Host/niche adaptation and specificity in *Escherichia coli*

by

Shuai Zhi

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Public Health

School of Public Health

University of Alberta

Abstract

Patterns of microbial host specificity have been observed at all host-related taxonomic levels, and several studies have demonstrated that *Escherichia coli* (*E. coli*) appears to display some level of host adaptation and specificity. Colonization of the gastrointestinal tract by *E. coli* largely depends on its ability to sense and respond to the physiological conditions of the gut - biological processes governed by the regulome. Consequently, it was hypothesized that genetic polymorphisms in the intergenic regions of the regulome may represent a unique target for assessing DNA sequence polymorphisms associated with host specificity. Supervised learning logic-regression-based analysis of DNA sequence variability in *E. coli* intergenic regions (ITGR) was used to identify single nucleotide polymorphism (SNP) biomarker patterns correlated with host origin. The results demonstrate that a significant proportion of the *E. coli* population found in a human/animal host appear to be host-specific. Various levels of host-specific information, as determined by sensitivity and specificity analysis, were encoded in different ITGRs, and not all ITGRs were informative across all animals examined. Whole genome discovery analysis revealed that certain ITGRs encode extremely high levels of host-specific information, reinforcing the finding that the majority of the *E. coli* populations from a particular animal are host-specialists. Moreover, an analysis of the genes regulated by these host-specific ITGRs revealed the selection forces potentially driving evolution of host-specificity. In *E. coli* derived from humans, many of the host-specific ITGRs regulated antibiotic resistant genes, whereas in cattle, ITGRs regulating environmental survival/stress genes were dominant. These bioinformatics tools were also used to assess whether certain populations of *E. coli* may have evolved to live outside their vertebrate host (i.e., water). To evaluate this, *E. coli* was isolated from chlorine-treated wastewater and subjected to ITGR logic regression-based biomarker

analysis. Interestingly, wastewater *E. coli* was found to be genetically distinct from human and animal strains. Moreover, these strains were infrequently observed in other water matrices (i.e., groundwater, surface water), suggesting that these strains were specifically adapted to survive in a wastewater environment and not a 'water-based' environment. Many of these wastewater strains (~59%) possessed a genetic insertion element (IS30) located within the *uspC-flhDC* intergenic region, and for which a PCR assay was developed to identify these strains in environmental samples. The occurrence of these wastewater strains in the environment correlated with other known markers of sewage/wastewater pollution (i.e., *Bacteroides*). The identification of naturalized wastewater *E. coli* strains offered an excellent opportunity to further explore the adaptive phenotype/genotype characteristics that allow for survival of these strains in a non-host environment. This was important to address in the thesis since it is well known that the majority of DNA sequence variability in biological systems is non-adaptive. To provide evidence of adaptive evolution, wastewater strains were characterized for phenotypic/genotypic characteristics that might reveal life history strategies for survival in this unique matrix. Naturalized wastewater strains were shown to: i) be chlorine-tolerant, ii) be capable of robust biofilm production, iii) possessed a vigorous generalized stress response (RopS), iv) possessed universal stress protein genes (*usp*), and v) carried the locus of heat resistance – elements known to be important for survival of *E. coli* in the non-host environment. These strains were also shown to differentially survive through the wastewater treatment process. The findings presented in this thesis advance our knowledge regarding microbial evolution and host specificity, and the approaches developed can also be used to characterize sources of microbial contamination and protect public health from contaminated water and food.

Preface

Chapter Two of this thesis has been published as Shuai Zhi, Qiaozhi Li, Yutaka Yasui, Thomas Edge, Edward Topp, and Norman F. Neumann, "Assessing host-specificity of *Escherichia coli* using a supervised learning logic-regression-based analysis of single nucleotide polymorphisms in intergenic regions", in *Molecular Phylogenetics and Evolution*, vol.92 (2015):72–81. Norman F. Neumann and Shuai Zhi conceived the study. Thomas Edge and Edward Topp provided the bacterial isolates. Shuai Zhi collected the data and Qiaozhi Li, Yutaka Yasui, and Shuai Zhi performed the statistical analysis. Norman F. Neumann, Shuai Zhi, and Qiaozhi Li drafted the manuscript. Thomas Edge, Edward Topp, and Yutaka Yasui contributed to revising the draft manuscript.

Chapter Three of this thesis has been published as Shuai Zhi, Qiaozhi Li, Yutaka Yasui, Graham Banting, Thomas A. Edge, Edward Topp, Tim McAllister, and Norman F. Neumann, "An Evaluation of Multi-spacer Sequence Typing and Biomarker Discovery in Predicting Host Specificity in *Escherichia coli*" in *Molecular Phylogenetics and Evolution*. Norman F. Neumann and Shuai Zhi conceived the study. Thomas Edge and Edward Topp provided the bacterial isolates. Yutaka Yasui and Qiaozhi Li provided statistical support for statistical analysis. Norman F. Neumann and Shuai Zhi drafted the manuscripts and all authors contributed to revision and review of the manuscript.

Chapter Four of this thesis has been published as Shuai Zhi, Graham Banting, Qiaozhi Li, Thomas A. Edge, Edward Topp, Mykola Sokurenko, Candis Scott, Shannon Braithwaite, Norma J. Ruecker, Yutaka Yasui, Tim McAllister, Linda Chui, Norman F. Neumann, "Evidence of Naturalized Stress-Tolerant Strains of *Escherichia coli* in Municipal Wastewater Treatment Plants", in *Applied Environmental and Microbiology*. Norman F. Neumann and Shuai Zhi

conceived the study. Thomas Edge, Edward Topp, Mykola Sokurenko, Candis Scott, Shannon Braithwaite, Norma J. Ruecker, Tim McAllister, and Graham Bating provided the bacterial isolates. Shuai Zhi collected the data and Qiaozhi Li, Yutaka Yasui, and Shuai Zhi performed the statistical analysis. Thomas Edge, Edward Topp, Yutaka Yasui, Linda Chui, and Graham Banting contributed to revising the draft manuscript.

Chapter Five of this thesis has been submitted as Shuai Zhi, Graham Banting, Norma J. Ruecker, Norman F. Neumann, "Characterization of Municipal Wastewater Contamination in the Environment Using an *E. coli* -Based Source-Tracking Marker" to *Environmental Science & Technology*. Norman F. Neumann and Shuai Zhi conceived the study. Norma J. Ruecker provided the water samples. Shuai Zhi and Graham Banting collected the data. Norman F. Neumann and Shuai Zhi drafted the manuscripts and all authors contributed to revision and review of the manuscript.

Acknowledgements

At the end of my journey of getting this long-dreamed PhD degree, the opportunity to say thank you to all the people that had helped me has finally come. First, I want extend my deepest thanks to my supervisor, Dr. Norman Neumann. He is the nicest person I've ever met. Being his student makes me the luckiest person in the world. Thank you for helping me to overcome all the challenges I have faced at a place far away from home. Thank you for being patient with me and not mad when I made mistakes. Most importantly, thank you for all your guidance on my study and research. I have learned so many things from you, which will be a valuable asset for life. Secondly, I want extend my gratefulness to the rest of my committee members. To Dr. Linda Chui who had been so supportive on my *E. coli* project. To Dr. Yutaka Yasui who had been extremely helpful on the development of appropriate statistical methods for my research. To Patrick Hanington whose enthusiasm on research inspired me a lot and also thank you for allowing me to use your lab which made my work became faster and easier.

Thanks to ProvLab for providing laboratory support for this thesis. Thanks to the wonderful laboratory staff working in the Department of Environmental Microbiology at ProvLab. Your help in the past few years are greatly appreciated. Thank you Edie Ashton, Michelle Brown, Cheryl Hilner, Moira Raposo, Jaqueline Truemner, Keiko Matthew, Catherine Lam, Annamarie Virag. I am also very grateful to Qiaozhi Li, Graham Banting, Shannon Braithwaite, and Candis Scott at University of Alberta site for all their supports. Thanks Qiaozhi for explaining all the R codes and helping solve the statistical problems for my research.

I would like to thank Dr. Tom Edge (Environment Canada), Dr. Ed Topp (Agriculture and Agri-Food Canada), and Dr. Tim McAllister (Agriculture and Agri-Food Canada) for generously providing *E. coli* isolates for my study. I also have to express my deepest gratitude to

Norma Ruecker and Theingi Maw from City of Calgary. Thanks so much for providing me all the valuable water samples.

I could not have had achieved so much without all the support and help from my family and friends. Special thanks to my mom (Fulan Li), dad (Xiuming Zhi), brother (Gang Zhi), Ting Li, Lei Sun, Fengping Wu, Min Duan, Xiaozhou Zhang, Yang Xu, and Minke Wang.

Funding for this thesis was provided by Natural Sciences and Engineering Research Council (NSERC), the Canadian Foundation for Innovation (CFI) and Alberta Innovates – Energy and Environment Solutions.

Table of Contents

List of Tables

List of Figures

List of Symbols

AFLP	Amplified fragment length polymorphism
AR	Acid resistance
ARA	Antibiotic resistance
ARCC	Average rate of correct classification
ATP	Alternate test procedure
CDC	Centers for Disease Control
CSOs	Combined sewer overflows
DNA	Deoxyribonucleic acid
E. coli	*Escherichia coli*
EAEC	Enteroaggregative *E. coli*
EHEC	Enterohaemorrhagic *E. coli*
EIEC	Enteroinvasive *E. coli*
EPEC	Enteropathogenic *E. coli*
EPS	Exopolysaccharide
ERIC-PCR	Enterobacterial Repetitive Intergenic Consensus-PCR
ESP	Enterococcal surface protein
ET	Electrophoresis types
ETEC	Enterotoxigenic *E. coli*
FC/FS	Fecal coliform/fecal streptococci
FIB	Fecal indicator bacteria
FM	Fitch-Margoliash
GI	Gastrointestinal
HPP	Host predictive power
HUS	Haemolytic uremic syndrome
IAC	Internal amplification control
IS	Insertion sequence
ISO	International Standard Orgnization
ITGR	Intergenic region
iTOL	Interactive tree of life
LHR	Locus of heat resistance
LOD	Limit of detection
LTIIa	Heat liable toxin IIA
MEA	Minimum evolution algorithms
mL	Milliliter
ML	Maximum likelihood
MLEE	Multilocus enzyme electrophoresis
MLST	Multilocus sequence typing
mM	Millimolar
MP	Maximum parsimony

MPN	Most probable number
MSS	Multi-spacer sequence
MSST	Multi-spacer sequence typing
MST	Microbial source tracking
NJ	Neighbour joining
OBGS	Octamer-based genome scanning
^{o}C	Degrees Celsius
PCR	Polymerase chain reaction
PFGE	Pulsed field gel electrophoresis
PMA	Propidium monoazide
ProvLab	Provincial Laboratory for Public Health
qPCR	Quantitative polymerase chain reaction
RAPD	Random amplified polymorphic DNA
RCC	Rate of correct classification
Rep-PCR	Repetitive sequence-based polymerase chain reaction
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
RpoS	General stress response
RT-PCR	Reverse transcriptase PCR
SNP	Single nucleotide polymorphism
STEC	Shiga toxin producing *E. coli*
STII	Heat stable toxin gene II
TAE	Tris acetate EDTA
TMAO	Trimethylamine N-oxide
TSB	Tryptic soy broth
U.S. EPA	United States Environmental Protection Agency
UPEC	Uropathogenic *E. coli*
UPGMA	Unweighted Pair Group Method with Arithmetic Means
Usp	Universal stress protein
UTI	Urinary tract infection
UV	Ultraviolet
V	Volts
VBNC	Viable but non-culturable cells
WWTP	Wastewater treatment plant
µL	Microliter
µm	Micrometer

Chapter One : INTRODUCTION

1.1 Microbes and their hosts

Microorganisms have colonized every environment in the world including extreme environments such as arctic ice, water in hot springs and thermal vents, dust particles, and any conceivable microenvironment where nutrients and resources can be extracted to produce the metabolic energy necessary to support life. It is estimated that there are about 10^{30} microorganisms on earth (Whitman *et al.,* 1998). They form diverse communities in different niches.

All microbes inhabiting a particular environmental niche are collectively known as a microbiome. Every member of the microbiome, together with the host, forms a dynamic, interactive metabolic network essential for mutual survival (Foster *et al.,* 2008). Some researchers suggest that humans and animals be considered as 'supra-organisms' - a collection of individuals that behave as a single unit for enhanced functioning (Glendinning and Free, 2014). The various microbial populations forming the microbiome have important functions for its host. Recent interest has focused on characterizing the microbiome of humans: the composite collection of all microorganisms that live in, and on, the human body. Take the gut microbiome for example. Firstly, the microbiome plays an important role in metabolism. In the human gut, the microbiome helps digest food components which are indigestible by human biochemical processes alone or that have escaped the human digestion system, permitting the acquisition of extra energy (Walter and Ley, 2011). Some compounds, such as Vitamin K, critical for human health, are also produced by the gut microbiome (Guarner and Malagelada, 2003). It has been demonstrated that the gut microbiome in mice effectively helps to detoxify xenobiotic compounds by promoting the expression of cytochrome P_{450} (Claus *et al.,* 2011). Secondly, the

gut microbiome also plays key roles in the development of host immunity. In a mouse model, the gut microbiome was shown to stimulate the mucosal immune system by raising the number of lymphoid cells, immunoglobulin-A secreting cells, and other immune factors (Klaasen *et al.,* 1993). In addition, the microbiome can also assist the host in developing extended periods of immunity to certain antigens (Moreau and Gaboriau-Routhiau, 1996). Thirdly, the microbiome can protect the host through acting as a barrier to pathogens attempting to colonize the gut. Evidence for this can be observed in the increased colonization of the pathogen *Clostridium difficile* in the intestinal tract of mice and humans following treatment with high doses of antibiotics: an effect which disturbs the composition of the original gut microbiome (Adams *et al.,* 2007). As further evidence, germ-free mice are highly susceptible to pathogen invasion (Lundin *et al.,* 2008). By competing for nutrients and attachment/colonization sites with the exogenous microorganisms, coupled with the production of bacteriocins, the microbiome provides a barrier for its host to defend against pathogen infection (Lundin *et al.,* 2008). The microbiome has been well accepted as a critical part of a host and it is claimed that the human microbiome should be regarded as a second genome in humans (Grice and Segre, 2012).

Based on the evidence above, it can be concluded that the microbiome greatly impacts a host's physiological and biological characteristics. Could, and to what extent, do the characteristics of a host's environment shape the population structure of the microbiome? As with all biological entities, evolutionary pressures drive natural selection and adaptation of the microbiome, ultimately governing the success of a microbial population to survive within an environment such as the gastrointestinal (GI) tract. Comparative metagenomic studies of the 16S rRNA gene in various animal species, demonstrate that gut microbiomes are variable among different host species (Ley *et al.,* 2008), among individuals within a species (Eckburg *et al.,*

2005), and among different body sites (Costello *et al.,* 2009). In one study, researchers reciprocally inoculated the gut microbiota of zebrafish and mice into germ-free zebrafish and mice and after a few days the microbiome composition of these animals changed to be similar to the normal microbiome of their host (Rawls *et al.,* 2006). Ley *et al.* (2008) studied the fecal microbiome of humans and several other mammals by sequencing the 16S rRNA gene. Among all prokaryotic phyla identified, only *Firmicutes* were present across all samples. On average, 62% of the bacteria identified at the genus level in each sample were unique. It was also found that the bacterial populations comprising the gut microbiome from carnivores were significantly different from omnivores and herbivores suggesting a correlation between diet and microbiome composition (Ley *et al.,* 2008). In another study, the microbiome from different humans and different intestinal sites were studied (Eckburg *et al.,* 2005). The greatest bacterial diversity was observed among different individuals, while the microbiome from different sites in the colon also showed high diversity but to a lesser extent. It has been observed that on human skin that the dominant microbes are comprised of bacteria from the phylum *Actinobacteria*, while the most abundant bacteria in the human GI tract belong to the phylum *Bacteroidetes* (Spor *et al.,* 2011). Interestingly, individuals with diabetes have fewer bacteria from the *Firmicutes* phylum but higher amount of bacteria from the *Bacteroidetes* phylum (Larsen *et al.,* 2010). All these observations suggest that the host environment is critical in shaping the microbiome structure. These environmental factors include different host physiological conditions, diet, and host genotypes.

It has been demonstrated that the composition of the bacterial microbiome in the GI tract varies among different individuals (Ley *et al.,* 2008; Eckburg *et al.,* 2005), but what evidence is there that specific microbial populations are associated with certain animal hosts? A study by

Fraune and Bosch (2007) examined the microbiota of two different *Hydra* species cultured under the same condition for more than 30 years in the laboratory and demonstrated that they have very different microbiomes. However, when comparing the microbiome of *Hydra* directly isolated from the wild, it was found that the same *Hydra* species had similar microbiome structures (Fraune and Bosch, 2007). Ellis *et al.* (2013) studied the fecal microbiome of human, chimpanzees and cattle. The results demonstrated that the microbiome of humans and chimpanzees were more similar than between the microbiome of primates and cattle suggesting that similarity in host physiology and/or diet between humans and chimpanzees may contribute this difference. Researchers also compared the microbiome of three different wild primate species (black-and-white colobus, red colobus, and red-tailed guenon). They demonstrated that the microbiome composition from the same primate species was more similar to each other than to the microbial composition from different primate species (Yildirim *et al.,* 2010). In studies comparing pig, mouse, cattle, and human GI microbiome data, it was found that these mammalian microbiomes were similar at the phylum level but distinct at the family and genus levels (Dethlefsen *et al.,* 2007). *Firmicutes* and *Bacteroidetes* bacteria were the predominant phylum in the microbiome of human and mouse while in zebra fish *Acidobacteria* are most frequently isolated (Ley *et al.,* 2006). These data suggest that the structure of gut microbiomes is reflective of the host species from which they originate and are therefore subject to the constraints associated with host physiology. Consequently this drives evolutionary adaptation and selection in the microbial populations comprising the microbiome in each host.

1.2 Microbial host specificity

Host specificity is a relative term, and refers to a particular microbe's ability to colonize and develop in a particular host (e.g., humans) or defined host group (e.g., mammals). Strains that only colonize the GI tract of one host can be regarded as specialists, while strains which have more than one host can be referred to as generalists. In terms of human disease, host generalists are considered as the group associated with zoonotic disease, whereas specialists are associated with anthropogenic disease.

Several species of bacteria have been shown to display a certain degree of host specificity. Sung *et al.* (2008) demonstrated that most *Staphylococcus aureus* strains from different animals (human, cows, horses, goats, sheep) can be grouped into various animal-specific lineages by microarray analysis, based on presence/absence of genes and base variations in certain genomic regions. Host specificity has also been observed in the Gram-negative *Bartonella* species (Kosoy *et al.,* 2000). When inoculating wild *Bartonella* strains into different rodent species, infection was only found in rats if the inoculated strains were isolated from same rodent species or from a phylogenetically-related species (Kosoy *et al.,* 2000). Several studies have revealed that species of *Bacteroidales* also show host specificity, characteristics that enable them to serve as indicators of fecal pollution in microbial source tracking studies (Bernhard and Field, 2000a; Bernhard and Field, 2000b; Tambalo *et al.,* 2012). Host specificity is not only found between microbes and animals, but also in plants, suggesting a universal paradigm of microbial evolution towards host specificity. A study on the symbiosis between *Rhizobium* species and their plants showed that *Rhizobium leguminosarum viciae* strains are only able to form nodules with plants from the genera *Pisum*, *Vicia*, *Lathyrus*, and *Lens* (Albrecht *et al.,* 1999). Perhaps the most studied organism with respect to host specificity is the protozoan parasite *Cryptosporidium.* Many

species are believed to be host-specific or capable of infecting a very limited and defined range of hosts. Certain species of *Cryptosporidium* are host-specific to major groupings of animals such as fish, amphibians, reptiles, birds or mammals (Fayer, 2004; Fayer and Xiao, 2007; Xiao *et al.,* 2004). Other species appear to have a very specific host range, such as *C. hominis* that appears to specifically infect humans and some primates (Fayer, 2004). Likewise, certain strains of *C. parvum* are also restricted to humans, whereas other strains infect only animals, while others appear to infect both humans and animals (i.e., zoonotic strains) [Fayer, 2004]. Infections in dogs or other canids are mostly associated with *Cryptosporidium canis* while cats are primarily infected by *Cryptosporidium felis* (Xiao *et al.,* 2004). It is widely believed that genetic diversity in *Cryptosporidium* is reflective of host-specificity and that *Cryptosporidium* is comprised genetically of many cryptic species.

For microbes, host specificity is largely governed at the genetic level, and host specificity can be encoded in a set of genes, a specific gene, or even single nucleotide polymorphisms (SNPs) within a single gene. In a study by Mandel *et al.* (2009), it was found that the addition of the gene, *RscS*, was sufficient to make a strain of *Vibrio fischeri* isolated from fish capable of colonizing *Euprymna scolopes,* a species of squid. *RscS* induces the expression of a transcriptional activator *sypG* for the production of exopolysaccharide (EPS), which facilitates robust biofilm formation during *Vibrio fischeri*'s initial infection in *Euprymna scolopes* and therefore helped to establish colonization (Mandel *et al.,* 2009). With regard to gene sets, it was found that an acquired type III secretion system is needed by *Sodalis glossinidius* to establish itself in the tsetse fly, indicating that a certain gene set may be able to determine host specificity of a microorganism (Dale *et al.,* 2001). In addition, it was found that *Rhizobia spp.*, a Gram-negative bacterium, can specifically trigger only leguminous plants to form root nodules (Fauvart

and Michiels, 2008). Single or certain combinations of *Nod* factor encoding genes determine *Rhizobia*'s nodulation on leguminous plants (Fauvart and Michiels, 2008; Lerouge *et al.,* 1990). In another study, genome comparison of an extremely host-specific insect fungus, *Ophiocordyceps unilateralis s.l,* with nineteen other fungi (that also had various levels of host-specificity), suggested that the gain and loss of several genes resulted in acquisition of host-specific characteristics. These gene sets included Class 1 hydrophobins, subtilisin proteases, gene families involved in membrane transport of sucrose, and bacterial-like toxins (Wichadakul *et al.,* 2015). SNP variation has also been found to be related to host specificity. It was reported that unique SNPs in the *sseC* gene were found in human adapted *Salmonella* serovars (Tracz *et al.,* 2006; Eswarappa *et al.,* 2008). In the influenza virus, such as H5N1, amino acid changes at position 627 of polymerase subunit II were found to be host-related. The amino acid lysine is located at this position in almost all human influenza strains while most avian viruses have glutamic acid at this position, except for the "Qinghai Lake" lineage of H5N1 viruses (Neumann *et al.,* 2009). These studies support the idea that a single gene, gene sets or a few SNPs can explain the host-specific patterns of certain microbes. However, it remains difficult to identify host-specific genetic markers in a microbial population. For example, in the study mentioned above regarding *Staphylococcus aureus* stains isolated from different animal hosts, it was found that all 2013 of the core genes found from human isolates (which existed in more than 95% of the strains) could also be found in animal isolates. Identification of these host-specific SNP patterns or genes is very much like finding the proverbial 'needle in a haystack'. More sophisticated molecular analysis and techniques are needed to identify and characterize host-specific genetic markers in microbial populations.

In addition to mutational variations within genes, the mutations in intergenic regions (ITGR) were also found to alter the phenotypic characteristics of microbes in terms of their adaptation to various environmental conditions. Intergenic regions carry DNA sequences that can regulate the expression of genes including promoters (Haugen *et al.,* 2008), transcriptional regulator binding sites (repressor/activator) (Browning and Busby, 2004), small regulatory RNAs (Gottesman, 2005), and transposable elements (Casacuberta and Gonzalez, 2013). In a long-term evolutionary experiment on *E. coli* populations *in vitro*, it was observed that under citrate abundance and glucose limiting conditions one *E.coli* population evolved the ability to use citrate as an energy source under aerobic conditions (Blount *et al.,* 2012). This was achieved by acquiring an aerobically-expressed promoter for the expression of a previously silent citrate transporter (Blount *et al.,* 2012). Intergenic regions have also been found to relate to host-specificity in plant-microbe interactions. Strains of *Rhizobium leguminosarum* biovar *trifolii* can from nodules on Caucasian clover and white clover. The strains that can form effective nodules on Caucasian clover usually form ineffective nodules on white clover and vice versa. By comparison of those two type strains, it was found that the intergenic regions of *nifH-fixA* of strains from Caucasian clover all carry an extra DNA region (111bps) while strains from white clover lacked this region (Miller *et al.,* 2007). Microbial adaptation to a specific host is a complex process; therefore the regulation of a number of genes and regulatory elements might be involved in this process. Analysis of more genes or intergenic regions might be a better approach for understanding host adaptation and specificity.

1.3 Microbial host specificity and application to microbial source tracking (MST)

Feces from humans and animals are a major source of microbial contamination of food and water (Field and Samadpour, 2007). Feces can carry a broad range of pathogenic bacteria, protozoan, and viruses (Reynolds *et al.,* 2008). Microbial source tracking methods aim to identifying the host sources of fecal contamination in order to track and control human, animal and environmental health risks. A significant effort over the last 20 years in developing microbial source tracking methods has greatly contributed to our knowledge about microbial host-specificity.

Most current MST methods are based on two assumptions (Blanch *et al.,* 2011; Sadowsky and Santo Domingo, 2007). First, selective physiological and resource competition pressures in the gut microenvironment force adaptation of microbial populations within the GI tract. Second, this adaptation leads to natural selection, resulting in the preferential colonization, survival and replication of the most genetically fit microbial populations within the gastrointestinal tract. Based on these two assumptions it is believed that at least some of the microbial community members of the gastrointestinal tract are host-specific and that these microbes (or their associated genetic/phenotypic properties) can be used to track microbial sources of pollution in the environment (e.g., food and water).

MST methods can be divided into library-dependent and library-independent methods. Library-dependent methods require a 'library' or collection of specific microorganisms (i.e., *E. coli*) from potential animal hosts, and for which the library is intended to provide an accurate composite picture of the genotypic/phenotypic variability of strains associated with a particular animal host. The genotypic/phenotypic characteristics of host-specific strains provide a composite "fingerprint" by which unknown isolates can be compared to in order to identify

potential sources of contamination in food, water or the environment. In contrast, library-independent methods usually use host-specific microorganisms or host-specific genetic markers to track contamination sources and no library of isolates is required.

1.3.1 MST library-dependent methods

1.3.1.1 Pulsed field gel electrophoresis (PFGE)

PFGE is a DNA fingerprinting method that uses rare cutting restriction enzymes to cleave genomic DNA (10 to 800kb in length) and then electrophoresed under alternating electrical currents to produce DNA fingerprints for each isolate. PFGE has been proven to be superior to most biochemical and molecular typing methods (Barbier *et al.,* 1996; Grundmann *et al.,* 1995; Noble *et al.,* 2003) and has often been regarded as a "gold standard" for molecular typing. PFGE has been used by the PulseNet program, which was established by the Centers for Disease Control and Prevention (CDC) of the United States to study widespread outbreaks of bacterial foodborne illness (Swaminathan *et al.,* 2001). In a MST study by Casarez *et al.* (2007), it was demonstrated that PFGE had higher rate of correct classification (RCC) than Enterobacterial Repetitive Intergenic Consensus-PCR (ERIC-PCR), ribotyping, and antibiotic resistance analysis. PFGE was also found to have a higher discriminatory power for resolving populations of cattle and human *E. coli* O157: H7 than multilocus sequence typing (MLST) and repetitive element PCR (Rep-PCR) (Foley *et al.,* 2004). In another study, it was shown that PFGE could correctly assign 88% of the isolates to their sources (Myoda *et al.,* 2003). However, in some other MST studies, there was no, or poor, associations found between PFGE profiles and isolate sources (Parveen *et al.,* 2001).

The advantages of PFGE are: i) the procedures are simple and standard but for which strict quality control is required; ii) the method has high reproducibility, making the results between different laboratories comparable; and iii) it is considered to have high discriminatory power since PFGE can even detect the difference between closely related strains. However, when applying PFGE in source tracking, a large library of isolates (i.e., thousands) from various animal hosts is required which makes this technique time-consuming and labour intensive. Furthermore, it usually takes 2-3 days to obtain PFGE results and the numbers of isolates that can be processed simultaneously are limited, which reduce its ability to analyze large numbers of isolates. Most importantly, in PFGE, only a limited number of DNA sites are represented by the restriction cut across the genome, and therefore, only a small number of sequence variations are actually used in fingerprinting analysis. Thus, the utility of this tool in reflecting the overall diversity within a bacterial population, such as *E. coli*, is extremely low, and its utility as a source-tracking tool remains in question.

1.3.1.2 Ribotyping

Ribotyping is a DNA fingerprint method that uses restriction endonucleases to cut bacteria genomic DNA, electrophorese the fragments on an agarose gel, and hybridize the rRNA gene fragments with oligonucleotide probes. Ribotyping has been proven to be effective for microbial source tracking in several studies (Carson *et al.,* 2001; Scott *et al.,* 2003; Hartel *et al.,* 2002). In a study by Scott *et al.* (2003), the RCC for human and animal derived isolates were 84.5% and 78.6% respectively. In another study by Carson *et al.* (2001), ribotyping was used to classify *E. coli* isolates from 8 known sources (human, cattle, pig, horse, dog, chicken, turkey, and goose) and the average rate of correct classification (ARCC) was 73.6%. When discriminant

analysis was restricted to three sources (human, dog and horse), ARCC was improved to 94.2%, and the highest ARCC of 97.1% was observed when isolates were grouped into human and non-human sources. Although ribotyping could effectively differentiate isolates from some sources in some studies, in several other studies ribotyping performed less effectively in differentiating isolates from a broader range of hosts (Carson *et al.,* 2001; Scott *et al.,* 2003). In another study, only 27% of the indicator strains were correctly assigned to their sources by ribotyping (Moore *et al.,* 2005). In an evaluation study, although 81%, 100% and 86% of the isolates were correctly assigned to its sources by three investigators, the false positive rates were as high as 23%, 57% and 19% respectively (Myoda *et al.,* 2003).

Ribotyping is a reproducible molecular typing method as rRNA is highly conserved. A major drawback of ribotyping is its variations in methodology, as different restriction endonucleases may perform different discrimination variability (Hartel *et al.,* 2002; Myoda *et al.,* 2003) and this makes ribotyping a non-standard MST method. In addition, ribotyping is expensive and labor-intensive since ribotyping includes many procedures.

1.3.1.3 Multilocus sequence typing (MLST)

MLST is a DNA sequence-based molecular typing method in which the sequences from several genes (usually housekeeping genes) are compared for genetic variations to classify strains, identify clonal groups and determine phylogenetic relationship. Several studies found that MLST had similar levels of discriminatory power as PFGE (Peacock *et al.,* 2002; Nallapareddy *et al.,* 2002), while some others found MLST to have more discriminatory power than PFGE and serotyping (Kotetishvili *et al.,* 2002; Revazishvili *et al.,* 2004). In other studies, MLST performed poorly compared to other methods. For instance, MLST showed the least

discriminatory power in an evaluation study which used PFGE, rep-PCR and MLST to type *E .coli* O157:H7 isolates from cattle, food, and infected humans (Foley *et al.,* 2004). In a study by Litrup *et al.* (2007), MLST was used to type 150 *Campylobacter coli* isolates from human, pig, cattle and food products. The result showed that only 68% of pig isolates had sequencing types that were present in pigs; only 53% of human isolates carried sequencing types that were present in humans. In another study, Adiri *et al.* (2003) used MLST to study the *E. coli* O78 strains from human, avian and cattle and no host specificity distribution was observed.

MLST data can easily be stored in a database accessible via the internet, and such databases include www.mlst.net and www.pubmlst.org (PubMLST). Researchers can compare their sequences to the sequences in the database, making the procedures of MLST relatively simple. Excellent reproducibility of MLST can be achieved due to availability of high quality DNA sequencing. However, as MLST's discrimination power depends on the sequence variations on genes, low-level variation in highly conserved housekeeping genes impairs the overall level of discrimination at the strain level.

1.3.1.4 Antibiotic Resistance Analysis (ARA)

ARA is based on comparing the multiple antibiotic resistance profiles of the isolates. The underlying assumption of this method is that microbial populations of the gastrointestinal tract of humans and animals are subject to different types, frequencies, and concentrations of antibiotics; therefore, overtime the selective pressure would make the bacteria in certain hosts develop a unique antibiotic resistance profile. Antibiotic resistance analysis is the most commonly used phenotypic method in MST to date. In a study on the differentiation of fecal streptococci isolates from human and animal sources, Wiggins (1996) reported an ARCC of 74% for six animal

sources and 92% for human isolates based on antibiotic resistance profiles. In another study, an ARCC of 93% was observed when combining the antibiotic resistance profiles of fecal coliform and fecal streptococci strains isolated from human, cattle, poultry, and swine (Evenson and Strevett, 2006). To the contrary, in some comparative studies, ARA had poor performance (Griffith *et al.,* 2003; Harwood *et al.,* 2003; Moore *et al.,* 2005; Samadpour *et al.,* 2005). For example, it was found ARA was less successful than ribotyping at assigning 120 *E. coli* isolates to correct sources (Samadpour *et al.,* 2005). In another study, only 44% of *E. coli* were assigned to correct sources by ARA (Moore *et al.,* 2005).

ARA is an inexpensive method that only requires basic microbiology techniques. It has received lots of attention as a phenotypic MST method. However, antibiotic resistance is often carried on plasmids that can be lost or transferred among different strains. As a result, this affects the stability of the microbe's antibiotic resistance profiles. In addition, antibiotics used for humans and domestic animals often belong to same classes and share same resistance mechanisms, therefore, microbes may develop cross resistance which would impair the discrimination power of ARA (Giedraitiene *et al.,* 2011). Furthermore, ARA is not useful when limited resistance patterns are observed, especially among the microbes isolated from wildlife that are under minor antibiotic selection pressure.

1.3.2 Library-independent MST methods

1.3.2.1 Bacterial culture-based MST methods

A method based on the ratio of fecal coliform/fecal streptococci (FC/FS) was used to assess the sources of fecal pollution, in which it was determined that a FC/FS >4 indicated human fecal pollution, a FC/FS between 0.1 and 0.6 indicated domestic animal fecal pollution,

and a FC/FS ＜0.1 indicated wild animal fecal pollution (Geldreich and Litsky, 1976). However, studies have subsequently showed that this method was not reliable due to different survival rates of these two bacteria in the environment, making the ratio unstable and therefore unreliable in an environmental setting (Sinton *et al.,* 1998).

Several MST methods based on host-specific bacteria have been described. For example, *Bifidobacterium*, a common genus of bacteria in the human intestine, was regarded as a good indicator for human fecal contamination based on the fact that it was rarely found in animals (Resnick and Levin, 1981; Gavini *et al.,* 1991; Rhodes and Kator, 1999). A human Bifid Sorbitol Agar was developed by Mara and Oragui (1983) to isolate sorbitol-fermenting Bifidobacteria, such as *B. adolescentis* and *B. breve,* which were only isolated from human fecal samples. However, a study by Blanch *et al.* (2006) found that human specific Bifidobacteria were present in some animal samples and missed in some of the human samples, which challenge their role in MST. Moreover, the environmental survival rate of these bacteria is influenced by temperature as they could not be detected in warm water (Rhodes and Kator, 1999; Bonjoch *et al.,* 2005; Carrillo *et al.,* 1985).

Rhodococcus coprophilus is another bacteria species that was found only in farm animals including cow, sheep, horse and deer (Mara and Oragui, 1981; Rowbotham and Cross, 1977). Its culture time is up to 21 days; therefore PCR methods were developed and subsequently used for its detection (Savill *et al.,* 2001; Wicki *et al.,* 2012). In addition, this microbe was only found in farm animals, impairing its broad applicability in MST.

1.3.2.2 Pathogen-based MST Methods

1.3.2.2.1 Host-specific viruses

Using viruses as fecal contamination indicators has also been reported by several studies. Currently, over 150 different types of enteric viruses have been identified (Wong *et al.,* 2012), among which human specific adenoviruses and enteroviruses are two commonly used viruses in the identification and assessment of human fecal contamination (Jiang *et al.,* 2001; Lee and Kim, 2002). Human polyomavirus has been suggested as a useful human fecal contamination indicator (McQuaig *et al.,* 2012). In addition, several animal specific viruses have been identified. For instance, bovine and porcine adenoviruses were used to identify cow and pig fecal contamination (Hundesa *et al.,* 2006; Maluquer *et al.,* 2004); and porcine teschoviruses were used to identify pig fecal contamination (Jimenez-Clavero *et al.,* 2003). In a study by Maluquer *et al.* (2004) porcine adenoviruses were detected in 70% of swine samples while bovine adenoviruses were present in 75% of cattle fecal samples.

Enteric viruses are difficult to cultivate and usually exist in the environment at low concentrations; therefore a large amount of water often needs to be filtered to collect them (Maluquer *et al.,* 2004). During this filtration, some other substance will also be collected and concentrated, which can often affect analysis. In addition, it has been observed that human specific viruses are more useful for identification of sewage contamination rather than individual fecal sources (Noble *et al.,* 2003).

1.3.2.2.2 F+ RNA coliphages

Coliphages are viruses that infect Gram-negative bacteria. F-specific coliphages attach to the F-pili encoded by the F plasmid of bacteria and can be further divided into F+DNA coliphage

and F+RNA coliphage. There are four subgroups of F+RNA, among which groups II and III F+RNA coliphages are usually isolated from human feces and wastewaters, while group I and IV are usually found in sources contaminated by animal feces (Long *et al.,* 2005). However, the F+RNA coliphage is not very specific. A study by Wolf *et al.* (2010) demonstrated that human associated group I F+RNA coliphage can also be detected in the feces from duck and pig. Another disadvantage of this method is that the numbers of coliphages present in the environment are low, so enrichment procedures or other techniques are needed to improve the sensitivity and stability of this method.

1.3.2.2.3 Parasite-based MST methods

Parasites, such as *Cryptosporidium*, have also been used in microbial source tracking to estimate health risk and track pollution sources (Ruecker *et al.,* 2007; Jellison *et al.,* 2009). It has been found that *Cryptosporidium* can infect more than 150 mammalian host and several species and genotypes have been identified (Fayer and Xiao, 2007; Xiao *et al.,* 2004). These species have their preferred host. For example, *C. parvum* is the dominant species infecting cattle less than 2 months old (Xiao and Fayer, 2008) while *C. andersoni* is frequently found in older animals (Kvac and Vitovec, 2003; Langkjaer *et al.,* 2007). *C. hominis* is predominantly found in humans (Xiao and Fayer, 2008; Xiao *et al.,* 2004). *C. canis* is mostly isolated from dogs, foxes, and coyotes (Xiao and Fayer, 2008). In one study by Gibson and Wagner (1986), it was found that *C. wrairi* only infects guinea pigs.

1.3.2.3 Host specific markers

Currently, the most commonly used host-specific genetic markers are based on the family *Bacteroidetes*. *Bacteroides* spp.are obligate anaerobes abundant in the feces of humans and animals (Wexler, 2007). In a study by Carson *et al.* (2005), a 542-bp amplicon specific to *Bacteroides thetaiotaomicron* were detected from 92% of human fecal samples, 100% sewage samples and 16% of dog fecal samples but with no detection in fecal samples from six other animals. One of the most commonly used *Bacteroides*-based human markers is HF183 due to its sensitivity and specificity (Harwood *et al.,* 2014). It has been demonstrated that HF183 can be detected in 100% sewage samples and 63% of human fecal samples (van de Werfhorst *et al.,* 2011) although it has also been shown to cross-react with feces from dog, chicken, cow, goat and sheep (Ahmed *et al.,* 2012; Odagiri *et al.,* 2015). Human and ruminant specific 16S rDNA genetic makers based on *Bacteroides-Prevotella*, which were 119 bp and 227 bp in length, respectively, were identified by Bernhard and Field (2000a) for which PCR assays were subsequently developed to discriminate human and ruminant fecal pollution sources. Their method has proven be successful, however, the cow-specific genetic markers were also detected in some other animal samples such as sheep, goat, elk and deer (Bernhard and Field, 2000b). In a study by Dick *et al.* (2005), pig and horse specific sequences were identified; and PCR primers were subsequently designed to identify fecal pollution from pig and horses; however, observed overlaps of 16S rDNA sequence among other hosts made it impossible to develop a host-specific marker for them (Dick *et al.,* 2005). In addition, several other host specific *Bacteroides*-based markers have been identified, such as a cow marker (CowM3) (Shanks *et al.,* 2008), muskrat marker (MuBac) (Marti *et al.,* 2011), and Canada goose marker (CGO1) (Fremaux *et al.,* 2010).

Host-specific toxin genes in *E. coli* have also been identified and have the potential to be used as MST markers (Scott *et al.,* 2005; Khatib *et al.,* 2002; Khatib *et al.,* 2003). In a study by Khatib *et al.* (2002), heat liable toxin IIA (LTIIa) gene from enterotoxigenic *E .coli* (ETEC) was used in a cattle-specific PCR assay and 87% of environmental samples from cattle waste and lagoons were LTIIa positive. In another study, a pig specific heat stable toxin gene II (STII) was identified by Khatib *et al.* (2003), which proved useful in distinguishing *E .coli* isolates between swine wastes and other animal sources. Unfortunately, the fact that many *E. coli* strains do not have these toxin markers hinders the application of these methods for identifying host sources of fecal contamination. Enterococcal surface protein (esp), a virulence factor from *Enterococcus faecium*, has also been proposed as a human specific marker as this virulence factor was believed to exist only in *Enterococcus* isolated from humans (Scott *et al.,* 2005), but this gene has also been observed in the feces from other animals (Layton *et al.,* 2009).

It is important to note that although evidence of host-specificity within microbial populations of the gut has been presented above, some MST studies have failed to provide this same evidence. For example, no or poor association was found between PFGE profiles and isolate sources (Parveen *et al.,* 2001). In an MST evaluation study using PFGE, rep-PCR, and ribotyping, although a high percentage of the isolates were correctly assigned to their host sources by three investigators, the false positive rates were also very high (Myoda *et al.,* 2003). , No host specificity was observed in a separate study examing *E. coli* O78 strains from human, avian and cattle using MLST (Adiri *et al.* 2003). Some host-specific *Bacteroides* markers were also found to cross-react with feces from other animals (Ahmed *et al.,* 2012; Odagiri *et al.,* 2015). Therefore more studies need to be done to understand microbial host specificity and find the most appropriate method in identifying host-related characteristic in microorganisms.

1.4 General biology of *E. coli*

E. coli, a Gram negative, facultative anaerobe, is widely distributed in the intestine of human and warm-blooded animals and is one of the best studied model microorganisms since its discovery in 1885. This organism has been proposed to have two principal habitats: the intestinal tracts of mammals and birds; and the non-host environment (water/sediment). Several studies have found that *E. coli* can grow outside the gastrointestinal tract of its animal hosts in tropical and subtropical environments (Byappanahalli and Fujioka, 1998; Solo-Gabriele *et al.,* 2000; Anderson *et al.,* 2005). Naturalized *E. coli* strains have also been described in sand, sediment and water from temperate climates (Power *et al.,* 2005; Tymensen *et al.,* 2015; Kon *et al.,* 2007; Chandrasekaran *et al.,* 2015). Survival and colonization of *E. coli* in these distinct habitats are mediated by adaptations to nutrient availability, temperature, pH, osmolality, chemicals, solar radiation and the presence of competitive microflora.

The conditions in the non-host environment are very different compared to the gastrointestinal tract of warm-blooded animals. In order to survive in these alternate environments and their associated stress-related conditions (i.e., animal host to environment and transmission back to host), various strategies have been utilized by *E. coli* for enhanced survival. RpoS is one sigma factor (subunit of RNA polymerase) in *E. coli* that triggers a general stress response under stressful conditions (Battesti *et al.,* 2011). The increasing levels of RpoS alter the expression of several genes resulting in resistance to conditions like starvation, low or high pH, low/high osmolarity, oxidative stress, and DNA damage (Battesti *et al.,* 2011). RpoS also regulates genes that are involved in biofilm formation that allow bacteria to grow and survive under hostile conditions (Sheldon *et al.,* 2012). Biofilms are composed of an extracellular matrix

of different types of biopolymers (polysaccharides, proteins, nucleic acids, lipids, etc.). Biofilms provide a physical barrier that can protect bacteria from various stresses such as oxidative stress, disinfection and antibiotics (Stewart *et al.,* 2013). In addition to the general stress response triggered by RpoS, stress responses controlled by other genes have also been identified in *E. coli*. For example, it was found that the heat shock response in *E. coli* was mediated by sigma factor σ^{32} that can induce the expression of heat shock proteins such as GroE, DnaK, GrpE. (Schlesinger, 1990; Arsene *et al.,* 2000). Recently, a heat resistance genomic island (LHR) encoding putative heat shock proteins and proteases was identified in *E. coli* (Mercer *et al.,* 2015). The presence of this LHR was found to be correlated with *E. coli* heat resistance, enabling strains to survive 5 minutes of heat treatment at 60 ºC. Acid resistance is another stress response that enables certain *E. coli* strains to survive acidic environments such as the stomach environment of humans or animals, whose pH ranges from 1.5 to 3.5 (Marieb and Hoehn, 2010). Several acid resistance mechanisms have been identified such as the hydrogen-gas-producing formate hydrogen lyase (FHL) complex as well as a four amino acid dependent acid resistance (AR) systems (i.e. glutamate dependent system [AR1], arginine dependent system [AR2], lysine dependent system [AR3], and ornithine dependent system [AR4])[Kanjee and Houry, 2013]. In these systems, cellular protons (H+) can be decreased by consumption of the amino acid-based acid resistance system (De Biase and Lund, 2015). As a facultative anaerobe, oxidative stress is another challenging condition for *E. coli* as the reactive oxygen species such as superoxide (O_2^-), hydrogen peroxide (H_2O_2) and hydroxyl radical ($HO^{\cdot}$) can damage bacterial cells (Fu *et al.,* 2015). Superoxide dismutase is able to transform superoxide into hydrogen peroxide and catalase can then degrade hydrogen peroxide into H_2O and O_2 (Baez and Shiloach, 2013). In addition to heat resistance, acid resistance, and oxidative resistance, other stress response systems such as

osmotic stress responses (Record *et al.,* 1998) and bile stress (Kwan *et al.,* 2015) also exist in *E. coli*. Stress responses of *E. coli* are complex. Under stressful environments, complex regulatory systems ensure the general stress response as well as other specific stress responses to be activated, which can protect the bacteria from damage and/or death in both the host and the external environment.

Most *E. coli* living in the gastrointestinal tract of humans and animals are commensal strains. Only a small proportion of *E. coli* strains are pathogenic and which can cause intestinal or extra-intestinal disease. Six well characterized pathovars of *E. coli* have been described: enteroinvasive (EIEC), enteropathogenic (EPEC), enterohaemorrhagic (EHEC), enterotoxigenic (ETEC), enteroaggregative (EAEC), and uropathogenic (UPEC) (Nguyen *et al.,* 2005), of which the first five pathovars are associated with gastrointestinal disease. EIEC share closely related characteristics with *Shigella* and can cause invasive inflammatory colitis and sometimes dysentery. The pathogenesis of EIEC seems to be determined by a type III secretion system encoding proteins IpaA, IpaB, IpaC, and IpaD, which enable the bacteria to enter host cells (Kaper *et al.,* 2004). EPEC can cause diarrhoea by attaching to epithelial cells and effacing the microvilli of the intestine (McDaniel *et al.,* 1995). The pathogenesis of EPEC is found coded on a pathogenicity island known as the *locus of enterocyte effacement* (LEE) that is also found in EHEC strains (McDaniel *et al.,* 1995). EHEC are associated with bloody diarrhoea, non-bloody diarrhoea, and haemolytic uremic syndrome (HUS). *E. coli* O157:H7 belongs to the EHEC pathotype. The most important virulence factor of EHEC is a toxin known as Shiga toxin (Kaper and O'Brien, 1998). ETEC is the major cause of traveller's diarrhoea and infant diarrhoea in developing countries (Qadri *et al.,* 2005). Heat-labile and heat stable toxins are the major contributing virulence factors (Qadri *et al.,* 2005). EAEC are an aggregative form of *E. coli* that

adheres to the surface of Hep-2 cells and to each other in a "stacked-brick" pattern (Kaper *et al.,* 2004). Several virulence factors have been found in EAEC, however, none of them are exclusively associated with EAEC (Kaper *et al.,* 2004). UPEC can cause urinary tract infections (UTI) in human. It was found that about 80-90% of uncomplicated UTI cases were caused by UPEC (Martinez and Hultgren, 2002). Several virulence and fitness factors are involved in its pathogenesis, however, no UPEC specific virulence genes have been found as of yet (Luthje and Brauner, 2014). Virulence genes and their associated proteins can be considered as host-adaptation factors – elements necessary for adherence, colonization, invasion, and/or intracellular survival of their respective host. *E. coli* isolates originating from various animal hosts possess a broad spectrum of virulence genes, but patterns of association are often specific to certain animal hosts and not others (White *et al.,* 2011).

The incredible phenotypic diversity observed in *E. coli* as a bacterial species is also reflected in its genetic diversity. The first complete DNA sequence of the *E. coli* genome (K12 derivative MG1655) was published in 1997. The genome size was reported as 4.64 Mbp and contained about 4,500 genes. By 2016, there were 4,618 *E. coli* genomes available from NCBI. Among the *E. coli* genomes that have known host source information, the majority were isolated from humans (>1,100) and only 75 originated from other animals. The genome size of the various *E. coli* strains also varies significantly. For example, the genome of *E. coli* strain BL21 is 4.56 Mbp while *E. coli* strain EC4115 has a genome size of 5.70 Mbp (Lukjancenko *et al.,* 2010). With regard to genome content, in a comparative study of 61 *E. coli* genomes (Lukjancenko *et al.,* 2010), it was found that only 20% of genes were conserved across all studied strains and those conserved genes made up less than 10% of the pan-genome of this species, indicating that there is considerable genetic variation in the *E. coli* genome. As a single species, having such

genomic diversity is surprising, and some researchers suggest that *E. coli* may represent a species complex at the genetic level (Lukjancenko *et al.,* 2010).

Host-specificity is not well understood in *E. coli*, even though a considerable effort has been made to evaluate this. Enzymatic electrophoresis mobility was used as early as 1973 to study population genetic diversity in *E. coli* (Milkman, 1973). The technique was named multi-locus enzyme electrophoresis (MLEE) and which became a popular technique in bacterial population genetic studies thereafter. Based on the results of MLEE, *E. coli* isolates were assigned into four main phylogenetic groups: A, B1, B2, and D. This phylogenetic grouping has been supported by several other techniques, such as MLST (Gordon *et al.,* 2008), random amplified polymorphic DNA [RAPD] (Desjardins *et al.,* 1995), and restriction fragment length polymorphism [RFLP] (Desjardins *et al.,* 1995). Among these techniques, a triplex-PCR developed by Clermont *et al.* (2000) is most popularly used by the field due to its robustness and simplicity. In this method, the presence and/or absence of three targets *chuA, yjaA*, and the TSPE4.C2 DNA fragment were used to group *E. coli* into four phylogenetic groups (Clermont *et al.,* 2000). An updated phylogenetic grouping method (Clermont *et al.,* 2013) has been developed since then, and which is able to group *E. coli* into phylogroups A, B1, B2, C, D, E, F, and a cryptic clade I.

It has been demonstrated that different phylogroups of *E. coli* differ in their ability of exploiting sugars, their antibiotic resistance profiles, and growth-rate temperature relationships (Gordon, 2004). These phylogroups also differ in pathogenicity as most of the extra-intestinal pathogenic *E. coli* belong to the B2 and D phylogroups, while most of the intestinal non-pathogenic strains belong to group A and B1 phylogroups (Johnson and Stell, 2000; Pupo *et al.,* 1997; Picard *et al.,* 1999). The four phylogenetic groups also differ in their ecological niches. It

was observed that group A and B1 are more frequently isolated from the non-host environment than from animals (Walk *et al.,* 2007). In addition, group A and B2 were more frequently isolated from humans while group B1 was most prevalent in animals (Tenaillon *et al.,* 2010). Gordon and Cowling (2003) analyzed 497 *E. coli* strains isolated from different animal species, and it was found that group B1 was the most predominant phylogenetic group in bird isolates, while group B2 were more abundantly distributed in omnivores. The distributions of *E. coli* phylogenetic groups in different hosts indicate that *E. coli* may display a certain level of host preference or specificity, which challenge the general opinion that *E. coli* is a generalist being able to colonize different host species.

As discussed in the previous section (Section 1.3), a significant amount of research has focused on microbial source tracking applications using *E. coli* as a model organism, the results of which have yielded conflicting results. Some additional evidence to support the idea that *E. coli* exhibits a certain degree of host specificity has been observed by several other studies. It was found that avian septicemia *E. coli* strains were more virulent to chicks than *E. coli* strains isolated from newborn meningitis (Eliora, 2005). A human-specific *E. coli* clone of the B2 lineage was reported by Clermont *et al.,* (2008), which was found in human *E. coli* isolated from several continents but not found in any animal isolates. In another study, Kim *et al.* (1999) used a DNA based typing method named octamer-based genome scanning (OBGS) to group *E. coli* O157 isolates from human and cattle. It was found that lineage I contained mostly human isolates (81.8%) whereas most bovine strains fell into lineage II (78%). These studies suggest that *E .coli* populations, including pathogenic strains, may adapt and are selected for in the gastrointestinal tract of specific hosts.

1.5 Statistical tools for host specificity discovery

Various laboratory tools have been used to identify host-specific genetic patterns of *E. coli*, including fingerprinting methods such as pulsed field gel electrophoresis (PFGE)[Stoeckel *et al.,* 2004], repetitive sequence-based polymerase chain reaction (rep-PCR)[Carson *et al.,* 2003], amplified fragment length polymorphism (AFLP)[Leung *et al.,* 2004], and ribotyping (Stoeckel *et al.,* 2004). DNA sequence-based methods such as gene-specific identification (Khatib *et al.,* 2003; Khatib *et al.,* 2002), multilocus sequence typing (MLST)[Clermont *et al.,* 2011] and multi-spacer sequence typing (MSST) [White *et al.,* 2011] have also been used for studying these host-specific relationships in *E. coli*. For fingerprinting methods, various statistical methods have been used to evaluate host-specificity such as discriminant analysis (Mohapatra *et al.,* 2008), logistic regression (Carlos *et al.,* 2010), maximum similarity (Dombek *et al.,* 2000), average similarity (Hassan *et al.,* 2005), and nearest neighbour (Lyautey *et al.,* 2010). For DNA sequence-based methods, host specificity has most often been evaluated based on unique allelic profiles (such as sequence types used in MLST) [Miller *et al.,* 2006; Ram *et al.,* 2004] or through phylogenetic analysis (Clermont *et al.,* 2011; Ivanetich *et al.,* 2006). Commonly used statistical methods for DNA sequence phylogenetic analysis include unweighted pair group methods such as Unweighted Pair Group Method with Arithmetic Means (UPGMA), Neighbour Joining (NJ), Fitch-Margoliash (FM), Minimum Evolution Algorithms (MEA), Maximum Parsimony (MP), and Maximum Likelihood (ML) [Xiong, 2006]. UPGMA (Sokal and Michener, 1958) and NJ (Saitou and Nei, 1987) cluster samples based on the pairwise similarities of samples computed on the basis of fingerprints or DNA sequence alignments, and construct a tree to reflect the structure presented in the pairwise similarities of samples. FM (Fitch and Margoliash, 1967) and MEA (Rzhetsky and Nei, 1992) compare many alternative tree

topologies and selecting one that has the best fit between estimated distances in the tree and the actual evolutionary distances. A major drawback of these methods is that the actual sequence information is lost when all the sequence variation is reduced to pairwise similarities (Xiong, 2006). MP and ML methods are based directly on the sequence characters rather than on the pairwise similarities of samples. These methods count mutational events accumulated on the sequences and study evolutionary dynamics of each character. Searching from all possible tree topologies and considering every position in an alignment, MP chooses a tree that requires the minimum evolutionary changes to evolve to the current sequences, and ML uses probabilistic models to choose the best tree that has the highest probability or likelihood of reproducing the observed data.

These statistical clustering-type methods are all considered *unsupervised learning* methods, which aim to discover patterns in data and then classify samples into discreet groups (i.e., clusters). A key feature of unsupervised learning is that the group labels (e.g., the animal host origin of an individual *E. coli* isolate) are not observed as data. Unsupervised learning is useful for finding hidden structures (i.e., clusters) in unlabelled data. The other type of classification methods, that includes discriminant analysis (Mohapatra *et al.,* 2008) and logistic regression (Carlos *et al.,* 2010), are known as *supervised learning*, which infers patterns in observed data associated with group labels (Mohri *et al.,* 2012), with each sample having an input set of data and a label of the grouping. For data with known group labels, supervised learning methods are more powerful for finding the patterns/structures that are related to the group labels. In other words, supervised learning and unsupervised learning approaches have some clear distinctions: 1) the former aims to infer a classification function that can be used to classify new samples into the groups that are observed and known, while the latter aims to

classify samples into groups where the groups and their labels are not observed; and 2) the former requires labelled data, while the latter uses unlabelled data.

It is hypothesized that supervised learning methods may represent novel approaches for understanding population dynamics and host-specificity*;* this based on the following rationale: 1) the group label of each sample or host source (i.e., the animal host from which the microbe was isolated from), could be provided as part of data in supervised learning methods; and 2) supervised learning could identify the most informative sites and patterns in the microbe's DNA sequence polymorphisms as they pertain to their host source label, and consequently the patterns formed by these sites can be used to infer the hosts of unknown samples. Unsupervised learning methods that have been used in the past to try and characterize host specificity only cluster similar samples based on similarities computed from information on all sites in DNA sequence, and include information about single nucleotides that may be irrelevant to host-specificity. Irrelevant information across *all* single nucleotides to infer information about the host origin of a microbe, may result in misleading, poor-performing classification outcomes, the results of which will not necessarily be informative in the context of which animal hosts are found within a set of new samples. Conceptually, the genetic information on only a small number of genes/ITGRs, pieces of genes/ITGRs, or SNPs may be decisive for determining host-specificity, and supervised learning methods could be used to select the most relevant information regarding host specificity from the labeled data. Application of supervised learning tools to genome data are becoming more common in cancer and chronic disease epidemiology in which SNP biomarkers are being used to infer health outcomes (Onay *et al.,* 2006; Dinu *et al.,* 2012).

Many supervised learning methods are available for classification problems and that give simple yet explicit analytical functions for classification. Methods include logistic regression

(McCullagh and Nelder, 1989), log-binomial regression (McCullagh and Nelder, 1989), and logic regression (Ruczinski *et al.,* 2004) with each of these methods having good interpretability. Other non-analytical methods giving complex classification functions that are usually not interpretable include Boosting (Freund *et al.,* 1999), Artificial Neural Network (Anderson, 1995), and Support Vector Machine Learning (Cortes and Vapnik, 1995). An interpretable classification function is generally preferable for assessing a scientific hypothesis.

1.6 Research rationale and hypotheses

Patterns of microbial host specificity have been observed at all host-related taxonomic levels in biological systems. This observation supports the first over-arching hypothesis of this thesis that survival in competitive microenvironments, such as the GI tract, leads to natural evolutionary processes that drive microbes towards host adaptation and specificity. As previously discussed, several studies have demonstrated that a certain degree of host-specificity appears to exist in *E .coli* (Clermont *et al.,* 2008; Kim *et al.,* 1999; Eliora, 2005), although microbial source tracking methods have not been able to adequately resolve or discriminate genetic variability down to this level. In order for an *E .coli* strain to survive in the GI tract of a particular host, it must: 1) out-compete less adapted forms by rapidly sensing and responding to the physiological conditions of the GI tract (pH, nutrients, osmolarity, temperature, antibiotics, etc.), and consequently mobilize nutrients for growth and replication; 2) overcome/escape/ or avoid host immune defence and other stress factors that may inhibit growth within the microenvironment; and 3) replicate at a sufficient rate to avoid complete elimination through the GI tract. The extreme genetic diversity observed in *E. co*li as a species likely reflects the broad range of host species (mammals, birds, etc.) for which this microbe is known to colonize. As

such it is hypothesized that strains within this diverse species are actually adapted to survive within a specific animal host (or small group of related hosts) and therefore should be considered as host-specialists (as opposed to host-generalists). The corollary to this hypothesis is that genetic signatures across the *E. coli* pangenome may encode host-specific information (see below), and consequently be used to track sources of *E. coli* contamination in food, water, or the environment. These genetic signatures may also be important in informing public health about the potential emergence of pathogenic strains from one host to another (i.e., zoonooses).

The second major hypothesis of this thesis is that the ability of *E. coli* to colonize the gastrointestinal system of a specific animal host will largely depend on its ability to respond to the sensory stimuli (i.e. regulatory transcriptome [regulome]) perceived within the gastrointestinal system of a specific host and for which the microbe must outcompete other conspecific and heterospecific microbes for limited nutrients and colonization/replication sites within the gastrointestinal environment. This intense intra-and inter-specific bacterial competition within the gut, coupled with the enormous diversity in the cellular and molecular physiology of gastrointestinal microenvironments across animal species, will drive the evolution of the *regulome* of *E. coli* towards host-adaptation and specificity. For these reasons it is hypothesized that host-specific structuring in *E .coli* is governed primarily at the regulatory transcriptome level. In turn, host-specific information may be encoded in sequence polymorphisms within the regulome, and in particular, ITGR sequences of the genome. Therefore, I believe intergenic regions represent a unique target for assessing DNA sequence polymorphisms associated with host-specificity. In addition, I hypothesize that the application of supervised learning methods to DNA sequence analysis in these regions may represent a better bioinformatics approach for understanding population genetics and host-specificity in *E. coli*

rather than the traditional cluster-based phylogenetic approaches using unsupervised learning methods.

Intergenic regions containing promoters/transcription factors (repressor and activator)/enhancers (Buck *et al.,* 2000; Browning and Busby, 2004) are non-coding sequences assumed to be subject to less adaptive selection pressures (i.e., genetic drift) and therefore contain more sequence variability than coding regions (Drancourt *et al.,* 2004). However, the opposite may be true. Selection pressure in these non-coding regions may be greater than in structural genes due to their functional importance in sensory regulation: a microorganism must first sense its environment before it can respond to the conditions of that environment. Intergenic spacer regions containing promoter/enhancer genetic elements have proven to be useful in typing of some microorganisms such as *Bartonella henselae* (Li *et al.,* 2006), *Rickettsia* spp. (Fournier and Raoult, 2007), *Mycobacterium tuberculosis* (Djelouadji *et al.,* 2008) and *Coxiella burnetii* (Glazunova *et al.,* 2005). Li *et al.* (2006), used multispacer typing to study 126 *Bartonella henselae* isolates and 39 multi-spacer types were identified by analyzing 9 intergenic spacers sequences, which demonstrates that multi-spacer sequencing typing (MSST) is a powerful method for genotyping *B. henselae* at the strain level. In a study by Djelouadji *et al.* (2008), 32 multispacer sequence patterns had been generated based on 8 spacer sequences; and the distribution of MSST patterns of 101 observed *Mycobacterium tuberculosis* correlated well with the assignment of their phylogeographical lineages, providing further evidence of the reliability of using multispacer sequences as a genotyping method for microorganisms. In addition, it was shown that MSST demonstrated higher typing power than MLST. In a study by White *et al.,* (2011), analysis of only 3 intergenic spacers (2064 nucleotides in length) across 24 *E. coli* strains by MSST was able to generate almost similar phylogenetic trees as by MLST based on 7

housekeeping gene sequences (3423 nucleotides in length), validating the role of MSST in genotyping. Furthermore, in this study, MSST effectively grouped most of 248 *E .coli* isolates from different animal sources into their corresponding phylogenetic groups and displayed some degree of host adaptation and clustering. Thus, intergenic sequences likely play an important role in *E. coli* evolution towards niche specificity.

1.7 Thesis objectives and overview

The overall objectives for this thesis were to:

i. Assess host-specificity of *E. coli* using a supervised learning logic-regression-based analysis of single nucleotide polymorphisms in intergenic regions (Chapter Two)

ii. Evaluate different intergenic regions in host source prediction in *E.coli* (Chapter Three and Chapter Four)

iii. Evaluate whether genetic variations associated with host-adaptation/ specificity in intergenic regions yield biologically relevant phenotypes important for survival of the *E. coli* genotypes in a particular niche (Chapters Four, Five and Six).

In total, there are seven chapters in this thesis. Five of these chapters (Chapters Two-Six) represent papers published, accepted, or submitted for publication. The five data chapters constitute two major Sections (I and II) of the thesis. Section I (encompassing Chapters Two, Three, and Four) focuses on addressing population genetics of *E. coli* based on intergenic sequence polymorphisms and the use of logic regression as a means of evaluating host-specificity. Section II of this thesis (encompassing Chapters Four, Five and Six) attempts to evaluate whether the host-specific variations observed in ITGRs contribute to the adaptive evolution of certain phenotypes in *E. coli*. The central hypothesis of this thesis - that genetic

variation in the regulome is important for driving evolution of host-specificity in *E. coli* - necessitates that the genetic variation observed in ITGRs be adaptive in nature and subject to evolutionary selection. Since the vast majority of genetic variation in biological systems is associated with non-adaptive and non-deterministic evolution (i.e., neutral selection theory, genetic drift), Section II attempts to link genetic variation in ITGRs to functionally-important and biologically plausible phenotypic outcomes that may be important for *E. coli* survival in a particular niche.

Chapter Two (Section I) of this thesis specifically addresses the application of logic regression as a supervised learning method for bioinformatic analysis of sequence variation in three intergenic regions (ITGRs) as a means of understanding host-specificity in *E. coli*. The three ITGRs considered in this analysis were previously shown to provide some degree of survival fitness in the host gut microenvironment (White *et al.,* 2011). Polymorphisms within these ITGRs were examined among 780 strains of *E. coli* isolated from 15 different animal hosts, and for which unique SNP patterns related to host-specificity were observed. In Chapter Three (Section I), the degree of host-specific information encoded among six ITGRs was compared among *E. coli* isolates obtained from human/animal feces, as well as 80 different ITGRs from published whole genome sequences. The data demonstrated that a high degree of host-specificity was observed in *E. coli*, but that not all ITGRs encode the same degree of host-specific information. In Chapter Four, (Sections I/II), logic regression analysis of ITGRs was used to identify naturalized strains of *E. coli* in non-host environments (i.e. wastewater), suggesting that some *E. coli* strains may have evolved towards adaptation and survival in wastewater. Naturalized wastewater strains of *E. coli* were shown to possess a number of stress-related adaptive mechanisms known to be important for survival of *E. coli* in non-host

environments (generalized stress response, universal stress response and heat resistance). In Chapter Five (Section II), it was demonstrated that these strains appear to be: i) phenotypically resistant to wastewater treatment, ii) have a widespread geographical distribution, and iii) appear to be specifically adapted to survival in wastewater. In fact, the ITGR polymorphisms in naturalized wastewater *E. coli* strains were so distinct from human and animal strains of *E. coli* that PCR assays were developed and shown to be valuable as a microbial source tracking targets for wastewater pollution. Further phenotypic characterization of the these wastewater strains was carried out in Chapter Six (Section II), demonstrating that ITGR variations in these naturalized strains correlated with the phenotypic properties of increased biofilm production and bacterial flagellar expression - mechanism known to be important for survival of bacteria in the non-host environment and during water treatment. Chapter Seven is a general discussion of the research finding of this thesis, highlighting the important of these findings as well as a discussion on the limitations and direction of future studies.

Chapter Two : ASSESSING HOST-SPECIFICITY OF *E.COLI* USING SUPERVISED LEARNING LOGIC-REGRESSION-BASED ANALYSIS OF SINGLE NUCLEOTIDE POLYMORPHISMS IN INTERGENIC REGIONS [1]

2.1 Abstract

Host specificity in *E. coli* is widely debated. This Chapter used supervised learning logic-regression-based analysis of intergenic DNA sequence variability in *E. coli* in an attempt to identify single nucleotide polymorphism (SNP) biomarkers of *E. coli* that are associated with natural selection and evolution towards host specificity. Seven-hundred and eighty strains of *E. coli* were isolated from 15 different animal hosts. Logic regression was utilized for analyzing DNA sequence data of three intergenic regions (flanked by the genes *uspC-flhDC*, *csgBAC-csgDEFG*, and *asnS-ompF*) to identify genetic biomarkers that could potentially discriminate *E. coli* based on host sources. Across 15 different animal hosts, logic regression successfully discriminated *E. coli* based on animal host source with relatively high specificity (i.e., among the samples of the non-target animal host, the proportion that correctly did not have the host-specific marker pattern) and sensitivity (i.e., among the samples from a given animal host, the proportion that correctly had the host-specific marker pattern), even after fivefold cross validation. Permutation tests confirmed that for most animals, host-specific intergenic biomarkers identified by logic regression in *E. coli* were significantly associated with animal host source. The highest level of biomarker sensitivity was observed in deer isolates, with 82% of all deer *E. coli* isolates displaying a unique SNP pattern that was 98% specific to deer. Fifty-three percent of human

[1] A version of this chapter has been published, the citation of which is: Zhi, S., Q. Li, Y. Yasui, T. Edge, E. Topp, and **N. F. Neumann**. 2015. Assessing host-specificity of *Escherichia coli* using a supervised learning logic-regression-based analysis of single nucleotide polymorphisms in intergenic regions. *Molecular Phylogenetics and Evolution,* 92:72-81.

isolates displayed a unique biomarker pattern that was 98% specific to humans. Twenty-nine percent of cattle isolates displayed a unique biomarker that was 97% specific to cattle. Interestingly, even within a related host group (i.e., Family: Canidae [domestic dogs and coyotes]), highly specific SNP biomarkers (98% and 99% specificity for dog and coyotes, respectively) were observed, with 21% of dog *E. coli* isolates displaying a unique dog biomarker and 61% of coyote isolates displaying a unique coyote biomarker. Application of a supervised learning method, such as logic regression, to DNA sequence analysis at certain intergenic regions demonstrates that some *E. coli* strains may have evolved to become host-specific.

2.2 Introduction

E. coli, a common Gram-negative facultative anaerobe, is widely distributed in the intestine of humans and animals and is one of the best-studied model microorganisms since its discovery in 1885 (Escherich, 1885). This bacterial species is normally regarded as a host generalist being able to colonize a wide variety of warm-blooded vertebrate hosts and some reptiles (Tenaillon *et al.,* 2010). An increasing number of studies, however, demonstrate that host preference and/or specificity is not uncommon in the microbial world (Poveda *et al.,* 1994; Xiao *et al.,* 2004; Taylor *et al.,* 2004; Mandel *et al.,* 2009; Fauvart and Michiels, 2008), and as outlined in Chapter One of this thesis, some evidence exists to suggest that *E. coli* may also exhibit some degree of host preference/specificity (Clermont *et al.,* 2008; Eliora, 2005; Kim *et al.,* 1999). For instance, the four main phylogenetic groups (A, B1, B2, and D) of *E. coli* differ in their ecological niches. In humans the most dominant phylogenetic group of *E. coli* is group A (40.5%) followed by B2 (25.5%), while in animals group B1 (41%) is most prevalent (Tenaillon

et al., 2010). Indeed, a human-specific *E. coli* clone of the B2 lineage has been previously reported (Clermont *et al.,* 2008).

Various laboratory tools have been used to identify host-specific genetic patterns of *E. coli*, including fingerprinting methods PFGE (Stoeckel *et al.,* 2004), rep-PCR (Carson *et al.,* 2003), AFLP (Leung *et al.,* 2004), and ribotyping (Stoeckel *et al.,* 2004), but with limited success. DNA sequence-based methods such as gene-specific identification (Khatib *et al.,* 2003; Khatib *et al.,* 2002), multilocus sequence typing (MLST) (Clermont *et al.,* 2011) and multi-spacer sequence typing (MSST) (White *et al.,* 2011) have also been used for studying these host-specific relationships in *E. coli*. For fingerprinting methods, various statistical methods have been used to evaluate host-specificity such as discriminant analysis (Mohapatra *et al.,* 2008), logistic regression (Carlos *et al.,* 2010), maximum similarity (Dombek *et al.,* 2000), average similarity (Hassan *et al.,* 2005), and nearest neighbor (Lyautey *et al.,* 2010). For DNA sequence-based methods, host specificity has most often been evaluated based on unique allelic profiles (such as sequence types used in MLST) [Miller *et al.,* 2006; Ram *et al.,* 2004] or through phylogenetic analysis (Clermont *et al.,* 2011; Ivanetich *et al.,* 2006), and statistical methods for DNA sequence phylogenetic analysis include Unweighted Pair Group Method with Arithmetic Means, neighbor joining, Fitch-Margoliash, minimum evolution algorithms , maximum parsimony, and maximum likelihood (Xiong, 2006).

As discussed in the previous chapter, these statistical clustering-type methods are all considered *unsupervised learning* methods, which aim to discover patterns in data and then classify samples into discreet groups (i.e., clusters) and which are independent of host source labels. Supervised learning methods may represent better approaches for understanding population genetic diversity and host-specificity in *E. coli,* particularly in the context of using

these statistical tools to analyze DNA sequence information since the group label of each sample or host source (i.e., the animal host from which the *E. coli* was isolated from) could be provided as part of data in supervised learning methods indentifying informative sites and patterns in *E. coli* DNA sequence polymorphisms (i.e., SNP) as they pertain to their host source label. This is unlike unsupervised learning methods that may use information from all sites in the DNA sequence including nucleotides that may be irrelevant to host-specificity. Application of supervised learning tools to genome data are common in cancer (Onay *et al.,* 2006) and chronic-disease epidemiology (Dinu *et al.,* 2012) in which SNP biomarkers are being used to infer disease risk and outcomes. Prior to starting this thesis, supervised learning methods had not been used in ITGR DNA sequence-based analytical methods to study microbial host adaptation/specificity.

In this chapter of the thesis, logic regression was used as a supervised learning method for analysis of multi-spacer sequencing (MSS) data from *E. coli* isolates collected from various animal hosts. The rationale for using MSS relates to the fact that intergenic spacer regions in the genome often contain promoter and enhancer elements that regulate the expression of genes and therefore relate to changes in cell phenotypes, adaptive functioning, and sensation (He and Saedler, 2005; Wray *et al.,* 2003; Schofield and Watson, 1986; Smits *et al.,* 2006). It is hypothesized that the ability of *E. coli* to colonize the gastrointestinal system of a specific animal host will largely depend on its ability to respond to the sensory stimuli (i.e., regulatory transcriptome) perceived within the gastrointestinal system of a specific host, and consequently, determine whether the microbe can outcompete other conspecific and heterospecific microbes for limited nutrients and replication sites within the host's gastrointestinal environment. In turn, host-specificity may be encoded in sequence polymorphisms within intergenic sequences.

2.3 Material and methods

2.3.1 *E. coli* isolates

In total, 845 *E. coli* strains isolated from 15 different host sources were used in this study (Table 2-2). Animal *E. coli* isolates were sourced from three bacterial libraries originally established from geographically segregated areas in Canada: a) the Hamilton/Toronto region in Ontario as described by Edge and Hill (2007), b) from Ottawa (Ontario), Lennoxville (Quebec) and Brandon (Manitoba) as described by Lyautey *et al.* (2010), and c) from southern Alberta as described by White *et al.* (2011) [bovine isolates]. Human *E. coli* strains were isolated from clinical fecal swabs submitted to the Edmonton site of the Alberta Provincial Laboratory for Public Health (ProvLab) for routine microbiological testing, and adhered to all ethics requirements (File #: Pro00005478_CLS3 at the University of Alberta). In the studies of Edge and Hill (2007) and Lyautey *et al.* (2010), *E. coli* isolates had been previously screened by rep-PCR (BOX elements and enterobacterial repetitive intergenic consensus sequences [ERIC]), and unique rep-PCR isolates were selected for this study in order to reduce likelihood of clonal representation of *E. coli* from single animals. In samples collected from Alberta (bovine and human fecal samples), the likelihood of clonal representation was reduced by only collecting one *E. coli* isolate from individual animals or people. *E. coli* was isolated from animal fecal samples as described by Edge and Hill (2007), Lyautey *et al.* (2010) and White *et al.* (2011). All presumptive human and animal *E. coli* isolates were confirmed as *E. coli* through biochemical analysis using a Vitek Bacterial Identification System (BioMerieux) according to manufacturer's instructions and protocols at the ProvLab. Host labels were assigned to *E. coli* isolates based on the animal host from which they were isolated.

2.3.2 PCR and sequence alignment

All *E. coli* isolates were grown in Tryptic Soy Broth (TSB) and genomic DNA extracted from *E. coli* TSB cultures using DNeasy Blood & Tissue kits (QIAGEN) according to the manufacturer's instructions. Three intergenic regions were selected for analysis and similar to those identified in a previous study (White *et al.,* 2011): 1) the *csgBAC-csgDEFG* region which regulates the transcription of regulators of the synthesis of curli fimbriae; 2) *uspC-flhDC* region which regulates the transcription of the master regulator of flagellum biosynthesis; and 3) the *asnS-ompF* region which regulates the transcription of the regulator of an outer membrane protein related to low-osmolality conditions. Intergenic targets were amplified separately, but using the same PCR conditions: initial denaturation at 95^{o}C for 4 min, 33 cycles of 95 ^{o}C for 30 s, annealing at 58 ^{o}C for 30 s, and 72 ^{o}C for 1 min, and followed by a 7 min extension at 72 ^{o}C.

Primers used for PCR amplification are listed in Table 2-1, and were designed according to Zaslaver *et al.* (2006). The total volume of each PCR reaction was 25 µL and contained 5uL of *E. coli* genomic DNA template, 2.5 µL of 10×PCR buffer, 1.75 U *Taq* polymerase, 0.5 µL of 10 mM dNTP mix, 1.25 µL of each primer (0.5 pmol/ uL), 2.5 µL of 25mM Mg^{2+}, and 11.75 µL of molecular biology degree water.

All PCR products were sequenced bidirectionally using Sanger sequencing by either Macrogen Inc. (South Korea) or at the University of Calgary Genetic Analysis Laboratory in Calgary, Alberta Canada (http://www.ucalgary.ca/dnalab/sequencing). SeqMan Pro (DNASTAR) was used to align the bidirectional sequences. Sequences were aligned using the ClustalX 2.0 (Larkin *et al.,* 2007) program. The multiple sequence alignment was manually edited to trim the 5'and 3' regions that contained missing data. Sequences from the three intergenic regions were

assembled into a concatenated single file for each *E. coli* strain. Multiple sequence alignment files are available upon request.

Table 2-1. PCR primers used in this study.

Target	Primer	Primer sequence (5'-3')	Reference
asnS-ompF	ompF-F	TACGTGATGTGATTCCGTTC	(Zaslaver *et al.,* 2006)
	ompF-R	TGTTATAGATTTCTGCAGCG	
csgBAC-csgDEFG	csgD-1	GGACTTCATTAAACATGATG	(Zaslaver *et al.,* 2006)
	csgD-2	TGTTTTTCATGCTGTCAC	
uspC-flhDC	flhDC-F	GAGGTATGCATTATTCCCACCC	(White *et al*., 2011)
	flhDC-R	TGGAGAAACGACGCAATC	

2.3.3 Phylogenetic analysis

Maximum likelihood (ML) phylogenetic analysis was performed on concatenated sequences using RAxML Blackbox (Stamatakis *et al.,* 2008) using default settings (http://phylobench.vital-it.ch/raxml-bb/) based on the multiple sequence alignment generated. The phylogenetic tree was edited to map against different animal host groupings (i.e., color coded) using the Interactive Tree Of Life (iTOL) online tool (Letunic and Bork, 2007; Letunic and Bork, 2011).

2.3.4 Logic-regression-based statistical analysis

In this study, logic regression (Ruczinski *et al.,* 2004) was used as the method of supervised learning classification for distinguishing between *E. coli* isolates obtained from various animal host species based on informative SNPs in concatenated sequences. In the context of logic regression, SNP intersection captures a situation where two or more *E. coli* SNPs jointly influence biological functions related to host-selection, while SNP union captures a situation where two or more *E. coli* SNPs are redundant in their biological effects (i.e., genetic heterogeneity). In an SNP intersection, all of the relevant SNPs in the intersection set must exist in their respective specific genotypes for *E. coli* to live in a certain host, where one, or a subset, of the set is insufficient. In an SNP union, any SNP in the union set taking a specific genotype is sufficient for *E. coli* to live in a certain host.

SNP intersections and unions were expressed mathematically as Boolean logics such as $(X_1 \wedge X_2) \vee X_3^{\,c}$, where X's are indicators of SNP genotypes. "$\wedge$", "$\vee$" and "^{c}" represents intersection (AND), union (OR), and complement (NOT), respectively. The systematic part of the logic regression model applied here takes the form:

$$\text{logit}\,(E[Y]) = \beta_0 + \beta_1 L_1 + \beta_2 L_2 + \ldots + \beta_p L_p$$

where,

- Y is a binary variable, an indicator for being from one *E. coli* host versus being from the other hosts,
- $\beta_0, \beta_1, \ldots \beta_p$ are the parameters indicating the degrees of association between the SNP patterns (L's) and the host indicator (Y), and
- $L_1, L_2, \ldots, L_p$ are SNP patterns that are Boolean combinations (called "trees") of indicators of SNP genotypes (called "leaves") in the *E. coli* genes.

A massive number of potential models can be built with varying sizes (i.e., number of trees and leaves). Thus, the model building requires significant computational resources. A simulated annealing algorithm is used in logic regression to select the trees and leaves adaptively based on deviance as the model fit measure for finding the best model. To limit computational burdens, the maximum size of the model was limited to two trees and 10 leaves.

A logic regression model was built for each host distinguishing between the samples from the animal host under analysis and those from the other animals. In each of the analyses, there was a target host and a 0/1 binary outcome set up for each sample according to its host label: for host 'X' as a target, the outcome was either 1 (i.e., sample that is collected from the host X) or 0 (i.e., a sample that is collected from the other hosts).

Across the dataset, each *E. coli* sample was labeled with the name of the host from which it was collected. Although each sample was collected from the labeled host, theoretically the *E. coli* isolate obtained may be able to live in other animal host(s). For example, *E. coli* isolates collected from human specimens could be either host-specialists (i.e., only survive in humans), or host generalists (i.e., could live within a specific group of hosts that include humans). Considering the potential for host specificity and host generality to co-exist in this bacterial species from the same host, the focus of this study was to identify a highly host-specific SNP pattern for a given host (i.e., a SNP pattern that is rarely found among the samples from hosts other than the specific host of interest) and evaluate what proportion of all *E. coli* isolates from that specific host carried the host-specific SNP pattern. The proportion of samples from the specific host that carry the host-specific SNP pattern is referred to as *sensitivity*, while the proportion of samples from hosts other than the specific host of interest that do not carry the

host-specific pattern specific to the host of interest is referred to as *specificity*. Since there is a tradeoff between sensitivity and specificity, our scheme for selecting a classification pattern was to set a specificity criterion to as close to 100% as possible (i.e., we achieved ≥96% across all host species surveyed, Table 2-2) and select the SNP pattern that had the highest sensitivity. This reduced the likelihood that genetic marker patterns discovered represented those of host generalists.

2.3.5 Assessing biomarker validity and statistical significance

Since logic-regression attempts to fit a very flexible model to a set of observations and their known outcomes, the 'overfitting' problem, a common issue among supervised-learning methods, is of concern. Overfitting refers to a situation where a flexible method models the observed data at hand too faithfully and incorporates features of the observed data at hand that are not generalizable to other sets of data on the same natural phenomenon. Permutation test and fivefold cross-validation were used as an attempt to prevent overfitting and/or evaluate the performance of the final model unbiasedly.

A permutation test (10,000 runs) was performed for each host to assess the validity of the identified host-specific genetic patterns as determined by logic regression and evaluate the significance of this association. The host labels were randomly permuted and the data analyzed by the logic regression method. The number of cases that had higher performance value (measured by the mean of sensitivity and specificity) than the result based on original data was counted. The *p*-value was calculated by dividing this number by 10,000.

Cross-validation involves randomly partitioning the observed data into mutually exclusive subsets and fitting and testing a model with different subsets. In this study, a fivefold

cross-validation was applied for evaluating the performance of the identified SNP patterns so as to ensure that model performance measures (sensitivity and specificity estimates) were not inflated as a result of overfitting. The data sequences from one host were randomly divided (i.e., computer randomization) into five subsets each having an equal number of samples. Four subsets of data were used as training data for logic regression analysis and the last data subset were used as testing data for evaluating the logic regression model that was identified based on the training data.

2.4 Results

2.4.1 PCR amplification of selected intergenic regions

The percent of *E. coli* strains yielding positive PCR results for the three intergenic regions from various animal hosts are shown in Table 2-2. In total, among the 845 *E. coli* isolates used in this study, 97.5%, 98.3%, and 98.1% of all *E. coli* isolates were PCR positive for the intergenic sequences *uspC-flhDC*, *csgBAC-csgDEFG*, and *asnS-ompF*, respectively. No apparent pattern of host-specificity was observed directly from PCR results in that none of the target regions failed to amplify from *E. coli* isolates collected from a specific animal host group. However, an interesting observation was made in a subset of pig *E. coli* isolates, in which insertion sequence (IS) elements were found in the *csgBAC-csgDEFG* region of 14 *E. coli* isolates obtained from pigs, (i.e., 21% of pig isolates). Thirteen of the 14 *E. coli* isolates carried the IS1 element, while the remaining isolate had an IS4 element in this intergenic region. These insertion element sequences within the *csgBAC-csgDEFG* locus were unique to pig isolates. Two chicken *E. coli* isolates were found to carry IS1 and IS903 insertional elements but in the *uspC-flhDC* locus. *E. coli* isolates with insertion sequences were not included in subsequent logic regression analysis for SNP biomarker pattern analysis.

Table 2-2. *E. coli* isolates collected from different host animals and PCR results for targeted intergenic regions.

Host	No. of isolates	No. of PCR positive isolates (%)			No. of isolates used for logic regression model building
		uspC-flhDC	*csgBAC-csgDEFG*	*asnS-ompF*	
Bovine	124	123 (99.2)	124 (100)	121 (97.6)	120
Cat	21	21 (100)	21 (100)	21 (100)	21
Dog	63	62 (98.4)	63 (100)	62 (98.4)	61
Deer	52	52 (100)	50 (96.2)	50 (96.2)	48
Goose	61	57 (93.4)	59 (96.7)	59 (96.7)	54
Human	105	105 (100)	105 (100)	105 (100)	105
Chicken	66	62 (93.9)	65 (98.5)	63 (95.5)	59
Moose	15	15 (100)	14 (93.3)	15 (100)	14
Muskrat	58	56 (96.6)	58 (100)	58 (100)	56
Horse	47	46 (97.8)	44 (93.6)	47 (100)	44
Pig	68	67 (98.5)	66 (97.1)	65 (95.6)	49
Coyote	52	49 (94.2)	51 (98.1)	50 (96.2)	44
Gull	20	20 (100)	20 (100)	20 (100)	18
Beaver	46	42 (91.3)	44 (95.7)	46 (100)	40
Sheep	47	47 (100)	47 (100)	47 (100)	47
Total	845	824 (97.5)	831 (98.3)	829 (98.1)	780

Samples in which all three intergenic regions were present (780 isolates in total) were selected for multiple sequence alignment. Multiple sequence alignment for all intergenic regions was concatenated together for each of the isolates and followed the order of *asnS-ompF*, *uspC-flhDC*, and *csgBAC-csgDEFG*. The final sequence alignment contained 2033 nucleotide sites including gaps introduced during multiple sequence alignment, among which 583 sites were polymorphic (defined as one isolate's sequence being different from all others at that base

position) among all the isolates. One hundred and sixty-seven polymorphic sites were found in the 611 bp *asnS-ompF* region; 277 polymorphic sites were found in the 699 bp *uspC-flhDC* region; and 139 polymorphic sites were found in the 723bp *csgBAC-csgDEFG* region. These polymorphic sites were then used for logic regression model building.

2.4.2 Logic regression analysis of DNA sequences from *E. coli* isolates from various hosts

For each animal host, a logic regression model was firstly fit for classifying *E. coli* isolated from the animal host vs. the rest of *E. coli* samples isolated from hosts other than the host of interest (Table 2-3). The measures of model performance associated with these logic models (i.e., specificity and sensitivity estimated from the fivefold cross-validation) are shown in Table 2-4. A graphical explanation of one example of the models is shown in Figure 2-1. High sensitivity as well as high specificity was observed in several animal host groups. Under fivefold cross-validation the highest sensitivity was observed in deer, with 82% of isolates displaying the SNP biomarker pattern with high specificity to deer (i.e., 98%). The second and third highest sensitivity of 77% and 67% were observed in muskrat and moose, respectively, with their specificity all equal to 99%. In humans, 53% of isolates carried a human specific SNP biomarker pattern with 98% specificity after fivefold cross validation. In general, the independent fivefold cross-validation exercise supported the potential presence of host-specific *E. coli* isolates within many of the representative animal hosts. In some cases, notable reductions in sensitivity were observed after fivefold cross-validation in some animal hosts. For example, only 4% of *E. coli* isolates from cats displayed a host-specific biomarker pattern after fivefold cross-validation, compared to 62% when all isolates were used in the original analysis. Although this result may imply a flaw in overfitting of the model, it is important to note that the 3 intergenic loci used for

Table 2-3. Host-specific SNP biomarker patterns identified by logic-regression-based analysis*#

Host Source	SNP pattern
Bovine	-2.08 -22.3 * (flh837_G and flh1243_A) +3.15 * ((((not omp132_T) or csg1938_C) or (flh1206_A or (not omp541_T))) or (((not omp593_A) and (not csg1505_A)) or (flh807_T or (not omp264_T))))
Cat	0.08 -19.3 * ((not omp396_T) or (not omp259_G)) -4.26 * (((omp168_T and omp54_T) and (flh1012_G and omp445_C)) and ((flh642_C or (not flh760_T)) and (flh632_A and flh761_C)))
Dog	-7.35 +4.19 * (((omp143_A and (not flh1196_A)) or flh814_C) or omp259_C) +3.37 * ((((flh1247_A or csg363_T) or flh1170_G) and ((not omp195_T) and ((not flh965_A) or (not csg1381_-))))
Deer	-5.49 +7.93 * ((((not flh745_C) and (not csg1840_A)) or (not flh1183_G)) or (((not flh1273_-) or (not flh925_G)) or omp256_T)) +6.97 * (((omp376_A and flh719_A) or flh1069_T) or (not omp375_G))
Goose	-2.2 +5.76 * ((((not omp438_T) or csg1878_G) or (csg1678_C or flh556_T)) or ((csg1530_A or csg1933_T) or ((not omp133_T) or omp440_C))) -3.27 * ((not csg1472_G) and (not omp7_T))
Human	-3.73 +6.46 * ((((not csg2029_T) or flh788_C) or (csg1519_C or flh902_C)) or ((not csg1608_C) or (omp17_A or flh747_T))) +2.13 * ((csg1679_T or (not omp587_A)) or csg1539_A)
Chicken	-1.52 -3.22 * (((not omp314_G) or flh1156_A) and (not csg1473_T)) +6.3 * (((omp290_T or csg1842_A) or flh1042_A) or ((omp181_T or flh1074_C) or ((not omp321_T) or (not omp458_T))))
Moose	-0.296 +4.6 * ((not flh1302_T) or ((omp259_C or flh635_C) or flh762_C)) -6.47 * ((((not flh1169_C) and (not flh1302_A)) and (csg1389_G and omp130_C)) or ((not omp396_T) or flh1166_G))
Muskrat	-3.79 -5.81 * ((not csg1836_C) and ((not csg1679_T) or omp216_T)) +6.46 * ((((not omp434_C) and flh32_-) and ((not csg1398_A) and (not flh1147_G))) and (csg1430_G and ((not omp203_G) and (not omp564_A))))
Horse	2.94 -7.32 * (((not flh772_G) and (flh792_C and (not omp155_A))) and (((not flh876_G) and (not csg1647_G)) or omp4_T)) +4.26 * (omp564_C and ((flh745_C and csg1840_G) or (not flh1105_A)))
Pig	-1.73 -5.97 * ((((not csg1698_A) and flh1213_T) and (not omp147_G)) and ((csg1398_G and (not omp284_T)) and ((not csg1548_G) and csg1691_G))) +3.73 * ((omp120_T or flh581_A) or omp314_T)
Coyotes	-2.01 -2.68 * ((omp564_C and ((not omp197_G) and flh1247_G)) or (not flh716_A)) +7.99 * (((flh965_A or omp377_C) or (omp336_- or (not flh889_C))) and ((not csg1574_A) and flh1153_A))
Gull	-8.76 +5.39 * (((flh871_T or (not omp576_-)) and (not flh673_-)) or ((omp591_G or (not omp582_C)) or (flh726_G or omp211_C))) +5.4 * (((not csg1656_T) or flh614_C) and flh745_C)
Beaver	-1.23 -5.52 * ((flh873_C or (not csg1430_A)) and (not omp195_T)) +5.11 * (((omp195_T or (not csg1719_C)) and omp259_G) and ((omp564_C and csg1486_C) and (csg1840_A and (not flh1247_T))))
Sheep	3.76 -6.69 * ((((not omp7_A) and (not flh1304_G)) and (csg1720_A and (not flh1214_C))) and ((omp564_A or (not flh1247_T)) and ((not omp334_G) and (not flh652_A)))) -4.42 * (csg1719_C or (not omp145_G))

* Analysis was carried out in the intergenic regions of the *csgBAC-csgDEFG* (listed as csg in models above), *uspC-flhDC* (listed as flh in models above), and *asnS-ompF* (listed as omp in models above) as described in Materials and Methods.

In reading the model output a value designated by 'flh226_G' refers to the guanine base at nucleotide position 226 in the 5' to 3' multisequence alignment of the *uspC-flhDC* intergenic loci.

typing were also highly polymorphic in cats compared to most other animals (Table 2-5). Fivefold cross-validation required that the 21 cat isolates be partitioned into 5 random groups, and as such the low sample sizes within the 5 groups (i.e., 4), coupled with the high polymorphism in these intergenic regions of cat *E. coli* isolates, likely resulted in low sensitivity outcomes associated with cross validation. Similar observations in reduced sensitivity were observed for goose and gull *E. coli* isolates (Table 2-4), with these animal hosts also having considerable polymorphism within these intergenic loci (Table 2-5). However, *E. coli* isolates from moose displayed the greatest SNP variation per host in these loci (Table 2-5) but robust models of specificity were still observed in this animal species after fivefold cross-validation and with only 14 isolates comprising the pool of isolates examined. Interestingly, even within a related host group (i.e., Family: Canidae [domestic dogs and coyotes]), highly specific SNP biomarkers (98% and 99% specificity for dog and coyotes, respectively) were observed, with 21% of dog *E. coli* isolates displaying a unique dog biomarker and 61% of coyote *E. coli* isolates displaying a unique coyote biomarker.

To assess the statistical significance of the identified host-specific genetic patterns in logic regression models, a permutation test (10,000-runs) were performed for each animal hosts (Table 2-4), permuting the host labels randomly. It was demonstrated that for most of the animal hosts there is statistical evidence ($p < 0.05$) that host-specific biomarker patterns identified by logic regression models were associated with the host source, except for goose, beaver, and gull (p-values > 0.05).

Table 2-4. Logic regression analysis of *E. coli* samples classified according to host animal of origin and evaluation of logic models using fivefold cross validation and permutation testing.

Host Source	Number of Samples	Logic regression		Fivefold cross validation		Permutation test
		Sensitivity	Specificity	Sensitivity	Specificity	*P*-value
Bovine	120	0.53	0.97	0.29	0.97	0.0001
Cat	21	0.62	0.98	0.04	0.99	0.03
Dog	61	0.59	0.96	0.21	0.98	0.0001
Deer	48	0.94	0.99	0.82	0.98	0.0001
Goose	54	0.31	0.99	0.05	0.99	0.06
Human	105	0.53	1	0.53	0.98	0.0001
Chicken	59	0.66	1	0.54	0.99	0.0001
Moose	14	0.86	1	0.67	0.99	0.003
Muskrat	56	0.77	1	0.77	0.99	0.0001
Horse	44	0.43	1	0.36	1	0.009
Pig	49	0.43	1	0.44	0.99	0.003
Coyote	44	0.61	1	0.61	0.99	0.0001
Gull	18	0.39	1	0.05	0.99	0.5
Beaver	40	0.35	1	0.35	1	0.09
Sheep	47	0.64	0.98	0.46	0.99	0.0001

2.4.3 A comparison of logic regression and maximum likelihood phylogeny for evaluating host specificity in *E. coli*

A comparison between supervised (logic regression) and unsupervised (ML phylogeny) methods for evaluating host specificity of *E. coli* based on DNA sequence analysis was performed (Figure 2-2). Figure 2-2 illustrates an ML phylogenetic tree of all 780 *E. coli* isolates constructed based on the concatenated intergenic sequences and associated with host source

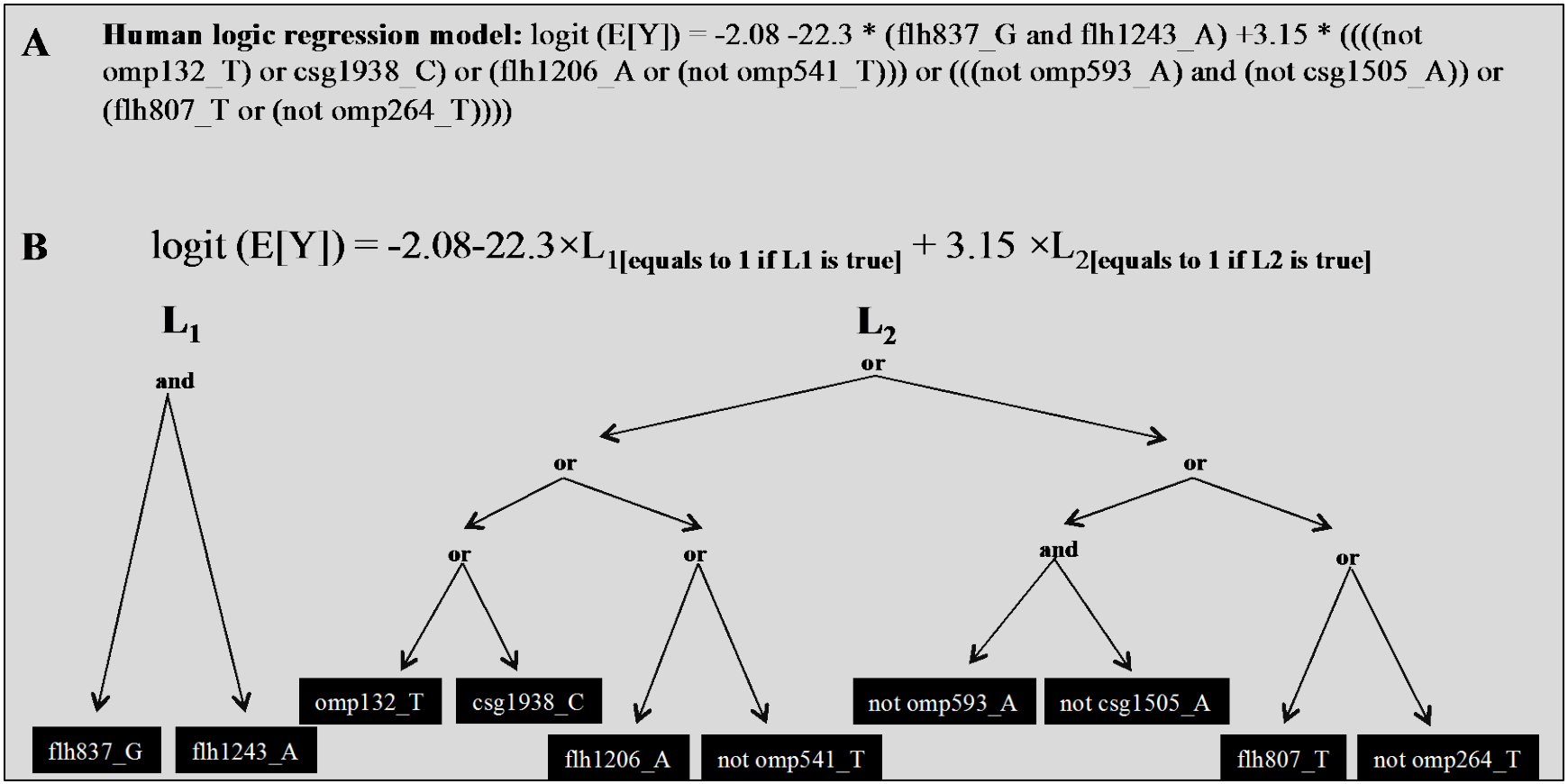

Figure 2-1. Human logic regression model manifested in a logic tree format. The original human logic regression model from Table 3 is provided in Panel A. In this model E[Y] is the probability of an *E. coli* isolate originating from a human source. The model can be broken down into two trees: L_1 and L_2 (Panel B). As an interpretive example, if the nucleotide sequence at position 837 is a "G" in the *uspC-flhDC* intergenic region of the concatenated sequence (labeled flh837_G) and is an "A" at position 1243 in this same region (labeled flh1243_A), then L_1 equals 1. If either of these condition are not met than L1 in the model is equal to 0. The position is coded in accordance with the nucleotide position in the multiple sequence alignment of the concatenated three intergenic sequences ((*asnS-ompF , uspC-flhDC, and csgBAC-csgDEFG*).

origin (as represented by 15 different colors [hosts] in the inner circle). Overall, a mosaic clustering of different host source types was observed, with *E. coli* from an individual animal host group scattered in different sections of the phylogenetic tree. For some hosts, such as muskrat and horse, many strains clustered close to each other; however strains from other animal hosts were found inserted in these clusters. In addition, some of the muskrat and horse *E. coli* strains were located further away from their main cluster. *E. coli* of human origin were scattered throughout the phylogenetic tree, with varying degrees of clustering occurring (i.e., examples of discreet clusters are provided in Figure 2-2).

Human strains were used as an example to compare the performance of logic regression and maximum likelihood phylogeny for DNA sequence analysis (Figure 2-2). The colors (red and black) represented in the outer circle of the phylogenetic tree represent all *E. coli* strains isolated from human feces. The red colored strains are considered human-specific (i.e., specialists) based on the logic regression model, while the black colored strains are not human specific (i.e., generalists). By classical ML phylogeny, human strains were scattered throughout the tree, with several small human clusters occurring but with each of these segregated from each other by various animal *E. coli* strains. To illustrate this point, two human clusters (Cluster 1 and Cluster 2 in Figure 2-2), were identified by ML phylogenetic analysis, and were comprised of 20 and 12 *E. coli* isolates each, respectively. These two clusters were segregated from each other on the phylogenetic tree by *E. coli* isolates originating from dogs, pigs, muskrats, moose, beavers, cats and coyotes. By comparison, logic regression classified 31 of the 32 *E. coli* isolates in these two clusters as human-specific (i.e., note that only one isolate [a black-labelled strain in the outer circle of the larger cluster]) was not considered human-specific by logic regression analysis). Interestingly, human isolates segregating even further away by ML phylogenetic analysis (i.e.,

Cluster 3) were also shown to carry the human-specific biomarker by logic regression (i.e., as represented by red color designations in the outer circle in this cluster). The data and associated statistical significance of models (fivefold cross validation and permutation test) indicate that the logic regression method may be a more powerful tool than phylogenetic analysis for identifying host-specific patterns of DNA sequence data in *E. coli*.

Table 2-5. Number of SNPs in intergenic regions (*uspC-flhDC*, *csgBAC-csgDEFG*, and *asnS-ompF*) observed among different hosts

Host	Total Number of SNPs in the Intergenic regions	Number of Isolates	Total number of SNPs/number of *E. coli* isolates analyzed in the host group
Bovine	329	120	2.74
Cat	140	21	6.67
Dog	217	61	3.56
Deer	154	48	3.21
Goose	185	54	3.43
Human	242	105	2.3
Chicken	129	59	2.19
Moose	142	14	10.14
Muskrat	89	56	1.59
Horse	107	44	2.43
Pig	190	49	3.88
Coyote	157	44	3.57
Gull	108	18	6
Beaver	73	40	1.83
Sheep	63	47	1.34

2.5 Discussion

As with all living organisms, natural selection and adaptation are critically important in a microbe's ability to survive and reproduce in an environment. Like any environment, the gastrointestinal system of warm-blooded animals is a highly competitive (e.g., inter- and intraspecific competition), variable (i.e., diet, pH, physiological metabolites [bile salts], molecular receptor affinity between host and microbe, etc.) and hazardous (e.g., immunological responses, microbial predation) environment. Given that such diverse gut physiologies exist at the molecular level in warm-blooded vertebrates, it is argued that these selective pressures may drive microbial evolution in the gut towards host adaption. In some cases, the pressures may lead to host specialization, whereas in other cases, and depending on microbial traffic between two host species (i.e., humans and their pets for example), microbes may adapt their survival to more than one host (i.e., host generalist). It is hypothesized that evolution towards host- specialization for gut survival may be driven largely at the regulatory transcriptome level since microbes must first 'sense' the gut microenvironment conditions and respond in a way that enhances their survival in this challenging and competitive microenvironment.

Herein, a supervised learning, logic-regression-based DNA analysis of SNPs in intergenic regions of *E. coli* is presented in order to study host specificity. The rationale for this method selection is its biological sensibility: the model uses specific forms of SNP-SNP interactions (i.e., SNP intersections and unions) which have biologically plausible interpretations (Dinu *et al.,* 2012). An SNP intersection requires that all of the SNPs in a specified group of SNPs must take their respective high-risk genotypes in order to increase the outcomes (i.e., disease) risk. This relationship is consistent with biological plausibility where sequential mutations must accumulate before a cell transforms into a cancerous phenotype in the multistage carcinogenesis.

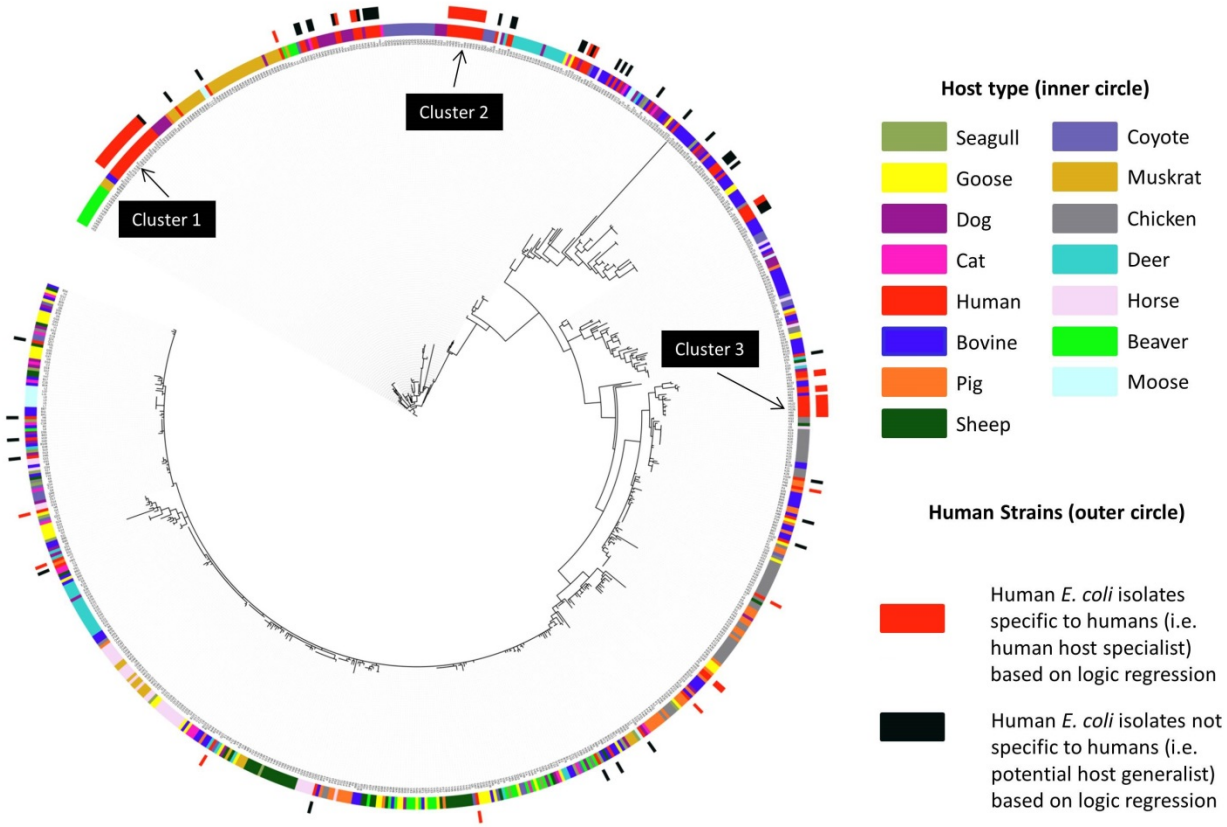

Figure 2-2. An unrooted maximum-likelihood (ML) phylogenetic tree of 780 *E. coli* isolates based on the concatenated intergenic sequences of *uspC-flhDC*, *csgBAC-csgDEFG*, and *asnS-ompF*. The interactive iTOL online tool was used for tree editing to overlay animal host grouping information onto the tree. The colors of the inner circle represent the 15 different animal hosts from which that representative *E. coli* was isolated. The outer circle (black and red colors only) represents all human isolates analyzed in this study. The red color in the outer circle represents strains of human origin that were grouped as human-specific strains (i.e., specialists) based on SNP biomarkers identified by logic regression, while the black colored strains, although isolated from humans, did not carry the human-specific biomarker (i.e., generalists). Two human clusters identified by ML phylogeny (Cluster 1 [containing 20 isolates] and Cluster 2 [containing 12 isolates]) are highlighted with Cluster 1 segregated from Cluster 2 by intervening animal isolates/clusters in the ML phylogenetic tree. Note that most isolates (31 of the 32) in these two clusters were shown to carry the human-specific biomarker identified by logic regression (i.e., as indicated by the red coloration of the strains in outer circle). Cluster 3 is another human phylogenetic clade that is distal to Clusters 1 and 2, but also segregated from these clusters by other animal isolates. Note that even within Cluster 3 all isolates are considered as human-specific based on biomarkers identified by logic regression. Moreover, a number of human isolates scattered throughout the ML phylogenetic tree, also carry the human-specific biomarker

theory. Likewise, accrued mutations within regulatory elements of microbes may allow for host-adaptation and colonization in novel gut microenvironments. SNP union, on the other hand, allows disease risk to be evaluated through multiple independent ways: this form captures genetic heterogeneity

The results indicate that the combination of logic regression and DNA sequence information from intergenic regions may provide a novel and robust approach to explore host-specific SNP biomarker patterns in *E. coli* populations. Conversely, unsupervised, cluster-based statistical methods that are commonly used in microbial systematics (White *et al.,* 2011; Furukawa *et al.,* 2011; Carson *et al.,* 2001; Ruecker *et al.,* 2012) may be insufficient and inappropriate to address this question. In fact, it is not surprising that *E. coli* isolates obtained from the same animal host do not always cluster together, the reason of which may be caused by the confounding presence of both host-generalists and host-specialists within a single animal host, and for which cluster-based algorithms are insufficient in resolving these populations.

In the original logic regression model building exercise, robust SNP biomarker patterns were observed in *E. coli* isolates collected from all 15 animal hosts. In the case of deer, 82% of all *E. coli* isolates collected from deer carried a SNP biomarker pattern highly specific for deer (98% specificity). Arguably, the ~18% of *E. coli* isolates that did not display the SNP-host-specific biomarker pattern may encompass the *E. coli* host-generalist population in deer (i.e., capable of transmission between deer and other animal hosts). Alternatively, this small population may, in fact, be truly host-specific to deer, but for which the intergenic sequences used in our analysis may be insufficient in resolving this association. Examples of where SNPs have been shown to alter the host-specific nature of bacteria include studies in *Vibrio fischeri* (Mandel *et al.,* 2009) and in certain *Salmonella* serovars (Tracz *et al.,* 2006; Eswarappa *et al.,*

2008). Since evolutionary trajectories towards host specificity are not linear, genetic analysis using supervised learning methods on a limited number of intergenic sequences (i.e., 3 intergenic regions in this study) should not be expected to contain all possible mutations encoding host-specificity in *E. coli*. In support of this idea, analysis of the isolates described by White *et al.* (White *et al.,* 2011) at 3 different intergenic sequences (*cutC-torY, metQ-rcsF, araH-otsB*) using logic regression (data shown in Chapter Three), resulted in poorer sensitivities and specificities in SNP biomarker patterns than the intergenic sequences used in this present chapter, suggesting that different intergenic regions contain differential information about host-specificity. Nevertheless, it is quite remarkable that such robust host-specificity was observed in this study across such a limited number of loci (3 intergenic regions only) and suggests that whole genome analysis approaches, coupled with supervised learning methods (i.e., logic regression), may shed even more light on the host-specific nature of *E. coli*. It is possible that the extreme genetic heterogeneity observed across the entire genome of this bacterial species [i.e., 10% similarity across the pangenome (Lukjancenko *et al.,* 2010)] may be a reflection of the wide host range for which this bacterial species can colonize, and encompass the large array of evolutionary trajectories for which this bacterial species may evolve towards host-specificity. The challenge rests with the vast computational resources needed to carry out logic regression-based biomarker analysis across entire genomes (i.e., the requirement for cluster-computers) and the need to perform whole genome sequencing across extensive host libraries of *E. coli* isolates. The logic regression model used in this study had a fixed tree size and the maximum size of the model was limited to two trees and 10 leaves (i.e. only 10 SNPs can be included in a model) in order to reduce the computational burden. However, in reality, it is possible that more than 10 SNPs (i.e.,

leaves) and more complex tree structures may be involved in the determination of host specificity in *E. coli*.

To avoid the common issue of over-fitting in the supervised learning method, both fivefold cross-validation and permutation tests were performed on our dataset, with the results generally supporting the concept of potential host-specificity of *E. coli* in most of the animal hosts surveyed. For three animals (goose, gull, and beaver), the permutation test demonstrated that the genetic patterns identified by logic regression were not significant. It is interesting to note that geese, gulls and beaver share a common aquatic niche, and we speculate that common shared environments may allow for sufficient microbial traffic/exposure between hosts to facilitate the expansion of *E. coli* generalist populations (e.g., enhanced transmission among these animal host populations due to deposition of feces in water and greater exposure in this common environment). Alternatively, and as mentioned above, the genetic regions we used to evaluate host specificity may not be relevant drivers of host adaptation in these animals. In the case of *E. coli* isolated from cats, the permutation test demonstrated significance of the logic model, but fivefold cross-validation yielded reduced sensitivities for that particular biomarker. *E. coli* from cats had the greatest degree of sequence polymorphism within the intergenic sequences analyzed, suggesting that greater representation of cat isolates in the library may improve the overall sensitivity of this statistically significant cat-biomarker in *E. coli*. An alternative hypothesis may be that *E. coli* populations in cats may be comprised of more host-generalist *E. coli* or that these intergenic regions may not be informative of host adaptation in cats.

When compared to traditional bioinformatic approaches (i.e., maximum likelihood phylogeny), logic regression analysis of DNA sequence revealed more robust associations with *E. coli* host specificity than did ML phylogeny. For example, using ML-phylogeny, *E. coli* isolates

obtained from humans segregated in dispersed clusters throughout the phylogenetic tree, segregated from each other by an array of animal isolates. In contrast, human biomarkers identified by logic regression analysis demonstrated that 53% of all human isolates appeared to be human specific (specialists), and that these human specific isolates partitioned themselves throughout all human clusters/isolates in a ML phylogenetic tree. Consequently, logic regression is able to use host label information in the analysis for grouping, outperforming the traditional phylogenetic approach and representing a novel and superior approach for characterizing host specificity/adaptation in microbes.

Although the logic regression results suggest that *E. coli* may display a certain level of host-specificity, the results of this study also cannot rule out the existence of host-generality in *E. coli* since the identified host-specific SNP patterns were not identified in all strains of that host group. Consequently, it is possible that host-generality exists in *E. coli* if sufficient microbial traffic between two animal hosts may drive *E. coli* to adapt and survive within multiple hosts. Examples of this could include predator-prey relationships (e.g., coyote and rodents), shared environments between different animal hosts (e.g., aquatic birds such as seagulls and geese), or close association between two host species (e.g., humans and pets, humans and domestic farm animals).

Although the data suggests that a certain degree of host-specificity may exist in *E. coli*, many different factors may affect the genetic population structure of *E. coli* in a single animal host or across various host individuals within an animal group. Temporal variations in *E. coli* population genetics have been studied in some host species (Caugant *et al.,* 1981; Jenkins *et al.,* 2003; Gordon, 1997). In one study (Caugant *et al.,* 1981), 550 *E. coli* were isolated from one human host over a period of 11 months, and based on multilocus enzyme electrophoresis (MLEE)

results, two electrophoresis types (ET) were observed frequently over extended periods while most ETs appeared transiently. This diversity of the *E. coli* populations seems to suggest a generalist lifestyle for *E. coli*; however one cannot exclude the fact that host-specific genetic signatures may be universally present in isolates but that techniques that use unsupervised learning methods for grouping such as MLEE, ribotyping, randomly amplified polymorphic methods, etc., may be unable to resolve this host-specificity. Host-specificity may reside at the SNP level and/or are determined by a combination of several SNPs across intergenic or gene loci. Nevertheless, factors that may affect the stability of biomarker SNP patterns may include diet, age, geography, health status and host-genetics, and that these biomarker patterns may change spatially and temporally. For example, all human isolates in this study were collected from stool samples from clinically ill patients (i.e., bacterial, viral or parasitic gastroenteritis). It is uncertain as to whether healthy human patients carry *E. coli* populations having similar host-specific biomarkers as those from clinically-ill patients. As another example, although *E. coli* strains isolated from cattle were collected from different geographical locations in Canada, there is some uncertainty in context of: a) whether isolates from animals on the same environment (i.e., a farm) carry a clonal lineage, and b) whether environments (i.e., farms) segregated geographically but in the same general area (i.e., southern Ontario) might have common circulating strains, whereas those isolated from animals across large geographical distances carry different host-specific biomarkers (i.e., cattle from Ontario compared to Alberta, Canada). The rationale for choosing only one *E. coli* isolate per individual of an animal host in the present study was to, a) reduce the chances of clonality and its effects in biomarker analysis, and b) to try and identify a stable biomarker that was present across multiple individuals in an animal host group.

It is also important to note that the reported sensitivity of the host-specific biomarkers in this study do not necessarily imply the overall prevalence of that *E. coli* biomarker in that particular animal host species (i.e., what proportion of the dogs actually carry the dog-specific *E. coli* biomarker?). To answer this question, more isolates from a single animal need to be analyzed. For example, in samples where no host-specific biomarker was observed from the single *E. coli* isolate collected from an animal, would the same result be got if a thousand *E. coli* isolates from that same animal was analyzed? Consequently, it is interesting to know how stable these markers are in a single animal host over time, and what proportion of *E. coli* in a single host may carry these biomarkers. The limitation of selecting one *E. coli* isolate per animal host rests with the fact that several separately evolved host-specific lineages may exist in the *E. coli* gut population of a single animal. Conversely, even though sensitivity may be low for some animal groups, replicate isolates collected from a single animal host may demonstrate the existence of *E. coli* isolates that carry the unique biomarkers, and consequently sensitivity of biomarker would be expected to increase (i.e., increased prevalence would result in increased sensitivity).

The approaches used in this study also have important implications for understanding emerging *E. coli* zoonotic disease in humans. For example, 53% of *E. coli* isolates obtained from human subjects in our study carried a human-specific SNP biomarker pattern. Consequently, pathogenic isolates within this group may be considered host-specific and consequently to cause disease in humans only. Of the 47% of isolates that did not carry the human-specific SNP biomarker pattern, these isolates may possibly represent host-generalists (as discussed above), and consequently comprise the potentially zoonotic population of *E. coli*. Identification of SNP biomarker patterns specific to humans in outbreak isolates would suggest

that human-to-human transmission was the most probable cause of disease outbreak. Interestingly, the use of supervised learning methods such as logic regression may also allow for the discovery of novel zoonotic SNP biomarkers. For example, group labels such as 'human-bovine' can be incorporated in the logic regression model building exercise to reflect populations that may colonize both humans and bovines. To do so requires that *E. coli* isolates that partition into the host-specific groups be removed from the dataset and the remaining isolates (i.e., generalists) be re-analyzed for human-bovine SNP patterns. Human-bovine SNP biomarker patterns with high specificity may be useful for determining zoonotic *E. coli* associated with bovine contamination sources (or other animal sources of human disease transmission).

2.6 Conclusion

A supervised-learning method of logic regression was applied for analysis of multi-spacer sequencing (MSS) data from *E. coli* isolates isolated from various animal hosts to study the concept of host specificity in *E. coli*. The results demonstrate host related genetic SNP biomarker patterns in *E. coli*. High sensitivity and specificity were achieved in most of the host animal groups verified by fivefold validation and permutation testing.

Host-specificity of *E. coli* is an important scientific question for studying microbial evolution. Understanding the interactions between microbes and their environment and identifying the driving forces for microbial evolution will advance our knowledge in several areas including the tracking of sources of fecal pollution in the environment and characterizing animal reservoirs of emerging enteric disease. In future studies, the robustness of the findings needs to be further assessed across larger libraries of *E. coli* isolates from an expanded range of

animal hosts, and incorporating the statistical approaches to supervised learning methods to whole genome analysis of *E. coli*.

Chapter Three : AN EVALUATION OF LOGIC REGRESSION-BASED BIOMARKER DISCOVERY ACROSS MULTIPLE INTERGENIC REGIONS FOR PREDICTING HOST SPECIFICITY IN *E. COLI*[2]

3.1 Abstract

Several studies have demonstrated that *E. coli* appears to display some level of host adaptation and specificity. Evolution towards host-specificity is not likely to follow a single trajectory arising from a single set of defined mutation, rather the evolutionary trajectory is likely to be multidirectional, driven by diverse combinations and permutations of various mutations across the transcriptome. As such, in order to determine the degree of host-specific information encoded in various ITGRs across a library of animal *E. coli* isolates, both whole genome analysis and a targeted ITGR sequencing approach were used. The results demonstrated that ITGRs across the genome encode various degrees of host-specific information. Incorporating multiple ITGRs (i.e., concatenation) into logic regression model building resulted in greater host-specificity and sensitivity outcomes in biomarkers, but the overall level of polymorphism in an ITGR did not correlate with the degree of host-specificity encoded in the ITGR. This suggests that distinct SNPs in ITGRs may be more important in defining host-specificity than overall sequence variation, explaining why traditional unsupervised learning phylogenetic approaches may be less informative in terms of revealing host-specific information encoded in DNA sequence. *In silico* analysis of 80 candidate ITGRs from publically available *E. coli* genomes was performed as a tool for discovering highly host-specific ITGRs. In one ITGR (*ydeR-yedS*) a SNP biomarker that was 98% specific for cattle was identified and for which 92% of all *E. coli*

[2] A version of this chapter has been published, the citation of which is: Zhi, S., Q. Li, Y. Yasui, G. Banting, T. A. Edge, E. Topp. T. A. McAllister and **N. F. Neumann**. 2016. An evaluation of logic regression-based biomarker discovery across multiple intergenic regions for predicting host-specificity in *Escherichia coli*. *Molecular Phylogenetics and Evolution*.

isolates originating from cattle carried this unique biomarker. In the case of humans, a host-specific biomarker (98% specificity) was identified in the concatenated ITGR sequences of *rcsD-ompC*, *ydeR-yedS*, and *rclR-ykgE*, and for which 78% of *E. coli* originating from humans carried this biomarker. Interestingly, human-specific biomarkers were dominant in ITGRs regulating antibiotic resistance, whereas in cattle host-specific biomarkers were found in ITGRs involved in stress regulation. The data suggests that evolution towards host specificity may be driven by different natural selection pressures on the regulome of *E. coli* among different animal hosts.

3.2 Introduction

In the previous chapter of this thesis, logic regression analysis of DNA sequences in intergenic regions (ITGRs) from human, animal and environmental *E.coli* was used as a novel bioinformatics approach to identify host-specific single nucleotide polymorphic (SNP) biomarkers in *E. coli*. It was found that *E. coli* populations in humans and animals consists of a mixture of host-specialists and host-generalists, and highly specific SNP biomarkers among host-specialist populations were identified across 15 different animal hosts. For example, 53% percent of all human isolates displayed a unique intergenic biomarker pattern that was 98% specific to humans, suggesting that a significant proportion of the populations of *E. coli* in humans may be host-specialists.

In order to survive under the diverse microenvironmental conditions associated with various animal host gut physiologies or non-host environments (i.e. nutrient availability, pH, temperature, predation, UV, high oxygen), bacteria have evolved a number of adaptive coping strategies. The acquisition of specific genes [e.g., antibiotic resistance genes (Giedraitiene *et al.,* 2011) or virulence genes (Mellata *et al.,* 2010)] is one strategy whereby bacteria can achieve

environmental fitness, but bacteria are also able to phenotypically adapt to adverse environments through the regulation of core genes (Prüss *et al.,* 2006; Ziebuhr *et al.,* 1999). Gene regulation can be altered through inactivation and/or activation of gene regulators (Baez and Shiloach, 2013) as well as through mutations in promoter sequences (Ando *et al.,* 2011). For example, mutations in the promoter regions of the *katG* gene can alter expression resulting in phenotypic changes affecting susceptibility of *Mycobacterium tuberculosis* to isoniazid (Ando *et al.,* 2011). By acquiring an aerobically expressed promoter for the expression of a previously silent citrate transporter, *E. coli* acquired the capacity to use citrate as an energy source under aerobic conditions (Blount *et al.,* 2012). In *Bordetella pertussis* a new allele in the promoter of the pertussis toxin gene caused a dramatic increase in pertussis in humans in the Netherlands (Mooi *et al.,* 2009).

It is hypothesized (Chapter One) that the intense intra-and inter-species microbial competition within the gut, coupled with the diversity in the cellular and molecular physiology of gastrointestinal microenvironments across animal species, will drive the evolution of the regulatory transcriptome of *E. coli* towards host-adaptation and specificity. Although the research described in the previous chapter supports this hypothesis (Zhi *et al.,* 2015), the original findings were limited to identifying host-specific SNP patterns in only three intergenic regions across a library of 780 *E. coli* strains isolated from 15 different animals. Evolution towards host-specificity is not likely to follow a single trajectory arising from a single set of defined mutations, rather the evolutionary trajectory is likely to be multidirectional, driven by diverse combinations and permutations of various mutations across the transcriptome. As such, this Chapter is aimed at determining the degree of host-specific information encoded among various ITGRs using two independent approaches to biomarker discovery: a) a targeted ITGR sequence-based approach,

examining 6 different ITGRs across a diverse library of *E. coli* isolates obtained from different animal hosts; and b) publically-available whole genome data of 160 *E. coli* isolates isolated from different animals. The advantage of using a targeted ITGR sequencing approach was that a large number of isolates from diverse *E. coli* strains collected from an animal-host library could be readily examined. The disadvantage of this approach was that relatively few ITGRs could be examined at once, and for which little *a priori* knowledge on host-specificity existed. Conversely, the advantage of using whole genome data was that a large number of ITGRs could be examined simultaneously (i.e., n=80 in this study), but with the disadvantage that relatively few *E. coli* whole genomes are available in NCBI from hosts other than humans.

3.3 Material and methods

3.3.1 *E. coli* isolates

In total, 356 *E. coli* strains isolated from eight different host sources (Table 3-2) were selected from the previously established *E. coli* library containing 845 isolates in Chapter Two (Zhi *et al.,* 2015). All 356 isolates that were PCR positive for the three ITGRs (*csgBAC-csgDEFG, uspC-flhDC, asnS-ompF*), as well as for three additional ITGRs (*cutC-torYZ, metQ-rcsF, araH-otsB*), were chosen for logic-regression-based biomarker analysis. The genes regulated by these ITGRs are as follows: 1) the *csgBAC-csgDEFG* region, regulating synthesis of curli fimbriae; 2) the *uspC-flhDC* region, regulating the expression of the master regulator of flagellum biosynthesis and universal stress response gene C; 3) the *asnS-ompF* region, regulating the expression of a regulator for an outer membrane protein related to low-osmolarity conditions; 4) the *cutC-torYZ* region, regulating expression of a trimethylamine N-oxide (TMAO) reductase complex capable of reducing nitrogenous and sulphurous-oxide compounds; 5) the *metQ-rcsF*

region, regulating the expression of an outer membrane lipoprotein that can stimulate colanic acid production; and 6) the *araH-otsB* region, regulating the biosynthesis of trehalose-6-phosphate phosphatase as it relates to osmotic stress.

All *E. coli* isolates were grown in tryptic soy broth (TSB) and genomic DNA extracted from *E. coli* TSB cultures using DNeasy Blood & Tissue kits (QIAGEN) according to the manufacturer's instructions.

3.3.2 PCR and DNA sequence analysis of targeted ITGRs

The targeted ITGRs were PCR amplified separately using the primers listed in Table 3-1. The PCR conditions for *csgBAC-csgDEFG, uspC-flhDC, asnS-ompF,* and *cutC-torYZ* regions were as follows: initial denaturation at 95^{o}C for 4 min, 33 cycles of 95 ^{o}C for 30 s, 58 ^{o}C for 30 s, and 72 ^{o}C for 1 min, followed by a 7 min extension at 72 ^{o}C. The PCR conditions for *metQ-rcsF* and *araH-otsB* were as follows: initial denaturation at 95^{o}C for 4 min, 35 cycles of 95 ^{o}C for 30 s, annealing at 56 ^{o}C for 20 s, and 72 ^{o}C for 15 s, followed by a 7 min extension at 72 ^{o}C. The total volume of each PCR reaction was 25 µL and contained 5 µL of template, 1X Maxima Hotstart Mastermix (ThermoSceintific), and each primer at a concentration of 500 nM.

All PCR products were sequenced bi-directionally using Sanger sequencing by either Macrogen Incorporated (South Korea) or at the University of Calgary Genetic Analysis Laboratory in Calgary, Alberta Canada (http://www.ucalgary.ca/dnalab/sequencing). SeqMan Pro (DNASTAR) was used to construct the bidirectional sequences and the sequences were aligned using the ClustalX 2.0 (Larkin *et al.,* 2007) program. The multiple sequence alignment of each of the six ITGRs was manually edited to trim the 5'and 3' regions that contained missing data.

Table 3-1. PCR primers used in this study.

Target	Primer	Primer sequence (5'-3')	Reference
araH-otsB	otsB-R	AGGATGCGGTTTGATTTCCG	(Zaslaver *et al.,* 2006)
	otsB-F	GCGAAACGCACTGTCTGA	
cutC-torYZ	torY-F	AGACTTTTTGCGCCAGCAAT	(Zaslaver *et al.,* 2006)
	torY-R	TGAACACGCGGACGAGTA	
metQ-rcsF	rcsF-F	ACGCCGAGAACGTGAAGA	(Zaslaver *et al.,* 2006)
	rcsF-R	GGGTTCGACAGGGGATCT	
asnS-ompF	ompF-F	TACGTGATGTGATTCCGTTC	(Zaslaver *et al.,* 2006)
	ompF-R	TGTTATAGATTTCTGCAGCG	
csgBAC-csgDEFG	csgD-1	GGACTTCATTAAACATGATG	(Zaslaver *et al.,* 2006)
	csgD-2	TGTTTTTCATGCTGTCAC	
uspC-flhDC	flhDC-F	GAGGTATGCATTATTCCCACCC	(White *et al.,* 2011)
	flhDC-R	TGGAGAAACGACGCAATC	

3.3.3 Logic regression analysis

Logic regression analysis was performed on the DNA sequence data from all six ITGRs in order to identify SNP biomarkers of host-specificity as described in Chapter Two (Zhi *et al.,* 2015). Briefly, logic regression analysis searches for logic regression models consisting of SNPs that can best predict the host source. The logic regression model was as follows:

$$\text{logit}\,(E[Y]) = \beta_0 + \beta_1 L_1 + \beta_2 L_2 + \ldots + \beta_p L_p$$

where,

- Y is a binary variable, an indicator for being from one *E. coli* host versus being from the other hosts,
- β_0, β_1,… β_p are the parameters indicating the degrees of association between the SNP patterns (L's) and the host indicator (Y), and

- $L_1, L_2, \ldots, L_p$ are SNP patterns that are Boolean combinations (called "trees") of indicators of SNP genotypes (called "leaves") in the *E. coli* genes.

In this analysis, sensitivity and specificity were used as measurements for the goodness of fit for the identified model. Similar to the definitions used in Chapter Two (Zhi *et al.,* 2015), sensitivity was defined as the proportion of samples from a target animal host or environmental sample (i.e., wastewater), that carried a specific SNP pattern. Specificity was defined as the proportion of samples from hosts/environments other than the target host/environment (i.e., all other human and animal hosts) that did not carry the SNP biomarker of interest. To compare the host source prediction ability between different intergenic regions, we used the sum of the sensitivity and specificity values as a simple index-based measure of *host predictive power* (HPP).

3.3.4 Host-Specific biomarker analysis from *E. coli* genomes

3.3.4.1 *E. coli* genome database

A hundred and sixty *E. coli* genomes obtained from NCBI genome database were used in this study (Appendix Table 1). These genomes were selected based on their host sources of isolation. Most of the *E. coli* genomes deposited in NCBI with a known host source origin were from humans. At the time this analysis was done only 35 *E. coli* genomes were from *E. coli* originating from bovine sources and 29 from other animal hosts (pig, chicken, turkey, horse, mouse, dog, deer, rabbit, and goat). All *E. coli* genomes from bovine and other animal sources were used in this study and 96 human *E. coli* genomes selected from the NCBI database to be representative of human strains. The genomes were downloaded and a local genome database was generated using BLAST-2.2.29 (Camacho *et al.,* 2009).

3.3.4.2 *In silico* logic regression analysis of ITGR in *E. coli* genomes

Ninety ITGRs were selected based on several key publications (Bauchart *et al.,* 2010; Chen *et al.,* 2006; Miskinyte *et al.,* 2013) and examined for host-specific biomarkers using logic regression analysis across the various *E. coli* genomes (Appendix Table 2). The ITGRs chosen for analysis were based on their potential role in regulating genes associated with host adaptation / pathogenesis and specificity. Due to the limited number of representative *E. coli* genomes from hosts other than bovines and humans, our *E. coli* genomes were grouped into three sources: human (Group H), bovine (Group B), and all other animals (Group D). As Group D consisted of *E. coli* isolated from various other animals it was used as a control group during logic regression analysis in the search for host-specialist biomarkers in humans and cattle. The ITGRs of these selected genes were derived from the genome of *E. coli* K12 MG1655 and blasted against our local genome database to identify homologies with the other *E. coli* genomes. Only sequences that had 100% coverage over the query sequence were kept for subsequent analysis. In addition, ITGRs represented by less than 100 genomes in our customized database were excluded from the analysis. In the end, 80 ITGRs out of the previously selected 90 passed our criteria and were used for logic regression analysis. For each selected ITGR, the extracted DNA sequences were aligned using the multiple sequence alignment tool Clustal X (Larkin *et al.,* 2007). The multiple sequence alignment for each gene was then used for further logic regression analysis.

3.3.4.3 Traditional phylogenetic analysis

During our genome-wide, supervised learning logic regression-based biomarker analysis, we discovered a number of potentially highly host-specific ITGRs. As such, maximum likelihood phylogenetic analysis was used to evaluate whether traditional unsupervised learning-

based bioinformatic approaches could reveal patterns of host-specificity in these loci. Maximum likelihood (ML) phylogenetic analysis was performed on the DNA sequences of the intergenic region *ydeR-yedS* as well as concatenated DNA sequences of intergenic regions *ydeR-yedS*, *rcsD-ompC*, and *rclR-ykgE* by RAxML Blackbox (Stamatakis *et al.,* 2008) using its default settings. The phylogenetic tree was edited to map against different animal host groupings using the Interactive Tree Of Life (iTOL) online tool (Letunic and Bork, 2007; Letunic and Bork, 2011).

3.4 Results

3.4.1 PCR amplification of selected intergenic regions

The proportion of *E. coli* isolates positive for PCR amplification of intergenic regions *cutC-torYZ, metQ-rcsF, araH-otsB* is shown in Table 3-2. Among the 356 *E. coli* strains that

Table 3-2. *E. coli* isolates collected from different host animals and PCR results for targeted intergenic regions.*

Host	No. of isolates	No. of PCR positive isolates (%)			No. of isolates used for logic regression model building
		cutC-torYZ	*metQ-rcsF*	*araH-otsB*	
Bovine	85	79 (92.9)	83 (97.6)	81 (95.3)	73
Cat	21	19 (90.1)	20 (95.2)	18 (85.7)	17
Dog	61	61 (100)	60 (98.4)	56 (91.8)	55
Goose	20	19 (95)	19 (95)	18 (90)	17
Human	105	101 (96.2)	105 (100)	104 (99.0)	100
Chicken	2	2 (100)	2 (100)	2 (100)	2
Pig	42	40 (95.2)	42 (100)	41 (97.6)	39
Gull	20	18 (90)	20 (100)	17 (85)	15
Total	356	339 (95.2)	351 (98.6)	337 (94.7)	318

* All isolates represented in the second column of this table were also PCR positive for *csgBAC-csgDEFG, uspC-flhDC, asnS-ompF* as determined in Chapter Two.

were PCR positive for *csgDEFG*, *uspC-flhDC*, and *asnS-ompF* intergenic loci, 95.2%, 98.6% and 94.7% were also PCR positive for *cutC-torYZ, metQ-rcsF*, and *araH-otsB*, respectively. In total, 318 *E. coli* isolates were PCR positive for all six intergenic regions. As shown in Table 3-3, 123 single nucleotide polymorphic sites (SNPs) were found in the 502 bp *cutC-torYZ* region; 12 in the 240 bp *metQ-rcsF* region; 31 in the 239 bp *araH-otsB* region; 206 in the 699 bp *uspC-flhDC* region; 110 in the 723 bp *csgBAC-csgDEFG* region and 132 in the 611 bp *asnS-ompF* region. The proportion of variable SNPs ranged from 5.0% for the *metQ-rcsF* region to 29.47% for the *uspC-flhDC* region.

Table 3-3. Number of SNPs in six intergenic regions observed among different hosts.

Intergenic regions	Total number of SNPs	Sequence length (including gaps)	Total number of SNPs/Number of sequence length
cutC-torYZ	123	502	24.50%
metQ-rcsF	12	240	5.00%
araH-otsB	31	239	12.97%
uspC-flhDC	206	699	29.47%
csgBAC-csgDEFG	110	723	15.21%
asnS-ompF	132	611	21.60%
Total	614	3014	20.37%

3.4.2 Identification of host-specific biomarker based in logic regression of six targeted ITGRs

Logic regression analysis was performed on six ITGRs (*cutC-torYZ*, *metQ-rcsF*, *araH-otsB, csgBAC-csgDEFG, uspC-flhDC,* and *asnS-ompF*) in the 318 *E. coli* isolates to assess whether these regions encode different degree of host specificity. Due to the limited number of *E. coli* isolates represented from cats, chickens, geese, and gulls in our library, only biomarker

results from human, bovine, pig, and dog were used to assess the ability of the six ITGRs to assess host specificity. However, strains from cat, chicken, goose, and gull were still included in logic regression analysis as negative, non-target animal controls.

Table 3-4 summarizes the individual logic regression results obtained from each of the ITGRs, as well as logic regression outcomes from concatenated sets of intergenic loci from the four different animal hosts (human, bovine, pig and dog). Individually, some ITGRs were more informative of host source than others. For example, in humans, the *cutC-torYZ* locus was the most host informative ITGR, with 41% of all human isolates possessing a biomarker that was 97% specific generating an HPP index of 1.37. Conversely, this locus was not particularly host-informative for dogs (HPP = 1.0). The most host informative locus of the six analyzed in cattle, pigs and dogs was the *uspC-flhDC* intergenic region. This locus demonstrated strong host-sensitivity and specificity in pigs where 74% of all *E. coli* isolated from pigs contained a biomarker that was 97% specific (HPP=1.71). The most 'uninformative' intergenic locus in pig was *metQ-rcsF* as it only generated an HPP of 1 (Table 3-4).

For all four animal hosts, concatenating the ITGRs for logic regression biomarker searching increased the robustness (defined as increased sensitivity and specificity) of prediction in host-specific models as compared to using only a single locus (Table 3-4). When the ITGR combination of *csgBAC-csgDEFG, uspC-flhDC, asnS-ompF* was compared to the three new ITGRs (*cutC-torYZ*, *metQ-rcsF*, *araH-otsB*), the sequences of *csgBAC-csgDEFG, uspC-flhDC,* and *asnS-ompF* had greater host predictive power across all animal groups (Table 3-4).

Table 3-4. Logic regression analysis of 318 *E. coli* isolates based on six intergenic regions.

Intergenic regions	Human			Bovine			Pig			Dog		
	Sens.[a]	Spec.[b]	PP[c]	Sens.	Spec.	PP	Sens.	Spec.	PP	Sens.	Spec.	PP
araH-otsB	0.39	0.97	1.36	0.15	0.98	1.13	0.46	0.96	1.42	0.24	0.98	1.22
metQ-rcsF	0.11	0.99	1.1	0	1	1	0	1	1	0	1	1
cutC-torYZ	0.41	0.96	1.37	0.3	0.98	1.28	0.1	1	1.1	0	1	1
asnS-ompF	0.33	1	1.33	0.27	1	1.27	0.28	1	1.28	0.36	0.97	1.33
uspC-flhDC	0.35	0.99	1.34	0.29	1	1.29	0.74	0.97	1.71	0.6	0.95	1.55
csgBAC-csgDEFG	0.3	1	1.3	0.22	0.99	1.21	0.28	1	1.28	0.4	0.99	1.39
araH-otsB, metQ-rcsF, cutC-torYZ	0.51	0.95	1.46	0.32	0.98	1.3	0.69	0.91	1.6	0.53	0.95	1.48
asnS-ompF, uspC-flhDC, csgBAC-csgDEFG	0.6	0.98	1.58	0.4	1	1.4	0.79	0.97	1.76	0.63	0.93	1.56

3.4.3 *In silico* biomarker searching using *E. coli* genome data

Based on the results described above, it is hypothesized that intergenic loci across the genome likely experience varying selective pressures in different animal hosts, and as such, certain ITGRs may possess differential dominance in driving host adaptation/specificity. It is important to evaluate whether these observations and associations held true across a larger repertoire of ITGRs in the genome, and subsequently whether certain ITGRs may be highly host-specific in some animals.

Table 3-5 provides a list of ITGRs with reasonably good host predictive performance (HPP index of > 1.4) based on logic regression analysis of our *E. coli* genome database. For the human group, only 6 of the 80 ITGRs achieved a HPP index of > 1.4, while in the bovine group 20 of the 80 ITGRs met this criteria (Table 3-5). In humans, the highest HPP index observed was 1.55 for the intergenic region *ydeR-ydeS*, for which 56% of all human *E. coli* isolates displayed a unique SNP biomarker pattern that was 99% specific (Table 3-5 and Table 3-6). A SNP biomarker identified in the *rcsD-ompC* intergenic region was the second most host-informative locus in human-derived *E. coli*, possessing 54% sensitivity and 99% specificity. The third most host-informative ITGR for human *E. coli* was the *aldB-yiaW*, followed by *paoA-yagU*, *ydeO-ydeN*, and *rclR-ykgE*.

As was observed in humans, the *ydeR-ydeS* locus was also the most host-informative for cattle, having an HPP index of 1.9 and with this biomarker being 98% specific (Table 3-5 and Table 3-6). In cattle, the second best performing host-specific ITGR was *rcsD-ompC*, providing a sensitivity of 49% and a specificity of 97%. In addition, the ITGRs *aldB-yiaW* and *rclR- ykgE* were also among the most informative host-specific biomarker regions in bovine *E. coli*. However, although *paoA-yagU* and *ydeO-ydeN* were found to be host-informative in *E. coli* from

Table 3-5. ITGRs for which good host predictive performance (i.e., ≥1.4 HPP) was observed by logic regression biomarker analysis.

	Name	Sensitivity	Specificity	Host Prediction power	No. of *E. coli* Genomes Used Across All Animal Groups for Model Building	Sequence length (bp)	No. of SNPs	No. of SNPs/Sequence length (bp)
Human (H)	*ydeR-yedS*	0.56	0.99	1.55	132	579	201	0.35
	rcsD-ompC	0.54	0.99	1.53	155	772	113	0.15
	aldB-yiaW	0.43	0.98	1.41	102	546	71	0.13
	paoA-yagU	0.41	0.99	1.4	102	392	67	0.17
	ydeO-ydeN	0.41	0.99	1.4	144	403	111	0.28
	rclR-ykgE	0.41	0.99	1.4	148	532	120	0.23
Bovine (B)	*ydeR-yedS*	0.92	0.98	1.9	132	579	201	0.35
	pgaA-ycdT	0.86	0.96	1.82	102	607	114	0.19
	ychE-oppA	0.79	0.98	1.77	155	761	115	0.15
	pagP-cspE	0.63	0.96	1.59	156	175	21	0.12
	flgB-flgA	0.54	0.99	1.53	158	155	19	0.12
	rclR-ykgE	0.54	0.99	1.53	148	532	120	0.23
	cysZ-cysK	0.56	0.96	1.52	156	182	15	0.08
	tdcG-yhaO	0.51	0.99	1.5	154	278	48	0.17
	nrdF-proV	0.53	0.95	1.48	157	357	43	0.12
	yhjC-yhjB	0.51	0.96	1.47	154	520	68	0.13
	rcsD-ompC	0.49	0.97	1.46	155	772	113	0.15
	ibpA-yidQ	0.51	0.95	1.46	156	306	16	0.05
	yjbM-yjbN	0.47	0.99	1.46	118	365	70	0.19
	ilvY-ilvC	0.46	0.99	1.45	158	142	12	0.08
	aldB-yiaW	0.47	0.96	1.43	102	546	71	0.13
	betT-betI	0.43	0.99	1.42	156	128	13	0.1
	yjjI-deoC	0.46	0.96	1.42	158	258	20	0.08
	kdgK- yhjH	0.46	0.96	1.42	158	229	23	0.1
	nrdA-yfaL	0.41	0.99	1.4	152	718	192	0.27

humans, they were not found to be host-informative for bovine *E. coli*. Except for the four shared informative ITGRs between human and bovine isolates, all the other reasonably good performing intergenic regions in bovine *E. coli* isolates were not very informative for *E. coli* isolates from humans.

As was observed in the results from targeted ITGR sequencing (Section 3.4.2), concatenation of whole genome-identified loci yielded even greater host predictive power than each of the individual loci. Consequently, three new ITGRs (*rcsD-ompC*, *ydeR-yedS*, and *rclR-ykgE*) were chosen as a gene set to evaluate their performance on host source prediction of *E.coli* from cattle and humans. These three ITGRs were selected based on: a) the fact that they were highly host-predictive of both bovine and human-derived *E. coli*, and b) all 123 *E. coli* genomes in our local genome database contained all three ITGRs. Logic regression analysis of these concatenated sequences resulted in the identification of a SNP biomarker that was carried by 78% of all human *E. coli* isolates and with a 98% specificity to human isolates (HPP = 1.76) [Table 3-6]. This HPP value was higher than the individual HPP values for *ydeR-yedS* (HPP = 1.55), *rcsD-ompC* (HPP = 1.53) and *rclR-ykgE* (HPP = 1.4) [Table 3-5]. Concatenation resulted in only a minor increase in HPP for cattle isolates (99% specificity with 92% sensitivity) [Table 3-6].

In order to evaluate how well the targeted ITGR sequencing approach (Section 3.4.2) related to the whole genome approach for biomarker discovery (Section 3.4.3), logic regression analysis of the six targeted ITGRs (*csgBAC-csgDEFG, uspC-flhDC, asnS-ompF*, *cutC-torYZ, metQ-rcsF,* and *araH-otsB*) were performed using publically available *E. coli* whole genome data. For human-derived *E. coli* the whole genome biomarker discovery approach yielded

Table 3-6. Logic regression models of host-specific SNP biomarkers found within the specified ITGRs and using publically available whole genome data.

Host source	Intergenic region	SNP pattern [a,b]
Bovine	*ydeR-yedS*	-3.96 +6.85 * ((yedS_457_G or (not yedS_443_T)) or (((not yedS_441_T) or (not yedS_292_C)) or (yedS_347_G or yedS_450_G))) +5.57 * ((((not yedS_207_C) or yedS_419_A) and yedS_347_C) or (not yedS_404_G))
Human	*ydeR-yedS*	0.226 +4.3 * (((yedS_459_A and (not yedS_443_C)) and yedS_410_T) and ((yedS_571_G and (not yedS_457_G)) and (yedS_291_C and (not yedS_373_G)))) -3.71 * (((not yedS_151_T) and (not yedS_69_G)) and (not yedS_420_G))
Bovine	*rcsD-ompC, ydeR-yedS,* and *rclR-ykgE*	-1.1 -23.2 * ((((not yedS_441_C) and (not ykgE_397_C)) and (not ykgE_120_G)) and (((not yedS_403_C) and ompC_620_A) and (yedS_434_G and yedS_457_A))) +4.23 * ((ykgE_190_A and (not ykgE_107_T)) and yedS_523_-)
Human	*rcsD-ompC, ydeR-yedS,* and *rclR-ykgE*	-2.95 -4.45 * ((((not ykgE_172_T) or (not yedS_48_T)) and (not yedS_385_G)) and (not ompC_181_A)) +7.01 * (((not ykgE_286_A) or (yedS_362_T and (not yedS_462_C))) and (ykgE_215_T and ((not ykgE_14_G) and ykgE_396_T)))

[a] Analysis was carried out in the intergenic regions of the *ydeR-yedS* (listed as yedS in models above), *rcsD-ompC* (listed as ompC in models above), and *rclR-ykgE* (listed as ykgE in models above) as described in Materials and Methods.

[b] In reading the model output a value designated by '*ydeR-yedS*_457_G' refers to the guanine base at nucleotide position 457 in the multisequence alignment of the *ydeR-yedS* intergenic loci. A value designated by 'yedS_523_-' refers to the gap at nucleotide position 523.

similar outcomes regarding host predictive power as the targeted ITGR discovery approach at all loci except the *cut*C-*tor*YZ locus (Figure 3-1). For cattle-derived *E. coli*, whole genome approaches tended to over-estimate host predictive power (Figure 3-1) compared to target ITGR biomarker approaches. The results suggest that although whole genome approaches forbiomarker discovery may be useful for identifying host-informative ITGRs. These outcomes may need to be validated by sequencing these ITGRs in large number of *E. coli* isolates that originate from diverse animal host sources.

Since logic regression was used to search for SNP biomarkers reflective of host origin of *E. coli*, it is important to know whether highly polymorphic ITGRs had better host predictive power than less polymorphic ITGRs. To do this the number of SNPs in each of the ITGRs that had reasonable host predictive power (i.e., $\geq$1.4, Table 3-5) were examined, and correlation analysis were performed against SNP polymorphism. There was little or no correlation between the number of SNPs in an ITGR and host predictive power (Figure 3-2). As a specific example, the ITGR locus *nrdA-yfaL* possessed 192 152 SNPs and ranked as one of the most polymorphic ITGRs in *E. coli* derived from cattle (Table 3-5), but the host predictive power for this ITGR locus was amongst the lowest examined. These data suggest that host-specificity is not necessarily linked to the degree of polymorphism in ITGRs.

Since the degree of polymorphism in ITGRs did not correlate with host predictive power, it is hypothesized that unsupervised learning-based bioinformatics approaches (i.e., clustering) may not be capable of detecting these host-specific patterns in DNA sequence. Therefore, maximum likelihood phylogenetic analysis was performed on whole genome DNA sequence data in some ITGRs to compare results with logic regression biomarker searching approach. Figure 3-3 illustrates two maximum likelihood phylogenetic trees, in which DNA sequences of

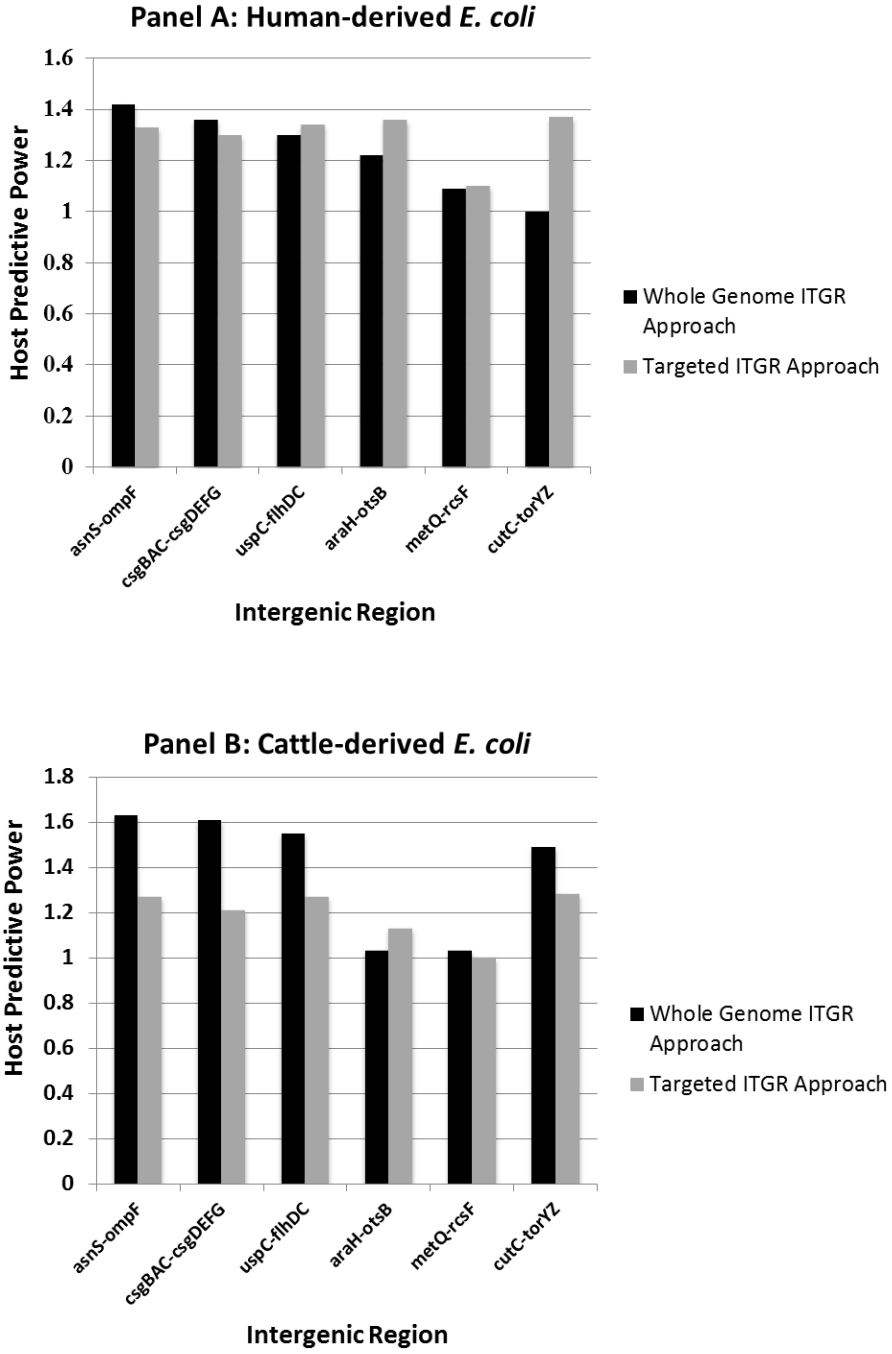

Figure 3-1. Comparison between a library-based ITGR sequencing approach and whole genome data approach for identifying logic regression–based biomarkers of host-specificity in humans (Panel A) and cattle (Panel B). The respective ITGRs used in the analysis (X-axis) were sequenced from a library of *E. coli* isolates collected from different animal hosts or extracted from whole genome sequence databases and analyzed by logic regression.

the intergenic region *ydeR-yedS* and concatenated ITGRs *rcsD-ompC* / *ydeR-yedS* / *rclR-ykgE* were carried out, respectively (Figure 3-3). Overall, phylogenetic analysis failed to differentiate (i.e., cluster) cattle *E. coli* according to host source at the *ydeR-yedS* locus (Figure 3-3), as only 13 of the 25 *E. coli* strains derived from cattle clustered with all other cattle strains interspersed

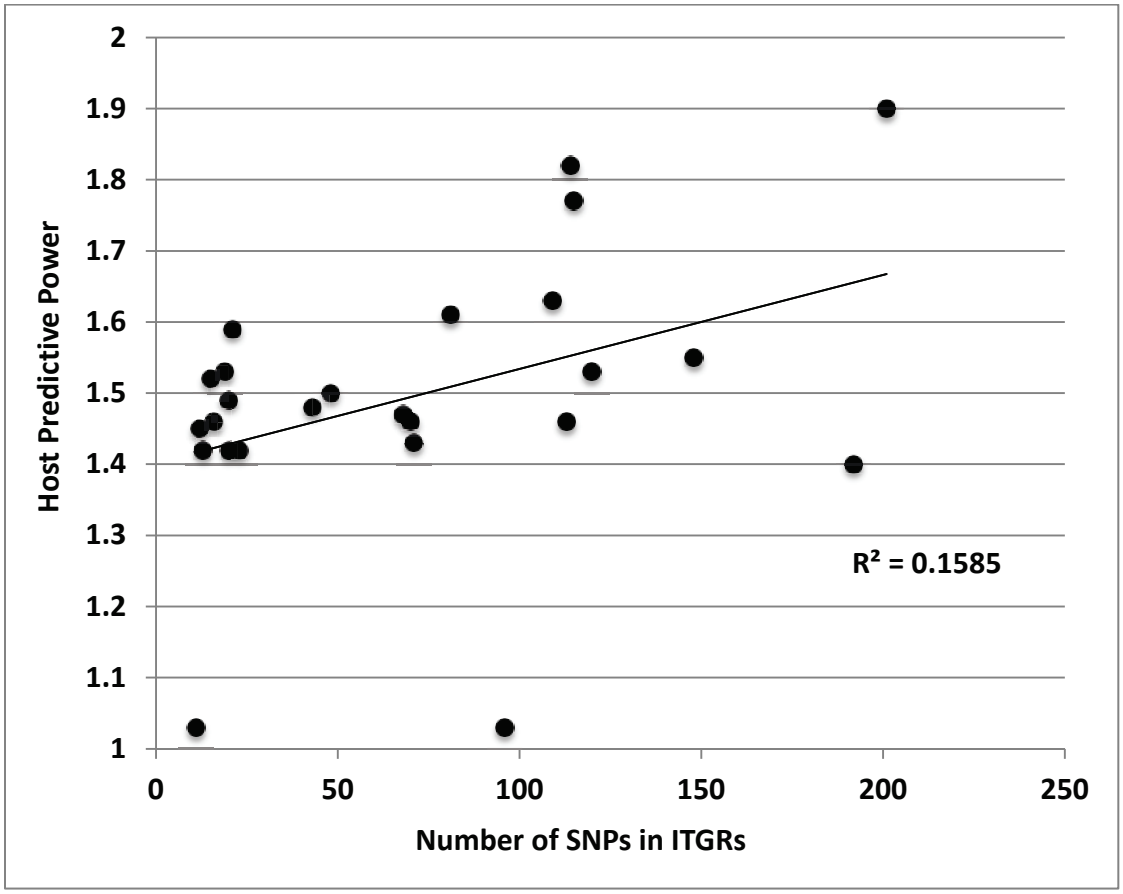

Figure 3-2. Correlation between the number of SNPs in an ITGR and the host-predictive power associated with that ITGR, as determined by logic regression. The data demonstrates a poor correlation between HPP and the number of SNPs in an ITGR, suggesting that the overall degree of polymorphism observed in an ITGR has little of no relation to host-specificity.

throughout the tree. Conversely, logic regression-based biomarker analysis revealed that at this locus 92% of strains carried a unique SNP biomarker that was 98% specific to cattle (Table 3-5). Likewise, human strains did not consistently partition out by phylogenetic analysis at this locus (Figure 3-3). Several phylogenetic clusters of human-derived *E. coli* were observed but these were interspersed among animal isolates. However, biomarker analysis revealed that many of the human isolates in these different phylogenetic clusters possess a common human-specific biomarker (Figure 3-3). Concatenating multiple ITGR sequences together (i.e., Tree B in Figure 3-3) did not result in greater clustering of host-specific strains by maximum likelihood phylogenetic analysis, but analysis by biomarkers revealed an even greater degree of host specificity after concatenation (Figure 3-3). This data supports the concept that traditional phylogenetic approaches may not be appropriate for evaluating host-specificity in bacteria.

3.5 Discussion

E. coli is a bacterial species with high genetic diversity. In a comparison of 61 *E. coli* genomes analyzed by Lukjancenko *et al.* (2010), the pangenome represented only 10% of the total genome. It is believed that the remaining 90% of accessary genes in the pangenome may reflect the evolutionary trajectory of *E. coli* to adapt and survive in various niches, including animal hosts and non-host environments. Intergenic regions located between functional genes contain promoter/repressor sequences that can regulate the expression of these genes, and are known to be important in niche adaptation, colonization, and exploitation of new environments by bacteria (Mooi *et al.,* 2009; Blount *et al.,* 2012; Ando *et al.,* 2011). As a corollary to this statement, as bacteria evolve to exploit new host environments, the DNA sequences in ITGRs may be genetically embossed with SNP biomarkers that are reflective of their evolutionary

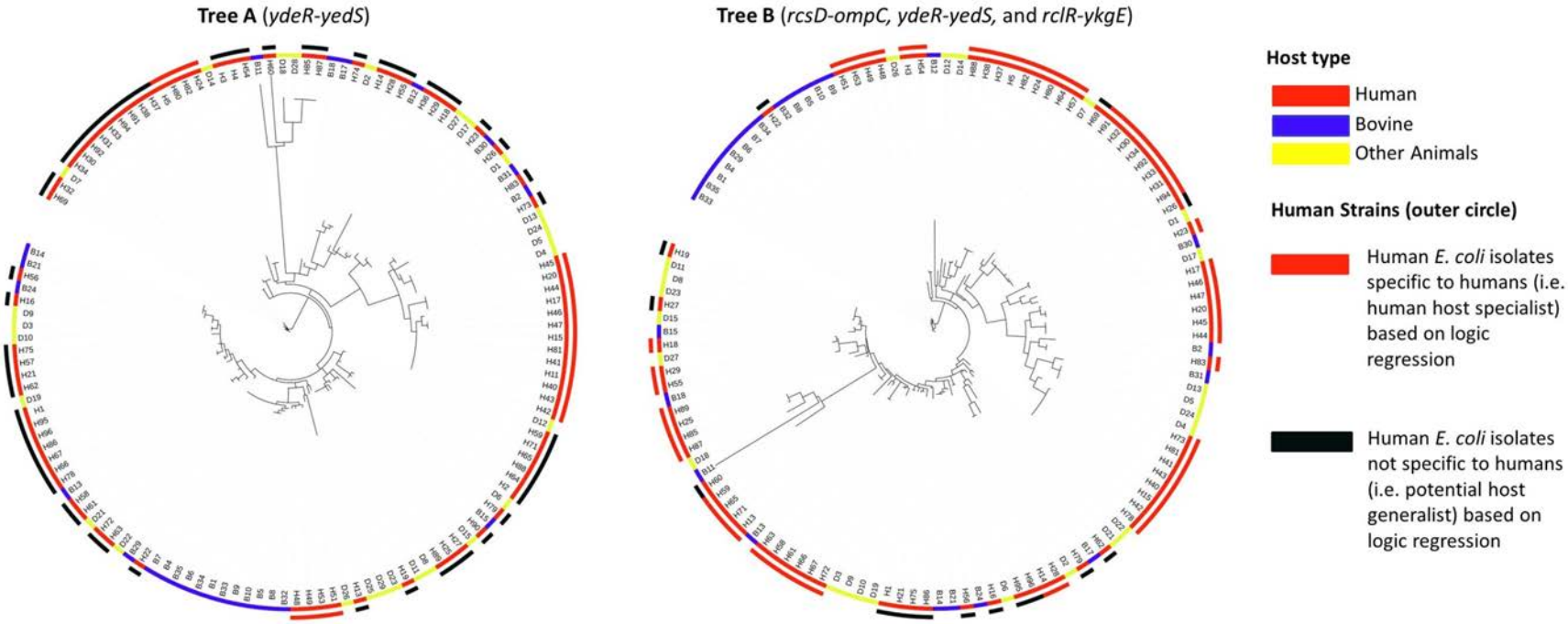

Figure 3-3. Two unrooted maximum-likelihood (ML) phylogenetic trees (A and B) of *E. coli* based on the intergenic sequences *ydeR-yedS* (Tree A) and *rcsD-ompC*, *ydeR-yedS*, and *rclR-ykgE* (Tree B), respectively. The interactive tree of life (iTOL) online tool was used for tree editing. The colors of the inner circle represent the three different host groups from which the *E. coli* strains originated (human [red], cattle [blue], all other animals [yellow]) and for which the ITGRs were analyzed by maximum likelihood phylogeny. The outer circle represents the human *E. coli* strains used in this study, with the red color representing human-specific strains (i.e., specialists) and the black color representing the potential host generalists, as determined by SNP biomarkers analysis using logic regression. The data demonstrates that unsupervised learning algorithms commonly used in bioinformatics (i.e., cluster-based analysis) fail to group *E. coli* together based on host-origin, whereas supervised learning methods (logic regression) group strains across these scattered phylogenetic clusters into specialists and generalist

trajectory towards host-specificity. ITGRs play an important role in the transcriptome, and for *E. coli*, ITGR polymorphisms may allow them to survive and compete for the limited nutrients and resources in the diverse gastrointestinal systems of animal host as well as non-host environments.

The published data in Chapter Two (Zhi *et al.,* 2015) demonstrated that some *E. coli* strains appear to have evolved to become host-specialists, capable of transmission and survival in a particular host species, as revealed by the presence of host-specific SNP biomarkers in ITGRs. The biomarker searching approach described in Chapter Two used supervised learning methods for DNA sequence analysis to reveal patterns associated with *E. coli* isolates obtained from certain animal hosts – representing the first study of its kind for evaluating the concept of host-specificity in bacteria. Building from this knowledge, one objective of this Chapter was to assess the degree to which ITGRs encode host-specific information in *E. coli.* Two independent approaches to biomarker discovery were used: a) a targeted ITGR sequence-based approach encompassing a diverse library of *E. coli* isolates obtained from different animal hosts; and b) publically-available whole genome data of 160 *E. coli* isolates isolated from different animals.

Under the targeted ITGR approach, six loci were examined and shown to encode different levels of host-specific information. When analyzed individually, the most host-informative locus for *E. coli* originating from cattle, pigs and dogs was the *uspC-flhDC* locus. In humans, the most host-informative locus was *cutC-torYZ*. When SNP biomarkers were generated from concatenated ITGR sequences, a higher level of host-predictive performance was observed as compared to any individual ITGR sequence. This suggested that greater SNP variability across multiple ITGRs may lead to more robust host-specific logic regression models. However, no correlation between the number of polymorphic sites in an ITGR and the overall level of host-predictive performance was found, suggesting that a small number of critical polymorphisms

across diverse ITGRs may be a more important determinant of host-specificity than the overall number of SNP polymorphisms in an ITGR locus.

As seen in Chapter Two, traditional bioinformatics approaches failed to resolve host-specific patterns in ITGR sequences among the *E. coli* isolates used, a finding re-iterated in this chapter using additional ITGR loci. Statistical methods for traditional DNA sequence phylogenetic analysis commonly use unsupervised learning approaches, and although valuable for finding hidden structures in sequence data (i.e., related clusters), a key feature of unsupervised learning is that group labels (such as animal host origin) are not observed as data in the analysis (Mohri *et al.,* 2012). Another potential pitfall of unsupervised learning methods is that these methods use information from all sites in the DNA sequence, including nucleotides that may be irrelevant to host-specificity. This likely results in misleading, poor-performing classification schemes in terms of predicting which animal hosts that *E. coli* may originate from. Conceptually, the genetic information on only a small number of ITGRs or SNPs within ITGRs may be decisive for determining host-specificity, and supervised learning methods may be more effective in selecting the most relevant SNP patterns pertaining to host specificity from the labeled data.

In order to identify ITGRs that were highly informative of host origin, publically-available whole genome data was used to evaluate how well 80 different ITGRs across the *E. coli* genome performed in terms of encoding host-specific information. Of the 80 ITGRs analyzed across 160 genomes (human, cattle, and other animals), only 6 had a reasonable level of host-predictive performance (i.e., HPP $\geq$ 1.4) across both cattle and humans. One ITGR locus in cattle appeared to be very host informative – the *ydeR-yedS* ITGR for which 92% of all cattle *E. coli* carried a biomarker that was 98% specific to cattle. In a concatenated sequence

encompassing the ITGRs of *rcsD-ompC / ydeR-yedS / rclR-ykgE*, a SNP biomarker that was 98% specific to humans and was possessed by 78% of all *E. coli* isolates of human origin was identified. These observations suggest that a considerable proportion of the *E. coli* in any animal host may be highly specific to that host (i.e., specialist), and that relatively few *E. coli* generalists comprise the population in the gastrointestinal tract of any specific animal host. This level of host-specificity in *E. coli* is substantially greater than that reported in Chapter Two.

Based on the results from whole genome analysis, there are some interesting observations regarding which of the 80 ITGRs examined in this study were identified as being potentially important for encoding host-specific information. For human *E. coli* strains, many of the ITGRs having the greatest host predictive power regulated genes involved in antibiotic resistance. For example, the *ydeR-yedS* locus regulates the expression of *yedS*, a gene encoding an outer membrane protein involved in carbapenem resistance (Warner *et al.,* 2013). Carbapenem resistance has been found in human clinical isolates and has been observed in food and companion animals (Guerra *et al.,* 2014). Of all 80 ITGRS examined by whole genome analysis, the *ydeR-yedS* locus had the greatest host predictive power for humans and cattle. Other ITGRs for which host-specific biomarkers were found in human *E. coli* strains included those that regulate membrane proteins (*ompC, yiaW, yagU*). The gene, *ompC*, encodes an outer membrane porin protein, important for antibiotic resistance to cephalosporins and fluoroquinolones (Tran *et al.,* 2013). *yiaW* encodes an inner membrane protein, and the disruption of its expression has also been shown to increase sensitivity to several antibiotics (Hu and Coates, 2005). Unlike the antibiotic-related ITGRs associated with host specificity in human *E. coli*, the most host-informative ITGRs in cattle-derived *E. coli* appeared to be those important in stress responses. For example, the second most host-informative ITGR in *E. coli* isolated from cattle was the

pgaA-ycdT locus. This ITGR regulates the gene, *ycdT*, a protein involved in bacterial motility. Motility has been found to be associated with biofilm formation (Wood *et al.,* 2006) and the formation of a biofilm is a strategy often used by bacteria for survival in harsh environments (Ryu and Beuchat, 2005; Uhlich *et al.,* 2006). The third most host-informative ITGR in cattle-derived *E. coli* was *ychE-oppA*. The deletion of *oppA* makes *E. coli* more sensitive to heat shock (Krisko *et al.,* 2014). Other host-informative ITGRs for the bovine group included those that regulate genes involved in bacteria adaptation to various environmental stresses, such oxidative stress [*cysZ-cysK*] (Ackerley *et al.,* 2006), UV resistance [*pagP-cspE*] (Mangoli *et al.,* 2001), motility (*flgB-flgA*) (Nambu and Kutsukake, 2000), and adhesion [*nrdA-yfaL*] (Roux *et al.,* 2005). These observations suggest that environmental persistence may be a key selective pressure for *E. coli* transmission among cattle, whereas in humans the selective pressure driving host-adaptation and specialization may be driven by antibiotic use.

When targeted ITGR sequencing approach (6 loci sequenced in house [*araH-otsB*, *metQ-rcsF*, and *cutC-torYZ*, *asnS-ompF, uspC-flhDC, csgBAC-csgDEFG*]) was compared to the whole genome approach for biomarker discovery (same 6 loci but found in NCBI whole genome databases), very similar results in context of the degree of host-specific information encoded in these loci was observed. This data suggests that whole genome approaches may be particularly important in discovering robust host-informative biomarkers, and for which validation of these biomarkers can occur through targeted ITGR sequencing of a large number of *E. coli* isolates collected from different animal and non-animal host environments. Based on work by Zaslaver *et al.* (2006), in which ~1820 promoters in *E. coli* K12 were characterized for transcriptional regulation, which of these ITGRs encodes the greatest level of host-specific information in *E. coli* was being examined. With the advent of routine genome sequencing being commonplace in

research laboratories, the number of sequenced *E. coli* genomes will greatly increase over the next few years and for which better representation from isolates collected from a wide variety of animals will likely occur. This will make logic regression biomarker analysis a more powerful approach in understanding population genetics of *E. coli*.

Nevertheless, many different factors may affect the genetic population structure of *E. coli*. Temporal variations in *E. coli* population genetics have been studied in some host species (Caugant *et al.,* 1981; Jenkins *et al.,* 2003; Gordon, 1997). In one study (Caugant *et al.,* 1981), 550 *E. coli* were isolated from a single human host over a period of 11 months, and based on multilocus enzyme electrophoresis (MLEE) results, two electrophoresis types (ET) were observed frequently over extended periods while most ETs appeared transiently. Factors that may affect the stability of SNP biomarker patterns within a single host may include diet, age, geography, health status and host-genetics, and for which these biomarker patterns may change spatially and temporally. In order to alleviate these potential confounders, it is important that *E. coli* libraries be composed of isolates collected from animal hosts over wide geographic areas, at different ages and under varying nutritional regimes. However, it is hypothesized that robust host-specific biomarkers exist in the *E. coli* genome for all major animal host species for which this bacteria has colonized, and that within each of these individual animal host species, additional biomarkers related to age, diet, and geography may also exist.

A potential implication of these findings also relates to understanding zoonosis in pathogenic *E. coli*. Based on the approach used in this study, *E. coli* strains possessing host-related biomarkers are considered host-specialists, restricted to survival in that animal host, and by statistical definition are not found in other animals (to a certain level of confidence [i.e., 95% or greater]). As a consequence, *E. coli* strains possessing animal-specific biomarkers may not be

zoonotic for humans. As an example, the data in this chapter suggests that 92% of the *E. coli* population in the GI tract of a cow are host-specialists, and although they may contain certain virulence genes associated with disease, the likelihood for zoonotic transmission to humans may be low.

Although these approaches may shed light on the evolutionary emergence of zoonoses, it does not rule out the possibility that zoonotic *E. coli* pathogens may emerge from the 'specialist population' – a potential result of the sudden acquisition of genetic elements known to be involved in pathogenesis in more than one particular host (i.e., individual virulence genes or pathogenicity islands). Although acquisition of virulence genes among strains of *E. coli* has been studied extensively, relatively few studies have simultaneously examined changes in host distribution/adaptation. Skyberg *et al.* (2006) eloquently demonstrated that an avian commensal strain of *E. coli* transformed with a virulence plasmid originating from an avian pathogenic *E. coli* (APEC) isolate, but for which the plasmid displayed virulence characteristics similar to a plasmid/pathogenicity island in uropathogenic *E. coli*, was able to outcompete non-transformed commensal strains in colonization of the urinary tracts of mice as well as induce pathogenesis in chicks. Thus, pathogen emergence may occur from both the specialist and generalist populations, the mechanisms of which may be different.

3.6 Conclusion

E. coli appears to display a high degree of host-specificity. Biomarkers of host-specificity are prominent in some, but not all, ITGRs of the *E. coli* genome. The degree of host-specific information encoded in these regions does not correlate with the overall level of sequence polymorphism, but rather with distinct SNPs alternations within these ITGRs. This

finding helps explain why traditional phylogenetic approaches (i.e., unsupervised learning-based methods) often fail to reveal host-specific patterns in DNA sequence. Traditional phylogenetic approaches make use all polymorphisms in DNA sequence, including those that may be host-informative and those that are not host informative, to describe genetic relationships (i.e., clustering). On the contrary, supervised learning methods such as logic regression search for the most informative polymorphisms in an ITGR as they relate to host origin. These novel approaches also shed light on which areas of the regulome are subject to natural selection evolutionary pressures driving host-specificity. In humans, host-specific ITGRs in *E. coli* were dominated by those genes related to antibiotic resistance, whereas in cattle *E. coli* host-specific ITGRs were dominated by those regulating stress-resistance.

Chapter Four : EVIDENCE OF NATURALIZED STRESS-TOLERANT STRAINS OF *E.COLI* IN MUNICIPAL WASTEWATER TREATMENT PLANTS [3]

4.1 Abstract

If genetic variation in the regulome is important for driving evolution of host-specificity in *E. coli*, polymorphisms in ITGRs must be deterministic and adaptive. Since *E. coli* has been proposed to have two habitats - the intestine of mammals/birds and the non-host environment, the objective of this Chapter was to assess whether certain strains of *E. coli* may have evolved towards adaptation and survival in the non-host environment, in particular, wastewater. In this study, raw sewage samples from different treatment plants were subjected to chlorine-stress, and ~59 % of the surviving *E. coli* was found to contain a genetic insertion (IS30) element located within the *uspC-flhDC* intergenic region. The positional location of the IS30 element was not observed across a library of 845 *E. coli* isolates collected from various animal hosts, nor within Genbank, or whole genome reference databases from human and animal *E. coli* isolates (n>1100). Phylogenetics clustered the IS30-containing wastewater *E. coli* isolates into a distinct clade, and biomarker analysis revealed that these wastewater isolates contained a SNP biomarker pattern that was specific for wastewater. These isolates belonged to phylogroup A, possessed a generalized stress response (RpoS), and carried the locus of heat resistance, features likely relevant for non-host environmental survival. Isolates were screened for 28 virulence genes but only carried the *fimH* marker. The data suggests that wastewater contains a naturalized resident population of *E. coli*. An end-point PCR targeting the IS30 element within the *uspC-flhDC*

[3] A version of this chapter has been published, the citation of which is: Zhi, S., G. Banting, Q. Li, T. Edge, E. Topp, M. Sokurenko, C. Scott, S. Braithwaite, N. J. Ruecker, Y. Yasui, T. McAllister, Linda Chui, and **N. F. Neumann**. 2016. Evidence of naturalized stress-tolerant *Escherichia coli* in municipal wastewater treatment plants. *Applied and Environmental Microbiology.*82: 5505-5518.

intergenic region was developed for which all raw sewage samples (n=21 [100%]), and the majority of secondary treated (16/17 [94%]) and UV-treated wastewater samples (9/12 [75%]) were positive. Conversely, their prevalence in *E. coli*-positive surface and groundwater samples was low (≤5%). This simple PCR assay may represent a convenient microbial source-tracking tool for identifying water samples impacted by municipal wastewater.

4.2 Introduction

The central hypothesis of this thesis - that genetic variation in the regulome is important for driving evolution of host-specificity in *E. coli* - necessitates that the genetic variation observed in ITGRs be adaptive and deterministic. However, most genetic variation in biological systems is believed to be non-adaptive and non-determinsitic (Kimura, 1983; Nielsen, 2005). Thus, although statistically-revelant ITGR biomarkers of host-specificity were observed in Chapters Two and Three of this thesis, it would be judicious to validate these findings in terms of phenotypic relevance. Empirically validating the phenotypic and biological relevance of ITGR biomarker variation is a daunting task in animal model systems, requiring cross-infectivity studies between conspecific and xenospecific recipient animal hosts, and for which individual host-specific strains of *E. coli* would need to be traced throughout the GI tract of these animals.

E. coli has two principal habitats: the intestinal tracts of mammals and birds; and the non-host environment (water/sediment). Survival and colonization in these distinct habitats are mediated by adaptations to nutrient availability, temperature, pH, osmolality, solar radiation and the presence of competitive microflora. Although *E. coli* has been used as a water quality indicator for fecal contamination for years, several studies have found that *E. coli* can grow outside the gastrointestinal tract of its animal hosts in tropical and subtropical environments

(Byappanahalli and Fujioka, 1998; Solo-Gabriele *et al.,* 2000; Anderson *et al.,* 2005). Naturalized *E. coli* strains have also been described in sand, sediments and water from temperate climates (Power *et al.,* 2005; Tymensen *et al.,* 2015; Kon *et al.,* 2007; Chandrasekaran *et al.,* 2015; Winfield and Groisman, 2003). The presence of *E. coli* in these environments indicates that some *E. coli* strains may survive and replicate outside animal host reservoirs.

In bacteria, survival and adaptation in different environments are necessary for their continued evolutionary success ultimately leading to genetic differences between populations. These differences may arise through mutations in regulatory sequences (Blount *et al.,* 2012), acquisition of functional genes through transduction or horizontal gene transfer (Nakata *et al.,* 1993; Dobrindt, 2005), random point mutations (Weissman *et al.,* 2003), homologous recombination (Seifert, 1996), and/or conjugation (Tobe *et al.,* 1999). Mutations that confer a fitness advantage and enhance survival in certain environments can lead to natural selection of sub-populations.

In preliminary experiments evaluating host-specific biomarkers of *E. coli* from humans, animals (Chapters Two and Three) and non-host environments, it was observed that *E. coli* isolates obtained from wastewater appeared to be genetically different than human and animal isolates. This serendipitous finding allowed for the potential interrogation of both genotype/ phenotype interactions as a function of ITGR polymorphisms in *E. coli* and its relation to niche specificity. The following Chapters (Four, Five, and Six) attempt to link genetic variation in ITGRs to functionally important and biologically plausible phenotype/genotype outcomes that may be important for *E. coli* survival in this non-host environmental niche.

Although human fecal wastes are the dominant microbial input source in sewage, municipal wastewater represents a very different environmental niche for *E. coli* compared to the

gastrointestinal tract. Wastewater has a very unique microbiome (Wang *et al.,* 2012), and its chemical composition encompasses a wide range of organic (detergents, antibiotics, personal care products) and inorganic (i.e., nutrients [nitrogen and phosphorus] and metals [mercury, lead, copper]) compounds (Henze, 2008). Consequently, based on the preliminary findings, it was hypothesized that the evolutionary selection forces imposed by wastewater treatment may have driven certain *E. coli* strains to adopt survival strategies in the non-host wastewater environment (e.g., treatment-resistance), potentially leading to the evolution of naturalized populations of *E. coli* in this niche. The objective of this study was to assess whether certain strains of *E. coli* in wastewater may be genetically distinct from fecal populations of this bacterium, based on ITGR biomarker analysis.

4.3 Material and methods

4.3.1 *E. coli* isolates

In total, 1426 *E. coli* strains were used in this study (Table 4-4). Among them, 845 *E. coli* strains were collected from 15 different animal host species including humans. Seventy *E. coli* strains were isolated from chlorine bleach-treated sewage (described below), and 187 strains were collected from surface water and groundwater.

Animal *E. coli* isolates were sourced from three bacterial libraries originally established from geographically disparate areas in Canada: a) the Hamilton/Toronto region in Ontario as described by Edge and Hill (2007); from Ottawa (Ontario), Lennoxville (Quebec) and Brandon (Manitoba) as described by Lyautey *et al.* (2010); and c) from Alberta as described by White *et al.* (2011) [human isolates]. For further details regarding bacterial libraries please refer to Chapter Two.

For *E. coli* from surface water and groundwater samples, the water samples were processed by membrane filtration and incubated on X-Gluc (Dalynn, Calgary, Canada) agar plates at 44.5 ^{o}C for 18-24 hours. One blue colony was picked from each sample and streaked on MacConkey (Dalynn Calgary, Canada) agar then incubated at 35 ^{o}C for 18-24 hours for further isolation. All presumptive human, animal, surface water and groundwater *E. coli* isolates were confirmed as *E. coli* through biochemical analysis using a Vitek Bacterial Identification System (BioMerieux Canada Inc., St. Laurent, Canada) according to manufacturer's instructions and protocols at the Provincial Laboratory for Public Health (ProvLab) in Edmonton, Alberta, Canada.

E. coli isolates from wastewater were obtained from a separate study involving an evaluation of ColiTag® as a suitable growth media for *E. coli*, with methods performed according to the United States Environmental Protection Agency's (U.S. EPA) Alternate Test Procedure (ATP) (EPA, 2010). This procedure incorporates a step in which bacteria are treated with chlorine bleach in order to stress the bacterial population. Briefly, raw sewage samples from different sewage treatment plants in Alberta, Canada, were collected and sent to the ProvLab for analysis. Each raw sewage sample was treated with 3% sodium hypochlorite until the free chlorine residual reached 0.3-0.5 ppm, causing a ~2-4 $\log_{10}$ reduction in the culturable concentration of *E. coli*, according to the ATP protocol (EPA, 2010). Chlorine reactivity was neutralized by addition of a 10% solution of sodium thiosulfate. Chlorine-treated wastewater samples were then used to inoculate either ColiTag® or in lauryl trypticase broth (LTB)/BCG media according to Method 9221.F in *Standard Methods for the Examination of Water and Wastewater* (Rice *et al*., 2012). *E. coli* was isolated from ColiTag® or LTB/BCG positive cultures by selective plating onto X-Gluc agar plates (Ciebin *et al.,* 1995) and incubated at 44.5

^{o}C for 24 h. Blue colonies were picked and streaked onto non-selective blood agar plates and incubated at 35 ^{o}C for 24 h. All presumptive wastewater *E. coli* isolates were confirmed through biochemical analysis using a Vitek Bacterial Identification System (BioMerieux Canada Inc., St. Laurent, Canada) according to manufacturer's instructions and protocols at the ProvLab. From this large collection of chlorine-tolerant *E. coli* isolates, 70 isolates originating from four geographically separated wastewater treatment plants (WWTPs) were chosen for further phenotypic and genetic analysis.

4.3.2 Environmental water samples

In addition to the four WWTPs for which chlorine-tolerant *E. coli* were isolated and the microbes genetically characterized, wastewater samples (treated and untreated) were also collected from the City of Calgary at two of their WWTPs (Bonnybrook and Pine Creek facilities). City of Calgary wastewater samples were used to: a) determine what proportion of wastewater samples (reflecting the entire *E. coli* population in that wastewater sample) possessed a novel genetic marker, *uspC-IS30-flhDC* (identified in this present Chapter and described below), at different points in the treatment process; and b) what proportion of the total numbers of *E. coli* present in untreated wastewater possessed the *uspC-IS30-flhDC* marker. In the first analysis, 100 ml of wastewater (untreated and treated) was cultured in Colilert® vessels (IDEXX Laboratories, Inc., Westbrooke, Maine, USA) and incubated at 35 °C for 24 h to enrich the population of *E. coli* in the sample. Samples positive for *E. coli* by Colilert® were saved and 1 mL of the Colilert® culture was centrifuged at 6,300 x *g* for 10 min to pellet bacteria for DNA extraction and PCR detection of the *uspC-IS30-flhDC* marker (see below). In experiments aimed at determining what proportion of *E. coli* isolates in a sample carry the marker, eight untreated

wastewater samples were collected from the City of Calgary at different dates and at their two WWTPs and the total count of *E. coli* quantified using 10-fold serial dilutions of the samples analyzed by most probable number (MPN) in a Colilert® QuantiTray® format (IDEXX). The total number of *E. coli* in a sample were determined based on number of wells having both β-glucosidase (yellow) and β-glucuronidase activity (fluorescence). To determine the number of *E. coli* possessing the *uspC-IS30-flhDC* marker in that sample, the contents of each *E. coli* positive well in the QuantiTray® were aseptically removed with a syringe and transferred to a 2 ml centrifuge tube. Samples were centrifuged at 10,000 x*g* for 10 min to collect bacteria pellets. DNA was extracted (according to methods described below) and PCR was performed (described below). The number of *E. coli* positive wells (yellow and fluorescent) and the number of those wells that were PCR positive for the *uspC-IS30-flhDC* marker in all eight wastewater samples were used to estimate the number of *E. coli* that were marker-positive in wastewater.

Environmental water samples also included groundwater and surface water samples. Groundwater samples were collected through routine testing at the ProvLab. Drinking water samples from privately owned groundwater wells were tested at the ProvLab for *E. coli* using Colilert® (IDEXX Laboratories Inc, Westbrooke, Maine, USA). Groundwater samples positive for *E. coli* by Colilert® were saved and 1 mL of the Colilert® culture was centrifuged at 6,300 x *g* for 10 min to collect bacterial pellets for DNA extraction (described below) and PCR analysis for the *uspC-IS30-flhDC* marker (described below). Colilert® positive samples represented the *E. coli* population present in the groundwater. Surface water samples were also obtained through routine submissions to ProvLab and processed by either membrane filtration on either X-gluc agar plates (Ciebin *et al.,* 1995) or by Colilert®. From X-gluc agar plates positive for *E. coli,* one presumptive *E. coli* colony was picked, verified as *E. coli* using a Vitek Bacterial Identification

System (BioMerieux Canada Inc., St-Laurent, Canada) and colony PCR was performed to determine the occurrence of the *uspC-IS30-flhDC* marker (described below). Other surface water samples were processed by Colilert® presence/absence testing (similar to groundwater samples above) in order to assess prevalence of the *uspC-IS30-flhDC* marker in the *E. coli* population of surface water samples.

4.3.3 Phenotypic stress response (RpoS) activity

The *rpoS* regulated stress response was evaluated in individual *E. coli* isolates using the glycogen and catalase tests, as described by White *et al.* (2011). *E. coli* were grown on LB agar and a single colony was transferred to tryptic soy broth (TSB) and grown overnight at 35 ^{o}C. One µL of the culture was inoculated on LB agar for colony growth and the plate incubated at 28^{o}C for 48 h. For detection of glycogen production, 5 mL of iodine solution (0.1 M I_2, 0.03M KI) was added to each plate which was allowed to stand for 5 min (White *et al.,* 2011; White *et al.,* 2010). Dark brown colonies were considered positive, while pale brown or white colonies were considered attenuated or null for glycogen production (Hengge-Aronis and Fischer, 1992). Catalase activity was tested by applying a 6% (wt/vol) hydrogen peroxide solution to individual colonies and a vigorous bubbling reaction was considered positive for catalase production. RpoS-positive [S. enterica serovar Typhymurium ATCC 14028] (White *et al.,* 2011), or RpoS-negative [S. enterica serovar Typhymurium ΔrpoS] (White *et al.,* 2011) isolates were used as control strains. *E. coli* isolates were considered as RpoS positive if both the glycogen and catalase tests were positive.

4.3.4 DNA extraction from individual *E. coli* isolates and cultured water samples

All *E. coli* isolates used in this study were grown in tryptic soy broth (TSB) and genomic DNA extracted from *E. coli* cultures using DNeasy Blood & Tissue kits (Qiagen, Waltham, Massachusettes) according to the manufacturer's instructions. Genomic DNA was then stored at -20 °C. For wastewater and raw water samples testing positive for *E. coli* by Colilert® or membrane filtration (i.e., section above), the DNA from bacteria was extracted by suspending the pellets in boiling molecular biology grade water for 10 min, followed by centrifugation at 10,000 x*g* for 10 min. The supernatant containing the DNA was then stored at -20 °C until used for PCR.

4.3.5 Detection of virulence genes in wastewater *E. coli* isolates

Ten *E. coli* isolates, possessing the wastewater-specific *uspC-IS30-flhDC* marker, were tested for 28 virulence genes associated with *E. coli* pathogenesis in humans and animals (a complete list of virulence genes is provided in Table 4-1). The 10 isolates selected were from four different wastewater treatment plants in Alberta, Canada. The primers and PCR conditions are provided in Table 4-2. Real-time PCR was used to detect Shiga toxin 1 (*stx*1) and Shiga toxin 2 (*stx*2), while all other virulence genes were amplified using end-point PCR. For all real-time PCR reactions, the reaction volume was 20 µL, and contained 10 µL TaqMan® Fast Advanced Master Mix (Applied Biosystems, Foster City, California, USA), 0.9 µM of each primer, 0.25 µM Taqman probe, 5 µL DNA template, and molecular biology grade water added to a total volume of 20 µL. The real-time PCR conditions were 50 °C for two minutes, 95°C for 30 seconds followed by 40 cycles of 95°C for three seconds and 60°C for 30 seconds. For all end-point PCR reactions, the reaction volume contained 5µL of *E. coli* genomic DNA template

Table 4-1. PCR primers for virulence genes

Primer Set	Gene Name	Primer name	Primer sequence (5'–3')	Product size (bp)	Reference
I	*aer*	aer-F	TACCGGATTGTCATATGCAGACCGT	601	(White *et al.,* 2011)
		aer-R	AATATCTTCCTCCAGTCCGGAGAAG		
	papC	papC-F	GTGGCAGTATGAGTAATGACCGTTA	202	
		papC-R	ATATCCTTTCTGCAGGGATGCAATA		
	traT	traT-F	GGTGTGGTGCGATGAGCACAG	287	
		traT-R	CACGGTTCAGCCATCCCTGAG		
II	*PAI*	PAI-F	GGACATCCTGTTACAGCGCGCA	921	(White *et al.,* 2011)
		PAI-R	TCGCCACCAATCACAGCCGAAC		
	fimH	fimH-F	TGCAGAACGGATAAGCCGTGG	505	
		fimH-R	GCAGTCACCTGCCCTCCGGTA		
III	*iroN*	iroN-F	AAGTCAAAGCAGGGGTTGCCCG	667	(White *et al.,* 2011)
		iroN-R	GACGCCGACATTAAGACGCAG		
	iutA	iutA-F	GGCTGGACATCATGGGAACTGG	301	
		iutA-R	CGTCGGGAACGGGTAGAATCG		
	ibeA	ibeA-F	AGGCAGGTGTGCGCCGCGTAC	169	
		ibeA-R	TGGTGCTCCGGCAAACCATGC		
IV	*cnfl*	cnfl-F	AAGATGGAGTTTCCTATGCAGGAG	497	(White *et al.,* 2011)
		cnfl-R	CATTCAGAGTCCTGCCCTCATTATT		
	papGII	papGII-F	GGGATGAGCGGGCCTTTGAT	189	
		papGII-R	CGGGCCCCCAAGTAACTCG		
V	*fuyA*	fuyA-F	TGATTAACCCCGCGACGGGAA	784	(White *et al.,* 2011)
		fuyA-R	CGCAGTAGGCACGATGTTGTA		
	papGIII	papGIII-F	GGCCTGCAATGGATTTACCTGG	257	

		papGIII-R	CCACCAAATGACCATGCCAGAC		
VI	*sfa-foc*	sfa/foc-F sfa/foc-R	CTCCGGAGAACTGGGTGCATCTTAC CGGAGGAGTAATTACAAACCTGGCA	407	(White *et al.,* 2011)
	hlyA	hlyA-F hlyA-R	AACAAGGATAAGCACTGTTCTGGCT ACCATATAAGCGGTCATTCCCGTCA	1176	
VII	*iha*	iha-F iha-R	CTGGCGGAGGCTCTGAGATCA TCCTTAAGCTCCCGCGGCTGA	826	(White *et al.,* 2011)
VIII	*aidA*	AIDA-I-F AIDA-I-R	TGCAAACATTAAGGGCTCG CCGGAAACATTGACCATACC	370	(Chapman *et al.,* 2006)
	aidA	aidA-F aidA-R	CAGTTTATCAATCAGCTCGGG CCACCGTTCCGTTATCCTC	450	
	aah	aah-F aah-R	CTGGGTGACATTATTGCTTGG TTTGCTTGTGCGGTAGACTG	543	
IX	*LT*	LTA-F LTA-R	GGCGACAGATTATACCGTGC CCGAATTCTGTTATATATGTC	696	(Chapman *et al.,* 2006)
	Stb	STb-F STb-R	ATCGCATTTCTTCTTGCATC GGGCGCCAAAGCATGCTCC	172	
X	*hra*	hra-F hra-R	CAGAAAACAACCGGTATCAG ACCAAGCATGATGTCATGAC	257	(Chapman *et al.,* 2006)
XI	*eaeA*	eaeA-F eaeA-R	GACCCGGCACAAGCATAAGC CCACCTGCAGCAACAAGAGG	384	(Chapman *et al.,* 2006)
XII	*Sta*	Sta-F Sta-R	TCTTTCCCCTCTTTTAGTCAG ACAGGCAGGATTACAACAAAG	166	(Chapman *et al.,* 2006)

XIII	*ipaH*	ipaH-III ipaH-IV	GTTCCTTGACCGCCTTTCCGATACCGTC GCCGGTCAGCCACCCTCTGAGATAC	600	(Chapman *et al.,* 2006)
XIV	*usp*	usp-F usp-R	CGGCTCTTACATCGGTGCGTTG GACATATCCAGCCAGCGAGTTC	614	(White *et al.,* 2011)
XV	*irp2*	irp2-F irp2-R	AAGGATTCGCTGTTACCGGAC6 TCGTCGGGCAGCGTTTCTTCT	285	(White *et al.,* 2011)
XVI	*stx1*	stx1-F stx1-R stx1-P	CATCGCGAGTTGCCAGAAT GCGTAATCCCACGGACTCTTC [FAM]-CTGCCGGACACATAGAAGGAAACTCATCA-[TAMARA]	78	(Chui *et al.,* 2010)
XVII	*stx2*	stx2-F stx2-R stx2-P	CCGGAATGCAAATCAGTC CAGTGACAAAACGCAGAACT [FAM]-ACTGAACTCCATTAACGCCAGATATGA-[TAMARA]	113	(Chui *et al.,* 2010)

and 12.5 µL of Fermentas Maxima Hotstart 2X master mix (ThermoFisher Scientific, Waltham, Massachusettes), 0.5 µM of each primer, and molecular biology grade water added to a total volume of 25 µL. The PCR products were run on a 2% agarose gel in 1X TAE buffer (Promega, Madison, Wisconsin, USA) at 140 V for 45 minutes.

4.3.6 Detection of an environmental resistance genomic island in wastewater *E. coli* isolates

The presence of a heat resistance genomic island, known as the locus of heat resistance [LHR] and previously identified in highly heat-resistant *E. coli* strains (Mercer *et al.,* 2015), was also assessed in chlorine-tolerant wastewater strains. PCR reactions targeting three fragments of this genomic island were performed on all 70 wastewater *E. coli* isolates (Table 4-3). PCR conditions for Fragment-A were as follows: initial denaturation at 95^{o}C for 5 min, 30 cycles of 95 ^{o}C for 45 s, annealing at 63 ^{o}C for 1 min, and 72 ^{o}C for 1.5 min, and followed by a 7 min extension at 72 ^{o}C. PCR conditions for Fragment-B were as follows: initial denaturation at 95^{o}C for 5 min, 30 cycles of 95 ^{o}C for 45 s, annealing at 64 ^{o}C for 1 min, and 72 ^{o}C for 2.5 min, and followed by a 7 min extension at 72 ^{o}C. PCR conditions for Fragment-C were the same as the conditions for Fragment-B except the annealing temperature was 58 ^{o}C. The PCR products of Fragment-A, B, and C were run on a 2% agarose gel in 1X TAE buffer (Promega, Madison, Wisconsin, USA) at 140 V for 45 minutes.

Table 4-2. PCR conditions of virulence genes.

Primer Set[a]	Initial denaturation		Number of Cycles	Denaturation		Annealing		Extension		Final extension	
	Temp (°C)	Time (minutes)		Temp (°C)	Time (seconds)	Temp (°C)	Time (seconds)	Temp (°C)	Time (seconds)	Temp (°C)	Time (minutes)
I, II, III, IV, V, VI, VII, XIV, XV	94	5	33	94	30	63	30	72	60	72	7
VIII	94	5	33	94	60	63	60	72	30	72	10
IX	94	5	30	94	60	58	60	72	60	72	7
X	94	5	33	94	30	55	30	72	30	72	7
XI	94	5	30	94	60	58	60	72	40	72	7
XII	94	5	30	94	60	53	60	72	30	72	7
XIII	94	5	33	94	30	50	30	68	60	72	7

a. The primer set in this table corresponds to the primer set listed in Table 4-1.

4.3.7 Genetic characterization of *E. coli* isolates obtained from humans, animals and wastewater

4.3.7.1 Phylogrouping

Phylogrouping of *E. coli* isolates was performed by using PCR assays developed by Clermont *et al.* (2013). The PCR assays targeting *chuA*, *yjaA*, and the TSPE4.C2 loci were adopted from a previous version of this phylogrouping method (Clermont *et al.,* 2000). The total volume for all PCR reactions was 25 µL, which contained 5µL of *E. coli* genomic DNA template and 12.5 µL of Fermentas Maxima Hotstart 2X master mix (ThermoFisher Scientific, Waltham, Massachusettes), 1.25 µL of each primer (10 pmol/ µL), and 5 µL molecular water. Each PCR mixture contained 1.25 U of *Taq* polymerase, dNTPs at a concentration of 200 µM, 1X PCR buffer, and 2 mM Mg^{2+}. The PCR conditions used for phylogrouping were the same as those previously described (Clermont *et al.,* 2013; Clermont *et al.,* 2000).

4.3.7.2 Phylogenetic analysis of human, animal and wastewater isolates

For evaluating genetic similarities among human, animal and wastewater *E. coli* isolates, maximum likelihood phylogenetic analysis was performed on three intergenic regions: i) *uspC-flhDC*, ii) *csgBAC-csgDEFG*, and iii) *asnS-ompF*. The selection of these regions was based on the results in Chapter Two in which host-specific SNP biomarkers were identified in these three intergenic regions across 15 animal species (Zhi *et al.,* 2015) and for which these intergenic sequences correlated with adaptive phenotypes in *E. coli* (White *et al.,* 2011). The PCR conditions targeting intergenic regions *uspC-flhDC*, *csgBAC-csgDEFG*, and *asnS-ompF* were as follows: initial denaturation at 95^{o}C for 4 min, and 33 cycles of 95 ^{o}C for 30 s, 58 ^{o}C for 30 s, and

72 °C for 1 min, followed by a 7 min extension at 72 °C. Electrophoresis was performed on 2% agarose gel in 1X TAE buffer at 140 V for 45 minutes.

Table 4-3. PCR primers used in this study.

Target	Primer	Primer sequence (5'-3')	Reference
asnS-ompF	ompF-F	TACGTGATGTGATTCCGTTC	Zaslaver *et al.* (2006)
	ompF-R	TGTTATAGATTTCTGCAGCG	
csgBAC-csgDEFG	csgD-1	GGACTTCATTAAACATGATG	Zaslaver *et al.* (2006)
	csgD-2	TGTTTTTCATGCTGTCAC	
uspC-flhDC	flhDC-F	GAGGTATGCATTATTCCCACCC	White *et al.* (2011)
	flhDC-R	TGGAGAAACGACGCAATC	
uspC-IS30-flhDC	flh-IS-F	CGGGGAACAAATGAGAACAC	This study
	flh-IS-R	TGGAGAAACGACGCAATC	White *et al.* (2011)
Fragment A	HR-F1	TTAGGTACCGCTGTCCATTGCCTGA	Mercer *et al.* (2015)
	HS-R1	AGACCAATCAGGAAATGCTCTGGACC	
Fragment B	HR-F2.2	GAGGTACCTGTCTTGCCTGACAACGTTG	Mercer *et al.* (2015)
	HR-R2	TATCTAGAATGTCATTTCTATGGAGGCATGAATCG	
Fragment C	HS-F1	GCAATCCTTTGCCGCAGCTATT	Mercer *et al.* (2015)
	HR-R3	GTCAAGCTTCTAGGGCTCGTAGTTCG	

The PCR products were purified using an Agencourt AMPure XP – PCR Purification kit (Beckman, Brea, California, USA). DNA sequencing was performed by Macrogen Inc. (Seoul, South Korea) bi-directionally. The ClustalX 2.0 (Larkin *et al.,* 2007) program was used for sequence alignment, and the resulting alignments were manually edited to trim the 5'and 3' regions that contained missing data.

Due to the presence of unique DNA insertion elements in chlorine-tolerant wastewater *E.coli* isolates within the *uspC-flhDC* intergenic region (described below), our phylogenetic analysis was based on the unrelated *csgBAC-csgDEFG* and *asnS-ompF* intergenic regions. The sequences were assembled into a concatenated single file for each *E. coli* strain, aligned, and analyzed by RAxML (Stamatakis *et al.,* 2008) to generate a maximum likelihood phylogenetic tree based on the GAMMA+Invar model (Stamatakis *et al.,* 2008). The tree was edited to map to different groupings using the Interactive Tree Of Life (iTOL) online tool (Letunic and Bork, 2007; Letunic and Bork, 2011).

4.3.7.3 SNP biomarker analysis

As an alternative method for evaluating the genetic uniqueness of chlorine-tolerant wastewater *E. coli*, a novel logic regression-based biomarker approach described in Chapter Two was used to identify SNP patterns in intergenic regions (*csgBAC-csgDEFG* and *asnS-ompF*) that were highly specific to wastewater *E. coli* isolates. This entailed using logic regression as a supervised learning classification method for distinguishing between *E. coli* isolates from different animal host/non-host environmental reservoirs using host-specific informative SNP patterns in intergenic DNA sequences. This analysis sought to determine whether a highly specific SNP pattern in chlorine-tolerant wastewater *E. coli* existed, and if so, was it distinct

from other human and animal isolates. In this analysis, biomarker sensitivity was defined as the proportion of samples from a targeted host or wastewater sample that carried a specific SNP pattern. Biomarker specificity was defined as the proportion of samples from hosts/ non-host environments other than the target host/ non-host environment (i.e., all other human and animal hosts) that did not carry the SNP biomarker of interest. Model building and statistical evaluations were done according to Chapter Two.

Since logic-regression attempts to fit a very flexible model to a set of observations, *over fitting*, a common problem in supervised learning, is of concern. A fivefold cross-validation was used as an attempt to prevent overfitting and unbiasedly evaluate the performance of the final logic SNP biomarker models for human, animal and wastewater *E. coli* isolates as outlined previously in Chapter Two (Zhi *et al.,* 2015). For performing fivefold cross-validation, data sequences from each host/ non-host environment were randomly divided (i.e., computer-based randomization) into five subsets each having an equal number of samples. Four subsets of data were used as training data for logic regression analysis and the last data subset used as testing data for evaluating the logic regression model that was identified based on the training data. This was repeated for all possible subsets for training and testing, and the results were averaged over the five subsets.

4.3.8 Development of a PCR assay specific to naturalized wastewater *E. coli*

4.3.8.1 Development of PCR assay

During the genetic screening and comparison of human, animal and wastewater *E. coli* isolates, it was observed that many of the wastewater isolates surviving chlorine bleach treatment contained an insertional element (IS30) located specifically in the *uspC-flhDC* intergenic region

(herein referred to as *uspC-IS30-flhDC* marker). As such, this study sought to develop a sensitive and specific endpoint PCR to target the IS30 element in the *uspC-flhDC* intergenic region. Primers were designed (Table 4-3) and tested against genomic extracts of chlorine-tolerant wastewater-*E. coli* isolates. The PCR conditions used to amplify the *uspC-IS30-flhDC* marker were as follows: initial denaturation at 95 ^{o}C for 4 min, and 35 cycles of 95 ^{o}C for 30 s, 60 ^{o}C for 30 s, and 72 ^{o}C for 30 s, and followed by a 7 min extension at 72 ^{o}C.

To test PCR sensitivity, a plasmid containing the genetic marker was constructed. Specifically, genomic DNA from a chlorine-tolerant *E. coli* isolate containing the *uspC-IS30-flhDC* marker was used as a DNA template and PCR was used to amplify the *uspC-flhDC* intergenic region (primer set flh-IS-F and flh-IS-R in Table 4-3). The PCR product was resolved on a 2% agarose gel in 1X TAE buffer (Promega) at 140 V for 45 minutes and purified from the gel using a QIAquick Gel Extraction Kit (Qiagen) and the amplicon cloned using a TOPO® TA Cloning Kit (Invitrogen). Isolation of recombinant plasmid DNA was performed using QIAprep Miniprep Kit (Qiagen) and the presence of the correct insert was confirmed by PCR screening and DNA sequencing of the cloned inserts. DNA sequencing was performed by Macrogen Inc. (Seoul, South Korea).

PCR sensitivity was tested using both cloned plasmid DNA and *E. coli* genomic DNA. The concentration of plasmid and genomic DNA was quantified using a Qubit® 2.0 Fluorometer (Invitrogen). Plasmid copy number was determined from cloned targets, and the number of targets in the genome was calculated assuming a genome size of 4.7 Mbp and single copy of each gene within each genome. Ten-fold serial dilutions, in replicates of eight, of both plasmid and genomic DNA were made and the *uspC-IS30-flhDC* marker amplified by PCR. The limit of

detection (with 95% confidence intervals [LOD_{95}]) was calculated using Excel (Wilrich and Wilrich, 2009).

4.3.8.2 Evaluating the specificity of the *uspC-IS30-flhDC* PCR

The specificity of the PCR primers targeting the *uspC-IS30-flhDC* marker in chlorine-tolerant wastewater *E. coli* was evaluated against DNA obtained from: a) 845 *E. coli* isolates from humans and animals - to determine if the IS30 element in the *uspC-flhDC* intergenic region was specific to chlorine-tolerant wastewater *E. coli;* and b) in 178 environmental water samples (*E. coli* positive surface water, groundwater and wastewater samples) - to determine the prevalence of the marker in *E. coli* populations from wastewater compared to *E. coli* in groundwater and surface water sources.

4.4 Results

4.4.1 Evidence for genetically unique strains of chlorine-tolerant *E. coli* in wastewater

The *uspC-flhDC* intergenic region was amplified in 824 of 845 *E. coli* isolates from 15 different animal species with all of them generating a PCR amplicon that was 739 bp long. In comparison, when the same PCR was performed on 70 *E. coli* strains isolated from chlorine-treated wastewater samples, 43 of the 70 isolates (61.4%) produced PCR amplicons with sizes of 1223bp or 1155bp. The abnormally long PCR products were bi-directionally sequenced by Sanger sequencing and aligned with the sequences of *E. coli* from human and animals and for which two large insertion sequences were found in the *uspC-flhDC* intergenic region of these isolates. These large internal sequences were analyzed by BLAST and searched against the IS Finder Database (Siguier *et al.,* 2006). Fifty-nine percent (41/70) of the chlorine-tolerant

wastewater isolates contained an insertional sequence element designated IS30 (Caspers *et al.,* 1984), while 3% (2 of the 70) contained another insertional element sequence ISEc33 (Poirel *et al.,* 2011). PCR amplicons of 1223 bp were shown to contain the IS30 element, whereas those wastewater *E. coli* isolates possessing the ISEc33 element produced amplicons of 1155 bp in length. The IS30 insertion sequence element is one of the smallest known bacterial transposon encoding a motility gene transposase (Caspers *et al.,* 1984).

The entire *uspC-IS30-flhDC* sequence was processed through BLAST analysis in GenBank, and no similar sequences were identified in the NCBI database, suggesting that the positional location of the IS30 element within the *uspC-flhDC* intergenic region may be unique to chlorine-tolerant wastewater *E. coli* isolates. As further evidence of this, none of the human or animal *E. coli* isolates analyzed in this study possessed the position-specific IS30 element (Table 4-4). Furthermore, BLAST analysis of the entire *uspC-IS30-flhDC* sequence against *E. coli* genomes from isolates originating from humans (n=1107), cattle (n=41) and other animals (n=29) (Table 4-4), revealed no sequence similarities in these genomic databases.

In order to provide further evidence of the genetic uniqueness of these wastewater isolates, a maximum likelihood phylogenetic analysis as well as logic regression biomarker analysis (Zhi *et al.,* 2015) were performed at intergenic regions (*csgBAC-csgDEFG* and *asnS-ompF*) unrelated to the *uspC-flhDC* region for all human, animal and wastewater isolates. Of the 70 chlorine-tolerant wastewater *E. coli* isolates, 68 were amplifiable at both the *csgBAC-csgDEFG* and *asnS-ompF* regions. The intergenic regions of *csgBAC-csgDEFG* and *asnS-ompF* in these *E. coli* isolates as well as 780 *E. coli* isolates from 15 different animal species (Table 4-5) were sequenced. Intergenic sequence data was concatenated and a multiple sequence alignment was then used to construct a maximum likelihood phylogenetic tree (Figure 4-1). All

Table 4-4. Presence of the *uspC-IS30-flhDC* marker in *E. coli* isolates or populations from various animal and environmental sources.

	Source	No. of Isolates Tested	No. of Isolates with IS30 element in *uspC-flhDC* intergenic region
Animal, human or environmental source of E. coli isolates	**Bovine**	120	0 (0%)
	Bovine (*in silico*) [a]	41	0 (0%)
	Cat	21	0 (0%)
	Dog	61	0 (0%)
	Deer	48	0 (0%)
	Goose	54	0 (0%)
	Human	105	0 (0%)
	Human (*in silico*)[a]	1107	0 (0%)
	Chicken	59	0 (0%)
	Moose	14	0 (0%)
	Muskrat	56	0 (0%)
	Horse	44	0 (0%)
	Pig	49	0 (0%)
	Coyote	44	0 (0%)
	Gull	18	0 (0%)
	Beaver	40	0 (0%)
	Sheep	47	0 (0%)
	Other animals (*in silico*) [a,b]	29	0 (0%)
	Wastewater (chlorine treated)[c]	70	41 (59%)
	Wastewater (raw)[d]	319	16 (5%)
	Surface water [e]	187	1 (0.5%)

[a] *In silico* refers to BLAST sequence analysis done of *E. coli* genomes in GenBank.
[b] Other animal genomes include chicken, pig,
[c] Chlorine-treated *E. coli* isolates in this category originated from four different wastewater treatment plants in Alberta, Canada.
[d] Raw wastewater isolates were collected from various wastewater plants in Alberta.
[e] Isolates from surface water were collected from various rivers and lakes in Alberta, and represented by one isolate per source.

chlorine-tolerant wastewater *E. coli* isolates possessing the IS30 insertion sequence in *uspC-flhDC* intergenic region clustered together in the *csgDEFG /asnS-ompF*-derived phylogenetic tree, providing evidence for their genetic difference from animal and human strains.

Microbial host-specific biomarker analysis, as described in Chapters Two and Three, was also used to determine genetic uniqueness based on SNP patterns in DNA sequences in *E. coli* that relate to animal host origin or non-host environmental origin. Logic regression analysis of SNP patterns in the concatenated *csgDEFG /asnS-ompF* intergenic sequences revealed host-specific patterns across most human and animal *E. coli* isolates (Table 4-5). In humans, for which logic regression models were subject to fivefold cross validation, 41% of all human *E. coli* isolates carried a SNP biomarker pattern that was unique to humans (i.e., specificity of 1 [Table 4-5]). Other animals for which a sensitive host-specific biomarker could be found (i.e., greater than 35% of isolates in the animal group carrying the specific biomarker and based on fivefold cross validation) included deer (37%), chickens (56%), muskrats (77%), coyote (59%), beaver (35%) and sheep (37%), all with specificity $\geq$ 99%. Remarkably, one of the greatest sensitivities observed in SNP biomarker analysis was in the chlorine-tolerant wastewater *E. coli* group, with 82% of the 68 isolates possessing a SNP biomarker with very high specificity to wastewater ($\geq$99%) based on grouped analysis. Of the 41 chlorine-tolerant *E. coli* isolates possessing the IS30 element in *uspC-flhDC* intergenic region, all isolates possessed the same SNP biomarker unique to wastewater in the *csgDEFG /asnS-ompF* intergenic sequences, providing further evidence of a common genetic background among the wastewater strains but that these were genetically different than human and animal isolates.

Table 4-5. Logic regression-based SNP analysis of *E. coli* samples classified according to isolation source.

Host Source	Number of Samples	Logic regression		Fivefold cross validation	
		Sensitivity	Specificity	Sensitivity	Specificity
Wastewater	68	0.82	1	0.76	0.99
Bovine[a]	120	0.38	0.98	0.21	0.96
Cat	21	0.29	1	0	1
Dog	61	0.57	0.96	0.23	0.97
Deer	48	0.88	0.95	0.37	1
Goose	54	0.24	0.99	0.09	0.99
Human	105	0.46	1	0.41	1
Chicken	59	0.68	1	0.56	1
Moose	14	0.29	0.99	0	1
Muskrat	56	0.77	0.99	0.77	0.99
Horse	44	0.11	0.99	0	1
Pig	49	0.39	1	0.22	1
Coyote	44	0.64	0.99	0.59	0.99
Gull	18	0.44	0.99	0.05	0.99
Beaver	40	0.45	0.99	0.35	1
Sheep	47	0.49	1	0.37	1

1. [a] Analysis was carried out in the intergenic regions of the *csgBAC-csgDEFG* (listed as csg in models above) and *asnS-ompF* (listed as omp in models above) as described in Materials and Methods.
2. [b] The logic regression-based biomarker model for *E. coli* from wastewater was as follows: logic (E[Y]) = 3.33 - 19.9 * ((omp47_T or csg1122_G) or ((not omp564_C) or (not omp335_A))) -7.01 * ((((not csg1180_G) and omp440_T) and (not omp81_T)) and ((csg993_G and omp481_G) and (not omp423_A))). In this wastewater logic regression model, E[Y] is the probability of an *E. coli* isolate originating from a wastewater source. In reading the model output, a value designated by 'omp47_T' refers to the thymine base at nucleotide position 47 in the multisequence alignment of the *asnS-ompF* intergenic region. For further interpretation of model variables, see Chapter Two.

A PCR based phylogrouping method was also used to also characterize all 70 chlorine-tolerant *E. coli* wastewater isolates, including those that did not contain either the IS30 or ISEc33 elements. Phylogroup A was the most prevalent group (74.3%, 52/70) followed by group D (11.4%, 8/70), B1 (7.1%, 5/70), B2 (4.3%, 3/70), and E (2.9%, 2/70). Noticeably, all chlorine-tolerant wastewater *E. coli* strains that carried the IS30 element in *uspC-flhDC* region belonged to phylogroup A.

To evaluate whether these chlorine-tolerant wastewater *E. coli* strains were simply adapted to survival to water in general, 187 *E. coli* isolates collected from various other water sources (rivers, lakes, etc.) were examined for the presence of the IS30 element in the *uspC-flhDC* intergenic sequence. Only 1 of the 187 *E. coli* isolates (0.5%) collected from surface water was positive, suggesting that strains possessing the *uspC-IS30-flhDC* sequence may be specifically adapted for wastewater survival. This was further supported by the observation that the prevalence of the IS30 element in the *uspC-flhDC* region was far greater (59% [41/70]) in *E. coli* obtained from chlorine-treated sewage than in *E. coli* directly obtained from untreated sewage (5% [16/319]) (Table 4-4).

DNA sequence analysis, phylogenetic analysis, biomarker analysis, phylogrouping, and comparative prevalence assessments all supported the evidence that the *E. coli* isolates that carry the IS30 insertion element in *uspC-flhDC* region appear to be genetically distinct from *E. coli* strains isolated from humans and animals, and that they may be specifically adapted to survival in wastewater. It is worth noting that the 41 wastewater isolates containing the IS30 element in the *uspC-flhDC* region were actually isolated from four geographically segregated WWTPs in the province of Alberta. This data suggests that not only are these strains unique genetically but that its genetic signature is consistent across geographical regions in Alberta.

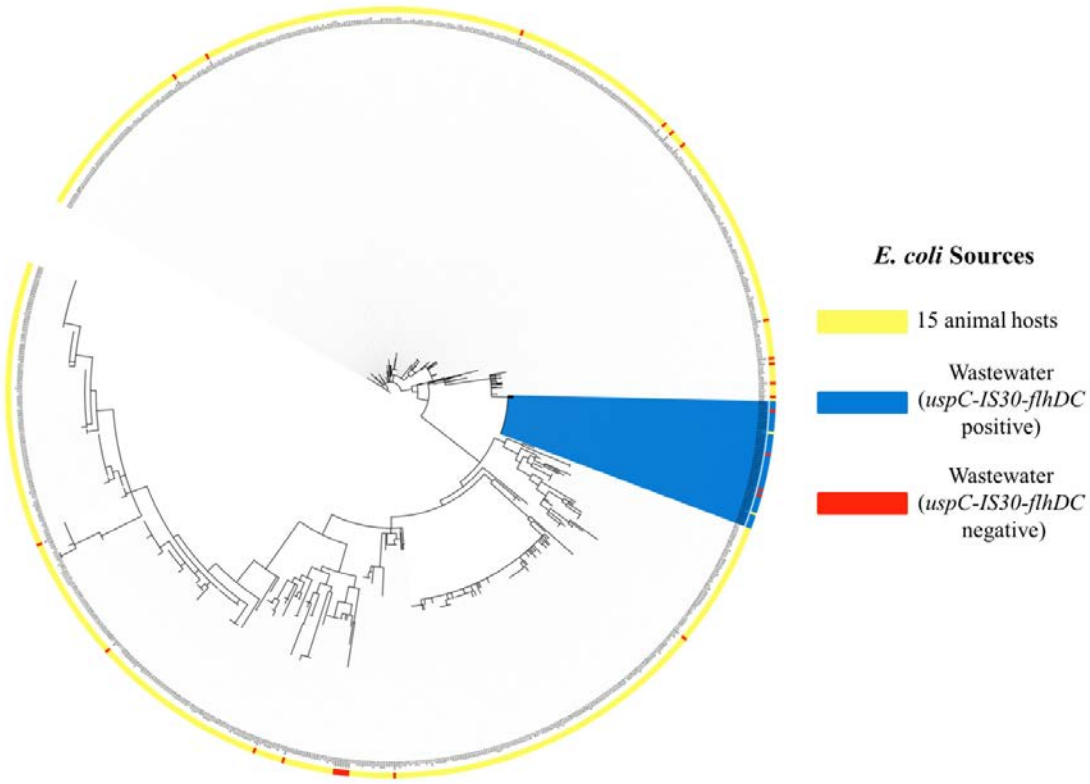

Figure 4-1. An unrooted, maximum likelihood phylogenetic tree encompassing 848 *E. coli* strains obtained from chlorine-treated wastewater (68 isolates) and 15 animal host groups (780 isolates, see Table 4-5), and based on on an analysis of the concantenated DNA sequences of two intergenic regions (*csgBAC-csgDEFG* and *asnS-ompF*). Of the 70 chlorine-tolerant wastewater *E. coli*, two failed to produce PCR products for *csgBAC-csgDEFG.* The colored ring that overlays the unrooted maximum-likelihood tree indicates different *E. coli* groups, as defined by: i) human/animal host sources (yellow), ii) wastewater isolates possessing the *uspC-IS30-flhDC* marker (blue), and iii) wastewater isolates that do not possess the *uspC-IS30-flhDC* marker (red). Note that all chlorine-tolerant wastewater isolates that possess the *uspC-IS30-flhDC* marker (and which were isolated from 4 different WWTP) group together within a single clade.

4.4.2 Determination of RpoS Stress response activity and presence of a heat resistance genomic island in wastewater *E. coli*

Experiment was performed to determine whether the chlorine-tolerant wastewater *E. coli* isolates possessing the *uspC-IS30-flhDC* marker also possessed the phenotypic properties of a generalized stress response (RpoS), important for environmental survival. The rationale was to determine whether the unique genetic backgrounds of these isolates correlated with known adaptive survival strategies for environmental persistence in *E. coli.* All 41 wastewater isolates possessing the *uspC-IS30-flhDC* marker were positive for both glycogen production and catalase activity (Figure 4-2), suggesting that these strains were phenotypically-adapted to survive in the environment.

PCR amplifications of three genetic fragments from a heat-resistance genomic island were performed on all 70 wastewater *E. coli* isolates. All 41 *uspC-IS30-flhDC* positive *E. coli* isolates were positive for the three PCR fragments (A, B, and C) typically found in this large genomic island.

4.4.3 Prevalence of virulence genes in wastewater *E. coli*

Ten *uspC-IS30-flhDC* marker positive wastewater *E. coli* strains were analyzed for the presence of 28 virulence genes. Twenty-seven of the 28 virulence genes were not found in any of these wastewater *E. coli* strains, the exception being for *fimH* gene, in which all 10 isolates were positive for *fimH* - an adhesive protein localized at the tip of type 1 fimbriae.

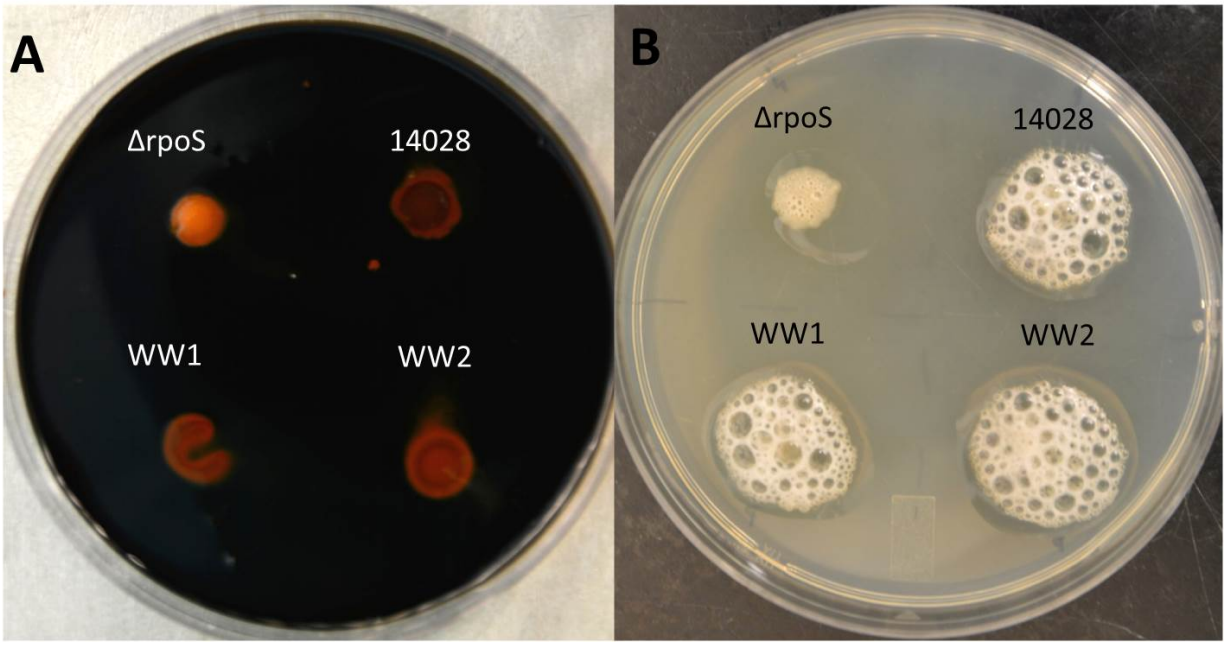

Figure 4-2. Glycogen (Panel A) and catalase activity (Panel B) of *E. coli* strains on LB agar. WW1 and WW2 are *E. coli* strains isolated from wastewater and possess the *uspC-IS30-flhDC* marker. The strain labelled14028 is a RpoS positive strain (*S. Typhimurium* ATCC 14028) and *ΔrpoS* is a negative strain (*S. Typhimurium* Δ*rpoS*). In Panel A, strains with dark brown color are considered positive for glycogen. In Panel B, colonies with vigorous bubbles were considered as positive for catalase activity.

4.4.4 Evaluation of a PCR targeting the *uspC-IS30-flhDC* region in chlorine-tolerant wastewater *E. coli*

The high prevalence of the *uspC-IS30-flhDC* intergenic signature in chlorine-tolerant wastewater *E. coli*, and its apparent absence in the genomes of *E. coli* isolated from other human, animal and even non-host environmental sources (i.e., surface water), suggested that this sequence may be useful as a genetic marker of municipal wastewater contamination in the environment. Consequently, a sensitive and specific endpoint PCR for detection of these

wastewater *E. coli* strains was developed. The PCR was designed to target the site-specific location of the IS30 element in the *uspC-flhDC* intergenic region. The forward primer was identical to the primer used to amplify the *uspC-flhDC* intergenic region, but the reverse primer targeted the IS30 sequence (Table 4-3). The position-specific PCR produced an amplicon of 386 bp. The LOD_{95} for the optimized PCR was 13 copies (95% Confidence Interval [CI_{95}] of 6-28 copies) based on cloned plasmid DNA templates, while for genomic DNA the LOD_{95} of the PCR was 19 copies (with a CI_{95} of 9-42 copies).

Table 4-6. Prevalence of the *uspC-IS30-flhDC* marker in *E. coli*-positive surface water, drinking water and wastewater samples.

***E. coli*-positive water sample from various sources**	**No. of Samples**	**Marker Positive Samples (%)**
Wastewater (Total)	50	43 (92%)
• Untreated wastewater	21	21 (100%)
• Secondary-treated wastewater	17	16 (94%)
• UV-treated wastewater	12	9 (75%)
Surface water	71	2 (3%)
Drinking water	57	3 (5%)

Because the previous work in Chapter Two confirmed that the 845 animal/human *E. coli* isolates used in this study did not possess the IS30 element inside the *uspC-flhDC* intergenic sequence, specificity of the PCR was only tested against a random selection of 90 *E. coli* isolates.

The rationale for testing specificity at this level was based on the fact that, although the IS30 element may exist in human and animal *E. coli* isolates in different areas of the genome, it is important to determine whether its specific location in the *uspC -flhDC* intergenic region was unique to chlorine-tolerant wastewater *E. co*li. The 386 bp PCR amplicon was not observed from any of the other human and animal *E. coli* isolates tested, whereas all of the 41 chlorine-tolerant wastewater *E. co*li isolates were positive by this endpoint PCR.

4.4.5 Prevalence of the *uspC-IS30-flhDC* marker in environmental water samples

The *uspC-IS30-flhDC* specific PCR was subsequently tested against environmental water samples in order to verify the presence of this marker in environmental wastewater samples, and evaluate the potential applicability of the *uspC-IS30-flhDC* PCR for routine public health screening of water supplies. The specificity and prevalence of this marker was evaluated in wastewater (treated and untreated [City of Calgary WWTPs]), groundwater, and surface water samples (rivers, lakes, etc.). Samples that were culture-positive for *E. coli* (based on presence/absence testing in Colilert®) were screened for *uspC-IS30-flhDC* marker by PCR. The *uspC-IS30-flhDC* marker was found in 92% of 50 wastewater samples tested (untreated or treated [Table 4-6 and Figure 4-3]), indicating a high prevalence of the strains in *E. coli* populations found in wastewater. All untreated wastewater samples from the City of Calgary's two WWTPs (n=21 [100%]), as well as most secondary treated samples (16/17 [94%]) and UV-treated wastewater samples (9/12 [75%]) were positive for *E. coli* strains carrying the *uspC-IS30-flhDC* marker. Among 57 *E. coli* culture-positive groundwater samples by Colilert®, only 3 samples (5%) were positive for the *uspC-IS30-flhDC* marker; while of the 71 surface water samples that were positive for *E. coli* by culture in Colilert®, only 2 samples (3%) carried the

this marker. The high prevalence of this marker in *E. coli* populations in wastewater, compared to those populations in contaminated groundwater and surface water, demonstrates that the *uspC-IS30-flhDC* marker appears to be specific to human municipal wastewater contamination.

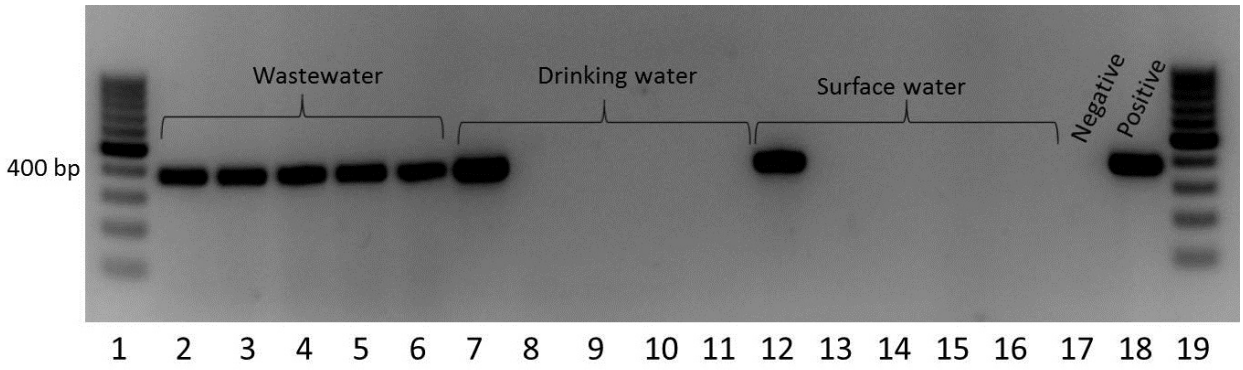

Figure 4-3. PCR amplification of the *uspC-IS30-flhDC* marker in Colilert® enriched *E. coli* positive wastewater, drinking water, and surface water samples. Lanes 1 and 19: molecular weight marker (GeneRuler 100 bp, Thermo Scientific). Lanes 2-16 represent the PCR results of individual water samples. Lanes 2-6: wastewater; lanes 7-11: drinking water; lanes 12-16: surface water. Lane 17 is a positive control using genomic DNA from an *E. coli* isolate possessing the *uspC-IS30-flhDC* marker. Lane 18 is a negative control using genomic DNA from an *E. coli* isolate that did not possess the *uspC-IS30-flhDC* marker. The PCR amplicon targeting the *uspC-IS30-flhDC* marker is 386 bp in length.

4.5 Discussion

E. coli, although defined as a single species, shares only 10% of its pangenome among individual members (Lukjancenko *et al.,* 2010). This genetic diversity, largely explained by an abundance of accessory genes, may account for their ability to survive and adapt to various niches including a diverse array of gastrointestinal microenvironments from warm-blooded animals as well as the non-host environment. Naturalized strains of *E. coli* have been identified

in various water sources (Tymensen *et al.,* 2015; Kon *et al.,* 2007; Chandrasekaran *et al.,* 2015), but to date, evidence for the presence of naturalized strains of *E. coli* in a wastewater environment has not been reported. It was hypothesized that the intrinsic selection pressures imposed by wastewater treatment will likely drive adaptation of certain strains of *E. coli* towards survival in wastewater and lead to the evolutionary selection of naturalized endogenous populations of this bacteria in wastewater matrices, and for which these strains may be genetically different than those found in human and animal hosts.

Evaluating the biological relevance (i.e., phenotypic host-specificity) of ITGR biomarkers in humans and animals is a complex question. Phenotypic verification of host specificity based on biomarker SNP patterns in ITGRs would require cross-infectivity studies between conspecific and xenospecific recipient hosts, and for which individual host-specific strains of *E. coli* strains could be tracked throughout the GI tract of these animals. In order to evaluate this, *E. coli* populations surviving typical wastewater treatment stressors (i.e., chlorination) were examined based on the biomarker searching approaches described in Chapters Two and Three. Sewage samples were collected from geographically-segregated WWTP in Alberta, Canada, and subjected to the U.S. EPA ATP procedures for evaluating recovery of chlorine-stressed organisms in selective media. Of the 70 *E. coli* isolates collected from chlorine-treated sewage in the laboratory (i.e., deemed chlorine-tolerant), 59% of the *E. coli* isolates possessed an IS30 insertion element within the intergenic region between the *uspC* and *flhDC* genes. The IS30 element, a 1.2 kb long DNA fragment typically found in various locations across the *E. coli* genome, was originally described by Caspers *et al.* (1984), but to date, has not been reported within the intergenic region of the *uspC* and *flhDC* genes. The specific location of the IS30 element in the *uspC-flhDC* intergenic region was unique to these chlorine-tolerant

wastewater *E. coli*, the positional location of the insertion sequence was not observed in a library of 845 *E. coli* isolates from 15 different animal species (including humans), nor in any GenBank submitted sequences, nor across 1177 *E. coli* genomes for which whole genome sequences were available from NCBI (1107 human, 41 bovine, and 29 other animals).

To further evaluate their genetic uniqueness, *E. coli* strains were subject to phylogrouping, phylogenetic analysis and host-specific biomarker analysis (Zhi *et al.,* 2015). All wastewater chlorine-tolerant *E. coli* isolates possessing the *uspC-IS30-flhDC* belonged to phylogroup A. In a separate study (Walk *et al.,* 2007), multilocus sequence typing (MLST) was used to identify a specific sequence type that was over-represented by *E. coli* strains isolated from the non-host environment. However, in contrast to the findings of this study, these strains were assigned to phylogroup B1. Other studies have also identified the B1 phylogroup as being more prevalent from isolates collected from water (Ratajczak *et al.,* 2010; Berthe *et al.,* 2013). However, it is interesting to note that isolates collected in these studies were obtained from surface water (i.e., rivers, lakes) as opposed to wastewater. In this study, only 1 of 187 *E. coli* isolates obtained from surface water sources possessed the *uspC-IS30-flhDC* marker. Even in this instance, it is possible that this isolate originated from wastewater as it was collected downstream from a municipal wastewater treatment plant. Indeed, in agreement with this study, *E. coli* from phylogroup A was found to be more abundant in wastewater in other studies (Figueira *et al.,* 2011; Sabate *et al.,* 2008).

In order to address whether these chlorine-tolerant strains may be genetically unique, maximum likelihood phylogenetic analysis and SNP biomarker analysis was carried out using DNA sequences at two alternate intergenic regions (*csgBAC-csgDEFG* and *asnS*-ompF) and unrelated to the *uspC* and *flhDC* genes, and compared across 780 human/animal *E. coli* isolates.

Based on the previous studies (Chapter Two and Three) examining these intergenic loci (Zhi *et al.,* 2015), traditional phylogenetic approaches failed to reveal host-specific clustering of animal and human *E. coli* strains, whereas biomarker analysis demonstrated a clear delineation in host-specific DNA biomarker patterns. Chlorine-tolerant wastewater *E.* coli isolates possessing the *uspC-IS30-flhDC* formed a single clade by traditional bioinformatics that was distinct from human and animal isolates, suggesting a common genetic background in these strains. SNP biomarker analysis at this same intergenic loci provided further evidence that these *E. coli* were genetically unique. Logic regression-based biomarker analysis revealed that 82% of all chlorine-tolerant wastewater *E. coli* isolates possessed a SNP biomarker that was highly specific to wastewater and not found in *E. coli* isolates collected from 15 different animals species, including humans. In addition, all of the 41 chlorine-tolerant wastewater isolates possessing the IS30 element in the *uspC-flhDC* also carried this wastewater-specific SNP biomarker in the *csgBAC-csgDEFG* / *asnS-ompF* intergenic region.

Of particular importance was the finding that *uspC-IS30-flhDC* chlorine-tolerant *E. coli* strains were observed in a number of geographically separated WWTP plants in Alberta. In fact, all wastewater treatment plants that have been tested to date in the province possess *E. coli* strains containing the *uspC-IS30-flhDC* marker, suggesting a wide spread dispersion of these strains, and suggesting a functional role of the site-specific insertion of the IS30 element in this intergenic region and which seems to favor bacterial survival and persistence in a broad spectrum of municipal wastewater environments.

Phenotypically, all chlorine-tolerant strains possessing the *uspC-IS30-flhDC* marker are also positive for RpoS generalized stress response activity. The transcriptional regulator, σ^S, is a key regulator of the RpoS stress response in *E. coli* initiating transcription of the genes essential

for stress resistance (Battesti *et al.,* 2011; Lacour and Landini, 2004; Weber *et al.,* 2005). A resistance phenotype regulated by σ^S is the rdar phenotype (red, dry and rough when grown on medium containing the dye Congo red) enhancing long-term survival under harsh conditions (White and Surette, 2006). Rdar cells secrete an extracellular matrix comprised of curli fimbriae, cellulose and polysaccharides (Zogaj *et al.,* 2001; Romling *et al.,* 1998) and for which the matrix provides increased resistance to disinfection (Ryu and Beuchat, 2005; Uhlich *et al.,* 2006). Chiang *et al.* (2011) observed that *E. coli* strains found in wastewater had the largest percentage of RpoS positive *E. coli* compared to strains found in natural water, beach sand and animal feces. In addition, a heat resistance genomic island was found in all *uspC-IS30-flhDC* positive *E. coli* strains. This genomic island in *E. coli*, also called the locus of heat resistance (LHR), was reported in a recent study by Mercer *et al.* (2015), and contains 16 open reading frames which are predicted to code for proteins that are associated with heat shock, cell envelop maintenance, and turnover of misfolded proteins (Mercer *et al.,* 2015). *E. coli* strains possessing the LHR have been shown to survive heat shock temperatures of 60^{o}C for longer than 5 minutes (Mercer *et al.,* 2015). The LHR in *E. coli* is flanked by transposable elements, suggesting genetic mobility of the LHR among *E. coli* strains and raising concerns for heat-resistant *E. coli* in food safety (Mercer *et al.,* 2015). Interestingly, this LHR in *E. coli* is homologous to a heat resistance plasmid observed in nosocomial pathogenic strains of *Klebsiella pneumoniae* (Bojer *et al.,* 2010). Bojer *et al.* (2010 and 2012) have demonstrated the conjugative transmissibility of the plasmid-mediated heat-resistance phenotype to naïve strains, and the co-localization of heat resistance with antibiotic-resistance on these plasmids. Moreover, biofilm producing *K. pneumoniae* have enhanced heat resistance compared to planktonic cells in cells carrying the heat resistance plasmid, suggesting that biolfilm formation and heat-resistance promote environmental stability

of these nosocomially-persistent clones (Bojer *et al.,* 2011). It is extremely interesting to note that all chlorine-tolerant *E. coli* carrying the *uspC-IS30-flhDC* locus also possessed the LHR. The fact that wastewater temperatures in Alberta fluctuate between 4^{o}C and 20^{o}C suggests that the LHR, and the proteins encoded therein, may have a broader spectrum of activity in promoting environmental resistance to more than just heat stress. This heat resistance characteristic in combination with positive RpoS stress response activity may contribute to the survival of these wastewater strains in the harsh environment.

The IS30 element, as well as other insertional elements, has been previously found to alter gene expression and consequently bacterial phenotypes (Neuwald and Stauffer, 1990; Dalrymple, 1987). In one study, insertion of an IS element appeared to reduce transcriptional repression resulting in increased expression of the *flhD* operon (Barker *et al.,* 2004). It was also noted that several IS elements were located upstream of the *flhD* promoter in *E. coli* isolates that had a high swarming rate, while no IS elements were found in strains with poor motility (Barker *et al.,* 2004). In turn, motility has been found to be associated with biofilm formation (O'Toole and Kolter, 1998). Wood *et al.* (2006) compared biofilm formation and motility of several *E. coli* strains, the result of which demonstrated that strains with the greatest motility also had the best biofilm formation capacity. Formation of a biofilm is a strategy often used by bacteria for survival in harsh environments (Ryu and Beuchat, 2005; Uhlich *et al.,* 2006). Consequently, it is hypothesized that the site-specific insertion of the IS30 element in the *uspC-flhDC* may alter flagellar expression and consequently the motility and biofilm-forming capacity of these strains, ultimately leading to treatment resistance and environmental persistence of these strains in a wastewater environment, and potentially complementing the RpoS generalized stress response and environmental resistance phenotypes encoded by the LHR. At present, there is no function-

based evidence that a casual relationship exists between carriage of the *uspC-IS30-flhDC* genetic marker and enhanced survival of *E. coli* under stressed conditions.

Interestingly, *uspC* and *flhDC* are divergently transcribed across the *E. coli* chromosome, and consequently the IS30 element may also play a role in regulating *uspC* transcription (also known as *yecG*). The universal stress protein (usp) superfamily, for which 6 proteins exist in *E. coli* (usp A, C, D, E, F and G), is involved in cellular stress responses to osmotic, oxidative, antibiotic and UV-induced stressors (Gustavsson *et al.,* 2002; Nachin *et al.,* 2005). More specifically, the *uspC* gene product has been shown to enhance flagellar production and motility in *E. coli* (Nachin *et al.,* 2005). Although other *usp* gene products have been shown to be important in dealing with oxidative stressors such as hydrogen peroxide and superoxide (and were not specifically examined in this study), the *uspC* gene has also been shown to be important in enhancing UV-resistance in *E. coli* (Gustavsson *et al.,* 2002). Both flagellar and universal stress response genes are known to be important virulence factors for pathogenic bacteria (Badea *et al.,* 2009; Yang *et al.,* 2014b). For example, mice infected with isogenic *uspA* mutants of *Salmonella typhimurium* have better survival outcomes (i.e., delayed death) than those infected with a pathogenic parent strain of the bacteria (Liu *et al.,* 2007). Thus, in addition to possessing the RpoS and LHR stress-adaptive phenotypes/genotypes, these naturalized strains of *E. coli* in wastewater may also have an enhanced universal stress response.

It is important to note that wastewater treatment processes at the City of Calgary's WWTPs incorporate oxidative aerobic digestion, biological nutrient removal and UV-treatment, but not chlorine treatment. It is hypothesized that these chlorine-tolerant, wastewater *E. coli* strains possess an array of stress-resistance determinants important for their overall survival in wastewater, irrespective of the specific wastewater treatment algorithms used at the WWTP, and

for which these strains become widely distributed throughout various municipal WWTPs. A common mechanism of resistance in these strains may center around biofilm formation, coupled with an enhanced generalized stress response (RpoS), universal stress protein upregulation (usp) and heat-shock stress response (LHR).

Only one virulence gene, *fimH*, out of 28 virulence genes surveyed, was found in the wastewater *E. coli* isolates possessing the *uspC-IS30-flhDC* marker. It is not surprising that naturalized strains of *E. coli* in wastewater lack virulence determinants, since virulence genes are, essentially, host-adaption genes – i.e., genes important in adherence, colonization, invasion, intracellular survival, and/or toxin production in a particular host. Since wastewater represents a very unique environment compared to the gastrointestinal system of an animal, these virulence genes may not be necessary for their survival in this matrix. It has been reported that type 1 fimbriae are critical for biofilm formation in *E. coli* (Rodrigues and Elimelech, 2009; Vila and Soto, 2012), and important for bacterial survival in harsh environments. The *fimH* gene encodes an adhesive protein localized at the tip of type 1 fimbriae, and is considered a virulence gene for uropathogenic *E. coli* (UPEC), as it has been demonstrated to be important for colonizing the urinary tract (Connell *et al.,* 1996). However, the absence of other UPEC-related virulence factors (Wang *et al.,* 2009) in these wastewater strains, such as uropathogenic-specific protein (*usp*), PapG adhesin genes (*papG* class I–III) of P-fimbriae, P fimbriae (*papC*), S and F1C fimbriae (*sfa/foc*), hemolysin (*hlyA*), iron-regulated gene A homologue adhesin (*iha*), cytotoxic necrotizing factor 1 (*cnf1*), catecholate siderophore receptor (*iroN*), and aerobactin receptor (*iutA*), in the wastewater *E. coli* isolates suggests that these strains are not likely UPEC. However, these strains are not the only *E. coli* present in finished wastewater, and other research has demonstrated that *E. coli* strains from these matrices can carry a high frequency of virulence

genes (Hamelin *et al.,* 2007). Therefore the absence of most virulence genes in these strains cannot exclude the possibility that virulence genes can be acquired by these strains through transduction or horizontal gene transfer, or that these strains may act as reservoirs of environmental and treatment-resistance determinants (i.e., heat-resistant genomic islands) for pathogenic strains (Dobrindt, 2005).

Wastewater treatment processes mimic the osmotic (dilution of feces in water), oxidative (aerobic digestion and chlorination) and UV (medium or low pressure UV treatment)-associated challenges commonly encountered by bacteria in the non-host environment. The data of this study provides credence to a hypothesis that strains of *E. coli* may have evolved over time to enhance their survival across the wastewater treatment process and that these strains appear to be genetically distinct from most human and animal isolates. Further studies are required to determine: a) what mechanistic role the site-specific insertion of the IS30 element has on phenotypic resistance in *E. coli* by comparing survival rates in response to osmotic, oxidative and/or UV stress; b) whether regulatory anomolies exists in other *usp* intergenic regions (A, D, E, F, G) or stress response elements in these strains, and what role they play in enhancing treatment-resistance/ environmental-persistence; c) whether these naturalized genetic strains can acquire virulence properties through transformation, transduction, or horizontal gene transfer in a wastewater matrix (or whether they can act as reservoirs for treatment-resistant genes to be transmitted to pathogenic strains); and d) the role of the RpoS and LHR for survival in a wastewater matrix.

This study raises some potentially important public health concerns about the evolution of environmentally-persistent/ treatment-resistant resident populations *E. coli* in wastewater. Firstly, there is concern that treatment-resistant bacteria from upstream wastewater treatment

plants may pose a risk to downstream drinking water utilities. Secondly, standards for wastewater treatment or reuse of wastewater warrant a careful consideration of the potential public health risks that may be associated with treatment-resistant or environmentally-persistent bacteria present in effluents. It is generally assumed that the *E. coli* populations present in raw sewage are homogeneous in terms of their ability to survive the wastewater treatment processes. Consequently, and for simplicity from a regulatory perspective, the *E. coli* entering a wastewater treatment plant are considered biologically similar to the *E. coli* exiting the plant after treatment, and WWTP performance is determined based on bacterial reduction across the treatment chain. However, the data of this study suggests that naturalized *E. coli* populations exist in wastewater and that they possess environmental/ treatment resistance properties that may allow them to preferentially survive the treatment process. Consequently, the influent population of *E. coli* is likely to be very different from the *E. coli* population exiting a WWTP, raising concerns about *E. coli* as a suitable indicator of WWTP performance. Moreover, if these stress-tolerant *E. coli* acquire virulence genes or act as reservoirs for treatment resistance, then the human health risks associated with WWTP effluent exposure (i.e., reuse) may be underestimated. Based on the observation that virulence markers are common in *E. coli* from wastewater effluents (Hamelin *et al.,* 2007) it will be interesting to determine whether pathogenic strains exiting a treatment plant also possess a sophisticated stress response that enhances their survival through the treatment train.

The genetic uniqueness of the *uspC-IS30-flhDC E. coli* strains identified in the present study allowed for development of a sensitive end-point PCR targeting the site-specific locations of the IS30 element in the *uspC-flhDC* intergenic region, and representing a potential novel marker of municipal wastewater contamination. At an *E. coli* population level, all raw

wastewater samples were shown to be positive for the *uspC-IS30-flhDC* PCR marker, but only a proportion of the *E. coli* isolates present in untreated wastewater were shown carry this marker (i.e., 5%). When raw wastewater was treated with chlorine, the prevalence of the *uspC-IS30-flhDC* marker in the surviving population jumped to 59%. The data suggests that a significant portion of the culturable *E. coli* biomass entering a WWTP is comprised of these novel, stress-resistant strains, and that they persist throughout the treatment stream (i.e., even after chlorine or UV treatment). It has been previously reported by Anastasi and colleagues (Anastasi *et al.,* 2013) that within the raw sewage influent only a portion of the *E. coli* population survive the treatment process when either UV or chlorine are used for disinfection. Future study needs to examine whether these strains differentially survive across the wastewater treatment process *in situ*.

The description of a novel PCR specifically targeting a chlorine-tolerant wastewater *E. coli* sub-population provides an interesting opportunity for adapting the assay to tracking sources of human municipal wastewater pollution in water samples routinely submitted for public health purposes. In this study, 1 ml aliquots of Colilert® positive *E. coli* samples were centrifuged, lysed by boiling, re-centrifuged, and the supernatant analyzed for the presence of the *uspC-IS30-flhDC* marker by PCR. All raw wastewater samples were positive, whereas 94% of secondary-treated wastewater samples and 75% of *E. coli* culture positive UV-treated samples were also positive for the *uspC-IS30-flhDC* marker. Conversely, only 2/71 and 3/57 of *E. coli* culturable surface and groundwater samples, respectively, were positive for the *uspC-IS30-flhDC* marker. In cases where the *uspC-IS30-flhDC* marker was observed in surface water it is possible that human wastewater may have impacted these sources, as collection sites were known to be downstream of wastewater treatment plants within the watershed. In the case of *E. coli*-contaminated groundwater samples, it is possible that septic field discharges may have impacted

groundwater sources in these samples. The advantage of using the *uspC-IS30-flhDC* marker for evaluating whether a water source is impacted by human municipal wastewater (over and above other source tracking tools such as *Bacteroides*) is that it can be immediately applied to routine water samples already positive for *E. coli* (i.e., Colilert® positive samples). This alleviates the need for duplicate processing of samples, and focuses the testing on only those samples positive for *E. coli*.

4.6 Conclusion

In conclusion, the findings described herein suggest that some *E. coli* strains may have adapted themselves to survive and grow within a wastewater matrix and that these populations compromise a significant proportion of the *E. coli* biomass entering a WWTP. The data also suggest that the PCR methods developed in this study may be useful as an *E. coli*-based biomarker of wastewater contamination in the environment. The findings consequently raise some important considerations about the utility of *E. coli* as an indicator of wastewater treatment performance and impact our understanding of adaptation and evolution in this bacterial species.

Chapter Five : EVIDENCE OF ADAPTIVE SPECIFICITY OF NATURALIZED WASTEWATER *E.COLI* BASED ON OCCURRENCE AND PERSISTENCE IN MUNICIPAL WASTEWATER[4]

5.1 Abstact

As demonstrated in Chapter Four, naturalized strains of *E. coli* exist in municipal sewage, possessing: a) SNP biomarker patterns distinct from those observed in humans and animals, and b) a transposable element (IS30) located specifically in the *uspC-flhDC* intergenic region (*uspC-IS30-flhDC*) of the genome. The present Chapter aims to assess the prevalence/occurrence of these naturalized strains in wastewater and compare their occurrence against other tracers of human sewage pollution, as a means of better understanding the phenotypic specificity of these strains to a non-host wastewater niche. A quantitative PCR (qPCR) assay targeting the *uspC-IS30-flhDC* marker was developed and the strains were routinely detected by qPCR throughout the wastewater treatment process, including treated effluents. Correlations between the *uspC-IS30-flhDC* marker and human fecal/sewage markers (HF183 and HumM2) and animal fecal markers (bovine [CowM3], seagull [LeeSG], and Canada goose [COG1]) were evaluated in surface water and storm water samples, and a significant correlation was observed between the *uspC-IS30-flhDC* marker and human-associated HF183 and HumM2 markers, but not with the animal markers. Seventeen of ninety-three surface/storm water samples possessed the *uspC-IS30-flhDC E. coli* marker, and of these 16 and 15 also contained HF183 and HumM2 markers,

[4] A version of this chapter has been submitted for publication, the citation of which is: Zhi, S., G. Banting, N. J. Ruecker, and **N. F. Neumann**. 2016. Characterization of municipal wastewater contamination in the environment using a novel *E. coli*-based source-tracking marker. *Environmental Science & Technology* (submitted September 2016)

respectively. Although shown to be less prevalent in water samples, the overall strength of the correlation to *uspC-IS30-flhDC* increased as HF183 and HumM2 concentrations increased. The advantage of the *uspC-IS30-flhDC* marker is the direct applicability to cultured *E. coli* samples (i.e., Colilert®), making this a potentially valuable screening tool for detecting municipal sewage pollution in water samples routinely submitted for public health screening for bacteriological water quality.

5.2 Introduction

The occurrence of niche-specific strains of *E. coli* in a variety of human, animal and wastewater *E. coli* populations was characterized using novel supervised learning approaches (logic regression) to bioinformatic analysis of intergenic DNA sequences, as described in Chapters Two, Three and Four of this thesis. Chapter Four described the existence of naturalized strains of *E. coli* in municipal wastewater, and which possessed a unique genetic signature in the form of an insertion sequence (IS30) specifically located in the *uspC-flhDC* intergenic region (*uspC-IS30-flhDC*) of the genome. This genetic signature was not observed in: a) *E. coli* strain libraries originating from a wide range of animals, b) whole genome sequence databases for *E. coli* originating from human or animals, or c) in any *E. coli* DNA sequence data represented in Genbank.

To date, these naturalized wastewater strains have been found in all wastewater treatment plants (WWTP) tested in Alberta, suggesting that they have a widespread geographic distribution. Their unique genetic background coupled with widespread occurrence, suggests that variations in the ITGRs may be adaptive to survival of these strains in a wastewater niche. Occurrence of these naturalized strains was evaluated across the wastewater treatment plant as a means of better

understanding phenotypic adaptation related to survival in these matrices. As such their occurrence was assessed across two wastewater treatment plants using a quantitative polymerase chain reaction (qPCR) assay targeting the *uspC-IS30-flhDC* biomarker. This *E. coli*-based marker was subsequently compared against the human *Bacteroides* source-tracking markers HF183 (Harwood *et al.,* 2014) and HumM2 (Harwood *et al.,* 2014), in addition to other animal fecal source-tracking markers (cattle, seagulls, and Canada goose) in surface/storm water sources as an evaluation of specificity of this marker to wastewater and wastewater contamination in the environment.

5.3 Methods

5.3.1 Development of a qPCR for detection of *uspC-IS30-flh* carrying *E. coli*

A TaqMan® qPCR assay was developed to detect the *uspC-IS30-flhDC* biomarker in naturalized wastewater strains of *E. coli*. Primers and probes for the *uspC-IS30-flhDC* marker are shown in Table 5-1. All reactions were performed in a 20 µl volume containing 5 µl of DNA template, 10 µl of TaqMan® Fast Advanced Master Mix (Applied Biosystems, Foster City, CA), 900 nM of each primer, and 250 nM of TaqMan probe. The cycling conditions were 95°C for 30 seconds, 40 cycles of 95 °C for 3 seconds and 60°C for 30 seconds.

To test qPCR sensitivity, a plasmid containing the qPCR target region was constructed. Specifically, genomic DNA from a naturalized wastewater *E. coli* isolate containing the *uspC-IS30-flhDC* marker was used as a DNA template, and PCR was used to amplify the *uspC-IS30-flhDC* intergenic region (primer set ZIS-F and ZIS-R in Table 5-1). The PCR product was resolved on a 2% agarose gel in 1X TAE buffer (Promega, Madison, Wisconsin) at 140V for 45 minutes and purified from the gel using a QIAquick Gel Extraction Kit (QIAGEN, Inc., Valencia,

CA). The amplicon was cloned using a TOPO® TA Cloning Kit according to manufacturer's instructions (Invitrogen, Inc., Carlsbad, CA). Isolation of recombinant plasmid DNA was performed using QIAprep Miniprep Kit (QIAGEN, Inc., Valencia, CA). The presence of the correct insert was confirmed by PCR screening and DNA sequencing of the cloned inserts. DNA sequencing was performed by the Applied Genomic Core (TAGC) facilities at the University of Alberta, Edmonton, Canada.

qPCR sensitivity was tested using cloned plasmid DNA constructs. The concentration of plasmid was quantified using a Qubit® 2.0 Fluorometer (Invitrogen, Carlsbad, CA). Plasmid copy number was determined from cloned targets. Ten-fold serial dilutions, in replicates of 12, of plasmid DNA were made and the *uspC-IS30-flhDC* marker amplified by qPCR. The 10-fold serial dilutions of plasmid ranged from 100,000 copies/ µl to 0.1 copies/ µl. The limit of detection (with 95% confidence intervals [LOD_{95}]) was calculated using a program based on Microsoft Office Excel (Wilrich and Wilrich, 2009). The specificity of the qPCR was tested against 41 *uspC-IS30-flh* positive and 29 *uspC-IS30-flh* negative *E. coli* strains, which had been confirmed by sequencing of the intergenic *uspC-flhDC* region in Chapter Four.

5.3.2 Proportion of naturalized *E. coli* in sewage and wastewater

To better understand occurrence of the naturalized populations of *E. coli* possessing the *uspC-IS30-flhDC* marker in municipal WWTPs, water samples were collected from two different WWTPs in Calgary, Alberta, Canada, at different points along the treatment train with an attempt to determine what proportion of the *E. coli* population in sewage and wastewater was comprised of these naturalized populations.

In total, 12 wastewater samples were collected from the two WWTPs by sampling three times at each plant between February and May 2015. At each sampling time, samples were collected after grit removal (post-grit) and after primary treatment (primary effluent). Twenty mL of post-grit and 200 mL of primary effluent were centrifuged at 8500 × *g* for 10 minutes to collect the bacterial pellets. The DNA was extracted from the pellet using a PowerWater® DNA Isolation Kit (Mo Bio Inc., Carlsbad, CA) according to the manufacturer's instructions. The final DNA extract volume was 100 µl.

An estimation of population prevalence for naturalized strains of *E. coli* in wastewater was assessed by comparing the genomic concentrations of *uspC-IS30-flhDC* (naturalized wastewater *E*. coli population) to *uidA* gene (a universal *E. coli* marker [(Yang *et al.,* 2014a)]) in post-grit and primary effluent samples. Plasmids containing the *uidA* gene target used in this study were constructed in a similar fashion as described for the *uspC-IS30-flhDC* plasmid. The qPCR methods and conditions for the *uidA* gene were the same as those used for the *uspC-IS30-flhDC* qPCR assay, with the primers and probes used for *uidA* shown in Table 5-1. Standard curves were constructed for each qPCR assay to determine the quantity of the two markers in each sample. The estimated genomic copies for *uspC-IS30-flhDC* were divided by the estimated genomic copies of the *uidA* gene to determine the population prevalence of naturalized *E. coli* in each wastewater sample.

Most probable number (MPN) *E. coli* concentrations were also determined for wastewater samples collected post-grit removal and after primary, secondary and ultraviolet (UV) treatment, and at 5 different times for the Bonnybrook WWTP and 6 different times for the Pine Creek WWTP. Serial 1:10 dilutions of wastewater samples were made using sterile buffered water and 100 mL of the samples at different dilutions were poured into a Colilert® vessel. The

Table 5-1. PCR primers and probes used in this study

Target	Primers/probes	Primer and probe sequence (5'-3')	Reference
HF183	HF183-F	ATCATGAGTTCACATGTCCG	(Haugland *et al.,* 2010)
	HF183-R	CGTAGGAGTTTGGACCGTGT	
	HF183-P	**FAM-** CTGAGAGGAAGGTCCCCCACATTGGA- **TAMRA**	
HumM2	HumM2-F	CGTCAGGTTTGTTTCGGTATTG	(Shanks *et al.,* 2009)
	HumM2-R	TCATCACGTAACTTATTTATATGCATTAGC	
	HumM2-P	**FAM-**TATCGAAAATCTCACGGATTAACTCTTGTGTACGC-**TAMRA**	
uspC-IS30-flhDC	ZIS-F	CAGACCGAGAAAGACACTGAA	This study
	ZIS-R	TGTACTGCATTCCCCGTATT	
	ZIS-P	**FAM**-TTGAAAGGGGTGTTGCATTGACAGA-**TAMRA**	
uidA	uidA-F	CGCAAGGTGCACGGGAATA	This study
	uidA-R	CAGGCACAGCACATCAAAGAGA	
	uidA-P	**FAM**-ACCCGACGCGTCCGATCACCT-**NFQMGB**	
CGO1	CGO1-F	GTAGGCCGTGTTTTAAGTCAGC	(Fremaux *et al.,* 2010)
	CGO1-R	AGTTCCGCCTGCCTTGTCTA	
	CGO1-P	**FAM**-CCGTGCCGTTATACTGAGACACTTGAG-**TAMRA**	
CowM3	CowM3-F	CCTCTAATGGAAAATGGATGGTATCT	(Shanks *et al.,* 2008)
	CowM3-R	CCATACTTCGCCTGCTAATACCTT	
	CowM3-P2	**FAM-** GGAAAGCAGGAACTTA-**NFQMGB**	This study
LeeSG	LeeSG-F	AGGTGCTAATACCGCATAATACAGAG	(Lee *et al.,* 2013)
	LeeSG-R	ATCTGCCACTCCATTGCCG	
	LeeSG-P	**FAM-** TTCTCTGTTGAAAGGCGCTT-**NFQMGB**	

FAM: 6-carboxyfluorescein; NFQMGB: Non-fluorescent quencher minor grove binder.

samples were mixed well to dissolve all powder in the vessel, and poured into a 51-well Colilert® Quanti-tray (IDEXX Laboratories, Westbrook, ME), sealed, and incubated at 35 °C for 24 h. The MPN of *E. coli* in water samples were estimated by counting the *E. coli* positive wells fluorescent under UV and cross-referenced to the MPN table provided by IDEXX Laboratories.

5.3.3 Determining source specificity of *uspC-IS30-flhDC* marker in sewage/wastewater, surface water, and groundwater

To determine the specificity of the naturalized strains to sewage and wastewater, their occurrence in wastewater was compared against *E. coli*-contaminated surface water and groundwater samples across Alberta. One hundred mL of water were processed for *E. coli* using Colilert® presence/absence testing, according to standard operating procedures in the Alberta Provincial Laboratory for Public Health [ProvLab], Edmonton, Alberta. ProvLab is the centralized water testing facility in Alberta, accredited to the International Standard Organization (ISO) 17025 requirements. In total, 76 wastewater samples, cultured by Colilert®, were used in the comparison (at different points in the treatment train), along with 71 *E. coli*-positive surface water samples and 57 *E. coli*-positive ground water samples, both of which were also cultured by Colilert®. From all Colilert® culture-positive samples, 1 mL of culture liquid was removed from the vessel after vigorous mixing, centrifuged at 6,300 × *g* for 10 minutes to collect bacterial pellets and DNA was extracted by using a DNeasy Blood & Tissue kits (QIAGEN, Inc., Valencia, CA), or by boiling bacterial pellets in 750 µL molecular biology-grade water for 10 min, followed by centrifugation at 10,000 × *g* for 10 min, and removal of the liquid suspension. The qPCR conditions for detecting *uspC-IS30-flhDC* marker in these samples were the same as those outlined in the previous section.

The existence of PCR inhibitors in these water samples were tested by the amplification of an internal amplification control (IAC) as described by Deer *et al.*(2010). Specifically, the 198 bp DNA sequence of IAC was synthesized and cloned by Integrated DNA Technolgies (IDT). The generated internal control plasmid DNA was transformed into *E. coli* and purified using QIAprep Miniprep Kit (QIAGEN, Inc., Valencia, CA). Each qPCR reaction of the PCR inhibition test contained 12.5 µl of TaqMan® Fast Advanced Master Mix (Applied Biosystems, Foster City, CA), 400 nM of each primer, and 100 nM of TaqMan probe, 5 µl of DNA template, 5 µl of internal control (100 copies per 5 µl), and molecular biology grade water to a total volume of 25 µl . For the IAC reference sample, the 5 µl DNA template was replaced by 5 µl molecular biology grade water. The cycling conditions were 50 °C for 2 minutes, 95 °C for 30 seconds, 40 cycles of 95 °C for 3 seconds and 60 °C for 30 seconds. The qPCR threshold was set to 0.2. DNA extracts were considered inhibited if their IAC Ct value was ≥3 cycles higher than the Ct value of the IAC reference sample.

5.3.4 Correlation between *uspC-IS30-flhDC* marker and other human/animal fecal contamination source-tracking markers in water.

Ninety-three water samples collected from various surface water and storm water systems in southern Alberta, Canada, were used to compare the correlation between the *uspC-IS30-flhDC* marker and two human markers of fecal sewage pollution (*Bacteroides* HF183 and HumM2). Additional animal fecal contamination markers [cow (CowM3), seagull (LeeSG), and Canada goose (CGO1)] were also tested. The primers and probes used for these markers are provided in Table 5-1.

Water samples (100 mL) were filtered onto 0.4 µm polycarbonate filters followed by DNA extraction using a PowerWater® DNA Isolation Kit (Mo Bio Inc., Carlsbad, CA) according to the manufacturer's instructions. The final volume of DNA extraction was 100 µl. All qPCR reactions were performed in a 20 µl volume containing 5 µl of templates and 10 µl of TaqMan® Fast Advanced Master Mix (Applied Biosystems, Foster City, CA). The final primer and probe concentrations for each qPCR reaction are shown in Table 5-2. The cycling conditions for qPCR of all markers were 95°C for 30 seconds, 40 cycles of 95 °C for 3 seconds and 60°C for 30 seconds. These water samples were also tested for existence of PCR inhibitors using the IAC control as described in the previous section.

Table 5-2. Primer and probe concentration in qPCR reactions for detection of different markers

	HF183	HumM2	CowM3	LeeSG	CGO1	*uspC-IS30-flhDC*
Primer (final concentration)	1000 nM	200 nM	200 nM	300 nM	300 nM	900 nM
Probe (final concentration)	80 nM	125nM	125nM	100 nM	100 nM	250 nM

5.3.5 Statistical analyses

Statistical tests were performed using R software Version 3.0.0. Fisher's exact test was used to compare the association between *uspC-IS30-flhDC* and other human and animal makers. A student's T-test was used to evaluate significance of proportional prevalence estimates of

naturalized *E. coli* in wastewater. In all cases, a p-value < 0.05 was considered statistically significant.

5.4 Results

5.4.1 Determination of LOD and standard curve construction of a qPCR assay targeting *uspC-IS30-flhDC* marker in *E. coli*

A qPCR for detection of the *uspC-IS30-flhDC* was developed. The qPCR was designed to target the site-specific location of the IS30 element in the *uspC-flhDC* intergenic region by targeting the forward primer in the *uspC* gene while the reverse primer targeted the IS30 sequence. The qPCR amplicon size was 151 bp.

A typical standard curve, based on plasmid copy numbers ranging from 500,000 to 5, is shown in Figure 5-1. The limit of detection with 95% confidence (LOD_{95}) for the qPCR was 7 copies per 5 µl plasmid DNA template, with a 95% Confidence Interval [CI_{95}] of 4-14 copies. This qPCR was tested against 70 *E. coli* isolates that were either negative or positive for the *uspC-IS30-flhDC* marker. All 41 *E. coli* strains possessing the *uspC-IS30-flhDC* marker were positive by qPCR whereas all 29 *E. coli* strains lacking this marker were negative by qPCR assay.

5.4.2 Proportion of naturalized *E. coli* in sewage and wastewater

To determine what proportion of the total population of *E. coli* in sewage is represented by naturalized strains of *E. coli*, the ratio between genomic copy numbers obtained for the *uspC-IS30-flhDC* marker and *uidA* genes was examined. The estimated genomic biomass of naturalized *E. coli* strains in raw sewage was 0.29%, but significantly increased during primary treatment to 0.46% (p= 0.007, Table 5-3). An increase in the proportion of naturalized *E. coli*

strains during primary treatment was associated with a 62% reduction in the overall level of culturable *E. coli* after primary treatment (Table 5-4).

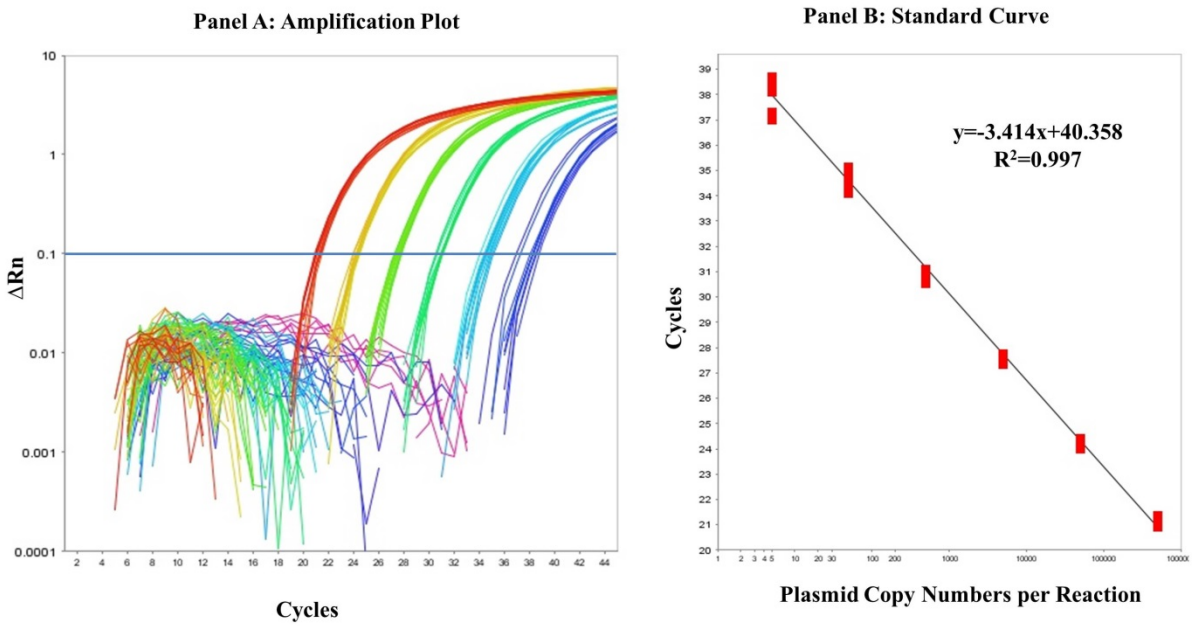

Figure 5-1. Analysis of *uspC-IS30-flhDC* quantitative PCR assay. Panel A. Fluorescence detection (ΔRn) was plotted against cycle numbers. The blue horizontal line represents the threshold for this assay. Panel B- A standard curve constructed based on plasmid copy numbers ranging from 500000 to 5 in each reaction.

5.4.3 Determination of the specificity of *uspC-IS30-flhDC* marker in sewage/wastewater, surface water, and groundwater

The presence of naturalized strains of *E. coli* possessing the *uspC-IS30-flhDC* marker was tested in wastewater (post-grit removal as well as primary, secondary, and UV-treated), and *E. coli*-contaminated surface water and groundwater samples in previous chapter using an end-point PCR. In this chapter, in order to evaluate the specificity of the newly developed *uspC-IS30-flhDC* qPCR assay to *E. coli* populations found in wastewater, expanded set of water samples were tested. For this comparison, *E. coli* cultured samples from each of these sources (Colilert®) were used for comparison. As shown in Table 5-5, 89% (68/76) of the wastewater samples (treated and untreated) were positive by qPCR. Among these wastewater samples, 100% of untreated sewage or primary treated wastewater (31/31), 92% of secondary-treated wastewater (23/25), and 70% of UV-treated wastewater (14/20) samples were positive for *uspC-IS30-flhDC* marker. The reduced occurrence of naturalized *E. coli* strains possessing the *uspC-IS30-flhDC* marker in UV-treated wastewater (i.e., 70%) as compared to primary sewage (i.e., 100%), may be accounted for by the fact that finished UV-treated wastewater effluent only contained a mean MPN value of 5.5 ±3.4 *E. coli* /100 mL and 16 ±25 *E. coli*/100 mL from the Bonnybrook and Pine Creek facilities, respectively (Table 5-4). By comparison, only 4% of the 71 *E. coli*-positive surface water samples were also positive by *uspC-IS30-flhDC* qPCR, and only 5% of the 57 *E. coli*-positive ground water samples were also found to be positive by *uspC-IS30-flhDC* qPCR.

Table 5-3. Prevalence of naturalized *E. coli* in sewage (post-grit removal) and primary-treated wastewater effluent, based on qPCR ratios between *uspC-IS30-flhDC* and *uidA*, and a comparison to the occurrence of the human *Bacteroides* marker (HF183) in these same samples.

		Post-Grit Removal				Primary Effluent			
Wastewater Treatment Plant	**Sampling Date**	*HF183* copy number (per 100 mL)	*uidA* copy number (per 100 mL)	*uspC-IS30-flhDC* copy number (per 100 mL)	% of total *E. coli* population carrying *uspC-IS30-flhDC* marker[a]	*HF183 copy number* (per 100 mL)	*uidA* copy number (per 100 mL)	*uspC-IS30-flhDC* copy number (per 100 mL)	% of total *E. coli* population carrying *uspC-IS30-flhDC* marker[a]
Bonnybrook Wastewater Treatment Plant	2015-02-23	1.3×10^8	6.1×10^8	1.6×10^6	0.25	7.9×10^6	7.6×10^7	4.4×10^5	0.58
	2015-03-16	1.2×10^8	6.6×10^8	2.1×10^6	0.32	1.1×10^7	1.1×10^8	7.4×10^5	0.65
	2015-04-20	9.7×10^7	5.4×10^8	1.6×10^6	0.29	9.4×10^6	6.2×10^7	2.6×10^5	0.42
Pine Creek Wastewater Treatment Plant	2015-03-03	8.9×10^7	5.3×10^8	1.5×10^6	0.28	3.1×10^7	1.9×10^8	6.6×10^5	0.36
	2015-04-14	7.7×10^7	4.6×10^8	1.6×10^6	0.34	2.0×10^7	1.5×10^8	5.9×10^5	0.40
	2015-05-05	5.6×10^7	4.1×10^8	1.0×10^6	0.25	1.4×10^7	1.3×10^8	4.9×10^5	0.37
	Mean (±SD)	9.6×10^7 ($\pm 2.8 \times 10^7$)	5.3×10^8 ($\pm 9.5 \times 10^7$)	1.5×10^6 ($\pm 3.5 \times 10^5$)	0.29 (0.04)	1.5×10^7 ($\pm 8.8 \times 10^6$)	1.2×10^8 ($\pm 4.6 \times 10^7$)	5.3×10^5 ($\pm 1.7 \times 10^5$)	0.46[b] (0.12)

[a] Calculated based on genomic copies of *uspC-IS30-flhDC* marker divided by genomic copies of *uidA*.

[b] Prevalence results between post-grit and primary effluent are significantly different based on student T-test (p=0.007)

5.4.4 Correlation between *uspC-IS30-flhDC* marker and other human/animal fecal contamination source-tracking markers in water.

As a potential marker for municipal sewage contamination, the *uspC-IS30-flhDC E. coli* marker was compared with the human-*Bacteroides* associated HF183 and the HumM2 markers in 93 surface water and storm water samples. The relationships between *uspC-IS30-flhDC E. coli* marker and a variety of animal fecal source tracking markers in these same samples were

Table 5-4. Wastewater treatment performance based on culturable (Colilert®) *E. coli* concentrations across the treatment train at the Bonnybrook and Pine Creek wastewater treatment plants.

		***E. coli* MPN / 100mL**			
Wastewater Treatment Plant	**Sampling Date**	**Post Grit**	**Primary Effluent**	**Pre-UV**	**Post-UV**
Bonnybrook Wastewater Treatment Plant	2015-01-19	4.6×10^6	2.2×10^6	1.6×10^4	7.5
	2015-02-23	9.2×10^6	2.0×10^6	2.8×10^4	1
	2015-03-16	3.9×10^6	2.4×10^6	7.7×10^3	3.1
	2015-04-20	5.2×10^6	2.4×10^6	4.1×10^4	6.3
	2015-05-25	5.8×10^6	2.2×10^6	3.3×10^4	9.6
	Mean (±SD)	5.7×10^6 ($\pm 2.1 \times 10^6$)	2.2×10^6 ($\pm 1.7 \times 10^5$)	2.5×10^4 ($\pm 1.3 \times 10^4$)	5.5 (±3.4)
Pine Creek Wastewater Treatment Plant	2015-01-12	3.9×10^6	3.1×10^6	1.7×10^4	6.3
	2015-02-02	4.4×10^6	3.6×10^6	3.0×10^3	5.2
	2015-03-03	6.5×10^6	4.6×10^6	2.9×10^4	4.1
	2015-04-14	7.7×10^6	4.4×10^6	5.2×10^4	68
	2015-05-05	5.8×10^6	4.1×10^6	1.6×10^3	11
	2016-06-01	7.7×10^6	3.7×10^6	1.6×10^3	3
	Mean (±SD)	6.0×10^6 ($\pm 1.6 \times 10^6$)	3.9×10^6 ($\pm 5.5 \times 10^5$)	1.7×10^4 ($\pm 2.0 \times 10^4$)	16 (±25)

also evaluated. The results are shown in Table 5-6. Statistical analysis (Fisher's exact test) demonstrated a significant correlation ($p<0.05$) between the *uspC-IS30-flhDC* marker and HF183 and HumM2 markers - based on the null hypothesis that these markers do not correlate with each other. The percentage of surface and storm water samples that were positive for HF183 and HumM2 was 61.3% (57/93) and 38.7% (36/93), respectively. By comparison, only 18.3% (17/93) of the samples were positive for *uspC-IS30-flhDC*. However, among the 17 *uspC-IS30-flhDC* positive samples, 16 out of 17 were positive for HF183 and 15 of 17 were also positive for HumM2.

Table 5-5. Prevalence of the *uspC-IS30-flhDC* marker by qPCR in *E. coli*-positive surface water, drinking water and wastewater samples cultured by Colilert®.

E. coli-positive water sample sources	Number of Samples	Marker Positive Samples (%)
Wastewater		
• Untreated (post-grit) or Primary treated wastewater	31	31 (100%)
• Secondary-treated wastewater	25	23 (92%)
• UV-treated wastewater	20	14 (70%)
Surface water	71	3 (4%)
Drinking water (groundwater)	57	3 (5%)

The observation that fewer surface/storm water samples contained the *uspC-IS30-flhDC* marker compared to the human HF183 *Bacterorides* marker, was reflected in the overall occurrence of these markers in raw and primary treated sewage. HF183 was shown to be 10-100 fold more abundant in raw and primary treated sewage compared to the *uspC-IS30-flhDC* marker

(Table 5-3). Correlation significance increased in surface/storm water samples as the concentration of HF183 and HumM2 increased (Table 5-6). Consequently, when the *uspC-IS30-flhDC* positive samples were compared with those containing greater than 4,200 copies of human HF183 per 100 mL of water – the upper value of which is considered important for human health risk assessments for swimming in sewage contaminated waters (Boehm *et al.,* 2015) – the discordance between the HF183 and the *uspC-IS30-flhDC* markers decreased substantially, to only 4.3% (4/93), and with the statistical *p*-value decreasing considerably (Table 5-6). Similarly, at a human health risk target of 2800 copies/100 mL for the human *Bacteroides* M2 target (Boehm *et al.,* 2015), the discordance between the HumM2 and the *uspC-IS30-flhDC* markers decreased from 24.7% (23/93) to 7.5% (7/93) and was also accompanied by a substantial decrease in the *p*-value based on Fisher's exact test (Table 5-6).

These 93 water samples were also tested for three additional animal markers (cow [CowM3], seagull [LeeSG] and Canada goose [CGO1]). Two samples were found positive for CowM3. One of the two samples was positive for *uspC-IS30-flhDC* but this same sample was also positive for HF183 and HumM2, suggesting that the water sample contained fecal contamination originating from both humans and cattle. Among the 27 seagull positive samples, 5 samples were positive for *uspC-IS30-flhDC*, but all 5 samples were also positive for HF183 and 4 were positive for the HumM2 marker, suggesting that these samples contained both seagull and human fecal contamination. In the 11 samples for which the Canada goose marker was observed, only one sample was positive for *uspC-IS30-flhDC* and this sample was also positive for both human HF183 and HumM2. Overall, statistical analysis (Table 5-6) showed no significant relationship between *uspC-IS30-flhDC* and the cow marker CowM3 (p=0.33), seagull marker LeeSG (p=1), and the Canada goose marker COG1 (p=0.68).

Table 5-6. Comparison of various microbial source-tracking markers for the detection of sewage contamination in environmental water samples (*n* = 93).

Source	Marker		*uspC-IS30-flhDC*		*P*-values
			Positive	Negative	
Human	HF183	Positive	16	41	0.002
		Negative	1	35	
	HF183 (> 1000 copies/100mL)	Positive	14	4	1.2×10^{-10}
		Negative	3	72	
	HF183 (> 4200 copies/100mL)	Positive	14	1	6.6×10^{-13}
		Negative	3	75	
	HumM2	Positive	15	21	5.4×10^{-6}
		Negative	2	55	
	HumM2 (> 2800 copies/100mL)	Positive	11	1	2.3×10^{-9}
		Negative	6	75	
Animal	LeeSG (seagull)	Positive	5[a]	22	1.00
		Negative	12	54	
	CGO1 (Canada goose)	Positive	1[a]	10	0.68
		Negative	16	66	
	CowM3 (cow)	Positive	1[a]	1	0.33
		Negative	16	75	

[a] These samples were also positive for human HF183 and M2 markers.

5.5 Discussion

Chapter Four characterized the presence of naturalized strains of *E. coli* in wastewater, identifiable by SNP biomarker patterns across various intergenic regions of the *E. coli* genome and for which these SNP patterns were distinct from *E. coli* isolates originating from human and animal strains. These strains also possessed a site-specific insertion of a transposon (IS30) in the *uspC-flhDC* intergenic region, not observed in any human and animal strains. These wastewater strains were shown to exist in WWTPs throughout Alberta, suggesting a wide spread occurrence

and geographic distribution. As such, this Chapter sought to characterize the occurrence and persistence of these naturalized *E. coli* in WWTPs in Calgary, Alberta. This *E. coli*-based marker was subsequently compared against the human *Bacteroides* source-tracking markers HF183 (Harwood *et al.,* 2014) and HumM2 (Harwood *et al.,* 2014), in addition to other animal fecal source-tracking markers (cattle, seagulls, and Canada goose) in surface/storm water sources as an evaluation of specificity of this marker (and therefore naturalized *E. coli*) to wastewater and wastewater contamination in the environment.

Naturalized wastewater strains of *E. coli* were shown to comprise a small percentage (0.29%) of the overall *E. coli* population in post-grit filtered sewage, as determined by a comparative qPCR analysis of the genome copy number ratios between *uspC-IS30-flhDC* and the universal *E. coli* marker, *uidA*. This finding contrasts the work in Chapter Four in which 5% of the *E. coli* in post-grit wastewater samples were shown to possess the *uspC-IS30-flhDC* marker. This discrepancy is likely due to the fact that molecular-based assays measure DNA in vegetative cells as well as dead and viable but non-culturable cells (VBNC), consequently over-estimating occurrence of total culturable cells in a sample. However, the presence of live, dead and VBNC cells likely contributes to an overall enhanced sensitivity of detection of these markers in water samples, since genome copies of *uspC-IS30-flhDC* exceed culturable *E. coli* levels in this study by >10 fold. These naturalized strains of *E. coli* in sewage represent a significant proportion of the overall *culturable E. coli* biomass after post-grit removal (i.e., 5%). Based on *uspC-IS30-flhDC:uidA* genomic ratios, the proportion of naturalized strains comprising the total *E. coli* population increased significantly through primary treatment, even though the overall mean levels of culturable *E. coli* were reduced by 62% after primary treatment. These data suggest that naturalized strains of *E. coli* in wastewater may be more resistant to wastewater treatment

processes than non-naturalized strains, corroborating the data in Chapter Four. In Chapter Four, it was demonstrated that the treatment of raw sewage with chlorine (i.e., bleach), at levels inducing a 2-4 $\log_{10}$ kill in overall *E. coli* levels (as performed according to the U.S. EPA Alternate Test Procedure [EPA, 2010]), resulted in 59% of the surviving culturable *E. coli* population possessing the *uspC-IS30-flhDC* marker, a significant increase compared to 5% seen in post-grit treated sewage. Furthermore, biochemical and genetic analysis of these strains done in Chapter Four demonstrated that they possess important elements necessary for survival in harsh environments: a) a vigorous generalized stress response (RpoS); b) universal stress-protein genes (*usp*) associated with enhanced survival under oxidative, UV and antibiotic-induced stress (Gustavsson *et al.,* 2002; Nachin *et al.,* 2005); and c) the locus of heat resistance (LHR)[Mercer *et al.,* 2015], a transposon encoding 16 different proteins shown to enhance heat-resistance in both *E. coli* [i.e., 60 ^{o}C treatment for 5 minutes (Mercer *et al.,* 2015)] and nosocomial strains of *Klebsiella pneumoniae* (Bojer *et al.,* 2010). These genetic properties likely promote phenotypic persistence in the non-host wastewater environment and suggest that these naturalized *E. coli* strains may represent excellent microbial source tracking targets for detecting municipal sewage contamination in the environment. The fact that these strains are also culturable in routine selective media for *E. coli* makes these strains amenable as source-tracking targets for routine public health water quality analysis that incorporates *E. coli* as an FIB.

Sewage contamination of surface water sources used for drinking, recreation, and/or irrigation is a global public health challenge (Leclerc *et al.,* 2002; Theron and Cloete, 2002). The United States Environmental Protection Agency (U.S. EPA) has estimated that as many as 40,000 sewer overflows occur each year in the U.S., and up to 500,000 km of coastlines, rivers and streams do not currently meet ambient water quality guidelines due to human and animal

waste contamination (United States Environmental Protection Agency, 2007a; United States Environmental Protection Agency, 2007b). Contamination can enter environmental waters through several routes: i) inadequately treated sewage discharged directly into surface water sources; ii) combined sewer overflows (CSOs); and iii) leaking septic tanks or sewer lines. As human sewage may impact water bodies used for human recreation, determining source attribution of fecal contamination in these sites is important for evaluating public health risks (Soller *et al.,* 2010).

Monitoring fecal indicator bacteria (FIB) levels, such as coliform bacteria, *E. coli*, and enterococci (Wymer, 2007), provides no information regarding the sources of pollution in water, as these microorganisms are also shed by other animals. Several human fecal/sewage specific indicators have been characterized, including human specific viruses [adenovirus, enterovirus and polyoma virus (van der Sanden *et al.,* 2013; Jiang *et al.,* 2001; Rusinol *et al.,* 2016)], protozoan parasites [*Cryptosporidium*(Ruecker *et al.,* 2007)], as well as bacterial markers [*Bacteroides* HF183/HumM2 (Harwood *et al.,* 2014) and the *Enterococcus esp* gene (Harwood *et al.,* 2014)]. However, each of these markers has advantages and disadvantages. For example, the low concentration of viruses and protozoans in water requires water samples to be concentrated for detection; for some markers there is questionable specificity due to their presence in other animal hosts (Jimenez-Clavero *et al.,* 2005; Lee and Kim, 2002). Similarly, human bacterial markers such as the *Bacteroides*-based marker HF183 and HumM2 have been reported to cross-react with canine feces, albeit at lower levels than found in humans (Ahmed *et al.,* 2012; Boehm *et al.,* 2013). Therefore, a "toolbox" strategy, which integrates several methods, has been proposed as most appropriate approach for identifying sources of human fecal/sewage contamination in the water environment (Santo Domingo *et al.,* 2007; Molina *et al.,*

2014). Another major drawback of all current source-tracking technologies is the requirement for separate pre-analytical processes for each of microbial targets examined and for which these processes are different than those used for routine monitoring of water quality (i.e., FIB such as culturable *E. coli*). A significant improvement in the integration of source tracking applications to water quality monitoring would be the development of a robust source-tracking tool centered around *E. coli* as a common FIB target for water quality assessments.

The qPCR developed in the present study was highly specific for naturalized strains of *E. coli* in wastewater, as only 4% and 5% of *E. coli*-contaminated surface water samples and groundwater samples, respectively, had *E. coli* that possessed the *uspC-IS30-flhDC* marker. This was compared to a 100%, 92%, and 70% prevalence of the marker in untreated and primary treated wastewater, secondary treated wastewater and post-UV treated effluent, respectively, similar that seen in Chapter Four with the end-point PCR. The observation that a lower percentage of UV-treated effluents contained *E. coli* carrying the *uspC-IS30-flhDC* marker (70%) compared to raw or primary treated sewage (100%) is reflected in the very low mean concentrations of *E. coli* in final finished effluent (5.5 *E. coli* /100mL in Bonnybrook and 16 *E. coli*/ 100mL in Pine Creek) compared to raw or primary treated wastewater (~ 6 x 10^6 *E. coli*/100 mL in both WWTPs). The observation that very low concentrations of *E. coli* exist in the final wastewater effluent but that 70% of these finished wastewater samples possess culturable *E. coli* carrying the *uspC-IS30-flhDC* marker further supports the concept of differential survival of these strains across the wastewater treatment process. The much lower prevalence of these strains in *E. coli*- contaminated surface and ground water sources suggests that not all *E. coli* that are present in water carry this marker, and therefore, the *uspC-IS30-flhDC* marker cannot be considered a generalized marker of *E. coli* strains capable of surviving in the water environment.

Wholistically and conceptually, the data from this chapter and the previous chapter imply that naturalized *E. coli* populations may have evolved specifically to survive and grow in a wastewater environment, and that these phenotypic/genotypic adaptive properties did not solely evolve in response to enhancing their survival in water-based matrix and likelihood of transmission to other hosts.

In the surface and groundwater samples that were positive for the *uspC-IS30-flhDC* marker we could not exclude the possibility of contamination of these sources with wastewater, as some surface sites were known to be downstream of wastewater treatment plants. Unfortunately, duplicate samples were not collected or processed for these surface water and groundwater sources so that alternate source-tracking markers could be examined (i.e., HF183, HumM2). However, herein lies the challenge with implementation of source tracking tools for which pre-analytical processes are different from those methods commonly used for water quality analysis (i.e., culturable *E. coli*). Although only 57 *E. coli*-positive groundwater samples were incorporated in this study, it is important to note that the overall prevalence of *E. coli*-contaminated groundwater wells in the province of Alberta (in any given year) is only about 2-3% (Neumann, unpublished data). Thus, it is estimated that approximately 2280 groundwater wells (based on a 2.5% *E. coli* contamination rate) were screened by routine public health testing at ProvLab in order to identify the 57 wells contaminated with *E. coli*. In order to assess human fecal contamination in these wells using *Bacteroides*-based HF183 or HumM2 markers, all 2280 groundwater wells would have had to been filtered, DNA extracted and qPCR performed on each of the targets independently – a costly endeavour. The distinct advantage of using the *uspC-IS30-flhDC* locus as a marker of wastewater contamination is that the qPCR assay can be performed directly on cultured *E. coli* positive samples only (i.e., the 57 samples), potentially

acting as a cost efficient screening process to identify source waters contaminated with municipal wastewater.

The applicability of the *uspC-IS30-flhDC* qPCR assay for tracking wastewater contamination in the environment was validated by comparing this marker with two human fecal/sewage markers, *Bacteroides*-based HF183 and the HumM2 markers, as well as other animal fecal markers (cattle, Canada goose, and seagull) in 93 surface and storm water samples collected in southern Alberta. In the 17 samples found positive for the *uspC-IS30-flhDC* marker, 16/17 also contained human HF183, and 15/17 also carried the HumM2 marker. Statistical analysis demonstrated a significant correlation between the *uspC-IS30-flhDC* marker and the two other human/sewage markers (HF183 and HumM2) in these 93 samples. However, the overall sensitivity of the *uspC-IS30-flhDC* marker was less than that observed for either the HF183 or HumM2 markers, as determined by the level of discordance between the samples (i.e., more samples contained HF183 and HumM2 than the *uspC-IS30-flhDC* marker). Nevertheless, recent studies employing quantitative microbial risk assessment approaches have demonstrated a substantial and unacceptable level of health risk associated with recreational exposure to water samples exceeding 4200 copies/100 mL of HF183 and 2,800 copies/100 mL of HumM2. Our study demonstrated that as levels of HF183 and HumM2 increased in water samples, stronger correlations were observed with the *uspC-IS30-flhDC* marker. Interestingly, the strongest correlations observed between the two human fecal markers (HF183 and HumM2) and the *E. coli*-based *uspC-IS30-flhDC* marker were in samples in which the HF183 concentration was >4,200 copies/100 mL and HumM2 concentration >2,800 copies/100 mL, suggesting that the *uspC-IS30-flhDC* qPCR assay may be an extremely valuable tool for screening water of unacceptable health risk.

The greater prevalence of HF183 and HumM2 *Bacteroides* markers in natural water samples compared to the *E. coli uspC-IS30-flhDC* marker may be explained as follows. Firstly, the prokaryotic order of *Bacteroidales* constitutes a great proportion of the gut microbiota in humans (Dominianni, *et al*., 2015; Andersson *et al.,* 2008; Wang *et al.,* 2004), and overall genomic copy numbers have been shown to approximate 10^8 per 100 mL of raw sewage (Staley *et al.,* 2012). This is compared to the 10^6 genomic copies of the *uspC-IS30-flhDC* marker per 100 mL of sewage that we observed in the present study. This observation possibly explains why there is increased strength of the correlations between HF183 and *uspC-IS30-flhDC* in samples for which HF183 exceed 1000 copies/mL. Although the overall PCR sensitivity of both assays was similar (i.e., LOD_{95} of ~7 copies/reaction), the 100-fold greater copy number associated with HF183 in sewage and wastewater likely contributes to its increased detection/occurrence. Secondly, HF183 and HumM2 can be found detected directly from human feces whereas the *uspC-IS30-flhDC* marker does not appear to be a constituent of the *E. coli* population in human or animal feces (Chapter Four). This marker appears to only exist in naturalized strains of *E. coli* found in wastewater (Chapter Four), and which are ubiquitously prevalent in wastewater treatment plants throughout Alberta, Canada. To date, the *uspC-IS30-flhDC* marker has not been found in any *E. coli* isolates originating from humans or animals, and nor is it represented in Genbank or extensive genome databases of *E. coli* from humans and animals (Chapter Four). As such, HF183, HumM2 and the *uspC-IS30-flhDC* markers are fundamentally very different from each other – the HF183 and HumM2 markers are indicative of human fecal contamination whereas the *uspC-IS30-flhDC* marker is indicative of municipal sewage/wastewater contamination and therefore, an indirect marker of human waste. At present, it is unknown whether these naturalized wastewater *E. coli* strains that possess the *uspC-IS30-*

flhDC marker are also constituents of small-scale waste management systems (i.e., septic systems). Therefore, a third possibility is that the surface and storm water samples positive for HF183 and HumM2 may have been contaminated by small scale septic systems rather than by municipal wastewater. Fourthly, it is possible that at least some of the discordance observed between human *Bacteroides* markers and the *E. coli*-based *uspC-IS30-flhDC* marker may be due to the non-specificity of the *Bacteroides*-based markers, as these markers have also been shown to cross react with dog, chicken, cow, goat and sheep feces (Ahmed *et al.,* 2012; Odagiri *et al.,* 2015). Although the *uspC-IS30-flhDC* marker appears to be highly specific for municipal wastewater treatment plants, the specificity of the *uspC-IS30-flhDC* marker has not been evaluated against fecal samples collected from a wide range of animals - a project currently underway in our laboratory. Nevertheless, the fact that stronger correlations occurred between increasing concentrations of HF183/HumM2 and the *uspC-IS30-flhDC* marker suggests that the relative abundance of genome copies in wastewater is the primary reason in the discrepancy between these markers.

5.6 Conclusion

The proportion of the total population of *E. coli* in sewage represented by naturalized strains of *E. coli* were estimated in raw sewage and after primary treatment. An increase in the proportion of naturalized *E. coli* strains during primary treatment was observed. Naturalized strains possess an insertion element (IS30) in the *uspC-flhDC* intergenic region, for which a qPCR assay targeting the *uspC-IS30-flhDC* was developed. A strong correlation was observed between this marker and the human *Bacteroides* HF183 and human M2 marker in water samples. No correlation was observed with animal markers. Although occurrence of the *uspC-IS30-flhDC* marker was 100-fold less compared to the human HF183 marker in raw wastewater, the

advantage of this assay is that naturalized wastewater strains readily grow in media bases commonly used for routine culture of *E. coli* from water (i.e., Colilert®, *E. coli* selective agar). Consequently, samples submitted for water quality analysis for culturable *E. coli* can be directly screened for wastewater contamination. This is unlike other genetic markers that require different pre-analytical processing steps and that add significant costs for analysis.

Chapter Six AN EXPLORATION OF THE POTENTIAL ADAPTIVE MECHANISMS UTILIZED BY NATURALIZED WASTEWATER *E.COLI* FOR SURVIVAL IN WASTEWATER

6.1 Abstract

In Chapters Two and Three of this thesis, novel logic regression-based SNP analysis of ITGRs was used as a mean of characterizing the host-specific nature of *E. coli* in animals, for which these same methods were used in Chapter Four to reveal genetic evidence of naturalized strains of wastewater-specific *E. coli*. In Chapter Four, naturalized wastewater *E. coli* strains were characterized by unique SNP biomarker patterns in various ITGRs, and many also possessed an insertional transposon (IS30 element) found in the *uspC-flhDC* ITGR. The prevalence/occurrence/specificity of this naturalized wastewater *E. coli* (*uspC-IS30-flhDC* positive) in wastewater was then evaluated by a qPCR assay in Chapter Five. The purpose of this chapter was to explore the potential adaptive cellular mechanisms used by naturalized wastewater *E. coli* for their survival in wastewater. In this chapter, a serial stress experiment (mimicking nutrient deprivation, osmotic stressors, followed by chlorine stress) of four *E. coli* strains was performed. Survival was measured by quantifying bacteria after each stress challenge. Biofilm formation was assessed among eight human fecal strains (not possessing the *uspC-IS30-flhDC*) and ten naturalized wastewater strains possessing the *uspC-IS30-flhDC* marker. The expression of the flagellar regulator gene, *flhDC*, between a human fecal strain H51 (not possessing the *uspC-IS30-flhDC*) and two *uspC-IS30-flhDC* wastewater positive strains (WW10 and WW63) under the serial stress treatments were compared to assess bacterial motility. For the human *E. coli* strain not possessing the *uspC-IS30-flhDC* or an RpoS generalized stress response, a 4.1 $\log_{10}$ mean reduction in bacterial cell counts was observed. This was in contrast

to another human strain lacking the *uspC-IS30-flhDC* but possessing the RpoS stress response as well as the naturalized wastewater *E. coli* strains (WW10 and WW63) for which stress-treatment reduced bacterial numbers by only ~2 $\log_{10}$. Biofilm forming capacity of the naturalized wastewater strains was significantly greater (p-value <0.1) than that of the human fecal strains. However not all wastewater strains had high biofilm forming capacity. There was no significant difference in *flhDC* expression between wild type human strain H51 and *uspC-IS30-flhDC* positive strains WW10 or WW63 after 24 hour of TSB culture followed by nutrient deprivation/osmotic stress (p-value>0.1). However, a significant difference (p-value <0.1) in *flhDC* expression between the wild type human strain H51 and the WW10 (p-value=0.06) and WW63 strain (p-value=0.08) was observed after chlorine treatment. In conclusion, the results in this Chapter indicate that although wastewater strains possess adaptive phenotypes to survival in wastewater, including chlorine tolerance, there does not appear to be a direct association between stress tolerance and the presence or absence of the *uspC-IS30-flhDC* marker, even though biofilm production and *flhDC* expression was enhanced in these strains. Further studies will need to be done to better understand the mechanisms used by these naturalized wastewater strains for their adaptation/survival in wastewater environment.

6.2 Introduction

As outlined in Chapter Four, naturalized wastewater *E. coli* strains were characterized by unique SNP biomarker patterns in various ITGRs, and many also possessed an insertional transposon (IS30 element) found in the *uspC-flhDC* ITGR (Chapter Four). Given its relative abundance in wastewater as well as its widespread distribution throughout wastewater treatment plants throughout Alberta (Chapters Four and Five), it was hypothesized that this unique ITGR

polymorphism may be functionally important for the survial and persistence of these strains in the wastewater environment.

Insertional elements have been shown to modify expression of certain genes (Barker *et al.,* 2004), and as such, the site-specific location of the IS30 element between the *uspC-flhDC* intergenic region is interesting. The *usp* gene family encodes universal stress proteins (usp A-G) that play an important role in promoting resistance to oxidative damage, UV radiation and even antibiotics (Gustavsson *et al.,* 2002; Nachin *et al.,* 2005). The *flhDC* gene encodes the master regulator for flagellar biosynthesis, acting as an activator for expression of bacterial flagellar proteins (Smith and Hoover, 2009). Flagella are required for bacterial motility, which in turn, has been positively associated with biofilm formation (Wood *et al.,* 2006). Biofilm production is an important strategy used by bacteria to survive unfavourable environmental conditions, and the formation of biofilms has shown to promote resistance to chlorine, UV radiation, oxidative damage and even predation (Vogeleer *et al.,* 2014; Ryu and Beuchat, 2005; Elasri and Miller, 1999) –important microbial reduction strategies used during wastewater treatment. Consequently, the *uspC-IS30-flhDC* polymorphism may have been important for the adaptive evolution of this bacteria towards colonization of this nutrient rich, but microbially and physicochemicallystressful environment. Demonstrating the adaptive significance of this genetic polymorphism in an ITGR would support the hypothesis that the regulome is important in the evolutionary dissemination of bacteria into new environments (i.e., animals or non-host environments), and also support a corollary hypothesis that ITGRs are embossed with SNP biomarkers indicative of host/environment specificity.

E. coli is a commensal strain of bacteria found in the intestine of warm blood animals and birds as well as the non-host environment such as soil, sand, and water (Power *et al.,* 2005;

Tymensen *et al.,* 2015; Kon *et al.,* 2007; Chandrasekaran *et al.,* 2015; Winfield and Groisman, 2003). In order to survive in these diverse habitats, *E. coli* has adapted and evolved various strategies for surviving stressful abiotic (nutrient availability, temperature, pH) and biotic conditions (microbial niche competitors). For example, when exposed to high temperature, *E. coli* can produce heat shock proteins that prevent aggregation and refolding of proteins (Arsene *et al.,* 2000). Under low pH, the bacteria can pump out protons (H^+) from the cytoplasm through an amino acid-based acid resistance system (De Biase and Lund, 2015). Although it is considered a facultative anaerobe, *E. coli* can also adapt to oxygen stress by utilizing catalase (Iuchi and Weiner, 1996) and superoxide dismutases (Baez and Shiloach, 2013) to prevent oxidative damage to cellular components. Certain strains of *E. coli* are also efficient biofilm producers allowing for enhanced survival under extreme environmental conditions (Vogeleer *et al.,* 2014). A wide repertoire of genes/pathways are known to regulate the stress response in *E. coli* and include: i) the generalized stress response (RpoS) which facilitates survival against nutrient derivation, oxidative and osmotic stressors (Battesti *et al.,* 2011); ii) the universal stress proteins (usp A-G), known to be important for oxidative stress, UV-radiation as well as antibiotic-resistance (Gustavsson *et al.,* 2002; Nachin *et al.,* 2005) ; and iii) the more recently discovered, locus of heat resistance (LHR), important for survival under high temperature conditions [i.e., 60^o C(Mercer *et al.,* 2015)]. The enormous genetic diversity observed in this bacterial species, along with diverse phenotypic mechanisms for dealing with stress, has been a major factor in facilitating the evolutionary radiation of this organism into diverse animal host and non-animal host environments.

The purpose of this study was to explore the potential correlation between the presence of the *uspC-IS30-flhDC* ITGR polymorphism observed in naturalized wastewater *E. coli* strains and

the adaptive phenotypes of motility, biofilm formation, and chlorine resistance as adaptive strategies promoting survival in a wastewater environment.

6.3 Methods

6.3.1 Stress experiment

Two naturalized *uspC-IS30-flhDC* marker positive wastewater *E. coli* isolates (WW10 and WW63) and two *uspC-IS30-flhDC* marker negative wild type *E. coli* (H51and H54) isolated from human feces were used in stress treatment experiment. The *E. coli* strains were cultured in tryptic soy broth (TSB) for 24 hours after which the culture was diluted 1:10 using sterile distilled water and incubated for another 24 hours at room temperature (nutrient deprivation/osmotic stress). The cultures were then treated with chlorine by adding 0.8% bleach to the culture. The mixture was shaken for 2 minutes and free chlorine was measured using ChemMets (CHEMetrics, Midland, VA). The bleach volume added to the culture was adjusted until the free residual chlorine reached 0.3-0.5 ppm. The culture was allowed to incubate for 5 minutes at room temperature and 10% sodium thiosulfate was added to neutralize the free chlorine. Free chlorine was measured again to ensure it was neutralized. Survivability of the *E. coli* strains was immediately determined based on culture (see below). Triplicate samples were performed for this experiment.

Bacterial cell counts were performed at three time points: after the initial 24 hour culture in TSB; after 24 hours incubation of the 1:10 dilution in distilled water under room temperature; and immediately after chlorine treatment. A tenfold serial dilution was made for each culture and 100 µl of each dilution was plated on LB agar plates, incubated at 35 °C overnight and colony counting performed 24 hours later.

6.3.2 *flhDC* experiment studies

For measurement of *flhDC* expression, *E. coli* cultures were collected for RNA extraction at the same three time points as described above for the stress experiment (Section 6.3.1). RNA extraction was performed using RNAProtect Bacteria Reagent (QIAGEN, Inc., Valencia, CA) and RNeasy mini kit (QIAGEN, Inc., Valencia, CA). Specifically, one volume of bacterial culture was mixed with two volumes of RNA protection reagent. The mixture was incubated for 5 minutes at room temperature and centrifuged at 6300 x *g* for 10 minutes to collect the pellet. The pellet was then used for RNA extraction using a RNeasy Mini Kit according to the manufacturer's instructions. RNA was treated with OPTIZYME™ DNase I (Fisher Scientific, Waltham, Massachusetts) for DNA removal.

Table 6-1. PCR primers used in this study

Target	Primers and probes	Primer and probe sequence (5'-3')	Reference
rpoA	rpoA-F	GGCTTGACGATTTCGACATC	(Sharma and Bearson, 2013)
	rpoA-R	GGTGAGAGTTCAGGGCAAAG	
	rpoA-P	**FAM-**TGAAGTTATTCTTACCTTGAATAAATCTGGCATTG-**TAMRA**	
flhDC	flhDC-F	ACAACATTAGCGGCACTGAC	(Sharma and Bearson, 2013)
	flhDC-R	AGAGTAATCGTCTGGTGGCTG	
	flhDC-P	**FAM**-AAACGGAAGTGACAAACCAGCTGATTG-**TAMRA**	

Quantitative reverse transcription PCR (RT-qPCR) was performed to measure the expression of *flhDC* gene using TaqMan® RNA-to-Ct™ 1-Step Kit (Thermo Fisher Scientific, Waltham, Massachusetts). All reactions were carried out in a 20 µl volume which contained 200ng RNA template, 10 µl TaqMan® RT-PCR Master Mix, 0.5 µl TaqMan® RT Enzyme Mix, 0.9 µM each primer, 0.25 µM TaqMan probe and molecular biology degree water. The primers used are listed in Table 6-1. The RT-qPCR conditions were as follows: a holding stage at 48 °C for 15 minutes for reverse transcription, another holding stage at 95 °C for 10 minutes for activation of enzymes, and 40 cycles of denaturation at 95 °C for 15 seconds and annealing/extension at 60 °C for 1 minute. The expression of *flhDC* in two *uspC-IS30-flhDC* positive wastewater strains (WW10 and WW63) was compared to the expression of the *flhDC* in wild-type human strain H51. The expression data of *flhDC* was normalized to endogenous levels of the housekeeping gene *rpoA*.

6.3.3 Bacterial biofilm formation assay

6.3.3.1 Bacterial strains

Eighteen *E. coli* strains were used in this portion of the study. Ten were naturalized wastewater *E. coli* strains (*uspC-IS30-flhDC* positive) originating from four different wastewater treatment plants in Alberta, Canada. Eight of them were human *E. coli* strains (*uspC-IS30-flhDC* negative) isolated from different patient fecal swabs submitted to the Edmonton site of the Alberta Provincial Laboratory for Public Health (ProvLab) for routine microbiological testing, and sampling adhered to all ethics requirements (File #: Pro00005478_CLS3 at the University of Alberta). All strains were confirmed as *E. coli* by biochemical analysis using a Vitek Bacterial

Identification System (BioMerieux Canada Inc., St. Laurent, Canada) according to the manufacturer's instructions and protocols at ProvLab.

6.3.3.2 Biofilm formation assay

Biofilm formation was evaluated as previously described by O'Toole (2011) with some modifications. Briefly, strains of *E. coli* were grown in TSB overnight at 35°C. The optical densities of the cultures were adjusted to OD600 nm=1.0 and 100 µl of the cultures were added into microtiter plates (CoStar 3595, NY) to test for their ability to form biofilms. The lid of the microtiter plate was replaced by a sterile 96-well PCR plate (Greiner Bio-One, Frickenhausen, Germany) by inserting the PCR plate wells into the wells of the culture microtiter plate to help biofilm binding. The plate was incubated at 35 °C for 24 hours. The culture media was then discarded. For each well, 125 µl of 0.1% crystal violet was added to stain the biofilm and the whole plate was washed with distilled water four times to get rid of excess cells and dye. Biofilm formation was quantified by adding 125 µl of 30% acetic acid to dissolve the crystal violet adhering to biofilms on the PCR plate, and the absorbance at 550 nm was measured using a microplate reader FLUOstar Omega (Thermofisher, Whitby, Ontario, Canada).

6.3.4 Statistics

Statistical analysis of the data for *flhDC* expression test and biofilm formation test was performed using R software version 3.0.0. For the biofilm formation test, an unpaired Student's t-test was performed. For *flhDC* gene expression test, a paired Student's t-test was performed. A significance value of $p < 0.1$ was deemed relevant.

6.4 Results

6.4.1 *E. coli* survivability under stressed conditions

Four *E. coli* strains, including: a) H51 – a human wildtype fecal strain lacking a generalized stress response (RpoS) and the *uspC-IS30-flhDC* biomarker; b) H54 – a human fecal strain possessing a strong RpoS generalized stress response but lacking the *uspC-IS30-flhDC* biomarker; and c) WW10 and WW63 - naturalized wastewater strains possessing strong RpoS activity and the *uspC-IS30-flhDC* biomarker, were analyzed for survivability against serial stress conditions. Strains were grown in TSB for 24 hrs, diluted 1:10 in distilled water (mimicking nutrient deprivation and osmotic stressors) for 24 hrs at room temperature and subsequently treated with chlorine bleach. All four *E. coli* strains displayed similar replicative potential in TSB, reaching levels of ~10^9cfu/ml (Table 6-2) after 24 hours. When TSB cultures were diluted 1:10 in distilled water and incubated for 24 hrs, no decline in survivability was observed among the four *E. coli* strains. However, survivability among the strains was significantly affected by treatment with chlorine. A 4.1 $\log_{10}$ mean reduction in bacterial cell counts were observed for the H51 negative control strain, whereas for the human H54 strain and the naturalized wastewater *E. coli* strains (WW10 and WW63) bacterial concentrations were only reduced by ~2 $\log_{10}$, suggesting differential survival among *E .coli* strains in response to chlorine treatment (Table 6-2).

6.4.2 Motility and biofilm formation

To better understand phenotypic alterations associated with differential resistance to chlorine, phenotypic characteristics such as biofilm formation and motility (as determined by *flhDC* gene expression assays) were examined as survival phenotypes for *E coli* in the non-host

Table 6-2. Survival of human and naturalized wastewater *E. coli* strains after nutrient deprivation/osmotic stress and chlorine treatment.

		Treatment				
		Control	Nutrient deprivation/ osmotic stress [a]		Chlorine Treatment [b]	
***E. coli* strains**	***E. coli* source**	***E. coli* numbers/mL after 24 hours culture in TSB [c]**	***E. coli* numbers/ mL after osmotic stress [c]**	**Log_{10} reduction after osmotic stress [c]**	***E. coli* numbers/mL after chlorine treatment [c]**	**Log_{10} reduction after chlorine treatment [c]**
H51	Human (rpoS *negative / uspC-IS30-flhDC* negative, control strain)	$2.0 \pm 0.28 \times 10^9$	$3.2 \pm 0.62 \times 10^9$	-0.20 ± 0.15 [d]	$3.3 \pm 0.8 \times 10^5$	4.1 ± 0.3 [e]
H54	Human (rpoS *positive uspC-IS30-flhDC* negative, control strain)	$1.1 \pm 0.19 \times 10^9$	$3.0 \pm 0.27 \times 10^9$	-0.46 ± 0.05 [d]	$2.2 \pm 0.12 \times 10^7$	2.1 ± 0.10 [e,f]
WW10	Wastewater	$1.0 \pm 0.09 \times 10^9$	$2.1 \pm 0.15 \times 10^9$	-0.32 ± 0.06 [d]	$1.7 \pm 0.15 \times 10^7$	2.1 ± 0.03 [e,f]
WW63	Wastewater	$9.3 \pm 1.5 \times 10^8$	$8.8 \pm 1.0 \times 10^8$	-0.01 ± 0.10 [d]	$5.6 \pm 0.91 \times 10^6$	2.2 ± 0.13 [e,f]

[a] Nutrient deprivation/osmotic shock performed by diluting TSB cultures 1:10 in distilled water an incubating for 24 hrs at room temperature.

[b] After nutrient deprivation/osmotic shock, cells were treated with 0.3-0.5 ppm residual free chlorine for 5 minutes.

[c] *E. coli* concentrations and log_{10} reductions are presented as mean ± SE (n=3).

[d] No significant reduction (p-value>0.05) in *E. coli* log_{10} survival when comparing bacterial levels between TSB control cultures and nutrient deprivation/ osmotic stress treatment.

[e] Significant reduction (p-value<0.05) in *E. coli* log_{10} survival when comparing bacterial levels between nutrient deprivation/osmotic stress and chlorine treatment.

[f] Significant difference in log_{10} survival outcomes between the strain indicated and the negative control human strain (H51- strain lacking the generalized rpoS stress response).

environment. Biofilm formation was assessed among eight human fecal strains (not possessing the *uspC-IS30-flhDC*) and ten naturalized wastewater strains possessing the *uspC-IS30-flhDC* marker. Biofilm forming capacity of the naturalized wastewater strains was significantly greater (p-value=0.04) than that of the human fecal strains, although large variation was observed within each of the groups (Figure 6-1). Biofilm formation, as measured by crystal violet absorption to the polysaccharide matrix, was 3.2 times greater in wastewater strains than human strains (i.e., compare mean absorbance values of wastewater strains [0.295±0.27] compared to human fecal strains [0.092±0.09]). However not all wastewater strains had high biofilm forming capacity. For example, wastewater strain WW53, although possessing the *uspC-IS30-flhDC* marker had lower biofilm-forming capacity than most of the human strains.

To determine if motility could be indirectly altered by the IS30 element in naturalized wastewater *E. coli* strains (WW10 and WW63) the expression of the flagellar regulator gene, *flhDC*, was compared between the human fecal strain H51 and the two *uspC-IS30-flhDC* positive wastewater strains (WW10 and WW63) under the serial stress treatments described above. Relative expression levels between the housekeeping gene *rpoA* and *flhDC* were used as a measure of gene transcription, since relative expression controlled for the greater inactivation of the H51 strain compared to wastewater strains treated with chlorine. The results demonstrated that there was no significant difference in *flhDC* expression between wild type strain H51 and *uspC-IS30-flhDC* positive strains WW10 or WW63 after 24 hour of TSB culture followed by nutrient deprivation/osmotic stress (Figure 6-2, p-value>0.1). However, a significant difference (p-value<0.1) in *flhDC* expression between the wild type human strain H51 and the WW10 (p-value=0.06) and WW63 strain (p-value=0.08) was observed after chlorine treatment. The relative

expression of *flhDC* of WW10 and WW64 increased 3.6 and 2.6 fold compared to the H51 human strain, respectively, after chlorine exposure.

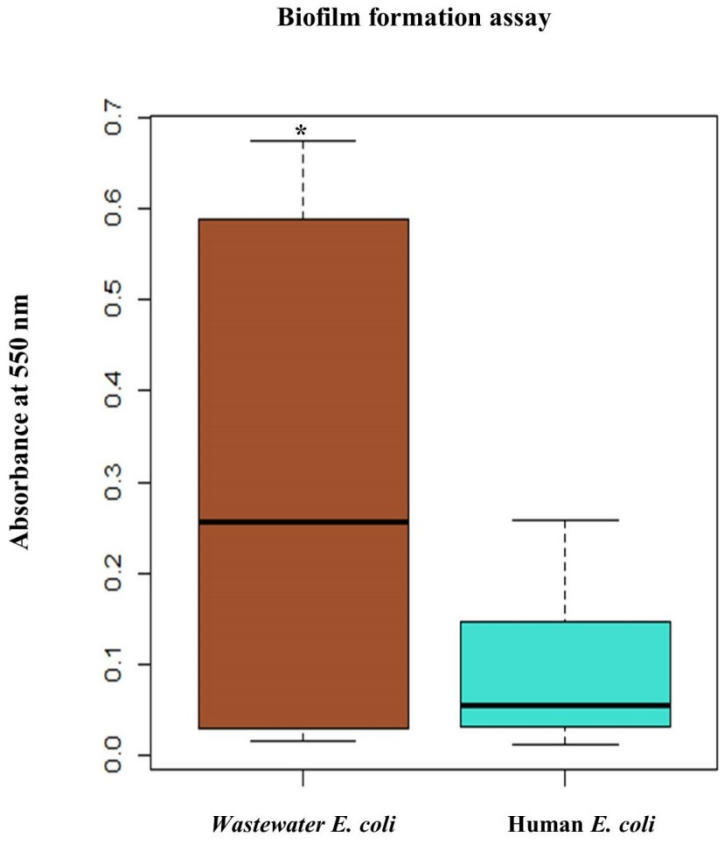

Figure 6-1. Comparison of biofilm formation capacity in ten *uspC-IS30-flhDC* positive and eight *uspC-IS30-flhDC* negative *E. coli* strains. Absorbance at 550 nm was used as a measure of biofilm formation capacity in biofilm formation assay. The result demonstrated that biofilm forming capacity of the naturalized wastewater strains (*uspC-IS30-flhDC* positive) was significantly greater (asterix representative of *p*-value<0.1) than that of the human fecal strains (*uspC-IS30-flhDC* negative).

6.5 Discussion

Several studies have demonstrated that although wastewater is largely composed of human and animal fecal wastes, the wastewater matrix itself possesses a very unique microbiome (Wang *et al.,* 2012). Although fecal bacteria may comprise the majority of microorganisms entering the wastewater treatment plant, certain strains of *E. coli* appear to have evolved to become naturalized populations - acquiring adaptive stress-resistant characteristics selected through evolution for enhanced survival in this non-host environment. Naturalized populations of *E. coli* possess distinct SNP biomarkers in various ITGRs, with many strains possessing an insertional element (IS30) within the *uspC-flhDC* locus (Chapter Four). The fact that the *uspC-IS30-flhDC* marker is highly specific to wastewater and is geographically disseminated throughout wastewater treatment plants across Alberta suggested that this genetic polymorphism may play a functional role in survival of these naturalized strains in the wastewater environment.

To better understand the potential role of this marker in the adaptive selection and evolution of naturalized *E. coli* populations in wastewater, the survival of *uspC-IS30-flhDC* positive and negative *E. coli* isolates were examined after serial stress conditions that mimicked typical wastewater treatment plant environments (nutrient deprivation, osmotic shock and chlorine treatment). Other phenotypic outcomes related to stress, including bacterial motility and biofilm formation were also examined, both of which have been shown to be important for *E. coli* survival in a non-host environment. Barker *et al.* (2004) observed that the placement of an insertion sequence known as IS5 upstream of *flhDC* in *E. coli* activated transcription of *flhDC* by potentially releasing transcriptional repression, leading to increased flagellar synthesis and bacterial motility; an effect that is known to promote biofilm formation in bacteria (O'Toole and Kolter, 1998). Sharma and Bearson (2013) specifically demonstrated that differential regulation

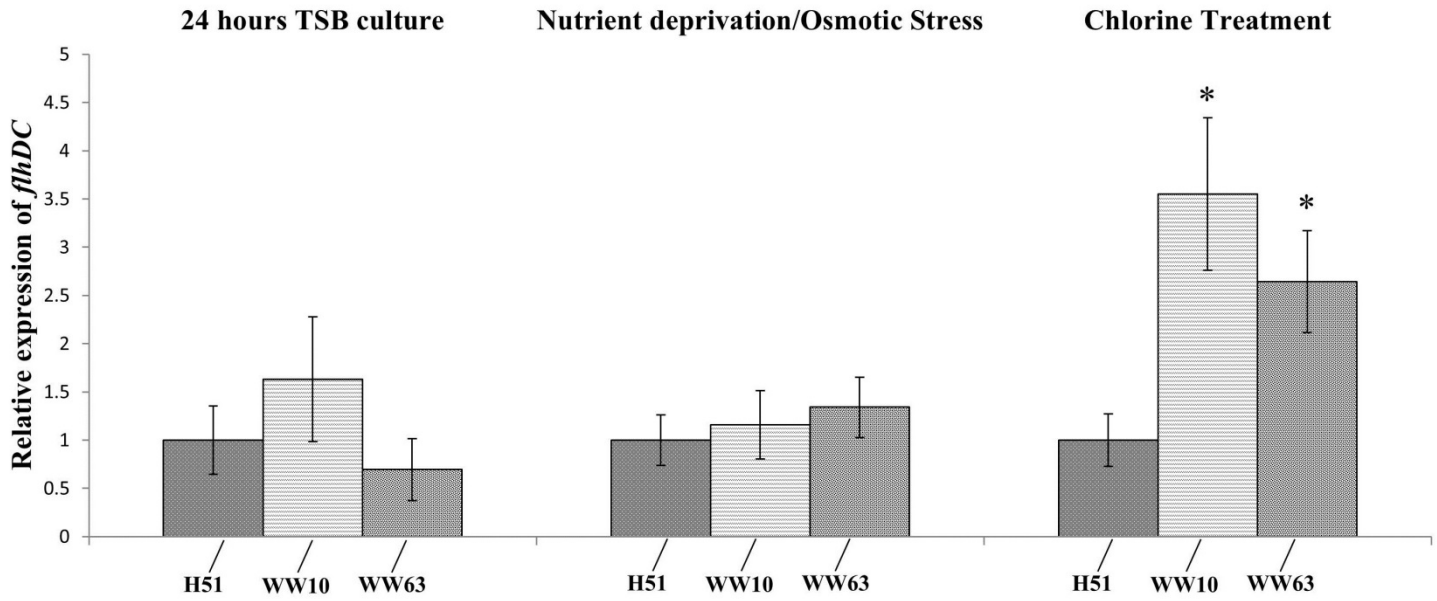

Figure 6-2. Expression of *flhDC* in response to nutrient deprivation/osmotic shock and chlorine treatment. Relative expression of *flhDC* was normalized to the expression of the reference gene, *rpoA*. The expression of *flhDC* was compared between *uspC-IS30-flhDC* negative *E. coli* strain H51 and two *uspC-IS30-flhDC* positive strains WW10 and WW63 after 24 hour TSB culture, nutrient deprivation/osmotic stress, and chlorine treatment. The results demonstrated that there was no significant difference in *flhDC* expression between wild type strain H51 and *uspC-IS30-flhDC* positive strains WW10 or WW63 after 24 hour TSB culture or after nutrient deprivation/osmotic stress (*p*-value>0.1). However, a significant difference (*p*-value<0.1 [asterix]) in *flhDC* expression between the wild type human strain H51 and the WW10 (*p*-value=0.06) or WW63 (*p*-value=0.08) after chlorine treatment was observed. The relative expression of *flhDC* of WW10 and WW64 increased 3.6 and 2.6 fold compared to the H51 human strain respectively.

of *flhDC* can influence biofilm formation, and Wood *et al.* (2006) demonstrated that bacterial motility was positively related to its biofilm capacity. Although flagellar expression is highly complex, involving more than forty genes (McCarter, 2006), it was hypothesized that the naturalized wastewater strains may increase levels of *flhDC* expression in response to environmental stressors and which would result in increased production of biofilms. It has been previously shown that biofilm-producing *E. coli* strains are more resistant to chlorine (Ryu and Beuchat, 2005; De Beer *et al.,* 1994) and that biofilm formation is important for protecting bacteria in harsh non-host environments (Stewart *et al.,* 2013). To examine this, one human strain (*uspC-IS30-flhDC* negative) and two wastewater strains (*uspC-IS30-flhDC* positive) were subjected to nutrient deprivation/osmotic shock and chlorine treatment and expression of *flhDC* examined. Wastewater strains had significantly greater relative expression of *flhDC* compared to the human strain after chlorine treatment. In addition, biofilm production across ten naturalized wastewater strains (all possessing the *uspC-IS30-flhDC* marker) was greater than that observed for eight human fecal strains lacking the *uspC-IS30-flhDC* marker, suggesting that both motility and biofilm formation may be important strategies for survival in a nutrient rich wastewater matrix.

To determine whether these specific alternations were associated with increased survival, culture-based methods were used as means of quantifying recovery of stressed bacteria. Four *E. coli* strains were assayed for this portion of the study. One human strain, H51 (also used in the *flhDC* expression studies), lacked an RpoS stress response, whereas another human strain, H54, possessed a vigorous rpoS stress response – a cellular response shown to be important for environment survival in *E. coli* (White *et al.,* 2011). Two wastewater strains, each possessing the *uspC-IS30-flhDC* marker and having a strong RpoS response, were also selected for

evaluation. Both wastewater strains as well as the human H54 wildtype strain were shown to be highly resistant to chlorine, displaying 100 times more resistance to chlorine when compared to the H51 strain. Interestingly, there was no difference in survivability between the human wildtype H54 strain and the wastewater strains. Consequently, there does not appear to be a direct association between chlorine tolerance and the presence or absence of the *uspC-IS30-flhDC* marker, even though biofilm production and *flhDC* expression was enhanced in these strains. Nevertheless, biofilm formation and motility are not the only strategies utilized by bacteria to survive chlorine-stress. Several inducible genes/regulators are also produced in response to exposure to reactive chlorine species in *E. coli* (Parker *et al.,* 2013; Gray *et al.,* 2013), the mechanisms of which can lead to increased intracellular glutathione levels (Chesney *et al.,* 1996; Saby *et al.,* 1999) as well as other antioxidants such as catalases (Dukan and Touati, 1996), methionine sulfoxide reductase (Rosen *et al.,* 2009), and even oxidative repair mechanisms (Gray *et al.,* 2013). It is also possible that the *E. coli* strains subjected to the high-stress conditions in this experiment entered into viable but non-culturable states (VBNC). This is a common survival strategy used by *E. coli* to deal with adverse environmental conditions (Trevors, 2011), but was not examined in the present study. Thus, although similar survival outcomes were observed between the human H54 strain and the naturalized wastewater *E. coli* strains based on cell culture, it is plausible that significant differences in overall survival may have existed if VBNC were also are accounted for. PCR assays incorporating propidium monoazide (PMA) have been used to evaluate VBNC survival under stress conditions (van Frankenhuyzen *et al.,* 2011). PMA binds to DNA preventing PCR amplification of target sequences, but can only penetrate cells in which cellular membrane of the bacteria has been compromised (Nocker *et al.,* 2006). During transition into a VBNC state, bacteria condense cellular material, often becoming

much smaller in size, yet their membrane structures remain intact and prevent PMA from penetrating (Nocker *et al.,* 2006). Quantitative PCR can be used to evaluate differences between PMA-treated and untreated cultures to determine the proportion of cells in a VBNC state within a sample. Future studies should entail an evaluation of survival based on culturable bacteria and VBNC and incorporate assays examining diverse chlorine tolerance mechanisms.

The fact that naturalized wastewater strains of *E. coli* displayed chlorine tolerance in this Chapter helps to explain the findings in Chapter Four where it was observed that a substantially greater proportion of the culturable *E. coli* population found in wastewater treated with chlorine were shown to carry the *uspC-IS30-flhDC* marker (59%) when compared to the *E. coli* population in untreated wastewater (5%). This finding also corroborates the findings of Chapter Five that demonstrate differential survival of naturalized wastewater *E. coli* strains across the wastewater treatment process, providing further circumstantial evidence of the potentially important role that *uspC-IS30-flhDC* may play in survival in a wastewater matrix. However, a definitive understanding of the role of the *uspC-IS30-flhDC* locus as an adaptive genotype for survival of naturalized *E. coli* populations in wastewater will require the generation of isogenic mutants in which this locus is altered or eliminated using site-directed mutagenesis. A major contribution to knowledge for this chapter of the thesis is the description of a suitable culture-based, serial stress bioassay that could be used for evaluating the survivability of isogenic mutants against parent strains, and for determining what effect these mutations may have on *flhDC* gene expression and biofilm formation.

The observation that chlorine treatment selects for survival of naturalized strains of wastewater *E. coli* has some important implications for the water industry. Firstly, the U.S. EPA Alternate Test Procedure – a protocol used for evaluating the suitability of culture media on the

recovery of chlorine-stressed fecal microbes such as *E. coli* - actually appears to select for survival and growth of naturalized *E. coli* populations in these samples. This finding has important implications. It suggests that the protocol may bias experimental outcomes towards the culture of naturalized wastewater strains of *E.coli* as opposed fecal strains of *E. coli*. Whether this has any impact on interpretation regarding the suitability of approved culture media for public health testing of fecal contamination of water remains to be determined. Also, since these approved culture media are used to evaluate wastewater treatment performance, growth bias towards treatment-resistant naturalized strains may underestimate true inactivation efficiency against fecal strains. As a corollary, the presence of naturalized treatment-resistant *E. coli* in wastewater raises some concerns regarding the use of *E. coli* as a suitable *fecal* indicator of water quality.

Secondly, the wide variation in survivability of *E. coli* strains to chlorine raises some direct concerns for public health. It was observed as early as 1949 that certain populations of bacteria, including *E. coli*, appear to be resistant to chlorine treatment (Allen and Brooks, 1949; Farkas-Himsley, 1964). Since that time, various mechanisms of chlorine resistance have been characterized in *E. coli* (Gray *et al.,* 2013). It is plausible that genetic determinants of chlorine tolerance from naturalized strains may act as reservoirs for horizontal transfer to pathogenic strains, or vice versa (i.e., virulence genes may transfer from pathogenic strains to naturalized wastewater strains). Chlorine tolerant *E. coli* generated from wastewater treatment plants located upstream of drinking water treatment plants may cause problems if treatment at the drinking water plant is compromised and virulence maintained in chlorine-tolerant strains. It would be interesting to determine if the *E. coli* strains often observed in chlorinated drinking water distribution systems carry the *uspC-IS30-flhDC* marker, as this would be suggestive of

wastewater contamination in drinking water treatment system. Interestingly, it has been demonstrated that chlorine can promote antibiotic resistance (Murray *et al.,* 1984). Khan *et al.* (2016) observed that chlorine-tolerant strains of bacteria isolated from chlorinated drinking water systems were more resistant to tetracycline, sulphamethoxozole and amoxicillin, and Jia *et al.* (2015) observed that chlorine residual was an important parameter is driving antibiotic resistance in bacterial populations found in drinking water distribution systems, including *E. coli*. Given that antibiotic-resistant *E. coli* (AR-*E. coli*) are relatively abundant in finished wastewater (Iwane *et al.,* 2001), as are *uspC-IS30-flhDC* positive *E. coli* strains (Chapter Four and Five), it would be interesting to evaluate whether *uspC-IS30-flhDC* positive *E. coli* are antibiotic resistant. At present, the characterization of antibiotic resistance in *uspC-IS30-flhDC* positive wastewater strains has not been evaluated.

6.6 Conclusion

To better understand the potential role of *uspC-IS30-flhDC* in the adaptive selection and evolution of naturalized *E. coli* populations in wastewater, the survival of *uspC-IS30-flhDC* positive and negative *E. coli* isolates were examined after serial stress conditions that mimicked typical wastewater treatment plant environments (nutrient deprivation, osmotic shock and chlorine treatment), and other phenotypic outcomes related to stress, including bacterial motility and biofilm formation were also examined, both of which have been shown to be important for *E. coli* survival in a non-host environment. The results demonstrate that there does not appear to be a direct association between chlorine tolerance and the presence or absence of the *uspC-IS30-flhDC* marker, even though biofilm production and *flhDC* expression was enhanced in these strains. Nevertheless, biofilm formation and motility are not the only strategies utilized by bacteria to survive chlorine-stress. Several inducible genes/regulators are also produced in

response to exposure to reactive chlorine species in *E. coli.* A definitive understanding of the role of the *uspC-IS30-flhDC* locus as an adaptive genotype for survival of naturalized *E. coli* populations in wastewater will require the generation of isogenic mutants in which this locus is altered or eliminated using site-directed mutagenesis.

Chapter Seven : GENERAL DISCUSSION

The central theme of this thesis was to better understand the microbial genetics of host specificity. *E. coli* was selected as the target microorganism for study since certain strains of *E. coli* have been shown to be host/niche specific (as reviewed in the introductory chapter of this thesis). Intergenic regions may represent unique and informative targets for studying host specificity, since these regions are important in the regulation of gene expression. The ability of *E. coli* to adapt and colonize a host/niche will largely depend on its ability to respond to the environmental conditions perceived within the host/niche, and to ultimately outcompete other microbes for limited nutrients, colonization sites and avoid predation. Thus, the regulome plays a key role in microbial survival, and it was hypothesized that DNA sequence polymorphisms in ITGRs may result in adaptive and deterministic phenotype variations that would drive natural selection and evolution of microbial host/niche-specificity.

This research study sought to answer three critical questions:

a. Is *E. coli* host/niche specific?
b. Do the intergenic regions of *E. coli* contain host/niche specific information?
c. Do genetic variations in ITGRs that correlate with *E. coli* host/niche-adaptation/specificity yield biologically relevant phenotypes important for survival in a particular niche?

Below, I present the main findings of this thesis, as they pertain to these three questions, and expound on the impact that this research has had, not only on our understanding of microbial host specificity but also in context of the evolution of disease emergence. Finally, I will also discuss the limitations of this research and highlight important directions for future work.

7.1 Significant findings

The main findings presented in the five data chapters were as follows:

I. Humans and animals appear to possess *E. coli* strains that are host-specific as identified using logic regression analysis of intergenic regions.

II. Incorporating multiple ITGRs (i.e., concatenation) into logic regression model building resulted in greater host-specificity and sensitivity outcomes in biomarkers.

III. Application of logic regression analysis to whole genome data (*in silico*) can be used as a valuable discovery tool for identifying ITGRs embossed with a high degree of host-specific information.

IV. The occurrence of naturalized strains of *E. coli* in wastewater.

V. Naturalized wastewater *E. coli* have the potential to be an important source-tracking marker for characterizing wastewater pollution in an environment.

7.1.1 Humans and animals appear to possess *E. coli* strains that are host-specific as identified using logic regression analysis of intergenic regions.

Based on the literature outlined in Chapter One, microbial host specificity has been observed across various microbial species, and *E. coli* is no exception. Indeed, the results of this thesis suggest that *E. coli* isolates collected from the feces of warm-blooded animals (birds and mammals) and even non-host environments (wastewater) appear to contain host/niche specific populations. Across 15 different animal hosts, *E. coli* were successfully discriminated based on animal host source with relatively high specificity (i.e., among the samples of the non-target animal host, the proportion that correctly did not have the host-specific marker pattern) and sensitivity (i.e., among the samples from a given animal host, the proportion that correctly had

the host-specific marker pattern). In Chapter Two it was shown that 82% of all deer *E. coli* isolates displayed a unique intergenic SNP biomarker pattern that was 98% specific to deer when evaluated using only three intergenic regions (*uspC-flhDC*, *csgBAC-csgDEFG*, and *asnS-ompF*). Host-specific biomarker sensitivities of 77% and 67% were observed in muskrat and moose, respectively, with specificity all equal to 99%. The data presented in Chapter Three suggest that host specificity may be quite high in any individual animal, particularly when host informative ITGRs are selected for analysis. Biomarker discovery using whole genome data demonstrated that as much as 78% of the *E. coli* originating from humans may be host-specific. Remarkably, for cattle, this number may be as high as 92%.

Intergenic regions regulate the expression of genes. They possess promoter sequences (Haugen, *et al*, 2008), transcriptional regulator binding sites (repressor/activator) (Browning and Busby, 2004), small regulatory RNAs (Gottesman, 2005), and transposable elements (Casacuberta and Gonzalez, 2013), and control changes in cell phenotypes, adaptive functioning, and sensation (Smits, *et al*, 2006). Consequently, host-specific population structures in *E .coli* may be governed at the regulatory transcriptome level. At present, most studies use DNA sequences of genes for inferring relationships between bacterial strains while relatively few studies have used the non-coding intergenic regions. The finding of this thesis demonstrated DNA sequences polymorphisms in intergenic regions contain valuable information that can be used for bacterial typing. In agreement with the research presented in this thesis, intergenic regions have been shown to possess robust typing resolution by various other studies (Glazunova, *et al*, 2005; Djelouadji, *et al*, 2008; Fournier and Raoult, 2007; Li, *et al*, 2006; Yanagihara, *et al*, 2010). None of these studies, however, have examined this in the context of host specificity, but rather have used traditional phylogenetic analysis (Yanagihara, *et al*, 2010).

The findings presented in this thesis demonstrate that supervised learning algorithms, such as logic regression, represent better approaches for understanding population genetic diversity and host-specificity in *E. coli* when compared to unsupervised learning algorithms. Supervised learning methods infer patterns in observed data associated with a group label, whereas the commonly used statistical methods for DNA sequence phylogenetic analysis, such as UPGMA, NJ, FM, MEA, MP, and ML - considered as *unsupervised learning* methods,- do not associate data with group labels. For data with known group labels (e.g., animal host origin of an individual *E. coli* isolate), supervised learning methods are more powerful for finding the patterns/structures that are related to the group labels. The results of Chapter Two, Three and Four demonstrate that logic regression could be used to identify ITGR biomarkers in *E. coli* with high host/niche specificity and sensitivity in isolates collected from various animal and non-animal environments. The validity of the host-specific biomarker patterns identified by logic regression analysis was verified through fivefold cross validation and permutation testing. In addition, a comparison between logic regression and maximum likelihood phylogenetic methods for characterizing host specificity patterns in ITGRs demonstrated that logic regression analysis revealed more robust associations with *E. coli* host specificity than did ML phylogeny. For example, using ML-phylogeny, *E. coli* isolates obtained from humans were represented in dispersed clusters throughout the ML phylogenetic tree, segregated from each other by an array of animal isolates. In contrast, human biomarkers were identified by logic regression analysis with the majority of human isolates appearing to be human specific. These human-specific isolates partitioned themselves throughout all human clusters identified by ML phylogenesis. The results suggest that logic regression is able to use host label information in the analysis for host grouping, outperforming the traditional phylogenetic approach and representing a novel and

superior approach for characterizing host specificity/adaptation in microbes. It should be noted that in Chapter Four ML phylogenetic analysis did cluster wastewater isolates together based on analysis at two concatenated ITGRs. However, these same isolates were also shown to have ITGR biomarkers with extremely high sensitivity and specificity values (i.e., 82% of wastewater isolates had a distinct logic regression-based biomarker that was 100% specific to wastewater), values substantially higher than those observed even among human/animal isolates in the library. These data suggest that unsupervised learning methods have lower resolution capacity compared to supervised learning methods for discriminating host/niche specific patterns in DNA sequence. At the same time, when unsupervised learning methods actually reveal patterns of host specificity in DNA sequence data, as they did in Chapter Four for wastewater isolates, it can be argued that the genetic background of these strains is very unique compared to the others used in the analysis (i.e., human and animal strains).

7.1.2 Incorporating multiple ITGRs (i.e., concatenation) into logic regression model building resulted in greater host-specificity and sensitivity outcomes in biomarkers

In Chapter Three of this thesis, it was demonstrated that when SNP biomarkers were generated from concatenated ITGR sequences, a higher level of host-predictive performance was observed as compared to any individual ITGR sequence. For example, higher HPP indices were observed by logic regression analysis when *csgBAC-csgDEFG, uspC-flhDC,* and *asnS-ompF* or *cutC-torYZ*, *metQ-rcsF*, and *araH-otsB* were concatenated. Although it can be argued that higher HPP indices result from increased DNA polymorphisms across multiple ITGRs, the results of this thesis demonstrated that there was no correlation between the number of SNPs in an ITGR and the overall level of host-predictive information encoded in the ITGR (see Figure 3-2). As a

specific example, among the 80 ITGRs used in the *in silico* genome analysis, the intergenic region of the *nrdA-yfaL* locus possessed 192 SNPs and ranked as one of the most polymorphic ITGRs in *E. coli*, but the HPP for this ITGR was amongst the lowest observed. These data suggest that host-specificity is not necessarily linked to the degree of polymorphism in ITGRs, but rather, that a small number of critical polymorphisms across diverse ITGRs may be a more important determinant of host-specificity. This would also help explain why unsupervised learning methods were less discriminatory than supervised methods for revealing patterns of host specificity. Unsupervised learning methods use all DNA sequence variation for evaluating statistical significance, including polymorphisms that are not related to host specificity (i.e., polymorphisms potentially associated with non-adaptive, non-deterministic variation [i.e., neutral selection/genetic drift]). The fact that supervised learning tools specifically search for DNA sequence data related to group labels (host/niche) suggests that the SNP biomarker patterns identified are inherently adaptive and deterministic, and therefore may play a role in the regulation of flanking genes. Logic regression biomarker searching from the concatenated ITGR loci of *rcsD-ompC*, *ydeR-yedS*, and *rclR-ykgE* obtained from whole genome data demonstrated that as much as 78% of the *E. coli* population found in humans may be host-specific.

7.1.3 Application of logic regression analysis to whole genome data (*in silico*) can be used as a valuable discovery tool for identifying ITGRs embossed with a high degree of host-specific information.

In order to identify ITGRs that were highly informative of host origin, a whole genome based approach was used to evaluate how well 80 different ITGRs across 160 *E. coli* genomes (human, cattle, and other animals) performed in terms of encoding host-specific information. The

whole genome based approach represented a convenient and cost-effective tool for searching for host-informative ITGRs. The disadvantage of using whole genome approaches is that there are relatively few *E. coli* genomes from bacterial strains originating from vertebrate host other than humans. Nevertheless, the results demonstrated that, among the host-informative ITGRs identified by whole genome based analysis, *ydeR-ydeS* was the most informative intergenic region in humans, with 56% of all human *E. coli* isolates displaying a unique SNP biomarker pattern that was 99% specific. *rcsD-ompC* was the second most host-informative locus in human-derived *E. coli*, possessing 54% sensitivity and 99% specificity. As was observed in humans, the *ydeR-ydeS* locus was also the most host-informative in cattle, for which 92% of all cattle *E. coli* carried a biomarker that was 98% specific to cattle. The second best performing host-specific ITGR for cattle was *rcsD-ompC*, possessing a sensitivity of 49% and a specificity of 97%. The findings suggest that *in silico* analysis represents a powerful approach for identifying host-informative ITGRs which can then be evaluated by targeted ITGR sequencing approach across a larger library of *E. coli* isolates obtained from different animal hosts. In addition, the results from Chapter Three also demonstrated that different ITGRs encode different levels of host-specific information, with some intergenic regions showing high HPP values (i.e. *ydeR-ydeS* in human) while others showed relatively low HPP values (see Table 3-5). The varied levels of host-specific information were not only observed among various intergenic regions in the same host animal but also on the same intergenic region across different host animals. For example, in Chapter Three it was shown that *cutC-torYZ* was the most host-informative ITGR among the six ITGRs tested in humans, while this locus was the least host-informative for dogs. Consequently, ITGRs that are informative for *E. coli* from one animal host are not necessarily informative for *E. coli* obtained from other host animals. These data suggest that evolution of

host specificity in *E. coli* is not linear (i.e., defined as a set of mutations in a distinct set of ITGRs). There are likely to be a multitude of evolutionary drivers for host adaptation and specificity in *E. coli* and these drivers may vary in different host animals (i.e. diet, pH, temperature, immunological responses, and physiological metabolites). These natural selection pressures will likely drive *E. coli* strains to regulate different sets of genes for its adaptation and survival in different animals. This concept was reinforced by the observation that among the 80 ITGRs examined across the whole genome of *E. coli*, those ITGRs having the highest level of host predictive power in humans regulated genes associated with antibiotic resistance, whereas in *E. coli* strains derived from cattle the most host predictive ITGRs regulated environmental stress genes.

7.1.4 The occurrence of naturalized strains of *E. coli* in wastewater

Escherichia coli has been proposed to have two habitats - the intestine of mammals/birds and the non-host environment. Naturalized *E. coli* strains have been described in sand, sediments and water (Power, *et al*, 2005; Tymensen, *et al*, 2015; Kon, *et al*, 2007; Chandrasekaran, *et al*, 2015; Winfield and Groisman, 2003), and the presence of *E. coli* in these environments indicates that some strains may have evolved to survive and replicate outside their animal host reservoirs. Municipal wastewater represents a very different environment for *E. coli* compared to the gastrointestinal tract, and it is widely believed that the *E. coli* present in wastewater originates from the fecal wastes generated by humans and animals. The discovery of naturalized strains of *E. coli* in wastewater (Chapter Four) was extraordinary, and provided an excellent opportunity to better understand the relationship between ITGR variation and the adaptive/deterministic phenotypes needed for niche specificity. Since most mutations in biological systems are non-

adaptive and non-deterministic (i.e., genetic drift) it was important to understand whether variations in ITGRs correlated with adaptive phenotypic properties of these niche-specific strains. This type of research would be particularly difficult to demonstrate in animal models, but the presence of naturalized strains in wastewater allowed for an interrogation of the phenotypic/genotypic properties that may have promoted the evolution of these strains to naturally survive and grow outside their vertebrate hosts. Therefore Chapters Four, Five, and Six of this thesis focused on characterizing the adaptive phenotypic and genotypic properties of naturalized wastewater *E. coli* strains.

Naturalized wastewater strains of *E. coli* were originally isolated from chlorine-treated sewage (Chapter Four) and identified by possessing a unique logic regression based biomarker that was very distinct from human and animal strains of *E coli*. Approximately 82% of chlorine-tolerant wastewater *E. coli* isolates possessed a unique SNP biomarker pattern which was 100% specific for wastewater. The majority of these *E. coli* isolates also possessed an IS30 insertion element located within the ITGR between the *uspC* and *flhDC* genes. Interestingly, the positional location of the insertion sequence was not observed in *E. coli* isolates from a library of non-wastewater sources (i.e. human and other animal hosts), nor was it represented in any GenBank submitted sequences or *E. coli* genomes deposited in NCBI genome database. All wastewater chlorine-tolerant *E. coli* isolates possessing the *uspC-IS30-flhDC* belonged to phylogroup A and formed a single clade in the ML phylogenetic tree, suggesting a common genetic background of these strains.

Surprisingly, these naturalized strains possessed the Locus of Heat Resistance [LHR] (Mercer, *et al*, 2015) – a genomic island containing 16 open reading frames that encode proteins associated with heat shock, cell envelop maintenance, and turnover of misfolded proteins

(Mercer, *et al*, 2015). *E. coli* strains possessing the LHR have been shown to survive heat shock temperatures of 60^{o}C for 5 minutes (Mercer, *et al*, 2015). It is interesting to note that wastewater temperatures do not exceed 20^{o}C in Alberta, and therefore, it is hypothesized that the LHR may play an important role in environmental stress survival against wastewater disinfection (i.e., chlorine, UV, ozone, etc.) – treatment processes targeting the molecular destruction of biomolecules (i.e., lipids, proteins, etc.,). Therefore it is speculated that the LHR may promote maintenance and turnover of lipid and protein structures damaged by chlorination, ozonation or UV-induced hydrolysis.

All chlorine-tolerant strains possessing the *uspC-IS30-flhDC* marker also possessed the RpoS-mediated generalized stress response – a bacterial mechanism known to be important for survival under limiting nutrient conditions, osmotic shock and oxidative stress (Battesti, *et al*, 2011; Lacour and Landini, 2004; Weber, *et al*, 2005). The resistance phenotype, known as *rdar*, has been shown to enhance long-term survival of *E coli* under harsh conditions (White and Surette, 2006). Rdar cells secrete an extracellular matrix comprised of curli fimbriae, cellulose and polysaccharides (Zogaj, *et al*, 2001; Romling, *et al*, 1998) and for which the matrix provides increased resistance to disinfection (Ryu and Beuchat, 2005; Uhlich, *et al*, 2006). In addition to possessing the LHR and RpoS environmental stress elements, these naturalized strains also appeared to possess universal stress proteins. The usp superfamily, represented by 6 different proteins in *E. coli* (usp A, C, D, E, F and G), is involved in cellular responses to osmotic, oxidative, antibiotic and UV-induced stressors (Gustavsson, *et al*, 2002; Nachin, *et al*, 2005). More specifically, the *uspC* gene, upstream of the IS30 element in naturalized wastewater strains of *E. coli*, has been shown to enhance flagellar production and motility in *E. coli* (Nachin, *et al*, 2005) and has also been shown to be important in enhancing UV-resistance in *E. coli*

(Gustavsson, *et al*, 2002). The presence of a multitude of stress-related pathways and their associated biological relevance in terms of promoting survival during wastewater treatment, points to an adaptive survival strategy that may have evolved in these naturalized strains to live and survive in this harsh environment.

Not surprisingly, these naturalized wastewater strains of *E. coli* were found to be a100 times more resistant to chlorine than a strain of *E. coli* directly isolated from human feces. Moreover, naturalized wastewater strains had better biofilm forming capacity than human-derived *E. coli* strains. It has been previously shown that biofilm-producing *E. coli* strains are more resistant to chlorine (Ryu and Beuchat, 2005; De Beer, *et al*, 1994) and that biofilm formation is important for protecting bacteria in harsh environments (Stewart, *et al*, 2013). The presence of the IS30 element in the *uspC-flhDC* ITGR was shown to upregulate *flhDC* gene expression under oxidative stress [chlorine] conditions (Chapter Six), and upregulation of flagellar synthesis is a feature believed to be important in biofilm formation. Barker, *et al* (2004) observed that an insertion sequence known as IS5 upstream of *flhDC* in *E. coli* activated transcription of *flhDC* by potentially releasing transcriptional repression, leading to increased flagellar synthesis and bacterial motility; an effect promoting biofilm production in bacteria (O'Toole and Kolter, 1998). Sharma and Bearson (2013) specifically demonstrated that differential regulation of *flhDC* can influence biofilm formation.

To further evaluate the phenotypic specificity of these naturalized *E. coli* strains to wastewater, water samples from various sources were tested for the presence of these naturalized strains using a (q)PCR targeting the *uspC-IS30-flhDC* ITGR. The results demonstrated that these naturalized strains were frequently found in wastewater samples but rarely in surface water and ground water contaminated with *E. coli*. A particularly important finding was that these

naturalized strains were found in geographically separate WWTP plants across Alberta, and that these naturalized strains also displayed enhanced survival across the wastewater treatment process *in situ*. Overall, the data strongly suggest that these strains are specifically adapted to survive in wastewater, and that the adaptations present are not simply associated with survival in water matrix.

Collectively, Chapters Four, Five and Six, provide valuable phenotypic/genotypic evidence for niche adaptation and specificity related to ITGR variation in *E. coli*, and support the hypothesis that genetic variation in the regulome is an important driver of host/niche specificity in microbes. These data also provide credence to the corollary hypothesis that ITGRs are embossed with host-specific information, useful for identifying host or environmental sources of *E. coli* contamination in food and water.

7.1.5 Naturalized wastewater *E. coli* are so unique genetically, that they have the potential to be an important source-tracking marker for characterizing wastewater pollution in an environment.

In this thesis, a qPCR assay targeting the *uspC-IS30-flhDC* ITGR marker from naturalized wastewater *E. c*oli strains was compared against commonly used human/sewage source-tracking markers, *Bacteroides* HF183 (Harwood, *et al*, 2014) and HumM2 (Harwood, *et al*, 2014), for detection of wastewater pollution in the environment. The results demonstrated that a significant correlation was observed between the *uspC-IS30-flhDC* marker and human/sewage-associated HF183 and HumM2 markers, but not with the animal markers (cattle, seagulls, and Canada goose). These marker comparison results added further evidence to support the niche specificity of *uspC-IS30-flhDC* positive *E. coli* to wastewater, and also demonstrated the

potential use of this marker for tracking wastewater contamination in the environment. Since naturalized wastewater strains of *E. coli* were shown to readily grow in culture-based systems commonly used for detection of this bacterium in water (Colilert®, Modified Fecal Coliform [m-FC] agar), the methods described in this thesis allow for the immediate application of these tools into routine public health screening of water quality. As outlined in Chapter Five, sewage contamination of surface water sources used for drinking, recreation, and/or irrigation is a universal public health challenge (Leclerc, *et al*, 2002; Theron and Cloete, 2002), and as many as 40,000 sewer overflows occur each year in the U.S. alone, resulting in up to 500,000 km of coastlines, rivers and streams not meeting ambient water quality guidelines (United States Environmental Protection Agency, 2007a; United States Environmental Protection Agency, 2007b).

This aspect of the thesis is a particularly important and relevant finding. It demonstrates how very basic scientific questions (i.e., the genetics of host-specificity in *E. coli*) can lead to important translational public health outcomes and applications (i.e, using this information for source tracking pollution in the environment). This is one of the most important lessons learned from this thesis research.

7.2 Limitations and future work

Herein, I discuss several limitations associated with the research carried out in this thesis.

7.2.1 *E. coli* library size

In this thesis, host-specific biomarkers were successfully identified using logic regression based on 845 *E. coli* isolates obtained from 15 different vertebrate hosts. Although the number of strains and their animal hosts were considered statistically relevant for analysis, for some animal

hosts the number of strains was relatively low. For example, only 21 cat *E. coli* isolates were included in this study. This may explain why the sensitivity and specificity of cat biomarkers decreased substantially after fivefold cross validation. Biomarkers identified based on a limited number of isolates may not be robust due to the fact that the collection of strains may not represent the genetic variability possessed by all strains found within that host, and therefore the likelihood of finding the true host-specific biomarker is decreased. However, although the sample size was small for some animal hosts, inclusion of them in the dataset was still important as they were used as a control group during logic regression analysis in the search for host-specific biomarkers for other animals. In order to increase the robustness of the identified host-specific biomarkers, the *E. coli* library should be expanded by adding more isolates into the host groups and including *E. coli* isolates from animal hosts not currently represented in the library.

Although the data suggests that a certain degree of host-specificity may exist in *E. coli*, many different factors may affect the genetic population structure of *E. coli* in a single animal or across various individuals within an animal grouping. Temporal variations in *E. coli* population genetics have been studied in some host species (Caugant, *et al*, 1981; Jenkins, *et al*, 2003; Gordon, 1997), and some strains appear to dominate whereas others appear transiently (Caugant, *et al*, 1981). Factors that may affect the stability of biomarker SNP patterns in any given host may include diet, age, geography, health status and host-genetics, and these biomarker patterns could potentially change spatially and temporally. Consequently, future studies should focus on expanding representation of *E. coli* isolates in the library that include representation of these aforementioned factors as categorical variables.

7.2.2 Number of intergenic regions used and *E. coli* genome database size

The number of intergenic regions used in this thesis was relatively small, compared to the estimated number of ITGRs across the *E. coli* genome. In Chapter Two only three intergenic regions were used in targeted biomarker analysis. In Chapter Three this number was increased to six, and whole genome analysis subsequently increased this number to 80. By comparison, Zaslaver, *et al* (2006) identified upwards of 1820 ITGRs in a laboratory strain of *E. coli*. The *E. coli* genome contains about 4500 genes of which 1500 of them are conserved (Lukjancenko, *et al*, 2010). Consequently, an extensive number of intergenic regions are still unexplored. Some intergenic regions may out-perform those used in this thesis, and therefore the true level of host-specificity in *E. coli* strains originating from any given host remains to be determined. The values reported in this thesis likely represent the minimal levels of host specificity/sensitivity, since relatively few ITGRs were examined (and assuming the factors that affect stability of the biomarkers [as mentioned above] are well represented in the current libraries). Indeed, in the *in silico* genome analysis, several intergenic regions had higher HPP values than those identified by targeted sequencing, and an evaluation of all 1820 ITGRs across *E. coli* may reveal very host-specific ITGRs. However, at present acquiring DNA sequences of all intergenic regions in a large collection (i.e. ~1,000) of *E. coli* isolates from different animals, by either targeted ITGR sequencing approach or whole genome based approach, is unrealistic in terms of the amount of work and cost for a single laboratory. For the targeted ITGR sequencing approaches, relatively few ITGRs could be examined at once, and for which little *a priori* knowledge on host-specificity exists. For whole genome based approach, a large number of ITGRs could be examined simultaneously to search for the most host-informative intergenic regions but a number of *E.coli* genomes would have to be sequenced from a diverse repertoire of animal hosts not

currently represented in NCBI databases. As of June 2016, about 4600 publically available *E.coli* genomes were deposited in NCBI database, but for which the vast majority are human strains. With the advent of routine genome sequencing being commonplace in research laboratories, the number of sequenced *E. coli* genomes will increase greatly over the next few years and for which better representation of *E. coli* isolates from various other host sources will inevitably occur. Eventually targeted ITGR approaches to biomarker discovery may become redundant. Nevertheless, in the meantime, genome discovery tools can be used to identify host informative ITGRs that can be subsequently validated by targeted ITGR sequencing. As an example, an immediate research study could focus on the validation of the *ydeR-yedS* locus identified from whole genome analysis as a cattle-specific ITGR, by targeted sequencing of this locus in all 845 strains of *E. coli* in the current library and performing logic regression analysis.

7.2.3 Mechanistic role the site-specific insertion of the IS30 element has on phenotypic resistance in *E. coli*

As outlined previously, the naturalized wastewater strains of *E. coli* identified in this thesis contained an insertion element (IS30) located specifically in the *uspC-flhDC* intergenic region. In Chapter Six, experiments were performed to understand the mechanistic role of the site-specific insertion of the IS30 element on phenotypic resistance in *E. coli*, but the results only provided circumstantial evidence and not direct evidence of its adaptive importance to survival in a wastewater matrix. There are several limitations in this Chapter.

Firstly, only expression of the *flhDC* gene was measured but not the *uspC* gene. *uspC* and *flhDC* are divergently transcribed across the *E. coli* chromosome, and consequently the IS30 element may also play a role in regulating *uspC* transcription. The *uspC* gene product has been

shown to enhance flagellar production and motility in *E. coli* (Nachin, *et al*, 2005) and important in enhancing UV-resistance in *E. coli* (Gustavsson, *et al*, 2002). In a future study, *uspC* expression could be measured to better understand the phenotypic resistance of the naturalized wastewater *E. coli* to UV-disinfection in a serial stress experiment (i.e., use UV instead of chlorine as the final stress condition).

Secondly, bacterial motility was measured by a qPCR test on *flhDC* gene expression but not using a phenotypic test. The *flhDC* gene encodes the master regulator for flagellar biosynthesis, acting as an activator for expression of bacterial flagellar proteins (Smith and Hoover, 2009). Barker, *et al* (2004) observed that increased *flhDC* expression is associated with an increase in the expression of *flhB* which encodes a structural component of flagellar. Although *flhDC* expression is shown to be able to reflect bacterial motility, a phenotypic bacterial motility test is still needed. Motility of *E. coli* is under complex regulation, of which various environmental factors may impact the phenotypic outcomes. It is possible the wastewater strains need certain stress conditions (i.e. environmental factors in wastewater environment) to initiate global stress responses so that bacterial motility is increased and which can ultimately lead to increased biofilm formation and resistance to stress conditions. In a future study, phenotypic bacterial motility should be performed, in which the *E. coli* strains should be stress induced first by using conditions that mimic the wastewater environment.

Thirdly, the number of *E. coli* strains used in the stress experiment is limited. Only four strains were used in stress experiment. During the stress experiment of this thesis, three replicates were processed for each of the four *E. coli* strains in parallel and at each step of the stress experiment triplicates were performed for both *E. coli* colony counting and RNA extraction. Therefore the scale of this stress experiment is relatively large and two people had to

work cooperatively to finish it. Due to these reasons, only four strains were chosen for the stress experiment. In future studies, more strains should be tested for the stress resistance to ensure the phenomenon observed is universal to other *uspC-IS30-flhDC* positive and negative *E. coli* strains.

Lastly, future studies should focus on generating isogenic mutants lacking the IS30 element in the *uspC-flhDC* locus. This thesis provides circumstantial evidence of the functional role of the IS30 element in wastewater survival. Chapters Four, Five and Six provide a comprehensive evaluation of the genotypic/phenotypic properties of naturalized strains of wastewater *E. coli* possessing the *uspC-IS30-flhDC* marker, but a definitive mechanistic understanding of the role of the *uspC-IS30-flhDC* ITGR as an adaptive genotype for survival of naturalized *E. coli* populations in wastewater will require the generation of isogenic strains. This thesis provides the laboratory methods necessary to evaluate this (i.e., serial stress experiment conditions), in which the survivability of isogenic mutants can be determined against parent strains as well as what effect these mutations may have on *flhDC* gene expression and biofilm formation.

7.3 Impact of this thesis

I believe the findings of this thesis have greatly impacted our understanding of the microbial evolution of host specificity in *E. coli*. Host specificity is not well understood in *E. coli*, and it is widely believed that this bacterium is considered to be a host generalist, able to colonize and transmit between various animal hosts. Although evidence of host specificity in *E. coli* has been observed in several studies, the concept of host specificity of *E. coli* has not been studied using logic regression statistical analysis of DNA sequence in intergenic regions. The results of

this thesis demonstrate host related genetic SNP biomarker patterns were successfully identified in *E. coli* from various animal hosts.

Strains which only colonize the GI tract of one host can be regarded as specialists, while strains which have more than one host can be referred to as generalists. As stated in the hypothesis of this thesis, interactions between microbes and their host environment will eventually lead certain strains to become adapted to survive in a specific host environment. However, the research presented in this thesis does not exclude the fact that host generalist *E. coli* strains may also exist in any given host. High levels of microbial traffic between different host species may lead to the emergence of host generalists that are able to adapt and survive in multiple hosts. This may happen between different animal hosts that share the same physical environment. For example, microbial traffic is likely to be intense between aquatic birds sharing the same habitat, due to deposition of feces into the same water body used for drinking and eating of aquatic vegetation/food. Under these circumstances, some microbes may adapt to live in multiple host species in order to enhance their success of transmission and survival in an avian gastrointestinal system. A similar situation may occur between humans and companion animals, in which sufficient microbial traffic may lead to the emergence of host generalist strains capable of transmitting between multiple hosts. It is these host generalists that are likely to be associated with zoonotic disease transmission as they can transmit between different host species. Based on the statistical definition of a host-specialist used in this thesis, host specialists cannot be transmitted between animal hosts. This has important implications for understanding emerging *E. coli* zoonotic disease in humans. In Chapter Three of this thesis, 78% of *E. coli* isolates obtained from human subjects carried a human-specific SNP biomarker pattern. As such, *E. coli* isolates within this group may be considered host-specialists and capable of causing disease in humans

only (i.e., anthropogenic disease). Of the 22% of isolates that did not carry the human-specific SNP biomarker pattern, these isolates may represent host-generalists, and consequently comprise the potentially zoonotic population of *E. coli*. A similar argument can be made for cattle isolates - 92% of cattle isolates were considered host specialists, and therefore only 8% of the *E. coli* population in cattle may be considered as potentially zoonotic for humans. Although this value may seem low, it is important to note that the overall numbers of *E. coli* in any given host can approximate 10^7-10^9 bacteria/gram of feces. As such, any individual animal may carry zoonotic strains, but these strains may be numerically inferior to the host-specialist population in any given host. Alternatively, it is possible that only a certain proportion of cattle in a herd may carry zoonotic *E. coli* strains – a concept reflected in the 'super-shedder hypotheses' (Matthews, *et al*, 2006). Both of these divergent interpretations represent biologically plausible explanations for zoonoses - one cannot simply imply what the *prevalence* of *E. coli* host-generalists is in the overall population given the data presented in this thesis. This too represents an area of future research.

The data presented in this thesis do not negate the possibility that host specificity can also transpire through the acquisition of new genes, such as virulence genes. In fact, virulence genes can be considered as host-adaptation genes, since virulence genes in pathogenic strains of bacteria are identified based on their requisite for mediating attachment, invasion and/or intracellular survival in their host. However, a proper genetic background, characterized by host-specific ITGRs, may be needed to accommodate the introduction of new genes. For example, horizontal gene transfer is most efficient among closely related species (Stecher, *et al*, 2013). Incompatibility between the new gene and the microbe will likely disrupt the microbe's equilibrium state and affect its fitness in the original host environment. In one study, it was

demonstrated that the acquisition of the beta-lactamase gene, which is an antibiotic resistance gene, affected *Salmonella*'s ability to grow intracellularly in its host (Morosini, *et al*, 2000). As such, even if a host-specific gene can be introduced into a microbe, the biological cost incurred may affect its competitive genetic fitness and therefore the strain could be outcompeted by other host-specialists. Like all genes, virulence genes are subject to regulation, and the expression of their products (i.e., cytotoxins such as Shiga-toxin) is complex. As such, it is hypothesized that the genetic background of the microbe, as dictated by regulatory controls in host-specific ITGRs, may be equally as important as the acquisition of virulence genes themselves for zoonotic disease emergence, and leads to the following question: what is more important for zoonotic disease emergence: the virulence gene or the host-specific regulatory elements controlling expression of the virulence genes? This question is analogous to the proverbial adage - what came first: the chicken or the egg? Research in this area, by and large, has focused on the acquisition of virulence genes themselves, and little attention is paid to the ITGR genetic background of the aquiring strain. The data in this thesis suggests that ITGRs are extremely important in host adaptation/specificity and as such more research is needed in this area to better understand zoonotic disease emergence.

This thesis demonstrated that supervised-learning methods (logic regression) may be more powerful than traditional phylogenetic analysis for identifying host-specific patterns of DNA sequence data in *E. coli.* Application of supervised learning tools to genome data are becoming more common in cancer and chronic disease epidemiology in which SNP biomarkers are being used to infer health outcomes (Onay, *et al*, 2006; Dinu, *et al*, 2012). Applying supervised learning tools to DNA sequence analysis of ITGRs for evaluating host specificity is an extremely novel contribution and one that could be widely applied to various other microbial

systems to better understand host-specificity in terms of disease emergence in humans and animals.

References

Ackerley, D.F., Barak, Y., Lynch, S.V., Curtin, J., and Matin, A., 2006. Effect of chromate stress on *Escherichia coli* K-12. J. Bacteriol. 188, 3371-3381.

Adams, D.A., Riggs, M.M., and Donskey, C.J., 2007. Effect of fluoroquinolone treatment on growth of and toxin production by epidemic and nonepidemic *Clostridium difficile* strains in the cecal contents of mice. Antimicrobial Agents Chemother. 51, 2674-2678.

Adiri, R.S., Gophna, U., and Ron, E.Z., 2003. Multilocus sequence typing (MLST) of *Escherichia coli* O78 strains. FEMS Microbiol. Lett. 222, 199-203.

Ahmed, W., Masters, N., and Toze, S., 2012. Consistency in the host specificity and host sensitivity of the *Bacteroides* HF183 marker for sewage pollution tracking. Lett. Appl. Microbiol. 55, 283-289.

Albrecht, C., Geurts, R., and Bisseling, T., 1999. Legume nodulation and mycorrhizae formation two extremes in host specificity meet. EMBO J. 18, 281-288.

Allen, L.A. and Brooks, E., 1949. Destruction of bacteria in sewage and other liquids by chlorine and by cyanogen chloride. J. Hyg. (Lond) 47, 320-336.

Anastasi, E.M., Wohlsen, T.D., Stratton, H.M., and Katouli, M., 2013. Survival of *Escherichia coli* in two sewage treatment plants using UV irradiation and chlorination for disinfection. Water Res. 47, 6670-6679.

Anderson, J.A., 1995. An introduction to neural networks. MIT Press, Cambridge.

Anderson, K.L., Whitlock, J.E., and Harwood, V.J., 2005. Persistence and differential survival of fecal indicator bacteria in subtropical waters and sediments. Appl. Environ. Microbiol. 71, 3041-3048.

Andersson, A.F., Lindberg, M., Jakobsson, H., Backhed, F., Nyren, P., and Engstrand, L., 2008. Comparative analysis of human gut microbiota by barcoded pyrosequencing. PLoS One 3, e2836.

Ando, H., Kitao, T., Miyoshi-Akiyama, T., Kato, S., Mori, T., and Kirikae, T., 2011. Downregulation of katG expression is associated with isoniazid resistance in *Mycobacterium tuberculosis*. Mol. Microbiol. 79, 1615-1628.

Arsene, F., Tomoyasu, T., and Bukau, B., 2000. The heat shock response of *Escherichia coli*. Int. J. Food Microbiol. 55, 3-9.

Badea, L., Beatson, S.A., Kaparakis, M., Ferrero, R.L., and Hartland, E.L., 2009. Secretion of flagellin by the LEE-encoded type III secretion system of enteropathogenic *Escherichia coli*. BMC Microbiol. 9, 30.

Baez, A. and Shiloach, J., 2013. *Escherichia coli* avoids high dissolved oxygen stress by activation of SoxRS and manganese-superoxide dismutase. Microb. Cell. Fact. 12, 23.

Barbier, N., Saulnier, P., Chachaty, E., Dumontier, S., and Andremont, A., 1996. Random amplified polymorphic DNA typing versus pulsed-field gel electrophoresis for epidemiological typing of vancomycin-resistant *enterococci*. J. Clin. Microbiol. 34, 1096-1099.

Barker, C.S., Pruss, B.M., and Matsumura, P., 2004. Increased motility of *Escherichia coli* by insertion sequence element integration into the regulatory region of the *flhD* operon. J. Bacteriol. 186, 7529-7537.

Battesti, A., Majdalani, N., and Gottesman, S., 2011. The RpoS-mediated general stress response in *Escherichia coli*. Annu. Rev. Microbiol. 65, 189-213.

Bauchart, P., Germon, P., Bree, A., Oswald, E., Hacker, J., and Dobrindt, U., 2010. Pathogenomic comparison of human extraintestinal and avian pathogenic *Escherichia coli*-search for factors involved in host specificity or zoonotic potential. Microb. Pathog. 49, 105-115.

Bernhard, A.E. and Field, K.G., 2000a. Identification of nonpoint sources of fecal pollution in coastal waters by using host-specific 16S ribosomal DNA genetic markers from fecal anaerobes. Appl. Environ. Microbiol. 66, 1587-1594.

Bernhard, A.E. and Field, K.G., 2000b. A PCR assay To discriminate human and ruminant feces on the basis of host differences in *Bacteroides-Prevotella* genes encoding 16S rRNA. Appl. Environ. Microbiol. 66, 4571-4574.

Berthe, T., Ratajczak, M., Clermont, O., Denamur, E., and Petit, F., 2013. Evidence for coexistence of distinct *Escherichia coli* populations in various aquatic environments and their survival in estuary water. Appl. Environ. Microbiol. 79, 4684-4693.

Blanch, A.R., Belanche-Munoz, L., Bonjoch, X., Ebdon, J., Gantzer, C., Lucena, F., Ottoson, J., Kourtis, C., Iversen, A., Kuhn, I. *et al.,* 2006. Integrated analysis of established and novel microbial and chemical methods for microbial source tracking. Appl. Environ. Microbiol. 72, 5915-5926.

Blanch, A.R., Hagedorn, C., and Harwood, V.J., 2011. Microbial source tracking: methods, applications, and case studies. Springer, New York.

Blount, Z.D., Barrick, J.E., Davidson, C.J., and Lenski, R.E., 2012. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. Nature 489, 513-518.

Boehm, A.B., Van De Werfhorst, L.C., Griffith, J.F., Holden, P.A., Jay, J.A., Shanks, O.C., Wang, D., and Weisberg, S.B., 2013. Performance of forty-one microbial source tracking methods: a twenty-seven lab evaluation study. Water Res. 47, 6812-6828.

Boehm, A.B., Soller, J.A., and Shanks, O.C., 2015. Human-associated fecal quantitative polymerase chain reaction measurements and simulated risk of gastrointestinal illness in recreational waters contaminated with raw sewage. Environ. Sci. Technol. Lett. 2, 270-275.

Bojer, M.S., Hammerum, A.M., Jorgensen, S.L., Hansen, F., Olsen, S.S., Krogfelt, K.A., and Struve, C., 2012. Concurrent emergence of multidrug resistance and heat resistance by CTX-M-15-encoding conjugative plasmids in *Klebsiella pneumoniae*. APMIS 120, 699-705.

Bojer, M.S., Krogfelt, K.A., and Struve, C., 2011. The newly discovered ClpK protein strongly promotes survival of *Klebsiella pneumoniae* biofilm subjected to heat shock. J. Med. Microbiol. 60, 1559-1561.

Bojer, M.S., Struve, C., Ingmer, H., Hansen, D.S., and Krogfelt, K.A., 2010. Heat resistance mediated by a new plasmid encoded Clp ATPase, ClpK, as a possible novel mechanism for nosocomial persistence of *Klebsiella pneumoniae*. PLoS One 5, e15467.

Bonjoch, X., Balleste, E., and Blanch, A.R., 2005. Enumeration of bifidobacterial populations with selective media to determine the source of waterborne fecal pollution. Water Res. 39, 1621-1627.

Browning, D.F. and Busby, S.J., 2004. The regulation of bacterial transcription initiation. Nat. Rev. Microbiol. 2, 57-65.

Buck, M., Gallegos, M.T., Studholme, D.J., Guo, Y., and Gralla, J.D., 2000. The bacterial enhancer-dependent sigma(54) (sigma(N)) transcription factor. J. Bacteriol. 182, 4129-4136.

Byappanahalli, M.N. and Fujioka, R.S., 1998. Evidence that tropical soil environment can support the growth of *Escherichia coli*. Water Sci. Technol. 38, 171-174.

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

Carlos, C., Pires, M.M., Stoppe, N.C., Hachich, E.M., Sato, M.I., Gomes, T.A., Amaral, L.A., and Ottoboni, L.M., 2010. *Escherichia coli* phylogenetic group determination and its application in the identification of the major animal source of fecal contamination. BMC Microbiol. 10, 161.

Carrillo, M., Estrada, E., and Hazen, T.C., 1985. Survival and enumeration of the fecal indicators *Bifidobacterium adolescentis* and *Escherichia coli* in a tropical rain forest watershed. Appl. Environ. Microbiol. 50, 468-476.

Carson, C.A., Shear, B.L., Ellersieck, M.R., and Asfaw, A., 2001. Identification of fecal *Escherichia coli* from humans and animals by ribotyping. Appl. Environ. Microbiol. 67, 1503-1507.

Carson, C.A., Christiansen, J.M., Yampara-Iquise, H., Benson, V.W., Baffaut, C., Davis, J.V., Broz, R.R., Kurtz, W.B., Rogers, W.M., and Fales, W.H., 2005. Specificity of a *Bacteroides thetaiotaomicron* marker for human feces. Appl. Environ. Microbiol. 71, 4945-4949.

Carson, C.A., Shear, B.L., Ellersieck, M.R., and Schnell, J.D., 2003. Comparison of ribotyping and repetitive extragenic palindromic-PCR for identification of fecal *Escherichia coli* from humans and animals. Appl. Environ. Microbiol. 69, 1836-1839.

Casacuberta, E. and Gonzalez, J., 2013. The impact of transposable elements in environmental adaptation. Mol. Ecol. 22, 1503-1517.

Casarez, E.A., Pillai, S.D., and Di Giovanni, G.D., 2007. Genotype diversity of *Escherichia coli* isolates in natural waters determined by PFGE and ERIC-PCR. Water Res. 41, 3643-3648.

Caspers, P., Dalrymple, B., Iida, S., and Arber, W., 1984. IS30, a new insertion sequence of *Escherichia coli* K12. Mol. Gen. Genet. 196, 68-73.

Caugant, D.A., Levin, B.R., and Selander, R.K., 1981. Genetic diversity and temporal variation in the *E. coli* population of a human host. Genetics 98, 467-490.

Chandrasekaran, R., Hamilton, M.J., Wang, P., Staley, C., Matteson, S., Birr, A., and Sadowsky, M.J., 2015. Geographic isolation of *Escherichia coli* genotypes in sediments and water of the Seven Mile Creek - A constructed riverine watershed. Sci. Total Environ. 538, 78-85.

Chapman, T.A., Wu, X.Y., Barchia, I., Bettelheim, K.A., Driesen, S., Trott, D., Wilson, M., and Chin, J.J., 2006. Comparison of virulence gene profiles of *Escherichia coli* strains isolated from healthy and diarrheic swine. Appl. Environ. Microbiol. 72, 4782-4795.

Chen, S.L., Hung, C.S., Xu, J., Reigstad, C.S., Magrini, V., Sabo, A., Blasiar, D., Bieri, T., Meyer, R.R., Ozersky, P. *et al.,* 2006. Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. Proc. Natl. Acad. Sci. U. S. A. 103, 5977-5982.

Chesney, J.A., Eaton, J.W., and Mahoney, J.R., 1996. Bacterial glutathione: a sacrificial defense against chlorine compounds. J. Bacteriol. 178, 2131-2135.

Chiang, S.M., Dong, T., Edge, T.A., and Schellhorn, H.E., 2011. Phenotypic diversity caused by differential RpoS activity among environmental *Escherichia coli* isolates. Appl. Environ. Microbiol. 77, 7915-7923.

Chui, L., Couturier, M.R., Chiu, T., Wang, G., Olson, A.B., McDonald, R.R., Antonishyn, N.A., Horsman, G., and Gilmour, M.W., 2010. Comparison of Shiga toxin-producing *Escherichia coli* detection methods using clinical stool samples. J. Mol. Diagn. 12, 469-475.

Ciebin, B.W., Brodsky, M.H., Eddington, R., Horsnell, G., Choney, A., Palmateer, G., Ley, A., Joshi, R., and Shears, G., 1995. Comparative evaluation of modified m-FC and m-TEC media for membrane filter enumeration of *Escherichia coli* in water. Appl. Environ. Microbiol. 61, 3940-3942.

Claus, S.P., Ellero, S.L., Berger, B., Krause, L., Bruttin, A., Molina, J., Paris, A., Want, E.J., de Waziers, I., Cloarec, O. *et al.,* 2011. Colonization-induced host-gut microbial metabolic interaction. MBIO 2(2), e00271-10.

Clermont, O., Bonacorsi, S., and Bingen, E., 2000. Rapid and simple determination of the *Escherichia coli* phylogenetic group. Appl. Environ. Microbiol. 66, 4555-4558.

Clermont, O., Christenson, J.K., Denamur, E., and Gordon, D.M., 2013. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. Environ. Microbiol. Rep. 5, 58-65.

Clermont, O., Olier, M., Hoede, C., Diancourt, L., Brisse, S., Keroudean, M., Glodt, J., Picard, B., Oswald, E., and Denamur, E., 2011. Animal and human pathogenic *Escherichia coli* strains share common genetic backgrounds. Infect. Genet. Evol. 11, 654-662.

Clermont, O., Lescat, M., O'Brien, C.L., Gordon, D.M., Tenaillon, O., and Denamur, E., 2008. Evidence for a human-specific *Escherichia coli* clone. Environ. Microbiol. 10, 1000-1006.

Connell, I., Agace, W., Klemm, P., Schembri, M., Marild, S., and Svanborg, C., 1996. Type 1 fimbrial expression enhances *Escherichia coli* virulence for the urinary tract. Proc. Natl. Acad. Sci. U. S. A. 93, 9827-9832.

Cortes, C. and Vapink, V., 1995. Support-vector networks. Mach. Learning 20, 273-297.

Costello, E.K., Lauber, C.L., Hamady, M., Fierer, N., Gordon, J.I., and Knight, R., 2009. Bacterial community variation in human body habitats across space and time. Science 326, 1694-1697.

Dale, C., Young, S.A., Haydon, D.T., and Welburn, S.C., 2001. The insect endosymbiont *Sodalis glossinidius* utilizes a type III secretion system for cell invasion. Proc. Natl. Acad. Sci. U. S. A. 98, 1883-1888.

Dalrymple, B., 1987. Novel rearrangements of IS 30 carrying plasmids leading to the reactivation of gene expression. Mol. Gen. Genet. 207(2-3), 413-420.

De Beer, D., Srinivasan, R., and Stewart, P.S., 1994. Direct measurement of chlorine penetration into biofilms during disinfection. Appl. Environ. Microbiol. 60, 4339-4344.

De Biase, D. and Lund, P.A., 2015. The *Escherichia coli* acid stress response and its significance for pathogenesis. Adv. Appl. Microbiol. 92, 49-88.

Deer, D.M., Lampel, K.A., and Gonzalez-Escalona, N., 2010. A versatile internal control for use as DNA in real-time PCR and as RNA in real-time reverse transcription PCR assays. Lett. Appl. Microbiol. 50, 366-372.

Desjardins, P., Picard, B., Kaltenbock, B., Elion, J., and Denamur, E., 1995. Sex in *Escherichia coli* does not disrupt the clonal structure of the population: evidence from random amplified polymorphic DNA and restriction-fragment-length polymorphism. J. Mol. Evol. 41, 440-448.

Dethlefsen, L., McFall-Ngai, M., and Relman, D.A., 2007. An ecological and evolutionary perspective on human-microbe mutualism and disease. Nature 449, 811-818.

Dick, L.K., Bernhard, A.E., Brodeur, T.J., Santo Domingo, J.W., Simpson, J.M., Walters, S.P., and Field, K.G., 2005. Host distributions of uncultivated fecal *Bacteroidales* bacteria reveal genetic markers for fecal source identification. Appl. Environ. Microbiol. 71, 3184-3191.

Dinu, I., Mahasirimongkol, S., Liu, Q., Yanai, H., Sharaf Eldin, N., Kreiter, E., Wu, X., Jabbari, S., Tokunaga, K., and Yasui, Y., 2012. SNP-SNP interactions discovered by logic regression explain Crohn's disease genetics. PLoS One 7(10), e43035.

Djelouadji, Z., Arnold, C., Gharbia, S., Raoult, D., and Drancourt, M., 2008. Multispacer sequence typing for *Mycobacterium tuberculosis* genotyping. PLoS One 3(6), e2433.

Dobrindt, U., 2005. (Patho-)Genomics of *Escherichia coli*. Int. J. Med. Microbiol. 295, 357-371.

Dombek, P.E., Johnson, L.K., Zimmerley, S.T., and Sadowsky, M.J., 2000. Use of repetitive DNA sequences and the PCR to differentiate *Escherichia coli* isolates from human and animal sources. Appl. Environ. Microbiol. 66, 2572-2577.

Dominianni, C., Sinha, R., Goedert, J.J., Pei, Z., Yang, L., Hayes, R.B., and Ahn, J., 2015. Sex, body mass index, and dietary fiber intake influence the human gut microbiome. PLoS One 10, e0124599.

Drancourt, M., Roux, V., Dang, L.V., Tran-Hung, L., Castex, D., Chenal-Francisque, V., Ogata, H., Fournier, P.E., Crubezy, E., and Raoult, D., 2004. Genotyping, Orientalis-like *Yersinia pestis*, and plague pandemics. Emerg. Infect. Dis. 10, 1585-1592.

Dukan, S. and Touati, D., 1996. Hypochlorous acid stress in *Escherichia coli*: resistance, DNA damage, and comparison with hydrogen peroxide stress. J. Bacteriol. 178, 6145-6150.

Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E., and Relman, D.A., 2005. Diversity of the human intestinal microbial flora. Science 308, 1635-1638.

Edge, T.A. and Hill, S., 2007. Multiple lines of evidence to identify the sources of fecal pollution at a freshwater beach in Hamilton Harbour, Lake Ontario. Water Res. 41, 3585-3594.

Elasri, M.O. and Miller, R.V., 1999. Study of the response of a biofilm bacterial community to UV radiation. Appl. Environ. Microbiol. 65, 2025-2031.

Eliora, Z.R., Host specificity of septicemic *Escherichia coli*: human and avian pathogens. Curr. Opin. Microbiol. 9, 28-32.

Ellis, R.J., Bruce, K.D., Jenkins, C., Stothard, J.R., Ajarova, L., Mugisha, L., and Viney, M.E., 2013. Comparison of the distal gut microbiota from people and animals in Africa. PLoS One 8(1), e54783.

EPA, U.S., 2010. EPA Microbiological Alternate Test Procedure (ATP) Protocol for Drinking Water, Ambient Water, Wastewater and Sewage Sludge Monitoring Methods, US EPA, Washington, DC.

Escherich, T. 1885. Die darmbakterien des neugeboren und sauglingen. Fortschr. Med. 3:515–528.

Eswarappa, S.M., Janice, J., Nagarajan, A.G., Balasundaram, S.V., Karnam, G., Dixit, N.M., and Chakravortty, D., 2008. Differentially evolved genes of *Salmonella* pathogenicity islands: insights into the mechanism of host specificity in *Salmonella*. PLoS One 3(12), e3829.

Evenson, C.J. and Strevett, K.A., 2006. Discriminant analysis of fecal bacterial species composition for use as a phenotypic microbial source tracking method. Res. Microbiol. 157, 437-444.

Farkas-Himsley, H., 1964. Killing of chlorine-resistant bacteria by chlorine-bromine solutions. Appl. Microbiol. 12, 1-6.

Fauvart, M. and Michiels, J., 2008. Rhizobial secreted proteins as determinants of host specificity in the rhizobium–legume symbiosis. FEMS Microbiol. Lett. 285, 1-9.

Fayer, R., 2004. *Cryptosporidium*: a water-borne zoonotic parasite. Vet. Parasitol. 126, 37-56.

Fayer, R. and Xiao, L., 2007. *Cryptosporidium* and Cryptosporidiosis, second ed. CRC Press, Boca Raton, Florida.

Field, K.G. and Samadpour, M., 2007. Fecal source tracking, the indicator paradigm, and managing water quality. Water Res. 41, 3517-3538.

Figueira, V., Serra, E., and Manaia, C.M., 2011. Differential patterns of antimicrobial resistance in population subsets of *Escherichia coli* isolated from waste- and surface waters. Sci. Total Environ. 409, 1017-1023.

Fitch, W.M. and Margoliash, E., 1967. Construction of phylogenetic trees. Science 155, 279-284.

Foley, S.L., McDermott, P.F., Zhao, S., White, D.G., Simjee, S., and Meng, J., 2004. Evaluation of molecular typing methods for *Escherichia coli* O157:H7 isolates from cattle, food, and humans. J. Food Prot. 67, 651-657.

Foster, J.A., Krone, S.M., and Forney, L.J., 2008. Application of ecological network theory to the human microbiome. Interdiscip. Perspect. Infect. Dis. 2008, Article 839501.

Fournier, P.E. and Raoult, D., 2007. Identification of rickettsial isolates at the species level using multi-spacer typing. BMC Microbiol. 7, 72.

Fraune, S. and Bosch, T.C.G., 2007. Long-term maintenance of species-specific bacterial microbiota in the basal metazoan *Hydra*. Proc. Natl. Acad. Sci. U. S. A. 104, 13146-13151.

Fremaux, B., Boa, T., and Yost, C.K., 2010. Quantitative real-time PCR assays for sensitive detection of Canada goose-specific fecal pollution in water sources. Appl. Environ. Microbiol. 76, 4886-4889.

Freund, Y., Schapire, R., and Abe, N., 1999. A short introduction to boosting. J. Jpn. Soc. Artif. Intel. 14, 771-780.

Fu, H., Yuan, J., and Gao, H., 2015. Microbial oxidative stress response: novel insights from environmental facultative anaerobic bacteria. Arch. Biochem. Biophys. 584, 28-35.

Furukawa, T., Yoshida, T., and Suzuki, Y., 2011. Application of PFGE to source tracking of faecal pollution in coastal recreation area: a case study in Aoshima Beach, Japan. J. Appl. Microbiol. 110, 688-696.

Gavini, F., Pourcher, A.M., Neut, C., Monget, D., Romond, C., Oger, C., and Izard, D., 1991. Phenotypic differentiation of bifidobacteria of human and animal origins. Int. J. Syst. Bacteriol. 41, 548-557.

Geldreich, E.E. and Litsky, W., 1976. Fecal coliform and fecal streptococcus density relationships in waste discharges and receiving waters. Crit. Rev. Environ. Sci. Technol. 6, 349-369.

Gibson, S.V. and Wagner, J.E., 1986. Cryptosporidiosis in guinea pigs: a retrospective study. J. Am. Vet. Med. Assoc. 189, 1033-1034.

Giedraitiene, A., Vitkauskiene, A., Naginiene, R., and Pavilonis, A., 2011. Antibiotic resistance mechanisms of clinically important bacteria. Medicina (Kaunas) 47, 137-146.

Glazunova, O., Roux, V., Freylikman, O., Sekeyova, Z., Fournous, G., Tyczka, J., Tokarevich, N., Kovacava, E., Marrie, T.J., and Raoult, D., 2005. *Coxiella burnetii* genotyping. Emerg. Infect. Dis. 11, 1211-1217.

Glendinning, L. and Free, A., 2014. Supra-organismal interactions in the human intestine. Front. Cell. Infect. Microbiol. 4, 47.

Gordon, D.M., 2004. The influence of ecological factors on the distribution and genetic structure of *Escherichia coli.* EcoSal Plus.

Gordon, D.M., 1997. The genetic structure of *Escherichia coli* populations in feral house mice. Microbiology 143, 2039-2046.

Gordon, D.M. and Cowling, A., 2003. The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. Microbiology 149, 3575-3586.

Gordon, D.M., Clermont, O., Tolley, H., and Denamur, E., 2008. Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. Environ. Microbiol. 10, 2484-2496.

Gottesman, S., 2005. Micros for microbes: non-coding regulatory RNAs in bacteria. Trends Genet. 21, 399-404.

Gray, M.J., Wholey, W.Y., and Jakob, U., 2013. Bacterial responses to reactive chlorine species. Annu. Rev. Microbiol. 67, 141-160.

Grice, E.A. and Segre, J.A., 2012. The human microbiome: our second genome. Annu. Rev. Genomics Hum. Genet. 13, 151-170.

Griffith, J.F., Weisberg, S.B., and McGee, C.D., 2003. Evaluation of microbial source tracking methods using mixed fecal sources in aqueous test samples. J. Water. Health. 1, 141-151.

Grundmann, H., Schneider, C., Hartung, D., Daschner, F.D., and Pitt, T.L., 1995. Discriminatory power of three DNA-based typing techniques for *Pseudomonas aeruginosa*. J. Clin. Microbiol. 33, 528-534.

Guarner, F. and Malagelada JR., 2003. Gut flora in health and disease. The Lancet 361, 512-519.

Guerra, B., Fischer, J., and Helmuth, R., 2014. An emerging public health problem: acquired carbapenemase-producing microorganisms are present in food-producing animals, their environment, companion animals and wild birds. Vet. Microbiol. 171, 290-297.

Gustavsson, N., Diez, A., and Nystrom, T., 2002. The universal stress protein paralogues of *Escherichia coli* are co-ordinately regulated and co-operate in the defence against DNA damage. Mol. Microbiol. 43, 107-117.

Hamelin, K., Bruant, G., El-Shaarawi, A., Hill, S., Edge, T.A., Fairbrother, J., Harel, J., Maynard, C., Masson, L., and Brousseau, R., 2007. Occurrence of virulence and antimicrobial resistance genes in *Escherichia coli* isolates from different aquatic ecosystems within the St. Clair River and Detroit River areas. Appl. Environ. Microbiol. 73, 477-484.

Hartel, P.G., Summer, J.D., Hill, J.L., Collins, J.V., Entry, J.A., and Segars, W.I., 2002. Geographic variability of *Escherichia coli* ribotypes from animals in Idaho and Georgia. J. Environ. Qual. 31, 1273-1278.

Harwood, V.J., Staley, C., Badgley, B.D., Borges, K., and Korajkic, A., 2014. Microbial source tracking markers for detection of fecal contamination in environmental waters: relationships between pathogens and human health outcomes. FEMS Microbiol. Rev. 38, 1-40.

Harwood, V.J., Wiggins, B., Hagedorn, C., Ellender, R.D., Gooch, J., Kern, J., Samadpour, M., Chapman, A.C., Robinson, B.J., and Thompson, B.C., 2003. Phenotypic library-based microbial source tracking methods: efficacy in the California collaborative study. J. Water. Health. 1, 153-166.

Hassan, W.M., Wang, S.Y., and Ellender, R.D., 2005. Methods to increase fidelity of repetitive extragenic palindromic PCR fingerprint-based bacterial source tracking efforts. Appl. Environ. Microbiol. 71, 512-518.

Haugen, S.P., Ross, W., and Gourse, R.L., 2008. Advances in bacterial promoter recognition and its control by factors that do not bind DNA. Nat. Rev. Microbiol. 6, 507-519.

Haugland, R.A., Varma, M., Sivaganesan, M., Kelty, C., Peed, L., and Shanks, O.C., 2010. Evaluation of genetic markers from the 16S rRNA gene V2 region for use in quantitative detection of selected *Bacteroidales* species and human fecal waste by qPCR. Syst. Appl. Microbiol. 33, 348-357.

He, C. and Saedler, H., 2005. Heterotopic expression of MPF2 is the key to the evolution of the Chinese lantern of *Physalis*, a morphological novelty in *Solanaceae*. Proc. Natl. Acad. Sci. U. S. A. 102, 5779-5784.

Hengge-Aronis, R. and Fischer, D., 1992. Identification and molecular analysis of *glgS*, a novel growth-phase-regulated and rpoS-dependent gene involved in glycogen synthesis in *Escherichia coli*. Mol. Microbiol. 6, 1877-1886.

Henze, M., 2008. Biological Wastewater Treatment : Principles, Modelling and Design. IWA Pub, London.

Hu, Y. and Coates, A.R.M., 2005. Transposon mutagenesis identifies genes which control antimicrobial drug tolerance in stationary-phase *Escherichia coli*. FEMS Microbiol. Lett. 243, 117-124.

Hundesa, A., Maluquer de Motes, C., Bofill-Mas, S., Albinana-Gimenez, N., and Girones, R., 2006. Identification of human and animal adenoviruses and polyomaviruses for determination of sources of fecal contamination in the environment. Appl. Environ. Microbiol. 72, 7886-7893.

Iuchi, S. and Weiner, L., 1996. Cellular and molecular physiology of *Escherichia coli* in the adaptation to aerobic environments. J. Biochem. 120, 1055-1063.

Ivanetich, K.M., Hsu, P.H., Wunderlich, K.M., Messenger, E., Walkup, W.G.,4th, Scott, T.M., Lukasik, J., and Davis, J., 2006. Microbial source tracking by DNA sequence analysis of the *Escherichia coli* malate dehydrogenase gene. J. Microbiol. Methods 67, 507-526.

Iwane, T., Urase, T., and Yamamoto, K., 2001. Possible impact of treated wastewater discharge on incidence of antibiotic resistant bacteria in river water. Water Sci. Technol. 43, 91-99.

Jellison, K.L., Lynch, A.E., and Ziemann, J.M., 2009. Source tracking identifies deer and geese as vectors of human-infectious *Cryptosporidium* genotypes in an urban/suburban watershed. Environ. Sci. Technol. 43, 4267-4272.

Jenkins, M.B., Hartel, P.G., Olexa, T.J., and Stuedemann, J.A., 2003. Putative temporal variability of *Escherichia coli* ribotypes from yearling steers. J. Environ. Qual. 32, 305-309.

Jia, S., Shi, P., Hu, Q., Li, B., Zhang, T., and Zhang, X.X., 2015. Bacterial community shift drives antibiotic resistance promotion during drinking water chlorination. Environ. Sci. Technol. 49, 12271-12279.

Jiang, S., Noble, R., and Chu, W., 2001. Human adenoviruses and coliphages in urban runoff-impacted coastal waters of Southern California. Appl. Environ. Microbiol. 67, 179-184.

Jimenez-Clavero, M.A., Escribano-Romero, E., Mansilla, C., Gomez, N., Cordoba, L., Roblas, N., Ponz, F., Ley, V., and Saiz, J.C., 2005. Survey of bovine enterovirus in biological and environmental samples by a highly sensitive real-time reverse transcription-PCR. Appl. Environ. Microbiol. 71, 3536-3543.

Jimenez-Clavero, M.A., Fernandez, C., Ortiz, J.A., Pro, J., Carbonell, G., Tarazona, J.V., Roblas, N., and Ley, V., 2003. Teschoviruses as indicators of porcine fecal contamination of surface water. Appl. Environ. Microbiol. 69, 6311-6315.

Johnson, J.R. and Stell, A.L., 2000. Extended virulence genotypes of *Escherichia coli* strains from patients with urosepsis in relation to phylogeny and host compromise. J. Infect. Dis. 181, 261-272.

Kanjee, U. and Houry, W.A., 2013. Mechanisms of acid resistance in *Escherichia coli*. Annu. Rev. Microbiol. 67, 65-81.

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

Kaper, J.B. and O'Brien, A.D., 1998. *Escherichia coli* 0157:H7 and other Shiga toxin-producing *E. coli strains*. ASM Press, Washington, DC.

Khan, S., Beattie, T.K., and Knapp, C.W., 2016. Relationship between antibiotic- and disinfectant-resistance profiles in bacteria harvested from tap water. Chemosphere 152, 132-141.

Khatib, Tsai, and Olson, 2002. A biomarker for the identification of cattle fecal pollution in water using the LTIIa toxin gene from enterotoxigenic *Escherichia coli*. Appl. Microbiol. Biotechnol. 59, 97-104.

Khatib, L.A., Tsai, Y.L., and Olson, B.H., 2003. A biomarker for the identification of swine fecal pollution in water, using the STII toxin gene from enterotoxigenic *Escherichia coli*. Appl. Microbiol. Biotechnol. 63, 231-238.

Kim, J., Nietfeldt, J., and Benson, A.K., 1999. Octamer-based genome scanning distinguishes a unique subpopulation of *Escherichia coli* O157:H7 strains in cattle. Proc. Natl. Acad. Sci. U. S. A. 96, 13288-13293.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, New York.

Klaasen, H.L., Van der Heijden,P.J., Stok, W., Poelma, F.G., Koopman, J.P., Van den Brink,M.E., Bakker, M.H., Eling, W.M., and Beynen, A.C., 1993. Apathogenic, intestinal, segmented, filamentous bacteria stimulate the mucosal immune system of mice. Infect. Immun. 61, 303-306.

Kon, T., Weir, S.C., Trevors, J.T., Lee, H., Champagne, J., Meunier, L., Brousseau, R., and Masson, L., 2007. Microarray analysis of *Escherichia coli* strains from interstitial beach waters of Lake Huron (Canada). Appl. Environ. Microbiol. 73, 7757-7758.

Kotetishvili, M., Stine, O.C., Kreger, A., Morris, J.G.,Jr, and Sulakvelidze, A., 2002. Multilocus sequence typing for characterization of clinical and environmental salmonella strains. J. Clin. Microbiol. 40, 1626-1635.

Krisko, A., Copic, T., Gabaldon, T., Lehner, B., and Supek, F., 2014. Inferring gene function from evolutionary change in signatures of translation efficiency. Genome Biol. 15, R44.

Kvac, M. and Vitovec, J., 2003. Prevalence and pathogenicity of *Cryptosporidium andersoni* in one herd of beef cattle. J. Vet. Med. B Infect. Dis. Vet. Public Health 50, 451-457.

Kwan, B.W., Lord, D.M., Peti, W., Page, R., Benedik, M.J., and Wood, T.K., 2015. The MqsR/MqsA toxin/antitoxin system protects *Escherichia coli* during bile acid stress. Environ. Microbiol. 17, 3168-3181.

Lacour, S. and Landini, P., 2004. SigmaS-dependent gene expression at the onset of stationary phase in *Escherichia coli*: function of sigmaS-dependent genes and identification of their promoter sequences. J. Bacteriol. 186, 7186-7195.

Langkjaer, R.B., Vigre, H., Enemark, H.L., and Maddox-Hyttel, C., 2007. Molecular and phylogenetic characterization of *Cryptosporidium* and *Giardia* from pigs and cattle in Denmark. Parasitology 134, 339-350.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R. *et al.,* 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947-2948.

Larsen, N., Vogensen, F.K., van, d.B., Nielsen, D.S., Andreasen, A.S., Pedersen, B.K., Abu Al-Soud, W., Sorensen, S.J., Hansen, L.H., and Jakobsen, M., 2010. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. PLoS One 5, e9085 .

Layton, B.A., Walters, S.P., and Boehm, A.B., 2009. Distribution and diversity of the enterococcal surface protein (esp) gene in animal hosts and the Pacific coast environment. J. Appl. Microbiol. 106, 1521-1531.

Leclerc, H., Schwartzbrod, L., and Dei-Cas, E., 2002. Microbial agents associated with waterborne diseases. Crit. Rev. Microbiol. 28, 371-409.

Lee, C., Marion, J.W., and Lee, J., 2013. Development and application of a quantitative PCR assay targeting *Catellicoccus marimammalium* for assessing gull-associated fecal contamination at Lake Erie beaches. Sci. Total Environ. 454-455, 1-8.

Lee, S.H. and Kim, S.J., 2002. Detection of infectious enteroviruses and adenoviruses in tap water in urban areas in Korea. Water Res. 36, 248-256.

Lerouge, P., Roche, P., Faucher, C., Maillet, F., Truchet, G., Prome, J.C., and Denarie J., 1990. Symbiotic host-specificity of *Rhizobium meliloti* is determined by a sulphated and acylated glucosamine oligosaccharide signal. Nature 344, 781-784.

Letunic, I. and Bork, P., 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 39, W475-478.

Letunic, I. and Bork, P., 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23, 127-128.

Leung, K.T., Mackereth, R., Tien, Y.C., and Topp, E., 2004. A comparison of AFLP and ERIC-PCR analyses for discriminating *Escherichia coli* from cattle, pig and human sources. FEMS Microbiol. Ecol. 47, 111-119.

Ley, R.E., Hamady, M., Lozupone, C., Turnbaugh, P.J., Ramey, R.R., Bircher, J.S., Schlegel, M.L., Tucker, T.A., Schrenzel, M.D., Knight, R. *et al.,* 2008. Evolution of mammals and their gut microbes. Science 320, 1647-1651.

Ley, R.E., Peterson, D.A., and Gordon, J.I., 2006. Review: ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell 124, 837-848.

Li, W., Chomel, B.B., Maruyama, S., Guptil, L., Sander, A., Raoult, D., and Fournier, P.E., 2006. Multispacer typing to study the genotypic distribution of *Bartonella henselae* populations. J. Clin. Microbiol. 44, 2499-2506.

Litrup, E., Torpdahl, M., and Nielsen, E.M., 2007. Multilocus sequence typing performed on *Campylobacter coli* isolates from humans, broilers, pigs and cattle originating in Denmark. J. Appl. Microbiol. 103, 210-218.

Liu, W.T., Karavolos, M.H., Bulmer, D.M., Allaoui, A., Hormaeche, R.D., Lee, J.J., and Khan, C.M., 2007. Role of the universal stress protein UspA of *Salmonella* in growth arrest, stress and virulence. Microb. Pathog. 42, 2-10.

Long, S.C., El-Khoury, S., Oudejans, S.J.G., Sobsey, M.D., and Vinjé, J., 2005. Assessment of sources and diversity of male-specific coliphages for source tracking. Environ. Eng. Sci. 22, 367-377.

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

Lundin, A., Bok, C.M., Aronsson, L., Bjorkholm, B., Gustafsson, J.A., Pott, S., Arulampalam, V., Hibberd, M., Rafter, J., and Pettersson, S., 2008. Gut flora, Toll-like receptors and nuclear receptors: a tripartite communication that tunes innate immunity in large intestine. Cell. Microbiol. 10, 1093-1103.

Luthje, P. and Brauner, A., 2014. Virulence factors of uropathogenic *E. coli* and their interaction with the host. Adv. Microb. Physiol. 65, 337-372.

Lyautey, E., Lu, Z., Lapen, D.R., Berkers, T.E., Edge, T.A., and Topp, E., 2010. Optimization and validation of rep-PCR genotypic libraries for microbial source tracking of environmental *Escherichia coli* isolates. Can. J. Microbiol. 56, 8-17.

Kosoy, M.Y, Saito, E.K, Green, D., Marston, E.L, Jones, D.C, and Childs, J.E, 2000. Experimental evidence of host specificity of *Bartonella* infection in rodents. Comp. Immunol. Microbiol. Infect. Dis. 23, 221-238.

Maluquer de Motes, C., Clemente-Casares, P., Hundesa, A., Martin, M., and Girones, R., 2004. Detection of bovine and porcine adenoviruses for tracing the source of fecal contamination. Appl. Environ. Microbiol. 70, 1448-1454.

Mandel, M.J., Wollenberg, M.S., Stabb, E.V., Visick, K.L., and Ruby, E.G., 2009. A single regulatory gene is sufficient to alter bacterial host range. Nature 458, 215-218.

Mangoli, S., Sanzgiri, V.R., and Mahajan, S.K., 2001. A common regulator of cold and radiation response in *Escherichia coli*. J. Environ. Pathol. Toxicol. Oncol. 20, 23-26.

Mara, D.D. and Oragui, J.I., 1983. Sorbitol-fermenting bifidobacteria as specific indicators of human faecal pollution. J. Appl. Bacteriol. 55, 349-357.

Mara, D.D. and Oragui, J.I., 1981. Occurrence of *Rhodococcus coprophilus* and associated actinomycetes in feces, sewage, and freshwater. Appl. Environ. Microbiol. 42, 1037-1042.

Marieb, E.N. and Hoehn, K., 2010. Human Anatomy & Physiology. Benjamin Cummings, San Francisco.

Marti, R., Zhang, Y., Lapen, D.R., and Topp, E., 2011. Development and validation of a microbial source tracking marker for the detection of fecal pollution by muskrats. J. Microbiol. Methods 87, 82-88.

Martinez, J.J. and Hultgren, S.J., 2002. Requirement of Rho-family GTPases in the invasion of Type 1-piliated uropathogenic *Escherichia coli*. Cell. Microbiol. 4, 19-28.

Matthews, L., McKendrick, I.J., Ternent, H., Gunn, G.J., Synge, B., and Woolhouse, M.E., 2006. Super-shedding cattle and the transmission dynamics of *Escherichia coli* O157. Epidemiol. Infect. 134, 131-142.

McCarter, L.L., 2006. Regulation of flagella. Curr. Opin. Microbiol. 9, 180-186.

McCullagh, P. and Nelder, J.A., 1989. Generalized Linear Models, 2nd ed. Chapman and Hall, London.

McDaniel, T.K., Jarvis, K.G., Donnenberg, M.S., and Kaper, J.B., 1995. A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens. Proc. Natl. Acad. Sci. U. S. A. 92, 1664-1668.

McQuaig, S., Griffith, J., and Harwood, V.J., 2012. Association of fecal indicator bacteria with human viruses and microbial source tracking markers at coastal beaches impacted by nonpoint source pollution. Appl. Environ. Microbiol. 78, 6423-6432.

Mellata, M., Ameiss, K., Mo, H., and Curtiss III, R., 2010. Characterization of the contribution to virulence of three large plasmids of avian pathogenic *Escherichia coli* chi7122 (O78:K80:H9). Infect. Immun. 78, 1528-1541.

Mercer, R.G., Zheng, J., Garcia-Hernandez, R., Ruan, L., Ganzle, M.G., and McMullen, L.M., 2015. Genetic determinants of heat resistance in *Escherichia coli*. Front. Microbiol. 6, Article 932.

Milkman, R., 1973. Electrophoretic variation in *Escherichia coli* from natural sources. Science 182, 1024-1026.

Miller, S.H., Elliot, R.M., Sullivan, J.T., and Ronson, C.W., 2007. Host-specific regulation of symbiotic nitrogen fixation in *Rhizobium leguminosarum* biovar *trifolii*. Microbiology 153, 3184-3195.

Miller, W.G., Englen, M.D., Kathariou, S., Wesley, I.V., Wang, G., Pittenger-Alley, L., Siletz, R.M., Muraoka, W., Fedorka-Cray, P.J., and Mandrell, R.E., 2006. Identification of host-associated alleles by multilocus sequence typing of *Campylobacter coli* strains from food animals. Microbiology 152, 245-255.

Miskinyte, M., Sousa, A., Ramiro, R.S., de Sousa, J.A., Kotlinowski, J., Caramalho, I., Magalhaes, S., Soares, M.P., and Gordo, I., 2013. The genetic basis of *Escherichia coli* pathoadaptation to macrophages. PLoS Pathog. 9, e1003802.

Mohapatra, B.R., Broersma, K., and Mazumder, A., 2008. Differentiation of fecal *Escherichia coli* from poultry and free-living birds by (GTG)5-PCR genomic fingerprinting. Int. J. Med. Microbiol. 298, 245-252.

Mohri, M., Talwalkar, A., and Rostamizadeh, A., 2012. Foundations of Machine Learning. MIT Press, Cambridge, MA.

Molina, M., Hunter, S., Cyterski, M., Peed, L.A., Kelty, C.A., Sivaganesan, M., Mooney, T., Prieto, L., and Shanks, O.C., 2014. Factors affecting the presence of human-associated and fecal indicator real-time quantitative PCR genetic markers in urban-impacted recreational beaches. Water Res. 64, 196-208.

Mooi, F.R., van Loo, I.H., van Gent, M., He, Q., Bart, M.J., Heuvelman, K.J., de Greeff, S.C., Diavatopoulos, D., Teunis, P., Nagelkerke, N. *et al.,* 2009. *Bordetella pertussis* strains with increased toxin production associated with pertussis resurgence. Emerg. Infect. Dis. 15, 1206-1213.

Moore, D.F., Harwood, V.J., Ferguson, D.M., Lukasik, J., Hannah, P., Getrich, M., and Brownell, M., 2005. Evaluation of antibiotic resistance analysis and ribotyping for identification of faecal pollution sources in an urban watershed. J. Appl. Microbiol. 99, 618-628.

Moreau, M.C. and Gaboriau-Routhiau, V., 1996. The absence of gut flora, the doses of antigen ingested and aging affect the long-term peripheral tolerance induced by ovalbumin feeding in mice. Res. Immunol. 147, 49-59.

Morosini, M.I., Ayala, J.A., Baquero, F., Martinez, J.L., and Blazquez, J., 2000. Biological cost of AmpC production for *Salmonella enterica* serotype Typhimurium. Antimicrob. Agents Chemother. 44, 3137-3143.

Murray, G.E., Tobin, R.S., Junkins, B., and Kushner, D.J., 1984. Effect of chlorination on antibiotic resistance profiles of sewage-related bacteria. Appl. Environ. Microbiol. 48, 73-77.

Myoda, S.P., Carson, C.A., Fuhrmann, J.J., Hahm, B.K., Hartel, P.G., Yampara-Lquise, H., Johnson, L., Kuntz, R.L., Nakatsu, C.H., Sadowsky, M.J. *et al.,* 2003. Comparison of genotypic-based microbial source tracking methods requiring a host origin database. J. Water. Health. 1, 167-180.

Nachin, L., Nannmark, U., and Nystrom, T., 2005. Differential roles of the universal stress proteins of *Escherichia coli* in oxidative stress resistance, adhesion, and motility. J. Bacteriol. 187, 6265-6272.

Nakata, N., Tobe, T., Fukuda, I., Suzuki, T., Komatsu, K., Yoshikawa, M., and Sasakawa, C., 1993. The absence of a surface protease, OmpT, determines the intercellular spreading ability of *Shigella*: the relationship between the ompT and kcpA loci. Mol. Microbiol. 9, 459-468.

Nallapareddy, S.R., Duh, R.W., Singh, K.V., and Murray, B.E., 2002. Molecular typing of selected *Enterococcus faecalis* isolates: pilot study using multilocus sequence typing and pulsed-field gel electrophoresis. J. Clin. Microbiol. 40, 868-876.

Nambu, T. and Kutsukake, K., 2000. The *Salmonella* FlgA protein, a putativeve periplasmic chaperone essential for flagellar P ring formation. Microbiology 146 (Pt 5), 1171-1178.

Neumann, G., Noda, T., and Kawaoka, Y., 2009. Emergence and pandemic potential of swine-origin H1N1 influenza virus. Nature 459, 931-939.

Neuwald, A.F. and Stauffer, G.V., 1990. IS30 activation of an *smp'-Lacz* Gene fusion in *Escherichia coli*. FEMS Microbiol. Lett. 68, 13-18.

Nguyen, T.V., Le Van, P., Le Huy, C., Gia, K.N., and Weintraub, A., 2005. Detection and characterization of diarrheagenic *Escherichia coli* from young children in Hanoi, Vietnam. J. Clin. Microbiol. 43, 755-760.

Nielsen, R., 2005. Molecular signatures of natural selection. Annu. Rev. Genet. 39, 197-218.

Noble, R.T., Allen, S.M., Blackwood, A.D., Chu, W., Jiang, S.C., Lovelace, G.L., Sobsey, M.D., Stewart, J.R., and Wait, D.A., 2003. Use of viral pathogens and indicators to differentiate between human and non-human fecal contamination in a microbial source tracking comparison study. J. Water. Health. 1, 195-207.

Nocker, A., Cheung, C.Y., and Camper, A.K., 2006. Comparison of propidium monoazide with ethidium monoazide for differentiation of live vs. dead bacteria by selective removal of DNA from dead cells. J. Microbiol. Methods 67, 310-320.

Odagiri, M., Schriewer, A., Hanley, K., Wuertz, S., Misra, P.R., Panigrahi, P., and Jenkins, M.W., 2015. Validation of *Bacteroidales* quantitative PCR assays targeting human and animal fecal contamination in the public and domestic domains in India. Sci. Total Environ. 502, 462-470.

Onay, V.U., Briollais, L., Knight, J.A., Shi, E., Wang, Y., Wells, S., Li, H., Rajendram, I., Andrulis, I.L., and Ozcelik, H., 2006. SNP-SNP interactions in breast cancer susceptibility. BMC Cancer 6, 114.

O'Toole, G.A., 2011. Microtiter dish biofilm formation assay. J. Vis. Exp. 47, 2437.

O'Toole, G.A. and Kolter, R., 1998. Flagellar and twitching motility are necessary for *Pseudomonas aeruginosa* biofilm development. Mol. Microbiol. 30, 295-304.

Parker, B.W., Schwessinger, E.A., Jakob, U., and Gray, M.J., 2013. The RclR protein is a reactive chlorine-specific transcription factor in *Escherichia coli*. J. Biol. Chem. 288, 32574-32584.

Parveen, S., Hodge, N.C., Stall, R.E., Farrah, S.R., and Tamplin, M.L., 2001. Phenotypic and genotypic characterization of human and nonhuman *Escherichia coli*. Water Res. 35, 379-386.

Peacock, S.J., de Silva, G.D., Justice, A., Cowland, A., Moore, C.E., Winearls, C.G., and Day, N.P., 2002. Comparison of multilocus sequence typing and pulsed-field gel electrophoresis as tools for typing *Staphylococcus aureus* isolates in a microepidemiological setting. J. Clin. Microbiol. 40, 3764-3770.

Picard, B., Garcia, J.S., Gouriou, S., Duriez, P., Brahimi, N., Bingen, E., Elion, J., and Denamur, E., 1999. The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. Infect. Immun. 67, 546-553.

Poirel, L., Bonnin, R.A., and Nordmann, P., 2011. Analysis of the resistome of a multidrug-resistant NDM-1-producing *Escherichia coli* strain by high-throughput genome sequencing. Antimicrob. Agents Chemother. 55, 4224-4229.

Poveda, J.B., Giebel, J., Flossdorf, J., and Meier, J., 1994. *Mycoplasma buteonis sp.* nov., *Mycoplasma falconis sp.* nov., and *Mycoplasma gypis sp.* nov., Three Species from Birds of Prey. Int. J. Syst. Bacteriol. 44, 94.

Power, M.L., Littlefield-Wyer, J., Gordon, D.M., Veal, D.A., and Slade, M.B., 2005. Phenotypic and genotypic characterization of encapsulated *Escherichia coli* isolated from blooms in two Australian lakes. Environ. Microbiol. 7, 631-640.

Prüss, B.,M., Besemann, C., Denton, A., and Wolfe, A.J., 2006. A complex transcription network controls the early stages of biofilm development by *Escherichia coli*. J. Bacteriol. 188, 3731-3739.

Pupo, G.M., Karaolis, D.K., Lan, R., and Reeves, P.R., 1997. Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and mdh sequence studies. Infect. Immun. 65, 2685-2692.

Qadri, F., Svennerholm, A., Faruque, A.S.G., and Sack, R.B., 2005. Enterotoxigenic *Escherichia coli* in developing countries: epidemiology, microbiology, clinical features, treatment, and prevention. Clin. Microbiol. Rev. 18, 465-483.

Ram, J.L., Ritchie, R.P., Fang, J., Gonzales, F.S., and Selegean, J.P., 2004. Sequence-based source tracking of *Escherichia coli* based on genetic diversity of beta-glucuronidase. J. Environ. Qual. 33, 1024-1032.

Ratajczak, M., Laroche, E., Berthe, T., Clermont, O., Pawlak, B., Denamur, E., and Petit, F., 2010. Influence of hydrological conditions on the *Escherichia coli* population structure in the water of a creek on a rural watershed. BMC Microbiol. 10, 222.

Rawls, J.F., Mahowald, M.A., Ley, R.E., and Gordon, J.I., 2006. Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. Cell 127, 423-433.

Record, M.T.,Jr, Courtenay, E.S., Cayley, D.S., and Guttman, H.J., 1998. Responses of *E. coli* to osmotic stress: large changes in amounts of cytoplasmic solutes and water. Trends Biochem. Sci. 23, 143-148.

Resnick, I.G. and Levin, M.A., 1981. Assessment of Bifidobacteria as indicators of human fecal pollution. Appl. Environ. Microbiol. 42, 433-438.

Revazishvili, T., Kotetishvili, M., Stine, O.C., Kreger, A.S., Morris, J.G.,Jr, and Sulakvelidze, A., 2004. Comparative analysis of multilocus sequence typing and pulsed-field gel electrophoresis for characterizing *Listeria monocytogenes* strains isolated from environmental and clinical sources. J. Clin. Microbiol. 42, 276-285.

Reynolds, K.A., Mena, K.D., and Gerba, C.P., 2008. Risk of waterborne illness via drinking water in the United States. Rev. Environ. Contam. Toxicol. 192, 117-158.

Rhodes, M.W. and Kator, H., 1999. Sorbitol-fermenting bifidobacteria as indicators of diffuse human faecal pollution in estuarine watersheds. J. Appl. Microbiol. 87, 528-535.

Rice E.W., Baird, R.B., Eaton, A.D., and Clesceri L.S., 2012. Standard Methods for the Examination of Water and Wastewater, American Public Health Association, Washington, DC.

Rodrigues, D.F. and Elimelech, M., 2009. Role of type 1 fimbriae and mannose in the development of *Escherichia coli* K12 biofilm: from initial cell adhesion to biofilm formation. Biofouling 25, 401-411.

Romling, U., Sierralta, W.D., Eriksson, K., and Normark, S., 1998. Multicellular and aggregative behaviour of *Salmonella typhimurium* strains is controlled by mutations in the *agfD* promoter. Mol. Microbiol. 28, 249-264.

Rosen, H., Klebanoff, S.J., Wang, Y., Brot, N., Heinecke, J.W., and Fu, X., 2009. Methionine oxidation contributes to bacterial killing by the myeloperoxidase system of neutrophils. Proc. Natl. Acad. Sci. U. S. A. 106, 18686-18691.

Roux, A., Beloin, C., and Ghigo, J.M., 2005. Combined inactivation and expression strategy to study gene function under physiological conditions: application to identification of new *Escherichia coli* adhesins. J. Bacteriol. 187, 1001-1013.

Rowbotham, T.J. and Cross, T., 1977. Ecology of *Rhodococcus coprophilus* and associated actinomycetes in fresh water and agricultural habitats. J. Gen. Microbiol. 100, 231-240.

Ruczinski, I., Kooperberg, C., and L. LeBlanc, M., 2004. Exploring interactions in high-dimensional genomic data: an overview of logic regression with applications. J. Multivar. Anal. 90, 178-195.

Ruecker, N.J., Braithwaite, S.L., Topp, E., Edge, T., Lapen, D.R., Wilkes, G., Robertson, W., Medeiros, D., Sensen, C.W., and Neumann, N.F., 2007. Tracking host sources of *Cryptosporidium spp.* in raw water for improved health risk assessment. Appl. Environ. Microbiol. 73, 3945-3957.

Ruecker, N.J., Matsune, J.C., Wilkes, G., Lapen, D.R., Topp, E., Edge, T.A., Sensen, C.W., Xiao, L., and Neumann, N.F., 2012. Molecular and phylogenetic approaches for assessing sources of *Cryptosporidium* contamination in water. Water Res. 46, 5135-5150.

Rusinol, M., Moriarty, E., Lin, S., Bofill-Mas, S., and Gilpin, B., 2016. Human-, ovine-, and bovine-specific viral source tracking tools to discriminate between the major fecal sources in agricultural waters. Food Environ. Virol. 8, 34-45.

Ryu, J.H. and Beuchat, L.R., 2005. Biofilm formation by *Escherichia coli* O157:H7 on stainless steel: effect of exopolysaccharide and Curli production on its resistance to chlorine. Appl. Environ. Microbiol. 71, 247-254.

Rzhetsky, A. and Nei, M., 1992. A simple method for estimating and testing minimum-Evolution trees. Mol.Biol. Evol. 9, 945.

Sabate, M., Prats, G., Moreno, E., Balleste, E., Blanch, A.R., and Andreu, A., 2008. Virulence and antimicrobial resistance profiles among *Escherichia coli* strains isolated from human and animal wastewater. Res. Microbiol. 159, 288-293.

Saby, S., Leroy, P., and Block, J.C., 1999. *Escherichia coli* resistance to chlorine and glutathione synthesis in response to oxygenation and starvation. Appl. Environ. Microbiol. 65, 5600-5603.

Sadowsky, M.J. and Santo Domingo, J.W., 2007. Microbial Source Tracking. ASM Press, Washington, D.C.

Saitou, N. and Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406-425.

Samadpour, M., Roberts, M.C., Kitts, C., Mulugeta, W., and Alfi, D., 2005. The use of ribotyping and antibiotic resistance patterns for identification of host sources of *Escherichia coli* strains. Lett. Appl. Microbiol. 40, 63-68.

Santo Domingo, J.W., Bambic, D.G., Edge, T.A., and Wuertz, S., 2007. Quo vadis source tracking? Towards a strategic framework for environmental monitoring of fecal pollution. Water Res. 41, 3539-3552.

Savill, M.G., Murray, S.R., Scholes, P., Maas, E.W., McCormick, R.E., Moore, E.B., and Gilpin, B.J., 2001. Application of polymerase chain reaction (PCR) and TaqMan PCR techniques to the detection and identification of *Rhodococcus coprophilus* in faecal samples. J. Microbiol. Methods 47, 355-368.

Schlesinger, M.J., 1990. Heat shock proteins. J. Biol. Chem. 265, 12111-12114.

Schofield, P.R. and Watson, J.M., 1986. DNA sequence of *Rhizobium trifolii* nodulation genes reveals a reiterated and potentially regulatory sequence preceding nodABC and nodFE. Nucleic Acids Res. 14, 2891-2903.

Scott, T.M., Jenkins, T.M., Lukasik, J., and Rose, J.B., 2005. Potential use of a host associated molecular marker in *Enterococcus faecium* as an index of human fecal pollution. Environ. Sci. Technol. 39, 283-287.

Scott, T.M., Parveen, S., Portier, K.M., Rose, J.B., Tamplin, M.L., Farrah, S.R., Koo, A., and Lukasik, J., 2003. Geographical variation in ribotype profiles of *Escherichia coli* isolates from humans, swine, poultry, beef, and dairy cattle in Florida. Appl. Environ. Microbiol. 69, 1089-1092.

Seifert, H.S., 1996. Questions about gonococcal pilus phase- and antigenic variation. Mol. Microbiol. 21, 433-440.

Shanks, O.C., Atikovic, E., Blackwood, A.D., Lu, J., Noble, R.T., Domingo, J.S., Seifring, S., Sivaganesan, M., and Haugland, R.A., 2008. Quantitative PCR for detection and enumeration of genetic markers of bovine fecal pollution. Appl. Environ. Microbiol. 74, 745-752.

Shanks, O.C., Kelty, C.A., Sivaganesan, M., Varma, M., and Haugland, R.A., 2009. Quantitative PCR for genetic markers of human fecal pollution. Appl. Environ. Microbiol. 75, 5507-5513.

Sharma, V.K. and Bearson, B.L., 2013. Hha controls *Escherichia coli* O157:H7 biofilm formation by differential regulation of global transcriptional regulators FlhDC and CsgD. Appl. Environ. Microbiol. 79, 2384-2396.

Sheldon, J.R., Yim, M.S., Saliba, J.H., Chung, W.H., Wong, K.Y., and Leung, K.T., 2012. Role of rpoS in *Escherichia coli* O157:H7 strain H32 biofilm development and survival. Appl. Environ. Microbiol. 78, 8331-8339.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M., 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 34, D32-36.

Sinton, L., Finlay, R., and Hannah, D., 1998. Distinguishing human from animal faecal contamination in water: a review. N. Z. J. Mar. Freshwat. Res. 32, 323-348.

Skyberg, J.A., Johnson, T.J., Johnson, J.R., Clabots, C., Logue, C.M., and Nolan, L.K., 2006. Acquisition of avian pathogenic *Escherichia coli* plasmids by a commensal *E. coli* isolate enhances its abilities to kill chicken embryos, grow in human urine, and colonize the murine kidney. Infect. Immun. 74, 6287-6292.

Smith, T.G. and Hoover, T.R., 2009. Deciphering bacterial flagellar gene regulatory networks in the genomic era. Adv. Appl. Microbiol. 67, 257-295.

Smits, W.K., Kuipers, O.P., and Veening, J.W., 2006. Phenotypic variation in bacteria: the role of feedback regulation. Nat. Rev. Microbiol. 4, 259-271.

Sokal, R.R. and Michener, C.D., 1958. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin 28, 1409-1438.

Soller, J.A., Schoen, M.E., Bartrand, T., Ravenscroft, J.E., and Ashbolt, N.J., 2010. Estimated human health risks from exposure to recreational waters impacted by human and non-human sources of faecal contamination. Water Res. 44, 4674-4691.

Solo-Gabriele, H.M., Wolfert, M.A., Desmarais, T.R., and Palmer, C.J., 2000. Sources of *Escherichia coli* in a coastal subtropical environment. Appl. Environ. Microbiol. 66, 230-237.

Spor, A., Koren, O., and Ley, R., 2011. Unravelling the effects of the environment and host genotype on the gut microbiome. Nat Rev Microbiol 9, 279-290.

Staley, C., Gordon, K.V., Schoen, M.E., and Harwood, V.J., 2012. Performance of two quantitative PCR methods for microbial source tracking of human sewage and implications for microbial risk assessment in recreational waters. Appl. Environ. Microbiol. 78, 7317-7326.

Stamatakis, A., Hoover, P., and Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML Web servers. Syst. Biol. 57, 758-771.

Stecher, B., Maier, L., and Hardt, W., 2013. 'Blooming' in the gut: how dysbiosis might contribute to pathogen evolution. Nat. Rev. Microbiol. 11, 277-284.

Stewart, E.J., Satorius, A.E., Younger, J.G., and Solomon, M.J., 2013. Role of environmental and antibiotic stress on *Staphylococcus epidermidis* biofilm microstructure. Langmuir 29, 7017-7024.

Stoeckel, D.M., Mathes, M.V., Hyer, K.E., Hagedorn, C., Kator, H., Lukasik, J., O'Brien, T.L., Fenger, T.W., Samadpour, M., Strickler, K.M. *et al.,* 2004. Comparison of seven protocols to identify fecal contamination sources using *Escherichia coli*. Environ. Sci. Technol. 38, 6109-6117.

Sung, J.M., Lloyd, D.H., and Lindsay, J.A., 2008. *Staphylococcus aureus* host specificity: comparative genomics of human versus animal isolates by multi-strain microarray. Microbiology 154, 1949-1959.

Swaminathan, B., Barrett, T.J., Hunter, S.B., Tauxe, R.V., and CDC PulseNet Task Force, 2001. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. Emerg. Infect. Dis. 7, 382-389.

Tambalo, D.D., Fremaux, B., Boa, T., and Yost, C.K., 2012. Persistence of host-associated *Bacteroidales* gene markers and their quantitative detection in an urban and agricultural mixed prairie watershed. Water Res. 46, 2891-2904.

Taylor, D.L., Bruns, T.D., and Hodges, S.A., 2004. Evidence for mycorrhizal races in a cheating orchid. Proc. Biol. Sci. 271, 35-43.

Tenaillon, O., Skurnik, D., Picard, B., and Denamur, E., 2010. The population genetics of commensal *Escherichia coli*. Nat. Rev. Microbiol. 8, 207-217.

Theron, J. and Cloete, T.E., 2002. Emerging waterborne infections: contributing factors, agents, and detection tools. Crit. Rev. Microbiol. 28, 1-26.

Tobe, T., Hayashi, T., Han, C.G., Schoolnik, G.K., Ohtsubo, E., and Sasakawa, C., 1999. Complete DNA sequence and structural analysis of the enteropathogenic *Escherichia coli* adherence factor plasmid. Infect. Immun. 67, 5455-5462.

Tracz, D.M., Tabor, H., Jerome, M., Ng, L.K., and Gilmour, M.W., 2006. Genetic determinants and polymorphisms specific for human-adapted serovars of *Salmonella enterica* that cause enteric fever. J. Clin. Microbiol. 44, 2007-2018.

Tran, Q.T., Williams, S., Farid, R., Erdemli, G., and Pearlstein, R., 2013. The translocation kinetics of antibiotics through porin OmpC: insights from structure-based solvation mapping using WaterMap. Proteins 81, 291-299.

Trevors, J.T., 2011. Viable but non-culturable (VBNC) bacteria: Gene expression in planktonic and biofilm cells. J. Microbiol. Methods 86, 266-273.

Tymensen, L.D., Pyrdok, F., Coles, D., Koning, W., McAllister, T.A., Jokinen, C.C., Dowd, S.E., and Neumann, N.F., 2015. Comparative accessory gene fingerprinting of surface water *Escherichia coli* reveals genetically diverse naturalized population. J. Appl. Microbiol. 119, 263-277.

Uhlich, G.A., Cooke, P.H., and Solomon, E.B., 2006. Analyses of the red-dry-rough phenotype of an *Escherichia coli* O157:H7 strain and its role in biofilm formation and resistance to antibacterial agents. Appl. Environ. Microbiol. 72, 2564-2572.

United States Environmental Protection Agency, 2007a. Protocol for developing nutrient TMDLs. United States Environmental Protection Agency, Washington, D.C.

United States Environmental Protection Agency, 2007b. Protocol for developing pathogen TMDLs. United States Environmental Protection Agency, Washington, D.C.

van de Werfhorst, L.C., Sercu, B., and Holden, P.A., 2011. Comparison of the host specificities of two *Bacteroidales* quantitative PCR assays used for tracking human fecal contamination. Appl. Environ. Microbiol. 77, 6258-6260.

van der Sanden, S.M., Koopmans, M.P., and van der Avoort, H.G., 2013. Detection of human enteroviruses and parechoviruses as part of the national enterovirus surveillance in the Netherlands, 1996-2011. Eur. J. Clin. Microbiol. Infect. Dis. 32, 1525-1531.

van Frankenhuyzen, J.K., Trevors, J.T., Lee, H., Flemming, C.A., and Habash, M.B., 2011. Molecular pathogen detection in biosolids with a focus on quantitative PCR using propidium monoazide for viable cell enumeration. J. Microbiol. Methods 87, 263-272.

Vila, J. and Soto, S.M., 2012. Salicylate increases the expression of marA and reduces in vitro biofilm formation in uropathogenic *Escherichia coli* by decreasing type 1 fimbriae expression. Virulence 3, 280-285.

Vogeleer, P., Tremblay, Y.D., Mafu, A.A., Jacques, M., and Harel, J., 2014. Life on the outside: role of biofilms in environmental persistence of Shiga-toxin producing *Escherichia coli*. Front. Microbiol. 5, 317.

Walk, S.T., Alm, E.W., Calhoun, L.M., Mladonicky, J.M., and Whittam, T.S., 2007. Genetic diversity and population structure of *Escherichia coli* isolated from freshwater beaches. Environ. Microbiol. 9, 2274-2288.

Walter, J. and Ley, R., 2011. The human gut microbiome: ecology and recent evolutionary changes. Annu. Rev. Microbiol. 65, 411-429.

Wang, M.C., Tseng, C.C., Wu, A.B., Huang, J.J., Sheu, B.S., and Wu, J.J., 2009. Different roles of host and bacterial factors in *Escherichia coli* extra-intestinal infections. Clin. Microbiol. Infect. 15, 372-379.

Wang, R.F., Beggs, M.L., Erickson, B.D., and Cerniglia, C.E., 2004. DNA microarray analysis of predominant human intestinal bacteria in fecal samples. Mol. Cell. Probes 18, 223-234.

Wang, X., Hu, M., Xia, Y., Wen, X., and Ding, K., 2012. Pyrosequencing analysis of bacterial diversity in 14 wastewater treatment systems in China. Appl. Environ. Microbiol. 78, 7042-7047.

Warner, D.M., Yang, Q., Duval, V., Chen, M., Xu, Y., and Levy, S.B., 2013. Involvement of MarR and YedS in carbapenem resistance in a clinical isolate of *Escherichia coli* from China. Antimicrob. Agents Chemother. 57, 1935-1937.

Weber, H., Polen, T., Heuveling, J., Wendisch, V.F., and Hengge, R., 2005. Genome-wide analysis of the general stress response network in *Escherichia coli*: sigmaS-dependent genes, promoters, and sigma factor selectivity. J. Bacteriol. 187, 1591-1603.

Weissman, S.J., Moseley, S.L., Dykhuizen, D.E., and Sokurenko, E.V., 2003. Enterobacterial adhesins and the case for studying SNPs in bacteria. Trends Microbiol. 11, 115-117.

Wexler, H.M., 2007. *Bacteroides*: the good, the bad, and the nitty-gritty. Clin. Microbiol. Rev. 20, 593-621.

White, A.P., Sibley, K.A., Sibley, C.D., Wasmuth, J.D., Schaefer, R., Surette, M.G., Edge, T.A., and Neumann, N.F., 2011. Intergenic sequence comparison of *Escherichia coli* isolates reveals lifestyle adaptations but not host specificity. Appl. Environ. Microbiol. 77, 7620-7632.

White, A.P. and Surette, M.G., 2006. Comparative genetics of the rdar morphotype in *Salmonella*. J. Bacteriol. 188, 8395-8406.

White, A.P., Weljie, A.M., Apel, D., Zhang, P., Shaykhutdinov, R., Vogel, H.J., and Surette, M.G., 2010. A global metabolic shift is linked to *Salmonella* multicellular development. PLoS One 5, e11814.

Whitman, W.B., Coleman, D.C., and Wiebe, W.J., 1998. Prokaryotes: The unseen majority. Proc. Natl. Acad. Sci. U. S. A. 95, 6578-6583.

Wichadakul, D., Kobmoo, N., Ingsriswang, S., Tangphatsornruang, S., Chantasingh, D., Luangsa-ard, J.J., and Eurwilaichitr, L., 2015. Insights from the genome of *Ophiocordyceps polyrhachis-furcata* to pathogenicity and host specificity in insect fungi. BMC Genomics 16, 881.

Wicki, M., Auckenthaler, A., Felleisen, R., Liniger, M., Loutre, C., Niederhauser, I., Tanner, M., and Baumgartner, A., 2012. Improved detection of *Rhodococcus coprophilus* with a new quantitative PCR assay. Appl. Microbiol. Biotechnol. 93, 2161-2169.

Wiggins, B.A., 1996. Discriminant analysis of antibiotic resistance patterns in fecal streptococci, a method to differentiate human and animal sources of fecal pollution in natural waters. Appl. Environ. Microbiol. 62, 3997-4002.

Wilrich, C. and Wilrich, P.T., 2009. Estimation of the POD function and the LOD of a qualitative microbiological measurement method. J. AOAC Int. 92, 1763-1772.

Winfield, M.D. and Groisman, E.A., 2003. Role of nonhost environments in the lifestyles of *Salmonella* and *Escherichia coli*. Appl. Environ. Microbiol. 69, 3687-3694.

Wolf, S., Hewitt, J., and Greening, G.E., 2010. Viral multiplex quantitative PCR assays for tracking sources of fecal contamination. Appl. Environ. Microbiol. 76, 1388-1394.

Wong, K., Fong, T.T., Bibby, K., and Molina, M., 2012. Application of enteric viruses for fecal pollution source tracking in environmental waters. Environ. Int. 45, 151-164.

Wood, T.K., Gonzalez Barrios, A.F., Herzberg, M., and Lee, J., 2006. Motility influences biofilm architecture in *Escherichia coli*. Appl. Microbiol. Biotechnol. 72, 361-367.

Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V., and Romano, L.A., 2003. The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. 20, 1377-1419.

Wymer, L.J., 2007. The evolution of water quality criteria in the United States, 1922-2003, in Wymer, L.J. (Eds), Statistical Framework for Recreational Water Quality Criteria and Monitoring. Wiley, Hoboken, New Jersey, pp.1-12.

Xiao, L. and Fayer, R., 2008. Molecular characterisation of species and genotypes of *Cryptosporidium* and *Giardia* and assessment of zoonotic transmission. Int. J. Parasitol. 38, 1239-1255.

Xiao, L., Fayer, R., Ryan, U., and Upton, S.J., 2004. *Cryptosporidium* taxonomy: recent advances and implications for public health. Clin. Microbiol. Rev. 17, 72-97.

Xiong, J., 2006. Essentials of Bioinformatics. Cambridge University Press, Cambridge, New York.

Yanagihara, M., Tsuneoka, H., Sugasaki, M., Nojima, J., and Ichihara, K., 2010. Multispacer typing of *Bartonella henselae* isolates from humans and cats, Japan. Emerg. Infect. Dis. 16, 1983-1985.

Yang, K., Pagaling, E., and Yan, T., 2014a. Estimating the prevalence of potential enteropathogenic *Escherichia coli* and intimin gene diversity in a human community by monitoring sanitary sewage. Appl. Environ. Microbiol. 80, 119-127.

Yang, Y., Zhou, M., Hou, H., Zhu, J., Yao, F., Zhang, X., Zhu, X., Hardwidge, P.R., and Zhu, G., 2014b. Quorum-sensing gene luxS regulates flagella expression and Shiga-like toxin production in F18ab *Escherichia coli*. Can. J. Microbiol. 60, 355-361.

Yildirim, S., Yeoman, C.J., Sipos, M., Torralba, M., Wilson, B.A., Goldberg, T.L., Stumpf, R.M., Leigh, S.R., White, B.A., and Nelson, K.E., 2010. Characterization of the fecal microbiome from non-human wild primates reveals species specific microbial communities. PLoS One 5, e13963.

Zaslaver, A., Bren, A., Ronen, M., Itzkovitz, S., Kikoin, I., Shavit, S., Liebermeister, W., Surette, M.G., and Alon, U., 2006. A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. Nat. Methods 3, 623-628.

Zhi, S., Li, Q., Yasui, Y., Edge, T., Topp, E., and Neumann, N.F., 2015. Assessing host-specificity of *Escherichia coli* using a supervised learning logic-regression-based analysis of single nucleotide polymorphisms in intergenic regions. Mol. Phylogenet. Evol. 92, 72-81.

Ziebuhr, W., Krimmer, V., Rachid, S., Lossner, I., Gotz, F., and Hacker, J., 1999. A novel mechanism of phase variation of virulence in *Staphylococcus epidermidis*: evidence for control of the polysaccharide intercellular adhesin synthesis by alternating insertion and excision of the insertion sequence element IS256. Mol. Microbiol. 32, 345-356.

Zogaj, X., Nimtz, M., Rohde, M., Bokranz, W., and Romling, U., 2001. The multicellular morphotypes of *Salmonella typhimurium* and *Escherichia coli* produce cellulose as the second component of the extracellular matrix. Mol. Microbiol. 39, 1452-1463.

Appendix

Appendix Table 1. NCBI *E. coli* genomes used in this study

Name (by this study)	Organism Name	Animal Source	Size (MB)	Gene	Sequencing Completeness
H1	*Escherichia coli DH1*	Human	4.6	4578	Complete Genome
H2	*Escherichia coli KO11FL*	Human	5.0	5037	Complete Genome
H3	*Escherichia coli O145:H28 str. RM13514*	Human	5.7	5901	Complete Genome
H4	*Escherichia coli O145:H28 str. RM13516*	Human	5.6	5683	Complete Genome
H5	*Escherichia coli ST2747*	Human	5.1	4920	Complete Genome
H6	*Escherichia coli MS 198-1*	Human	5.3	5383	Scaffold
H7	*Escherichia coli MS 84-1*	Human	5.3	5383	Scaffold
H8	*Escherichia coli MS 115-1*	Human	4.8	4795	Scaffold
H9	*Escherichia coli MS 182-1*	Human	5.0	5054	Scaffold
H10	*Escherichia coli MS 146-1*	Human	4.7	4734	Scaffold
H11	*Escherichia coli MS 45-1*	Human	5.0	4977	Scaffold
H12	*Escherichia coli MS 69-1*	Human	5.2	5191	Scaffold
H13	*Escherichia coli MS 187-1*	Human	4.4	4338	Scaffold
H14	*Escherichia coli O104:H4 str. ON2010*	Human	5.1	5211	Scaffold
H15	*Escherichia coli LCT-EC106*	Human	5.2	5142	Scaffold
H16	*Escherichia coli 95JB1*	Human	5.3	5699	Scaffold
H17	*Escherichia coli 2362-75*	Human	5.2	5364	Contig
H18	*Escherichia coli 3431*	Human	5.2	5470	Contig
H19	*Escherichia coli E128010*	Human	5.2	5621	Contig
H20	*Escherichia coli RN587/1*	Human	5.1	5228	Contig
H21	*Escherichia coli NCCP15647*	Human	5.1	5268	Contig
H22	*Escherichia coli NCCP15658*	Human	5.5	5648	Contig
H23	*Escherichia coli 541-15*	Human	5.0	5211	Contig
H24	*Escherichia coli 576-1*	Human	5.2	5442	Contig
H25	*Escherichia coli 75*	Human	4.6	4638	Contig
H26	*Escherichia coli HM605*	Human	5.1	5299	Contig
H27	*Escherichia coli 541-1*	Human	5.0	5119	Contig
H28	*Escherichia coli O104:H4 str. E112/10*	Human	5.3	5517	Contig
H29	*Escherichia coli ONT:H33 str. C48/93*	Human	5.0	4985	Contig
H30	*Escherichia coli TOP382-1*	Human	4.9	4892	Contig
H31	*Escherichia coli TOP382-2*	Human	5.1	5057	Contig
H32	*Escherichia coli TOP382-3*	Human	5.0	4962	Contig

H33	*Escherichia coli TOP550-2*	Human	5.0	4992	Contig
H34	*Escherichia coli TOP550-3*	Human	5.0	4942	Contig
H35	*Escherichia coli TOP550-4*	Human	5.0	4943	Contig
H36	*Escherichia coli TOP2396-1*	Human	4.4	4385	Contig
H37	*Escherichia coli TOP2396-2*	Human	4.7	4753	Contig
H38	*Escherichia coli TOP2396-3*	Human	4.9	5066	Contig
H39	*Escherichia coli TOP2522-1*	Human	4.7	4562	Contig
H40	*Escherichia coli TOP2662-1*	Human	4.9	4864	Contig
H41	*Escherichia coli TOP2662-2*	Human	4.9	4878	Contig
H42	*Escherichia coli TOP2662-3*	Human	5.0	4931	Contig
H43	*Escherichia coli TOP2662-4*	Human	5.0	4928	Contig
H44	*Escherichia coli C639_08*	Human	4.9	5029	Contig
H45	*Escherichia coli C844_97*	Human	4.7	4707	Contig
H46	*Escherichia coli O127:H6 str. E2348/69 substr. CVDNalr*	Human	4.9	5110	Contig
H47	*Escherichia coli O127:H6 str. E2348/69 substr. UMD753*	Human	4.9	5115	Contig
H48	*Escherichia coli C12_92*	Human	5.1	5217	Contig
H49	*Escherichia coli C1244_91*	Human	5.3	5501	Contig
H50	*Escherichia coli C1214_90*	Human	5.6	5998	Contig
H51	*Escherichia coli C154_11*	Human	5.2	5376	Contig
H52	*Escherichia coli C155_11*	Human	5.5	5923	Contig
H53	*Escherichia coli C157_11*	Human	5.5	5804	Contig
H54	*Escherichia coli C161_11*	Human	5.3	5436	Contig
H55	*Escherichia coli C213_10*	Human	4.7	4765	Contig
H56	*Escherichia coli OK1114*	Human	5.5	5979	Contig
H57	*Escherichia coli 2-005-03_S4_C3*	Human	5.2	5597	Contig
H58	*Escherichia coli 1-182-04_S4_C3*	Human	5.1	5201	Contig
H59	*Escherichia coli 1-176-05_S4_C3*	Human	5.2	5396	Contig
H60	*Escherichia coli 1-250-04_S4_C1*	Human	5.2	5458	Contig
H61	*Escherichia coli 1-182-04_S4_C1*	Human	5.1	5236	Contig
H62	*Escherichia coli STEC O174:H2 str. 02-04446*	Human	5.1	5252	Contig
H63	*Escherichia coli STEC O174:H8 str. 02-07607*	Human	5.3	5392	Contig
H64	*Escherichia coli O174:H21*	Human	5.0	5138	Contig
H65	*Escherichia coli 1-176-05_S4_C2*	Human	5.1	5269	Contig
H66	*Escherichia coli 4541-1*	Human	5.0	5000	Contig
H67	*Escherichia coli 4552-1*	Human	5.1	5147	Contig
H68	*Escherichia coli 10810*	Human	5.0	5243	Contig
H69	*Escherichia coli 7996-1*	Human	5.3	5351	Contig
H70	*Escherichia coli 11117*	Human	5.3	5437	Contig
H71	*Escherichia coli 3-267-03_S3_C1*	Human	5.1	5339	Contig
H72	*Escherichia coli 3-105-05_S1_C1*	Human	5.2	5376	Contig

H73	*Escherichia coli 3-105-05_S4_C2*	Human	5.4	5442	Contig
H74	*Escherichia coli 2-011-08_S1_C3*	Human	5.0	5122	Contig
H75	*Escherichia coli 2-052-05_S4_C3*	Human	4.9	4999	Contig
H76	*Escherichia coli 2-156-04_S1_C1*	Human	5.7	5380	Contig
H77	*Escherichia coli 2-156-04_S1_C2*	Human	5.7	5351	Contig
H78	*Escherichia coli 2-156-04_S1_C3*	Human	5.0	5146	Contig
H79	*Escherichia coli 2-156-04_S3_C1*	Human	5.0	5095	Contig
H80	*Escherichia coli 2-316-03_S1_C2*	Human	5.3	5503	Contig
H81	*Escherichia coli 8-415-05_S4_C1*	Human	5.3	5385	Contig
H82	*Escherichia coli 2-316-03_S1_C1*	Human	5.4	5612	Contig
H83	*Escherichia coli 2-460-02_S1_C3*	Human	5.5	5737	Contig
H84	*Escherichia coli 3-020-07_S3_C2*	Human	5.6	5915	Contig
H85	*Escherichia coli 6-319-05_S4_C2*	Human	5.1	5302	Contig
H86	*Escherichia coli 6-537-08_S1_C1*	Human	5.3	5514	Contig
H87	*Escherichia coli 6-319-05_S4_C3*	Human	5.2	5556	Contig
H88	*Escherichia coli 6-175-07_S1_C3*	Human	5.4	5469	Contig
H89	*Escherichia coli 6-175-07_S4_C3*	Human	4.8	4905	Contig
H90	*Escherichia coli CS03*	Human	5.0	4870	Contig
H91	*Escherichia coli TOP293-2*	Human	5.5	-	Contig
H92	*Escherichia coli TOP293-3*	Human	5.5	-	Contig
H93	*Escherichia coli TOP293-4*	Human	5.5	-	Contig
H94	*Escherichia coli TOP498*	Human	5.1	-	Contig
H95	*Escherichia coli O104:H21 str. CFSAN002237*	Human	4.9	4543	Contig
H96	*Escherichia coli O104:H21 str. CFSAN002236*	Human	4.9	4616	Contig
B1	*Escherichia coli O157:H7 str. SS17*	Cattle	5.7	5847	Complete Genome
B2	*Escherichia coli AA86*	Cow	5.0	5039	Scaffold
B3	*Escherichia coli EC4196*	Cattle	5.4	5748	Scaffold
B4	*Escherichia coli EC4203*	Cattle	5.4	5776	Scaffold
B5	*Escherichia coli FRIK1996*	Cattle	5.4	5872	Scaffold
B6	*Escherichia coli FRIK1985*	Cattle	5.5	6007	Scaffold
B7	*Escherichia coli 93-001*	Cattle	5.4	5719	Scaffold
B8	*Escherichia coli FRIK1990*	Cattle	5.5	5954	Scaffold
B9	*Escherichia coli O157:H7 str. FRIK966*	Bovine	5.4	5616	Contig
B10	*Escherichia coli O157:H7 str. FRIK2000*	Bovine	5.4	5614	Contig
B11	*Escherichia coli 1.2741*	Cow	5.7	5806	Contig
B12	*Escherichia coli 97.0246*	Cow	5.5	5821	Contig
B13	*Escherichia coli 97.0264*	Cow	5.2	5178	Contig
B14	*Escherichia coli 4.0522*	Cow	5.8	6208	Contig
B15	*Escherichia coli 99.0741*	Cow	5.5	5617	Contig
B16	*Escherichia coli 900105 (10e)*	Calf	5.6	5832	Contig
B17	*Escherichia coli 5.0588*	Cow	5.0	5049	Contig

B18	*Escherichia coli 3.3884*	Cow	5.2	5313	Contig
B19	*Escherichia coli O111:H11 str. CVM9534*	Cow	5.5	5957	Contig
B20	*Escherichia coli O111:H11 str. CVM9545*	Cow	5.6	6330	Contig
B21	*Escherichia coli O111:H8 str. CVM9570*	Cow	5.5	6107	Contig
B22	*Escherichia coli O26:H11 str. CVM9942*	Cow	5.6	6158	Contig
B23	*Escherichia coli O26:H11 str. CVM10026*	Cow	5.6	6040	Contig
B24	*Escherichia coli O111:H8 str. CVM9634*	Cow	5.8	6262	Contig
B25	*Escherichia coli O26:H11 str. CVM10030*	Cow	5.5	5938	Contig
B26	*Escherichia coli O111:H11 str. CVM9553*	Cow	5.6	6159	Contig
B27	*Escherichia coli O26:H11 str. CVM10021*	Cow	5.5	5927	Contig
B28	*Escherichia coli O111:H11 str. CFSAN001630*	Cow	5.6	6056	Contig
B29	*Escherichia coli C842_97*	Cattle	5.2	5416	Contig
B30	*Escherichia coli ECC-Z*	Bovine	4.9	4909	Contig
B31	*Escherichia coli LAU-EC2*	Bovine	5.2	5347	Contig
B32	*Escherichia coli O157:H7 str. 09BKT048303*	Cattle	5.2	5810	Contig
B33	*Escherichia coli O157:H7 str. T1543_06*	Cattle	5.3	5755	Contig
B34	*Escherichia coli O157:H7 str. 08BKT061141*	Cattle	5.3	5917	Contig
B35	*Escherichia coli O157:H7 str. 09BKT002497*	Cattle	5.1	-	Contig
D1	*Escherichia coli APEC O1*	Turkey	5.5	5572	Complete Genome
D2	*Escherichia coli UMNK88*	Pig	5.7	5863	Complete Genome
D3	*Escherichia coli UMNF18*	Pig	5.6	5907	Complete Genome
D4	*Escherichia coli SWW33*	Mouse	4.9	4777	Scaffold
D5	*Escherichia coli K02*	Mouse	4.8	4791	Scaffold
D6	*Escherichia coli 1.2264*	Goat	5.5	5570	Contig
D7	*Escherichia coli 9.1649*	Pig	5.1	5057	Contig
D8	*Escherichia coli 4.0967*	Rabbit	5.9	6149	Contig
D9	*Escherichia coli 2.3916*	Pig	5.6	6041	Contig
D10	*Escherichia coli B41*	Pig	5.0	5051	Contig
D11	*Escherichia coli 3.2608*	Horse	5.5	5622	Contig
D12	*Escherichia coli AI27*	Pig	4.9	5014	Contig
D13	*Escherichia coli KD1*	Dog	4.8	4824	Contig
D14	*Escherichia coli KD2*	Dog	5.0	4963	Contig
D15	*Escherichia coli CUMT8*	Mouse	4.7	4791	Contig
D16	*Escherichia coli O26:H11 str. CVM9952*	Pig	5.5	5915	Contig
D17	*Escherichia coli AD30*	Chicken	5.1	5118	Contig
D18	*Escherichia coli O08*	Chicken	5.1	5109	Contig
D19	*Escherichia coli S17*	Broiler chick	4.6	4602	Contig
D20	*Escherichia coli SEPT362*	Laying Hen	5.3	5342	Contig

D21	*Escherichia coli IMT8073*	Pig	5.1	-	Contig
D22	*Escherichia coli C527_94*	Rabbit	5.0	5125	Contig
D23	*Escherichia coli C900_01*	Pig	5.1	5417	Contig
D24	*Escherichia coli MP1*	Mouse	4.8	4761	Contig
D25	*Escherichia coli E455*	Pig	4.6	4526	Contig
D26	*Escherichia coli 48*	Deer	6.3	6397	Contig
D27	*Escherichia coli 77302533*	Pig	5.2	5447	Contig
D28	*Escherichia coli 77300132*	Pig	5.0	5195	Contig
D29	*Escherichia coli 77300095*	Pig	5.3	5526	Contig

Appendix Table 2. Genes selected for ITGR *in silico* analysis.

Gene name	Annotation
aer	fused signal transducer for aerotaxis sensory
agaI	putative galactosamine-6-phosphate isomerase
amiA	probable N-acetylmuramoyl-L-alanine amidase
araC	AraC-type regulatory protein
argI	orthithine carbamoyltransferase chain I
argT	lysine/arginine/ornithine transporter subunit
betI	transcriptional regulator *BetI*
cedA	cell division regulatory protein
codB	cytosine deaminase
csgC	predicted curli production protein
cspA	stress protein, member of the CspA family
cspE	cold shock protein E
cstA	carbon starvation protein
cycA	serine/alanine/glycine APC transporter
cysD	sulfate adenylyltransferase subunit 2
cysJ	sulfite reductase, flavoprotein subunit complex
cysK	cysteine synthase A, O-acetylserine sulfhydrolase A subunit
cysP	subunit of sulfate/thiosulfate/selenite/selenate ABC transporter
deoC	deoxyribose-phosphate aldolase
dgoR	predicted DNA-binding transcriptional regulator
dhaK	dihydroxyacetone kinase subunit K
fdnG	α subunit of formate dehydrogenase N
fepA	outer membrane receptor for ferric enterobactin
fhuF	ferric iron reductase involved in ferric hydroximate transport
fiu	predicted iron outer membrane transporter
flgA	flagellar basal-body P-ring biosynthesis protein A
flhB	flagellar biosynthesis protein B

fliA	flagellar biosynthesis sigma factor
fliC	flagellin
fliF	flagellar M-ring protein
fruK	subunit of 1-phosphofructokinase
galE	subunit of UDP-glucose 4-epimerase
hdeA	stress response protein acid-resistance protein
hipB	HipB antitoxin and DNA-binding transcriptional repressor
ilvC	ketol-acid reductoisomerase
ilvL	ilvGEDA operon leader peptide
katE	Catalase HPII
lon	ATP-dependent protease
map	methionine aminopeptidase
mdtA	MdtABC-TolC multidrug efflux transport system - putative membrane fusion protein
motA	proton conductor component of motor
nanA	N-acetylneuraminate lyase
napF	ferredoxin-type protein
ompC	outer membrane porin protein C
oppA	peptide ABC transporter - periplasmic binding protein
paaA	ring 1,2-phenylacetyl-CoA epoxidase, monooxygenase subunit
pepQ	Xaa-Pro dipeptidase
potA	ATP-dependent polyamine transporter
ppdA	exodeoxyribonuclease V λ chain
proV	glycine betaine / proline ABC transporter - ATP binding subunit
purE	N5-carboxyaminoimidazole ribonucleotide mutase
pyrL	*Pyr* operon leader peptide
rpsL	30S ribosomal subunit
sfmA	putative chaperone-usher fimbrial operon
slp	outer membrane lipoprotein
tar	methyl-accepting chemotaxis protein II
tsx	regulator for d*eo* operon, *udp, cdd, tsx, nupC,* and *nupG*
wza	lipoprotein required for capsular polysaccharide translocation through the outer membrane
xseA	exodeoxyribonuclease VII large subunit
yaeP	conserved protein
yagU	inner membrane protein
ycdT	hypothetical protein
ycgR	protein involved in flagellar function
ydeN	predicted sulfatase
ydjR	hypothetical protein
yedS	outer membrane protein
yedU	chaperone protein HchA
yefM	YefM-antitoxin DNA-binding transcriptional repressor
yegH	membrane protein
yfaL	hypothetical protein

yfeA	predicted diguanylate cyclase
ygdI	hypothetical protein
yggB	putative transport protein
yhaO	predicted transporter
yhgA	hypothetical protein
yhjB	DNA-binding response regulator in two-component regulatory system
yhjH	EAL domain containing protein involved in flagellar function
yiaW	inner membrane protein
yicM	putative transport protein
yidQ	hypothetical protein
yihV	6-deoxy-6-sulfofructose kinase
yjbN	hypothetical protein
yjiY	predicted inner membrane protein
yjjN	hypothetical zin-type alcohol dehydrogenase-like protein
ykgE	predicted oxidoreductase
yliE	conserved inner membrane protein
ynbA	predicted inner membrane protein
ynbB	predicted CDP-diglyceride synthase
ynjH	hypothetical protein
yrfF	inner membrane protein