs): utility and problems in molecular ecology. *Molecular Ecology*, *7*, 1441–1455.

MacDonald, D. M., Fyfe, M., Paccagnella, A., Trinidad, A., Louie, K., & Patrick, D. (2004). *Escherichia coli* O157:H7 outbreak linked to salami, British Columbia, Canada, 1999. *Epidemiology and Infection*, *132*, 283–289.

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America*, *95*, 3140–3145.

Majowicz, S. E., McNab, W. B., Sockett, P., Henson, T. S., Doré, K., Edge, V. L., Buffett, M. C., Fazil, A., Read, S., McEwen, S., et al. (2006). Burden and cost of gastroenteritis in a Canadian community. *Journal of Food Protection*, *69*, 651–659.

Manning, G., Dowson, C. G., Mary, C., Ahmed, I. H., West, M., Newell, D. G., & Bagnall, M. C. (2003). Multilocus sequence typing for comparison of veterinary and human isolates of *Campylobacter jejuni* multilocus sequence typing for comparison of veterinary and human isolates of *Campylobacter jejuni*. *Applied and Environmental Microbiology*, *69*, 6370–6379.

Maslow, J., Mulligan, M., & Arbeit, R. (1993). Molecular epidemiology: application of contemporary techniques to the typing of microorganisms. *Clinical Infectious Diseases*, *17*, 153–162.

Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics and Development*, *15*, 589–594.

Muellner, P., Marshall, J. C., Spencer, S. E. F., Noble, A. D., Shadbolt, T., Collins-Emerson, J. M., Midwinter, A. C., Carter, P. E., Pirie, R., Wilson, D. J., et al. (2011). Utilizing a combination of molecular and spatial tools to assess the effect of a public health intervention. *Preventive Veterinary Medicine*, *102*, 242–253.

Müllner, P., Collins-Emerson, J. M., Midwinter, A. C., Carter, P., Spencer, S. E. F., Van Der Logt, P., Hathaway, S., & French, N. P. (2010). Molecular epidemiology of *Campylobacter jejuni* in a geographically isolated country with a uniquely structured poultry industry. *Applied and Environmental Microbiology*, *76*, 2145–2154.

Müllner, P., Spencer, S. E. F., Wilson, D. J., Jones, G., Noble, A. D., Midwinter, A. C., Collins-Emerson, J. M., Carter, P., Hathaway, S., & French, N. P. (2009). Assigning the source of human campylobacteriosis in New Zealand: a comparative genetic and epidemiological approach. *Infection, Genetics and Evolution*, *9*, 1311–1319.

Murray, P. R. (2010). Matrix-assisted laser desorption ionization time-of-flight mass spectrometry: usefulness for taxonomy and epidemiology. *Clinical Microbiology and Infection*, *16*, 1626–1630.

Musgrove, M. T., Berrang, M. E., Byrd, J. A., Stern, N. J., & Cox, N. A. (2001). Detection of *Campylobacter* spp. in ceca and crops with and without enrichment. *Poultry Science*, *80*, 825–828.

Mutschall, S., Cook, A., Hetman, B., Kruczkiewicz, P., Pintar, K., Marshall, B., Pollari, F., & Taboada, E. N. (2013). Molecular epidemiology of *Campylobacter* isolates from multiple sources within a Canadian sentinel surveillance site. *17th International Workshop on Campylobacter, Helicobacter and Related Organisms. Aberdeen, Scotland*.

Nachamkin, I. (2002). Chronic effects of *Campylobacter* infection. *Microbes and Infection*, *4*, 399–403.

Nachamkin, I., Allos, B., & Ho, T. (1998). *Campylobacter* species and Guillain-Barré syndrome. *Clinical Microbiology Reviews*, *11*, 555–567.

Nachamkin, I., Bohachick, K., & Patton, C. M. (1993). Flagellin gene typing of *Campylobacter jejuni* by restriction fragment length polymorphism analysis. *Journal of Clinical Microbiology*, *31*, 1531–1536.

Neimann, J., Engberg, J., Mølbak, K., & Wegener, H. C. (2003). A case-control study of risk factors for sporadic *Campylobacter* infections in Denmark. *Epidemiology and Infection*, *130*, 353–366.

Nesbitt, A. & Ravel, A. (2012). Integrated surveillance and potential sources of *Salmonella enteritidis* in human cases in Canada from 2003 to 2009. *Epidemiology and Infection*, *140*, 1757–72.

Neuwirth, E. (2015). RColorBrewer: ColorBrewer palettes. Retrieved from https://cran.r-project.org/web/packages/RColorBrewer/index.html. Accessed 15 Jul, 2015.

Newell, D. G., Elvers, K. T., Dopfer, D., Hansson, I., Jones, P., James, S., Gittins, J., Stern, N. J., Davies, R., Connerton, I., et al. (2011). Biosecurity-based interventions and strategies to reduce *Campylobacter* spp. on poultry farms. *Applied and Environmental Microbiology*, *77*, 8605–14.

Nichols, G. L. (2005). Fly transmission of *Campylobacter*. *Emerging Infectious Diseases*, *11*, 361–364.

Nichols, G. L., Richardson, J. F., Sheppard, S. K., Lane, C., & Sarran, C. (2012). *Campylobacter* epidemiology: a descriptive study reviewing 1 million cases in England and Wales between 1989 and 2011. *BMJ Open*, *2*, e001179.

Nylen, G., Dunstan, F., Palmer, S. R., Andersson, Y., Bager, F., Cowden, J., Feierl, G., Galloway, Y., Kapperud, G., Megraud, F., et al. (2002). The seasonal distribution of *Campylobacter* infection in nine European countries and New Zealand. *Epidemiology and Infection*, *128*, 383–390.

Ochman, H., Whittam, T. S., Caugant, D. A., & Selander, R. K. (1983). Enzyme polymorphism and genetic population structure in *Escherichia coli* and *Shigella*. *Journal of General Microbiology*, *129*, 2715–2726.

On, S. L. & Holmes, B. (1995). Classification and identification of campylobacters, helicobacters and allied taxa by numerical analysis of phenotypic characters. *Systematic and Applied Microbiology*, *18*, 374–390.

Parkhill, J., Wren, B., Mungall, K., & Ketley, J. (2000). The genome sequence of the foodborne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, *403*, 665–668.

Pitkänen, T. (2013). Review of *Campylobacter* spp. in drinking and environmental waters. *Journal of Microbiological Methods*, *95*, 39–47.

Public Health Agency of Canada (2006). FoodNet Canada: Sentinel sites - Public Health Agency Canada. Retrieved from http://www.phac-aspc.gc.ca/foodnetcanada/necessity-importance-eng.php. Accessed 12 Jan, 2016.

Public Health Agency of Canada (2010). National Integrated Enteric Disease Surveillance Program: Sample Collection, Preparation and Laboratory Methodologies. Retrieved from http://www.phac-aspc.gc.ca/foodnetcanada/niedsp10-pnisme10/index-eng.php. Accessed 12 Jan, 2016.

Public Health Agency of Canada (2016). Canadian Communicable Disease Report : Antimicrobial Stewardship. Retrieved from http://www.phac-aspc.gc.ca/publicat/ccdr-rmtc/15vol41/dr-rm41s-4/assets/pdf/15vol41-s4-eng.pdf. Accessed 12 Jan, 2016.

R Core Team (2015). R: a language and environment for statistical computing. Retrieved from http://www.r-project.org. Accessed 15 Jun, 2015.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*, 846–850.

Rasko, D. & Webster, D. (2011). Origins of the *E. coli* strain causing an outbreak of hemolyticuremic syndrome in Germany. *New England Journal of Medicine*, *365*, 709–717.

Rohde, H., Qin, J., Cui, Y., Li, D., Loman, N. J., Hentschke, M., Chen, W., Pu, F., Peng, Y., Li, J., et al. (2011). Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *New England Journal of Medicine*, *365*, 718–724.

Sabat, A. J., Budimir, A., Nashev, D., Sá-Leão, R., van Dijl, J., Laurent, F., Grundmann, H., Friedrich, A. W., & the ESCMID Study Group of Epidemiological Markers (ESGEM) (2013). Overview of molecular typing methods for outbreak detection and epidemiological surveillance. *Euro Surveillance : European Communicable Disease Bulletin*, *18*.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., & Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, *239*, 487–491.

Sails, A. D., Swaminathan, B., & Fields, P. I. (2003). Utility of multilocus sequence typing as an epidemiological tool for investigation of outbreaks of gastroenteritis caused by *Campylobacter jejuni* utility of multilocus sequence typing as an epidemiological tool for investigation of outbreaks of gastroenteritis. *Journal of Clinical Microbiology*, *41*, 4733–4739.

Salmonella-Subcommittee of the International Society for Microbiology (1934). The genus *Salmonella Lignières*, 1900: issued by the *Salmonella* subcommittee of the nomenclature committee of the International Society for Microbiology. *Journal of Hygiene*, *34*, 333–350.

Savelkoul, P., Aarts, H., de Haas, J., & Dijkshoorn, B. (1999). Amplified-fragment length polymorphism analysis: the state of an art. *Journal of Clinical Microbiology*, *37*, 3083–3091.

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

Sears, A., Baker, M. G., Wilson, N., Marshall, J., Muellner, P., Campbell, D. M., Lake, R. J., & French, N. P. (2011). Marked campylobacteriosis decline after interventions aimed at poultry, New Zealand. *Emerging Infectious Diseases*, *17*, 1007–1015.

Selander, R. K. & Levin, B. R. (1980). Genetic Diversity and Structure in *Escherichia coli* Populations. *Science*, *210*, 545–547.

Seng, P., Drancourt, M., Gouriet, F., La Scola, B., Fournier, P.-E., Rolain, J. M., & Raoult, D. (2009). Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Clinical Infectious Diseases*, *49*, 543–551.

Severiano, A., Pinto, F. R., Ramirez, M., & Carriço, J. A. (2011). Adjusted Wallace coefficient as a measure of congruence between typing methods. *Journal of Clinical Microbiology*, *49*, 3997–4000.

Shanker, S., Lee, A., & Sorrell, T. (1990). Broiler Chicks : Experimental Studies. *Epidemiology and Infection*, pp. 101–110.

Sheppard, S. K., Cheng, L., Méric, G., De Haan, C. P. A., Llarena, A. K., Marttinen, P., Vidal, A., Ridley, A., Clifton-Hadley, F., Connor, T. R., et al. (2014). Cryptic ecology among host generalist *Campylobacter jejuni* in domestic animals. *Molecular Ecology*, *23*, 2442–2451.

Sheppard, S. K., Colles, F. M., McCarthy, N. D., Strachan, N. J. C., Ogden, I. D., Forbes, K. J., Dallas, J. F., & Maiden, M. C. J. (2011). Niche segregation and genetic structure of *Campylobacter jejuni* populations from wild and agricultural host species. *Molecular Ecology*, *20*, 3484–3490.

Sheppard, S. K., Dallas, J. F., MacRae, M., McCarthy, N. D., Sproston, E. L., Gormley, F. J., Strachan, N. J. C., Ogden, I. D., Maiden, M. C. J., & Forbes, K. J. (2009a). *Campylobacter* genotypes from food animals, environmental sources and clinical disease in Scotland 2005/6. *International Journal of Food Microbiology*, *134*, 96–103.

Sheppard, S. K., Dallas, J. F., Strachan, N. J. C., MacRae, M., McCarthy, N. D., Wilson, D. J., Gormley, F. J., Falush, D., Ogden, I. D., Maiden, M. C. J., et al. (2009b). *Campylobacter* genotyping to determine the source of human infection. *Clinical Infectious Diseases*, *48*, 1072–1078.

Sheppard, S. K., Didelot, X., Jolley, K. A., Darling, A. E., Pascoe, B., Meric, G., Kelly, D. J., Cody, A., Colles, F. M., Strachan, N. J. C., et al. (2013). Progressive genome-wide introgression in agricultural *Campylobacter coli*. *Molecular Ecology*, *22*, 1051–1064.

Sheppard, S. K., Jolley, K. A., & Maiden, M. C. J. (2012). A gene-by-gene approach to bacterial population genomics: whole genome MLST of *Campylobacter*. *Genes*, *3*, 261–277.

Sheppard, S. K., McCarthy, N. D., Falush, D., & Maiden, M. C. J. (2008). Convergence of *Campylobacter* species: implications for bacterial evolution. *Science*, *320*, 237–239.

Sheth, A. N., Hoekstra, M., Patel, N., Ewald, G., Lord, C., Clarke, C., Villamil, E., Niksich, K., Bopp, C., Nguyen, T.-A., et al. (2011). A national outbreak of *Salmonella* serotype Tennessee infections from contaminated peanut butter: a new food vehicle for salmonellosis in the United States. *Clinical Infectious Diseases*, *53*, 356–362.

Silva, J., Leite, D., Fernandes, M., Mena, C., Gibbs, P. A., & Teixeira, P. (2011). *Campylobacter* spp. as a foodborne pathogen: a review. *Frontiers in Microbiology*, *2*, 1–12.

Simpson, E. H. (1949). Measurement of Diversity. *Nature*, *163*, 688–688.

Singleton, P. & Sainsbury, D. (2007). Serotyping. In *Dictionary of Microbiology and Molecular Biology*, pp. 682–757. John Wiley & Sons, Ltd.

Sinnott, R. W. (1984). Virtues of the haversine. *Sky and Telescope*, *68*, 158.

Skirrow, M. & Benjamin, J. (1980). Differentiation of enteropathogenic *Campylobacter*. *Journal of Clinical Pathology*, *33*, 1122.

Soloman, E. & Hoover, D. (1999). *Campylobacter jejuni*: a bacterial paradox. *Journal of Food Safety*, *19*, 121–136.

Stanley, K. N., Wallace, J. S., Currie, J. E., Diggle, P. J., & Jones, K. (1998). The seasonal variation of thermophilic campylobacters in beef cattle, dairy cattle and calves. *Journal of Applied Microbiology*, *85*, 472–480.

Stanley, T. G. & Wilson, I. (2003). Multilocus enzyme electrophoresis: a practical guide. *Molecular Biotechnology*, *24*, 203–220.

Stevenson, L. G., Drake, S. K., Shea, Y. R., Zelazny, A. M., & Murray, P. R. (2010). Evaluation of matrix-assisted laser desorption ionization - time of flight mass spectrometry for identification of clinically important yeast species. *Journal of Clinical Microbiology*, *48*, 3482–3486.

Swaminathan, B., Barrett, T. J., Hunter, S. B., & Tauxe, R. V. (2001). PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerging Infectious Diseases*, *7*, 382–389.

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

Taboada, E., Ross, S., Mutschall, S., MacKinnon, J., Roberts, M., Buchanan, C., Kruczkiewicz, P., Jokinen, C., Thomas, J., Nash, J., et al. (2012). Development and validation of a comparative genomic fingerprinting method for high-resolution genotyping of *Campylobacter jejuni*. *Journal of Clinical Microbiology*, *50*, 788–797.

Taboada, E. N., Mackinnon, J. M., Luebbert, C. C., Gannon, V. P. J., Nash, J. H. E., & Rahn, K. (2008). Comparative genomic assessment of multi-locus sequence typing: rapid accumulation of genomic heterogeneity among clonal isolates of *Campylobacter jejuni*. *BMC Evolutionary Biology*, *8*, 229.

Taboada, E. N., van Belkum, A., Yuki, N., Acedillo, R. R., Godschalk, P. C., Koga, M., Endtz, H. P., Gilbert, M., & Nash, J. H. (2007). Comparative genomic analysis of

Campylobacter jejuni associated with Guillain-Barré and Miller Fisher syndromes: neuropathogenic and enteritis-associated isolates can share high levels of genomic similarity. *BMC Genomics*, *8*, 359.

Tam, C. C., Higgins, C. D., Neal, K. R., Rodrigues, L. C., Millership, S. E., O'Brien, S. J., & on behalf of the *Campylobacter* Case-Control Study Group (2009). Chicken consumption and use of acid-suppressing medications as risk factors for *Campylobacter* enteritis, England. *Emerging Infectious Diseases*, *15*, 1402–1408.

Tannock, I., Hill, R., Bristow, R., & Harrington, L. (2013). *Basic Science of Oncology, Fifth Edition*. McGraw Hill Professional.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 13950–13955.

The European Food Safety Authority (2015). The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2013. *European Food Safety Authority Journal*, *13*, 1–162.

Thomas, M. K., Murray, R., Flockhart, L., Pintar, K., Pollari, F., Fazil, A., Nesbitt, A., & Marshall, B. (2013). Estimates of the burden of foodborne illness in Canada for 30 specified pathogens and unspecified agents, circa 2006. *Foodborne Pathogens and Disease*, *10*, 639–48.

Ugarte-Ruiz, M., Gómez-Barrero, S., Porrero, M., Álvarez, J., García, M., Comerón, M., Wassenaar, T., & Domínguez, L. (2012). Evaluation of four protocols for the detection and isolation of thermophilic *Campylobacter* from different matrices. *Journal of Applied Microbiology*, *113*, 200–8.

Ugarte-Ruiz, M., Wassenaar, T. M., Gómez-Barrero, S., Porrero, M. C., Navarro-Gonzalez, N., & Domínguez, L. (2013). The effect of different isolation protocols on detection and molecular characterization of *Campylobacter* from poultry. *Letters in Applied Microbiology*, *57*, 427–35.

Vavrek, M. M. J. (2015). Fossil: Palaeoecological and Palaeogeographical Analysis Tools. Retrieved from http://matthewvavrek.com/programs-and-code/fossil/. Accessed 05 May, 2015.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., & Kuiper, M. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research*, *23*, 4407–4414.

Vranakis, I., Chochlakis, D., Sandalakis, V., Tselentis, Y., & Psaroulaki, A. (2012). Cost- and time-effectiveness of application of MALDI-TOF mass spectrometry methodology in a food and water microbiology laboratory. *Archives of Hellenic Medicine*, *29*, 477–479.

Wallace, D. L. (1983). A method for comparing two hierarchical clusterings: comment. *Journal of the American Statistical Association*, *78*, 569–576.

Wang, G., Whittam, T. S., Berg, C. M., & Berg, D. E. (1993). RAPD (arbitrary primer) PCR is more sensitive than multilocus enzyme electrophoresis for distinguishing related bacterial strains. *Nucleic Acids Research*, *21*, 5930–5933.

Warnes, A. G. R., Bolker, B., Bonebakker, L., Huber, W., Liaw, A., Lumley, T., Magnusson, A., Moeller, S., & Schwartz, M. (2015). Gplots: various R programming tools for plotting data.

Wassenaar, T. M., Geilhausen, B., & Newell, D. G. (1998). Evidence of genomic instability in *Campylobacter jejuni* isolated from poultry. *Applied and Environmental Microbiology*, *64*, 1816–1821.

Wetterstrand, K. (2015). DNA sequencing costs: data from the NHGRI genome sequencing program. Retrieved from http://www.genome.gov/sequencingcosts/. Accessed 28 Jan, 2015.

Wheeler, J. G., Sethi, D., Cowden, J. M., Wall, P. G., Rodrigues, L. C., Tompkins, D. S., Hudson, M. J., & Roderick, P. J. (1999). Study of infectious intestinal disease in England: rates in the community, presenting to general practice, and reported to national surveillance. The Infectious Intestinal Disease Study Executive. *British Medical Journal*, *318*, 1046–1050.

Whittam, T. S., Wachsmuth, I. K., & Wilson, R. A. (1988). Genetic evidence of clonal descent of *Escherichia coli* O157:H7 associated with hemorrhagic colitis and hemolytic uremic syndrome. *The Journal of Infectious Diseases*, *157*, 1124–1133.

Whittam, T. S. & Wilson, R. A. (1988). Genetic relationships among pathogenic *Escherichia coli* of serogroup 0157. *Infection and Immunity*, *56*, 2467–2473.

Wickham, H. & Winston, C. (2015). Ggplot2: An Implementation of the Grammar of Graphics. Retrieved from http://ggplot2.org. Accessed 15 Jun, 2015.

Williams, L. K., Sait, L. C., Cogan, T. A., Jørgensen, F., Grogono-Thomas, R., & Humphrey, T. J. (2012). Enrichment culture can bias the isolation of *Campylobacter* subtypes. *Epidemiology and Infection*, *140*, 1227–1235.

Wilson, D. J., Gabriel, E., Leatherbarrow, A. J. H., Cheesbrough, J., Gee, S., Bolton, E., Fox, A., Fearnhead, P., Hart, C. A., & Diggle, P. J. (2008). Tracing the source of campylobacteriosis. *PLoS Genetics*, *4*, e1000203.

Wilson, D. J., Gabriel, E., Leatherbarrow, A. J. H., Cheesbrough, J., Gee, S., Bolton, E., Fox, A., Hart, C. A., Diggle, P. J., & Fearnhead, P. (2009). Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*. *Molecular Biology and Evolution*, *26*, 385–397.

Wyklicky, H. & Skopec, M. (1983). Philipp Semmelweis, the Prophet of Bacteriology. *Infection Control: Proceedings of the First International Symposium on Hospital-Acquired Infections*, *4*, 367–370.

Young, K. T., Davis, L. M., & Dirita, V. J. (2007). *Campylobacter jejuni*: molecular biology and pathogenesis. *Nature Reviews Microbiology*, *5*, 665–79.

Chapter 6

Sequenced *C. jejuni* Strain Information

6.1 Metadata for Sequenced *C. jejuni* isolates

Table 6.1: Strain list and metadata for isolates of *C. jejuni* selected from the CGF Database for whole genome sequencing. *Project indicates either funding body or facility responsible for sampling.

Strain	Source	Province	City	Latitude	Longitude	Day	Month	Year	Project*
06_2866	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	26	9	2005	FNC
06_3245	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	6	6	2006	FNC
06_3569	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	19	6	2006	FNC
06_3783	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	4	7	2006	FNC
06_3849	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	22	6	2006	FNC
06_3851	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	5	7	2006	FNC
06_3852	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	5	7	2006	FNC
06_4734	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	10	8	2006	FNC
06_4911	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	15	8	2006	FNC
06_5176	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	18	8	2006	FNC
06_5790	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	13	9	2006	FNC
06_6211	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	19	9	2006	FNC
06_6212	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	24	9	2006	FNC
06_6554	A_RuminantFaecalA_Cow	ON	Waterloo	43.4643	-80.5204	4	7	2006	FNC
06_7331	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	7	11	2006	FNC
06_7332	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	29	10	2006	FNC
06_7515	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	23	11	2006	FNC

(Table 6.1 Continued)

Strain	Source	Province	City	Latitude	Longitude	Day	Month	Year	Project*
06_7656	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	4	12	2006	FNC
07_0549	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	5	2	2007	FNC
07_0675	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	13	2	2007	FNC
07_0971	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	2	3	2007	FNC
07_1009	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	6	3	2007	FNC
07_1493	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	29	3	2007	FNC
07_1875	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	17	4	2007	FNC
07_2174	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	7	5	2007	FNC
07_2680	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	30	5	2007	FNC
07_3238	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	19	6	2007	FNC
07_3508	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	6	7	2007	FNC
07_3647	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	10	7	2007	FNC
07_3853	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	16	7	2007	FNC
07_4076	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	20	7	2007	FNC
07_4268	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	31	7	2007	FNC
07_4269	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	31	7	2007	FNC
07_4428	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	7	8	2007	FNC
07_5038	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	20	8	2007	FNC

(Table 6.1 Continued)

Strain	Source	Province	City	Latitude	Longitude	Day	Month	Year	Project*
07_5039	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	21	8	2007	FNC
07_5041	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	21	8	2007	FNC
07_5581	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	5	9	2007	FNC
07_5583	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	6	9	2007	FNC
07_6017	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	23	9	2007	FNC
07_6066	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	25	9	2007	FNC
07_6215	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	3	10	2007	FNC
07_7314	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	10	4	2007	FNC
07_7324	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	28	5	2007	FNC
07_7331	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	11	6	2007	FNC
08_0096	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	16	7	2007	FNC
08_0099	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	21	7	2007	FNC
08_0100	A_PorcineRetailA_Pig	ON	Waterloo	43.4643	-80.5204	23	7	2007	FNC
08_1700	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	7	8	2007	FNC
08_1709	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	18	9	2007	FNC
08_1711	A_PorcineRetailA_Pig	ON	Waterloo	43.4643	-80.5204	18	9	2007	FNC
08_1714	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	9	10	2007	FNC
08_4456	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	26	6	2008	FNC

(Table 6.1 Continued)

Strain	Source	Province	City	Latitude	Longitude	Day	Month	Year	Project*
08_4460	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	2	7	2008	FNC
08_4461	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	2	7	2008	FNC
08_4466	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	2	7	2008	FNC
08_4468	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	2	7	2008	FNC
08_4472	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	2	7	2008	FNC
08_4474	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	2	7	2008	FNC
08_4603	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	2	7	2008	FNC
08_4696	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	7	7	2008	FNC
08_4697	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	7	7	2008	FNC
08_4913	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	15	7	2008	FNC
08_5176	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	22	7	2008	FNC
08_5490	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	28	7	2008	FNC
08_5603	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	12	8	2008	FNC
08_5925	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	15	8	2008	FNC
08_6160	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	22	8	2008	FNC
08_6208	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	28	8	2008	FNC
08_6877	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	7	9	2008	FNC
08_7016	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	12	9	2008	FNC

(Table 6.1 Continued)

Strain	Source	Province	City	Latitude	Longitude	Day	Month	Year	Project*
08_7017	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	17	9	2008	FNC
08_7039	HumanClinicalHuman	ON	Waterloo	43.4643	-80.5204	23	9	2008	FNC
CE_M_09_2085	A_RuminantFaecalA_Cow	ON	Waterloo	43.4643	-80.5204	27	7	2009	FNC
CE_M_09_3054	A_RuminantFaecalA_Cow	ON	Waterloo	43.4643	-80.5204	19	5	2009	FNC
CE_M_09_3081	A_RuminantFaecalA_Cow	ON	Waterloo	43.4643	-80.5204	20	7	2009	FNC
CE_M_09_4099	A_AvianFaecalA_Chicken	ON	Waterloo	43.4643	-80.5204	21	9	2009	FNC
CE_M_10_2096	A_RuminantFaecalA_Cow	ON	Waterloo	43.4643	-80.5204	29	11	2010	FNC
CE_M_10_2108	A_RuminantFaecalA_Cow	ON	Waterloo	43.4643	-80.5204	6	12	2010	FNC
CE_M_10_2113	A_RuminantFaecalA_Cow	ON	Waterloo	43.4643	-80.5204	6	12	2010	FNC
CE_M_10_3062	A_RuminantFaecalA_Cow	ON	Waterloo	43.4643	-80.5204	29	6	2010	FNC
CE_M_10_3107	A_RuminantFaecalA_Cow	ON	Waterloo	43.4643	-80.5204	15	11	2010	FNC
CE_M_10_4054	A_AvianFaecalA_Chicken	ON	Waterloo	43.4643	-80.5204	7	6	2010	FNC
CE_M_10_4091	A_AvianFaecalA_Chicken	ON	Waterloo	43.4643	-80.5204	20	9	2010	FNC
CE_R_10_0273	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	15	6	2010	FNC
CE_R_10_0305	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	29	6	2010	FNC
CE_R_10_0306	A_PorcineRetailA_Pig	ON	Waterloo	43.4643	-80.5204	29	6	2010	FNC
CE_R_11_0077	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	22	3	2011	FNC
CE_R_11_0100	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	19	4	2011	FNC

(Table 6.1 Continued)

Strain	Source	Province	City	Latitude	Longitude	Day	Month	Year	Project*
CE_R_11_0114	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	3	5	2011	FNC
CE_R_11_0170	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	9	7	2011	FNC
CE_R_11_0178	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	9	7	2011	FNC
CE_R_11_0192	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	12	7	2011	FNC
CE_R_11_0238	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	30	8	2011	FNC
CE_R_11_0240	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	30	8	2011	FNC
CE_R_11_0251	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	13	9	2011	FNC
CE_R_11_0270	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	4	10	2011	FNC
CE_R2_11_0134	A_AvianRetailA_Chicken	ON	Waterloo	43.4643	-80.5204	8	3	2011	FNC
CE_R2_11_0350	A_AvianRetailA_Turkey	ON	Waterloo	43.4643	-80.5204	2	8	2011	FNC
CE_R2_11_0374	A_AvianRetailA_Turkey	ON	Waterloo	43.4643	-80.5204	23	8	2011	FNC
CE2_R_11_1063	A_AvianRetailA_Chicken	BC	Chilliwack	49.1579	-121.9515	23	8	2011	FNC
CE2_R_11_3009	A_AvianRetailA_Chicken	BC	Chilliwack	49.1579	-121.9515	4	1	2011	FNC
CE2_R_11_3021	A_AvianRetailA_Chicken	BC	Chilliwack	49.1579	-121.9515	1	2	2011	FNC
CE2_R_11_3039	A_AvianRetailA_Chicken	BC	Chilliwack	49.1579	-121.9515	1	3	2011	FNC
CE2_R_11_3113	A_AvianRetailA_Chicken	BC	Chilliwack	49.1579	-121.9515	12	7	2011	FNC
CE2_R_11_3131	A_AvianRetailA_Chicken	BC	Chilliwack	49.1579	-121.9515	6	9	2011	FNC
CE2_R2_11_1006	A_AvianRetailA_Chicken	BC	Chilliwack	49.1579	-121.9515	5	4	2011	FNC

(Table 6.1 Continued)

Strain	Source	Province	City	Latitude	Longitude	Day	Month	Year	Project*
CE2_R2_11_1009	A_AvianRetailA_Chicken	BC	Chilliwack	49.1579	-121.9515	10	5	2011	FNC
CE2_R2_11_1027	A_AvianRetailA_Chicken	BC	Chilliwack	49.1579	-121.9515	19	7	2011	FNC
CE2_R2_11_1033	A_AvianRetailA_Chicken	BC	Chilliwack	49.1579	-121.9515	19	7	2011	FNC
CE2_R2_11_2018	A_AvianRetailA_Chicken	BC	Chilliwack	49.1579	-121.9515	5	7	2011	FNC
CE2_R2_11_2022	A_AvianRetailA_Turkey	BC	Chilliwack	49.1579	-121.9515	5	7	2011	FNC
CE2_R2_11_3023	A_AvianRetailA_Chicken	BC	Chilliwack	49.1579	-121.9515	15	3	2011	FNC
CE2_R2_11_3081	A_AvianRetailA_Chicken	BC	Chilliwack	49.1579	-121.9515	12	7	2011	FNC
CE2_R2_11_3085	A_AvianRetailA_Turkey	BC	Chilliwack	49.1579	-121.9515	12	7	2011	FNC
CGY_HR_009	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	14	4	2005	CGY_HR
CGY_HR_013	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	18	4	2005	CGY_HR
CGY_HR_020	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	21	4	2005	CGY_HR
CGY_HR_022	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	23	4	2005	CGY_HR
CGY_HR_026	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	26	4	2005	CGY_HR
CGY_HR_027	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	28	4	2005	CGY_HR
CGY_HR_058	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	22	5	2005	CGY_HR
CGY_HR_061	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	25	5	2005	CGY_HR
CGY_HR_073	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	31	5	2005	CGY_HR
CGY_HR_074	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	1	6	2005	CGY_HR

(Table 6.1 Continued)

Strain	Source	Province	City	Latitude	Longitude	Day	Month	Year	Project*
CGY_HR_076	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	1	6	2005	CGY_HR
CGY_HR_080	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	2	6	2005	CGY_HR
CGY_HR_082	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	3	6	2005	CGY_HR
CGY_HR_083	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	3	6	2005	CGY_HR
CGY_HR_090	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	8	6	2005	CGY_HR
CGY_HR_098	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	12	6	2005	CGY_HR
CGY_HR_101	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	14	6	2005	CGY_HR
CGY_HR_103	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	14	6	2005	CGY_HR
CGY_HR_108	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	18	6	2005	CGY_HR
CGY_HR_109	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	18	6	2005	CGY_HR
CGY_HR_118	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	21	6	2005	CGY_HR
CGY_HR_121	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	1	1	2005	CGY_HR
CGY_HR_139	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	27	6	2005	CGY_HR
CGY_HR_140	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	27	6	2005	CGY_HR
CGY_HR_170	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	2	7	2005	CGY_HR
CGY_HR_174	HumanClinicalHuman	AB	Calgary	51.0453	-114.0581	4	7	2005	CGY_HR
CHR_022	HumanClinicalHuman	AB	Lethbridge	49.6935	-112.8418	20	6	2004	CHR
CHR_023	HumanClinicalHuman	AB	Lethbridge	49.6935	-112.8418	21	6	2004	CHR

(Table 6.1 Continued)

Strain	Source	Province	City	Latitude	Longitude	Day	Month	Year	Project*
CHR_026	HumanClinicalHuman	AB	Lethbridge	49.6935	-112.8418	8	7	2004	CHR
CHR_028	HumanClinicalHuman	AB	Lethbridge	49.6935	-112.8418	12	7	2004	CHR
CHR_053	HumanClinicalHuman	AB	Lethbridge	49.6935	-112.8418	13	9	2004	CHR
CHR_090	HumanClinicalHuman	AB	Lethbridge	49.6935	-112.8418	26	6	2005	CHR
CHR_119	HumanClinicalHuman	AB	Lethbridge	49.6935	-112.8418	6	6	2005	CHR
CHR_127	HumanClinicalHuman	AB	Lethbridge	49.6935	-112.8418	16	6	2005	CHR
CHR_130	HumanClinicalHuman	AB	Lethbridge	49.6935	-112.8418	24	6	2005	CHR
CHR_151	HumanClinicalHuman	AB	Lethbridge	49.6935	-112.8418	26	8	2005	CHR
CHR_159	HumanClinicalHuman	AB	Lethbridge	49.6935	-112.8418	27	6	2005	CHR
CI_0094	E_WaterWaterE_Water	AB	N/A	53.9333	-116.5765	16	6	2004	Misc
CI_0136	A_RuminantFaecalA_Cow	AB	N/A	53.9333	-116.5765	2	6	2004	Misc
CI_0165	A_RuminantFaecalA_Cow	AB	N/A	53.9333	-116.5765	2	6	2004	Misc
CI_0168	A_RuminantFaecalA_Cow	AB	N/A	53.9333	-116.5765	2	6	2004	Misc
CI_0182	A_RuminantFaecalA_Cow	AB	N/A	53.9333	-116.5765	8	6	2004	Misc
CI_0292	E_WaterWaterE_Water	AB	N/A	53.9333	-116.5765	14	7	2004	Misc
CI_0322	A_RuminantFaecalA_Cow	AB	N/A	53.9333	-116.5765	12	7	2004	Misc
CI_0325	A_RuminantFaecalA_Cow	AB	N/A	53.9333	-116.5765	12	7	2004	Misc
CI_0334	A_RuminantFaecalA_Cow	AB	N/A	53.9333	-116.5765	21	7	2004	Misc

(Table 6.1 Continued)

Strain	Source	Province	City	Latitude	Longitude	Day	Month	Year	Project*
CI_0346	A_RuminantFaecalA_Sheep	AB	N/A	53.9333	-116.5765	21	7	2004	Misc
CI_0392	E_SewageSewageE_Sewage	AB	Fort Macleod	49.7217	-113.4	28	7	2004	Misc
CI_0405	A_CompanionFaecalA_Cat	AB	N/A	53.9333	-116.5765	10	8	2004	Misc
CI_0450	A_AvianFaecalA_Goose	AB	N/A	53.9333	-116.5765	19	8	2004	Misc
CI_0453	A_AvianFaecalA_Goose	AB	N/A	53.9333	-116.5765	19	8	2004	Misc
CI_0458	A_AvianFaecalA_Goose	AB	N/A	53.9333	-116.5765	19	8	2004	Misc
CI_0532	E_WaterWaterE_Water	AB	N/A	53.9333	-116.5765	1	9	2004	Misc
CI_0609	E_WaterWaterE_Water	AB	N/A	53.9333	-116.5765	15	9	2004	Misc
CI_0637	E_SewageSewageE_Sewage	AB	Fort Macleod	49.7217	-113.4	28	9	2004	Misc
CI_0677	A_AvianFaecalA_Chicken	AB	N/A	53.9333	-116.5765	5	10	2004	Misc
CI_0685	A_AvianFaecalA_Chicken	AB	N/A	53.9333	-116.5765	5	10	2004	Misc
CI_0697	E_WaterWaterE_Water	AB	N/A	53.9333	-116.5765	6	10	2004	Misc
CI_0699	E_WaterWaterE_Water	AB	N/A	53.9333	-116.5765	6	10	2004	Misc
CI_0765	A_AvianFaecalA_Goose	AB	N/A	53.9333	-116.5765	15	11	2004	Misc
CI_0783	A_RuminantFaecalA_Cow	AB	N/A	53.9333	-116.5765	15	11	2004	Misc
CI_0884	A_RuminantFaecalA_Cow	AB	Pincher Creek	49.4863	-113.9503	17	5	2005	Misc
CI_0893	A_RuminantFaecalA_Cow	AB	Pincher Creek	49.4863	-113.9503	17	5	2005	Misc
CI_0898	A_RuminantFaecalA_Cow	AB	Pincher Creek	49.4863	-113.9503	17	5	2005	Misc

(Table 6.1 Continued)

Strain	Source	Province	City	Latitude	Longitude	Day	Month	Year	Project*
CI_0918	A_RuminantFaecalA_Cow	AB	Lethbridge	49.6935	-112.8418	25	5	2005	Misc
CI_0927	A_RuminantFaecalA_Cow	AB	Pincher Creek	49.4863	-113.9503	17	5	2005	Misc
CI_0955	A_PorcineFaecalA_Pig	AB	Lethbridge	49.6935	-112.8418	6	6	2005	Misc
CI_0973	A_RuminantFaecalA_Cow	AB	Fort Macleod	49.7217	-113.4	22	6	2005	Misc
CI_0987	A_RuminantFaecalA_Cow	AB	Pincher Creek	49.4863	-113.4	22	6	2005	Misc
CI_1092	A_RuminantFaecalA_Sheep	AB	Lethbridge	49.6935	-112.8418	20	7	2005	Misc
CI_1096	A_RuminantFaecalA_Cow	AB	Lethbridge	49.6935	-112.8418	20	7	2005	Misc
CI_1109	A_RuminantFaecalA_Cow	AB	Fort Macleod	49.7217	-113.4	10	8	2005	Misc
CI_1117	A_RuminantFaecalA_Sheep	AB	Lethbridge	49.6935	-112.8418	10	8	2005	Misc
CI_1415	E_WaterWaterE_Water	AB	N/A	53.9333	-116.5765	28	9	2005	Misc
CI_1636	E_WaterWaterE_Water	BC	Salmon River	49.1464	-122.5978	6	3	2006	Misc
CI_1653	E_WaterWaterE_Water	AB	Fort Macleod	49.7217	-113.4	13	3	2006	Misc
CI_1660	E_WaterWaterE_Water	NB	Grand Falls	47.048	-67.7399	21	3	2006	Misc
CI_1799	E_WaterWaterE_Water	BC	Sumas River	49.0406	-122.1837	24	4	2006	Misc
CI_1845	E_WaterWaterE_Water	ON	N/A	51.2538	-85.3232	15	5	2006	Misc
CI_1874	A_RuminantFaecalA_Cow	AB	Lethbridge	49.6935	-112.8418	29	5	2006	Misc
CI_1875	A_RuminantFaecalA_Cow	AB	Lethbridge	49.6935	-112.8418	29	5	2006	Misc
CI_1884	A_RuminantFaecalA_Cow	AB	Lethbridge	49.6935	-112.8418	29	5	2006	Misc

(Table 6.1 Continued)

Strain	Source	Province	City	Latitude	Longitude	Day	Month	Year	Project*
CI_1915	E_WaterWaterE_Water	AB	N/A	53.9333	-116.5765	7	6	2006	Misc
CI_1920	A_AvianFaecalA_Goose	AB	Fort Macleod	49.7217	-113.4	20	6	2006	Misc
CI_1943	A_CompanionFaecalA_Dog	AB	Lethbridge	49.6935	-112.8418	26	6	2006	Misc

*Project Definitions:

CGY_HR	Calgary Health Region
CHR	Chinook Health Region
FNC	FoodNet Canada
Misc	Miscellaneous Projects

6.2 Metrics for Comparison of *in-silico* typing systems

Table 6.2: Simpson's Index of Diversity for Typing Methods generated *in-silico* on the dataset of 274 *C. jejuni* genomes.

Method	Partitions	SID	95% CI
CGF_100.0%	65	0.969	(0.963-0.975)
CGF_97.5%	28	0.943	(0.931-0.946)
CGF_95.0%	21	0.886	(0.748-0.824)
CGF_92.5%	15	0.837	(0.179-0.313)
CGF_90.0%	11	0.794	(0.099-0.215)
CGF_87.5%	7	0.685	(0.000-0.000)
CGF_85.0%	5	0.627	(0.000-0.000)
CGF_82.5%	3	0.585	(0.000-0.000)
MLST_100.0%	58	0.927	(0.905-0.948)
MLST_85.7%	36	0.839	(0.809-0.870)
MLST_71.4%	20	0.784	(0.751-0.817)
MLST_57.1%	11	0.703	(0.665-0.740)
MLST_42.9%	6	0.547	(0.500-0.594)
MLST_28.6%	3	0.064	(0.023-0.105)
rMLST_100.0%	129	0.982	(0.976-0.987)
rMLST_98.1%	77	0.970	(0.963-0.976)
rMLST_96.2%	62	0.955	(0.944-0.965)
rMLST_94.2%	53	0.940	(0.926-0.954)
rMLST_92.3%	49	0.934	(0.918-0.949)
rMLST_90.4%	45	0.912	(0.891-0.934)
rMLST_88.5%	40	0.909	(0.888-0.931)
rMLST_86.5%	36	0.893	(0.871-0.916)
rMLST_84.6%	32	0.893	(0.870-0.915)
rMLST_80.8%	30	0.877	(0.852-0.901)
rMLST_78.8%	29	0.864	(0.837-0.890)
rMLST_76.9%	26	0.863	(0.837-0.889)
rMLST_75.0%	23	0.854	(0.828-0.881)
rMLST_73.1%	20	0.848	(0.821-0.875)
rMLST_71.2%	17	0.827	(0.800-0.854)
rMLST_69.2%	16	0.827	(0.800-0.854)
rMLST_67.3%	14	0.812	(0.784-0.840)
rMLST_63.5%	14	0.812	(0.784-0.840)
rMLST_61.5%	14	0.812	(0.784-0.840)
rMLST_57.7%	14	0.812	(0.784-0.840)

(Table 6.2 Continued)

Method	Partitions	SID	95% CI
rMLST_51.9%	12	0.775	(0.746-0.805)
rMLST_50.0%	11	0.743	(0.711-0.776)
rMLST_48.1%	10	0.738	(0.705-0.770)
rMLST_46.2%	9	0.666	(0.627-0.706)
rMLST_44.2%	7	0.659	(0.619-0.699)
rMLST_42.3%	6	0.644	(0.606-0.681)
rMLST_38.5%	5	0.641	(0.605-0.677)
rMLST_34.6%	4	0.629	(0.595-0.663)
rMLST_30.8%	2	0.091	(0.044-0.137)
cgMLST_100.0%	269	1.000	(1.000-1.000)
cgMLST_95.1%	109	0.979	(0.972-0.985)
cgMLST_90.1%	83	0.973	(0.966-0.979)
cgMLST_84.9%	67	0.966	(0.959-0.973)
cgMLST_80.5%	60	0.959	(0.951-0.967)
cgMLST_75.7%	57	0.958	(0.950-0.966)
cgMLST_70.4%	47	0.940	(0.928-0.952)
cgMLST_67.5%	45	0.938	(0.926-0.951)
cgMLST_60.1%	37	0.863	(0.835-0.892)
cgMLST_56.9%	34	0.861	(0.832-0.890)
cgMLST_54.9%	33	0.861	(0.832-0.889)
cgMLST_45.7%	32	0.850	(0.821-0.879)
cgMLST_40.5%	25	0.822	(0.791-0.852)
cgMLST_35.9%	20	0.787	(0.754-0.820)
cgMLST_30.5%	14	0.780	(0.749-0.811)
cgMLST_27.8%	13	0.778	(0.746-0.809)
cgMLST_20.0%	11	0.750	(0.716-0.783)
cgMLST_17.8%	9	0.710	(0.675-0.745)
cgMLST_10.3%	5	0.641	(0.605-0.677)
cgMLST_06.6%	2	0.091	(0.044-0.137)

Table 6.3: Adjusted Rand statistic for the comparison of cgMLST cluster thresholds against *in-silico* CGF, MLST and rMLST. Jacknife pseudo-values for estimates of 95% CI indicated in parentheses.

cgMLST_Threshold	CGF	MLST	rMLST
cgMLST_100%	0.010 (0.000-0.024)	0.004 (0.000-0.010)	0.011 (0.000-0.032)
cgMLST_099%	0.196 (0.113-0.279)	0.096 (0.056-0.136)	0.257 (0.165-0.350)
cgMLST_098%	0.334 (0.246-0.421)	0.175 (0.124-0.225)	0.392 (0.310-0.474)
cgMLST_097%	0.441 (0.363-0.520)	0.250 (0.178-0.323)	0.514 (0.424-0.604)
cgMLST_096%	0.615 (0.516-0.714)	0.396 (0.301-0.491)	0.707 (0.612-0.802)
cgMLST_095%	0.644 (0.550-0.738)	0.408 (0.313-0.503)	0.712 (0.620-0.805)
cgMLST_094%	0.670 (0.579-0.761)	0.425 (0.333-0.518)	0.701 (0.607-0.794)
cgMLST_093%	0.705 (0.620-0.790)	0.457 (0.360-0.553)	0.723 (0.634-0.813)
cgMLST_092%	0.710 (0.628-0.793)	0.461 (0.365-0.557)	0.727 (0.638-0.815)
cgMLST_091%	0.730 (0.651-0.808)	0.479 (0.380-0.577)	0.702 (0.612-0.792)
cgMLST_090%	0.721 (0.643-0.800)	0.494 (0.399-0.589)	0.694 (0.604-0.785)
cgMLST_089%	0.721 (0.644-0.799)	0.496 (0.402-0.591)	0.690 (0.600-0.781)
cgMLST_088%	0.718 (0.635-0.800)	0.499 (0.407-0.590)	0.712 (0.619-0.804)
cgMLST_087%	0.715 (0.633-0.796)	0.505 (0.416-0.595)	0.691 (0.596-0.786)
cgMLST_086%	0.710 (0.629-0.791)	0.519 (0.429-0.609)	0.678 (0.582-0.775)
cgMLST_085%	0.723 (0.644-0.802)	0.536 (0.443-0.629)	0.685 (0.589-0.780)
cgMLST_084%	0.724 (0.645-0.802)	0.536 (0.443-0.629)	0.686 (0.590-0.781)
cgMLST_083%	0.736 (0.655-0.816)	0.509 (0.424-0.594)	0.615 (0.516-0.713)
cgMLST_082%	0.731 (0.651-0.812)	0.516 (0.430-0.601)	0.612 (0.514-0.710)
cgMLST_081%	0.731 (0.651-0.812)	0.516 (0.430-0.601)	0.612 (0.514-0.710)
cgMLST_080%	0.734 (0.655-0.814)	0.515 (0.430-0.601)	0.610 (0.512-0.707)
cgMLST_079%	0.734 (0.655-0.814)	0.515 (0.430-0.601)	0.610 (0.512-0.707)
cgMLST_078%	0.734 (0.654-0.813)	0.515 (0.429-0.600)	0.609 (0.511-0.706)
cgMLST_077%	0.734 (0.654-0.813)	0.515 (0.429-0.600)	0.609 (0.511-0.706)
cgMLST_076%	0.734 (0.654-0.813)	0.515 (0.429-0.600)	0.609 (0.511-0.706)
cgMLST_075%	0.745 (0.668-0.823)	0.512 (0.427-0.597)	0.602 (0.504-0.699)
cgMLST_074%	0.706 (0.627-0.786)	0.546 (0.461-0.630)	0.564 (0.468-0.661)
cgMLST_073%	0.706 (0.627-0.786)	0.546 (0.461-0.630)	0.564 (0.468-0.661)
cgMLST_072%	0.706 (0.627-0.786)	0.546 (0.461-0.630)	0.564 (0.468-0.661)
cgMLST_071%	0.683 (0.603-0.763)	0.531 (0.450-0.613)	0.540 (0.446-0.635)
cgMLST_070%	0.589 (0.499-0.678)	0.598 (0.523-0.673)	0.455 (0.362-0.549)
cgMLST_069%	0.584 (0.495-0.673)	0.594 (0.520-0.669)	0.451 (0.358-0.545)
cgMLST_068%	0.584 (0.495-0.673)	0.594 (0.520-0.669)	0.451 (0.358-0.545)
cgMLST_067%	0.576 (0.486-0.666)	0.605 (0.531-0.679)	0.444 (0.351-0.538)
cgMLST_066%	0.576 (0.486-0.666)	0.605 (0.531-0.679)	0.444 (0.351-0.538)
cgMLST_065%	0.576 (0.486-0.666)	0.605 (0.531-0.679)	0.444 (0.351-0.538)

(Table 6.3 Continued)

cgMLST_Threshold	CGF	MLST	rMLST
cgMLST_064%	0.576 (0.486-0.666)	0.605 (0.531-0.679)	0.444 (0.351-0.538)
cgMLST_063%	0.567 (0.477-0.656)	0.611 (0.538-0.684)	0.433 (0.340-0.526)
cgMLST_062%	0.380 (0.296-0.464)	0.634 (0.552-0.716)	0.278 (0.198-0.359)
cgMLST_061%	0.335 (0.265-0.405)	0.577 (0.495-0.660)	0.240 (0.174-0.306)
cgMLST_060%	0.298 (0.230-0.366)	0.649 (0.562-0.736)	0.212 (0.152-0.273)
cgMLST_059%	0.297 (0.230-0.365)	0.648 (0.560-0.735)	0.212 (0.151-0.272)
cgMLST_058%	0.297 (0.230-0.365)	0.648 (0.561-0.735)	0.212 (0.151-0.272)
cgMLST_057%	0.293 (0.226-0.360)	0.658 (0.571-0.745)	0.208 (0.148-0.268)
cgMLST_056%	0.293 (0.226-0.360)	0.658 (0.571-0.745)	0.208 (0.148-0.268)
cgMLST_055%	0.293 (0.226-0.360)	0.658 (0.571-0.745)	0.208 (0.148-0.268)
cgMLST_054%	0.293 (0.226-0.360)	0.657 (0.570-0.744)	0.208 (0.148-0.268)
cgMLST_053%	0.293 (0.226-0.360)	0.657 (0.570-0.744)	0.208 (0.148-0.268)
cgMLST_052%	0.293 (0.226-0.360)	0.657 (0.570-0.744)	0.208 (0.148-0.268)
cgMLST_051%	0.293 (0.226-0.360)	0.657 (0.570-0.744)	0.208 (0.148-0.268)
cgMLST_050%	0.293 (0.226-0.360)	0.657 (0.570-0.744)	0.208 (0.148-0.268)
cgMLST_049%	0.293 (0.226-0.360)	0.657 (0.570-0.744)	0.208 (0.148-0.268)
cgMLST_048%	0.293 (0.226-0.360)	0.657 (0.570-0.744)	0.208 (0.148-0.268)
cgMLST_047%	0.293 (0.226-0.360)	0.657 (0.570-0.744)	0.208 (0.148-0.268)
cgMLST_046%	0.293 (0.226-0.360)	0.657 (0.570-0.744)	0.208 (0.148-0.268)
cgMLST_045%	0.281 (0.220-0.342)	0.618 (0.526-0.711)	0.192 (0.136-0.248)
cgMLST_044%	0.240 (0.187-0.293)	0.540 (0.440-0.641)	0.162 (0.114-0.209)
cgMLST_043%	0.240 (0.187-0.293)	0.540 (0.440-0.641)	0.162 (0.114-0.209)
cgMLST_042%	0.240 (0.188-0.293)	0.540 (0.439-0.640)	0.161 (0.114-0.209)
cgMLST_041%	0.240 (0.188-0.293)	0.540 (0.439-0.640)	0.161 (0.114-0.209)
cgMLST_040%	0.251 (0.196-0.307)	0.535 (0.435-0.635)	0.159 (0.112-0.206)
cgMLST_039%	0.252 (0.196-0.307)	0.535 (0.435-0.634)	0.159 (0.112-0.206)
cgMLST_038%	0.252 (0.196-0.307)	0.535 (0.435-0.634)	0.159 (0.112-0.206)
cgMLST_037%	0.253 (0.197-0.308)	0.533 (0.433-0.632)	0.159 (0.112-0.205)
cgMLST_036%	0.210 (0.166-0.254)	0.455 (0.351-0.559)	0.131 (0.091-0.170)
cgMLST_035%	0.210 (0.166-0.253)	0.454 (0.350-0.557)	0.130 (0.091-0.170)
cgMLST_034%	0.207 (0.164-0.249)	0.445 (0.343-0.547)	0.127 (0.089-0.166)
cgMLST_033%	0.206 (0.164-0.249)	0.444 (0.342-0.546)	0.127 (0.088-0.165)
cgMLST_032%	0.206 (0.163-0.248)	0.443 (0.341-0.545)	0.127 (0.088-0.165)
cgMLST_031%	0.206 (0.163-0.248)	0.443 (0.341-0.545)	0.127 (0.088-0.165)
cgMLST_030%	0.204 (0.162-0.246)	0.439 (0.338-0.540)	0.125 (0.087-0.163)
cgMLST_029%	0.204 (0.162-0.246)	0.439 (0.338-0.540)	0.125 (0.087-0.163)
cgMLST_028%	0.204 (0.162-0.246)	0.439 (0.338-0.540)	0.125 (0.087-0.163)
cgMLST_027%	0.201 (0.160-0.243)	0.434 (0.333-0.535)	0.123 (0.086-0.161)

(Table 6.3 Continued)

cgMLST_Threshold	CGF	MLST	rMLST
cgMLST_026%	0.201 (0.160-0.243)	0.434 (0.333-0.535)	0.123 (0.086-0.161)
cgMLST_025%	0.201 (0.160-0.243)	0.434 (0.333-0.535)	0.123 (0.086-0.161)
cgMLST_024%	0.201 (0.160-0.243)	0.434 (0.333-0.535)	0.123 (0.086-0.161)
cgMLST_023%	0.201 (0.160-0.243)	0.434 (0.333-0.535)	0.123 (0.086-0.161)
cgMLST_022%	0.201 (0.159-0.242)	0.433 (0.332-0.533)	0.123 (0.086-0.160)
cgMLST_021%	0.201 (0.159-0.242)	0.433 (0.332-0.533)	0.123 (0.086-0.160)
cgMLST_020%	0.175 (0.137-0.213)	0.383 (0.284-0.483)	0.106 (0.073-0.140)
cgMLST_019%	0.175 (0.137-0.213)	0.383 (0.284-0.483)	0.106 (0.073-0.140)
cgMLST_018%	0.148 (0.116-0.179)	0.330 (0.235-0.424)	0.089 (0.061-0.117)
cgMLST_017%	0.146 (0.114-0.177)	0.325 (0.232-0.419)	0.088 (0.061-0.116)
cgMLST_016%	0.146 (0.114-0.177)	0.325 (0.232-0.419)	0.088 (0.061-0.116)
cgMLST_015%	0.146 (0.114-0.177)	0.325 (0.232-0.419)	0.088 (0.061-0.116)
cgMLST_014%	0.146 (0.114-0.177)	0.325 (0.232-0.419)	0.088 (0.061-0.116)
cgMLST_013%	0.120 (0.095-0.146)	0.273 (0.188-0.359)	0.072 (0.049-0.096)
cgMLST_012%	0.119 (0.093-0.144)	0.270 (0.185-0.354)	0.071 (0.048-0.094)
cgMLST_011%	0.115 (0.090-0.140)	0.263 (0.179-0.346)	0.069 (0.047-0.092)
cgMLST_010%	0.108 (0.085-0.131)	0.248 (0.170-0.325)	0.065 (0.044-0.086)
cgMLST_009%	0.103 (0.081-0.125)	0.237 (0.164-0.309)	0.062 (0.042-0.081)
cgMLST_008%	0.026 (0.017-0.035)	0.063 (0.040-0.086)	0.015 (0.009-0.022)
cgMLST_007%	0.026 (0.017-0.035)	0.063 (0.040-0.086)	0.015 (0.009-0.022)
cgMLST_006%	0.006 (0.003-0.010)	0.016 (0.006-0.025)	0.004 (0.001-0.006)
cgMLST_005%	0.006 (0.003-0.010)	0.016 (0.006-0.025)	0.004 (0.001-0.006)
cgMLST_004%	0.006 (0.003-0.010)	0.016 (0.006-0.025)	0.004 (0.001-0.006)
cgMLST_003%	0.006 (0.003-0.010)	0.016 (0.006-0.025)	0.004 (0.001-0.006)

Chapter 7

Detailed information from quantitative epidemiologic modelling

7.1 Calculating spheroidal distances using the Haversine

The Haversine formula, as described in (Sinnott, 1984) is used to calculate the distance across a spherical surface, using longitudinal and latitudinal coordinates. The equation is as follows:

$$\begin{aligned} dlon &= lon2 - lon1 \\ dlat &= lat2 - lat1 \\ a &= (sin(dlat/2))^2 + cos(lat1) * cos(lat2) * (sin(dlon/2))^2 \\ c &= 2 * a(sin(\sqrt{a})) \\ d &= R * c \end{aligned} \tag{7.1}$$

Where

- R = radius of the Earth (6373 km)
- lon1/lon2 = longitudinal coordinates of locations 1 and 2, respectively
- lat1/lat2 = latitudinal coordinates of locations 1 and 2, respectively

7.2 Histograms and colour scales for heatmap analyses

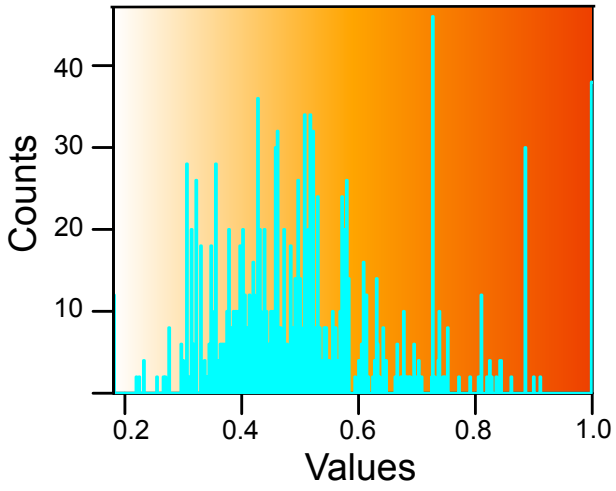

Figure 7.1: Frequency counts of pairwise similarity values presented in the source clustering heatmap in Figure 3.3. Darker colours indicate higher similarity, (n = 1444).

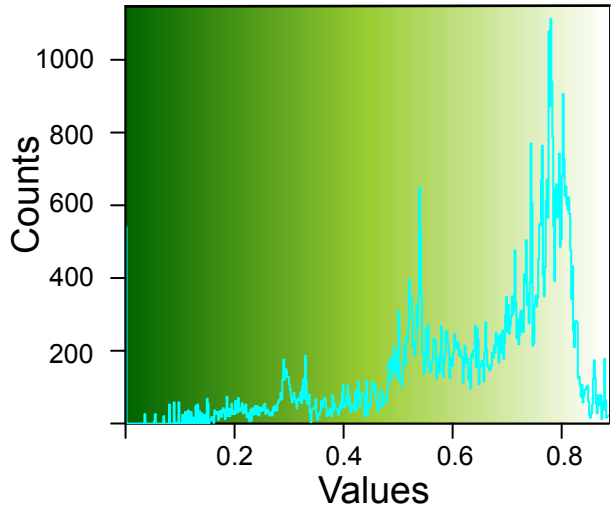

Figure 7.2: Frequency counts of pairwise similarity values presented in the epidemiological clustering heatmap in Figure 3.4. Darker colours indicate higher similarity, (n = 75076).

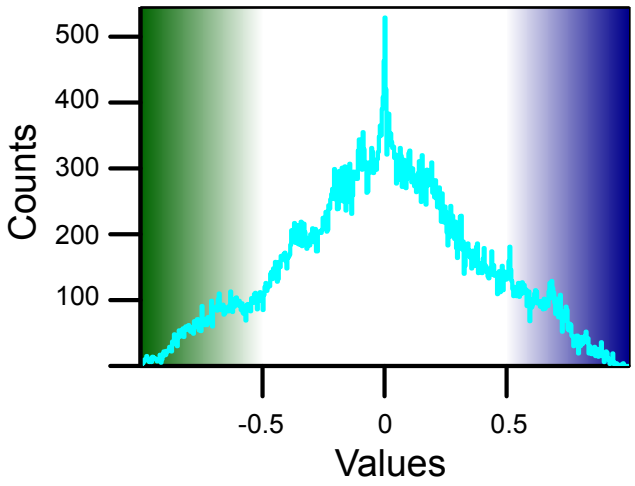

Figure 7.3: Frequency counts of pairwise values presented in the hierarchical rank clustering analysis heatmap in Figure 3.5, (n = 75076). Green and blue colour scales indicate higher similarity pairings via epidemiological and genomic relevance, respectively.