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 *m*EpiLab 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 *E. coli*
DEC	Diarrheagenic *E. coli*
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonucleotide triphosphates
EAEC	Enteroaggregative *E. coli*
EHEC	Enterohaemorrhagic *E. coli*
EIEC	Enteroinvasive *E. coli*
EPEC	Enteropathogenic *E. coli*
ESS	Effective sample size
ETEC	Enterotoxigenic *E. coli*
ExPEC	Extraintestinal *E. coli*
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 *Escherichia coli*
“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 *E. coli*
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

List of figures

List of tables

1. Introduction

Shiga toxin-producing *Escherichia coli* (STEC) are a global public health concern and can cause human disease with a broad range of symptoms from abdominal cramps and diarrhoea to life threatening haemolytic uraemic syndrome and in severe cases, death [2,3]. The infectious dose of some STEC serotypes is reported to be very low [4], therefore infection can be caused by minimal exposure to STEC. The number of STEC notifications has been steadily increasing in New Zealand since the late 1990's, and in 2016, the number of STEC notifications was 8.9 cases per 100,000 population (Figure 1.1) [5]. More intensive laboratory screening methods for STEC were introduced in Auckland in 2015, which may partially contribute to this increase [5]. In 2016, 491 STEC isolates were identified from clinical cases, of which 41.8% were *E. coli* O157:H7, 21.4% were identified as non-O157 serogroups and for 21.4% of cases an isolate was not obtained [5]. Due to the lack of standardised culture-based detection methods for non-O157 serogroups, it has been suggested they are likely to be under-reported [6].

Ruminants, particularly calves, are asymptomatic reservoirs of STEC shedding this pathogen in their faeces [7], providing an important source of both food and environmentally acquired STEC infections. In New Zealand, risk factors for sporadic STEC infections include contact with cattle, animal manure, or recreational waters [8]. No food related risk factors were identified [8], unlike overseas where STEC has been identified as the causative pathogen in disease outbreaks associated with contaminated food products such as romaine lettuce [9], ice-cream [10], and hamburger patties [11].

Due to a large outbreak of human disease associated with undercooked beef in North America, *E. coli* O157:H7 was declared an adulterant of ground beef in 1994 by the United States Department of Agriculture Food Safety Inspection Services (USDA-FSIS) [12]. This 'zero tolerance' led to the development of standardised detection methods for this pathogen [13]. The association of the "super six" serogroups (O26, O45, O103, O111, O121 and O145) with human disease has resulted in these serogroups also being declared adulterants of

ground beef in the USA in 2012 [14]. Even with good hygienic practices during meat processing, STEC can contaminate carcasses, therefore STEC are a major regulatory concern for New Zealand's meat industry. Recent research has shifted to the development of detection methods for the non-O157 serogroups due to their 'zero tolerance' in ground beef, high prevalence in cattle [15], and their association with human disease [5]. However, development of a standardised detection method for non-O157 serogroups has been hindered by the lack of differential characteristics between STEC and non-pathogenic *E. coli*, and also the phenotypic variation observed within these serogroups [16,17].

A cross-sectional study investigating the prevalence of STEC in young calves throughout New Zealand identified STEC O145 as the most prevalent serogroup (43%) at the dairy farm level compared to the other Top 7 (O157 and the "super six") serogroups [18]. This high prevalence suggests *E. coli* serogroup O145 is both a public health concern and a regulatory issue for New Zealand's meat export industry. Therefore, a standardised detection method for serogroup O145 would provide a valuable tool for maintaining and monitoring public health and providing food quality and safety assurances that the New Zealand meat for export is free of this pathogen.

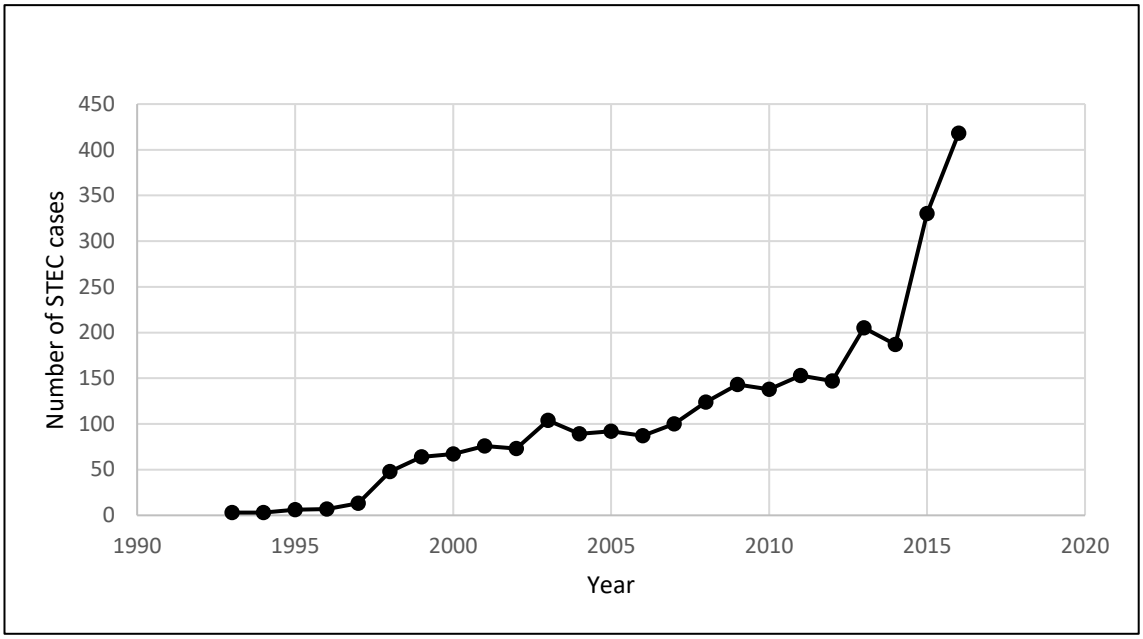

Figure 1.1: The number of notified STEC cases per year in New Zealand from 1993-2016.

The number of STEC notifications per year was obtained from New Zealand public health surveillance reports from 1993-2016 provided by The Institute of Environmental Science and Research Ltd [19].

1.1 Classification and pathotypes of *E. coli*

E. coli are a diverse species which are naturally found as commensals in the intestinal tract of healthy humans and animals, however a range of *E. coli* pathotypes can also cause severe human disease [20]. *E. coli* are classified serologically according to O (O-specific polysaccharide component of lipopolysaccharide), H (flagellar) and K (capsular) antigens [21,22]. O and H serotyping is often considered the 'gold standard' of *E. coli* characterisation [23]. Currently, 188 O-serogroups and 53 H antigens are included in the *E. coli* serotyping scheme [23]. Specific combinations of certain O and H antigens are defined as a serotype [22]. Culture-based serotyping of *E. coli* is time-consuming and laborious, however is essential for outbreak detection and identifying clinically significant serogroups such as the Top 7 [23].

Many bacterial species, including *E. coli*, can also be classified using multi-locus sequence typing (MLST). MLST takes advantage of nucleotide variation in

regions from multiple house-keeping genes (usually seven for *E. coli*). Sequence variation from within each of these regions is assigned a number and each unique combination of these regions is specified as the sequence type (ST) [24]. MLST is traditionally determined using polymerase chain reaction (PCR) amplification and Sanger sequencing, however, databases are currently available for the *in silico* analysis of MLST from whole genome sequencing (WGS) data [24]. MLST is a valuable bacterial typing scheme allowing the classification of outbreak isolates and also provides a standardised classification scheme for a global comparison of strains [24].

Pathogenic *E. coli* can cause a broad spectrum of disease including meningitis and sepsis, urinary tract infections and diarrhoea [22], due to the acquisition of a number of virulence factors. Diarrheagenic *E. coli* (DEC) have been classified into six main pathotypes according to the symptoms they cause and pathotype-specific virulence factors [22,25]. These pathotypes include: (i) enteropathogenic *E. coli* (EPEC), (ii) enterotoxigenic *E. coli* (ETEC), (iii) enteroinvasive *E. coli* (EIEC), (iv) enteroaggregative *E. coli* (EAEC), (v) diffuse-adherent *E. coli* (DAEC) and (vi) STEC [25]. Enterohaemorrhagic *E. coli* (EHEC) are a subgroup of STEC [25].

Although there are defined pathotypes, they can be difficult to distinguish as many of these groups cannot be differentiated by metabolic properties, including from other non-pathogenic strains, and individual serogroups are often associated with multiple DEC pathotypes [22]. Currently there is no standardised definition to characterise the DEC pathotypes, and proposed definitions often rely on the presence or absence of specific virulence factors [25]. The proposed defining markers for these pathotypes include the locus of enterocyte effacement (LEE) pathogenicity island for EPEC [25], and this pathotype is further divided into typical EPEC and atypical EPEC, according to the presence and absence of bundle-forming pili (*bfp*), respectively [26]. The defining markers for ETEC are the presence of heat-labile and heat-stabile enterotoxins; EIEC the plasmid pINV; EAEC the presence of aggregative adhesion factors such as *aggR*; and for DAEC the proposed marker is AIDA-1, an adhesin involved in diffuse adherence [25]. The defining characteristics of STEC is the production of Shiga toxins [22,25]. However, these defining features have been widely debated as DEC pathotypes

can acquire additional virulence factors by horizontal gene transfer (HGT), which would traditionally belong to another DEC pathotype. For example, many STEC also carry the LEE pathogenicity island.

The lack of definitive markers for each specific DEC pathotype highlights the genetic heterogeneity of *E. coli* and their ability to readily acquire new genetic material via HGT which may result in strains being described as the incorrect pathotype, or distantly related strains being grouped together [25]. Additionally, hybrid strains may emerge which contain pathogenic determinants of multiple DEC pathotypes [25], such as the O104:H4 EAEC/ STEC hybrid strain which caused a large outbreak of human disease in Germany in 2011 [27]. To overcome the current limitations of defining DEC pathotypes, the use of a typing scheme based on whole genome sequence data has been suggested, which would incorporate genes in both the core and accessory genomes [25]. Due to the advances in WGS and the increase in publicly available *E. coli* whole genome sequences, this represents an area warranting subsequent research.

1.2 Pathogenicity of STEC

1.2.1 Shiga toxins

The defining characteristic of STEC is the production of Shiga toxins [22]. STEC may express two types of Shiga toxins, Stx1 and Stx2 which are immunologically distinct [28]. STEC pathogenesis in humans involves adhesion and intestinal colonisation followed by Stx production, resulting in impaired intestinal epithelial cell barrier function and diarrhoea [29]. Stx can also cause haemolytic uraemic syndrome and other sequelae, when the toxin enters the bloodstream and disseminates to other tissues, including the brain [29,30]. The Shiga toxin genes, *stx1* and *stx2*, are encoded by bacteriophage [31], which infect *E. coli* and can result in the insertion of bacteriophage genes into the host chromosome that can be maintained in a lysogenic state, leading to the expression of the subsequent Shiga toxins [32]. There are a number of *stx*-encoding bacteriophage insertion sites in the bacterial genome, for example *stx* insertion has been commonly identified in the *wrbA*, *yehV*, *yecE* and *argW* genes [33]. Shiga toxins have two main subunits A and B; a subunit B complex binds to the host cell receptor

globotriaosylceramide (Gb3) and the subunit A protein enters the cell, blocking protein synthesis via the cleavage of an adenine residue from the 28S rRNA of the 60S ribosome [28]. Allelic variants have been described for both Stx1 (Stx1, Stx1c, Stx1d) and at least 11 variants for Stx2 (such as Stx2a, Stx2c, Stx2d and Stx2f) [28,34]. In comparison to Stx1, Stx2 has been associated with more severe disease in humans [35]. However, the full mechanisms by which the Shiga toxins cause disease in humans is not yet fully understood.

1.2.2 Intimin

Other virulence factors, such as intimin, are involved in the pathogenicity of STEC. Intimin is an outer membrane adhesin, encoded by the *eae* gene on the LEE pathogenicity island [36]. The LEE pathogenicity island is at least 35 kb in length and also encodes a type three secretion system (T3SS), the translocated intimin receptor (*tir*), a variety of other effector proteins and regulatory elements [37]. The genes encoding the T3SS are relatively conserved, whereas genes encoding intimin and other effector proteins are less conserved, according to DNA sequence similarity [36]. Intimin and other proteins encoded by the LEE are involved in the formation of attaching and effacing (A/E) lesions [36]. A/E lesions are formed by microvilli effacement and attachment of bacteria to the intestinal epithelial cells, resulting in the formation of actin pedestals [36]. The *eae* gene is ~2.8 kb and consists of a conserved 5’ end and a variable 3’ end [38]. According to the variable C-terminal, the *eae* gene can be categorised into different subtypes [38,39], and at least 28 different *eae* subtypes denoted by Greek letters (such as α, β, γ) have been identified [38]. Some STEC serotypes such as O157:H7 (γ), O26:H11 (β), O103:H2 (ε), O111:H8 (θ) and O145:H28 (γ) are characterised by a single *eae* subtype [40], however the association of multiple *eae* subtypes with specific serogroups has been observed [36,41].

It has been hypothesised that the LEE was acquired by STEC (and EPEC) through HGT as it has a lower GC content (38.3%) compared to the remaining *E. coli* genome (~50.5%) [42]. The LEE has also been identified in other bacteria such as *Citrobacter rodentium* and *Escherichia albertii*, which can also cause A/E lesions [43].

The LEE consists of five polycistronic operons (*LEE 1-5*) and 41 open reading frames [44]. Regulation of the expression of genes encoded by the LEE is complex involving physical, environmental, host and microbiota associated cues [37]. However, the transcriptional regulator Ler, also encoded on the LEE, is one of the main regulators involved in the expression of these operons [37].

LEE pathogenicity islands are inserted in the *E. coli* genome near the tRNA genes *selC*, *pheV* and *pheU* [36] and the genetic similarity of different LEE regions has been proposed to be classified according to these different insertion sites [36]. According to sequence analysis of conserved genes within the LEE, it has been suggested that there are two types of LEE core clusters; one which is found in strains with *eae* subtypes ε (*pheV*) and β (*pheU*), and the other in strains with *eae* subtypes γ (*selC*) and α (*selC*) [36]. The LEE inserted at *pheV* and *pheU* are very similar and LEE inserted at *pheV* contain traces of *pheU* regions suggesting that once inserted in an *E. coli* genome, the LEE may have shifted from the *pheU* to *pheV* locus, potentially via site-specific recombination [36].

Phylogenetically similar *E. coli* serotypes have been demonstrated to harbour the same LEE, according to the *eae* subtype and insertion site, and in addition, the same LEE has been detected in different phylogenetic lineages [36]. These findings suggest that the independent acquisition of the LEE pathogenicity island into distinct *E. coli* phylogenetic lineages has occurred multiple times and that intimin subtypes may be a good marker for *E. coli* phylogenetic lineages [36].

The LEE is involved in adhesion of bacteria to intestinal epithelial cells and is required for STEC pathogenesis. However, LEE-negative STEC strains have been implicated in human disease [30]. Recently an 86 kb pathogenicity island, the locus of adhesion and autoaggregation (LAA), has been described which contains a number of virulence factors associated with adhesion [30]. Currently, the LAA has only been detected in LEE-negative strains and it has been suggested to be a recent acquisition with LAA-positive STEC representing an emerging STEC subgroup [30].

1.2.3 Enterohaemolysin

Enterohaemolysin is a pore-forming virulence factor encoded by the *ehxA* gene [45]. Enterohaemolysin is plasmid-associated and such plasmids have been

identified in a number of STEC strains including O145:H28 [46], O157, O26, O103 and O111 [47]. Enterohaemolysin causes haemolysis of washed sheep erythrocytes, which can be easily identified in culture [48] and carriage of enterohaemolysin is associated with Shiga toxin-production for some *E. coli* serotypes [48]. Therefore, enterohaemolysin has been suggested as a marker for STEC [45]. The role of enterohaemolysin in STEC pathogenesis remains to be fully elucidated, however, enterohaemolysin derived from serotype O128:H28 has been shown to increase cytokine interleukin-1β *in vivo* [49]. Interleukin-1β promotes inflammation and is involved in the regulation of expression of Gb3, the Stx host cell receptor [49]. At least six *ehxA* subtypes (A-F) have been identified in *E. coli* using PCR and restriction fragment length polymorphism (RFLP) based methods [50]. Phylogenetic analysis of these subtypes suggests that A and D are associated with *eae*-negative strains and the remaining subtypes are usually associated with *eae*-positive strains [45,50]. Genetic analysis of the *ehxA* subtype D encoding plasmid, phylogenetically the most distinct subtype compared to the others [45], has revealed additional virulence genes associated with other DEC pathotypes [51]; highlighting that further sequence analysis of the *ehxA* encoding plasmids may provide additional insight into the evolution of these plasmids and their role in STEC pathogenesis.

1.3 Epidemiology

1.3.1 Epidemiology and detection methods for serogroup O157

In 1994 *E. coli* O157:H7 was declared an adulterant of ground beef by the USDA-FSIS, following a large foodborne outbreak in North America associated with undercooked beef [12]. The first identified case of STEC associated with human illness in New Zealand was in 1993 and was caused by serotype O157:H7 [52]. In 2016, the majority of notified STEC cases in New Zealand were caused by *E. coli* O157:H7 (41.8% of 491 isolates) [5]. Due to the severity of human disease caused by O157:H7, and it being declared an adulterant of meat, many optimised detection methods for this serogroup have been developed [13,53]. Properties of serotype O157:H7 such as the inability to ferment sorbitol after 24 hours [54] and resistance to antimicrobials such as potassium tellurite [55], enable this serotype

to be selectively-cultured and distinguished from other STEC. Sorbitol MacConkey agar supplemented with cefixime and tellurite (CT-SMAC) was developed utilising these properties, allowing for improved isolation rates of O157 colonies which appear grey compared to sorbitol fermenting colonies that are purple/pink. The inability to ferment sorbitol is also the basis for detection of O157:H7 colonies on media containing 4-methylumbelliferyl-β-D-glucuronide (MUG), in which this serotype does not fluoresce [56]. Recommended detection methods for O157:H7 include (i) the International Organisation for Standardisation method based on indole production, growth on selective media and serogroup-specific latex agglutination test [53]; and (ii) the USDA-FSIS recommended method for isolating this serotype from food products including immunomagnetic separation (IMS), acid treatment and plating onto modified Rainbow agar. Latex agglutination tests are performed on suspect colonies, and once isolated, the colonies are confirmed using either biochemical identification kits, agglutination kits targeting the O157 and H7 antigens or using PCR methods for identifying the O157 serogroup, *stx* and *eae* genes [13]. Due to the increased awareness of non-O157 serogroups implicated in human disease, much work has focused on standardising a detection method for these additional serogroups.

1.3.2 Epidemiology of non-O157 serogroups

The prevalence of non-O157 serogroups in cattle has been widely studied [57,58,59]. A cross-sectional study of 102 dairy farms in spring in New Zealand that used the NeoSeek (Section 1.4.2) detection method, found the prevalence of STEC serogroups O121, O111, O157, O45, O103, O26 and O145 in calves (n=1508) to be 0%, 0.2%, 1.9%, 2.9%, 5%, 7.2% and 9.8%, respectively (Figure 1.2) [1]. A separate study of New Zealand bobby calves that used real-time (RT) PCR as the detection method found the prevalence of *E. coli* serogroups O26, O103, O111 and O145 to be 44.8%, 22.7%, 0% and 15.7%, respectively [60].

A study of feedlot cattle in the USA determined that the summer prevalence for STEC O157, O103, O26 and O145 from pen floor faecal samples (n=24) was 43.1% 1.7%, 1.2%, and 1.0% respectively, during which time STEC O45, O111 and O121 were not isolated [59]. No Top 7 STEC serogroups were detected in winter [59]. Seasonal variation was also observed in feedlot cattle in Canada with serogroups O26, O45, O103, O121 and O157 less prevalent in winter, whereas

the prevalence of serogroups O111 and O145 increased during this time period [61]. The prevalence of the “super six” serogroups varied between geographic location, however little variation was observed for serogroup O157 [61]. In addition to seasonal and geographic variation, detection of the Top 7 serogroups was enhanced using pooled faecal samples; from feedlot cattle in Canada where serogroups O111 and O145 were detected in <10% samples, O121 in >50-<70%, O26 and O157 in >75-<85% and serogroups O45 and O103 detected in >90% of samples [61]. The difference in prevalence between studies could be due to a variety of factors such as study population and different detection methods [15]. Also the most prevalent STEC serotypes causing human disease have been shown to vary between geographic regions [62].

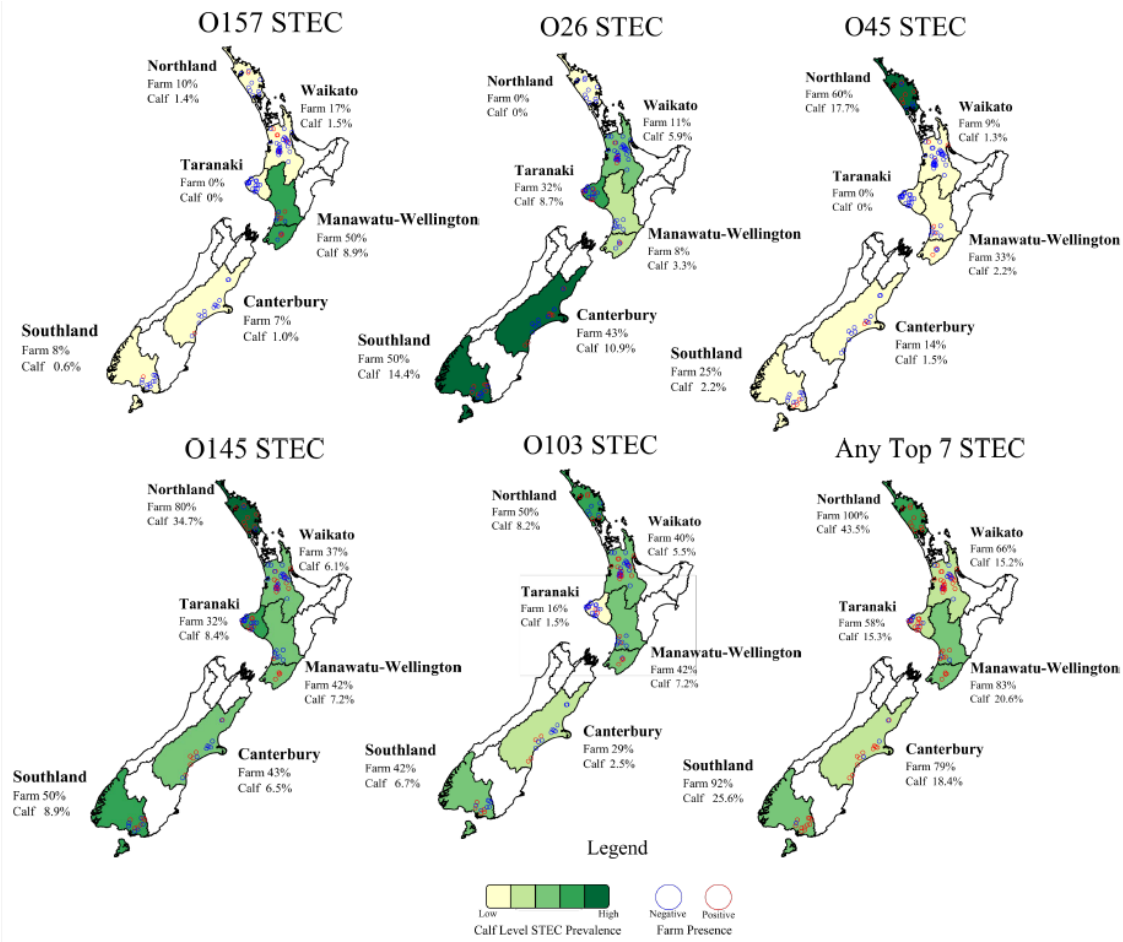

Figure 1.2: Prevalence of the Top 7 STEC serogroups on dairy farms in New Zealand in spring 2014

Shades of green indicate calf level prevalence and a farm was positive if at least one calf was positive on the farm, negative farms had no calves detected as positive. Approximately 15 calves were sampled per farm. Figure provided by Browne et al. [1].

1.4 Current detection methods

1.4.1 Culture-based detection methods

An ideal selective culture medium should have two key features: (i) to enrich the growth of the target pathogen using differential characteristics and (ii) to inhibit the growth of non-target bacteria [63,64]. Current culture-based detection methods for non-O157 STEC are problematic due to the lack of consistent differential characteristics between STEC and non-pathogenic *E. coli* and the phenotypic variation between non-O157 serogroups [16,17]. A variety of selective media are available for the detection and isolation of STEC that utilise carbohydrate fermentation patterns to detect the serogroups based on colony colour [17,65]. Rainbow agar O157 contains potassium tellurite and novobiocin to reduce background flora and chromogenic substrates for *E. coli* associated enzymes causing colour changes allowing for serogroup-specific detection [66]. Recently carbon substrates have been proposed to differentiate serogroups O26, O103, O111, O145 and O157 [67], in conjunction with other previously developed methods [65]. However, the efficacy of a media utilising these substrates is yet to be evaluated [67].

STEC recovery rates on agar media vary between studies [63,68]. CHROMagar™ STEC was able to support the growth of 86.5% of STEC strains tested [58], and of five agar media, CHROMagar™ STEC and modified Posse medium had the highest STEC detection rate of artificially inoculated cattle faecal samples, although these media did not detect all strains analysed [69].

The use of a single media is problematic due to inconsistent serogroup colony colour and some media being too selective [63,64]. The use of two media in parallel, a highly selective media paired with a media that supports the growth of a wide range of STEC, could be more beneficial for the detection of the non-O157 serogroups [58,63,64,68]. Culture methods for non-O157 serogroups generally consist of enrichment of the target pathogen, IMS using serogroup-specific beads and plating onto selective media [17]. Recommended enrichment protocols often differ in the advised selection and enrichment broths, temperature and incubation times [70]. IMS involves immunomagnetic beads (IMBs) that consist of paramagnetic particles coated in antibodies against the target pathogen [71]. A complex forms when IMBs bind to the target and this is separated by a magnetic

field [71] which is then plated onto selective media for further identification. The IMS system relies on the quality of the antibodies and their binding to the target pathogen with monoclonal antibodies having greater stability and specificity compared to polyclonal antibody preparations [71].

The specificity of IMS beads varies between serogroups; in one study at least 88% of IMS complexes contained the target serogroups for O26, O103 and O45 whereas there was low detection of serogroups O121 (50%) and O111 (20%) [16]. A recent study identified that the recovery rate for the “super six” serogroups from artificially inoculated matrices varied between serogroup, matrix and bead brands [72]. The most effective IMS bead preparation recovered serogroups O26, O45, O103, O111, O121 and O145 from cattle faeces at a rate of 67.7%, 100%, 72.2%, 94.4%, 88.9% and 66.7%, respectively [72]. When the growth of the “super six” serogroups was examined with competing bacteria in EC broth, serogroups O26, O111 and O145 grew to a lesser extent, suggesting that these serogroups may be out-competed in enrichments, which may consequently decrease their isolation rate [72]. Sample to sample variation due to the presence of highly competitive naturally occurring bacteria had an impact on isolation efficiency which may account for some of the discordance in IMS between studies [65]. In addition, the efficacy of IMS has been suggested to vary according to the concentration of bacterial cells present, as high cell concentrations may reduce the specificity of IMS beads by favouring non-specific binding [73].

Acid treatment of samples has been reported to reduce non-target bacteria, increasing STEC recovery [74]. Many *E. coli*, including STEC, are intrinsically acid resistance or tolerant [75,76]. No difference in STEC recovery for the “super six” serogroups was identified with and without acid treatment on modified Rainbow agar. However, acid treatment reduced non-target bacteria thus decreasing the number of colonies for subsequent tests (e.g. latex agglutination) [77]. Varying effectiveness of acid treatment for different food test matrices have been observed [64].

Although culture-based methods may be laborious and slower than molecular methods [78,79], the identification of an individual isolate for further analysis is a major advantage [17,68,79]. However, the lack of a refined culture method is associated with highly variable levels of isolates recovered between studies [16,63,68,77].

1.4.2 Molecular based detection methods

The use of molecular methods for the detection of non-O157 serogroups is essential due to the limitations of culture-based methods [80]. Conventional, multiplex or RT-PCR methods have been developed targeting components of the lipopolysaccharide biosynthesis gene cluster including the *wzx* O-antigen flippase and *wzy* O-antigen polymerase genes for the Top 7 serogroups, as well as STEC-associated virulence factors including the *stx* and *eae* genes [79,81]. However the detection of a Top 7 O-antigen gene only confirms the presence of the serogroup and does not indicate the presence of an *stx*-positive target organism [80]. RT-PCR is a rapid method that provides an indication of whether a target organism is present, which is essential for prompt patient treatment and outbreak interventions [79].

Studies evaluating serogroup specific PCRs have demonstrated varying levels of effectiveness, as a result of the methods used, study population and matrices tested. A 7-serogroup specific multiplex PCR (mPCR) was evaluated using spiked cattle faeces and the PCR detected 4.1×10^5 colony forming units (CFU)/g of pooled STEC culture, and two thirds of positive PCR samples were negative using culture methods [80]. Another 7-serogroup mPCR detected all serogroups at 10 CFU/g from spiked cattle faeces, and the primers used were 100% specific for the detection of the “super six” serogroups [16], in contrast other primers used for the identification of STEC were only 30-100% specific in spiked human stool samples [79].

Compared to culture-based methods, a new PCR/ mass spectrometry (NeoSeek) method detected a higher proportion of samples positive for one or more of the “super six” or O157 serogroups [57], although some NeoSeek negative samples were found to be culture positive [57]. Further analysis indicated a statistically significant difference in the performance of the two methods [57]. In comparison, the proportion of faecal samples identified as positive for one or more of the “super six” serogroups did not differ between culture and PCR methods [17]. This may be due to the culture-based methods used in this study which involved the testing of presumptive pooled colonies. However, both detection methods identified positive samples that were recognised as negative by the other method [17]. Using NeoSeek and a culture-based method involving multiple media and

IMS [82] the mean prevalence for the detection of the Top 7 serogroups was 25.9% and 6.5%, respectively [82]. The majority of samples which were non-O157 positive according to the NeoSeek method were negative by culture, highlighting the need for more sensitive and specific culture-based methods for these serogroups [82]. Enhanced detection of the Top 7 STEC serogroup genetic markers with the NeoSeek method may be associated with the concurrent assessment of a large number of bacterial cells from an enrichment. However for culture-based methods only a limited number of bacterial colonies are tested [82]. This discordance highlights the importance of using multiple detection methods in parallel [17,57].

Although molecular based methods are often reliable, they do not provide a bacterial isolate for further analysis which is a major disadvantage. Furthermore, 82 different STEC O-serogroups have been reported in association with human disease [63], but the currently described molecular methods only detect the Top 7 serogroups. This may negatively impact how rapidly novel STEC pathogens may be detected [16,63,79].

1.5 Genomic epidemiology of STEC

1.5.1 Sequencing technologies

A range of next generation sequencing (NGS) platforms are currently available and with the reduced cost and increased speed of such sequencing platforms, the use of WGS is likely to increase [83]. Each sequencing platform has advantages and limitations. Pacific Biosciences sequencing provides longer reads compared to other technologies which is beneficial for genome assembly, however this platform has a higher error rate (~13%) [83,84]. Ion Torrent has a relatively low raw error rate (1.78%) but is not ideal for detecting and interpreting homopolymers [83,84]. Illumina sequencing platforms are routinely used for NGS and have a low error rate (<0.4%) which is a major advantage of these platforms, however the use of short-read data and an increased number of contigs can be difficult for genome assembly [83]. It is estimated that the number of complete sequenced genomes is low, and due to the cost and technical skills required many genomes remain unfinished [85]. The use of a specific sequencing platform

should be evaluated individually for each sequencing project and the platform used depends on cost, biological factors of the organism such as GC (guanine-cytosine) content, expertise of the laboratory and the application of the data, for example whether a low error rate is essential [83,85].

1.5.2 Previous comparative genomic studies

Due to the decreased cost of NGS, increased availability of sequencing platforms and improved downstream analysis software, there has been a rapid increase in the number of bacterial WGS data. WGS data is now available for pathogenic *E. coli* and STEC strains [47,86,87] and these comparative genomic studies often investigate genome composition (including genome length, GC content, number of genes), virulence factors present, HGT events, MLST and a phylogenetic comparison with other public available genome sequences, usually comparing factors such as the core genome [47,87]. There is currently no standardised analysis pipeline for genomic studies, which is an issue when comparing studies as the different parameters used can influence downstream analyses, and in extreme cases the inferences drawn. For example, the use of different assembly and annotation software was shown to influence the number of genes identified for *Vibrio vulnificus* genomes [88]. Current STEC genomic studies have often analysed a small number of strains belonging to a specific serogroup [46,87,89] which may cause strain-specific diversity to be over-looked, although some studies have analysed a large number of STEC strains [90]. With WGS becoming cheaper and the number of publicly available genomes increasing, comparative genomic studies analysing a large and diverse number of bacterial genomes are likely to become more prevalent, providing further insight into the genomic epidemiology of STEC and potentially identifying strain-specific diversity or variation associated with rare STEC serotypes.

1.6 Conclusion

The number of STEC notifications has been steadily increasing in New Zealand, with non-O157 serogroups accounting for an increasing number of cases. A refined detection method for non-O157 serogroups is an essential requirement for both improved public health and maintenance and growth of the New Zealand

meat export industry due to the increase in STEC cases associated with the “super six” serogroups and their classification as adulterants of meat in 2012. There are well-noted benefits of both culture and molecular based detection methods, however their respective limitations routinely result in studies adopting multiple methods in parallel.

From this literature review, future work could investigate differential characteristics of non-O157 serogroups, in particular serogroup O145 which was shown to be highly prevalent at the dairy farm level in New Zealand. For example carbohydrate fermentation or antimicrobial resistance determinants could be investigated. Lastly, a standardised and refined method, like the use of CT-SMAC for O157:H7, could be developed and evaluated for non-O157 serogroups. This would allow for better comparisons between epidemiological studies and for identification of potential STEC interventions along the food value chain. A comparative genomic analysis of a large and diverse panel of serogroup O145 strains would also provide important information regarding the genomic epidemiology of this serogroup.

1.7 Objectives of this study

The purpose of this study was to identify metabolic characteristics which would enable the development of a differential media for serogroup O145, and to further understand the genomic epidemiology of this serogroup. This was achieved by examining carbon utilisation profiles of a diverse range of serogroup O145 strains and using WGS analysis.

The specific objectives of this study were:

(i) To investigate the metabolic characteristics of *E. coli* O145 to facilitate the differential culture of this serogroup.

(ii) To understand the genomic epidemiology of serogroup O145 using WGS.

(iii) Identify phenotype and genotype correlations between the carbon utilisation data and whole genome sequences.

2. Materials and methods

2.1 Subculture

2.1.1 Hopkirk Research Institute culture collection

Bacterial strains used in this study are listed in Appendix A. Bacterial strains used in subsequent analyses were either isolated in this study (described in section 2.2), or were obtained from the Hopkirk Institute culture collection (Appendix A) which contains a range of bacterial strains from many sources and countries, stored at -80°C in glycerol broth.

2.1.2 Subculture from glycerol broth

Bacteria were resuscitated from frozen glycerol broths stored at -80°C and inoculated on either Columbia Sheep Blood agar (5% blood) or CHROMagar™ STEC agar (Fort Richard, Auckland, New Zealand) and streaked for isolated colonies. Agar plates were incubated at 37°C for 18 hours.

2.1.3 Subculture from agar

An individual colony was sub-cultured from one agar plate to another, and incubated at 37°C for 18 hours.

2.1.4 Glycerol broth inoculation

Pure isolates used in this study were resuspended in brain heart infusion (BHI) (Oxoid, Hampshire, United Kingdom) broth containing glycerol (33% [v/v]) and stored at -80°C.

2.2 Culture-based methods

2.2.1 Calf faecal enrichments used in this study

Calf recto-anal mucosal swabs (RAMS) used in this work originated from a New Zealand cross-sectional study across 102 dairy farms [91] that had been previously enriched in modified tryptone soya broth (mTSB) (Oxoid Limited, Hampshire, United Kingdom) for 18 hours at 42°C. Approximately 1 mL of the post-enrichment broth sample was stored at -80°C with glycerol (final

concentration 16.66% [v/v]). Samples were screened for the Top 7 STEC serogroups and associated STEC virulence genes using the NeoSeek confirmation assay (NeoSeek STEC Confirmation, NeoGen Corporation, Lansing, USA). This is a commercial molecular screening test that uses MALDI-TOF mass spectrometry and PCR to identify the Top 7 STEC serogroups according to the presence or absence of a set of target genes [92].

2.2.2 *E. coli* serogroup O145 latex agglutination tests

Serogroup-specific O145 latex agglutination tests were performed according to the manufacturer's instructions (Oxoid, Hampshire, United Kingdom) on ten mauve colonies removed from each respective CHROMagar™ STEC plate. STEC colonies appear mauve, other *Enterobacteriaceae* are blue or colourless and gram positive bacteria are inhibited on CHROMagar™ STEC. Latex positive isolates were sub-cultured onto CHROMagar™ STEC agar and incubated at 37°C for 18 hours. O145 latex positive isolates were stored as described in section 2.1.3.

2.2.3 Immunomagnetic separation (IMS)

IMS was performed from semi-thawed frozen RAMS enrichments from the aforementioned national STEC study [91] which had been identified as serogroup O145-positive using the NeoSeek method. Concentration of *E. coli* serogroup O145 was performed using immunomagnetic beads coated with a polyclonal antibody preparation raised against the O-antigen from *E. coli* O145. IMS was adapted from the manufacturer's instructions (Abraxis, Warminster, USA) with the following changes. To re-enrich the broth, 100 µL of semi-thawed frozen enrichment was inoculated in 10 mL mTSB broth and incubated for 2 hours at 37°C followed by 15 hours at 42°C. Re-enriched culture (1 mL) was removed, added to 20 µL IMBs and processed according to the manufacturer's instructions (Abraxis, Warminster, USA). After the final wash step, the O145 beads were re-suspended in 100 µL E buffer (100 mL buffered peptone water mixed with 0.5 g bovine serum albumin and 50 µL Tween-20; filter sterilised, pH of 7.2 +/- 0.2) and 50 µL of this bead solution was inoculated onto CHROMagar™ STEC agar for individual colonies. The inoculated plates were incubated at 37°C for 18 hours,

and ten mauve colonies per plate were subjected to O145 latex agglutination tests, as described in section 2.2.2.

2.3 DNA extraction

2.3.1 Crude DNA extraction

DNA for subsequent PCR reactions was extracted from *E. coli* strains by adding 3-4 well-spaced colonies to 400 µL sterile molecular biology-grade water. The bacteria were heated at 100°C for 10 min in a heating block and stored at -20°C.

2.3.2 QIAamp DNA mini kit extraction

E. coli serogroup O145 isolates to be genome sequenced in this project were inoculated on Columbia Sheep Blood agar (Fort Richard, Auckland, New Zealand) and incubated at 37°C for 18 hours. A single colony from these plates was sub-cultured onto a fresh Columbia Sheep Blood agar plate and incubated at 37°C for 18 hours to ensure purity of a single clonal isolate. DNA was extracted from these cultures using the QIAamp DNA mini kit (QIAGEN, Hilden, Germany), according to the manufacturer's instructions for "DNA purification from tissues" with several modifications to optimise the protocol for *E. coli*. Briefly, approximately one third of a 10 µL loop of bacteria were suspended in 180 µL ATL buffer, 20 µL proteinase K was added, mixed, and the sample incubated at 56°C for 1.5 hours on a shaking heating block (Provocell™, ESCO, Singapore). Next, 10 µL RNase A was added and incubated for 1 hour at 37°C. Buffer AL (200 µL) was added, mixed and incubated for 10 min at 70°C. After incubation, 200 µL ethanol (96-100%) was added and the sample (including the precipitate) was added to a QIAamp Mini spin column. The spin column was centrifuged for 1 min at 6200 *g*, the filtrate discarded and the spin column placed in a new collection tube. Buffer AW1 (500 µL) was added, centrifuged for 1 min at 6200 *g*, the filtrate discarded and spin column placed in a new collection tube. This step was repeated once more using Buffer AW2. The spin column was centrifuged for 2 min at 17,000 *g*, the filtrate discarded and the spin column placed in a new collection tube. Molecular biology-grade water (75 µL) was added to the spin column which was incubated at 37°C for 5 min. The spin column was centrifuged at 6200 *g* for 1 min to elute the DNA. The DNA sample (filtrate) was then stored at -20°C.

2.4 DNA quantification

2.4.1 Nanodrop

The Nanodrop microvolume spectrophotometer (Nanodrop 3300, Thermo Fisher Scientific, New Zealand) was used to provide an approximate DNA concentration for purified PCR products prior to Sanger dideoxy sequencing. The Nanodrop was initialised with 2 µL molecular biology-grade water, and blanked using 2 µL molecular biology-grade water, as this was the elution buffer used. Each DNA sample was then quantified by adding 2 µL to the Nanodrop, and the approximate DNA concentration measured.

2.4.2 Qubit

DNA extractions of isolates for WGS were quantified using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, New Zealand) on a Qubit 2.0 fluorometer (Thermo Fisher Scientific, New Zealand), according to the manufacturer's instructions. Briefly, the Qubit working solution was prepared by diluting the dsDNA HS reagent 1:200 with dsDNA HS buffer. The Qubit 2.0 was calibrated using 190 µL Qubit working solution mixed with 10 µL standard, for both standards provided. To analyse the samples, 198 µL of Qubit working solution was mixed with 2 µL sample DNA, the samples were vortex mixed and incubated at room temperature for 2 min before reading the results. The DNA concentrations were reported in µg/mL.

2.5 Polymerase chain reaction (PCR)

2.5.1 Multiplex virulence PCR and O145 serogroup-specific PCR

PCR amplifications were performed in 20 µL reaction volumes which contained 10 µL KAPA HiFi HotStart ReadyMix (KAPA BioSystems, Wilmington, USA), 0.6 µL of each primer (10µM stocks) (Table 2.1), 1 µL crude DNA template and 4.2 µL or 7.8 µL molecular biology-grade water for the STEC virulence mPCR and O145 serogroup-specific PCR, respectively. PCR was undertaken on a Bio-Rad T100 Thermal Cycler using the following conditions: initial denaturation at 95°C for 3 min, then 30 cycles of 98°C for 20 sec, 65°C for 15 sec, and 72°C for 30

sec, with a final extension step at 72°C for 1 min. PCR products underwent electrophoresis on a 2% agarose gel and stained as described in section 2.6.1.

2.5.2 PCR amplification of *eae*

Amplification of a 1.85 kb region of the *eae* gene for subsequent sequence analysis was performed in 25 µL reaction volumes, with 12.5 µL KAPA HiFi HotStart ReadyMix, 0.75 µL of each primer (10µM stock), 10 µL molecular biology-grade water and 1 µL crude DNA template. PCR was undertaken on a Bio-Rad T100 Thermal Cycler using the following conditions: initial denaturation of 95°C for 3 min, then 30 cycles of 98°C for 20 sec, 56°C for 30 sec, and 72°C for 90 sec, with a final extension step of 72°C for 5 min. PCR products underwent electrophoresis on a 0.8% agarose gel and stained as described in section 2.6.2.

Table 2.1: PCR primer sequences and resulting amplicon lengths

Primer	Primer Sequence (5' → 3')	Amplicon bp	PCR used in	Reference
stx1 F	ATAAATCGCCATTCGTTGACTAC	180	2.5.1 virulence	[93]
stx1 R	AGAACGCCCACTGAGATCATC			
stx2 F	GGCACTGTCTGAAACTGCTCC	255	2.5.1 virulence	[93]
stx2 R	TCGCCAGTTATCTGACATTCTG			
eae F	GACCCGGCACAAGCATAAGC	384	2.5.1 virulence	[93]
eae R	CCACCTGCAGCAACAAGAGG			
ehxA F	GCATCATCAAGCGTACGTTCC	534	2.5.1 virulence	[93]
ehxA R	AATGAGCCAAGCTGGTTAAGCT			
O145 F (*wzy*/ *wzx* gene)	GCGGGTGTTGCCCGTTCTGT	766	2.5.1 O145 serogroup	[16]
O145 R (*wzy*/ *wzx* gene)	ACGGCATTCCGCTGCGAGTT			
intRFLP F	GATTCWAAACTRTTAACTCA	1848	2.5.2 *eae*	[39]
intRFLP R	AGCHTTAATCTCAGTAATRCT			

2.6 Gel electrophoresis

2.6.1 2% w/v agarose

A 2% agarose gel was prepared by adding 1.5 g of agarose (Invitrogen, Auckland, New Zealand) to 75 mL of 0.5x Tris-borate-EDTA (TBE) buffer. The agarose was mixed and dissolved in a microwave upon heating. Once dissolved, 3.75 µL RedSafe (Invitrogen, Auckland, New Zealand) was added, and when cooled the agarose was poured into a gel electrophoresis tray. Once solidified, the electrophoresis tray was submerged in an electrophoresis gel tank tray containing 0.5x TBE buffer. PCR product (2 µL) was mixed with 1 µL gel loading buffer (BlueJuice™, ThermoFisher Scientific, New Zealand), and 1.5 µL of this solution was loaded into a well in the agarose. A 1kb+ ladder (Invitrogen, Auckland, New Zealand) (1.5 µL) was electrophoresed as a size standard. Electrophoresis was undertaken at 80 V for 40 min, and gel images captured and stored using a GelDoc XR$^+$ (BioRad, New Zealand).

2.6.2 0.8% w/v agarose

A 0.8% agarose gel was prepared using a similar method as described in section 2.6.1, except that 0.6 g agarose was added to 75 mL 0.5x TBE buffer.

2.7 Intimin (*eae*) subtyping

2.7.1 *eae* PCR and PCR product visualisation

Approximately 1848 bp of the *eae* gene was amplified using the intRFLP F and intRFLP R primers as described in section 2.5.2, and the PCR products were visualised on a 0.8% agarose gel as described in section 2.6.2.

2.7.2 PCR product purification and quantification

The remaining PCR product was purified using the QIAquick PCR purification kit (QIAGEN, Hilden, Germany) according to the manufacturer's instructions. PCR products were quantified using spectrophotometry (Nanodrop 3300, Thermo Fisher Scientific) as described in section 2.4.1 and each PCR product was diluted using molecular biology-grade water to a concentration of 12.5 ng/µL.

2.7.3 Sanger dideoxy sequencing PCR

Separate sequencing reactions were performed in 10 µL reaction volumes, using the same intRFLP F and intRFLP R PCR primers, with 1.75 µL buffer, 0.5 µL Big Dye™ Terminator v3.1 (Thermo Fisher, New Zealand), 1 µL primer (3.2 pm/µL), 4.75 µL sterile molecular biology-grade water and 2 µL DNA template (12.5 ng/µL). The sequencing PCR was undertaken on a Bio-Rad T100 Thermal Cycler using the following conditions: 95°C for 1 min, then 25 cycles of 95°C for 10 sec, 50°C for 10 sec, and 60°C for 90 sec.

2.7.4 Determining the *eae* subtype

The purified *eae* PCR products were separately sequenced using the intRFLP F and intRFLP R primers according to the PCR described above. Sanger dideoxy sequencing of single strand amplicons was performed by the Massey Genome Service (Massey University, Palmerston North, New Zealand) using an ABI3730 DNA analyser (Applied Biosystems). ABI traces were trimmed at the proximal 5' and distal 3' end to remove poor quality sequence using Geneious 8.1 [94], and any nucleotide matches identified using BLASTN [95].

2.8 Biolog phenotypic microarray assays

2.8.1 Inoculation of microarray assay plates

E. coli serogroup O145 strains were inoculated on Columbia Sheep Blood agar (Fort Richard, Auckland, New Zealand) and incubated at 37°C for 18 hours. Omnilog inoculation fluid was prepared by combining 20.83 mL IF-0a solution (Biolog Inc, California, USA), 0.30 mL Redox Dye A (Biolog Inc, California, USA) and 3.87 mL molecular biology-grade water. Bacterial growth from well-separated colonies was added to 5 mL IF-0a solution and 1 mL sterile molecular biology-grade water to reach 42% transmittance using a turbidimeter (Biolog Inc, California, USA). Bacterial suspension (5 mL) was added to the Omnilog inoculation fluid and 100 µL of this was inoculated into each PM MicroPlate™ well (Biolog Inc, California, USA). PM1 and PM2A MicroPlates™ were used, each having 96 wells, with the A1 position on each plate not containing a carbon source, which was used as an internal negative control and the remaining 95 wells containing a different carbon substrate [96]. The inoculated MicroPlates™

were incubated in the Omnilog instrument at 37°C for 24 hours. The dye intensity, which is proportional to the level of cell respiration, was measured at 15 min intervals and recorded in Omnilog Units. A PM1 and PM2A MicroPlate™ inoculated with the bacterial diluent solution only was used as a negative control to ensure no false positive reactions between the dye and the diluent. Replicate and duplicate plates were also included to check reproducibility. Replicate plates were inoculated on different days from the same -80°C glycerol stocks, whereas duplicate plates were inoculated from separate cultures on the same day.

2.8.2 Analysis of phenotypic microarray assays

The raw Omnilog data was converted to OKA files, and imported into the proprietary Omnilog software (Biolog Inc, California, USA) where the level of utilisation of each carbon substrate during incubation could be analysed per strain. Using the Omnilog software, the files could be exported as comma separated value (csv) files which were analysed using R version 3.3.1 [97] and the packages ‘opm’ [98] and ‘gplots’ [99]. High resolution individual XY plots for each strain were produced indicating the level of utilisation of each carbon substrate during incubation. XY plots for comparing replicate strains were also produced. Cluster analysis was performed to produce a dendrogram, as were heat-maps with strains clustered according to the metabolic profile similarities, allowing comparison between the isolates. For the dendrogram the end-point dye intensity values for each carbon substrate were used. For the heat map, end-point dye intensity values representing carbon substrate utilisation (ranging from 0-300) were grouped into three ranges: 0-≤50 indicating a no substrate utilisation; 51-≤150 indicating moderate utilisation; and 151-≥400 indicating extensive substrate utilisation. A heat map was generated with these three groupings rather than a traditional gradient range to provide a clearer indication of which substrates could or could not be utilised. Patterns were visualised by eye using the heat-map and dendrogram, according to the metabolic profiles that allowed clustering of a certain subset of serogroup O145 strains (for example, the same virulence profile or the same *eae* subtype). The R code used in this analysis is included in Appendix B.

2.9 Whole genome sequencing (WGS)

2.9.1 DNA extraction, quantification and dilutions

The QIAamp DNA mini kit (QIAGEN, Hilden, Germany) extraction method was used for isolates to be genome sequenced in this project, as described in section 2.3.2. The DNA concentrations were determined using a Qubit 2.0 fluorometer (Thermo Fisher Scientific, New Zealand), as described in section 2.4.2 and the concentrations were reported in µg/mL. The DNA samples were sequentially diluted to the desired concentration for making library preparations of 0.16 µg/mL using molecular biology-grade water. Diluted DNA was stored at -20°C.

2.9.2 Library preparations

In-house library preparations were made using the Nextera XT DNA library preparation kit (Illumina, San Diego, USA), according to the manufacturer's instructions with modifications to optimise the protocol for *E. coli*. The Nextera XT DNA library preparation kit uses an engineered transposome to tagment the genomic DNA and also tag the DNA with adapters [100]. For the tagmentation process, 10 µL of tagment DNA buffer was added to 5 µL amplicon tagment mix in a PCR tube. To this, 5 µL genomic DNA (normalised to 0.16 µg/mL) was added and the samples were run through a tagmentation step at 55°C for 12 min. Immediately after the tagmentation protocol was finished, 5 µL of neutralise tagment buffer was added to stop the process. To amplify the libraries, a limited-cycle PCR protocol was used which utilised the adapters added during tagmentation to amplify the DNA and also added the unique index adapters to the libraries; which enabled sequencing of pooled libraries on Illumina sequencing platforms [100]. Two unique index adapter combinations, 5 µL of both index primer one and index primer two, were added to each reaction and 15 µL of Nextera PCR master mix was added. The PCR was run on a SensoQuest Lightcycler PCR (SensoQuest, Germany) machine using the following conditions: 72°C for 3 min, 95°C 30 sec, then 12 cycles of 95°C for 10 sec, 55°C for 30 sec, 72°C 30 sec, followed by 72°C for 5 min.

To clean the library and remove excess adapters, 50 µL of PCR product was mixed with 50 µL AmPure XP beads (Beckman Coulter, USA). The tubes were incubated for 5 min at 25°C on a shaking heating block (Provocell™, ESCO,

Singapore) at 1500 rpm. The tubes were placed on a magnetic stand and the supernatant was removed once the suspension became clear. The pellet was washed twice with 80% molecular grade ethanol and the pellet air dried. The pellets were re-suspended in 55 µL resuspension buffer for 5 min on a shaking heating block (Provocell™, ESCO, Singapore) at 1800 rpm. The tubes were placed back on the magnetic stand and the supernatant was transferred to a new tube and stored at -20°C.

2.9.3 Pooling individual library preparations

Individual library preparations were combined (pooled) in equimolar ratios to ensure even sequencing coverage across all samples. Sample molarity and therefore pooling volumes were determined according to an average double-strand DNA molecule size of 1000 bp within each library preparation, determined from the results of the Bioanalyzer 2100 quality control described in section 2.9.4. The minimum library concentration required for Illumina HiSeq sequencing is 4.24 ng/µL, with the equivalent molar concentration of 6.24 nM. Samples were therefore combined and diluted (using molecular biology-grade water) into a total volume of 300 µL, and the required volumes were determined using the calculations shown below. The final concentration and average molecular size of the pooled library was verified using a Qubit 2.0 fluorometer (described in section 2.4.2) and assessed using the high-sensitivity DNA kit (Agilent Technologies, Santa Clara, USA) on a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, USA) (described in section 2.9.4), respectively, to ensure that the correct molar concentration of DNA was used in the sequencing reaction.

Sample calculations:

Concentration of individual library required = a / (b x c)

Volume of individual library required = d x e / f

a = Individual library concentration (µg/mL)

b = 4.24 µg

c = 6.24 nM

d = Desired concentration (nM)

e = Desired library volume (µL)

f = library concentration (nM)

2.9.4 Library preparation quality controls

To ensure the library preparations produced fragments of the desired size (250-1500 bp) and had low RNA and protein levels (RNA must be <10% of the total nucleic acid present in the sample and for the same for the level of protein per sample), aliquots of the first four library preparations and an aliquot of the pooled libraries, were sent to New Zealand Genomics Limited (NZGL, Massey Genome Service, Massey University, Palmerston North, New Zealand) and analysed as a quality control using the Bioanalyzer 2100 (Agilent Genomics, Santa Clara, USA).

2.9.5 Whole genome sequencing (WGS)

The WGS was performed by New Zealand Genomics Limited (University of Otago, Dunedin, New Zealand) using the Illumina HiSeq platform with v4 chemistry (2 x 125 bp). Illumina HiSeq uses sequencing by synthesis technology which relies on reversible terminator fluorescently labelled deoxyribonucleotide triphosphates (dNTPs). In the first step of the sequencing pipeline, the pooled libraries are applied to a flow cell where the fragments attach to surface bound oligonucleotides, complementary to the adapters on the libraries [101]. Via bridge amplification, fragments are amplified into clonal clusters. Fluorescently labelled dNTPs are incorporated by DNA polymerase and fluorophore excitation can be used to determine which base is incorporated during multiple cycles of DNA synthesis [101]. Raw error rates and incorporation bias are reduced as all four dNTPs are present during the sequencing cycles. Using lllumina HiSeq technology, the libraries can be read in parallel, providing an efficient and cost-effective sequencing method [101].

2.10 Whole genome sequencing quality assessment and genome assembly

2.10.1 Proprietary Illumina sequencing report

The quality of the sequencing reads was first analysed using the proprietary Illumina sequencing report, MultiQC v.01, provided by NZGL. This report utilises FastQC v0.11.5 [102] and provided information on the sequence quality length

and per base content, GC content, adapter contamination, sequence length distribution, and any duplication issues or over-represented sequences. This report identifies any potential issues with individual libraries that would be investigated further using more thorough quality assessment tools.

2.10.2 QCtool

The raw sequencing reads were run through an in-house quality control software (QCtool) [103] which gives an indication of the quality of the sequencing reads. The QCtool utilises base quality scores to trim and filter reads, determine primer contamination and GC content and runs FastQC [102], SolexaQA++ v3.1.7.1 [104], and FastQScreen v0.11.1 [105]. Fast QC [102] provides quality control checks and indicates any areas of the sequencing data which may have quality issues by providing summary graphs indicating the sequencing quality scores, GC content, per base N content, sequence lengths, sequence duplications and any over-represented sequences [102]. SolexaQA++ [104] is used to evaluate the output reads, assess sequence quality and provide sequence quality statistics. For example, SolexaQA++ provides information on base call errors, the mean quality of the sequences in the dataset, and identifies regions of the sequences with poor quality which can be further investigated using other analysis software [104]. SolexaQA++ also contains the program DynamicTrim, which enables the sequences to be trimmed to their longest continuous read segment at a given quality threshold [104]. FastQScreen [105] is a quality control tool which allows you to screen Fastq files against a set of known genomes to confirm the composition of the library matches to other genomes of the organism being sequenced. FastQScreen can also check for other contaminants such as vector sequences and Illumina adapters [105].

2.10.3 SPAdes

The WGS were *de novo* assembled using SPAdes v3.9.1 [106]. SPAdes assembles the sequences in four stages: assembly graph construction, *k*-bimer adjustment, paired assembly graph and contig construction [106]. Stage one, assembly graph construction, involves implementing algorithms to produce multi-sized de Bruijn graphs, allowing for different *k*-mer sizes. This produces distance histograms of aggregated bi-read information. Stage two, *k*-bimer adjustment,

utilises distance histograms to estimate the distance between *k*-mers in the genome. Stage three constructs the paired assembly graph and stage four is the construction of contigs [106].

SPAdes uses a reference genome, which the assembled genome can be compared to, to identify significant assembly errors. When selecting a reference genome, two main factors should be considered: (i) quality of the genome and (ii) similarity of the reference genome compared the genomes of interest. Ideally, a reference genome would be a complete sequence, sequenced using technology with a low error rate. Secondly, it is important to have a reference that is genetically similar to the genomes of interest for certain post-assembly analyses, such as Snippy [107]. When selecting a reference genome, it is important to understand and acknowledge the limitations. In this analysis, RM12761 was used as the reference genome, which is a STEC O145:H28 strain isolated from ice-cream in Belgium (accession no. NZ_CP007133) [46]. This isolate was selected as the reference as it only has three contigs (a chromosome and two large plasmids), has a virulence profile similar to the majority of the isolates in this study (*stx*-positive, *eae* subtype γ) and is a published genome. A limitation of this reference genome is that it is not a New Zealand isolate, however, there are currently no publicly available genome sequences from STEC O145 isolated in New Zealand.

2.10.4 QUAST

The quality assessment tool QUAST v4.4 was used to assess and compare the quality of the genome assemblies and can be used with or without a reference genome [108]. QUAST identifies a range of assembly statistics that give an indication of the quality of the assemblies, and provides an indication of any errors in an assembly that may warrant further investigation. The assemblies are aligned to a reference genome using the Nucmer aligner from MUMmer (v3.0) [109], and then QUAST identifies assembly statistics in the four categories of (i) contig size, (ii) misassemblies and structural variations, (iii) genome representations and (iv) assembly statistics based on the contigs such as N_{50} (an assembly statistic that is the length of the smallest contig which represents at least 50% of the assembly) [108] that can be used to draw inferences on the quality of the assemblies. Contig sizes includes factors such as the total number of contigs, the largest contig and

the total number of bases. QUAST identifies structural errors in the assemblies by reporting either unaligned contigs or ‘misassembled’ contigs, which could be either due to misassemblies or structural variations such as rearrangements. The genome representations category includes information such as the GC content, the number of genes and number of operons.

2.11 Single nucleotide polymorphism (SNP) analysis

2.11.1 SNP analysis

SNPs were identified in the paired-end sequencing reads using Snippy v3.0 [107], executed using in-house Perl scripts (Appendix C) [110]. Snippy uses a reference genome to help identify a set of core SNPs and indels (insertions/deletions) between the sequencing reads and a reference genome. If the reference genome is very different it will reduce the size of the core SNPs being analysed, highlighting the importance of selecting a reference genome similar to the genomes of interest. Snippy first identified SNPs individually for each pair of sequence reads and the reference genome, and then Snippy can be run to identify a core SNP alignment for all of the genomes (when compared to the same reference genome) to infer phylogeny [107]. The core SNP alignment was viewed in SplitsTree v4.14.4 [111] using Neighbor-joining methods.

2.12 Genome annotation and comparative analysis

2.12.1 Prokka

The assembled genomes were annotated using Prokka v1.12 [112], executed using in-house Perl scripts (Appendix D) [113]. Prokka uses external programmes for predicting genomic features, and annotation of the genomes includes predicting co-ordinates of coding regions and their putative products [112]. The external programmes used are: Prodigal (v2.6.3) [114] which predicts coding sequences; RNAmmer (v1.2) [115], ribosomal RNA genes; ARAGORN (v1.2) [116], transfer RNA genes; SignalP (v4.1) [117], signal leader peptides and Infernal (v1.1.2) [118], non-coding RNA. Prokka annotates the genomes in two steps, (i) Prodigal [114] predicts co-ordinates of coding genes and (ii) identifies

the putative gene product by comparing the sequence to known sequences in a database [112]. Prokka searches the databases to assign an annotation using a hierarchical system, initially searching from reliable databases through to domain-specific databases. For each database the *e*-value threshold of 10^{-6} is used [112]. Examples of databases utilised by Prokka include user-provided databases, UniProt [119], RefSeq, Pfam [120], TIGRFAMs [121], and if the sequence does not match any known proteins in these databases it is labelled as a hypothetical protein. Prokka outputs a range of files such as fasta files of the genomic sequences, summary statistics and an annotated GenBank file [112].

2.12.2 Center for Genomic Epidemiology

The Center for Genomic Epidemiology (CGE) [122] is an online service that contains a range of databases and tools for WGS analysis. The Bacterial Analysis Pipeline [123] was used to upload the assembled genomes and corresponding metadata, which then ran the genomes through various other CGE databases. These databases included the SpeciesFinder v1.2 [124] and SerotypeFinder v1.1 [23] which were used to confirm that the WGS belonged *E. coli* serogroup O145 and identify the H antigens carried by each genome, if present. The CGE VirulenceFinder v1.5 [125], PlasmidFinder [126] and MLST v1.8 [24] were also used to identify various virulence factors, *E. coli* plasmid types and the ST type, respectively.

2.12.3 Comparison of virulence genes

The presence or absence of 31 virulence genes (identified using the VirulenceFinder, section 2.12.2) were used to make a Neighbor-Net tree according to Euclidean distance (A/Prof. Patrick Biggs). The tree was subsequently edited using the Interactive Tree of Life (iTOL) v4.1 and isolate metadata was included for *eae* subtype, ST and isolation source as indicated by the colour keys. The presence and absence of the virulence factors that differ between isolates (n=23) was indicated by the matrix. Virulence factors that were either present or absent in all of the strains were not shown in the matrix but were used to construct the tree.

2.12.4 Ribosomal multi-locus sequence typing (rMLST)

In silico analysis of rMLST was performed based on classification described by Jolley et al. [127] using in-house Perl scripts (Appendix D) [113]. Briefly, *in silico* rMLST analysis uses the BIGSdb database which contains all of the known diversity among the ribosomal protein genes. New alleles are periodically added to the database, and checks are in place to ensure there is no redundancy. Unique sequences in the database are designated an arbitrary allele number [127], and a combination of alleles are used to distinguish and identify different rMLST types. rMLST provides an indication of phylogenetic relationships using SNPs identified in the 53 genes encoding the ribosome protein subunits (*rps*, *rpm* and *rpl*). These genes are found in all bacteria and are under constant selection for functional conservation and therefore are used to resolve bacterial phylogenies [127]. This method is also robust against horizontal genetic exchange. The *in silico* rMLST analysis produced a range of outputs, and the alignments can be viewed in SplitsTree v4.14.4 [111].

2.12.5 Identification of the locus of enterocyte effacement (LEE) pathogenicity island integration sites and *stx*-bacteriophage insertion sites

The LEE pathogenicity island integration sites were identified using two methods: (i) visualisation of the GenBank files in Geneious [94] and identification of the *eae* coding sequences (CDS), *tir* and other LEE-encoded genes including a prophage integrase next to the likely tRNA gene (*selC*, *pheU* or *pheV*) integration site, or (ii) the contigs were assembled to a reference genome and the likely tRNA integration site predicted based on the mapped contigs and gene synteny. The reference genomes used are shown in Table 2.2. The *stx*-bacteriophage insertion sites were also detected by identifying either (i) the insertion of *stx* genes in known insertion sites for *stx*-positive strains, or (ii) lack off disruption of known insertion sites to indicate these were available in *stx*-negative strains.

Table 2.2: Reference genomes used to identify LEE pathogenicity island integration sites.

Serotype	Strain	LEE integration site	Accession no.	Publication
O145:H28	RM13514	tRNA *selC*	NZ_CP006027	[46]
O145:H28	RM12761	tRNA *selC*	NZ_CP007133	[46]
O26:H11	11368	tRNA *pheU*	AP010953	[47]
O103:H2	12009	tRNA *pheV*	AP010958	[47]

2.12.6 Download of publicly available serogroup O145 raw read data

Publicly available serogroup O145 WGS data (Appendix E) were identified to offer a global comparison to New Zealand serogroup O145 strains sequenced. Serogroup O145 strains were identified from NCBI [128], EnteroBase [129] and published papers. Only whole genome sequences in which the raw read sequence data was publicly available were analysed in this study using the same pipeline (quality assessment and assembly) for improved comparison. The raw read sequence data was downloaded using the SRA toolkit [130]. Publicly available WGS data was excluded from the analysis if any discrepancies were identified during the quality assessment such as an unexpected genome size or GC content indicative of potential contamination.

2.12.7 Identification of orthologous groups

GET_HOMOLOGUES (v20170302) [131], a program used for comparative genome analysis, was used to identify orthologous groups from WGS data. The sequence-clustering algorithms COGtriangles [132] and OrthoMCL [133] were used to identify orthologous groups and distinguish paralogues, according to BLAST reciprocal best hit results [131]. Both sequencing-clustering algorithms were used to obtain a consensus of the cluster sets. GET_HOMOLOGUES was also used to determine the genome composition of the strains, by estimating the core and pan genome size by random sampling of the genomes [131] and by calculating the cloud, shell and core genome as defined by Tettelin et al. [134]. The definition of these genome statistics are: core, genes found in all genomes being investigated; soft-core, genes found in 95% genomes [135]; cloud, genes

found in a few genomes (the second most populated gene cluster); and shell, the remaining genes present in multiple genomes [131]. The core and pan genome identified by GET_HOMOLOGUES [131] was compared to the core and pan genome identified by Roary (v3.8.2) [136] as described in section 2.13.1.

2.12.8 BEAST

BEAST 2.0 (v2.4.6) [137], which uses Bayesian evolutionary analysis, was used to estimate the time of most recent common ancestor (TMRCA) of aligned core genomes based on mutations rates, and was calibrated using the isolation date of each strain. Briefly, the core SNP alignment obtained using Snippy v3.0 [107] was analysed using Gubbins (v2.2.2) [138], an algorithm that removes regions containing increased densities of base substitutions. This core SNP alignment, with removed recombination regions, was used as the input for BEAUti, a program to set the parameters for BEAST 2.0. For BEAST 2.0, the following parameters were used: coalescent constant population, the Hasegawa-Kishino-Yano (HKY) substitution model, a relaxed molecular clock model, effective sample sizes (ESS) greater than 100 were accepted and the Markov Chain Monte Carlo (MCMC) chain was run for 50,000,000 iterations. A 10% burn-in was used and the data was stored every 10,000 iterations. Summary statistics were analysed using the BEAST 2.0 program Tracer, and the tree estimates were visualised using FigTree v1.4.3 [139].

2.13 Comparison of phenotypic and genotypic data

2.13.1 Identification of the core and pan genome

Roary (v3.8.2) [136] was used to identify the pan genome and the core and accessory genes in the *E. coli* serogroup O145 strains. Prokka [112], GFF3 output files were used as the input for Roary. Roary uses protein sequences, which are iteratively pre-clustered with CD-Hit [140], and subsequently subject to an all-against-all comparison using BLASTP (at as 95% identity). These results were combined using the Markov cluster (MCL) algorithm [141] and homologous groups identified using BLAST. Roary analysis then removes paralogous genes using the conserved gene neighbourhood for each gene. From this, the

presence/absence matrix of the accessory genome is produced, as well as identification of the core genes [136].

2.13.2 Interrogation of the pan genome

Scoary (v1.6.16) [142] is a program that identifies associations between genes in the accessory genome, as identified in the pan genome presence/absence matrix from Roary [136], and traits of interest defined by the user. Scoary also calculates the strength of these associations and reports various statistics such as p-values, adjusted p-values and odds ratios per trait, as well as annotation information for each gene [142]. Scoary was used to calculate the association between genes in the pan genome matrix and the *eae* subtypes carried by the strains (either γ or non-γ), their isolation source (either human or animal origin), and whether they were toxigenic (either *stx*-positive or *stx*-negative). For carbon substrate utilisation, each 'trait' was defined as whether the strains were either able or unable to utilise the specific carbon substrate. Scoary was then used to identify genes in the pan genome which were associated with the carbon utilisation observed using the Omnilog phenotypic microarray system.

2.13.3 Identification of genes associated with carbohydrate metabolism

To distinguish the function of genes identified by Scoary described in section 2.13.2, BlastKOALA (v2.1) [143] was used. BlastKOALA is an online tool which assigns KEGG (Kyoto Encyclopedia of Genes and Genomes) Orthology (KO) to individual genes. The genes can then be separated into functional categories that can be used to reconstruct KEGG pathways and BRITE hierarchies to determine gene function [143]. BlastKOALA was used as a tool to filter the genes identified in section 2.13.2 into functional groups, with a focus on genes involved in carbohydrate metabolism.

3. Results - Isolation of *E. coli* serogroup O145

3.1 Isolation of *E. coli* serogroup O145

Serogroup O145 strains were isolated from calf faecal enrichments using a combination of molecular and culture-based methods to identify additional strains for WGS (Chapter 5). The calf faecal enrichments were collected as part of a spring 2014 cross-sectional study on 102 dairy farms to investigate the prevalence of the Top 7 STEC serogroups in young calves throughout New Zealand [1]. The calf faecal enrichments were undertaken by Browne et al. [1] as described in section 2.2.1.

A total of 1508 calf faecal enrichments were collected, and screened for the Top 7 STEC serogroups and associated virulence genes using the NeoSeek STEC confirmation assay (NeoSeek STEC Confirmation, NeoGen Corporation, Lansing, USA) [91]. This molecular method detects the Top 7 STEC serogroups based on the presence or absence of a set of target genes, using MALDI-TOF mass spectrometry multiplexing and PCR [92]. The NeoSeek STEC confirmation assay identified 148 of 1508 faecal enrichments as STEC O145 positive (9.8%) where the limit of detection (LOD) was approximately 10^3 CFU/mL [144]. In parallel, the enrichments were also investigated using RT-PCR for the detection of the Top 7 serogroups and STEC-associated virulence genes (*stx1*, *stx2*, *eae* and *ehxA*) [1]. In-house validation of the serogroup O145 RT-PCR revealed a LOD of 10^2 CFU/ml [145].

As there were a large number of STEC O145 positive calf faecal enrichments, not all were able to be screened for bacterial isolation via culture methods. Therefore, for inclusion in this study three criteria were used: (i) that the enrichment was identified as STEC O145 positive by the NeoSeek STEC confirmation assay, (ii) that the enrichment generated a cycle threshold (C_t) value ≤30 for the serogroup O145 RT-PCR, and (iii) that only one enrichment per calf per shed per farm was included. These criteria were used to enhance successful

isolation of serogroup O145 through the selection of enrichments with a higher abundance of target serogroup, and also to screen enrichments from different animals to maximise the isolation of a broad diversity of strains. In addition to the above criteria, five calf faecal enrichments were excluded from the screening process as serogroup O145 strains had been previously isolated from these enrichments [91]. Consequently, a total of 37 calf faecal enrichments were included for screening (Figure 3.1).

E. coli serogroup O145 strains were isolated in this study using a range of culture-based methods, including direct plating onto CHROMagar™ STEC, and IMS-culture followed by serogroup-specific latex agglutination tests and PCR confirmation (described in sections 2.2.2 and 2.5.1, respectively). Using the criteria described above, 37 calf faecal enrichments were screened using direct plating (n=31) and/ or IMS (n=17). Of the 31 enrichments screened using direct plating, 11 enrichments (in which O145 isolation was initially unsuccessful) were subsequently screened using IMS due to a presumptive low prevalence of O145 according to high RT-PCR C_t values. Fewer enrichments were screened using IMS due to cost, therefore in general this method was used to screen low abundance O145 enrichments with higher O145 RT-PCR C_t values. Details of all enrichments screened are listed in Appendix F.

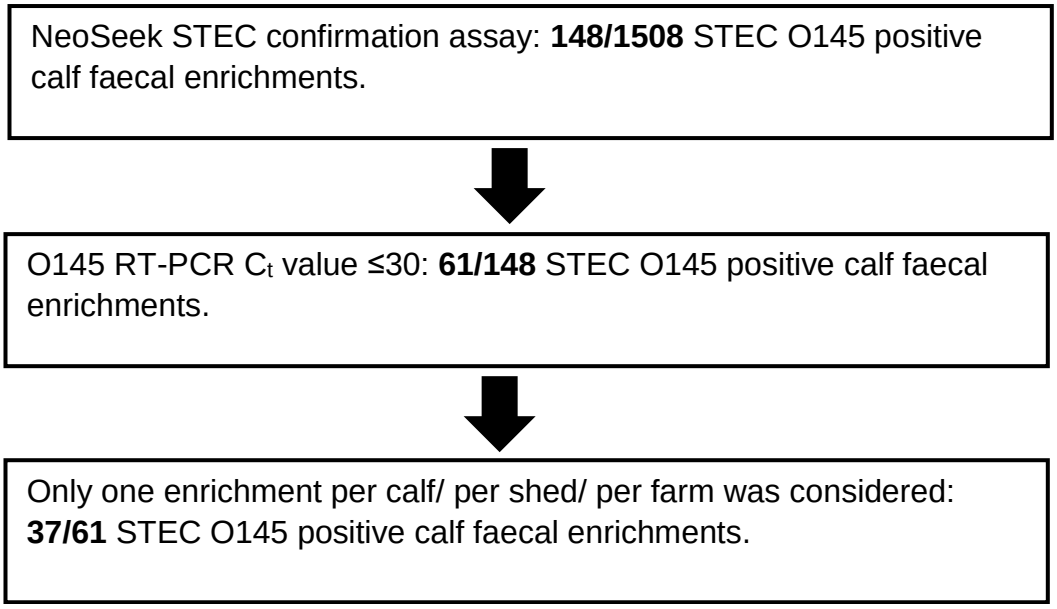

Figure 3.1: Calf faecal enrichment screening process [1]

3.2 Culture-based isolation

Using direct plating onto CHROMagar™ STEC, *E. coli* serogroup O145 strains were isolated from five of 31 (16.1%) enrichments, however when IMS was used in combination with CHROMagar™ STEC, the isolation rate increased to eight of 17 enrichments (47.0%) (Table 3.1). Using direct plating, ten O145 strains were isolated from five enrichments, compared to 22 O145 isolated from eight enrichments using IMS methods and selective plating (Table 3.2). Although no *stx*-positive isolates were identified using direct plating, one *stx*-positive isolate was identified using IMS (3.1% of the total isolates). All serogroup O145 strains isolated (n=32) were both *eae* and *ehxA* positive. The number of *E. coli* serogroup O145 strains isolated using culture-based methods was relatively low compared to the number of enrichments identified as STEC O145 positive using the RT-PCR and NeoSeek molecular based methods (32 isolates were obtained from 13 out of 37 calf faecal enrichments screened, 35.1%).

The average O145 C_t values for contrasting enrichments screened using direct plating and IMS were not significantly different (t-test, p = 0.3378). RAMS enrichments from which O145 was successfully isolated had an average C_t value of 21.86 for direct plating, compared to that of IMS which was 25.20 (t-test p = 0.0563).

In addition to the 32 *E. coli* serogroup O145 strains isolated in this study, 35 O145 strains were obtained from the Hopkirk Research Institute culture collection for

subsequent experiments. Bacterial strains used in this study are listed in Appendix A.

Table 3.1: Serogroup O145 isolation from calf faecal enrichments using culture-based methods

Isolation method	Number of enrichments screened	Enrichments isolation successful†	O145 PCR confirmed isolates‡	O145 STEC‡	O145 isolates *eae*, *ehxA* positive‡
Direct plating	31	5 *(16.1%)*	10	0 *(0%)*	10 *(100%)*
IMS	17	8 *(47.0%)*	22	1 *(4.5%)*	22 *(100%)*
Total	37	13 *(35.1%)*	32	1 *(3.1%)*	32 *(100%)*

†: Culture-based isolation was successful if ≥1 serogroup O145 strain was isolated from the calf faecal enrichment.

‡: Serogroup confirmation and virulence profile were determined using mPCR for the O145 *wzy*/*wzx* gene, *stx1* and/or *stx2*, *eae* and *ehxA*, respectively.

Table 3.2: Comparison of the number of serogroup O145 isolates confirmed for each enrichment for both culture-based methods

	No. of enrichments	
No. of isolates confirmed/ enrichment	Direct plating	IMS
1	2	2
2	2	1
3	0	2
4	1	3

3.3 Serogroup O145 characterisation

Serogroup O145 strains isolated in this study and from the Hopkirk Research Institute culture collection (n=73) (Appendix A) were characterised using an O145 serogroup-specific PCR and the mPCR for the four virulence genes *stx1*, *stx2*, *eae* and *ehxA* (described in section 2.5.1). All of the strains were confirmed as serogroup O145. According to the mPCR, 13 strains were *stx*-positive (2 *stx1*-

positive, 11 *stx2*-positive), all strains were *eae* positive (n=73) and a large proportion were *ehxA* positive (69 out of 73) (Appendix A).

The intimin subtypes for some serogroup O145 strains (n=38) was determined (described in sections 2.5.2 and 2.7) to prioritise which strains should be analysed using the Omnilog phenotypic microarray system (Chapter 4); as previous work had identified a potential association between different *eae* subtypes and carbon substrate utilisation [146]. For the 38 serogroup O145 strains tested, the *eae* subtypes β (n=1), ε (n=4), ι (n=4) and γ (n=29) were identified (Table 3.3). Of these strains, 37 subsequently underwent WGS and the intimin subtypes confirmed by *in silico* analysis of the *eae* gene (described in section 2.12.2). Using this method, the intimin subtypes of the additional strains which underwent whole genome sequenced was determined (Appendix A).

Table 3.3: Intimin subtypes determined according to best match using BLASTN

***eae* subtype identified**	**Genbank accession no. match**	**Identity (%)**	**No. strains identified**
β	KT591225.1	99	1
ι	KT591302.1	99-100	4
ε	KT591278.1	99-100	4
γ	KT591261.1	99-100	29

3.4 Discussion

According to the number of calf faecal enrichments identified as STEC O145 positive using molecular methods, the isolation rate was relatively low. Low isolation recovery rates for serogroup O145 have been reported in other studies [17,147]. However many of the recommended enrichment protocols for non-O157 serogroups differ in the selection and enrichment broths, temperature and incubation times described [70]. Evans et al. [147] reported an isolation rate of 0.1% for serogroup O145 using enrichment and IMS methods on sheep faecal samples. *E. coli* serogroup O145 were isolated using IMS and selective media with an isolation rate of 3.0% by Noll et al. [17]. Interestingly, of the serogroup O145 strains isolated (n=19), 29.4% were isolated using non-target IMS beads [17]. Higher recovery rates have been reported for studies detecting STEC

serogroups in artificially inoculated food matrices, for example Hara-Kudo et al. [148] detected serogroup O145 from artificially inoculated ground beef and radish sprouts using IMS and selective media with a sensitivity of 68.2-100%, depending on the culture media used. Posse et al. [65] detected serogroup O145 with an isolation efficiency of 84.6% from artificially inoculated dairy and meat matrices. Although, few (n=2) serogroup O145 strains were tested [65]. This increased recovery rate when using artificially inoculated food matrices may potentially be due to reduced competition with highly competitive microbial consortia associated with naturally-occurring biological samples.

A recent study by Hallewell et al. [73] suggested the efficiency of IMS is influenced by the number of cells present and also any competing serogroups. This was suggested to be a factor in the lower isolation of serogroup O111, O121 and O145 compared to other Top 7 serogroups [73]. Serogroups O111, O26 and O145 grew in a slightly lower proportion in modified EC broth compared to the other “super six” serogroups [72], which may be important when isolating these serogroups from mixed cultures. Kraft et al. [72] also found, when using IMS and selective media, serogroup O145 had the lowest recovery rate from both sterile and non-sterile faeces. To overcome these limitations of IMS, it has been suggested the development of differential media for non-O157 serogroups would be advantageous to increase their successful isolation rate [72,73].

Stromberg et al. [57] compared the detection of the Top 7 serogroups using culture and the NeoSeek method and demonstrated that the NeoSeek method detected a higher proportion of positive samples, compared to culture-based methods, a statistically significant difference [57]. In contrast, Noll et al. [17] found no difference between the proportion of “super six” serogroups detected using culture and PCR; both methods detected positive samples that were identified as negative by the other method [17]. This discordance highlights the importance of using multiple methods in parallel for the detection of STEC.

A number of factors may have contributed to the low isolation rate from this study: (i) the lack of differential culture media specific for the detection of serogroup O145 [16], (ii) the freezing and thawing of the calf faecal enrichments which may have reduced the number of viable serogroup O145 cells during resuscitation, (iii) serogroup O145 isolates may have been present in low abundance in the

enrichments and (iv) the NeoSeek method may have been detecting free DNA rather than DNA associated with viable cells which may have resulted in the identification of false positive RAMS enrichments. PCR based methods coupled with propidium monoazide (PMA) treatment have been developed to prevent amplification of free DNA, as PMA penetrates compromised cells and intercalates with the DNA making it non-amplifiable during subsequent PCR [149]. Variations of PMA PCR based methods have been shown to be effective in detecting viable and non-viable DNA from *E. coli* in treated waste water [150], from *E. coli* O157:H7 and *Salmonella* spp. in milk [151], and from *E. coli* O157:H7 in ground beef [149]. Other factors that may have affected the isolation success rate include the screening criteria or the culture methods used. Proceeding with a subset of calf faecal enrichments is unlikely to have reduced the isolation rate as only enrichments with ≤30 O145 RT-PCR C_t values were tested. Although different protocols are available for the detection of STEC [70], the culture-based methods utilised in this project are consistent with STEC isolation strategies used in other studies, with modifications to factors such as the media used [59,152]. Additional culture-based treatments have been shown to increase STEC recovery such as acid treatment of samples [74] and using multiple selective media in parallel [63,68].

The isolation rate for serogroup O145 was enhanced when using IMS coupled with direct plating compared to using direct plating alone. The improved isolation success rate was likely due to the IMS methodology first concentrating serogroup O145 bacteria prior to plating onto selective CHROMagar™ STEC. Hallewell et al. [73] suggested that the specificity of IMS beads coated with polyclonal antisera raised against O145 LPS was decreased when other competing serogroups were present, such as in a faecal enrichment, which may have contributed to the low overall isolation rate. The recovery of >1 serogroup O145 isolate was more associated with IMS compared to direct plating (Table 3.2). Mostly, these isolates had identical virulence profiles using the mPCR (described in section 2.5.1), with the exception of two isolates VC237m and VC237o. Comparison of the O145 RT-PCR C_t values for the enrichments screened using the different culture-based methods was not significantly different ($p = 0.3378$). However, the power of this comparison was likely to be low due to the small sample size. Screening further

enrichments may provide an improved indication of the relationship between low RT-PCR C_t value and isolation success. However, these comparisons may be confounded with exogenous DNA.

3.5 Summary

The isolation rate for serogroup O145 using culture-based methods was low compared to the number of calf faecal enrichments identified as STEC O145 positive using molecular based methods. This low success rate may be due to a range of factors such as the lack of differential culture media, low abundance of serogroup O145 present in the enrichments, or the presence of free DNA/non-viable O145. The use of IMS coupled with direct plating onto CHROMagar™ STEC improved the rate of O145 isolation, however this was still much lower than the prevalence detected using molecular methods. Additional steps such as acid treatment and using multiple selective media in parallel could be used to improve the isolation of serogroup O145, however, the additional time and cost of using these subsequent methods would need to be evaluated.

Most of the serogroup O145 strains isolated in this study were a similar genotype (*stx*-negative, *eae* subtype γ and *ehx*A-positive, n=14 out of 15 strains with full genotype information), with only a single *stx*-positive, *eae* subtype γ and *ehx*A-positive O145 identified.

4. Results - Utilisation of carbon substrates

To investigate the metabolic characteristics of serogroup O145 strains and identify carbon substrates for the use in a differential culture media, the Omnilog phenotypic microarray system (Biolog Inc, Hayward, California, USA) was used. This system allows the rapid determination of bacterial growth associated with specific conditions (pH, osmolarity, antibiotics etc.) or metabolites, such as the utilisation of specific carbon substrates. The serogroup O145 strains were inoculated and examined on either PM1 or PM1 and PM2A MicroPlates™ (Appendix G), each containing 95 unique carbon substrates, and incubated for 24 hours at 37°C with the chemical reduction of the dye assessed at 15 min intervals (as described in section 2.8.1). Serogroup O145 strains to be examined using the Omnilog phenotypic microarray system were selected using random stratified sampling according to separate variables, such as *eae* subtype, ST, geographic origin and a toxigenic or non-toxigenic genotype (Appendix A). If multiple isolates were obtained from the same enrichment, only one isolate was included.

4.1 Utilisation of carbon substrates (PM1 MicroPlates™)

The metabolic profiles of 28 *E. coli* serogroup O145 strains was determined using the Omnilog phenotypic microarray system (Appendix H). The utilisation of carbon substrates on the PM1 MicroPlates™ suggested significant metabolic variation between serogroup O145 strains (Figure 4.1).

4.1.1 Clustering broadly correlates with *eae* subtype

To compare the carbon utilisation profiles of serogroup O145 strains, a heat-map was produced in R using the pre-defined Omnilog cut-off ranges, which grouped the bacterial utilisation values (ranging from 0-300 Omnilog Units) into the following three categories: 0-50 representing no utilisation, 51-150 representing moderate utilisation and 151-400 representing extensive utilisation, as described in section 2.8.2.

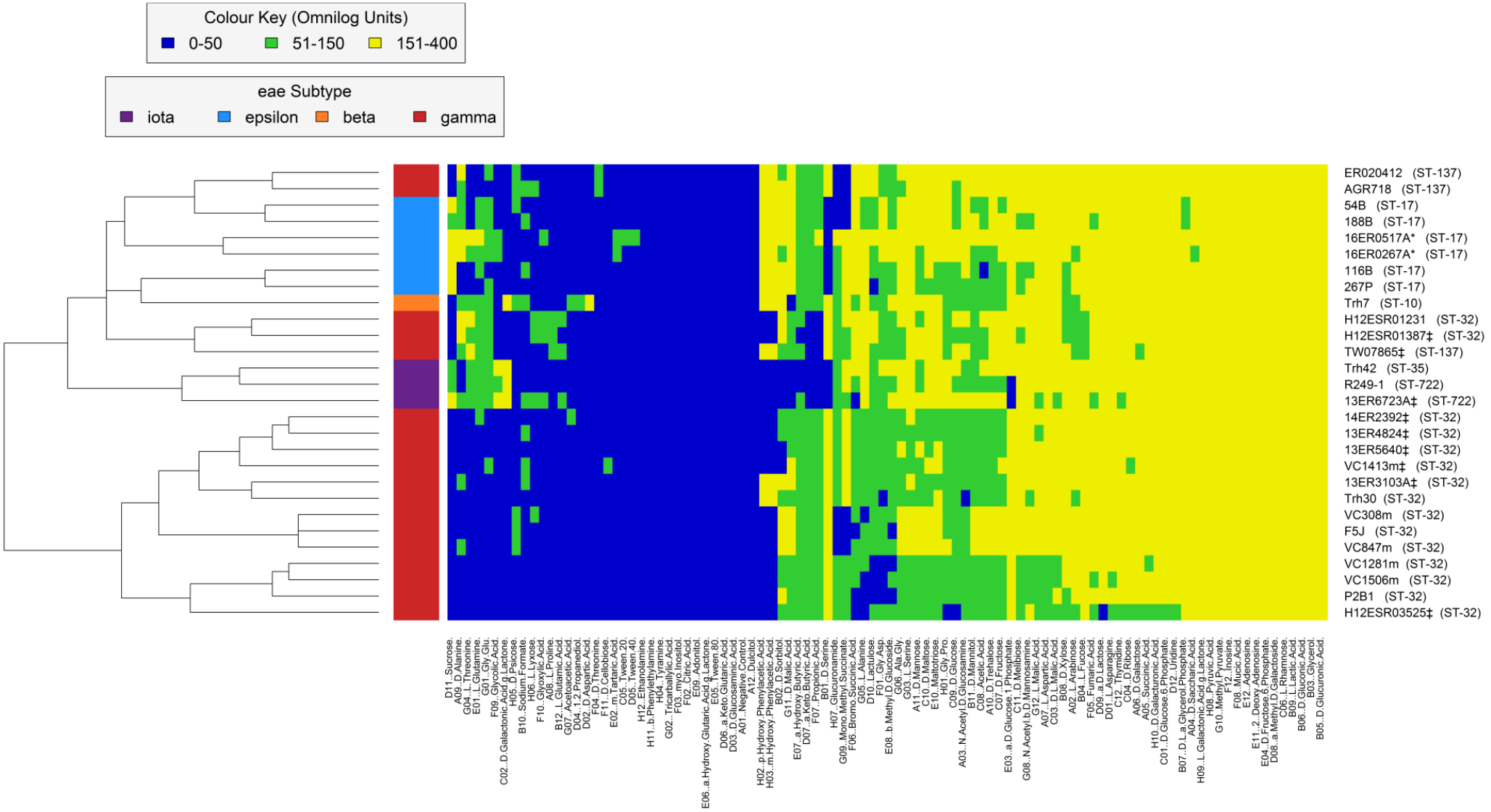

Figure 4.1: Heat-map showing *E. coli* serogroup O145 strains carbon utilisation profiles (PM1 MicroPlates™)

Heat-map of PM1 carbon substrate metabolism over a 24 hour incubation period at 37°C by serogroup O145 strains. The end-point utilisation values (Omnilog Units) were grouped into the following three categories: 0-50 representing no utilisation, 51-150 representing moderate utilisation and 151-400 representing extensive utilisation, as indicated by the colour key. Each strain (n=28) is indicated on the right and the 95 carbon substrates listed along at the foot of the figure. Metadata is included for *eae* subtype, sequence type and whether the strains were toxigenic. *eae* subtype on the left is indicated by the colour key, sequence type is shown in brackets, *stx1* positive as * and *stx2* positive as ‡.

The clustering of serogroup O145 strains by carbon utilisation is broadly associated with *eae* subtype, a gene encoding the protein intimin (Figure 4.1) which is involved in bacterial attachment to intestinal epithelial cells. The 28 strains analysed using the Omnilog system contain the *eae* subtypes β (n=1), ι (n=3), ε (n=6) and γ (n=18). Although the serogroup O145 strains were only represented by four *eae* subtypes, over 28 different *eae* alleles have been identified [38].

Cluster analysis was used to group the serogroup O145 strains into two distinct clusters using heat-map data (Figure 4.1). Most strains clustered according to *eae* subtype, except *eae* γ which was present in both of the main clusters. Thirteen *eae* γ strains isolated from a range of bovine and human sources clustered together. The remaining five serogroup O145 *eae* γ strains clustered separately with the *eae* subtype β (n=1), ι (n=3) and ε (n=6) O145 strains.

There was no association between the utilisation of specific carbon substrates and whether a strain was either toxigenic (*stx*-positive) or non-toxigenic (*stx*-negative). Similarly, strains of human and bovine origin are distributed throughout the dendrogram.

A dendrogram was produced using hierarchical clustering of the PM1 carbon substrate utilisation data (Figure 4.2). The general clustering was comparable to that of the heat-map (Figure 4.1) with the strains in similar clusters correlated with *eae* subtype and STs, but with a different cluster order due to the contrasting methodology involved, i.e. the heat-map was generated using the Omnilog value cut-off ranges (described in section 2.8.2), whereas the dendrogram uses the actual Omnilog value endpoint for each substrate.

4.1.2 Clustering broadly correlates with sequence type

In addition to *eae* subtype, the metabolic profiles of serogroup O145 strains also broadly correlated with ST (Figure 4.1). Serogroup O145 strains analysed in this study belong to six different STs with a correlation between the STs and the *eae* subtype each strain carries. For example, strains carrying *eae* subtype ε belong

to ST-17, β to ST-10, ı to ST-35 (n=1) and ST-722 (n=2), and γ to ST-32 (n=15) and ST-137 (n=3).

From the heat-map data (Figure 4.1) the strains are separated into two main clusters, one containing ST-32 (n=13 strains), and the other containing strains that are ST-10 (n=1), ST-17 (n=6), ST-35 (n=1) ST-137 (n=3), ST-722 (n=2) and the remaining two ST-32 strains. In contrast to the *eae* subtype γ ST-32 and ST-137 strains that clustered separately, the *eae* subtype ı clustered together even though they consisted of distinct STs (ST-35 and ST-722).

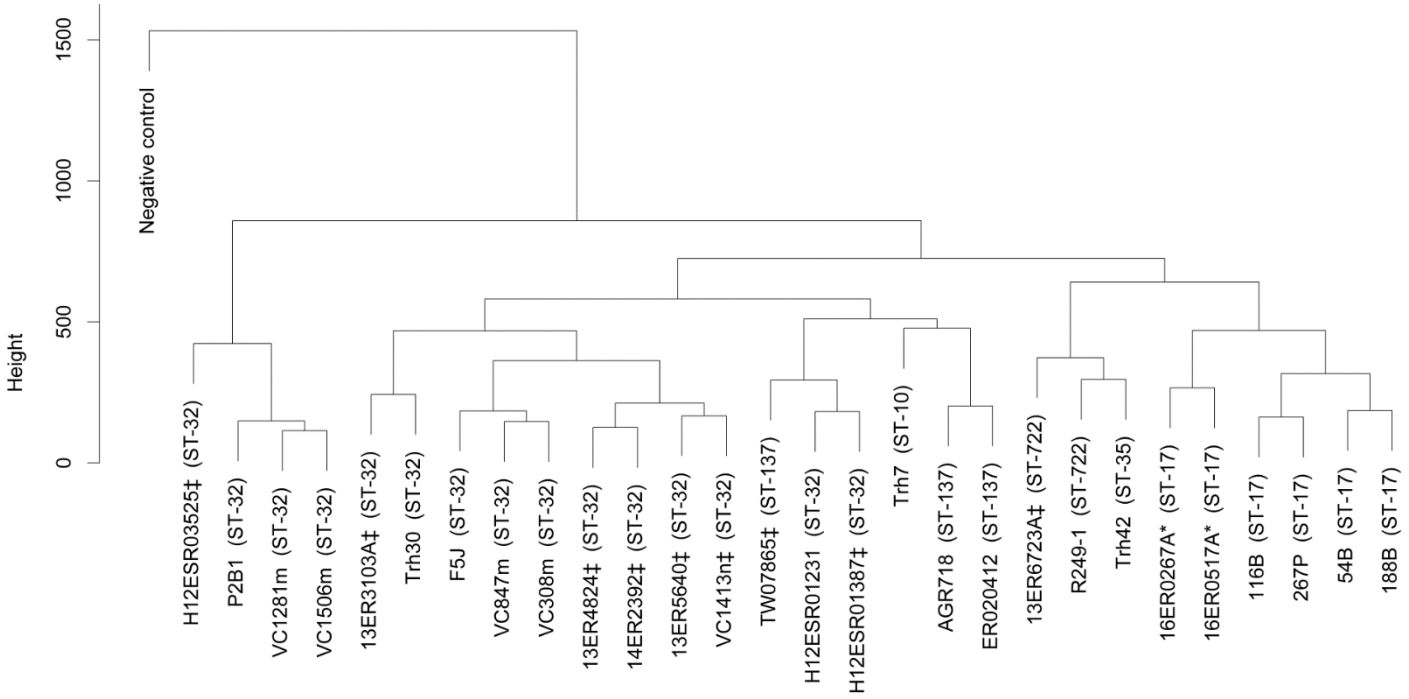

Figure 4.2: Cluster dendrogram showing the similarities of *E. coli* serogroup O145 strains based on their carbon utilisation profile (PM1 MicroPlate™)

The end-point values per serogroup O145 strain (n=28) for each carbon substrate on the phenotypic microarray plates (n=95) was recorded, and used to produce a cluster dendrogram using hierarchical clustering. Metadata is included for sequence type and whether the strains were toxigenic or non-toxigenic. Sequence type is shown in brackets, *stx1* positive as * and *stx2* positive as ‡.

4.1.3 Reproducibility of serogroup O145 carbon utilisation on PM1 MicroPlates™

Half of the isolates (n=14, 50%) were tested in replicate experiments and two in duplicate using PM1 MicroPlates™ to determine the biological reproducibility of specific O145 strains (Appendix H). The isolates were selected using random stratified sampling of the serogroup O145 strains according to *eae* subtype.

With the inclusion of the duplicates and replicates in the carbon utilisation analyses, the clustering observed remained broadly comparable to that seen previously with correlation according to *eae* subtype and ST (Figure 4.3). With replicates included in the heat-map, the strains also grouped into two main clusters with one containing *eae* subtype γ strains (n=13) and replicates (n=4) and the other containing β, ε, ι and the remaining γ isolates, together with respective duplicates and replicates.

There was contrasting utilisation of some carbon substrates (11.6%; 11/95) by the duplicate and replicates (Table 4.1). Some substrates utilised during an original phenotypic microarray experiment were not utilised during the subsequent experiment (orange), and vice versa (blue). Notably, the utlisation of D-alanine (6/16), lactulose (5/16), D-psicose (9/16), L-lyxose (5/16) and glucuronamide (8/16) differed for multiple isolates indicating the utilisation of these substrates by serogroup O145 strains may be inconsistent. The time duration between replicate phenotypic microarrays may contribute to carbon utilisation inconsistencies as the replicate sets which were conducted <1 month for the PM1 data showed less variation in substrate utilisation compared to replicate sets which were carried out >1 month apart (Table 4.1). Comparison of the original and duplicate phenotypic microarrays (n=2) indicated contrasting utilisation of only one substrate (L-lyxose).

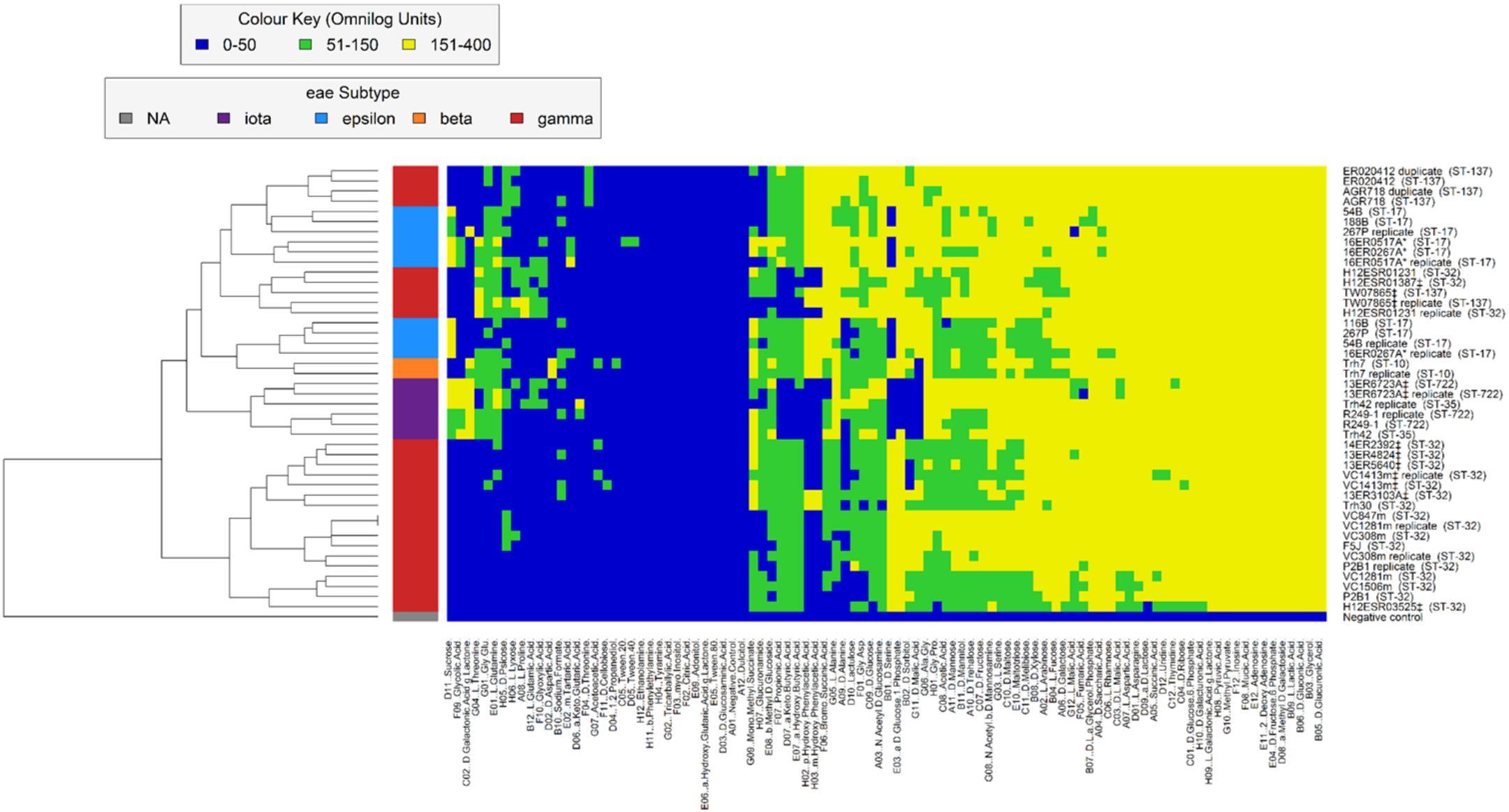

Figure 4.3: Heat-map showing *E. coli* serogroup O145 strains carbon utilisation profiles (PM1 MicroPlate™) with replicates and duplicates

Heat-map of PM1 carbon substrate metabolism over a 24 hour incubation period at 37°C by serogroup O145 strains. The end-point utilisation values (Omnilog Units) were grouped into the following three categories: 0-50 representing no utilisation, 51-150 representing moderate utilisation and 151-400 representing extensive utilisation, as indicated by the colour key. Each strain (n=28, n=14 replicates, n=2 duplicates) is indicated on the right and the 95 carbon substrates listed along at the foot of the figure. Metadata is included for *eae* subtype, sequence type and whether the strains were toxigenic. *eae* subtype on the left is represented by the colour key, NA is not applicable, sequence type is shown in brackets, *stx1* positive as * and *stx2* positive as ‡.

Table 4.1: Comparison of carbon substrates from PM1 MicroPlates™ (n=11) which differ between ≥1 set of replicates (n=14) or duplicates (n=2)

Substrate	Well position	Trh7†	267P†	54B	16ER0517A†	16ER0267A	R249-1†	Trh42†	13ER6723A†	ER020412‡	AGR718‡	TW07865†	H12ESR01231†	VC1413m	VC1281m†	VC308m	P2B1†
		1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2
D-alanine	A09	+/+	-/+	+/-	+/+	+/+	-/+	-/+	+/+	+/+	+/+	+/+	+/+	-/-	-/+	-/+	-/-
D-galactonic-acid-γ-lactone	C02	+/+	-/+	-/-	-/-	-/-	+/+	+/+	+/+	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
1,2- propandiol	D04	+/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
Lactulose	D10	+/+	-/+	+/+	+/+	+/-	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/-	-/+	+/+	-/+
Glycolic acid	F09	-/-	-/-	-/-	+/+	+/-	+/+	+/+	+/+	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
L-threonine	G04	+/+	-/-	-/-	+/+	+/+	+/+	+/-	+/-	-/-	-/-	+/+	+/+	-/-	-/-	-/-	-/-
Mono-methyl succinate	G09	+/+	-/-	-/-	+/+	+/+	+/+	+/-	+/+	-/-	-/-	+/+	+/+	-/-	-/-	-/-	-/-
L-malic acid	G12	+/+	+/-	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
D-psicose	H05	+/-	-/+	+/+	-/+	-/-	-/-	-/+	-/+	+/+	+/+	-/+	-/+	-/-	-/+	+/-	-/-
L-lyxose	H06	-/-	-/-	-/-	-/+	-/-	-/-	-/-	+/-	-/+	+/+	-/+	+/+	-/-	-/-	+/-	-/-
Glucuronamide	H07	+/+	+/-	-/-	+/-	+/+	+/+	+/-	+/-	-/-	-/-	+/-	+/-	+/+	+/-	-/-	+/-

†: >1 month between replicate

1: Original phenotypic microarray assay

2: Replicate phenotypic microarray assay

Orange shading: Contrasting substrate utilisation in original plate and not in replicate plate

Blue shading: Contrasting no substrate utilisation in original plate and utilisation in replicate plate

‡: Duplicate phenotypic microarray assay

4.2 Utilisation of carbon substrates (PM2A MicroPlates™)

PM2A MicroPlates™ were used to examine the growth of serogroup O145 strains on a further 95 carbon substrates using the Omnilog phenotypic microarray system (Appendix H). Twenty *E. coli* serogroup O145 strains were analysed on PM2A MicroPlates™, with observations indicating fewer carbon substrates (22 of 95 substrates) on the PM2A MicroPlates™ were utilised by ≥1 *E. coli* strain (Figure 4.4). Although fewer strains were analysed on these plates, a similar clustering pattern of carbon utilisation correlated with *eae* subtype and ST was observed as seen with the PM1 carbon source data. Additionally, there appeared to be no correlation between the metabolic profiles and isolation source and whether the strains were *stx*-positive or *stx*-negative.

4.2.1 Clustering broadly correlates with *eae* subtype and sequence type

Twenty serogroup O145 strains were examined for their metabolism of the 95 PM2A carbon sources. The strains separated into three main clusters by substrate utilisation sources which, like the utilisation of PM1 substrates, broadly correlates with *eae* subtype and ST. The first cluster contains 11 *eae* subtype γ strains (ST-32). The second cluster contains four *eae* subtype ε strains (ST-17), the single *eae* subtype β strain (ST-10) and the final *eae* subtype γ strain (ST-137). The remaining cluster contains three *eae* subtype ι strains (ST-35 and ST-722).

Strain TW07865 is the only *eae* subtype γ strain that clusters separately from the remaining 11 γ strains, including H12ESR01231 and H12ESR01387, with which it clusters with according to PM1 carbon utilisation data. This is likely due to different carbon substrates utilised by TW07865 on the PM2A MicroPlates™ compared to the other *eae* subtype γ strains. TW07865 was also isolated from Germany whereas the latter two strains are New Zealand isolates.

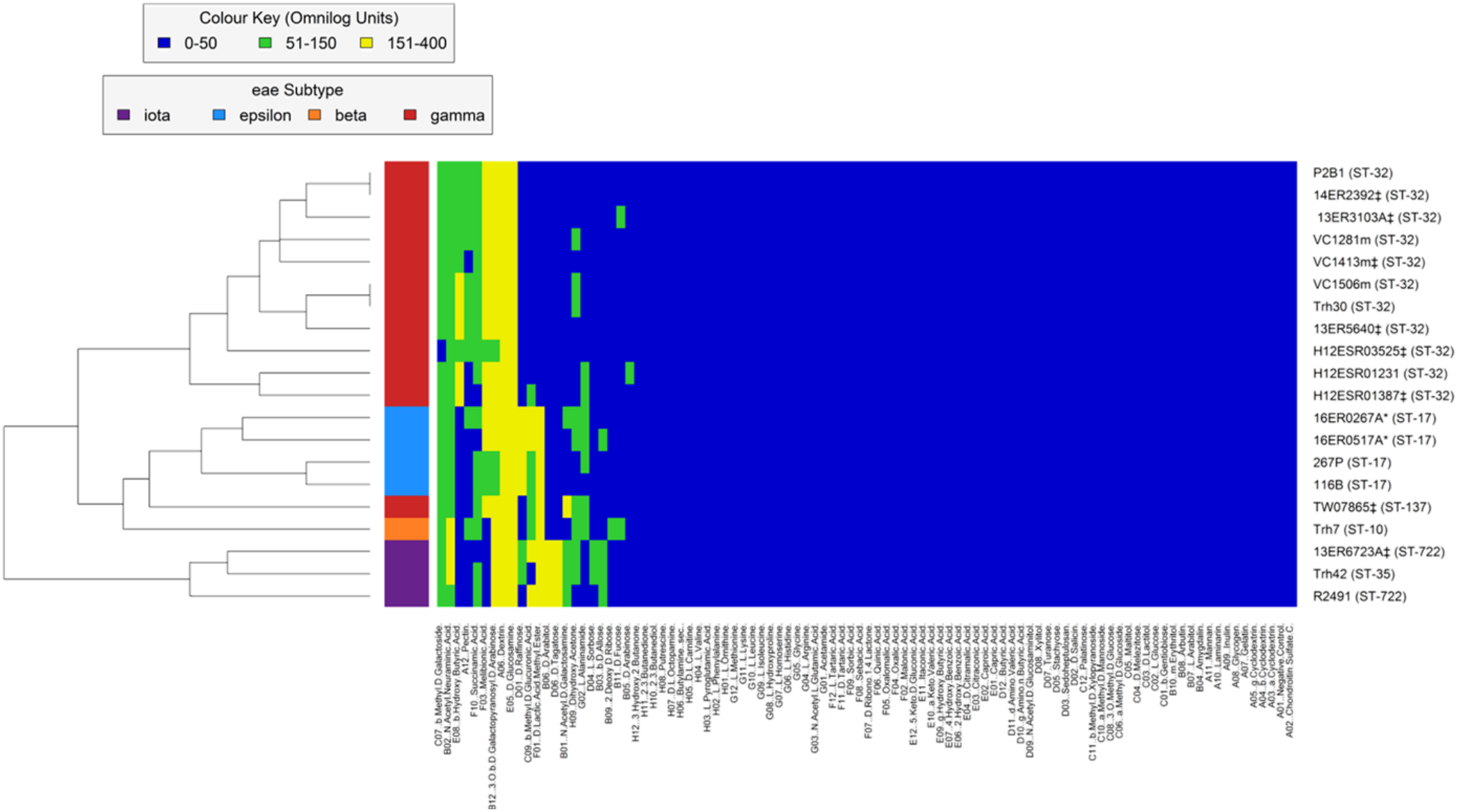

Figure 4.4: Heat-map showing *E. coli* serogroup O145 strains carbon utilisation profiles (PM2A MicroPlates™)

Heat-map of PM2A carbon substrate metabolism over a 24 hour incubation period at 37°C by serogroup O145 strains. The end-point utilisation values (Omnilog Units) were grouped into the following three categories: 0-50 representing no utilisation, 51-150 representing moderate utilisation and 151-400 representing extensive utilisation, as indicated by the colour key. Each strain (n=20) is indicated on the right and the 95 carbon substrates listed along at the foot of the figure. Metadata is included for *eae* subtype, sequence type and whether the strains were toxigenic. *eae* subtype on the left is represented by the colour key, sequence type is shown in brackets, *stx1* positive as * and *stx2* positive as ‡.

4.2.2 Reproducibility of serogroup O145 carbon utilisation on PM2A MicroPlates™

The four respective serogroup O145 biological replicates (one of each of the *eae* subtype β, ε, γ and ι strains) clustered closely with the original phenotypic microarray assay (Figure 4.5). However, contrasting utilisation of carbon substrates (4.2%, 4/95) was noted by replicates (Table 4.2). Some substrates utilised during an original phenotypic microarray experiment were not utilised during the subsequent experiment (orange), indicating that like some of the substrates included in the PM1 MicroPlates™ the utilisation of substrates by serogroup O145 strains may be inconsistent. However, in contrast to the PM1 data obtained, there appears to be no association between the length of time between inoculation of the original and replicate phenotypic microarray plates and variation in carbon substrate utilisation. Further replicates may be required to test this hypothesis as data is only available for four replicate phenotypic microarray experiments.

Table 4.2: Comparison of carbon substrates from PM2A MicroPlates™ (n=4) which differ between ≥1 set of replicates (n=4)

	Well position	Trh7†	R249-1†	16ER 0267A	VC1413m
		1/2	1/2	1/2	1/2
Pectin	A12	**+/-**	**-/-**	**+/-**	**-/-**
N-acetyl-D-galactosamine	B01	**-/-**	**+/+**	**+/-**	**-/-**
D-fucose	B11	**+/-**	**-/-**	**-/-**	**-/-**
Dihydroxy-acetone	H09	**+/-**	**-/-**	**+/-**	**-/-**

†: >1 month between replicate

1: Original phenotypic microarray

2: Replicate phenotypic microarray

Orange: Contrasting substrate utilisation in original plate and not in duplicate plate

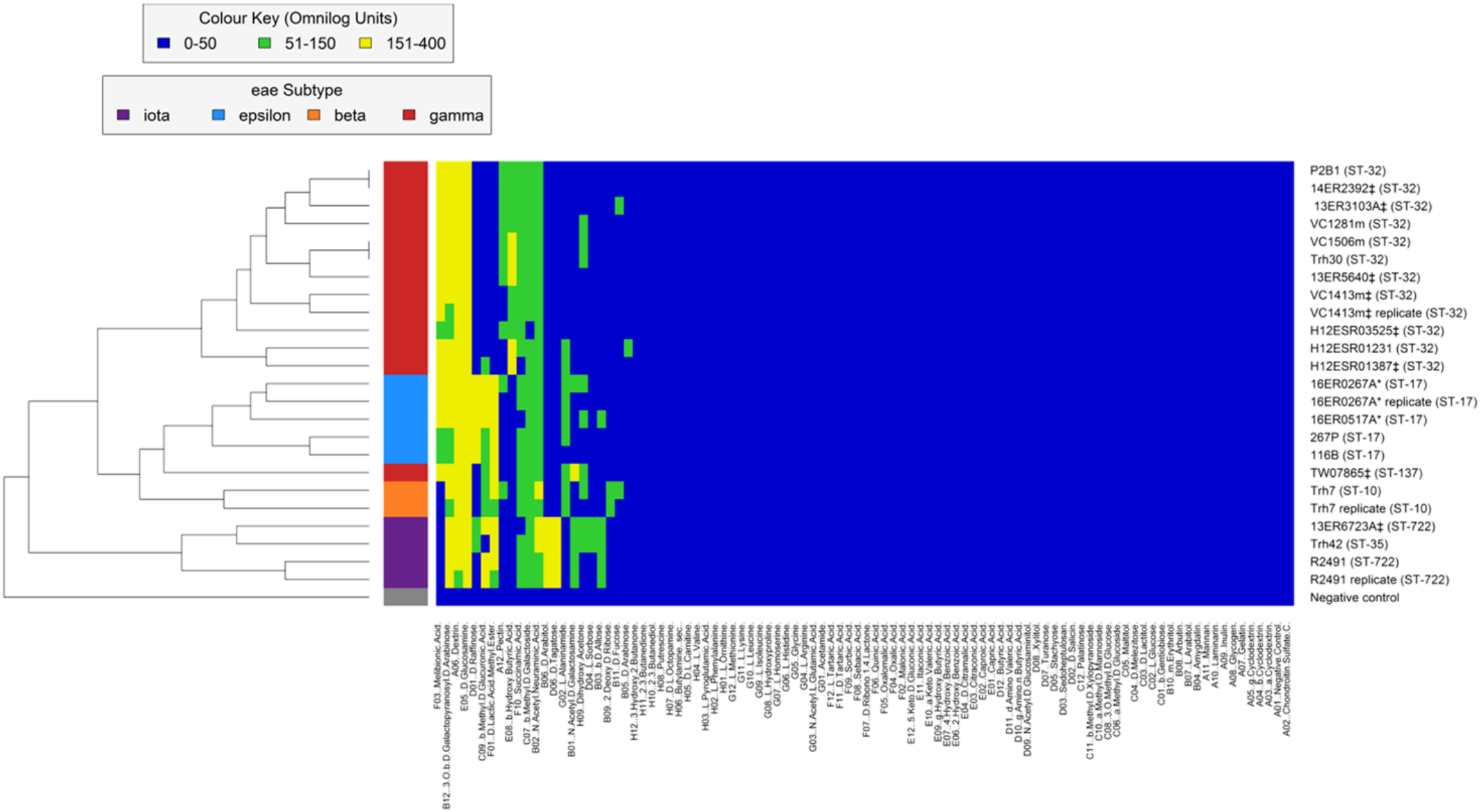

Figure 4.5: Heat-map showing *E. coli* serogroup O145 strains carbon utilisation profiles (PM2A MicroPlates™) with replicates

Heat-map of PM2A carbon substrate metabolism over a 24 hour incubation period at 37°C by serogroup O145 strains. The end-point utilisation values (Omnilog Units) were grouped into the following three categories: 0-50 representing no utilisation, 51-150 representing moderate utilisation and 151-400 representing positive utilisation, indicated by the colour key. Each strain (n=20 and four replicates) is indicated on the right and the 95 carbon substrates listed along the bottom. Metadata is included for *eae* subtype, sequence type and whether the strains were toxigenic. *eae* subtype on the left is represented by the colour key, sequence type is shown in brackets, *stx1* positive as * and *stx2* positive as ‡.

4.3 Candidate substrates for use in a differential media

It is unlikely that any carbon substrate examined in this study would clearly distinguish serogroup O145 from all other types of *E. coli*. However, it could be coupled with molecular methods, such as RT-PCR or NeoSeek to aid in the preliminary identification of serogroup O145-positive enrichments with the further analysis of these enrichments using culture-based isolation methods such as IMS. Ideally, a potential carbon substrate that would differentiate the majority of serogroup O145 strains would need to be both relatively inexpensive and easily incorporated into a culture media.

Examination of the metabolism of 190 carbon substrates failed to identify a candidate that could be used to differentiate *stx*-positive and *stx*-negative *E. coli* serogroup O145 strains, or strains isolated from human or bovine sources. However, carbon substrates were identified which could be used to differentiate specific subsets of serogroup O145 strains, such as certain *eae* subtypes and STs.

4.3.1 Identification of carbon substrates to differentiate certain *eae* subtypes and sequence types

Twelve carbon substrates were identified from the PM1 MicroPlates™ which differentiate subsets of serogroup O145 strains (Figure 4.6). Some substrates, such as D-aspartic acid, only differentiate one *eae* subtype and ST. Other carbon substrates can be used to differentiate multiple subsets of serogroup O145 strains. For example, sucrose differentiates *eae* subtypes ε (ST-17) and ι (ST-35, ST-722) and D-serine differentiates *eae* subtype γ (ST-32 and ST-137) and β (ST-10). Another notable substrate which differentiates the majority of serogroup O145 strains is D-malic acid which is utilised by *eae* subtypes γ (ST-32, ST-137) and ε (ST-17). Interestingly, propionic acid is not utilised by *eae* subtype ι (ST-35, ST-722) and the three ST-137 strains and therefore its metabolism distinguishes the two *eae* subtype γ clusters.

Six carbon substrates were identified from the PM2A MicroPlates™ which differentiate certain subsets of serogroup O145 strains (Figure 4.7). Of these, the substrates D-arabitol and D-tagatose are only utilised by *eae* subtype ι strains (ST-35, ST-722). The substrate D-lactic-acid-methyl-ester was utilised by *eae*

subtypes ι (ST-35, ST-722), *eae* subtype ε (ST-17) and the ST-137 *eae* subtype γ strain. D-raffinose was utilised by *eae* subtype ε strains (ST-17) and two out of three *eae* subtype ι strains (ST-35, ST-722). Melibionic acid was utilised by all (n=13) of the *eae* subtype γ strains (ST-32, ST-137) and β-hydroxy-butyric acid was utilised by 12 (ST-32) out of 13 *eae* subtype γ strains, the exception being TW07865 (ST-137).

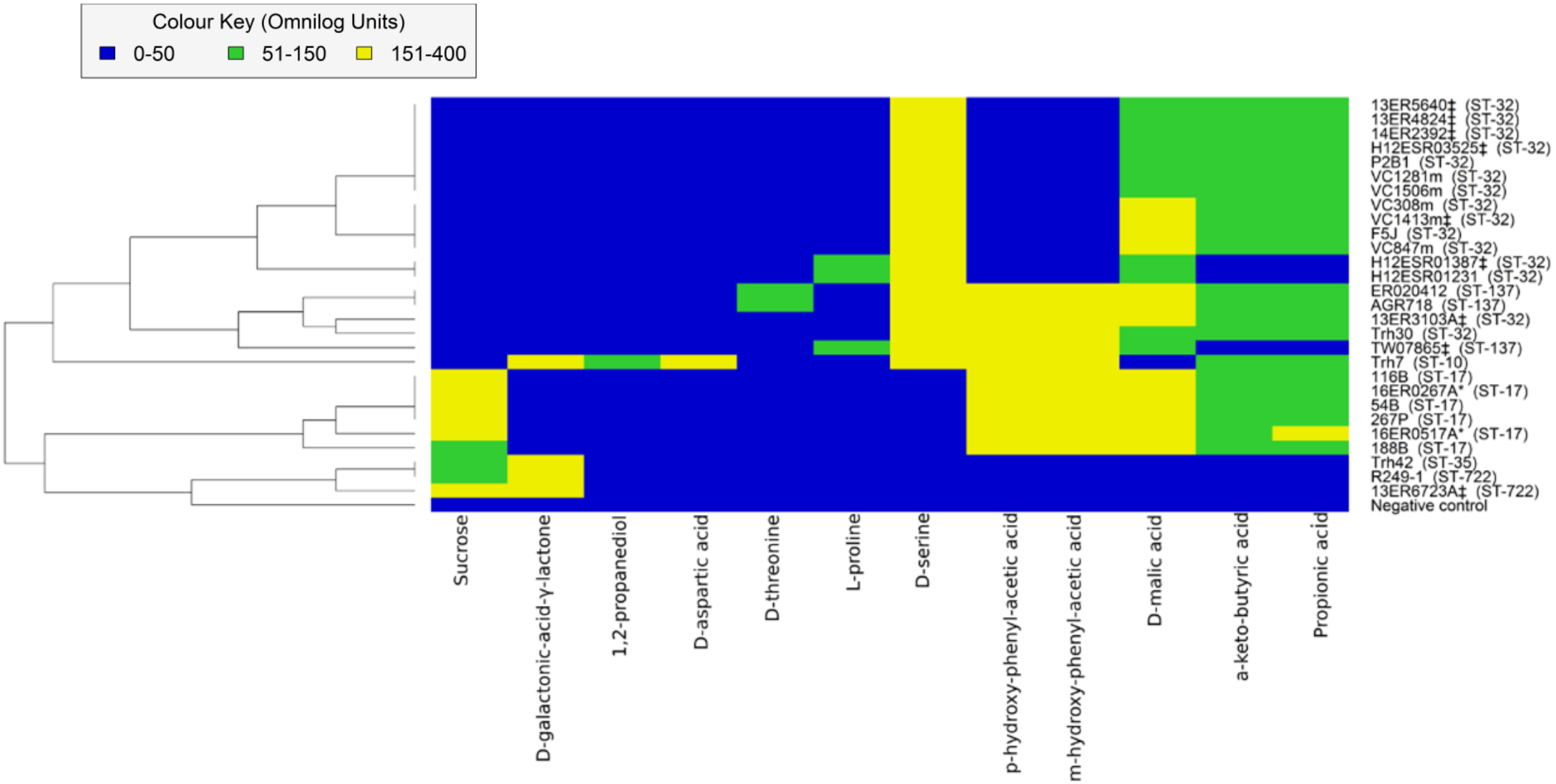

Figure 4.6: Heat-map showing *E. coli* serogroup O145 strains carbon utilisation profiles on selected PM1 carbon substrates

Heat-map of PM1 carbon substrate metabolism over a 24 hour incubation period at 37°C to differentiate serogroup O145 strains. The end-point utilisation values (Omnilog Units) were grouped into the following three categories: 0-50 representing no utilisation, 51-150 representing moderate utilisation and 151-400 representing extensive utilisation, as indicated by the colour key. Each strain (n=28) is indicated on the right and the 12 carbon substrates listed along at the foot of the figure. Metadata is included for sequence type and whether the strains were toxigenic. Sequence type is shown in brackets, *stx1* positive as * and *stx2* positive as ‡.

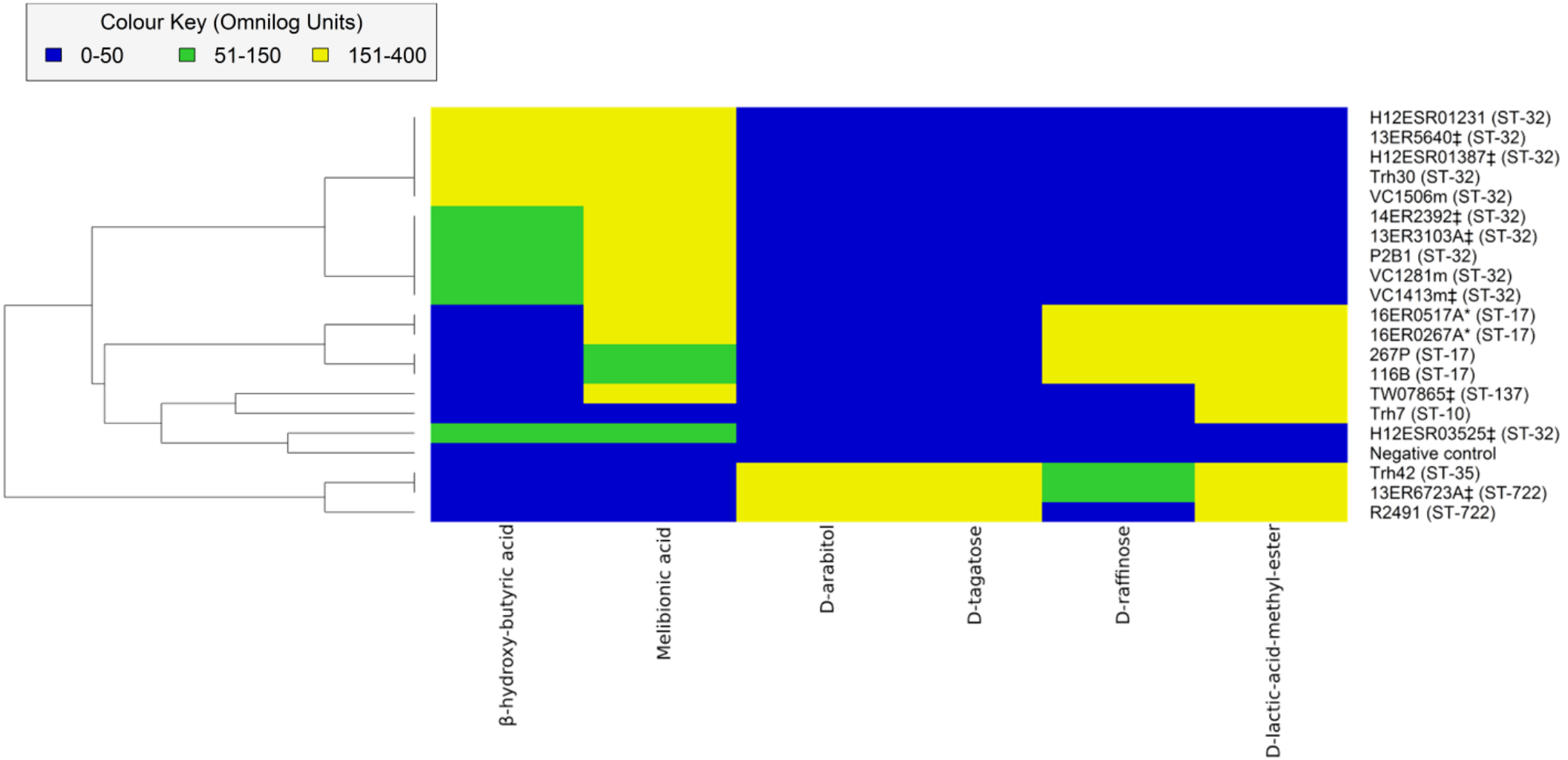

Figure 4.7: Heat-map showing *E. coli* serogroup O145 strains carbon utilisation profiles on selected PM2A carbon substrates

Heat-map of PM2A carbon substrate metabolism over a 24 hour incubation period at 37°C to differentiate serogroup O145 strains. The end-point utilisation values (Omnilog Units) were grouped into the following three categories: 0-50 representing no utilisation, 51-150 representing moderate utilisation and 151-400 representing extensive utilisation, as indicated by the colour key. Each strain (n=20) is indicated on the right and the six carbon substrates listed along at the foot of the figure. Metadata is included for sequence type and whether the strains were toxigenic. Sequence type is shown in brackets, *stx1* positive as * and *stx2* positive as ‡.

Table 4.3: Specific carbon substrates utilised by serogroup O145 strains that could be used to differentiate *eae* subtypes and sequence types

			***eae* subtypes**				**Sequence types**					
Substrate	**PM MicroPlate™‡**	**Well position**	**β**	**ε**	**ι**	**γ**	**ST-10**	**ST-17**	**ST-32**	**ST-35**	**ST-137**	**ST-722**
D-serine	PM1	B01	✓	✗	✗	✓	✓	✗	✓	✗	✓	✗
D-arabitol	PM2A	B06	✗	✗	✓	✗	✗	✗	✗	✓	✗	✓
D-tagatose	PM2A	D06	✗	✗	✓	✗	✗	✗	✗	✓	✗	✓
α-keto-butyric acid	PM1	D07	✓	✓	✗	✓ *(15/18)*	✓	✓	✓ *(13/15)*	✗	✓ *(2/3)*	✗
β-hydroxy-butyric acid	PM2A	E08	✗	✗	✗	✓ *(11/12)*	✗	✗	✓	✗	✗	✗
Melibionic acid	PM2A	F03	✗	✓	✗	✓	✗	✓	✓	✗	✓	✗
Propionic acid	PM1	F07	✓	✓	✗	✓ *(15/18)*	✓	✓	✓ *(13/15)*	✗	✓ *(2/3)*	✗
D-malic acid	PM1	G11	✗	✓	✗	✓	✗	✓	✓	✗	✓	✗

‡: Strains analysed on PM1 MicroPlate™: β (n=1), ε (n=6), ι (n=3) and γ (n=18)

Strains analysed on PM2A MicroPlate™: β (n=1), ε (n=4), ι (n=3) and γ (n=12)

4.4 Discussion

The contrasting carbon utilisation of serogroup O145 strains and broad correlation with *eae* subtype indicates that acquisition of distinct LEE pathogenicity islands has likely occurred on multiple occasions into different *E. coli* lineages [36,153,154]. The observation that both *eae* subtype γ and ε strains are included in multiple clusters may indicate subsequent evolution of distinct sub-lineages and highlights the multiple mechanisms through which carbon is utilised and the various factors that are thus involved. The differentiation of *eae* subtype γ strains into separate clusters using carbon substrate utilisation corresponds with their contrasting STs, as described below. The separation of *eae* subtype ε and γ strains into distinct clusters could not be unequivocally resolved using Omnilog data and was further investigated using genomic data (Chapter 5). In contrast to the *eae* subtype γ ST-32 and ST-137 strains that clustered separately, the *eae* subtype ι clustered together even though they consisted of distinct STs (ST-35 and ST-722). This may indicate that the housekeeping genes from which the ST scheme is derived from, are sufficiently dissimilar in the *eae* subtype γ ST-32 and ST-137 strains as well as the carbon metabolism loci, such that clustering is divergent. In contrast, the carbon metabolism described in this study of the ST-35 and ST-722 *eae* subtype ι strains is sufficiently similar for the three strains to cluster together.

Some of the potential carbon substrates (Figures 4.6 and 4.7) were deemed unsuitable for the use in a differential media due to the extent and consistency in utilisation by serogroup O145 strains. For example, carbon substrates that only differentiate a limited number of serogroup O145 strains, such as sucrose or D-tagatose, would be unsuitable as they are limited in their scope to select a broad range of serogroup O145 isolates. However, substrates such as these may be useful for the presumptive culture-based identification of different *eae* subtypes or STs for serogroup O145 strains. Carbon substrates such as propionic acid, α-keto-butyric acid and D-serine, could be suitable candidates as they differentiate a large proportion of *E. coli* serogroup O145 isolates investigated in this study. Two of the carbon substrates shown in Figure 4.6 (1,2-propanediol and D-galactonic-acid-γ-lactone) would also be unsuitable as utilisation of these substrates was inconsistent (Table 4.1).

Some carbon substrates may be suitable for the use in a differential media to aid the isolation of *E. coli* serogroup O145 strains (Table 4.3). These substrates differentiate a large proportion of the serogroup O145 strains analysed in this study and were not associated with any variation in substrate utilisation during examination of biological replicates. However, it is important to note both β-hydroxy-butyric acid and melibionic acid are substrates from PM2A MicroPlates™ and were not examined with the same rigour as the PM1 substrates. These potential carbon substrates warrant subsequent investigation through the testing of additional serogroup O145 strains on these substrates and by substituting them as the main energy source using basal MacConkey agar or in a minimal media.

It is unknown whether the interval between an original phenotypic microarray experiment and a subsequent repeat examining the carbon utilisation of the same O145 strain may affect carbon metabolism. However duplicate experiments involving PM1 MicroPlates™ which were conducted with an interval of <1 month exhibited less variation in substrate utilisation compared to experiments conducted with an interval of >1 month (Table 4.1). This apparent relationship between substrate utilisation/carbon metabolism profiles and time interval between individual experiments may be influenced by technical variations associated with the Omnilog machine (optical density measurements), freeze thawing of culture stocks or a changes in bacterial gene expression. The variation of carbon substrate utilisation between O145 replicates suggests that these substrates are unsuitable for use in a differential media as their utilisation per strain is inconsistent.

Few published studies have compared the carbon utilisation profiles of *E. coli* serogroups, and none have compared the carbon utilisation among a large number of genetically diverse *E. coli* serogroup O145 strains. A comparison of the carbon utilisation of *E. coli* O157:H7 (n=81), commensal *E. coli* (n=39) and two K-12 laboratory strains demonstrated that that O157:H7 strains were less metabolically diverse, based on carbon utilisation than the commensal strains tested [155]. Furthermore 27 out of 95 carbon substrates were utilised by commensal strains but not by O157:H7 strains [155]. However the substrate metabolism was scored in binary as either positive or negative, compared to

measuring the chemical reduction of the dye every 15 mins as in this study [155]. The carbon metabolism of STEC (n=37) representing ten serogroups was examined on the PM1 and PM2A carbon MicroPlates™ and several carbon substrates were identified that were able to differentiate defined STEC serogroups [67]. Although only three serogroup O145 strains were investigated, an identification strategy was proposed involving the metabolism of β-hydroxy-butyric acid and a lack of growth on dulcitol and D-galactonic acid-γ-lactone [67], together with the identification scheme proposed by Posse et al. [156]. Although, none of the serogroup O145 strains (n=28) examined in this study were able to grow on dulcitol, *eae* subtype β (n=1), ε (n=4), ι (n=3) and γ (n=1, ST-137) strains were unable to grow on β-hydroxy-butyric acid and *eae* subtype β (n=1) and ι (n=3) utilised D-galactonic acid-γ-lactone. This suggests that dulcitol may be suitable for the differentiation of serogroup O145, however β-hydroxy-butyric acid and D-galactonic acid-γ-lactone may not be suitable due to utilisation of these substrates by some serogroup O145 strains.

4.5 Summary

The carbon utilisation profiles, conducted using the Omnilog phenotypic microarray system, have demonstrated that *E. coli* serogroup O145 strains are metabolically diverse. There appears to be no correlation between carbon utilisation profiles and whether the strains are *stx*-positive or *stx*-negative, or whether the strains are isolated from bovine or human sources. However, there is a broad correlation between carbon utilisation and *eae* subtype and ST which is likely associated with other underlying factors such as the evolution of different *E. coli* lineages. Candidate carbon substrates have been identified for the use in a differential media to aid the isolation of serogroup O145, in addition to other culture-based methods such as IMS from enrichments identified as serogroup O145 positive using molecular based methods. The difference in utilisation of certain carbon substrates from biological replicates from the PM1 (n=11) and PM2A (n=4) MicroPlates™ has highlighted that certain substrates would not be suitable for use in a differential media, as their utilisation by *E. coli* strains is inconsistent.

5. Results - Whole genome sequencing and comparative analysis

5.1 Selection of *E. coli* serogroup O145 strains for whole genome sequencing

A total of 53 *E. coli* serogroup O145 strains underwent WGS in this study (Appendix A) using the Illumina HiSeq platform with v4 chemistry (2 x 125 bp) (as described in section 2.9.1-5). Due to the small sample size of O145 isolates (73 isolates), the selection criteria for WGS was only one strain isolated from the same enrichment sample, with the same virulence profile determined by mPCR (as described in section 2.5.1). In one case, two O145 strains from the same enrichment were identified having different virulence profiles and consequently both underwent WGS. Of the 32 serogroup O145 strains isolated in this study (Appendix A), 13 underwent WGS with the remaining isolates obtained from various collections within the Hopkirk Research Institute.

5.2 Comparative genomics

5.2.1 Genome composition

A summary of the genome composition for the *E. coli* serogroup O145 strains (n=53) is shown in Figure 5.1, indicating genome length, GC content, CDS and number of tRNAs for each strain. The genome length and GC content were determined using QUAST [108] and the remaining parameters (CDS and number of tRNAs) were determined using Prokka [112], as described in sections 2.10.4 and 2.12.1, respectively. These data represent the analysis of all the *E. coli* serogroup O145 strains that underwent WGS, therefore the overrepresentation of *eae* subtype γ (n=41), compared to *eae* subtypes β (n=1), ε (n=7), ι (n=4), should be considered when interpreting the box and whisker plots.

The genome length for the serogroup O145 strains had an interquartile range of 5,219,500-5,310,820 bp, with a median genome length of 5,236,588 bp. There are a number of strains identified as outliers on the box and whisker plot for genome length. The four *eae* subtype ι strains have shorter genome lengths ranging from 4,663,358-5,010,707 bp. Within the *eae* subtype ι strains the two

ST-35 and ST-526 strains have similar genome lengths, as do the two strains belonging to ST-722. In addition, five out of seven *eae* subtype ε strains were classified as outliers on the upper limit of genome length, with the remaining two strains very close to the upper range of the box and whisker plot. The *eae* subtype β (n=1) and γ (n=41) strains mostly fall within the interquartile range.

Overall, the GC content of all serogroup O145 strains is relatively similar (50.22-50.67%) with an interquartile range of 50.43–50.47%. A number of serogroup O145 strains were identified as outliers according to GC content. Firstly, the *eae* subtype β strain had the lowest GC content of 50.22%, and five *eae* subtype γ strains had GC content between 50.31-50.37%. Secondly, the four *eae* subtype ι strains were on the upper GC content limit and three were identified as outliers, with GC contents ranging from 50.51-50.67%.

The number of CDS for the serogroup O145 strains had an interquartile range of 5,258-5,524. The number of CDS per strain showed a similar pattern to the genome length, the four *eae* subtype ι strains were outliers at the lower limit with CDS ranging from 4,453-4,825. Similarly, the seven *eae* subtype ε strains were outliers at the upper limit with CDS ranging from 6,081-6,281. The *eae* subtype β (n=1) and γ (n=41) strains were mostly within the interquartile range.

The interquartile range for the number of tRNA for the serogroup O145 strains was 89-91. The number of tRNA varied for the different *eae* subtypes: ι 80-88, β 88, ε 84-91 and γ 87-97. There was no distinct separation between the number of tRNA per strain and *eae* subtype, as was seen for other parameters, however this may be attributable to different genome assemblies.

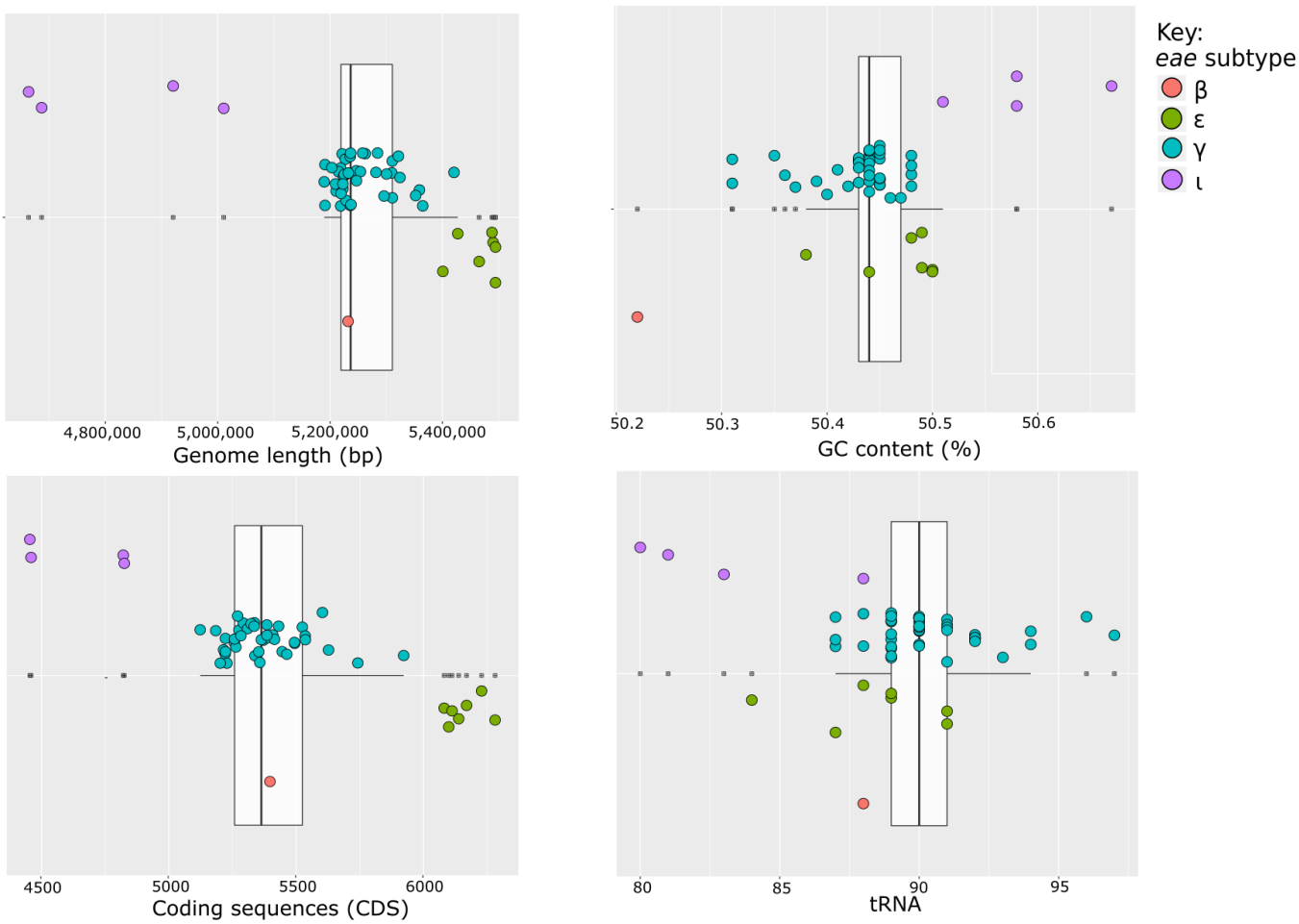

Figure 5.1: Box and whisker plots indicating the genome composition of *E. coli* serogroup O145 strains (n=53)

The box and whisker plots indicate the genome length (bp), GC content (%), number of coding sequences and number of tRNA for the serogroup O145 strains (n=53). Each data point is shown on the plots and has been colour coded according to *eae* subtype, as indicated by the figure key.

5.2.2 Virulence factors

The serogroup O145 genome sequences (n=53) were screened for 76 known *E. coli* virulence genes using the Center for Genomic Epidemiology (CGE) VirulenceFinder webserver [125]. A Neighbor-Net tree was produced using neighbor-joining methods according to the presence or absence of virulence factors (Appendix I) to visualise the relationship between virulence profile and metadata such as *eae* subtype, ST and isolation source (Figure 5.2).

The virulence profiles of the serogroup O145 isolates indicate that the strains cluster according to both *eae* subtype and ST. The *eae* subtype ι strains (n=4) cluster together, with variation associated with contrasting ST. In comparison to other serogroup O145 strains, the *eae* subtype ι strains carry only five or six of the 22 virulence factors listed in Figure 5.2. The *eae* subtype β strain also clusters separately having only five out of 22 virulence factors. The low number of virulence factors carried by *eae* subtype ι and β strains may be partially due to the absence of plasmid-acquired virulence factors, such as *etpD, ehxA* and *katP*, resolved using VirulenceFinder [125] (Appendix I). Using PlasmidFinder [126] on the CGE webserver (section 2.12.2), plasmids were detected in Trh7, Trh46 and R249-1 but appeared to be absent in Trh42 and 13ER6723A. In addition, the *eae* subtype ι (n=4), β (n=1) and two γ strains did not carry *ehxA*. The *eae* subtype ε strains (n=7, ST-17) carry between 14-16 virulence factors and cluster together with the two *stx1* positive ε strains present on a separate branch.

The *eae* subtype γ ST-137 (n=4) strains cluster separately from the *eae* subtype γ ST-32 strains. WGS data analysis of the four ST-137 strains indicate the presence of between 10-13 virulence factors (Figure 5.2). Notably, the plasmid-associated virulence factors *espP*, *iss*, *katP* and *toxB* were generally absent but *etpD* present in the four ST-137 strains, whereas the ST-32 *eae* subtype γ strains (n=37) are generally *espP*, *iss*, *katP* and *toxB* positive and *etpD* negative. The further 37 *eae* subtype γ strains comprised of 36 ST-32, and one unidentified ST, contain between 12 and 18 virulence factors and cluster together, including the *stx2* positive strains which together form their own monophyletic group. Although strains appeared to cluster together according to isolation source, further WGS data analysis of additional strains from different sources are required to confirm this hypothesis.

A comparison between the mPCR (described in section 2.5.1) and the virulence factors detected using the CGE VirulenceFinder (described in section 2.12.2) were consistent except for the isolate 13ER6723A which was identified as *stx*-negative by mPCR and *stx2*-positive by CGE VirulenceFinder analysis [125]. The *stx2* gene from the 13ER6723A WGS data was identified, and underwent BLAST analysis. The best BLAST hit was to *stx2f* (100% identity with top matches to *E. coli* Shiga toxin 2f genes serovar O145:H34 strain 65-4 (Accession no. AB499802.1) and to *E. coli* Shiga toxin 2f genes serovar OUT:HNM strain 17-8 (Accession no. AB499801.1)), a rare *stx2* variant, compared to the *stx2a* variant found in most serogroup O145 strains. The *stx2f* gene sequence was aligned with other *stx2a* gene variants and the forward and reverse *stx2* mPCR primers (Table 2.1) to determine any genetic differences which may have caused this variant to be undetected. This alignment identified a 5 bp mismatch within both the forward and reverse *stx2* mPCR primers, highlighting the absence of PCR amplification of this *stx2* variant in the mPCR. The *stx2f* variant was first described in pigeons [157] and has since been detected in humans, including in serogroup O145 isolates [158,159]. It has been suggested *stx2f*-positive *E. coli* are an emerging pathogen, causing milder infections compared to STEC serogroup O157 [158]. The anticipated prevalence of the *stx2f* variant among STEC populations should be considered when selecting PCR primers for detecting STEC virulence determinants. Alternative methods for the detection of *stx2f* have been outlined previously [160], as some PCR primers [93] such as those included in this study will not detect this *stx2* variant.

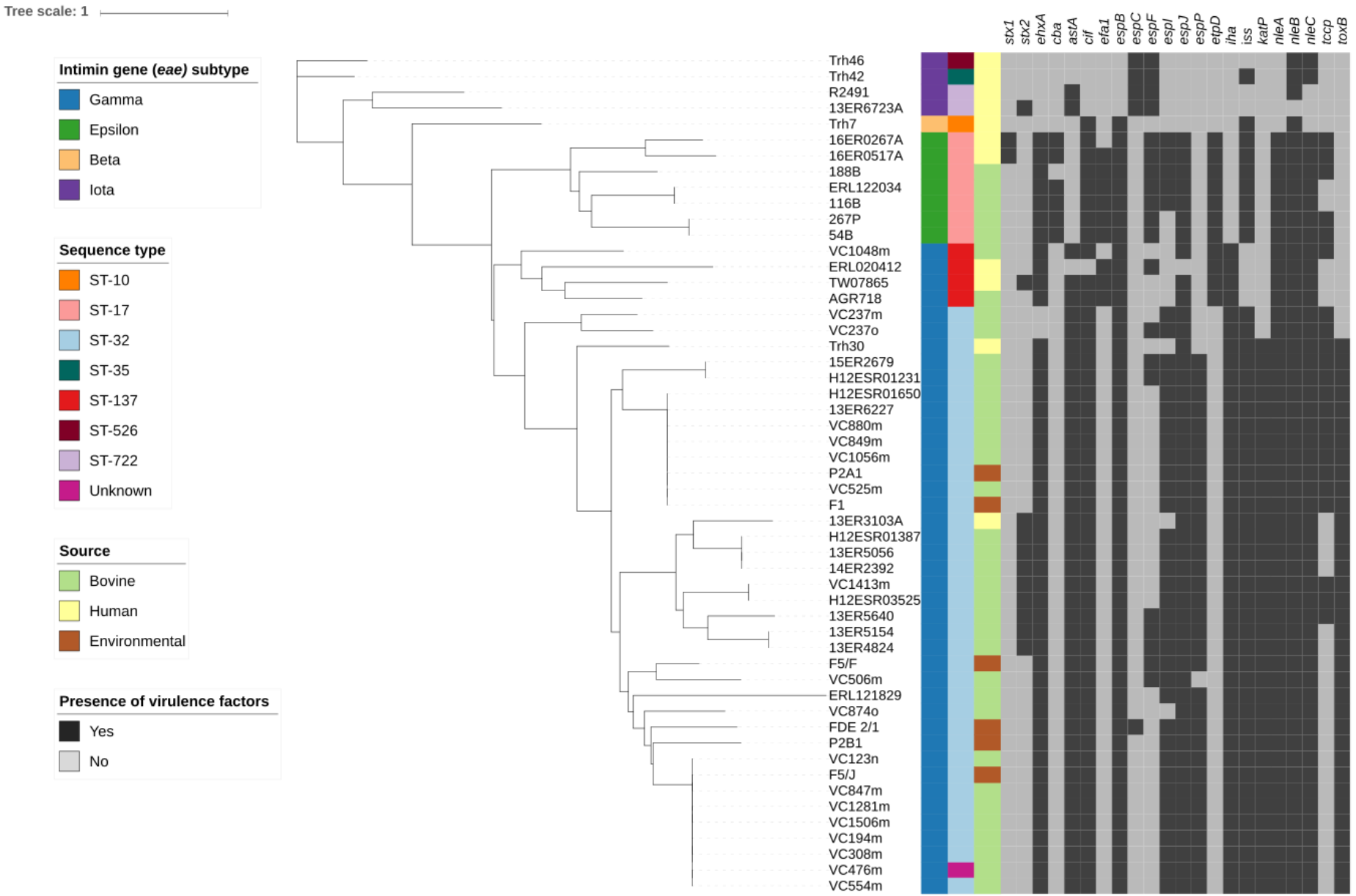

Figure 5.2: Neighbor-Net tree constructed using the presence or absence data from 31 virulence genes identified by the CGE VirulenceFinder webserver

The presence or absence of 31 virulence genes were used to make a Neighbor-Net tree using Euclidean distances. The tree was edited using the iTOL webserver. Isolate metadata is included for *eae* subtype, sequence type and isolation source, as indicated by the colour keys. The presence and absence of the virulence factors that differ between isolates (n=22) is indicated by the matrix. Virulence factors that were either present or absent in all of the strains were not shown in the matrix but were used to construct the tree.

5.2.3 *in silico* ribosomal multi-locus sequence typing

The rMLST analysis provides an indication of phylogenetic relationships using the SNP variations identified in 51 out of 53 genes encoding the ribosome protein subunits (*rps*, *rpm* and *rpl*). The *in silico* rMLST analysis (Figure 5.3) indicated several groups having identical rMLST profiles: the two *eae* subtype ι ST-722 strains, the two *eae* subtype ι strains belonging to ST-35 and ST-526, the *eae* subtype ε strains (n=7) and the *eae* subtype β strain (n=1). Among the *eae* subtype γ strains (n=41), four strains (TW07865 [ST-137], 13ER6227 [ST-32], Trh30 [ST-32] and 13ER3103A [ST-32]) have differing rMLST profiles, although there are very few SNP variations between these profiles.

5.2.4 Locus of enterocyte effacement pathogenicity island integration sites

The LEE pathogenicity island integration sites were identified in 40 of 53 serogroup O145 strains and are listed in Appendix J. The LEE is integrated near the tRNA *pheV* gene for the *eae* subtype ε strains (n=7) and β (n=1) strain. The LEE was integrated near the tRNA *selC* gene for the *eae* subtype ι (n=3) strains and γ (n=28) strains. The integration sites identified for specific *eae* subtypes are consistent with published studies identifying the LEE integration sites in other *E. coli* serogroups [36]. However, the *selC* LEE integration site for *eae* subtype ι strains has not been previously published. For one *eae* subtype ι strain, Trh42, the LEE integration site could not be precisely determined, however, the LEE was located near the tRNA *Leu* gene. The LEE insertion site could not be determined in fourteen strains, likely due to assembly inconsistencies arising during *in silico* analysis of incomplete genomes.

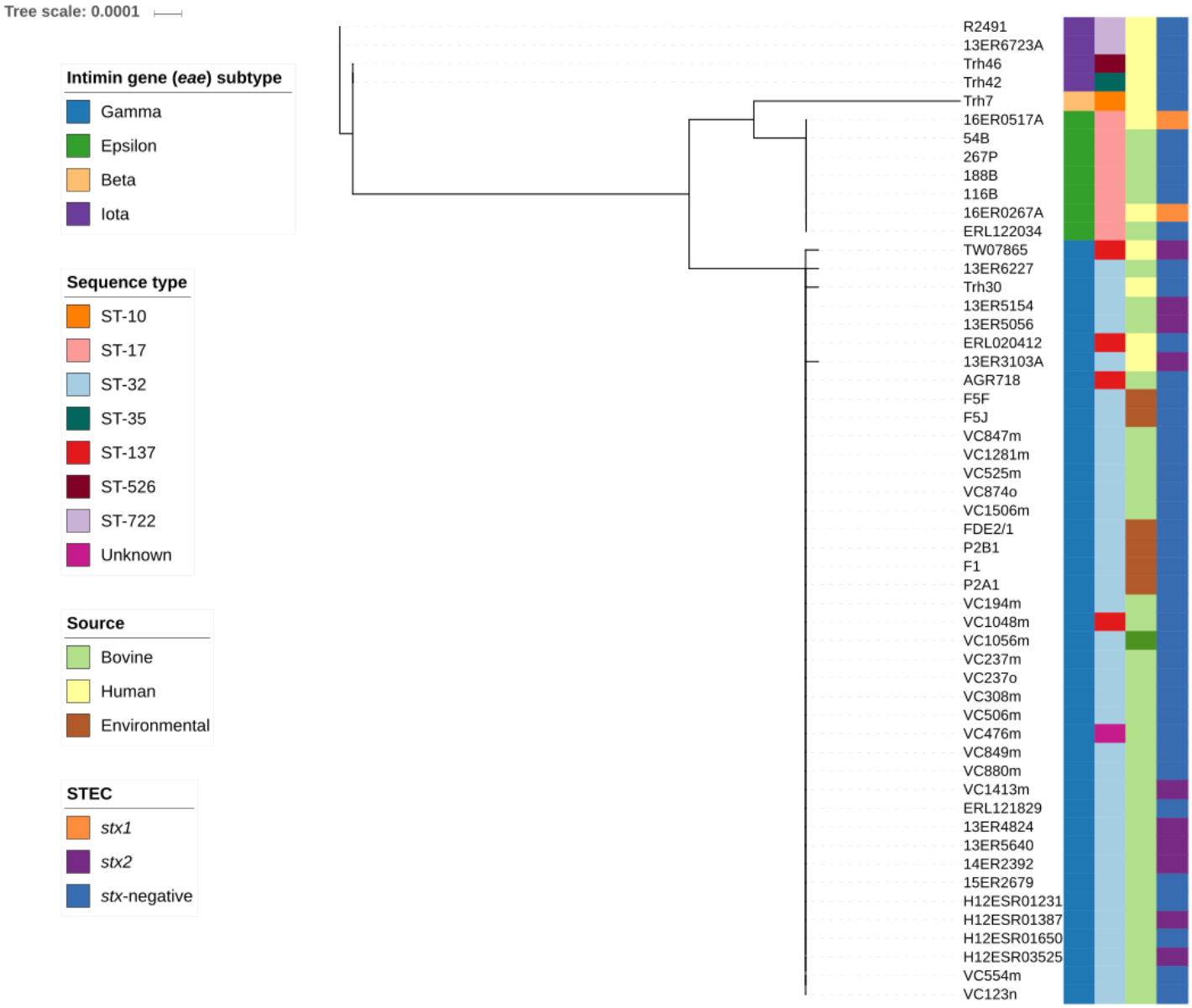

Figure 5.3: Neighbor-Net phylogeny constructed using *in silico* ribosomal multi-locus sequence typing

A Neighbor-Net tree was produced using rMLST and SNPs identified in the 53 genes encoding the ribosome protein subunits (*rps*, *rpm* and *rpl*). The *in silico* rMLST analysis was visualised using neighbor-joining methods in SplitsTree and edited using the iTOL webserver. Isolate metadata is included for ST, isolation source and whether the strains were *stx*-positive or *stx*-negative, as indicated by the colour keys.

5.3 Core single nucleotide polymorphism analysis

Core SNP variations were identified using Snippy [107] (as described in section 2.11.1), and the alignment from the core SNPs was visualised in SplitsTree [111] using neighbor-joining methods to examine the phylogeny of the serogroup O145 strains.

5.3.1 Core SNP analysis of serogroup O145 strains sequenced in this project (n=53)

Core SNP analysis of the 53 serogroup O145 strains WGS in this project (Figure 5.4) separated the strains into four main clusters, representing the four *eae* subtypes: subtype γ strains (n=41), subtype ι strains (n=4), subtype ε strains (n=7) and subtype β strain (n=1). For these analyses, an average of 93.18% bases aligned to the reference genome. Interestingly, a large number of SNPs (141,143 SNPs) were identified between the core genome of these strains, accounting for ~2.6% of the *E. coli* genome. This large proportion of core SNP variations suggests genetic heterogeneity among serogroup O145 strains and supports the hypothesis of the evolution of distinct *E. coli* phylogenetic lineages associated with different *eae* subtypes.

To resolve the phylogeny of the serogroup O145 strains, an additional core SNP alignment tree was produced examining the *eae* subtype γ strains. The core SNPs of the most numerous *eae* subtype (γ: n=41) provided a much higher resolution of the phylogeny of these strains (Figure 5.5) and revealed approximately a 30x reduction in the number of core SNPs identified, when compared with the core SNP alignment of all *eae* subtypes. For the *eae* subtype γ strains, 4,646 core SNPs were identified accounting for <0.1% of the *E. coli* genome. The number of aligned bases to the reference genome was an average of 96.20% for this analysis. The tree contained two major clusters which separated according to ST: one cluster contained the ST-32 strains (n=37) (VC476m ST unknown, but likely to be ST-32), and the other cluster contained the ST-137 strains (n=4). Within cluster variation was observed with seven out of ten *stx2*-positive strains clustering closely together and the environmental isolates clustering closely together. The environmental samples were all collected from the same two farms over different time periods, therefore it is difficult to determine whether these strains are genetically similar due to their isolation

source or their geographic origin. The range of isolation sources for these strains was human (n=4), environmental (n=6) and bovine (n=31) which is insufficient to provide any indication of similarities of strains according to isolation source. However, these preliminary analyses indicate that isolation source is not a significant determinant influencing the genetic similarity of serogroup O145 strains, and that other factors such as *eae* subtype and ST are more influential.

5.3.2 Core SNP analysis comparison with publicly available serogroup O145 strains

To further understand the genomic epidemiology of *E. coli* serogroup O145 strains, whole genome sequences from an additional 47 O145 strains were downloaded from publicly available databases (described in section 2.12.6, Appendix E) and compared with the 53 serogroup O145 strains sequenced in this project to offer a global comparison to New Zealand serogroup O145 strains sequenced. Core SNP analysis was performed for the 100 serogroup O145 strains (Figure 5.6) and in concordance with the previous analyses, a large number of SNPs (139,782) were identified accounting for ~2.5% of the *E. coli* genome. For these enhanced analyses, an average of 93.20% bases aligned to the reference genome. The strains were separated into four main clusters which correlated with *eae* subtype. One cluster contained the *eae* subtype β strains (n=2) and another *eae* subtype ε strains (n=7). The third cluster contained *eae* subtypes ι (n=6) and α2 (n=2) and the last contained *eae* subtype γ (n=83). These analyses including additional WGS data supports the hypothesis of independent acquisition of the LEE in distinct *E. coli* phylogenetic lineages.

To resolve the phylogeny of the strains sequenced in this project compared to publicly available serogroup O145 strains, core SNP analysis was performed on the *eae* subtype γ strains (n=83) (Figure 5.7). Core SNP analysis of the 83 *eae* subtype γ strains identified 7,272 SNPs, accounting for <0.1% of the genome. For these enhanced analyses on 83 *eae* subtype γ isolates, an average of 95.70% bases aligned to the reference genome. The ST-137 strains (n=5) clustered together as a separate branch from the main tree. Two strains AA053 (ST-137) and ewgs1003 (ST-32) cluster separately from the other isolates. In addition, the New Zealand serogroup O145 isolates on the top half of the tree are broadly

separated from the overseas isolates (UK, Germany, USA and Norway) at the bottom, with the exception of the ST-137 strains.

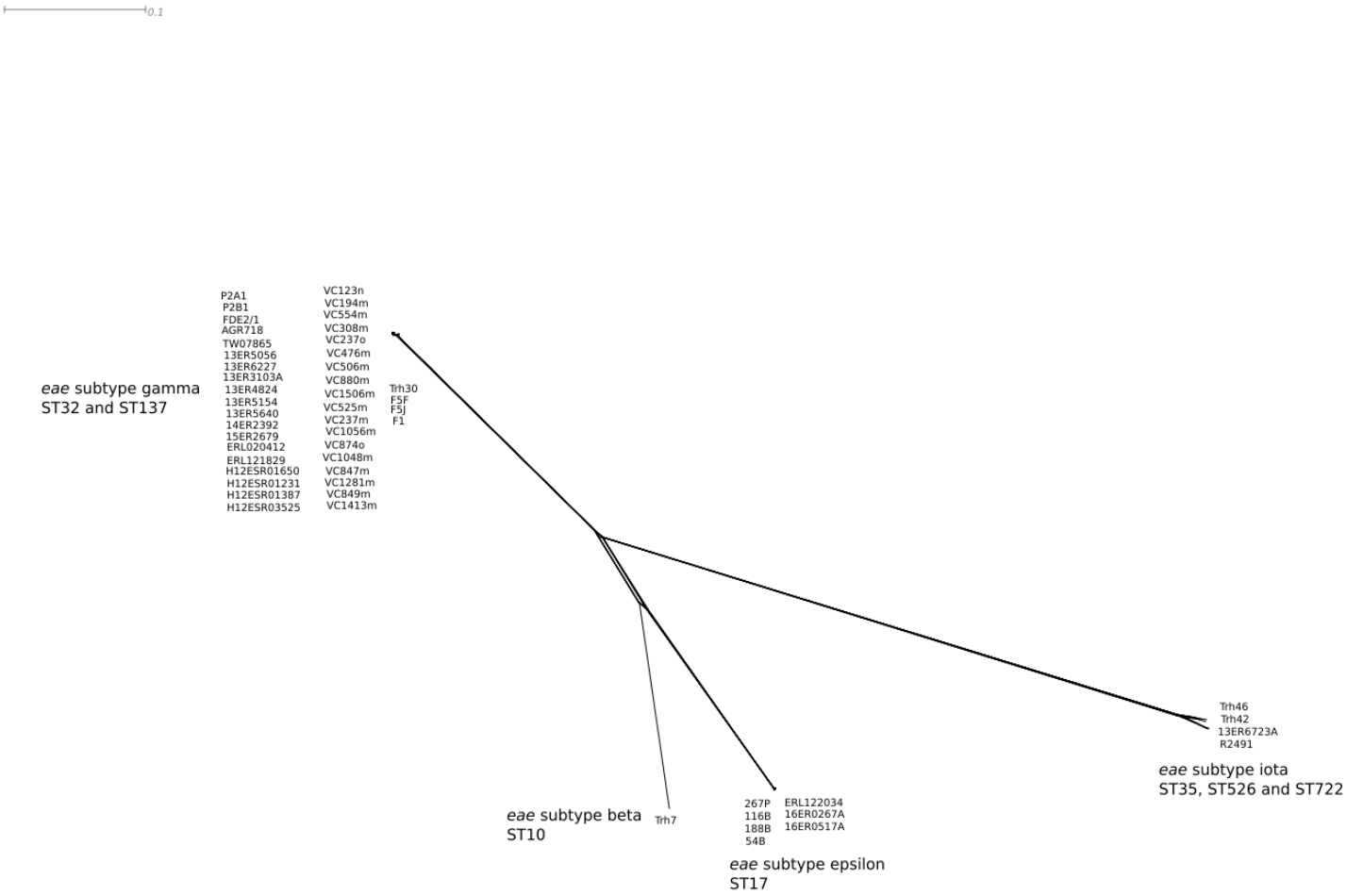

Figure 5.4: Neighbor-Net phylogeny of core SNP analysis from serogroup O145 strains sequenced in this study (n=53)

Neighbor-Net phylogeny of core SNP genome analysis from serogroup O145 genome sequences (n=53). The tree was generated using 141,143 core SNPs. Metadata is included for *eae* subtype and ST, and additional information for each isolate can be found in Appendix A.

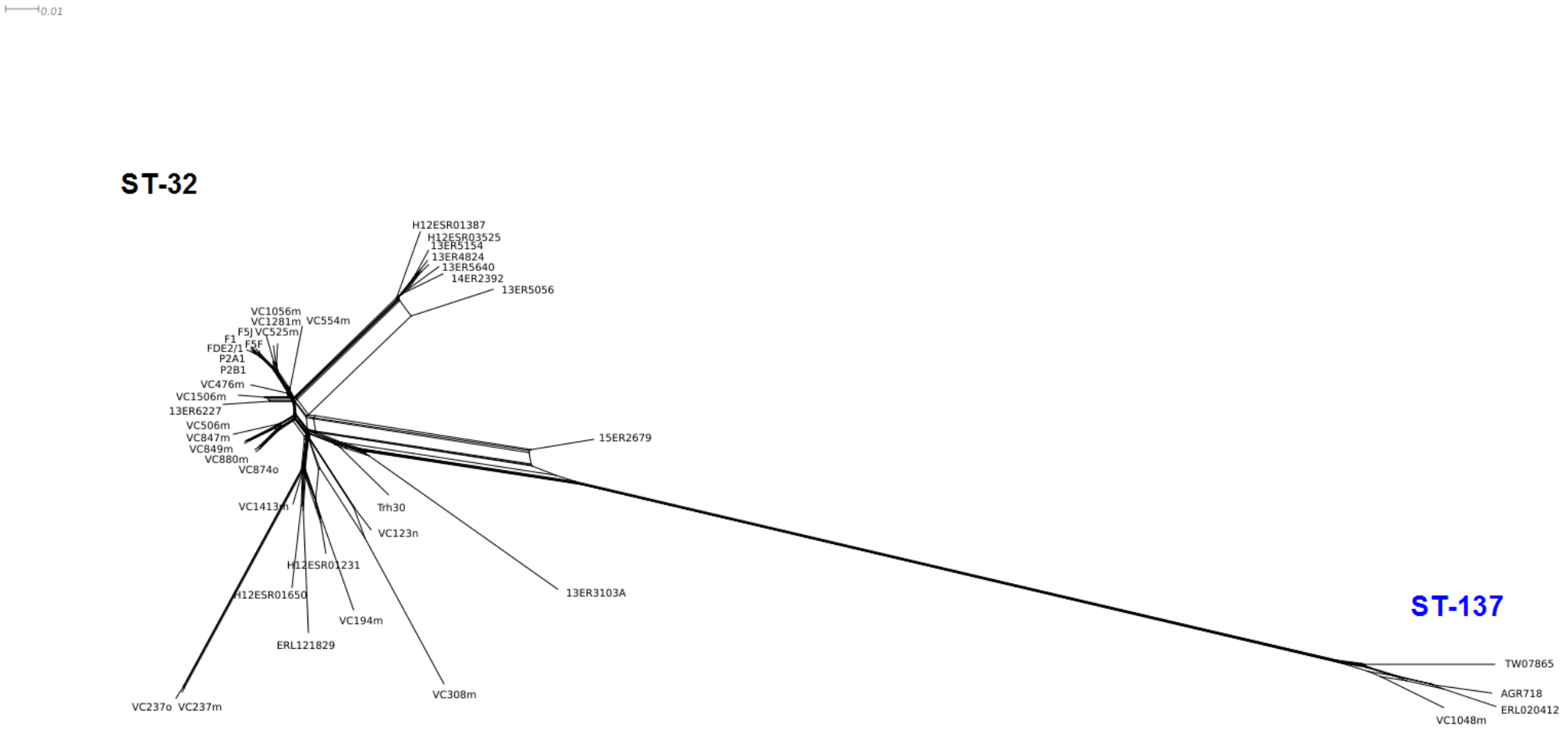

Figure 5.5: Neighbor-Net phylogeny of core SNP analysis from *eae* subtype γ serogroup O145 strains sequenced in this study (n=41)

Neighbor-Net phylogeny of core SNP analysis from serogroup O145 *eae* subtype γ genome sequences (n=41). The tree was generated using 4,646 core SNPs. Metadata is included for ST, and additional information for each isolate can be found in Appendix A.

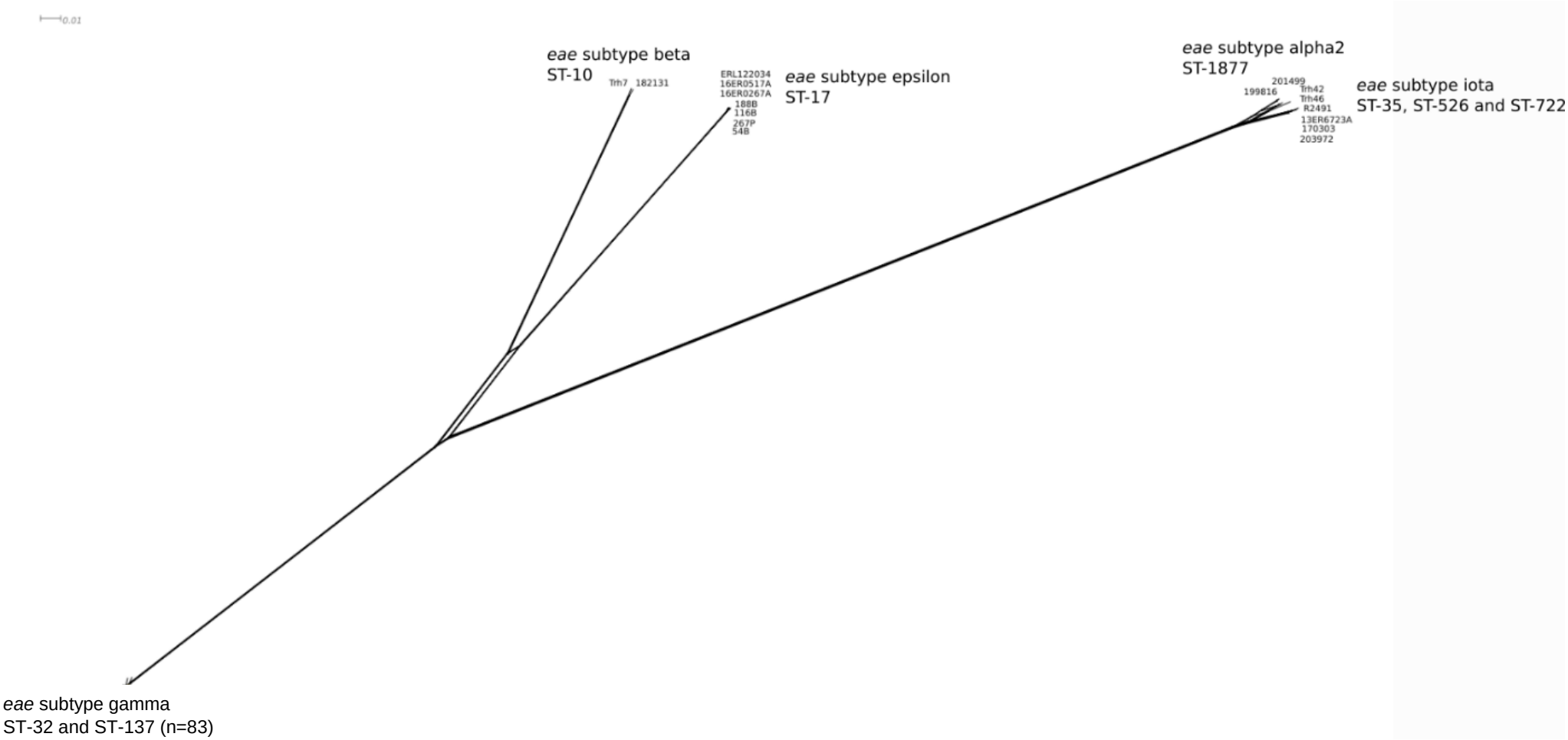

Figure 5.6: Neighbor-Net phylogeny of core SNP analysis of serogroup O145 strains sequenced in this study and publicly available serogroup O145 strains (n=100)

Neighbor-Net phylogeny of core SNP analysis from serogroup O145 isolates sequenced in this project (n=53) and publicly available whole genome sequences (n=47). The tree was generated using 139,782 core SNPs. Metadata for each isolate can be found in Appendices A and E.

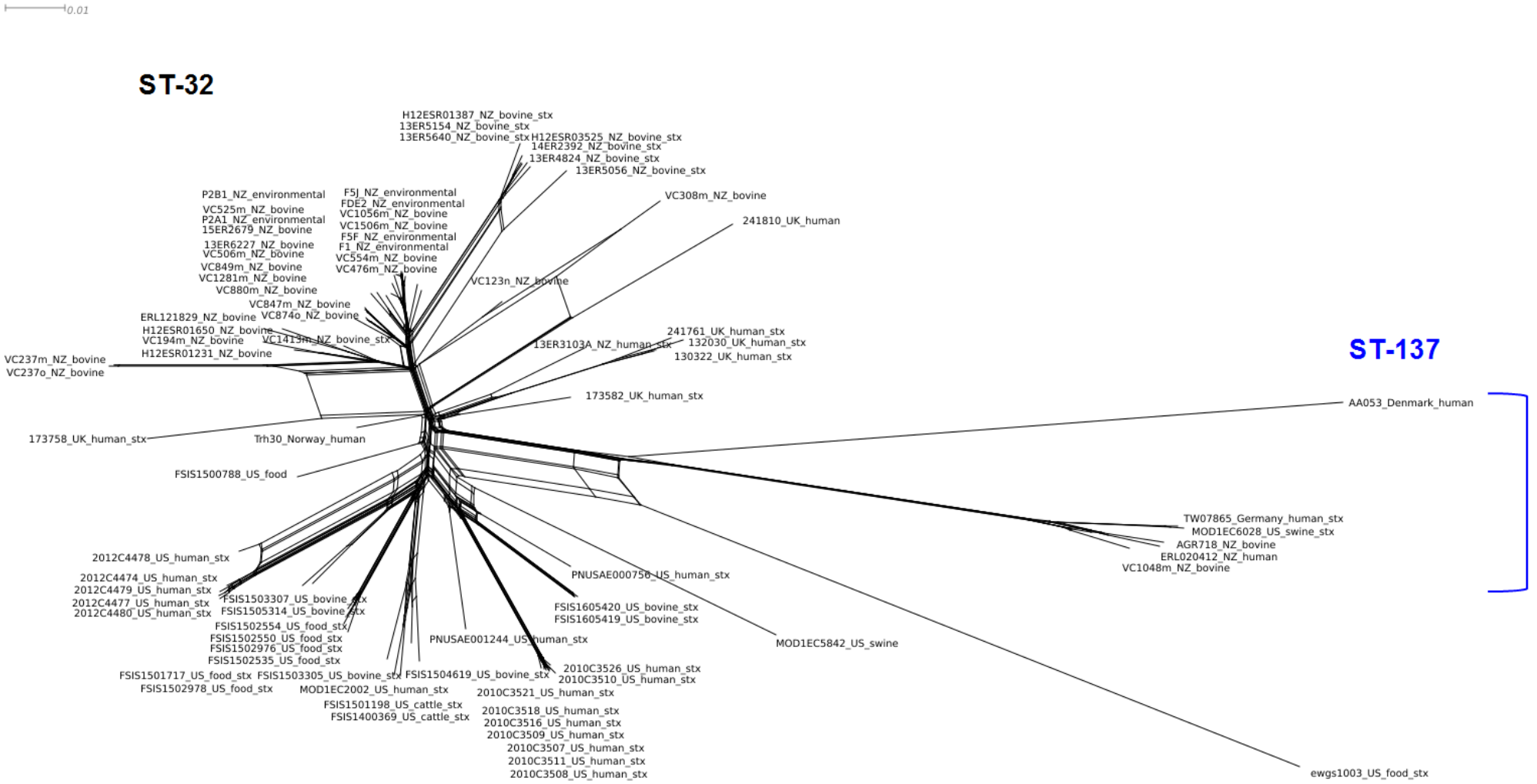

Figure 5.7: Neighbor-Net phylogeny of core SNP analysis of serogroup O145 *eae* subtype γ sequenced in this study and publicly available serogroup O145 *eae* subtype γ strains (n=83)

Neighbor-Net phylogeny of core SNP analysis from serogroup O145 *eae* subtype γ isolates sequenced in this project (n=41) and publicly available whole genome sequences (n=42). The tree was generated using 7,272 core SNPs. Metadata for each isolate can be found in Appendices A and E.

5.4 Core and pan genome analysis

5.4.1 Identification of the core and pan genome

The core and pan genome were identified for serogroup O145 strains sequenced in this study (n=53) using Roary [136] (as described in section 2.13.1). In this analysis the core genome is defined as genes present in all strains (n=53), the soft-core as genes present in ≥50 and ≤52 strains, the shell as genes present in ≥7 strains but <50 strains and the cloud as genes present in <7 strains. The pan genome is defined as the entire set of genes present in all strains in the dataset. The genome sequences were added in a random order for these analyses.

The number of conserved and total genes present in serogroup O145 strains (n=53) is shown in Figure 5.8. An increase in the number of genomes analysed is associated with an increased number of total genes, but the number of conserved genes decreased i.e. the core gene set decreasing and the number of unique genes increasing. The decrease of conserved genes as each genome is added is shown in Figure 5.9 and is associated with the heterogeneity of the serogroup O145 isolates. As each genome was added, the number of genes in the pan genome increased (Figure 5.10), suggestive of the pan genome for serogroup O145 strains (n=53) being open. For a given population, when additional genome sequences are included, an open pan genome will identify un-characterised genes, whereas a closed pan genome will have approached a constant [161]. The pan genome composition of the serogroup O145 strains (n=53) is summarised in Figure 5.11. The pan genome analysis suggested a core gene set of 3,242 genes, a soft-core of 105 genes, a shell of 2,982 genes and a cloud of 6,713 genes. The pan genome was also determined using OrthoMCL which predicted a similar size core gene set of 3,291 genes.

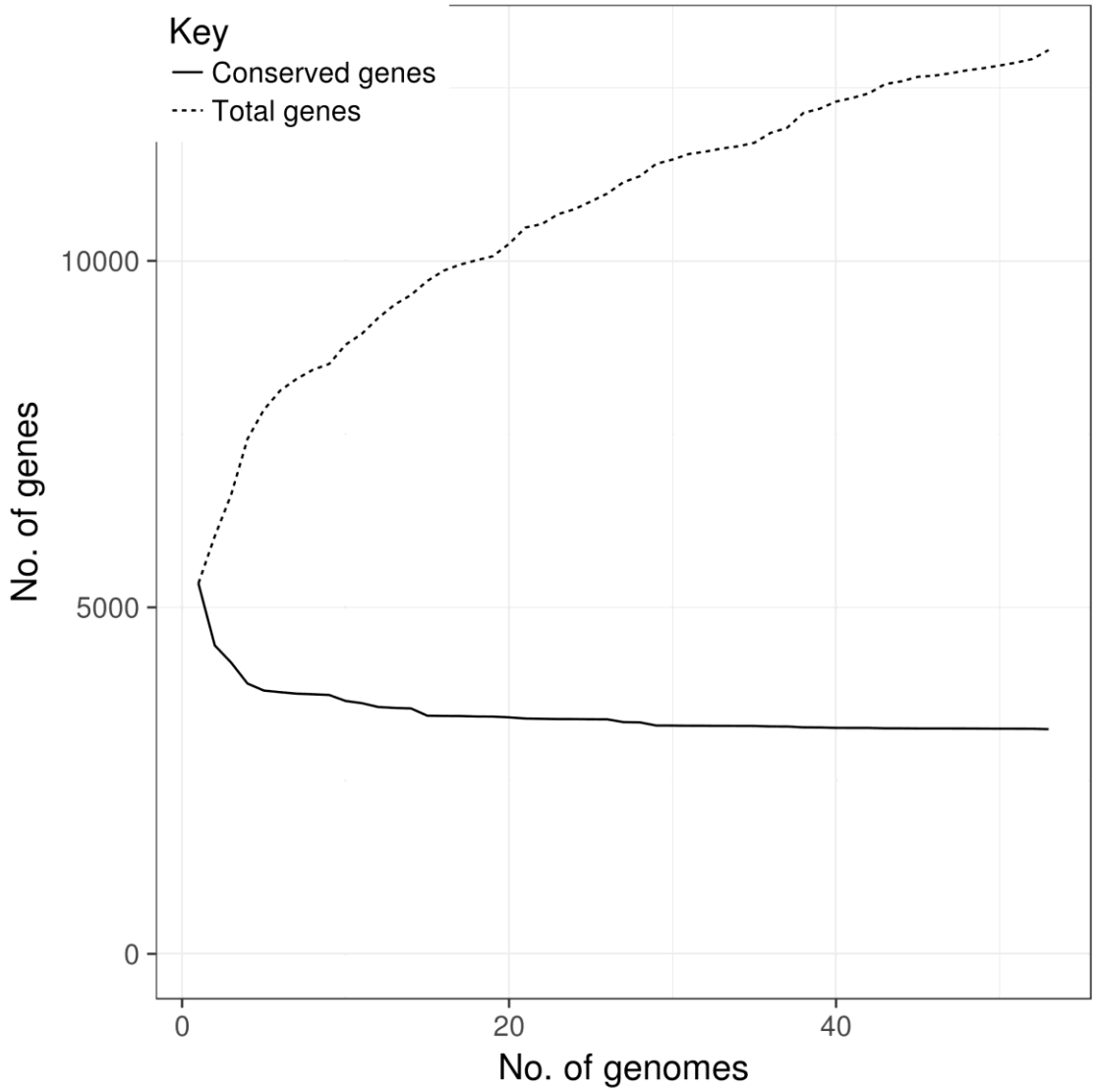

Figure 5.8: Comparison of the number of conserved and total genes in the serogroup O145 pan genome with increasing number of genomes

This analysis indicates the effect an increasing number of serogroup O145 genomes included in the analysis has on the number of conserved and total genes.

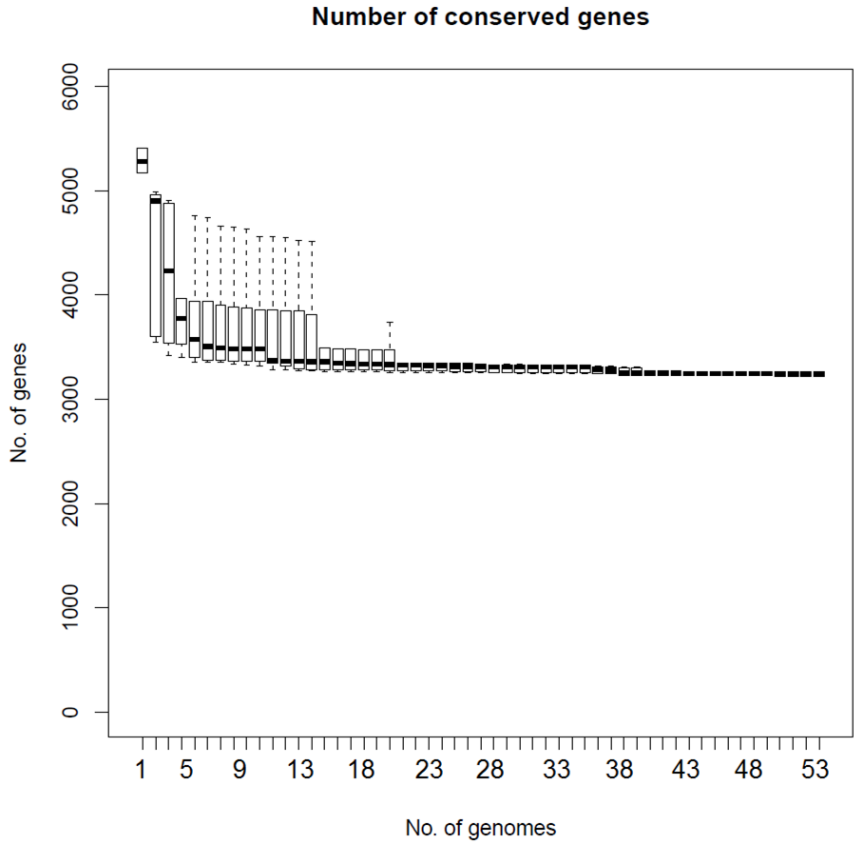

Figure 5.9: The effect the number of serogroup O145 genomes included in the analysis has on the number of conserved genes

This analysis indicates the effect an increasing number of serogroup O145 genomes included in the analysis has on the number of conserved genes.

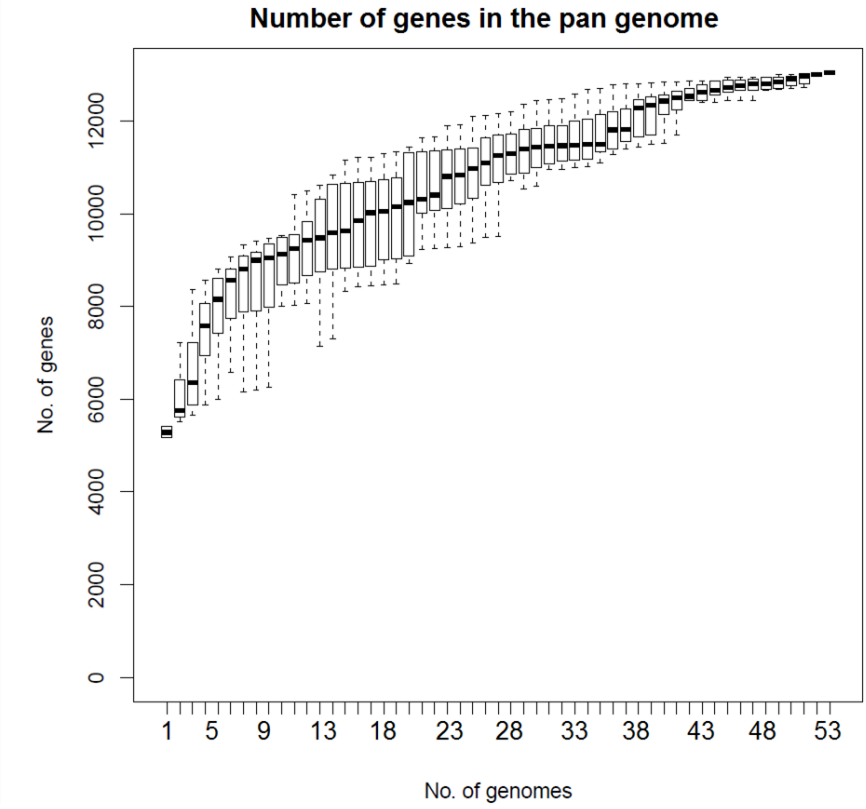

Figure 5.10: The effect the number of genomes included in the analysis has on the number of genes in the pan genome

The effect the number of genomes included in the analysis has on the number of genes in the pan genome.

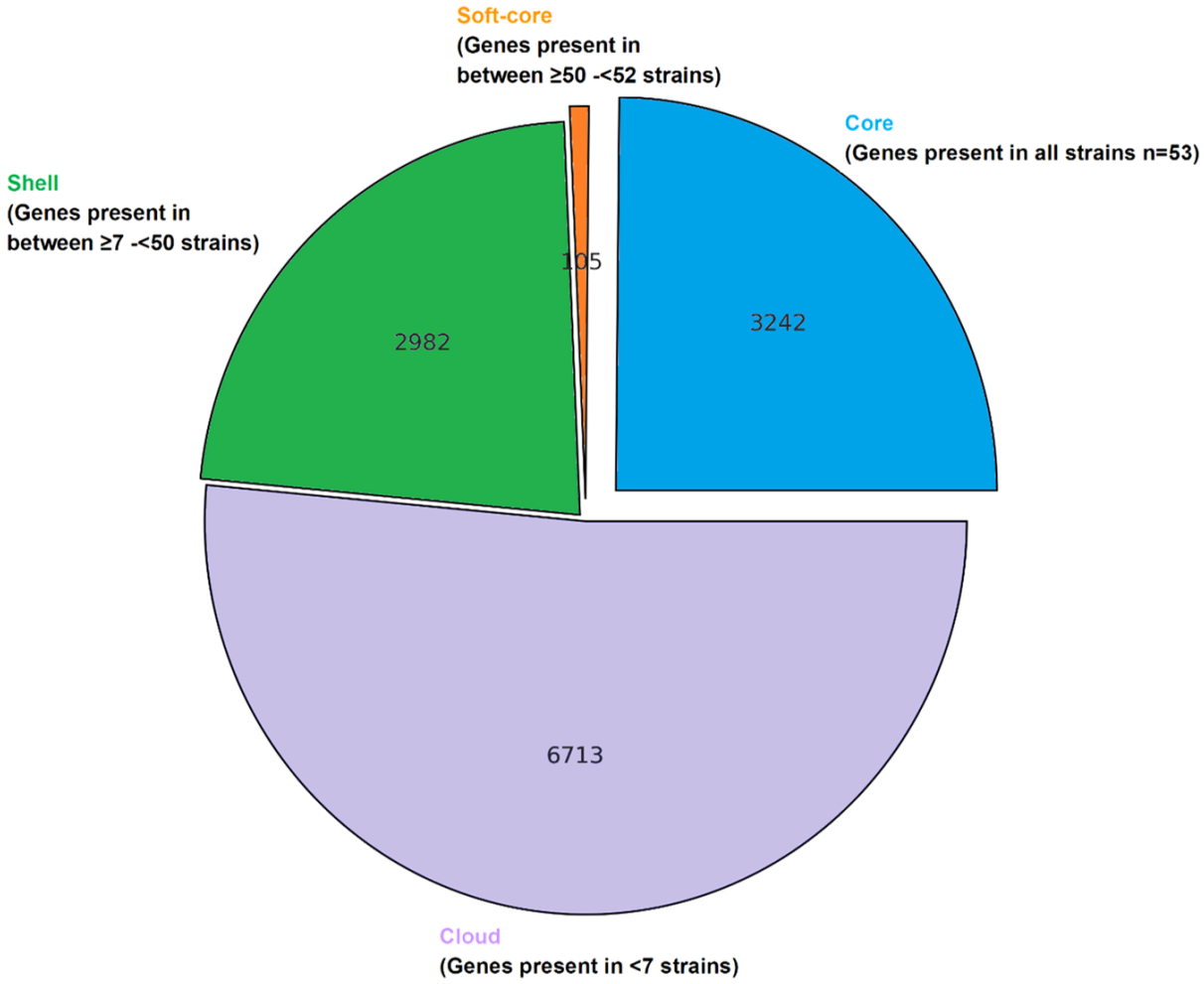

Figure 5.11: The pan genome composition of serogroup O145 strains (n=53)

Pan genome composition of serogroup O145 strains (n=53) showing the core, soft-core, shell and cloud gene sets as defined in the figure legends.

5.4.2 Association of pan genome with traits of interest

Scoary [142] (described in section 2.13.2) was used to calculate the association between genes in the pan genome and a number of traits of interest: (A) the '*eae* subtypes' carried by the strains (γ or non-γ), (B) 'isolation source' (human or animal origin), and (C) whether they were 'toxigenic' (*stx*-positive or *stx*-negative). The amino acid sequence of each gene identified associated with the three traits of interest (A, B and C described above) was obtained and the proteins were annotated into functional categories using BlastKOALA [143] (Figure 5.12). A number of proteins were unable to be annotated into functional groups by the system for *eae* subtype (16.3%) and whether the strains were toxigenic or non-toxigenic (18.2%). All proteins for the 'isolation source' trait were annotated. Hypothetical proteins identified in the Scoary analysis were not included in the functional analysis by BlastKOALA.

The diversity of functional groups associated with each trait varied. Proteins associated with '*eae* subtype' γ and non-γ strains belonged to the widest range of functional groups (Figure 5.12A). For the 'toxigenic' (Figure 5.12B) or isolation source (Figure 5.12C) traits, the functional categories 'cellular processes' and 'genetic information processing' were the predominant functional categories identified. For the trait '*eae* subtype', these functional categories are still present, however 'environmental information processing' and 'carbohydrate metabolism' were predominant.

For both 'isolation source' and 'toxigenic' traits, none of the proteins identified were explicitly involved with carbon metabolism, whereas three proteins associated with '*eae* subtype' were involved with carbon metabolism. These genes and proteins associated with carbon metabolism were *cysE* (serine acetyltransferase), *rpiB* (ribose-5-phosphate isomerase B), and *fbaA* (fructose-biphosphate adolase). The gene *cysE* was present in all '*eae* subtype' γ strains and absent in non-γ strains, whereas *rpiB* was present in all non-γ strains and absent in γ strains. The gene *fbaA* was present in all strains.

Interrogation of the pan genome to identify genes present in all *stx*-positive strains and absent in all *stx*-negative strains identified only the Shiga toxin genes. This finding highlights the genetic similarities between toxigenic and non-toxigenic

strains and suggests the insertion of the *stx*-encoding bacteriophage has occurred multiple times in *E. coli* phylogenetic lineages. Of the genes associated with ‘toxigenic’ strains, many were virulence associated genes, which were also found in *stx*-negative strains.

Although a number of genes associated with ‘isolation source’ were identified, none were found exclusively in human or animal origin strains. This suggests strains of human and animal origin are similar and cannot be definitively distinguished using this pan genome data. In comparison the majority of the genes associated with ‘*eae* subtype’ γ and non-γ strains were explicitly found in each group. For example, proteins were identified which were only present in *eae* subtype γ strains and were absent in non-γ strains and *vice versa*. This observation was true for genes found in all types of functional groups and highlights the genetic heterogeneity of serogroup O145. These data also demonstrate the genetic similarities within different *eae* subtypes and further supports hypothesis of the independent acquisition of LEE pathogenicity islands into distinct *E. coli* phylogenetic lineages.

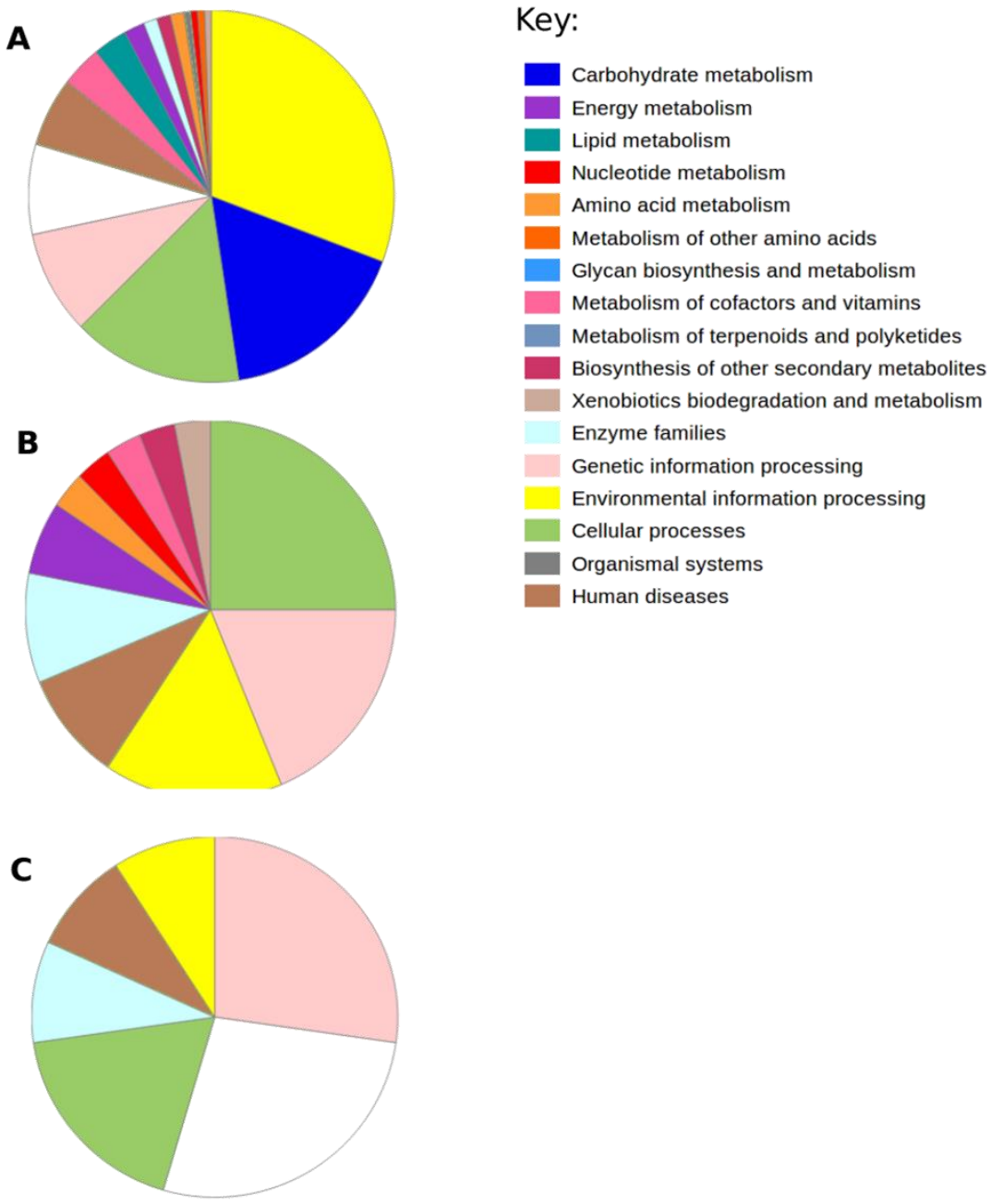

Figure 5.12: Functional analysis of proteins associated with traits of interest for serogroup O145

Scoary was used to calculate the association between genes in the pan genome and a number of traits of interest: the *eae* subtypes carried by the strains (γ or non-γ), isolation source (human or animal origin), and whether they were toxigenic (*stx*-positive or *stx*-negative). The amino acid sequences for each gene identified that was associated with a trait of interest were obtained and the proteins were annotated into functional categories using BlastKOALA. Functional groups of proteins, as indicated by the colour key, white indicates proteins which were unclassified, associated with whether strains **A.** carried *eae* subtype γ or non-γ; **B.** are of human or animal origin; or **C.** are *stx*-positive or *stx*-negative.

5.5 Evolutionary analysis of serogroup O145 strains

To estimate the TMRCA of the serogroup O145 strains sequenced in this project, BEAST 2.0 [137] was used (as described in section 2.12.5). Only *eae* subtype γ strains (n=40) were included in this analysis as the recombination regions could not be removed with Gubbins from non-γ *eae* subtypes due to the high level of diversity amongst these strains. In addition, one *eae* subtype γ strain (TW07865) was excluded from the analysis as an accurate isolation date, required for calibration of the tip dates, could not be obtained.

5.5.1 Mutation rate and estimated TMRCA of *E. coli* serogroup O145 *eae* subtype γ strains

Phylogenetic analysis using a relaxed molecular clock assumes all branches of the tree have different mutation rates. In this study the predicted mutation rate of the *eae* subtype γ isolates (n=40) was 7.41×10^{-7} substitutions/site/year. The ancestral date reconstruction analysis predicted that the isolates shared a most recent common ancestor in approximately 1828 (Figure 5.13) (95% highest posterior density (HPD) interval 1748-1897).

Since the estimated TMRCA, serogroup O145 strains further diverged according to ST. The ST-32 strains diverged approximately 40-80 years ago (95% HPD interval 1939-1981), with more recent divergent events among the different isolates, compared to the ST-137 strains which diverged approximately 30-50 years ago (95% HPD interval 1965-1988). The STEC 13ER3103A strain, a human clinical isolate from New Zealand, diverged earlier (95% HPD interval 1939-1981) compared to the remaining isolates which further diverged between 1954 and 1986 (95% HPD interval). These results also suggest that STEC serogroup O145 strains have emerged multiple times in the last 100 years. Identification of the common *stx*-bacteriophage insertion sites for serogroup O145 [33] were analysed (section 2.12.5) to identify whether these sites were occupied or available in *stx*-negative strains. Although, some sites were vacant in the majority of *stx*-negative strains, not all insertion sites could be detected in the genomes. This may be due to these genes being unannotated, the sites being occupied and the insertion site therefore disrupted or due to assembly issues. Only one overseas isolate (Trh30) was included in the analysis and therefore no conclusions can be drawn about geographic origin and time of convergence.

The TMRCA for the *eae* subtype γ strains cannot be applied to the non-γ *eae* subtypes. The core SNP genome analysis and virulence profiles have indicated that these strains are highly heterogeneous and have likely evolved from distinct phylogenetic lineages, and therefore are unlikely to share the same most recent common ancestor.

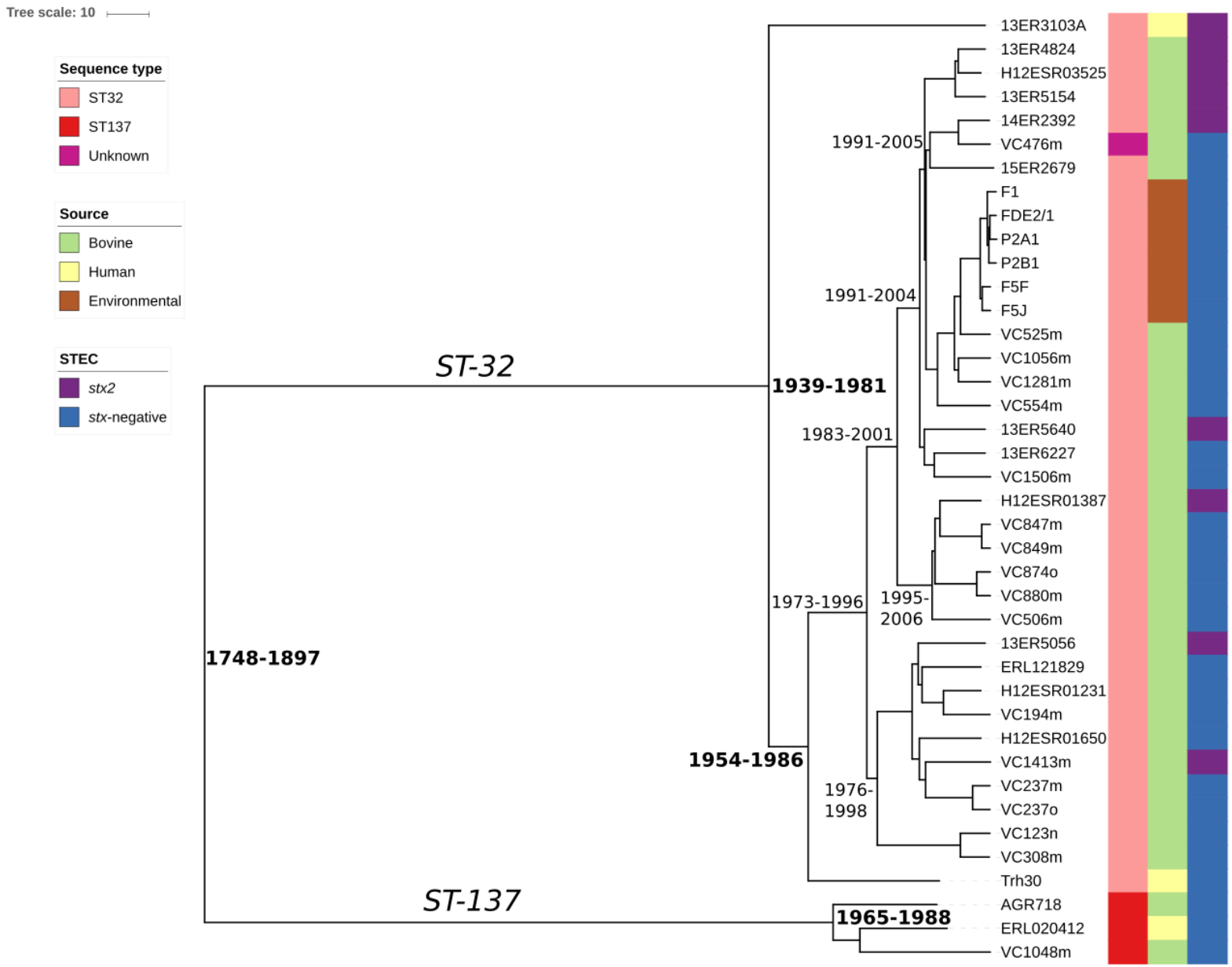

Figure 5.13: Maximum clade credibility tree showing predicted dates serogroup O145 *eae* subtype γ strains last shared a common ancestor

Maximum clade credibility tree based on BEAST 2.0 analysis for *eae* subtype γ strains (n=40) shows the predicted dates these strains last shared a common ancestor. The 95% HPD intervals are only indicated for key nodes on the tree. Metadata has been included for the sequence type, isolation source and whether the strains were toxigenic or non-toxigenic, as indicated by the colour keys. The tree scale is 10 years.

5.6 Discussion

The genome composition of the serogroup O145 strains sequenced in this study is consistent with other comparative genomic studies of *E. coli*. A study by Lorenz et al. [87] analysed the genomes of *E. coli* serogroups O145, O26, O103, O111, O157 and O165 and reported genome lengths (5,191,000-5,697,000 bp), number of CDS (5,179-5,780) and GC content (50.5-50.7%) [87], generally consistent with the genome composition described in this study (Figure 5.1). In contrast, the genome lengths for *eae* subtype ι strains (n=4) were smaller. These *eae* subtype ι strains had substantially fewer virulence factors compared to the other serogroup O145 strains, which may be indicative of fewer mobile genetic elements and the smaller genome size for these strains. The number of CDS was consistent for the *eae* subtype γ (n=41) and β (n=1) strains, however the number of CDS from published data [87] was inconsistent for the *eae* subtype ε and ι strains, which were greater than, and less than, data from this study, respectively. The GC content was also consistent for most O145 strains, however, the *eae* subtype β strain (n=1) had a slightly lower value (50.22%). A comparative genomic study of WGS data of O26, O103, O111 and O157 genomes also identified genome sizes (5,371,000-5,697,000 bp) and number of CDS (5,264-5,609) consistent with *eae* subtype γ (n=41) and β (n=1) strains (Figure 5.1) [47]. Comparison of O26, O103, O104, O111, O145 and O157 genomes indicated genome sizes between 5,273,000-5,697,000 bp, number of CDS between 4,972-5,613 and GC content of 50.4-50.7% [46]. In particular, the serogroup O145:H28 strains (n=2) had genome lengths of 5,737,294 and 5,559,008 bp, CDS of 5,776 and 5,512 and a GC content of 50.7% [46]; in comparison to the serogroup O145:H25 strains (n=2) which had genome lengths of 5,377,000 and 5,374,000 bp, they had 5179 and 5193 CDS and a GC content of 50.5% [87]. These findings were comparable to the serogroup O145 strains genome sequenced in this study (Figure 5.1). However, Lukjancenko et al. [162] suggested that genomes which are in multiple contigs, such as in this study, may have an overestimated genome size.

More tRNAs were reported by Lorenz et al. [87] (96-106), Cooper et al. [46] (93-110) and Ogura et al. [47] (98-106) compared to this study (80-97). However, a study by Iguchi et al. [86] comparing other pathogenic types of *E. coli* (EPEC,

EHEC, uropathogenic *E. coli* (UPEC) and commensal *E. coli*) noted a tRNA range (81-105) similar to that found in this study. Lower numbers of tRNA (81-94) were also found in other EPEC, ETEC, extraintestinal pathogenic *E. coli* (ExPEC) and commensal and laboratory *E. coli* strains [47].

Although the *eae* subtype ι strains (n=4) had smaller genome sizes in comparison to the other strains sequenced in this study, similar sized genomes have been reported for other pathogenic and commensal *E. coli* such as an EPEC O127, a UPEC O6, an ETEC O139 and a commensal O9 [86]. The number of CDS reported for these strains is also consistent with the *eae* subtype ι strains [86]. A STEC/ETEC bovine serogroup O2 strain of a similar genome size (4,907,103 bp) was also reported by Nyholm et al. [89]. Interestingly, two genome sequences described by Cooper et al. [46] reported the number of CDS for an O111 and O104 strain to be 4972 and 4975, respectively which were similar to the *eae* subtype ι strains. The smaller genome size of *eae* subtype ι strains may be due to the lower number of virulence factors carried by these strains (Figure 5.2), representative of these phylogenetic lineages (represented by ST-35, ST-526, and ST-722) or those strains carrying *eae* subtype ι. These hypotheses are difficult to test as many studies do not report the *eae* subtype carried by the isolates and this subtype has been reported to be rare [163].

Analysis of the core SNP phylogenies identified a large number of SNPs which represents considerable heterogeneity among the serogroup O145 strains. Genome wide core SNP analysis of 69 *E. coli* strains also identified significant genetic diversity with 86,350 SNPs identified across 1,371 core genes [87]. A reduced number of SNPs identified compared to this study may be partially attributed to the smaller core genome being analysed.

The pan genome analysis of serogroup O145 strains (n=53) suggested a core gene size of 3,242 genes, a soft-core of 105 genes, a shell of 2,982 genes and a cloud of 6,713 genes (Figure 5.11). The pan genome consisting of >10,000 genes is open (Figure 5.10) which demonstrates the heterogeneity of this dataset, and supports the findings of distinct phylogenetic lineages. A number of published studies have provided varying numbers of genes associated with the core and pan genome of *E. coli*. A study comparing commensal and laboratory strains of *E. coli* with pathogenic strains (various EPEC, ExPEC, ETEC and UPEC) (total

genome sample size n=17) predicted a core genome size of 2,344 genes [164]. Pan genome analysis of 53 *E. coli* strains identified a core gene set of 1,472 conserved gene families (genes were categorised as a gene family if they shared ≥50% amino acid identify against ≥50% of the length of the longest gene) [162]. The presentation of the core gene set as gene families makes it difficult to compare with other studies, as it is unclear how many genes are included in a gene family. In contrast, comparison of two *E. coli* serogroup O145 strains identified a large core gene set of 5,173 [46]. This supports previous findings that the core genome size decreases as the number of genomes analysed increases (Figure 5.9), and conversely the pan genome size increases as additional genomes are added for an open pan genome (Figure 5.10). When analysing the pan genome of 10 STEC strains, Cooper et al. [46] reported a core gene set of 4,192 genes, which is still comparatively larger compared to other studies. Lastly, a study of serogroup O26 isolates (n=373) from 11 countries identified a core genome size of 3,254 genes which is very similar to the core genome of serogroup O145 strains identified in this study [165].

Similar to this study, Rasko et al. [164] found the pan genome of the *E. coli* strains analysed to be open, as the inclusion of additional genome sequences continued to identify unique genes and that the pan genome size of *E. coli* is >13,000 genes [164]. Lukjancenko et al. [162] identified a pan genome set of 13,296 gene families when comparing 53 *E. coli* strains. Browne et al. [165] identified a large accessory genome of 40,117 genes which is likely due to the large number of genome sequences analysed (n=373) from a number of countries. Interestingly, the analysis of a large number of genome sequences (n=373) is still indicative of an open pan genome [165].

Pan genome analysis has indicated a significant proportion of the *E. coli* genome being comprised of diverse genes. This represents the heterogeneity of *E. coli* and its ability to alter its genome through acquiring new genetic material via HGT, incorporation of phage genetic material and through gene loss or duplication. The diversity of the *E. coli* genome may explain why this species can occupy many ecological niches and cause diseases of varying severity [162]. This heterogeneity of the *E. coli* genome supports the “Public goods hypothesis for the evolution of life on Earth” [166].

The discrepancies between the core and pan genomes identified in published studies is likely due to a variety of factors such as the number of genomes analysed. In general, it has been shown as the number of genomes analysed increases, the core genome size decreases and the pan genome size increases if it is open (Figures 5.9 and 5.10) and therefore the size of the dataset will have an impact on the pan genome. Secondly, the genetic similarity of the genomes being analysed is important. Strains that are genetically very similar will have a larger core genome size and a smaller accessory genome. These characteristics were noted during the analysis of two serogroup O145 strains by Cooper et al. [46]. Conversely, strains that are genetically distinct will have a small core genome size and a larger accessory genome.

Lastly, the software and parameters used to define the core and accessory genomes are essential factors to consider when comparing pan genomes between studies. For example, Cooper et al. [46] reported a large core genome size using a more relaxed definition of a core genome compared to other studies (a core sequence was defined as a protein present in all strains, with ≥75% identity across ≥75% of the sequences). In comparison, Rasko et al. [164] used BLAST score ratio analyses to identify the core and accessory genome and a threshold value of approximately >80% over the length of the protein. Lukjancenko et al. [162] used a combination of methods described by Friis et al. [167] and Snipen and Ussery [168] for the identification of the pan genome, which included construction of a pan genome tree using hierarchical clustering and Manhattan distance and adding genomes to the analysis non-randomly based on the pan genome tree with subsequent identification using a BLAST matrix [167]. The pan genome identified in this study and by Browne et al. [165] was identified by the software Roary, which uses a minimum identity for BLASTP of 95% and the definition of the core and accessory genes as defined in Figure 5.11. The different parameters used to define the pan genome in various studies make comparisons difficult and therefore may warrant the implementation of a standardised approach to defining pan genomes in comparative genomic studies. However, it is likely that all of these factors will have an impact on the core and accessory genomes identified in studies rather than a single factor alone, and should be considered when comparing studies.

The TMRCA of the *eae* subtype γ ST-32 and ST-137 strains was in 1828 (95% HPD interval 1748-1897) which is similar to the estimated TMRCA for other *E. coli* populations. Dallman et al. [169] estimated the TMRCA for *E. coli* O157 strains circulating in the UK prior to 2015 to be 1840 (95% HPD interval 1817-1855). Similarly, global ETEC isolates from the 1980s-2011 were analysed and these lineages were estimated to have emerged between the 1840s and 1970s [170]. The hypothesised TMRCA for New Zealand and global serogroup O26 ST21 strains was in 1865 [171] and more recent evolution of geographically distinct clones has been suggested to correlate with importations of livestock [171,172]. Due to ruminants being a well-known reservoir of STEC [7,173] it is likely that such importations of cattle are a potential source of STEC into New Zealand. However, there is insufficient information to suggest a likely specific introduction date of serogroup O145 strains now circulating in New Zealand due to the lack of O145 from other geographically distinct countries being included in this specific study for comparison.

The serogroup O145 *eae* subtype γ strains were estimated to mutate at a rate of 7.41×10^{-7} substitutions/site/year. This was consistent with substitution rates for global ETEC lineages of 1.0×10^{-6}-4.0×10^{-7} sites/ year [170], for STEC O157:H7 in the UK (2.6 mutations genome^{-1} year^{-1} [169], and for global *Shigella* species (6.0×10^{-7} substitutions/per year) [174]. Similar mutation rates between studies may indicate common selective pressures bacteria are under to adapt to their environment.

5.7 Summary

Comparative analysis of whole genome sequence data has demonstrated considerable genetic heterogeneity among serogroup O145 strains (n=53) and indicated that the O145 serogroup consists of distinct phylogenetic lineages. The genome composition, virulence profiles, rMLST and core SNP analysis indicate clustering of the strains is correlated with both *eae* subtype and ST, which is likely indicative of distinct phylogenetic lineages and independent acquisition of the LEE pathogenicity island. Inclusion of further WGS data from publicly available serogroup O145 strains and core SNP analysis underscores these data. Pan genome analysis suggests a core gene set of 3,242 genes and an open pan genome of more than 10,000 genes, highlighting the genetic diversity within the serogroup O145 strains. The identification of carbon metabolism genes found in certain *eae* subtypes and STs suggests there may be potential targets for the development of a differential culture media for serogroup O145 which will be investigated further (Chapter 6). Ancestral data reconstruction analysis of the serogroup O145 *eae* subtype γ strains predicted the isolates shared a date a TMRCA in approximately 1828 (95% HPD interval 1748-1897).

6. Results - Phenotype and genotype correlations

6.1 Association of genes in the pan genome and specific carbon substrate utilisation

Scoary [142] (described in section 2.13.2) was used to identify genes in the pan genome associated with specific carbon substrate metabolism by *E. coli* O145 using PM1 and PM2A MicroPlates™ and the Omnilog phenotypic microarray system (Chapter 4). Of the 18 carbon substrates identified as potential targets for use in a differential media (described in section 4.3), eight (Table 6.1) were selected for further investigation due to the differential metabolism and growth of O145 strains (e.g. substrates able to support the growth of all or defined *eae* subtypes) and the complexity of the pathways involved in the metabolism of each substrate. For each carbon substrate, the growth of serogroup O145 strains analysed using the Omnilog system (Appendix H) was scored as either substrate utilisation (≥51 Omnilog Units) or no substrate utilisation (≤50 Omnilog Units) the substrate. *E. coli* O145 growth and utilisation of the carbon substrate was defined as the 'trait' in the Scoary analysis, and associations with the utilisation of the specific substrate and genes in the pan genome were identified. For Scoary analysis of each candidate carbon substrate, output amino acid sequences were obtained and 100 proteins with a p-value <0.05 were functionally annotated using BlastKOALA [143] (described in section 2.13.3). Hypothetical proteins were not included in the functional analysis.

6.2 Diversity of protein functional groups associated with the utilisation of specific carbon substrates

For each carbon substrate investigated, the functional groups and number of proteins (transcribed *in silico*) within each group varies, however, the predominant functional groups are similar (Figure 6.1). For all carbon substrates, the functional groups 'environmental information processing' and 'cellular processes' predominate. However, the functional groups 'genetic information

processing’ and ‘amino acid metabolism’ are prevalent for some but not all of the substrates. These predominant functional groups are all essential processes. Proteins involved in carbohydrate metabolism were identified for all substrates examined.

Table 6.1: Carbon substrates selected for further investigation to identify phenotype and genotype correlations

Carbon substrate§	PM plate§‡	Well position§	Utilised by strains with *eae* subtypes§	Proteins annotated (%)†	Proteins involved in carbohydrate metabolism†	Proteins involved in carbon metabolism†	Gene involved in carbon metabolism†
D-serine	PM1	B01	β and γ	79.2	14	1	*fbaA*
D-galactonic acid-γ-lactone	PM1	C02	β and ι	83.8	12	1	*atoB*
D-tagatose	PM2A	D06	ι	80.6	13	4	*fbaA, lpd, kdgK, acnA*
Sucrose	PM1	D11	ε and ι	79.2	14	1	*fbaA*
β-hydroxy-butyric acid	PM2A	E08	γ *(11/12)*	84.7	12	1	*fbaA*
D-lactic acid methyl ester	PM2A	F01	β, ε, ι and γ (*1/12*, ST137)	84.5	12	1	*fbaA*
Melibionic acid	PM2A	F07	ε, ι *(1/3)* and γ	80.8	9	1	*atoB*
D-malic acid	PM1	G11	ε and γ	83.7	12	1	*atoB*

§: Identified using the Omnilog phenotypic microarray system.

‡: Strains analysed on PM1 MicroPlates™: β (n=1), ε (n=6), ι (n=3) and γ (n=18).

Strains analysed on PM2A MicroPlates™: β (n=1), ε (n=4), ι (n=3) and γ (n=12).

†: Identified using BlastKOALA functional analysis. 100 amino acids per substrate were analysed.

1

6.3 Proteins involved in carbon metabolism

Proteins involved with the carbohydrate metabolism functional group were identified for each carbon substrate trait and within this group, a subgroup of proteins involved with carbon metabolism (Table 6.1). A similar number of proteins involved in carbohydrate metabolism were identified for each substrate (12-14 proteins), with fewer associated with melibionic acid utilisation (9 proteins). Many of the identified proteins were common for each substrate and involved processes such as glycolysis and gluconeogenesis, fructose and mannose metabolism, enzymes involved in the phosphotransferase system and galactose metabolism.

Except for D-tagatose utilisation, the genes and associated proteins involved with carbon metabolism (Table 6.1) were either *fbaA* (fructose biphosphate aldolase) or *atoB* (Acetyl-CoA C-acetyltransferase). The four genes associated with D-tagatose were *fbaA*, *lpd* (dihydrolipoamide dehydrogenase), *kdgK* (2-dehydro-3-deoxygluconokinase) and *acnA* (aconitate hydratase). The gene *fbaA* was present in all strains (Chapter 5). The *atoB* gene is present in the *eae* subtype ι (n=3) and β (n=1) strains, and absent in the *eae* subtype ε (n=6) and γ (n=41) strains. Thus the presence of the respective genes in strains carrying different *eae* subtypes is generally related with the contrasting carbon substrate utilisation of the different *E. coli* O145 strains (Table 6.1).

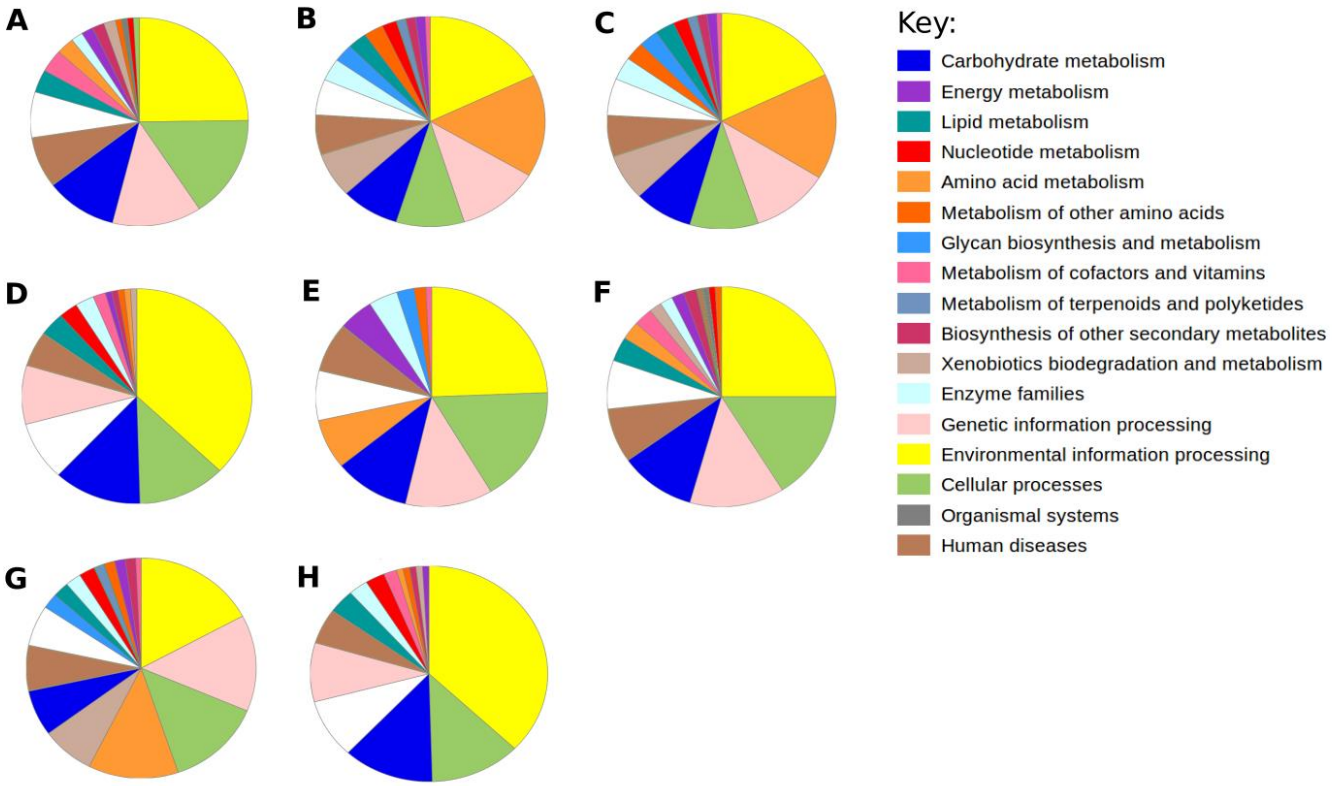

Figure 6.1: Functional analysis of proteins associated with the utilisation of specific carbon substrates

Association of genes in the pan genome and the utilisation of specific carbon substrates identified using the Omnilog phenotypic microarray system. The amino acid sequences for each gene identified were obtained and the proteins annotated into functional categories using BlastKOALA. Functional groups are indicated by the colour key, white indicates proteins which were unclassified. Functional groups of proteins associated with the utilisation of: **A.** β-hydroxy-butyric acid. **B.** D-galactonic acid-γ-lactone. **C.** D-malic acid. **D.** D-serine. **E.** D-tagatose. **F.** D-lactic acid methyl ester. **G.** Melibionic acid. **H.** Sucrose.

6.4 Discussion

6.4.1 Proteins and genes identified by genomics which are potentially associated with carbon substrate utilisation

Using Scoary, a number of genes involved in carbon metabolism were identified to be associated with carbon substrate utilisation (Table 6.1) including three (*fbaA*, *cysE* and *rpiB*) which were found to be associated with specific *eae* subtypes (Chapter 5).

The gene *fbaA* encoding fructose biphosphate aldolase class II was associated with many carbon substrate utilisation traits, and all *eae* subtypes. Fructose biphosphate aldolase is a ubiquitous enzyme which has a pivotal role in both glycolysis and gluconeogenesis [175], but is also hypothesised to have a diverse range of additional metabolic roles [176]. It has been suggested that fructose biphosphate aldolase class II enzymes are essential for the viability of *E. coli*, *Pseudomonas aeruginosa* and *Bacillus subtilis*, in which the disruption of the genes encoding this enzyme have been studied [176]. Fructose biphosphate aldolase has also been demonstrated to have an impact on the pathogenicity in other bacterial species such as *Mycobacterium tuberculosis* and *Toxoplasma gondii* [176]. Due to the essential function of this enzyme, *fbaA* is directly associated with carbon metabolism for many serogroup O145 strains.

The gene *cysE* encodes a serine acetyltransferase and was only present in *eae* subtype γ strains. The *cysE* gene is involved in the biosynthesis of cysteine from L-serine, and more specifically the conversion of serine to *O*-acetyl-L-serine [177,178]. Serine acetyltransferase can also use L-threonine as a substrate to convert to *O*-acetyl-L-threonine but at lower efficiency than the conversion with L-serine [177]. However, Omnilog phenotypic microarray data indicated that L-serine (PM1, G03) was utilised by all strains but that L-threonine (PM1, G04) was inconsistently utilised by serogroup O145 strains (*eae* subtypes ε (n=2), β (n=1), γ (n=3) and ι (n=3)), indicating other genes may be involved in the metabolism of these substrates. Sturgill et al. [177] identified another role of serine acetyltransferase; they identified an *E. coli cysE* mutant which formed biofilms more rapidly than wild-type strains, suggesting *cysE* may play a role in the regulation of biofilm development [177]. This suggests a selective advantage for strains carrying the *cysE* gene in some environments.

The gene *rpiB* encodes ribose-5-phosphate isomerase B, which was present only in non-γ *eae* subtype strains. This enzyme is involved in the interconversion of ribulose-5-phosphate and ribose-5-phosphate [179] and has been suggested to have additional unknown functions [179]. *E. coli* may also carry the *rpiA* gene encoding ribose-5-phosphate isomerase A, which catalyses the same reaction as ribose-5-phosphate isomerase B [180], but *rpiA* is constitutively expressed and is suggested to account for the majority of ribose phosphate isomerase activity

when bacterial strains are grown in nutrient broth [179]. The Omnilog phenotypic microarray data indicated D-ribose (PM1, C4) was utilised by all serogroup O145 strains tested, which may be due to multiple enzymes involved in the utilisation of this substrate.

The gene *atoB* encodes an acetyl-CoA C-acetyltransferase and is involved in the degradation of acetoacetate to acetyl coenzyme A [181]. Acetoacetic acid (PM1, G07) was only utilised by *eae* subtypes β (n=1) and γ (n=1). The genes and proteins associated with D-tagatose utilisation were *fbaA* (as described above), *lpd* (dihydrolipoamide dehydrogenase), *kdgK* (2-dehydro-3-deoxygluconokinase) and *acnA* (aconitate hydratase). Dihydrolipoamide dehydrogenase is part of the pyruvate dehydrogenase complex and is involved in the re-oxidation of dihydrolipoic acid by a thiol group [182]. The enzyme 2-dehydro-3-deoxygluconokinase is involved with the phosphorylation of 2-keto-3-deoxy-D-gluconic acid [183,184] and aconitate hydratase is involved in both the citric acid and glyoxylate cycle where it isomerises citrate to isocitrate [185].

6.4.2 Difficulties identifying genes involved in carbon substrate utilisation

Identifying phenotype and genotype correlations is essential for (i) providing an understanding of the observed phenotypes and identifying potential genetic causes for such differences, and (ii) linking genetic data with observed phenotypes as due to complex gene regulation pathways and other factors, the presence of a specific gene does not always indicate an observed phenotype.

Of the genes identified as being involved in carbon metabolism, none were specifically linked to the utilisation of certain carbon substrates (Table 6.1). This may be due to (i) the methodology used to identify phenotype and genotype correlations not being as comprehensive as required, (ii) a number of genes being involved with complex pathways and/or regulatory systems, rather than a single gene responsible for the utilisation of a specific carbon substrate, or (iii) the association of individual genes with the specific carbon substrate utilisation in the pan genome may be confounded by their association with certain *eae* subtypes. For example, *atoB* was only found in *eae* subtype ι and β strains and therefore, during analysis to identify genes associated with substrates utilised by these *E.*

coli O145 strains, only the gene *atoB* is identified. For each substrate, 100 amino acid sequences were analysed however, a larger sample size may be required to identify genes specific for the utilisation of a certain substrate. It is also important to note the identification of the proteins and genes from the pan genome was based on short-read sequence data. Future work from this project could involve using alternative approaches to identifying the phenotypic and genotypic correlations between the Omnilog and genomic data (Chapter 7). Potentially this additional analysis would identify and confirm candidate targets for further investigation and their use in differential culture media for serogroup O145. In addition, proteins which were associated with carbon substrate utilisation were also linked to other metabolic pathways such as mannose and galactose metabolism. These pathways may warrant further investigation for the identification of potential targets for the development of a differential media for *E. coli* serogroup O145.

6.5 Summary

Although a diverse range of functional groups for proteins associated with the utilisation of specific carbon substrates were identified, the categories ‘environmental information processing’ and ‘cellular processes’ predominate. The categories ‘genetic information processing’ and ‘amino acid metabolism’ were common in many but not all of the Scoary substrate analyses. Proteins involved in carbohydrate metabolism were identified and found to be associated with the utilisation of all carbon substrates investigated. Of the carbon metabolism genes identified, *fbaA* (fructose biphosphate aldolase) and *atoB* (Acetyl-CoA C-acetyltransferase) were the most common. Although genes and proteins associated with carbon metabolism were identified, they were not directly linked to the specific carbon substrate being utilised. This may be due to a variety of factors such as the methodology used, the complexity of carbon metabolism or the association between carbon substrate utilisation and *eae* subtype confounding the results. An additional and more extensive prospective approach to identifying phenotype and genotype correlations is outlined in Chapter 7.

7. General discussion

This thesis comprises two main research aims: firstly, to identify phenotypic characteristics of *E. coli* serogroup O145 which would enable the development of a differential media, and secondly to understand the genomic epidemiology of serogroup O145 strains. STEC O145 are a public health concern, causing severe disease ranging from diarrhoea to haemolytic uraemic syndrome [3], and are also a major issue for New Zealand's meat export industry as they have been declared adulterants of ground beef in the USA, along with the other Top 7 serogroups (O26, O45, O103, O111, O121 and O157) [12].

In New Zealand serogroup O145 has been isolated from ruminants including dairy cattle [186], calves [187], and it has been detected in deer using RT-PCR [145]. A cross-sectional study on New Zealand dairy farms identified STEC O145 as the most prevalent serogroup, at the dairy farm level, compared to the other Top 7 serogroups [1]. Overseas serogroup O145 has been isolated from a range of ruminant sources including cattle [188], deer [189] and sheep [147]. Serogroup O145 has also caused outbreaks of human disease linked to consumption of contaminated romaine lettuce [9] and ice-cream [10]. The widespread carriage and excretion of serogroup O145 in a range of ruminant sources is a potential public health concern and an issue to New Zealand's meat export industry.

Both culture independent (e.g. PCR) and culture dependent (e.g. selective media, IMS) methods are available for the identification and isolation of serogroup O145. Culture independent methods provide rapid indication of the presence of a target organism. However, the use of molecular methods does not provide a bacterial isolate for further analysis: PCR detection of virulence genes in mixed populations may not be associated with clinically significant serogroups and without the use of specific treatments such as with PMA, molecular methods cannot distinguish between 'free' DNA, or DNA from non-viable bacteria, with DNA associated with viable cells. Additionally, the emergence of novel STEC pathogens, such as the EAEC/STEC hybrid serogroup O104, may not be detected with the use of standard molecular techniques targeting a subset of STEC serogroups such as the Top 7 [16]. In contrast, culture dependent methods are often slow and

laborious, but the recovery of a bacterial isolate is a major advantage [17], and is required for many additional analyses such as WGS.

Differential media for the isolation of the Top 7 STEC serogroups, including O145, are currently available. However, separate agar media preparations have been associated with highly variable isolate recovery levels [17,59,147]. Studies by Posse et al. [156] and Kerangart et al. [67] have proposed carbon metabolism characteristics (serogroup O145 utilisation of D-arabinose and β-hydroxy-butyric acid; no utilisation of D-raffinose, dulcitol and D-galactonic acid γ-lactone) to distinguish serogroup O145, but the sensitivity and specificity of isolation is variable [65,69].

For this study, WGS was used as a tool to elucidate the genomic epidemiology of serogroup O145. A completed closed genome for each O145 strain was not obtained, nevertheless an assembly of >100 contigs provides unparalleled genetic resolution. In comparative genomics, consistent use of an analysis pipeline within a respective study is essential as the pipeline used may impact the outputs, and subsequently the inferences drawn. Morrison et al. [88] assembled and annotated *Vibrio vulnificus* genomes using a range of currently available software analyses methods and demonstrated significant differences in the number of gene features between the methods. Therefore, comparative genomic studies should make use of the same analysis pipeline to reduce any inconsistencies caused by different analysis parameters. Hence, only serogroup O145 genomes in which the raw read data was accessible were used to compare with WGS data generated in this study, as the same analysis pipeline could be applied to the two datasets (Chapter 5).

For comparative analysis, a closed and completed STEC O145 reference genome was available (section 2.10.3) for improved assembly, and the core SNPs were determined to provide a phylogenetic genome wide assessment of the similarities and differences between the strains. Virulence factors were identified (VirulenceFinder v1.5, [23]) to provide a more thorough understanding of the putative pathogenicity determinants associated with the strains. MLST, according to seven house-keeping genes, was used to further subtype the isolates [24] allowing for comparison of sequence types among *E. coli* strains,

and rMLST analysis was undertaken to provides additional phylogenetic analysis independently of HGT.

Pan genome analysis provides an additional approach to investigating the genetic diversity of a bacterial population by defining the core and accessory genes (Chapter 5). The core genes are often involved in essential functions and phenotypes and the accessory genome may contain genes which confer selective advantages such as virulence factors, antibiotic resistance, or genes involved in the adaption to a new ecological niche [190]. An open pan genome is indicative of a highly variable accessory genome and the occurrence of events increasing genetic diversity such as the acquisition of new genetic material. In this study, pan genome analysis enabled the identification of carbon metabolism genes associated with the utilisation of specific carbon substrates (Chapter 4).

The results of each piece of experimental work have been discussed in the relevant chapters, and subsequently, the following sections discuss the main findings of this research and how it contributes to the current scientific knowledge.

7.1 Culture-based isolation of serogroup O145 (Chapter 3)

Isolation of target bacteria from faecal samples is often more complex compared to other matrices such as carcass swabs or dairy products, due to the high levels of naturally competitive bacteria in faeces [65]. Multiple *E. coli* serogroups have been identified in faecal samples [17,61], highlighting the diversity of *E. coli* present in ruminant faeces. In addition, culture-based isolation methods for non-O157 serogroups have been demonstrated to be less effective for mixed cultures [73] including when used against cattle faeces compared to other sample matrices [65]. For example, serogroup O145, in comparison to the other “super six” serogroups, had the lowest recovery rate from cattle faeces [72].

In this study, the serogroup O145 recovery rate from RAMS was low (Chapter 3), and may have been impacted by the use of re-enriched cultures, which due to the freezing and thawing of enrichments, may have reduced the number of viable serogroup O145 cells during resuscitation. Also, the selection criteria to identify STEC O145 positive enrichments relied on the NeoSeek results (Chapter 3) which has a LOD of approximately 10^3 CFU/mL [144]. However, the NeoSeek

method may detect free DNA rather than DNA associated with viable cells, resulting in an over-estimation of STEC O145 positive RAMS enrichments.

The phenotypic variation within serogroups, in particular colony colour, and some media being overly selective and inhibiting the growth of the target organism, are major issues with culture-based isolation on currently available media [63,64]. The use of two media in parallel, a highly selective media paired with a media that supports the growth of a wide range of STEC, may provide an enhanced isolation rate for the non-O157 serogroups [58,63,64]. Therefore, the use of multiple media in parallel for the isolation of serogroup O145 in this study may have increased the recovery rate.

The use of selective media coupled with serogroup-specific IMS increased the isolation rate in this study (Chapter 3). However, the effectiveness of IMS varies widely between studies. For example, using IMS-based methods, serogroup O145 isolates were obtained from 71% of the spiked faecal samples (n=14) by Conrad et al. [16] and Posse et al. [65] detected serogroup O145 with an isolation efficiency of 84.6% from artificially inoculated dairy and meat matrices. The discordance in IMS between studies may be partially attributed to sample variation coupled with the presence of highly competitive background flora [65], the concentration of the target bacteria in the sample [73], the type of matrix used or competition with other STEC serogroups reducing bead specificity [73]. Kraft et al. [72] evaluated the efficacy of three IMS bead brands (Dynabeads®, Abraxis and Romer) to detect the “super six” serogroups in a variety of sample matrices. Serogroup O145 recovery using the three bead brands was consistently low across all matrices which included food and faecal samples [72]. Serogroups O145 (16.7-66.7%) and O111 (0-94.4%) had the lowest recovery from faeces [72]. Improved recovery of serogroup O145 with the Romer beads was observed over all matrices compared to other bead preparations [72]. Abraxis beads were used in this study, but whether enhanced recovery may have been associated with the use of serogroup-specific O145 beads of other manufacturers is unknown.

Only 1 out of 32 serogroup O145 strains isolated in this study were identified as *stx*-positive (Chapter 3); 13 of the remaining 21 strains that underwent WGS were *stx*-positive (2 *stx1*-positive, 11 *stx2*-positive) (Chapter 5). Compared to *stx*-

negative O145 in ruminants, STEC O145 may be less abundant. However, the instability of the *stx*-encoding bacteriophage during sub-cultivation [191] may contribute to the low isolation rate of STEC O145 in this study. Noll et al. [17] identified 6 out of 19 serogroup O145 isolates from cattle faeces as *stx*-positive (4 *stx1*-positive, 2 *stx2*-positive). Using the same set of calf-faecal enrichments as this study, Browne et al. [192] identified six serogroup O145 isolates which were all *stx*-negative, giving an overall recovery rate of 13.3%. Similarly Dewsbury et al. [59] identified 18 serogroup O145 strains from cattle with few virulent isolates (4 *stx1*-positive, 2 *stx2*-positive, 6 *eae*-positive).

Unlike *stx*1 and *stx*2, all strains that underwent WGS in this study were *eae* positive (n=53), and a high proportion (46 out of 53) were *ehxA*-positive (Chapters 3 and 5). In comparison to the other "super six" serogroups fewer serogroup O145 isolates were negative for the common STEC-associated virulence factors *stx*1, *stx*2, *eae* or *ehx*A; with 17 of 19 O145 both *eae* and *ehxA*-positive [17]. Noll et al. [17] identified one avirulent (lacking *stx*1, *stx*2, *eae* and *ehx*A) O145 strain from cattle faeces, indicating that serogroup O145 strains may on rare occasions be *eae*-negative. Another study Hofer et al. [189] examined serogroup O145 strains from chamois, ibex and deer in Switzerland; all eight O145 detected were *eae* negative and two *ehxA* positive [189]. However all eight O145 strains were *stx*-positive (2 *stx1c* and *stx2b*-positive, 6 strains *stx2b*-positive) [189] suggesting that wild ruminants carry different STEC populations, which are usually *eae*-positive, compared to cattle and other well-studied ruminants.

Some STEC serotypes such as O157:H7 (γ), O26:H11 (β), O103:H2 (ε), O111:H8 (θ) and O145:H28 (γ) are characterised by a single *eae* subtype [40], however the association of multiple *eae* subtypes with specific serogroups, such as O103, has been observed [41]. In this work the *eae* γ subtype was the most common (83%) including all 32 strains isolated from RAMS enrichments. However NeoSeek identification of STEC O145 includes the detection of *eae* subtype γ SNPs [193]. Therefore RAMS enrichments containing STEC O145 of other non-γ *eae* subtypes will not have been identified as candidate enrichments for further analysis in this work. Other studies have identified STEC O145 containing the *eae* subtype γ [194,195,196], and the relative abundance of *eae* subtype γ STEC

O145 whole genome sequences suggests that *eae* subtypes α, β, ε, ι are comparatively rare.

7.2 Carbon utilisation (Chapter 4)

Previous bacterial carbon utilisation studies have often involved the use of multiple broths or media each separately examining the fermentation of carbohydrates, with the growth according to a colour change of an indicator dye. A similar approach was utilised by Posse et al. [197] to examine carbohydrate utilisation of some of the Top 7 STEC serogroups. These methods are often laborious and time-consuming, cumbersome to examine multiple substrates and the interpretation of the growth based on a colour change may introduce bias. Therefore, the use of the Omnilog system is extremely valuable in examining carbon utilisation as this system is rapid and allows multiple carbon substrates to be examined in parallel resulting in less variation between experiments. The chemical reduction of the dye, which is proportional to the level of cell respiration, provided an absolute growth value allowing for construction of growth curves and reduced any bias introduced by interpretation of growth. The outputs of the Omnilog system are readily utilisable by a range of analysis software such as the R statistics package 'opm' [98]. However, a limitation of this method is that all substrates in the MicroPlates™ may not be readily available or suitable for use in a new modified media.

The Omnilog analysis revealed highly variable carbon substrate utilisation among serogroup O145 strains (Chapter 4) none of which were able to distinguish between *stx*-positive and *stx*-negative strains. In comparison to other STEC serotypes, other studies demonstrated serogroup O145 isolates (n=3) showed little variation in the number of carbon substrates utilised and identified β-hydroxy-butyric acid as a candidate metabolite for O145 differentiation [67]. The three strains were all O145:H28, likely to be *eae* subtype γ, and displayed similarities in carbon utilisation with O157:H7 strains, which may account for the limited variation observed in comparison to the heterogeneity seen in this study [67]. In contrast, carbon utilisation of a large number of *E. coli* and *Shigella* strains in another study was shown to be highly variable [198]. Kerangart et al. [67] highlighted further STEC strains would need to be tested to confirm the carbon

utilisation profiles observed. According to the results of the present study, it would be beneficial to study additional STEC strains of a range of *eae* subtypes and ST, as studying these strains may reveal carbon utilisation characteristics specific to each serogroup which are essential in developing discriminatory characteristics for a differential media. There was a remarkable association between *eae* subtype and ST and the clustering observed according to carbon utilisation (Chapters 4 and 5). This finding supported the hypothesis of the independent acquisition of *eae* by distinct *E. coli* O145 phylogenetic lineages.

On some occasions, the reproducibility of carbon substrate utilisation by defined serogroup O145 strains was examined using PM1 and PM2A MicroPlates™ (Chapter 4). Unexpectedly there was some variation in the carbon substrate utilisation of replicate strains examined on separate experimental days. In contrast the carbon substrate utilisation of strains examined in duplicate on the same day, were generally very similar (Chapter 4). This suggests that an experimental factor, influences the utilisation of carbon substrates, such as the functionality of the Omnilog machine, freeze thawing of culture stocks or a change of gene expression, which varies between days of inoculation (Chapter 4). It has been proposed that changes in bacterial strains, such as plasmid loss, metabolic changes due to stress, and antibiotic resistance can be influenced by to laboratory conditions and storage [199]. However additional duplicates would need to be compared to test this hypothesis. The variation of carbon substrate utilisation between O145 replicates and any possible changes of gene expression suggests it may have been beneficial to use the same bacterial clone for the carbon utilisation analysis and WGS.

This is the first study to examine the growth of genetically diverse serogroup O145 strains against a large number of carbon substrates (Chapter 4). A number of previous studies examined the carbon utilisation of the Top 7 *E. coli* serogroups [67,155], demonstrating the variability in carbon utilisation and how this variability has hindered the development of a differential media for many non-O157 serogroups. The carbon substrate utilisation diversity observed in this study suggests the development media that permits the selective growth of all O145 strains would be difficult. The high prevalence of serogroup O145 *eae* subtype γ strains in New Zealand cattle (Chapters 3 and 5) indicates this subset is an

important public health issue and a concern for New Zealand's meat export industry; therefore, development of a media solely for this subtype may be appropriate.

7.3 Comparative genomics of serogroup O145 (Chapter 5)

WGS analysis provides unparalleled genetic resolution which allows for high discriminative comparison between strains. Several WGS analysis methods including core SNP, virulence genes profiling, and rMLST differentiated strains according to *eae* subtype and ST (Chapter 5). However, due to a smaller number of genes being analysed, the rMLST resolution is less defined.

Core genome SNP evaluation of the O145 *eae* subtype γ WGS data was unable to distinguish between *stx*-positive and *stx*-negative strains indicating the close phylogenetic relationship between the different pathotypes. The *eae* subtype γ and ε strains carried between 10-18 and 14-16 virulence factors, respectively, compared to the *eae* subtype β (5) and ι (5-6) strains. *eae* subtype γ strains are not commonly associated with human disease in New Zealand, but have been commonly isolated from clinical cases overseas. Potentially, the additional virulence factors associated with *eae* subtype γ strains may indicate enhanced pathogenicity during human clinical infection. The *eae* subtype ε strains have been associated with human disease in New Zealand (n=2), however this *eae* subtype appears to be rarely associated with STEC O145. *E. coli* O145 have a number of similarities with the highly pathogenic serogroup O157 (both serogroups are typically *eae* subtype γ and can be *stx2*-positive), and O157 strains (n=11) were shown to carry a large number of virulence factors (n=20) using VirulenceFinder v1.5 [23] [200]. Although several other virulence factors are associated with STEC pathogenicity, the range found in *eae* subtype γ O145 strains may point to its enhanced virulence compared to other *eae* subtypes. Genome composition analysis suggested strains of the same *eae* subtype had a similar genome size (Chapter 5). The *eae* subtype ι strains had the smallest genome size and fewer virulence factors, suggesting that there may also be an association between genome size and the number of virulence factors acquired. Identification of mobile genetic elements may provide a more comprehensive link of this association.

Different *eae* subtypes have been suggested to influence bacterial tissue tropisms [201] and contrasting gastro-intestinal colonisation sites through the formation of A/E lesions. Bacteria with *eae* subtype α (O127:H6) preferentially colonise the mucosa of the small intestine [201], *eae* subtype γ (O157:H7) the follicle associated epithelium (FAE) of Peyer's patches [202], and *eae* subtype ε (O103) the FAE of the duodenum [203]. Whether these same *eae* subtype specific tissue tropisms are associated with other *E. coli* serogroups possessing the same *eae* subtype has not been established.

Evolutionary analysis predicted that serogroup O145 *eae* subtype γ strains shared a most recent common ancestor in approximately 1828 (95% HPD interval 1748-1897) (Chapter 5), but that after this time these clones evolved, potentially adapting to the New Zealand environment. This date is consistent with the TMRCA for New Zealand and global serogroup O26 (ST21) strains which was in 1865 [171]. Further evolutionary analysis of a larger and more diverse panel of *eae* subtype γ strains including those from overseas with precise isolation dates would be required to determine the likely time period for the introduction of O145 into New Zealand.

The association between genes in the pan genome and a number of traits of interest (*eae* subtype, isolation source and whether they were toxigenic) enabled the identification of three carbon metabolism genes which were associated with specific *eae* subtypes and ST. These associations highlight the use of *eae* and ST as good characteristics for the contrasting *E. coli* phylogenetic lineages and suggests carbon metabolism characteristics can be used to distinguish these different subtypes.

7.4 Phenotype and genotype correlations (Chapter 6)

The phenotype and genotype analysis did not provide any supplementary evidence to identify a potential carbon substrate for the use in a differential media. Carbon metabolism genes were identified from pan genome analysis, however, none could be linked to the carbon substrate utilised (Chapter 6). This may be due to the complexity of carbon metabolism and the multiple pathways and regulatory steps involved; therefore the association of a single gene with carbon

utilisation may be unlikely. However, these preliminary analyses will form a basis for subsequent phenotype and genotype correlations, to be investigated using an alternative approach (section 7.6.3). Previous work to investigate the association between carbon substrate utilisation and the presence or absence of specific metabolic pathways indicated that the utilisation of a specific substrate was not always correlated with the presence of the corresponding metabolic pathway [198]. Potentially this may be due to the involvement of currently unknown enzymes or metabolic pathways in carbon metabolism. Similarly, strains which had a metabolic pathway involved in the utilisation of a specific carbon substrate did not always grow on the corresponding substrate. Unidentified gene regulators or mutations in genes involved in the metabolic pathway are likely to affect whether metabolic activity takes place [198].

Pan genome analysis revealed the genetic heterogeneity of serogroup O145 strains and has indicated an open pan genome of more than 10,000 genes (Chapter 5), highlighting the diversity of *E. coli* strains. Further analysis of the *eae* subtype γ strains pan genome would be useful to provide a better understanding of these strains. Restricting the pan genome analysis to *eae* subtype γ strains, alone would increase the number of core genes and decrease the number of accessory genes. Additional analysis of the pan genome could also be used to investigate factors of interest such as the resistome, insights into niche adaptation and bacterial evolution.

7.5 Value of this research

The variation in carbon utilisation among serogroup O145 strains has been demonstrated and substrates have been suggested for subsequent analysis for the use in a differential media. The WGS analysis has demonstrated the value of the *eae* gene and ST as good markers to distinguish separate phylogenetic lineages and the WGS and pan genome represent a valuable future resource for subsequent research of *E. coli* serogroup O145. In summary, this study has provided significant progress in the identification of components that would likely influence the development of a differential media for serogroup O145. Such a medium would prove a valuable tool for maintaining and monitoring public health

and providing food quality and safety assurances that the NZ meat for export is free of this pathogen.

7.6 Areas for further research

7.6.1 Development of a differential media for serogroup O145

Candidate substrates (Table 4.3) require further testing to determine their effectiveness in a differential media for serogroup O145. Testing could include supplementing MacConkey agar with the candidate substrate or formulating a minimal media. During the development of the media a number of factors need to be addressed, such as the suitability of the carbon source in a culture media (e.g. storage conditions, pH) and cost effectiveness. The culture media also needs to be tested to: (i) confirm its effectiveness in growing a diverse range of serogroup O145 strains, (ii) determine the growth of other *E. coli* serotypes and *Enterobacteriaceae* on the media, and (iii) testing the media using artificially inoculated matrices (e.g. faecal and food samples) spiked with serogroup O145 and other *E. coli* serogroups of known concentrations to calculate the sensitivity and specificity of the media at detecting serogroup O145 from mixed cultures. The isolation rate for serogroup O145 using such a media should be compared with isolation success rates from this study and other published recovery rates [65,148] to determine whether the development of such a media offers any advantages in the isolation of serogroup O145.

7.6.2 Subsequent WGS analysis

Additional WGS analysis would provide further clues on the genomic epidemiology of serogroup O145 strains. Such analyses could include investigation of mobile genetic elements such as insertion sequences and integrated prophage, or analysis of the pan genome of *eae* subtype γ strains. The genomic dataset (n=100: n=53 strains from this study and n=47 publicly available genomes) could be compared with other *E. coli* Top 7 serogroups to identify any characteristics unique to serogroup O145. Comparative genomic studies for *E. coli* have been conducted [47,162], however, no studies have focused on a large number of serogroup O145 strains. Additional comparison of serogroup O145 strains from a range of isolation sources and geographic origins would be useful

for (i) evolutionary analysis to determine the introduction date of *E. coli* serogroup O145 into New Zealand and, (ii) to further identify any associations between these factors and genomic characteristics such as number of core SNPs.

7.6.3 An alternative approach for identifying phenotype and genotype correlations

A future strategy for identifying phenotype and genotype associations (Chapter 6) was identified using a method that was developed to link Omnilog phenotypic microarray data with whole genome sequence data for *Campylobacter jejuni* [204]. Briefly this method uses self-organising maps and hierarchical clustering to identify the smallest number of representative curve shapes from the Omnilog respiration curves (Chapter 4) to characterise all phenotypic responses observed. The random forests algorithm then compares these representative curve clusters with the genotype data to identify the Clusters of Orthologous Groups (COGs) (Chapter 5) which are associated with a specific respiration curve cluster. This approach would provide a thorough investigation of the phenotype and genotype correlations by providing an indication of which COGs are associated with specific carbon substrate utilisation that can subsequently analysed further.

7.7 Concluding statement

This research has demonstrated that a more sensitive and specific differential media for serogroup O145 is an important requirement if its true epidemiological role in human disease and food safety is to be determined. Despite the contrasting carbon utilisation profiles among serogroup O145 strains, carbon substrates have been identified as candidates for the development of a differential media and warrant subsequent testing. A differential media developed from the carbon substrates identified in this study would likely be coupled with current molecular and culture-based methods, such as serogroup-specific PCR and IMS, to aid in the isolation of this pathogen. The WGS has revealed the genetic heterogeneity of serogroup O145, and remarkably, has demonstrated in a number of analyses the use of the *eae* gene and ST as good markers for *E. coli* phylogenetic lineages.

8. Bibliography

1. Browne AS, Midwinter AM, Withers HL, Cookson AL, Biggs PJ, Marshall JC, Benschop J, Hathaway S, French NP (2015) Prevalence, risk factors, and spatial distribution of Shiga toxin-producing *E. coli* (STEC) on dairy farms in New Zealand. VTEC Conference Boston Massachusetts, USA Retrieved from:

2. Griffin PM, Ostroff SM, Tauxe RV, Greene KD, Wells JG, Lewis JH, Blake PA (1988) Illnesses associated with *Escherichia coli* O157-H7 infections: A broad clinical spectrum. *Annals of Internal Medicine* 109: 705-712.

3. Beutin L, Zimmermann S, Gleier K (1998) Human infections with Shiga toxin-producing *Escherichia coli* other than serogroup O157 in Germany. *Emerging Infectious Diseases* 4: 635-639.

4. Thorpe CM (2004) Shiga toxin-producing *Escherichia coli* infection. *Clinical Infectious Diseases* 38: 1298-1303.

5. The Institute of Environmental Science and Research. (2017) Notifiable diseases in New Zealand: Annual report 2016. Porirua, NZ. Retrieved from: https://surv.esr.cri.nz/PDF_surveillance/AnnualRpt/AnnualSurv/2016/2016AnnualNDReportFinal.pdf?m=1507246186.

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. *Journal of Infectious Diseases* 192: 1422-1429.

7. Bettelheim KA (2000) Role of non-O157 VTEC. *Journal of Applied Microbiology* 88: 38-50.

8. Jaros P, Cookson AL, Campbell DM, Besser TE, Shringi S, Mackereth GF, Lim E, Lopez L, Dufour M, Marshall JC, Baker MG, Hathaway S, Prattley DJ, French NP (2013) A prospective case-control and molecular epidemiological study of human cases of Shiga toxin-producing *Escherichia coli* in New Zealand. *BMC Infectious Diseases* 13: 450.

9. Taylor EV, Nguyen TA, Machesky KD, Koch E, Sotir MJ, Bohm SR, Folster JP, Bokanyi R, Kupper A, Bidol SA, Emanuel A, Arends KD, Johnson SA, Dunn J, Stroika S, Patel MK, Williams I (2013) Multistate outbreak of *Escherichia coli* O145 infections associated with romaine lettuce consumption, 2010. *Journal of Food Protection* 76: 939-944.

10. De Schrijver K, Buvens G, Possé B, Van den Branden D, Oosterlynck O, De Zutter L, Eilers K, Piérard D, Dierick K, Van Damme-Lombaerts R, Lauwers C, Jacobs R (2008) Outbreak of verocytotoxin-producing *E. coli* O145 and O26 infections associated with the consumption of ice cream produced at a farm, Belgium, 2007. *Eurosurveillance* 13: 9-10.

11. 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. *Journal of the American Medical Association* 272: 1349-1353.

12. U.S. Department of Agriculture FSIS (2012) Risk profile for pathogenic non-O157 shiga toxin-producing *Escherichia coli*. Retrieved from: http://www.fsis.usda.gov/shared/PDF/Non_O157_STEC_Risk_Profile_May2012.pdf.

13. U.S. Department of Agriculture FSIS (2015) Detection, isolation and identification of *Escherichia coli* O157:H7 from meat products and carcass and environmental sponges. Retrieved from: http://www.fsis.usda.gov/wps/wcm/connect/51507fdb-dded-47f7-862d-ad80c3ee1738/MLG-5.pdf?MOD=AJPERES.

14. U.S. Department of Agriculture FSIS (2011) Shiga toxin-producing *Escherichia coli* in certain raw beef products. Retrieved from: https://www.gpo.gov/fdsys/pkg/FR-2011-09-20/html/2011-24043.htm.

15. Gill A, Gill CO (2010) Non-O157 verotoxigenic *Escherichia coli* and beef: A Canadian perspective. *Canadian Journal of Veterinary Research* 74: 161-169.

16. Conrad CC, Stanford K, McAllister TA, Thomas J, Reuter T (2014) Further development of sample preparation and detection methods for O157 and the top 6 non-O157 STEC serogroups in cattle feces. *Journal of Microbiological Methods* 105: 22-30.

17. Noll LW, Shridhar PB, Dewsbury DM, Shi XR, Cernicchiaro N, Renter DG, Nagaraja TG (2015) A comparison of culture and PCR based methods to detect six major non-O157 serogroups of Shiga toxin-producing *Escherichia coli* in cattle feces. *PloS One* 10: 12.

18. Browne AS, Midwinter AC, Withers H, Cookson AL, Biggs PJ, Marshall JC, Benschop J, Hathaway S, Haack N, Akhter R, French NP (2018) Shiga toxin-producing *Escherichia coli* (STEC) on New Zealand dairy farms: application of a culture-independent assay to estimate national point prevalence of multiple serogroups. In revision for Applied Environmental Microbiology Retrieved from:

19. The Institute of Environmental Science and Research Ltd (n.d.) New Zealand public health surveillance reports 1993-2016. Porirua, New Zealand. Retrieved from: https://surv.esr.cri.nz/surveillance/NZPHSR.php.

20. Kaper JB, Nataro JP, Mobley HLT (2004) Pathogenic *Escherichia coli*. *Nature Reviews Microbiology* 2: 123-140.

21. Orskov I, Orskov F, Jann B, Jann K (1977) Serology, chemistry and genetics of O and K antigens of *Escherichia coli*. *Bacteriological Reviews* 41: 667-710.

22. Stenutz R, Weintraub A, Widmalm G (2006) The structures of *Escherichia coli* O-polysaccharide antigens. *FEMS Microbiology Reviews* 30: 382-403.

23. Joensen KG, Tetzschner AMM, Iguchi A, Aarestrup FM, Scheutz F (2015) Rapid and easy *in silico* serotyping of *Escherichia coli* isolates by use of whole genome sequencing data. *Journal of Clinical Microbiology* 53: 2410-2426.

24. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Ponten T, Ussery DW, Aarestrup FM, Lund O (2012) Multilocus sequence typing of total genome sequenced bacteria. *Journal of Clinical Microbiology* 50: 1355-1361.

25. Robins-Browne RM, Holt KE, Ingle DJ, Hocking DM, Yang J, Tauschek M (2016) Are *Escherichia coli* pathotypes still relevant in the era of whole-genome sequencing? *Frontiers in Cellular and Infection Microbiology* 6: 141.

26. Trabulsi LR, Keller R, Gomes TAT (2002) Typical and atypical enteropathogenic *Escherichia coli*. *Emerging Infectious Diseases* 8: 508-513.

27. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji YM, Zhang WL, McLaughlin SF, Henkhaus JK, Leopold B, Bielaszewska M, Prager R, Brzoska PM, Moore RL, Guenther S, Rothberg JM, Karch H (2011) Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PloS One* 6.

28. Melton-Celsa AR (2014) Shiga toxin (Stx) classification, structure, and function. *Microbiology Spectrum* 2.

29. Paton JC, Paton AW (1998) Pathogenesis and diagnosis of Shiga toxin-producing *Escherichia coli* infections. *Clinical Microbiology Reviews* 11: 450-479.

30. Montero DA, Velasco J, Del Canto F, Puente JL, Padola NL, Rasko DA, Farfan M, Salazar JC, Vidal R (2017) Locus of adhesion and autoaggregation (LAA), a pathogenicity island present in emerging Shiga toxin-producing *Escherichia coli* strains. *Scientific Reports* 7: 13.

31. 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.

32. Tozzoli R, Grande L, Michelacci V, Ranieri P, Maugliani A, Caprioli A, Morabito S (2014) Shiga toxin-converting phages and the emergence of new pathogenic *Escherichia coli*: A world in motion. *Frontiers in Cellular and Infection Microbiology* 4: 80.

33. Kruger A, Lucchesi PMA (2015) Shiga toxins and *stx* phages: Highly diverse entities. *Microbiology* 161: 451-462.

34. Bertin Y, Boukhors K, Pradel N, Livrelli V, Martin C (2001) Stx2 subtyping of Shiga toxin-producing *Escherichia coli* isolated from cattle in France: Detection of a new Stx2 subtype and correlation with additional virulence factors. *Journal of Clinical Microbiology* 39: 3060-3065.

35. 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. *Journal of Clinical Microbiology* 37: 497-503.

36. Jores J, Rumer L, Wieler LH (2004) Impact of the locus of enterocyte effacement pathogenicity island on the evolution of pathogenic *Escherichia coli*. *International Journal of Medical Microbiology* 294: 103-113.

37. Furniss RCD, Clements A (2018) Regulation of the locus of enterocyte effacement in attaching and effacing pathogens. *Journal of Bacteriology* 200: 12.

38. Lacher DW, Steinsland H, Whittam TS (2006) Allelic subtyping of the intimin locus (*eae*) of pathogenic *Escherichia coli* by fluorescent RFLP. *FEMS Microbiology Letters* 261: 80-87.

39. Cookson AL, Bennett J, Thomson-Carter F, Attwood GT (2007) Intimin subtyping of *Escherichia coli*: Concomitant carriage of multiple intimin subtypes from forage-fed cattle and sheep. *FEMS Microbiology Letters* 272: 163-171.

40. Bibbal D, Loukiadis E, Kerouredan M, de Garam CP, Ferre F, Cartier P, Gay E, Oswald E, Auvray F, Brugere H (2014) Intimin gene (*eae*) subtype-based real-time PCR strategy for specific detection of Shiga toxin-producing *Escherichia coli* serotypes O157:H7, O26:H11, O103:H2, O111:H8, and O145:H28 in cattle feces. *Applied and Environmental Microbiology* 80: 1177-1184.

41. Iguchi A, Iyoda S, Ohnishi M (2012) Molecular characterization reveals three distinct clonal groups among clinical Shiga toxin-producing *Escherichia coli* strains of serogroup O103. *Journal of Clinical Microbiology* 50: 2894-2900.

42. McDaniel TK, Kaper JB (1997) A cloned pathogenicity island from enteropathogenic *Escherichia coli* confers the attaching and effacing phenotype on *E-coli* K-12. *Molecular Microbiology* 23: 399-407.

43. McDaniel TK, Jarvis KG, Donnenberg MS, Kaper JB (1995) A genetic-locus of enterocyte effacement conserved among diverse enterobacterial pathogens. *PNAS* 92: 1664-1668.

44. Elliott SJ, Wainwright LA, McDaniel TK, Jarvis KG, Deng YK, Lai LC, McNamara BP, Donnenberg MS, Kaper JB (1998) The complete sequence of the locus of enterocyte effacement (LEE) from enteropathogenic *Escherichia coli* E2348/69. *Molecular Microbiology* 28: 1-4.

45. 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. *Applied and Environmental Microbiology* 79: 6301-6311.

46. Cooper KK, Mandrell RE, Louie JW, Korlach J, Clark TA, Parker CT, Huynh S, Chain PS, Ahmed S, Carter MQ (2014) Comparative genomics of enterohemorrhagic *Escherichia coli* O145:H28 demonstrates a common evolutionary lineage with *Escherichia coli* O157:H7. *BMC Genomics* 15: 17.

47. Ogura Y, Ooka T, Iguchi A, Toh H, Asadulghani M, Oshima K, Kodama T, Abe H, Nakayama K, Kurokawa K, Tobe T, Hattori M, Hayashi T (2009) Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* 106: 17939-17944.

48. 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*. *Journal of Clinical Microbiology* 27: 2559-2564.

49. Taneike I, Zhang HM, Wakisaka-Saito N, Yamamoto T (2002) Enterohemolysin operon of Shiga toxin-producing *Escherichia coli*: a virulence function of

inflammatory cytokine production from human monocytes. *FEBS Letters* 524: 219-224.

50. 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*. *Applied and Environmental Microbiology* 73: 6360-6369.

51. 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. *Applied and Environmental Microbiology* 82: 6367-6377.

52. Wright J, Fraser D, Baker M (1993) *Escherichia coli* O157:H7 infection: first New Zealand case report. *Communicable Diseases New Zealand* 93: 113-116.

53. International Organisation for Standardisation (2001) Microbiological methods, ISO 16654: Microbiology of Food and Animal Feeding Stuffs- Horizontal Method for the Detection of *Escherichia coli* O157. Geneva: International Organisation for Standardisation Retrieved from:

54. March SB, Ratnam S (1986) Sorbitol-MacConkey medium for detection of *Escherichia-coli* O157-H7 associated with hemorrhagic colitis. *Journal of Clinical Microbiology* 23: 869-872.

55. Zadik PM, Chapman PA, Siddons CA (1993) Use of tellurite for the selection of verocytoxigenic *Escherichia-coli* O157. *Journal of Medical Microbiology* 39: 155-158.

56. Thompson JS, Hodge DS, Borczyk AA (1990) Rapid biochemical test to identify verocytotoxin-positive strains of *Escherichia coli* serotype O157. *Journal of Clinical Microbiology* 28: 2165-2168.

57. Stromberg ZR, Baumann NW, Lewis GL, Sevart NJ, Cernicchiaro N, Renter DG, Marx DB, Phebus RK, Moxley RA (2015) Prevalence of enterohemorrhagic *Escherichia coli* O26, O45, O103, O111, O121, O145, and O157 on hides and preintervention carcass surfaces of feedlot cattle at harvest. *Foodborne Pathogens and Disease* 12: 631-638.

58. Wylie JL, Van Caeseele P, Sitter D, Guttek C, Giercke S, Gilmour MW (2013) Evaluation of a new chromogenic agar medium for detection of Shiga toxin-producing *Escherichia coli* (STEC) and relative prevalences of O157 and non-O157 STEC in Manitoba, Canada. *Journal of Clinical Microbiology* 51: 466-471.

59. Dewsbury DM, Renter DG, Shridhar PB, Noll LW, Shi XR, Nagaraja TG, Cernicchiaro N (2015) Summer and winter prevalence of Shiga toxin-producing *Escherichia col*i (STEC) O26, O45, O103, O111, O121, O145, and O157 in feces of feedlot cattle. *Foodborne Pathogens and Disease* 12: 726-732.

60. Irshad H, Cookson AL, Prattley DJ, Marshall J, French NP (2017) Epidemiology of *Escherichia coli* serogroups O26, O103, O111 and O145 in very young ('bobby') calves in the North Island, New Zealand. *Epidemiology and Infection* 145: 1606-1616.

61. Stanford K, Johnson RP, Alexander TW, McAllister TA, Reuter T (2016) Influence of season and feedlot location on prevalence and virulence factors of seven serogroups of *Escherichia coli* in feces of Western-Canadian slaughter cattle. *PloS One* 11: 18.

62. Johnson KE, Thorpe CM, Sears CL (2006) The emerging clinical importance of non-O157 Shiga toxin-producing *Escherichia coli*. *Clinical Infectious Diseases* 43: 1587-1595.

63. Gill A, Huszczynski G, Gauthier M, Blais B (2014) Evaluation of eight agar media for the isolation of Shiga toxin-producing *Escherichia coli*. *Journal of Microbiological Methods* 96: 6-11.

64. Verhaegen B, Van Damme I, Heyndrickx M, Botteldoorn N, Elhadidy M, Verstraete K, Dierick K, Denayer S, De Zutter L, De Reu K (2016) Evaluation of detection methods for non-O157 Shiga toxin-producing *Escherichia coli* from food. *International Journal of Food Microbiology* 219: 64-70.

65. Posse B, De Zutter L, Heyndrickx M, Herman L (2008) Quantitative isolation efficiency of O26, O103, O111, O145 and O157 STEC serotypes from artificially contaminated food and cattle faeces samples using a new isolation protocol. *Journal of Applied Microbiology* 105: 227-235.

66. BIOLOG (2008) Rainbow agar O157 technical information. Retrieved from: http://www.biolog.com/pdf/milit/00P%20037rC%20RBO157%2030gmTech%20FEB08.pdf.

67. Kerangart S, Cournoyer B, Loukiadis E (2017) C-source metabolic profilings of foodborne Shiga-toxin producing *E. coli* match serogroup differentiations and highlight functional adaptations. *International Journal of Food Microbiology* 266: 324-336.

68. Verhaegen B, De Reu K, Heyndrickx M, De Zutter L (2015) Comparison of six chromogenic agar media for the isolation of a broad variety of non-O157 shiga toxin-producing *Escherichia coli* (STEC) serogroups. *International Journal of Environmental Research and Public Health* 12: 6965-6978.

69. Stromberg ZR, Lewis GL, Moxley RA (2016) Comparison of agar media for detection and quantification of Shiga toxin-producing *Escherichia coli* in cattle feces. *Journal of Food Protection* 79: 939-949.

70. Brusa V, Pineyro PE, Galli L, Linares LH, Ortega EE, Padola NL, Leotta GA (2016) Isolation of Shiga toxin-producing *Escherichia coli* from ground beef using multiple combinations of enrichment broths and selective agars. *Foodborne Pathogens and Disease* 13: 163-170.

71. Xiong QR, Cui X, Saini JK, Liu DF, Shan S, Jin Y, Lai WH (2014) Development of an immunomagnetic separation method for efficient enrichment of *Escherichia coli* O157:H7. *Food Control* 37: 41-45.

72. Kraft AL, Lacher DW, Shelver WL, Sherwood JS, Bergholz TM (2017) Comparison of immunomagnetic separation beads for detection of six non-O157 Shiga toxin-producing *Escherichia coli* serogroups in different matrices. *Letters in Applied Microbiology* 65: 213-219.

73. Hallewell J, Alexander T, Reuter T, Stanford K (2017) Limitations of immunomagnetic separation for detection of the top seven serogroups of Shiga toxin-producing *Escherichia coli*. *Journal of Food Protection* 80: 598-603.

74. Fedio WM, Jinneman KC, Yoshitomi KJ, Zapata R, Weagant SD (2012) Efficacy of a post enrichment acid treatment for isolation of *Escherichia coli* O157:H7 from alfalfa sprouts. *Food Microbiology* 30: 83-90.

75. Bhagwat AA, Chan L, Han R, Tan J, Kothary M, Jean-Gilles J, Tall BD (2005) Characterization of enterohemorrhagic *Escherichia coli* strains based on acid resistance phenotypes. *Infection and Immunity* 73: 4993-5003.

76. Foster JW (2004) *Escherichia coli* acid resistance: tales of an amateur acidophile. *Nature Reviews Microbiology* 2: 898-907.

77. Tillman GE, Wasilenko JL, Simmons M, Lauze TA, Minicozzi J, Oakley BB, Narang N, Fratamico P, Cray WC (2012) Isolation of Shiga toxin-producing *Escherichia coli* serogroups O26, O45, O103, O111, O121, and O145 from ground beef using modified rainbow agar and post-immunomagnetic separation acid treatment. *Journal of Food Protection* 75: 1548-1554.

78. Almeida C, Sousa JM, Rocha R, Cerqueira L, Fanning S, Azevedo NF, Vieira MJ (2013) Detection of *Escherichia coli* O157 by peptide nucleic acid fluorescence in situ hybridization (PNA-FISH) and comparison to a standard culture method. *Applied and Environmental Microbiology* 79: 6293-6300.

79. Lin A, Sultan O, Lau HK, Wong E, Hartman G, Lauzon CR (2011) O serogroup specific real time PCR assays for the detection and identification of nine clinically relevant non-O157 STECs. *Food Microbiology* 28: 478-483.

80. Paddock Z, Shi XR, Bai JF, Nagaraja TG (2012) Applicability of a multiplex PCR to detect O26, O45, O103, O111, O121, O145, and O157 serogroups of *Escherichia coli* in cattle feces. *Veterinary Microbiology* 156: 381-388.

81. Shridhar PB, Noll LW, Shi X, An B, Cernicchiaro N, Renter DG, Nagaraja TG, Bai J (2016) Multiplex quantitative PCR assays for the detection and quantification of the six major non-O157 *Escherichia coli* serogroups in cattle feces. *Journal of Food Protection* 79: 66-74.

82. Stromberg ZR, Lewis GL, Aly SS, Lehenbauer TW, Bosilevac JM, Cernicchiaro N, Moxley RA (2016) Prevalence and level of enterohemorrhagic *Escherichia coli* in culled dairy cows at harvest. *Journal of Food Protection* 79: 421-431.

83. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13.

84. Buermans HPJ, den Dunnen JT (2014) Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta* 1842: 1932-1941.

85. Mavromatis K, Land ML, Brettin TS, Quest DJ, Copeland A, Clum A, Goodwin L, Woyke T, Lapidus A, Klenk HP, Cottingham RW, Kyrpides NC (2012) The fast changing landscape of sequencing technologies and their impact on microbial genome assemblies and annotation. *PloS One* 7.

86. Iguchi A, Thomson NR, Ogura Y, Saunders D, Ooka T, Henderson IR, Harris D, Asadulghani M, Kurokawa K, Dean P, Kenny B, Quail MA, Thurston S, Dougan G, Hayashi T, Parkhill J, Frankel G (2009) Complete genome sequence and comparative genome analysis of enteropathogenic *Escherichia coli* O127:H6 strain E2348/69. *Journal of Bacteriology* 191: 347-354.

87. Lorenz SC, Gonzalez-Escalona N, Kotewicz ML, Fischer M, Kase JA (2017) Genome sequencing and comparative genomics of enterohemorrhagic *Escherichia coli* O145:H25 and O145:H28 reveal distinct evolutionary paths and marked variations in traits associated with virulence and colonization. *BMC Microbiology* 17: 15.

88. Morrison SS, Pyzh R, Jeon MS, Amaro C, Roig FJ, Baker-Austin C, Oliver JD, Gibas CJ (2014) Impact of analytic provenance in genome analysis. *BMC Genomics* 15: 11.

89. Nyholm O, Halkilahti J, Wiklund G, Okeke U, Paulin L, Auvinen P, Haukka K, Siitonen A (2015) Comparative genomics and characterization of hybrid shiga toxigenic and enterotoxigenic *Escherichia coli* (STEC/ETEC) strains. *PloS One* 10: 17.

90. Haugum K, Johansen J, Gabrielsen C, Brandal LT, Bergh K, Ussery DW, Drablos 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.

91. Browne AS, Midwinter AC, Withers HL, Cookson AL, Biggs PJ, Marshall J, Benschop J, Hathaway S, French N. Epidemiology of Shiga toxin-producing *E. coli* (STEC) on New Zealand dairy farms using new molecular and genomic technologies; 2016 21-24 June; Hamilton, New Zealand.

92. Neogen Corporation (2013) NeoSeek approach to STEC detection and identification. Retrieved from: http://www.neogen.com/techlibrary/pdf/WhitePapers/EcoliSTECIdentification_0313.pdf.

93. Paton JC, Paton AW (1998) Pathogenesis and diagnosis of Shiga toxin-producing *Escherichia coli* infections. *Clinical Microbiology Reviews* 11: 450.

94. 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.

95. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal Molecular Biology* 215: 403-410.

96. BIOLOG (2012) PM 1-10 Plate Maps. Retrieved from: http://www.biolog.com/pdf/pm_lit/PM1-PM10.pdf.

97. R Core Team (2016) R: A language and environment for statistical computing. Retrieved from: www.r-project.org/.

98. Vaas LAI, Sikorski J, Hofner B, Fiebig A, Buddruhs N, Klenk HP, Goker M (2013) Opm: An R package for analysing omniLog (R) phenotype microarray data. *Bioinformatics* 29: 1823-1824.

99. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WH, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B (2016) gplots: Various R programming tools for plotting data. Retrieved from: http://CRAN.R-project.org/package=gplots.

100. Illumina (2016) Nextera XT DNA library prep reference guide. Retrieved from: http://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nextera-xt/nextera-xt-library-prep-guide-15031942-01.pdf.

101. Illumina (2016) MiSeq: Sequencing chemistry. Retrieved from: http://support.illumina.com/sequencing/sequencing_instruments/miseq/training.html.

102. Andrews S (2010) FastQC: A quality control tool for high throughput sequence data. Retrieved from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

103. Biggs PJ, Truglio M (2016) QCtool. Palmerston North, New Zealand: Massey University Retrieved from:

104. Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 11: 485.

105. Wingett S (2011) FastQ Screen. Retrieved from: http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/.

106. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19: 455-477.

107. Seemann T (2016) Snippy. 3.1 ed Retrieved from: https://github.com/tseemann/snippy/blob/master/README.md.

108. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* 29: 1072-1075.

109. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. *Genome Biology* 5: 9.

110. Biggs PJ (2016) SQS2 scripts. Palmerston North, New Zealand: Massey University Retrieved from:

111. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Journal Molecular Biology* 23: 254-267.

112. Seemann T (2014) Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30: 2068-2069.

113. Biggs PJ (2014) Prokka scripts. Palmerston North, New Zealand: Massey University Retrieved from:

114. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11.

115. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* 35: 3100-3108.

116. Laslett D, Canback B (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research* 32: 11-16.

117. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* 8: 785-786.

118. Kolbe DL, Eddy SR (2011) Fast filtering for RNA homology search. *Bioinformatics* 27: 3102-3109.

119. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang HZ, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LSL (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research* 32: D115-D119.

120. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, Eddy SR (2004) The Pfam protein families database. *Nucleic Acids Research* 32: D138-D141.

121. Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E (2013) TIGRFAMs and genome properties. *Nucleic Acids Research* 41: D387-D395.

122. Center for Genomic Epidemiology (2011) Overview of services. Denmark Retrieved from: http://www.genomicepidemiology.org/.

123. Thomsen MCF, Ahrenfeldt J, Cisneros JLB, Jurtz V, Larsen MV, Hasman H, Aarestrup FM, Lund O (2016) A Bacterial Analysis Platform: An integrated system for analysing cacterial whole genome sequencing data for clinical diagnostics and surveillance. *PloS One* 11.

124. Larsen MV, Cosentino S, Lukjancenko O, Saputra D, Rasmussen S, Hasman H, Sicheritz-Ponten T, Aarestrup FM, Ussery DW, Lund O (2014) Benchmarking of methods for genomic taxonomy. *Journal of Clinical Microbiology* 52: 1529-1539.

125. 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*. *Journal of Clinical Microbiology* 52: 1501-1510.

126. Carattoli A, Zankari E, Garcia-Fernandez 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. *Antimicrobial Agents and Chemotherapy* 58: 3895-3903.

127. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna H, Harrison OB, Sheppard SK, Cody AJ, Maiden MCJ (2012) Ribosomal multilocus sequence typing: Universal characterization of bacteria from domain to strain. *Microbiology-Sgm* 158: 1005-1015.

128. National Center for Biotechnology Information. (1988) National Library of Medicine. USA Retrieved from: https://www.ncbi.nlm.nih.gov/.

129. Enterobase.warwick.ac.uk EnteroBase. Retrieved from: http://enterobase.warwick.ac.uk.

130. Leinonen R, Sugawara H, Shumway M, on behalf of the International Nucleotide Sequence Database C (2011) The Sequence Read Archive. *Nucleic Acids Research* 39: D19-D21.

131. Contreras-Moreira B, Vinuesa P (2013) GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and Environmental Microbiology* 79: 7696-7701.

132. Kristensen DM, Kannan L, Coleman MK, Wolf YI, Sorokin A, Koonin EV, Mushegian A (2010) A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics* 26: 1481-1487.

133. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research* 13: 2178-2189.

134. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Ros IMY, 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 LW, 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". *Proceedings of the National Academy of Sciences of the United States of America* 102: 13950-13955.

135. Kaas RS, Friis C, Ussery DW, Aarestrup FM (2012) Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC Genomics* 13: 13.

136. 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.

137. Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ (2014) BEAST 2: A software platform for bayesian evolutionary analysis. *PLoS Computational Biology* 10.

138. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR (2015) Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research* 43.

139. Rambaut A (2016) FigTree. Edinburgh, UK.: University of Edinburgh Retrieved from: https://github.com/cdeanj/figtree/blob/master/README.txt.

140. Fu LM, Niu BF, Zhu ZW, Wu ST, Li WZ (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28: 3150-3152.

141. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* 30: 1575-1584.

142. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V (2016) Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biology* 17.

143. Kanehisa M, Sato Y, Morishima K (2016) BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *Journal of Molecular Biology* 428: 726-731.

144. Hosking E (personal communication) Retrieved from:

145. Browne AS (2017) personal communication. Retrieved from:

146. Collis RM, Midwinter AC, Cookson AL (2016) Metabolic characteristics of *E. coli* serogroups O121 and O103. Palmerston North, New Zealand: AgResearch Ltd Retrieved from:

147. Evans J, Knight H, McKendrick IJ, Stevenson H, Barbudo AV, Gunn GJ, Low JC (2011) Prevalence of *Escherichia coli* O157: H7 and serogroups O26, O103, O111 and O145 in sheep presented for slaughter in Scotland. *Journal of Medical Microbiology* 60: 653-660.

148. Hara-Kudo Y, Konishi N, Ohtsuka K, Iwabuchi K, Kikuchi R, Isobe J, Yamazaki T, Suzuki F, Nagai Y, Yamada H, Tanouchi A, Mori T, Nakagawa H, Ueda Y, Terajima J (2016) An interlaboratory study on efficient detection of Shiga toxin-producing *Escherichia coli* O26, O103, O111, O121, O145, and O157 in food using real-time PCR assay and chromogenic agar. *International Journal of Food Microbiology* 230: 81-88.

149. Liu YR, Mustapha A (2014) Detection of viable *Escherichia coli* O157:H7 in ground beef by propidium monoazide real-time PCR. *International Journal of Food Microbiology* 170: 48-54.

150. Kibbee RJ, Ormeci B (2017) Development of a sensitive and false-positive free PMA-qPCR viability assay to quantify VBNC *Escherichia coli* and evaluate disinfection performance in wastewater effluent. *Journal of Microbiological Methods* 132: 139-147.

151. Zhou BQ, Liang TB, Zhan ZX, Liu R, Li F, Xu HY (2017) Rapid and simultaneous quantification of viable *Escherichia coli* 0157:H7 and *Salmonella* spp. in milk through multiplex real-time PCR. *Journal of Dairy Science* 100: 8804-8813.

152. Jenkins C, Pearce MC, Smith AW, Knight HI, Shaw DJ, Cheasty T, Foster G, Gunn GJ, Dougan G, Smith HR, Frankel G (2003) Detection of *Escherichia coli* serogroups O26, O103, O111 and O145 from bovine faeces using immunomagnetic separation and PCR/DNA probe techniques. *Letters in Applied Microbiology* 37: 207-212.

153. Stevens MP, Frankel GM (2014) The locus of enterocyte effacement and associated virulence factors of enterohemorrhagic *Escherichia coli*. *Microbiology Spectrum* 2.

154. Donnenberg MS, Whittam TS (2001) Pathogenesis and evolution of virulence in enteropathogenic and enterohemorrhagic *Escherichia coli*. *Journal of Clinical Investigation* 107: 539-548.

155. Durso LM, Smith D, Hutkins RW (2004) Measurements of fitness and competition in commensal *Escherichia coli* and *E. coli* O157 : H7 strains. *Applied and Environmental Microbiology* 70: 6466-6472.

156. Posse B, De Zutter L, Heyndrickx M, Herman L (2008) Novel differential and confirmation plating media for Shiga toxin-producing *Escherichia coli* serotypes O26, O103, O111, O145 and sorbitol-positive and -negative O157. *FEMS Microbiology Letters* 282: 124-131.

157. Schmidt H, Scheef J, Morabito S, Caprioli A, Wieler LH, Karch H (2000) A new Shiga toxin 2 variant (Stx2f) from *Escherichia coli* isolated from pigeons. *Applied and Environmental Microbiology* 66: 1205-1208.

158. Friesema I, van der Zwaluw K, Schuurman T, Kooistra-Smid M, Franz E, van Duynhoven Y, van Pelt W (2014) Emergence of *Escherichia coli* encoding Shiga toxin 2f in human Shiga toxin-producing *E. coli* (STEC) infections in the Netherlands, January 2008 to December 2011. *Eurosurveillance* 19: 26-32.

159. Prager R, Fruth A, Siewert U, Strutz U, Tschape H (2009) *Escherichia coli* encoding Shiga toxin 2f as an emerging human pathogen. *International Journal of Medical Microbiology* 299: 343-353.

160. Scheutz F, Teel LD, Beutin L, Pierard 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* somenclature. *Journal of Clinical Microbiology* 50: 2951-2963.

161. Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology* 11: 472-477.

162. Lukjancenko O, Wassenaar TM, Ussery DW (2010) Comparison of 61 sequenced *Escherichia coli* genomes. *Microbial Ecology* 60: 708-720.

163. Creuzburg K, Middendorf B, Mellmann A, Martaler T, Holz C, Fruth A, Karch H, Schmidt H (2011) Evolutionary analysis and distribution of type III effector genes in pathogenic *Escherichia coli* from human, animal and food sources. *Environmental Microbiology* 13: 439-452.

164. 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. *Journal of Bacteriology* 190: 6881-6893.

165. Browne AS, Biggs PJ, Cookson AL, Midwinter AC, Marshall J, Benschop J, Bloomfield S, Wilkinson D, Roger L, Withers H, Hathaway S, George T, Jaros

P, Irshad H, French N. The local and global evolution and transmission of Shiga toxin-producing *E. coli* (STEC) serogroup O26; 2017; Wellington, New Zealand. Massey University.

166. McInerney JO, Pisani D, Bapteste E, O'Connell MJ (2011) The public goods hypothesis for the evolution of life on Earth. *Biology Direct* 6: 17.

167. Friis C, Wassenaar TM, Javed MA, Snipen L, Lagesen K, Hallin PF, Newell DG, Toszeghy M, Ridley A, Manning G, Ussery DW (2010) Genomic characterization of *Campylobacter jejuni* strain M1. *PloS One* 5: 12.

168. Snipen L, Ussery DW (2010) Standard operating procedure for computing pangenome trees. *Standards in Genomic Sciences* 2: 135-141.

169. Dallman TJ, Ashton PM, Byrne L, Perry NT, Petrovska L, Ellis R, Allison L, Hanson M, Holmes A, Gunn GJ, Chase-Topping ME, Woolhouse ME, Grant KA, Gally DL, Wain J, Jenkins C (2015) Applying phylogenomics to understand the emergence of Shiga toxin-producing *Escherichia coli* O157:H7 strains causing severe human disease in the UK. *Microbial Genomics* 1.

170. von Mentzer A, Connor TR, Wieler LH, Semmler T, Iguchi A, Thomson NR, Rasko DA, Joffre E, Corander J, Pickard D, Wiklund G, Svennerholm AM, Sjoling A, Dougan G (2014) Identification of enterotoxigenic *Escherichia coli* (ETEC) clades with long-term global distribution. *Nature Genetics* 46: 1321-1326.

171. Browne AS, Biggs PJ, Marshall J, Cookson AL, Midwinter AC, Benschop J, Withers H, Hathaway S, French N (2016) Evolution, virulence and source attribution of *Escherichia coli* serogroup O26 isolates from New Zealand cattle and humans. One Health Conference. Melbourne, Australia. Retrieved from:

172. Binney BM, Biggs PJ, Carter PE, Holland BM, French NP (2014) Quantification of historical livestock importation into New Zealand 1860-1979. *New Zealand Veterinary Journal* 62: 309-314.

173. Twardon J, Sobieszczanska B, Gonet A, Blaszkowska M (2005) Epidemiology of Shiga-like toxin-producing *Escherichia coli* strains (STEC). *Electronic Journal of Polish Agricultural Universities* 8.

174. Holt KE, Baker S, Weill FX, Holmes EC, Kitchen A, Yu J, Sangal V, Brown DJ, Coia JE, Kim DW, Choi SY, Kim SH, da Silveira WD, Pickard DJ, Farrar JJ, Parkhill J, Dougan G, Thomson NR (2012) *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nature Genetics* 44: 1056.

175. Hall DR, Leonard GA, Reed CD, Watt CI, Berry A, Hunter WN (1999) The crystal structure of *Escherichia coli* class II fructose-1,6-bisphosphate aldolase in complex with phosphoglycolohydroxamate reveals details of mechanism and specificity. *Journal of Molecular Biology* 287: 383-394.

176. Ziveri J, Tros F, Guerrera IC, Chhuon C, Audry M, Dupuis M, Barel M, Korniotis S, Fillatreau S, Gales L, Cahoreau E, Charbit A (2017) The metabolic enzyme fructose-1,6-bisphosphate aldolase acts as a transcriptional regulator in pathogenic *Francisella*. *Nature Communications* 8: 15.

177. Sturgill G, Toutain CM, Komperda J, O'Toole GA, Rather PN (2004) Role of CysE in production of an extracellular signaling molecule in *Providencia stuartii* and *Escherichia coli*: Loss of *cysE* enhances biofilm formation in *Escherichia coli*. *Journal of Bacteriology* 186: 7610-7617.

178. Pye VE, Tingey AP, Robson RL, Moody PCE (2004) The structure and mechanism of serine acetyltransferase from *Escherichia coli*. *Journal of Biological Chemistry* 279: 40729-40736.

179. Sorensen KI, Hove-Jensen B (1996) Ribose catabolism of *Escherichia coli*: Characterization of the *rpiB* gene encoding ribose phosphate isomerase B and of the *rpiR* gene, which is involved in regulation of *rpiB* expression. *Journal of Bacteriology* 178: 1003-1011.

180. Zhang RG, Andersson CE, Savchenko A, Skarina T, Evdokimova E, Beasley S, Arrowsmith CH, Edwards AM, Joachimiak A, Mowbray SL (2003) Structure of *Escherichia coli* ribose-5-phosphate isomerase: A ubiquitous enzyme of the pentose phosphate pathway and the Calvin cycle. *Structure* 11: 31-42.

181. Jenkins LS, Nunn WD (1987) Regulation of the *ato* operon by the *atoC* gene in *Escherichia coli*. *Journal of Bacteriology* 169: 2096-2102.

182. Schminckeott E, Bisswanger H (1981) Dihydrolipoamide dehydrogenase component of the pyruvate-dehydrogenase complex from *Escherichia coli* K-12: Comparative characterization of the free and complex-bound component. *European Journal of Biochemistry* 114: 413-420.

183. Pouyssegur J, Stoeber F (1974) Genetic control of the 2-keto-3-deoxy-D-gluconate metabolism in *Escherichia coli* K-12: *kdg* regulon. *Journal of Bacteriology* 117: 641-651.

184. Cynkin MA, Ashwell G (1960) Uronic acid metabolism in bacteria. 4. Purificaiton and properties of 2-keto-3-deoxy-D-gluconokinase in *Escherichia coli*. *Journal of Biological Chemistry* 235: 1576-1579.

185. Prodromou C, Haynes MJ, Guest JR (1991) The aconitase of *Escherichia coli:* Purification of the enzyme and molecular-cloning and map location of the gene (*acn*). *Journal of General Microbiology* 137: 2505-2515.

186. Cookson AL, Cao M, Bennett J, Nicol C, Thomson-Carter F, Attwood GT (2010) Relationship between virulence gene profiles of atypical enteropathogenic *Escherichia coli* and Shiga toxin- troducing *E. coli* isolates from cattle and sheep in New Zealand. *Applied and Environmental Microbiology* 76: 3744-3747.

187. Irshad H, Cookson AL, Ross CM, Jaros P, Prattley DJ, Donnison A, McBride G, Marshall J, French NP (2016) Diversity and relatedness of Shiga toxin-producing *Escherichia coli* and *Campylobacter jejuni* between farms in a dairy catchment. *Epidemiology and Infection* 144: 1406-1417.

188. Stromberg ZR, Lewis GL, Schneider LG, Erickson GE, Patel IR, Smith DR, Moxley RA (2018) Culture-based quantification with molecular characterization of non-O157 and O157 enterohemorrhagic *Escherichia coli* isolates from rectoanal mucosal swabs of feedlot cattle. *Foodborne Pathogens and Disease* 15: 26-32.

189. 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 Pathogens and Disease* 9: 792-795.

190. Chaudhari NM, Gupta VK, Dutta C (2016) BPGA- an ultra-fast pan-genome analysis pipeline. *Scientific Reports* 6: 10.

191. Karch H, Meyer T, Russmann H, Heesemann J (1992) Frequent loss of Shiga-like toxin genes in clinical isolates of *Escherichia coli* upon subcultivation. *Infection and Immunity* 60: 3464-3467.

192. Browne AS, Biggs PJ, Marshall J, Cookson AL, Midwinter AC, Benschop J, Withers H, Hathaway S, French N (2018) Transmission dynamics of STEC on New Zealand dairy farms and implications for transport, lairage, and processing of veal carcasses. Palmerston North, New Zealand: Massey University Retrieved from:

193. Bosilevac JM, Wang R, Luedtke BE, Hinkley S, Wheeler TL, Koohmaraie M (2017) Characterization of enterohemorrhagic *Escherichia coli* on veal hides and carcasses. *Journal of Food Protection* 80: 136-145.

194. 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. *Journal of Clinical Microbiology* 42: 311-319.

195. Blanco M, Padola NL, Kruger A, Sanz ME, Blanco JE, Gonzalez EA, Dahbi G, Mora A, Bernardez MI, Etcheverria 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. *International Microbiology* 7: 269-276.

196. Fierz L, Cernela N, Hauser E, Nuesch-Inderbinen M, Stephan R (2017) Characteristics of Shiga toxin-producing *Escherichia coli* strains isolated during 2010-2014 from human infections in Switzerland. *Frontiers in Microbiology* 8: 7.

197. Posse B, De Zutter L, Heyndrickx M, Herman L (2007) Metabolic and genetic profiling of clinical O157 and non-O157 Shiga-toxin-producing *Escherichia coli*. *Research in Microbiology* 158: 591-599.

198. Sabarly V, Bouvet O, Glodt J, Clermont O, Skurnik D, Diancourt L, de Vienne D, Denamur E, Dillmann C (2011) The decoupling between genetic structure and metabolic phenotypes in *Escherichia coli* leads to continuous phenotypic diversity. *Journal of Evolutionary Biology* 24: 1559-1571.

199. Desroches M, Royer G, Roche D, Mercier-Darty M, Vallenet D, Medigue C, Bastard K, Rodriguez C, Clermont O, Denamur E, Decousser JW (2018) The odyssey of the ancestral *Escherich* Strain through culture collections: An example of allopatric diversification. *mSphere* 3.

200. Cookson AL (2017) Whole genome sequencing of *E. coli* serogroup O157. Palmerston North: AgResearch Ltd. Retrieved from:

201. Phillips AD, Frankel G (2000) Intimin-mediated tissue specificity in enteropathogenic *Escherichia coli* interaction with human intestinal organ cultures. *Journal of Infectious Diseases* 181: 1496-1500.

202. Phillips AD, Navabpour S, Hicks S, Dougan G, Wallis T, Frankel G (2000) Enterohaemorrhagic *Escherichia coli* O157 : H7 target Peyer's patches in humans and cause attaching/effacing lesions in both human and bovine intestine. *Gut* 47: 377-381.

203. Fitzhenry RJ, Stevens MP, Jenkins C, Wallis TS, Heuschkel R, Murch S, Thomson M, Frankel G, Phillips AD (2003) Human intestinal tissue tropism of intimin epsilon O103 *Escherichia coli*. *FEMS Microbiology Letters* 218: 311-316.

204. Sammarro M, Vignes M, Biggs PJ, French NP, Marshall J (2016) Revealing the genetic basis for host-pathogen interaction using Machine Learning. Palmerston North, New Zealand: Massey University. 1-38 p. Retrieved from:

205. Irshad H (2013) Molecular epidemiology of Shiga toxin-producing *Escherichia coli* O157 and non-O157 STEC in calves in the North island of New Zealand. Palmerston North: Massey. 182 p.

206. The Institute of Environmental Science and Research Ltd (n.d.) The Enteric Reference Laboratory. Retrieved from: https://surv.esr.cri.nz/.

207. Ross C, Rapp D, Brightwell G (2017) Longitudinal study of the prevalence and genotypes of the “Top 7” STEC in the environment of two Waikato-based case-study farms 1-30 p. Retrieved from:

208. Nguyen RN, Taylor LS, Tauschek M, Robins-Browne RM (2006) Atypical enteropathogenic *Escherichia coli* infection and prolonged diarrhea in children. *Emerging Infectious Diseases* 12: 597-603.

209. Afset JE, Anderssen E, Bruant G, Harel J, Wieler L, Bergh K (2008) Phylogenetic backgrounds and virulence profiles of atypical enteropathogenic *Escherichia coli* strains from a case-control study using multilocus sequence typing and DNA microarray analysis. *Journal of Clinical Microbiology* 46: 2280-2290.

210. Public Health England (2016) Routine surveillance of *E. coli* and Shigella. Retrieved from: https://www.ncbi.nlm.nih.gov/sra/?term=SRR3581355.

211. Trees E, Strockbine N, Changayil S, Ranganathan S, Zhao K, Weil R, MacCannell D, Sabol A, Schmidtke A, Martin H, Stripling D, Ribot EM, Gerner-Smidt P (2014) Genome sequences of 228 Shiga toxin-producing *Escherichia coli* isolates and 12 isolates representing other diarrheagenic *E. coli* pathotypes. *Genome Announcements* 2: e00718-00714.

212. FDA Center for Food Safety and Applied Nutrition (2013) Foodborne pathogen survey. Retrieved from: https://www.ncbi.nlm.nih.gov/sra/?term=SRR975374.

213. FDA Center for Food Safety and Applied Nutrition (2014) GenomeTrakr Project. Retrieved from: https://www.ncbi.nlm.nih.gov/sra/?term=SRR1272534.

214. U.S Department of Agriculture FSaIS (2014) GenomeTrakr project: USDA-FSIS. Retrieved from: https://www.ncbi.nlm.nih.gov/sra/?term=SRR1693413.

215. U.S Department of Agriculture FSaIS (2015) GenomeTrakr Project: USDA-FSIS. Retrieved from: https://www.ncbi.nlm.nih.gov/sra/?term=SRR2126002.

216. U.S Department of Agriculture FSaIS (2016) GenomeTrakr project: USDA-FSIS. Retrieved from: https://www.ncbi.nlm.nih.gov/sra/?term=SRR3185246.

217. FDA Center for Food Safety and Applied Nutrition (2016) Genome Trakr project: U.S Food and Drug Adminstration. Retrieved from: https://www.ncbi.nlm.nih.gov/sra/?term=SRR3124089.

218. Centers for Disease Control and Prevention Enteric Diseases Laboratory Branch (2016) PulseNet *Escherichia coli* and *Shigella* genome sequencing. Retrieved from: https://www.ncbi.nlm.nih.gov/sra/?term=SRR3371785.

9. Appendices

Appendix A - Bacterial strains used in this study

Strain†	Serotype	Source	Origin	Virulence profile‡	*eae* subtype§	Sequence type§	Reference/ source
116B†	O145	Bovine	Taranaki, New Zealand	*eae, ehxA*	ε	ST-17	Irshad [205]
13ER3103A†	O145:HNM	Human	Auckland, New Zealand	*stx2, eae, ehxA*	γ	ST-32	The Institute of Environmental Science and Research Ltd [206]
13ER4824†	O145	Bovine	New Zealand	*stx2, eae, ehxA*	γ	ST-32	The Institute of Environmental Science and Research Ltd [206]
13ER5056†	O145	Bovine	New Zealand	*stx2, eae, ehxA*	γ	ST-32	The Institute of Environmental Science and Research Ltd [206]
13ER5154†	O145	Bovine	New Zealand	*stx2, eae, ehxA*	γ	ST-32	The Institute of Environmental Science and Research Ltd [206]
13ER5640†	O145	Bovine	New Zealand	*stx2, eae, ehxA*	γ	ST-32	The Institute of Environmental Science and Research Ltd [206]
13ER6227†	O145	Bovine	New Zealand	*eae, ehxA*	γ	ST-32	The Institute of Environmental Science and Research Ltd [206]
13ER6723A†	O145:H34	Human	Auckland, New Zealand	*stx2, eae*	ι	ST-722	The Institute of Environmental Science and Research Ltd [206]
14ER2392†	O145	Bovine	New Zealand	*stx2, eae, ehxA*	γ	ST-32	The Institute of Environmental Science and Research Ltd [206]
15ER2679†	O145	Bovine	New Zealand	*eae, ehxA*	γ	ST-32	The Institute of Environmental Science and Research Ltd [206]
16ER0267A†	O145:H2	Human	Auckland, New Zealand	*stx1, eae, ehxA*	ε	ST-17	The Institute of Environmental Science and Research Ltd [206]
16ER0517A†	O145:H2	Human	Auckland, New Zealand	*stx1, eae, ehxA*	ε	ST-17	The Institute of Environmental Science and Research Ltd [206]
188B†	O145	Bovine	Taranaki, New Zealand	*eae, ehxA*	ε	ST-17	Irshad [205]
267P†	O145	Bovine	Taranaki, New Zealand	*eae, ehxA*	ε	ST-17	Irshad [205]
54B†	O145	Bovine	Taranaki, New Zealand	*eae, ehxA*	ε	ST-17	Irshad [205]
AGR718†	O145:H46	Bovine	Manawatu, New Zealand	*eae, ehxA*	γ	ST-137	Cookson et al. [186]

Strain†	Serotype	Source	Origin	Virulence profile‡	*eae* subtype§	Sequence type§	Reference/ source
ERL020412†	O145:H-	Human	New Zealand	*eae, ehxA*	γ	ST-137	The Institute of Environmental Science and Research Ltd [206]
ERL121829†	O145	Bovine	New Zealand	*eae, ehxA*	γ	ST-32	The Institute of Environmental Science and Research Ltd [206]
ERL122034†	O145	Bovine	New Zealand	*eae, ehxA*	ε	ST-17	The Institute of Environmental Science and Research Ltd [206]
F1†	O145	Environmental	Waikato, New Zealand	*eae, ehxA*	γ	ST-32	Ross et al. [207]
F5F†	O145	Environmental	Waikato, New Zealand	*eae, ehxA*	γ	ST-32	Ross et al. [207]
F5J†	O145	Environmental	Waikato, New Zealand	*eae, ehxA*	γ	ST-32	Ross et al. [207]
FDE21†	O145	Environmental	Waikato, New Zealand	*eae, ehxA*	γ	ST-32	Ross et al. [207]
H12ESR01231†	O145	Bovine	New Zealand	*eae, ehxA*	γ	ST-32	The Institute of Environmental Science and Research Ltd [206]
H12ESR01387†	O145	Bovine	New Zealand	*stx2, eae, ehxA*	γ	ST-32	The Institute of Environmental Science and Research Ltd [206]
H12ESR01650†	O145	Bovine	New Zealand	*eae, ehxA*	γ	ST-32	The Institute of Environmental Science and Research Ltd [206]
H12ESR03525†	O145	Bovine	New Zealand	*stx2, eae, ehxA*	γ	ST-32	The Institute of Environmental Science and Research Ltd [206]
P2A1†	O145	Environmental	Waikato, New Zealand	*eae, ehxA*	γ	ST-32	Ross et al. [207]
P2B1†	O145	Environmental	Waikato, New Zealand	*eae, ehxA*	γ	ST-32	Ross et al. [207]
R249-1†	O145:H34	Human	Australia	*eae*	ι	ST-722	Nguyen et al. [208]
Trh30†	O145:H-	Human	Norway	*eae, ehxA*	γ	ST-32	Afset et al. [209]
Trh42†	O145:H34	Human	Norway	*eae*	ι	ST-35	Afset et al. [209]
Trh46†	O145:H25	Human	Norway	*eae*	ι	ST-526	Afset et al. [209]
Trh7†	O145:H40	Human	Norway	*eae*	β	ST-10	Afset et al. [209]
TW07865†	O145:H28	Human	Germany	*stx2, eae, ehxA*	γ	ST-137	The Institute of Environmental Science and Research Ltd [206]
VC1048m†	O145	Bovine	Taranaki, New Zealand	*eae, ehxA*	γ	ST-137	This study
VC1048n	O145	Bovine	Taranaki, New Zealand	*eae, ehxA*	ND	ND	This study

Strain†	Serotype	Source	Origin	Virulence profile‡	*eae* subtype§	Sequence type§	Reference/ source
VC1048o	O145	Bovine	Taranaki, New Zealand	*eae, ehxA*	ND	ND	This study
VC1048p	O145	Bovine	Taranaki, New Zealand	*eae, ehxA*	ND	ND	This study
VC1056m†	O145	Bovine	Taranaki, New Zealand	*eae, ehxA*	γ	ST-32	This study
VC123n†	O145	Bovine	Waikato, New Zealand	*eae, ehxA*	γ	ST-32	This study
VC1281m†	O145	Bovine	Canterbury, New Zealand	*eae, ehxA*	γ	ST-32	Browne et al. [18]
VC1413m†	O145	Bovine	Southland, New Zealand	*stx2, eae, ehxA*	γ	ST-32	This study
VC1414n	O145	Bovine	Southland, New Zealand	*stx2, eae, ehxA*	ND	ND	Browne et al. [18]
VC1506m†	O145	Bovine	Southland, New Zealand	*eae, ehxA*	γ	ST-32	Browne et al. [18]
VC194m†	O145	Bovine	Northland, New Zealand	*eae, ehxA*	γ	ST-32	This study
VC194n	O145	Bovine	Northland, New Zealand	*eae, ehxA*	ND	ND	This study
VC237m†	O145	Bovine	Northland, New Zealand	*eae, ehxA*	γ	ST-32	This study
VC237n	O145	Bovine	Northland, New Zealand	*eae, ehxA*	ND	ND	This study
VC237o†	O145	Bovine	Northland, New Zealand	*eae, ehxA*	γ	ST-32	This study
VC237p	O145	Bovine	Northland, New Zealand	*eae, ehxA*	ND	ND	This study
VC308m†	O145	Bovine	Northland, New Zealand	*eae, ehxA*	γ	ST-32	This study
VC308n	O145	Bovine	Northland, New Zealand	*eae, ehxA*	ND	ND	This study
VC308o	O145	Bovine	Northland, New Zealand	*eae, ehxA*	ND	ND	This study
VC308p	O145	Bovine	Northland, New Zealand	*eae, ehxA*	ND	ND	This study
VC476m†	O145	Bovine	Waikato, New Zealand	*eae, ehxA*	γ	Unknown	This study
VC506m†	O145	Bovine	Waikato, New Zealand	*eae, ehxA*	γ	ST-32	This study
VC506n	O145	Bovine	Waikato, New Zealand	*eae, ehxA*	ND	ND	This study
VC506o	O145	Bovine	Waikato, New Zealand	*eae, ehxA*	ND	ND	This study
VC508m†	O145	Bovine	Waikato, New Zealand	*eae, ehxA*	γ	ST-32	This study
VC508n	O145	Bovine	Waikato, New Zealand	*eae, ehxA*	ND	ND	This study
VC508o	O145	Bovine	Waikato, New Zealand	*eae, ehxA*	ND	ND	This study
VC508p	O145	Bovine	Waikato, New Zealand	*eae, ehxA*	ND	ND	This study

Strain†	Serotype	Source	Origin	Virulence profile‡	*eae* subtype§	Sequence type§	Reference/ source
VC525m†	O145	Bovine	Waikato, New Zealand	*eae, ehxA*	γ	ST-32	Browne et al. [18]
VC525o	O145	Bovine	Waikato, New Zealand	*eae, ehxA*	ND	ND	Browne et al. [18]
VC554m†	O145	Bovine	Waikato, New Zealand	*eae, ehxA*	γ	ST-32	This study
VC554n	O145	Bovine	Waikato, New Zealand	*eae, ehxA*	ND	ND	This study
VC847m†	O145	Bovine	Manawatu-Wellington, New Zealand	*eae, ehxA*	γ	ST-32	Browne et al. [18]
VC849m†	O145	Bovine	Manawatu-Wellington, New Zealand	*eae, ehxA*	γ	ST-32	This study
VC849n	O145	Bovine	Manawatu-Wellington, New Zealand	*eae, ehxA*	ND	ND	This study
VC874o†	O145	Bovine	Taranaki, New Zealand	*eae, ehxA*	γ	ST-32	Browne et al. [18]
VC880m†	O145	Bovine	Taranaki, New Zealand	*eae, ehxA*	γ	ST-32	This study
VC880n	O145	Bovine	Taranaki, New Zealand	*eae, ehxA*	γ	ST-32	This study

†: *E. coli* serogroup O145 strains which underwent whole genome sequencing in this study

‡: Virulence profile (*stx*1, *stx*2, *eae*, and/or *ehx*A) determined using mPCR. *stx*2 was not detected in strain 13ER6723A using mPCR (see section 5.2.2)

§: ND, not done. *eae* subtype and MLST is not available for strains for where whole genome sequence was not undertaken

Appendix B - R code for Omnilog analysis

R code for generating XY plots:

```
### Generate XY plots ###

library(opm) #load OPM package
data<-read_opm("Filename.csv") #load raw .csv data in working directory (one plate per .csv file)
metadata<-collect_template(data) #collect metadata
metadata[,"label"]<-sub("\\.[[:alnum:]]+$","",basename(metadata[,"File"]))
print(metadata) # show metadata in console to check for errors
data<-include_metadata(data, md=metadata) #combine the metadata with the original data vector

# print xy_plot. when only one data file is loaded remove legend to prevent errors.
if(length(data)==1){print(xy_plot(data))
}else {print(xy_plot(data,include=c("label")))
}


### Make XY plot high resolution ###

tiff(filename = "Filename.tiff", res=600, pointsize=2, width=12000,height=6750,compression="lzw")
print(xy_plot(data[], lwd=0.4, main="Main_title")) #draw graph data[FileNumber,DataPoints,Wells],
dev.off() # save image to file and end command
```

R code was written by Kevin Rouw and modified by Angie Reynolds and Rose Collis.

R code for generating heat-maps:

```
#########################################################
### A) Reading in data and transform it into matrix format
#########################################################

data <- read.csv("Filename.csv", comment.char="#")
rnames <- data[,1]                            # assign labels in column 1 to "rnames"
mat_data <- data.matrix(data[,2:ncol(data)])  # transform column 2-5 into a matrix
rownames(mat_data) <- rnames                  # assign row names

#########################################################
### B) Customizing and plotting the heat map
#########################################################

# creates a own color palette from red to green
my_palette <- colorRampPalette(c("mediumblue", "limegreen", "yellow2"))(n = 299)

# (optional) defines the color breaks manually for a "skewed" color transition
col_breaks = c(seq(1,length=100),  # for blue3
               seq(2,length=100),           # green
               seq(3,length=100))             # yellow


# creates a 5 x 5 inch image#edited for my file size#
tiff(filename = "Filename.tiff",    # create tiff for the heat map
     width = 13500,         # 12000 pixels
     height = 8500,
     res = 600,             # 600 pixels per inch,
     pointsize = 22,        # smaller font size
     compression="lzw")    # compression

heatmap.2(mat_data,
          main = "Main_title", # heat map title
          RowSideColors = c(    # grouping row-variables into different
            rep("gray52", 1),   #negative control
            rep("darkorchid4", 3),    # iota
            rep("dodgerblue", 4),    #epsilon
            rep("chocolate1", 1),    #beta
            rep("firebrick3", 12)),     #gamma)
          key= FALSE,
          notecol="black",      # change font color of cell labels to black
          density.info="none",  # turns off density plot inside color legend
          trace="none",         # turns off trace lines inside the heat map
          margins =c(15,12),   # widens margins around plot
          col=my_palette,       # use on color palette defined earlier
          dendrogram="row")     # only draw a row dendrogram
Colv="NA"            # turn off column clustering

legend(inset=.05,                      #inset into figure
       cex = 0.8,                        #size of legend
       x= 0.065,                        #x axis position on plot
       y=1,                             # y axis position on plot
       title="Colour Key (Omnilog Units)",  #title of legend
legend(inset=.05,                     #inset into figure
       cex = 0.8,                       #size of legend
       x= 0.01,                        #x axis position on plot
       y=0.9,                           #y axis position on plot
       horiz=TRUE,
       title="eae Subtype",  #title of legend
       c( "NA", "iota", "epsilon", "beta", "gamma"), # category labels
       fill=c("gray52", "darkorchid4", "dodgerblue", "chocolate1", "firebrick3"),
       bg="grey96")

dev.off()
```

Appendix C - SQS2 perl script

The SQS2 perl script can be accessed via the following Dropbox link: https://www.dropbox.com/sh/n81p1cf27gs887o/AACLwEuiykXAnNpHWSM1EEhQa?dl=0

Appendix D - Prokka perl script

The Prokka perl script can be accessed via the following Dropbox link: https://www.dropbox.com/sh/n81p1cf27gs887o/AACLwEuiykXAnNpHWSM1EEhQa?dl=0

Appendix E - Publicly available genome sequences analysed in this study

Isolate	Serotype	Source	Origin	STEC	*eae* subtype	Sequence type	SRA accession number	BioSample ID	Reference/ source
130322	O145	Human	UK	+	γ	ST-32	SRR3579383	SAMN05171053	Public Health England [210]
132030	O145	Human	UK	+	γ	ST-32	SRR3578591	SAMN05170684	Public Health England [210]
170303	O145:H34	Human	UK	-	ι	ST-722	SRR3578794	SAMN05170850	Public Health England [210]
173582	O145	Human	UK	+	γ	ST-32	SRR3581442	SAMN05171900	Public Health England [210]
173758	O145	Human	UK	+	γ	ST-32	SRR3581328	SAMN05171781	Public Health England [210]
182131	O145:H40	Human	UK	-	β	ST-10	SRR3581355	SAMN05171820	Public Health England [210]
199816	O145:H34	Human	UK	-	α2	ST-722	SRR3578986	SAMN05171019	Public Health England [210]
2010C-3507	O145	Human	USA	+	γ	ST-32	SRR3371785	SAMN02352965	Trees et al. [211]
2010C-3508	O145	Human	USA	+	γ	ST-32	SRR3371786	SAMN02352966	Trees et al. [211]
2010C-3509	O145	Human	USA	+	γ	ST-32	SRR3371787	SAMN02352967	Trees et al. [211]
2010C-3510	O145	Human	USA	+	γ	ST-32	SRR3371788	SAMN02352968	Trees et al. [211]
2010C-3511	O145	Human	USA	+	γ	ST-32	SRR3371789	SAMN02352969	Trees et al. [211]
2010C-3516	O145	Human	USA	+	γ	ST-32	SRR3371790	SAMN02352970	Trees et al. [211]
2010C-3518	O145	Human	USA	+	γ	ST-32	SRR3371792	SAMN02352972	Trees et al. [211]
2010C-3521	O145	Human	USA	+	γ	ST-32	SRR3371793	SAMN02352973	Trees et al. [211]
2010C-3526	O145	Human	USA	+	γ	ST-32	SRR3371794	SAMN02352974	Trees et al. [211]
2012C-4474	O145	Human	USA	+	γ	ST-32	SRR975374	SAMN02352667	FDA Center for Food Safety and Applied Nutrition [212]
2012C-4477	O145	Human	USA	+	γ	ST-32	SRR975375	SAMN02352668	FDA Center for Food Safety and Applied Nutrition [212]
2012C-4478	O145	Human	USA	+	γ	ST-32	SRR975376	SAMN02352669	FDA Center for Food Safety and Applied Nutrition [212]
2012C-4479	O145	Human	USA	+	γ	ST-32	SRR975377	SAMN02352670	FDA Center for Food Safety and Applied Nutrition [212]

Isolate	Serotype	Source	Origin	STEC	*eae* subtype	Sequence type	SRA accession number	BioSample ID	Reference/ source
2012C-4480	O145	Human	USA	+	γ	ST-32	SRR975378	SAMN02352671	FDA Center for Food Safety and Applied Nutrition [212]
201499	O145:H34	Human	UK	-	α2	ST1877	SRR3578586	SAMN05170678	Public Health England [210]
203972	O145	Human	UK	-	ι	ST-722	SRR4192081	SAMN05750680	Public Health England [210]
241761	O145	Human	UK	+	γ	ST-32	SRR3574267	SAMN05163744	Public Health England [210]
241810	O145	Human	UK	-	γ	ST-32	SRR3574240	SAMN05163717	Public Health England [210]
AA053	O145	Human	Denmark	-	γ	ST-137	ERR1010242	SAMEA3529328	Joensen et al. [23]
ewgs1003	O145	Food	USA	+	γ	ST-32	SRR1272534	SAMN02566897	FDA Center for Food Safety and Applied Nutrition [213]
FSIS1400369	O145	Cattle	USA	+	γ	ST-32	SRR1693413	SAMN03216751	U.S Department of Agriculture [214]
FSIS1500788	O145	Food	USA	-	γ	ST-32	SRR2126002	SAMN03922108	U.S Department of Agriculture [215]
FSIS1501198	O145	Cattle	USA	+	γ	ST-32	SRR3405428	SAMN04870300	U.S Department of Agriculture [216]
FSIS1501717	O145	Food	USA	+	γ	ST-32	SRR3405608	SAMN04870303	U.S Department of Agriculture [216]
FSIS1502535	O145	Food	USA	+	γ	ST-32	SRR3185246	SAMN04510513	U.S Department of Agriculture [216]
FSIS1502550	O145	Food	USA	+	γ	ST-32	SRR3175216	SAMN04497383	U.S Department of Agriculture [216]
FSIS1502554	O145	Food	USA	+	γ	ST-32	SRR3175217	SAMN04497385	U.S Department of Agriculture [216]
FSIS1502976	O145	Food	USA	+	γ	ST-32	SRR3175218	SAMN04497386	U.S Department of Agriculture [216]
FSIS1502978	O145	Food	USA	+	γ	ST-32	SRR3185253	SAMN04510516	U.S Department of Agriculture [216]
FSIS1503305	O145	Cattle	USA	+	γ	ST-32	SRR3441262	SAMN04901729	U.S Department of Agriculture [216]

Isolate	Serotype	Source	Origin	STEC	*eae* subtype	Sequence type	SRA accession number	BioSample ID	Reference/ source
FSIS1503307	O145	Cattle	USA	+	γ	ST-32	SRR3441301	SAMN04901731	U.S Department of Agriculture [216]
FSIS1504619	O145	Cattle	USA	+	γ	ST-32	SRR2826835	SAMN04208149	U.S Department of Agriculture [215]
FSIS1505314	O145	Cattle	USA	+	γ	ST-32	SRR3106214	SAMN04421068	U.S Department of Agriculture [216]
FSIS1605419	O145	Bovine	USA	+	γ	ST-32	SRR3106215	SAMN04421070	U.S Department of Agriculture [216]
FSIS1605420	O145	Bovine	USA	+	γ	ST-32	SRR3106213	SAMN04421071	U.S Department of Agriculture [216]
MOD1-EC2002	O145	Human	USA	+	γ	ST-32	SRR3124089	SAMN04256127	FDA Center for Food Safety and Applied Nutrition [217]
MOD1-EC5842	O145	Swine	USA	-	γ	ST-32	SRR3987499	SAMN05439376	FDA Center for Food Safety and Applied Nutrition [217]
MOD1-EC6028	O145	Swine	USA	+	γ	ST-137	SRR3988027	SAMN05439480	FDA Center for Food Safety and Applied Nutrition [217]
PNUSAE000756	O145	Human	USA	+	γ	ST-32	SRR3541143	SAMN03782146	Centers for Disease Control and Prevention Enteric Diseases Laboratory Branch [218]
PNUSAE001244	O145	Human	USA	+	γ	ST-32	SRR2177987	SAMN04002944	Centers for Disease Control and Prevention Enteric Diseases Laboratory Branch [218]

Appendix F – Calf faecal enrichments screened for serogroup O145 using culture-based methods

Calf ID†	Farm ID	Shed ID	C_t value‡	Direct plating isolation successful§	IMS and culture isolation successful§	Number of isolates identified
VC101	VCF7	20	22.84	-	-	0
VC1046	VCF71	188	21.47	-	-	0
VC1048	VCF71	187	20.41	+	ND	4
VC1051	VCF71	187	22.54	-	ND	0
VC1056	VCF72	190	17.80	+	ND	1
VC1057	VCF72	190	24.40	-	ND	0
VC1061	VCF72	190	16.77	-	ND	0
VC1140	VCF77	205	25.31	-	ND	0
VC123	VCF9	23	23.64	+	ND	1
VC1246	VCF84	224	21.75	-	-	0
VC1259	VCF85	226	21.20	-	-	0
VC1262	VCF85	226	29.52	-	ND	0
VC1279	VCF87	231	25.20	-	-	0
VC1394	VCF95	253	26.73	-	ND	0
VC1413	VCF96	254	26.56	ND	+	1
VC1502	VCF102	267	26.03	-	ND	0
VC194	VCF13	36	21.47	+	ND	2
VC227	VCF16	42	24.33	-	ND	0
VC229	VCF16	42	24.09	-	-	0
VC237	VCF16	44	22.01	-	+	4
VC290	VCF20	54	28.06	ND	-	0
VC307	VCF21	58	26.69	-	ND	0
VC308	VCF21	58	23.18	-	+	4
VC313	VCF21	59	30.77	-	-	0

Calf ID†	Farm ID	Shed ID	C_t value‡	Direct plating isolation successful§	IMS and culture isolation successful§	Number of isolates identified
VC476	VCF32	88	27.18	ND	+	1
VC506	VCF34	93	23.08	-	+	3
VC508	VCF34	93	23.53	-	+	4
VC522	VCF35	96	28.06	-	ND	0
VC554	VCF37	101	26.00	+	ND	2
VC59	VCF4	11	25.98	-	ND	0
VC835	VCF56	150	26.08	-	ND	0
VC849	VCF57	153	28.56	ND	+	3
VC864	VCF58	156	25.98	-	-	0
VC873	VCF59	157	25.30	-	ND	0
VC88	VCF6	17	27.58	ND	-	0
VC880	VCF59	158	27.57	ND	+	2
VC949	VCF64	168	25.99	-	ND	0

†: All enrichments were STEC O145 positive according to NeoSeek analysis

‡: C_t value recorded for serogroup-specific O145 RT-PCR

§: ND, not done

+, isolation successful

-, isolation unsuccessful

Appendix G - PM1 and PM2A MicroPlates™ carbon substrates

PM1 MicroPlate™ Carbon Sources

A1 Negative Control	A2 L-Arabinose	A3 N-Acetyl-D-Glucosamine	A4 D-Saccharic Acid	A5 Succinic Acid	A6 D-Galactose	A7 L-Aspartic Acid	A8 L-Proline	A9 D-Alanine	A10 D-Trehalose	A11 D-Mannose	A12 Dulcitol
B1 D-Serine	B2 D-Sorbitol	B3 Glycerol	B4 L-Fucose	B5 D-Glucuronic Acid	B6 D-Gluconic Acid	B7 D,L-α-Glycerol-Phosphate	B8 D-Xylose	B9 L-Lactic Acid	B10 Formic Acid	B11 D-Mannitol	B12 L-Glutamic Acid
C1 D-Glucose-6-Phosphate	C2 D-Galactonic Acid-γ-Lactone	C3 D,L-Malic Acid	C4 D-Ribose	C5 Tween 20	C6 L-Rhamnose	C7 D-Fructose	C8 Acetic Acid	C9 α-D-Glucose	C10 Maltose	C11 D-Melibiose	C12 Thymidine
D-1 L-Asparagine	D2 D-Aspartic Acid	D3 D-Glucosaminic Acid	D4 1,2-Propanediol	D5 Tween 40	D6 α-Keto-Glutaric Acid	D7 α-Keto-Butyric Acid	D8 α-Methyl-D-Galactoside	D9 α-D-Lactose	D10 Lactulose	D11 Sucrose	D12 Uridine
E1 L-Glutamine	E2 m-Tartaric Acid	E3 D-Glucose-1-Phosphate	E4 D-Fructose-6-Phosphate	E5 Tween 80	E6 α-Hydroxy Glutaric Acid-γ-Lactone	E7 α-Hydroxy Butyric Acid	E8 β-Methyl-D-Glucoside	E9 Adonitol	E10 Maltotriose	E11 2-Deoxy Adenosine	E12 Adenosine
F1 Glycyl-L-Aspartic Acid	F2 Citric Acid	F3 m-Inositol	F4 D-Threonine	F5 Fumaric Acid	F6 Bromo Succinic Acid	F7 Propionic Acid	F8 Mucic Acid	F9 Glycolic Acid	F10 Glyoxylic Acid	F11 D-Cellobiose	F12 Inosine
G1 Glycyl-L-Glutamic Acid	G2 Tricarballylic Acid	G3 L-Serine	G4 L-Threonine	G5 L-Alanine	G6 L-Alanyl-Glycine	G7 Acetoacetic Acid	G8 N-Acetyl-β-D-Mannosamine	G9 Mono Methyl Succinate	G10 Methyl Pyruvate	G11 D-Malic Acid	G12 L-Malic Acid
H1 Glycyl-L-Proline	H2 p-Hydroxy Phenyl Acetic Acid	H3 m-Hydroxy Phenyl Acetic Acid	H4 Tyramine	H5 D-Psicose	H6 L-Lyxose	H7 Glucuronamide	H8 Pyruvic Acid	H9 L-Galactonic Acid-γ-Lactone	H10 D-Galacturonic Acid	H11 Phenylethyl-amine	H12 2-Aminoethanol

PM2A MicroPlate™ Carbon Sources

A1 Negative Control	A2 Chondroitin Sulfate C	A3 α-Cyclodextrin	A4 β-Cyclodextrin	A5 γ-Cyclodextrin	A6 Dextrin	A7 Gelatin	A8 Glycogen	A9 Inulin	A10 Laminarin	A11 Mannan	A12 Pectin
B1 N-Acetyl-D-Galactosamine	B2 N-Acetyl-Neuraminic Acid	B3 β-D-Allose	B4 Amygdalin	B5 D-Arabinose	B6 D-Arabitol	B7 L-Arabitol	B8 Arbutin	B9 2-Deoxy-D-Ribose	B10 i-Erythritol	B11 D-Fucose	B12 3-0-β-D-Galacto-pyranosyl-D-Arabinose
C1 Gentiobiose	C2 L-Glucose	C3 Lactitol	C4 D-Melezitose	C5 Maltitol	C6 a-Methyl-D-Glucoside	C7 β-Methyl-D-Galactoside	C8 3-Methyl Glucose	C9 β-Methyl-D-Glucuronic Acid	C10 α-Methyl-D-Mannoside	C11 β-Methyl-D-Xyloside	C12 Palatinose
D1 D-Raffinose	D2 Salicin	D3 Sedoheptulosan	D4 L-Sorbose	D5 Stachyose	D6 D-Tagatose	D7 Turanose	D8 Xylitol	D9 N-Acetyl-D-Glucosaminitol	D10 γ-Amino Butyric Acid	D11 δ-Amino Valeric Acid	D12 Butyric Acid
E1 Capric Acid	E2 Caproic Acid	E3 Citraconic Acid	E4 Citramalic Acid	E5 D-Glucosamine	E6 2-Hydroxy Benzoic Acid	E7 4-Hydroxy Benzoic Acid	E8 β-Hydroxy Butyric Acid	E9 γ-Hydroxy Butyric Acid	E10 a-Keto-Valeric Acid	E11 Itaconic Acid	E12 5-Keto-D-Gluconic Acid
F1 D-Lactic Acid Methyl Ester	F2 Malonic Acid	F3 Melibionic Acid	F4 Oxalic Acid	F5 Oxalomalic Acid	F6 Quinic Acid	F7 D-Ribono-1,4-Lactone	F8 Sebacic Acid	F9 Sorbic Acid	F10 Succinamic Acid	F11 D-Tartaric Acid	F12 L-Tartaric Acid
G1 Acetamide	G2 L-Alaninamide	G3 N-Acetyl-L-Glutamic Acid	G4 L-Arginine	G5 Glycine	G6 L-Histidine	G7 L-Homoserine	G8 Hydroxy-L-Proline	G9 L-Isoleucine	G10 L-Leucine	G11 L-Lysine	G12 L-Methionine
H1 L-Ornithine	H2 L-Phenylalanine	H3 L-Pyroglutamic Acid	H4 L-Valine	H5 D,L-Carnitine	H6 Sec-Butylamine	H7 D.L-Octopamine	H8 Putrescine	H9 Dihydroxy Acetone	H10 2,3-Butanediol	H11 2,3-Butanedione	H12 3-Hydroxy 2-Butanone

Carbon substrates lists obtained from BIOLOG [96] (http://www.biolog.com/pdf/pm_lit/PM1-PM10.pdf).

Appendix H – Serogroup O145 strains analysed using the Omnilog phenotypic microarray system

Strain	MicroPlates™†	Serogroup	Source	Origin	Virulence profile	*eae* subtype
116B	PM1, PM2A	O145:H2	Bovine	Taranaki, New Zealand	*eae, ehxA*	ε
13ER3103A	PM1, PM2A	O145:HNM	Human	Auckland, New Zealand	*stx2, eae, ehxA*	γ
13ER4824	PM1	O145	Bovine	New Zealand	*stx2, eae, ehxA*	γ
13ER5640	PM1, PM2A	O145	Bovine	New Zealand	*stx2, eae, ehxA*	γ
13ER6723A	PM1+, PM2A	O145:H34	Human	Auckland, New Zealand	*stx2, eae*	ι
14ER2392	PM1, PM2A	O145	Bovine	New Zealand	*stx2, eae, ehxA*	γ
16ER0267A	PM1+, PM2A+	O145:H2	Human	Auckland, New Zealand	*stx1, eae, ehxA*	ε
16ER0517A	PM1+, PM2A	O145:H2	Human	Auckland, New Zealand	*stx1, eae, ehxA*	ε
188B	PM1	O145:H2	Bovine	Taranaki, New Zealand	*eae, ehxA*	ε
267P	PM1+, PM2A	O145:H2	Bovine	Taranaki, New Zealand	*eae, ehxA*	ε
54B	PM1+	O145:H2	Bovine	Taranaki, New Zealand	*eae, ehxA*	ε
AGR718	PM1‡	O145:H46	Bovine	Manawatu, New Zealand	*eae, ehxA*	γ
ERL020412	PM1‡	O145:H-	Human	New Zealand	*eae, ehxA*	γ
F5J	PM1	O145	Environmental	Waikato, New Zealand	*eae, ehxA*	γ
H12ESR01231	PM1+, PM2A	O145	Bovine	New Zealand	*eae, ehxA*	γ
H12ESR01387	PM1, PM2A	O145	Bovine	New Zealand	*stx2, eae, ehxA*	γ
H12ESR03525	PM1, PM2A	O145	Bovine	New Zealand	*stx2, eae, ehxA*	γ
P2B1	PM1+, PM2A	O145	Environmental	Waikato, New Zealand	*eae, ehxA*	γ
R249-1	PM1+, PM2A+	O145:H34	Human	Australia	*eae*	ι
Trh30	PM1, PM2A	O145:H-	Human	Norway	*eae, ehxA*	γ
Trh42	PM1+, PM2A	O145:H34	Human	Norway	*eae*	ι
Trh7	PM1, PM2A+	O145:H40	Human	Norway	*eae*	β
TW07865	PM1+, PM2A	O145:H28	Human	Germany	*stx2, eae, ehxA*	γ
VC1281m	PM1+, PM2A	O145	Bovine	Canterbury, New Zealand	*eae, ehxA*	γ

Strain	MicroPlates™†	Serogroup	Source	Origin	Virulence profile	*eae* subtype
VC1413m	PM1+, PM2A+	O145	Bovine	Southland, New Zealand	*stx2, eae, ehxA*	γ
VC1506m	PM1, PM2A	O145	Bovine	Southland, New Zealand	*eae, ehxA*	γ
VC308m	PM1+	O145	Bovine	Northland, New Zealand	*eae, ehxA*	γ
VC847m	PM1	O145	Bovine	Manawatu-Wellington, New Zealand	*eae, ehxA*	γ

†: +, MicroPlates™ were completed in replicate

‡: MicroPlates™ were completed in duplicate

Appendix I - Virulence factors identified from serogroup O145 whole genome sequence data in this study (n=53)†

Isolate‡	*stx1*	*stx2*	*ehxA*	*cba*	*astA*	*cif*	*efa1*	*espB*	*espC*	*espF*	*espI*	*espJ*	*espP*	*etpD*	*iha*	*iss*	*katP*	*mchB*	*mchC*	*mchF*	*mcmA*	*nleA*	*nleB*	*nleC*	*tccP*	*toxB*
116B	-	-	+	+	-	+	+	+	-	+	+	+	-	+	-	+	-	-	-	-	-	+	+	+	-	-
13ER3103A	-	+	+	-	+	+	-	+	-	-	-	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+
13ER4824	-	+	+	-	+	+	-	+	-	+	+	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+
13ER5056	-	+	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+
13ER5154	-	+	+	-	+	+	-	+	-	+	+	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+
13ER5640	-	+	+	-	+	+	-	+	-	+	+	+	+	-	+	+	+	-	-	-	-	+	+	+	+	+
13ER6227	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	+	+
13ER6723A	-	+	-	-	+	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
14ER2392	-	+	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+
15ER2679	-	-	+	-	+	+	-	+	-	+	+	+	+	-	+	+	+	-	-	-	-	+	+	+	+	+
16ER0267A	+	-	+	+	-	+	-	+	-	+	+	+	-	+	-	+	-	-	-	-	-	+	+	+	+	-
16ER0517A	+	-	+	+	-	+	+	+	-	+	+	+	-	+	-	+	-	-	-	-	-	+	+	+	+	-
188B	-	-	+	-	-	+	+	+	-	+	+	+	-	+	-	+	-	-	-	-	-	+	+	+	+	-
267P	-	-	+	+	-	+	+	+	-	+	-	+	-	+	-	+	-	-	-	-	-	+	+	+	+	-
54B	-	-	+	+	-	+	+	+	-	+	-	+	-	+	-	+	-	-	-	-	-	+	+	+	+	-
AGR718	-	-	+	-	+	+	+	+	-	-	-	+	-	+	+	-	-	-	-	-	-	+	+	+	-	-
ERL020412	-	-	+	-	-	-	+	+	-	+	-	-	-	+	+	-	-	-	-	-	-	+	+	+	-	-
ERL121829	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	+	+	+	+	+	+	+	-	+
ERL122034	-	-	+	+	-	+	+	+	-	+	+	+	-	+	-	+	-	-	-	-	-	+	+	+	-	-
F1	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	+	+
F5F	-	-	+	-	+	+	-	+	-	+	+	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+

Isolate‡	*stx1*	*stx2*	*ehxA*	*cba*	*astA*	*cif*	*efa1*	*espB*	*espC*	*espF*	*espI*	*espJ*	*espP*	*etpD*	*iha*	*iss*	*katP*	*mchB*	*mchC*	*mchF*	*mcmA*	*nleA*	*nleB*	*nleC*	*tccP*	*toxB*
F5J	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+
FDE21	-	-	+	-	+	+	-	+	+	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+
H12ESR01231	-	-	+	-	+	+	-	+	-	+	+	+	+	-	+	+	+	-	-	-	-	+	+	+	+	+
H12ESR01387	-	+	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+
H12ESR01650	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	+	+
H12ESR03525	-	+	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	+	+
P2A1	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	+	+
P2B1	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+
R249-1	-	-	-	-	+	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-
Trh30	-	-	+	-	+	+	-	+	-	-	-	+	-	-	+	+	+	-	-	-	-	+	+	+	+	+
Trh42	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	+	-	-	-	-	-	-	+	+	-	-
Trh46	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-
Trh7	-	-	-	-	-	+	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	+	-	-	-
TW07865	-	+	+	-	+	+	+	+	-	-	-	+	-	+	+	-	-	-	-	-	-	+	+	+	-	-
VC1048m	-	-	+	-	+	+	-	+	-	-	-	+	-	+	+	-	-	-	-	-	-	+	+	+	+	-
VC1056m	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	+	+
VC123n	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+
VC1281m	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+
VC1413m	-	+	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	+	+
VC1506m	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+
VC194m	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+
VC237m	-	-	-	-	+	+	-	+	-	-	+	+	-	-	+	+	-	-	-	-	-	+	+	+	+	-
VC237o	-	-	-	-	+	+	-	+	-	+	+	+	-	-	+	+	-	-	-	-	-	+	+	+	+	-
VC308m	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+

Isolate‡	*stx1*	*stx2*	*ehxA*	*cba*	*astA*	*cif*	*efa1*	*espB*	*espC*	*espF*	*espI*	*espJ*	*espP*	*etpD*	*iha*	*iss*	*katP*	*mchB*	*mchC*	*mchF*	*mcmA*	*nleA*	*nleB*	*nleC*	*tccP*	*toxB*
VC476m	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+
VC506m	-	-	+	-	+	+	-	+	-	+	+	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+
VC525m	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	+	+
VC554m	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+
VC847m	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+
VC849m	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	+	+
VC874o	-	-	+	-	+	+	-	+	-	-	-	+	+	-	+	+	+	-	-	-	-	+	+	+	-	+
VC880m	-	-	+	-	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	-	-	+	+	+	+	+

†: Virulence factors determined using the VirulenceFinder v1.5 from the CGE webserver [125].

‡: The genes *eae*, *tir*, *gad* and *espA* were present in all strains.

Appendix J - *E. coli* tRNA integration site for the locus for enterocyte effacement (LEE) pathogenicity island

Isolate	LEE insertion site†	*eae* subtype
Trh42	*Leu*‡	ι
16ER0267A	*pheV*	ε
ERL122034	*pheV*	ε
Trh30	*pheV*	γ
Trh7	*pheV*	β
VC847m	*pheV*	γ
116B	*pheV*	ε
188B	*pheV*	ε
267P	*pheV*	ε
54B	*pheV*	ε
16ER0517A	*pheV*	ε
13ER5154	*selC*	γ
13ER5056	*selC*	γ
TW07865	*selC*	γ
ERL020412	*selC*	γ
AGR718	*selC*	γ
Trh46	*selC*	ι
R2491	*selC*	ι
VC1281m	*selC*	γ
VC874o	*selC*	γ
VC1506m	*selC*	γ
FDE2/1	*selC*	γ
P2B1	*selC*	γ
F1	*selC*	γ
VC1048m	*selC*	γ
VC1056m	*selC*	γ
VC237m	*selC*	γ
VC237o	*selC*	γ
VC506m	*selC*	γ
VC880m	*selC*	γ
VC1413m	*selC*	γ
ERL121829	*selC*	γ
13ER6723A	*selC*	ι
13ER3103A	*selC*	γ
13ER5640	*selC*	γ
13ER6227	*selC*	γ
14ER2392	*selC*	γ
15ER2679	*selC*	γ
H12ESR01231	*selC*	γ
H12ESR03525	*selC*	γ
VC554m	*selC*	γ
VC506m	*selC*	γ

Isolate	LEE insertion site†	*eae* subtype
F5F	-	γ
F5J	-	γ
H12ESR01387	-	γ
H12ESR01650	-	γ
VC123n	-	γ
P2A1	-	γ
VC194m	-	γ
VC308m	-	γ
VC476m	-	γ
VC849m	-	γ
VC525m	-	γ
13ER4824	-	γ

†: -, LEE pathogenicity island integration site could not be precisely determined

‡: The LEE pathogenicity island integration site could not be precisely determined, however, the LEE was located near the tRNA *Leu* gene

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

Isolate‡	*stx1*	*stx2*	*ehxA*	*astA*	*cif*	*efa1*	*espB*	*espC*	*espF*	*espI*	*espJ*	*espP*	*etpD*	*iha*	*iss*	*katP*	*nleA*	*nleB*	*nleC*	*perA*	*tccP*	*toxB*
130322	-	+	+	+	-	-	+	-	+	+	+	+	-	+	+	+	+	+	+	-	+	+
132030	-	+	+	+	+	-	+	-	-	+	+	+	-	+	+	+	+	+	+	-	-	+
170303	-	-	-	+	-	-	-	+	+	-	-	-	-	-	-	-	-	+	-	-	-	-
173582	-	+	+	+	+	-	+	-	+	-	+	+	-	+	+	-	+	+	+	-	+	+
173758	+	-	+	+	+	-	+	-	+	-	+	+	-	+	+	+	+	+	+	-	+	-
182131	-	-	-	-	+	-	+	-	+	-	-	-	-	-	+	-	-	+	-	-	-	-
199816	-	-	-	-	+	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-
2010C-3507	-	+	+	+	-	-	+	-	-	+	+	+	-	+	+	-	+	+	+	-	+	+
2010C-3508	-	+	+	+	+	-	+	-	-	+	+	+	-	+	+	-	+	+	+	-	+	+
2010C-3509	-	+	+	+	+	-	+	-	+	+	+	+	-	+	+	-	+	+	+	-	+	+
2010C-3510	-	+	+	+	+	-	+	-	+	+	+	+	-	+	+	-	+	+	+	-	+	+
2010C-3511	-	+	+	+	+	-	+	-	+	+	+	+	-	+	+	-	+	+	+	-	+	+
2010C-3516	-	+	+	+	+	-	+	-	+	+	+	+	-	+	+	-	+	+	+	-	+	+
2010C-3518	-	+	+	+	+	-	+	-	-	+	+	+	-	+	+	-	+	+	+	-	+	+
2010C-3521	-	+	+	+	+	-	+	-	+	-	+	+	-	+	+	-	+	+	+	-	+	+
2010C-3526	-	+	+	+	+	-	+	-	+	+	+	+	-	+	+	-	+	+	+	-	+	+
2012C-4474	-	+	+	+	+	-	+	-	+	-	+	-	-	+	+	-	+	+	+	-	+	-
2012C-4477	-	+	-	+	+	-	+	-	-	-	+	-	-	+	+	-	+	+	+	-	+	-
2012C-4478	-	+	-	+	+	-	+	-	+	-	+	-	-	+	+	-	+	+	+	-	+	-
2012C-4479	-	+	-	+	+	-	+	-	+	-	+	-	-	+	+	-	+	+	+	-	-	-

Isolate‡	*stx1*	*stx2*	*ehxA*	*astA*	*cif*	*efa1*	*espB*	*espC*	*espF*	*espI*	*espJ*	*espP*	*etpD*	*iha*	*iss*	*katP*	*nleA*	*nleB*	*nleC*	*perA*	*tccP*	*toxB*
2012C-4480	-	+	-	+	+	-	+	-	+	-	+	-	-	+	+	-	+	+	+	-	+	-
201499	-	-	-	-	+	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-
203972	-	-	-	+	-	-	-	+	+	-	-	-	-	-	-	-	-	+	-	-	-	+
241761	-	+	+	+	+	-	+	-	-	+	+	+	-	+	+	+	+	+	+	-	-	+
241810	-	-	+	+	+	-	+	-	+	-	+	+	-	+	+	+	+	+	+	-	+	+
AA053	-	-	+	+	+	-	+	-	+	-	+	-	+	+	+	+	+	+	+	+	+	+
ewgs1003	-	+	+	+	+	-	+	-	-	+	+	+	-	+	+	-	+	+	+	-	+	-
FSIS1400369	+	-	+	+	+	-	+	-	-	-	+	+	-	+	+	-	+	+	+	-	-	+
FSIS1500788	-	-	+	+	+	-	+	-	+	-	+	+	-	+	+	+	+	+	+	-	+	-
FSIS1501198	+	-	+	+	+	-	+	-	+	-	+	+	-	+	+	+	+	+	+	-	+	+
FSIS1501717	+	+	+	+	+	-	+	-	+	-	+	+	-	+	+	+	+	+	+	-	+	+
FSIS1502535	+	+	+	+	+	-	+	-	+	-	+	+	-	+	+	+	+	+	+	-	-	-
FSIS1502550	+	+	+	+	+	-	+	-	+	-	+	+	-	+	+	+	+	+	+	-	+	+
FSIS1502554	+	+	+	+	+	-	+	-	-	-	+	+	-	+	+	+	+	+	+	-	+	+
FSIS1502976	+	+	-	+	+	-	+	-	-	-	+	-	-	+	+	-	+	+	+	-	+	-
FSIS1502978	+	+	+	+	+	-	+	-	+	-	+	+	-	+	+	+	+	+	+	-	+	-
FSIS1503305	+	-	+	+	+	-	+	-	+	-	+	+	-	+	+	+	+	+	+	-	+	-
FSIS1503307	-	+	+	+	+	-	+	-	+	-	+	+	-	+	+	+	+		+	-	+	-
FSIS1504619	+	-	+	+	+	-	+	-	+	-	+	+	-	+	+	+	+	+	+	-	-	+
FSIS1505314	-	+	+	+	+	-	+	-	-	-	+	+	-	+	+	+	+	+	+	-	-	+
FSIS1605419	-	+	+	+	+	-	+	-	+	+	+	+	-	+	+	+	+	+	+	-	+	+
FSIS1605420	-	+	+	+	+	-	+	-	+	+	+	+	-	+	+	+	+	+	+	-	-	+
MOD1-EC2002	+	+	+	+	+	-	+	-	+	-	+	+	-	+	+	+	+	+	+	-	+	+
MOD1-EC5842	-	-	-	+	+	-	+	-	-	+	+	-	-	-	+	-	+	+	+	-	+	-

Isolate‡	***stx1***	***stx2***	***ehxA***	***astA***	***cif***	***efa1***	***espB***	***espC***	***espF***	***espI***	***espJ***	***espP***	***etpD***	***iha***	***iss***	***katP***	***nleA***	***nleB***	***nleC***	***perA***	***tccP***	***toxB***
MOD1-EC6028	-	+	+	+	+	+	+	-	-	-	+	-	+	+	-	-	+	+	+	-	+	-
PNUSAE000756	-	+	+	+	+	-	+	-	+	+	+	+	-	+	+	+	+	+	+	-	+	+
PNUSAE001244	-	+	+	+	+	-	+	-	-	+	+	-	-	+	+	+	+	+	+	-	+	-

†: Virulence factors determined using the VirulenceFinder v1.5 from the CGE webserver [125].

‡: The genes *eae*, *tir, gad* and *espA* were present in all strains. The genes *cba*, *mchB*, *mchC*, *mchF* and *mcmA* were absent in all strains.