

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

**Metabolic characteristics and
genomic epidemiology of
Escherichia coli serogroup O145**

A thesis presented in partial fulfilment of the
requirements for the degree of
Master of Science
in Microbiology
at Massey University,
Palmerston North, New Zealand

Rose Collis

2018

Abstract

Shiga toxin-producing *Escherichia coli* (STEC) are a global public health concern, and can cause severe human disease. Ruminants are asymptomatic reservoirs of STEC, shedding this pathogen via their faeces. There is 'zero tolerance' for the Top 7 STEC serogroups (O26, O45, O103, O111, O121, O145 and O157) in ground beef products exported to the USA. STEC may contaminate carcasses during processing and therefore are a major regulatory concern for New Zealand's meat industry. A previous study investigating the prevalence of STEC in young calves (n=1508) throughout New Zealand identified STEC O145 as the most prevalent serogroup (43%) at the dairy farm level compared to the other Top 7 serogroups. This high prevalence underlines STEC O145 as a public health concern and an issue for the meat industry.

Current culture-based methods for STEC detection are not fully discriminatory due to the lack of consistent differential characteristics between STEC and non-pathogenic *E. coli*. This study aims to (i) investigate metabolic characteristics of *E. coli* O145 to facilitate the differential culture of this serogroup and (ii) understand the genomic epidemiology of *E. coli* O145 using whole genome sequencing (WGS).

E. coli O145 strains examined in this study were genetically and metabolically diverse, according to carbon utilisation. The metabolic and genomic analyses were unable to differentiate between *stx*-positive and *stx*-negative O145 strains and there was no association with isolation source. However, clustering of O145 strains was observed according to multi-locus sequence type and at the level of *eae* subtype, a gene encoding the protein intimin which is involved in bacterial attachment to intestinal epithelial cells. Carbon substrates such as D-serine and D-malic acid were identified as candidate metabolites to differentiate defined O145 sequence types and may assist with identification in conjunction with currently available molecular methods.

This research has demonstrated the genetic heterogeneity of serogroup O145 and has made significant progress in the identification of metabolites that may prove beneficial in the development of a differential media for certain subsets of serogroup O145. Such a medium would prove a valuable tool for maintaining and monitoring public health and providing food quality and safety assurances that New Zealand meat for export is free of this pathogen.

Acknowledgements

Firstly, I would like to thank my supervisors Dr Adrian Cookson, Dr Anne Midwinter, A/Prof Patrick Biggs and Springer Browne for their guidance and encouragement throughout my study. Their help and support has been invaluable: from answering my many questions, listening to my ideas, reading numerous thesis drafts and providing constructive feedback to improve my lab, writing and genomic analysis skills. I really appreciate it!

I would also like to thank all members of the AgResearch Food Assurance and Meat Quality team and the Massey University *mEpiLab* for their support and encouragement throughout my study. I am very grateful to have been able to work with a group of people who are very encouraging and always happy to share their knowledge- and a few laughs of course! A special thank you to Dr David Wilkinson for his guidance and help with the library preparations for WGS and genomic analysis; Dr Samuel Bloomfield for his help with the genomic evolutionary analysis; and Dr Sara Burgess for helping me with the formatting of this thesis.

Thank you to Dr Colleen Ross and Dr Delphine Rapp (AgResearch Ltd) and Hugo Strydom and Naveena Karki (The Institute of Environmental Science and Research) for generously providing serogroup O145 isolates for use in this study.

I would like to acknowledge the financial support of the Palmerston North Medical Research Foundation, the IVABS post-graduate research fund and the AgResearch Food Provenance and Assurance Strategic Science Investment (SSI) Fund programme for generously funding components of this research project; and to the AgResearch SSI Fund and Massey University for awarding me a Masterate Scholarship in my first and second year of study, respectively.

Finally, thank-you to my parents for encouraging me to follow my passion and continue studying microbiology, and inspiring me every day with your hard work and determination. A special thanks to my parents, sisters, Louis, Ellie, family and friends for your continued love and support throughout my study.

Declaration

The virulence factor tree (section 2.12.3) and perl scripts for genomic analyses (Appendices C and D) were provided by A/Prof Patrick Biggs. The remainder of the work in this thesis was conducted by the candidate with guidance from supervisors.

Abbreviations

°C	Degrees Celsius
µg	Microgram
µL	Microlitre
A/E lesions	Attaching and effacing lesions
BHI	Brain heart infusion
bp	Base pairs
CDS	Coding sequences
CFU	Colony forming units
CGE	Center for Genomic Epidemiology
COGs	Clusters of Orthologous Groups
C _t	Cycle threshold
CT-SMAC	Cefixime and tellurite sorbitol MacConkey agar
DAEC	Diffuse-adherent <i>E. coli</i>
DEC	Diarrheagenic <i>E. coli</i>
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonucleotide triphosphates
EAEC	Enteraggregative <i>E. coli</i>
EHEC	Enterohaemorrhagic <i>E. coli</i>
EIEC	Enteroinvasive <i>E. coli</i>
EPEC	Enteropathogenic <i>E. coli</i>
ESS	Effective sample size
ETEC	Enterotoxigenic <i>E. coli</i>
ExPEC	Extraintestinal <i>E. coli</i>
FAE	Follicle associated duodenum
GC	Guanine-cytosine
HGT	Horizontal gene transfer
HKY substitution model	Hasegawa-Kishino-Yano substitution model
HPD	Highest posterior density
IMBs	Immunomagnetic beads
IMS	Immunomagnetic separation
Indels	Insertions/deletions
iTOL	Interactive Tree of Life
kb	Kilobase
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Orthology
LAA pathogenicity island	Locus of adhesion and autoaggregation pathogenicity island
LEE pathogenicity island	Locus of enterocyte effacement pathogenicity island
LOD	Limit of detection

MCL	Markov cluster
MCMC	Markov Chain Monte Carlo
min	Minute
mL	Millilitres
MLST	Multi-locus sequence typing
mPCR	Multiplex polymerase chain reaction
mTSB	Modified tryptone soya broth
ng	Nanogram
nM	Nanomolar
PCR	Polymerase chain reaction
pm	Picomolar
PMA	Propidium monoazide
RAMS	Recto-anal mucosal swabs
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
rpm	Revolutions per minute
RT-PCR	Real-time polymerase chain reaction
sec	Seconds
SNP	Single nucleotide polymorphism
ST	Sequence type
STEC	Shiga toxin-producing <i>Escherichia coli</i>
“Super six” STEC serogroups	O26, O45, O103, O111, O121, O145
T3SS	Type three secretion system
TBE buffer	Tris-borate-EDTA buffer
TMRCA	Time of most recent common ancestor
Top 7 STEC serogroups	O26, O45, O103, O111, O121, O145 and O157
tRNA	Transfer RNA
UPEC	Uropathogenic <i>E. coli</i>
USDA-FSIS	United States Department of Agriculture Food Safety Inspection Services
V	Volt
v/v	Volume per volume
w/v	Weight per volume
WGS	Whole genome sequencing

Table of contents

Abstract.....	II
Acknowledgements	III
Declaration.....	IV
Abbreviations	V
Table of contents	VII
List of figures	XIII
List of tables.....	XV
1. Introduction.....	1
1.1 Classification and pathotypes of <i>E. coli</i>	3
1.2 Pathogenicity of STEC	5
1.2.1 Shiga toxins	5
1.2.2 Intimin	6
1.2.3 Enterohaemolysin	7
1.3 Epidemiology.....	8
1.3.1 Epidemiology and detection methods for serogroup O157.....	8
1.3.2 Epidemiology of non-O157 serogroups	9
1.4 Current detection methods	11
1.4.1 Culture-based detection methods.....	11
1.4.2 Molecular based detection methods	13
1.5 Genomic epidemiology of STEC	14
1.5.1 Sequencing technologies.....	14
1.5.2 Previous comparative genomic studies	15
1.6 Conclusion	15
1.7 Objectives of this study	16
2. Materials and methods	17
2.1 Subculture	17
2.1.1 Hopkirk Research Institute culture collection	17
2.1.2 Subculture from glycerol broth	17
2.1.3 Subculture from agar	17

2.1.4 Glycerol broth inoculation	17
2.2 Culture-based methods	17
2.2.1 Calf faecal enrichments used in this study	17
2.2.2 <i>E. coli</i> serogroup O145 latex agglutination tests	18
2.2.3 Immunomagnetic separation (IMS)	18
2.3 DNA extraction	19
2.3.1 Crude DNA extraction	19
2.3.2 QIAamp DNA mini kit extraction	19
2.4 DNA quantification	20
2.4.1 Nanodrop	20
2.4.2 Qubit	20
2.5 Polymerase chain reaction (PCR)	20
2.5.1 Multiplex virulence PCR and O145 serogroup-specific PCR	20
2.5.2 PCR amplification of <i>eae</i>	21
2.6 Gel electrophoresis	23
2.6.1 2% w/v agarose	23
2.6.2 0.8% w/v agarose	23
2.7 Intimin (<i>eae</i>) subtyping	23
2.7.1 <i>eae</i> PCR and PCR product visualisation	23
2.7.2 PCR product purification and quantification	23
2.7.3 Sanger dideoxy sequencing PCR	24
2.7.4 Determining the <i>eae</i> subtype	24
2.8 Biolog phenotypic microarray assays	24
2.8.1 Inoculation of microarray assay plates	24
2.8.2 Analysis of phenotypic microarray assays	25
2.9 Whole genome sequencing (WGS)	26
2.9.1 DNA extraction, quantification and dilutions	26
2.9.2 Library preparations	26
2.9.3 Pooling individual library preparations	27
2.9.4 Library preparation quality controls	28
2.9.5 Whole genome sequencing (WGS)	28
2.10 Whole genome sequencing quality assessment and genome assembly	28
2.10.1 Proprietary Illumina sequencing report	28
2.10.2 QCtool	29
2.10.3 SPAdes	29
2.10.4 QUAST	30

2.11	Single nucleotide polymorphism (SNP) analysis.....	31
2.11.1	SNP analysis	31
2.12	Genome annotation and comparative analysis	31
2.12.1	Prokka.....	31
2.12.2	Center for Genomic Epidemiology	32
2.12.3	Comparison of virulence genes	32
2.12.4	Ribosomal multi-locus sequence typing (rMLST)	33
2.12.5	Identification of the locus of enterocyte effacement (LEE) pathogenicity island integration sites and <i>stx</i> -bacteriophage insertion sites.....	33
2.12.6	Download of publicly available serogroup O145 raw read data	34
2.12.7	Identification of orthologous groups	34
2.12.8	BEAST	35
2.13	Comparison of phenotypic and genotypic data	35
2.13.1	Identification of the core and pan genome.....	35
2.13.2	Interrogation of the pan genome.....	36
2.13.3	Identification of genes associated with carbohydrate metabolism	36
3.	Results - Isolation of <i>E. coli</i> serogroup O145	37
3.1	Isolation of <i>E. coli</i> serogroup O145	37
3.2	Culture-based isolation.....	39
3.3	Serogroup O145 characterisation	40
3.4	Discussion	41
3.5	Summary.....	44
4.	Results - Utilisation of carbon substrates	45
4.1	Utilisation of carbon substrates (PM1 MicroPlates™)	45
4.1.1	Clustering broadly correlates with <i>eae</i> subtype	45
4.1.2	Clustering broadly correlates with sequence type	47
4.1.3	Reproducibility of serogroup O145 carbon utilisation on PM1 MicroPlates™	49
4.2	Utilisation of carbon substrates (PM2A MicroPlates™).....	52
4.2.1	Clustering broadly correlates with <i>eae</i> subtype and sequence type	52
4.2.2	Reproducibility of serogroup O145 carbon utilisation on PM2A MicroPlates™	54

4.3	Candidate substrates for use in a differential media	56
4.3.1	Identification of carbon substrates to differentiate certain <i>eae</i> subtypes and sequence types	56
4.4	Discussion	61
4.5	Summary	64
5.	Results - Whole genome sequencing and comparative analysis	65
5.1	Selection of <i>E. coli</i> serogroup O145 strains for whole genome sequencing	65
5.2	Comparative genomics	65
5.2.1	Genome composition	65
5.2.2	Virulence factors	68
5.2.3	<i>in silico</i> ribosomal multi locus sequence typing	71
5.2.4	Locus of enterocyte effacement pathogenicity island integration sites	71
5.3	Core single nucleotide polymorphism analysis	73
5.3.1	Core SNP analysis of serogroup O145 strains sequenced in this project (n=53)	73
5.3.2	Core SNP analysis comparison with publicly available serogroup O145 strains	74
5.4	Core and pan genome analysis	79
5.4.1	Identification of the core and pan genome	79
5.4.2	Association of pan genome with traits of interest	83
5.5	Evolutionary analysis of serogroup O145 strains	86
5.5.1	Mutation rate and estimated TMRCA of <i>E. coli</i> serogroup O145 <i>eae</i> subtype γ strains	86
5.6	Discussion	89
5.7	Summary	94
6.	Results - Phenotype and genotype correlations	95
6.1	Association of genes in the pan genome and specific carbon substrate utilisation	95
6.2	Diversity of protein functional groups associated with the utilisation of specific carbon substrates	95
6.3	Proteins involved in carbon metabolism	98
6.4	Discussion	99

6.4.1	Proteins and genes identified by genomics which are potentially associated with carbon substrate utilisation	99
6.4.2	Difficulties identifying genes involved in carbon substrate utilisation	101
6.5	Summary	102
7.	General discussion	104
7.1	Culture-based isolation of serogroup O145 (Chapter 3)	106
7.2	Carbon utilisation (Chapter 4)	109
7.3	Comparative genomics of serogroup O145 (Chapter 5)	111
7.4	Phenotype and genotype correlations (Chapter 6)	112
7.5	Value of this research	113
7.6	Areas for further research	114
7.6.1	Development of a differential media for serogroup O145	114
7.6.2	Subsequent WGS analysis	114
7.6.3	An alternative approach for identifying phenotype and genotype correlations	115
7.7	Concluding statement	115
8.	Bibliography	104
9.	Appendices	134
Appendix A	- Bacterial strains used in this study	135
Appendix B	- R code for Omnilog analysis	139
Appendix C	- SQS2 perl script	141
Appendix D	- Prokka perl script	141
Appendix E	- Publicly available genome sequences analysed in this study	142
Appendix F	- Calf faecal enrichments screened for serogroup O145 using culture-based methods	145
Appendix G	- PM1 and PM2A MicroPlates™ carbon substrates	147
Appendix H	- Serogroup O145 strains analysed using the Omnilog phenotypic microarray system	148
Appendix I	- Virulence factors identified from serogroup O145 whole genome sequence data in this study (n=53)	150
Appendix J	- <i>E. coli</i> tRNA integration site for the locus for enterocyte effacement (LEE) pathogenicity island	153

Appendix K - Virulence factors identified from publicly available serogroup
O145 whole genome sequence data (n=47)155

List of figures

Figure 1.1: The number of notified STEC cases per year in New Zealand from 1993-2016.	3
Figure 1.2: Prevalence of the Top 7 STEC serogroups on dairy farms in New Zealand in spring 2014.....	10
Figure 3.1: Calf faecal enrichment screening process.....	39
Figure 4.1: Heat-map showing <i>E. coli</i> serogroup O145 strains carbon utilisation profiles (PM1 MicroPlates™)	46
Figure 4.2: Cluster dendrogram showing the similarities of <i>E. coli</i> serogroup O145 strains based on their carbon utilisation profile	48
Figure 4.3: Heat-map showing <i>E. coli</i> serogroup O145 strains carbon utilisation profiles (PM1 MicroPlate™) with replicates and duplicates	50
Figure 4.4: Heat-map showing <i>E. coli</i> serogroup O145 strains carbon utilisation profiles (PM2A MicroPlates™)	53
Figure 4.5: Heat-map showing <i>E. coli</i> serogroup O145 strains carbon utilisation profiles (PM2A MicroPlates™) with replicates	55
Figure 4.6: Heat-map showing <i>E. coli</i> serogroup O145 strains carbon utilisation profiles on selected PM1 carbon substrates	58
Figure 4.7: Heat-map showing <i>E. coli</i> serogroup O145 strains carbon utilisation profiles on selected PM2A carbon substrates.....	59
Figure 5.1: Box and whisker plots indicating the genome composition of <i>E. coli</i> serogroup O145 strains (n=53)	67
Figure 5.2: Neighbor-Net tree constructed using the presence or absence data from 31 virulence genes identified by the CGE VirulenceFinder webserver	70
Figure 5.3: Neighbor-Net phylogeny constructed using <i>in silico</i> ribosomal multi-locus sequence typing	72
Figure 5.4: Neighbor-Net phylogeny of core SNP analysis from serogroup O145 strains sequenced in this study (n=53)	75
Figure 5.5: Neighbor-Net phylogeny of core SNP analysis from <i>eae</i> subtype γ serogroup O145 strains sequenced in this study (n=41)	76
Figure 5.6: Neighbor-Net phylogeny of core SNP analysis of serogroup O145 strains sequenced in this study and publicly available serogroup O145 strains (n=100)	77

Figure 5.7: Neighbor-Net phylogeny of core SNP analysis of serogroup O145 <i>eae</i> subtype γ sequenced in this study and publicly available serogroup O145 <i>eae</i> subtype γ strains (n=83)	78
Figure 5.8: Comparison of the number of conserved and total genes in the serogroup O145 pan genome with increasing number of genomes	80
Figure 5.9: The effect the number of serogroup O145 genomes included in the analysis has on the number of conserved genes.....	81
Figure 5.10: The effect the number of genomes included in the analysis has on the number of genes in the pan genome	81
Figure 5.11: The pan genome composition of serogroup O145 strains (n=53)	82
Figure 5.12: Functional analysis of proteins associated with traits of interest for serogroup O145	85
Figure 5.13: Maximum clade credibility tree showing predicted dates serogroup O145 <i>eae</i> subtype γ strains last shared a common ancestor	88
Figure 6.1: Functional analysis of proteins associated with the utilisation of specific carbon substrates	99

List of tables

Table 2.1: PCR primer sequences and resulting amplicon lengths	22
Table 2.2: Reference genomes used to identify LEE pathogenicity island integration sites.....	34
Table 3.1: Serogroup O145 isolation from calf faecal enrichments using culture-based methods	40
Table 3.2: Comparison of the number of serogroup O145 isolates confirmed for each enrichment for both culture-based methods.....	40
Table 3.3: Intimin subtypes determined according to best match using BLASTN ..	41
Table 4.1: Comparison of carbon substrates from PM1 MicroPlates™ (n=11)	51
Table 4.2: Comparison of carbon substrates from PM2A MicroPlates™ (n=4) which differ between ≥1 set of replicates (n=4)	54
Table 4.3: Specific carbon substrates utilised by serogroup O145 strains that could be used to differentiate <i>eae</i> subtypes and sequence types	60
Table 6.1: Carbon substrates selected for further investigation to identify phenotype and genotype correlations.....	97