

GENETIC DIVERSITY OF CLINICAL AND BOVINE
NON-O157 SHIGA TOXIN-PRODUCING *ESCHERICHIA COLI* (STEC)

By

Heather Marie Blankenship

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Microbiology and Molecular Genetics – Doctor of Philosophy

2019

ABSTRACT

GENETIC DIVERSITY OF CLINICAL AND BOVINE NON-O157 SHIGA TOXIN-PRODUCING *ESCHERICHIA COLI* (STEC)

By

Heather Marie Blankenship

Shiga toxin-producing *Escherichia coli* (STEC) is a leading cause of foodborne infection resulting in 265,000 illnesses and more than 3,600 hospitalizations annually. Since its identification in 1982 associated with an outbreak of haemorrhagic colitis, serotype O157:H7 has been the primary focus of research and surveillance. However, the increasing incidence of other serogroups, or non-O157 STEC, that are associated with clinical illness has since surpassed the incidence of O157 and has raised questions about the genetic diversity of this pathogen population. Six serogroups, O26, O45, O103, O111, O121, and O145, have been denoted as “big six” non-O157 STEC serogroups since they are frequently associated with clinical outcomes.

In this dissertation, 895 non-O157 STEC isolates recovered from patients in Michigan between 2001-2018 were analyzed using whole genome sequencing (WGS) to identify virulence gene profiles and apply new typing methods to better discriminate closely related strains. The recovery of a wide range of serogroups from cases presenting with symptoms ranging from mild diarrhea to hemorrhagic colitis, indicates that genetic diversity and variation may have an impact on disease outcomes. The number and richness of serogroups identified over the past 18 years has been steadily increasing and serogroup alone lacks the discriminatory capabilities to classify related isolates. Indeed, strains representing the same sequence types (ST) were often found to be unrelated by serogroup. Notably, some serogroups, STs, virulence gene profiles and alleles were associated with clinical outcomes and patient demographics. Contrast to national surveillance, cases between 11 and 29 years of age had the highest frequency of STEC infections in Michigan.

Additionally, a subset of 44 non-O157 STEC recovered from Michigan patients between 2000 and 2006 were examined more comprehensively while making comparisons to 114 clinical STEC isolates from Connecticut to examine the impact of geographic location on risk factors for non-O157 STEC infections. Lastly, a subset of STEC isolates associated with outbreaks in Michigan were examined to identify the impact of WGS on identification of strain relatedness for surveillance compared to pulsed-field gel electrophoresis.

While most of the work outlined in this dissertation focused on characterizing clinical non-O157 STEC isolates, a comparative analysis of cattle isolates was also performed since cattle are an important reservoir of STEC. Indeed, numerous outbreaks and illnesses have been traced back to contaminated cattle-based food products or fecal contamination of water and crops. The ability of STEC to persist in the cattle reservoir and farm environment may give rise to more pathogenic strains due to the accumulation of horizontally acquired genes. 66 STEC isolates recovered from a beef herd over four samplings were examined to identify the genetic diversity within the cattle population and longitudinal persistence. The ability of a strain to form a strong biofilm was associated with the ability to persist and be recovered at multiple sampling phases from the same animal. Further, to better understand the genetic diversity of STEC recovered from the cattle reservoir, an additional 12 STEC isolates from three bovine herds (n=78) and 241 clinical O157 STEC isolates (n=1,135) were included to identify shared profiles. The similarity in serogroups and virulence gene profiles warrant a continued surveillance of the cattle environment to better understand crossover events and the ability of strains to evolve into new virulent STEC lineages. The work described in this dissertation helped to elucidate the genetic characteristics important for clinical outcomes and identified targets for future surveillance to better understand lineages that may be important for disease.

‘The journey of a thousand miles begins with a single step.’
-Lao Tzu

This work is dedicated to my husband, Philip Blankenship, who has encouraged and supported every step of this journey.

ACKNOWLEDGEMENTS

Pursuing a PhD has shared many features of completing an ultramarathon including; highs and lows, a lot of nutrition (coffee) and a community to support the endeavor both near and far. None of this would have been possible without the continued support, encouragement and guidance from my mentor, Shannon Manning. She encouraged and pushed me to mature as a person and as a scientist, even when I doubted myself. She further encouraged me to explore bioinformatics and genomics and was patient as I stumbled through code in the first few years, but this opened new opportunities and research questions that I never thought I would pursue. More importantly, she taught me how to design, analyze, write and present in the field of molecular epidemiology. In addition, I am incredibly grateful that I was able to be a part of the Manning lab and work with such an amazing group of people. Especially, Rebekah Mosci and Samantha Carbonell for growing up hundreds of isolates and preliminary PCR typing, Brian Nohomovich for discussing bioinformatic methods with me, and Megan Shiroda for providing an outside perspective and humor in the office. I would also like to thank my committee members, Dr. Dan Grooms, Dr. Lixin Zhang, and Dr. Chris Waters for their support, guidance, and expertise that they brought to my project.

I would like to thank the staff at the Michigan Department of Health and Human Services for their help and support in completion of this project. Dr. Marty Soehnlen provided me with the public health perspective of my project, and I am thankful that she allowed me to immerse myself into the field and broaden my knowledge of infectious disease. Dr. James Rudrik, Jim Collins, Ben Hutton, Jason Wholehan, Steve Dietrich, Elizabeth Burgess, Kelly Jones and the staff at MDHHS helped with specimen processing, compiling epidemiological data, performing

DNA extractions, PFGE analysis, and whole genome sequencing of isolates included in the study. I would also like to thank Karen McWilliams, Scott Benko, Karen Pietrzen, and Ted Gatesy at the Michigan Department of Agriculture and Rural Development for performing whole genome sequencing of isolates that were included in the study.

Funding for this work came from the National Institutes of Health, US Department of Agriculture, National Institute of Food and Agriculture, and Michigan State University. Additional funding was provided through the MSU Enrichment Fellowship, College of Natural Science Dissertation Continuation Fellowship, Department of Microbiology and Molecular Genetics Bertina Wentworth Award, Rudolph Hugh Award, and Marvis A Richardson Award.

Lastly, I have to thank my friends and family who have continually supported my journey from both near and far. Whether it was through a text or phone call after a rough day, sharing some miles on the roads and trails together, a cup of coffee or a bottle of wine, all of the conversations and laughs helped to encourage and support me along the way. I would like to thank my parents and in-laws, my sisters, Erica and Kelley, and my extended family for their love, support, and encouragement throughout this journey. Lastly, my biggest cheerleader and deserving of his own “honorary” degree, my husband, Phil, has supported every step of this journey. I am so grateful to him for always making sure that my coffee mug was filled, for all of the late-night dinners he brought to the office, the perspective and humor that he provided when I was having a rough day and the endless miles he shared with me. I have to express an appreciation for Revy, Roxy, and Rue, who each helped in their own way and provided me with daily playfulness: Revy reminded me to take time for myself and shared cuddles while I wrote and worked from home, Rue provided me with daily laughs at her antics and was the only one

who would wait up until 3am to provide me with nighttime kisses, and Roxy reminded me to work and play hard but to not forget the importance of taking time for myself.

It really does take a community to pursue a PhD and I am so thankful for everyone who has had an impact on this journey.

TABLE OF CONTENTS

LIST OF TABLES	xii
LIST OF FIGURES	xiii
KEY TO ABBREVIATIONS	xvi
CHAPTER 1 LITERATURE REVIEW: SHIGA TOXIN-PRODUCING ESCHERICHIA COLI (STEC) AND BACTERIAL TYPING METHODS.....	1
SHIGA TOXIN-PRODUCING ESCHERICHIA COLI.....	2
Non-O157 STEC.....	2
Detection and identification of STEC	3
STEC virulence factors.....	4
Demographic factors associated with STEC infections	6
STEC reservoirs and transmission to humans	7
BACTERIAL STRAIN TYPING METHODS.....	8
Phenotypic Typing Methods	10
Genotypic Typing Methods.....	11
Banding Pattern Based Typing Methods	12
Pulsed-Field Gel Electrophoresis (PFGE).....	12
Restriction Fragment Length Polymorphism (RFLP).....	13
Multiple Locus Variable-number Tandem Repeat Analysis (MLVA)	14
DNA Sequencing Typing Methods.....	15
Multilocus Sequence Typing (MLST).....	15
Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)	16
Single Nucleotide Polymorphism Typing (SNP)	18
Pan Genome Comparisons	19
SUMMARY	20
REFERENCES	22
CHAPTER 2 TRENDS OF SHIGA TOXIN-PRODUCING ESCHERICHIA COLI (STEC) STRAINS RECOVERED FROM PATIENTS IN MICHIGAN, 2001-2018.....	41
ABSTRACT	42
INTRODUCTION	43
MATERIALS AND METHODS	47
Bacterial strains and epidemiological data	47
Ethics Statement.....	47
DNA isolation and whole genome sequencing (WGS)	47
Bioinformatic and in silico analysis.....	48
Data analysis and visualization	48
RESULTS	49
Case demographics of patients with non-O157 STEC infections in Michigan	49
Distribution of serogroups and association with clinical outcomes.....	51

Trends of virulence genes and subtypes	52
Genetic diversity of non-O157 STEC and association of clusters with disease	53
DISCUSSION	56
APPENDIX	61
REFERENCES	78

CHAPTER 3 GENETIC DIVERSITY OF NON-O157 SHIGA TOXIN-PRODUCING <i>ESCHERICHIA COLI</i> (STEC) STRAINS RECOVERED FROM PATIENTS IN MICHIGAN AND CONNECTICUT	85
ABSTRACT	86
INTRODUCTION	88
MATERIALS AND METHODS	91
Bacterial strains and epidemiological data	91
Ethics statement	91
DNA isolation and whole genome sequencing (WGS)	91
Multilocus sequence typing (MLST) and in silico analysis of virulence genes.....	92
CRISPR-Cas sequence analysis	93
Data analysis	93
RESULTS	95
Characteristics of cases infected with non-O157 STEC by state.....	95
Distribution of serogroups and virulence genes and association with clinical outcomes	96
Genetic diversity of non-O157 STEC and association with disease.....	98
CRISPR profiling and phylogenetic analysis	100
CRISPR spacer content indicative of phage and plasmid transfer	101
DISCUSSION	103
APPENDIX	107
REFERENCES	119

CHAPTER 4 ANALYSIS OF WHOLE GENOME SEQUENCING FOR CHARACTERIZATION AND OUTBREAK IDENTIFICATION OF SHIGA TOXIN-PRODUCING <i>ESCHERICHIA COLI</i> (STEC) STRAINS, 2015-2018	126
ABSTRACT	127
INTRODUCTION	128
MATERIALS AND METHODS	131
Bacterial strains, DNA isolation and whole genome sequencing (WGS)	131
Pulsed-Field Gel Electrophoresis (PFGE)	131
Bioinformatic Analysis	131
Data Analysis and Visualization	132
RESULTS	133
Isolate identification and serogroup distributions	133
Phylogenetic analysis based on MLST loci	134

Core genome SNP (cgSNP) analysis differentiates outbreak strains that cluster by MLST	134
High quality SNP (hqSNP) analysis further differentiates outbreak isolates compared to the core genome analysis and PFGE	136
DISCUSSION	138
APPENDIX	142
REFERENCES	154
 CHAPTER 5 GENETIC FACTORS OF SHIGA TOXIN-PRODUCING <i>ESCHERICHIA COLI</i> (STEC) ASSOCIATED WITH PERSISTENCE AND BIOFILM FORMATION IN BEEF CATTLE FARMS	161
ABSTRACT	162
INTRODUCTION	163
METHODS	166
Bacterial strains	166
Biofilm assays	166
DNA isolation and whole genome sequencing (WGS)	167
Bioinformatic analysis	167
Data analysis	168
RESULTS	169
Herd demographics and prevalence of STEC	169
Genetic diversity of STEC	170
Biofilm formation and persistence of STEC isolates	170
Longitudinal examination of related isolates	172
DISCUSSION	174
APPENDIX	179
REFERENCES	187
 CHAPTER 6 COMPARATIVE GENOMICS OF SHIGA TOXIN-PRODUCING <i>ESCHERICHIA COLI</i> (STEC) CATTLE AND CLINICAL ISOLATES IN MICHIGAN	196
ABSTRACT	197
INTRODUCTION	199
MATERIALS AND METHODS	201
Bacterial Strains	201
DNA isolation and whole genome sequencing (WGS)	201
Bioinformatic analysis	202
RESULTS	203
Genetic relatedness of clinical and cattle isolates with shared gene profiles and serotypes using multilocus sequence typing (MLST)	203
Shared genomic profiles among STEC isolates from patients and cattle	204
Core genome SNP (cgSNP) analysis of clinical and cattle isolates with related STs, serotypes, and gene profiles	205
DISCUSSION	207
APPENDIX	211

REFERENCES	219
CHAPTER 7 CONCLUSIONS AND FUTURE DIRECTIONS	226
REFERENCES	233

LIST OF TABLES

Table 2.1.	Clinical outcomes associated with reported gender	62
Table 2.2.	Demographic, molecular characteristics, and clinical outcomes associated with age at time of infection relative to age group 11-29	63
Table 2.3.	Demographic, molecular characteristics and clinical outcomes associated with big six non-O157 STEC infections relative to other non-O157 serogroups	65
Table 2.4.	Association between big six non-O157 STEC serogroups and hospitalization	67
Table 2.5.	Association between big six non-O157 STEC serogroups and bloody diarrhea	68
Table 2.6.	Virulence gene profiles found in 894 non-O157 STEC isolates from patients with infections.	69
Table 2.7.	Demographic and molecular characteristics associated with MLST clusters compared to all other non-O157 isolates	70
Table 3.1.	Comparison of demographics and clinical outcomes among non-O157 STEC cases from Michigan and Connecticut between 2001 and 2006.	108
Table 3.2.	Demographic, molecular profiles and clinical outcomes associated with big-six non-O157 serogroups and all other non-O157 serogroups from cases in Michigan and Connecticut combined.	110
Table 3.3.	Demographic, molecular profiles and clinical outcomes associated with big-six non-O157 serogroups from cases in Michigan and Connecticut relative to infection with other non-O157 serogroups	112
Table 4.1.	STEC serogroups present in Michigan from <i>in silico</i> typing	143
Table 5.1.	Serogroups and virulence gene profiles among 63 cattle derived non-O157 STEC isolates	180
Table 6.1.	Number of clinical and cattle STEC isolates with shared serotypes and virulence gene profiles	212

LIST OF FIGURES

Figure 2.1.	Total number of non-O157 STEC isolates that were recovered for WGS (2001-2018) compared to the total number of non-O157 STEC cases reported by MDHHS (2001-2012, 2015-2018)	72
Figure 2.2.	Frequency of age groups that reported a non-O157 STEC infection, 2001-2018	73
Figure 2.3.	Prevalence of non-O157 big-six STEC infections in Michigan, 2001-2018	74
Figure 2.4.	Distribution and changes in the non-O157 serogroups reported in Michigan, 2001-2018	75
Figure 2.5.	Distribution and gene frequency across non-O157 big-six and other STEC serogroups	76
Figure 2.6.	MLST-based phylogeny of 894 non-O157 STEC isolates examined using the neighbor-joining algorithm with 1000 bootstrap replication	77
Figure 3.1.	Prevalence of serogroups detected in Michigan and Connecticut, 2001-2006	114
Figure 3.2.	Distribution and gene frequency of virulence genes in STEC serogroups	115
Figure 3.3.	Neighbor-joining phylogenetic analysis constructed using seven gene MLST in 155 clinical STEC isolates from Michigan (n=44, green circles) and Connecticut (n=111, blue circles) with 1000 bootstrap replication to establish genetic relatedness	116
Figure 3.4.	CRISPR spacer content for strains belonging to Clusters 1 and 2 as determined using MLST	117
Figure 3.5.	Unweighted pair group method with arithmetic averages (UPGMA) clustered using a Jaccard similarity index to compare the spacer patterns of the CRISPR profiles of 149 total isolates from Michigan (n=40) and Connecticut (n=109)	118
Figure 4.1.	Percent of all STEC isolates sequenced (n=510) at the MDHHS per year (black line) and the frequency of non-O157 and O157 serogroups in sequenced STEC isolates	144

Figure 4.2.	Neighbor-joining phylogenetic tree constructed based on seven MLST loci for 509 STEC isolates from 2015-2018 with 1000 bootstrap replication	145
Figure 4.3.	Core genome SNP analysis of 135 STEC isolates belonging to the multilocus sequence type (ST)-106/104 cluster, including serogroup O26 and NT outbreak associated isolates	146
Figure 4.4.	Core genome SNP analysis of 188 STEC isolates belonging to the multilocus sequence type (ST)-119 cluster, including serogroup O103 outbreak associated isolates	147
Figure 4.5.	Core genome SNP analysis of 99 O157 STEC isolates belonging to the multilocus sequence type (ST)-66 cluster	148
Figure 4.6.	Core genome SNP analysis of 17 STEC isolates belonging to the multilocus sequence type (ST)-175 cluster, including serogroup O5 outbreak associated isolates	149
Figure 4.7.	Phylogeny based on hqSNP analysis of ST-175 isolates that clustered with outbreak strains using cgSNP analysis	150
Figure 4.8.	Phylogeny based on hqSNP analysis of ST-66-O1 isolates that clustered with outbreak strains using cgSNP analysis	151
Figure 4.9.	Phylogeny based on hqSNP analysis of ST-66-O2/O3 isolates that clustered with outbreak strains using cgSNP analysis	152
Figure 4.10.	Phylogeny based on hqSNP analysis of 60 ST-119 isolates that clustered with outbreak strains using cgSNP analysis	153
Figure 5.1.	Maximum likelihood phylogeny with 1,000 bootstrap replications constructed using 2,933 concatenated genes. Serogroup and multilocus sequence type (ST) designations are indicated for each cluster and high (open circles) and low (colored circles) biofilm formation. Bootstrap values >0.98 support for clustering by ST and serogroup	181
Figure 5.2.	Frequency of STEC serogroups stratified by the level of biofilm production in 66 isolates recovered from cattle	182
Figure 5.3.	Longitudinal overview of STEC isolates by cow, sampling period, serogroup and strength of biofilm	183
Figure 5.4.	High quality SNP (hqSNP) analysis of ten O26 STEC isolates recovered from ten cattle at different sampling periods	184

Figure 5.5.	High quality SNP (hqSNP) analysis of eight O168 STEC isolates recovered from cattle over multiple samplings. All isolates recovered from the same cow (760, blue boxes) are indicated as well as those isolates with high levels of biofilm production (up triangle). All other O168 isolates came from other cattle at varying sampling points	185
Figure 5.6.	High quality (hqSNP) analysis of 36 O6 STEC isolates identified from cattle in the study. Low biofilm formers (inverted triangle) and isolates from the same cow (752, blue; 761, red; 763, green; 764, orange; 767, purple; 768, grey; 773, yellow) are denoted with similar shading	186
Figure 6.1.	Neighbor joining phylogeny with 1000 bootstrap replication constructed with concatenated seven gene MLST profiles from 78 cattle isolates and 1,135 clinical isolates. All cattle STs are denoted with: **COW and blue shading of the node	214
Figure 6.2.	Core genome SNP analysis of 279 clinical and 10 cattle isolates within clade ST-106 (blue shading of shared clade)	215
Figure 6.3.	Core genome SNP analysis of 419 clinical and 5 cattle isolates within clade ST-119 (blue shading of shared clade). Two shared clinical isolates within the cattle clade are denoted with closed circles	216
Figure 6.4.	Core genome SNP analysis of 236 clinical and 7 cattle isolates (cattle isolates denoted as blue boxes) identified as ST-66	217
Figure 6.5.	Core genome SNP analysis of 13 clinical and 6 cattle isolates within clade ST-157/158/159 (blue shading of shared clade). One shared clinical isolate within the cattle clade is denoted with a closed circle	218

KEY TO ABBREVIATIONS

A ₅₉₅	Absorbance 595nm
BLAST	Basic Local Alignment Search Tool
cgMLST	Core Genome Multilocus Sequence Typing
cgSNP	Core Genome Single Nucleotide Polymorphism
CDC	Centers for Disease Control and Prevention
CI	Confidence Interval
CIDT	Culture Independent Test
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CTDPH	Connecticut Department of Public Health
CV	Crystal Violet
EcMLST	<i>E.coli</i> Multilocus Sequence Typing
EAEC	Enterotoaggregative <i>E.coli</i>
EHEC	Enterohemorrhagic <i>E.coli</i>
EPEC	Enteropathogenic <i>E.coli</i>
FDA	Food and Drug Administration
FoodNet	Foodborne Disease Active Surveillance Network
GB ₃	Globotriaosylceramide
HUS	Hemolytic Uremic Syndrome
hqSNP	High Quality Single Nucleotide Polymorphism
LEE	Locus of Enterocyte Effacement
LPS	Lipopolysaccharide
MDHHS	Michigan Department of Health and Human Services
MDSS	Michigan Disease Surveillance System
MLST	Multilocus Sequence Typing
MLVA	Multiple Locus Variable-number Tandem Repeat Analysis
NCBI	National Center for Biotechnology Information
NNDSS	National Notifiable Disease Surveillance System

OR	Odds Ratio
PCR	Polymerase Chain Reaction
PFGE	Pulsed Field Gel Electrophoresis
PHL	Public Health Laboratory
QC	Quality Control
REP-PCR	Repetitive Extragenic Palindromic PCR
RFLP	Restriction Fragment Length Polymorphism
SES	Socioeconomic Status
SNP	Single Nucleotide Polymorphism
SMAC	Sorbitol-MacConkey
ST	Sequence Type
Stx	Shiga Toxin
STEC	Shiga Toxin-Producing <i>E.coli</i>
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
USDA	United States Department of Agriculture
VNTR	Variable Number Tandem Repeat
WGS	Whole Genome Sequencing
wgMLST	Whole Genome Multilocus Sequence Typing

CHAPTER 1

LITERATURE REVIEW: SHIGA TOXIN-PRODUCING *ESCHERICHIA COLI* (STEC) AND BACTERIAL TYPING METHODS

SHIGA TOXIN-PRODUCING *ESCHERICHIA COLI*

Shiga toxin-producing *Escherichia coli* (STEC) is a leading foodborne pathogen that results in 265,000 illnesses and 3,600 hospitalizations annually (1). Clinical outcomes present as diarrhea and hemorrhagic colitis; however, hemolytic uremic syndrome, kidney failure and ultimately death may occur in severe cases (2). An outbreak of hemorrhagic colitis in 1982 identified O157:H7 as the etiologic agent in contaminated hamburgers in Michigan and Oregon (3–5). However, STEC was not recognized nationally as an important foodborne pathogen until 1993, when a multi-state outbreak of Jack in the Box hamburgers resulted in 732 illnesses and four deaths, all deaths were in children under the age of 18 (6). Due to a lack of structured surveillance, it took 39 days for public health officials to identify that there was an outbreak occurring across multiple states and to begin implementing measures to control the outbreak (7). This outbreak was seminal for identifying a need to implement food safety measures, a national surveillance, and understanding of STEC pathogenesis. The Foodborne Diseases Active Surveillance Network (FoodNet) was established in 1995 by the Centers for Disease Control and Prevention, the US Department of Agriculture Food Safety and Inspection Service and the Food and Drug Administration to provide active surveillance at 10 sites in the US.

Non-O157 STEC

O157:H7 has been the predominant focus for research and food safety measures due to being associated with HUS and multiple outbreaks, however, a wide range of other serogroups (non-O157) have the potential to cause clinical outcomes (4, 8, 9). Since the addition of non-O157 STEC to the list of nationally notifiable diseases, the incidence has increased from 0.12 to 1.65 per 100,000 people from 2000-2015 and has since surpassed the incidence of O157 (10, 11).

In the US, six serogroups are frequently identified from clinical cases and have been denoted the “big six” non-O157 serogroups; O26, O45, O103, O111, O121, and O145 (12). Between 1983 and 2002, 71% of the non-O157 serogroups were attributed to the big six, with 55 additional serogroups that were identified and associated with clinical outcomes (12). Regardless of serogroup, all strains are characterized by the presence of Shiga toxin (*stx*) genes. Numerous outbreaks have been attributed to these serogroups from various food sources including serogroup O26 associated with ground beef and O121 and O26 associated with flour (13–15). However, serogroups that are less frequently isolated can still result in severe clinical outcomes and hospitalization (15, 16).

Detection and identification of STEC

The inability of O157 to ferment sorbitol is contrast to other *E.coli* organisms including non-O157 STEC (17). This biochemical difference led to the initial use of sorbitol-MacConkey (SMAC) agar to differentially identify O157 isolates (18). However, there is no biochemical difference that is shared among all non-O157s to differentiate these serogroups from commensal *E.coli*, which is important for the under reporting of non-O157 cases. The use of culture independent tests (CIDT) as an alternative to culture, rely on sequence-based identification of the *stx* genes or enzyme immunoassay to detect the Shiga toxin. Changes to CIDTs may be responsible for some of the increases in non-O157 STEC due to the ability to better identify non-O157 STEC cases. An increase in the number of non-O157 isolates identified correlated with the number of labs that were performing enzyme immunoassays for the detection of STEC ($p < 0.001$) (10). In the recent FoodNet report, over 500 cases were identified to be STEC by CIDTs and were either culture negative or culturing was not attempted (19). While CIDTs allow for better

and quicker detection of pathogens, lack of an isolate prevents further analysis to identify antimicrobial resistance profiles, bacterial subtypes or identify relatedness of strains for outbreak identification.

STEC virulence factors

The main virulence factor of STEC is the lambdoid bacteriophage encoded Shiga toxin (Stx) that was originally identified in *Shigella* (3, 20–22). Stx is an AB₅ toxin that binds to the endothelial cell surface receptor globotriaosylceramide (Gb₃) (23, 24). Five B subunits help with binding of the toxin to Gb₃ receptors and the A subunit halts eukaryotic protein synthesis in the cells (25, 26). Two immunologically distinct toxins have been identified, Stx1 and Stx2, with a similar mode of action (26). Further subtypes for each toxin were originally identified when immunological differences were observed. The use of whole genome sequencing (WGS) has enabled the identification of additional subtypes with no immunological differences. Stx1a and Stx2a are the most common subtypes (27). Three Stx1 subtypes have been identified, Stx1a, Stx1c, and Stx1d, while Stx1a is the only subtype that has been associated with severe clinical outcomes (5, 28). Eight Stx2 subtypes have been identified and include: Stx2a, Stx2b, Stx2c, Stx2d, Stx2e, Stx2f, Stx2g and Stx2h (29, 30). Genes encoding the Stx2 subtypes, *stx2a*, *stx2c*, and *stx2d*, have been linked to more severe disease outcomes (31–33). Subtype Stx2b and Stx1c have been associated with environmental sources and tend to cause asymptomatic or milder infections in humans (34, 35). Similarly, Stx2e has been associated with mild outcomes in humans but can cause edema in pigs resulting in ataxia and death (31, 36–39).

Other virulence factors of importance are encoded within the 35.5kb locus of enterocyte effacement (LEE) pathogenicity island and through the acquisition of plasmids. The LEE

encodes genes for intimin (*eae*), the translocated intimin receptor (Tir), a type III secretion system, and other putative virulence factors (40–43). These genes are required for the attaching and effacement of epithelial cells, also known as attaching and effacing lesions. The Tir protein is translocated into epithelial cells and localized to the cell membrane of host cells triggering the formation of F-actin pedestals (44, 45). Intimin binding to the Tir provides the intimate attachment of STEC to epithelial cells (46). Thirty distinct subtypes of intimin have been identified based on differences in the C-terminal region (34, 47, 48). These differences may influence the host cell tropism and specificity of intimin binding (49). Intimin subtypes have been associated with serogroups; O157 and O145 predominantly harbor subtype gamma and a diverse range of subtypes are found amongst the other big six serogroups O26 (beta), O103 (epsilon), and O111 (theta) (50–54). Other putative attachment genes have been identified in LEE-negative STEC that may provide other means of attachment to epithelial cells; these include autoagglutinating adhesins (*saa*), long polar fimbriae (*lpf*), and autotransporters (*ehaA* and *sab*) (55–58).

Lastly, a range of virulence factors can be acquired on plasmids. The pO157 is a non-conjugative F-like plasmid that ranges in size from 92-104kb found in O157 (59, 60). The structure of the pO157 plasmid is different from plasmids found within non-O157 serogroups, however, the virulence genes encoded on pO157 is conserved in plasmids of non-O157 serogroups (61–63). Putative virulence factors such as catalase-peroxidase for colonization in the absence of oxygen (*katP*), adhesions to increase the production and secretion of type III secretion systems (*toxB*), and an enterohemolysin (*ehxA*) (64–66). Importantly, the presence of *ehxA* has been associated with the presence of Stx and HUS development (67, 68), suggesting that it may be an indicator of pathogenicity and impact disease severity. Six genetically distinct *ehxA*

subtypes (A-F) have been identified (69, 70). Subtypes C and F were found to be more common in clinical isolates, while subtype A was suggested to be more common in environmental sources (69–71). A number of other virulence factors are encoded within the chromosome or are acquired through plasmids, bacteriophages, and other mobile genetic elements via horizontal gene transmission to facilitate survival in the acidic pH of the stomach as well as colonization.

Demographic factors associated with STEC infections

The ability of STEC to differentially cause infections in individuals with varying demographics has been reported by the FoodNet surveillance sites. The highest incidence of STEC occurs in children <5 years old (1.89 per 100,000 people) and decreases until age 60, which is followed by an increase in incidence among the elderly (10, 72). A similar incidence of non-O157 infection is seen in adults and elderly, while O157 STEC infections are mostly responsible for the increasing incidence observed in the elderly population (10). In children younger than 10 years of age, O157 infection is the leading risk factor for development of HUS in the US and is the most common cause of renal failure and HUS worldwide (73–75).

Incidence rates for non-O157 STEC range from 0.11-1.14 per 100,000 people among the ten FoodNet surveillance sites (10). This wide range suggests that geographic factors may influence the number of infections that occur within an area at a given time. Regions that have high cattle densities have been associated with an increase in O157 and non-O157 infections (38, 76). Similarly, direct contact with cattle and other animals that may asymptotically harbor STEC in petting zoos have been associated with infections (77, 78). Variation in the socioeconomic status (SES) of a region may also be responsible for differences in incidence rates. A person with a high SES has a consistently higher risk of acquiring an O157 or non-O157

STEC infection (79, 80). While it was hypothesized that individuals with a higher SES would have access to better healthcare and seek medical advice even for minor infections, a 2000-2003 study in FoodNet sites identified that lower income categories were more likely to seek care (79). A higher SES may also allow for international travel to countries with higher rates of non-O157 infections and a larger range of serogroups commonly associated with severe clinical outcomes. In Europe, serogroups O26, O91, O103, O111, O145, and O146, for example, are the top six non-O157 serogroups commonly associated with infections according to the European Centre for Disease Prevention and Control (81). Similarly, in Germany, serogroup O91 is the most common serogroup found in adult infections and has caused severe clinical infections resulting in HUS (82, 83).

STEC reservoirs and transmission to humans

Cattle are the main reservoir for STEC and have been implicated as the source for numerous STEC outbreaks (84–86). The lack of Gb₃ vascular receptors in cattle allows for asymptomatic colonization without any clinical indicators. Other ruminants and farm animals have been identified to harbor STEC including deer and pigs (87–90). However, Stx2e positive STEC isolates can cause edema in pigs through binding with the globotetraosylceramide (Gb₄) receptor in epithelial or vascular endothelial cells (91). Environmental factors and farm practices play a role in the prevalence of STEC within the farm environment. Studies have shown that an increase in temperature during summer months was associated with a higher prevalence of STEC (92–95). Similarly, a number of studies have examined the association between STEC shedding in cattle and age (96–99). Though there is no age that is strongly supported by all studies for the

highest risk of STEC, the majority of studies have shown that younger cows are more likely to be colonized, albeit asymptotically.

Transmission of STEC to humans can occur through direct contact with colonized animals, consumption of contaminated food or contact with contaminated water, and through person to person contact. Improper hand washing after contact with animals at petting zoos and direct contact with cattle have been associated with STEC infection (77, 78, 100). The most common means of transmission is foodborne through incorrect handling and preparation of cattle products (101). STEC has been associated with outbreaks of leafy greens through the contamination of crop irrigation water from fecal runoff of nearby farms (102). Further, transmission of STEC from person to person can occur through fecal-oral route. The median duration of asymptomatic shedding of STEC after infection in childcare outbreaks ranges from 20 to 50 days (103–105). This extended period of shedding along with a low infectious dose may result in a higher rate of secondary transmission (106).

BACTERIAL STRAIN TYPING METHODS

The ability to reliably differentiate and type bacterial strains is critical for public health surveillance and identifying related strains of importance for research. Historic typing of strains relied on the classification of organisms based on biochemical differences, changes in the cell surface or structure, and other phenotypic determinants that were specific to an organism. The resolution of bacterial strains has been increased through the introduction of molecular analysis and typing methods (107, 108). Recent years have seen a transition in molecular typing methods away from traditional gel-based methods to targeted gene identification and whole genome

sequencing. This shift has allowed for a larger amount of information to be readily available and has changed typing isolates at the genus level to the species and subspecies level (109, 110).

The usefulness of a typing method relies on its ability to discriminate strains in a meaningful way and distinguish among unrelated isolates within a species (111). All typing methods will aim to target something that is specific for a genus or species for classification but will not be so specific as to exclude strains that should be included within the classification. Newer typing methods have allowed a better ability to identify epidemiologically related isolates from those that are within the same species (108, 112). Molecular strain typing relies on the evolution within a species to accurately and reproducibly differentiate strains. However, this will be species dependent as the rate for evolutionary events may vary, thus, the typing methods selected will need to ensure that it accurately identifies related isolates within the evolution rate of the species.

The assessment of typing methods should ensure that the discriminatory power and reproducibility can be obtained for the organism of interest. Discriminatory ability of a typing method should discriminate between strains at a high resolution while maintaining epidemiological concordance (113). A quantitative indicator of discriminatory power can be expressed as a probability defined by the Simpson's index of diversity (114). The method should also be reproducible independent of when, where and by whom the method was performed (113). Inability to produce consistent result will limit the usefulness of the test and comparisons that can be performed across studies and locations. While a method may reproducibly discriminate strains, the time and labor requirements may limit the feasibility of the method if there is a high labor, monetary or time cost. The throughput of a method should allow for a high number of isolates to be analyzed and sustained over an extended period.

Changes in molecular biology over the recent years has led to a shift in the methods that were once gold standards for typing and identification of related organisms (115–117). With the accessibility and cost of sequencing decreasing, several sequencing methods have been developed that focus on whole genome analysis or targeted identification of genes (117–121). This shift away from phenotypic and gel-based typing methods is allowing for a higher discriminative power, however, access to the whole genome of bacterial isolates is providing an exponential amount of data that will need to be sorted for identification of new methods.

Phenotypic Typing Methods

Traditional microbiology focused on the use of phenotypic properties to identify bacterial organisms and determine the relatedness of strains. Serotyping of organisms based on somatic and flagellar antigens allows for the differentiation of isolates into subspecies. Serotyping of *E.coli* has been historically used to help with epidemiological investigations and identification of clinically relevant strains based on O- H- and K-antigen determinants (122, 123). Recent examination of serogroup typing is identifying that this may not be indicative of a strain's relatedness, due to the exchange of horizontal elements allowing for the rise of related organisms with different serogroups (124).

Examination of the antimicrobial resistance profiles are currently used as ways to cluster large groups of organisms, particularly carbapenem resistant organisms to examine the transmission of these organisms in epidemiological investigations; carbapenem-resistant Enterobacteriaceae (CRE), *Acinetobacter baumannii* (CRAB), and *Pseudomonas aeruginosa* (CRPsA) (125, 126). The generation of antibiogram (antibiotic susceptibility testing) profiles from the resistance and susceptibility profiles has been examined in chromosomal and plasmid

acquired resistance (127). Stability of the resistance profile in an organism and utilizing the antibiogram may incorrectly group together unrelated strains due to the mobility of antimicrobial resistance genes.

Lastly, traditional identification of an organism is based on biochemical tests to narrow down the genus or species. Differences in growth on differential or selective media and the presence of enzymes within an organism will provide visual identification. Differentiation of O157 STEC relies on the inability of O157 to ferment sorbitol resulting in colorless colonies on SMAC agar (128).

Genotypic Typing Methods

Molecular methods of typing have been rapidly evolving over the past 30 years as better means of examining the bacterial genome have become available. The main advantage for using molecular methods is the ability to identify related organisms with a higher sensitivity and quicker turn around than phenotypic methods. For the past 23 years, pulsed-field gel electrophoresis (PFGE) has been the gold standard for disease surveillance and identifying related organisms for enteric pathogens as part of the PulseNet surveillance network at the CDC (129, 130). Recent transitions have introduced WGS as the new standard for surveillance and identification of related strains (131). Various molecular methods have been developed over the past 30 years that rely on differences in the genomic profile to identify related isolates. These methods can be grouped into three main categories: methods that rely on electrophoresis to generate banding patterns for comparison, methods that rely on DNA-DNA hybridization, and methods that utilize targeted gene or whole genome sequencing.

DNA hybridization methods are not discussed in extensive details below, however, they rely on the binding of DNA to a labeled cDNA or oligonucleotide in microarrays and have been reviewed elsewhere (132, 133). The use of cDNA as the labeled probe aids in the identification of housekeeping genes or virulence genes to help identify or characterize organisms. However, microarrays with oligonucleotides as the labeled probe, can be utilized to identify specific SNPs that may differ in related organisms.

Banding Pattern Based Typing Methods

Pulsed-Field Gel Electrophoresis (PFGE)

Various methods that separate out DNA fragments and compare patterns have been developed, however, PFGE has been adopted for epidemiological surveillance for a range of pathogens (116, 134–137). PFGE was originally developed in 1982 by Schwartz and Cantor to examine the chromosomal DNA of yeast. It has since been adopted and standardized for surveillance and tracking of outbreaks in gram negative and positive organisms (116, 135). Rare cutting restriction enzymes are used to cut the DNA into fragments, which are run on an agarose gel through a switching electrical field. However, changes in the secondary DNA structures and the methylation status of the DNA may impact the ability of a restriction enzyme to cut the DNA resulting in shifts from the expected banding pattern (138, 139). Further, bacterial organisms, such as STEC, that harbor lysogenic bacteriophages in their genomes may have shifts in the banding patterns due to the loss or acquisition of the bacteriophage (140). Similarly, PFGE is not useful for identification of granular changes in the DNA that do not impact a restriction enzyme site or change the length of the DNA fragment. Further, the process of PFGE is time consuming and labor intensive with a minimal throughput of 11-12 isolates analyzed on a single gel. Lastly,

differences in banding patterns can be observed due to deviations from protocol which may influence the resolution and interpretation of the banding patterns.

Prior to WGS, the use of PFGE was able to discriminate strains with a high discriminative power based on banding patterns for outbreak surveillance. The standardization of the method allowed for a national surveillance of foodborne pathogens, including O157 and non-O157 STEC, by PulseNet through the comparison of gel banding patterns (116, 141). This standardization also allowed for international comparison to identify international transmissions and isolates that may be travel related (142). While distinct banding patterns were generated for STEC isolates, the high similarity within serogroups and related isolates, may fail to identify epidemiologically unrelated isolates. The inability to discriminate unrelated isolates was seen in O157 STEC banding patterns, identifying a need for a more discriminative typing method for related isolates (112, 129). The benefit of PFGE is its ability to discriminate strains that may be, however, subtle changes in DNA are not observed and further discrimination of closely related subspecies cannot be elucidated.

Restriction Fragment Length Polymorphism (RFLP)

Similar to PFGE, the comparison of banding patterns generated through the use of restriction enzymes are used to identify the relatedness of isolates. However, the DNA is digested with restriction enzymes that frequently cut the DNA. To minimize the number of DNA fragments that are run on an agarose gel, RFLP is commonly paired with polymerase chain reactions (PCR, PCR-RFLP) to examine the banding pattern of a specific region or gene of interest.

PCR-RFLP has been generalized for methods in a range of bacterial organisms that target a number of genes of interest. The identification of intimin and enterohemolysin subtype in STEC were originally identified and typed through the use of PCR-RFLP to provide a higher resolution of the genes (47, 69, 142). Similarly, PCR-RFLP has been applied to the 16S rDNA of bacterial isolates from a diverse group of organisms that were associated with endophthalmitis and were able to differentiate all species, except for discriminating between *E.coli* and *Serratia marcescens* (143).

Multiple Locus Variable-number Tandem Repeat Analysis (MLVA)

Amplification of variable number tandem repeat (VNTR) arrays through PCR are performed to generate DNA fragments. The size of the fragments is identified through the use of capillary electrophoresis, which runs the DNA fragments through a gel matrix and generates an electropherogram. The electropherogram gets converted into allele types that will correlate with the relatedness of strains. This technique is not widely used due to the need of generating PCR primers and allele databases for each organism, instead, it has been used for some organisms as a supplemental analysis for strains with similar PFGE patterns. Within STEC subtyping, the use of MLVA was proposed as a secondary subtyping method for STEC isolates with shared PFGE patterns to further discriminate strains. A preliminary study aimed at examining the feasibility of MLVA in public health laboratories as a secondary subtyping tool for O157 STEC, 200 isolates were discriminated into 162 MLVA patterns and 139 unique PFGE patterns (144). MLVA provides a high discriminatory resolution among related isolates and has been identified as the standard for bacterial subtyping in a range of bacterial organisms including: *Bacillus anthracis*, *Yersinia pestis*, and *Mycobacterium tuberculosis* (145–147).

DNA Sequencing Typing Methods

Multilocus Sequence Typing (MLST)

Traditional six to eight gene MLST examines 400-600bp regions of housekeeping genes. Originally designed for PCR and Sanger sequencing, MLST allows for a higher resolution and discrimination than serotyping alone to observe evolutionary changes in a broad population and identify epidemiological associations between groups of isolates (148, 149). MLST is advantageous for studies of large population due to a high level of strain differentiation and stable allele calls that can be compared across studies (150–152). The use of standardized databases assures that the data has a high level of repeatability. Additionally, databases are internationally available making it feasible to compare studies that use the same database (153, 154). Allelic variation is assigned for each gene identified and sequence types assigned based on the allelic profile. Examination of allelic profiles and sequence type designation does not provide any insight into the evolutionary changes that may have occurred within an allele or how many evolutionary differences are between two alleles. At the same time, only conserved coding regions are included in the MLST analysis, which may lack the ability to discriminate related isolates. Examination of O157 STEC has typed most of the isolates by Whittam seven gene MLST as sequence type 66, however there is a large diversity in the genetic composition of O157 isolates that have been associated with differences in disease outcomes when a method with higher discriminatory power such as SNP-typing is used (112, 152).

The introduction and accessibility of WGS has changed the limits of MLST and allowed for the development of more discriminative methods. The use of core genome (cgMLST) and whole genome (wgMLST) methods allow for a national surveillance with a high discriminative power allowing for comparisons with other labs or studies since the allele codes/profiles will not

change regardless of the isolates that are added (155). Similar to the six to eight gene MLST typing schemes, the database and allele calls are standardized allowing for reproducibility across labs, studies and time periods. Regardless, there is a lot of up-front development and validations that are required to generate and upkeep the database to include new alleles and profiles.

The difference between wgMLST and cgMLST is the database that will be used to assign allele codes. WgMLST will compare genomes of interest gene by gene with a database that is comprised of all genes from a genus or species of interest and represents a diverse genetic background, including horizontally acquired genes. Due to more genes included in the analysis, wgMLST will have a higher resolution of isolates than cgMLST, however, the inclusion of horizontally acquired genes in the analysis may influence the phylogenetic relationship of isolates that are related (156, 157). Conversely, cgMLST will compare on a gene by gene basis, but it will only include genes or portions of genes that are common across all species resulting in a smaller database. While wgMLST will provide a higher resolution, cgMLST is more stable and will not change due to all strains encoding genes that are present within the MLST database. Publicly available databases for *Salmonella*, *Escherichia/Shigella*, *Yersinia*, *Campylobacter*, and *Listeria* can be found, however, validations have only been performed for *Listeria* (156, 158).

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)

The CRISPR array contains repeat sequences that vary in length from 20-50 nucleotides and are separated by highly conserved direct repeats (159). Transcription of the CRISPR array along with the CRISPR associated (*cas*) genes forms a complex of crRNA and Cas proteins, which can target foreign DNA that is complementary to the crRNA for degradation (160). While the repeats that separate the spacers are conserved, the spacer itself is highly variable and has

been found complementary to phage or plasmid sequences (159, 161, 162). These sequences are used by the bacteria as an adaptive immune system and can help the bacteria target and degrade invading genetic material that may be detrimental to the cell (163, 164). New spacers are added to the distal end of the CRISPR array allowing for evolutionary analysis of isolates to identify potential divergence of isolates and generate stepwise evolutionary experiments (161, 165–167). Similarly, spacer composition within the CRISPR array can be used to examine the relatedness of isolates based on the presence and absence of spacer sequences (161, 168–170). The ability to utilize the CRISPR region is only beneficial if the organism is not rapidly acquiring new spacers and if the CRISPR array is heterogenous within the population of interest. Rapid acquisition could result in CRISPR array profiles differing within an outbreak, thus excluding isolates that may be outbreak-associated. Changes within the CRISPR array can also be caused by microevolution occurring within the loci resulting in loss or duplication of the spacers (166, 167, 171, 172).

The subtyping of isolates by spacer-oligonucleotide typing (spoligotyping), which examines heterogeneity within the CRISPR region, has been used since 1993 for *Mycobacterium tuberculosis* (173, 174). The use of DNA hybridization arrays has generated over 3,000 known spoligotype patterns to differentiate related strains based on the knowledge of known spacers (175). Various experiments have shown that the stability and conservation of CRISPR array in STEC can be used as a subtyping marker because it is not acquiring new spacers and there is little diversity within subtypes (165, 176, 177). As a result, pairing the CRISPR sequences along with presence of *stx* and *eae*, and the intimin adherence protein gene commonly found in enterohemorrhagic *E. coli*, identified polymorphisms that provide a more specific typing profile (178, 179).

Single Nucleotide Polymorphism Typing (SNP)

The identification and systematic analysis of SNPs are useful for outbreak investigations, and can resolve closely related bacterial genotypes, provide insights into the microevolutionary history of the genome divergence, and contribute to an epidemiologic assessment of associations between bacterial genotypes and disease. SNP analysis allows for every nucleotide in the genome to be analyzed, however, a reference strain is required for comparison and analysis.

The resolution and discriminatory power of methods have gotten increasingly stronger with the accessibility of WGS. Prior to WGS, known SNPs could be examined within the genome using PCR based assays or DNA hybridization. The use of Sanger sequencing and further WGS has allowed for an easier identification of known and unknown SNPs to be extracted for analyses. Development of a 32 SNP loci typing for O157 was developed to differentiate O157 isolates with the same PFGE pattern (112). A MLST-based neighbor joining phylogeny clustered isolates into eight distinct clades that were associated with varying disease outcomes. Notably, strains belonging to Clade 8 were significantly more likely to develop HUS (112). The increased use of WGS has allowed for the examination of SNPs that may be present outside of the loci identified in O157 and applied to non-O157 STEC.

Core genome (cgSNP) and high quality (hqSNP) SNP analyses allows for the examination of all nucleotides that are shared among isolates for comparison to a reference strain. These analyses allow for a high discriminative power to examine relatedness and considers all mutations and evolutionary events that may occur to produce a new pathogenic strain (157). This method can also be generalized and adapted to all organisms without any additional up-front development. Since SNP analysis is not standardized, changes in the strains present or the reference strain can result in differences within the phylogenetic relationship of the

isolates. Analyses are limited by appropriate selection of a reference strain that is related and generated from quality data (180). While the analyses is reproducible within the set of strains and reference chosen due to the lack of ambiguity in sequencing data, comparisons across studies with different parameters, isolates, and reference is limited.

HqSNP is similar in analysis to cgSNP, however, it utilizes raw reads when mapped to a reference to provide a confidence of the SNP identified. Three main points are taken into account when performing hqSNP with various pipelines such as Lyve-SET: quality, coverage, and frequency of SNPs (181). The quality of sequencing data will greatly influence the interpretation of the data and ability to identify the relatedness of strains. Coverage will examine how many reads are present at a given location to identify if there are enough reads to accurately identify a potential SNP. Lastly, frequency indicates the confidence in the SNP call dependent on the percentage of reads that support the SNP at that location. The additional parameters on hqSNP analysis allows for the interpretation of SNP distance with a high confidence. However, the identification of clusters and SNP distances among isolates are only one piece of supporting evidence for strain relatedness. Outbreak investigations require additional epidemiological support and linkage between the isolates.

Pan Genome Comparisons

Extraction and alignment of the pan genome shared across isolates in an analysis allows for a genome wide comparison without the need of a reference strain. Differing from a core genome analysis by potentially including a larger number of genes in the analysis since it can encompass all core genes in addition to accessory genes and strain-specific genes. This analysis

relies on a high bioinformatic knowledge and will vary across analyses depending on the isolates that are included, however, the general clustering of isolates should remain unchanged.

The use of a pan genome comparison has been used in pathogenic *E. coli* isolates of different pathotypes to identify putative virulence genes that may be important for pathogenesis in those specific pathotypes (182). Comparison with a commensal strain of *E. coli* to identify the pan genome then allows for the identification of genes that are only present in specific pathotypes and clades. Analyses in other foodborne organisms such as, *Campylobacter jejuni*, have utilized the pan genome to identify epidemiologically related isolates belonging to a similar outbreak (131). The benefit of utilizing the pan genome for analysis of isolates is that it allows for a high discriminatory power without the need for identifying a reference strain for comparison or the use of a database, thus, it is generalizable and can be applied to any organism of interest (183).

SUMMARY

STEC is considered a prominent foodborne pathogen with multiple outbreaks every year that affect thousands of people. As a prominent foodborne pathogen, the ability to identify related organisms that may be outbreak associated is essential for surveillance. A range of serogroups and virulence factors have been identified in clinical isolates presenting with disease outcomes. Specifically, the continued increase in non-O157 STEC incidence supports the need to further understand the diversity within this subpopulation. This work also examined the potential transmission of STEC between the cattle reservoir and human population to identify targets for future surveillance and intervention. While O157 is commonly isolated from infections associated with cattle products, it is not understood which non-O157 profiles are associated with

clinical outcomes. The use of WGS methods will allow for subtyping and examination of genetic relatedness of isolates with a higher discriminatory power than traditional methods. In an effort to better understand the genetic diversity and the ability to use WGS methods for subtyping, the primary objective of this dissertation was to examine the genetic characteristics and diversity within populations of non-O157 STEC to elucidate characteristics of importance for disease outcomes and transmission into the human population.

REFERENCES

REFERENCES

1. Scallan E, Hoekstra RM, Angulo FJ, Tauxe R V., Widdowson M-A, Roy SL, Jones JL, Griffin PM. 2011. Foodborne Illness Acquired in the United States—Major Pathogens. *Emerg Infect Dis* 17:7–15. doi: 10.3201/eid1701.P11101.
2. Karmali MA, Petric M, Lim C, McKeough PC, Arbus GS, Lior H. 1985. The association between idiopathic hemolytic uremic syndrome and infection by verotoxin-producing *Escherichia coli*. *J Infect Dis* 151:775–782. doi: 10.1093/infdis/151.5.775.
3. O’Brien a D, LaVeck GD, Thompson MR, Formal SB. 1982. Production of *Shigella dysenteriae* type 1-like cytotoxin by *Escherichia coli*. *J Infect Dis* 146:763–769. doi: 10.1093/infdis/146.6.763.
4. Rangel JM, Sparling PH, Crowe C, Griffin PM, Swerdlow DL. 2005. Epidemiology of *Escherichia coli* O157:H7 outbreaks, United States, 1982-2002. *Emerg Infect Dis* 11:603–609. doi: 10.3201/eid1104.040739.
5. Lee W. Riley, M.D., Robert S. Remis, M.D., M.P.H., Steven D. Helgerson, M.D., M.P.H., Harry B. McGee, M.P.H., Joy G. Wells, M.S., Betty R. Davis, M.S., Richard J. Hebert, M.D., Ellen S. Olcott, R.N., Linda M. Johnson, R.N., M.S., Nancy T. Hargrett, Ph.D., MD, Riley L, Remis R, Helgerson S, McGee H, Wells JG, Davis B, Hebert R, Olcott E, Johnson L, Hargrett N, Blake P, Cohen M. 1983. Hemorrhagic Colitis Associated with a Rare *Escherichia coli* Serotype. *N Engl J Med* 308:661–685.
6. Bell BP, Goldoft M, Griffin PM, Davis MA, Gordon DC, Tarr PI, Bartleson CA, Lewis JH, Barrett TJ, Wells JG, Baron R, Kobayashi J. 1994. A Multistate Outbreak of *Escherichia coli* O157:H7—Associated Bloody Diarrhea and Hemolytic Uremic Syndrome From Hamburgers: The Washington Experience. *JAMA J Am Med Assoc* 272:1349–1353. doi: 10.1001/jama.1994.03520170059036.
7. CDC. 2016. PulseNet: 20 years of making food safer to eat. Centers for Disease Control and Prevention. Atlanta, GA.
8. Mody RK, Luna-Gierke RE, Jones TF, Comstock N, Hurd S, Scheftel J, Lathrop S, Smith G, Palmer A, Strockbine N, Talkington D, Mahon BE, Hoekstra RM, Griffin PM. 2012. Infections in pediatric postdiarrheal hemolytic uremic syndrome: Factors associated with identifying Shiga toxin-producing *Escherichia coli*. *Arch Pediatr Adolesc Med* 166:902–909. doi: 10.1001/archpediatrics.2012.471.
9. Espié E, Grimont F, Mariani-Kurkdjian P, Bouvet P, Haeghebaert S, Filliol I, Loirat C, Decludt B, Minh NNT, Vaillant V, De Valk H. 2008. Surveillance of hemolytic uremic syndrome in children less than 15 years of age, a system to monitor o157 and non-o157 shiga toxin-producing *Escherichia coli* infections in France, 1996-2006. *Pediatr Infect Dis J* 27:595–601. doi: 10.1097/INF.0b013e31816a062f.

10. Gould LH, Mody RK, Ong KL, Clogher P, Cronquist AB, Garman KN, Lathrop S, Medus C, Spina NL, Webb TH, White PL, Wymore K, Gierke RE, Mahon BE, Griffin PM, Patricia M. Griffin for the EIPFWG, Griffin PM. 2013. Increased recognition of non-O157 Shiga toxin-producing *Escherichia coli* infections in the United States during 2000-2010: epidemiologic features and comparison with *E. coli* O157 infections. *Foodborne Pathog Dis* 10:453–60. doi: 10.1089/fpd.2012.1401.
11. Crim S, Iwamoto M, Huang J, Griffin P, Gilliss D, Cronquist A, Cartter M, Tobin-D'Angelo M, Blythe D, Smith K, Lathrop S, Zansky S, Cieslak P, Dunn J, Holt K, Lance S, Tauxe R, Henao O. 2015. Preliminary Incidence and Trends of Infection with Pathogens Transmitted Commonly Through Food — Foodborne Diseases Active Surveillance Network, 10 U.S. Sites, 2006–2014. *Morb Mortal Wkly Rep* 64.
12. Brooks JT, Sowers EG, Wells JG, Greene KD, Griffin PM, Hoekstra RM, Strockbine NA. 2005. Non-O157 Shiga Toxin–Producing *Escherichia coli* Infections in the United States, 1983–2002 . *J Infect Dis* 192:1422–1429. doi: 10.1086/466536.
13. Crowe SJ, Bottichio L, Shade LN, Whitney BM, Corral N, Melius B, Arends KD, Donovan D, Stone J, Allen K, Rosner J, Beal J, Whitlock L, Blackstock A, Wetherington J, Newberry LA, Schroeder MN, Wagner D, Trees E, Viazis S, Wise ME, Neil KP. 2017. Shiga toxin–producing *E. coli* infections associated with flour. *N Engl J Med* 377:2036–2043. doi: 10.1056/NEJMoa1615910.
14. CDC. 2015. *Escherichia coli* O26 Infections Linked to Chipotle Mexican Grill Restaurants (Final Update).Centers for Disease Control and Prevention. Atlanta, GA.
15. Luna-Gierke RE, Griffin PM, Gould LH, Herman K, Bopp CA, Strockbine N, Mody RK. 2014. Outbreaks of non-O157 Shiga toxin-producing *Escherichia coli* infection: USA. *Epidemiol Infect* 142:2270–2280. doi: 10.1017/S0950268813003233.
16. Kuehne A, Bouwknecht M, Havelaar A, Gilsdorf A, Hoyer P, Stark K, Werber D, Amon O, Büscher R, Hampel T, Fehrenbach H, Habbig S, Pohl M, Häffner K, Hoppe B, Klaus G, Konrad M, Latta K, Leichter H, Loos S, Montoya C, Müller D, Galiano M, Muschiol E, Pape L, Staude H, Wühl E, Henn M, Pohl M, Wygoda S. 2016. Estimating true incidence of O157 and non-O157 Shiga toxin-producing *Escherichia coli* illness in Germany based on notification data of haemolytic uraemic syndrome. *Epidemiol Infect* 144:3305–3315. doi: 10.1017/S0950268816001436.
17. Wells JG, Davis BR, Wachsmuth IK, Riley LW, Remis RS, Sokolow R, Morris GK. 1983. Laboratory investigation of hemorrhagic colitis outbreaks associated with a rare *Escherichia coli* serotype. *J Clin Microbiol* 18:512–520.
18. Remis RS, MacDonald KL, Riley LW, Puhf ND, Wells JG, Davis BR, Blake PA, Cohen ML. 1984. Sporadic cases of hemorrhagic colitis associated with *Escherichia coli* O157:H7. *Ann Intern Med* 101:624–626. doi: 10.7326/0003-4819-101-5-624.

19. Marder EP, Cieslak PR, Cronquist AB, Dunn J, Lathrop S, Rabatsky-Ehr T, Ryan P, Smith K, Tobin-D'angelo M, Vugia DJ, Zansky S, Holt KG, Wolpert BJ, Lynch M, Tauxe R, Geissler AL. 2017. Incidence and trends of infections with pathogens transmitted commonly through food and the effect of increasing use of culture-independent diagnostic tests on surveillance — Foodborne diseases active surveillance network, 10 U.S. Sites, 2013-2016. *Morb Mortal Wkly Rep* 66:397–403. doi: 10.15585/mmwr.mm6615a1.
20. O'Brien AD, LaVeck GD. 1983. Purification and characterization of a *Shigella dysenteriae* 1-like toxin produced by *Escherichia coli*. *Infect Immun* 40:675–683.
21. Shiga K. 1898. Ueber den Dysenterie-bacillus (*Bacillus dysenteriae*). *Zentralbl Bakteriolog Orig* 24:913–918.
22. Strockbine NA, Jackson MP, Sung LM, Holmes RK, O'Brien AD. 1988. Cloning and sequencing of the genes for Shiga toxin from *Shigella dysenteriae* type 1. *J Bacteriol* 170:1116–22. doi: 10.1128/jb.170.3.1116-1122.1988.
23. Karve SS, Weiss AA. 2014. Glycolipid binding preferences of shiga toxin variants. *PLoS One* 9:e101173. doi: 10.1371/journal.pone.0101173.
24. Hale TL, Formal SB. 1980. Cytotoxicity of *Shigella dysenteriae* 1 for cultured mammalian cells. *Am J Clin Nutr* 33:2485–2490. doi: 10.1093/ajcn/33.11.2485.
25. Fraser ME, Chernaia MM, Kozlov Y V., James MNG. 1994. Crystal structure of the holotoxin from *shigella dysenteriae* at 2.5 Å resolution. *Nat Struct Biol* 1:59–64. doi: 10.1038/nsb0194-59.
26. Fraser ME, Fujinaga M, Cherney MM, Melton-Celsa AR, Twiddy EM, O'Brien AD, James MNG. 2004. Structure of Shiga toxin type 2 (Stx2) from *Escherichia coli* O157:H7. *J Biol Chem* 279:27511–27517. doi: 10.1074/jbc.M401939200.
27. Melton-Celsa AR. 2014. Shiga Toxin (Stx) Classification, Structure, and Function. *Microbiol Spectr* 2:EHEC-0024-2013. doi: 10.1128/microbiolspec.ehec-0024-2013.
28. Friedrich AW, Borell J, Bielaszewska M, Fruth A, Tschäpe H, Karch H. 2003. Shiga toxin 1c-producing *Escherichia coli* strains: Phenotypic and genetic characterization and association with human disease. *J Clin Microbiol* 41:2448–2453. doi: 10.1128/JCM.41.6.2448-2453.2003.
29. Scheutz F, Teel LD, Beutin L, Piérard D, Buvens G, Karch H, Mellmann A, Caprioli A, Tozzoli R, Morabito S, Strockbine NA, Melton-Celsa AR, Sanchez M, Persson S, O'Brien AD. 2012. Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing Stx nomenclature. *J Clin Microbiol* 50:2951–2963. doi: 10.1128/JCM.00860-12.
30. Bai X, Fu S, Zhang J, Fan R, Xu Y, Sun H, He X, Xu J, Xiong Y. 2018. Identification and pathogenomic analysis of an *Escherichia coli* strain producing a novel Shiga toxin 2

- subtype. *Sci Rep* 8:1–11. doi: 10.1038/s41598-018-25233-x.
31. Friedrich AW, Bielaszewska M, Zhang W, Pulz M, Kuczius T, Ammon A, Karch H. 2002. *Escherichia coli* Harboring Shiga Toxin 2 Gene Variants: Frequency and Association with Clinical Symptoms . *J Infect Dis* 185:74–84. doi: 10.1086/338115.
 32. Persson S, Olsen KEP, Ethelberg S, Scheutz F. 2007. Subtyping method for *Escherichia coli* Shiga toxin (Verocytotoxin) 2 variants and correlations to clinical manifestations. *J Clin Microbiol* 45:2020–2024. doi: 10.1128/JCM.02591-06.
 33. Donohue-Rolfe A, Kondova I, Oswald S, Hutto D, Tzipori S. 2000. *Escherichia coli* O157:H7 Strains That Express Shiga Toxin (Stx) 2 Alone Are More Neurotropic for Gnotobiotic Piglets Than Are Isotypes Producing Only Stx1 or Both Stx1 and Stx2 . *J Infect Dis* 181:1825–1829. doi: 10.1086/315421.
 34. Mora A, López C, Dhabhi G, López-Beceiro AM, Fidalgo LE, Díaz EA, Martínez-Carrasco C, Mamani R, Herrera A, Blanco JE, Blanco M, Blanco J. 2012. Seropathotypes, phylogroups, stx subtypes, and intimin types of wildlife-carried, shiga toxin-producing *Escherichia coli* strains with the same characteristics as human-pathogenic isolates. *Appl Environ Microbiol* 78:2578–2585. doi: 10.1128/AEM.07520-11.
 35. Hofer E, Cernela N, Stephan R. 2012. Shiga toxin subtypes associated with shiga toxin-producing *Escherichia coli* strains isolated from red deer, roe deer, chamois, and ibex. *Foodborne Pathog Dis* 9:792–795. doi: 10.1089/fpd.2012.1156.
 36. Meng Q, Bai X, Zhao A, Lan R, Du H, Wang T, Shi C, Yuan X, Bai X, Ji S, Jin D, Yu B, Wang Y, Sun H, Liu K, Xu J, Xiong Y. 2014. Characterization of Shiga toxin-producing *Escherichia coli* isolated from healthy pigs in China. *BMC Microbiol* 14:1–14. doi: 10.1186/1471-2180-14-5.
 37. Tseng M, Fratamico PM, Bagi L, Manzinger D, Funk JA. 2015. Shiga toxin-producing *E. Coli* (STEC) in swine: Prevalence over the finishing period and characteristics of the STEC isolates. *Epidemiol Infect* 143:505–514. doi: 10.1017/S0950268814001095.
 38. Friesema IHM, Van De Kasstele J, De Jager CM, Heuvelink AE, Van Pelt W. 2011. Geographical association between livestock density and human Shiga toxin-producing *Escherichia coli* O157 infections. *Epidemiol Infect* 139:1081–1087. doi: 10.1017/S0950268810002050.
 39. Muniesa M, Recktenwald J, Bielaszewska M, Karch H, Schmidt H. 2000. Characterization of a Shiga toxin 2e-converting bacteriophage from an *Escherichia coli* strain of human origin. *Infect Immun* 68:4850–4855. doi: 10.1128/IAI.68.9.4850-4855.2000.
 40. Mcdaniel TK, Jarvis KG, Donnenberg MS, Kaper JB. 1995. A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens. *Proc Natl Acad Sci USA* 92:1664–1668. doi: 10.1073/pnas.92.5.1664.

41. Jerse AE, Yu J, Tall BD, Kaper JB. 1990. A genetic locus of enteropathogenic *Escherichia coli* necessary for the production of attaching and effacing lesions on tissue culture cells. *Proc Natl Acad Sci U S A* 87:7839–7843. doi: 10.1073/pnas.87.20.7839.
42. Louie M, De Azavedo JCS, Handelsman MYC, Clark CG, Ally B, Dytoc M, Sherman P, Brunton J. 1993. Expression and characterization of the *eaeA* gene product of *Escherichia coli* serotype O157:H7. *Infect Immun* 61:4085–4092.
43. Franzin FM, Sircili MP. 2015. Locus of enterocyte effacement: A pathogenicity island involved in the virulence of enteropathogenic and enterohemorrhagic *Escherichia coli* subjected to a complex network of gene regulation. *Biomed Res Int* 2015:1–10. doi: 10.1155/2015/534738.
44. Hartland EL, Batchelor M, Delahay RM, Hale C, Matthews S, Dougan G, Knutton S, Connerton I, Frankel G. 1999. Binding of intimin from enteropathogenic *Escherichia coli* to Tir and to host cells. *Mol Microbiol* 32:151–158. doi: 10.1046/j.1365-2958.1999.01338.x.
45. Ogura Y, Ooka T, Whale A, Garmendia J, Beutin L, Tennant S, Krause G, Morabito S, Chinen I, Tobe T, Abe H, Tozzoli R, Caprioli A, Rivas M, Robins-Browne R, Hayashi T, Frankel G. 2007. TccP2 of O157:H7 and non-O157 enterohemorrhagic *Escherichia coli* (EHEC): Challenging the dogma of EHEC-induced actin polymerization. *Infect Immun* 75:604–612. doi: 10.1128/IAI.01491-06.
46. Lai Y, Rosenshine I, Leong JM, Frankel G. 2013. Intimate host attachment: Enteropathogenic and enterohaemorrhagic *Escherichia coli*. *Cell Microbiol* 15:1796–1808. doi: 10.1111/cmi.12179.
47. Ramachandran V, Brett K, Hornitzky MA, Dowton M, Bettelheim KA, Walker MJ, Djordjevic SP. 2003. Distribution of Intimin Subtypes among *Escherichia coli* Isolates from Ruminant and Human Sources. *J Clin Microbiol* 41:5022–5032. doi: 10.1128/JCM.41.11.5022-5032.2003.
48. Gardette M, Le Hello S, Mariani-Kurkdjian P, Fabre L, Gravey F, Garrivier A, Loukiadis E, Jubelin G. 2019. Identification and prevalence of in vivo-induced genes in enterohaemorrhagic *Escherichia coli*. *Virulence* 10:180–193. doi: 10.1080/21505594.2019.1582976.
49. Phillips AD, Frankel G. 2000. Intimin-Mediated Tissue Specificity in Enteropathogenic *Escherichia coli* Interaction with Human Intestinal Organ Cultures. *J Infect Dis* 181:1496–1500. doi: 10.1086/315404.
50. Delannoy S, Chaves BD, Ison SA, Webb HE, Beutin L, Delaval JJ, Billet I, Fach P. 2016. Revisiting the STEC testing approach: Using *espK* and *espV* to make enterohemorrhagic *Escherichia coli* (EHEC) detection more reliable in beef. *Front Microbiol* 7:1–10. doi: 10.3389/fmicb.2016.00001.

51. Madic J, Peytavin De Garam C, Vingadassalon N, Oswald E, Fach P, Jamet E, Auvray F. 2010. Simplex and multiplex real-time PCR assays for the detection of flagellar (H-antigen) fliC alleles and intimin (eae) variants associated with enterohaemorrhagic *Escherichia coli* (EHEC) serotypes O26:H11, O103:H2, O111:H8, O145:H28 and O157:H7. *J Appl Microbiol* 109:1696–1705. doi: 10.1111/j.1365-2672.2010.04798.x.
52. Oswald E, Schmidt H, Morabito S, Karch H, Marchès O, Caprioli A. 2000. Typing of intimin genes in human and animal enterohemorrhagic and enteropathogenic *Escherichia coli*: Characterization of a new intimin variant. *Infect Immun* 68:64–71. doi: 10.1128/IAI.68.1.64-71.2000.
53. Bugarel M, Beutin L, Fach P. 2010. Low-density macroarray targeting non-locus of enterocyte effacement effectors (nle genes) and major virulence factors of Shiga toxin-producing *Escherichia coli* (STEC): A new approach for molecular risk assessment of STEC isolates. *Appl Environ Microbiol* 76:203–211. doi: 10.1128/AEM.01921-09.
54. Tarr CL, Whittam TS. 2002. Molecular evolution of the intimin gene in O111 clones of pathogenic *Escherichia coli*. *J Bacteriol* 184:479–487. doi: 10.1128/JB.184.2.479-487.2002.
55. Paton AW, Srimanote P, Woodrow MC, Paton JC. 2001. Characterization of Saa, a novel autoagglutinating adhesin produced by locus of enterocyte effacement-negative Shiga-toxigenic *Escherichia coli* strains that are virulent for humans. *Infect Immun* 69:6999–7009. doi: 10.1128/IAI.69.11.6999-7009.2001.
56. Jenkins C, Perry NT, Cheasty T, Shaw DJ, Frankel G, Dougan G, Gunn GJ, Smith HR, Paton AW, Paton JC. 2003. Distribution of the saa gene in strains of Shiga toxin-producing *Escherichia coli* of human and bovine origins. *J Clin Microbiol* 41:1775–8. doi: 10.1128/jcm.41.4.1775-1778.2003.
57. Croxen MA, Finlay BB. 2010. Molecular mechanisms of *Escherichia coli* pathogenicity. *Nat Rev Microbiol* 8:26–38. doi: 10.1038/nrmicro2265.
58. Low AS, Holden N, Rosser T, Roe AJ, Constantinidou C, Hobman JL, Smith DGE, Low JC, Gally DL. 2006. Analysis of fimbrial gene clusters and their expression in enterohaemorrhagic *Escherichia coli* O157:H7. *Environ Microbiol* 8:1033–1047. doi: 10.1111/j.1462-2920.2006.00995.x.
59. Makino K, Ishii K, Yasunaga T, Hattori M, Yokoyama K, Yutsudo CH, Kubota Y, Yamaichi Y, Iida T, Yamamoto K, Honda T, Han CG, Ohtsubo E, Kasamatsu M, Hayashi T, Kuhara S, Shinagawa H. 1998. Complete nucleotide sequences of 93-kb and 3.3-kb plasmids of an enterohemorrhagic *Escherichia coli* O157:H7 derived from Sakai outbreak. *DNA Res* 5:1–9. doi: 10.1093/dnares/5.1.1.
60. Lim JY, Yoon J, Hovde CJ. 2010. A brief overview of *Escherichia coli* O157:H7 and its plasmid O157. *J Microbiol Biotechnol* 20:5–14.

61. Fratamico PM, Yan X, Caprioli A, Esposito G, Needleman DS, Pepe T, Tozzoli R, Cortesi ML, Morabito S. 2011. The complete DNA sequence and analysis of the virulence plasmid and of five additional plasmids carried by Shiga toxin-producing *Escherichia coli* O26:H11 strain H30. *Int J Med Microbiol* 301:192–203. doi: 10.1016/j.ijmm.2010.09.002.
62. Lorenz SC, Monday SR, Hoffmann M, Fischer M, Kase JA. 2016. Plasmids from Shiga toxin-producing *Escherichia coli* strains with rare enterohemolysin gene (*ehxA*) subtypes reveal pathogenicity potential and display a novel evolutionary path. *Appl Environ Microbiol* 82:6367–6377. doi: 10.1128/AEM.01839-16.
63. Ogura Y, Ooka T, Asadulghani, Terajima J, Nougayrède JP, Kurokawa K, Tashiro K, Tobe T, Nakayama K, Kuhara S, Oswald E, Watanabe H, Hayashi T. 2007. Extensive genomic diversity and selective conservation of virulence-determinants in enterohemorrhagic *Escherichia coli* strains of O157 and non-O157 serotypes. *Genome Biol* 8:R138. doi: 10.1186/gb-2007-8-7-r138.
64. Brunder W, Schmidt H, Karch H. 1996. KatP, a novel catalase-peroxidase encoded by the large plasmid of enterohaemorrhagic *Escherichia coli* O157:H7. *Microbiology* 142:3305–3315. doi: 10.1099/13500872-142-11-3305.
65. Tozzoli R, Caprioli A, Morabito S. 2005. Detection of *tox*B, a plasmid virulence gene of *Escherichia coli* O157, in enterohemorrhagic and enteropathogenic *E. coli*. *J Clin Microbiol* 43:4052–4056. doi: 10.1128/JCM.43.8.4052-4056.2005.
66. Beutin L, Montenegro MA, Orskov I, Orskov F, Prada J, Zimmermann S, Stephan R. 1989. Close association of verotoxin (shiga-like toxin) production with enterohemolysin production in strains of *Escherichia coli*. *J Clin Microbiol* 27:2559–2564.
67. De Rauw K, Jacobs S, Piérard D. 2018. Twenty-seven years of screening for shiga toxin-producing *Escherichia coli* in a university hospital. Brussels, Belgium, 1987-2014. *PLoS One* 13:e0199968. doi: 10.1371/journal.pone.0199968.
68. Schmidt H, Karch H. 1996. Enterohemolytic phenotypes and genotypes of shiga toxin-producing *Escherichia coli* O111 strains from patients with diarrhea and hemolytic-uremic syndrome. *J Clin Microbiol* 34:2364–2367.
69. Cookson AL, Bennett J, Thomson-Carter F, Attwood GT. 2007. Molecular subtyping and genetic analysis of the enterohemolysin gene (*ehxA*) from Shiga toxin-producing *Escherichia coli* and atypical enteropathogenic *E. coli*. *Appl Environ Microbiol* 73:6360–6369. doi: 10.1128/AEM.00316-07.
70. Lorenz SC, Son I, Maounounen-Laasri A, Lin A, Fischer M, Kase JA. 2013. Prevalence of hemolysin genes and comparison of *ehxA* subtype patterns in Shiga toxin-producing *Escherichia coli* (STEC) and Non-STEC strains from clinical, food, and animal sources. *Appl Environ Microbiol* 79:6301–6311. doi: 10.1128/AEM.02200-13.
71. Fu S, Bai X, Fan R, Sun H, Xu Y, Xiong Y. 2018. Genetic diversity of the

- enterohaemolysin gene (ehxA) in non-O157 Shiga toxin-producing *Escherichia coli* strains in China. *Sci Rep* 8:1–8. doi: 10.1038/s41598-018-22699-7.
72. Gould LH, Demma L, Jones TF, Hurd S, Vugia DJ, Smith K, Shiferaw B, Segler S, Palmer A, Zansky S, Griffin PM. 2009. Hemolytic Uremic Syndrome and Death in Persons with *Escherichia coli* O157:H7 Infection, Foodborne Diseases Active Surveillance Network Sites, 2000–2006 . *Clin Infect Dis* 49:1480–1485. doi: 10.1086/644621.
 73. Tarr PI, Gordon CA, Chandler WL. 2005. Shiga-toxin-producing *Escherichia coli* and haemolytic uraemic syndrome. *Lancet* 365:1073–1086. doi: 10.1016/S0140-6736(05)71144-2.
 74. Wong CS, Mooney JC, Brandt JR, Staples AO, Jelacic S, Boster DR, Watkins SL, Tarr PI. 2012. Risk factors for the hemolytic uremic syndrome in children infected with *Escherichia coli* O157:H7: A multivariable analysis. *Clin Infect Dis* 55:33–41. doi: 10.1093/cid/cis299.
 75. Williams DM, Sreedhar SS, Mickell JJ, Chan JCM. 2002. Acute kidney failure: A pediatric experience over 20 years. *Arch Pediatr Adolesc Med* 156:893–900. doi: 10.1001/archpedi.156.9.893.
 76. Frank C, Kapfhammer S, Werber D, Stark K, Held L. 2008. Cattle density and Shiga toxin-producing *Escherichia coli* infection in Germany: increased risk for most but not all serogroups. *Vector Borne Zoonotic Dis* 8:635–643. doi: 10.1089/vbz.2007.0237.
 77. O’Brien SJ, Adak GK, Gilham C. 2001. Contact with farming environment as a major risk factor for Shiga toxin (Vero cytotoxin)-producing *Escherichia coli* O157 infection in humans. *Emerg Infect Dis* 7:1049–1051. doi: 10.3201/eid0706.010626.
 78. Schlager S, Lepuschitz S, Ruppitsch W, Ableitner O, Pietzka A, Neubauer S, Stöger A, Lassnig H, Mikula C, Springer B, Allerberger F. 2018. Petting zoos as sources of Shiga toxin-producing *Escherichia coli* (STEC) infections. *Int J Med Microbiol* 308:927–932. doi: 10.1016/J.IJMM.2018.06.008.
 79. Hadler JL, Clogher P, Hurd S, Phan Q, Mandour M, Bemis K, Marcus R. 2011. Ten-Year Trends and Risk Factors for Non-O157 Shiga Toxin-Producing *Escherichia coli* Found Through Shiga Toxin Testing, Connecticut, 2000–2009. *Clin Infect Dis* 53:269–276. doi: 10.1093/cid/cir377.
 80. Whitney BM, Mainero C, Humes E, Hurd S, Niccolai L, Hadler JL. 2015. Socioeconomic status and foodborne pathogens in Connecticut, USA, 2000–2011. *Emerg Infect Dis* 21:1617–1624. doi: 10.3201/eid2109.150277.
 81. ECDC. 2019. Shiga toxin/verocytotoxin-producing *Escherichia coli* (STEC/VTEC) infection. European Centre for Disease Prevention and Control. Stockholm.

82. Werber D, Beutin L, Pichner R, Stark K, Fruth A. 2008. Shiga toxin-producing *Escherichia coli* serogroups in food and patients, Germany. *Emerg Infect Dis* 14:1803–1806. doi: 10.3201/eid1411.080361.
83. Mellmann A, Fruth A, Friedrich AW, Wieler LH, Harmsen D, Werber D, Middendorf B, Bielaszewska M, Karch H. 2009. Phylogeny and disease association of Shiga toxin-producing *Escherichia coli* O91. *Emerg Infect Dis* 15:1474–1477. doi: 10.3201/eid1509.090161.
84. Marouani-Gadri N, Augier G, Carpentier B. 2009. Characterization of bacterial strains isolated from a beef-processing plant following cleaning and disinfection - Influence of isolated strains on biofilm formation by Sakai and EDL 933 *E. coli* O157:H7. *Int J Food Microbiol* 133:62–67. doi: 10.1016/j.ijfoodmicro.2009.04.028.
85. Jay MT, Garrett V, Mohle-Boetani JC, Barros M, Farrar JA, Rios R, Abbott S, Sowadsky R, Komatsu K, Mandrell R, Sobel J, Werner SB. 2004. A Multistate Outbreak of *Escherichia coli* O157:H7 Infection Linked to Consumption of Beef Tacos at a Fast-Food Restaurant Chain. *Clin Infect Dis* 39:1–7. doi: 10.1086/421088.
86. Ferens WA, Hovde CJ. 2011. *Escherichia coli* O157:H7: animal reservoir and sources of human infection. *Foodborne Pathog Dis* 8:465–87. doi: 10.1089/fpd.2010.0673.
87. Zschock M, Hamann HP, Kloppert B, Wolter W. 2000. Shiga-toxin-producing *Escherichia coli* in faeces of healthy dairy cows, sheep and goats: Prevalence and virulence properties. *Lett Appl Microbiol* 31:203–208. doi: 10.1046/j.1365-2672.2000.00789.x.
88. Grauke LJ, Kudva IT, Yoon JW, Hunt CW, Williams CJ, Hovde CJ. 2002. Gastrointestinal tract location of *Escherichia coli* O157:H7 in ruminants. *Appl Environ Microbiol* 68:2269–2277. doi: 10.1128/AEM.68.5.2269-2277.2002.
89. Ishii S, Meyer KP, Sadowsky MJ. 2007. Relationship between phylogenetic groups, genotypic clusters, and virulence gene profiles of *Escherichia coli* strains from diverse human and animal sources. *Appl Environ Microbiol* 73:5703–5710. doi: 10.1128/AEM.00275-07.
90. Beutin L, Geier D, Steinruck H, Zimmermann S, Scheutz F. 1993. Prevalence and some properties of verotoxin (Shiga-like toxin)-producing *Escherichia coli* in seven different species of healthy domestic animals. *J Clin Microbiol* 31:2483–2488.
91. Steil D, Bonse R, Meisen I, Pohlentz G, Vallejo G, Karch H, Müthing J. 2016. A topographical atlas of shiga toxin 2e receptor distribution in the tissues of weaned piglets. *Toxins (Basel)* 8:E357. doi: 10.3390/toxins8120357.
92. Venegas-Vargas C, Henderson S, Khare A, Mosci RE, Lehnert JD, Singh P, Ouellette LM, Norby B, Funk JA, Rust S, Bartlett PC, Grooms D, Manning SD. 2016. Factors associated with Shiga toxin-producing *Escherichia coli* shedding by dairy and beef cattle.

- Appl Environ Microbiol 82:5049–5056. doi: 10.1128/AEM.00829-16.
93. Menrath A, Wieler LH, Heidemanns K, Semmler T, Fruth A, Kemper N. 2010. Shiga toxin producing *Escherichia coli*: identification of non-O157:H7-Super-Shedding cows and related risk factors. *Gut Pathog* 2:7. doi: 10.1186/1757-4749-2-7.
 94. Callaway TR, Carr MA, Edrington TS, Anderson RC, Nisbet DJ. 2009. Diet, *Escherichia coli* O157:H7, and cattle: A review after 10 years. *Curr Issues Mol Biol* 11:67–80. doi: 10.21775/cimb.011.067.
 95. Dunn JR, Keen JE, Thompson RA. 2004. Prevalence of shiga-toxigenic *Escherichia coli* O157:H7 in adult dairy cattle. *J Am Vet Med Assoc* 224:1151–1158. doi: 10.2460/javma.2004.224.1151.
 96. Hussein HSS, Sakuma T. 2005. Invited review: Prevalence of Shiga toxin-producing *Escherichia coli* in dairy cattle and their products. *J Dairy Sci* 88:450–465. doi: 10.3168/jds.S0022-0302(05)72706-5.
 97. Mir RA, Weppelmann TA, Kang M, Bliss TM, DiLorenzo N, Lamb GC, Ahn S, Jeong KC. 2015. Association between animal age and the prevalence of Shiga toxin-producing *Escherichia coli* in a cohort of beef cattle. *Vet Microbiol* 175:325–331. doi: 10.1016/j.vetmic.2014.12.016.
 98. Mir RA, Weppelmann TA, Elzo M, Ahn S, Driver JD, Jeong KCC. 2016. Colonization of beef cattle by Shiga toxin producing *Escherichia coli* during the first year of life: A cohort study. *PLoS One* 11:1–16. doi: 10.1371/journal.pone.0148518.
 99. Cho S, Fossler CP, Diez-Gonzalez F, Wells SJ, Hedberg CW, Kaneene JB, Ruegg PL, Warnick LD, Bender JB. 2013. Herd-level risk factors associated with fecal shedding of Shiga toxin-encoding bacteria on dairy farms in Minnesota, USA. *Can Vet J* 54:693–697.
 100. Rivas M, Sosa-Estani S, Rangel J, Caletti MG, Vallés P, Roldán CD, Balbi L, Marsano de Mollar MC, Amoedo D, Miliwebsky E, Chinen I, Hoekstra RM, Mead P, Griffin PM. 2008. Risk factors for sporadic Shiga toxin-producing *Escherichia coli* infections in children, Argentina. *Emerg Infect Dis* 14:763–71. doi: 10.3201/eid1405.071050.
 101. Kintz E, Brainard J, Hooper L, Hunter P. 2017. Transmission pathways for sporadic Shiga-toxin producing *E. coli* infections: A systematic review and meta-analysis. *Int J Hyg Environ Health* 220:57–67. doi: 10.1016/j.ijheh.2016.10.011.
 102. Berry ED, Wells JE, Bono JL, Woodbury BL, Kalchayanand N, Norman KN, Suslow T V., López-Velasco G, Millner PD. 2015. Effect of proximity to a cattle feedlot on *Escherichia coli* O157:H7 contamination of leafy greens and evaluation of the potential for airborne transmission. *Appl Environ Microbiol* 81:1101–1110. doi: 10.1128/AEM.02998-14.
 103. MacDonald E, Dalane PK, Aavitsland P, Brandal LT, Wester AL, Vold L. 2014.

- Implications of screening and childcare exclusion policies for children with Shiga-toxin producing *Escherichia coli* infections: lessons learned from an outbreak in a daycare centre, Norway, 2012. *BMC Infect Dis* 14:673. doi: 10.1186/s12879-014-0673-2.
104. Dabke G, Le Menach A, Black A, Gamblin J, Palmer M, Boxall N, Booth L. 2014. Duration of shedding of Verocytotoxin-producing *Escherichia coli* in children and risk of transmission in childcare facilities in England. *Epidemiol Infect* 142:327–334. doi: 10.1017/S095026881300109X.
 105. Brown JA, Hite DS, Gillim-Ross LA, Maguire HF, Bennett JK, Patterson JJ, Comstock NA, Watkins AK, Ghosh TS, Vogt RL. 2012. Outbreak of shiga toxin-producing *Escherichia coli* serotype O26: H11 infection at a child care center in colorado. *Pediatr Infect Dis J* 31:379–383. doi: 10.1097/INF.0b013e3182457122.
 106. Snedeker KG, Shaw DJ, Locking ME, Prescott RJ. 2009. Primary and secondary cases in *Escherichia coli* O157 outbreaks: A statistical analysis. *BMC Infect Dis* 9:144. doi: 10.1186/1471-2334-9-144.
 107. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM. 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *J Clin Microbiol* 52:1501–1510. doi: 10.1128/JCM.03617-13.
 108. Ronholm J, Nasheri N, Petronella N, Pagotto F. 2016. Navigating microbiological food safety in the era of whole-genome sequencing. *Clin Microbiol Rev* 29:837–857. doi: 10.1128/CMR.00056-16.
 109. Woese CR, Stackebrandt E, Macke TJ, Fox GE. 1985. A Phylogenetic Definition of the Major Eubacterial Taxa. *Syst Appl Microbiol* 6:143–151. doi: 10.1016/S0723-2020(85)80047-3.
 110. Clarridge JE. 2004. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin Microbiol Rev* 17:840–862. doi: 10.1128/CMR.17.4.840-862.2004.
 111. Foxman B, Zhang L, Koopman JS, Manning SD, Marrs CF. 2005. Choosing an appropriate bacterial typing technique for epidemiologic studies. *Epidemiol Perspect Innov* 2:10. doi: 10.1186/1742-5573-2-10.
 112. Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, Ouellette LM, Mladonicky JM, Somsel P, Rudrik JT, Dietrich SE, Zhang W, Swaminathan B, Alland D, Whittam TS. 2008. From the Cover: Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks. *Proc Natl Acad Sci* 105:4868–4873. doi: 10.1073/pnas.0710834105.
 113. Hunter PR. 1990. Reproducibility and indices of discriminatory power of microbial typing methods. *J Clin Microbiol* 28:1903–1905.

114. Gaston MA, Hunter PR. 1989. Efficient selection of tests for bacteriological typing schemes. *J Clin Pathol* 42:763–766. doi: 10.1136/jcp.42.7.763.
115. Carleton HA, Gerner-Smidt P. 2016. Whole-genome sequencing is taking over foodborne disease surveillance: Public health microbiology is undergoing its biggest change in a generation, Replacing traditional methods with whole-genome sequencing. *Microbe* 11:311–317. doi: 10.1128/microbe.11.311.1.
116. Ribot EM, Fair MAA, Gautom R, Cameron DNN, Hunter SBB, Swaminathan B, Barrett TJ. 2006. Standardization of pulsed-field gel electrophoresis protocols for the subtyping of *Escherichia coli* O157:H7, *Salmonella*, and *Shigella* for PulseNet. *Foodborne Pathog Dis* 3:59–67. doi: 10.1089/fpd.2006.3.59.
117. Van Belkum A. 1994. DNA fingerprinting of medically important microorganisms by use of PCR. *Clin Microbiol Rev* 7:174–184. doi: 10.1128/CMR.7.2.174.
118. Adzitey F, Huda N, Ali GRR. 2013. Molecular techniques for detecting and typing of bacteria, advantages and application to foodborne pathogens isolated from ducks. *3 Biotech* 3:97–107. doi: 10.1007/s13205-012-0074-4.
119. Ludwig W, Klenk H-P. 2001. Overview: A Phylogenetic Backbone and Taxonomic Framework for Procaryotic Systematics. *Bergey's Manual® Syst Bacteriol* 49–65. doi: 10.1007/978-0-387-21609-6_8.
120. Tang YW, Ellis NM, Hopkins MK, Smith DH, Dodge DE, Persing DH. 1998. Comparison of phenotypic and genotypic techniques for identification of unusual aerobic pathogenic gram-negative bacilli. *J Clin Microbiol* 36:3674–3679.
121. Service RF. 2006. The race for the \$1000 genome. *Science* 311:1544–1546. doi: 10.1126/science.311.5767.1544.
122. Wolf MK. 1997. Occurrence, distribution, and associations of O and H serogroups, colonization factor antigens, and toxins of enterotoxigenic *Escherichia coli*. *Clin Microbiol Rev* 10:569–584. doi: 10.1128/cmr.10.4.569.
123. Fratamico PM, DebRoy C, Liu Y, Needleman DS, Baranzoni GM, Feng P. 2016. Advances in molecular serotyping and subtyping of *Escherichia coli*. *Front Microbiol* 7:644. doi: 10.3389/fmicb.2016.00644.
124. Eichhorn I, Heidemanns K, Semmler T, Kinnemann B, Mellmann A, Harmsen D, Anjum MF, Schmidt H, Fruth A, Valentin-Weigand P, Heesemann J, Suerbaum S, Karch H, Wieler LH. 2015. Highly virulent non-O157 enterohemorrhagic *Escherichia coli* (EHEC) serotypes reflect similar phylogenetic lineages, providing new insights into the evolution of EHEC. *Appl Environ Microbiol* 81:7041–7047. doi: 10.1128/AEM.01921-15.
125. Willey BM, McGeer AJ, Ostrowski MA, Kreiswirth BN, Low DE. 1994. The Use of Molecular Typing Techniques in the Epidemiologic Investigation of Resistant

- Enterococci. *Infect Control Hosp Epidemiol* 15:548–556. doi: 10.2307/30148408.
126. Sader HS, Pignatari AC, Leme IL, Burattini MN, Tancredi R, Hollis RJ, Jones RN. 1993. Epidemiologic typing of multiply drug-resistant *Pseudomonas aeruginosa* isolated from an outbreak in an intensive care unit. *Diagn Microbiol Infect Dis* 17:13–18. doi: 10.1016/0732-8893(93)90063-D.
 127. Rossney AS, Coleman DC, Keane CT. 1994. Antibigram-resistogram typing scheme for methicillin-resistant *Staphylococcus aureus* 41:430–440.
 128. March SB, Ratnam S. 1986. Sorbitol-MacConkey medium for detection of *Escherichia coli* O157:H7 associated with hemorrhagic colitis. *J Clin Microbiol* 23:869–872.
 129. Gerner-Smidt P, Kincaid J, Kubota K, Hise K, Hunter SB, Fair MA, Norton D, Woo-Ming A, Kurzynski T, Sotir MJ, Head M, Holt K, Swaminathan B. 2005. Molecular surveillance of Shiga toxigenic *Escherichia coli* O157 by PulseNet USA. *J Food Prot* 68:1926–1931. doi: 10.4315/0362-028X-68.9.1926.
 130. Barrett TJ, Gerner-Smidt P, Swaminathan B. 2006. Interpretation of pulsed-field gel electrophoresis patterns in foodborne disease investigations and surveillance. *Foodborne Pathog Dis* 3:20–31. doi: 10.1089/fpd.2006.3.20.
 131. Oakeson KF, Wagner JM, Rohrwasser A, Atkinson-Dunn R. 2018. Whole-genome sequencing and bioinformatic analysis of isolates from foodborne illness outbreaks of *Campylobacter jejuni* and *Salmonella enterica*. *J Clin Microbiol* 56:e00161-18. doi: 10.1128/JCM.00161-18.
 132. Bumgarner R. 2013. DNA microarrays: types, applications, and their future, p. 22.1.1-22.1.11. *In* *Current Protocols in Molecular Biology*. doi: 10.1002/0471142727.mb2201s101.
 133. Heller MJ. 2002. DNA Microarray Technology: Devices, Systems, and Applications. *Annu Rev Biomed Eng* 4:129–153. doi: 10.1146/annurev.bioeng.4.020702.153438.
 134. Fitzgerald C, Helsel LO, Nicholson MA, Olsen SJ, Swerdlow DL, Flahart R, Sexton J, Fields PI. 2001. Evaluation of methods for subtyping *Campylobacter jejuni* during an outbreak involving a food handler. *J Clin Microbiol* 39:2386–2390. doi: 10.1128/JCM.39.7.2386-2390.2001.
 135. McDougal LK, Steward CD, Killgore GE, Chaitram JM, McAllister SK, Tenover FC. 2003. Pulsed-Field Gel Electrophoresis Typing of Oxacillin-Resistant *Staphylococcus aureus* Isolates from the United States: Establishing a National Database. *J Clin Microbiol* 41:5113–5120. doi: 10.1128/JCM.41.11.5113-5120.2003.
 136. Sandt CH, Krouse DA, Cook CR, Hackman AL, Chmielecki WA, Warren NG. 2006. The key role of pulsed-field gel electrophoresis in investigation of a large multiserotype and multistate food-borne outbreak of *Salmonella* infections centered in Pennsylvania. *J Clin*

- Microbiol 44:3208–3212. doi: 10.1128/JCM.01404-06.
137. Neoh HM, Tan XE, Sapri HF, Tan TL. 2019. Pulsed-field gel electrophoresis (PFGE): A review of the “gold standard” for bacteria typing and current alternatives. *Infect Genet Evol* 74:103935. doi: 10.1016/j.meegid.2019.103935.
 138. Sobral BWS, McClelland M. 1992. Methyltransferases as Tools to Alter the Specificity of Restriction Endonucleases, p. 159–172. *In Methods in Molecular Biology*. Humana Press. doi: 10.1385/0-89603-229-9:159.
 139. Nelson M, Schildkraut I. 1987. The Use of DNA Methylases to Alter the Apparent Recognition Specificities of Restriction Endonucleases. *Methods Enzymol* 155:41–48. doi: 10.1016/0076-6879(87)55008-X.
 140. Murase T, Yamai S, Watanabe H. 1999. Changes in pulsed-field gel electrophoresis patterns in clinical isolates of enterohemorrhagic *Escherichia coli* O157:H7 associated with loss of Shiga toxin genes. *Curr Microbiol* 38:48–50.
 141. Jaros P, Dufour M, Gilpin B, Freeman MM, Ribot EM. 2015. PFGE for Shiga toxin-producing *Escherichia coli* O157:H7 (STEC O157) and non-O157 STEC. *Methods Mol Biol* 1301:171–189. doi: 10.1007/978-1-4939-2599-5_15.
 142. Leotta GA, Miliwebsky ES, Chinen I, Espinosa EM, Azzopardi K, Tennant SM, Robins-Browne RM, Rivas M. 2008. Characterisation of Shiga toxin-producing *Escherichia coli* O157 strains isolated from humans in Argentina, Australia and New Zealand. *BMC Microbiol* 8:46. doi: 10.1186/1471-2180-8-46.
 143. Okhravi N, Adamson P, Matheson MM, Towler HMA, Lightman S. 2000. PCR-RFLP–Mediated Detection and Speciation of Bacterial Species Causing Endophthalmitis. *Invest Ophthalmol Vis Sci* 41:1419–1447.
 144. Hyytiä-Trees E, Smole SC, Fields PA, Swaminathan B, Ribot EM. 2006. Second generation subtyping: A proposed PulseNet protocol for multiple-locus variable-number tandem repeat analysis of Shiga toxin-producing *Escherichia coli* O157 (STEC O157). *Foodborne Pathog Dis* 3:118–131. doi: 10.1089/fpd.2006.3.118.
 145. Lista F, Faggioni G, Valjevac S, Ciammaruconi A, Vaissaire J, Le Doujet C, Gorgé O, De Santis R, Carattoli A, Ciervo A, Fasanella A, Orsini F, D’Amelio R, Pourcel C, Cassone A, Vergnaud G. 2006. Genotyping of *Bacillus anthracis* strains based on automated capillary 25-loci Multiple Locus Variable-Number Tandem Repeats Analysis. *BMC Microbiol* 6:33. doi: 10.1186/1471-2180-6-33.
 146. Pourcel C, André-Mazeaud F, Neubauer H, Ramisse F, Vergnaud G. 2004. Tandem repeats analysis for the high resolution phylogenetic analysis of *Yersinia pestis*. *BMC Microbiol* 4:22. doi: 10.1186/1471-2180-4-22.
 147. Le Flèche P, Fabre M, Denoeud F, Koeck J-L, Vergnaud G. 2002. High resolution, on-line

- identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing. *BMC Microbiol* 2:37.
148. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA* 95:3140–3145. doi: 10.1073/pnas.95.6.3140.
 149. Maiden MCJ. 2006. Multilocus Sequence Typing of Bacteria. *Annu Rev Microbiol* 60:561–588. doi: 10.1146/annurev.micro.59.030804.121325.
 150. Enright MC, Spratt BG. 1998. A multilocus sequence typing scheme for *Streptococcus pneumoniae*: Identification of clones associated with serious invasive disease. *Microbiology* 144:3049–3060. doi: 10.1099/00221287-144-11-3049.
 151. Ruiz-Garbajosa P, Bonten MJM, Robinson DA, Top J, Nallapareddy SR, Torres C, Coque TM, Cantón R, Baquero F, Murray BE, Del Campo R, Willems RJL. 2006. Multilocus sequence typing scheme for *Enterococcus faecalis* reveals hospital-adapted genetic complexes in a background of high rates of recombination. *J Clin Microbiol* 44:2220–2228. doi: 10.1128/JCM.02596-05.
 152. Qi W, Lacher DW, Bumbaugh AC, Hyma KE, Ouellette LM, Large TM, Tarr CL, Whittam TS. 2004. EcMLST: An online database for multi locus sequence typing of pathogenic *Escherichia coli*. *Comput Syst Bioinform Conf* 2004:520–521. doi: 10.1109/csb.2004.1332482.
 153. Feil EJ, Spratt BG. 2001. Recombination and the Population Structures of Bacterial Pathogens. *Annu Rev Microbiol* 55:561–590. doi: 10.1146/annurev.micro.55.1.561.
 154. Aanensen DM, Spratt BG. 2005. The multilocus sequence typing network: *mlst.net*. *Nucleic Acids Res* 33:W728–W733. doi: 10.1093/nar/gki415.
 155. Moura A, Tourdjman M, Leclercq A, Hamelin E, Laurent E, Fredriksen N, van Cauteren D, Bracq-Dieye H, Thouvenot P, Vales G, Tessaud-Rita N, Maury MM, Alexandru A, Criscuolo A, Quevillon E, Donguy MP, Enouf V, de Valk H, Brisse S, Lecuit M. 2017. Real-time whole-genome sequencing for surveillance of *Listeria monocytogenes*, France. *Emerg Infect Dis* 23:1462–1470. doi: 10.3201/eid2309.170336.
 156. Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A, Tarr C, Björkman JT, Dallman T, Reimer A, Enouf V, Larssonneur E, Carleton H, Bracq-Dieye H, Katz LS, Jones L, Touchon M, Tourdjman M, Walker M, Stroika S, Cantinelli T, Chenal-Francisque V, Kucerova Z, Rocha EPC, Nadon C, Grant K, Nielsen EM, Pot B, Gerner-Smidt P, Lecuit M, Brisse S. 2016. Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. *Nat Microbiol* 2:16185. doi: 10.1038/nmicrobiol.2016.185.

157. Pearce ME, Alikhan NF, Dallman TJ, Zhou Z, Grant K, Maiden MCJ. 2018. Comparative analysis of core genome MLST and SNP typing within a European *Salmonella* serovar Enteritidis outbreak. *Int J Food Microbiol* 274:1–11. doi: 10.1016/j.ijfoodmicro.2018.02.023.
158. Zhou Z, Alikhan N-F, Mohamed K, Group the AS, Achtman M. 2019. The user's guide to comparative genomics with EnteroBase. Three case studies: micro-clades within *Salmonella enterica* serovar Agama, ancient and modern populations of *Yersinia pestis*, and core genomic diversity of all *Escherichia*. *bioRxiv* 613554. doi: 10.1101/613554.
159. Pul Ü, Wurm R, Arslan Z, Geißen R, Hofmann N, Wagner R. 2010. Identification and characterization of *E. coli* CRISPR-cas promoters and their silencing by H-NS. *Mol Microbiol* 75:1495–1512. doi: 10.1111/j.1365-2958.2010.07073.x.
160. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315:1709–1712. doi: 10.1126/science.1138140.
161. Di H, Ye L, Yan H, Meng H, Yamasak S, Shi L. 2014. Comparative analysis of CRISPR loci in different *Listeria monocytogenes* lineages. *Biochem Biophys Res Commun* 454:399–403. doi: 10.1016/j.bbrc.2014.10.018.
162. Barrangou R, Horvath P. 2012. CRISPR: New Horizons in Phage Resistance and Strain Identification. *Annu Rev Food Sci Technol* 3:143–162. doi: 10.1146/annurev-food-022811-101134.
163. Garneau JE, Dupuis MÈ, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadán AH, Moineau S. 2010. The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* 468:67–71. doi: 10.1038/nature09523.
164. Horvath P, Barrangou R. 2010. CRISPR/Cas, the immune system of Bacteria and Archaea. *Science* 327:167–170. doi: 10.1126/science.1179555.
165. Toro M, Cao G, Ju W, Allard M, Barrangou R, Zhao S, Brown E, Meng J. 2013. Association of CRISPR elements with serotypes and virulence potential of Shiga toxin-producing *Escherichia coli*. *Appl Environ Microbiol*.
166. Yin S, Jensen MA, Bai J, DebRoy C, Barrangou R, Dudley EG. 2013. The Evolutionary Divergence of Shiga Toxin-Producing *Escherichia coli* Is Reflected in Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) Spacer Composition. *Appl Environ Microbiol* 79:5710–5720. doi: 10.1128/AEM.00950-13.
167. Fricke WF, Mammel MK, McDermott PF, Tartera C, White DG, LeClerc JE, Ravel J, Cebula TA. 2011. Comparative genomics of 28 *Salmonella enterica* isolates: Evidence for crispr-mediated adaptive sublineage evolution. *J Bacteriol* 193:3556–3568. doi: 10.1128/JB.00297-11.

168. McGhee GC, Sundin GW. 2012. *Erwinia amylovora* CRISPR elements provide new tools for evaluating strain diversity and for microbial source tracking. *PLoS One* 7:e41706. doi: 10.1371/journal.pone.0041706.
169. Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajj SA, Allix C, Aristimuño L, Arora J, Baumanis V, Binder L, Cafrune P, Cataldi A, Cheong S, Diel R, Ellermeier C, Evans JT, Fauville-Dufaux M, Ferdinand S, De Viedma DG, Garzelli C, Gazzola L, Gomes HM, Gutierrez MC, Hawkey PM, Van Helden PD, Kadival G V., Kreiswirth BN, Kremer K, Kubin M, Kulkarni SP, Liens B, Lillebaek T, Ho ML, Martin C, Martin C, Mokrousov I, Narvskaja O, Yun FN, Naumann L, Niemann S, Parwati I, Rahim Z, Rasolofo-Razanamparany V, Rasolonavalona T, Rossetti ML, Rüsche-Gerdes S, Sajduda A, Samper S, Shemyakin IG, Singh UB, Somoskov A, Skuce RA, Van Soolingen D, Streicher EM, Suffys PN, Tortoli E, Tracevska T, Vincent V, Victor TC, Warren RM, Sook FY, Zaman K, Portaels F, Rastogi N, Sola C. 2006. *Mycobacterium tuberculosis* complex genetic diversity: Mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiol* 6:23. doi: 10.1186/1471-2180-6-23.
170. Mokrousov I, Limeschenko E, Vyazovaya A, Narvskaya O. 2007. *Corynebacterium diphtheriae* spoligotyping based on combined use of two CRISPR loci. *Biotechnol J* 2:901–6. doi: 10.1002/biot.200700035.
171. Fabre L, Zhang J, Guigon G, Le Hello S, Guibert V, Accou-Demartin M, de Romans S, Lim C, Roux C, Passet V, Diancourt L, Guibourdenche M, Issenhuth-Jeanjean S, Achtman M, Brisse S, Sola C, Weill F-X. 2012. CRISPR typing and subtyping for improved laboratory surveillance of *Salmonella* infections. *PLoS One* 7:e36995. doi: 10.1371/journal.pone.0036995.
172. Touchon M, Charpentier S, Clermont O, Rocha EPC, Denamur E, Branger C. 2011. CRISPR distribution within the *Escherichia coli* species is not suggestive of immunity-associated diversifying selection. *J Bacteriol* 193:2460–2467. doi: 10.1128/JB.01307-10.
173. Groenen PMA, Bunschoten AE, Soolingen D van, Erftink JDA va. 1993. Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. *Mol Microbiol* 10:1057–1065. doi: 10.1111/j.1365-2958.1993.tb00976.x.
174. Van der Zanden AGM, Kremer K, Schouls LM, Caimi K, Cataldi A, Hulleman A, Nagelkerke NJD, Van Soolingen D. 2002. Improvement of differentiation and interpretability of spoligotyping for *Mycobacterium tuberculosis* complex isolates by introduction of new spacer oligonucleotides. *J Clin Microbiol* 40:4628–4639. doi: 10.1128/JCM.40.12.4628-4639.2002.
175. Sola C, Filliol I, Gutierrez MC, Mokrousov I, Vincent V, Rastogi N. 2001. Spoligotype Database of *Mycobacterium tuberculosis*: Biogeographic Distribution of Shared Types and Epidemiologic and Phylogenetic Perspectives - Volume 7, Number 3—June 2001 -

Emerging Infectious Disease journal - CDC. *Emerg Infect Dis* 7:390–396. doi: 10.3201/EID0703.017304.

176. Delannoy S, Beutin L, Fach P. 2015. Improved traceability of Shiga-toxin-producing *Escherichia coli* using CRISPRs for detection and typing. *Environ Sci Pollut Res* 23:8163–8174. doi: 10.1007/s11356-015-5446-y.
177. Jiang Y, Yin S, Dudley EG, Cutter CN. 2015. Diversity of CRISPR loci and virulence genes in pathogenic *Escherichia coli* isolates from various sources. *Int J Food Microbiol* 204:41–46. doi: 10.1016/j.ijfoodmicro.2015.03.025.
178. Delannoy S, Beutin L, Fach P. 2012. Use of clustered regularly interspaced short palindromic repeat sequence polymorphisms for specific detection of enterohemorrhagic *Escherichia coli* strains of serotypes O26:H11, O45:H2, O103:H2, O111:H8, O121:H19, O145:H28, and O157:H7 by real-time PCR. *J Clin Microbiol* 50:4035–4040. doi: 10.1128/JCM.02097-12.
179. Shariat N, Dudley EG. 2014. CRISPRs: Molecular Signatures Used for Pathogen Subtyping. *Appl Environ Microbiol* 80:430–439. doi: 10.1128/AEM.02790-13.
180. Pightling AW, Petronella N, Pagotto F. 2014. Choice of reference sequence and assembler for alignment of *Listeria monocytogenes* short-read sequence data greatly influences rates of error in SNP analyses. *PLoS One* 9:e104579. doi: 10.1371/journal.pone.0104579.
181. Katz LS, Griswold T, Williams-Newkirk AJ, Wagner D, Petkau A, Sieffert C, Domselaar G Van, Deng X, Carleton HA. 2017. A comparative analysis of the Lyve-SET phylogenomics pipeline for genomic epidemiology of foodborne pathogens. *Front Microbiol* 8:375. doi: 10.3389/fmicb.2017.00375.
182. Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V, Ravel J. 2008. The pangenome structure of *Escherichia coli*: Comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol* 190:6881–6893. doi: 10.1128/JB.00619-08.
183. Tettelin H, Massignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli S V., Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit Y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial “pan-genome.” *Proc Natl Acad Sci USA* 102:13950–13955. doi: 10.1073/pnas.0508532102.

CHAPTER 2

TRENDS OF SHIGA TOXIN-PRODUCING *ESCHERICHIA COLI* (STEC) STRAINS RECOVERED FROM PATIENTS IN MICHIGAN, 2001-2018

ABSTRACT

Shiga toxin-producing *Escherichia coli* (STEC) is a leading foodborne pathogen, resulting in 265,000 illnesses and 3,600 hospitalizations annually. STEC can be classified into over 180 different O serogroups, with O157 being the primary focus of most research studies. Other non-O157 serogroups, however, are associated with clinical symptoms and are increasing in frequency. Indeed, the incidence of non-O157 serogroups has now surpassed the incidence of O157 and has identified a gap in our knowledge of the genetic diversity of STEC and non-O157 serogroups. In this study, whole genome sequencing was used to examine the virulence profiles and genetic diversity of non-O157 STEC recovered from clinical cases between 2001 and 2018 (n=894). A total of 69 serogroups were identified over the entire time period, though the diversity increased from an average of 5 (2001-2006) to 18.3 (2008-2018) per year. These strains represented 52 distinct multilocus sequence types (STs) and 14 of these STs were novel with new gene variants or allele profiles. The MLST-based phylogeny identified four clusters with >80% bootstrapping support. The clusters were associated with specific *stx*, *eae*, and *ehxA* profiles. This comprehensive analysis of non-O157 STEC strains in Michigan over an 18-year period is helpful to understand changes in circulating genotype distributions and to better understand the genotypes that may be associated with disease or specific clinical outcomes.

INTRODUCTION

Shiga toxin producing *Escherichia coli* (STEC) is a foodborne pathogen that results in 265,000 illnesses annually (1). STEC can present as a wide range of gastrointestinal symptoms including cramps, diarrhea, hemorrhagic colitis and in severe cases, hemolytic uremic syndrome (HUS) or kidney failure (2–5). The identification of O157 STEC in 1982 as the etiological agent associated with contaminated hamburgers led to research and surveillance focusing on O157 STEC (6, 7). However, the increasing incidence of non-O157 serogroups associated with clinical illness and outbreaks has led to the classification of non-O157 STEC as a nationally notifiable disease (8–10). Within the non-O157 STEC serogroups associated with illness in the US, six serogroups, commonly referred to as the “big six” non-O157 serogroups, predominate and include: O26, O45, O103, O111, O121 and O145 (11). Although these serogroups are commonly isolated in the US associated with illness, other serogroups outside of the big six such as, O91 and O146, have been more frequently isolated in other countries (9, 12). Despite the wide range of non-O157 STEC serogroups associated with disease, little research has been done to understand the genetic characteristics and to identify epidemiological factors that may be associated with these non-O157 serogroups (9, 11, 13, 14).

STEC isolates are characterized by the presence of the Shiga toxin genes (*stx1* and/or *stx2*) that are found on lysogenic bacteriophages (15). Over the past two decades, STEC identification has changed from culturing STEC to using culture independent tests (CIDTs) (16). Initially, sorbitol-MacConkey (SMAC) agar, which is specific for O157 STEC as it relies on the ability of O157 but not non-O157 serogroups to ferment sorbitol, was used in most public health laboratories. In 2009, the Centers for Disease Control and Prevention (CDC) recommended use

of specific CIDs that directly identify the Shiga toxin genes or the presence of the Shiga toxin (8, 17, 18), thereby enhancing detection of non-O157 STEC. The ability to better detect *stx* positive isolates in addition to those that are O157 based on SMAC culturing may influence the increasing incidence rate of non-O157 STEC.

Shiga toxin variants for both *stx1* and *stx2* have been identified and associated with various clinical outcomes. Variants of *stx1*, *stx1a*, *stx1c* and *stx1d*, are antigenically distinct and strains with *stx1c* and *stx1d* are infrequently recovered from patients and have been associated with milder clinical symptoms (19). By contrast, there are seven variants of *stx2* (a-g) that have been identified; *stx2a* and *stx2d* were found to be associated with more severe infections resulting in HUS (20–23). The *stx2e*, *stx2f* and *stx2g* variants, however, have been more commonly recovered from environmental sources and animal reservoirs (24). Examination of other virulence factors and their variants such as *eae* (intimin) and *ehxA* (enterohemolysin), have been associated with varying degrees of disease severity and clinical outcomes (25–27). The *eae* gene, which is encoded within the locus of enterocyte effacement (LEE) pathogenicity island in the chromosomal DNA, is responsible for the attachment and effacement of intestinal epithelial cells. Patients infected with STEC strains possessing *eae* are more likely to present with bloody diarrhea as these strains were found to be more virulent than LEE-negative STEC (11). Enteropathogenic *E. coli* (EPEC), a *stx*-negative, *eae*-positive *E. coli* pathotype, as well as other *stx*-negative pathotypes have also been isolated from patients with diarrhea and have been linked to outbreaks. It is therefore clear that other virulence factors besides the Shiga toxin play a role in *E. coli*-associated enteric infections (28). Indeed, fourteen *eae* subtypes have been characterized, thereby adding to the diversity of the pathogenic *E. coli* population. Analysis of *eae* subtypes has identified that ruminants are a significant source of diverse subtypes (29).

Several hemolysin genes including the *ehxA* gene encoded on the large enterohemorrhagic virulence plasmid, have also been found in STEC isolates associated with severe disease such as HUS (26). While there are other hemolysins that have been found in *E.coli* such as alpha-hemolysins (HlyA) and silent hemolysin (SheA), *ehxA* is correlated with the presence of the *stx* genes indicating that it may be a virulence factor and play a role in disease outcomes for STEC-specific illnesses (19, 25, 30, 31). Six *ehxA* subtypes (A-F) have been characterized and can be distinguished by sequencing and restriction fragment length polymorphism analysis (29). Subtype A is commonly isolated from environmental and animal samples, while subtype C is commonly associated with clinical cases (25, 26). Virulence genes have not been found to be serotype specific, thus, serotype similarity alone is not enough to identify whether isolates have the potential to cause clinical outcomes or share a genetic profile for outbreak investigations.

Over 180 STEC serogroups have been identified, however, the relatedness of these serogroups and the spectrum in which they are able to cause disease has not been fully elucidated (10, 11, 32, 33). Various typing methods have been used to examine relationships between strains belonging to non-O157 serogroups including pulsed-field gel electrophoresis (PFGE), multilocus variable number of tandem repeat analysis (MLVA) and MLST (34, 35). A key limitation with these methods is their reduced ability to further discriminate strains that cluster together to define their genetic relatedness with a higher discriminatory power.

Through this study, we sought to examine the changes in genetic factors in STEC isolates that were isolated in Michigan over an 18-year time period. Changes within groups of isolates that are classified as the same serotype and/or ST were examined to identify genomic traits that

are important for more severe clinical infections. The trends that are examined within Michigan over an extended time period will help with future public health interventions and understanding of the genetic diversity and factors that are important for STEC infections.

MATERIALS AND METHODS

Bacterial strains and epidemiological data

The Michigan Department of Health and Human Services (MDHHS) recovered 1,926 STEC isolates from patient samples from 2001-2018. All non-O157 isolates (n=894) recovered during this period were examined using whole genome sequencing (WGS). Epidemiological data for a subset of 552 cases were extracted from the Michigan Disease Surveillance System (MDSS) for cases with infections reported between 2001 and 2015; no epidemiological data was available for cases with isolates submitted after 2015. All data were maintained in Microsoft Access and Excel.

Ethics Statement

Data collection from human subjects was approved by the Institutional Review Board at Michigan State University (MSU; Lansing, MI, USA; IRB #10-736SM) and the MDHHS (842-PHALAB).

DNA isolation and whole genome sequencing (WGS)

Overnight cultures were grown aerobically at 37°C in Luria-Bertani broth. DNA isolation was performed with various methods depending on the time period the isolates were obtained. DNA for isolates recovered between 2001 and 2006 were extracted with the Wizard[®] Genomic DNA purification kit, while isolated recovered from 2007-2018 were extracted with the Qiagen DNAeasy spin column kit (Qiagen, Valencia, CA, USA) following PulseNet protocols established at the Centers for Disease Control and Prevention (CDC) (36–38). All isolates were

prepped with the Nextera XT kit (Illumina, San Diego, CA, USA) and sequenced on the Illumina MiSeq platform (2x250 reads).

Bioinformatic and in silico analysis

Raw sequencing reads were processed with Trimmomatic to trim adapters and remove sequences with a quality score less than 20 (Q20) or less than 100 nucleotides in length before assessment with FastQC for a quality check of the reads (39, 40). *De novo* assembly was performed with Spades 3.10.1, using kmers 21, 33, 55, 77, 99, 127 with error correction to minimize mismatches present in the final contigs (41).

Extraction of serotyping and virulence genes were performed using Abricate with the Center for Genomic Epidemiology databases for *wzx/wzy* (O-antigen), *fliC* (H-antigen) and *stx* genes (42) (www.genomicepidemiology.com). Curated databases were compiled for *ehxA* (enterohemolysin) and *eae* (intimin) gene variants. In house bioinformatic scripts were developed to extract the seven multilocus sequence typing (MLST) loci and sequence types (STs) were assigned using the EcMLST v1.2 database (<http://www.shigatox.net>)(43).

Data analysis and visualization

MEGA X was utilized to concatenate and align MLST alleles using CLUSTALW and to construct a neighbor-joining tree with 1000 bootstrap replication (44). Epidemiological associations between demographic variables, serogroups and virulence profiles were examined using Chi-Square (χ^2) and Mantel-Haenszel Chi-Square test; sample sizes less than five were examined using the Fisher's exact test. Statistical tests were performed in SAS v9.3 (SAS Institute, Cary, NC); $p < 0.05$ was considered significant and was reported along with the odds ratio (OR) and 95% confidence interval (CI) in a univariate analysis.

RESULTS

Case demographics of patients with non-O157 STEC infections in Michigan

A total of 894 non-O157 STEC isolates were recovered from MDHHS during 2001-2018 as part of various surveillance studies. For the period between 2001 and 2012, the total number of non-O157 isolates reported by MDHHS was determined based on records listed in MDSS in a previous study (10). Large discrepancies (>5 isolate difference) were observed for 2001, 2007, 2008, and 2010 to 2012 between the number of cases reported in MDSS and the number of isolates that were submitted for WGS (Figure 2.1). 2001 was the only year in which more non-O157s were recovered and typed with WGS than were reported by MDHHS. Overall, the period between 2010 and 2012 had more non-O157 case reports in MDSS when compared to number of isolates submitted for WGS.

Examination of the demographic data associated with isolates identified 504 (59.3%) female cases out of the 850 (95%) non-O157 STEC cases with gender data available. The proportion of female cases was significantly higher ($p < 0.0001$) than the proportion of men over the entire time period, and several factors were associated with gender. For example, women were significantly more likely to present with body aches than men (OR: 1.6, 95% CI: 1.10-2.47) (Table 2.1). Women were also significantly more likely to be hospitalized (OR: 1.7, 95% CI: 1.10-2.47). Across the four age groups, the gender proportion for cases between 11-29 (female: 219/383, $p < 0.0001$), 30-64 (female: 165/230, $p = 0.008$), and ≥ 65 years (female: 49/75, $p = 0.0048$) were significantly different. The proportion of cases identified as women ranged from 57.2%-71.7% (Table 2.2).

Age frequencies over the past 18 years fluctuated throughout the first 10 years before becoming more stable from 2013-2018 (Figure 2.2). The average age over the time period was

29.0 years, while the highest frequency of cases between 11-29 years of age was observed in 13 of the 18 years examined. In 2007, only three non-O157 STEC isolates were received, and all were from patients in the 11-29 age group. Across all age groups, significant differences were identified when stratified by serogroup (Mantel-Haenszel χ^2 $p=0.04$), *stx* variants (Mantel-Haenszel χ^2 $p<0.0001$) and *eaeA* presence (Mantel-Haenszel χ^2 $p<0.0001$) (Table 2.2). All age groups had a high proportion of O103 (23.1-28.2%), while O26 was highest in younger children between 0 and 10 years (17.9%) and in the 11-29 (17.3%) age group. The frequency of non-O157 big six serogroups was significantly higher in children and young adults (0-29 years) (OR: 1.5, 95%CI: 1.10, 2.13). Differences in the presence of key STEC virulence genes was also observed, particularly among the adult and elderly population compared to the children and young adults. For instance, *stx2* frequency increased in proportion from 8.0% among the 0-10 age group to 28.2% among elderly patients ≥ 65 years ($p<0.0001$). Conversely, the presence of *eaeA* decreased from 98.8% in cases 0-10 years to 80.8% in cases ≥ 65 years ($p<0.0001$).

Clinical outcomes for abdominal pain and body aches were reported along with age for 503 cases, while hospitalization status was reported for 515 (57.6%). A significant difference in proportions across all age groups was detected for all three clinical outcomes; abdominal pain ($p<0.0001$), body ache ($p=0.0079$), and hospitalization ($p<0.0001$). A univariate analysis was performed within each outcome using 11-29 years of age as the reference category to identify associations by age group. Over the entire time period, young children were significantly less likely to be hospitalized (OR: 0.4; 95% CI: 0.20-0.71), while elderly cases were 2.7 times more likely to require hospitalization (95% CI: 1.41-5.44).

Distribution of serogroups and association with clinical outcomes

Over the past 18 years, non-O157 STEC representing the big six serogroups have been more frequently isolated than other non-O157 serogroups; between 60-80% of the isolates were classified as a big six serogroup for 16 of the past 18 years (Figure 2.3). Importantly, excluding 2007, other non-O157 isolates were identified from at least 20% or more of the cases each year. Isolates classified as one of the big six serogroups were significantly more likely to be recovered from males (OR: 1.5, 95%CI: 1.07-4.46) than females (Table 2.3). The virulence gene profiles of non-O157 big six isolates also differed compared to the remainder of the isolates representing all other serogroups. Big six strains were significantly more likely to have *stx1* or *stx1,2* (OR: 2.9, 95% CI: 1.90, 4.46), *eaeA* (Fisher's χ^2 $p < 0.0001$), and *ehxA* (OR: 4.7, 95%CI: 2.64, 8.45) compared to other serogroups (Table 2.3). While there was a significant difference in the virulence profiles between big six and other non-O157 serogroups, the only association with clinical outcomes identified was between big six cases and history of bloody diarrhea (OR: 1.9, 95%CI: 1.19, 3.08), which is an indicator for more severe infections.

Further breaking the non-O157 serogroup distribution by year identified an increase in the number and diversity of serogroups that were identified over time (Figure 2.4). A total of 69 different serogroups were identified over the 18-year period. Since 2001 there has been a steadily increasing trend in the number of serogroups that have been identified each year. From 2001 to 2006, an average of 5 serogroups were reported each year compared to an average of 18.3 serogroups per year from 2008-2018. While a large number of different serogroups were identified, associations between clinical outcomes were only evaluated for the big six serogroups due to the small sample sizes of certain serogroups. Most notably, serogroup O45 was associated with case hospitalization (OR: 1.8, 95% CI: 1.18-2.72) (Table 2.4) as well as bloody diarrhea

(OR: 1.5, 95% CI: 1.01, 2.30) relative to all other serogroups (Table 2.5). Bloody diarrhea was also significantly more common among O111 infections (OR: 2.0, 95% CI: 1.05, 3.66), however, cases with O111 infections were not more likely to be hospitalized. No other associations between specific serogroups and different clinical outcomes were observed.

Trends of virulence genes and subtypes

Eleven different virulence gene profiles were identified from all non-O157 STEC based on the presence or absence of *stx*, *eae* and *ehxA* (Table 2.6). Most of the isolates (89.1%) had both *eae* and *ehxA* present along with one or more *stx* genes; however, 2.0% of the isolates were missing *eae* and *ehxA*, while the remaining 8.9% of the isolates were missing either *eae* or *ehxA*. The 10.9% of the isolates that lacked *eae* and/or *ehxA* suggests that there are other virulence factors that play a role in disease progression, specifically in the serogroups outside of the big six.

Within the big six serogroups, the virulence genes can be further differentiated into gene variants or alleles (Figure 2.5). The *stx1a* variant predominated among strains belonging to serogroups O26 (n=127; 97.7%), O45 (n=189; 99.5%), O103 (n=220; 100.0%) and O111 (n=94; 100.0%). Conversely, serogroups O121 (n=48; 98.0%) and O145 (n=14; 87.5%) predominantly encoded the *stx2a* variant (Figure 2.5A). A subset of the isolates (n=57, 6.4%) encoded the *stx1a2a* combination. The *stx2c* variant was isolated from one O145 and three O177 isolates, one of the O177 isolates also encoded the *stx2a* variant. A single isolate, serogroup O8, encoded the rare *stx2e* variant. All other non-O157 serogroups possessed a large range of *stx* variants including *stx2d*, which was not isolated from any of the big six serogroups.

Similar to the *stx* variants, strains belonging to a given serogroup only possessed one type of *eae* variant, or rather, specific serogroups were mainly comprised of strains with a single *eae* variant (Figure 2.5B). Serogroup O45 (n= 190, 100%), O103 (n=207, 94.1%), and O121 (n= 48, 98.0%), for instance, mainly encoded the epsilon variant, while O26 (n= 130, 100%) strains only encoded the beta variant. O111 (n= 93, 98.9%) and O145 (n= 15, 93.8%) strains possessed theta and gamma variants, respectively. The largest diversity of gene variants was identified in those other serogroups outside of the big six. For *ehxA*, the big-six serogroups primarily had *ehxA-C*, which were mainly found in O26 (n= 126, 96.9%), O111 (n=87, 92.6%), O121 (n= 46, 93.9%) and O145 (n= 16, 100%) strains (Figure 2.5C). The *ehxA-F* variants, however, were predominantly identified in O45 (n= 185, 97.4) and O103 (n= 131, 59.6%) strains. Similar to other gene variants that were examined, non-O157 serogroups besides the big six had the largest diversity of *ehxA* variants.

Genetic diversity of non-O157 STEC and association of clusters with disease

MLST was used to examine the genetic diversity of 883 non-O157 STEC with data available (Figure 2.6). In all, 66 STs were identified and 14 of these were newly identified STs; nine of these new STs had new allele combinations and five had new SNPs in existing gene alleles including ST-NEW4 (*uidA4*), NEW6 (*aspC7*), NEW7 (*lysP1*), NEW8 (*uidA2*), and NEW12 (*aspC6*). Non-O157 big six serogroups are found throughout the phylogenetic tree and clustered together with other non-O157 serogroups. Four clusters of STs were identified with >0.80 bootstrap support.

Cluster 1 is comprised of ST-106/104 and four additional STs representing 279 isolates. Big six serogroups O26 (n=122) and O111 (n=94) are represented with the highest frequency

within this cluster (77.4%). Eight other serogroups also clustered within Cluster 1 including six strains belonging to the big six serogroup, O103, which were classified as ST-106. Some STs were extremely diverse representing multiple serogroups; ST-104 and ST-106, for instance, contain strains with five and seven serogroups, respectively. The virulence profile of an isolate recovered from Cluster 1 is significantly more likely to possess *stx1* or *stx1/stx2* (Fisher's χ^2 $p < 0.0001$), *eae* (Fisher's χ^2 $p < 0.0001$), and *ehxA* (OR: 3.9, 95% CI: 1.78, 8.33) relative to all other isolates not clustered. Only three *stx2*-positive and 2 *eae*-negative isolates were identified (Table 2.7).

Cluster 2, which represents the largest of the four clusters, includes 419 isolates, with a large proportion of isolates typed as big six serogroup O45 (n=188) and O103 (n=209) (Figure 2.6). The remaining 22 isolates were typed as O123 or O151. All four of these serogroups were classified as ST-119. An additional three STs, all O103, were identified as ST-119, ST-526 and a new ST due to a gene allele variant in *uidA2*. Virulence gene profiles of isolates found within the cluster were significantly more likely to present with *stx1* or *stx1/stx2* (Fisher's χ^2 $p < 0.0001$), *eae* (Fisher's χ^2 $p < 0.0001$), and *ehxA* (OR: 7.0, 95% CI: 3.13, 15.58) relative to all other isolates not clustered (Table 2.7). Only one Cluster 2 isolate was found to possess *stx2*. Three *eae*-negative isolates were also identified.

Compared to Clusters 1 and 2, the remaining two clusters were smaller, though both Clusters 3 and 4 have strains representing one big six serogroup (Figure 2.6). Among Cluster 3 strains (n=57), O121, ST-182 strains predominated (n=48; 84.2%); however, three additional STs were also identified within the cluster with isolates representing serogroups O38, O28ac/O42, and O113, along with two non-typeable (NT) isolates. A new gene allele in *uidA4* was identified in one NT isolate and denoted as NEW-4. Converse to the prior two clusters,

isolates in Cluster 3 were significantly more likely to harbor the *stx2* gene (OR: 0.04, 95% CI: 0.02, 0.11) and *eae* (OR: 2.7, 95% CI: 1.19, 6.16) (Table 2.7). Lastly, Cluster 4 was the smallest (n=37) but was the most diverse with the largest number of STs (n=17) and serogroups (n=14) present in a cluster. Big six serogroup O145 was present with the highest frequency (n=15; 39.5%) within this cluster and represented two distinct STs, ST-78 and ST-80, that were exclusive to serogroup O145. Cluster 4 isolates were significantly more likely to harbor *stx2* (OR: 14.5, 95% CI: 7.27, 28.83), while being *eaeA*-(OR: 0.6, 95% CI: 0.26, 1.20) and *ehxA*-(OR: 0.8, 95% CI: 0.31, 2.22) negative (Table 2.7).

DISCUSSION

Non-O157 STEC is an important foodborne pathogen that has been steadily increasing in incidence over the past 18 years since identification as a nationally notifiable condition (8, 10, 45). Young children and elderly have been identified nationally as being more susceptible to STEC illness (45, 46). Conversely, in Michigan age 11-29 years was consistently the highest frequency age group. Socioeconomic factors may play a role in the age differences if care is only sought during severe clinical outcomes. Differences in census tract socioeconomic status were associated with STEC infections, however, access to health care may not be the only reason that the socioeconomic status would differ among age groups (47, 48). While a higher socioeconomic status has been associated with STEC and HUS, the ability for these cases to travel or frequently eat out may be the underlying behaviors that result in STEC illness (48, 49). Similarly, Michigan is largely an agricultural state employing almost 25% of the state's workforce along with over 12,000 cattle farms which may be asymptotically colonized by STEC and present as occupational risks (50–52). Lastly, STEC illness and outbreaks are frequently associated with the improper handling of food or the consumption of contaminated foods. The CDC National Health and Nutrition Examination Survey identified that 44.9% of young adults age 20-39 are more likely to eat out or consume fast food, higher than 37.7% and 24.1% of 40-59 and >60 years of age respectively (53). Thus, young adults may be more likely to have foodborne associated STEC illness.

Across age groups, there was also a difference observed in the distribution of serogroups. Older adults and the elderly had higher percentages of other serogroups outside of the big six serogroups identified. International travel may be associated with this increase since other countries have reported higher frequencies of serogroups not commonly found in the US. In

Europe, O146 is included in the five most common non-O157 serogroups, and serogroup O91 is the fourth most common STEC serogroup identified in Germany (12, 54). Future studies will need to examine the travel status of cases to identify if serogroups outside of the big six are associated with travel in the US.

An examination of the non-O157 STEC isolates in Michigan from 2001-2018 identified a continued increasing trend of total non-O157 STEC reports. These trends support the need to further understand the genetic composition and ability of these strains to cause disease. An enhanced diversity of the serogroups was seen from 2001 to 2018 with six and 29 serogroups identified, respectively, which could be partly due to the transitions in surveillance that occurred over this time period. Isolates from the earlier years were recovered as part of a sentinel surveillance primarily using culture-based methods and isolates outside of the sentinel sites may not have been as effective at classifying each serogroup. A prior study identified that the use of enzyme immunoassays increased the number of non-O157 STEC identified from 1.6% on SMAC agar to 48% within Michigan during 2001-2005 (55). In our study, the frequency of big six serogroups varied between 33.3% to 100%, while O45 was the only serogroup isolated in every year. This finding is notable given that O45 is the least common serogroup reported by FoodNet surveillance and the CDC relative to the other big six serogroups (45, 56). Increased diversity of STEC serogroups may also be due to the transition from traditional plating on SMAC agar to the use of culture independent tests that directly amplify the genes or interact with the toxin. The inability to differentiate non-O157 STEC from non-pathogenic *E.coli* on SMAC agar may have resulted in under reporting of non-O157 incidence as well as specific serogroups. The change to culture independent tests have been shown in the US to increase the number of STEC that are identified within a given year (8). Nonetheless, it is also possible that evolutionary

changes have occurred within the STEC population, or in the Shiga toxin-encoding bacteriophage populations, that impact the emergence of new serogroups.

The phylogenetic analysis of non-O157 STEC did not identify any big six specific clusters; however, since the big six were frequently isolated, these serogroups were predominant in the four clusters that were identified. The ST allele profiles in each cluster were identified in a range of differing serogroups. Cluster 1, comprising serogroups O26 and O111, and Cluster 2 with O45 and O103 strains, were significantly associated with the *stx1* or *stx1stx2*, *eae*, and *ehxA* gene profile. Within Cluster 1, predominated by ST-106, seven serogroups were identified and four of those were not members of the big six; O118, O123, O151, and O177. Similarly, Cluster 2 comprised four serogroups that were classified as ST-119. Notably, two rare serogroups, O123 and O151, and the big six O103 serogroup were found in both Clusters 1 and 2. These shared serogroups across two distinct MLST clades demonstrates that serogroup alone is not enough for classification of strains. At the same time, no clinical outcomes were associated with cluster 1 or 2, however, O45 and O111, which were part of clusters 2 and 1, respectively, were associated with a more severe disease marker, presence of bloody diarrhea. Further, O45 was associated with cases that were hospitalized. An increase in the number of cases within the clusters may be needed to detect and associations of clinical outcomes.

Conversely, Cluster 3 (O121) and Cluster 4 (O145) were significantly more likely to have *stx2*, while only Cluster 3 was significantly associated with the presence of *eae*. These two clusters were both diverse in serogroup composition and contained a large number of isolates with ST-serogroup combinations only isolated once over the 18-year period. The cytotoxicity of *stx2* has been demonstrated to be more virulent than *stx1* and it was shown to have an enhanced

ability to attach to epithelial cells (21, 57). Subtypes *stx2a*, *2c* and *2d* have been identified in a range of serogroups and have been linked to the development of bloody diarrhea and HUS (58, 59). Frequent carriage of *stx2* and *eae* have been reported associated with HUS from O157 and non-O157 STEC (11, 60). In Germany and Austria, cases of pediatric HUS was significantly associated with strains that carried both *stx2* and *eae* (60). Similarly, multivariate analysis in a study in Denmark identified that cases resulting in HUS were associated with *stx2* and *eae*, regardless of the serogroup that is carrying both genes (61). Even though strains within certain phylogenetic clusters were associated with *stx* subtypes, further associations with cluster did not identify any clinical outcomes of interest.

Further, the distribution of gene variants was different across serogroups, however, the distribution across clinical outcomes appeared similar due to small frequencies of the gene variants and missing epidemiological data. Similarly, isolates negative for one of the genes may be due to loss of plasmid during culturing or inability to extract the gene sequences due to sequencing quality for that region. The rarity of some gene variants resulted in sample sizes that were too small to statistically analyze. Future studies are needed to increase the sample size of isolates in order to identify associations between variants and clinical outcomes. However, frequencies of the variants were examined within the big six serogroups and each serogroup group was predominated by a single *eae* and *ehxA* variant. Serogroups O45 and O103 shared similar frequencies for *stx1a* (O45: 98.4%, O103: 97.3%) and *eae-epsilon* (O45: 100%, O103: 94.1%). While there were slight differences in *ehxA* subtype frequencies, *ehxA-F* (O45: 97.4%, O103: 59.6%), this subtype was the most frequently isolated in both serogroups. This shared profile and clustering within MLST clades provides support for evolutionary events that may have occurred and given rise to two similar molecular profiles with distinct serogroups. The

potential O-antigen switching that is seen in the *rfb*-like region in O26 and O111, may also be possible among O103 and O45 isolates (13). Similar evolutionary events such as the evolution of O157 from O55 via exchange of *rfb* genes, may be occurring in a similar pathway with different serogroups (62). Further, the transmissibility of these virulence factors may influence the subtypes that are present within a serogroup. Additional studies will need to be performed to understand the carriage of virulence genes and their subtypes.

The continued surveillance and examination of non-O157 molecular profiles is needed to further the understanding of the diversity within the heterogeneous non-O157 STEC. While non-O157 STEC incidence has surpassed that of O157 in Michigan and nationally, the under-reporting of cases may impact the associations and serogroups that are identified (10, 45). The distribution of demographics within Michigan may vary from national statistics and result in state specific interventions to minimize the number of STEC infections in certain age groups. However, further research is needed to identify if gene variants or molecular profiles are commonly associated with specific clinical outcomes. The continued improvements in STEC identification and the use of WGS will allow future studies to gain a more complete profile of non-O157 STEC present within a population.

APPENDIX

Table 2.1. Clinical outcomes associated with reported gender.

Characteristics	No. of isolates	Male (n: 346) No. (%)	Female (n: 504) No. (%)	OR (95% CI [†])	P value [‡]
Clinical Outcomes					
Case hospitalization					
Yes	149	48(32.2)	101(67.8)	1.0	-
No	357	157(44.0)	200(56.0)	1.6 (1.10, 2.47)	0.014
Body ache					
Yes	74	31(29.5)	74(70.5)	1.0	-
No	220	169(43.4)	220(56.6)	1.8 (1.15, 2.92)	0.009

*Number of isolates may not add up to the total for some variables due to missing data in case reports

[†] 95% confidence interval for the odds ratio (OR)

[‡] p-value for statistical significance calculated using the Likelihood Ratio Chi-Square or Fisher's exact test for variables with n < 5 in at least on cell; ORs were not calculated for variables with <5 per cell.

Table 2.2. Demographic, molecular characteristics, and clinical outcomes associated with age at time of infection relative to age group 11-29.

Characteristic*	0-10 years (n=162) No (%)	11-29 years (n=393) No (%)	30-64 years (n=241) No (%)	≥65 years (n=78) No (%)	$\chi^{2\ddagger}$	$p^\ddagger$
Case Demographics						
Sex						
Male	89/159 (56.0)	164/383 (42.8)	65/230 (28.3)	26/75 (34.7)	25.02	<0.0001
Female	70/159 (44.0)	219/383 (57.2)	165/230 (71.7)	49/75 (65.33)		
Serogroups and Virulence Factors						
Serogroup						
O26	29/162 (17.9)	68/393 (17.3)	22/241 (9.1)	7/78 (9.0)	4.3	0.037
O45	29/162 (17.9)	97/393 (24.7)	51/241 (21.1)	12/78 (15.4)		
O103	38/162 (23.5)	96/393 (24.4)	68/241 (28.2)	18/78 (23.1)		
O111	22/162 (13.6)	36/393 (9.2)	24/241 (10.0)	10/78 (12.8)		
O121	10/162 (6.2)	22/393 (5.6)	12/241 (5.0)	5/78 (6.4)		
O145	3/162 (1.8)	4/393 (1.0)	6/241 (2.5)	3/78 (3.8)		
other	31/162 (19.1)	70/393 (17.8)	58/241 (24.1)	23/78 (29.5)		
Shiga toxin						
stx1	138/162 (85.2)	336/393 (85.5)	188/241 (78.0)	48/78 (61.5)	32.6	<0.0001
stx1/stx2	11/162 (6.8)	25/393 (6.4)	20/241 (8.3)	8/78 (10.3)		
stx2	13/162 (8.0)	32/393 (8.1)	33/241 (13.7)	22/78 (28.2)		
eaeA						
Yes	160/162 (98.8)	379/393 (96.4)	213/241 (88.4)	63/78 (80.8)	39.2	<0.0001
No	2/162 (1.2)	14/393 (3.6)	28/241 (11.6)	15/78 (19.2)		
ehxA						
Yes	154/162 (95.1)	373/393 (94.9)	227/241 (94.2)	70/78 (89.7)	2.1	0.15
No	8/162 (4.9)	20/393 (5.1)	14/241 (5.8)	8/70 (10.3)		

Table 2.2 (cont'd)

Clinical Outcomes						
Abdominal pain/cramps						
Yes	64/100 (64.0)	206/230 (89.6)	104/133 (78.2)	30/40 (75.0)	30.38	<0.0001
No	36/100 (36.0)	24/230 (10.4)	29/133 (21.8)	10/40 (25.0)		
Body ache						
Yes	9/100 (9.0)	59/230 (25.6)	30/133 (22.6)	8/40 (20.0)	11.87	0.0079
No	91/100 (91.0)	171/230 (74.4)	103/133 (77.4)	32/40 (80.0)		
Case hospitalization						
Yes	14/102 (8.6)	69/234 (29.5)	47/138 (34.1)	22/41 (53.7)	25.08	<0.0001
No	88/102 (86.3)	165/234 (70.5)	91/138 (65.9)	19/41 (46.3)		

*Total isolates for each variable examined may not add up to the total per column due to missing data in case reports

‡ p-value for statistical significance calculated using Mantel-Haenszel Chi-Square (df=1) for the association between each characteristic and O-type

Table 2.3. Demographic, molecular characteristics and clinical outcomes associated with big six non-O157 STEC infections relative to other non-O157 serogroups

Characteristics	No. of isolates	Non-O157 “big six” No. (%)	OR (95% CI [†])	P value [*]
Demographic				
Age				
0-10	162	131(80.7)	0.9 (0.57, 1.46)	0.71
11-29	393	323(82.19)	1.0	-
30-64	241	183(75.9)	0.68 (0.46, 1.01)	0.057
≥65	78	55(70.5)	0.52 (0.30, 0.90)	0.02
Sex				
Male	346	288(83.2)	1.5 (1.07, 2.15)	0.018
Female	504	386(76.6)	1.0	-
Virulence Genes				
Shiga toxin				
stx1 and stx1/stx2	789	637(80.7)	2.9 (1.90, 4.46)	<0.0001
stx2	105	62(59.1)	1.0	-
<i>eaeA</i> (intimin)				
Yes	828	694(83.8)	-	<0.0001
No	66	5(7.6)		
<i>ehxA</i> (enterohemolysin)				
Yes	844	676(80.1)	4.7 (2.64, 8.45)	<0.0001
No	50	23(46.0)	1.0	-
Clinical Outcomes				
Bloody diarrhea				
Yes	293	256(87.4)	1.9 (1.19, 3.08)	0.006
No	212	166(78.3)	1.0	-

*Number of isolates may not add up to the total for some variables due to missing data in case reports

[†] 95% confidence interval for the odds ratio (OR)

Table 2.3 (cont'd)

‡ p-value for statistical significance calculated using the Likelihood Ratio Chi-Square or Fisher's exact test for variables with $n < 5$ in at least one cell; ORs were not calculated for variables with <5 per cell.

Table 2.4. Association between big six non-O157 STEC serogroups and hospitalization.

Characteristics	No. of isolates	Case Hospitalization No. (%)	OR (95% CI [†])	P value [‡]
Serogroups				
O26				
Yes	84	17(20.2)	0.5 (0.31, 0.98)	0.038
No	432	136(31.5)	1.0	-
O45				
Yes	153	52(39.1)	1.8 (1.18, 2.72)	0.006
No	383	101(26.4)	1.0	-
O103				
Yes	125	24(19.2)	0.5 (0.29, 0.79)	0.003
No	391	129(33.0)	1.0	-
O111				
Yes	52	19(36.5)	1.4 (0.78, 2.58)	0.25
No	464	134(28.9)	1.0	-
O121				
Yes	28	11(39.3)	1.6 (0.72, 3.45)	0.25
No	488	142(29.1)	1.0	-
O145				
Yes	10	5(50.0)	-	0.17
No	506	148(29.3)		

*Number of isolates may not add up to the total for some variables due to missing data in case reports

† 95% confidence interval for the odds ratio (OR)

‡ p-value for statistical significance calculated using the Likelihood Ratio Chi-Square or Fisher's exact test for variables with $n < 5$ in at least on cell; ORs were not calculated for variables with <5 per cell.

Table 2.5. Association between big six non-O157 STEC serogroups and bloody diarrhea.

Characteristics	No. of isolates	Bloody Diarrhea No. (%)	OR (95% CI [†])	P value [‡]
Serogroups				
O26				
Yes	81	44(54.3)	0.8 (0.52, 1.35)	0.46
No	424	249(58.7)	1.0	-
O45				
Yes	133	87(65.4)	1.5 (1.01, 2.30)	0.04
No	372	206(55.4)	1.0	-
O103				
Yes	120	62(51.7)	0.7 (0.47, 1.08)	0.11
No	385	231(60.0)	1.0	-
O111				
Yes	53	38(71.7)	2.0 (1.05, 3.66)	0.03
No	452	255(56.4)	1.0	-
O121				
Yes	26	17(65.4)	1.4 (0.61, 3.18)	0.43
No	479	276(57.6)	1.0	-
O145				
Yes	9	8(88.9)	-	0.058
No	496	285(57.5)		

*Number of isolates may not add up to the total for some variables due to missing data in case reports

[†] 95% confidence interval for the odds ratio (OR)

[‡] p-value for statistical significance calculated using the Likelihood Ratio Chi-Square or Fisher's exact test for variables with n < 5 in at least one cell; ORs were not calculated for variables with <5 per cell.

Table 2.6. Virulence gene profiles found in 894 non-O157 STEC isolates from patients with infections.

Virulence Profile	No.	(%)
<i>stx1, ehxA, eaeA</i>	676	(75.6)
<i>stx1, eaeA</i>	26	(2.9)
<i>stx1, ehxA</i>	10	(1.1)
<i>stx1</i>	10	(1.1)
<i>stx2, ehxA, eaeA</i>	73	(8.2)
<i>stx2, eaeA</i>	3	(0.3)
<i>stx2, ehxA</i>	21	(2.4)
<i>stx2</i>	8	(0.9)
<i>stx12, ehxA, eaeA</i>	47	(5.3)
<i>stx12, eaeA</i>	3	(0.3)
<i>stx12, ehxA</i>	17	(1.9)

Table 2.7. Demographic and molecular characteristics associated with MLST clusters compared to all other non-O157 isolates.

		other	Cluster 1			Cluster 2			Cluster 3			Cluster 4		
Characteristics	No. of isolates	No.	No.(%)	OR (95% CI [†])	<i>p</i> -value [‡]	No.(%)	OR (95% CI [†])	<i>p</i> -value [‡]	No.(%)	OR (95% CI [†])	<i>p</i> -value [‡]	No.(%)	OR (95% CI [†])	<i>p</i> -value [‡]
Demographic														
Age														
0-10	162	15	65(40.1)	1.1 (0.58, 2.23)	0.72	68(42.0)	0.8 (0.40, 1.51)	0.45	10(6.2)	1.1 (0.41, 2.84)	0.88	4(2.5)	-	1.0
11-29	393	34	130(33.1)	1.0	-	199(50.6)	1.0	-	21(5.3)	1.0	-	9(2.3)	1.0	-
30-64	241	27	57(23.7)	0.6 (0.31, 0.99)	0.05	120(49.8)	0.8 (0.44, 1.32)	0.33	19(7.9)	1.1 (0.51, 2.54)	0.75	18(7.5)	2.5 (0.98, 6.49)	0.05
≥65	78	15	19(24.4)	0.3 (0.15, 0.72)	0.004	31(39.7)	0.3 (0.17, 0.72)	0.003	7(9.0)	0.7 (0.26, 2.16)	0.60	6(7.7)	1.5 (0.46, 5.01)	0.50
Virulence Genes														
Shiga toxin														
stx1 or stx1/stx2	789	74	276(35.0)	-	<0.0001	418(53.0)	-	<0.0001	6(0.8)	0.04 (0.02, 0.11)	<0.0001	15(1.9)	0.2 (0.11, 0.52)	0.0002
stx2	105	27	3(2.9)			1(1.0)			51(48.6)			23(21.9)		
<i>eaeA</i> (intimin)														
Yes	828	67	277(33.5)	-	<0.0001	416(50.2)	-	<0.0001	48(5.8)	2.9 (1.30, 6.67)	0.007	20(2.4)	0.6 (0.26, 1.20)	0.14
No	66	34	2(3.0)			3(4.6)			9(13.6)			18(27.3)		
<i>ehxA</i> (enterohemolysin)														
Yes	844	85	266(31.5)	3.9 (1.78, 8.33)	0.0002	408(48.3)	7.0 (3.13, 15.58)	<0.0001	54(6.4)	-	0.07	31(3.7)	0.8 (0.31, 2.22)	0.72
No	50	16	13(26.0)			11(22.0)			3(6.0)			7(14.0)		

Table 2.7 (cont'd)

*Number of isolates may not add up to the total for some variables due to missing data in case reports

† 95% confidence interval for the odds ratio (OR)

‡ p-value for statistical significance calculated using the Likelihood Ratio Chi-Square or Fisher's exact test for variables with $n < 5$ in at least one cell; ORs were not calculated for variables with <5 per cell.

Figure 2.1. Total number of non-O157 STEC isolates that were recovered for WGS (2001-2018) compared to the total number of non-O157 STEC cases reported by MDHHS (2001-2012, 2015-2018).

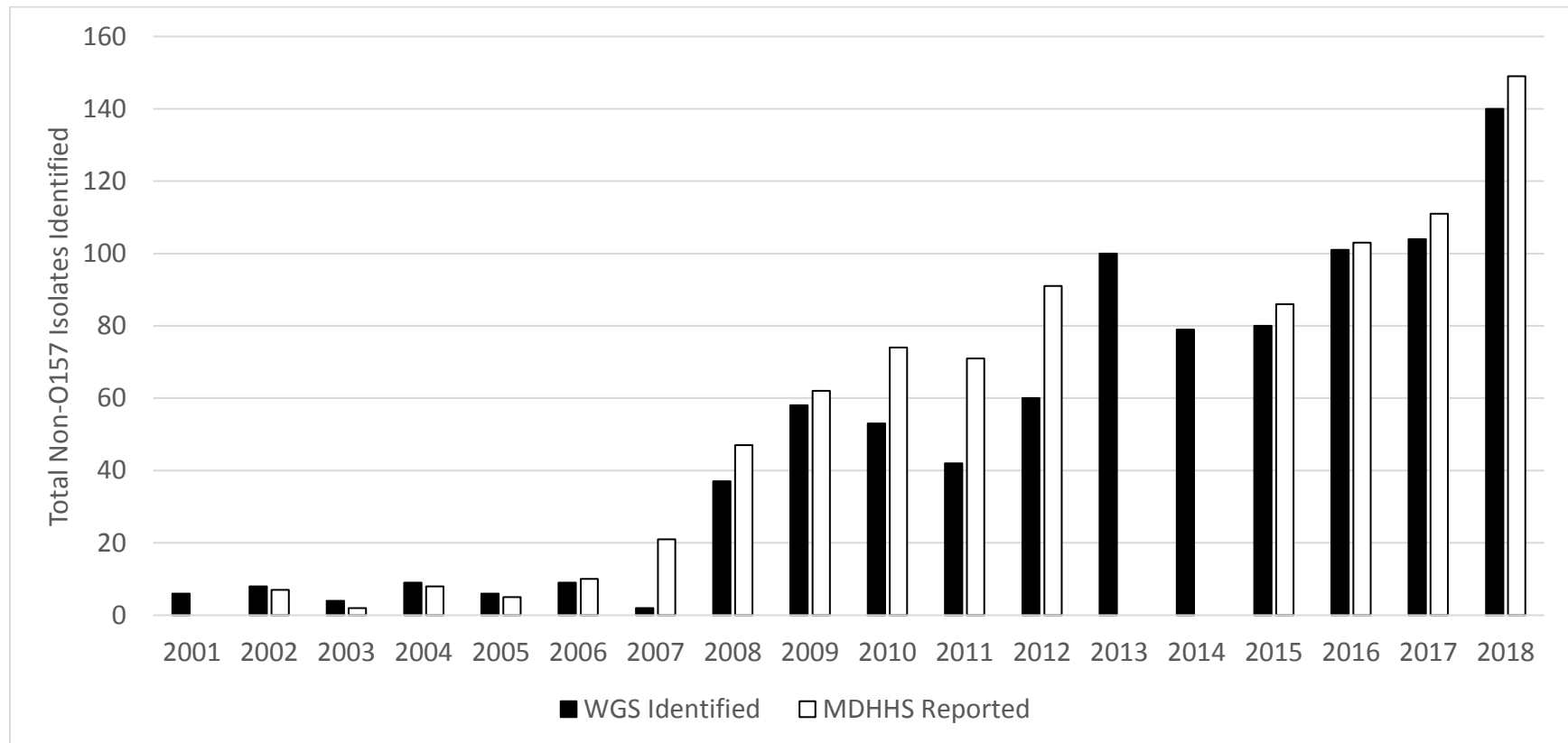


Figure 2.2. Frequency of age groups that reported a non-O157 STEC infection, 2001-2018.

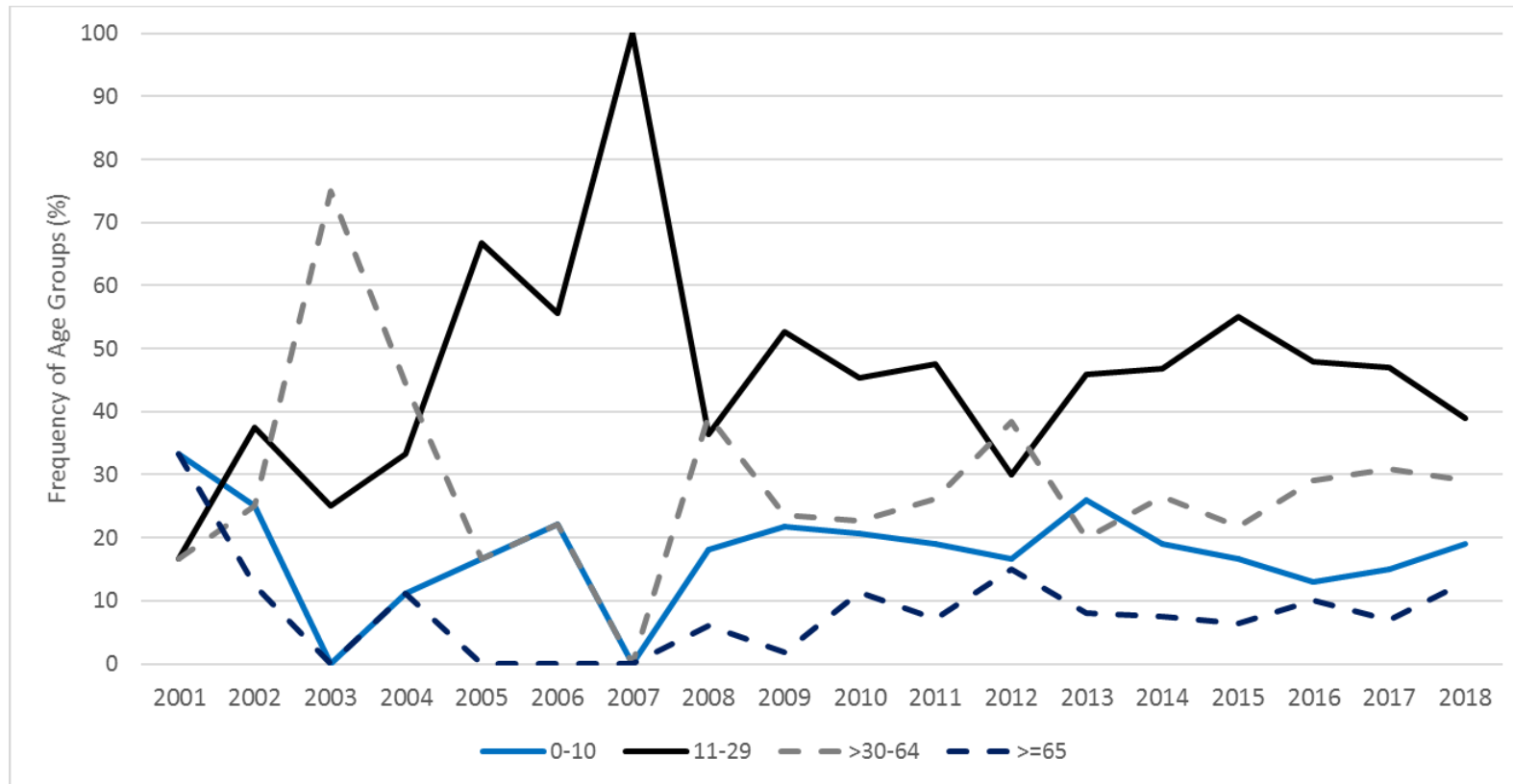


Figure 2.3. Prevalence of non-O157 big-six STEC infections in Michigan, 2001-2018.

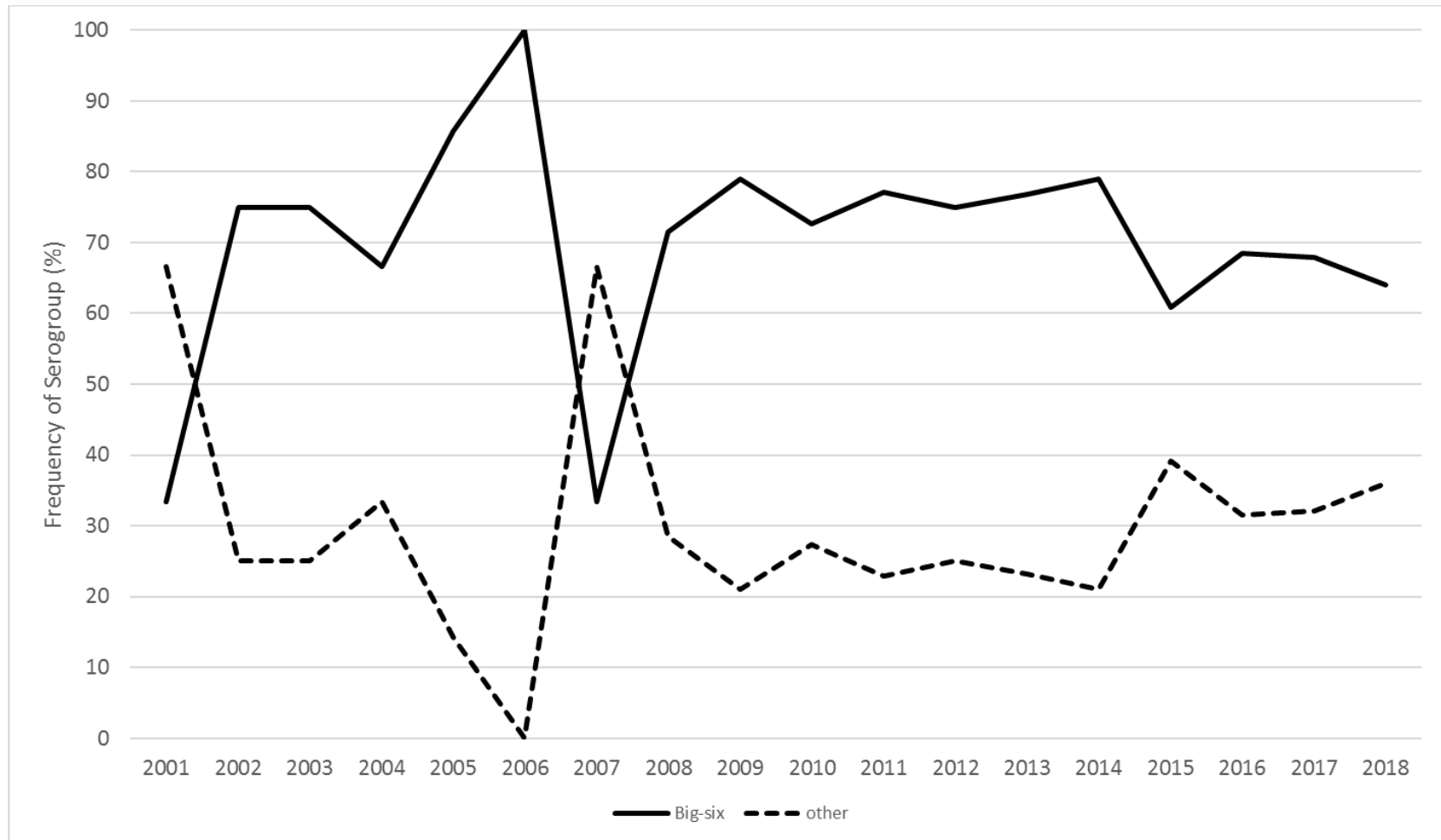


Figure 2.4. Distribution and changes in the non-O157 serogroups reported in Michigan, 2001-2018.

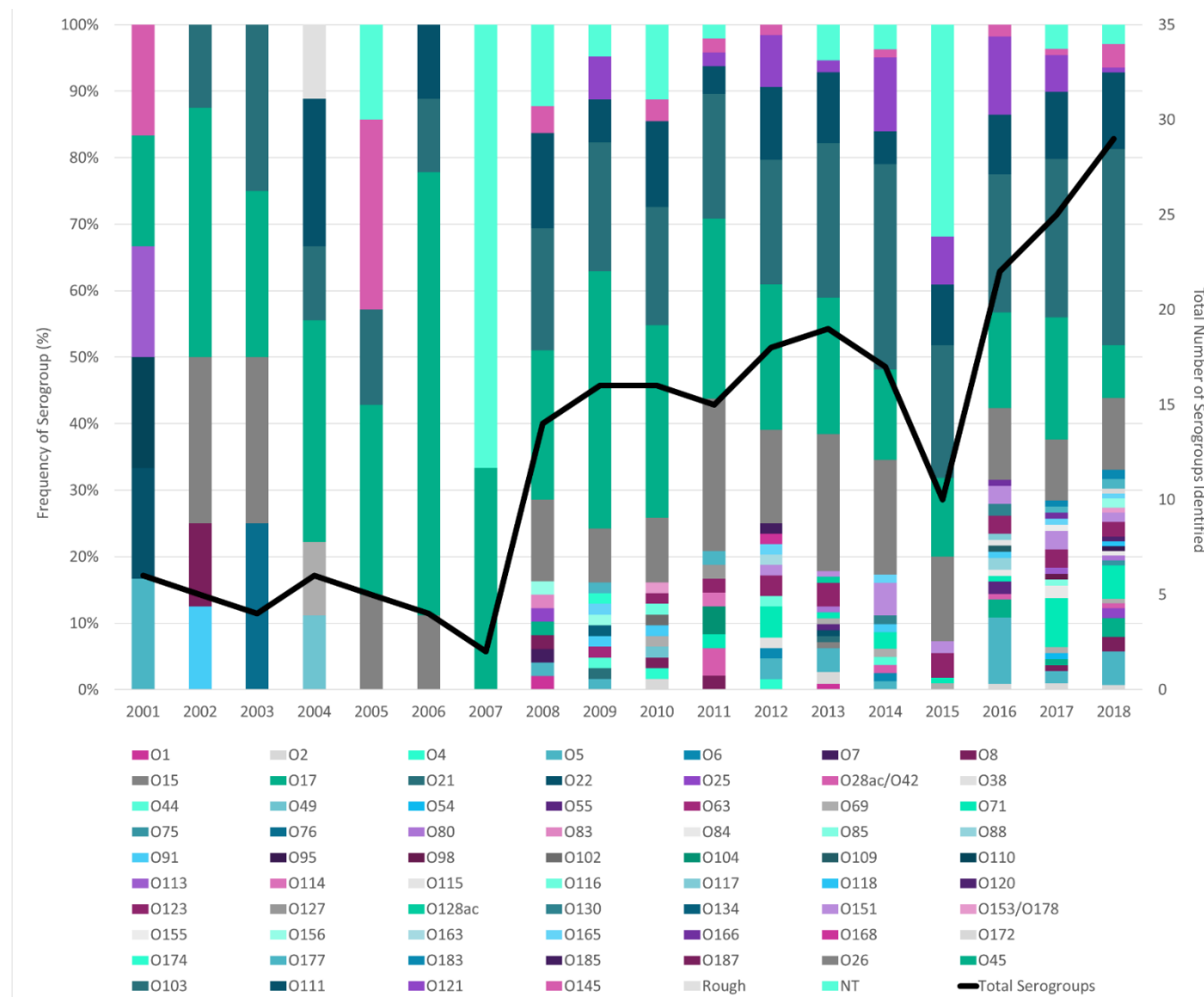


Figure 2.5. Distribution and gene frequency across non-O157 big-six and other STEC serogroups.

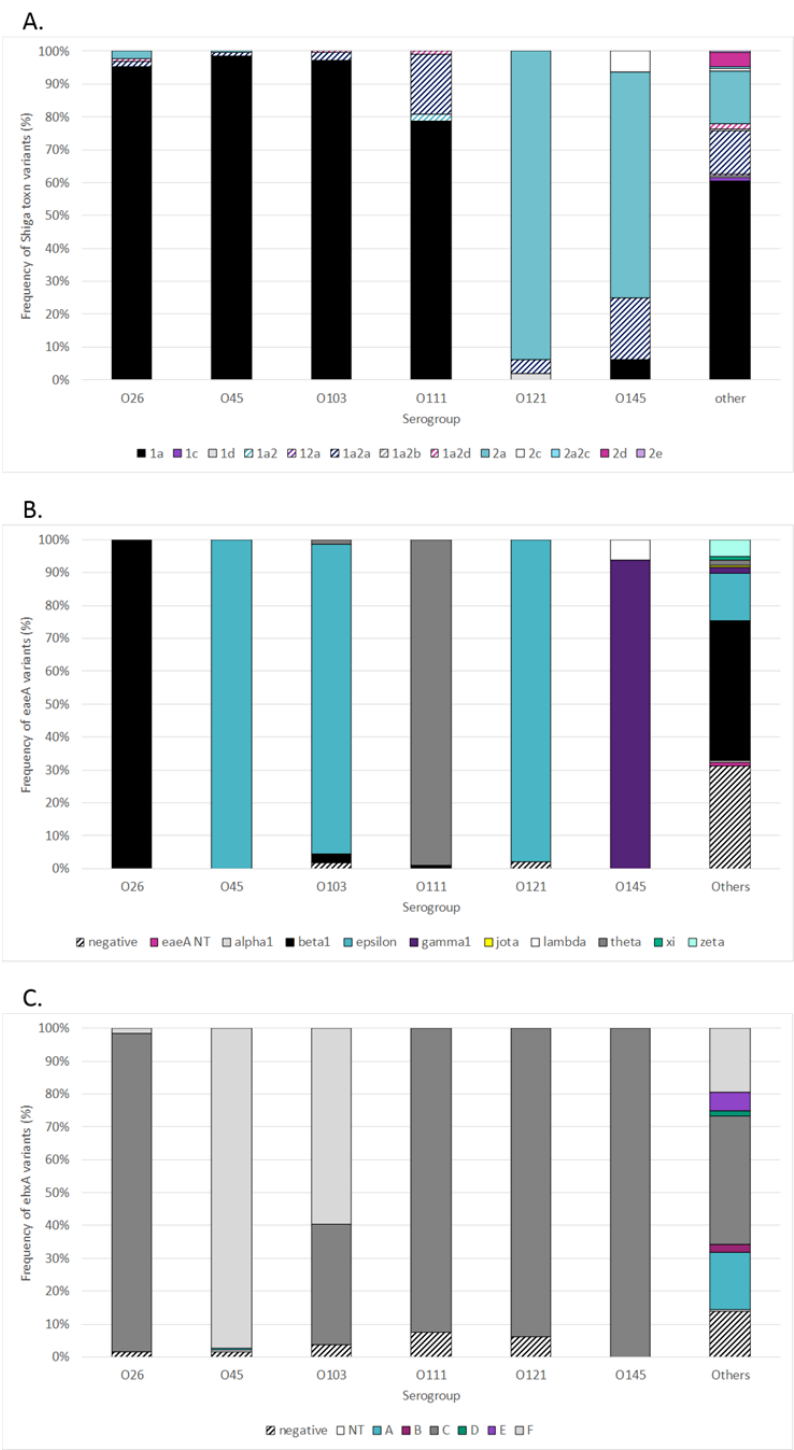
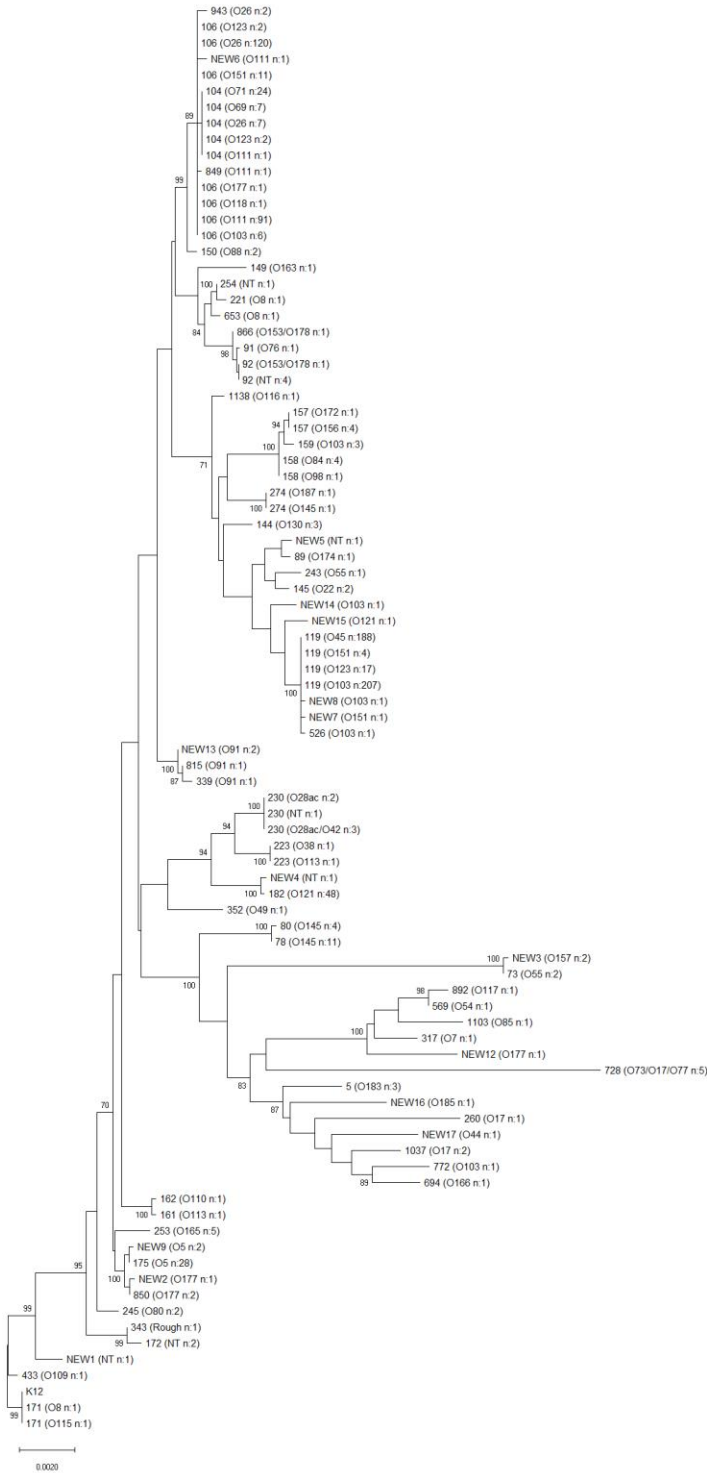


Figure 2.6. MLST-based phylogeny of 894 non-O157 STEC isolates examined using the neighbor-joining algorithm with 1000 bootstrap replication.



REFERENCES

REFERENCES

1. Scallan E, Hoekstra RM, Angulo FJ, Tauxe R V., Widdowson M-A, Roy SL, Jones JL, Griffin PM. 2011. Foodborne Illness Acquired in the United States—Major Pathogens. *Emerg Infect Dis* 17:7–15. doi: 10.3201/eid1701.P11101.
2. Nguyen Y, Sperandio V. 2012. Enterohemorrhagic *E. coli* (EHEC) pathogenesis. *Front Cell Infect Microbiol* 2:90. doi: 10.3389/fcimb.2012.00090.
3. Mead PS, Griffin PM. 1998. *Escherichia coli* O157:H7. *Lancet* 352:1207–1212. doi: 10.1016/S0140-6736(98)01267-7.
4. Besser RE, Griffin PM, Slutsker L. 1999. *Escherichia coli* O157:H7 Gastroenteritis and the Hemolytic Uremic Syndrome: An Emerging Infectious Disease. *Annu Rev Med* 50:355–367. doi: 10.1146/annurev.med.50.1.355.
5. Karmali MA, Petric M, Lim C, McKeough PC, Arbus GS, Lior H. 1985. The association between idiopathic hemolytic uremic syndrome and infection by verotoxin-producing *escherichia coli*. *J Infect Dis* 151:775–782. doi: 10.1093/infdis/151.5.775.
6. O’Brien AD, LaVeck GD. 1983. Purification and characterization of a *Shigella dysenteriae* 1-like toxin produced by *Escherichia coli*. *Infect Immun* 40:675–683.
7. Rangel JM, Sparling PH, Crowe C, Griffin PM, Swerdlow DL. 2005. Epidemiology of *Escherichia coli* O157:H7 outbreaks, United States, 1982-2002. *Emerg Infect Dis* 11:603–609. doi: 10.3201/eid1104.040739.
8. Marder EP, Griffin PM, Cieslak PR, Dunn J, Hurd S, Jervis R, Lathrop S, Muse A, Ryan P, Smith K, Tobin-D’Angelo M, Vugia DJ, Holt KG, Wolpert BJ, Tauxe R, Geissler AL. 2018. Preliminary incidence and trends of infections with pathogens transmitted commonly through food - foodborne diseases active surveillance network, 10 U.S. sites, 2006-2017. *Morb Mortal Wkly Rep* 67:324–328. doi: 10.15585/mmwr.mm6711a3.
9. Carroll KJ, Harvey-Vince L, Jenkins C, Mohan K, Balasegaram S. 2019. The epidemiology of Shiga toxin-producing *Escherichia coli* infections in the South East of England: November 2013–March 2017 and significance for clinical and public health. *J Med Microbiol* 68:930–939. doi: 10.1099/jmm.0.000970.
10. Tseng M, Sha Q, Rudrik JT, Collins J, Henderson T, Funk JA, Manning SD. 2016. Increasing incidence of non-O157 Shiga toxin-producing *Escherichia coli* (STEC) in Michigan and association with clinical illness. *Epidemiol Infect* 144:1394–1405. doi: 10.1017/S0950268815002836.
11. Brooks JT, Sowers EG, Wells JG, Greene KD, Griffin PM, Hoekstra RM, Strockbine NA. 2005. Non-O157 Shiga Toxin–Producing *Escherichia coli* Infections in the United States, 1983–2002 . *J Infect Dis* 192:1422–1429. doi: 10.1086/466536.

12. Werber D, Beutin L, Pichner R, Stark K, Fruth A. 2008. Shiga toxin-producing *Escherichia coli* serogroups in food and patients, Germany. *Emerg Infect Dis* 14:1803–1806. doi: 10.3201/eid1411.080361.
13. Eichhorn I, Heidemanns K, Semmler T, Kinnemann B, Mellmann A, Harmsen D, Anjum MF, Schmidt H, Fruth A, Valentin-Weigand P, Heesemann J, Suerbaum S, Karch H, Wieler LH. 2015. Highly virulent non-O157 enterohemorrhagic *Escherichia coli* (EHEC) serotypes reflect similar phylogenetic lineages, providing new insights into the evolution of EHEC. *Appl Environ Microbiol* 81:7041–7047. doi: 10.1128/AEM.01921-15.
14. Hedican E, Medus C, Besser J, Juni B, Koziol B, Taylor C, Smith K. 2009. Characteristics of O157 versus non-O157 Shiga toxin-producing *Escherichia coli* infections in Minnesota, 2000-2006. *Clin Infect Dis* 49:358–64. doi: 10.1086/600302.
15. O'Brien AD, Laveck GD, Thompson MR, Formal SB. 1982. Production of shigella dysenteriae type j-like cytotoxin by *Escherichia coli*. *J Infect Dis* 146:763–769. doi: 10.1093/infdis/146.6.763.
16. Gould LH, Bopp C, Strockbine N, Atkinson R, Baselski V, Body B, Carey R, Crandall C, Hurd S, Kaplan R, Neill M, Shea S, Somsel P, Tobin-D'Angelo M, Griffin PM, Gerner-Smidt P. 2012. Update: Recommendations for Diagnosis of Shiga Toxin-Producing *Escherichia coli* Infections by Clinical Laboratories. *Clin Microbiol News* 34:75–83. doi: 10.1016/j.clinmicnews.2012.04.004.
17. CDC. 2017. Foodborne Diseases Active Surveillance Network (FoodNet): FoodNet 2015 Surveillance Report (Final Data). Centers for Disease Control and Prevention. Atlanta, GA.
18. Tack DM, Marder EP, Griffin PM, Cieslak PR, Dunn J, Hurd S, Jervis R, Lathrop S, Muse A, Ryan P, Smith K, Tobin-D'Angelo M, Vugia DJ, Holt KG, Wolpert BJ, Tauxe R, Geissler AL. 2019. Preliminary incidence and trends of infections with pathogens transmitted commonly through food - foodborne diseases active surveillance network, 10 U.S. sites, 2006-2017. *Morb Mortal Wkly Rep* 68:369–373. doi: 10.15585/mmwr.mm6816a2.
19. Slanec T, Fruth A, Creuzburg K, Schmidt H. 2009. Molecular analysis of virulence profiles and Shiga toxin genes in food-borne Shiga toxin-producing *Escherichia coli*. *Appl Environ Microbiol* 75:6187–6197. doi: 10.1128/AEM.00874-09.
20. Bielaszewska M, Mellmann A, Zhang W, Köck R, Fruth A, Bauwens A, Peters G, Karch H. 2011. Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: A microbiological study. *Lancet Infect Dis* 11:671–676. doi: 10.1016/S1473-3099(11)70165-7.
21. Melton-Celsa AR. 2014. Shiga Toxin (Stx) Classification, Structure, and Function. *Microbiol Spectr* 2:EHEC-0024-2013. doi: 10.1128/microbiolspec.ehec-0024-2013.

22. Melton-Celsa AR, Darnell SC, O'Brien AD. 1996. Activation of Shiga-like toxins by mouse and human intestinal mucus correlates with virulence of enterohemorrhagic *Escherichia coli* O91:H21 isolates in orally infected, streptomycin-treated mice. *Infect Immun* 64:1569–76.
23. Bielaszewska M, Friedrich AW, Aldick T, Schurk-Bulgrin R, Karch H. 2006. Shiga Toxin Activatable by Intestinal Mucus in *Escherichia coli* Isolated from Humans: Predictor for a Severe Clinical Outcome. *Clin Infect Dis* 43:1160–1167. doi: 10.1086/508195.
24. Friedrich AW, Bielaszewska M, Zhang W, Pulz M, Kuczius T, Ammon A, Karch H. 2002. *Escherichia coli* Harboring Shiga Toxin 2 Gene Variants: Frequency and Association with Clinical Symptoms . *J Infect Dis* 185:74–84. doi: 10.1086/338115.
25. Blanco JE, Blanco M, Alonso MP, Mora A, Dahbi G, Coira MA, Blanco J. 2004. Serotypes, Virulence Genes, and Intimin Types of Shiga Toxin (Verotoxin)-Producing *Escherichia coli* Isolates from Human Patients: Prevalence in Lugo, Spain, from 1992 through 1999. *J Clin Microbiol* 42:311–319. doi: 10.1128/JCM.42.1.311-319.2004.
26. Lorenz SC, Son I, Maounounen-Laasri A, Lin A, Fischer M, Kase JA. 2013. Prevalence of hemolysin genes and comparison of ehxA subtype patterns in Shiga toxin-producing *Escherichia coli* (STEC) and Non-STEC strains from clinical, food, and animal sources. *Appl Environ Microbiol* 79:6301–6311. doi: 10.1128/AEM.02200-13.
27. Dong H-J, Lee S, Kim W, An J-U, Kim J, Kim D, Cho S. 2017. Prevalence, virulence potential, and pulsed-field gel electrophoresis profiling of Shiga toxin-producing *Escherichia coli* strains from cattle. *Gut Pathog* 9:22. doi: 10.1186/s13099-017-0169-x.
28. Friedrich AW, Zhang W, Bielaszewska M, Mellmann A, Kock R, Fruth A, Tschape H, Karch H. 2007. Prevalence, Virulence Profiles, and Clinical Significance of Shiga Toxin-Negative Variants of Enterohemorrhagic *Escherichia coli* O157 Infection in Humans. *Clin Infect Dis* 45:39–45. doi: 10.1086/518573.
29. Ramachandran V, Brett K, Hornitzky MA, Dowton M, Bettelheim KA, Walker MJ, Djordjevic SP. 2003. Distribution of Intimin Subtypes among *Escherichia coli* Isolates from Ruminant and Human Sources. *J Clin Microbiol* 41:5022–5032. doi: 10.1128/JCM.41.11.5022-5032.2003.
30. Blanco M, Padola NL, Krüger A, Sanz ME, Blanco JE, González EA, Dahbi G, Mora A, Bernárdez MI, Etcheverría AI, Arroyo GH, Lucchesi PMA, Parma AE, Blanco J. 2004. Virulence genes and intimin types of Shiga-toxin-producing *Escherichia coli* isolated from cattle and beef products in Argentina. *Int Microbiol* 7:269–276. doi: 10.2436/im.v7i4.9482.
31. Lorenz SC, Monday SR, Hoffmann M, Fischer M, Kase JA. 2016. Plasmids from Shiga toxin-producing *Escherichia coli* strains with rare enterohemolysin gene (ehxA) subtypes reveal pathogenicity potential and display a novel evolutionary path. *Appl Environ Microbiol* 82:6367–6377. doi: 10.1128/AEM.01839-16.

32. Cobbold RN, Rice DH, Szymanski M, Call DR, Hancock DD. 2004. Comparison of shiga-toxigenic *Escherichia coli* prevalences among dairy, feedlot, and cow-calf herds in Washington State. *Appl Environ Microbiol* 70:4375–4378. doi: 10.1128/AEM.70.7.4375-4378.2004.
33. Delannoy S, Beutin L, Burgos Y, Fach P. 2012. Specific detection of enteroaggregative hemorrhagic *Escherichia coli* O104:H4 strains by use of the CRISPR locus as a target for a diagnostic real-time PCR. *J Clin Microbiol* 50:3485–3492. doi: 10.1128/JCM.01656-12.
34. Fratamico PM, DebRoy C, Liu Y, Needleman DS, Baranzoni GM, Feng P. 2016. Advances in molecular serotyping and subtyping of *Escherichia coli*. *Front Microbiol* 7:644. doi: 10.3389/fmicb.2016.00644.
35. Foxman B, Zhang L, Koopman JS, Manning SD, Marrs CF. 2005. Choosing an appropriate bacterial typing technique for epidemiologic studies. *Epidemiol Perspect Innov* 2:10. doi: 10.1186/1742-5573-2-10.
36. Centers for Disease Control and Prevention. Laboratory Standard Operating Procedure for Whole Genome Sequencing on MiSeq.
37. Centers for Disease Control and Prevention. Laboratory Standard Operating Procedure for PulseNet Total DNA Extraction and Quality Control of Purified DNA Extracts.
38. Centers for Disease Control and Prevention. Laboratory Standard Operating Procedure for PulseNet Nextera XT Library Prep and Run Setup for the Illumina MiSeq.
39. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. doi: 10.1093/bioinformatics/btu170.
40. Andrews S. 2010. FASTQC, a quality control tool for the high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
41. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V., Sirotkin A V., Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477. doi: 10.1089/cmb.2012.0021.
42. Seemann T. Abricate. Github. doi: Available online at: <https://github.com/tseemann/abricate>.
43. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410. doi: 10.1016/S0022-2836(05)80360-2.
44. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* 35:1547–1549. doi: 10.1093/molbev/msy096.

45. Gould LH, Mody RK, Ong KL, Clogher P, Cronquist AB, Garman KN, Lathrop S, Medus C, Spina NL, Webb TH, White PL, Wymore K, Gierke RE, Mahon BE, Griffin PM, Patricia M. Griffin for the EIPFWG, Griffin PM. 2013. Increased recognition of non-O157 Shiga toxin-producing *Escherichia coli* infections in the United States during 2000-2010: epidemiologic features and comparison with *E. coli* O157 infections. *Foodborne Pathog Dis* 10:453–60. doi: 10.1089/fpd.2012.1401.
46. Gould LH, Demma L, Jones TF, Hurd S, Vugia DJ, Smith K, Shiferaw B, Segler S, Palmer A, Zansky S, Griffin PM. 2009. Hemolytic Uremic Syndrome and Death in Persons with *Escherichia coli* O157:H7 Infection, Foodborne Diseases Active Surveillance Network Sites, 2000–2006 . *Clin Infect Dis* 49:1480–1485. doi: 10.1086/644621.
47. Simonsen J, Frisch M, Ethelberg S. 2008. Socioeconomic risk factors for bacterial gastrointestinal infections. *Epidemiology* 19:282–290. doi: 10.1097/EDE.0b013e3181633c19.
48. Whitney BM, Mainero C, Humes E, Hurd S, Niccolai L, Hadler JL. 2015. Socioeconomic status and foodborne pathogens in Connecticut, USA, 2000–2011. *Emerg Infect Dis* 21:1617–1624. doi: 10.3201/eid2109.150277.
49. Quinlan JJ. 2013. Foodborne illness incidence rates and food safety risks for populations of low socioeconomic status and minority race/ethnicity: a review of the literature. *Int J Environ Res Public Health*. doi: 10.3390/ijerph10083634.
50. Pruimboom-Brees IM, Morgan TW, Ackermann MR, Nystrom ED, Samuel JE, Cornick NA, Moon HW. 2000. Cattle lack vascular receptors for *Escherichia coli* O157:H7 Shiga toxins. *Proc Natl Acad Sci USA* 97:10325–10329. doi: 10.1073/pnas.190329997.
51. Economic Research Service United States Department of Agriculture. 2006. Changes in the Size and Location of U.S. Dairy Farms.Profits, Costs, and the Changing Structure of Dairy Farming.
52. Friesema IHM, Van De Kasstele J, De Jager CM, Heuvelink AE, Van Pelt W. 2011. Geographical association between livestock density and human Shiga toxin-producing *Escherichia coli* O157 infections. *Epidemiol Infect* 139:1081–1087. doi: 10.1017/S0950268810002050.
53. Fryar CD, Hughes JP, Herrick KA, Ahluwalia N. 2018. Fast Food Consumption Among Adults in the United States, 2013-2016. *NCHS Data Brief* 1–8.
54. European Food Safety Authority and European Centre for Disease Prevention and Control. 2018. The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2017. *EFSA J* 16:5500. doi: 10.2903/j.efsa.2018.5500.
55. Manning SD, Madera RT, Schneider W, Dietrich SE, Khalife W, Brown W, Whittam TS,

- Somsel P, Rudrik JT. 2007. Surveillance for Shiga toxin-producing *Escherichia coli*, Michigan, 2001-2005. *Emerg Infect Dis* 13:318–321. doi: 10.3201/eid1302.060813.
56. Valilis E, Ramsey A, Sidiq S, DuPont HL. 2018. Non-O157 Shiga toxin-producing *Escherichia coli*—A poorly appreciated enteric pathogen: Systematic review. *Int J Infect Dis* 76:82–87. doi: 10.1016/j.ijid.2018.09.002.
 57. Fuller CA, Pellino CA, Flagler MJ, Strasser JE, Weiss AA. 2011. Shiga toxin subtypes display dramatic differences in potency. *Infect Immun* 79:1329–1337. doi: 10.1128/IAI.01182-10.
 58. Kawano K, Ono H, Iwashita O, Kurogi M, Haga T, Maeda K, Goto Y. 2012. Stx genotype and molecular epidemiological analyses of Shiga toxin-producing *Escherichia coli* O157:H7/H- in human and cattle isolates. *Eur J Clin Microbiol Infect Dis* 31:119–127. doi: 10.1007/s10096-011-1283-1.
 59. Orth D, Grif K, Khan AB, Naim A, Dierich MP, Würzner R. 2007. The Shiga toxin genotype rather than the amount of Shiga toxin or the cytotoxicity of Shiga toxin in vitro correlates with the appearance of the hemolytic uremic syndrome. *Diagn Microbiol Infect Dis* 59:235–242. doi: 10.1016/j.diagmicrobio.2007.04.013.
 60. Werber D, Fruth A, Buchholz U, Prager R, Kramer MH, Ammon A, Tschäpe H. 2003. Strong Association between Shiga Toxin-Producing *Escherichia coli* O157 and Virulence Genes stx2 and eae as Possible Explanation for Predominance of Serogroup O157 in Patients with Haemolytic Uraemic Syndrome. *Eur J Clin Microbiol Infect Dis* 22:726–730. doi: 10.1007/s10096-003-1025-0.
 61. Ethelberg S, Olsen KEP, Scheutz F, Jensen C, Schiellerup P, Engberg J, Petersen AM, Olesen B, Gerner-Smidt P, Mølbak K. 2004. Virulence Factors for Hemolytic Uremic Syndrome, Denmark. *Emerg Infect Dis* 10:842–847. doi: 10.3201/eid1005.030576.
 62. Feng P, Lampel KAA, Karch H, Whittam TSS. 1998. Genotypic and Phenotypic Changes in the Emergence of *Escherichia coli* O157:H7. *J Infect Dis* 177:1750–1753. doi: 10.1086/517438.

CHAPTER 3

GENETIC DIVERSITY OF NON-O157 SHIGA TOXIN-PRODUCING *ESCHERICHIA COLI* (STEC) STRAINS RECOVERED FROM PATIENTS IN MICHIGAN AND CONNECTICUT

ABSTRACT

STEC is a leading cause of foodborne infections in the U.S. While O157:H7 strains have been linked to more severe infections, non-O157 serotypes have gradually increased in frequency. Unlike O157 strains, non-O157 STEC are diverse and can be further classified by serotyping and multilocus sequence typing (MLST). Because clustered regularly interspaced repeats (CRISPR) spacers were shown to comprise horizontally acquired DNA elements, the CRISPR region does not actively acquire spacers in STEC, it represents an ideal target to examine the evolutionary history of STEC. Therefore, we sought to examine genetic variation in all clinical non-O157 isolates identified via sentinel surveillance in Michigan between 2001 and 2005 (n=41) and make comparisons to the 114 isolates recovered in Connecticut between 2000 and 2005. Whole genome sequencing was performed and genomic elements were extracted for serotyping (O and H antigen), MLST and CRISPR analysis through the use of bioinformatic scripts, CRISPRFinder and Geneious. Phylogenetic analysis performed using the Neighbor-joining algorithm and the unweighted pair group method with arithmetic mean (UPGMA) with Jaccard similarity coefficient. A total of 29 serogroups were identified among the two states, 8 and 13 were unique to Michigan and Connecticut respectively, while the “big-six” non-O157 serogroups were similar between states (MI: 73.2%, CT: 81.6%). MLST classified 5 of the 29 serogroups to different STs located on distinct branches of the phylogeny; 38 STs were represented in all. In addition, 23 unique CRISPR spacer profiles were found in the subset of 149 strains evaluated. The UPGMA tree defined 9 unique clusters based on CRISPR profiles and exhibited similar clustering of strains as identified in MLST analysis. Two CRISPR spacers, 231 and 317, were isolated from 79.2% (n=118) and 59.1% (n=88) of strains respectively, regardless of serogroup and ST. These data illustrate the high degree of diversity among STEC isolates

linked to clinical infections and demonstrate that CRISPR profiling can be used to further discriminate strains along with MLST. Understanding the diversity of non-O157 strains associated with disease is required to help identify characteristics and lineages associated with disease and identify new ways to combat infections.

INTRODUCTION

Shiga toxin-producing *Escherichia coli* (STEC) is a leading foodborne pathogen in the United States that was estimated to cause 265,000 illnesses and more than 3,600 hospitalizations each year (1). STEC strains are classified based on the presence of Shiga toxin genes encoded on lambdoid bacteriophages that result in the production of Shiga toxin (2). Patients with STEC often present with hemorrhagic colitis or bloody diarrhea and in severe cases, hemolytic uremic syndrome (HUS), kidney failure and death can occur (3).

Historically, STEC O157 strains have predominated in clinical infections, causing the greatest number of outbreaks and the most severe clinical outcomes; however, an increase in the incidence of infections caused by strains belonging to other serogroups, or non-O157 strains has been reported in recent years (4). In the years between 2000 and 2015, FoodNet reported an increase in the incidence of non-O157 infections from 0.12 to 1.65 per 100,000, while more recently, a decrease in O157 incidence from 2.17 to 0.95 per 100,000 has been documented (4, 5). The emergence of other serogroups associated with disease has resulted in the classification of the “big-six”, the predominant non-O157 serogroups comprising O26, O45, O103, O111, O121 and O145 (6). These six serogroups accounted for 83% of non-O157 cases reported to FoodNet from 2000-2010 (4). Although a wide range of other serogroups are responsible for the remaining infections, less is known about the epidemiology and genetic diversity of these strains relative to O157 STEC.

Multiple methods have been used to examine the genetic diversity of STEC. Multilocus sequence typing (MLST) and pulsed-field gel electrophoresis (PFGE) allow for the differentiation of isolates, however, both typing tools are unable to distinguish closely related isolates with high discriminatory power (7, 8). For O157, MLST was found to inadequately

differentiate strains (9) resulting in the development of schemes such as single nucleotide polymorphism (SNP) genotyping that can differentiate strains into distinct clades (10). A SNP genotyping scheme has yet to be developed for non-O157 strains and hence, additional genotyping methods that utilize widely available genomes are needed to evaluate the genetic diversity and evolutionary history of STEC as well as discriminate among epidemiologically linked isolates.

In prior studies, clustered regularly interspaced palindromic repeat (CRISPR) loci have been used to characterize and subtype foodborne pathogens like *Salmonella enterica* and *Campylobacter jejuni* (11, 12); however, they are not routinely used to evaluate the population structure of STEC. CRISPR loci, which are important for adaptive immunity, have been found in up to 50% of bacteria (13). These loci comprise a series of direct repeats separated by diverse spacer sequences, which range in size between 21 and 72 bp and are located next to CRISPR associated sequence (*cas*) genes (14, 15). The high degree of diversity in CRISPR-Cas systems is primarily due to the variation within these spacer sequences (16).

It was previously shown that Cas proteins allow for the integration of invasive or foreign DNA fragments as spacers into the CRISPR region (17, 18). These foreign DNAs were found to be derived from phages, plasmids or other mobile genetic elements (19–21). Transcription of this CRISPR-Cas region results in the assembly of CRISPR RNAs with Cas effector proteins to recognize foreign DNAs (22–24) for cleavage and degradation (25, 26). In *E. coli*, four CRISPR loci have been identified and characterized as CRISPR 1, 2, 3 and 4; these loci are classified as Type I-E or Type I-F depending on the presence of the associated *cas* genes (for a review, see (27)). *E. coli* can also possess CRISPR loci that lack *cas* genes. CRISPR 1 and 2 were defined as

having the *iap/cas* and *ygcE/ygcF* genes, respectively, while CRISPR3-4 show little variation within the spacer region (28, 29).

Although the impact of CRISPRs on immune function has not been established in *E. coli* in natural conditions, it has been suggested these systems may have alternative functions (30). Nonetheless, the degree of variability within the CRISPR loci were suggested to be adequate for subtyping (29, 31). One study of STEC, for example, identified an association between the CRISPR region and the H-antigen (31), which is notable given that serotyping based on the O- and H-antigen is the primary classification scheme for STEC. Consequently, we sought to apply CRISPR subtyping to a diverse set of clinical non-O157 strains isolated from patients in two geographic locations. Because non-O157 STEC strains are more diverse than O157 strains, have increased in frequency, and are difficult to characterize antigenically without prior knowledge of the O-antigen type, more accurate and rapid subtyping platforms based on whole genome sequencing data require validation. Use of these standardized tools will allow researchers to examine diversity across strain populations, better understand evolutionary relationships, track related strains, and identify epidemiological associations with specific genotypes.

MATERIALS AND METHODS

Bacterial strains and epidemiological data

The Michigan Department of Health and Human Services (MDHHS) recovered 41 isolates from patient fecal samples during the years 2001-2006 as part of a sentinel surveillance developed specifically for non-O157 STEC (32). During an overlapping time period between 2000 and 2005, the Connecticut Department of Public Health (CTDPH) recovered 114 isolates from patient fecal samples as part of the Centers for Disease Control and Prevention (CDC) Foodborne Disease Active Surveillance Network (FoodNet). Epidemiological data was obtained from the Michigan Disease Surveillance System as part of MDHHS, and the CTDPH as part of the FoodNet program.

Ethics statement

All protocols used in this study were previously approved by the Institutional Review Boards at Michigan State University (MSU; Lansing, MI, USA; IRB #10-736SM), the MDHHS (842-PHALAB) and the CTDPH.

DNA isolation and whole genome sequencing (WGS)

Isolates were grown aerobically overnight in Luria-Bertani broth at 37°C. DNA was isolated using the Wizard® Genomic DNA purification kit and subsequently prepped for sequencing using the Nextera XT kit (Illumina, San Diego, CA, USA) following manufacturer's instructions. Libraries were sequenced at the MSU Research Technology Support Facility as paired end reads on the Illumina MiSeq platform (2x250 reads). *De novo* genome assembly was performed using Spades, 3.10.1 (33) following trimming and quality checking with

Trimmomatic (34) and FastQC (35), respectively. Multiple k-mers (21, 33, 55, 77, 99, 127) were used and k-mers that passed quality control were cross-assembled to generate the assembly used for downstream analyses. Error correction was performed during the assembly process to minimize the number of mismatches present in the assembled contigs.

Multilocus sequence typing (MLST) and in silico analysis of virulence genes

Bioinformatic scripts were developed to extract virulence genes and MLST alleles from the assembled genomes using a local Basic Local Alignment Search Tool (BLAST) (36). Sequences specific to the query were extracted from the genomes using an E-value = 0.0001 to ensure specificity of the sequences obtained. The EcMLST v1.2 (www.shigatox.net) was used to assign alleles to seven housekeeping genes and classify strains into sequence types (STs). A new sequence type allele was identified during the analysis and is denoted as ST-NEW while the ST designation is pending.

Additional bioinformatic scripts that utilize BLAST were also developed to extract key virulence genes and to determine the molecular serotypes, which is based on the *wzy* and *wzx* (O-antigen lipopolysaccharide) genes and *flhC* (flagellar H-antigen). Virulence gene sequences included the Shiga toxin gene variants, *stx1* and *stx2*, as well as genes encoding intimin (*eae*) and enterohemolysin (*ehxA*). To quantify the abundance of prophages embedded in the genomes, Phaster (37) was used to extract prophage-specific sequences, while the Center for Genomic Epidemiology plasmid database was used to quantify the number of plasmids present (38) with a bioinformatic script that utilized BLAST. Any genes missing from the WGS data were verified using PCR for confirmation. If a strain was positive for a gene based on PCR, then Sanger sequencing was performed at the MSU Research Technology Support Facility for additional

confirmation. All scripts and computing workflows developed and used for data analyses can be accessed on GitHub (<https://github.com/ManningLab>).

CRISPR-Cas sequence analysis

Preliminary spacer sequences were identified using CRISPRFinder (39) and verified manually in Geneious (40) to confirm that each spacer sequence was flanked by the respective CRISPR associated genes. Any CRISPR loci that were missing from the genomes were verified by PCR before concluding that a given strain was negative for one or both loci. If these strains were found to be positive for the CRISPR loci based on PCR, then Sanger sequencing was performed for confirmation at the MSU Research Technology Support Facility. PCR primers for CRISPR1 loci were 5'-TGGTGAAGGAGTTGGCGAAGG-3' and 5'-AAAATGTCCCTCCGCGCTTACG-3', which annealed *iap* and *cas2* and amplified as described in a prior study (41). CRISPR2 loci were amplified using primers 5'-TACACGCCCTTACGAACACA-3' and 5'-CCTGGGAAAAGCTTGAGGAT-3' targeting *ygcE* and *ygcF*, respectively, using the following conditions: 95°C for 3 min followed by 30 cycles of 95°C for 15 s, 69°C for 15 s and 72°C for 30 s, ending with 72°C for 3 min.

Data analysis

MLST alleles were concatenated and aligned using CLUSTALW, and phylogenetic trees were generated using the Neighbor-joining algorithm with 1000 bootstrap replication in MEGA7 (42). CRISPR spacer profiles were converted into a binary code representing the presence and absence of individual spacers. An unweighted pair group method with arithmetic mean (UPGMA) and Jaccard similarity index tree were assembled from the spacer profiles using Past3 (43). Associations between source, serogroup, epidemiological and molecular data were

identified using the Chi-Square (χ^2) and Mantel-Haenszel Chi-Square test, while the Fisher's exact test was used for sample sizes less than five. The t-test was used to identify differences in means for continuous variables such as the number of CRISPR spacers. SAS v9.3 (SAS Institute, Cary, NC) was used for the epidemiological analysis; $p < 0.05$ was considered significant and was reported along with the odds ratio (OR) and 95% confidence interval.

RESULTS

Characteristics of cases infected with non-O157 STEC by state

From 2000 to 2005, there were 146 confirmed non-O157 STEC reported to either the MDHHS (n=32) or the CTDPH (n=114); an additional nine isolates were identified in Michigan in 2006 and were included in the analysis. Among the cases, no significant difference in the gender distribution was observed between the two states, though more females were affected in Michigan (64.9%) than Connecticut (54.2%) despite the greater sample size in the latter (Table 3.1). A significant difference, however, was observed in the age group distribution between states (Mantel-Haenszel χ^2 $p=0.02$). Most Michigan cases were between the ages of 11 and 29 (32.4%) or 30 and 64 (40.5%); only 6 (16.2%) cases were less than 10 years of age. Connecticut had a similar proportion of cases between 11 and 29 years of age (36.4%) but the number of cases under the age of 10 was greater (35.5%) relative to Michigan. Both states had a similar proportion of elderly cases over the age of 65 (Michigan: 10.8%, Connecticut: 9.3%).

Differences in the proportion of the 134 cases reporting specific symptoms (n=134) were also observed between the two states (Table 3.1). Notably, a greater proportion of Michigan cases were hospitalized (n=14; 51.9%) compared to Connecticut cases (n=12; 11.2%) ($p<0.0001$). Among the 26 hospitalized cases, those between 19 and 64 years of age were significantly more likely to be hospitalized (57.7%) compared to those under the age of 18 (23.1%) and over the age of 65 (19.2%) combined (Odds ratio (OR): 3.5; 95% Confidence interval (CI): 1.46, 8.59). Gender was not significantly associated with hospitalization, though more females (n=17; 65.4%) than males (n=9; 34.6%) were hospitalized. Among a subset of 93 cases with data available, no significant difference was observed in the proportion of cases reporting bloody diarrhea between states although slightly more Michigan (70.4%) cases were

affected than Connecticut (51.5%) cases ($p=0.09$). In all, only one Michigan case presented with HUS, which was caused by a *stx1*-positive strain belonging to serotype O103:H2.

Distribution of serogroups and virulence genes and association with clinical outcomes

A total of 29 serogroups were recovered from the two states; 8 (27.6%) and 13 (44.8%) of these 29 serogroups were found solely in Michigan and Connecticut, respectively; while the remaining 8 serogroups were found in both locations (Figure 3.1). Among these eight serogroups, most ($n=5$; 62.5%) belonged to the predominant “big-six” serogroups except for O121, which was not detected in Michigan during this time period. The remaining three serogroups found in both locations were O5, O76 and O91. Notably, differences in the distribution of some serogroups were observed between states. O45 strains, for instance, were significantly more common in Michigan (95% CI: 1.02, 5.28) than Connecticut. Although the frequency of O111 strains was much higher in Connecticut (89.3%) than Michigan (10.7%), this difference was not statistically significant (Fisher’s exact test $p=0.056$). No differences were observed in the distribution of O26, O103, O121 and O145 by state. Similarly, the virulence gene profiles between the two states were similar based on the presence of *stx*, *eaeA* or *ehxA* (Table 3.1). The presence of *stx1* alone or a combination of *stx1* and *stx2* were the most commonly identified profiles in both Michigan (87.8%) and Connecticut (90.3%).

Among all cases in both states, examination of demographic characteristics revealed no significant differences in gender among cases infected with the big-six serogroups or other non-O157 serogroups (Table 3.2). Cases infected with big-six serogroups from Michigan, however, were significantly more likely to be over 30 years of age relative to Connecticut cases (OR: 2.7; 95% CI: 1.08, 6.56). Cases with big-six STEC infections were also significantly more likely to

report abdominal cramps and diarrhea with blood relative to cases infected with other non-O157 strains (Table 3.2). Differences were also observed in symptom reporting by state. Connecticut cases with infections caused by STEC representing the big-six serogroups were significantly more likely to report diarrhea with blood (Fisher's exact test $p=0.03$) compared to the big-six cases in Michigan. By contrast, big-six cases from Michigan were more likely to be hospitalized compared to Connecticut cases with big-six infections (OR: 6.3; 95% CI: 2.18, 18.41). No difference was observed for abdominal cramping by state ($p=0.80$).

Several associations were identified when the big-six serogroups were analyzed individually and compared to the other non-O157 serogroups (Table 3.3). Most notably, the O45 cases were significantly more likely to be hospitalized (OR: 2.6; 95% CI: 1.02, 6.87) when compared to cases infected with all other serogroups. In addition, children younger than 18 years old were significantly more likely to have O111 STEC infections (OR: 4.2; 95% CI: 1.48, 11.95), while cases over 19 years of age were significantly more likely to have O45 infections (OR: 3.3; 95% CI: 1.40, 7.96) compared to all other non-O157 serogroups. When stratified by state, 87.0% of the O111 cases in Connecticut were reported in children under 18 years of age (Mantel-Haenszel χ^2 $p=0.03$). A difference by sex was also observed as males were more likely to have O111 infections than females; however, this difference was not statistically significant ($p=0.06$).

When the virulence gene profiles were examined, the big-six serogroups were more likely to have *stx1* (OR: 6.5; 95% CI: 2.19, 19.18) alone or in combination with *stx2*, compared to all other serogroups. Similarly, the big-six serogroups had a significantly higher frequency of *eaeA* (OR: 58.5; 95% CI: 15.22, 224.49) and *ehxA* (OR: 13.5; 95% CI: 3.89, 49.99) compared to all other non-O157 serogroups. Those isolates representing serogroups O26, O45, O103, and

O111 had either *stx1* (n=105) or *stx1, stx2* (n=8) profiles, while all of the O121 isolates had *stx2a* only (n=3). Isolates belonging to the O145 serogroup, however, varied and contained multiple *stx* gene profiles (Table 3.3). Comparatively, a wider range of nine *stx* variants/profiles were observed among the other non-O157 serogroups, further highlighting the heterogeneity of the non-O157 strain population (Figure 3.2). Although the *eaeA* gene profiles were relatively homogeneous within a serogroup, seven different *eaeA* variants were identified. Many (59.4%) of the non-O157 strains outside of the big-six group were negative for *eaeA*.

Genetic diversity of non-O157 STEC and association with disease

MLST was utilized to examine the genetic diversity of the STEC strains isolated from both states. A total of 38 STs were identified in all; 17 STs were recovered in Michigan (MI) and 27 STs were collected in Connecticut (CT) (Figure 3.3). Six of the STs were shared and found in both locations. The shared STs comprised 75.5% of the cases in the two states, with ST-106 (MI: n=8 (19.5%); CT: n=38 (33.3%)) and ST-119 (MI: n=16 (39.0%); CT: n=41 (36.0%)) predominating. One isolate from Connecticut was classified as a new ST with a unique allele profile and was denoted as ST-NEW in the analysis.

A neighbor-joining phylogenetic tree with bootstrapping (n=1000) grouped the strains into two clusters with greater than 90% bootstrap support. The first cluster, Cluster 1, contains ST-104, 106, 150, 310, 849, and 852, while Cluster 2 contains STs 89, 119, 145, 286, 526, 845, 846, 851, and NEW (Figure 3.3). All strains not grouping within these two clusters were considered as an “other” group for the epidemiological analyses regardless of placement within the tree. Strains within Cluster 1 contained eight different serotypes including O88:H25, O26:H11, O118:H16, O111:H11, O69:H11, O111:H8, O103:H2, and O151:H8, whereas Cluster

2 included serotypes O103:H2, O153:H2, O45:H2, O22:H8, O13:H21, O146:H21, O174:H21, O8:H14, and O174:H8. Only one serotype, O103:H2, was found in both Clusters 1 and 2 as well as a smaller unrelated cluster. Multiple serotypes are represented by genetically unrelated STs and were found across different branches of the tree. O103:H2 strains, for example, represented STs 772, 106, 851, 526, and 119, while O26:H11 strains comprised STs 338, 104, 106, and 844. Notably, strains of the same serogroup belonged to multiple STs and clustered separately on different branches of the phylogenetic tree.

Examination of the Shiga toxin genes associated with clusters did not identify any *stx* combinations to be significantly different between the clusters. Variants *ehxA-F* were significantly more common in Cluster 2 (OR: 31.5; 95% CI: 12.28, 80.83), while *ehxA-C* was more common in Cluster 1 (Fisher's exact test $p < 0.0001$) relative to all other isolates in different clusters. Strains with other *ehxA* variants or that lacked it altogether were not associated with a specific Cluster. Similarly, the *eaeA* variants, beta (OR: 11.6; 95% CI: 4.5, 29.5) and epsilon (Fisher's exact test $p < 0.0001$), were the only two variants found in Cluster 1. The epsilon variant, however, predominated (86.6%) Cluster 2 and was significantly more common relative to all other Clusters (Fisher's exact test $p < 0.0001$). Strains harboring other *eaeA* variants or that lacked *eaeA* were present throughout the phylogenetic tree with exception of *eaeA*-xi, which was only found in Cluster 2 in an O103:H2, ST-119 strain from CT.

No clustering of strains was observed by state; however, age was significantly associated with cluster designation. Specifically, cases with non-O157 STEC belonging to Cluster 1 (n=35; 70.0%) were significantly more likely to be young, or less than 18 years of age, compared to the 27 (44.3%) strains belonging to Cluster 2 (OR: 2.9; 95% CI: 1.34, 6.46) or all other Clusters

(n=16; 48.5%; OR: 2.5; 95% CI: 1.0, 6.17). No associations were identified between Cluster and more severe clinical outcomes like hospitalization and presence of blood in the stool.

CRISPR profiling and phylogenetic analysis

CRISPR1 and CRISPR2 loci were identified in 149 of the 155 strains. Because the CRISPR2 loci were not detected in sequences from six strains, those strains were excluded from the downstream analyses. Two of these strains, TW14929 (O103:H2) and TW10122 (O26:H11), were missing the CRISPR2 loci entirely and both lacked CRISPR spacers and repeats in the region between *ygcE* and *ygcF*. Another strain, TW14904 (O111:H8), had an interrupted CRISPR2 locus with a potential insertion element lacking any spacers or repeats.

In all, the total number of CRISPR spacers in CRISPR1 and CRISPR2 ranged from six to 49 spacers, while the individual CRISPR1 loci ranged from one to 30 spacers and CRISPR2 ranged from zero to 21 spacers. Each strain had an average of 14 spacers and no difference was observed in the average number of spacers by Cluster. Most strains belonging to Cluster 1 (n=36; 69.2%) had between 11-20 spacers, while 53.0% (n=35) of the Cluster 2 strains had 11-20 spacers. In all, only 13 (8.7%) strains had more than 20 spacers. By contrast, the number of spacers was significantly different across serogroups (Mantel-Haenszel χ^2 $p < 0.0001$) with the big-six serogroups having fewer spacers than all other serogroups. The average number of spacers was 12.8 for the 120 strains belonging to the big-six serogroups compared to 20.2 for the 29 strains representing other serogroups (t-test $p = 0.0006$).

A total of 361 unique spacers were identified that grouped into 79 different CRISPR profiles. The presence of spacers 56 (n=62), 231 (n=118) and 317 (n=88) were identified in multiple strains regardless of serogroup or ST (Figure 3.4). The spacers for both CRISPR loci

were concatenated in each strain and compared for presence or absence, and clusters were examined based on a Jaccard similarity of >45%. Overall, the CRISPR profiles of the strains clustered relative to the STs regardless of serogroup (both O and H-type) and geographic location (Figure 3.5).

Since twice as many CRISPR profiles were identified when compared to the 39 STs identified using the MLST scheme, the Simpson's diversity index was calculated for each genotyping method. The discriminatory power of MLST and CRISPR profiling was 0.76 and 0.96, respectively, while the combination of both methods was 0.97. Nonetheless, several discrepancies were observed between the methods. TW15008, a ST-119 serotype O103:H2 strain, belongs to Cluster 2 in the MLST-based neighbor-joining tree but groups together with other Cluster 1 strains in the UPGMA tree based on CRISPR spacer profiling (Figure 3.5). Another O103:H2 strain (TW14919) was classified as ST-106 within Cluster 1 in the MLST phylogeny but had a CRISPR profile that was more similar to other strains within Cluster 1 and not to the other O103:H2 strains belonging to Cluster 2.

CRISPR spacer content indicative of phage and plasmid transfer

Notably, 5.5% (n=20) of all 361 spacers belonged to known or putative phages and plasmids when using BLAST against the NCBI database with a minimum of 3 nucleotide differences. Spacers that matched to CRISPR spacers in other organisms such as *Shigella sonnei*, were not noted. Spacer 356 was of interest because 30 of the 32 nucleotides matched the *E.coli* O157 T7 typing phage, which is a common *E. coli* phage (44). Although spacers 56, 231 and 317 were common to strains regardless of genetic relatedness, these spacers did not match any published phage or plasmid sequences available in the NCBI database. The total horizontal gene

transfer value was assigned to each genome by adding together the number of phages and plasmids present in the genomes. Strains with ≥ 20 spacers had a significantly higher number of horizontally acquired elements (total plasmids and phages ≥ 8) (OR: 4.9; 95% CI: 1.50, 16.36) when compared to strains with a low (<20) spacer content.

DISCUSSION

Although the number of non-O157 STEC infections has been steadily increasing in the U.S. since 2000 (4, 45), little is known about the molecular epidemiology and genetic diversity of these pathogens in different geographic locations. Through this analysis, we have shown that a wide range of strain types are linked to human infection in two states and that strains representing one of the six (“big-six”) most abundant serogroups predominated in each. Variation in epidemiological factors among cases from each state was also observed as well as variation in the molecular characteristics of the STEC populations.

In all, a greater number of cases were detected in Connecticut compared to Michigan over the same time period, which could be due to differences in surveillance activities. Connecticut participated in the FoodNet active surveillance system, while Michigan utilized a sentinel surveillance system established by the MDHHS (32). The age distribution also varied among cases from each state. Most STEC cases reported by FoodNet occur in young children or the elderly (4), which was similar to the age distribution in Connecticut. In Michigan, however, most cases were between 19 and 64 years of age. Such differences could be due to varying environmental factors, behavioral practices, or occupational risks. Indeed, Michigan has a larger number of dairy cattle farms (46) and prior studies have linked high cattle densities to STEC infections caused by specific STEC serogroups (47, 48).

Cases in Michigan were also more likely to be hospitalized compared to cases from Connecticut. Because gastrointestinal infections are underreported and Michigan was not participating in active surveillance, cases with less severe infections may have been less likely to be screened for non-O157 STEC relative to hospitalized cases. Higher hospitalization rates in

Michigan could also be due to a lower threshold for hospital admission or may indicate variation in virulence of the STEC strains recovered from each state.

Each big-six serogroup, except O121, was found in both Michigan and Connecticut as were strains belonging to serogroups O5, O76 and O91, which have been linked to human infections in Europe (49–51). Furthermore, serogroup O91 is among the most frequently isolated serogroups in food and human infections in Europe (51). The close proximity of Connecticut and Michigan to international airports or borders may indicate that some of these infections were travel-associated as associations between infection with O111, O103 and O26 strains have been linked to international travel (4, 52, 53). These data, however, were not available for cases in either state and therefore, future studies are needed to establish relationships between travel and risk of infection with specific STEC strain types.

Phylogenetic analysis of the isolates identified that there was no geographic clustering present and 75.5% of the isolates belonged to STs that were present in both states. The highest frequency ST that were reported among the clinical isolates were ST-106 and ST-119. ST-106 was primarily composed of O26 and O111 serogroups, both belonging to the big six serogroups and similarly within ST-119, serogroups O45 and O103 were predominant. Using a different MLST scheme, another study also identified that O26 and O111 serogroups were clustering together potentially indicating a lateral gene transfer event of the *rfb*-like region similar to what occurred within the O157 lineage (54, 55). Outside of the big-six serogroups, there were a wide range of serogroups that were clustering together along with the big-six serogroups into two main clusters. Examination of these clusters identified that cluster 1 was significantly associated with cases that were younger than 18 years old. This association may be driven by the presence of serogroup O111:H8 which was also significantly associated with age less than 18 years and

composes a large percentage of cluster 1 isolates (42.6%). However, the genetic relatedness of strains with different serogroups may indicate that serogroup alone will not be an indication of disease outcome. The presence of different serogroups in multiple genetically unrelated branches on the tree, such as O103:H2 being characterized as ST-106, 119, 526, 772 and 851, further supports the genetic diversity that is seen within a single serogroup. Similarly, a large number of isolates with different disease outcomes are genotyped as the same serogroup and ST further supporting the need for a more discriminate subtyping method to provide more information about the strain while maintaining epidemiological concordance. With the rise of whole genome sequencing analysis for surveillance, the development of methods that will rapidly identify and provide information on the strain is needed.

Through the addition of CRISPR loci analysis, the discriminatory power when used in combination with MLST was increased from 0.763 to 0.968. CRISPR spacer analysis has been previously shown to help with discrimination of *Salmonella enterica* and *Campylobacter jejuni* outbreak isolates (12, 56). The amplification of two CRISPR regions, the MLST genes and the serogroup genes are less expensive and time consuming than current methods of PFGE and reveal more about the genetic relatedness and variation that is present among the strains in a collection (7). Utilizing both MLST and CRISPR typing, identified an O103:H2 strain that clustered with ST-119 based on the MLST typing scheme, however, when subtyping with CRISPR loci, it was more similar to strains that were clustering with ST-106. A potential evolutionary event may have occurred that would have been missed if only examining the MLST of these strains. The use of CRISPR spacers has been previously used to examine the evolutionary divergence of O55:H7 to O157:H7 (31).

Analysis of the CRISPR spacers that were isolated, identified 5.54% of the spacers to originate from known phages or plasmids, which is similar to what has been reported in other studies that have examined the spacer content (19, 31). While the putative function of the CRISPR loci is to provide adaptive immunity, in laboratory conditions, STEC is not provided with immunity when subjected to plasmids or phages that have corresponding spacers in the CRISPR loci (57–59). However, the number of spacers that are present in the CRISPR loci may be indicative of the strain living in an environment that is subject to high horizontal gene transfer. Strains with a higher number of CRISPR spacers were significantly more likely to encode a higher number of plasmids or phages in their genome. This helps to further support that STEC may have an active CRISPR loci in specific conditions outside of the laboratory or that a recent event caused STEC to turn off the CRISPR loci, leaving it available to uptake more plasmids and phages without the foreign DNA being targeted by the CRISPR system.

In all, this study helps to further understand the genetic composition of non-O157 STEC and the wide number of gene profiles that are present in strains isolated from patients. A limitation of the study is the lack of epidemiological information that was present and the lack of overlap in some of the epidemiological variables that were collected between the two states since they were a part of two different surveillance systems. At the same time, Michigan non-O157 STEC may be under estimated due to only having a sentinel surveillance in place. However, the associations that were identified can be further explored in a future study to determine if these associations hold true. The ability to subtype strains based on the CRISPR loci and identifying that it overlaps with MLST typing, will help to rapidly evaluate strains, identify the genetic relatedness and determine if strains are from the same outbreak.

APPENDIX

Table 3.1. Comparison of demographics and clinical outcomes among non-O157 STEC cases from Michigan and Connecticut between 2001 and 2006.

Characteristic	Total no. Michigan	No (%) Michigan	Total no. Connecticut	No (%) Connecticut	Odds Ratio (95% CI†)	p-value‡
Case Demographics						
Sex	37		107			
Male		13 (35.1)		49 (45.8)	1.5 (0.72, 3.38)	0.26
Female		24 (64.9)		58 (54.2)		
Age in years	37		107			
0-10		6 (16.2)		38 (35.5)	0.5 (0.17, 1.51)	0.22
11-29		12 (32.4)		39 (36.5)	1.0	-
30-64		15 (40.6)		20 (18.7)	2.4 (0.96, 6.18)	0.06
≥ 65		4 (10.8)		10 (9.4)	-	0.73
Clinical Outcomes						
Abdominal pain/cramps	26		62			
No		5 (19.2)		12 (19.4)	1.0 (0.32, 3.22)	1.0
Yes		21 (80.8)		50 (80.7)		
Any bloody diarrhea	27		66			
No		8 (29.6)		32 (48.5)	2.2 (0.86, 5.82)	0.096
Yes		19 (70.4)		34 (51.5)		
Hospitalization	27		107			
No		13 (48.5)		95 (88.8)	8.5 (3.25, 22.37)	<0.0001
Yes		14 (51.9)		12 (11.2)		

Table 3.1 (cont'd)

*Number of isolates may not add up to the total (n=155) for some variables due to missing data in case reports.

† 95% confidence interval for the odds ratio (OR)

‡ p-value for statistical significance calculated using the Likelihood Ratio Chi-Square or Fisher's exact test for variables with $n \leq 5$ in at least one cell; ORs were not calculated for variables with <5 per cell.

Table 3.2. Demographic, molecular profiles and clinical outcomes associated with big-six non-O157 serogroups and all other non-O157 serogroups from cases in Michigan and Connecticut combined.

Characteristic	Total no. non-O157 big-six	No (%) non-O157 big-six	Total no. non- O157 other	No (%) non- O157 other	OR (95% CI) [†]	<i>p</i> -value [‡]
Case Demographics						
State	123		32			
Michigan		30 (24.4)		11 (34.4)	0.6 (0.27, 1.42)	0.26
Connecticut		93 (75.6)		21 (65.6)		
Sex	114		30			
Male		51 (44.7)		11 (36.7)	1.4 (0.61, 3.20)	0.43
Female		63 (55.3)		19 (63.3)		
Age in years	114		30			
0-10		34 (29.8)		10 (33.3)	1.9 (0.51, 6.94)	0.48
11-29		43 (37.7)		8 (26.7)	3.0 (0.79, 11.27)	0.13
30-64		28 (24.6)		7 (23.3)	2.2 (0.56, 8.76)	0.29
≥65		9 (7.9)		5 (16.7)	1.0	-
Virulence Genes						
Shiga toxin	123		32			
<i>stx1</i> and <i>stx1/stx2</i>		116 (94.3)		23 (71.9)	6.5 (2.19, 19.18)	0.0002
<i>stx2 only</i>		7 (5.7)		9 (28.1)		
<i>eaeA</i>	123		32			
No		3 (2.4)		19 (59.4)	58.5 (15.22, 224.49)	<0.0001
Yes		120 (97.6)		13 (40.6)		
<i>ehxA</i>	123		32			
No		4 (3.2)		10 (31.2)	13.5 (3.89, 49.99)	<0.0001
Yes		119 (96.8)		22 (68.8)		

Table 3.2 (cont'd)

Clinical Outcomes						
Abdominal pain/cramps	71		17			
No		10 (14.1)		7 (41.2)	4.3 (1.32, 13.82)	0.01
Yes		61 (85.9)		10 (58.8)		
Any bloody diarrhea	75		18			
No		27 (36.0)		5 (27.8)	4.6 (1.49, 14.37)	0.005
Yes		48 (64.0)		13 (72.2)		
Hospitalization	108		26			
No		86 (79.6)		22 (84.6)	-	0.78
Yes		22 (20.4)		4 (15.4)		

*Total isolates for each variable examined may not add up to the total (n=155) due to missing epidemiological information in case reports

† 95% confidence interval (CI) for the odds ratio (OR) reported

‡ p-value for statistical significance calculated using Chi-Square test or Fisher's exact test for variables with $n \leq 5$ in at least on cell;

ORs were not calculated for variables with <5 per cell.

Table 3.3. Demographic, molecular profiles and clinical outcomes associated with big-six non-O157 serogroups from cases in Michigan and Connecticut relative to infection with other non-O157 serogroups.

Characteristic*	O26 (n=24) No (%)	O45 (n=32) No (%)	O103 (n=29) No (%)	O111 (n=28) No (%)	O121 (n=3) No (%)	O145 (n=7) No (%)	Other (n=32) No (%)	χ^2 [‡]	p [‡]
Case Demographics									
State									
Michigan	6 (25.0)	13 (40.6)	5 (17.2)	3 (10.7)	0 (0.0)	3 (42.9)	11 (34.4)	1.89	0.17
Connecticut	18 (75.0)	19 (59.4)	24 (82.8)	25 (89.3)	3 (100.0)	4 (57.1)	21 (65.6)		
Sex									
Male	6 (27.3)	14 (48.3)	12 (42.9)	15 (60.0)	0 (0.0)	4 (57.1)	11 (36.7)	0.001	0.97
Female	16 (72.7)	15 (51.7)	16 (57.1)	10 (40.0)	3 (100.0)	3 (42.9)	19 (63.3)		
Age in years									
0-10	8 (36.4)	4 (13.8)	6 (21.4)	12 (48.0)	1 (33.3)	3 (42.9)	10 (33.3)	2.31	0.13
11-29	6 (27.3)	12 (41.4)	14 (50.0)	8 (32.0)	0 (0.0)	3 (42.9)	8 (26.7)		
30-64	5 (22.7)	10 (34.5)	8 (28.6)	3 (12.0)	1 (33.3)	1 (14.2)	7 (23.3)		
≥65	3 (13.6)	3 (10.3)	0 (0.0)	2 (8.0)	1 (33.3)	0 (0.0)	5 (16.7)		
Virulence Factors									
Shiga toxin									
stx1	24	32	29	20 (71.4)	0 (0.0)	1 (14.3)	16 (51.6)	0.16	0.69
stx1/stx2	(100.0)	(100.0)	(100.0)	8 (28.6)	0 (0.0)	2 (28.6)	6 (19.4)		
stx2	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	3 (100.0)	4 (57.1)	9 (29.0)		
0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)						
eaeA									
No	2 (8.3)	0 (0.0)	1 (3.4)	0 (0.0)	0 (0.0)	0 (0.0)	19 (59.4)	25.68	<0.0001
Yes	22 (91.7)	32 (100.0)	28 (96.6)	28 (100.0)	3 (100.0)	7 (100.0)	13 (40.6)		

Table 3.3 (cont'd)

<i>ehxA</i>									
No	2 (8.3)	1 (3.1)	1 (3.4)	0 (0.0)	0 (0.0)	0 (0.0)	10 (31.3)	9.86	0.0017
Yes	22 (91.7)	31 (96.9)	28 (96.6)	28 (100.0)	3 (100.0)	7 (100.0)	22 (68.7)		
Clinical Outcomes									
Abdominal pain/cramps									
No	4 (30.8)	2 (10.0)	1 (5.9)	3 (20.0)	0 (0.0)	0 (0.0)	7 (41.2)	1.26	0.26
Yes	9 (69.2)	18 (90.0)	16 (94.1)	12 (80.0)	1 (100.0)	5 (100.0)	10 (58.8)		
Diarrhea with blood									
No	5 (38.5)	6 (28.6)	8 (42.1)	7 (43.8)	0 (0.0)	1 (20.0)	13 (72.2)	3.38	0.07
Yes	8 (61.5)	15 (71.3)	11 (57.9)	9 (56.2)	1 (100.0)	4 (80.0)	5 (27.8)		
Case Hospitalization									
No	16 (80.0)	18 (66.7)	22 (84.6)	23 (92.0)	3 (100.0)	4 (57.1)	22 (84.6)	0.005	0.94
Yes	4 (20.0)	9 (33.3)	4 (15.4)	2 (8.0)	0 (0.0)	3 (42.9)	4 (15.4)		

*Total isolates for each variable examined may not add up to the total per column due to missing data in case reports

‡ p-value for statistical significance calculated using Mantel-Haenszel Chi-Square (df=1) for the association between each characteristic and O-type

Figure 3.1. Prevalence of serogroups detected in Michigan and Connecticut, 2001-2006.

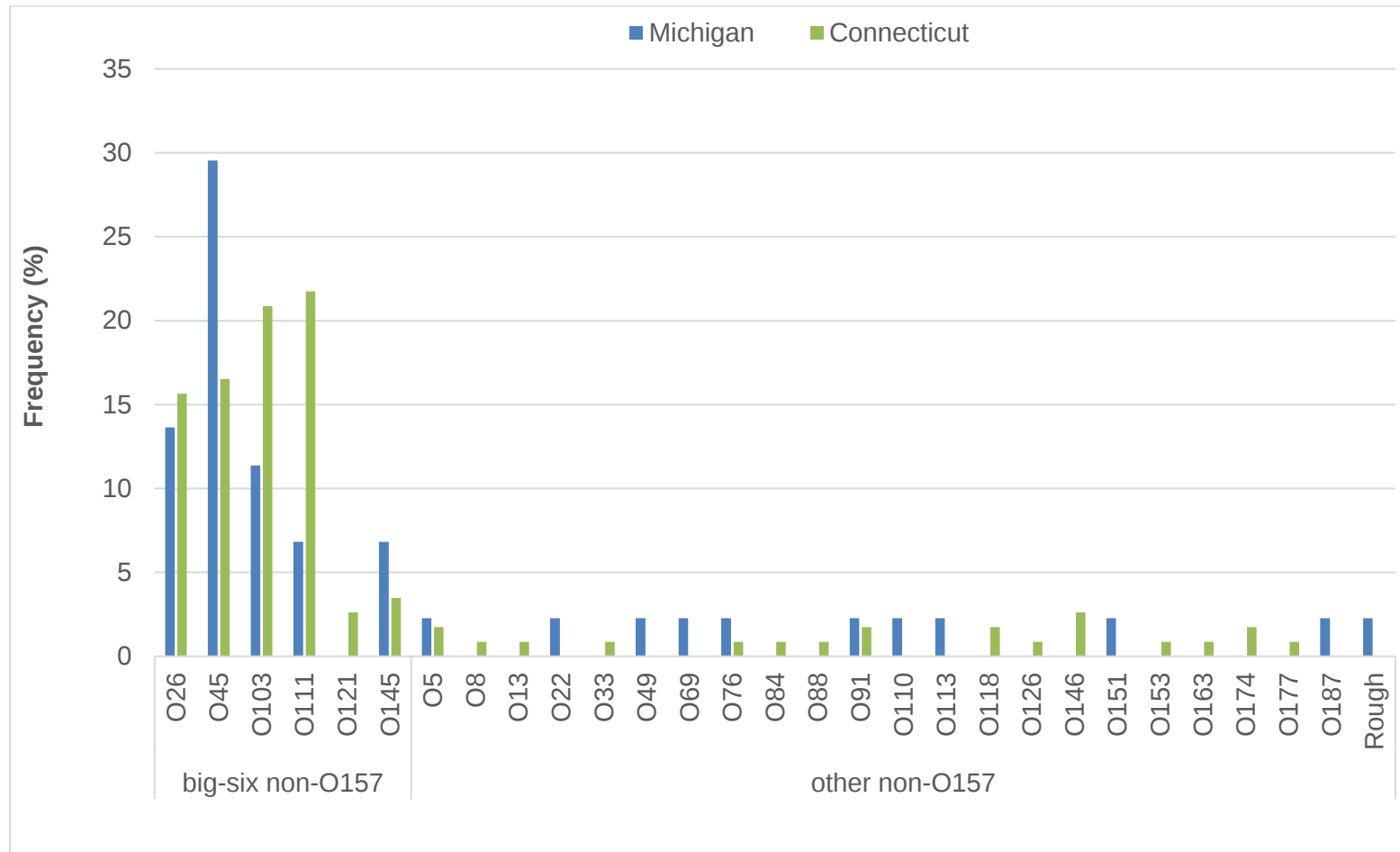


Figure 3.2. Distribution and gene frequency of virulence genes in STEC serogroups.

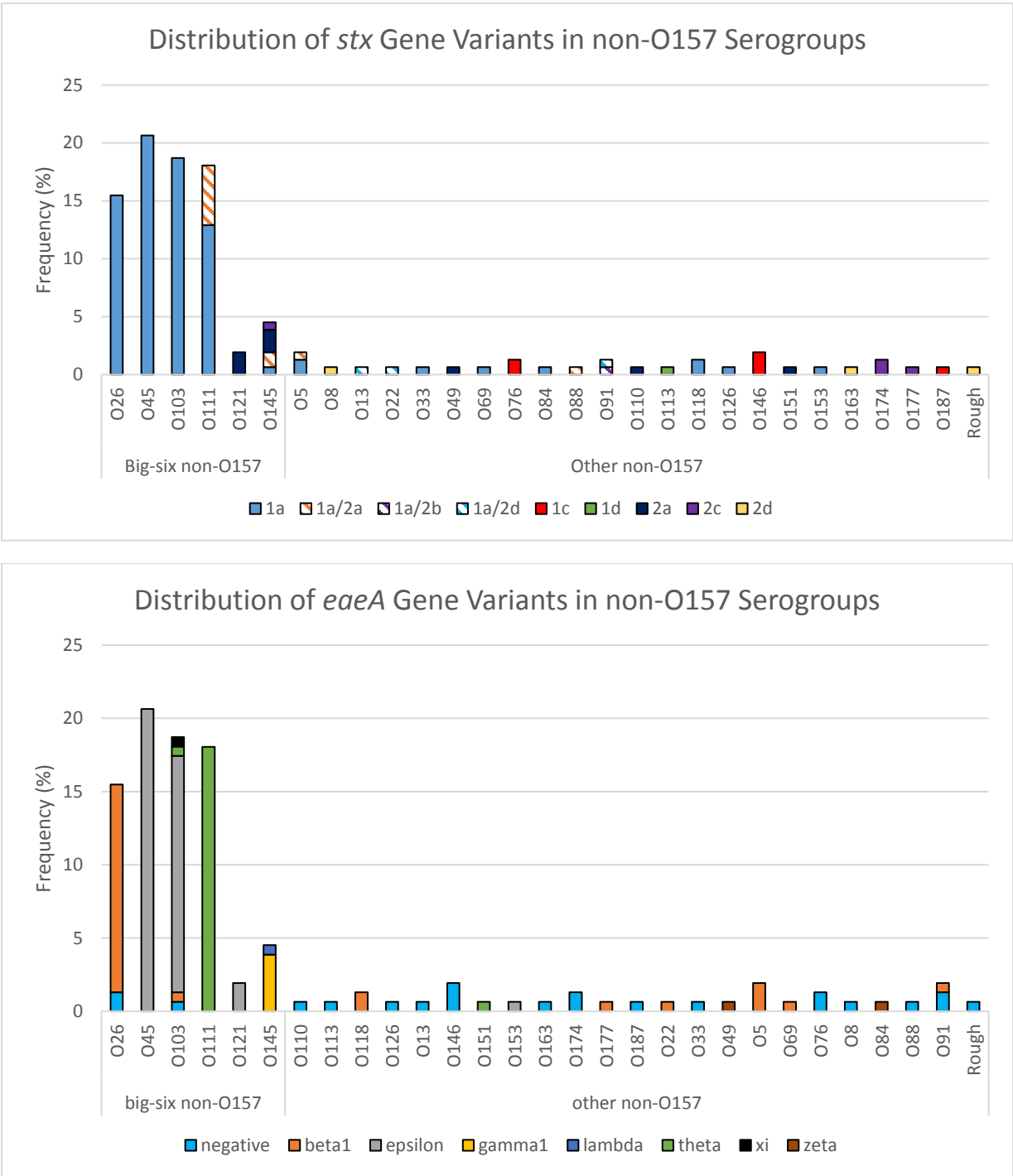


Figure 3.3. Neighbor-joining phylogenetic analysis constructed using seven gene MLST in 155 clinical STEC isolates from Michigan (n=44, green circles) and Connecticut (n=111, blue circles) with 1000 bootstrap replication to establish genetic relatedness.

* Clusters 1 and 2 represent STs that grouped together with >90% bootstrap values. STs shared across the two geographic locations are indicated with pink stars.

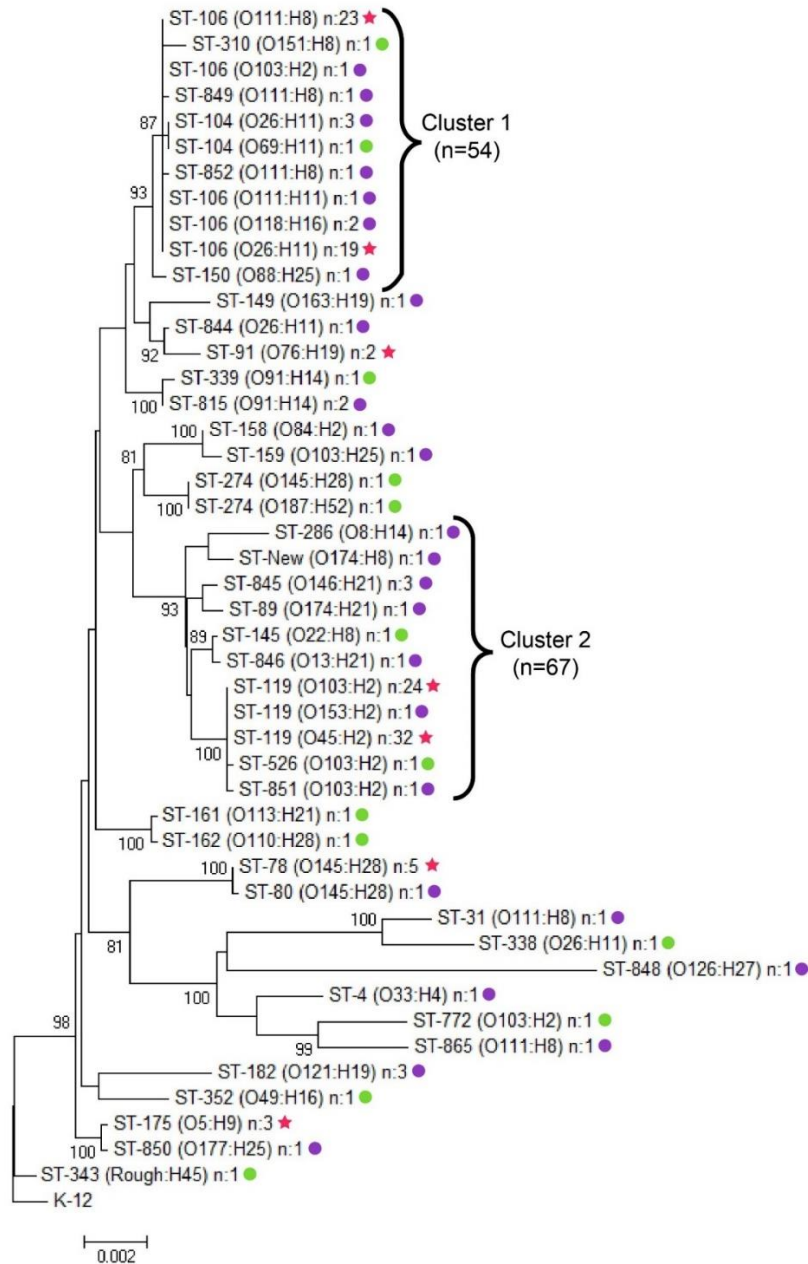


Figure 3.4. CRISPR spacer content for strains belonging to Clusters 1 and 2 as determined using MLST.

* Spacer number denotes a specific DNA sequence, color gradient is used to identify similar spacer numbers.

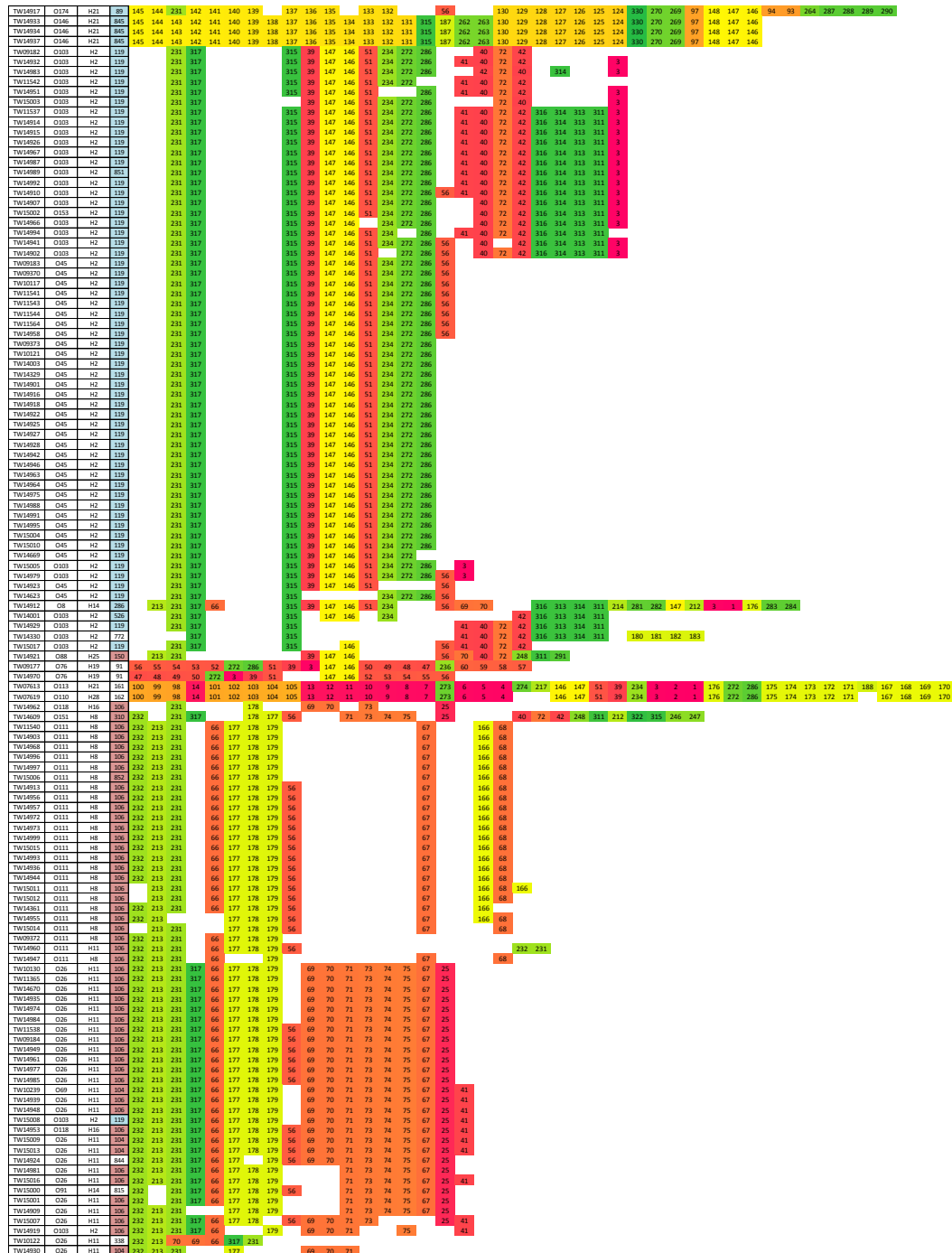
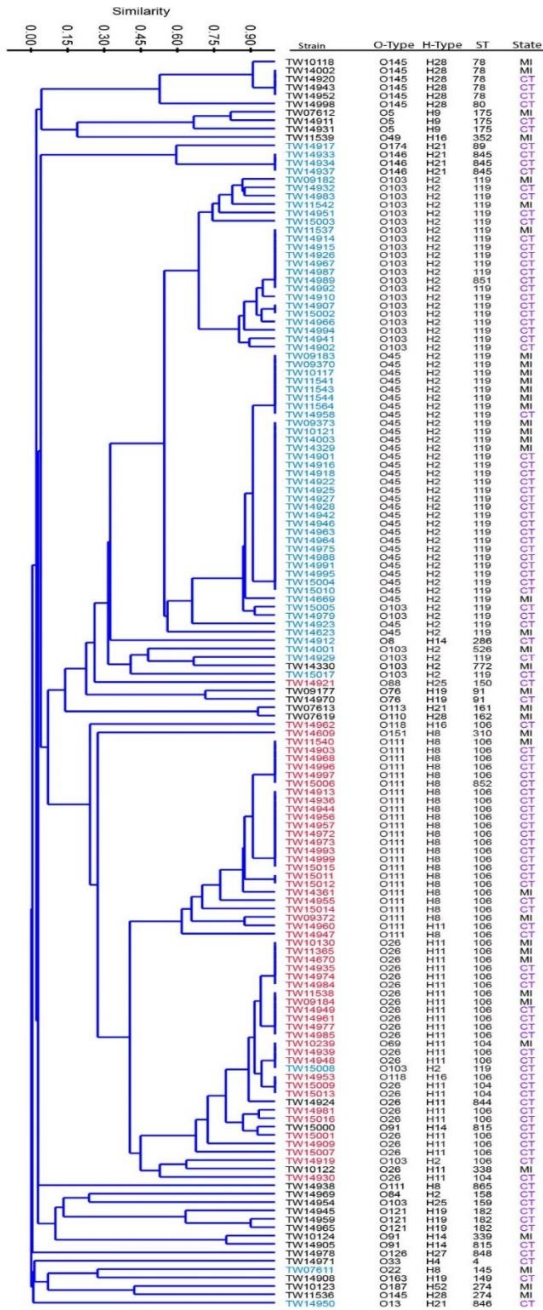


Figure 3.5. Unweighted pair group method with arithmetic averages (UPGMA) clustered using a Jaccard similarity index to compare the spacer patterns of the CRISPR profiles of 149 total isolates from Michigan (n=40) and Connecticut (n=109).

* Strains belonging to Cluster 1 (blue) and Cluster 2 (red) are differentiated based on color. All sequences are labeled with the strain name, O-type, H-type, ST and state.



REFERENCES

REFERENCES

1. Scallan E, Hoekstra RM, Angulo FJ, Tauxe R V., Widdowson M-A, Roy SL, Jones JL, Griffin PM. 2011. Foodborne Illness Acquired in the United States—Major Pathogens. *Emerg Infect Dis* 17:7–15. doi: 10.3201/eid1701.P11101.
2. O’Brien AD, Newland JW, Miller SF, Holmes RK, Smith HW, Formal SB. 1984. Shiga-like toxin-converting phages from *Escherichia coli* strains that cause hemorrhagic colitis or infantile diarrhea. *Science* 226:694–696. doi: 10.1126/science.6387911.
3. Karmali MA, Petric M, Lim C, McKeough PC, Arbus GS, Lior H. 1985. The association between idiopathic hemolytic uremic syndrome and infection by verotoxin-producing *Escherichia coli*. *J Infect Dis* 151:775–782. doi: 10.1093/infdis/151.5.775.
4. Gould LH, Mody RK, Ong KL, Clogher P, Cronquist AB, Garman KN, Lathrop S, Medus C, Spina NL, Webb TH, White PL, Wymore K, Gierke RE, Mahon BE, Griffin PM. 2013. Increased recognition of Non-O157 shiga toxin-producing *Escherichia coli* infections in the United States during 2000-2010: Epidemiologic features and comparison with *E. coli* O157 infections. *Foodborne Pathog Dis* 10:453–460. doi: 10.1089/fpd.2012.1401.
5. Crim SM, Griffin PM, Tauxe R, Marder EP, Gilliss D, Cronquist AB, Cartter M, Tobin-D’angelo M, Blythe D, Smith K, Lathrop S, Zansky S, Cieslak PR, Dunn J, Holt KG, Wolpert B, Henao OL. 2015. Preliminary incidence and trends of infection with pathogens transmitted commonly through food — Foodborne diseases active surveillance network, 10 U.S. sites, 2006–2014. *Morb Mortal Wkly Rep* 64:495–499.
6. Brooks JT, Sowers EG, Wells JG, Greene KD, Griffin PM, Hoekstra RM, Strockbine NA. 2005. Non-O157 Shiga Toxin–Producing *Escherichia coli* Infections in the United States, 1983–2002 . *J Infect Dis* 192:1422–1429. doi: 10.1086/466536.
7. Ribot EM, Fair MAA, Gautom R, Cameron DNN, Hunter SBB, Swaminathan B, Barrett TJ. 2006. Standardization of pulsed-field gel electrophoresis protocols for the subtyping of *Escherichia coli* O157:H7, *Salmonella*, and *Shigella* for PulseNet. *Foodborne Pathog Dis* 3:59–67. doi: 10.1089/fpd.2006.3.59.
8. Sabat AJ, Budimir A, Nashev D, Sá-Leão R, van Dijk JM, Laurent F, Grundmann H, Friedrich AW, on behalf of the ESCMID Study Group. 2013. Overview of molecular typing methods for outbreak detection and epidemiological surveillance. *Eurosurveillance* 18:pil=20380. doi: 10.2807/ese.18.04.20380-en.
9. Noller AC, McEllistrem MC, Stine OC, Morris JG, Boxrud DJ, Dixon B, Harrison LH. 2003. Multilocus sequence typing reveals a lack of diversity among *Escherichia coli* O157:H7 isolates that are distinct by pulsed-field gel electrophoresis. *J Clin Microbiol* 41:675–679. doi: 10.1128/JCM.41.2.675-679.2003.

10. Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, Ouellette LM, Mladonicky JM, Somsel P, Rudrik JT, Dietrich SE, Zhang W, Swaminathan B, Alland D, Whittam TS. 2008. Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks. *Proc Natl Acad Sci USA* 105:4868–4873. doi: 10.1073/pnas.0710834105.
11. Shariat N, DiMarzio MJ, Yin S, Dettinger L, Sandt CH, Lute JR, Barrangou R, Dudley EG. 2013. The combination of CRISPR-MVLST and PFGE provides increased discriminatory power for differentiating human clinical isolates of *Salmonella enterica* subsp. *enterica* serovar Enteritidis. *Food Microbiol* 34:164–173. doi: 10.1016/j.fm.2012.11.012.
12. Kovanen SM, Kivistö RI, Rossi M, Hänninen ML. 2014. A combination of MLST and CRISPR typing reveals dominant *Campylobacter jejuni* types in organically farmed laying hens. *J Appl Microbiol* 117:249–257. doi: 10.1111/jam.12503.
13. Grissa I, Vergnaud G, Pourcel C. 2007. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* 8:172. doi: 10.1186/1471-2105-8-172.
14. Jansen R, Van Embden JDA, Gastra W, Schouls LM. 2002. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* 43:1565–1575. doi: 10.1046/j.1365-2958.2002.02839.x.
15. Horvath P, Barrangou R. 2010. CRISPR/Cas, the immune system of Bacteria and Archaea. *Science* 327:167–170. doi: 10.1126/science.1179555.
16. Koonin E V., Makarova KS, Zhang F. 2017. Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol* 37:67–78. doi: 10.1016/j.mib.2017.05.008.
17. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315:1709–1712. doi: 10.1126/science.1138140.
18. Nuñez JK, Lee ASY, Engelman A, Doudna JA. 2015. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* 519:193–198. doi: 10.1038/nature14237.
19. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Soria E. 2005. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* 60:174–182. doi: 10.1007/s00239-004-0046-3.
20. Bolotin A, Quinquis B, Sorokin A, Dusko Ehrlich S. 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* 151:2551–2561. doi: 10.1099/mic.0.28048-0.

21. Pourcel C, Salvignol G, Vergnaud G. 2005. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* 151:653–663. doi: 10.1099/mic.0.27437-0.
22. Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, Snijders APL, Dickman MJ, Makarova KS, Koonin E V., van der Oost J. 2008. Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science* 321:960–964. doi: 10.1126/science.1159689.
23. Carte J, Wang R, Li H, Terns RM, Terns MP. 2008. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev* 22:3489–3496. doi: 10.1101/gad.1742908.
24. Jackson RN, Wiedenheft B. 2015. A Conserved Structural Chassis for Mounting Versatile CRISPR RNA-Guided Immune Responses. *Mol Cell* 58:722–728. doi: 10.1016/j.molcel.2015.05.023.
25. Marraffini LA, Sontheimer EJ. 2008. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* 322:1843–1845. doi: 10.1126/science.1165771.
26. Garneau JE, Dupuis MÈ, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadán AH, Moineau S. 2010. The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* 468:67–71. doi: 10.1038/nature09523.
27. Xue C, Sashital DG. 2019. Mechanisms of Type I-E and I-F CRISPR-Cas Systems in Enterobacteriaceae. *EcoSal Plus* 8:ESP-0008-2018. doi: 10.1128/ecosalplus.esp-0008-2018.
28. Touchon M, Charpentier S, Clermont O, Rocha EPC, Denamur E, Branger C. 2011. CRISPR distribution within the *Escherichia coli* species is not suggestive of immunity-associated diversifying selection. *J Bacteriol* 193:2460–2467. doi: 10.1128/JB.01307-10.
29. Díez-Villaseñor C, Almendros C, García-Martínez J, Mojica FJM. 2010. Diversity of CRISPR loci in *Escherichia coli*. *Microbiology* 156:1351–1361. doi: 10.1099/mic.0.036046-0.
30. Babu M, Beloglazova N, Flick R, Graham C, Skarina T, Nocek B, Gagarinova A, Pogoutse O, Brown G, Binkowski A, Phanse S, Joachimiak A, Koonin E V., Savchenko A, Emili A, Greenblatt J, Edwards AM, Yakunin AF. 2011. A dual function of the CRISPR-Cas system in bacterial antiviral immunity and DNA repair. *Mol Microbiol* 79:484–502. doi: 10.1111/j.1365-2958.2010.07465.x.
31. Yin S, Jensen MA, Bai J, DebRoy C, Barrangou R, Dudley EG. 2013. The Evolutionary Divergence of Shiga Toxin-Producing *Escherichia coli* Is Reflected in Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) Spacer Composition. *Appl Environ Microbiol* 79:5710–5720. doi: 10.1128/AEM.00950-13.

32. Manning SD, Madera RT, Schneider W, Dietrich SE, Khalife W, Brown W, Whittam TS, Somsel P, Rudrik JT. 2007. Surveillance for Shiga toxin-producing *Escherichia coli*, Michigan, 2001-2005. *Emerg Infect Dis* 13:318–321. doi: 10.3201/eid1302.060813.
33. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V., Sirotkin A V., Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477. doi: 10.1089/cmb.2012.0021.
34. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. doi: 10.1093/bioinformatics/btu170.
35. Andrews S. 2010. FASTQC, a quality control tool for the high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410. doi: 10.1016/S0022-2836(05)80360-2.
37. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. 2016. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 44:W16–W21. doi: 10.1093/nar/gkw387.
38. Carattoli A, Zankari E, Garcíá-Fernández A, Larsen MV, Lund O, Villa L, Aarestrup FM, Hasman H. 2014. In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* 58:3895–3903. doi: 10.1128/AAC.02412-14.
39. Grissa I, Vergnaud G, Pourcel C. 2008. CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* 36:52–57. doi: 10.1093/nar/gkn228.
40. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28:1647–1649. doi: 10.1093/bioinformatics/bts199.
41. Sheludchenko MS, Huygens F, Stratton H, Hargreaves M. 2015. CRISPR Diversity in *E. coli* Isolates from Australian Animals, Humans and Environmental Waters. *PLoS One* 10:e0124090. doi: 10.1371/journal.pone.0124090.
42. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol Biol Evol* 33:msw054. doi: 10.1093/molbev/msw054.

43. Hammer Ø, Harper DA, Ryan PD. 2001. Past: Paleontological Statistics Software Package for Education and Data Analysis. *Palaeontol Electron* 4:1–9.
44. Cowley LA, Beckett SJ, Chase-topping M, Perry N, Dallman TJ, Gally DL, Jenkins C. 2015. Analysis of whole genome sequencing for the *Escherichia coli* O157 : H7 typing phages. *BMC Genomics* 16:271. doi: 10.1186/s12864-015-1470-z.
45. Marder EP, Griffin PM, Cieslak PR, Dunn J, Hurd S, Jervis R, Lathrop S, Muse A, Ryan P, Smith K, Tobin-D'Angelo M, Vugia DJ, Holt KG, Wolpert BJ, Tauxe R, Geissler AL. 2018. Preliminary incidence and trends of infections with pathogens transmitted commonly through food - foodborne diseases active surveillance network, 10 U.S. sites, 2006-2017. *Morb Mortal Wkly Rep* 67:324–328. doi: 10.15585/mmwr.mm6711a3.
46. Economic Research Service United States Department of Agriculture. 2006. Changes in the Size and Location of U.S. Dairy Farms. Profits, Costs, Chang Struct Dairy Farming ERR-47:2–4.
47. Frank C, Kapfhammer S, Werber D, Stark K, Held L. 2008. Cattle density and Shiga toxin-producing *Escherichia coli* infection in Germany: increased risk for most but not all serogroups. *Vector Borne Zoonotic Dis* 8:635–643. doi: 10.1089/vbz.2007.0237.
48. Friesema IHM, Van De Kastele J, De Jager CM, Heuvelink AE, Van Pelt W. 2011. Geographical association between livestock density and human Shiga toxin-producing *Escherichia coli* O157 infections. *Epidemiol Infect* 139:1081–1087. doi: 10.1017/S0950268810002050.
49. Messens W, Bolton D, Frankel G, Liebana E, McLauchlin J, Morabito S, Oswald E, Threlfall EJ. 2015. Defining pathogenic verocytotoxin-producing *Escherichia coli* (VTEC) from cases of human infection in the European Union, 2007-2010. *Epidemiol Infect* 143:1652–1661. doi: 10.1017/S095026881400137X.
50. Beutin L, Krause G, Zimmermann S, Kaulfuss S, Gleier K. 2004. Characterization of Shiga Toxin-Producing *Escherichia coli* Strains Isolated from Human Patients in Germany over a 3-Year Period. *J Clin Microbiol* 42:1099–1108. doi: 10.1128/JCM.42.3.1099-1108.2004.
51. European Centre for Disease Prevention and Control. 2016. Annual Epidemiological Report 2016 – Shigatoxin/verocytotoxin-producing *Escherichia coli* infection. Stockholm: ECDC.
52. Lathrop S, Edge K, Bareta J. 2009. Shiga toxin-producing *Escherichia coli*, New Mexico, USA, 2004-2007. *Emerg Infect Dis* 15:1289–1291. doi: 10.3201/eid1508.081515.
53. Tseng M, Sha Q, Rudrik JT, Collins J, Henderson T, Funk JA, Manning SD. 2016. Increasing incidence of non-O157 Shiga toxin-producing *Escherichia coli* (STEC) in Michigan and association with clinical illness. *Epidemiol Infect* 144:1394–1405. doi: 10.1017/S0950268815002836.

54. Eichhorn I, Heidemanns K, Semmler T, Kinnemann B, Mellmann A, Harmsen D, Anjum MF, Schmidt H, Fruth A, Valentin-Weigand P, Heesemann J, Suerbaum S, Karch H, Wieler LH. 2015. Highly virulent non-O157 enterohemorrhagic *Escherichia coli* (EHEC) serotypes reflect similar phylogenetic lineages, providing new insights into the evolution of EHEC. *Appl Environ Microbiol* 81:7041–7047. doi: 10.1128/AEM.01921-15.
55. Feng P, Lampel KAA, Karch H, Whittam TSS. 1998. Genotypic and Phenotypic Changes in the Emergence of *Escherichia coli* O157:H7. *J Infect Dis* 177:1750–1753. doi: 10.1086/517438.
56. Shariat N, Timme RE, Pettengill JB, Barrangou R, Dudley EG. 2015. Characterization and evolution of *Salmonella* CRISPR-Cas systems. *Microbiology* 161:374–386. doi: 10.1099/mic.0.000005.
57. Touchon M, Charpentier S, Pognard D, Picard B, Arlet G, Rocha EPC, Denamur E, Branger C. 2012. Antibiotic resistance plasmids spread among natural isolates of *Escherichia coli* in spite of CRISPR elements. *Microbiol (United Kingdom)* 158:2997–3004. doi: 10.1099/mic.0.060814-0.
58. Mojica FJM, Díez-Villaseñor C. 2010. The on-off switch of CRISPR immunity against phages in *Escherichia coli*. *Mol Microbiol* 77:1341–1345. doi: 10.1111/j.1365-2958.2010.07326.x.
59. Edgar R, Qimron U. 2010. The *Escherichia coli* CRISPR system protects from λ lysogenization, lysogens, and prophage induction. *J Bacteriol* 192:6291–6294. doi: 10.1128/JB.00644-10.

CHAPTER 4

ANALYSIS OF WHOLE GENOME SEQUENCING FOR CHARACTERIZATION AND OUTBREAK IDENTIFICATION OF SHIGA TOXIN-PRODUCING *ESCHERICHIA COLI* (STEC) STRAINS, 2015-2018

ABSTRACT

Shiga toxin-producing *Escherichia coli* (STEC) is a leading cause of foodborne infections in both developed and underdeveloped countries. In the US, STEC is responsible for 265,000 illnesses and numerous outbreaks each year. The use of pulsed-field gel electrophoresis (PFGE) was the gold standard for surveillance until the recent transition to whole genome sequencing (WGS). Retrospective analysis of 510 clinical STEC isolates were examined in Michigan to further understand the genetic diversity and relatedness of outbreak associated isolates. In all, 34 typeable serogroups were identified including those belonging to the big six non-O157 STEC serogroups (59.6%). Core genome analysis was able to differentiate clusters of isolates with similar PFGE patterns and multilocus sequence types (STs). Two isolates belonging to serogroup O26 and NT, which were classified as outbreak-associated by PFGE and clustered together within ST-106/104 by MLST, were found to be distantly related via core genome analysis. Conversely, core genome analysis clustered six outbreak-associated serogroup O5 isolates within the same clade along with five other ST-175 serogroup O5 isolates. The use of high-quality single nucleotide polymorphism analysis could further discriminate the outbreak-associated ST-175 O5 strains into a single cluster. Indeed, use of WGS has identified genetic differences that are important for grouping strains thought to be genetically related via a given typing method. Implementation of WGS in public health labs will allow for further differentiation of related strains in addition to classifying virulence genes and serogroups, particularly for isolates that may be considered non-typeable using conventional methods.

INTRODUCTION

Shiga toxin-producing *Escherichia coli* (STEC) is a prominent foodborne pathogen that is the etiological agent for 265,000 illnesses annually and has been responsible for numerous outbreaks since identification (1–4). A diverse range of STEC serogroups including O157 and non-O157, have been associated with disease outcomes including diarrhea, hemolytic uremic syndrome (HUS), and kidney failure (5–7). Since non-O157 STEC was identified as a nationally notifiable disease, the incidence of non-O157 has been steadily increasing and has since surpassed O157 (2, 8, 9).

STEC outbreaks have been associated with various food items ranging from chicken and beef products to flour and lettuce and have been caused by a wide range of serogroups, both O157 and non-O157 (1, 10–14). Foodborne transmission is estimated to account for 85% of O157 cases annually (15). The ability to accurately track the emergence of an outbreak is crucial to prevent further infections associated with contaminated food items. Until recently, the use of pulsed-field gel electrophoresis (PFGE) has been the gold standard for STEC surveillance by PulseNet at the Centers for Disease Control and Prevention (CDC) (16–18). PFGE standardization allowed for the comparison of banding patterns nationally (19). However, usage of PFGE is time and labor intensive and does not allow for a high enough discriminative power to identify if strains with similar patterns are related (16). Analysis of PFGE patterns have grouped together strains with the same H-type regardless of the O-type, further supporting the need to discriminate strains using whole genome sequencing (WGS) to identify genetic factors (20). The 2011 O104:H4 German outbreak strain was indistinguishable from other O104 strains when using conventional epidemiological typing methods such as PFGE, serotyping, multilocus

sequence typing (MLST), optical mapping, and REP-PCR (repetitive extragenic palindromic PCR) (21, 22). The use of sequencing technology, however, enabled the identification of a novel O104:H4 strain, which was classified as Shiga-toxin producing enteroaggregative *Escherichia coli* (EAEC) and could be differentiated from other O104 isolates (23). Further sequencing of outbreak strains from two regions identified that 19 Single Nucleotide Polymorphisms (SNPs) occurred among the cases during the outbreak (24).

Increasing use of WGS has allowed for a better understanding of the genetic diversity within non-O157 and O157 STEC strains. The use of WGS allows for public health laboratories (PHL) to examine the relatedness of strains and identify other genetic factors such as antibiotic resistance, virulence genes, plasmids, and serotyping genes, that may be crucial for surveillance. WGS has the ability to identify, type, and characterize pathogens more quickly and precisely than traditional microbiological methods and with a higher resolution than other molecular methods (25, 26). The use of WGS in other foodborne organisms has already identified more outbreak clusters than other conventional methods, and more outbreaks have been solved or linked to a source since the implementation of WGS (27). WGS also allows for the ability to examine pathogens that are rapidly evolving to detect virulence genes and antibiotic resistance markers, which are easily transmitted among pathogens.

Retrospective analysis of strains isolated from patients during 2015-2018 in Michigan will allow for a complete genomic assessment of STEC isolates associated with illness. The ability to overlay PFGE data onto genomic data will also help identify strains that may not have been included in outbreak investigations or would not be considered related using PFGE analysis. The use of WGS will allow for a stricter discrimination of isolates associated with

outbreaks and an enhanced understanding of genetic variation, which will allow for better detection and identification of strains with pathogenic potential.

MATERIALS AND METHODS

Bacterial strains, DNA isolation and whole genome sequencing (WGS)

The Michigan Department of Health and Human Services (MDHHS) recovered and sequenced 625 clinical isolates during 2015-2018 that were identified to be STEC or *Shigella*. Isolates were grown overnight at 37°C and prepped for sequencing using standard operating procedures established for PulseNet at the CDC (<https://www.cdc.gov/pulsenet/pathogens/wgs>). DNA was extracted using the Qiagen DNeasy Kit (Qiagen, Valencia, CA, USA) and libraries were prepared using the Nextera XT library prep kit (Illumina, San Diego, CA, USA). Sequencing was performed on the Illumina MiSeq platform (2x250 reads) (Illumina, San Diego, CA).

Pulsed-Field Gel Electrophoresis (PFGE)

PFGE was performed on clinical isolates as part of the PulseNet national surveillance according to the standard operating procedure for PulseNet PFGE of O157:H7 and non-O157 STEC isolates developed at the CDC. Patterns were analyzed in BioNumerics (Applied Maths, Austin, TX) and assigned outbreak codes by the CDC if they matched other isolates in the database with the same banding pattern.

Bioinformatic Analysis

Prior to read processing and analysis, Kraken (28) was utilized to identify *Shigella* isolates and removed from the analysis. Preprocessing of the reads were performed with Trimmomatic (29) to remove adapters, reads with a phred quality score lower than 20 (Q20), and reads with lengths less than 100 nucleotides. Quality checking of the sequences was performed with FastQC (30) and *de novo* assemblies were performed with Spades 3.10.1 using kmers 21,

33, 55, 77, 99, and 127 for the assemblies (31). Error correction was performed during the assemblies to minimize the number of mismatches in the final contigs.

Abricate (<https://github.com/tseeman/abricate>) was used for *in silico* serotyping utilizing databases downloaded from the Center for Genomic Epidemiology (<http://www.genomicepidemiology.org/>) for the *wzy/wzx* (O-antigen), *fliC* (H-antigen), and *stx1* (Shiga toxin 1) and *stx2* (Shiga toxin 2) genes. Seven MLST gene sequences were extracted using in-house scripts developed with a Basic Local Alignment Search Tool (BLAST) platform available at the National Center for Biotechnology Information (NCBI) (32). Sequence types (STs) were assigned using EcMLST v1.2 via the STEC Center at Michigan State University (<http://www.shigatox.net>). Preliminary core genome single nucleotide polymorphism (cgSNP) analysis was performed with Parsnp on large clusters from the MLST phylogeny and strains with similar serogroups (33). Using parameters described by Katz et al. for STEC (34), Lyve-SET was used to examine high quality SNPs (hqSNP) within clusters identified in Parsnp that were associated with outbreak isolates to better define SNP differences between strains.

Data Analysis and Visualization

MLST alleles were concatenated and aligned using CLUSTALW, and a phylogenetic tree was generated using the Neighbor-joining algorithm with 1000 bootstrap replication in MEGA X (35). Core genome SNP trees were also generated in Parsnp using FastTree to infer approximate-maximum-likelihood phylogenies from SNP nucleotide alignments. High quality SNP trees were generated with Lyve-SET using RAxML to infer the maximum likelihood phylogenies from SNP alignments (34). MEGA X, TreeGraph2 (36), and FigTree (<http://tree.bio.ed.ac.uk/software/figtree>) were used to visualize each phylogeny.

RESULTS

Isolate identification and serogroup distributions

WGS was introduced into MDHHS protocols for STEC in 2015 to enhance surveillance activities and outbreak investigations. A total of 625 probable *Shigella* and diarrheagenic *E.coli* isolates were sequenced and given PNUSAE identifiers. Using WGS analysis, 97 of these isolates were classified as *Shigella* spp. and 18 isolates were *stx*-negative, resulting in removal from the analysis and leaving 510 isolates for analysis. Since its introduction, use of WGS for STEC isolates in Michigan has risen from 70.6% to 96.3% in 2015 and 2018, respectively (Figure 4.1). Over the four-year time period, 14.6% of the isolates (87 total isolates) were not sequenced due to potential duplication of isolates or low prioritization of isolates at the start of WGS introduction since PFGE remained the gold standard for surveillance activities until the beginning of 2019. Of the isolates not sequenced, 35 and 34 were recovered in 2015 and 2016 respectively, and 60 of those isolates were typed as O157.

The frequency of non-O157 serogroups isolated and sequenced from 2015-2018 decreased from 94.0% to 73.1%, while the frequency of sequenced O157 strains increased from 6.0% to 26.9%. In all, 34 typeable serogroups were identified over the four-year period and non-O157 strains belonging to the “big-six” serogroups comprised 59.6% of all serogroups identified (Table 4.1). Other serogroups that were highly prevalent during this time period included O5, O71, O123, and O151. Eleven isolates were O-antigen untypeable due to incomplete or missing *wzx/wzy* genes and one isolate was H antigen negative [H-] due to an incomplete *fliC*.

Phylogenetic analysis based on MLST loci

PFGE analysis identified 352 unique *XbaI* PFGE patterns, while MLST analysis typed 509 STEC isolates into 46 STs, with 60 serogroup and ST combinations (Figure 2). Ten new STs were identified (NEW1-10) with variants of *aspC* (*aspC7*), *fadD* (*fadD13*), *lysP* (*lysP1*), *mdh* (*mdh2*), and *uidA* (*uidA2*) in STs NEW1,2,4,6-8, while STs NEW3,5,9,10 had novel allele combinations. A single O103:H2 isolate had incomplete sequencing data for all seven genes and was removed from further analysis. A Neighbor-joining tree identified five clusters of STs that grouped together with significant bootstrap support (> 0.90) and contained more than 15 isolates per cluster (Figure 4.2). Within the five clusters, strains comprising ST-104/106 and ST-119 included serogroups that were identified to be outbreak associated by PFGE, however, other serogroups that were not outbreak associated shared the ST or clustered with high bootstrap support. In all, the predominant STs identified were ST-66 (18.9%), ST-106/104 (24.8%), and ST-119 (36.1%). The O157 isolates were all typed as ST-66, except two O157 isolates belonged to new STs, NEW-3 and NEW-10, which clustered together with ST-66.

Core genome SNP (cgSNP) analysis differentiates outbreak strains that cluster by MLST

Four clusters from the MLST phylogeny were selected for cgSNP analysis because each cluster contained known outbreak isolates previously found to be identical by PFGE. A total of 26 outbreak isolates were evaluated representing the following serotype/genotype combinations: ST-66 O157 (n=16), ST-106/104 O26/NT (n=2), ST-175 O5 (n=6), and ST-119 O103 (n=2).

The cgSNP analysis of the 135 ST-106/104 O26/NT isolates did not cluster the two outbreak isolates together. Rather, the two isolates, PNUSAE001592 and PNUSAE001586, grouped into two distinct clades indicating that they are not genetically related as was previously

determined by PFGE (Figure 4.3). Similarly, cgSNP analysis of the ST-119 O103 isolates did not cluster the two outbreak isolates within the same clade, indicating that isolates PNUSAE004161 and PNUSAE004654 are also not related as was indicated by PFGE and outbreak designation (Figure 4.4). Nonetheless, the ST-119 O103 cgSNP analysis could discriminate the strains into more clusters than MLST, which would be more informative for source tracking and identifying epidemiological associations.

For the 99 STEC O157 strains belonging to ST-66 O157, three previously defined outbreaks were examined to determine the relatedness of the strains using cgSNPs for comparison to PFGE and MLST (Figure 4.5). Outbreak one (ST-66-O1) comprised six isolates that clustered into a single clade along with nine non-outbreak associated isolates, only three of the nine isolates had the same *XbaI* PFGE pattern as outbreak associated isolates. Outbreaks two (ST-66-O2, 3 isolates) and three (ST-66-O3, 6 isolates) clustered within the same clade along with four non-outbreak associated isolates. Similar to the ST-119 cgSNP analysis, ST-66 isolates could be differentiated into smaller clusters that allows for epidemiological investigations to be performed with smaller groups of isolates that may be more closely related. Indeed, only two distinct clades were observed with the clustering of the outbreak associated isolates (n=6 isolates) along with nine non-outbreak associated isolates comprising one clade.

The cgSNP analysis was also performed on 14 strains belonging to ST-175 O5 due to the rarity of the serogroup in Michigan along with the high number of outbreak associated isolates (Figure 4.6).

High quality SNP (hqSNP) analysis further differentiates outbreak isolates compared to the core genome analysis and PFGE

The hqSNP analysis of ST-175 O5 isolates clustered the six outbreak isolates into one distinct clade and detected only 0-1 SNP differences among them; the isolates were also identical by PFGE (Figure 4.7). Three additional isolates with a distinct *Xba*I PFGE pattern also clustered together on a separate branch of the hqSNP phylogeny and differed by 0-11 SNPs. Within this cluster, there is one isolate (PNUSAE007117) that did not have the shared PFGE pattern because it was misclassified as O157, however, its molecular serotype based on WGS was O5:H9 and it differed by 0-77 SNPs from the other isolates.

The 99 O157 ST-66 isolates were split into two hqSNP analyses due to the distinct clustering observed in the cgSNP phylogeny. The first analysis included the six outbreak ST-66-O1 isolates and the nine other isolates from the same cgSNP cluster (Figure 4.8). *Xba*I PFGE patterns for these isolates were similar with a shift or change of a single band being the only difference from the common pattern. All ST-66-O1 outbreak isolates clustered together and differed by 0-24 SNPs. Two non-outbreak associated isolates were also closely related to outbreak isolates and exhibited an identical PFGE banding pattern. One outbreak associated isolate (PNUSAE013456) had one extra band in the PFGE pattern but was clustered within the same clade as other outbreak associated isolates.

The second analysis of O157 ST-66 outbreak isolates included those belonging to ST-66-O2 and ST-66-O3 as well as the four non-outbreak associated isolates that clustered together in the cgSNP analysis (Figure 4.9). The *Xba*I PFGE patterns for isolates representing both outbreaks differed from the first group (ST-66-O1) of outbreak isolates that were analyzed. Within this group, there were two isolates (PNUSAE000698 and PNUSAE020868) that had

multiple band differences compared to the rest of the isolates and did not cluster with the other isolates in the hqSNP phylogeny. Notably, ST-66-O2 O157 and ST-66-O3 O157 isolates clustered together with 0-7 SNP differences and very similar banding patterns; all ST-66-O2 O157 isolates had a slight shift of the first band. Within the outbreak clade, there were two isolates with few SNP differences (0-3 SNPs) compared to the outbreak isolates with identical banding patterns.

While cgSNP was able to differentiate two ST-119 O103 outbreak associated isolates into distinct clades, hqSNP was performed on the large grouping of isolates (n=60) to examine the relatedness of the isolates with other non-outbreak associated isolates that had the same ST (Figure 4.10). The two isolates fall within different clades on the hqSNP phylogeny and differ by 120 SNPs even though the *Xba*I PFGE patterns were identical. Since the cgSNP analysis distinguished the outbreak strains belonging to the ST-106/106 O26/NT a direct genome comparison was performed instead of hqSNP analysis. While the *Xba*I PFGE pattern for the two ST-106/104 O26/NT isolates were identical and they clustered together based on the MLST phylogeny, there are 25,037 SNPs that differ between them along with the serogroup and clustering based on cgSNPs.

DISCUSSION

The introduction of WGS into public health laboratories across the United States has allowed for improved surveillance and ability to detect enteric pathogens that may be epidemiologically related or from a specific food source. WGS allows for a complete genomic analysis of the strains to be performed to give insight into important genetic characteristics such as antibiotic resistance genes and virulence factors, while replacing traditional microbiological methods in a shorter turnaround time (37–40). The use of WGS has allowed for typing of strains that were previously unable to be typed due to novel serogroups, antigen cross reactivity or the unavailability of antibodies for the specific serogroup (41, 42).

It is important to note that the implementation of WGS has some limitations. While the library preparation and sequencing methodology has been standardized by the CDC for public health laboratories, the analysis of the sequencing data has been limited to those laboratories with skilled bioinformaticians on staff or with standard pipelines in place. With the switch from PFGE to WGS, the CDC has been analyzing all WGS data for national outbreaks until BioNumerics is fully functional and validated; PFGE has been used simultaneously to prevent a lapse in surveillance activities. Because use of WGS has enhanced our ability to cluster isolates with a high discriminatory power, the focus on creating cutoffs to identify clusters or excluding isolates with a slightly higher SNP differences may lead to the identification of missing epidemiological links. Conversely, the lack of cutoffs may result in the identification of many smaller clusters forming that classify some isolates as outbreak-associated even with the lack of epidemiological linkage. Identifying cutoffs will continue to evolve and vary by pathogen and analysis to determine which isolates and patients should be examined more comprehensively in an outbreak situation.

This retrospective study has allowed for an analysis of outbreak isolates by WGS for comparison to PFGE patterns to identify whether similar clustering and differentiation is observed by both methods. It also allowed for the identification of isolates that were misclassified by PFGE and should have been included in epidemiological investigations or identification of isolates that were genetically unrelated to each other but shared a similar PFGE banding pattern. The use of PFGE as the gold standard for surveillance has allowed for standardization across many labs so that national surveillance is possible to compare banding patterns (16, 43). Although PFGE allows for the clustering of potential isolates, it lacks discriminatory capacity and prevents the ability to perform more advanced phylogenetic analyses (16, 25, 44, 45).

Since WGS is still relatively new to public health laboratories, the isolates and serogroups being reported solely by WGS may not be representative of the true frequencies within a state or region. The prioritization and identification of certain serogroups may differ from the original implementation dates to the present. The recent trends of STEC non-O157 and O157 frequencies differ from what has been reported by the CDC through FoodNet, with STEC O157 decreasing nationally but increasing in Michigan and inversely occurring with non-O157 (2, 8, 46). This discrepancy may be due to fewer O157 isolates getting sequenced in the first two years with prioritization given to suspected outbreak and non-O157 isolates. During the 2015 and 2016 time period, multiple non-O157 outbreaks occurred in Michigan and elsewhere, which included the 2016 O5:H9 (ST-175) outbreak associated with contaminated cheese and a 2016 multistate O26/O121 outbreak linked to contaminated flour (47, 48).

In our analysis, the use of MLST grouped all 509 STEC isolates into large clusters. Although MLST is beneficial to examine genetic diversity within a bacterial population, the

discriminatory power is low (49–51). Indeed, within the clusters (bootstrap support > 0.90) defined by MLST, isolates from one serogroup were represented by multiple STs, while eleven serogroups were grouped within the same part of the phylogeny. This result further suggests that the O-antigen may not be discriminative enough for classifying isolates that are epidemiologically related. Alternatively, it may also indicate changes within the *rfb* region leading to a change in the O-antigen (52, 53).

Further analysis of the isolates that clustered together by MLST was performed using cgSNP and hqSNP analyses to get a better understanding of the relatedness of outbreak associated isolates as well as other isolates that were collected during the same time period. While the clustering of isolates with few SNP differences may indicate that they are related, epidemiological investigation is still required to confirm the link between the isolates and to identify potential outbreak sources. Isolates with similar PFGE patterns that group in different parts of the cgSNP and hqSNP phylogenies can occur because of mutations in the genome that do not affect the *Xba*I restriction sites or drastically change the size of the fragments; insertions/deletions of a few nucleotides are too minute to be accurately detected by gel electrophoresis (44). At the same time, isolates with distinct PFGE patterns that clustered together in the WGS analyses could be due to changes at the restriction enzyme sites or methylation of the DNA (54). The ability for PFGE to accurately identify strains that are similar is reliant on restriction enzyme sites remaining unmodified by genetic mutations such as insertions, deletions, or point mutations.

Collectively, our data highlights the need for transitioning to WGS to enhance outbreak surveillance activities and to more accurately identify isolates that should be pursued in epidemiological investigations. Both MLST and PFGE were often found to cluster strains

together, but WGS showed that some of these isolates were not related and is due to the higher discriminatory power of WGS. For example, the evaluation of two ST-106 isolates (PNUSAE001592 and PNUSAE001586) with a similar PFGE pattern demonstrated that they were distinct in the cgSNP phylogeny and differed by >25,000 SNPs. Conversely, the hgSNP phylogeny of ST-119 isolates identified a cluster of isolates that differed by 19-90 SNPs even though all strains had distinct PFGE patterns. These isolates clustered together by MLST and cgSNP but exhibited slight differences in the hgSNP analysis, which were reflected by the PFGE patterns. Most importantly, WGS accurately differentiated 17 O5 ST-175 and 99 O157 ST-66 outbreak isolates that clustered together based on both SNP analyses and were identical by PFGE.

Implementing WGS in public health laboratories will allow for more rapid characterization of foodborne pathogens and will facilitate the extraction of virulence factors, such as toxin and O-antigen genes, as well as antibiotic resistance genes to develop a preliminary assessment of virulence. At the same time, the genetic relatedness of strains can be deduced to identify isolates that should be examined for epidemiological links and outbreak investigations. Future surveillance of STEC in Michigan with WGS will allow the continual monitoring of emergent serogroups and strain types that are associated with clinical illness and allow for examination of virulence gene variants to identify factors that may impact disease severity.

APPENDIX

Table 4.1. STEC serogroups present in Michigan from *in silico* typing.

	O type	2015	2016	2017	2018	total
non-O157 other	O5	0	6	2	7	15
	O17	0	3	0	3	6
	O28ac	0	1	0	1	2
	O55	0	2	0	0	2
	O69	1	0	1	1	3
	O71	1	1	8	7	17
	O73 or O17/O77	0	0	1	1	2
	O8	0	0	0	3	3
	O80	0	0	0	1	1
	O84	0	1	2	1	4
	O85	0	0	1	0	1
	O98	0	0	0	1	1
	O109	0	1	0	0	1
	O113	0	0	1	0	1
	O115	0	1	0	0	1
	O117	0	1	0	0	1
	O118	0	0	0	1	1
	O123	4	3	3	3	13
	O130	0	1	0	0	1
	O151	2	3	3	2	10
	O153/O178	0	0	0	1	1
	O156	0	0	0	2	2
	O165	0	0	1	1	2
	O166	0	1	0	0	1
	O172	0	0	0	1	1
	O177	0	0	0	3	3
	O183	0	0	1	2	3
non-O157 big six	O26	14	12	10	15	51
	O45	13	16	21	11	61
	O103	21	23	26	41	111
	O111	10	10	11	16	47
	O121	8	13	6	1	28
	O145	0	2	0	4	6
	O157	5	11	33	49	98
	NT	5	0	1	3	9
total		84	112	132	182	510

Figure 4.1. Percent of all STEC isolates sequenced (n=510) at the MDHHS per year (black line) and the frequency of non-O157 and O157 serogroups in sequenced STEC isolates.

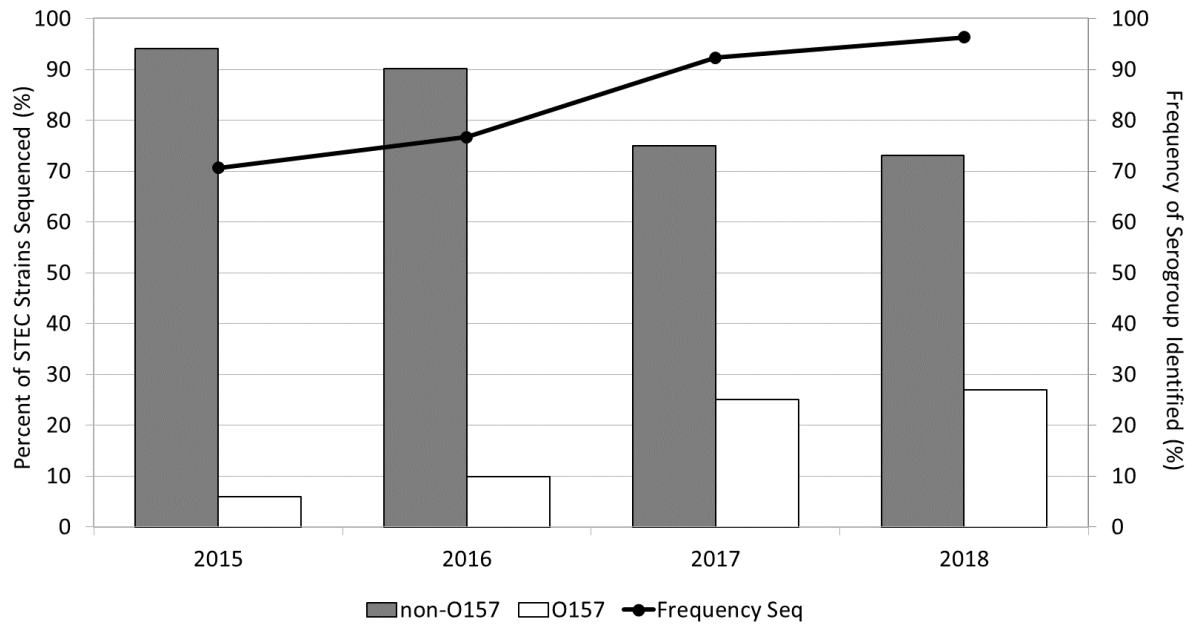


Figure 4.2. Neighbor-joining phylogenetic tree constructed based on seven MLST loci for 509 STEC isolates from 2015-2018 with 1000 bootstrap replication.

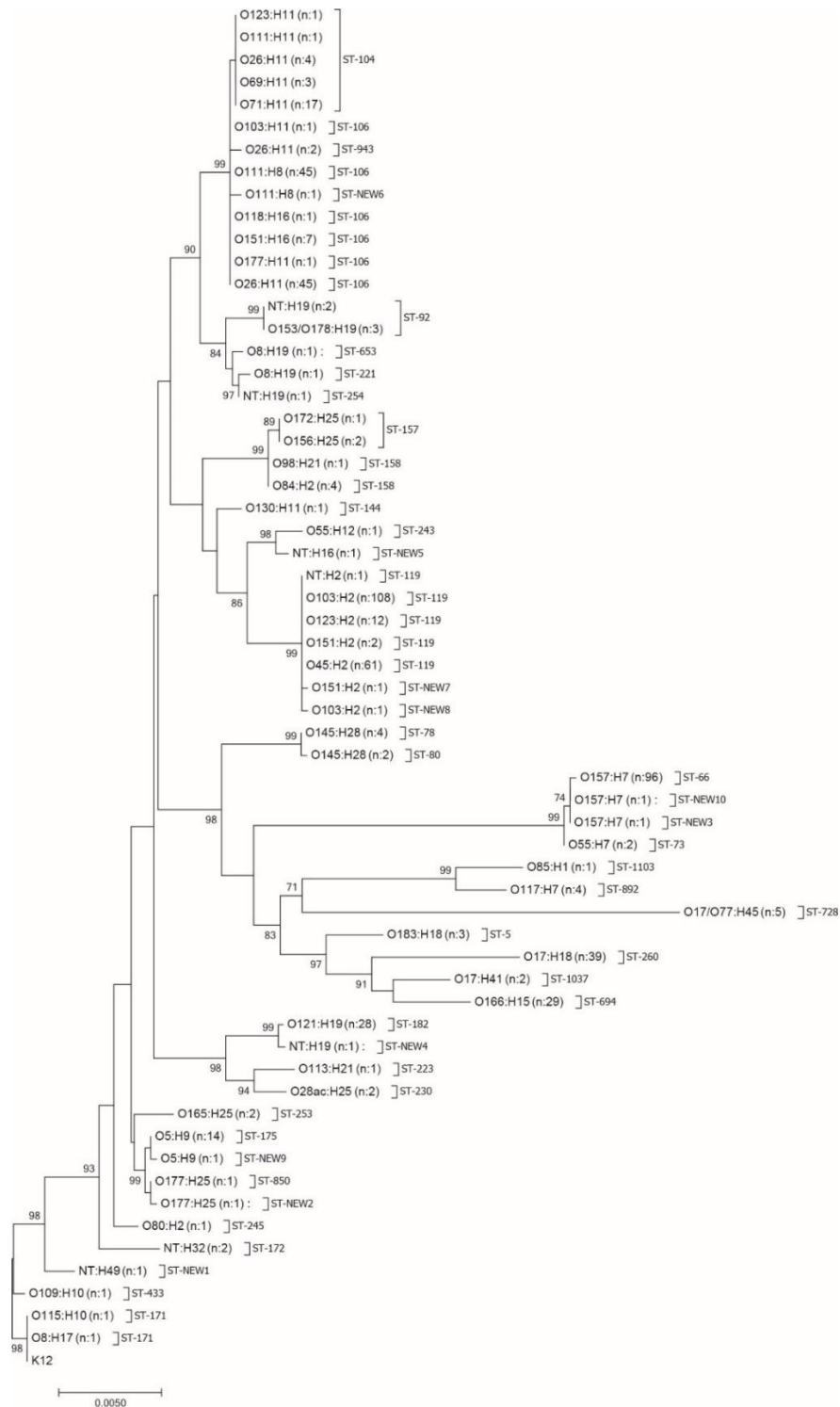


Figure 4.3. Core genome SNP analysis of 135 STEC isolates belonging to the multilocus sequence type (ST)-106/104 cluster, including serogroup O26 and NT outbreak associated isolates.



Figure 4.4. Core genome SNP analysis of 188 STEC isolates belonging to the multilocus sequence type (ST)-119 cluster, including serogroup O103 outbreak associated isolates.

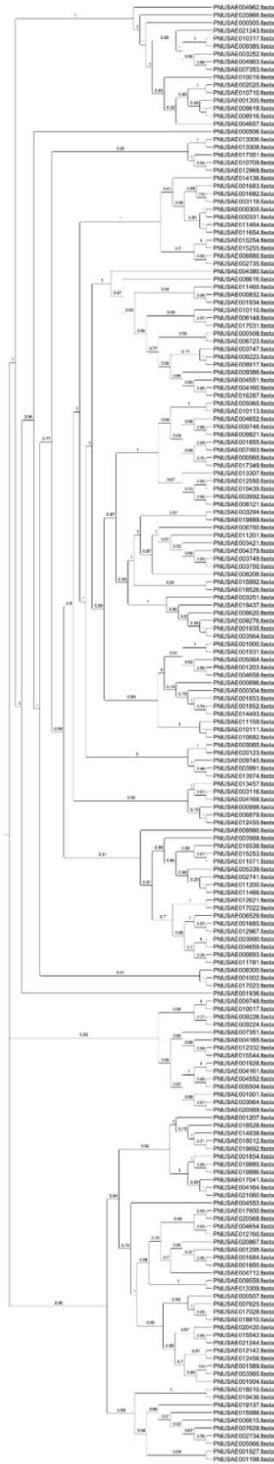


Figure 4.5. Core genome SNP analysis of 99 O157 STEC isolates belonging to the multilocus sequence type (ST)-66 cluster

* Three outbreaks denoted O1 (open star), O2 (colored star), and O3 (open triangle).

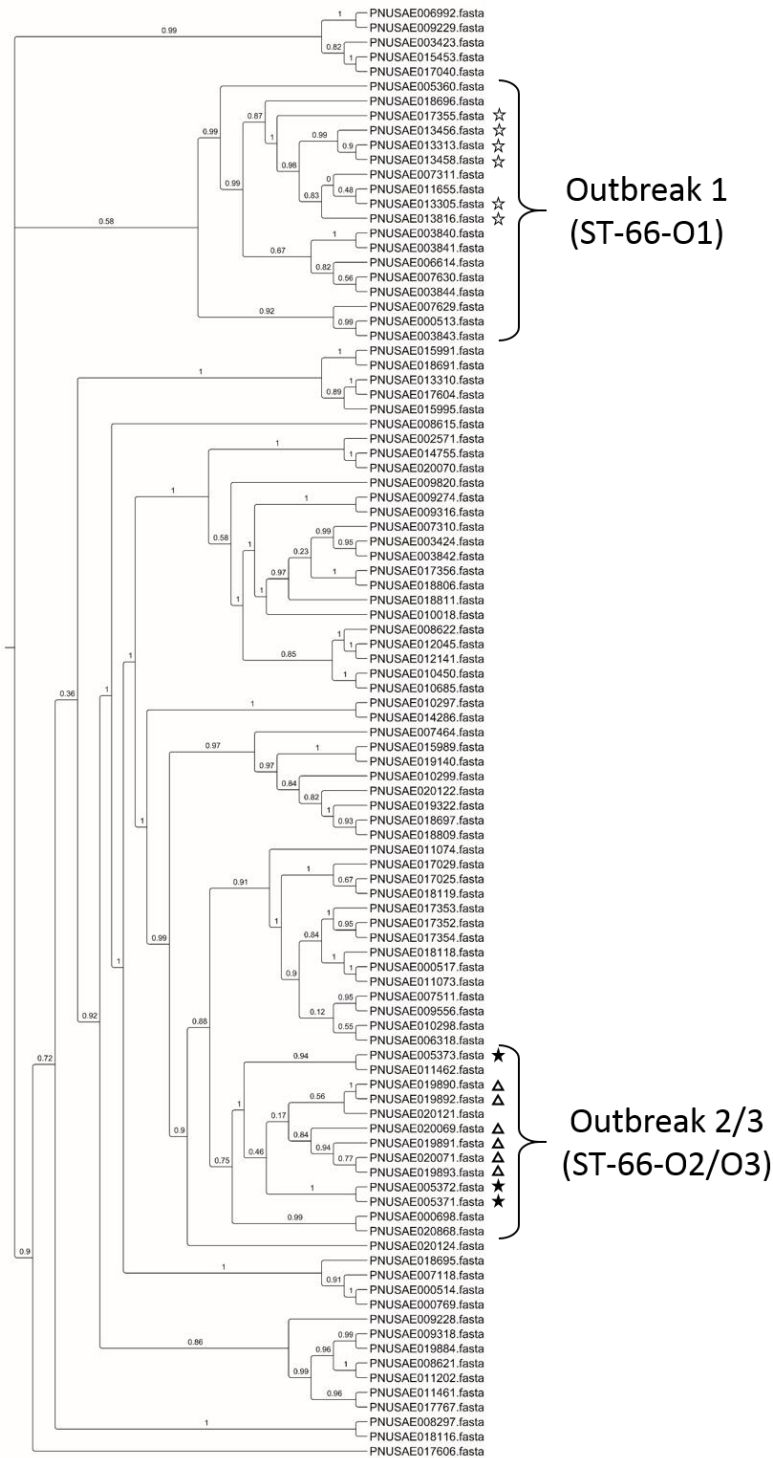


Figure 4.6. Core genome SNP analysis of 17 STEC isolates belonging to the multilocus sequence type (ST)-175 cluster, including serogroup O5 outbreak associated isolates.

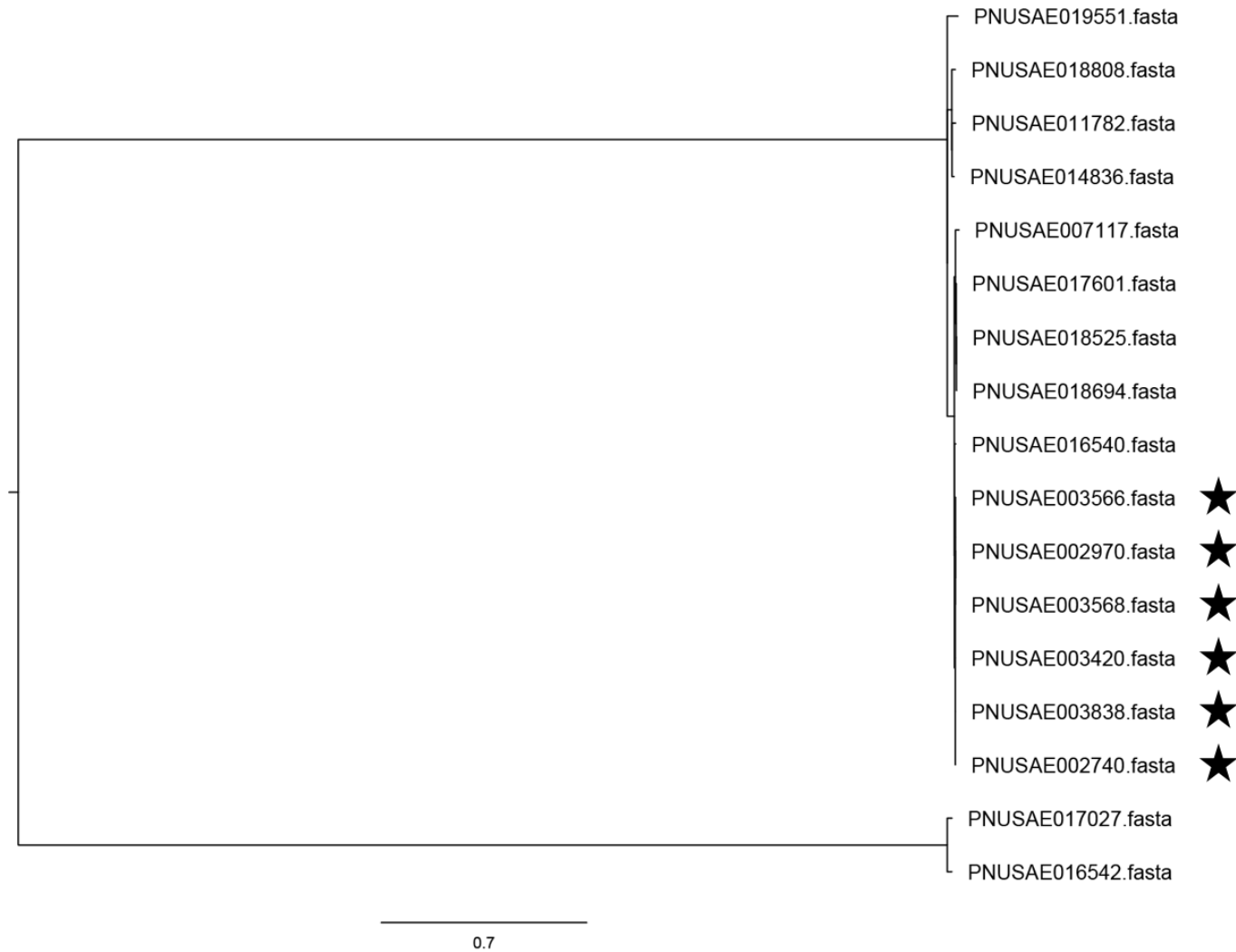


Figure 4.7. Phylogeny based on hqSNP analysis of ST-175 isolates that clustered with outbreak strains using cgSNP analysis.

* PFGE patterns (XbaI) for STEC isolates included in the hqSNP analysis and outbreak associated isolates denoted with stars.

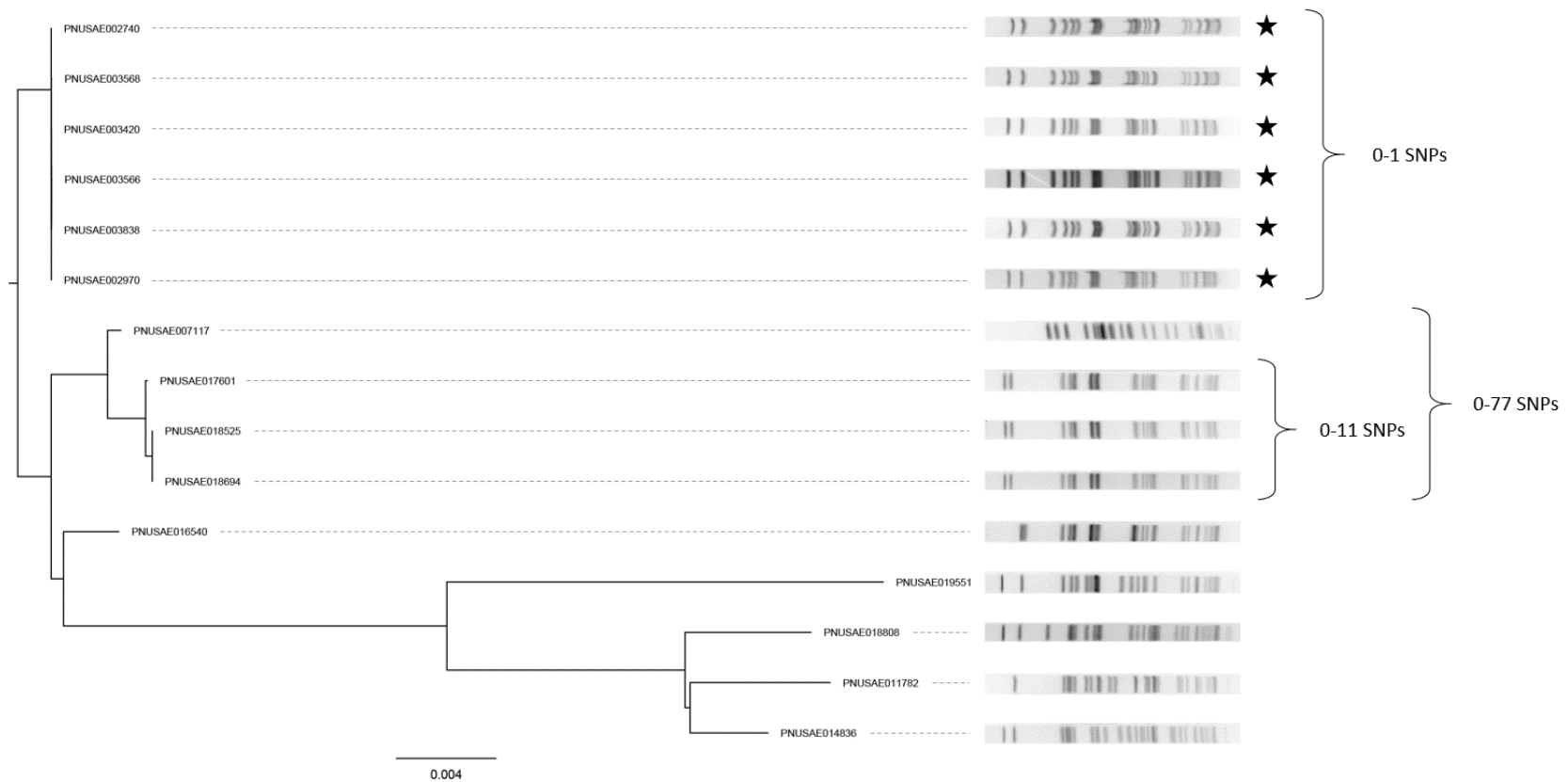


Figure 4.8. Phylogeny based on hqSNP analysis of ST-66-O1 isolates that clustered with outbreak strains using cgSNP analysis.

* PFGE patterns (XbaI) for STEC isolates included in the hqSNP analysis and outbreak associated isolates denoted with stars.

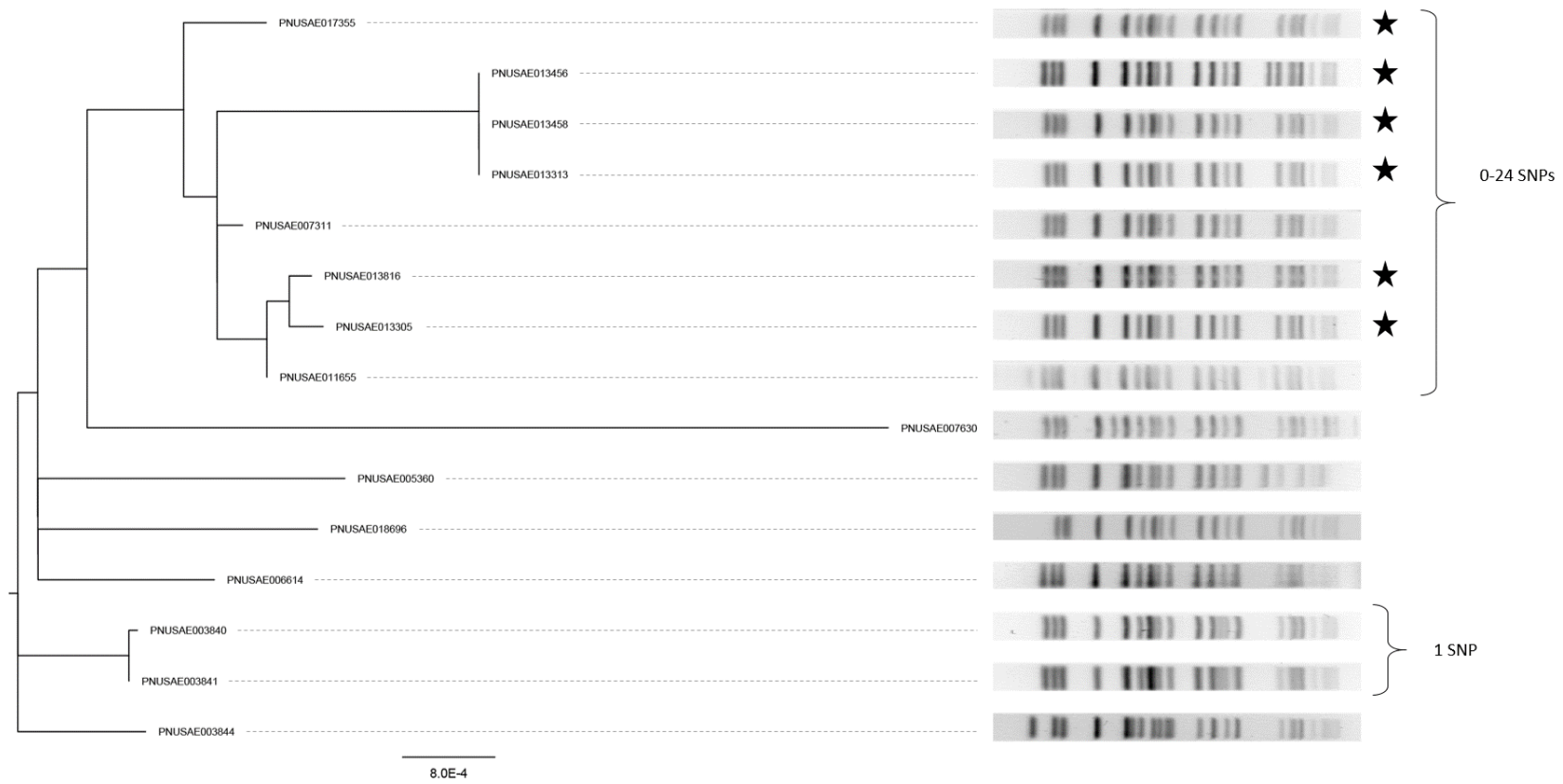


Figure 4.9: Phylogeny based on hqSNP analysis of ST-66-O2/O3 isolates that clustered with outbreak strains using cgSNP analysis.

* PFGE patterns (XbaI) for STEC isolates included in the hqSNP analysis. The two outbreaks are denoted by colored star (ST-66-O2) and open triangle (ST-66-O3).

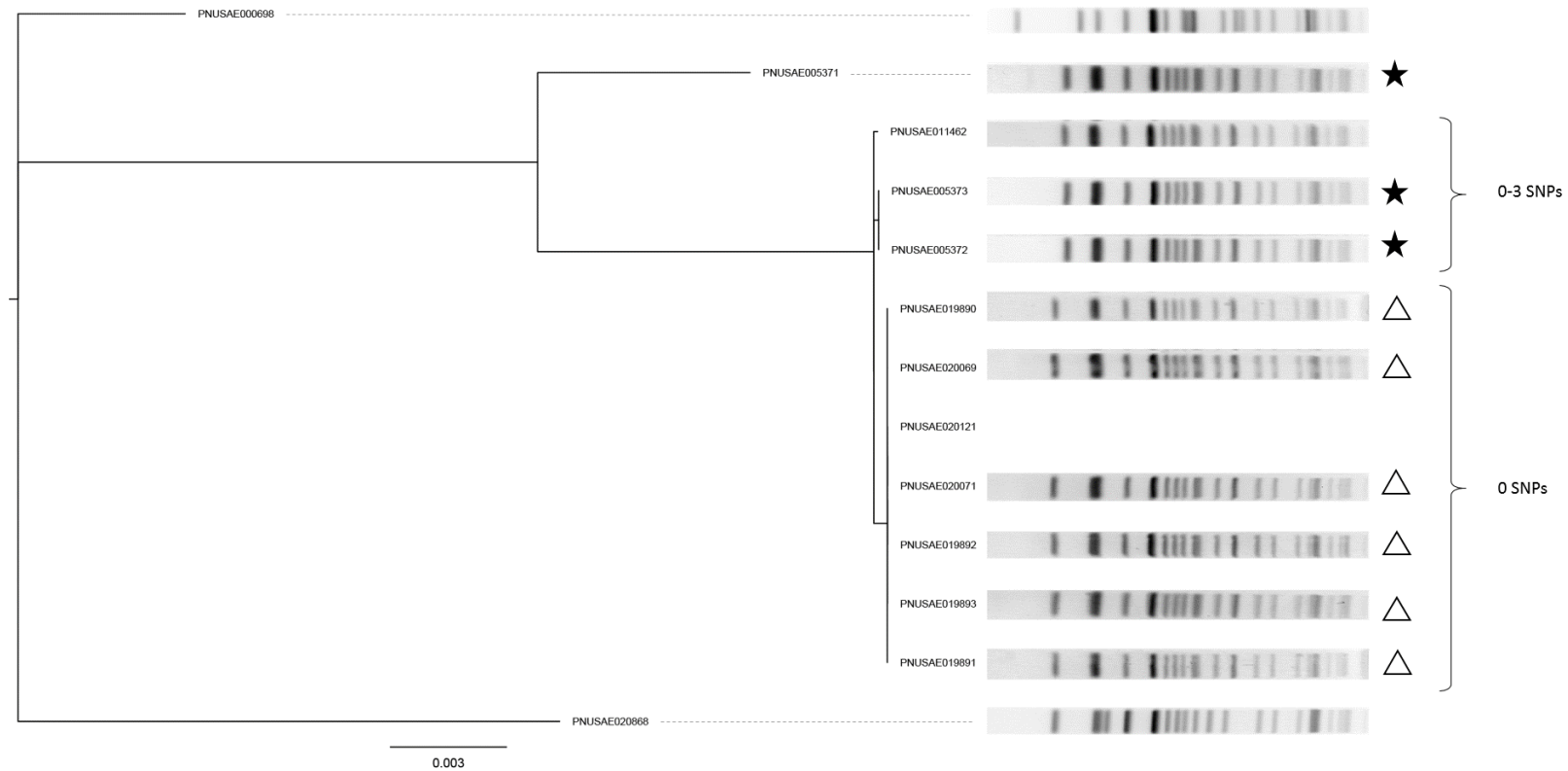
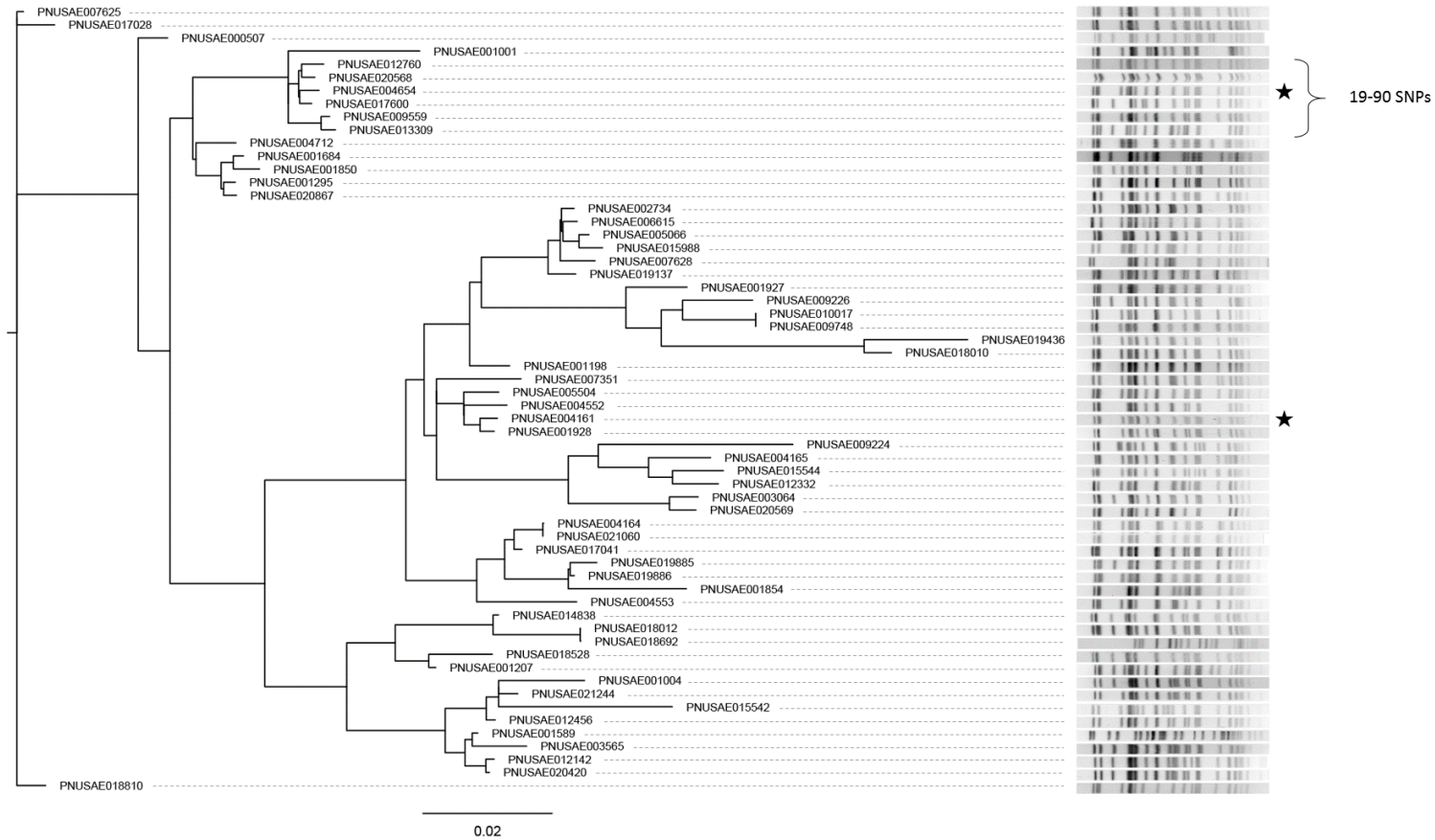


Figure 4.10. Phylogeny based on hqSNP analysis of 60 ST-119 isolates that clustered with outbreak strains using cgSNP analysis.

* PFGE patterns (*Xba*I) for STEC isolates included in the hqSNP analysis and outbreak associated isolates denoted with stars.



REFERENCES

REFERENCES

1. Luna-Gierke RE, Griffin PM, Gould LH, Herman K, Bopp CA, Strockbine N, Mody RK. 2014. Outbreaks of non-O157 Shiga toxin-producing *Escherichia coli* infection: USA. *Epidemiol Infect* 142:2270–2280. doi: 10.1017/S0950268813003233.
2. Gould LH, Mody RK, Ong KL, Clogher P, Cronquist AB, Garman KN, Lathrop S, Medus C, Spina NL, Webb TH, White PL, Wymore K, Gierke RE, Mahon BE, Griffin PM. 2013. Increased recognition of Non-O157 shiga toxin-producing *Escherichia coli* infections in the United States during 2000-2010: Epidemiologic features and comparison with *E. coli* O157 infections. *Foodborne Pathog Dis* 10:453–460. doi: 10.1089/fpd.2012.1401.
3. Rangel JM, Sparling PH, Crowe C, Griffin PM, Swerdlow DL. 2005. Epidemiology of *Escherichia coli* O157:H7 outbreaks, United States, 1982-2002. *Emerg Infect Dis* 11:603–609. doi: 10.3201/eid1104.040739.
4. Scallan E, Hoekstra RM, Angulo FJ, Tauxe R V., Widdowson M-A, Roy SL, Jones JL, Griffin PM. 2011. Foodborne Illness Acquired in the United States—Major Pathogens. *Emerg Infect Dis* 17:7–15. doi: 10.3201/eid1701.P11101.
5. Brooks JT, Sowers EG, Wells JG, Greene KD, Griffin PM, Hoekstra RM, Strockbine NA. 2005. Non-O157 Shiga Toxin-Producing *Escherichia coli* Infections in the United States, 1983–2002. *J Infect Dis* 192:1422–1429. doi: 10.1086/466536.
6. Hedican E, Medus C, Besser J, Juni B, Koziol B, Taylor C, Smith K. 2009. Characteristics of O157 versus non-O157 Shiga toxin-producing *Escherichia coli* infections in Minnesota, 2000-2006. *Clin Infect Dis* 49:358–64. doi: 10.1086/600302.
7. Haugum K, Johansen J, Gabrielsen C, Brandal LT, Bergh K, Ussery DW, Drabløs F, Afset JE. 2014. Comparative genomics to delineate pathogenic potential in Non-O157 Shiga toxin-producing *Escherichia coli* (STEC) from patients with and without haemolytic uremic syndrome (HUS) in Norway. *PLoS One* 9. doi: 10.1371/journal.pone.0111788.
8. Tseng M, Sha Q, Rudrik JT, Collins J, Henderson T, Funk JA, Manning SD. 2016. Increasing incidence of non-O157 Shiga toxin-producing *Escherichia coli* (STEC) in Michigan and association with clinical illness. *Epidemiol Infect* 144:1394–1405. doi: 10.1017/S0950268815002836.
9. CDC. 2017. Foodborne Diseases Active Surveillance Network (FoodNet): FoodNet 2015 Surveillance Report (Final Data). Centers for Disease Control and Prevention. Atlanta, GA.
10. Slayton RB, Turabelidze G, Bennett SD, Schwensohn CA, Yaffee AQ, Khan F, Butler C, Trees E, Ayers TL, Davis ML, Laufer AS, Gladbach S, Williams I, Gieraltowski LB. 2013. Outbreak of Shiga Toxin-Producing *Escherichia coli* (STEC) O157:H7 Associated

- with Romaine Lettuce Consumption, 2011. PLoS One 8:e55300. doi: 10.1371/journal.pone.0055300.
11. Mäde D, Geuthner A-C, Imming R, Wicke A. 2017. Detection and isolation of Shiga-Toxin producing *Escherichia coli* in flour in Germany between 2014 and 2017. *J Consum Prot Food Saf* 12:245–253. doi: 10.1007/s00003-017-1113-1.
 12. CDC. 2015. *Escherichia coli* O157:H7 Infections Linked to Costco Rotisserie Chicken Salad (Final Update). Centers for Disease Control and Prevention. Atlanta, GA.
 13. CDC. 2014. Multistate Outbreak of Shiga toxin-producing *Escherichia coli* O157:H7 Infections Linked to Ground Beef (Final Update). Centers for Disease Control and Prevention. Atlanta, GA.
 14. Erickson MC, Doyle MP. 2007. Food as a vehicle for transmission of Shiga toxin-producing *Escherichia coli*. *J Food Prot* 70:2426–49.
 15. Mead PS, Slutsker L, Dietz V, McCaig LF, Bresee JS, Shapiro C, Griffin PM, Tauxe R V. 1999. Food-related illness and death in the United States. *Emerg Infect Dis* 5:607–25. doi: 10.3201/eid0505.990502.
 16. Ribot EM, Fair MA, Gautom R, Cameron DN, Hunter SB, Swaminathan B, Barrett TJ. 2006. Standardization of Pulsed-Field Gel Electrophoresis Protocols for the Subtyping of *Escherichia coli* O157:H7, *Salmonella*, and *Shigella* for PulseNet. *Foodborne Pathog Dis* 3:59–67. doi: 10.1089/fpd.2006.3.59.
 17. Swaminathan B, Barrett TJ, Hunter SB, Tauxe R V. 2001. PulseNet: The Molecular Subtyping Network for Foodborne Bacterial Disease Surveillance, United States. *Emerg Infect Dis* 7:382–389. doi: 10.3201/eid0703.010303.
 18. Scharff RL, Besser J, Sharp DJ, Jones TF, Peter GS, Hedberg CW. 2016. An Economic Evaluation of PulseNet: A Network for Foodborne Disease Surveillance. *Am J Prev Med* 50:S66–S73. doi: 10.1016/j.amepre.2015.09.018.
 19. Carleton HA, Gerner-Smidt P. 2016. Whole-genome sequencing is taking over foodborne disease surveillance: Public health microbiology is undergoing its biggest change in a generation, Replacing traditional methods with whole-genome sequencing. *Microbe* 11:311–317. doi: 10.1128/microbe.11.311.1.
 20. Ju W, Cao G, Rump L, Strain E, Luo Y, Timme R, Allard M, Zhao S, Brown E, Meng J. 2012. Phylogenetic analysis of non-O157 Shiga toxin-producing *Escherichia coli* strains by whole-genome sequencing. *J Clin Microbiol* 50:4123–4127. doi: 10.1128/JCM.02262-12.
 21. Mariani-Kurkdjian P, Bingen E, Gault G, Jourdan-Da Silva N, Weill F-X. 2011. *Escherichia coli* O104:H4 south-west France, June 2011. *Lancet Infect Dis* 11:732–3. doi: 10.1016/S1473-3099(11)70266-3.

22. Gault G, Weill FX, Mariani-Kurkdjian P, Jourdan-da Silva N, King L, Aldabe B, Charron M, Ong N, Castor C, Macé M, Bingen E, Noël H, Vaillant V, Bone A, Vendrely B, Delmas Y, Combe C, Bercion R, D'Andigné E, Desjardin M, Rolland P, de Valk H. 2011. Outbreak of haemolytic uraemic syndrome and bloody diarrhoea due to *Escherichia coli* O104:H4, south-west France, June 2011. *Eurosurveillance* 16:19905. doi: 10.2807/ese.16.26.19905-en.
23. Rohde H, Qin J, Cui Y, Li D, Loman NJ, Hentschke M, Chen W, Pu F, Peng Y, Li J, Xi F, Li S, Li Y, Zhang Z, Yang X, Zhao M, Wang P, Guan Y, Cen Z, Zhao X, Christner M, Kobbe R, Loos S, Oh J, Yang L, Danchin A, Gao GF, Song Y, Li Y, Yang H, Wang J, Xu J, Pallen MJ, Wang J, Aepfelbacher M, Yang R. 2011. Open-Source Genomic Analysis of Shiga-Toxin–Producing *E. coli* O104:H4. *N Engl J Med* 365:718–724. doi: 10.1056/NEJMoa1107643.
24. Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, FitzGerald M, Godfrey P, Haas BJ, Murphy CI, Russ C, Sykes S, Walker BJ, Wortman JR, Young S, Zeng Q, Abouelleil A, Bochicchio J, Chauvin S, DeSmet T, Gujja S, McCowan C, Montmayeur A, Steelman S, Frimodt-Møller J, Petersen AM, Struve C, Krogfelt KA, Bingen E, Weill F-X, Lander ES, Nusbaum C, Birren BW, Hung DT, Hanage WP. 2012. Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. *Proc Natl Acad Sci* 109:3065–3070. doi: 10.1073/pnas.1121491109.
25. Ronholm J, Nasheri N, Petronella N, Pagotto F. 2016. Navigating microbiological food safety in the era of whole-genome sequencing. *Clin Microbiol Rev* 29:837–857. doi: 10.1128/CMR.00056-16.
26. Gwinn M, MacCannell DR, Khabbaz RF. 2017. Integrating advanced molecular technologies into public health. *J Clin Microbiol* 55:703–714. doi: 10.1128/JCM.01967-16.
27. Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A, Carleton H, Katz LS, Stroika S, Gould LH, Mody RK, Silk BJ, Beal J, Chen Y, Timme R, Doyle M, Fields A, Wise M, Tillman G, Defibaugh-Chavez S, Kucerova Z, Sabol A, Roache K, Trees E, Simmons M, Wasilenko J, Kubota K, Pouseele H, Klimke W, Besser J, Brown E, Allard M, Gerner-Smidt P. 2016. Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation. *Clin Infect Dis* 63:380–386. doi: 10.1093/cid/ciw242.
28. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 15:R46. doi: 10.1186/gb-2014-15-3-r46.
29. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. doi: 10.1093/bioinformatics/btu170.
30. Andrews S. 2010. FASTQC, a quality control tool for the high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.

31. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V., Sirotkin A V., Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477. doi: 10.1089/cmb.2012.0021.
32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410. doi: 10.1016/S0022-2836(05)80360-2.
33. Treangen TJ, Ondov BD, Koren S, Phillippy AM. 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol* 15:524. doi: 10.1186/s13059-014-0524-x.
34. Katz LS, Griswold T, Williams-Newkirk AJ, Wagner D, Petkau A, Sieffert C, Domselaar G Van, Deng X, Carleton HA. 2017. A comparative analysis of the Lyve-SET phylogenomics pipeline for genomic epidemiology of foodborne pathogens. *Front Microbiol* 8:375. doi: 10.3389/fmicb.2017.00375.
35. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* 35:1547–1549. doi: 10.1093/molbev/msy096.
36. Stöver BC, Müller KF. 2010. TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics* 11:7. doi: 10.1186/1471-2105-11-7.
37. Lindsey RL, Pouseele H, Chen JC, Strockbine NA, Carleton HA. 2016. Implementation of Whole Genome Sequencing (WGS) for Identification and Characterization of Shiga Toxin-Producing *Escherichia coli* (STEC) in the United States. *Front Microbiol* 7:766. doi: 10.3389/fmicb.2016.00766.
38. Rumore J, Tschetter L, Kearney A, Kandar R, McCormick R, Walker M, Peterson C-L, Reimer A, Nadon C. 2018. Evaluation of whole-genome sequencing for outbreak detection of Verotoxigenic *Escherichia coli* O157:H7 from the Canadian perspective. *BMC Genomics* 19:870. doi: 10.1186/s12864-018-5243-3.
39. Chattaway MA, Dallman TJ, Gentle A, Wright MJ, Long SE, Ashton PM, Perry NT, Jenkins C. 2016. Whole Genome Sequencing for Public Health Surveillance of Shiga Toxin-Producing *Escherichia coli* Other than Serogroup O157. *Front Microbiol* 7:258. doi: 10.3389/fmicb.2016.00258.
40. Parsons BD, Zelyas N, Berenger BM, Chui L. 2016. Detection, Characterization, and Typing of Shiga Toxin-Producing *Escherichia coli*. *Front Microbiol* 7:478. doi: 10.3389/fmicb.2016.00478.
41. Jenkins C, Willshaw GA, Evans J, Cheasty T, Chart H, Shaw DJ, Dougan G, Frankel G, Smith HR. 2003. Subtyping of virulence genes in verocytotoxin-producing *Escherichia coli* (VTEC) other than serogroup O157 associated with disease in the United Kingdom. *J*

- Med Microbiol 52:941–947. doi: 10.1099/jmm.0.05160-0.
42. Byrne L, Vanstone GL, Perry NT, Launders N, Adak GK, Godbole G, Grant KA, Smith R, Jenkins C. 2014. Epidemiology and microbiology of Shiga toxin-producing *Escherichia coli* other than serogroup O157 in England, 2009-2013. *J Med Microbiol* 63:1181–1188. doi: 10.1099/jmm.0.075895-0.
 43. Sabat AJ, Budimir A, Nashev D, Sá-Leão R, van Dijl JM, Laurent F, Grundmann H, Friedrich AW, on behalf of the ESCMID Study Group. 2013. Overview of molecular typing methods for outbreak detection and epidemiological surveillance. *Eurosurveillance* 18:pii=20380. doi: 10.2807/ese.18.04.20380-en.
 44. Barrett TJ, Gerner-Smidt P, Swaminathan B. 2006. Interpretation of pulsed-field gel electrophoresis patterns in foodborne disease investigations and surveillance. *Foodborne Pathog Dis* 3:20–31. doi: 10.1089/fpd.2006.3.20.
 45. Oakeson KF, Wagner JM, Rohrwasser A, Atkinson-Dunn R. 2018. Whole-genome sequencing and bioinformatic analysis of isolates from foodborne illness outbreaks of *Campylobacter jejuni* and *Salmonella enterica*. *J Clin Microbiol* 56:e00161-18. doi: 10.1128/JCM.00161-18.
 46. Marder EP, Griffin PM, Cieslak PR, Dunn J, Hurd S, Jervis R, Lathrop S, Muse A, Ryan P, Smith K, Tobin-D'Angelo M, Vugia DJ, Holt KG, Wolpert BJ, Tauxe R, Geissler AL. 2018. Preliminary incidence and trends of infections with pathogens transmitted commonly through food - foodborne diseases active surveillance network, 10 U.S. sites, 2006-2017. *Morb Mortal Wkly Rep* 67:324–328. doi: 10.15585/mmwr.mm6711a3.
 47. Hainstock L, Donovan D. 2017. The Cheese Stood Alone. 17th Annu Michigan Commun Dis Conf.
 48. Crowe SJ, Bottichio L, Shade LN, Whitney BM, Corral N, Melius B, Arends KD, Donovan D, Stone J, Allen K, Rosner J, Beal J, Whitlock L, Blackstock A, Wetherington J, Newberry LA, Schroeder MN, Wagner D, Trees E, Viazis S, Wise ME, Neil KP. 2017. Shiga toxin-producing *E. coli* infections associated with flour. *N Engl J Med* 377:2036–2043. doi: 10.1056/NEJMoa1615910.
 49. Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, Ouellette LM, Mladonicky JM, Somsel P, Rudrik JT, Dietrich SE, Zhang W, Swaminathan B, Alland D, Whittam TS. 2008. Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks. *Proc Natl Acad Sci USA* 105:4868–4873. doi: 10.1073/pnas.0710834105.
 50. Noller AC, McEllistrem MC, Stine OC, Morris JG, Boxrud DJ, Dixon B, Harrison LH. 2003. Multilocus sequence typing reveals a lack of diversity among *Escherichia coli* O157:H7 isolates that are distinct by pulsed-field gel electrophoresis. *J Clin Microbiol* 41:675–679. doi: 10.1128/JCM.41.2.675-679.2003.

51. Zhang W. 2006. Probing genomic diversity and evolution of *Escherichia coli* O157 by single nucleotide polymorphisms. *Genome Res* 16:757–767. doi: 10.1101/gr.4759706.
52. Eichhorn I, Heidemanns K, Semmler T, Kinnemann B, Mellmann A, Harmsen D, Anjum MF, Schmidt H, Fruth A, Valentin-Weigand P, Heesemann J, Suerbaum S, Karch H, Wieler LH. 2015. Highly virulent non-O157 enterohemorrhagic *Escherichia coli* (EHEC) serotypes reflect similar phylogenetic lineages, providing new insights into the evolution of EHEC. *Appl Environ Microbiol* 81:7041–7047. doi: 10.1128/AEM.01921-15.
53. Feng P, Lampel KAA, Karch H, Whittam TSS. 1998. Genotypic and Phenotypic Changes in the Emergence of *Escherichia coli* O157:H7. *J Infect Dis* 177:1750–1753. doi: 10.1086/517438.
54. Tenover FC, Arbeit RD, Goering R V, Mickelsen PA, Murray BE, Persing DH, Swaminathan B. 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. *J Clin Microbiol* 33:2233–2239.

CHAPTER 5

GENETIC FACTORS OF SHIGA TOXIN-PRODUCING *ESCHERICHIA COLI* (STEC) ASSOCIATED WITH PERSISTENCE AND BIOFILM FORMATION IN BEEF CATTLE FARMS

ABSTRACT

Shiga toxin-producing *Escherichia coli* (STEC) is a foodborne pathogen that is characterized by the presence of bloody diarrhea that can lead to hemolytic uremic syndrome (HUS) or kidney failure in some cases. Cattle can be asymptotically colonized with STEC and are considered a main reservoir. Biofilm formation may result in the persistence of STEC in the farm environment, though few studies have examined whether biofilms enhance the ability of some strains to persist. Indeed, the diversity and genetic variation of strains from the cattle reservoir and the farming environment and the ability to persist in this environment may contribute to the emergence of new strains in the clinical setting. In all, 26 cattle were sampled at four time points over a period of three months. Thirteen cows (50.0%) were positive for STEC at multiple phases. Among these 26 animals, 66 STEC isolates were recovered and the level of biofilm production was determined using crystal violet assays. A total of seven typeable serogroups were identified including serogroup O157 and two big six non-O157 serogroups, O26 and O103, with serogroup O6 predominating in 73.1% of the 26 animals at any given time point. The *stx2a* gene that encodes Shiga toxin 2, was present in 77.8% (n=49) of the isolates and the highest biofilm levels were observed for strains belonging to serogroup O6, which was the only serogroup that persistently colonized multiple cows throughout the study. This longitudinal study will help understand the genetic diversity of isolates in a beef herd and to better understand the role that biofilm formation may play in serogroup persistence.

INTRODUCTION

Shiga toxin-producing *Escherichia coli* (STEC) is a foodborne pathogen that can cause a wide range of disease outcomes from hemorrhagic colitis to hemolytic uremic syndrome (HUS) (1–3). Since the first STEC O157:H7 outbreak in 1982, numerous O157 and non-O157 STEC outbreaks have also been linked to cattle products such as beef, milk, and cheese (4–8). In the US, six non-O157 serogroups, O26, O45, O103, O111, O121, and O145, denoted the big six non-O157, predominate in human infections and have been associated with multiple outbreaks (8). Although cattle have been implicated as an important reservoir for STEC, other ruminants and farm animals have also been identified to harbor STEC (9–13). While consumption of contaminated food is the main transmission mode for STEC; occupational and recreational contact with cattle has been identified as a risk factor for STEC infections (14–16). The high prevalence of STEC in the cattle farming environment supports the need to understand the environmental niche that is occupied in order to develop programs and measures aimed at minimizing the risk of transmission to humans.

Environmental factors and farm practices may play a role in the prevalence of STEC within a farm regardless of the serotypes and genetic composition of the strains. Warmer temperatures in the summer months have been linked to a higher prevalence of STEC which varies considerably across herds and geographic locations (17, 18). Indeed, STEC prevalence ranges from 44.4% and 4.1-10.5% in beef feedlots and 12.6% and 9.2-18.3% in dairy farms in South Korea and Nebraska, respectively (19, 20). A prior study in Michigan also showed that beef feedlot farms had a higher prevalence of STEC compared to dairy farms, which was contrary to a study performed at Washington State (17, 21). Additionally, a study on Midwestern

cattle farms determined that non-O157 serogroups (19.3%) were more prevalent than O157 (12.9%), which was also confirmed in the prior Michigan study (17, 18).

Variation in prevalence estimates among cattle derived non-O157 and O157 STEC in prior studies as well as the identification of hundreds of serogroups suggests that there is a considerable degree of genetic variation among isolates recovered in these environments. Genetic variation also contributes to differences in STEC phenotypes including biofilm formation, which is an important survival mechanism against antibiotics, bacteriophages, and environmental stressors. Biofilm formation in STEC has been hypothesized to contribute to persistence and has been shown to help STEC survive in cattle water troughs (22). Similarly, the ability to persist within a biofilm, allows for a higher likelihood of gene transfer to occur within strains (23, 24). STEC can survive in the farming environment as well as on food processing equipment, indicating that biofilms may play an important role in the ability of STEC to persist and impact both human health and the food industry (25–27).

To form a biofilm, STEC was shown to utilize different classes of fimbriae including, type 1 fimbriae, curli, and type 4 pili, to initiate attachment to surfaces (28, 29), while surface proteins are needed for biofilm maturation. Specifically, lipopolysaccharide (LPS), a glycolipid that is abundant on the surface of STEC, has been shown to aid in the formation of biofilms (30, 31). The production of the exopolysaccharide matrix is necessary for further maturation of the biofilm (32). Although it is logical to assume that biofilm formation may play a role in STEC persistence in the farm environment, few studies have examined the ability of different serotypes to form biofilms or have determined the role they play within a given environment or reservoir.

Cattle have been previously implicated as a reservoir for STEC in a number of previous studies; however, little research on the transmission and persistence of strains within this

reservoir has been performed (17, 33). Examining STEC isolates from cattle collected at multiple time points will enable identification of virulence characteristics that are important for persistence within a herd. Prior studies have identified interesting epidemiological associations with cattle-derived STEC isolates, including a high prevalence of *stx2*, which encodes Shiga toxin 2 and has been linked to more severe clinical outcomes (17, 34–36). Investigating persistence is important not only because cattle are a major reservoir of STEC, but also because persistence and constant recontamination of animals can result in the emergence of virulent or antibiotic-resistant strains due to the increased risk of horizontal gene transfer from STEC and other bacterial populations. Overall, this study will examine the genotypic and phenotypic profiles of strains to give insight on the persistence of certain serotypes in the farm environment and identify genetic similarities between bovine and clinical isolates.

METHODS

Bacterial strains

A total of 66 STEC isolates, which were recovered in 2012 from a Michigan beef farm that was previously found to have a high prevalence of STEC over multiple samplings (17). The isolates used in this study are a subset of the original isolates reported, due to exclusion of strains that changed *stx* profiles from bacteriophage loss or inability to recover isolates from fecal samples. A total 26 cattle were sampled over four sampling periods (1-4), which were roughly 3-4 weeks apart. Multiplex PCR was performed on the isolates to verify that they were *stx1* and/or *stx2* positive using primers and cycling conditions previously published (37). Approval to conduct the study was obtained by the Michigan State University Institutional Animal Care and Use Committee (AN12/10-223-00).

Biofilm assays

Isolates were grown in Luria-Bertani (BD Diagnostics) broth and incubated at 37°C overnight with shaking. Overnight cultures were diluted 1:10 in prewarmed Luria-Bertani no salt (LB-NS) broth before 30µl were plated in a 96-well microtiter plate (TPP, Techno Plastic Products AG) with 100µl LB-NS broth and incubated at 25°C for 48 hours. The 96-well plate was washed thrice with phosphate buffered saline (PBS) to remove any unattached cells that had not formed a biofilm. Biofilms were fixed by adding 100% methanol and incubating for ten minutes. Following removal of the methanol and air drying, the biofilms were stained with crystal violet (CV) and incubated for 15min at room temperature. PBS was used to wash the biofilms and remove any CV that had not been absorbed by the biofilm. After air drying, 200µl of 33% glacial acetic acid was used to solubilize the CV and the absorbance (A_{595}) was

determined using a plate reader. The data was normalized to the blank control of LB-NS broth for each biological replicate. At least four technical replicates were averaged on a plate and repeated for at least three biological replicates.

DNA isolation and whole genome sequencing (WGS)

DNA was extracted from 63 *E. coli* isolates using the Qiagen DNeasy Kit (Qiagen, Valencia, CA, USA) and was confirmed to be STEC using a multiplex PCR assay targeting *stx1*, *stx2* and *eae* using primers: *eae*_FP 5' TCAATGCAGTTCGGTTATCAGTT 3' *eae*_RP 5' GTAAAGTCCGTTACCCCAACCTG 3'; *stx1*_FP 5' CGATGTTACGGTTTGTTACTG TGACAGC 3' *stx1*_RP 5' AATGCCACGCTTCCCAGAATTG 3'; and *stx2*_FP 5' GTTTTG ACCATCTTCGTCTGATTATTGAG 3' *stx2*_RP 5' AGCGTAAGGCTTCTGCTGTGAC 3'. Following an initial denaturation step at 95°C for 10 min, cycling conditions included 30 cycles of 95°C for 15s, 65°C for 15s and 72°C for 30s, ending with 72°C for 3min. Sequencing libraries were prepared for all 63 *stx*-positive isolates using the Nextera XT library prep kit (Illumina, San Diego, CA, USA) and sequenced at 2x250 bp on the Illumina MiSeq platform (Illumina, San Diego, CA, USA) at the Michigan State University Research Technology Support Facility (RTSF) and the Michigan Department of Agriculture and Rural Development. All sequencing reads used in the analysis were deposited in the NCBI SRA database and the associated accession numbers are pending.

Bioinformatic analysis

Sequences were preprocessed with Trimmomatic to remove sequencing adapters, sequences <100 nucleotides, and reads with a phred quality score less than 20 (Q20) (38).

FastQC was used to quality check the reads prior to assembling and analysis (39). Assemblies were generated with Spades 3.10.1 using kmers 21, 33, 55, 77, 99 and 127 (40).

Molecular serotyping and identification of Shiga toxin gene profiles were performed using Abricate (<https://github.com/tseeman/abricate>) and databases generated by the Center for Genomic Epidemiology (www.genomicepidemiology.org) for *wzy/wzx* (O-antigen), *fliC* (H-antigen) and *stx* genes. In house bioinformatic scripts were developed using a Basic Local Alignment Search Tool (NCBI-BLAST) to extract the seven housekeeping genes commonly used for multilocus sequence typing (MLST). Sequence types (STs) were classified using EcMLST v1.2 database (<http://www.shigatox.net>) (41), while pan genome analysis was performed using a pipeline developed by Oakeson *et al* (42). Briefly, assemblies were annotated using Prokka v1.14.0 and the pan genomes were extracted from annotated genomes using Roary v3.11.2 (43, 44). Lyve-SET was utilized for a subset of strains to generate high quality single nucleotide polymorphisms (hqSNPs) to further examine strains of the same serogroups and to determine whether identical strains were persisting in cattle using parameters for STEC (45). All bioinformatic pipelines are available upon request.

Data analysis

Concatenated MLST alleles were aligned with CLUSTALW and the phylogeny was constructed using the Neighbor-joining algorithm with 1000 bootstrap replication in MEGA X (46). RAxML v8 was used to infer maximum likelihood phylogenies for both SNP profiles determined by Lyve-SET and pan genome concatenated gene sequences generated by Roary (47). All phylogenies were visualized using MEGA X, TreeGraph2 or FigTree (<http://tree.bio.ed.ac.uk/software/figtree>) (46, 48).

RESULTS

Herd demographics and prevalence of STEC

A total of 54 cows were sampled four times over a period of 3 months and 26 (48.1%) cows were identified as positive for STEC in at least one sampling period. In all, 19 (35.2%) cows were positive for at least one STEC isolate in the first sampling period, while an additional seven (13.0%) cows acquired STEC by sampling periods 2 or 3. In addition, a subset of nine (16.7%) animals had more than one distinct isolate available for characterization at a given sampling. The prevalence of STEC varied across the sampling periods; the first sampling had the highest colonization rate (n=19, 35.2%) and the second sampling had the lowest (n=8, 14.8%). Among all four samplings, a total of 66 STEC isolates were recovered. The number of STEC isolates obtained was similar across samplings with an average of 16.5 isolates per period. Twenty isolates were recovered during the first sampling period followed by 16, 19, and 11 isolates during samplings 2-4, respectively.

Among all 66 isolates, seven typeable serogroups were identified including O157 (n=1) and two big six non-O157 serogroups, O26 and O103 (Table 5.1). One isolate was classified as non-typeable (NT) due to incomplete or missing *wzy/wzx* genes, which were used for O-antigen typing. In addition, three NT isolates lacking WGS data were removed from the analysis. Overall, the most common serogroup was O6 (n=36; 57.1%) followed by O26 (n=11; 17.5%) and O168 (n=8; 12.7%). The virulence gene distribution also varied by serogroup. Among the 63 isolates, *stx2* was identified in 77.8% (n=49) of the isolates, with serogroups O6 (n=36) and O168 (n=1) harboring the *stx2c* variant. The remaining O168 isolates (n=7) as well as the O8 (n=1), O185 (n=1) and O157 (n=1) isolates contained the *stx2a* variant. The *stx* variant could not be determined for two isolates (NT and O103) because of the variable regions missing from the

sequencing data. Only three serogroups, O26 (n= 10, 100%, *eaeA*: beta), O103 (n= 5, 100%, *eaeA*: epsilon) and O157 (n= 1, 100%, *eaeA*: gamma) harbored the *stx1a* variant; all three had distinct *eaeA* alleles and were positive for *ehxA*, except one O103 was negative for *ehxA*.

Genetic diversity of STEC

MLST loci sequences were extracted to examine the genetic diversity of the STEC isolates and a reference-free pan genome analysis was performed for comparison. Eight STs were identified from 62 isolates; one isolate was excluded due to incomplete sequencing of MLST genes. All O6 isolates were classified as a new ST due to a SNP present in *mdh* (300, G-T), thereby generating a new *mdh* allele. This ST is denoted as ST-NEW in the analysis as the allele designation is pending.

Construction of a maximum likelihood phylogeny with bootstrapping (n=1,000) of 2,933 core, concatenated genes, which were shared across all isolates, clustered the isolates into four clades that comprised strains with the same STs and serogroups (Figure 5.1). Four additional isolates were singletons that were not included within any clades on the phylogeny. Isolates with similar virulence gene profiles also clustered together. While high bootstrap values supported the clustering of isolates, the long branch length, specifically for O157 (TW17220), indicates that there it is more distantly related to the other non-O157 STEC isolates.

Biofilm formation and persistence of STEC isolates

Static biofilm assays performed on all 66 STEC isolates resulted in a range of absorbance values from 0.15 to 5.93. Plotting the absorbance values identified a distinct break in the data at an absorbance of 2.0, which is close to the average of 2.48. Therefore, high biofilm production was classified by A_{595} values greater than 2.0 and low production was classified by A_{595} values

less than 2.0. A total of 38 (57.6%) isolates were classified as high biofilm formers, while the remaining 28 (42.4%) were classified as low biofilm formers. In all, there was a range in biofilm production across and within the serogroups (Figure 5.2). Serogroups O26, O103, and O168 predominantly were low biofilm formers, however, there was one isolate in all three serogroups (n=1, 1.5%) that were high biofilm formers. Similarly, serogroup O6 was significantly more likely to form a high biofilm relative to all other serogroups (OR: 55.0; 95% CI: 12.00, 252.18), however, three isolates (4.5%) were classified as low biofilm formers.

Any isolate with the same serogroup and virulence gene profile from a single animal found in a subsequent phase was classified as persistent (Figure 5.3). Among the 26 cows that were STEC positive, eight (30.8%) were persistently colonized by a single serogroup and profile. A total of 15 (23.8%) isolates were classified as persistent and had an isolate of the same serogroup and profile isolated at a later phase. High biofilm formers were significantly more likely to persistently colonize animals within the herd (Fishers $\chi^2 p = 0.016$). Indeed, 13 of the 38 (34.2%) high biofilm formers were recovered from the same animal over the sampling periods compared to only 2 of the 28 (7.4%) low biofilm formers. This association is primarily driven by serogroup O6, which was found in two or more phases in seven of the cows that were positive STEC (26.9%); all NT isolates were excluded from this analysis. Three cows that were negative for O6 at the first sampling acquired a persistent, high biofilm forming isolate. Cow 760 was persistently colonized with O168 at three consecutive samplings and was the only cow with persistence that did not belong to serogroup O6. Serogroup O6 and O168 were isolated from all the samplings and in 19 (73.1%) and 5 (19.2%) animals respectively, while serogroup O26 was isolated from three of the samplings in ten different animals.

Longitudinal examination of related isolates

The hqSNP analysis was performed to determine the genetic relatedness of the most closely related STEC isolates within the herd. Serogroup O103 was examined due to being isolated from multiple cows sampled at one time point. A phylogenetic tree, however, could not be generated due to a lack of informative sites that met the confidence standards (95% of reads agree with SNP designation and 20 read depth at a site). Similarly, O26 was found among 10 animals within three sampling periods even though it never persisted within a single cow (Figure 5.4). Interestingly, SNP analysis demonstrated that all O26 isolates were closely related differing by only 0-8 SNPs and that there was no distinct clustering of isolates. The O26 isolate, TW17255, was removed from this analysis due to low quality sequencing and lack of supported informative sites.

Serogroups O6 and O168 were also selected for hqSNP analysis due to the high frequency of persistence observed among these isolates. Hence, the goal of this analysis was to determine whether one strain of each serogroup was present throughout the sampling period or if there were subtle changes (e.g., SNPs) within the serogroups that may indicate evolutionary events. All O168 serogroup isolates that were isolated from the same cow (760) clustered with only 0-2 SNP differences, while all 8 isolates recovered over the four samplings were related differing by 0-18 SNPs (Figure 5.5). One O168 isolate, TW19599, that clustered with 0-2 SNPs was classified as a high-biofilm former. A direct comparison was performed to identify SNP differences in specific genes that may account for the variation in biofilm formation compared to all other isolates from the same cow. Three SNPs in the *rsxC* gene, an electron transport subunit, differed in TW19599 relative to all other isolates.

For serogroup O6 isolates, which were highly prevalent throughout the herd and persisted in nineteen cows, SNP analysis was performed on 36 isolates except TW17231 and TW17186 due to poor sequencing quality (Figure 5.6). One cluster, which comprised isolates with 0-1 SNP differences, originated from ten cows, suggesting transmission among the herd. Two O6 isolates, TW19669 and TW19601, were low biofilm formers, contrary to the remaining 34 O6 isolates. A direct comparison of TW19669 with the other three isolates from the same cow (761) identified two genes, a putative outer membrane autotransporter barrel (*icsA*) and a type VI secretion protein (*vgrG*), that had SNP differences across all isolates compared to TW19669. Similarly, TW19601 when compared to all other isolates from the same cow (763) that formed high biofilms had mutations in five genes of interest. These include: putative D-alanyl-D-alanine endopeptidase (F7D04_07630), putative paraquat-inducible protein A (FPV29_09865), DNA repair (*radC*), putative phage tail protein, and type VI secretion protein (*vgrG*). Extraction of the *vgrG* gene from all isolates identified that multiple copies of the gene are present within the genome. Any SNPs identified within an isolate were split in some instances with SNPs in two of the *vgrG* genes but not in the third copy.

DISCUSSION

Environmental reservoirs are important to the pathogenic cycle for STEC to share horizontally acquired genes and disseminate throughout the environment through carriage by asymptomatic animals. The main reservoir implicated in high prevalence rates of STEC and multiple foodborne outbreaks are cattle (10, 17, 49–51). Lack of the globotriaosylceramide (Gb3) receptor allows for colonization and shedding of STEC by cattle, resulting in contaminated beef and dairy products while creating the potential for contamination of downstream crops and water sources from fecal runoff into water systems (52, 53). Examination of the genetic diversity and prevalence of serogroups within a herd provides a better understanding of the bacterial population that is present in this niche. More importantly, the longitudinal study of animals within a single herd provides insight into the transmission and persistence of strains within the herd and enhances the understanding of strain types that more effectively colonize the cow and exist in the farm environment.

Our previous study in Michigan identified the prevalence of STEC in 12 cattle herds to be variable ranging from 10.9-53.7% with beef herds having higher prevalence rates than dairy farms (17). This study utilized the Michigan beef herd with the highest STEC prevalence rate and identified a range of non-O157 serogroups including those representing both the big-six non-O157 and other serogroups. Only one isolate was typed as O157. This low prevalence of O157 within beef herds is consistent with other cattle studies and surveys of farm animals such as pigs (54–56). When this study was being conducted in 2012, the incidence of clinical infections caused by non-O157 STEC had surpassed the incidence of O157 infections in Michigan and other locations throughout the US (57, 58). Calculating prevalence rates within cattle herds,

however, is difficult due to variation in isolation methods and the inability to recover all *stx*-positive isolates via culture.

Among STEC, various *stx* subtypes can cause varying degrees of cytotoxicity and differences in clinical outcomes (52, 59). Subtypes *stx2a*, *stx2c*, and *stx2d*, for example, has been shown to exhibit stronger Gb3 receptor binding *in vivo* when compared to *stx1* subtypes (60). All isolates typed as other non-O157 serogroups were *stx2a* or *stx2c* positive and *eaeA* negative, which is important because prior studies have linked *stx2a* and *stx2c* variants to severe clinical outcomes (35, 36, 61). Surprisingly, these isolates were all negative for *eaeA*, a critical STEC virulence factor that when found with *stx2*, is associated with enhanced virulence (62–64). Conversely, the O103 and O26 isolates were all positive for *stx1*, *eaeA*, and *ehxA* except for a single O103 isolate profile that was positive for both *stx1* and *stx2*. While these serogroups lacked the more virulent *stx2* gene, the presence of the other virulence factors in addition to *stx1* has been isolated from clinical cases, further supporting the crossover potential between the two environments.

Although identification of *stx* subtypes is important for estimating disease potential, other genetic factors have also been shown to play a role in disease severity. For example, some isolates have been obtained from patients presenting with diarrhea that were *stx*-negative but positive for *eaeA* (intimin) and *ehxA* (enterohemolysin) (65), though the Stx bacteriophages encoding the Shiga toxins are occasionally lost *in vivo* (66). The presence of *eaeA* and *ehxA* in addition to various subtypes may influence the ability of a strain to effectively colonize ruminants and cause clinical outcomes (67–69). While these three virulence factors are important for the presentation of clinical outcomes in humans, various other factors as well as phenotypic

differences may play a role in the ability of a strain to persistently colonize cattle and the environment.

The ability of a strain to form a biofilm in the cattle environment allows for an increased potential for transmission. STEC isolates have been isolated from water troughs resulting in the continual reinfection of the cattle as well as processing equipment, causing contamination of beef products (22, 70). In this study, serogroup O6 was the predominant profile isolated from multiple cows and multiple samplings over the course of the study. Consistent with a prior study by Barth *et al.* that examined persistent and transient STEC colonization within herds in Germany, a high prevalence of O6 was observed. The German O6:H49 isolate, however, was a transient colonizer unlike the O6:H34 observed in our study, which persisted across all samplings and in multiple animals (71). Future comparative genome analyses are needed within serogroups that exhibit differences in colonization to elucidate genomic factors that are important for persistent colonization. One phenotypic factor, biofilm production, appears to be important for O6 STEC from beef cattle in our study as most isolates belonging to this serogroup were high biofilm producers. The remainder of isolates from other serogroups predominantly exhibited low biofilm formation.

To control for variation in our biofilm assays, temperature was kept steady to model the air temperature (25C, 77F) that would have been present on the farms during the time of each sampling. Indeed, decreasing temperature has been shown to reduce the ability of an isolate to form a biofilm (72). While significant for meat processing plants that have control over the environment, the ability to form a biofilm is strain-dependent and temperature may not influence biofilm formation equally among all isolates. Other physiological factors may also play a role in the ability of a strain to form a biofilm. Extracellular structures formed by the bacteria are thought to help with biofilm formation, for example, as studies have shown that the production

of curli is variable even with the gene present and it does not correlate the biofilm formation (73). Similarly, other adhesion factors such as fimbriae and autotransporters, are associated with the ability of a strain to form a biofilm (74); these factors were not evaluated in our study but future studies could focus on extracting genes important for biofilm production for characterization and comparative studies.

While serogroup O6 persisted throughout the study, serogroup O103 was present only at the second sampling before being cleared from the environment. By contrast, serogroup O26 was present throughout the samplings but never from the same cow. Transient colonization by these two serogroups may indicate that they are not drivers of horizontal gene transfer but may evolve and acquire virulence genes in the short time that they are present within the environment. Serogroup O168 was the only other serogroup besides O6 that was found to persist and O168 isolates were predominantly classified as low biofilm formers. The persistent O6 colonization may have been due to differences in the gut microbiota of the cow, which did not allow it to clear the O168 colonization (75). While changes in the microbiome of cattle have been associated with age, there are other environmental factors that could also play a role (76, 77).

Genomic analyses allowed for a within serogroup and within cow analysis between high and low biofilm formers to identify differences that may influence biofilm formation by isolate. The *rsxC* gene, which was identified in serogroup O168 isolates, is part of a membrane associated complex that interacts with *soxRS*, a superoxide response regulon (78, 79). A single SNP was identified in all three low biofilm forming O168 isolates when compared to the isolate with high biofilm production. Mutagenesis studies have shown that deletion of genes within the *rsx* gene cluster, including *rsxC*, will result in constitutive *soxS* activation to protect against

reactive oxygen species (ROS) (78, 80). Within a biofilm, ROS accumulate, and microorganisms need to have mechanisms to combat the oxidative stress and neutralize any ROS (81).

A similar analysis, performed with the O6 serogroups, identified heterogenous SNPs in *vgrG*, a multi copy type VI secretion system, between high and low biofilm formation. The *vgrG* gene is widely distributed within gram negative organisms (82). Similar to a bacteriophage tail spike protein, VgrG is secreted from the cell and essential for the secretion of virulence factors into other cells (83). Studies in *Acidovorax citrulli* showed that *vgrG* mutants exhibited a decreased biofilm formation, whereas *Acinetobacter baumannii vgrG* mutants were not affected in the ability to form a biofilm, but instead had reduced attachment to epithelial cells (84, 85). The SNP that was identified across isolates changes an amino acid from Tyrosine (aromatic) to a Histidine (basic). Potential differences in expression of one *vgrG* variant over another within a cell may result in a more efficient type VI secretion system that may influence the ability of a cell to form a biofilm. A larger sample of isolates with shared profiles and differing biofilms would provide more insight into genes of interest that may differ across the high and low biofilm former populations.

Understanding the genetic profiles and the persistent colonization of STEC within the cattle environment will allow for the development and implementation of targeted practices to minimize STEC prevalence within the cattle environment. Genomic analysis to identify transmission networks among the animals with the same serogroup may identify areas of importance to minimize transmission. Subsequently, minimizing the diversity of isolates that are present in the cattle environment will minimize the risk of new zoonotic pathogens emerging from evolution and gain of virulence factors in the environment.

APPENDIX

Table 5.1. Serogroups and virulence gene profiles among 63 cattle derived non-O157 STEC isolates.

Serogroup	Total (%)*	<i>stx</i> profile	<i>eaeA</i> profile	<i>ehxA</i> profile
O6	36 (57.1%)	2c	negative	negative
O8	1 (1.6%)	2a	negative	A
O168	8 (12.7%)	2a	negative	negative
O185	1 (1.6%)	2a	negative	negative
O26	10 (15.9%)	1a	beta	C
O103	5 (7.9%)	1a (n=4), 1a,2 (n=1)	epsilon	C (n=4)**
O157	1 (1.6%)	1a,2a	gamma	B
NT	1 (1.6%)	2	negative	A

*Whole genome sequencing data was only available for 63 of the 66 STEC isolates.

**One O103 *stx1a* isolate lacked *ehxA*

Figure 5.2. Frequency of STEC serogroups stratified by the level of biofilm production in 66 isolates recovered from cattle.

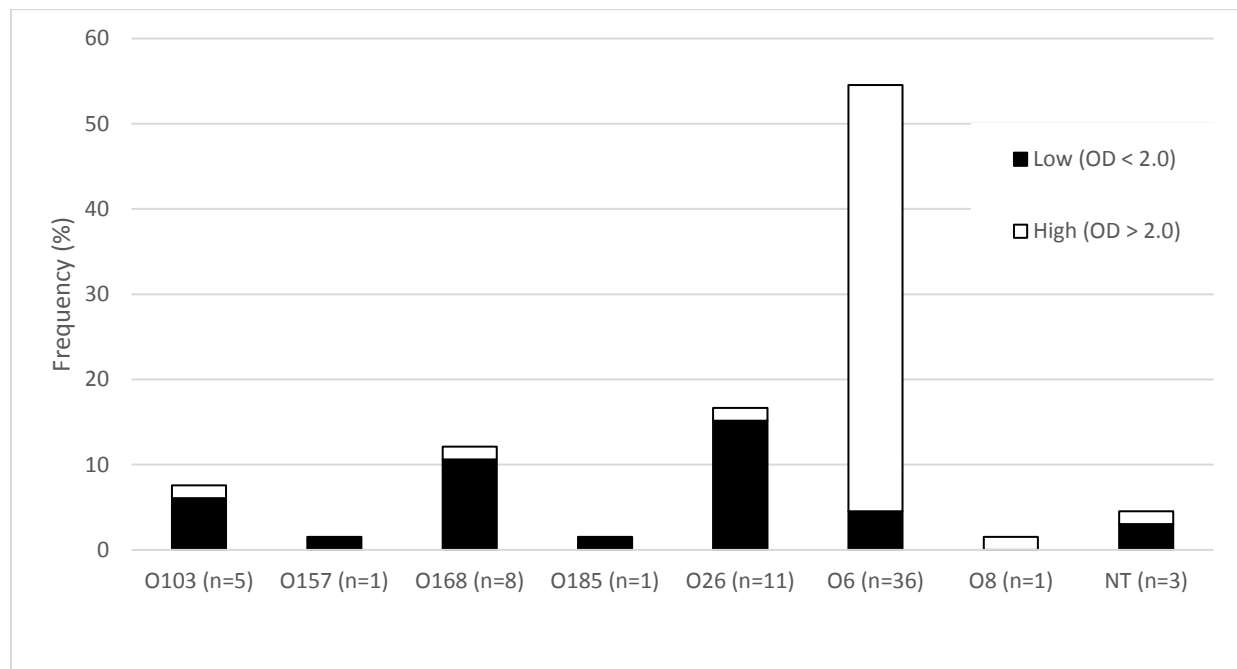


Figure 5.3. Longitudinal overview of STEC isolates by cow, sampling period, serogroup and strength of biofilm.

Cow ID	1		2				3				4	
728	O168											
729	O168											
731	O8											
735	O6											
740	O6											
742	O6											
751							O6				O26	
752	O6		O103*	O6	O103*	O6	O6				O6	
753			O103	O6			O26					
754	NT											
755	O6											
756	O6											
758			O6				O26					
759							O168				O6	
760	O168		O168	O168			O168					
761	O6						O26	O157	O6	O6	O6	
763			O6	NT	O6						O6	
764	O6		O6				O6	O26			O6	
765	O6											
767			O6	O103			O6	O6	O26			
768	O6		O6				O26	O26			O6	
770	O6											
771	O185						O26				O168	O6
773							O6				O26	O6
775	NT											
776	O103	O26										

High (OD > 2.0)
 Low (OD < 2.0)

*Profiles of O103 isolates differed within the same animal: *stx1a2*, *eae-epsilon*, *ehxA-C* and *stx1a*, *eae-epsilon*, *ehxA-C*

Figure 5.4. High quality SNP (hqSNP) analysis of ten O26 STEC isolates recovered from ten cattle at different sampling periods.

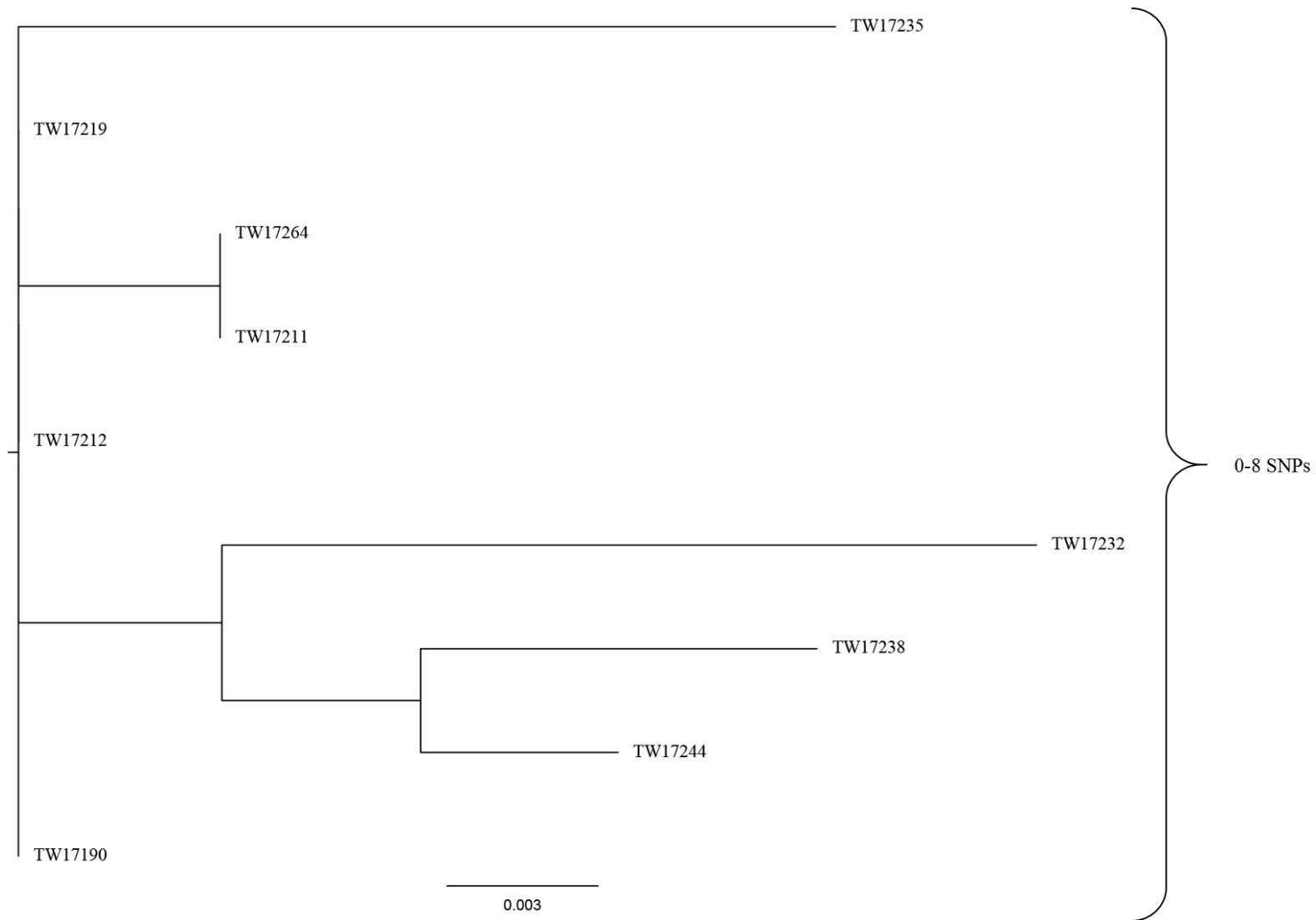


Figure 5.5. High quality SNP (hqSNP) analysis of eight O168 STEC isolates recovered from cattle over multiple samplings. All isolates recovered from the same cow (760, blue boxes) are indicated as well as those isolates with high levels of biofilm production (up triangle). All other O168 isolates came from other cattle at varying sampling points.

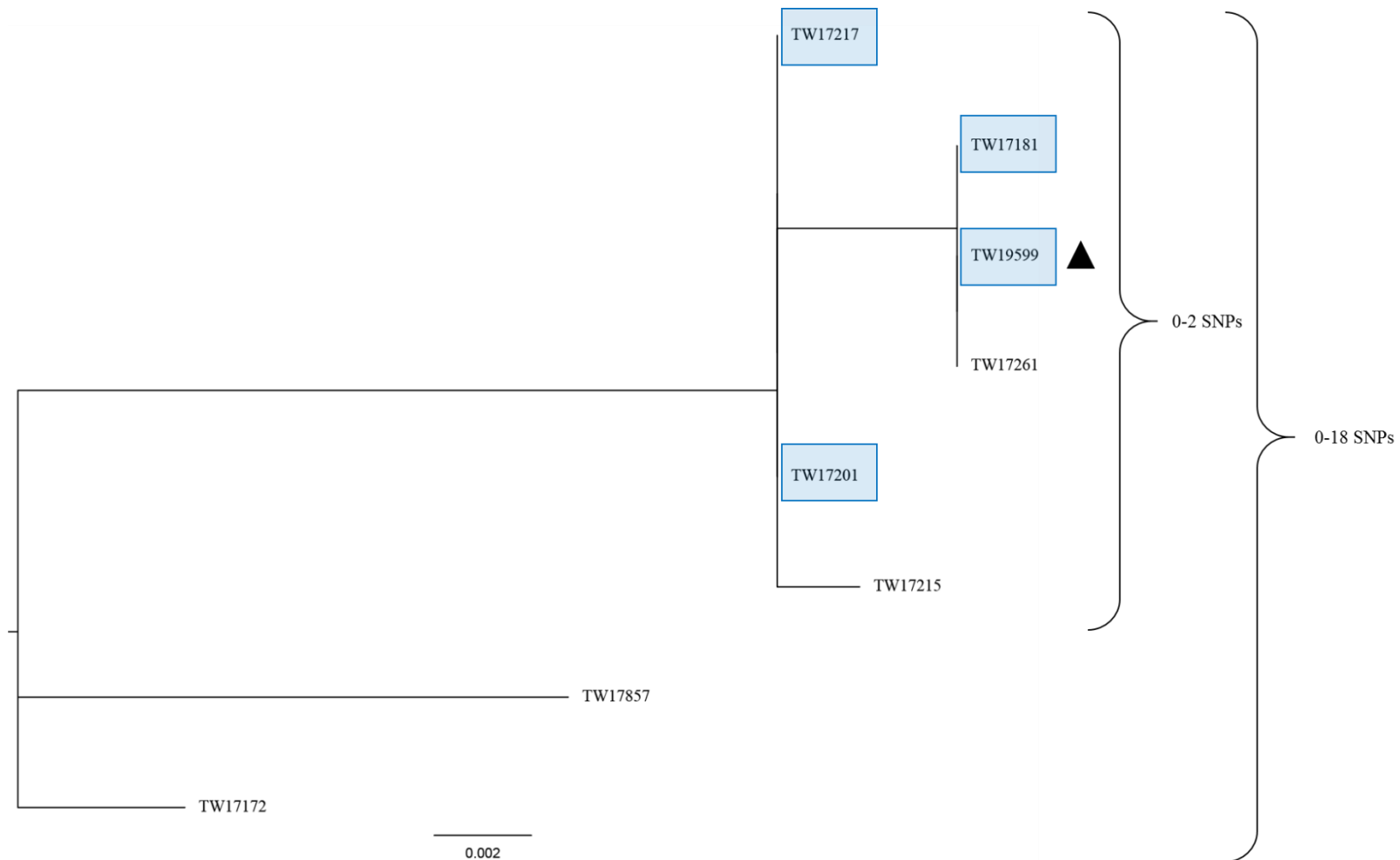
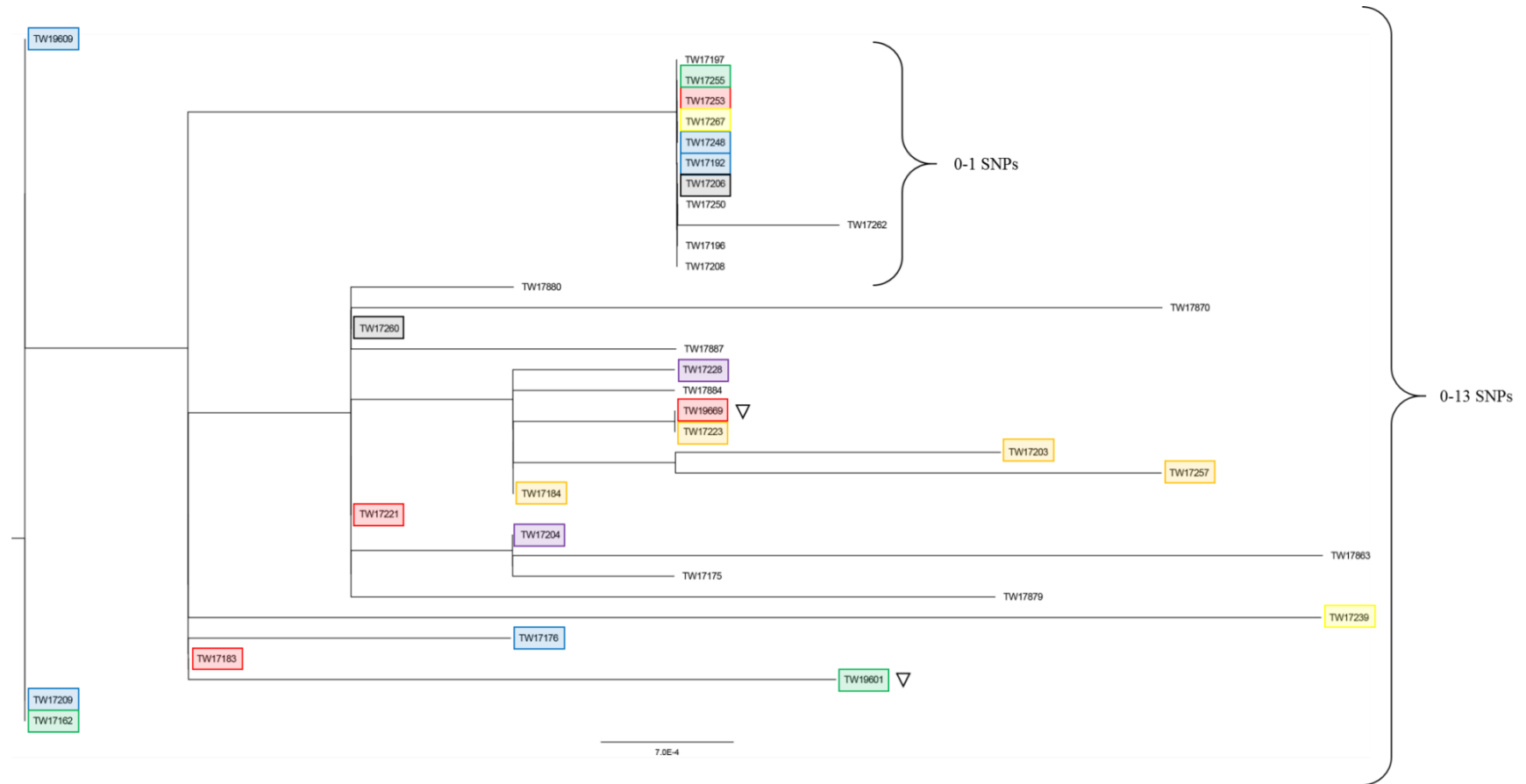


Figure 5.6. High quality (hqSNP) analysis of 36 O6 STEC isolates identified from cattle in the study. Low biofilm formers (inverted triangle) and isolates from the same cow (752, blue; 761, red; 763, green; 764, orange; 767, purple; 768, grey; 773, yellow) are denoted with similar shading.



REFERENCES

REFERENCES

1. Karmali MA, Petric M, Lim C, McKeough PC, Arbus GS, Lior H. 1985. The association between idiopathic hemolytic uremic syndrome and infection by verotoxin-producing *Escherichia coli*. *J Infect Dis* 151:775–782. doi: 10.1093/infdis/151.5.775.
2. Gault G, Weill FX, Mariani-Kurkdjian P, Jourdan-da Silva N, King L, Aldabe B, Charron M, Ong N, Castor C, Macé M, Bingen E, Noël H, Vaillant V, Bone A, Vendrely B, Delmas Y, Combe C, Bercion R, D’Andigné E, Desjardin M, Rolland P, de Valk H. 2011. Outbreak of haemolytic uraemic syndrome and bloody diarrhoea due to *Escherichia coli* O104:H4, south-west France, June 2011. *Eurosurveillance* 16:19905. doi: 10.2807/ese.16.26.19905-en.
3. Karch H, Tarr PI, Bielaszewska M. 2005. Enterohaemorrhagic *Escherichia coli* in human medicine. *Int J Med Microbiol* 295:405–418. doi: 10.1016/j.ijmm.2005.06.009.
4. Rangel JM, Sparling PH, Crowe C, Griffin PM, Swerdlow DL. 2005. Epidemiology of *Escherichia coli* O157:H7 outbreaks, United States, 1982–2002. *Emerg Infect Dis* 11:603–609. doi: 10.3201/eid1104.040739.
5. O’Brien AD, Newland JW, Miller SF, Holmes RK, Smith HW, Formal SB. 1984. Shiga-like toxin-converting phages from *Escherichia coli* strains that cause hemorrhagic colitis or infantile diarrhea. *Science* 226:694–696. doi: 10.1126/science.6387911.
6. Wells JG, Shipman LD, Greene KD, Sowers EG, Green JH, Cameron DN, Downes FP, Martin ML, Griffin PM, Ostroff SM. 1991. Isolation of *Escherichia coli* serotype O157:H7 and other Shiga-like-toxin-producing *E. coli* from dairy cattle. *J Clin Microbiol* 29:985–989.
7. Hainstock L, Donovan D. 2017. The Cheese Stood Alone. 17th Annu Michigan Commun Dis Conf.
8. Brooks JT, Sowers EG, Wells JG, Greene KD, Griffin PM, Hoekstra RM, Strockbine NA. 2005. Non-O157 Shiga Toxin–Producing *Escherichia coli* Infections in the United States, 1983–2002. *J Infect Dis* 192:1422–1429. doi: 10.1086/466536.
9. Blanco JE, Blanco M, Alonso MP, Mora A, Dahbi G, Coira MA, Blanco J. 2004. Serotypes, Virulence Genes, and Intimin Types of Shiga Toxin (Verotoxin)-Producing *Escherichia coli* Isolates from Human Patients: Prevalence in Lugo, Spain, from 1992 through 1999. *J Clin Microbiol* 42:311–319. doi: 10.1128/JCM.42.1.311-319.2004.
10. Marouani-Gadri N, Augier G, Carpentier B. 2009. Characterization of bacterial strains isolated from a beef-processing plant following cleaning and disinfection - Influence of isolated strains on biofilm formation by Sakai and EDL 933 *E. coli* O157:H7. *Int J Food Microbiol* 133:62–67. doi: 10.1016/j.ijfoodmicro.2009.04.028.

11. Kassenborg HD, Hedberg CW, Hoekstra M, Evans MC, Chin AE, Marcus R, Vugia DJ, Smith K, Ahuja SD, Slutsker L, Griffin PM. 2004. Farm visits and undercooked hamburgers as major risk factors for sporadic *Escherichia coli* O157:H7 infection: data from a case-control study in 5 FoodNet sites. *Clin Infect Dis* 38:S271-8. doi: 10.1086/381596.
12. Zschock M, Hamann HP, Kloppert B, Wolter W. 2000. Shiga-toxin-producing *Escherichia coli* in faeces of healthy dairy cows, sheep and goats: Prevalence and virulence properties. *Lett Appl Microbiol* 31:203–208. doi: 10.1046/j.1365-2672.2000.00789.x.
13. Grauke LJ, Kudva IT, Yoon JW, Hunt CW, Williams CJ, Hovde CJ. 2002. Gastrointestinal tract location of *Escherichia coli* O157:H7 in ruminants. *Appl Environ Microbiol* 68:2269–2277. doi: 10.1128/AEM.68.5.2269-2277.2002.
14. Schlager S, Lepuschitz S, Ruppitsch W, Ableitner O, Pietzka A, Neubauer S, Stöger A, Lassnig H, Mikula C, Springer B, Allerberger F. 2018. Petting zoos as sources of Shiga toxin-producing *Escherichia coli* (STEC) infections. *Int J Med Microbiol* 308:927–932. doi: 10.1016/J.IJMM.2018.06.008.
15. Rivas M, Chinen I, Miliwebsky E, Masana M. 2014. Risk Factors for Shiga Toxin-Producing *Escherichia coli*-Associated Human Diseases. *Microbiol Spectr* 2:1–14. doi: 10.1128/microbiolspec.ehec-0002-2013.
16. O'Brien SJ, Adak GK, Gilham C. 2001. Contact with farming environment as a major risk factor for Shiga toxin (Vero cytotoxin)-producing *Escherichia coli* O157 infection in humans. *Emerg Infect Dis* 7:1049–1051. doi: 10.3201/eid0706.010626.
17. Venegas-Vargas C, Henderson S, Khare A, Mosci RE, Lehnert JD, Singh P, Ouellette LM, Norby B, Funk JA, Rust S, Bartlett PC, Grooms D, Manning SD. 2016. Factors associated with Shiga toxin-producing *Escherichia coli* shedding by dairy and beef cattle. *Appl Environ Microbiol* 82:5049–5056. doi: 10.1128/AEM.00829-16.
18. Barkocy-Gallagher GA, Arthur TM, Rivera-Betancourt M, Nou X, Shackelford SD, Wheeler TL, Koohmaraie M. 2003. Seasonal Prevalence of Shiga Toxin-Producing *Escherichia coli*, Including O157:H7 and Non-O157 Serotypes, and *Salmonella* in Commercial Beef Processing Plants. *J Food Prot* 66:1978–1986. doi: 10.4315/0362-028X-66.11.1978.
19. Renter DG, Morris JG, Sargeant JM, Hungerford LL, Berezowski J, Ngo T, Williams K, Acheson DWK. 2005. Prevalence, risk factors, O serogroups, and virulence profiles of Shiga toxin-producing bacteria from cattle production environments. *J Food Prot* 68:1556–1565. doi: 10.4315/0362-028X-68.8.1556.
20. Dong H-J, Lee S, Kim W, An J-U, Kim J, Kim D, Cho S. 2017. Prevalence, virulence potential, and pulsed-field gel electrophoresis profiling of Shiga toxin-producing *Escherichia coli* strains from cattle. *Gut Pathog* 9:22. doi: 10.1186/s13099-017-0169-x.

21. Cobbold RN, Rice DH, Szymanski M, Call DR, Hancock DD. 2004. Comparison of shiga-toxigenic *Escherichia coli* prevalences among dairy, feedlot, and cow-calf herds in Washington State. *Appl Environ Microbiol* 70:4375–4378. doi: 10.1128/AEM.70.7.4375-4378.2004.
22. LeJeune JT, Besser TE, Hancock DD. 2001. Cattle Water Troughs as Reservoirs of *Escherichia coli* O157. *Appl Environ Microbiol* 67:3053–3057. doi: 10.1128/AEM.67.7.3053-3057.2001.
23. Król JE, Nguyen HD, Rogers LM, Beyenal H, Krone SM, Top EM. 2011. Increased transfer of a multidrug resistance plasmid in *Escherichia coli* biofilms at the air-liquid interface. *Appl Environ Microbiol* 77:5079–5088. doi: 10.1128/AEM.00090-11.
24. Maeda S, Ito M, Ando T, Ishimoto Y, Fujisawa Y, Takahashi H, Matsuda A, Sawamura A, Kato S. 2006. Horizontal transfer of nonconjugative plasmids in a colony biofilm of *Escherichia coli*. *FEMS Microbiol Lett* 255:115–120. doi: 10.1111/j.1574-6968.2005.00072.x.
25. Wang R, Bono JL, Kalchayanand N, Shackelford S, Harhay DM. 2012. Biofilm formation by shiga toxin-producing *Escherichia coli* O157:H7 and non-O157 strains and their tolerance to sanitizers commonly used in the food processing environment. *J Food Prot* 75:1418–1428. doi: 10.4315/0362-028X.JFP-11-427.
26. Silagyi K, Kim S-HH, Martin Lo Y, Wei C i., Lo YM, Wei C i. 2009. Production of biofilm and quorum sensing by *Escherichia coli* O157:H7 and its transfer from contact surfaces to meat, poultry, ready-to-eat deli, and produce products. *Food Microbiol* 26:514–519. doi: 10.1016/j.fm.2009.03.004.
27. Rice EW, Johnson CH. 2000. Short communication: survival of *Escherichia coli* O157:H7 in dairy cattle drinking water. *J Dairy Sci* 83:2021–2023. doi: 10.3168/jds.S0022-0302(00)75081-8.
28. Saldaña Z, Xicohtencatl-Cortes J, Avelino F, Phillips AD, Kaper JB, Puente JL, Girón JA. 2009. Synergistic role of curli and cellulose in cell adherence and biofilm formation of attaching and effacing *Escherichia coli* and identification of Fis as a negative regulator of curli. *Environ Microbiol* 11:992–1006. doi: 10.1111/j.1462-2920.2008.01824.x.
29. Pratt LA, Kolter R. 1998. Genetic analysis of *Escherichia coli* biofilm formation: Roles of flagella, motility, chemotaxis and type I pili. *Mol Microbiol* 30:285–293. doi: 10.1046/j.1365-2958.1998.01061.x.
30. Genevaux P, Bauda P, DuBow MS, Oudega B. 1999. Identification of Tn 10 insertions in the *rfaG*, *rfaP*, and *gaLU* genes involved in lipopolysaccharide core biosynthesis that affect *Escherichia coli* adhesion. *Arch Microbiol* 172:1–8. doi: 10.1007/s002030050732.
31. Puttamreddy S, Cornick NA, Minion FC. 2010. Genome-wide transposon mutagenesis reveals a role for *pO157* genes in biofilm development in *Escherichia coli* O157:H7

- EDL933. *Infect Immun* 78:2377–2384. doi: 10.1128/IAI.00156-10.
32. Danese PN, Pratt L a, Kolter R. 2000. Exopolysaccharide Production Is Required for Development of *Escherichia coli* K-12 Biofilm Architecture Exopolysaccharide Production Is Required for Development of *Escherichia coli* K-12 Biofilm Architecture. *J Bacteriol* 182:3593–3596. doi: 10.1128/JB.182.12.3593-3596.2000.Updated.
 33. Hussein HS, Sakuma T. 2005. Invited Review: Prevalence of Shiga Toxin-Producing *Escherichia coli* in Dairy Cattle and Their Products. *J Dairy Sci* 88:450–465. doi: 10.3168/jds.S0022-0302(05)72706-5.
 34. Riordan JT, Viswanath SB, Manning SD, Whittam TS. 2008. Genetic differentiation of *Escherichia coli* O157:H7 clades associated with human disease by real-time PCR. *J Clin Microbiol* 46:2070–2073. doi: 10.1128/JCM.00203-08.
 35. Persson S, Olsen KEP, Ethelberg S, Scheutz F. 2007. Subtyping method for *Escherichia coli* Shiga toxin (Verocytotoxin) 2 variants and correlations to clinical manifestations. *J Clin Microbiol* 45:2020–2024. doi: 10.1128/JCM.02591-06.
 36. Friedrich AW, Bielaszewska M, Zhang W, Pulz M, Kuczius T, Ammon A, Karch H. 2002. *Escherichia coli* Harboring Shiga Toxin 2 Gene Variants: Frequency and Association with Clinical Symptoms . *J Infect Dis* 185:74–84. doi: 10.1086/338115.
 37. Reischl U, Youssef MT, Kilwinski J, Lehn N, Zhang WL, Karch H, Strockbine NA. 2002. Real-time fluorescence PCR assays for detection and characterization of Shiga toxin, intimin, and enterohemolysin genes from Shiga toxin-producing *Escherichia coli*. *J Clin Microbiol* 40:2555–2565. doi: 10.1128/JCM.40.7.2555-2565.2002.
 38. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. doi: 10.1093/bioinformatics/btu170.
 39. Andrews S. 2010. FASTQC, a quality control tool for the high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
 40. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V., Sirotkin A V., Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477. doi: 10.1089/cmb.2012.0021.
 41. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410. doi: 10.1016/S0022-2836(05)80360-2.
 42. Oakeson KF, Wagner JM, Mendenhall M, Rohrwasser A, Atkinson-Dunn R. 2017. Bioinformatic analyses of whole-genome sequence data in a public health laboratory. *Emerg Infect Dis* 23:1441–1445. doi: 10.3201/eid2309.170416.

43. Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069. doi: 10.1093/bioinformatics/btu153.
44. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31:3691–3693. doi: 10.1093/bioinformatics/btv421.
45. Katz LS, Griswold T, Williams-Newkirk AJ, Wagner D, Petkau A, Sieffert C, Domselaar G Van, Deng X, Carleton HA. 2017. A comparative analysis of the Lyve-SET phylogenomics pipeline for genomic epidemiology of foodborne pathogens. *Front Microbiol* 8:375. doi: 10.3389/fmicb.2017.00375.
46. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* 35:1547–1549. doi: 10.1093/molbev/msy096.
47. Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313. doi: 10.1093/bioinformatics/btu033.
48. Stöver BC, Müller KF. 2010. TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics* 11:7. doi: 10.1186/1471-2105-11-7.
49. Jay MT, Garrett V, Mohle-Boetani JC, Barros M, Farrar JA, Rios R, Abbott S, Sowadsky R, Komatsu K, Mandrell R, Sobel J, Werner SB. 2004. A Multistate Outbreak of *Escherichia coli* O157:H7 Infection Linked to Consumption of Beef Tacos at a Fast-Food Restaurant Chain. *Clin Infect Dis* 39:1–7. doi: 10.1086/421088.
50. Awadallah MA, Ahmed HA, Merwad AM, Selim MA. 2016. Occurrence, genotyping, shiga toxin genes and associated risk factors of *E. coli* isolated from dairy farms, handlers and milk consumers. *Vet J* 217:83–88. doi: 10.1016/j.tvjl.2016.09.014.
51. Ferens WA, Hovde CJ. 2011. *Escherichia coli* O157:H7: Animal reservoir and sources of human infection. *Foodborne Pathog Dis* 8:465–487. doi: 10.1089/fpd.2010.0673.
52. Melton-Celsa AR. 2014. Shiga Toxin (Stx) Classification, Structure, and Function. *Microbiol Spectr* 2:EHEC-0024-2013. doi: 10.1128/microbiolspec.ehec-0024-2013.
53. Pruimboom-Brees IM, Morgan TW, Ackermann MR, Nystrom ED, Samuel JE, Cornick NA, Moon HW. 2000. Cattle lack vascular receptors for *Escherichia coli* O157:H7 Shiga toxins. *Proc Natl Acad Sci* 97:10325–10329. doi: 10.1073/pnas.190329997.
54. Meng Q, Bai X, Zhao A, Lan R, Du H, Wang T, Shi C, Yuan X, Bai X, Ji S, Jin D, Yu B, Wang Y, Sun H, Liu K, Xu J, Xiong Y. 2014. Characterization of Shiga toxin-producing *Escherichia coli* isolated from healthy pigs in China. *BMC Microbiol* 14:1–14. doi: 10.1186/1471-2180-14-5.
55. Fan R, Shao K, Yang X, Bai X, Fu S, Sun H, Xu Y, Wang H, Li Q, Hu B, Zhang J, Xiong

- Y. 2019. High prevalence of non-O157 Shiga toxin-producing *Escherichia coli* in beef cattle detected by combining four selective agars. *BMC Microbiol* 19:213. doi: 10.1186/s12866-019-1582-8.
56. Cha W, Fratamico PM, Ruth LE, Bowman AS, Nolting JM, Manning SD, Funk JA. 2018. Prevalence and characteristics of Shiga toxin-producing *Escherichia coli* in finishing pigs: Implications on public health. *Int J Food Microbiol* 264:8–15. doi: 10.1016/j.ijfoodmicro.2017.10.017.
 57. Tseng M, Sha Q, Rudrik JT, Collins J, Henderson T, Funk JA, Manning SD. 2016. Increasing incidence of non-O157 Shiga toxin-producing *Escherichia coli* (STEC) in Michigan and association with clinical illness. *Epidemiol Infect* 144:1394–1405. doi: 10.1017/S0950268815002836.
 58. Gould LH, Mody RK, Ong KL, Clogher P, Cronquist AB, Garman KN, Lathrop S, Medus C, Spina NL, Webb TH, White PL, Wymore K, Gierke RE, Mahon BE, Griffin PM. 2013. Increased recognition of Non-O157 shiga toxin-producing *Escherichia coli* infections in the United States during 2000-2010: Epidemiologic features and comparison with *E. coli* O157 infections. *Foodborne Pathog Dis* 10:453–460. doi: 10.1089/fpd.2012.1401.
 59. Fuller CA, Pellino CA, Flagler MJ, Strasser JE, Weiss AA. 2011. Shiga toxin subtypes display dramatic differences in potency. *Infect Immun* 79:1329–1337. doi: 10.1128/IAI.01182-10.
 60. Karve SS, Weiss AA. 2014. Glycolipid binding preferences of shiga toxin variants. *PLoS One* 9:e101173. doi: 10.1371/journal.pone.0101173.
 61. Kawano K, Ono H, Iwashita O, Kurogi M, Haga T, Maeda K, Goto Y. 2012. Stx genotype and molecular epidemiological analyses of Shiga toxin-producing *Escherichia coli* O157:H7/H- in human and cattle isolates. *Eur J Clin Microbiol Infect Dis* 31:119–127. doi: 10.1007/s10096-011-1283-1.
 62. Boerlin P, McEwen SA, Boerlin-Petzold F, Wilson JB, Johnson RP, Gyles CL. 1999. Associations between virulence factors of Shiga toxin-producing *Escherichia coli* and disease in humans. *J Clin Microbiol* 37:497–503.
 63. Werber D, Fruth A, Buchholz U, Prager R, Kramer MH, Ammon A, Tschäpe H. 2003. Strong Association between Shiga Toxin-Producing *Escherichia coli* O157 and Virulence Genes *stx2* and *eae* as Possible Explanation for Predominance of Serogroup O157 in Patients with Haemolytic Uraemic Syndrome. *Eur J Clin Microbiol Infect Dis* 22:726–730. doi: 10.1007/s10096-003-1025-0.
 64. Gerber A, Karch H, Allerberger F, Verweyen HM, Zimmerhackl LB. 2002. Clinical Course and the Role of Shiga Toxin-Producing *Escherichia coli* Infection in the Hemolytic-Uremic Syndrome in Pediatric Patients, 1997–2000, in Germany and Austria: A Prospective Study. *J Infect Dis* 186:493–500. doi: 10.1086/341940.

65. Croxen MA, Finlay BB. 2010. Molecular mechanisms of *Escherichia coli* pathogenicity. *Nat Rev Microbiol* 8:26–38. doi: 10.1038/nrmicro2265.
66. Murase T, Yamai S, Watanabe H. 1999. Changes in pulsed-field gel electrophoresis patterns in clinical isolates of enterohemorrhagic *Escherichia coli* O157:H7 associated with loss of Shiga toxin genes. *Curr Microbiol* 38:48–50.
67. Dean-Nystrom EA, Bosworth BT, Moon HW, O'Brien AD. 1998. *Escherichia coli* O157:H7 requires intimin for enteropathogenicity in calves. *Infect Immun* 66:4560–4563.
68. Mundy R, Schüller S, Girard F, Fairbrother JM, Phillips AD, Frankel G. 2007. Functional studies of intimin in vivo and ex vivo: Implications for host specificity and tissue tropism. *Microbiology* 153:959–967. doi: 10.1099/mic.0.2006/003467-0.
69. Pradel N, Etienne-Mesmin L, Thévenot J, Cordonnier C, Blanquet-Diot S, Livrelli V. 2015. In vitro adhesion properties of Shiga toxin-producing *Escherichia coli* isolated from cattle, food, and humans. *Front Microbiol* 6. doi: 10.3389/fmicb.2015.00156.
70. Marouani-Gadri N, Firmesse O, Chassaing D, Sandris-Nielsen D, Arneborg N, Carpentier B. 2010. Potential of *Escherichia coli* O157:H7 to persist and form viable but non-culturable cells on a food-contact surface subjected to cycles of soiling and chemical treatment. *Int J Food Microbiol* 144:96–103. doi: 10.1016/j.ijfoodmicro.2010.09.002.
71. Barth SA, Menge C, Eichhorn I, Semmler T, Wieler LH, Pickard D, Belka A, Berens C, Geue L. 2016. The accessory genome of Shiga toxin-producing *Escherichia coli* defines a persistent colonization type in cattle. *Appl Environ Microbiol* 82:5455–5464. doi: 10.1128/AEM.00909-16.
72. Ma Z, Bumunang EW, Stanford K, Bie X, Niu YD, McAllister TA. 2019. Biofilm Formation by Shiga Toxin-Producing *Escherichia coli* on Stainless Steel Coupons as Affected by Temperature and Incubation Time. *Microorganisms* 7:95. doi: 10.3390/microorganisms7040095.
73. Chen CY, Hofmann CS, Cottrell BJ, Strobaugh TP, Paoli GC, Nguyen LH, Yan X, Uhlich GA. 2013. Phenotypic and genotypic characterization of biofilm forming capabilities in non-O157 Shiga toxin-producing *Escherichia coli* strains. *PLoS One* 8:e84863. doi: 10.1371/journal.pone.0084863.
74. Vogeeler P, Tremblay YDN, Mafu AA, Jacques M, Harel J. 2014. Life on the outside: role of biofilms in environmental persistence of Shiga-toxin producing *Escherichia coli*. *Front Microbiol* 5:317. doi: 10.3389/fmicb.2014.00317.
75. Mir RA, Weppelmann TA, Elzo M, Ahn S, Driver JD, Jeong KCC. 2016. Colonization of beef cattle by Shiga toxin producing *Escherichia coli* during the first year of life: A cohort study. *PLoS One* 11:1–16. doi: 10.1371/journal.pone.0148518.
76. Shanks OC, Kelty CA, Archibeque S, Jenkins M, Newton RJ, McLellan SL, Huse SM,

- Sogin ML. 2011. Community structures of fecal bacteria in cattle from different animal feeding operations. *Appl Environ Microbiol* 77:2992–3001. doi: 10.1128/AEM.02988-10.
77. Durso L, Wells JE, Seok M. 2013. Diversity of Microbiomes in Beef Cattle. *Encycl Metagenomics* 129–138. doi: 10.1007/978-1-4614-6418-1.
 78. Koo M-S, Lee J-H, Rah S-Y, Yeo W-S, Lee J-W, Lee K-L, Koh Y-S, Kang S-O, Roe J-H. 2003. A reducing system of the superoxide sensor SoxR in *Escherichia coli*. *EMBO J* 22:2614–22. doi: 10.1093/emboj/cdg252.
 79. Tsaneva IR, Weiss B. 1990. soxR, a locus governing a superoxide response regulon in *Escherichia coli* K-12. *J Bacteriol* 172:4197–205. doi: 10.1128/jb.172.8.4197-4205.1990.
 80. Krapp AR, Humbert MV, Carrillo N. 2011. The soxRS response of *Escherichia coli* can be induced in the absence of oxidative stress and oxygen by modulation of NADPH content. *Microbiology* 157:957–965. doi: 10.1099/mic.0.039461-0.
 81. Gambino M, Cappitelli F. 2016. Mini-review: Biofilm responses to oxidative stress. *Biofouling* 32:167–178. doi: 10.1080/08927014.2015.1134515.
 82. Gallique M, Bouteiller M, Merieau A. 2017. The Type VI Secretion System: A Dynamic System for Bacterial Communication? *Front Microbiol* 8:1454. doi: 10.3389/fmicb.2017.01454.
 83. Dudley EG, Thomson NR, Parkhill J, Morin NP, Nataro JP. 2006. Proteomic and microarray characterization of the AggR regulon identifies a pheU pathogenicity island in enteroaggregative *Escherichia coli*. *Mol Microbiol* 61:1267–1282. doi: 10.1111/j.1365-2958.2006.05281.x.
 84. Wang J, Zhou Z, He F, Ruan Z, Jiang Y, Hua X, Yu Y. 2018. The role of the type VI secretion system vgrG gene in the virulence and antimicrobial resistance of *Acinetobacter baumannii* ATCC 19606. *PLoS One* 13:e0192288. doi: 10.1371/journal.pone.0192288.
 85. Tian Y, Zhao Y, Wu X, Liu F, Hu B, Walcott RR. 2015. The type VI protein secretion system contributes to biofilm formation and seed-to-seedling transmission of *Acidovorax citrulli* on melon. *Mol Plant Pathol* 16:38–47. doi: 10.1111/mpp.12159.

CHAPTER 6

COMPARATIVE GENOMICS OF SHIGA TOXIN-PRODUCING *ESCHERICHIA COLI* (STEC) CATTLE AND CLINICAL ISOLATES IN MICHIGAN

ABSTRACT

Shiga toxin-producing *Escherichia coli* (STEC) is a prominent foodborne pathogen that results in numerous cases and outbreaks that can be traced back to the cattle environment. The ability for STEC to asymptomatically colonize cattle allows for the potential of this zoonotic pathogen to give rise to new virulent strains. A total of 1,212 isolates recovered from two sources, cattle (n=77) and patients with clinical infections (n=1,135), were examined to identify shared multilocus sequence types (STs), serogroups and/or virulence gene profiles among strains from the two sources. Three large clusters characterized by the ST and serogroups of highest frequency were identified to cluster 28.6% (n=22) cattle and 82.3% (n=934) clinical isolates; ST-119 (O103/O45), ST-106 (O26/O111), and ST-66 (O157). A single clade (bootstrap= 0.6) encompassed 58.4% (n=45) of all cattle isolates, which was driven by the high frequency of serogroup O6 (n= 36). In all, six shared serotype and ST profiles, including O157, O26, O103, O8, O98, and O109 were found in both patients and cattle. The latter three serogroups, however, were considerably less common in clinical cases than cattle. Core genome single nucleotide polymorphism (cgSNP) analysis clustered the cattle-derived O26, O103 and O98 isolates into distinct clades and importantly, one to two clinical isolates from young children were also included in these clusters. Serogroup O157 isolates from both sources were spread throughout the core genome phylogeny, suggesting that crossover events may occur in Michigan and that some strain types with specific profiles may be more likely to cross over than others. Together, these data are informative for future studies and intervention practices since the recovery of isolates with similar STs and virulence gene profiles from both humans and cattle can provide clues about those STEC strains that are most capable of causing human infections. These data

can be used to guide new farm practices aimed at minimizing the contamination of food products.

INTRODUCTION

Shiga toxin-producing *Escherichia coli* (STEC) is a foodborne pathogen that results in 265,000 illnesses annually and can present as diarrhea, hemorrhagic colitis and hemolytic uremic syndrome in severe cases (1, 2). While O157 serogroup is commonly implicated in outbreaks worldwide, a wide range of serogroups, non-O157, have been associated with varying degrees of clinical outcomes (3, 4). Six non-O157 serogroups have been identified as the predominant serogroups isolated: O26, O45, O103, O111, O121, O145 (5). With over 100 serogroups identified associated with clinical outcomes, they are all characterized by the presence of the Shiga toxin gene (*stx1* and/or *stx2*) (6–8). Other virulence factors have been identified and associated with various clinical outcomes such as *eaeA* (intimin gene used for attachment and effacement) and *ehxA* (plasmid encoded enterohemolysin) (7, 9–11). However, serogroup alone is not an indicator of disease outcome since a range of serogroups and virulence gene variants have been associated with human illness. Many of the non-O157 strains associated with disease share virulence factors with O157 due to horizontal gene transfer. This sharing of genetic elements leads to the acquisition of virulence genes that can result in the emergence of new zoonotic strains with the potential to cause severe disease and outbreaks.

Numerous O157 and non-O157 outbreaks and illnesses can be traced back to the contamination of cattle products or fecal contamination of nearby water and food (12–14). As the main reservoir for O157 and non-O157 STEC, cattle epithelial cells lack the globotriaosylceramide (Gb₃) receptor which allows for asymptomatic colonization (15). As a result, direct contact with the farm environment and occupational contact with this environment have been identified as risk factors for development of STEC infection (16–18). STEC is not specific to cattle and have been isolated from a wide range of other ruminants and animals (19).

Studies in dairy cattle have reported prevalence rates of 0.17-19.0% throughout the US and 0.2-48.8% for O157 and 0.4-74.0% for non-O157 worldwide (20). Surveillance of STEC in dairy farms has identified over 193 serogroups, including those representing the big-six non-O157 and O157; 24 of the serogroups that have been identified in dairy farms worldwide have been isolated from patients presenting with HUS, suggesting that crossover of some strain types is likely (14).

While STEC contamination has been found in cattle products and the farming environment, it is not fully understood whether certain serotypes and virulence factors are only circulating within the farming environment or if they have been previously isolated from clinical cases. Examination and comparison of the genetic profiles among STEC isolates recovered from both patients and from cattle can facilitate the identification of virulence factors and strain types with an enhanced ability to cause clinical infection. Indeed, this study will help to identify various genetic backgrounds that could be targeted in future surveillance studies to minimize the likelihood of STEC transmission.

MATERIALS AND METHODS

Bacterial Strains

A total of 77 cattle derived STEC isolates were examined. These isolates were recovered in 2012 from three beef herds (Herds 8B (n= 63), 11B (n= 3), and 12B (n= 4)) and one dairy farm (Herd 9D (n= 7)) as part of a prior study (21). The isolates included in this study were confirmed to be STEC by screening using multiplex PCR (22). All 77 STEC isolates were recovered from 38 cows; 70 isolates were from 32 cows in the three beef herds and seven isolates were from five cows in the dairy herd. The 63 STEC isolates from Herd 8B were previously characterized in chapter 5 and are included in this analysis.

For comparison, we examined 1,135 STEC isolates from patients with infections in Michigan between 2001 and 2018. All non-O157 isolates (n= 894) were characterized in Chapters 2, 3, and 4, while a subset of O157 isolates (n= 98) recovered in 2015-2018 were characterized in Chapter 4. An additional 143 O157 STEC isolates (2007-2014) were sequenced and included in the analysis for comparison.

DNA isolation and whole genome sequencing (WGS)

DNA was extracted from all cattle and clinical isolates that had not been characterized previously using the EZNA Bacterial DNA Kit (Omega Bio-Tek, Norcross, GA, USA). Library preps were performed with the Nextera XT kit (Illumina, San Diego, CA, USA) and libraries were sequenced on the Illumina MiSeq platform (2x250bp reads). New isolates that were not characterized previously were sequenced by the Michigan Department of Agriculture and Rural Development.

Bioinformatic analysis

Raw data was processed and cleaned with Trimmomatic to remove sequencing adapters, sequences with a phred score less than 20 and length less than 100bp (23). Quality control was performed with FastQC to examine the quality of the sequencing data (24). *De novo* assembly was performed with Spades 3.10.1 using kmers 21, 33, 55, 77, 99, and 127 (25).

Serotyping and virulence gene profiling was performed using Abricate and databases downloaded and curated from Center for Genomic Epidemiology (26) (www.genomicepidemiology.com). Multi locus sequence typing (MLST) alleles and sequence type (ST) assignment was performed using in house scripts with a Basic Local Alignment Search Tool backbone and EcMLST (27) (<http://shigatox.net>). MLST alleles were extracted, concatenated and aligned with ClustalW in MegaX (28). Neighbor-joining phylogenetic trees were generated using bootstrapping with 1000 repetitions. Core genome single nucleotide polymorphism (cgSNP) analysis was performed with Parsnp to identify clusters within shared serotypes and STs (29). cgSNP trees were visualized in FigTree (<http://tree.bio.ed.ac.uk/software/figtree>).

RESULTS

Genetic relatedness of clinical and cattle isolates with shared gene profiles and serotypes using multilocus sequence typing (MLST)

Application of MLST groups the 77 cattle isolates and 1,135 clinical isolates into three main clusters of importance, labeled based on ST with the highest frequency in the cluster: ST-119, which includes the O103 and O45 serogroups; ST-106 that groups the O26 and O111 serogroups; and ST-66, which includes O157 isolates (Figure 6.1). These three clusters account for 28.6% (n=22) of all 77 cattle isolates and 82.3% (n=934) of all 1,135 clinical isolates. Albeit low bootstrapping values (0.60), most cattle isolates (n=45, 58.4%) cluster together in a single clade with few clinical isolates (n=2). This clustering is likely driven by the high frequency of serogroup O6 (n=36, 46.8%) isolates, which was shown to persistently colonize herd 8B in Chapter 5. A fourth cluster with high bootstrap support (0.99) was also identified, which contains cattle-derived isolates representing two different serogroups that clustered together with clinical isolates. Serogroup O98 (ST-157, n= 5) and O182 (ST-158, n=1) were recovered from cattle, however, only one O98 isolate was identified in the clinical environment. Six different serogroups, three STs, and 19 isolates comprised this diverse cluster. The remaining cattle isolates (n=4) representing rare serogroups O8, O91, and O109 as well as one non-typeable (NT) isolate, were found throughout the phylogeny either within clades that contained a few clinical isolates (< 3 isolates) or singletons. Serogroup O91 did not cluster with the clinical O91 isolates and had a different ST (cattle: 653, clinical: 339, 815, NEW13).

Shared genomic profiles among STEC isolates from patients and cattle

A total of 28 cattle-derived isolates (37.2%) and 313 clinical isolates (27.6%) shared STs, virulence gene profiles and serotypes and represented the ideal population for the comparative analysis (Table 6.1). An additional 246 of the 1,135 clinical isolates (21.7%) were also included due to clustering by MLST with high bootstrap support and sharing similar virulence profiles, yielding 559 (49.2%) shared clinical isolates for comparison. These 246 isolates were considered potentially related or shared given slight differences in virulence gene profiles due to a high genetic relatedness and clustering based on MLST. For instance, a strain with a difference in *ehxA* presence but that shared a serotype and *eae* profile, would be designated as a potentially shared profile due to possible differences in sequencing quality. Incomplete sequencing of specific genes, for instance, could limit detection or identify slight variations even though the isolate may be related.

Among all 559 clinical and 28 cattle isolates with shared profiles and serotypes, six serotypes and nine combined serotype/virulence gene profiles were represented (Table 1). Serotype O26:H11 predominated among the shared serogroups with frequencies of 35.7% (n=10) and 21.8% (n=122) in cattle and patients with infection, respectively. Serotype O157:H7 was found with the second highest frequency for both cattle (n=7; 25.0%) and clinical (n=108; 19.3%) sources. Serotype O103:H2 was also found in both environments and is one of the big six non-O157 serogroups along with O26. The remaining three shared serotypes, O8:H19, O98:H21 and O109:H10, were less frequently isolated and only one clinical isolate was identified for each serotype. The shared virulence gene profiles that were found in isolates from both sources had a range of virulence gene subtypes including presence of *stx1* and/or *stx2*. For example, in shared O103:H2 strains, all isolates were identified to have *stx1a*, however, 4 (14.3%) cattle and 74

(13.2%) clinical isolates were identified as *ehxA:C*, while 131 (23.4%) clinical isolates were *ehxA:F*. Similarly, in the same subset, 1 (3.6%) cattle isolate and 4 (0.7%) clinical isolates were *stx2* positive in addition to *stx1a*.

Core genome SNP (cgSNP) analysis of clinical and cattle isolates with related STs, serotypes, and gene profiles

Four clusters of isolates containing ST-119 (O103), ST-66 (O157), ST-106 (O26), and ST-157/158 (O182/O98) were selected for core genome (cgSNP) analysis to examine the relatedness of the strains using a tool with a higher discriminatory power than MLST and virulence profiles. All isolates comprising clusters with high bootstrap support irrespective of serotype and ST, were included in the analysis.

The core genome analysis for ST-106 cluster (cattle: n=10, clinical: n=279) grouped all cattle isolates into one clade that lacked any clinical isolates (Figure 6.2). This grouping was also seen for ST-119 isolates (cattle: n=5, clinical: n=419) as all cattle isolates clustered together in a single clade, however, there were two clinical isolates that were related (Figure 6.3) All profiles, *stx1a*, *eae:epsilon*, *ehxA: C*, are shared among the 5 (100%) cattle and 2 (0.5%) clinical isolates that clustered together on the same clade, except for one cattle isolate was *ehxA* negative and another isolate was *stx1a2*. Conversely, the ST-66 cluster analysis (cattle: n=7, clinical: n=236) did not cluster all cattle isolates into a single clade, instead, they were dispersed throughout the phylogeny in smaller clusters (Figure 6.4). Two cattle isolates were grouped together into single clade, otherwise, the cattle isolates were all more closely related to clinical isolates than to other cattle isolates. Lastly, the analysis of ST-157/158 (cattle: n= 6, clinical: n= 13) cluster excluded cattle isolate TW17286 due to low sequencing quality for core genome analysis. Other ST-157

isolates were excluded from the final analysis due to the MUMi distance ≤ 0.01 indicative of unrelatedness between the isolates. The cattle isolates all clustered into one clade, along with a single clinical isolate that was identified previously and has a shared profile with a cattle isolate (Figure 6.5). Identical virulence profiles were identified for *stx1a* and *eae:zeta* across the cattle and clinical isolates that clustered. Differences were seen in *ehxA* subtypes, the cattle isolates were all negative (n:4, 66.7%) except for a single isolate that shared the *ehxA* subtype with the clinical isolate, *ehxA:F*, and a cattle isolate that was as *ehxA:B*. All serogroups clustered together, except for one O156 isolate that was more related to the O103 isolates than the other serogroups even though there was a difference in the virulence profiles of O156, *stx1a*, *eae:zeta*, *ehxA:C*, and the O103, *stx1a*, *eae:theta*, *ehxA:C* (n:3, 75%), isolates.

DISCUSSION

STEC has been isolated from a wide range of environmental sources including soil, water, deer and pigs (30–32). However, cattle are the main reservoir that harbors a diverse genetic pool of STEC with the ability to contaminate other environmental sources and serve as centers of horizontal gene transfer for the potential rise of more pathogenic zoonotic organisms. The interconnectedness of these environments and the ability for STEC to successfully survive and colonize humans, cattle and the water/soil emphasizes the One Health initiative. The core concept of One Health is that the health of all three facets (humans, animals and the environments) are interconnected and dependent upon the other (33). By examining the potential transmission events that may be occurring between the two environments, it will allow for future studies to examine how various preventative and surveillance methods may be used to help minimize the presence of specific serogroups and/or gene profiles.

Cattle food products have been implicated in several outbreaks since its initial identification as the etiological agent of hemorrhagic colitis associated with contaminated hamburgers in 1982 (34–37). O157:H7 was identified with this initial outbreak, since this time, other serogroups associated with the non-O157 big-six serogroups have been identified in outbreak associated cattle products (38, 39). Of the six shared profiles that were identified, three are rare serogroups that are not commonly isolated in humans and one clinical isolate was identified for each of these serogroups. Serogroup O8 and more specifically O8:H19 have been previously associated with clinical cases presenting with HUS, however, the majority of the cases are *stx2* positive and *eaeA* and *hlyA* (hemolysin) negative and cause mild disease (40, 41). Serotype O8:H19 has also been isolated from other environmental sources including porcine (42). Similarly, serotype O98:H21 has been isolated from patients presenting with HUS, as well

as, from other environmental sources such as deer (32, 43). The last serotype, O109:H10, is rarely found in the clinical environment and not commonly isolated from patients with severe clinical outcomes. The other three serogroups, two were non-O157 big six serogroups and O157, had a large number of clinical isolates with a shared profile.

The MLST genetic diversity of all the STEC isolates did not cluster any cattle isolates into cattle specific clusters. Every ST that was associated with a cattle isolate was either identified in a clinical isolate as well or a closely related ST was identified. Further analysis was performed to obtain a core genome analysis of groups that clustered based on similar ST and shared profiles across the two environments to identify the relatedness between the cattle and clinical isolates. Four analyses were performed with different clusters that had the following shared ST's present: ST-66 (O157), ST-119 (O103), ST-106 (O26), and ST-158 (O98). All ST's and serogroups that were within the clade supported with high bootstrap support were included to ensure that potential relatedness between isolates were not missed. Core genome analysis of the ST-66 clade did not cluster the cattle isolates together within one clade. Instead they were throughout the tree and grouped with different clinical isolates. Recent reports by the Centers for Disease Control and Prevention, US Food and Drug Administration and the US Department of Agriculture Food Safety and Inspection Service, estimated that 20-40% of O157 illnesses can be attributed to consumption of beef products (44). Similarly, ST-119 core genome analysis clustered all cattle isolates together and included two clinical isolates with the same serotype and virulence gene profile. Both clinical isolates were infants isolated around the same time period. Young children and elderly are more susceptible to STEC infection due to a weakened immune system (45). STEC isolates that rarely result in clinical outcomes have been identified to cause severe outcomes in young children. A case report described the transmission and infection of a

newborn that acquired an O146:H28 isolate during delivery that was *stx1a* positive and negative for intimin and hemolysin genes (46). The newborn subsequently developed HUS and severe neurological symptoms including epileptic episodes. The rarity of this serotype causing clinical outcomes supports the need to examine the rarer serotypes that are present in crossover events. ST-158 cluster contained 14 isolates that were typed into four serogroups and two ST's. Core genome analysis clustered the clinical and cattle isolates into different clades, however, one clinical isolate was found clustered within the cattle clade. The clustering of the O156 isolate with the O103 isolates suggests that serogroup alone may not be indicative of strain relatedness and was also observed in Chapter 2, 3 and 4. The last core genome analysis of ST-106 clustered all cattle and clinical isolates into distinct clades similar to ST-119 analysis. The closest related clinical isolate that was on the same clade as the cattle isolates was isolated from a young child. The young ages from the ST-119 and ST-106 may be due to contact with animals and has previously been shown that petting zoos and contact with farm animals are associated with STEC infection (17). Adults that have come in contact with colonized animals and did not wash their hands afterwards before handling young children may also be responsible to human to human transmission of STEC (47).

Some virulence gene profiles that were reported were not exact matches between the two environments, however, they were still included in the table as associated virulence gene profiles. In the core genome analysis, strains with differences in virulence gene subtypes were found to cluster together. In the ST-158/159 cluster, isolates with different epsilon subtypes, zeta and theta, were found on the same clade. Similarly, in the ST-119 analysis, one cattle isolate differed from the other profiles, *stx1a*, by the presence of *stx2* resulting in a *stx1a2* profile. Differences in the presence of *stx1* or *stx2* has been previously identified that *stx2* is more

toxigenic to epithelial cells and associated with severe disease (48–50). However, the differences between some of the gene variants such as, *stx2a* and *stx2c*, has not been fully elucidated and both have been associated with severe clinical outcomes, whereas *stx2e* has been found to be less virulent in humans (49, 51, 52). The ability of strains to acquire different Shiga toxin genes through the loss and acquisition of bacteriophages may shift the genetic variants of Shiga toxins that are present, thus supporting the need to include all isolates that may be related. Similarly, *ehxA* is plasmid mediated and related strains may lose and acquire different plasmids with varying gene variants. Any future studies would be able to identify whether some of the potentially associated gene variants are also present in the cattle environment.

Geographic location to farms and the cattle population density may further enhance crossover events that may occur between cattle and humans (53). Complete zip code or county data is not available for all clinical isolates to identify whether they were geographically located near one or multiple farms that were sampled. At the same time, the interactions of the cattle and humans with other animals colonized with STEC will further enhance the transmission and crossover events, since shared profiles have been reported in other animals (17, 18, 31, 54). Future surveillance of the cattle, clinical and environmental sources will help to further identify the transmission of STEC profiles among the niches that STEC is able to colonize. The health of all three sources and the application of preventative strategies in one may have an impact on the other two sources and is embodied by the One Health concept. Additional studies that examine other environmental sources and more cattle are needed to identify whether the shared profiles are shifting over time and whether the implementation of different cattle management practices may influence the profiles that are present and the transmission into other environments or humans.

APPENDIX

Table 6.1. Number of clinical and cattle STEC isolates with shared serotypes and virulence gene profiles.

Serogroup	Sequence Type	Profile	Total Cattle (n=28)	Total Clinical (n=559)*	Associated Clinical Profiles (n=246)**
O26:H11	106	<i>stx1a, eaeA-beta, ehxA-C</i>	10 (35.7%)	122 (39.0%)	<i>stx1a, eaeA-beta</i> (n= 2) <i>stx1a2a, eaeA-beta, ehxA-C</i> (n=2) <i>stx1a2d, eaeA-beta, ehxA-C</i> (n=1) <i>stx2a, eaeA-beta, ehxA-C</i> (n=1)
O103:H2	119	<i>stx1a, eaeA-epsilon, ehxA-C</i>	3 (10.7%)	70 (22.4%)	<i>stx1a, eaeA-epsilon, ehxA-F</i> (n=127) <i>stx1a, ehxA-F</i> (n=3) <i>stx1a2d, eaeA- epsilon, ehxA-F</i> (n=1)
		<i>stx1a, eaeA-epsilon</i>	1 (3.6%)	6 (1.9%)	
		<i>stx1a2, eaeA-epsilon, ehxA-C</i>	1 (3.6%)	4 (1.3%)	
O157:H7	66	<i>stx1a2a, eaeA-gamma, ehxA-B</i>	4 (14.3%)	86 (27.5%)	<i>stx1a, eaeA-gamma, ehxA-B</i> (n=3) <i>stx1a2c, eaeA-gamma, ehxA-B</i> (n=11) <i>stx1a2d, eaeA-gamma, ehxA-B</i> (n=1) <i>stx2a, eaeA-gamma, ehxA-B</i> (n=63) <i>stx2a, eaeA-gamma</i> (n=1) <i>stx2c, ehxA-B</i> (n=1) <i>stx2c, eaeA-gamma</i> (n=1) <i>stx2a2c, eaeA-gamma, ehxA-B</i> (n=28)
		<i>stx2c, eaeA-gamma, ehxA-B</i>	3 (10.7%)	22 (7.0%)	
O8:H19	653	<i>stx2a, ehxA-A</i>	1 (3.6%)	1 (0.3%)	
O98:H21	158	<i>stx1a, eaeA-zeta</i>	5 (17.9%)	1 (0.3%)	
O109:H10	433	<i>stx2a, eaeA-jota, ehxA-E</i>	1 (3.6%)	1 (0.3%)	

*Total number of clinical isolates that have a shared and associated virulence profile

Table 6.1 (cont'd)

****Associated clinical profiles do not share an exact match with cattle isolates, which may be due to the loss and acquisition of virulence genes or the lack of gene presence in WGS data due to sequencing quality at that location in the genome, however, they represent strains that may share genetic relatedness**

Figure 6.1. Neighbor joining phylogeny with 1000 bootstrap replication constructed with concatenated seven gene MLST profiles from 78 cattle isolates and 1,135 clinical isolates. All cattle STs are denoted with: **COW and blue shading of the node.

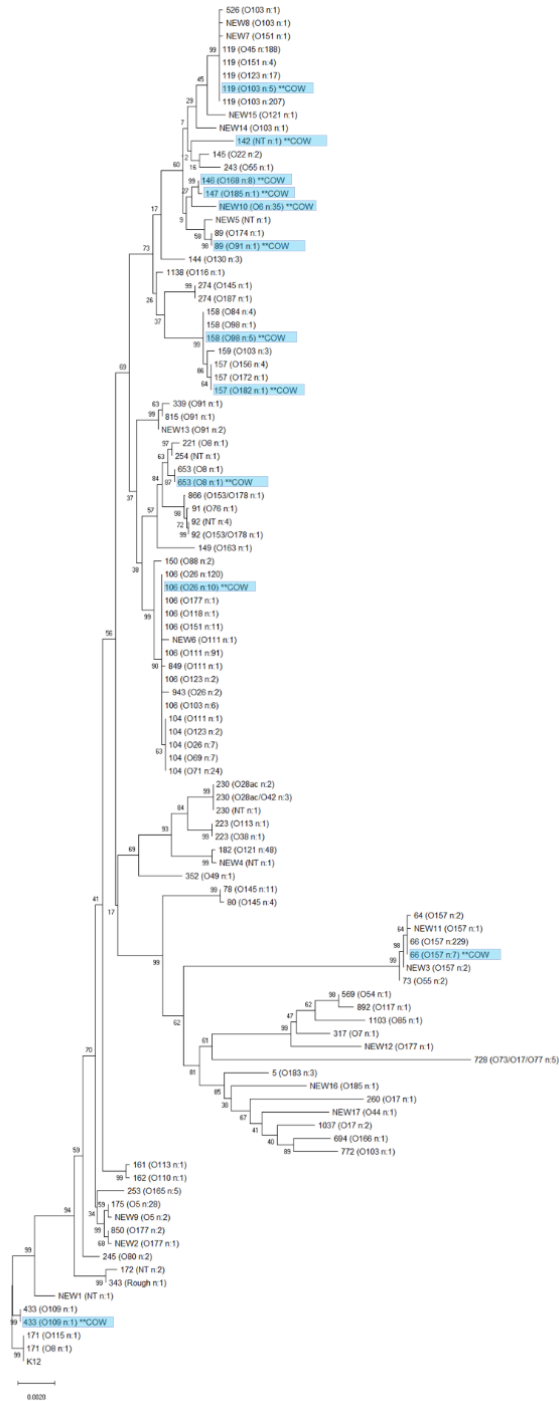


Figure 6.2. Core genome SNP analysis of 279 clinical and 10 cattle isolates within clade ST-106 (blue shading of shared clade).

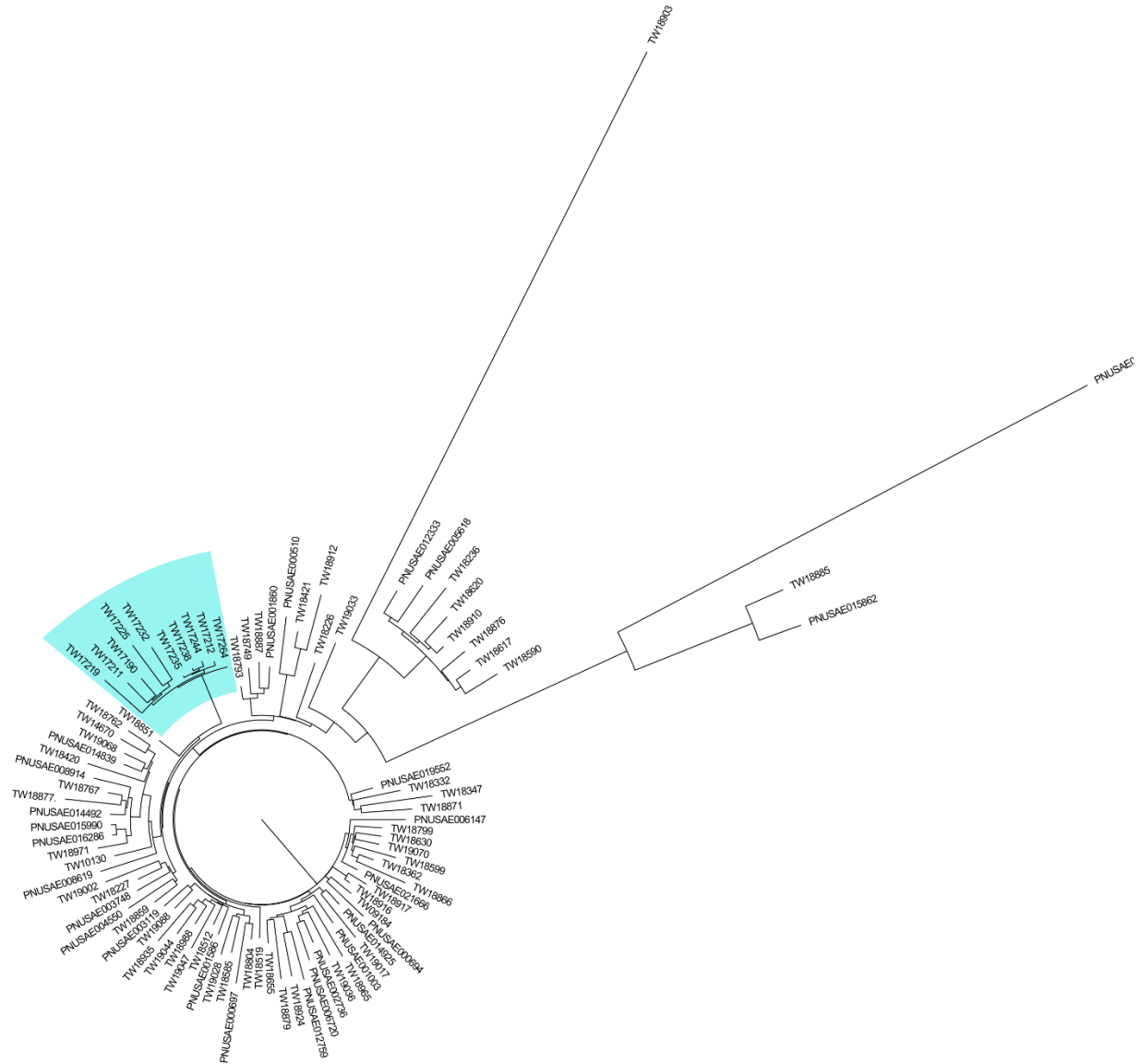
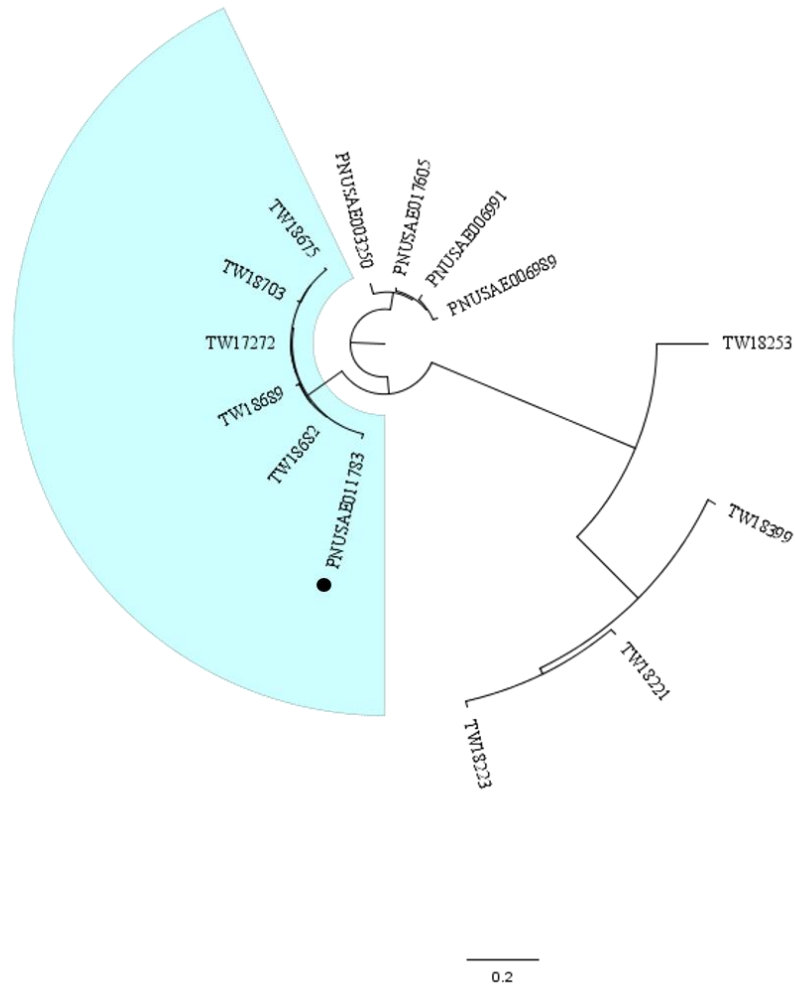


Figure 6.5. Core genome SNP analysis of 13 clinical and 6 cattle isolates within clade ST-157/158/159 (blue shading of shared clade). One shared clinical isolate within the cattle clade is denoted with a closed circle.



REFERENCES

REFERENCES

1. Scallan E, Hoekstra RM, Angulo FJ, Tauxe R V., Widdowson M-A, Roy SL, Jones JL, Griffin PM. 2011. Foodborne Illness Acquired in the United States—Major Pathogens. *Emerg Infect Dis* 17:7–15. doi: 10.3201/eid1701.P11101.
2. Karmali MA, Petric M, Lim C, McKeough PC, Arbus GS, Lior H. 1985. The association between idiopathic hemolytic uremic syndrome and infection by verotoxin-producing *Escherichia coli*. *J Infect Dis* 151:775–782. doi: 10.1093/infdis/151.5.775.
3. Gould LH, Mody RK, Ong KL, Clogher P, Cronquist AB, Garman KN, Lathrop S, Medus C, Spina NL, Webb TH, White PL, Wymore K, Gierke RE, Mahon BE, Griffin PM. 2013. Increased recognition of Non-O157 shiga toxin-producing *Escherichia coli* infections in the United States during 2000-2010: Epidemiologic features and comparison with *E. coli* O157 infections. *Foodborne Pathog Dis* 10:453–460. doi: 10.1089/fpd.2012.1401.
4. Crim SM, Griffin PM, Tauxe R, Marder EP, Gilliss D, Cronquist AB, Cartter M, Tobin-D'angelo M, Blythe D, Smith K, Lathrop S, Zansky S, Cieslak PR, Dunn J, Holt KG, Wolpert B, Henao OL. 2015. Preliminary incidence and trends of infection with pathogens transmitted commonly through food — Foodborne diseases active surveillance network, 10 U.S. sites, 2006–2014. *Morb Mortal Wkly Rep* 64:495–499.
5. Brooks JT, Sowers EG, Wells JG, Greene KD, Griffin PM, Hoekstra RM, Strockbine NA. 2005. Non-O157 Shiga Toxin–Producing *Escherichia coli* Infections in the United States, 1983–2002 . *J Infect Dis* 192:1422–1429. doi: 10.1086/466536.
6. O'Brien AD, LaVeck GD. 1983. Purification and characterization of a *Shigella dysenteriae* 1-like toxin produced by *Escherichia coli*. *Infect Immun* 40:675–683.
7. Blanco JE, Blanco M, Alonso MP, Mora A, Dahbi G, Coira MA, Blanco J. 2004. Serotypes, Virulence Genes, and Intimin Types of Shiga Toxin (Verotoxin)-Producing *Escherichia coli* Isolates from Human Patients: Prevalence in Lugo, Spain, from 1992 through 1999. *J Clin Microbiol* 42:311–319. doi: 10.1128/JCM.42.1.311-319.2004.
8. Eklund M, Scheutz F, Siitonen A, Stec A, States U, Pathogens E, Toxin S. 2001. Clinical Isolates of Non-O157 Shiga Toxin-Producing *Escherichia coli*: Serotypes, Virulence Characteristics, and Molecular Profiles of Strains of the Same Serotype. *J Clin Microbiol* 39:2829–2834. doi: 10.1128/JCM.39.8.2829.
9. Jenkins C, Willshaw GA, Evans J, Cheasty T, Chart H, Shaw DJ, Dougan G, Frankel G, Smith HR. 2003. Subtyping of virulence genes in verocytotoxin-producing *Escherichia coli* (VTEC) other than serogroup O157 associated with disease in the United Kingdom. *J Med Microbiol* 52:941–947. doi: 10.1099/jmm.0.05160-0.
10. Slanec T, Fruth A, Creuzburg K, Schmidt H. 2009. Molecular analysis of virulence

- profiles and Shiga toxin genes in food-borne Shiga toxin-producing *Escherichia coli*. *Appl Environ Microbiol* 75:6187–6197. doi: 10.1128/AEM.00874-09.
11. Ramachandran V, Brett K, Hornitzky MA, Dowton M, Bettelheim KA, Walker MJ, Djordjevic SP. 2003. Distribution of Intimin Subtypes among *Escherichia coli* Isolates from Ruminant and Human Sources. *J Clin Microbiol* 41:5022–5032. doi: 10.1128/JCM.41.11.5022-5032.2003.
 12. Wells JG, Shipman LD, Greene KD, Sowers EG, Green JH, Cameron DN, Downes FP, Martin ML, Griffin PM, Ostroff SM. 1991. Isolation of *Escherichia coli* serotype O157:H7 and other Shiga-like-toxin-producing *E. coli* from dairy cattle. *J Clin Microbiol* 29:985–989.
 13. Hainstock L, Donovan D. 2017. The Cheese Stood Alone. 17th Annu Michigan Commun Dis Conf.
 14. Hussein HS, Sakuma T. 2005. Invited Review: Prevalence of Shiga Toxin-Producing *Escherichia coli* in Dairy Cattle and Their Products. *J Dairy Sci* 88:450–465. doi: 10.3168/jds.S0022-0302(05)72706-5.
 15. Pruijboom-Brees IM, Morgan TW, Ackermann MR, Nystrom ED, Samuel JE, Cornick NA, Moon HW. 2000. Cattle lack vascular receptors for *Escherichia coli* O157:H7 Shiga toxins. *Proc Natl Acad Sci* 97:10325–10329. doi: 10.1073/pnas.190329997.
 16. Frank C, Kapfhammer S, Werber D, Stark K, Held L. 2008. Cattle density and Shiga toxin-producing *Escherichia coli* infection in Germany: increased risk for most but not all serogroups. *Vector Borne Zoonotic Dis* 8:635–643. doi: 10.1089/vbz.2007.0237.
 17. Schlager S, Lepuschitz S, Ruppitsch W, Ableitner O, Pietzka A, Neubauer S, Stöger A, Lassnig H, Mikula C, Springer B, Allerberger F. 2018. Petting zoos as sources of Shiga toxin-producing *Escherichia coli* (STEC) infections. *Int J Med Microbiol* 308:927–932. doi: 10.1016/J.IJMM.2018.06.008.
 18. Warshawsky B, Gutmanis I, Henry B, Dow J, Reffle J, Pollett G, Ahmed R, Aldom J, Alves D, Chagla A, Ciebin B, Kolbe F, Jamieson F, Rodgers F. 2002. Outbreak of *Escherichia coli* O157:H7 Related to Animal Contact at a Petting Zoo. *Can J Infect Dis* 13:175–181. doi: 10.1155/2002/873832.
 19. Martin A, Beutin L. 2011. Characteristics of Shiga toxin-producing *Escherichia coli* from meat and milk products of different origins and association with food producing animals as main contamination sources. *Int J Food Microbiol* 146:99–104. doi: 10.1016/j.ijfoodmicro.2011.01.041.
 20. Renter DG, Morris JG, Sargeant JM, Hungerford LL, Berezowski J, Ngo T, Williams K, Acheson DWK. 2005. Prevalence, risk factors, O serogroups, and virulence profiles of Shiga toxin-producing bacteria from cattle production environments. *J Food Prot* 68:1556–1565. doi: 10.4315/0362-028X-68.8.1556.

21. Venegas-Vargas C, Henderson S, Khare A, Mosci RE, Lehnert JD, Singh P, Ouellette LM, Norby B, Funk JA, Rust S, Bartlett PC, Grooms D, Manning SD. 2016. Factors associated with Shiga toxin-producing *Escherichia coli* shedding by dairy and beef cattle. *Appl Environ Microbiol* 82:5049–5056. doi: 10.1128/AEM.00829-16.
22. Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, Ouellette LM, Mladonicky JM, Somsel P, Rudrik JT, Dietrich SE, Zhang W, Swaminathan B, Alland D, Whittam TS. 2008. Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks. *Proc Natl Acad Sci USA* 105:4868–4873. doi: 10.1073/pnas.0710834105.
23. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. doi: 10.1093/bioinformatics/btu170.
24. Andrews S. 2010. FASTQC, a quality control tool for the high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
25. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V., Sirotkin A V., Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477. doi: 10.1089/cmb.2012.0021.
26. Seemann T. Abricate. Github. doi: Available online at: <https://github.com/tseemann/abricate>.
27. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410. doi: 10.1016/S0022-2836(05)80360-2.
28. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* 35:1547–1549. doi: 10.1093/molbev/msy096.
29. Treangen TJ, Ondov BD, Koren S, Phillippy AM. 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol* 15:524. doi: 10.1186/s13059-014-0524-x.
30. Ferens WA, Hovde CJ. 2011. *Escherichia coli* O157:H7: Animal reservoir and sources of human infection. *Foodborne Pathog Dis* 8:465–487. doi: 10.1089/fpd.2010.0673.
31. Cha W, Fratamico PM, Ruth LE, Bowman AS, Nolting JM, Manning SD, Funk JA. 2018. Prevalence and characteristics of Shiga toxin-producing *Escherichia coli* in finishing pigs: Implications on public health. *Int J Food Microbiol* 264:8–15. doi: 10.1016/j.ijfoodmicro.2017.10.017.
32. Singh P, Sha Q, Lacher DW, Del Valle J, Mosci RE, Moore JA, Scribner KT, Manning SD. 2015. Characterization of enteropathogenic and Shiga toxin-producing *Escherichia*

- coli in cattle and deer in a shared agroecosystem. *Front Cell Infect Microbiol* 5:29. doi: 10.3389/fcimb.2015.00029.
33. CDC. 2013. About One Health - One Health. Centers for Disease Control and Prevention. Atlanta, GA.
 34. O'Brien AD, Laveck GD, Thompson MR, Formal SB. 1982. Production of shigella dysenteriae type j-like cytotoxin by escherichia coli. *J Infect Dis* 146:763–769. doi: 10.1093/infdis/146.6.763.
 35. Erickson MC, Doyle MP. 2007. Food as a vehicle for transmission of Shiga toxin-producing *Escherichia coli*. *J Food Prot* 70:2426–2449. doi: 10.4315/0362-028X-70.10.2426.
 36. Kassenborg HD, Hedberg CW, Hoekstra M, Evans MC, Chin AE, Marcus R, Vugia DJ, Smith K, Ahuja SD, Slutsker L, Griffin PM. 2004. Farm visits and undercooked hamburgers as major risk factors for sporadic *Escherichia coli* O157:H7 infection: data from a case-control study in 5 FoodNet sites. *Clin Infect Dis* 38:S271-8. doi: 10.1086/381596.
 37. Jay MT, Garrett V, Mohle-Boetani JC, Barros M, Farrar JA, Rios R, Abbott S, Sowadsky R, Komatsu K, Mandrell R, Sobel J, Werner SB. 2004. A Multistate Outbreak of *Escherichia coli* O157:H7 Infection Linked to Consumption of Beef Tacos at a Fast-Food Restaurant Chain. *Clin Infect Dis* 39:1–7. doi: 10.1086/421088.
 38. Robbins A, Anand M, Nicholas DC, Egan JS, Musser KA, Giguere S, Prince H, Beaufait HE, Sears SD, Borda J, Dietz D, Collaro T, Evans P, Seys SA, Kissler BW. 2014. Ground Beef Recall Associated with Non-O157 Shiga Toxin-producing *Escherichia coli*, United States. *Emerg Infect Dis* 20:165–167. doi: 10.3201/eid2001.130915.
 39. Pihkala N, Bauer N, Eblen D, Evans P, Johnson R, Webb J, Williams C, the FSIS non-O157 working Group. 2012. Risk Profile for Pathogenic Non-O157 Shiga Toxin-Producing *Escherichia coli*. Office of Public Health Science, Office of Policy and Program Development, USDA-FSIS. Available from: <http://www.fsis.usda.gov/wps/wcm/connect/92de038d-c30e-4037-85a6-65c3a709435>.
 40. Friesema IHM, Schotsborg M, Heck MEOC, Van Pelt W. 2015. Risk factors for sporadic Shiga toxin-producing *Escherichia coli* O157 and non-O157 illness in the Netherlands, 2008-2012, using periodically surveyed controls. *Epidemiol Infect* 143:1360–1367. doi: 10.1017/S0950268814002349.
 41. Saupe A, Edel B, Pfister W, Löffler B, Ehricht R, Rödel J. 2017. Acute diarrhoea due to a Shiga toxin 2e-producing *Escherichia coli* O8:H19. *JMM Case Reports* 4:4–6. doi: 10.1099/jmmcr.0.005099.
 42. Baranzoni GM, Fratamico PM, Gangiredla J, Patel I, Bagi LK, Delannoy S, Fach P, Boccia F, Anastasio A, Pepe T. 2016. Characterization of shiga toxin subtypes and

- virulence genes in porcine shiga toxin-producing *Escherichia coli*. *Front Microbiol* 7:1–10. doi: 10.3389/fmicb.2016.00574.
43. Bai X, Mernelius S, Jernberg C, Einemo IM, Monecke S, Ehricht R, Löfgren S, Matussek A. 2018. Shiga toxin-producing *Escherichia coli* infection in Jönköping county, Sweden: Occurrence and molecular characteristics in correlation with clinical symptoms and duration of stx shedding. *Front Cell Infect Microbiol* 8:1–8. doi: 10.3389/fcimb.2018.00125.
 44. Interagency Food Safety Analytics Collaboration. 2018. Foodborne illness source attribution estimates for 2016 for *Salmonella*, *Escherichia coli* O157, *Listeria monocytogenes*, and *Campylobacter* using multi-year outbreak surveillance data, United States. GA and DC: U.S. Department of Health and Human Services, CDC, FDA, USDA-FSIS.
 45. Gould LH, Demma L, Jones TF, Hurd S, Vugia DJ, Smith K, Shiferaw B, Segler S, Palmer A, Zansky S, Griffin PM. 2009. Hemolytic Uremic Syndrome and Death in Persons with *Escherichia coli* O157:H7 Infection, Foodborne Diseases Active Surveillance Network Sites, 2000–2006. *Clin Infect Dis* 49:1480–1485. doi: 10.1086/644621.
 46. Stritt A, Tschumi S, Kottanattu L, Bucher BS, Steinmann M, Von Steiger N, Stephan R, Hächler H, Simonetti GD. 2013. Neonatal hemolytic uremic syndrome after mother-to-child transmission of a low-pathogenic stx2b harboring shiga toxin-producing *Escherichia coli*. *Clin Infect Dis* 56:114–116. doi: 10.1093/cid/cis851.
 47. Tozzi AE, Caprioli A, Minelli F, Gianviti A, De Petris L, Edefonti A, Montini G, Ferretti A, De Palo T, Gaido M, Rizzoni G. 2003. Shiga Toxin–Producing *Escherichia coli* Infections Associated with Hemolytic Uremic Syndrome, Italy, 1988–2000. *Emerg Infect Dis* 9:106–108. doi: 10.3201/eid0901.020266.
 48. Melton-Celsa AR. 2014. Shiga Toxin (Stx) Classification, Structure, and Function. *Microbiol Spectr* 2:EHEC-0024-2013. doi: 10.1128/microbiolspec.ehec-0024-2013.
 49. Shridhar PB, Siepker C, Noll LW, Shi X, Nagaraja TG, Bai J. 2017. Shiga Toxin Subtypes of Non-O157 *Escherichia coli* Serogroups Isolated from Cattle Feces. *Front Cell Infect Microbiol* 7:121. doi: 10.3389/fcimb.2017.00121.
 50. Fuller CA, Pellino CA, Flagler MJ, Strasser JE, Weiss AA. 2011. Shiga toxin subtypes display dramatic differences in potency. *Infect Immun* 79:1329–1337. doi: 10.1128/IAI.01182-10.
 51. Friedrich AW, Bielaszewska M, Zhang W, Pulz M, Kuczius T, Ammon A, Karch H. 2002. *Escherichia coli* Harboring Shiga Toxin 2 Gene Variants: Frequency and Association with Clinical Symptoms. *J Infect Dis* 185:74–84. doi: 10.1086/338115.
 52. Tseng M, Fratamico PM, Manning SD, Funk JA. 2014. Shiga toxin-producing *Escherichia*

- coli in swine: The public health perspective. *Anim Heal Res Rev* 15:63–75. doi: 10.1017/S1466252313000170.
53. Friesema IHM, Van De Kasstele J, De Jager CM, Heuvelink AE, Van Pelt W. 2011. Geographical association between livestock density and human Shiga toxin-producing *Escherichia coli* O157 infections. *Epidemiol Infect* 139:1081–1087. doi: 10.1017/S0950268810002050.
54. Miko A, Rivas M, Bentancor A, Delannoy S, Fach P, Beutin L. 2014. Emerging types of Shiga toxin-producing *E. coli* (STEC) O178 present in cattle, deer, and humans from Argentina and Germany. *Front Cell Infect Microbiol* 4:1–14. doi: 10.3389/fcimb.2014.00078.

CHAPTER 7

CONCLUSIONS AND FUTURE DIRECTIONS

Shiga toxin-producing *Escherichia coli* (STEC) is a leading foodborne pathogen that can be acquired through the consumption of contaminated food, water, or contact with colonized animals (1, 2). Since the addition of non-O157 STEC to the list of nationally notifiable conditions, the incidence rate has been steadily increasing over the past two decades, yet the knowledge of this diverse group of pathogens remains minimal in comparison to O157 STEC (3, 4). The increase in incidence may be due in part to increased surveillance efforts targeting non-O157 STEC and the transition from culturing on Sorbitol MacConkey (SMAC) agar to culture-independent tests that allow for better detection of non-O157 STEC strains (5). Additionally, the application of whole genome sequencing (WGS) has allowed for increased surveillance activities, an understanding of circulating strain types, and more accurate identification of strains (3). STEC has a diverse genetic background and has been classified into over 150 serogroups that have been isolated from humans with infection and multiple animal reservoirs including cattle, pigs, and other ruminants (6, 7). Despite extensive research on STEC diversity, gaps remain in our knowledge.

Minimal research, for instance, has been conducted on the range of molecular profiles in a diverse population of non-O157 STEC and on identifying relationships with specific clinical outcomes. Indeed, non-O157 STEC have been implicated in several outbreaks, but the genetic diversity has not been well characterized enough to define the role of genetic variation in strain pathogenicity. In spite of the common implication of beef and cattle products in foodborne infections, the genetic diversity of isolates in the farm environment and the extent by which cattle derived isolates cause human infections is also not fully understood (8). It is clear, however, that asymptomatic colonization in cattle allows for STEC to share horizontal elements, increasing the potential emergence of new pathogenic strains (9). Nonetheless, an assessment of

evolutionary changes in STEC from cattle has not been tracked in a longitudinal study. In an effort to better understand the genetic diversity of non-O157 STEC, the focus of this dissertation was to use WGS to examine the genetic relatedness and virulence profiles of 1,135 STEC isolates to identify associations with more severe clinical infections (n= 895) and persistence in cattle (n=77) and the farm environment.

The overall goal of the first study was to examine the genetic diversity and trends of non-O157 STEC in Michigan over the past 18 years, 2001-2018. Multiple transitions have occurred during the time period including sentinel to active surveillance, the emergence of culture independent tests (CIDTs), and the transition to WGS for *in silico* serotyping and detection of virulence genes. These transitions are apparent as the number of serogroups identified within a year increased from an average of 5 (2001-2006) to 18.3 (2008-2018) per year. Notably, O45 was the only serogroup associated with both hospitalization (n= 52, 39.1%) and bloody diarrhea (n= 87, 65.4%). Serogroup O111 was also associated with cases of bloody diarrhea (n= 38, 71.7%), but the lack of hospitalization suggests that other virulence factors may be playing a role in disease severity. Further, virulence gene profiles were associated with different MLST clusters. Specifically, cluster 3 and 4 were comprised of a range of other serogroups outside of big six and were associated with the presence of *stx2*. However, sample sizes may have been too small to identify associations with clinical outcomes. Additionally, future identification of gene variants associated with clinical outcomes will help direct public health intervention and surveillance. Further examination of similar serogroups can also be performed to identify any evolutionary events that may have occurred over the past 18 years, particularly for those rare and novel STs that appear to have diversified within Michigan. Extraction of the CRISPR loci and spacers could be used to examine whether there are longitudinal differences in highly similar

strain populations over time to help understand the stability of the CRISPR loci. With the increased accessibility to WGS, CRISPR typing is proving to be more labor intensive than utilizing a core genome analysis to examine large groups of strains. However, the use of the CRISPR loci and databases of CRISPR spacers that have been generated may be useful to extract bacterial genomes from metagenomic data for pathogen detection and classification.

Chapter 3 identified demographic and molecular differences of STEC isolates between two geographic locations, Michigan (n= 41) and Connecticut (n= 114), while exploring the potential use of the CRISPR loci as a genotyping tool. Prior research had identified that there were differences in the number and type of non-O157 STEC infections among FoodNet sites, suggesting that geographic location may play a role in non-O157 STEC diversity (3). In Chapter 3, we demonstrated that several similar serogroups outside of the big six non-O157 STEC including O5, O76 and O91, were shared among the two locations and that MLST did not identify any state specific clades. The common serogroups identified have been previously isolated from clinical cases and serogroup O91 is identified in high frequency in Europe (10). Further examination of foreign travel history for patients infected with these serogroups could help to elucidate whether any of the shared profiles were attributed to transmission while visiting Europe. Also, a difference in age groups was identified. Michigan cases were more likely to be between 11-29 years of age (n=12, 32.4%), which contrasts national reports by the CDC (3). Since Michigan contains more agricultural farms and has a higher cattle density than Connecticut, occupational differences may result in more adults being in contact with STEC colonized animals (11). Further epidemiological studies could examine whether occupational risks may be attributed to the increased number of adults that present with STEC infections. Similarly, socioeconomic factors could be examined to determine if the higher average salary

and better access to healthcare resources in Connecticut could partly explain why Michigan patients reported more severe clinical outcomes. In an effort to further discriminate isolates, an examination of the CRISPR spacer regions classified isolates as having up to 23 unique spacer profiles. Similar clustering of isolates was observed in the UPGMA phylogeny based on CRISPR profiles and the MLST neighbor-joining phylogeny. Further examination of CRISPR spacers is needed to identify if epidemiological concordance with related and non-related strains is still observed when examining the CRISPR spacers along with ST. Indeed, this methodology has been applied to STEC previously as well as other pathogens such as *Salmonella* and spoligotyping in *Mycobacterium tuberculosis* (12–14).

The next study, which is highlighted in Chapter 4, involved a retrospective examination of 510 outbreak-associated isolates to elucidate the relatedness of the strains and evaluate the ability of WGS to identify more informative clusters for epidemiological investigations. The use of WGS, core genome SNP (cgSNP), and high-quality SNP (hqSNP) analyses in increasing resolution and discrimination allowed for the differentiation of related isolates that were previously determined to be identical by pulsed-field gel electrophoresis (PFGE) and serogrouping. PFGE and WGS provided concordant results for six serogroup O5 outbreak-associated isolates; however, two isolates, O26 and NT, that were related by PFGE were distinct and found in unrelated clades in the cgSNP-based phylogeny. Importantly, hqSNP analysis was even more discriminatory, although this method is time consuming and computationally challenging and should only be used for highly related isolates previously classified with a sensitive tool such as cgSNP analysis. Future work should focus on the collection of more comprehensive epidemiological data to be used alongside WGS to examine whether epidemiological linkage was observed for cases excluded from outbreak investigation based on

PFGE. Identification of clusters using WGS can help determine whether a potential outbreak-associated isolate was missed and if epidemiological data supports the new clustering of isolates.

The last two Chapters, 5 and 6, focused on understanding STEC diversity in isolates recovered from cattle and the potential for some strain types to cross over and cause disease in humans. We also examined the ability of STEC to persist within a beef herd through an evaluation of biofilm production to determine if biofilms play a role in strain persistence and transmission. The highest biofilm formers, serogroup O6 (n=36, 57.1%), was the predominant serogroup identified throughout the herd and sampling periods. Other serogroups that have been implicated in severe clinical outcomes and outbreaks were O157 (n=1, 1.6%), O26 (n=11, 17.5%) and O103 (n=5, 7.9%), which were transiently identified within the herd. The cgSNP analysis was used to elucidate differences between high and low biofilm formers among strains with the same serogroups and from the same cow over time. Serogroup O6, for instance, persisted in cow 761 and 763, while serogroup O168 persisted in cow 760. WGS identified heterogeneous SNPs in *vgrG*, a multi copy type VI secretion system, across all O6 serogroups regardless of biofilm formation ability. This gene has been shown to facilitate biofilm formation in other gram-negative organisms (15, 16). Nonetheless, future transcriptomics work targeting this gene and other notable genes is important to identify whether specific gene variants are upregulated in strains with varying levels of biofilm production.

Finally, Chapter 6 aimed to elucidate the molecular characteristics of STEC isolates with similar properties or that are shared among clinical cases and cattle. Six shared serotypes, O26:H11, O103:H2, O157:H7, O8:H19, O98:H21, and O109:H10, and a total of nine shared virulence profiles were identified in strains from the two sources. Three serogroups were outside

of the O157 and big six non-O157 serogroups, further supporting the ability of rare profiles to cross over between cattle and humans. In addition, the cgSNP analysis clustered clinical cases within clades that mostly contained cattle isolates. Serogroup O157 was the exception as multiple cattle O157 isolates were identified throughout the phylogeny and grouped together isolates from clinical cases. Future work should include isolates from other STEC reservoirs for characterization to get a more complete understanding of STEC transmission and determine if serogroups and gene variants are more commonly isolated from specific reservoirs.

Overall, the findings in this dissertation illustrate that non-O157 STEC represent a diverse pathogen population and that WGS is advantageous for identifying the relatedness between strains using multiple methods. Additional studies, however, are needed on the non-O157 serogroups to determine how specific genetic characteristics are linked to disease severity and hospitalization risk. Surveillance of non-O157 STEC will not only help identify the distribution of strains and define the genetic variation among isolates in different geographic locations, but it will also identify the impact of selective pressures (e.g., antibiotic use) that promote or inhibit pathogen survival in different regions. Future projects will be aimed at examining larger sets of data that have more complete epidemiological information to increase the likelihood of identifying risk factors for both disease and persistent environmental colonization and to enhance understanding of the molecular profiles needed to cause clinical outcomes.

REFERENCES

REFERENCES

1. Scallan E, Hoekstra RM, Angulo FJ, Tauxe R V., Widdowson M-A, Roy SL, Jones JL, Griffin PM. 2011. Foodborne Illness Acquired in the United States—Major Pathogens. *Emerg Infect Dis* 17:7–15. doi: 10.3201/eid1701.P11101.
2. Caprioli A, Morabito S, Brugère H, Oswald E. 2005. Enterohaemorrhagic *Escherichia coli*: Emerging issues on virulence and modes of transmission. *Vet Res* 36:289–311. doi: 10.1051/vetres:2005002.
3. Gould LH, Mody RK, Ong KL, Clogher P, Cronquist AB, Garman KN, Lathrop S, Medus C, Spina NL, Webb TH, White PL, Wymore K, Gierke RE, Mahon BE, Griffin PM. 2013. Increased recognition of Non-O157 shiga toxin-producing *Escherichia coli* infections in the United States during 2000-2010: Epidemiologic features and comparison with *E. coli* O157 infections. *Foodborne Pathog Dis* 10:453–460. doi: 10.1089/fpd.2012.1401.
4. CDC. 2017. Foodborne Diseases Active Surveillance Network (FoodNet): FoodNet 2015 Surveillance Report (Final Data). Centers for Disease Control and Prevention. Atlanta, GA.
5. Manning SD, Madera RT, Schneider W, Dietrich SE, Khalife W, Brown W, Whittam TS, Somsel P, Rudrik JT. 2007. Surveillance for Shiga toxin-producing *Escherichia coli*, Michigan, 2001-2005. *Emerg Infect Dis* 13:318–321. doi: 10.3201/eid1302.060813.
6. Ramachandran V, Brett K, Hornitzky MA, Dowton M, Bettelheim KA, Walker MJ, Djordjevic SP. 2003. Distribution of Intimin Subtypes among *Escherichia coli* Isolates from Ruminant and Human Sources. *J Clin Microbiol* 41:5022–5032. doi: 10.1128/JCM.41.11.5022-5032.2003.
7. Miko A, Rivas M, Bentancor A, Delannoy S, Fach P, Beutin L. 2014. Emerging types of Shiga toxin-producing *E. coli* (STEC) O178 present in cattle, deer, and humans from Argentina and Germany. *Front Cell Infect Microbiol* 4:1–14. doi: 10.3389/fcimb.2014.00078.
8. Erickson MC, Doyle MP. 2007. Food as a vehicle for transmission of Shiga toxin-producing *Escherichia coli*. *J Food Prot* 70:2426–2449. doi: 10.4315/0362-028X-70.10.2426.
9. Pruimboom-Brees IM, Morgan TW, Ackermann MR, Nystrom ED, Samuel JE, Cornick NA, Moon HW. 2000. Cattle lack vascular receptors for *Escherichia coli* O157:H7 Shiga toxins. *Proc Natl Acad Sci* 97:10325–10329. doi: 10.1073/pnas.190329997.
10. Carroll KJ, Harvey-Vince L, Jenkins C, Mohan K, Balasegaram S. 2019. The epidemiology of Shiga toxin-producing *Escherichia coli* infections in the South East of England: November 2013–March 2017 and significance for clinical and public health. *J Med Microbiol* 68:930–939. doi: 10.1099/jmm.0.000970.

11. Friesema IHM, Van De Kasstele J, De Jager CM, Heuvelink AE, Van Pelt W. 2011. Geographical association between livestock density and human Shiga toxin-producing *Escherichia coli* O157 infections. *Epidemiol Infect* 139:1081–1087. doi: 10.1017/S0950268810002050.
12. Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajj SA, Allix C, Aristimuño L, Arora J, Baumanis V, Binder L, Cafrune P, Cataldi A, Cheong S, Diel R, Ellermeier C, Evans JT, Fauville-Dufaux M, Ferdinand S, De Viedma DG, Garzelli C, Gazzola L, Gomes HM, Gutierrez MC, Hawkey PM, Van Helden PD, Kadival G V., Kreiswirth BN, Kremer K, Kubin M, Kulkarni SP, Liens B, Lillebaek T, Ho ML, Martin C, Martin C, Mokrousov I, Narvskaja O, Yun FN, Naumann L, Niemann S, Parwati I, Rahim Z, Rasolofo-Razanamparany V, Rasolonalalana T, Rossetti ML, Rüschoff-Gerdes S, Sajduda A, Samper S, Shemyakin IG, Singh UB, Somoskov A, Skuce RA, Van Soolingen D, Streicher EM, Suffys PN, Tortoli E, Tracevska T, Vincent V, Victor TC, Warren RM, Sook FY, Zaman K, Portaels F, Rastogi N, Sola C. 2006. *Mycobacterium tuberculosis* complex genetic diversity: Mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiol* 6:23. doi: 10.1186/1471-2180-6-23.
13. Shariat N, Kirchner MK, Sandt CH, Trees E, Barrangou R, Dudley EG. 2013. Subtyping of *Salmonella enterica* serovar Newport outbreak isolates by CRISPR-MVLST and determination of the relationship between CRISPR-MVLST and PFGE results. *J Clin Microbiol* 51:2328–36. doi: 10.1128/JCM.00608-13.
14. Shariat N, Dudley EG. 2014. CRISPRs: Molecular Signatures Used for Pathogen Subtyping. *Appl Environ Microbiol* 80:430–439. doi: 10.1128/AEM.02790-13.
15. Wang J, Zhou Z, He F, Ruan Z, Jiang Y, Hua X, Yu Y. 2018. The role of the type VI secretion system *vgrG* gene in the virulence and antimicrobial resistance of *Acinetobacter baumannii* ATCC 19606. *PLoS One* 13:e0192288. doi: 10.1371/journal.pone.0192288.
16. Tian Y, Zhao Y, Wu X, Liu F, Hu B, Walcott RR. 2015. The type VI protein secretion system contributes to biofilm formation and seed-to-seedling transmission of *Acidovorax citrulli* on melon. *Mol Plant Pathol* 16:38–47. doi: 10.1111/mpp.12159.