Niche-Adaptive Evolution in *Campylobacter jejuni*

LAURA MORLEY

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Doctor of Philosophy

September 2014

The conclusions and analyses presented in this thesis are my own.

The experimental work contained in this thesis is original research carried out by the author at Nottingham Trent University, unless otherwise stated.

Genomic sequences were produced by Konrad Paszkiewicz and the sequencing services team at the University of Exeter using DNA samples produced by the author at Nottingham Trent University.

Experiments presented in Chapter Six of this thesis were performed by Jukka Corander and his colleagues at the University of Helsinki, Finland. The output from experiments carried out by J. Corander was converted to a new format for analysis by Simon Harris at the Sanger Centre.

Abstract

Campylobacter jejuni is the leading causative agent of human bacterial gastroenteritis. Human *C. jejuni* infection (campylobacteriosis) is frequently associated with poultry; through consumption of undercooked products, cross contamination from raw meats, or through direct contact with birds or their faecal matter, however it is established that poultry is not the sole cause of *C. jejuni* infection in humans.

This research reveals new information on the MLST ST403 Clonal Complex, a previously identified *C. jejuni* lineage associated with the porcine host. ST403CC *C. jejuni* have also been linked with other mammalian hosts to a lesser degree, and have been implicated in human campylobacteriosis, however to date this clonal complex has not been linked to poultry. The original hypothesis of this research predicted that due to sharing a host niche commonly associated with *C. coli*, the porcine ST403CC may show evidence of increased recombination with *C. coli*, however this was not observed.

Six ST403CC isolates of porcine origin were subjected to phenotype testing and whole genome sequencing; these isolates were capable of invasion *in vitro*, and were revealed both to have acquired seemingly lineage specific content, in the form of Restriction-Modification (R-M) system associated genes, and to have undergone degredation of certain loci. The ST403CC isolates also exhibited a distinct pattern of reduced genomic recombination compared to non-ST403CC *C. jejuni*, with evidence of lineage specific recombination events.

Both generalist & specialist lineages have previously been revealed in *C. jejuni*. The research presented here identifies a new specialist lineage which is associated with mammalian hosts, and not found in poultry.

Acknowledgements

Although this thesis is my work, there are several people without whom it would never have been produced, so I am pleased to take this opportunity to acknowledge their support.

Thanks to my supervisor, Dr Gina Manning, for her patience, kindness, support and dedication over these years, not to mention my gratitude for providing me with this fantastic opportunity. Also, my second supervisor Dr Alan McNally, for inspiration and assistance, and probably having more confidence in my ability than I did.

Thanks to the NTU PhD crew – without you this would have been a dull affair! Thanks to all from the micro lab, cels211 and our research group, I'll miss you all. Especially Miquette, for keeping things (me) crazy, and Alya – cheers for all the giggles!

I also would like to give a special mention to former NTU micro staff members Pam Horne and Mike Brice, who were there for me from undergraduate day one, and whose kindness I will never forget.

Of course I must give special thanks to my family: my fabulous Mum - thank you so much - you have always stood by me through thick and thin; and the wonderful Sue and Dennis, who have truly accepted me; each of whom have given so much to help me succeed.

And to my husband, Vince. Without you I would never have got this far. Your support and belief has been beyond compare – from the endless coffees and hugs to keeping me going when things got too much for me, thank you for everything.

Finally, I dedicate this thesis to Dad. I know if you were still here you'd have great fun teasing me over this big daft book, whilst internally bursting with pride. As you once put it:

"work hard, play hard, and most of all be happy"

I think I'm getting there, old man.

This research was funded through Nottingham Trent University

Contents

Abbreviations

BLAST	Basic Local Alignment Search Tool
bp	Base Pair
BRIG	BLAST Ring Image Generator
CC	Clonal Complex
CDS	Coding Sequence
cfu/ml	Colony Forming Units per Millilitre
CJIE	*Campylobacter jejuni* Integrated Element
DNA	Deoxyribonucleic Acid
DMEM	Dulbecco's Modified Eagle's Medium
EDGAR	Efficient Database framework for comparative Genome Analyses using BLAST score Ratios
GBS	Guillain Barré Syndrome
G:C	Guanine:Cytosine Base Content
GI	Genomic Island
HGT	Horizontal Gene Transfer
HPA	Health Protection Agency, now replaced by PHE
IS	Insertion Sequence
kbp	Kilobase Pair
LPS	Lipopolysaccharide
LOS	Lipooligosaccharide
MCCDA	Modified Charcoal Cefoperazone Deoxycholate Agar
MIC	Minimum Inhibitory Concentration
MLST	Multilocus Sequence Typing
MOI	Multiplicity of Infection
NCTC	National Collection of Type Cultures
OD	Optical Density
PBS	Phosphate Buffered Saline
PHAST	Phage Search Tool
PHE	Public Health England
SNP	Single Nucleotide Polymorphism
ST	Sequence Type
SVR	Short Variable Region
v/v	Volume per Volume
w/v	Weight per Volume

List of Tables

List of Figures

Chapter One: Introduction

1.1 History of *Campylobacter*

The bacteria we now know as *Campylobacter* have a complicated history, as perspectives have changed with the development of new techniques and technologies. In 2005, Moore *et al* produced a review which described the history of *Campylobacter* as a group of emerging pathogens; briefly, these organisms were first recognised as animal pathogens during the early twentieth century, and later - around the 1950s - begun to be occasionally identified from human blood cultures and were viewed as occasional opportunistic human pathogens. However, in later years, the development of improved isolation techniques led to the discovery that these bacteria in fact represent a substantial cause of human disease; most commonly infectious gastroenteritis. As such there are two important historical perspectives when considering *Campylobacter*; firstly the history of its relation to disease, and secondly the history of the nomenclature applied to *Campylobacter*.

1.1.1 History of *Campylobacter* Association with Disease

Campylobacters were first observed in the early 20th Century, and were initially described as '*Vibrio*-like species' (Debruyne, Gevers & Vandamme in Nachamkin, Szymanski & Blaser (Eds), 2008). These '*Vibrio*-like' bacteria were first recognised as veterinary pathogens; associated with abortion in sheep (McFadyean & Stockman, 1913 - cited by Vandamme & Goossens, 1992) and cows (Smith, 1918 - cited by Vandamme & Goossens, 1992), and were subsequently described as *'Vibrio fetus'* by Smith & Taylor (1919 - cited by Vandamme & Goossens, 1992). Later, these '*Vibrio*-likes' were also linked to diarrhoeal infection in cattle (Smith & Orcutt, 1927; Jones, Orcutt & Little, 1931) and were then referred to as *Vibrio jejuni*. In 1944 another '*Vibrio*-like species' was identified by Doyle in relation to diarrhoea in pigs, and later became known as *Vibrio coli* (Doyle, 1948 - cited in Vandamme & Goossens, 1992). Also during the 1940s, these '*Vibrio*-like species' began to be sporadically isolated from human cases, via body fluid samples, however it was as late as the 1970s when *Campylobacter* was first successfully isolated from faecal samples as described in the review by Butzler (2004).

The first significant work on human infection with the '*Vibrio*-like' later known to be *Campylobacter jejuni* or *C. coli* was that carried out by King in 1957. Due to the limitations in culture and isolation techniques, these isolates were from blood samples but were intellectually linked with diarrhoeal symptoms expressed by the patients; at this time culture from stools was unsuccessful, with overgrowth on plates due to out-competition by coliforms on growth media. King (1957) observed a distinct group of 'related *Vibrios*' with a higher optimal growth temperature (42°C), some biochemical differences, and a distinct diarrhoeal clinical presentation compared to previously described animal *Vibrio fetus* isolates. The newly observed human blood isolates thought to be *V. fetus* were morphologically indistinguishable from their assumed animal counterparts, and showed similar microaerophilic growth characteristics, however, the animal isolates grew best between 25-37°C with no or very little growth at 42°C whereas the human isolate 'related vibrios' grew at 37-42°C with optimal growth at 42°C and no or very little growth at 25°C. Since then the prevalence and clinical effects of *Campylobacter* has been increasingly studied, particularly due to the development of filtration techniques and selective procedures to isolate and culture *Campylobacters* from stool samples by Butzler *et al* (1973). These new techniques had a huge impact on the future study of human *Campylobacter* infections, and, for example, led to the work of Skirrow (1977), who observed that 7.1% (57) of 803 diarrhoeic patients tested positive for *C. jejuni* or *C. coli*, with zero observed from 194 non-diarrhoeic control individuals. The highest incidence of *Campylobacter* was observed in children (Skirrow, 1977), and all individuals who tested positive for *Campylobacter* described distinctive symptoms including abdominal pain. Skirrow (1977) demonstrated the link between *Campylobacter* presence and enteritis with diarrhoea and abdominal cramps, and drew attention to *Campylobacter* as a likely overlooked cause of enterocolitis.

A subsequent study (Skirrow & Benjamin, 1980) also supported the early work of King (1957), whereby growth at 25°C provided a clear cut-off between *V. fetus* and the 'other related *Vibrios*' (now identified as *C. jejuni*/*C. coli*) and other NARTC isolates (nalidixic acid resistant thermophilic *Campylobacter* isolates), however they found that distinguishing between *C. jejuni* and *C. coli* was less accurate using

temperature and biochemical tests. Skirrow & Benjamin (1980) also observed associations between *C. jejuni* and cattle and *C. coli* and pigs, however, poultry strains were found to exhibit 'intermediate' characteristics, and human isolates were observed which represented both groups, but more commonly resembled *C. jejuni*.

From the 1970s-1980s it was increasingly observed and accepted that the previously unculturable *Campylobacter* isolates in stools were a large cause of human gastroenteritis; these were a distinct group from animal pathogens with a higher optimal growth temperature despite morphological similarity. To develop the full picture this progression must be paired with the advances in systematics of *Campylobacter.*

1.1.2 Nomenclature of *Campylobacter*

Linked with the general difficulties in culturing and identifying Camyplobacters, similar complexity was encountered in the accurate nomenclature for these organisms. As described previously the organism was first known as '*Vibrio*' in the early 20th century; before being expanded and eventually recognised as a new genus. These bacteria were initially described as '*Vibrio*-like' species around 1910 with the subsequent addition of species such as *V. fetus* (Smith & Taylor, 1919 – cited in Vandamme & Goossens, 1992), *V. jejuni* (Smith & Orcutt, 1927; Jones, Orcutt & Little, 1931) and *V. coli* (Doyle, 1948 in Vandamme & Goossens, 1992). The genus *Campylobacter* was introduced in 1963 by Sebald and Véron, (cited in Debruyne, Gevers & Vandamme, in Nachamkin, Szymanski & Blaser (Eds), 2008), due to the unique characteristics of these isolates, such as their lack of fermentation, microaerophily, and low GC content. *Campylobacter* initially contained two of the previously described 'related *Vibrio'* species; *V. fetus* and *V. bubulus*, now designated *C. fetus* and *C. bubulus*, with *C. fetus* as the type species and the new genus placed within the family *Spirilliaceae* (Debruyne, Gevers & Vandamme, in Nachamkin, Szymanski & Blaser (Eds), 2008; Sebald & Véron, 1963). This was subsequently expanded a decade later by Véron & Chatelain (1973), based upon biochemical and serotyping tests combined with assessment of GC content, to include the transfer of four '*Vibrio*-like' species (*Vibrio coli*, *V. jejuni*, *V. sputorum*, *V.*

bubulus) to the genus *Campylobacter* previously created by Sebald & Véron (1963) as the new species *C. coli*; *C. jejuni* and *C. sputorum* (containing subspecies *C. sputorum sputorum* and *C. sputorum bubulus*), as well as the separation of the previously described type species *C. fetus* into two subspecies; *C. fetus fetus* and *C. fetus venerealis*. Numerous revisions of different species have been undertaken over the subsequent decades, and in 1991 Vandamme and De Ley introduced the family *Campylobacteraceae* to contain all species of *Campylobacter* and *Arcobacter*.

1.1.2.1 Determining Relatedness & Taxonomy

Determining the relationships between bacterial species has been a topic of consideration since microorganisms first began to be observed and understood. The systems for determining these relationships however have been turbulent, with frequent changes to classifications being necessary (Ludwig & Klenk, in Garrity (Ed), 2001). Initial methods for determining bacterial systematics relied upon observable physical characteristics such as morphology, motility and biochemical profiles, and in some cases these initial determinants remain relative; however in many cases the advent and constant improvement of molecular and genomic methods have altered judgements significantly. Early molecular or genomic methods such as comparing G:C nucleotide content, or analysis of cell wall or lipid content frequently provided better distinction. Later (around the 1960s), DNA-DNA hybridisation became the process of choice for determining systematic relationships, providing not just differentiating information as the earlier methods but also a direct means to interpret phylogeny, with the accepted standard being that organisms exhibiting greater than 70% hybridisation represent the same species (Ludwig & Klenk, in Garrity (Ed), 2001). Nucleotide sequencing became increasingly valuable, with the main example being the use of small subunit rRNA as a measure of relatedness.

1.1.2.2 Taxonomy of *Campylobacter*

In their review of *Camyplobacter* taxonomy, Vandamme & Goossens (1992) described that in the eighth Bergey's manual (in 1974), *Campylobacter* was grouped with the genus Spirillum into the family Spirillaceae, until in 1984 when the family Spirillaceae was discontinued and the orphaned groups including *Campylobacter*

were placed into a collection of unassigned taxa which showed some resemblance in physiology or morphology. Subsequently, in 1987 Romaniuk *et al* used partial rRNA sequencing of six *Campylobacter* strains and determined that they represented a distinct group of eubacteria; as described by Vandamme & Goossens (1992) these were later demonstrated to be a subgroup of what is now known as the *Proteobacteria*.

In the 1980s rRNA cistron similarity became the new vogue for determining bacterial classification. This in part led to the separation in 1989 of the then *C. pylori* and *C. mustelae* to the novel genus *Helicobacter* (*H. pylori* and *H. mustelae*). This marked the beginning of a period of significant study and reorganisation in the *Campylobacters* throughout the 1990s. This was followed by a large scale DNA-rRNA hybridisation study by Vandamme *et al* (1991) – leading to the description of what is now known as the epsilon sub-division of *Proteobacteria*.

Vandamme *et al*'s (1991) study of DNA-rRNA hybridisation determined that campylobacters and their relatives were a separate phylogenetic branch; rRNA superfamily VI. Several groups already exist within the *Eubacteria* including the *Proteobacteria.* Already characterised within the *Proteobacteria* were the alpha, beta, gamma and delta subclasses. Vandamme & De Ley (1991) observed that rRNA superfamily VI containing the *Campylobacters* and related organisms represents a separate *Proteobacteria* subclass, designated the epsilon *Proteobacteria*. Based upon rRNA-DNA hybdridisation, they proposed that *Campylobacter* and *Arcobacter* were within the new genera *Campylobacteraceae* with *Helicobacter*, *Wolinella* and other related organisms being closely related (superfamily VI) but distinct and excluded from the genera. The epsilon *Proteobacteria* consists of three clusters; cluster one consisting of *Bacteroides gracilis; B. ureolyticus; C. coli*; *C. concisus*; *C. fetus*; *C. hyointestinalis*; *C. jejuni*; *C. lari*; *C. mucosalis*; *C. sputorum*; *C. upsaliensis*; *Wolinella curva*; and *W. recta*. Cluster two consisting of 'aerotolerant *Campylobacter*-like organisms' which were reclassified as the novel genus *Arcobacter,* whilst the third cluster contained *C. cinaedi*; *C. fennelliae*; *Flexispira rappini*; *Helicobacter mustelae*; *H. pylori*; *W. succinogenes* and the unnamed 'Campylobacter-like-organism-3' isolate'. Due to the determination of these

clusters they proposed renaming *W. curva* and *W. recta* as *C. curvus* and *C. rectus*, and *C. cinaedi* and *C. fennelliae* as *Helicobacter cinaedi* and *H. fennelliae*.

In 1991 Vandamme & De Ley introduced the family *Campylobacteraceae*, comprising the genera *Arcobacter* and *Campylobacter*, within the *Proteobacteria* subgroup described previously. In a subsequent study *B. gracilis* and *B uerolyticus* were amended to the novel species *C. gracilis* (Vandamme *et al,* 1995), with the effect that the genus *Campylobacter* could be described as anything sharing rRNA 'cluster one' homology with the type species *C. fetus*.

rRNA studies in the early nineties led to reclassification and improved understanding of *Campylobacters*; the *Campylobacter* genus species must have homology at rRNA hybridization level with the type species, *C. fetus* (Debruyne, Gevers & Vandamme, in Nachamkin, Szymanski & Blaser (Eds), 2008). At this time the genus *Arcobacter* was introducted, for rRNA 'cluster two'; whilst two more previous 'campylobacters' were reassigned as *Helicobacters*, this time *H. cinaedi* and *H. fennelliae*. Vandamme *et al* (1991) further 'finessed' the categories and members of *Campylobacter* and related organisms, introducing genus *Arcobacter* and rearranging some species, as described above. The genera *Campylobacter and Arcobacter* were assigned to a new family, *Campylobacteriaceae*; which also contains *Sulfospurilium*, *Wolinella* and *Helicobacter* (Vandamme & De Ley, 1991; Debruyne, Gevers & Vandamme, in Nachamkin, Szymanski & Blaser (Eds), 2008)

1.2 The Clinical Significance of *Campylobacter*

Species of *Campylobacter* can be subdivided into two groups which are often referred to as the thermophilic and nonthermophilic groups. In the literature this is often used interchangeably with the alternative term thermotolerant campylobacters, which is more appropriate as true thermophily is characterised by growth/survival at higher temperatures, such as a growth range between approximately 20-70°C and an optimum temperature for growth usually in excess of 50°C (Madigan & Martinko, 2006).

Non-thermotolerant Campylobacters grow optimally in the range of 25-37°C, with little to no growth observed at 42°C; whereas thermotolerant *Campylobacter* species display optimum growth at 42°C, and strong growth also at 37°C but with little to no growth occurring at 25°C (King, 1957; Skirrow & Benjamin, 1980; Butzler, 2004).

1.2.1 Human Infections with *Campylobacter*

Both non-thermotolerant (*C. fetus fetus*; *C. fetus venerealis*) and thermotolerant (*C. jejuni*; *C. coli*; *C. upsaliensis*) campylobacters are potential human pathogens (Penner, 1988).

C. fetus fetus causes spontaneous abortion in sheep and cattle, whilst *C. fetus venerealis* causes infective infertility in cattle (Lastovica & Allos, in Nachamkin, Szymanski & Blaser (Eds), 2008; Florent, 1959 - cited in Vandamme & Goossens, 1992). Moore *et al* (2005) described that *C. fetus* is a potential but rare cause of gastroenteritis in humans, but is more commonly associated with systemic infections – typically in patients who are already unwell, as a complication. Human *C. fetus* infection can include fever and metastatic localisation, and has an estimated mortality rate of 15%. *C. fetus fetus* can cause diarhhoeal disease in humans, which is described by Lastovica & Allos (in Nachamkin, Szymanski & Blaser (Eds), 2008) as similar to *C. jejuni* infection. Penner (1988) stated that *C. fetus fetus* infections typically include septicaemia and occasionally meningitis occurring in patients with illness already presenting. Human foetal infections can occur but are thought to be rare.

Thermotolerant *Campylobacter* species, including *C. jejuni*, *C. coli* and *C. lari*, commonly cause infectious gastroenteritis in humans. As described in Butzler's review article (2004), several *Campylobacter* species are known to cause infective diarrhoea in humans (including *C. coli, C. concisus*, *C. fetus fetus*, *C. hyointestinalis*, *C. jejuni doylei, C. jejuni jejuni*, *C. lari*, *C. upsaliensis*) however *C. jejuni jejuni* is by a clear margin the most commonly associated *Campylobacter* with human gastroenteritis. Due to the similarities between infection and characteristics of *C. jejuni* and *C. coli*, many epidemiological studies do not differentiate between the two, either considering both as one group, or including *C. jejuni* only – and many

labs do not routinely carry out speciation between the two (Siemer, Nielsen & On, 2005).

1.2.2 Clinical Presentations

The most common presentation of human *Campylobacter* infection is a typical gastroenteritis illness including diarrhoea, abdominal cramps and fever (Butzler, 2004) which is often referred to as Campylobacteriosis. Infection with *C. jejuni* or *C. coli* causes, normally, an unpleasant but self-limiting diarrhoeal disease which only requires treatment in severe or unusual cases. Symptoms of infection are not distinguishable from other common gastroenteral pathogens without the need for laboratory confirmation. Campylobacteriosis typically has an incubation period of between two and five days before symptoms are presented, although in some cases the incubation period can last as much as ten days (Butzler, 2004; Moore *et al*, 2005). Often, fever, abdominal pain and general malaise begin before the onset of diarrhoeal symptoms, which then typically lasts for around forty-eight to seventy-two hours, after which point the other symptoms may persist longer (Butzler, 2004). Stools often contain blood, pus and or mucus due to the inflammation caused by the infection.

Symptoms develop generally between two and three days following the consumption of contaminated matter and the majority of cases are self-limiting, usually within a week (Moore *et al*, 2005); however, duration and severity of *Campylobacter* infection can vary greatly, and infection is often associated with fever and painful cramps (Zilbauer *et al*, 2008) and post infection complications can occur, and may be severe (Grant, Woodward & Maskell, 2006). The invasive ability of *Campylobacter* means that bacteraemia (bacteria in blood) is a potential complication. Rarely, *C. jejuni* infection can lead to bacteraemia; most commonly in patients who are already immunocompromised (Butzler, 2004). Other potential but uncommon complications of *C. jejuni* infection include reactive arthiritis, or other extra-intestinal infections such as meningitis or osteomyelitis (Butzler, 2004). The most serious risk post infection with *C. jejuni* is the development of autoimmune mimicry disease, including GBS (Guillain-Barré Syndrome). Other infections can be

caused if *C. jejuni* spreads from the site of infection (typically the jejunum, ileum or colon) into other regions of the gastrointestinal tract, or extraintestinal spread to other organs, even causing bacteraemia in some patients (in rare instances). Occasionally *C. jejuni* infection can lead to serious systemic illness which can end in sepsis and death, however these cases are very rare and *C. jejuni* has very low mortality rates (Allos, 2001).

GBS is the most important potential complication of *Campylobacter* infection. GBS is a form of flaccid paralysis which is caused by the molecular mimicry of a *C. jejuni* surface antigen of the human GM1 epitope which is expressed on motor neurones; the induced anti-GM1 antibody causes autoimmunity leading macrophages to enter and destroy the motor nerve axon, creating the flaccid paralysis exhibited in sufferers (Yuki, 2001). GBS incidence is low (estimated at less than 1 per 1000 *C. jejuni* cases), however, with the imminent elimination of poliomyelitis, it is now the most frequent cause of acute flaccid paralysis (Allos, 2001; Yuki, 2001). GBS can be associated with specific Penner scheme serotypes, however there is no observed link between the severity of the initial *C. jejuni* infection and the subsequent development of GBS (Allos, 2001; Yuki, 2001).

The symptoms of *C. jejuni* and *C. coli* infection are not so distinct that the organism can be differentiated from other gastroenteritis causing bacteria in clinical cases without microbiological investigation (Butzler, 2004). In the clinical setting, *C. jejuni* infection is indistinguishable not only from *C. coli* infection but also those caused by other bacterial diarrhoeal pathogens such as *Yersinia*, *Salmonella* or *Shigella* species as it typically manifests as a short-term illness with fever cramps and diarrhoea (Allos, 2001).

1.2.3 Treatment & Antimicrobial Resistance in *Campylobacter*

Generally campylobacteriosis is self-limiting and symptoms are resolved within two weeks, although bacteria may still be shed in faeces for up to eight weeks. The most common treatment required for Campylobacteriosis is simply the provision of hydration and electrolytes if necessary to prevent dehydration, however, if a patient is immunocompromised or otherwise in ill health, or if symptoms are

extreme or persist for over a week, antimicrobial chemotherapy is required (Butzler, 2004). Generally the decision to use antibiotics should not be taken unless special circumstances require it; such as severe or prolonged illness or for patients who are HIV positive, are currently pregnant, or suffer immunocompromising disorders (Allos, 2001). Where possible, antimicrobial treatments are avoided for *Campylobacter* infection, as Campylobacters are known to be naturally resistant to some antibiotics and have been demonstrated to acquire resistance to normally effective antibiotics under selective pressure. Additionally antibiotic treatments should be avoided where possible not just due to the risk of increasing antimicrobial resistance in *Campylobacter* but also in case of misdiagnosis (*E. coli* O157:H7 can cause a similar bloody diarrhoea, for example) and the risk of increasing antimicrobial fitness in other pathogens or in opportunistic commensals (Allos, 2001).

Traditionally, fluoroquinolones were commonly used to treat *C. jejuni* infection (Butzler, 2004; Allos, 2001), as they could be used freely without determining the specific cause; as a general treatment for bacterial gastroenteritis caused by organisms such as *Campylobacter*, *Shigella*, *Salmonella* (Allos, 2001) however this has led to an increase in fluoroquinolone resistance in *Campylobacters*. An additional factor in the prevalence of fluoroquinolone resistance in *Campylobacters* is the use of certain fluoroquinolones in animal husbandry in the UK, USA and Europe (Allos, 2001; Butzler, 2004).

Nachamkin, Ung & Li (2002) reviewed the issue of antibiotic, specifically fluoroquinolone, resistance in *C. jejuni*. Resistant strains of *C. jejuni* were first identified in the 1980s in Europe, and were linked with the use of fluoroquinolones in animal sources. They also cite research in the USA (Minnesota) showing that over a six year period from 1992 to 1998 *C. jejuni* fluoroquinolone resistance expanded from 1.3% to 10.2% (Smith *et al*, 1999; Nachamkin, Ung & Li, 2002). Their research found that although the presence of erythromycin resistance in tested *C. jejuni* isolates remained relatively consistent across the period between 1982 and 2001 (around 2-5%) whereas resistance to ciprofloxacin grew from around 10% in 1996 to approximately 40% in 2001. Nachamkin, Ung & Li (2002) also found that

when MIC (Minimum Inhibitory Concentrations) were taken into consideration, although the majority of isolates were inhibited by a low ciprofloxacin concentration (≤0.250µg/ml), the resistant isolates had high levels of resistance, with MICs in excess of ≥32ug/ml. Additionally, in their study, foreign travel seemed to be associated with acquisition of fluoroquinolone resistant *C. jejuni*. In a similar study to that by Nachmkin, Ung & Li (2002), Unicomb *et al* (2003) observed that fluoroquinolone resistance, despite being highly prevalent elsewhere, was not a significant issue in Australia. Only twelve of 370 *C. jejuni* isolates from human cases were found to be resistant to fluoroquinolones, and of these ten were demonstrated to be associated with foreign travel, although the other two could not be assigned, resistance was not observed in the 358 isolates which were locally acquired. They suggest the probable link between the lack of *C. jejuni* fluroquinolone resistance and the lack of use of these in food production animals. A study by Gaunt & Piddock (1996) demonstrated that 4.1% of 2209 isolates taken in 1991 in the UK were resistant to ciprofloxacin in lab tests, although none of these patients had been treated with any quinolones, 33% of them had recently travelled abroad. This suggested that the use of quinolones to treat broiler flocks has created a selective pressure for quinolone resistance in *Campylobacter*. Smith *et al* (1999) reported a dramatic increase (from 1.3% to 10.2%) in quinolone resistant *C. jejuni* isolates in Minnesota from 1992 to 1998. They observed that quinolone treatment could only be associated with about 15% of resistant cases, and observed that although foreign travel was associated with resistant isolates, this explained fewer cases as time went on, with 1996-1998 resistance being increasingly associated with domestic infections. They confirmed a link between chicken products and domestic infections.

Erythromycin is now the recommended treatment for *C. jejuni* infection, where necessary (Allos, 2001; Butzler, 2004; others). Modern macrolides (such as clarithromycin and its relatives) have been demonstrated to also represent an effective treatment against *Campylobacter*, however their use is not recommended due to the extra cost incurred with no improved success rates over cheaper treatments (Nachamkin, Ung & Li, 2002).

1.3 The Burden of *Campylobacter*

Skirrow (1977) suggested that if the results of the first major study of *C. jejuni* and *C. coli* in diarrhoeic patients could be extrapolated it was reasonable to suggest that *Campylobacter* may be the biggest single cause of infectious bacterial enterocolitis; this has subsequently been demonstrated in modern studies and *Campylobacter* is now accepted as the leading cause of bacterial gastroenteritis worldwide, causing more cases than *Salmonella* spp, *Shigella* spp or *E. coli* O157:H7 (Allos, 2001; Moore *et al*, 2005). According to the HPA (Health Protection Agency), in 2010 there were over 62000 reported incidences of *Campylobacter* infection in England and Wales; compared to just more than 9000 recorded cases of *Salmonella* in the same period. In a recent review of the levels of human *Campylobacter* infections in the UK (Strachan & Forbes, 2010) there were 64000 reported cases of *Campylobacter* gastroenteritis in England, Scotland and Wales combined during 2009 (Pollock *et al*, 2009, and anonymous HPA reports, 2010 - cited in Strachan & Forbes, 2010), however, under-reporting is a significant issue for *Campylobacter* and for gastroenteritis generally, the real figures for this period may be closer to 450000 (Wheeler *et al*, 1999; Strachan & Forbes, 2010). There have been fluctuations in the number of confirmed cases reported by the HPA in England & Wales since they began screening for *Campylobacter*, however the general trend across the studied time is an increase (from approximately 34 thousand in 1989 to approximately 62 thousand in 2010). In a study by Wheeler *et al* (1999), in England, *Campylobacter* species were the largest bacterial agent of infectious intestinal disease in the community and in general practice. In general, one in five people are affected by some kind of infectious intestinal disease every year; of these approximately one in six will report to a GP.

Although *Campylobacter* infection is not normally serious and does not normally require treatment, the sheer number of cases places a considerable burden on the workplace and economy through sick days. According to Strachan & Forbes (2010) around 10% of human *Campylobacter* cases in England, Wales and Scotland require hospitalisation. As stated previously it is also important to remember the large degree of un-reported cases meaning the burden is larger than predicted. Also the

expense of the rare complications is still considerable. Outbreaks of *Campylobacter* are uncommon (or at least, identifying them is), with the majority being sporadic cases (Janssen *et al*, 2008). Where recognised outbreaks do occur they are often associated with water or with raw milk products.

Cases of *Campylobacter* associated illness can occur at any age, but *C. jejuni* is particularly a cause of morbidity in children and in travellers (particularly holiday makers from 'developed countries' to less industrialised areas) (Allos, 2001; Crushell *et al*, 2004). *Campylobacter* incidence is also more common in rural areas, especially in children. This has been demonstrated in numerous studies, including those by: Ethelberg *et al* (2005) who carried out a longitudinal study over an eleven year period in Denmark and observed associations with low population density and rural housing and *Campylobacter* infection, especially in children; Green, Krause & Wylie (2006), who observed higher incidence of *Campylobacter* infection both in children, and in rural areas, in a Canadian province with both rural, agricultural and urban regions; Strachan *et al* (2009), who demonstrated that children under five years in rural areas in NorthEast Scotland had a higher prevalence of campylobacteriosis than those in urban areas; and Strachan *et al* (2013a) revealed that, in Scotland, children in rural areas specifically exhibited increased loads of non-chicken associated campylobacteriosis cases in the late spring period.

1.4 Sources of Campylobacteriosis

Despite the large volume of human infections, humans are not in fact a natural host or reservoir for thermotolerant *Campylobacter*. Thermotolerant *Campylobacter* is considered an accidental pathogen of humans and is typically a zoonotic infection – transmitted to humans from an animal host. In this section a brief review is prevented of the potential sources for human *Campylobacter* infection.

1.4.1 Common Sources of *Campylobacter*

The largest recognised source for human campylobacteriosis is poultry, and in particular, chickens (Janssen *et al*, 2008; Moore *et al*, 2005; Strachan & Forbes, 2010), with cases attributed to direct handling of poultry, contact or contamination

with poultry faeces, handling or preparation of meats, or consumption of inadequately cooked poultry products, or of foods which have become cross-contaminated (Janssen *et al*, 2008). A unique example evidencing the link between poultry meats and human *C. jejuni* (*C. jejuni*/*C. coli*) infection was created by the Dioxin crisis which occurred in Belgium during 1999. In this instance, contamination of livestock foodstuffs with Dioxin led to the removal from sale of all Belgian poultry products within Belgium and during this period where poultry was unavailable for consumption a 40% drop in documented *Campylobacter* infections was observed by the routine surveillance programme. The number of cases returned to the 'normal' level once poultry products returned to the shelves (Vellinga & Van Loock, 2002).

1.4.2 *Campylobacter* in Food Production Birds & Animals

Food production mammals, including cattle, pigs and sheep, are a known source of *Campylobacter* infection although the pervasiveness and human importance of these sources is debated in the literature (Moore *et al*, 2005). Cattle have been thought to have carriage levels anywhere between 0-80% (Atabay & Corry, 1998; Moore *et al*, 2005), sheep around 20% carry Campylobacters (Zweifel, Zychowska & Stephan, 2004; Moore *et al*, 2005), and pigs have accepted high incidence of *Campylobacter*, although this may be associated more with the nature of pig carcass processing than a truly higher carrier rate in the gut (Nesbakken *et al*, 2003; Moore *et al*, 2005). When Nesbakken *et al* (2003) investigated carriage of *Campylobacter* spp. in pigs at slaughter in Norway they observed 100% carriage in the gastrointestinal tract. *C. coli* is more commonly associated with pigs than is *C. jejuni* and pigs tend to carry *Campylobacters* more than sheep or cattle (Franco, 1988, cited by Nesbakken *et al*, 2003). As described by Nesbakken *et al* (2003), the levels of *Campylobacter* carriage in pigs varies considerably in literature from as low as 2.9% in Poland found by Kwiatek, Wolton & Stern (1990, cited in Nesbakken *et al*, 2003) to as much as 95% in Sweden by Svedhem and Kaijser (1981, cited in Nesbakken *et al*, 2003). Nesbakken *et al* (2003) found mostly *C. coli* in pigs by some margin, but a few examples of *C. jejuni* and *C. lari* were also observed.

Pezzotti *et al* (2003) investigated the occurrence of *C. jejuni* and *C. coli* from rectal swabs of food animals (broiler chickens, beef cattle and pigs) from farms in North-East Italy during a period in 2000-2001, observing either *C. jejuni* and or *C. coli* in 82.9% of chickens, 53.9% of cattle and 63.5% of pigs. They also tested meat, both submitted directly to the lab (n=444) or purchased at retail (n=37) observing *Campylobacter* from 81.3% of chicken meat samples, only 1.3% of beef and 10.3% of pork. Furthermore, Pezzotti *et al* (2003) also investigated the species of the *Campylobacter* isolates found: of the beef cattle with *Campylobacter*, 22% were *C. jejuni*, 26% were *C. coli* and the remainder (52%) were 'other thermophilic species'; however in the two retail beef samples which tested positive both (100%) were *C. jejuni*. In chickens Pezzotti *et al* (2003) observed 44% *C. jejuni* and 56% *C. coli* from rectal swabs, with 56% *C. jejuni* and 44% *C. coli* from meat. Whilst in pigs 1.3% of *Campylobacters* recovered by Pezzotti *et al* (2003) were *C. jejuni,* 63.5% were *C. coli* and the remaining 35.2% were 'other thermophilic species' from rectal swabs; from pork meat however 38.9% were *C. jejuni* and 55.5% were *C. coli*, with only one instance (5.6%) of 'other' species. Pezotti *et al* (2003) observed an unusually high level of *C. coli* in chicken and cattle compared to other studies; they associated this potentially with the high levels of antimicrobial resistance they observed in the *C. coli* isolates, related to routine antimicrobial treatments being used at the farm level, and also suggested that the high levels in pork meat could possibly be due to cross contamination occurring within butchers shops.

Siemer, Nielsen & On (2005) considered the epidemiology of *C. coli*. They found that despite the known prevalence of *C. coli* in pigs, the majority of human *C. coli* isolates in their study were more closely related to *C. coli* from poultry sources, and suggest that human *Campylobacter* infection may rarely be linked with pork products.

Atabay & Corry (1998) investigated the prevalence and species of *Campylobacters* from dairy cattle rectal swabs from 136 cattle (42 of which were calves) across three farms; they observed around 7% positive for *C. jejuni*, 11% positive for *C. fetus fetus*, 32% positive for *C. hyointestinalis* and 21% positive for *'C. sputorum* biovar *paraureolyticus'*.

Zweifel, Zychowska & Stephan (2004) studied caecum samples from 653 slaughtered sheep, of which 17.5% were positive for *Campylobacter* species, which were determined to be comprised of 64.9% *C. jejuni* and 35.1% *C. coli*.

Although poultry is the primary animal source of *Campylobacter*, it has been demonstrated that *C. jejuni* and *C. coli* are frequently carried by food production mammals. Calculations of the burden of *Campylobacter* in food production mammals has varied, however it is generally accepted that pigs carry considerable loads of *Campylobacter*, and are particularly associated with *C. coli*, whilst sheep and cattle are generally less commonly associated with *Campylobacter* infection. A survey of abattoirs carried out by Manning *et al* (2003) revealed an unusual potential association between an MLST clonal complex of *C. jejuni* isolates and the typically *C. coli* associated porcine host.

1.4.3 Alternative Sources of *Campylobacter*

An MLST survey of over 250 *C. jejuni* isolates was carried out in 2003 by Manning *et al*, demonstrating that *C. jejuni* are also recovered from diverse sources including giraffes and ostriches as well as environmental samples, whilst Sproston *et al* (2010) later demonstrated that flies and slugs can serve as carriers of *C. jejuni* and *C. coli* via ruminant faeces, and may represent a potential link in the transfer of *C. jejuni* and *C. coli* within and between food production mammals, or may be a risk factor for human infection through contact with food preparation surfaces.

Other recognised sources of *Campylobacter* infection include pets, especially dogs and cats. In one study, Hald & Madsen (1997) investigated the presence of faecal shedding of *Campylobacters* from healthy puppies and kittens (aged 11-17 weeks) and found that 29% of the 72 puppies tested positive for *Campylobacter*; 76% *C. jejuni*, 5% *C. coli*, 19% *C. upsaliensis*, whereas of the 42 kittens only 5% (two) were positive for Campylobacter, specifically *C. upsaliensis* in both cases. Parsons *et al* (2009) also investigated the potential link between dogs and human infection with *C. jejuni* – they found evidence that the same types of *C. jejuni* (by MLST) could be observed in humans and dogs, although this could indicate both are getting it from the same source (such as water), or it could be that dogs are transmitting to

humans, or even vice-versa, but suggest that it should be acknowledged that this is at least a potential route for zoonoses.

Water is another potential source for *Campylobacter* infection. Wilson & Moore (1996) 42% of 380 tested shellfish tested positive for *Campylobacter* spp., with seasonal variation observed – 6% tested May-August were positive, compared with 58% Feb-April and 81% October-January. Of these, the majority (57%) were 'urease-positive thermophilic *Campylobacters*' – not commonly associated with human infections, the rest being *C. lari* (24%) or 'other' *Campylobacter* spp (9%) with only 8% being *C. coli* and 2% *C. jejuni*. Szewzyk *et al* (2000) conducted a report on the link between waters and microbial infections. They observed that lake water had been implicated as the source for infection of *C. jejuni* into dairy cattle, and stated that Campylobacters have been observed in raw sewage as well as in faecally contaminated surface waters. The involvement of water also goes full circle in Stanley, Cunningham & Jones (1998), where dairy cattle were potentially responsible for the observation of *C. jejuni* contaminating groundwater. Outbreaks of Campylobacter from drinking water have also been observed.

Zoonotic illness is commonly associated with contaminated food but is also linked with water and wildlife. Wilson *et al* (2008) describe evidence for *C. jejuni* being a waterborne pathogen - the seasonality is similar to that in waterborne illnesses, and some studies of phylogeny have suggested that human cases are often caused by 'non-livestock' types; however their study of over 1000 *C. jejuni* from human patients showed that the majority of cases are poultry, then cattle, with only about 3% of *C. jejuni* infections caused by water/wildlife sources (Wilson *et al*, 2008).

1.5 Prevention Strategies

As described previously, human infection with *Campylobacter*, and particularly *C. jejuni* places a considerable burden on healthcare and the economy. *C. jejuni* is the largest cause of bacterial gastroenteritis in the 'developed' nations, so any efforts to reduce the impact of these infections could be of great benefit. Despite the recognition in recent times of the prevalence of *C. jejuni* infection it remains a significant issue. A number of prevention and control strategies have been

introduced or proposed to reduce the occurrence of *Campylobacter* related illness, with varying levels of success; this section provides an overview of the work in this area.

1.5.1 Targets & Techniques for Reducing *Campylobacter* Infection

Due to the association between poultry handling and consumption and *C. jejuni* infection the optimal way to reduce incidence of campylobacteriosis would be through reducing or eradicating the *C. jejuni* burden in poultry (Allos, 2001; Strachan & Forbes, 2010), however this is almost certainly a futile approach at present – even if flocks are cleared, there is ample opportunity for recontamination at slaughter house or during packing and distribution. Therefore other targets for reducing or eradicating *Campylobacter* include the slaughterhouse environment or at the latter stages by treating food products and their packaging.

1.5.1.1 Reducing *Campylobacter* in Poultry Flocks

Chicks do not normally hatch already infected with campylobacter – horizontal transmission is not a common occurrence – this is why the 'top-down' programme used to reduce Salmonella in poultry is not of use. *Campylobacter* usually enters a flock at or after the age of two weeks, and due to the lack of symptoms in the birds it can remain undetected unless diagnostic samples are taken (Wassenaar, 2011). Water can be a source of infection for the chicks, although acquisition from contaminated food stuffs is less common. Other sources include insects, flies and vermin, wild birds or other farm animals, the human handlers and so on. Vaccination of chickens has also been largely unsuccessful in *Campylobacter* – despite being a considerable success in *Salmonella* – this is likely due to the variation in *Campylobacter*, however research has shown there may still be hope for an effective vaccination programme, using attenuated *Salmonella* carrying specific *Campylobacter* proteins, such as *Cj0113* (omp18, CjaD) as described by Layton *et al* (2011). Competitive exclusion of *Campylobacter* in poultry using pre- or probiotics has also been investigated, and although eradication was not achieved it has been demonstrated that this may provide a significant reduction in *Campylobacter* burden in poultry flocks (Lin, 2009).

Newell *et al* (2011) considered biosecurity interventions to reduce campylobacter at poultry farm flocks. They demonstrated that improved biosecurity can reduce *Campylobacter* colonisation in broiler flocks; however, they acknowledged that this would need to be paired with other approaches to have a large impact on reducing *Campylobacter* prevalence at the farm level.

1.5.1.2 Reducing *Campylobacter* at the Slaughterhouse

The above are methods to reduce or eliminate colonisation of the birds themselves. The next opportunity would be to reduce the bacterial load at the slaughter-house or just prior to slaughter. Bacteriophage therapy could be applied prior to slaughter (Wassenaar, 2011; Connerton, Timms & Connerton, 2011; Lin, 2009) however there is a risk of resistance developing and this has not been demonstrated as being successful at industrial scale.

During movement to the slaughter-house and the subsequent slaughter and processing of carcasses, each flock is bringing its own bacterial load, and any flocks which remained previously uninfected with *Campylobacter* are likely to acquire it at this point through cross contamination; even birds that enter the slaughterhouse without *Campylobacter* are likely to leave as contaminated meat, and transmission via the vehicles, workers and poultry crates at the slaughter-house has the potential to result in *Campylobacter* being returned to flocks which were previously clear of *Campylobacter* (Ellerbroek, Lienau & Klein, 2010; Wassenaar, 2011).

1.5.1.3 Reducing *Campylobacter* during the Processing & Packaging Stages

Here we have meat from the slaughterhouse which is contaminated with Campylobacter. A significant issue here is that the environment designed to reduce spoilage of meat – reduced oxygen – actually increases the likelihood of *Campylobacter* survival (Wassenaar, 2011). Even once packaged there is still a risk; Patrick *et al* (2010) demonstrated that children riding in shopping carts were at increased risk of campylobacter if raw packaged chicken was in the cart.

Decontamination of meat and packaging would provide a useful means of reducing human *Campylobacter* infection, however various methods (including forced air

chilling, crust freezing, steam treatment, ultrasound, and electrolysed oxidising water) have varying levels of both cost and effectiveness and often are not as successful as freezing the meat would be (Wassenaar, 2011) although there is generally less demand for frozen poultry compared to the fresh meat market.

Irradiation of food products would be a useful method but is not generally approved by the consumers (Allos, 2001). Other modern methods for treating carcasses and packaging include high-intensity light pulses (Haughton *et al*, 2011) and lactic acid washes (Rajkovic *et al*, 2010). Rajkovic *et al* (2010) demonstrated that washing with a lactic acid/sodium lactate buffer reduces levels of *C. jejuni* on chicken carcasses, and this was even more effective when combined with subsequent packaging being carried out in high oxygen atmosphere. Additionally, in some countries, hyper-chlorinated water washes are used to clean carcasses post-harvest; in particular this method has had apparent success in New Zealand, however it is not used in Europe to legislative restrictions (Newell *et al*, 2010).

1.5.1.4 Reducing *Campylobacter* through Consumer Practices

The major approach currently in reducing *C. jejuni* infection is to educate and encourage the consumer to take part in good hygiene practices and thorough cooking of meat products, especially where poultry is concerned – certainly it should not be served pink, thermometers could be used, thorough cleaning of cutting surfaces, utensils and hands after handling (Allos, 2001). Thorough cooking is hugely important; the heat will kill any remaining *Campylobacters*; however the biggest risk factor in the home is cross-contamination of other surfaces and food-stuffs - via chopping boards, utensils, salads and raw foods – (Luber, 2009).

1.5.2 Successful Interventions for *Campylobacter*

Jore *et al* (2010) reviewed the incidence of *Campylobacter* across six European countries (Denmark, Finland, Iceland, Norway, Sweden and The Netherlands), both in human cases and in broiler flocks over a ten year period between 1997 and 2007. Each country showed the seasonality of *Campylobacter* infections, with an increase in numbers observed in broilers and in human campylobacteriosis cases reported

during summer months, peaking July and August. Across the time of their study a decrease in campylobacteriosis, and broiler carriage was observed in Denmark and Sweden (2001-2007), campylobacteriosis cases were also declining in Iceland from 1999-2007, with broiler carriage falling from 2004-2007. Norway had an increase in human cases across the period, and instances in The Netherlands and Finland remained relatively consistent across the period of study. The reduction in Iceland is associated with a testing and freezing program which has been implemented since 1999; where flocks are tested for *Campylobacter* presence, and all meat is frozen if it tests positive (Lowman *et al*, 2009 conference poster, cited by Strachan & Forbes, 2010).

Other success has been found reducing *Campylobacter* in New Zealand (Public Health Services report, New Zealand, 2009; Strachan & Forbes, 2010) and in the USA (Samuel *et al*, 2004; Strachan & Forbes, 2010). Samuel *et al* (2004) considered culture-confirmed cases of campylobacter in the USA during 1996 to 1999. Overall the average *Campylobacter* incidence observed was 21.9 cases per 100,000 of the population, however Samuel *et al* (2004) observed a reduction each year. This decrease may be potentially associated with improved prevention techniques, however it must be noted that Campylobacter infection still remains a significant issue, with around 2 million cases per year. The decrease in numbers observed by Samuel *et al* (2004) did coincide with improved methods in poultry treatment (increased water volume in washing and additional disinfection of water) with the introduction of "Pathogen Reduction (PR)/Hazard Analysis and Critical Control Points (HACCP) systems final rule", however improved behaviour by food production workers and/or the general public could not be ruled out as additional factors in the decrease in reported *Campylobacter* cases. The numbers and trend observed by Samuel *et al* (2004) appear to be consistent in recent figures; in 2012 the number of cases per 100,000 were 14.3 as recorded by FoodNet (http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6215a2.htm).

Some countries have successfully caused a reduction in the burden of bacterial infectious gastroenteritis (such as The Netherlands), however even these countries still face a large incidence of disease (Newell *et al*, 2010). The three bacterial agents

of food-borne gastroenteritis which receive the most focus are *E. coli*, *Salmonella* spp. and *Campylobacter* as they represent the majority of 'food-poisoning' cases, although in recent years attention is also being focused on other emerging pathogens, however despite the consistent research, surveillance and prevention strategies over multiple decades these three pathogens remain a significant burden (Newell *et al*, 2010).

Campylobacter interventions (in the UK) have largely focused on the control of the poultry industry – various methods in various countries including at the farm, slaughter and processing and packaging, and through addressing at home and in industry kitchen hygiene knowledge and standards - as this is known to be a significant source for *Campylobacter*. As Strachan & Forbes (2010) stated - it is important that in the UK we take responsibility for the high burden of *Campylobacter* and acknowledge that effective measures can be introduced with investment and commitment, and as Moore *et al* (2005) stated, the most likely route for success in reducing *Campylobacter* infections should be to apply interventions across the food chain; beginning with reducing colonisation of animals, improving control during transport, slaughter, dressing and packaging, and with improved education of the consumer.

1.6 Typing, Speciation, & Differentiation in *Campylobacter*

This section reviews the current and important earlier methods for typing and differentiation in *C. jejuni*. As mentioned earlier, in epidemiological studies many labs often do not distinguish between *C. jejuni* and *C. coli*; this is due to the difficulties in quickly and accurately differentiating between the two using standard laboratory based tests.

1.6.1 Biochemical & Serological Typing Methods

Traditionally, hippurate hydrolysis was used as a discriminatory test to distinguish *C. jejuni* from other thermo-tolerant *Campylobacter* species, notably *C. coli* and *C. lari* (Nicholson & Patton, 1995). The basis for this differentiation was that *C. jejuni* would be hippurate hydrolysis positive whilst *C. coli* and *C. lari* were negative for

hippurate hydrolysis; however it has been shown that some *C. jejuni* strains are phenotypically hippurate negative (Caner *et al*, 2008), although the *HipO* gene is still present. A further complication in the speciation of *C. jejuni* are the subspecies; as described by Parker *et al* (2007) the species *Campylobacter jejuni* is divided into two subspecies - *C. jejuni jejuni* and *C. jejuni doylei*. Subspecies *doylei* isolates are hippuricase positive (as are most *C. jejuni jejuni*) and are commonly distinguished from *C. jejuni jejuni* through their inability to reduce nitrate. *C. jejuni doylei* is frequently associated with clinical cases, and particularly with blood cultures from septicaemia and bacteraemia. Parker *et al* (2007) demonstrated that the *C. jejuni doylei* subspecies is phylogenetically distinct, and that it and *C. jejuni jejuni* most likely split from a common ancestor; due either to an ecological barrier preventing exchange, or due to some genetic barrier to recombination between the two subspecies if they do coexist within the host environment. Due to the different illness typically associated with *C. jejuni doylei* mistaken identity is not as common an issue as between *C. jejuni* and *C. coli*.

Typing techniques are also required as a tool in studying the relatedness of isolates, and to reveal sources and routes of transmission. A significant difficulty in understanding *Campylobacter* is the lack of an agreed typing system – a range of biochemical and serotyping methods have been used, but these are not always reliable, and it is not uncommon for isolates to return 'untypeable' results.

Serotyping was one of the earliest successful typing techniques; with two approaches - firstly Penner & Hennessy (1980) developed a serotyping scheme based upon heat-stable antigens using haemagglutination (later revealed to be the lipopolysaccharide *O*-antigen, as reviewed by Penner, 1988); and subsequently Lior *et al* (1982) developed a scheme using heat-labile antigens (using antisera-antibody-technique). The heat-labile method was easier and generally adopted in clinical labs (using coagulation). Serotyping was a useful tool particularly in the early days of *Campylobacter* research however the large number of untypeable strains became an issue as research progressed. Various other biochemical tests were also used in the early period of *C. jejuni* and *C. coli* study; including the production of H_2S, resistance to a group of reagents ('resistotyping') and various Carbon source

utilisation tests (Penner, 1988). A bacteriophage typing scheme for *C. jejuni* and *C. coli* was introduced by Grajewski, Kusek & Gelfand (1985) using fourteen lytic phages; creating a profile of susceptibility for each isolate. This was later updated by Sails *et al* (1998) who created a sixteen phage typing system, including six of the original phages from Grajewski, Kusek & Gelfand (1985).

Each biochemical, phage, or serological based technique, despite usefulness suffered some limitations, through variation, poor reliability of reproducibility, development of resistance and with the advances in DNA techniques these became preferred new methods for study.

1.6.2 Molecular Typing Techniques

In 1990, Chang & Taylor used PFGE (Pulsed Field agarose Gel Electrophoresis) to establish the size of the *Campylobacter* genome. PFGE requires digestion of the genome by restriction endonucleases in order to produce a small number of large fragments which can then be separated using pulsed field electrophoresis. Following on from this, Yan, Chang & Taylor (1991) then considered the potential for PFGE as an epidemiological tool for *C. jejuni* and *C. coli*, and demonstrated its use in not only species differentiation, but also in separating distinct isolates. This led to later studies such as that by Gibson, Lorenz & Owen (1997) who used PFGE to create 'fingerprints' for heat-stable serotype reference strains, and non-typeable strains and demonstrated its use as a method to improve understanding of serotypable strains, and determine relatedness of untypeable strains. In a similar study, Fitzgerald, Owen & Stanley (1996) had previously tried to determine genotype of heat stable penner reference serotypes and untypeable strains using ribotyping – typing based upon the RFLP of the 16S rRNA gene, where it was also shown that un-serotypeable isolates were related to serotypable isolates in most cases.

Meinersmann *et al* (1997) investigated the use of *flaA* gene sequencing in studying the relatedness of *C. jejuni* isolates. The *flaA* gene contains two regions of high variability - one between 700-1450bp region, and a short variable region between bases 450-600, with conserved regions maintained at both ends of the gene locus

(Harrington, Thomson-Carter & Carter, 1997; Meinersmann *et al*, 1997). Meinersmann *et al* (1997) established that the SVR (Short Variable Region) provided discrimination at almost the same level as using the whole *flaA* gene - they showed that *flaA* SVR typing could be combined with or replace other typing techniques such as ribotyping or serotyping. Harrington, Thomson-Carter & Carter (1997) also reviewed the *flaA* typing method, and observed evidence that recombination occurs within the *flaA* gene. In *Campylobacter jejuni* the flagellin subunit of the filament for flagella is coded by two homologous genes located in tandem but promoted/expressed separately - *flaA* and *flaB*. By comparing sequence of *flaA* in eighteen *C. jejuni* strains they observed that recombination between strains occurs within the *flaA* gene and also observed evidence of recombination between the *flaA* and *flaB* genes – a mechanism with potential for increased antigenic diversity and to protect from loss of the gene (the loss of *flaA* impairs motility, and it has been suggested that recombination from *flaB* may effectively 'repair' the *flaA* gene). The use of *flaA* typing is not suitable as a long term assessor of relatedness of *C. jejuni* isolates due to the large amount of recombination that occurs within the gene, however it can be highly discriminatory in closely related isolates, particularly when applied in combination with other typing strategies (Harrington, Thomson-Carter & Carter, 1997; Meinersmann *et al*, 1997; De Boer *et al*, 2000).

De Boer *et al* (2000) considered the relative merits of *flaA* typing, PFGE, AFLP and automated ribotyping for their discriminatory power and use for epidemiological analysis of campylobacters, using fifty poultry Campylobacters. They found AFLP to be more highly discriminatory than PFGE, *flaA* typing, or automated ribotyping (41, 38, 31 and 26 types determined, respectively by each method). They also showed that combining methods could improve differentiation results in isolates which were more closely related.

As Duim et al (1999) described, various phenotypic methods preceded genotyping with varying success – serotyping, phage typing, biotyping which relied on specific reagents, and so on – each had varying success in discriminating isolates and tended to encounter 'untypeable' strains. As such, genotyping quickly became the more desirable method for typing *Campylobacter* species, and AFLP (Amplified Fragment

Length Polymorphism) was one of the more successful 'early' methods for fingerprinting isolates – previously developed methods included PFGE, flaA typing and RFLP, ribotyping however, as described above, each of these also lack either discriminatory power (such as *flaA* typing) or may struggle with reproducible typing results. AFLP selectively amplifies restriction fragments which are cut from the target genome; AFLP restriction enzymes digest genomic DNA and then specific fragments are amplified by PCR to create bands which produce a 'fingerprint' on the gel. Using enzymes *HhaI* and *Hind*III with selective primers *Hind*A and *HhaA* Duim *et al* (1999) were able to obtain 40-50 bands which were well spread across the genome and evenly distributed with lengths ranging 50-450bp with acceptable levels of reproducibility and good discriminatory power: AFLP represented another successful fingerprinting technique, which readily distinguished distantly related isolates, but required combination with other techniques, such as *flaA* typing, in order to differentiate closely related strains.

1.6.3 MultiLocus Sequence Typing

MLST (MultiLocus Sequence Typing) is a useful tool for studying closely related organisms through the sequencing of essential, and therefore well conserved, housekeeping genes. The MLST scheme for *Campylobacter* was developed by Dingle and colleagues (2001) and compares the sequences of seven loci, which are each separated from each other by at least 70kb. The loci used are as follows: *aspA* (aspartase A); *glnA* (glutamine synthetase); *gltA* (citrate synthase); *glyA* (serine hydroxymethyltransferase); *pgm* (phosphoglucomutase); *tkt* (transketolase) and *uncA* (ATP synthase α subunit). Each distinct sequence is given an arbitrary number, so each isolate is assigned a seven-digit allelic profile, or Sequence Type (ST). The sequence types are then further grouped into ST-complexes where MLST profiles are identical at four or more of the seven loci.

Dingle *et al* (2005) created an MLST scheme for *C. coli* using the same loci as in the previously developed MLST scheme for *C. jejuni* (Dingle *et al*, 2001), but with new primers for the same genes. They also looked at *flaA* typing (sequencing the short variable region of the *flaA* gene) and found that it has been shared commonly

between *C. jejuni* and *C. coli* so it cannot be used to separate species but can be useful in distinguishing closely related isolates. As previously discussed, *flaA* typing alone is not useful for *Campylobacter*, but it can be used as an additional layer of information combined with a typing scheme such as MLST (Dingle *et al*, 2005. In 2008, Dingle *et al* developed an extended MSLT scheme, improving the discriminatory power of their original (Dingle *et al*, 2001) scheme by adding three antigen coding genes; *flaA* (SVR) *flaB* (SVR) and *porA*.

As described previously, the groupings determined by serotyping, although useful, can show very different results to typing based upon genetic analysis, and it has been demonstrated that different serotypes can be genetically closely related (Taboada *et al*, 2004). Genotyping studies such as MLST and PFGE replaced phenotypic tests as the 'gold standard' techniques for typing in *Campylobacter*, however these too face their limitations, notably in the limited view of considering only a small number of genes. Taboada *et al* (2004) created a full genome microarray using comparative genomic hybridisation (CGH) to investigate genome wide conservation. Microarray CGH compares the DNA of 2 strains competitively to a full genome on a slide. The work done on this report combines their fifty-one strains with the data gathered in three separate studies, given a total cumulative study of ninety-seven strains (Dorrell, 11 strains; Leonard, 16 strains and Pearson, 18). CGH/microarray allows study across the entire genome rather than being limited to a small number of genes, however it is limited as it can only illustrate what the query genome(s) share with the reference genome on the slide. Later microarrays have been created which hold 'pan arrays' for a species, for example however even these cannot inform on any novel content in the query genome.

1.7 MLST & Host Adaptation

Host adaptation has been evidenced in each group of the *Campylobacteriales* order – *Campylobacter*, *Helicobacter* and *Wolinella* (Young, Davis & DiRita, 2007) and numerous studies have investigated host association in *Campylobacter*, particularly using MLST and comparative genomic hybridisation (microarray). Sheppard and colleagues (2010) used MLST to investigate host association in *C. jejuni* and *C. coli,*

and found substantial differences between those isolated from poultry and ruminants. In 2007 McCarthy *et al* developed an MLST scheme to identify the origin of an isolate as cattle, sheep or chicken. Using microarray techniques it has been demonstrated that *C. jejuni* appears to form two distinct clades; livestock and non-livestock sources (Champion *et al*, 2005). Skirrow & Benjamin (1980) also observed associations between *C. jejuni* and cattle and *C. coli* and pigs. Various previous studies have linked specific MLST clonal complexes with a specific host, or host type, whilst other clonal complexes are associated with isolation from a wide range of sources.

French *et al* (2005) demonstrated that ST61CC *C. jejuni* isolates showed an association with cattle, and also found ST45CC isolates linked with water and wildlife samples, although ST45CC isolates were associated with a range of other sources including cattle. French *et al* (2005) also considered a specific allele in ST61 – the founder strain for ST61 has the *uncA* allele *uncA*17, and, as they describe, this is one of a group of unusual *uncA* alleles not commonly found in *C. jejuni* – it was suggested that this provides evidence of genetic transfer between *C. jejuni* and *C. coli* within the host as this is thought to be originally a *C. coli* allele. French *et al* (2005) also identified a range of new isolates with new sequence types which were associated with wildlife and water samples and not related to human pathogenic isolates. Manning *et al* (2003) also showed an association between ST61CC and cattle, combined with human isolates; however ST61CC has also been linked with sheep (Colles *et al*, 2003) and therefore is possibly a ruminant adapted isolate – it has been associated with human cases and is rarely isolated from poultry (French *et al*, 2005). Colles *et al* (2003) observed a possible link between ST45CC and poultry, similar to Manning *et al* (2003); although this has also been associated with numerous other host types. Colles *et al* (2003) found that both ST61CC and ST42CC were over represented in sheep and absent from poultry samples. In contrast, ST21CC doesn't display host specificity (Colles *et al*, 2003; Manning *et al*, 2003). Manning *et al* (2003) identified the ST403 clonal complex as a potential porcine adapted group, based upon a survey of 266 *C. jejuni* isolates from a range of souces.

Miller *et al* (2006) were able to associate specific MLST alleles with host types in *C. coli*. Miller *et al* (2006) identified 'common' alleles which were present in isolates from all four investigated host types (swine, turkey, cattle and chicken) and also identified alleles significantly associated with swine, chicken or turkey; they did not observed cattle specific alleles. Swine specific alleles in *C. coli* included *aspA*32, *aspA*53, *glnA*38, *gln*A153, *glt*A44, *glt*A134, *gly*A81, *gly*A167, *gly*A173, *pgm*118*, pgm152*, and *tkt*173.

Host association was also investigated by McCarthy *et al* (2007), using isolates from chickens and from cattle and sheep. McCarthy *et al* (2007) did not find successful separation based upon MLST or clonal complex, however they demonstrated that combining the allelic profiles of isolates provided discrimination; showing that host associated markers can be observed in *C. jejuni* and also demonstrating that *C. jejuni* isolates can share and import genetic content within a shared host. McCarthy *et al* (2007) found that it was much more difficult to separate cattle and sheep isolates from each other than from chicken isolates. This may be due to cattle and sheep isolates sharing a common gene pool, and transmission between these two sources. McCarthy *et al* (2007) also concluded that source association had a greater 'effect' than time or geographical features.

Sheppard *et al* (2010) further demonstrated that the host association of *C. jejuni* isolates over-rides the effect of geographical variation. Sheppard *et al* (2011a) showed considerable sharing of genetic content between *C. jejuni* and *C. coli*, with such a large proportion of the movement being from *C. jejuni* to *C. coli* they suggested that the two species were introgressing such that the species boundary is being eroded. Specifically, introgression was witnessed between *C. jejuni* and *C. coli* clade 1 – the 'farm' associated lineage, with no mosaic alleles observed in clades 2 and 3. Sheppard *et al* (2011b) demonstrated that some lineages can be grouped and ascribed to a host type, whilst others may be linked to several hosts and therefore not be associated with a specific host associated lineage.

Host adaptation also occurs in *Campylobacter* outside of food production animals; Griekspoor *et al* (2013) found that wild bird isolates of *C. jejuni* were genetically

distinct from those from human and food animal isolates. *C. jejuni* isolates from different wild bird species were also distinct from each other, and Williams *et al* (2010) identified a specialist niche-adapted *C. jejuni* clone was linked with the bank vole (*Myodes glareolus*).

Hepworth *et al* (2011) demonstrated that loss or divergence of certain genes were associated with niches, when comparing an isolate (1336, ST841) from the diverse water and wildlife lineage of *C. jejuni* against the host specialist isolate *C. jejuni* 414 (ST3704) which is niche adapted to the bank vole. In other instances, however, niche adaptation has also been associated with differential expression: Killiny & Almeida (2011) demonstrated in the plant pathogen *Xylella fastidiosa* that in this organism, host specificity is determined by changes in gene regulation.

It is accepted that the majority of human *Campylobacter* infections are caused by *Campylobacter jejuni* as demonstrated in case based analysis by Gillespie *et al* (2002). It is also acknowledged that the majority of human *C. jejuni* cases for which a cause can be identified are attributed to chicken as the source be it through handling of raw meat, consumption of undercooked meat, or through contact with the birds and their faeces, however this does not account for all cases of Campylobacteriosis and it is clear that other sources are important. *Campylobacter* is increasingly becoming considered a ubiquitous organism as it is discovered in more and more hosts and sources (Strachan & Forbes, 2010). *Campylobacter jejuni* and *Campylobacter coli* have also been found to occur in mammals including food production animals such as pigs and cattle as well as in companion animals including cats and dogs. Water has also been recognised as a significant reservoir for *Campylobacter*, including lakes, faecally contaminated surface waters and groundwater, as well as drinking water (Szewzyk *et al*, 2000; Moore *et al*, 2005.

Typically, *C. jejuni* is associated with poultry as the source, whereas *C. coli* is most frequently associated with food production mammals such as cattle and pigs. However, although *C. jejuni* is most commonly associated with poultry (either through consumption of poorly cooked food, contact with raw meats, or direct

contact with birds/faecal matter); it is accepted that poultry is not the sole source of *C. jejuni* infection in humans.

1.8 The Current Climate of *Campylobacter* Research

The history and description of *Campylobacter* has been considered, with some focus on its relevance to human illness and the sources from which it can cause this. The focus of this work has been and will remain centred on the thermo-tolerant *Campylobacter* species, and particularly on *C. jejuni*; the most commonly isolated bacterial cause of gastroenteritis. An overview will be provided of the current 'climate' in *Campylobacter* research; covering recent developments in understanding and current accepted knowledge.

Gaynor & Szymanski (2012) published a review conducted following the 16th CHRO meeting (Vancouver 2011) - combining new research discussed at the meeting and comparing new developments with the previous 30 years of CHRO research. New areas of research from recent years which were discussed by Gaynor & Szymanski (2012) included glycobiology; *C. jejuni* was the first bacterial species shown to possess an N-linked glycosylation pathway - leading to a new frontier in creating bacterial glycoconjugate vaccines (work by Brendan Wren); and advancements in new or improved animal models for the study of *C. jejuni* infection, particularly the technique of creating 'humanised' microbiomes in gnotobiotic mice to investigate the role of commensal bacteria in establishment or prevention of *C. jejuni* colonisation.

In the early to mid-2000s it was discovered that *C. jejuni* was able to produce recombinant glycoproteins (Terra *et al*, 2012). The N-linked protein glycosylation system observed in *C. jejuni* was the first instance of this being found in bacteria (Linton *et al*, 2005). The oligosaccharyltransferase involved (*pglB*) was subsequently expressed in *E. coli* and used to readily produce glycoconjugates for the development of vaccines in a much easier and more efficient way than previously possible; and developed even further using PGCT (protein glycan coupling technology; Langdon, Cuccui & Wren, 2009, cited by Terra *et al*, 2012).

Bereswill *et al* (2011) showed that murine microbiota prevents *C. jejuni* colonisation of the mouse model. Mice with an artificial 'human' microbiota, or gnotobiotic mice however remained colonised in Bereswill *et al*'s (2011) experiment for six weeks, compared to just two days required to clear *C. jejuni* in mice with mouse microbiota. The humanised or gnotobiotic mice showed a pro-inflammatory immune response and therefore provide a useful model for studying the immune regulatory response to *C. jejuni* infection, as well as providing an additional measure for comparing the pathogenicity of individual strains or mutants.

Previously, research was largely focused upon one strain or one single gene, whereas now *Campylobacter* is now rapidly becoming focused upon whole genome sequence and large scale investigations.

Despite *Campylobacter* being the largest cause of bacterial gastroenteritis in the developed world, and although significant research and advances have been made in the last few decades, *Campylobacter* remains a comparatively poorly understood organism, especially when compared to other human intestinal pathogens, such as *Salmonella* and *E. coli* which have been characterised to a great extent. Significant progress has been made across all aspects of *Campylobacter* - we now have a much greater understanding of sources, transmission and infection than when it was considered an emerging pathogen not too distant a time ago; however many questions remain in *Campylobacter* research. Although it is now generally accepted that poultry is a major source of human Campylobacteriosis in 'developed' nations, there is still much work to be done in controlling this source and preventing subsequent infection; additionally even were this to be achieved there remains the issue of alternative sources for Campylobacteriosis - as described above there are numerous potential reservoirs for *Campylobacter* each of which may play a varying role and may provide cross-talk with other hosts. Alongside the valuable work undertaken elsewhere to investigate means for reducing the burden of *Campylobacter* in poultry there is also a need to investigate other sources and their potential importance.

In summary, *C. jejuni* is the most commonly identified thermotolerant *Campylobacter* and is the largest recognised cause of human bacterial gastroenteritis. Its fastidious nature led to difficulties in recognising, describing and identifying *Campylobacter* as an important human pathogen, despite its prevalence. It is now established that *C. jejuni* has a reservoir in many sources including poultry, food mammals and water and represents an opportunistic human zoonosis which does not require humans as a host but will cause illness when acquired through contamination. Typically human infection with *C. jejuni* is an unpleasant but not serious 'food-poisoning' event which is cleared in a number of days-weeks characterised by fever, cramps and diarrhoea caused by the invasive behaviour of the bacteria within the epithelial cell lining of the intestinal tract; although much more serious sequelae can also occur, particularly in 'at-risk' groups. The difficulties in recovering and culturing *Campylobacters* from mixed cultures, coupled with the high levels of genomic variation, make the typing and observation of *C. jejuni* a challenging prospect also so studying groups of related and unrelated isolates is an important tool in learning more about this variation. Additionally, host association has been observed in Campylobacters and related organisms but more work is required to consider the role of hosts or reservoirs as potential melting pots for sharing genetic information within and between adapted groups.

1.9 Background & Aims of the Research

The research presented in this thesis developed from previous work by Manning *et al* (2003), which used MLST to analyse over 250 *C. jejuni* isolates from a range of veterinary and environmental sources as well as from human clinical cases. This study observed considerable overlap between clinical and veterinary isolates, with the majority (eighteen of nineteen) of the observed MLST clonal complexes containing both veterinary and clinical isolates. A single ST complex (the ST403CC) was identified however, which contained only veterinary isolates. Furthermore of the isolates within the ST403CC, 89% (sixteen of eighteen) were from pigs. These isolates were recovered from faecal samples taken from animals prior to slaughter during a survey of abattoirs conducted in England and Wales during 2000-2001. The remaining two ST403CC isolates were recovered from faecal samples taken from

cattle at slaughter. The ST403CC isolates were the only group observed in the study by Manning *et al* (2003) to show a strong association with veterinary only, mammalian sources, and were also shown to be hippurate hydrolysis negative in laboratory study. The hippurate hydrolysis negative phenotype of the identified ST403CC *C. jejuni* isolates was revealed by Stephen On, and announced by Manning *et al* (2003). As a result of these two factors, it was suggested that this group of isolates may represent a niche adapted clonal group of *C. jejuni*. The majority (nine of sixteen) were untypeable by the LEP method (Laboratory of Enteric Pathogens method; Frost *et al*, 1998), whilst those that were typeable were: one serotype 22 (ST270 isolate), one serotype 29 (ST270 isolate), four serotype 23 (one isolate each from ST435, ST550, ST552, ST553), and one serotype 35 (ST556). Of the ST403CC pig isolated *C. jejuni* strains, 4 share a common serotype, two are unique serotypes within the group, and 9 were untypeable.

Of the sixteen porcine ST403CC *C. jejuni* isolates, six were selected for in depth study, each representing a different sequence type within the ST403 clonal compex. These isolates are described in Table 1.1.

Species	Isolate ID*	Isolate Alias^	Sequence Type	MLST Clonal Complex	Serotype	Source
C. jejuni	PS857	857	270	403	NT◊	Pig
C. jejuni	PS549.1	549.1	403	403	NT◊	Pig
C. jejuni	PS623	623	552	403	NT◊	Pig
C. jejuni	PS304	304	551	403	NT◊	Pig
C. jejuni	PS484	484	435	403	23	Pig
C .jejuni	PS444	444	553	403	23	Pig

Table 1.1: ST403CC *C. jejuni* Isolates Included in Study
Information on the six ST403CC *C. jejuni* isolates selected for study. The isolates represent different sequence types within the ST403 clonal complex.
* Isolate ID used in Manning *et al* (2003)
^ Isolate alias ID used subsequently throughout the research.
◊ Not typeable

Manning *et al* (2003) reported, following a search of the *Campylobacter* MLST database, that other ST403CC isolates had been observed from food sources and

from infected humans - particularly from the Dutch Caribbean island Curaçao, although some of the specific STs within the complex were unique to pig isolates. Manning *et al* (2003) also cited work by Moore *et al* (2002), which had previously shown that *C. jejuni* and *C. coli* from pigs displayed considerable genomic diversity, but that the majority of isolates could be assigned to one of four genotypes based upon ribotyping and *flaA*-RFLP. Moore *et al* (2002) suggested that the population of *C. jejuni* and *C. coli* in pigs might consist mostly of a small number of dominant types, with a smaller number of variable subtypes.

The inconsistency in LEP serotype for the sixteen ST403CC pig isolates observed by Manning *et al* (2003) demonstrated the power of MLST over serotyping in this type of study as this subgroup would have been entirely overlooked by this method. It was also common for different serotypes to be found in other MLSTCCs and serves as evidence also of the fact that closely related isolates can express different antigens.

In March 2011 a search for ST403 complex isolates on the pubmlst.org database online (Jolley & Maiden, 2010: http://pubmlst.org/campylobacter/) returned 156 records in total from a variety of sources including 83 isolates associated with human gastroenteritis, 3 unspecified human isolates, 1 asymptomatic human carrier and 3 from GBS patients, plus isolates from beef (4), lamb (2), sheep (1) and pork (6) meats/offal, 2 from human blood culture (septicaemia), 18 from carrying cattle, 2 from dogs, 1 each from 'farm environment' and 'environmental waters' and several (26) from pigs as carriers (10 from Denmark, 16 from UK) a further 3 unspecified ST-403 complex isolates round out the group. A large proportion (45) of recorded ST-403 complex isolates associated with human gastroenteritis are from sporadic cases reported in Curaçao (all of which are sequence type 403 and were collected in 2000), plus 6 from Canada; 24 in the UK, 1 in South Africa, 1 in Egypt and 6 in the Netherlands. From this information it is clear that the ST-403 complex is a wide-spread group, but a tendency for prevalence in humans and pigs seems to be apparent. It also seems particularly interesting that this group has been frequently found both in the UK and in Curaçao (Duim *et al*, 2003; Endtz *et al*, 2003), given the markedly different environment and lifestyle between the two. It

was also observed that the ST403CC was not associated with poultry in the database, further suggesting that this may represent a niche-adapted group not associated with the most common host for *C. jejuni*.

These numbers can be compared with a more recent search of the same database (18/07/2014) by which point 287 isolates were logged as ST403CC isolates, from a total of 29011 isolates (approximately 1%). Human isolates included still just one from blood culture, associated with systemic disease, 159 from human stool – five GBS, one carrier, five unspecified (with no additional description provided) 148 gastroenteritis plus thirteen 'human unspecified' isolates, of which seven were associated with gastroenteritis, whilst the remaining six were not specified. The remaining 114 non-human isolates break down as fifty-three associated with cattle (carriers, meat or milk), three with sheep (animal or meat products), forty with pig (carrier, meat or offal), with six associated with dogs, one described as being from a 'farm environment', and eleven with no source description. Once again this more recent search maintains the lack of attribution of ST403CC *C. jejuni* isolates to poultry.

As described previously, typically *C. jejuni* is associated with poultry as the source, whereas *C. coli* is most frequently associated with food production mammals such as cattle and pigs. However, although *C. jejuni* is most commonly associated with poultry (either through consumption of poorly cooked food, contact with raw meats, or direct contact with birds/faecal matter); it is accepted that poultry is not the sole source of *C. jejuni* infection in humans.

The major goal of this research is to assess whether these closely related ST-403 complex porcine *C. jejuni* isolates may represent a host-adapted group, whilst also considering whether pigs may serve as an additional source for Campylobacteriosis. This will be achieved through: 'wet-lab' assessment of the ST403CC *C. jejuni* isolates, in order to consider the likely virulence potential of the isolates; and via whole genome sequence analysis to investigate potential unique features of this possibly host adapted group of isolates.

As described by Maiden and Dingle (2008, in Nachamkin, Szymanski & Blaser (Eds)), in order to truly understand an organism, the large scale population dynamics must be combined with comprehensive study focused on a single isolate or group of related isolates. This study aims to do just that; as it combines the established knowledge of *Campylobacter* with physiological and genomic study of this group of closely related isolates to inform on the adaptation of *C. jejuni*.

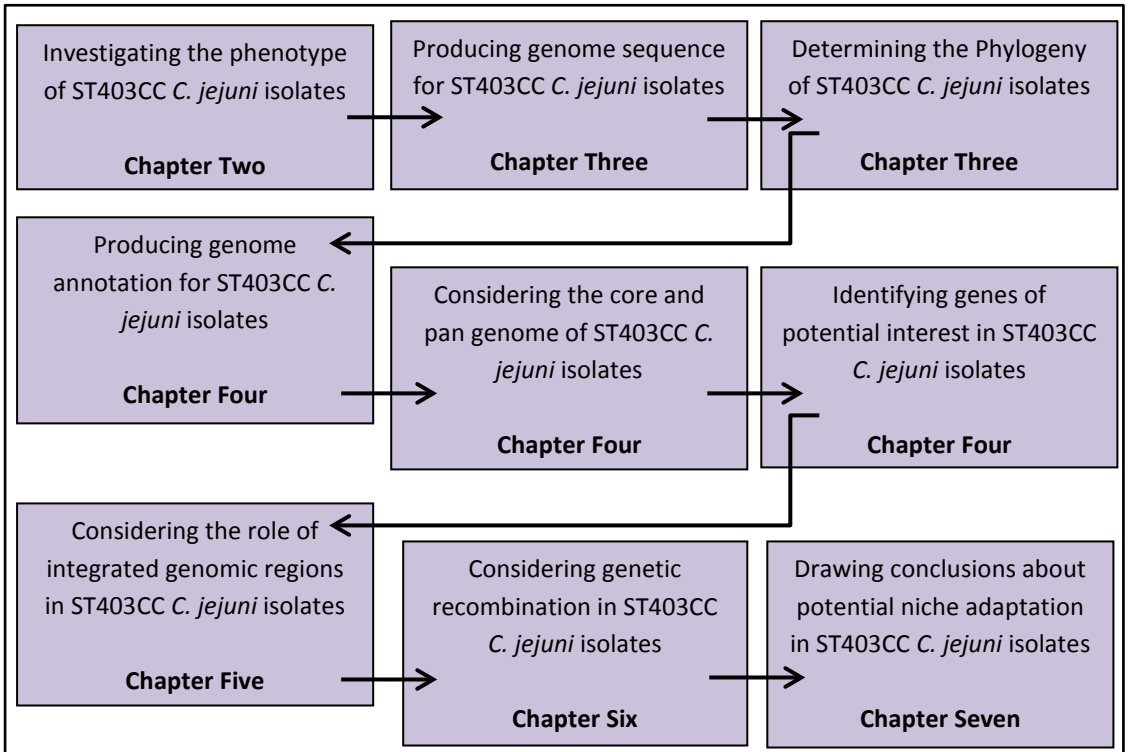

Figure 1.1: Flow of Study

Illustration of progression of study throughout the thesis.

Over the course of this thesis, a number of different bioinformatics techniques will be employed to investigate the genomics of the selected ST403CC *C. jejuni* isolates. Figure 1.1 provides a diagrammatic overview of the main views of the thesis, whilst in Table 1.2 brief information is provided regarding the major bioinformatics programs which are utilised in this study.

Chapter	Method	Function	Author(s)
Three	VELVET	Sequence 'shuffling' - mixing forward and reverse raw FastQ files to produce FastA consensus sequence for subsequent improvement and analysis	Zerbino & Birney (2008)
	ABACAS	Improves sequence quality by aligning, ordering and orientating a contiguated sequence against a reference genome	Assefa *et al* (2009)
	IMAGE	Improves draft sequence quality by removing gaps by aligning sequence against a reference genome	Tsai, Otto & Berriman (2010)
	SRST	Determines MLST from genome sequence	Inouye *et al* (2012)
	MUGSY	Produces multiple genome alignments for subsequent analysis including producing phylogenetic trees of relatedness	Angiuoli & Salzberg (2010)
	Mothur	Removes non-alphabetic characters from sequence files	Schloss *et al* (2009)
	RAxML	Produces maximum likelihood phylogenetic trees	Rokas (2011)
	FigTree	Phylogenetic tree editing software	Rambaut (2007)
	BRIG	Visualises regions of similarity and divergence between genetic sequences.	Alikhan *et al* (2011)
	SMALT	Produces mapped sequence alignments using a reference sequence	Ponstigl (2009)
	SamTools	Assesses SNP variation between aligned sequences	Li *et al* (2009)
Four	PROKKA	Genome annotation using a protein database and BLAST searching	Seemann (2014)
	RATT	Transfers annotation from a reference genome to a query genome	Otto *et al* (2011)
	EDGAR	Defines the core and pan genome of a population of sequences	Blom *et al* (2009)
Five	PHAST	Identifies prophage regions in genome sequences	Zhou *et al* (2011)
	AlienHunter	Identifies integrated genomic regions in genome sequences	Vernikos & Parkhill (2006)
Six	Heirarchical BAPS	Predicts the structure of a population of sequences	Corander *et al* (2012)
	BRATNextGen	Identifies recombination events in populations of closely related genome sequences	Marrtinen *et al* (2012)

Table 1.2: A Summary of the Major Bioinformatics Tools used in the Thesis
Provides an overview of the main bioinformatics methods used, with reference to the authors, and to the chapter in which they are used and described in greater detail.

Chapter Two: Determining the Phenotypic Characteristics of ST403CC *Campylobacter jejuni* Isolates

2.1 Introduction

Six *Campylobacter jejuni* isolates were selected from MLST clonal complex 403. It was hypothesised that these isolates potentially represent a host-adapted clonal group of *C. jejuni*, associated with recovery predominantly from pigs but also from cattle. The isolates were previously un-studied with the exception of being typed by MLST and being tested for hippurate hydrolysis, for which they were negative (Manning *et al*, 2003), as such initial phenotypic testing was required in order to provide a picture of the general characteristics and behaviour of the isolates. Due to the atypical porcine host of these ST403CC *C. jejuni* isolates, experiments were selected and undertaken in order to investigate the capacity of these veterinary isolates to cause disease in humans, in addition to considering whether they exhibit any evidence of host adaptation, such as stress resistance or motility, to allow survival or possible dominance in porcine or mammalian hosts which may explain the bias of isolation of ST403CC from mammalian hosts, and particularly from pigs.

2.1.1 Host Cell Invasion

Invasion is one of many routes used by intestinal pathogens to circumvent the host immune response. *C. jejuni* is known to be intestinally invasive; as has been demonstrated using tissue and cell culture methods, patient biopsies, and in animal model testing (Moore *et al*, 2005), and invasive ability is therefore considered an important virulence factor in *C. jejuni* (Hu & Kopecko, in Ketley & Konkel (Eds), 2008). *Campylobacter* is able to survive passage through the stomach, and is understood to first adhere to and later invade epithelial cells in the colon, ileum and jejunum, by inducing cytoskeletal rearrangement of the host cell (Krause-Gruszczynska *et al*, 2007), however the exact full route of invasion taken by *Campylobacter* has not been fully elucidated and does not follow the same pathway as model organisms such as *Salmonella* Typhi or *Escherichia coli* (Gilbreath *et al*, 2011). Epithelial cell lines have been particularly useful for the study of *C. jejuni* infection *in vivo* (Young, Davis & DiRita, 2007), including HeLa cells (human cervical

carcinoma cell line), INT407 (originally foetal intestinal epithelium, contaminated with HeLa, as reviewed by Lacroix, 2008), HEp2 (human epithelial type 2, originally human laryngeal carcinoma cells) and the human colonic carcinoma cell line Caco-2. *In vitro* investigation, although informative, has not generally been as successful in *C. jejuni* research due to the lack of an ideal small animal model (Young, Davis & DiRita, 2007). Various animal models have been used, with differing degrees of success; ferrets can undergo a similar disease response to *C. jejuni* as is seen in humans, however this is a high cost model, which is not commonly used and as such lacks the knock-out techniques available in more well characterised models such as mouse (Young, Davis & DiRita, 2007). The mouse model, which is commonly used in the laboratory study of bacterial pathogens, has not been as successful in the study of *C. jejuni* due to the natural colonisation resistance of the mouse against *C. jejuni*, and poor reproducibility found in some studies, as described by Young, Davis & DiRita in their review of *C. jejuni* pathogenesis (2007), although in recent years new advancements in techniques have meant that the murine model may be increasingly relevant in studying *C. jejuni* infection and immunology (Bereswill *et al*, 2011). Other animal models include the larvae of *Galleria mellonella* (Champion *et al*, 2009; Senior *et al*, 2011) and piglets (Babakhani, Bradley & Jeons, 1993). Chicken colonisation experiments have also commonly been used in *C. jejuni* research, particularly to study the effects of gene knock-outs, however the difference in the chicken and human host response and host environment reduce the value of the chicken model for predicting pathogenic potential in humans.

Several methods have been used to investigate the ability of bacteria to internalise into host cells, persist and cause disease, including 'predictive' tests – those which have been associated with particularly virulent isolates – as well as the *in vitro* and *in vivo* methods described briefly above. Cell culture models provide a means to obtain quantifiable data regarding the invasive capability of a given isolate; Fearnley *et al* (2008), demonstrated that although there is a substantial degree of strain-strain variation for invasiveness in *C. jejuni* results for individual strains using the gentamicin protection assay are reproducible. Fearnley *et al* (2008) also demonstrated that *C. jejuni* isolates could be identified as having low- high- or

hyper-invasive phenotype using the gentamicin protection assay with INT407 cells. *C. jejuni* 81116 was used as a low invasive control stain, and formed the basis for comparison for attribution to the three classes: a strain which was more than twenty-five times more invasive than *C. jejuni* 81116 was considered hyper invasive; any isolate which was ten times more invasive than *C. jejuni* 81116 was considered highly invasive; and any strain less than ten times more invasive than *C. jejuni* 81116 would be considered a low invader. Subsequent studies including that by Javed *et al* (2010) have confirmed these characteristics in the Caco-2 cell line.

2.1.2 Predictive Tests for Pathogenic Potential

As described previously, the pathogenic potential of *C. jejuni* isolates may be tested *in vitro* in cell and tissue culture experiments, and *in vivo* using whole animal models, however in some cases a rapid indication of pathogenic potential can be determined using 'predictive' methods. Two such predictive characteristics are motility and autoagglutination.

Motility is important in, but not essential for, host cell invasion by *C. jejuni*; Wassenaar *et al* (1991) illustrated this through a series of *flaA* mutant studies. Inactivation of *flaA* leads to an immotile bacterium which is less able to adhere to and subsequently invade intestinal epithelial cells experimentally, however these non-motile mutants were demonstrated to still be capable of adherence and invasion when brought into contact with the host cells via centrifugation. Further studies involving flagellar mutants have shown that the flagellum is unimportant in the adherence of *Campylobacter* to intestinal cells, but is important for internalisation (Grant *et al*, 1993). Grant *et al* (1993) demonstrated that aflagellate mutants were able to adhere to INT407 cells at levels comparable to the wild type, but were significantly less efficient at internalising into INT407 cells. The *Campylobacter* flagellum does not appear to have a role as a binding mechanism, however it has been demonstrated that the flagellar apparatus plays a role in secreting virulence proteins which may be involved in invasion (Song *et al*, 2004; Guerry, 2008; Neal-McKinney & Konkel, 2012). Specifically, in 2004, Song *et al* demonstrated that flagellar apparatus was required for the secretion of FlaC, and

that *flaC* null mutants were impaired for invasion into HEp-2 cells compared to the wild type *C. jejuni* TGH9011; whilst in 2012 Neal-McKinney & Konkel showed that the virulence associated protein CiaC required a functional flagellar hook mechanism for its secretion, and also highlighted a potential role of CiaC in the production of membrane ruffles, an important stage of bacterial internalisation. Motility may therefore be considered a useful indicator of virulence potential in *Campylobacter*; as strains with low motility are less likely to successfully colonise the human gut, due both to their reduced opportunity of obtaining contact with cells in order to adhere, and their reduced capacity for internalisation into the cells once adherence occurs.

Autoagglutination is also a potential indicator for virulence in *C. jejuni*, as described by Misawa & Blaser (2000). Autoagglutination was shown to be highly associated with hydrophobicity, and with adhesion to INT407 cells. No link was observed between autoagglutination and motility, although flagellation was thought to play a role in autoagglutination. As such, autoagglutination may be considered a quick test to predict likely pathogenic ability of *C. jejuni* isolates – a strain with high autoagglutination being more likely to successfully adhere to host cells and subsequently to invade into the cell.

For the purposes of this investigation, pathogenic potential was assessed using the predictive assays for motility and autoagglutination, and subsequently through quantifying the adhesion and invasion efficiency using a standard gentamicin protection assay.

2.1.3 Resistance to Environmental Stressors

In addition to the invasive potential of a pathogen, the ability of an isolate to survive stressful environments may contribute to its ability to colonise the host and therefore increase its pathogenic potential (Habib *et al*, 2010). *Campylobacter* is a microaerophilic organism and as such is vulnerable to oxygen, however *Campylobacter* must encounter reactive oxygen species (ROS) including superoxide radicals ($O_2^{\cdot-}$), hydrogen peroxide (H_2O_2) and hydroxide radicals (·OH) as a result of its own metabolism and during colonisation, both as part of the host immune

response and as a result of the metabolism of host microflora (Baillon *et al*, 1999; Palyada *et al*, 2009). ROS form a significant hazard for microorganisms as they can cause significant cell damage. Reactive oxygen species affect proteins, nucleic acids and lipids. The lipid membranes of bacterial cells are particularly affected by ROS: lipid peroxidation occurs when free radicals attack the polyunsaturated fatty acids found in membranes, and significantly affects the properties of the membrane itself, and the membrane bound proteins (Cabiscol, Tamarit & Ros, 2000; Palyada *et al*, 2009). ROS can also damage both the sugars and bases of nucleic acid, resulting in breaks in single or double strands, blocking of DNA replication, and loss of function through additional cross-links (Cabiscol, Tamarit & Ros, 2000). Many Gram negative bacteria utilise SoxRS and OxyR to regulate their stress response to superoxide and peroxide exposure, however these are absent in *C. jejuni*, which has an alternative peroxide sensor known as PerR (van Vliet *et al*, 1999). Oxidative stressors also play a role in the up-regulation of virulence associated genes (Harvey & Leach, 1998).

Bile, and more specifically the acids within bile, comprise a second environmental stress encountered by *Campylobacter* during the course of colonisation of the human host (Fox *et al*, 2007). Bile acids are formed in the human liver from cholesterol and are present in the intestine (Garrett & Grisham, 2009), providing two-fold functions for the human host; aiding in fat digestion, and by acting as a detergent with antimicrobial activity through damaging the lipid bilayers of bacterial cellular membrane (Raphael *et al*, 2005; Gunn, 2000). The two major primary bile acids produced in humans are cholic acid and chenodeoxycholic acid, which are produced from the liver, and circulated from the gallbladder to the large intestine, before being reabsorbed into the blood to return to the beginning of the cycle (Masanta *et al*, 2013). Within the intestine, commensal bacteria act to metabolise any cholic acid and chenodeoxycholic acid left behind into the 'secondary' bile acids deoxycholic acid and lithocholic acid, respectively, with deoxycholic acid being the most abundant secondary bile salt (Carey, in Nair & Kritchevsky, 1973; Masanta *et al*, 2013).

In addition to the obvious need to survive exposure to bile acids in order to survive the human gut, bile acid resistance also has a direct implied link with pathogenicity. Deactivation of bile acid resistance processes via mutation of the efflux pump CmeABC was demonstrated to increase bile salt sensitivity, and also prevented colonisation in the chick model (Lin *et al*, 2003). Later it was demonstrated that the two component regulator system *cbrR-cbrS* controls resistance to deoxycholic acid (sodium deoxycholate) with *cbrR* negative mutants exhibiting greatly increased sensitivity to deoxycholic acid and also having reduced chick colonisation ability (Raphael *et al*, 2005).

Chemotaxis plays a significant role in bacterial colonisation of the host (Ketley, 1997); allowing movement towards nutrients and target organs or surfaces, as well as movement away from potentially bactericidal chemicals. Herrmann & Burman (1983) previously showed that *E. coli* displayed positive chemotaxis towards urine – suggesting that this may have a role in its pathogenesis in UTIs (urinary tract infections), following this, Hugdahl, Beery & Doyle (1988) demonstrated a similar phenomenon for *C. jejuni*, which displayed positive chemotaxis towards mucin, L-serine and L-fucose, and negative chemotaxis away from bile acids. Subsequently, Takata, Fujimoto & Amako (1992), using the methods developed by Hugdahl, Beery & Doyle (1988), demonstrated that non-chemotactic mutants failed to colonise mice. Resistance to bile acids, and the ability to move away from them are therefore important factors in potential *C. jejuni* virulence – an isolate which is unable to move away from high concentrations of bile acids may not survive to colonise the host, whilst an isolate which is particularly resistant may be able to persist, leading to colonisation.

Temperature resistance may also play a role in host invasion; *C. jejuni* does not grow at temperatures below 25°C, however it has been demonstrated to survive at low temperatures for considerable time (Murphy, Carroll & Jordan, 2006) and recover upon consumption by the host. As with other thermotolerant *Campylobacter*, *C. jejuni* grows well at 37°C, with an optimum growth temperature of 42°C (King, 1957). In this project, the ST403CC *C. jejuni* isolates were cultured at

37°C and 42°C and displayed normal growth at these temperatures as described later in this chapter.

An isolate which is exceptionally resistant to bile acids and or oxidative stress may be able to colonise faster, or more successfully, and may cause increased levels of illness both due to the larger exposure and due to the increased invasive potential due to up-regulation of virulence genes in the presence of the stressors.

Resistance to ROS was tested for the ST403CC *C. jejuni* isolates using hydrogen peroxide and pyrogallol, both of which represent oxygen species encountered within the host. Hydrogen peroxide comprises the intermediate stage in the breakdown of superoxide radicals: NADPH oxidase catalyses the oxidation of NADPH to superoxide (NADPH and $2O_2$ to NADPH + $2O_2^-$ + H+); then superoxide dismutase catalyses $2H^+ + 2O_2^-$ to H_2O_2 and O_2 and finally catalase can convert $2H_2O_2$ to $2H_20$ and O_2^-. Hydrogen peroxide conversion from superoxide radicals occurs within phagosomes, raising the pH to around 7.8-8, activating antimicrobial peptides which break down the engulfed bacterium (Parham, 2009). Pyrogallol (1,2,3-trihydroxybenzene) is a superoxide (O_2^-) producer; it is oxidised spontaneously in air producing O_2^- and is used to create superoxide stress in laboratory experimentation (Champion *et al*, 2010). Superoxide and Hydrogen peroxide are produced by the human immune system as part of the defence response against pathogenic microorganisms (Parham, 2009), and sensitivity to both was tested for in the ST403CC *C. jejuni* isolates.

The ST403CC *C. jejuni* isolates were also examined for sensitivity to primary human bile acids cholic acid and deoxycholic acid as well as the major secondary bile acid chenodeoxycholic acid.

2.1.4 Summary & Aims

The overall aim of this chapter was to consider the virulence potential of potentially host adapted ST403CC *C. jejuni* isolates. This was addressed through predictive tests for pathogenicity, assessing the motility and autoagglutination of the isolates, as well as through *in vivo* experiments to establish the invasive capability of the

isolates. Indirect factors affecting the pathogenic potential of the isolates were also assessed, considering the sensitivity of the isolates to oxidative stress and bile acids. In tandem with the experiments being conducted on the six porcine ST403CC *C. jejuni* isolates, two 'control' *C. jejuni* isolates were also included; the low invasive strain *C. jejuni* 81116 and the high invading strain *C. jejuni* 11168. Additionally included for comparison were four porcine *C. coli* isolates; as *C. coli* is more commonly associated with pigs than *C. jejuni*. These 'control' isolates were included as a contrast to the porcine ST403CC *C. jejuni* isolates.

The major aims of this chapter were:

- To investigate the degree of motility across the ST403CC *C. jejuni* isolates and compare to the motility observed for known low and high invasive *C. jejuni* and for porcine *C. coli* isolates.
- To observe the level of autoagglutination for the ST403CC *C. jejuni* isolates, compared to those for *C. jejuni* reference isolates, and porcine *C. coli*.
- To determine the adherence and invasive capabilities of ST403CC *C. jejuni* isolates, to investigate whether these isolates have the potential to cause disease in humans, to categorise the ST403CC *C. jejuni* isolates are low- high- or hyper- invasive, and to compare against the results observed for porcine *C. coli* strains in the same test.
    - The relationship between motility, autoagglutination and adherence and invasion will also be considered.
- To observe the levels of sensitivity to oxygen stressors and bile acid exposure for ST403CC *C. jejuni* isolates and once again to relate this to the reference *C. jejuni* strains and *C. coli* from pigs, as well as to consider how these sensitivity levels may relate to the 'direct' virulence measures.
- To consider whether, across the phenotypic tests, the ST403CC *C. jejuni* isolates tend to behave in the same way as each other, and whether they can be distinguished on the basis of these characteristics as being different to the reference *C. jejuni* strains or sharing characteristics with *C. coli* isolates associated with the same host type.

2.2 Methods

2.2.1 Storage & General Culture Procedures

All *Campylobacter* strains were stored at -80°C in Mueller-Hinton broth (Oxoid (Thermo Scientific) Item CM0405) containing 20% v/v glycerol (Fisher Scientific Item G/0600/17) until required. Unless otherwise stated, when required, *Campylobacter* strains were cultured from the -80°C stocks onto a blood-free *Campylobacter* agar (MCCDA, Modified Charcoal Cefoperazone Deoxycholate Agar (Oxoid (Thermo Scientific) Item CM0739) and incubated microaerobically using anaerobic jars and 'CampyGen' atmosphere generating sachets (Fisher Scientific Item CN025A); developing a gas mixture of 5% O_2, 10% CO_2 and 85% N_2, for 48 hours at 37°C prior to use.

2.2.2 Bacterial Strains included in Phenotypic Study

To investigate the phenotypic evidence for potential host adaption, the six selected ST403CC pig origin *C. jejuni* isolates were tested alongside two human clinical *C. jejuni* isolates and four porcine *C. coli* isolates which were selected to provide an appropriate basis for comparison, as detailed in Table 2.1.

The isolates included in the study were each assessed for: motility; autoagglutination; adhesion; invasion; and sensitivity to hydrogen peroxide; superoxide (pyrogallol); and bile acids cholic acid, deoxycholic acid and chenodeoxycholic acid.

Species	Strain	Country of Isolation	Host	Date of Isolation	Sequence Type	Sequence Type Complex
C. jejuni	81116	UK	Human	1981	267	283
C. jejuni	11168	UK	Human	1977	43	21
C. jejuni	857	UK	Pig	2000	270	403
C. jejuni	549.1	UK	Pig	1999	403	403
C. jejuni	623	UK	Pig	1999	552	403
C. jejuni	304	UK	Pig	1999	551	403
C. jejuni	484	UK	Pig	1999	435	403
C. jejuni	444	UK	Pig	1999	553	403
C. coli	99/321	Denmark	Pig	1999	ND*	ND*
C. coli	03/121	UK	Pig	2003	ND*	ND*
C. coli	03/103	UK	Pig	2003	ND*	ND*
C. coli	03/317	UK	Pig	2003	ND*	ND*

Table 2.1: Isolates Included in Phenotypic Studies

C. jejuni isolates 81116 and 11168 were used as control strains due to being amongst the most well-characterised *C. jejuni* strains, both being from human cases and having been defined as low and high invasive strains, respectively. ST403CC *C. jejuni* isolates were all recovered from pig faecal samples during an abattoir survey, and each represent different sequence types within the ST403 clonal complex. *C. coli* isolates included are from the same host type and same survey style, and from around the time frame of recovery of the ST403CC *C. jejuni* isolates.

ND* these isolates have not yet been typed by MLST, therefore these details are currently unknown.

2.2.3 Motility

Two methods were used to observe motility, firstly the hanging drop method, which allows a simple positive or negative result for motility, and secondly an agar plating method which allows observation of the degree of swarming motility displayed by each isolate.

- Hanging Drop Method. Wells were created on glass slides (using petroleum jelly to create a barrier) into which was placed a drop of bacterial suspension. The bacterial suspension was created by emulsifying a single colony of growth from agar plate into sterile PBS (phosphate buffered saline Oxoid (Thermo Scientific) Item BR0053). Slides were viewed under x400 magnification in order to observe darting motility.
- Swarming Motility Method. A quantifiable method, adapted from Novik, Hofreuter & Galán (2010) was used to compare the relative motility of the isolates. Bacterial growth was recovered from the agar plate and suspended

in sterile PBS, bacterial suspensions were subsequently adjusted to an optical density of approximately 0.1 at 600nm. Subsequently 2µl of the standardised bacterial suspension was stabbed into the centre of a soft agar plate (Mueller-Hinton broth with 0.4% w/v agarose), before being incubated for 48 hours at 42°C, following which the diameter of the zone of growth was measured to observe the motility displayed by each isolate. Incubation was carried out at 42°C for this assay, due to inconsistent results being observed at 37°C during initial runs; in some cases isolates which had been demonstrated to be motile under microscopic observation were returning negative results for swarming motility. 42°C is the optimal temperature for *C. jejuni* and *C. coli* and when trialled as a condition for motility provided more consistent results in this case, despite 37°C being used in other work including Golden & Acheson (2002) and Novik, Hofreuter & Galán (2010).

2.2.4 Autoagglutination

Autoagglutination was determined using the method developed by Misawa & Blaser (2000): bacteria were grown for 48 hours before being swabbed from the plates and suspended in PBS. The bacterial suspension was adjusted to an optical density of 1.0 at 600nm (Absorbance at 600nm; A600nm) in 2ml PBS and subsequently incubated without agitation at 37°C for 24 hours. Following incubation, the upper layer of the solution was carefully aspirated, and the *A600nm* was measured again. The level of autoagglutination was observed by comparing the starting optical density value with the optical density of the upper layer of the bacterial suspension following incubation; the principle for autoagglutination is based upon the premise that when autoagglutination occurs the resulting 'clumps' of bacteria drop to the bottom of the solution. Some settling would occur regardless of autoagglutination, but a large change in optical density is caused by autoagglutination (Misawa & Blaser, 2000; Golden & Acheson, 2002). Autoagglutination results were observed as the change in optical density between the starting point and the 24 hour point.

2.2.5 Adhesion & Invasion

Adhesion and invasion efficiencies were assessed using a gentamicin protection assay with Caco-2 cells.

Bacteria were grown on blood agar comprised of 20% v/v lyophilised horse blood (Fisher Scientific Item SR050C) in Columbia Blood Agar Base (Oxoid (Thermo Scientific) Item CM0331) for 48 hours at 37°C before being suspended into PBS to be adjusted for introduction to Caco-2 cells.

Caco-2 cells were maintained in tissue culture flasks, splitting to new cells routinely (approximately once per week, as dictated by growth of cells) with media changed approximately every three days when cells were not split to new flasks. Two media types of slightly different composition were used to maintain and run assays with Caco-2 cells, all media and reagents were pre-warmed prior to use to avoid damage to the cells.

- Growth medium consisted of DMEM (Dulbecco's Modified Eagle's Medium; Sigma-Aldrich Item D6429), supplemented with 1% v/v non-essential amino acids (Sigma-Aldrich Item M7145)**;** 20% v/v foetal bovine serum (Sigma-Aldrich Item F7524); and 1% penicillin-streptomycin antibiotic solution (Sigma-Aldrich Item P0781).
- Infection medium was prepared in the same way as growth medium, with the exception that no antibiotics were added (DMEM, with 10% v/v foetal bovine serum, and 1% v/v non-essential amino acids solution added.

When either removing cells to new growth flasks, or to begin gentamicin protection assay, growth media was first removed from flask, before twice gently rinsing cells with sterile PBS (Dulbecco's Phosphate Buffer Saline solution; Sigma-Aldrich Item D8537) to remove dead cells. Cells were then disassociated from the flask using TrypLE Express (Stable Trypsin-like Enzyme liquid; Fisher Scientific Item VX12605028) by adding 5-10ml and incubating for approximately eight minutes until cells are in suspension. Cells were then transferred to a falcon tube and centrifuged to form a pellet (5 minutes at 150xG) before being re-suspended into

5ml of growth media and counted in order to maintain cells at $2x10^5$ cells/ml in $25cm^3$ flask, or $7x10^5$ cells/ml in $75cm^3$ flask in growth media, or for experimental use, seeded to a 24 well plate at 500µl per well with $2x10^5$ cells/ml in infection media. For experimental procedures Caco-2 cells were then grown to semi-confluency (approximately 80% coverage of well, achieved between 24-48 hours) before being ready for assay. In all cases, Caco-2 cells were incubated at 37C in 5% CO_2.

To run the gentamicin protection assay, bacterial cells were recovered from agar plates by swabbing, and suspended in PBS, then adjusted to an optical density of approximately 0.5 at 600nm before being diluted to provide an infection solution of 3.5ml with optical density of 0.1 at 600nm. This equated to a predicted viable count of approximately $1x10^8$ cfu/ml (colony forming units per milliletre); actual bacterial counts for the inoculum were determined retrospectively using Miles-Misra plate counts.

The standardised, diluted bacterial solution was then applied in 500µl aliquots to the Caco-2 cells in 24 well plates, following the removal of cell growth media and the cell layers being washed gently three times with PBS to remove any dead cells. In each assay, 500µl of bacterial suspension was applied in triplicate to wells on two plates – one for assessment of bacterial attachment and one for bacterial invasion into the Caco-2 cells.

The 24 well plates were then incubated for three hours at 37°C with 5% CO_2, following which the first plate was assessed for bacterial attachment to the cells; the infection media was careful aspirated from each well, which were then gently washed three times with PBS before being lysed with 1% Triton-X100 (Sigma-Aldrich Item T8787) so that bacterial counts could be then determined using Miles-Misra plate counts. At the same three hour time point, the invasion assay plate was also aspirated of media and washed with PBS, before being subjected to 500µl of the 'infection media' solution containing 250µg/ml of gentamicin and subsequently being incubated for a further two hours under the same conditions in order to kill any external bacteria. Following the additional two hour incubation, cells were

washed and lysed and bacterial counts determined as described for the association assay.

Each assay included three replicates of each isolate and assays were performed at least in triplicate. Each assay also included three wells of a no-inoculum blank to ensure that results were not confounded by contamination of the intestinal cells or the infection media, and three wells of the low invader isolate *C. jejuni* to confirm the reproducibility of the assay. All bacterial counts using the Miles-Misra technique were plated onto MCCDA agar and incubated at 37°C for 48 hours before counts were recorded.

Results were presented as percentage adhesion and percentage invasion; that is, the percentage of the original inoculum which was observed to have attached to or internalised into the Caco-2 cells. Percentage adhesion and percentage invasion were calculated as follows:

$$\% \text{ Adhesion} = \frac{\text{Number of bacteria adhered to cells}}{\text{Number of bacteria in inoculum}} * 100$$

[+] Where adhered bacteria = total number of bacteria associated with cells – number of bacteria invaded into cells

$$\% \text{ Invasion} = \frac{\text{Number of bacteria invaded into cells}}{\text{Number of bacteria in inoculum}} * 100$$

2.2.6 Resistance to Oxidative & Bile Acid Stressors

Resistance to hydrogen peroxide, and the bile acid stressors cholic acid, deoxycholic acid and chenodeoxycholic acid was observed using a disc diffusion assay adapted from Baillon *et al* (1999). Isolates were recovered from frozen stocks and incubated at 37°C for 24 hours before being sub-cultured in duplicate onto fresh MCCDA plates and incubated overnight to provide lawns of bacterial growth. Cultures were then re-suspended into Mueller-Hinton broth and adjusted to an optical density at 600nm of 0.4. For each assay, 200µl of bacterial suspension was added to 4ml of soft Mueller-Hinton agar (Oxoid (Thermo Scientific) Item CM0405; plus 0.4% w/v Agarose (Fisher Scientific Item BP1356) which had been cooled to approximately 35°C to avoid damaging the bacteria. The mixture was mixed thoroughly before

being poured over thin layer Mueller-Hinton agar plates and allowed to set. Sterile 3mm filter discs were then applied to the surface of the bacteria infused soft agar and treated with 3µl each of a range of six concentrations of the stressor; the concentration ranges of which are shown below in Table 2.2. Plates were then incubated as previously described and the zones of inhibition around the discs were observed after 48 hours. In each assay isolates were tested in triplicate and for each isolate the assays were repeated at least twice.

Chemical	Concentrations Tested					
Hydrogen Peroxide (%v/v)	1	3	5	10	15	20
Chenodeoxycholic Acid (mg/ml)	1	3	5	10	15	20
Deoxycholic Acid (mg/ml)	3	5	10	12	13	15
Cholic Acid (mg/ml)	0.5	1	2	3	5	10

Table 2.2: Concentrations of chemicals used for stress assay testing of *C. jejuni* and *C. coli* isolates.

Ranges were selected based upon previously described experiments in literature, so that the results for a single representative concentration of each stressor could be presented.

Resistance to the superoxide producer pyrogallol was tested using a protocol adapted from that of Champion *et al* (2009). Isolates were recovered from frozen stocks and then sub-cultured as described for the previously described stress assays, before being suspended in Mueller-Hinton broth and adjusted to an optical density of 0.4 at 600nm. 100µl of the bacterial suspension was then pipetted onto a dried Mueller-Hinton agar plate and spread to cover the surface of the plate. Once dried, a sterile filter disc was applied in the centre of the plate and inoculated with 3µl of 1M pyrogallol. Zones of inhibition created around the discs were observed after 48 hours incubation at 37°C and the assay was repeated at least three times for each isolate. In the assay used by Champion *et al* (2009), the agar was supplemented with catalase (1000 units/ml); this was initially emulated herein, however adding the catalase whilst ensuring sterility and complete dispersion through the agar was problematic, and parallel tests with and without the addition of catalase showed no noticeable variation in results and as such the addition of catalase was subsequently abandoned.

2.2.7 Statistical Analyses

In order to consider the impact of the results produced from these assays, statistical testing was applied. Statistical tests applied were the ANOVA test (one-way, unpaired, parametric Analysis of Variance, with multiple comparison testing against control values) and standard T-test (unpaired, parametric, two-tailed, with Welch's correction). These were selected to investigate any differences between the porcine origin ST403CC *C. jejuni* isolates, and non-porcine origin, non-ST403CC *C. jejuni* isolates, and porcine origin *C. coli*; and to accept or reject the following specific null hypotheses:

- ST403CC *C. jejuni* isolates do not show evidence of the potential to cause disease in humans
- ST403CC *C. jejuni* isolates do not exhibit higher pathogenic potential, as determined in laboratory tests, than non-ST403CC *C. jejuni* and *C. coli* isolates
- ST403CC *C. jejuni* isolates do not exhibit a distinct phenotypic profile compared to non-ST403CC *C. jejuni* and *C. coli*

Analysis using one-way ANOVA compared each category - each isolate, in this case - looking for significance in variation between the categories. Parametric ANOVA assumes that the data being tested fits a normal distribution. This assumption is often made when the true distribution of real-world results is not known, in relation to the 'central limit theorem', whereby random samples under mild conditions tend to have normal distribution. In this instance, distribution was checked for a sample of the experiments conducted and normal distribution was observed. Additionally, multiple comparisons were applied to the data as part of the analysis of variance in order to establish whether statistically significant variation exists between a selected control strain - *C. jejuni* 81116 - and each of the other individual isolates. In contrast, the T-test was used to compare 'condition a' vs 'condition b' and analyse the likelihood that the difference between groups is statistically significant, or due simply to chance. This was used to compare scores for groups of isolates: porcine *vs* non-porcine isolates; porcine *vs* non-porcine isolates for *C. jejuni* only; *C.*

coli vs C. jejuni regardless of host. All statistics presented refer to a confidence interval of 95% (P=0.05).

2.3 Results

A range of phenotypic tests were applied in order to assess the virulence potential of a group of potentially porcine host adapted *C. jejuni* isolates. Also included for comparison were two human *C. jejuni* isolates and four porcine *C. coli* isolates. Results are presented below for motility, autoagglutination, adherence to and invasion into Caco-2 cells, resistance to hydrogen peroxide and a superoxide producer, and resistance to bile acids cholic acid, deoxycholic acid and chenodeoxycholic acid. Unless otherwise stated, values presented are mean values calculated from representative scores; experiments were completed at least in triplicate, and where considerable variation was observed additional replicates were completed to determine more accurate, representative results. Results are presented with error bars using the standard deviation calculated for the mean score from the range observed for each isolate individually.

2.3.1 Motility

Motility was initially observed qualitatively using the hanging drop method. All isolates displayed motility under microscopic activity, including both reference strains and all test isolates. Subsequently, motility was subjected to a more quantitative method of study using swarming motility as described above. Results for swarming motility are presented in Figure 2.1, overleaf. Control strains *C. jejuni* 81116 and 11168 recorded motility diameters of 19mm and 21mm respectively. For the test strains, the observed range of score ranged from 3-30mm and showed notable variation between isolates. Overall *C. coli* showed lower motility (range 3-11.mm, mean 7.25mm) than the control strains (range 19-21mm, mean 20mm) and the ST403CC *C. jejuni* isolates (range 3-30mm, mean 14.79mm). The *C. coli* isolates recorded less variation in motility across the selected isolates (3-11.67mm) than was observed in *C. jejuni* (3-30mm), however the *C. coli* isolates displayed more intra-strain variation between tests (the mean standard deviation 2.36, range of standard deviation across four isolates 0.0-6.03) compared to *C. jejuni* ST403CC

isolates (mean standard deviation 1.67, range 0.0-4.0 across six isolates) and the reference *C. jejuni* isolates (zero standard deviation, each isolate replicated exact scores across repeat experiments). No clear trend emerged from the data presented in Figure 2.1 to indicate a consistent pattern of motility for ST403CC *C. jejuni* isolates; the ST403CC isolates were demonstrated to be motile but did not appear distinct from other *C. jejuni* or *C. coli* isolates.

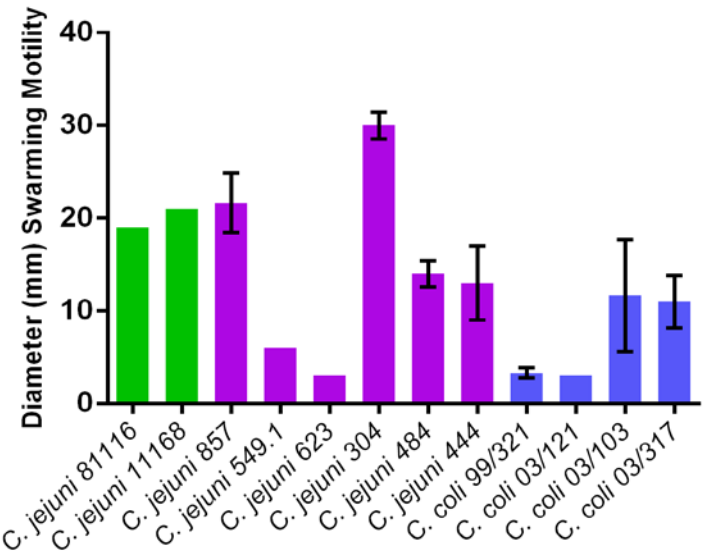

Figure 2.1: Swarming Motility of *C. jejuni* & *C. coli* Isolates.

Swarming motility was determined for *C. jejuni* and *C. coli* isolates using soft agar stabs. Motility was measured as the diameter of the zone of swarming motility in millimetres. *C. jejuni* control isolates are indicated by green colour and *C. coli* isolates by blue. ST403CC *C. jejuni* isolates are indicated by purple colour. Error bars show standard deviation from the mean motility observed; where no error bars are shown the same result was observed across repeated experiments.

One-way ANOVA for swarming motility was found to be statistically significant ($P<0.001$ (to be significant $P\leq0.05$)). Means *vs* control analysis revealed significant variation between the reference strain control *C. jejuni* 81116 and *C. jejuni* 549.1, 623, 304, plus *C. coli* 99/321 and 03/121 - all of which recorded significantly lower motility, except 304 which displayed higher motility than reference isolate. T-test analysis revealed that, despite appearances (Figure 2.1), variation between groups was also statistically significant, for source only (porcine *vs* non-porcine isolates, regardless of species; $P<0.0001$); for source within *C. jejuni* isolates (porcine *vs* non-

porcine isolates from *C. jejuni* only; P=0.0361); and for species (*C. jejuni vs C. coli*, regardless of source; P=0.0033). Considerable strain-strain variation was observed for degree of motility, though all tested isolates were motile, and the ST403CC *C. jejuni* isolates displayed, on average, lower motility than typical *C. jejuni* strains, but higher motility than porcine *C. coli* included in the study.

2.3.2 Autoagglutination

Figure 2.2 shows the observed results for autoagglutination, recorded as the change in OD600nm after 24 hour incubation without agitation; with a larger change indicating a higher level of autoagglutination. Variation was observed across the twelve tested isolates (range 0.26-0.92) with the lowest scores recorded for the reference strains *C. jejuni* 81116 and 11168 (mean scores of 0.26 and 0.35 respectively). Autoagglutination scores for the ST403CC *C. jejuni* isolates ranged from 0.36-0.90 (mean change 0.67) whilst scores for the *C. coli* isolates ranged from 0.49-0.92 (mean 0.70). The ST403CC *C. jejuni* isolate 304 exhibited notable lower autoagglutination than the other 'test' isolates, recording autoagglutination levels more consistent with those of the reference *C. jejuni* strains.

One-way ANOVA determined that the variation between samples for autoagglutination was statistically significant (P>0.0001) and multiple comparison determined that variation from control strain *C. jejuni* 81116 was statistically significant for every included isolate with the exception of *C. jejuni* isolates 11168 and 304. T-test analysis showed significant difference between groups for porcine vs non-porcine control isolates, both across both species and for *C. jejuni* only (P<0.0001 for both), but there was no significant difference between *C. jejuni* and *C. coli* groups (P=0.0820).

It was demonstrated that the studied ST403CC *C. jejuni* isolates all exhibited autoagglutination, and to a higher degree than the reference *C. jejuni* isolates.

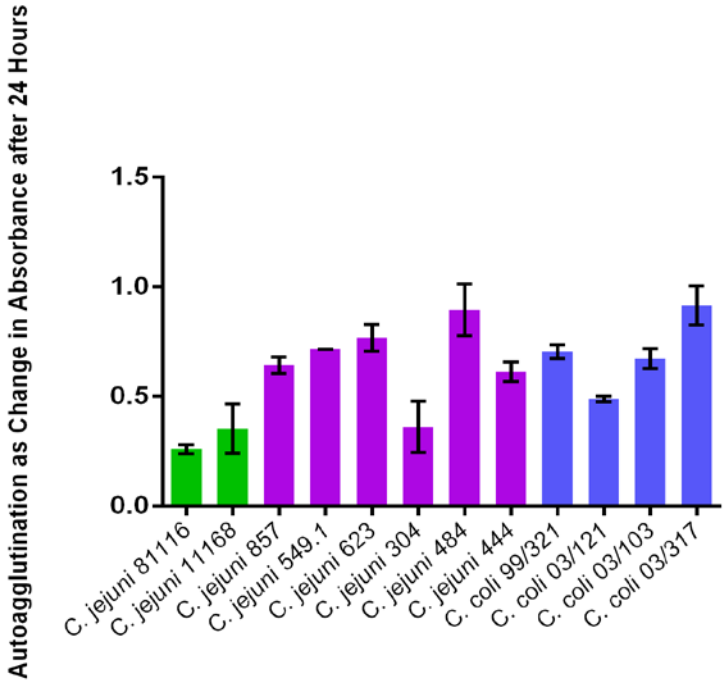

Figure 2.2: Autoagglutination of *C. jejuni* & *C. coli* Isolates.

The level of autoagglutination was observed by comparing the change in optical density between the start point of the experiment, and following 24 hours incubation without agitation; the values plotted are the mean changes in optical density for each isolate.

C. jejuni control isolates are indicated by green colour and *C. coli* isolates by blue. ST403CC *C. jejuni* isolates are indicated by purple colour. Error bars show standard deviation from the mean motility observed; where no error bars are shown the same result was observed across repeated experiments.

2.3.3 Adherence to and Invasion into Caco-2 Cells

Figure 2.3, overleaf, shows the adhesion to Caco-2 cell surfaces for *C. jejuni* and *C. coli* isolates, three hours post infection**.** Values are presented as the percentage of the inoculum adhered to the cells, calculated from the mean results for two representative values.

As described above, each isolate was tested three times per assay, and each assay was repeated at least in triplicate: from this, two representative results were selected, then the percentage adherence was calculated for each (from the mean inoculum), and the mean of these two values was taken for presentation from this.

Substantial strain-strain variation was observed; the range of mean values was $1.71x10^6$-$7.17x10^6$ equating to a range of percentages of initial inoculum of 0.24-3.33%.

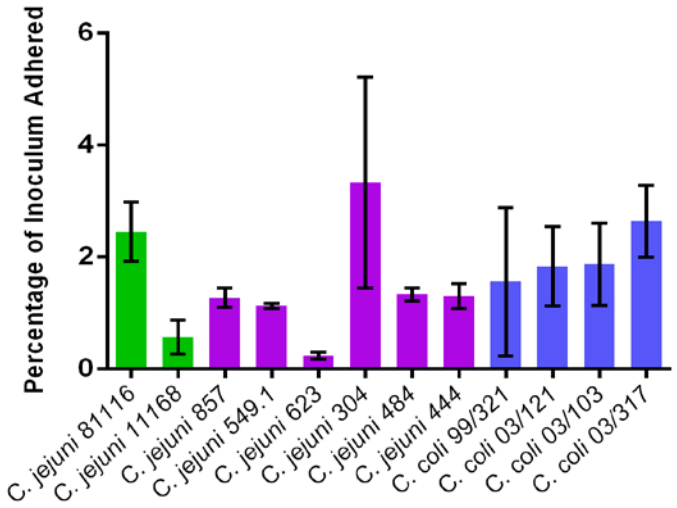

Figure 2.3: Adherence of *C. jejuni* & *C. coli* Isolates to Caco-2 Cells

Bacterial cells adhered to Caco-2 cells two hours post infection, as determined by retrospective bacterial counts and presented as a percentage of the initial inoculum for each isolate.

C. jejuni control isolates are indicated by green colour and *C. coli* isolates by blue. ST403CC *C. jejuni* isolates are indicated by purple colour. Error bars show standard deviation from the mean motility observed.

Both the highest and lowest observed percentage adherence were isolates within the ST403CC *C. jejuni* group. Reference isolates *C. jejuni* 81116 and 11168 recorded adherence percentages of 2.46 and 0.57 respectively, whilst the *C. coli* isolates had an average percentage of inoculum adhered of 1.98% (range 1.56-2.64%) and the ST403CC *C. jejuni* isolates average was 1.44% (range 0.24 to 3.33%). No obvious trend was observed from observation of Figure 2.3, and this was confirmed by statistical analyses: one-way ANOVA determined no significant variation between strains (P=0.0642), with no significant results from multiple comparison analysis. Furthermore no statistically significant variation was observed for adhesion between groups by T-test analysis (porcine/non-porcine (P=0.9122); porcine/non-

porcine *C. jejuni* (P=0.8312); *C. jejuni vs C. coli* (P=0.2016)). Although the statistical analysis denies significant variation between strains (range 0.24-3.33, with mean of 1.63), it is however observable from Figure 2.3, that considerable variation exists between strains, and that each of the tested isolates was able to successfully adhere to Caco-2 cells.

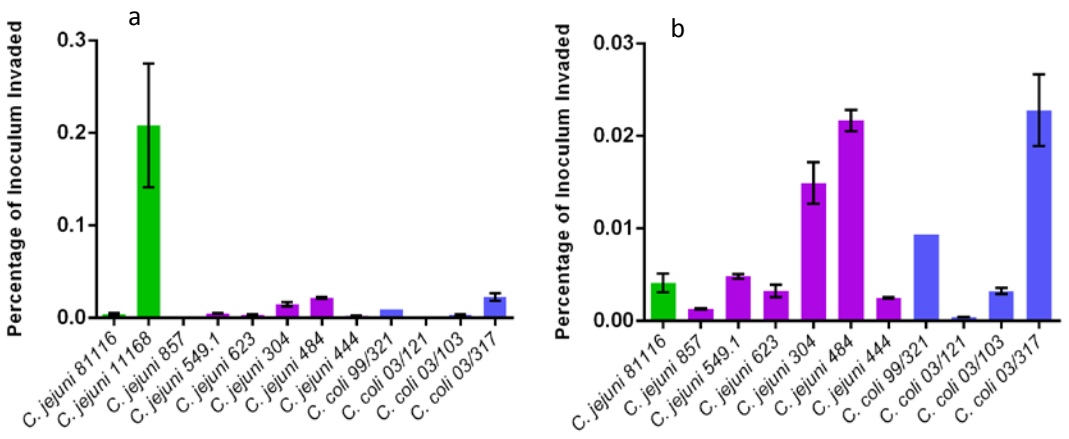

Figure 2.4: Internalisation of *C. jejuni* & *C. coli* Isolates into Caco-2 Cells

Bacterial cells internalised into Caco-2 cells five hours post infection, as determined by retrospective bacterial counts and presented as a percentage of the initial inoculum for each isolate.

a) includes all twelve isolates tested *C. jejuni* and *C. coli* isolates.

b) excludes the result for *C. jejuni* 11168 as it is so much more invasive than the others it makes viewing the variation between the remaining eleven isolates difficult.

C. jejuni control isolates are indicated by green colour and *C. coli* isolates by blue. ST403CC *C. jejuni* isolates are indicated by purple colour. Error bars show standard deviation from the mean motility observed; where no error bars are shown the same result was observed across repeated experiments.

Figure 2.4 displays retrospective bacterial counts for the number of bacteria internalised into Caco-2 cells a total of five hours post infection (three hour adhesion phase followed by two hour gentamicin protection phase to remove external bacteria) as a percentage of the initial inoculum. Figure 2.4a shows the results for all twelve isolates; whilst Figure 2.4b shows the same data with the exemption of the highly invasive isolate *C. jejuni* 11168, as the data in part is skewed by its presence, and its exclusion in Figure 2.4b allows for greater resolution in the remaining eleven datasets. Considerable variation was observed across the

twelve isolates, with reference strains *C. jejuni* 81116 and 11168 recording 0.04% and 0.21% respectively, whilst the ST403CC *C. jejuni* isolates ranged from 0.001-0.022% (mean 0.008%) and the *C. coli* isolates ranged from 0.0004-0.023% with a mean of 0.009% of inoculum internalised**.**

Invasion data was found to be statistically significant in variation across the group by one-way ANOVA (P=<0.0001), although via multiple comparison analysis the only strain to differ significantly from the low invasive reference isolate *C. jejuni* 81116 was the high invasive reference isolate *C. jejuni* 11168. Additionally, invasion efficiency was not found to show statistically significant variation between groups: by source only (P=0.2118); by source within *C. jejuni* (P=0.2130); by species (P=0.2087).

Each of the isolates was found to successfully invade Caco-2 cells *in vivo*. Considerable variation was observed between isolates however no association was observed between host or species and invasive ability. ST403CC *C. jejuni* isolates are capable of invasion into Caco-2 cells, with variation observed between strains within the group.

2.3.4 Sensitivity to Oxidative Stress

The sensitivity of ST403CC *C. jejuni* isolates, and other *C. jejuni* and *C. coli* isolates, to oxidative stresses were observed for hydrogen peroxide and superoxide, via the superoxide producer pyrogallol.

2.3.4.1 Sensitivity to Hydrogen Peroxide

Sensitivity to hydrogen peroxide was observed as zones of inhibition on plates when exposed to a range of concentrations. The concentration which produced the most consistent, reproducible results was 3% v/v hydrogen peroxide, and the mean diameters of the zones of inhibition for this concentration are presented in Figure 2.5.

Some strain-strain variation was observed across the twelve isolates, however no clear pattern was observed between the different groups of isolates. The range across the experiment was 18-31.67mm, with reference isolates *C. jejuni* 81116 and

11168 scoring 25mm and 31.67mm respectively. Zones of inhibition for the ST403CC *C. jejuni* isolates ranged from 23mm to 31.67mm (mean of 25.61mm) whilst the *C. coli* isolates ranged from 18-28mm (mean 23.75mm).

For the H_2O_2 assay, statistical analysis found that, by one-way ANOVA, variation between strains was statistically significant (P<0.0001), and that some isolates (*C. jejuni* 11168 and 623 and *C. coli* 03/121) varied significantly from the control strain *C. jejuni* 81116: *C. coli* 03/121 with a significantly smaller zone of inhibition - being more resistant, whilst *C. jejuni* 11168 and 623 were significantly larger zones - being more sensitive. T-test analysis confirmed that there was no significant difference between groups: host origin (P=0.1055); host origin within *C. jejuni* (P=0.2020); *C. jejuni vs C. coli* (P=0.1073).

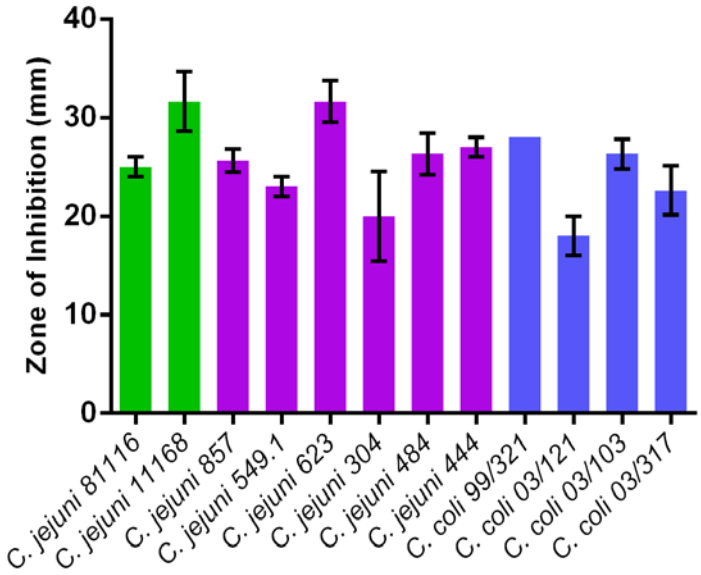

Figure 2.5: Sensitivity to Hydrogen Peroxide of *C. jejuni* & *C. coli* Isolates

H_2O_2 sensitivity, recorded as zones of inhibition of growth (diameter, mm) using 3% v/v Hydrogen Peroxide.

C. jejuni control isolates are indicated by green colour and *C. coli* isolates by blue. ST403CC *C. jejuni* isolates are indicated by purple colour. Error bars show standard deviation from the mean motility observed; where no error bars are shown the same result was observed across repeated experiments.

2.3.4.2 Sensitivity to Superoxide

Sensitivity to superoxide was assessed by observing zones of inhibition created when isolates were exposed to the superoxide producer pyrogallol (1M). Some

variation was observed between isolates, however no distinct pattern of sensitivity was observed for species or host type, as depicted in Figure 2.6. Zones of inhibition across the group ranged from 31.80-43.80mm, with diameters of 37.60mm and 32.40mm for *C. jejuni* 81116 and 11168. The range for ST403CC *C. jejuni* isolate sensitivity was larger than non-ST403CC *C. jejuni* (31.80-43.80mm, mean 37.93mm), with ST403CC isolates again representing both the highest and lowest recorded scores. Zones of inhibition for *C. coli* isolates ranged 32-38.40mm (mean 34.85mm).

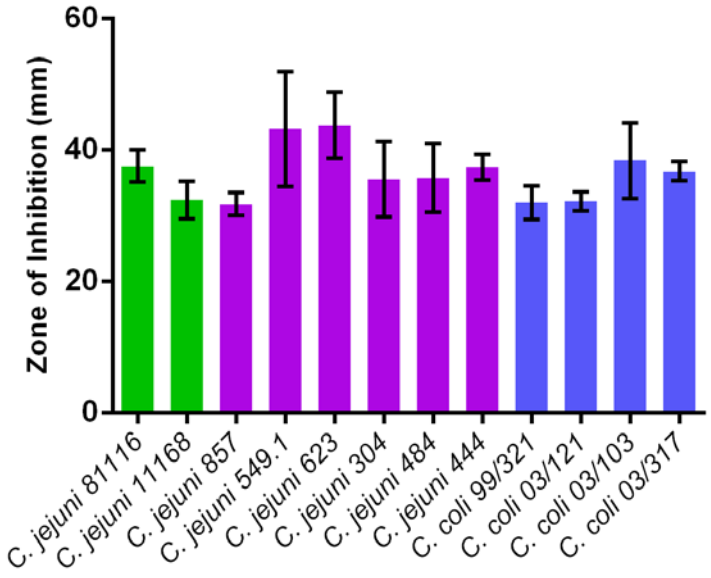

Figure 2.6: Sensitivity to Superoxide of *C. jejuni* & *C. coli* Isolates

Superoxide sensitivity, recorded as zones of inhibition (mm) created when exposed to the superoxide producer Pyrogallol (1M).

C. jejuni control isolates are indicated by green colour and *C. coli* isolates by blue. ST403CC *C. jejuni* isolates are indicated by purple colour. Error bars show standard deviation from the mean motility observed; where no error bars are shown the same result was observed across repeated experiments.

The between strain variation observed in the pyrogallol assay was confirmed to be significant by statistical analysis (one-way ANOVA, P=0.0001), although the multiple comparisons method found no statistically significant variation in any strain compared to the control strain *C. jejuni* 81116. Additionally there was no significant variation between groups: host origin (P=0.2507); host origin within *C. jejuni* (P=0.0899); *C. jejuni vs C. coli* (P=0.0842). ST403CC *C. jejuni* isolates showed similar

superoxide sensitivity to other *C. jejuni* and *C. coli* isolates when treated with pyrogallol.

2.3.5 Sensitivity to Bile Acids

The sensitivity of *C. jejuni* and *C. coli* isolates to the major bile acids in humans were assessed using disc diffusion methods.

2.3.5.1 Sensitivity to Cholic Acid

Sensitivity to cholic acid was assessed by examining zones of inhibition created upon exposure to cholic acid. A range of concentrations were used initially, as described in Table 2.2, from the cholic acid assay, the highest concentration used (10mg/ml) was selected for presentation, as it provided the most consistent results across repeated tests; mean results are shown in Figure 2.7.

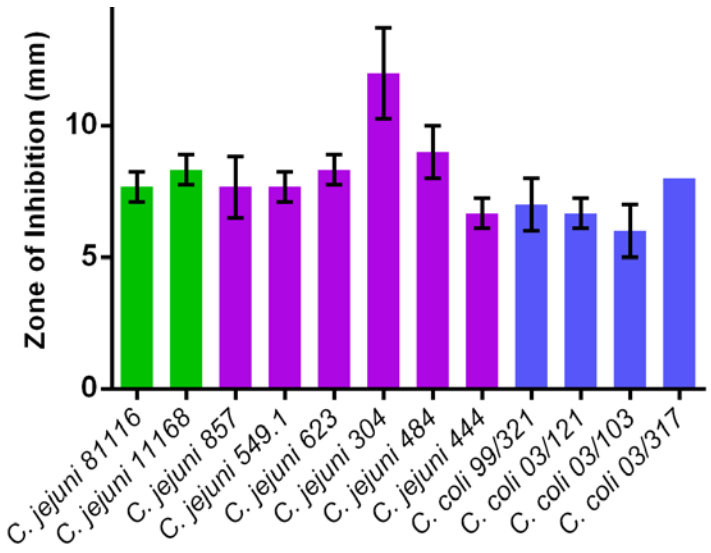

Figure 2.7: Sensitivity to Cholic Acid of *C. jejuni* & *C. coli* Isolates

Sensitivity to cholic acid (10mg/ml) as observed by diameter of zones of inhibition (mm) observed.

C. jejuni control isolates are indicated by green colour and *C. coli* isolates by blue. ST403CC *C. jejuni* isolates are indicated by purple colour. Error bars show standard deviation from the mean motility observed; where no error bars are shown the same result was observed across repeated experiments.

The control strains recorded similar zones of inhibition (*C. jejuni* 81116 7.67mm and *C. jejuni* 11168 8.33mm), which were also consistent with the zones observed for ST403CC *C. jejuni* isolates (range 6.67-12mm, mean 8.56mm), whereas the *C. coli* isolates recorded smaller zones of inhibition (6mm-8mm, mean 6.92mm).

This was confirmed by statistical analysis; one-way ANOVA was statistically significant for variation between strains ($P<0.0001$), and as would be expected, multiple comparison of test means against control revealed statistically significant variation from *C. jejuni* 81116 for *C. jejuni* 304 but not for any other isolate. T-test analysis revealed no significant variation between groups for pig origin *vs* non-pig origin across species ($P=0.8134$) or between pig origin *vs* non-pig origin within *C. jejuni* isolates only ($P=0.3029$), however statistically significant variation ($P=0.0023$) was observed between species (*C. jejuni* vs *C. coli*) irrespective of isolate origin.

With the exception of *C. jejuni* 304, which was notably more sensitive, the ST403CC *C. jejuni* isolates showed similar levels of cholic acid sensitivity to the *C. jejuni* reference isolates; all of which were less resistant to cholate stress than the *C. coli* isolates.

2.3.5.2 Sensitivity to Deoxycholic Acid

As for cholic acid, deoxycholic acid sensitivity was assessed using a range of concentrations and measured as the diameter of zones of inhibition. Figure 2.8 shows the mean diameters observed using the selected concentration 13mg/ml; this concentration was selected as the representative concentration due to the reproducibility of results.

The observed range across all isolates was 6.33-8mm, with reference isolates *C. jejuni* 81116 and 11168 recording zones of inhibition of 6.33mm and 7mm respectively. In comparison the range for the *C. jejuni* ST403CC isolates was 6.67-8mm with a mean of 7.45mm, and for the *C. coli* isolates was 6.67-7mm (mean 6.86mm).

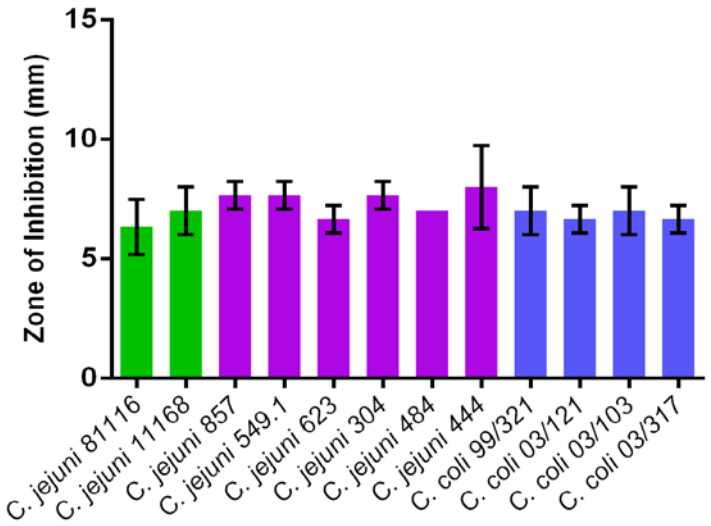

Figure 2.8: Sensitivity to Deoxycholic Acid of *C. jejuni* & *C. coli* Isolates

Sensitivity to deoxycholic acid (13mg/ml) as observed by diameter of zones of inhibition (mm) observed.

C. jejuni control isolates are indicated by green colour and *C. coli* isolates by blue. ST403CC *C. jejuni* isolates are indicated by purple colour. Error bars show standard deviation from the mean motility observed; where no error bars are shown the same result was observed across repeated experiments.

Results were similar across the twelve isolates, with the smallest diameter (most resistant) isolate being *C. jejuni* 81116. This observation was supported by statistical analysis which confirmed no significant variation between strains via one-way ANOVA (P=0.4453), with no significant variation from the control strain during multiple comparison analysis, as well as no significant variation between groups when analysed by T-test (pig origin *vs* non-pig origin across species (P=0.2771); pig origin *vs* non-pig origin within *C. jejuni* isolates only (P=0.1375); *C. jejuni vs C. coli* (P=0.1520)).

2.3.5.3 Sensitivity to Chenodeoxycholic Acid

Sensitivity to chenodeoxycholic acid was assessed in the same manner as cholic acid and deoxycholic acid, and the concentration selected for presentation was 15mg/ml. Chenodeoxycholic acid sensitivity results are displayed in Figure 2.9; zones of inhibition across the experiment ranged from 5.33-9mm, with diameters of 6mm for *C. jejuni* 81116 and 6.33mm for *C. jejuni* 11168, compared with a range of 5.33-7.33 (mean 6.28mm) for the ST403CC *C. jejuni* isolates, and 5.67-9.00mm (mean 7.17mm) for the *C. coli* isolates. Once again for chenodeoxycholic acid

minimal variation was observed across the isolates. Statistical analysis determined that variation between the strains was statistically significant (P<0.0001), however multiple comparison test against the control strain revealed that only most sensitive isolate *C. coli* 03/317 varied from the control strain *C. jejuni* 81116 statistically significantly. T-test analysis did not determine any significant variation between groups (pig origin *vs* non-pig origin across species (P=0.1146); pig origin *vs* non-pig origin within *C. jejuni* isolates only (P=0.7075); *C. jejuni vs C. coli* (P=0.0652)).

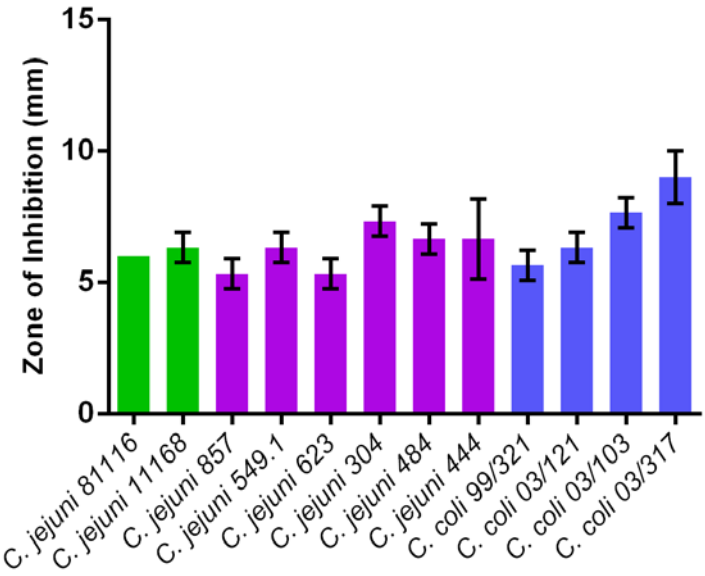

Figure 2.9: Sensitivity to Chenodeoxycholic Acid of *C. jejuni* & *C. coli* Isolates

Sensitivity to chenodeoxycholic acid (15mg/ml) as observed by diameter of zones of inhibition (mm) observed.

C. jejuni control isolates are indicated by green colour and *C. coli* isolates by blue. ST403CC *C. jejuni* isolates are indicated by purple colour. Error bars show standard deviation from the mean motility observed; where no error bars are shown the same result was observed across repeated experiments.

Overall, for the bile salt acid assays, variation was less than for other assays, however the difference across the group was statistically significant (95% C.I.) for both chenodeoxycholic acid and cholic acid. Species appeared to have an effect only for the cholate assay, and porcine *vs* non-porcine origin did not appear important in differences in bile acid sensitivity. In each of the three bile acid tests, the ST403CC *C. jejuni* group demonstrated some variation between isolates, but recorded sensitivity levels similar to other *C. jejuni* and *C. coli* isolates.

2.3.6 Summary of Experiments Characterising ST403CC *C. jejuni* Isolates

Figure 2.10 provides an overview of the phenotyping results as a 'heat-map'. The colour gradient indicates the virulence potential of the isolate; for each experiment, a darker colour signifies the more virulent phenotype (such as high invasion or motility, or low sensitivity to a stressor), whereas a paler colour indicates a less virulent phenotype.

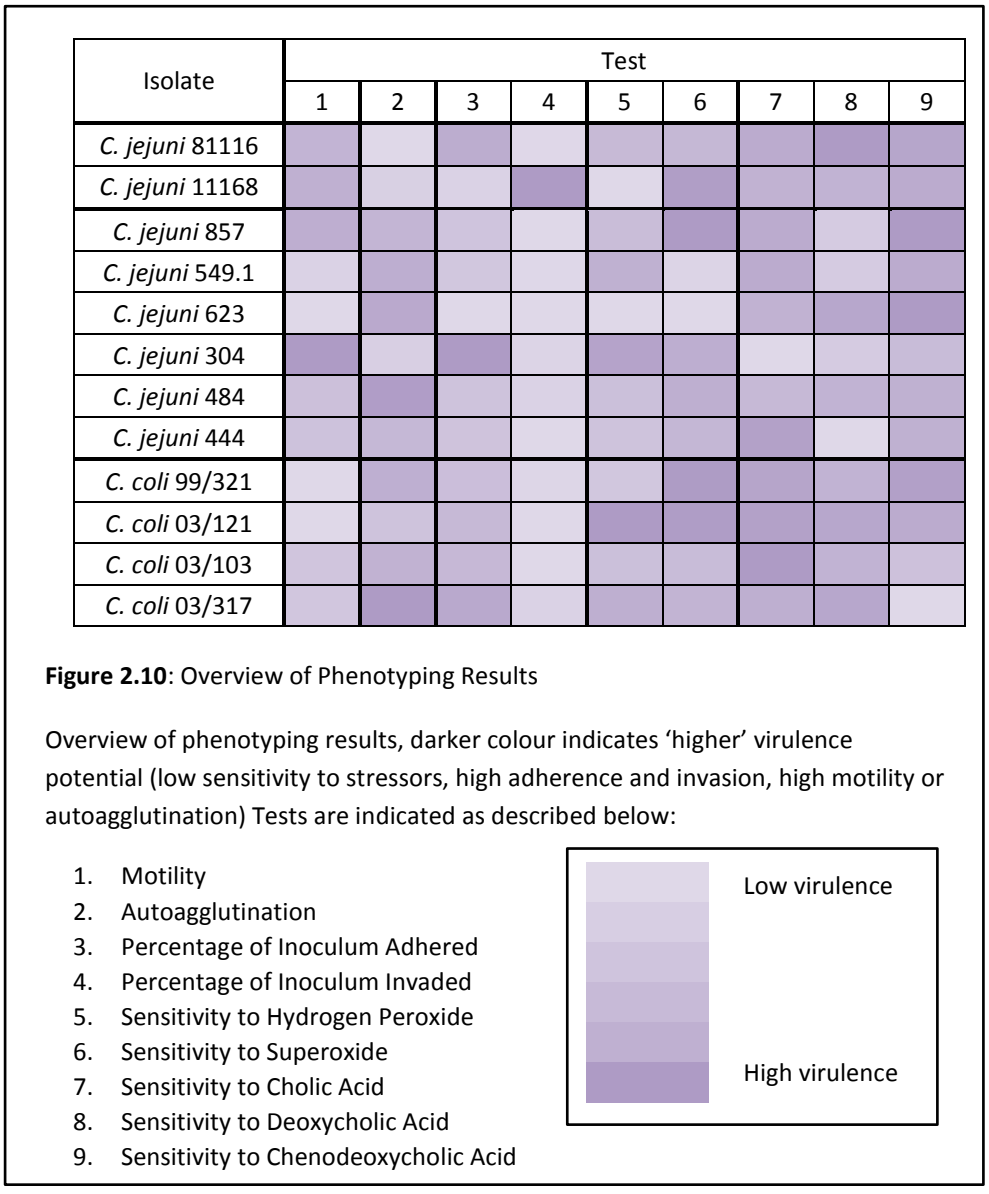

Figure 2.10: Overview of Phenotyping Results

Overview of phenotyping results, darker colour indicates 'higher' virulence potential (low sensitivity to stressors, high adherence and invasion, high motility or autoagglutination) Tests are indicated as described below:

1. Motility
2. Autoagglutination
3. Percentage of Inoculum Adhered
4. Percentage of Inoculum Invaded
5. Sensitivity to Hydrogen Peroxide
6. Sensitivity to Superoxide
7. Sensitivity to Cholic Acid
8. Sensitivity to Deoxycholic Acid
9. Sensitivity to Chenodeoxycholic Acid

Considerable variation in virulence potential was found across the group of isolates and range of tests included. Neither the ST403CC *C. jejuni* nor the other *C. jejuni* or *C. coli* formed a distinct pattern of behaviour, however it was observed that the

ST403CC *C. jejuni* isolates showed similar virulence potential to other *C. jejuni* and *C. coli* isolates.

2.4 Discussion

Phenotypic assays were used to assess the potential capacity for human disease of a group of *C. jejuni* isolates which were suspected as being adapted to the porcine host. The isolates were part of the MLST 403 clonal complex, and were unstudied prior to this study with the exception of being typed by MLST and having been shown to be hippurate negative (Manning *et al*, 2003). This chapter set out to investigate the phenotype and pathogenic potential of the ST403CC *C. jejuni* isolates in order to determine whether these isolates may theoretically be capable of causing illness in humans, and whether this potentially host adapted group of isolates exhibited a distinct phenotypic profile when compared to other *C. jejuni* and *C. coli* isolates. Phenotypic characterisation demonstrated that the ST403CC isolates did not have a specific phenotypic profile compared to other *C. jejuni* and *C. coli* isolates, however they have the potential to cause disease in humans.

2.4.1 Motility

Motility results for the reference strain *C. jejuni* 81116 were lower than in previously published work (Fearnley *et al*, 2008), however the results were consistent across repeated tests. This was observed despite the similarities between methodologies (stabbing inoculum into Mueller-Hinton with 0.4% agar, and incubating for 48 hours are 42°C); with diameters of motility between 5-5.8cm in contrast to the mean of 1.9cm observed here. Differences which may serve to explain this variation include the different media for prior culture (bacteria were grown at 42°C on BASA plates (Blood agar containing Skirrow's selective antibiotics in Fearnley *et al* 2008 for 24 hours then transferred to brain heart infusion broth supplemented with yeast extract for 20 hours - in contrast to recovery for 48 hours at 37°C on MCCDA plates) and a possible difference in concentration of the bacterial suspension; they adjusted theirs to a consistent optical density but did not state what; it may be that they were using a higher initial inoculum than was described here (optical density of approximately 0.1 at 600nm). These different culture

techniques may have led their isolates to express virulence associated genes and therefore behave differently in the assay compared to the version used here. Another potential factor would be the increasing lab adaptation of isolate *C. jejuni* 81116 leading to a reduction in motility, although passage was minimal in an attempt to avoid this risk. Fearnley *et al* (2008) found no significant differences in motility between low and hyper invasive isolates.

All of the tested isolates exhibited motility, and considerable variation was observed between the twelve studied isolates. The ST403CC *C. jejuni* isolates displayed on average lower motility than the human *C. jejuni* isolates, and higher motility than the porcine *C. coli* isolates. Aroori, Cogan & Humphrey (2013) investigated the effect of temperature on motility and invasion in *Campylobacter* species; they observed that *C. coli* isolates were less motile at 37°C than at 42°C; it is possible therefore that the low motility observed for *C. coli* was linked to this temperature effect, and that the lower motility of ST403CC isolates indicates a similarity in behaviour with *C. coli* isolates.

2.4.2 Autoagglutination

Considerable variation in autoagglutination was observed across the twelve *C. jejuni* and *C. coli* isolates. Interestingly, ST403CC *C. jejuni* isolate 304, which was the most motile of the strains, showed the lowest degree of autoagglutination of the test strains, being notably closer to the reference strains (reference strains 0.26 and 0.92, *C. jejuni* 304 0.36, remaining ST403CC isolates 0.61-0.90). This may infer a possible trend of inverse relationship between motility and autoagglutination, of which there is some further evidence amongst the dataset; of the six highest observed motility scores (*C. jejuni* strains 304, 857, 11168, 81116, 484 and 444), five are also amongst the lowest autoagglutination scores (81116, 11168, 304, 03/121, 444, 857) although there are also two clear exceptions amongst these groups – *C. jejuni* 484 recorded the fourth highest motility score, yet was also the second highest autoagglutination score, whilst *C. coli* 03/121 recorded the third lowest autoagglutination score whilst also having recorded the equal lowest motility score of the twelve included strains. Misawa & Blaser (2000) used mutant studies of *C.*

jejuni 81116 to investigate motility and autoagglutination and found that motility itself may not be essential for autoagglutination – a flagellated but non motile *flaA* variant retained autoagglutination close to that of the wild-type strain, whilst an aflagellate variant did not display autoagglutination. This was expanded upon by Golden & Acheson (2002) who showed that *flaA* is not the only determinant in autoagglutination levels – presence of *flaA* is necessary for autoagglutination but alone is not sufficient. Golden & Acheson (2002) showed that motility and autoagglutination (and subsequent invasion ability) are related, but that motility does not predict autoagglutination, or vice versa.

2.4.3 Adherence to & Invasion into Caco-2 Cells

Variation in invasion efficiency across the twelve isolates was statistically significant, although the only isolate which was significantly different from the low invader control strain *C. jejuni* 81116 was the high invader control strain *C. jejuni* 11168. This corroborates the work of Fearnley *et al* (2008) which first distinguished the categories of low- high- and hyper- invasive *C. jejuni* isolates. In the categories defined by Fearnley *et al* (2008), high invasive strains including *C. jejuni* 11168 exhibit 10 times higher invasive efficiency than the low invasive type strain *C. jejuni* 81116. By extension of this it was established that the ST403CC *C. jejuni* isolates, which all demonstrated invasive capabilities in the Caco-2 model, would all be considered low-invasive strains. The same description may also be applied to the four *C. coli* isolates, however it may be risky to extrapolate from an original study based solely upon *C. jejuni* isolates to apply to another albeit closely related species. It is however reasonable to state that both the tested *C. coli* and ST403CC *C. jejuni* isolates displayed a similar range of invasive capability, and it can therefore tentatively be suggested that, based upon these data, the porcine ST403CC *C. jejuni* isolates exhibit a similar level of potential for causing illness as do the porcine *C. coli* isolates.

Strain-strain variation is very common in 'wet-lab' tests of *Campylobacter*, and adhesion and invasion assays are no exception to this; however, the invasion efficiency results observed in this study were comparable to results observed in the

literature. For instance, Fearnley *et al* (2008) observed that the majority of the isolates they investigated showed an invasion efficiency percentage of between 0.0006 and 0.3% of the inoculum being internalised (into INT407 cells); this range included their result for the low invader control isolate *C. jejuni* 81116 which was used in this study and recorded an invasion efficiency of 0.04%, both this and all the included experimental isolates fit within the low invader efficiency range observed by Fearnley *et al* (2008).

Multiplicity of infection (MOI) may be considered an important factor in the gentamicin protection assay, and a range of MOIs have been utilised in the literature: Hu & Kopecko (1999) used a range of MOI for *C. jejuni* 81-176 and determined that an MOI of 0.2 was most efficient for internalisation percentage, with MOIs in the range 0.2-200 providing a steady increase in raw numbers of internalised bacteria in INT407 cells. In 1994, Russell & Blake found an MOI of ratio 1000:1 to be suitable when studying *C. jejuni* 81-176 invasion in Caco-2 cells, whilst in 2010 Javed *et al* utilised an MOI of 100 when using *C. jejuni* strains 11168, wild-type and mutant strains of hyper invasive isolate 01/51 and the *flaA/flaB* mutants of 81116. *C. jejuni* invasion in cell culture has been demonstrated to occur at a considerable range of MOIs: maximal efficiency for invasion occurs around 100-200:1, with an increase in the number of bacterial cells internalised but not in the percentage efficiency of invasion observed after this point. The most important factor is consistency to allow comparison between repeat tests and different isolates, initial runs to test the assay conditions and procedure were successful and were deemed to have an approximate MOI of 1000:1 following retrospective bacterial counts; as such an MOI of approximately 1000:1 was subsequently used for all tested strains; estimated by adjusting optical density to a standardised amount, and confirmed using retrospective counts.

Definite links were not observed between the 'predictor' tests and efficiency in adhesion or invasion. Although the most motile isolate (*C. jejuni* 304) was also the top isolate in the adhesion assay, of the other top five motility isolates, only one other (81116) was in the top six isolates for adhesion efficiency, whilst others (*C. jejuni* 11168, 857) were low efficiency adherence isolates. Likewise although one

isolate was a high scorer in both autoagglutination and adhesion efficiency (*C. coli* 03/317), the highest adhesion isolate (*C. jejuni* 304) was the third lowest scorer in the autoagglutination assay. Similar variation was observed between motility and invasion efficiency; three of the top six motility isolates were also in the top four for invasion efficiency (*C. jejuni* isolates 304, 11168 and 484), however the second highest motility isolate (*C. jejuni* 857) was also second lowest for invasion efficiency. Similar patterns were observed between autoagglutination and invasion efficiency; the second most efficient invasion isolate (*C. coli* 03/317) was also the highest scorer in the autoagglutination assay, however the highest invading isolate (*C. jejuni* 11168) was the second lowest scoring isolate in the autoagglutination assay. Additionally, adhesion efficiency did not appear to have a direct relationship with subsequent invasion efficiency; although some isolates, such as *C. jejuni* 304 and *C. coli* 03/317, were efficient at both adhesion and invasion, and others were mid-range for both, other isolates were poor in adhesion efficiency, yet amongst the most efficient invaders included in the study; most notably, the lowest adherence efficiency observed was that of *C. jejuni* 11168, the highly invasive control isolate, which was the most efficient invader studied by a considerable margin.

ST403CC *C. jejuni* isolates were demonstrated to be capable of adherence to and invasion into Caco-2 cells, as such they likely possess the capacity to cause gastrointestinal disease in humans. Considerable variation was observed across the six ST403CC isolates in efficiency for both adhesion and invasion, however it is common for variation to be observed in *C. jejuni*.

2.4.4 Sensitivity to Hydrogen Peroxide & Superoxide Stress

Baillon *et al* (1999) used agar pour plates and applied 6mm discs loaded with 3µl of 3% H_2O_2 and observed a 'zone of killing' for *C. jejuni* 81116 of 20mm, compared to the 25mm observed in this study, using a similar technique; applying 3µl of 3% H_2O_2 using 3mm discs in soft agar pour plates. This apparent increase in sensitivity may potentially be attributed to continuing lab adaptation of the *C. jejuni* 81116 isolate; although passage is kept to a minimum this isolate has been used in laboratory culture and study for over thirty years.

The ST403CC *C. jejuni* isolates 304 and 623 once again formed the outlier scores within the ST403CC group, with 304 being the second most resistant overall (lowest being *C. coli* 03/121), whilst 623 was the second most sensitive isolate after *C. jejuni* 11168.

No apparent link was observed between resistance to hydrogen peroxide and resistance to superoxide; some isolates were amongst the most sensitive to both (*C. jejuni* 623) and others resistant to both (*C. coli* 03/121) but others were amongst the most sensitive to one and most resistant to the other (*C. jejuni* 11168 and 549.1). There is possibly some relationship between adhesion efficiency and hydrogen peroxide resistance; the three most efficient adherers (*C. jejuni* 304, *C. coli* 03/317, *C. jejuni* 81116) were also amongst the most resistant to hydrogen peroxide stress, likewise poor adherence isolates such as *C. jejuni* 11168 and 623 were also the most sensitive to hydrogen peroxide exposure – although the picture was not complete as there were isolates which contradict this pattern, and this does not expand to include superoxide resistance, or invasion efficiency. A link between adhesion efficiency and hydrogen peroxide sensitivity would make sense as isolates which are more able to survive the H_2O_2 stress would potentially have a better opportunity to adhere to the cells, both *in vitro* and *in vivo*. Harvey & Leach (1998) proposed that oxidative stress can increase the invasive potential of *Campylobacter* isolates; a link was not observed between survival of oxidative stress and increased invasion in these experiments however as only one high-invasive strain and a majority of low invasive isolates were tested it may be likely that this trend wasn't observed due to the lack of additional high or hyper invasive strains.

Grant & Park (1995) observed *katA* presence in *Campylobacter* and its function as a potential oxidative stress defence system. Catalase is common to thermo-tolerant *Campylobacters* (Day *et al*, 2000), and it is known to play a role in surviving oxidative stress in the environment, as well as protecting against oxidative burst (hydrogen peroxide) in macrophage (Grant & Park, 1995). Grant & Park (1995) demonstrated that deactivation of the *katA* gene resulted in an oxidative stress sensitive isolate.

Direct testing for catalase activity was not carried out, as *C. jejuni* and *C. coli* are known to be catalase positive (Penner, 1988) although some examples of catalase negative or weak *C. jejuni* have been observed, these were recategorised as *C. upsaliensis*. Had any isolates failed to survive at all in the hydrogen peroxide assay then catalase testing would have been applied along with PCR confirmation of *katA* presence to confirm the results.

2.4.5 Bile Acid Sensitivity

Overall the variation between isolates appeared less than for other assays, however the variation across the group was nevertheless statistically significant (95% C.I.) for both cholic acid and chenodeoxycholic acid. Species appeared to have an effect only for the cholic acid assay, and host origin (porcine *vs* non-porcine) did not appear significant in bile salt sensitivity. *C. coli* isolates were significantly more resistant to cholic acid stress than *C. jejuni* isolates, but this was not reflected in the deoxycholic or chenodeoxycholic acid assays. This may seem slightly unexpected as the major bile acids in pigs are hyocholic acid, chenodeoxycholic acid, and hyodeoxycholic acid (Elliott, in Danielsson & Sjövall, 1985).

C. jejuni 304 was notably sensitive across the three bile acid assays, being the most sensitive of all twelve isolates to cholic acid, the joint most sensitive to deoxycholic acid, and third most sensitive, after *C. coli* isolates 03/103 and 03/317, to chenodeoxycholic acid. *C. jejuni* isolate 623 however was the equal most resistant isolate to both chenodeoxycholic acid and deoxycholic acid; although it recorded the third most sensitive result for cholic acid however the variation across the range of cholic acid results – with the exception of 304 – is small (range of just 6-8.33mm).

Exposure to bile acids is known to have regulatory effects on *C. jejuni*. Malik-Kale, Parker & Konkel (2008) described that exposure to sodium deoxycholate (deoxycholic acid) induced the expression of virulence genes (*ciaB*, *cmeABC*, *dccR*, *tlyA*) and increased invasive potential in INT407 cells, without affecting the motility or adherence for *C. jejuni* F38011. Fox *et al* (2007) found that culturing *C. jejuni* with above 'natural' bile concentrations (up to 5% ox bile w/v in brain heart infusion broth) induces a stress response involving GroEL GaIU and bacterioferritin.

Hugdahl, Beery & Doyle (1988) observed negative chemotaxis from cholic acid and deoxycholic acid, and positive chemotaxis towards bile (both bovine and hog), demonstrating that although bile is a chemoattractant for *C. jejuni*, its individual components act as chemorepellants.

Raphael *et al* (2005) used growth measurements using optical density observation following the addition of various antimicrobial reagents in 96 well plate growth assay. Using this method, they observed MIC concentrations of 12.5mg/ml for deoxycholic acid, 10mg/ml for chenodeoxycholic acid, and 3.125mg/ml for cholic acid for *C. jejuni* human disease isolate F38011. Raphael *et al* (2005) demonstrated that inactivation of *Cj0643* (which they designated *cbrR* – *Campylobacter* bile resistance regulator) prevents growth at 1% wt/vol deoxycholic acid, and inhibits growth in other bile acids, but does not affects sensitivity to antibiotics and other stressors, and causes decreased ability to colonised in the chicken model. This further supports the belief that bile resistance is important in host colonisation, and demonstrates that the response regulator *cbrR* modulates bile acid resistance in *C. jejuni*. The ranges of stressor concentrations used in this study were guided by the results of the study by Raphael *et al* (2005), and the single concentrations selected for inclusion in results and discussion (10mg/ml cholic acid; 13mg/ml deoxycholic acid; 15mg/ml chenodeoxycholic acid) reflected the patterns observed across the range, with the least intra-strain variation.

Varying degrees of sensitivity to bile acids were observed across the twelve included isolates, and across the ST403CC *C. jejuni* isolates. What was observed however is that these isolates can survive bile acid stresses, as they must if colonising the human or porcine host.

2.4.6 Summary & Overall Discussion of Phenotype

Gripp *et al* (2011) attempted to investigate host-specificity by using ST21 isolates from a variety of sources. As they stated, it had been frequently shown that closely related strains (MLST) can be found in a variety of sources, but that the limitation of MLST considering only conserved house-keeping genes might be overlooking important information about virulence or host specific genes. Gripp *et al* (2011)

observed considerable phenotypic diversity within the single ST however this was general variation between isolates and did not correlate with source. Gripp *et al* (2011) also completed genome sequences of five of the ST21 isolates and determined that the dominant STs which are found across a variety of sources have a 'generalist' phenotype which enables them to colonise humans and a range of domestic animals, however they also acknowledged that this does not rule out specialism in other *C. jejuni*. Examples of specific niche adaptation include the bank vole host adapted *C. jejuni* MLST3704 including isolate 414 which was distinct in both phenotype and accessory genome (Hepworth *et al*, 2011).

There is currently little understanding of why *Campylobacter* causes such different response in chickens and humans (Young, Davis & DiRita, 2007), and this phenomenon is also important with regards to *Campylobacters* in pigs, as piglets have been used as a model of human infection, despite the fact that some *C. jejuni* and *C. coli* strains are isolated from pigs which were not expressing symptoms.

In conclusion the ST403CC *C. jejuni* isolates from food production pigs, representing a potential example of *C. jejuni* niche adaptation, did not exhibit a notably different phenotypic profile during this series of wet-lab experiments, however they behaved similarly to both human *C. jejuni* isolates and porcine *C. coli* isolates, with considerable variation between the strains, and with results suggesting their likely ability to cause pathogenesis in humans. This indicates that these ST403CC *C. jejuni* isolates are capable of causing gastrointestinal illness, as were the ST403CC *C. jejuni* isolates observed by Duim *et al* (2003) in Curaçao. Whilst these results indicate that the tested ST403CC *C. jejuni* isolates are potentially capable of causing disease in humans this is an extrapolated conclusion. Additionally this does not inform on the potential for disease of these isolates in other host species; the isolates have been demonstrated to be capable of invasion using a human colonic cell line, and may or may not be capable of invasion in other host cell types. It is unlikely that these isolates are a common cause of illness in the porcine or bovine host, as they were recovered from an abattoir and not from suspected cases of disease. *C. jejuni* may rarely cause illness in other host species (poultry, cattle, pigs) but generally is a harmless coloniser in these hosts; it is therefore unlikely that ST403CC *C. jejuni*

isolates would cause disease in the host, although this may occur as an opportunistic infection, for example if an individual is otherwise immunocompromised.

Null hypotheses were outlined previously - in relation to these, based upon the results observed in this chapter, we can reject the first null hypothesis 'ST403CC C. jejuni isolates do not show evidence of the potential to cause disease in humans'; however we must accept the following two null hypotheses: 'ST403CC *C. jejuni* isolates do not exhibit higher pathogenic potential, as determined in laboratory tests, than non-ST403CC *C. jejuni* and *C. coli* isolates' and 'ST403CC *C. jejuni* isolates do not exhibit a distinct phenotypic profile compared to non-ST403CC *C. jejuni* and *C. coli*'.

It was demonstrated that the ST403CC *C. jejuni* isolates from pigs, previously identified as a potentially host adapted group and being unusual hippurate negative strains (Manning *et al*, 2003) behaved similarly to more 'typical' *C. jejuni* isolates in virulence and stress tests. Lack of distinction by these tests however cannot rule out evidence for adaptation in the genome of these isolates, it could be that host associated genes have been acquired which simply were not important in these tests. The subsequent chapters will focus upon acquiring and investigating the genome sequences of these six ST403CC *C. jejuni* isolates, beginning by confirming their phylogeny within known *C. jejuni* isolates.

Chapter Three: Investigating the Whole Genome Phylogeny of ST403CC *Campylobacter jejuni* isolates

3.1 Introduction

It was previously established that the selected ST403CC *C. jejuni* strains are closely related in terms of MLST, representing different Sequence Types within a Clonal Complex (Chapter One; Manning *et al*, 2003) and has now been demonstrated that they have the capacity to cause disease in humans (Chapter Two). The next significant objective was to determine how closely related the ST403CC *C. jejuni* isolates are at the whole genome level, where they fit within the *Campylobacter* phylogeny, and to begin investigating any genetic evidence of adaptation.

3.1.1 Methods for Investigating Phylogeny

Numerous methods exist to study the relatedness of bacterial isolates and introduced briefly below are some of the most important tools used in the study of *Campylobacter*; MLST, genomic hybridisation and microarray analyses, and whole genome sequence 'phylogenomics'.

3.1.1.1 MultiLocus Sequence Typing

As introduced in previous chapters, MultiLocus Sequence Typing (MLST) is a useful method for typing and studying the population structure of *Campylobacter*, as it is based upon the genetic sequence of seven 'house-keeping' genes across the genome (Dingle *et al*, 2001). MLST is informative as these genes are slowly changing, essential genes which cannot be deactivated or lost entirely and is particularly valuable as a starting point when investigating new isolates. MLST is also an important measure as it is readily shareable and allows rapid comparison between laboratories. MLST is limited however in that it only considers the seven genes in the scheme and cannot inform on any grander characteristics of an isolate.

As the efficiency of whole genome sequencing increases, and resulting costs decrease, bacterial research has reached an age where it is frequently more feasible to conduct whole genome sequencing than to use traditional MLST (Larsen *et al*, 2012; Inouye *et al*, 2012). As such an increasing number of researchers are using

and developing tools to glean MLST information from short-read sequences (Inouye *et al*, 2012). In some instances the MLST genes are identified from genomes which have already been produced to a draft quality level; that is the initial short-read sequences have been assembled and annotated to a sufficient level that the appropriate loci can be selected and typed, and may be concatenated to provide short segments of sequence representing the complete MLST profile of the isolate; such as that described in papers by McCarthy *et al* (2007) and Sheppard *et al* (2009) which are both described with more detail below. Other researchers have developed means to assess the MLST of newly sequenced isolates without the need for the 'completion' of the genome; Inouye *et al* (2012) developed a software tool which uses raw short-read sequence data to determine the sequence type with a high degree of accuracy, shortly after Larsen *et al* (2012) produced a browser based platform with similar capabilities.

MLST provides useful indications of relatedness between isolates (Dingle *et al*, 2001), however as it is limited to seven genes it cannot inform on other genetic characteristics of isolates.

3.1.1.2 Hybridisation Techniques

Hybridisation techniques have provided another important tool in assessing similarity and likely relatedness between bacterial isolates (Leonard *et al*, 2003). Comparative genomic hybridisation (CGH) allows direct comparison of a query isolate to a reference isolate at the full genome level, without the need for sequencing the query isolate, by observing hybridisation to fragments of sequence from the reference genome (Leonard *et al*, 2003; Luo, Lin & Xu in Xu (Ed), 2010). DNA microarrays can be produced based upon a full genome, displaying every open reading frame in a given isolate for comparative genomic hybridisation; providing detailed information on the genetic content of a query isolate by identifying ORFs common to both isolates, and ORFs which are present in the reference isolate but absent from the query strain (Taboada *et al*, 2004).

The use of DNA microarrays formed an important stage in the development of knowledge about genetic variation in *Campylobacter*. Following the sequencing and

annotation of the first *C. jejuni* genome (NCTC11168) in 2000 by Parkhill *et al*, microarray provided a framework for comparison of isolates. Dorrell *et al* (2001) developed a whole genome DNA microarray for the newly sequenced *C. jejuni* NCTC11168, allowing hybridisation experiments to inform on the presence or absence of each coding sequence in a query isolate.

Microarray analysis provided a tool which could give much more information than PFGE and RAPD, which were the main typing tools preceding its introduction (Leonard *et al*, 2003). However, microarray study was and is limited by the content on the slide – a *C. jejuni* 11168 microarray can assess only the similarity of a single strain to the genome content of *C. jejuni* 11168, whilst a pan genome array, although with much wider scope is still limited to what is present on the array. Leonard *et al* (2003) demonstrated that a DNA microarray can uncover both highly conserved or divergent regions between isolates or groups of isolates, however, microarray analysis cannot inform on genetic information present in the experimental isolate which is not found on the array, so any additional genetic content is overlooked. This limitation could be addressed using methods such as suppression subtractive hybridisation (SSH), wherein the DNA of a pair of strains are digested before being subjected to hybridisation leaving behind DNA present only in the test strain for subsequent analysis, revealing additional genetic content of that isolate (Winstanley, 2002); this process is valuable in determining variation between closely related isolates, however once again this technique is limited to one 'test' strain and one 'driver' (reference) strain. As described below, whole genome sequencing is now more commonly used to investigate variation in genomic content.

During the early 2000s, microarray analysis was more cost effective than full genome sequencing, however with the development of second generation sequencing techniques and the advent of high-throughput low cost methods it is now typically more appropriate to use whole genome sequencing methods to answer the questions previously addressed by DNA based hybridisation techniques (Croucher & Thomson, 2010).

3.1.1.3 Whole Genome Alignment

Sequence data is the most valuable and most commonly used basis for phylogenetics. Whole genome sequence based phylogeny, or phylogenomics, is increasingly being recognised as an important measure of relatedness between bacterial isolates (Luo, Lin & Xu in Xu (Ed), 2010).

Phylogenetic analysis can be attributed to the so called 'first fact' of biological sequence analysis, attributed to Dan Gusfield of the Unviersity of California in 1997 (as cited by Mushegian, 2007): outlining the concept that similarity between sequences often implies similarity in structure or function, and also, crucially, that high sequence similarity often suggests an evolutionary link.

Numerous approaches and techniques exist within whole genome studies of relatedness, including investigation of core and pan genomes, core genome alignment and maximum likelihood trees, comparative genome searching and alignment such as BLAST (Basic Local Alignment Search Tool; Altschul *et al*, 1990) and BRIG (BLAST Ring Image Generator; Alikhan *et al*, 2011), and analysis of single nucleotide polymorphisms (SNPs).

The core, accessory, and pan genome of a species are effectively the essential, non-essential, and total portions of the genome content of a bacterial species (Tettelin *et al*, 2005). The core genome is the shared genomic content, including essential and conserved genes, whilst the accessory genome is the additional content present in some or individual isolates, and the pan genome is the term for the total content of both core and accessory genome (Tettelin *et al*, 2005). These elements can also be studied at higher levels such as genera, or across smaller defined groups.

Whole genome phylogeny can be inferred from global alignment of sequences. Sequence alignment determines the evolutionary relationship of sequences based upon homology (Dewey & Pachter, 2006). From multiple whole genome sequence alignments, phylogenetic trees can be produced as a visualisation of evolutionary relationships. Phylogenetic trees show the evolutionary relationships of individuals;

the divergence of the closest common relative, rather than the overall similarity of the sequences (Rokas, 2011).

Comparative searching and alignment methods can be used to specifically investigate whether a particular gene or sequence region is found in other genome sequences. BLAST (Basic Local Alignment Search Tool, Altschul *et al*, 1990) searching can be carried out using the web-based repository of published sequences, or using a local database created with specific sequences of interest, and produces a score of similarity for query sequence against a database; allowing specific inferences on sequence homology. Various methods exist for visualising these comparisons, such as ACT (Artemis Comparison Tool, Carver *et al*, 2005; Carver *et al*, 2012) which uses Artemis sequence viewer (Rutherford *et al*, 2000) to provide visual results of 'A vs B' BLAST comparisons for pairs of sequences, and BRIG (BLAST Ring Image Generator, Alikhan *et al*, 2011) which produces circular diagrams of BLAST results for multiple query sequences against a reference sequence.

The detection of Single Nucleotide Polymorphisms (SNPs) is a useful tool for identifying variation in sequences. This method compares two sequences and identifies single nucleotide variations between a reference sequence and the query sequence. This approach can be used to consider evolution or variation in specific genes, or can provide an approximation of relatedness between isolates; a higher number of SNPs suggesting more divergence. SMALT (Ponstigl, 2010 – currently not published as a journal article, program and manual available at https://www.sanger.ac.uk/resources/software/smalt/) maps query sequencing reads against a reference sequence; the output of which can subsequently be assessed for the presence of SNPs using SAMtools (Li *et al*, 2009).

The species concept is problematic in bacteriology, and the degree of similarity within accepted species boundaries varies significantly depending on the levels of diversity and recombination observed in different bacterial species and genera (Hanage, Fraser & Spratt, 2005). In a study of species phylogeny using the sequence of MLST alleles in *Neisseria*, Hanage, Fraser & Spratt (2005), it was observed that in

such a highly recombinogenic species, there can exist 'fuzzy' species boundaries, wherein intermediary isolates may be located in between two existing accepted species. Further to this, in 2012, Corander *et al* demonstrated that the 'fuzzy species' within *Neisseria* demonstrate elevated rates of recombination between species.

Comparative genomics is the use of whole genome sequencing to study differences between isolates, between species, and larger scales. It can inform on adaptation, virulence and phenotype (Wren, 2000).

3.1.2 *Campylobacter* Phylogeny

Comparative genomic hybridisation has been utilised to investigate population structure in *Campylobacter*, and can provide data which is informative of what is present or absent in an experimental strain in relation to the clusters on the microarray slide. The microarray for comparative hybridisation may be based upon the genome of a single isolate, or a 'pan-genome' array with sequence data from a representative group of isolates, and may be used for determining genomic content, or for investigation gene expression (transcriptomics) (Luo, Lin & Xu, in Xu (Ed), 2010).

Champion *et al* (2005) used DNA microarray analysis based upon the genome sequence of *C. jejuni* NCTC11168 to demonstrate that *C. jejuni* forms two distinct clades, associated with livestock and non-livestock sources. Within this study Champion *et al* (2005) also observed a set of genes which were associated with isolates from the livestock clade, including a cluster of genes associated with the flagellin glycosylation locus.

In 2009, Lefébure & Stanhope created a phylogenetic tree based upon the core genome alignment of seventeen *Campylobacter* sequences; ten *C. jejuni* genomes, plus one each representing *C. coli*, *C. concisus*, *C. curvus*, *C. fetus*, *C. hominis*, *C. lari* and *C. upsaliensis*. It was observed that the genome based phylogeny distinctly separared the two major groups of *Campylobacter*; the thermo-tolerant and non-thermotolerant clades. Within the thermo-tolerant group, *C. lari* was shown to be

the most ancestral genome, with *C. coli* and *C. jejuni* the most recently evolved species. Figure 3.1 shows the maximum likelihood tree produced by Lefébure & Stanhope (2009) for illustration of the relationships of *Campylobacter* as determined by their investigation.

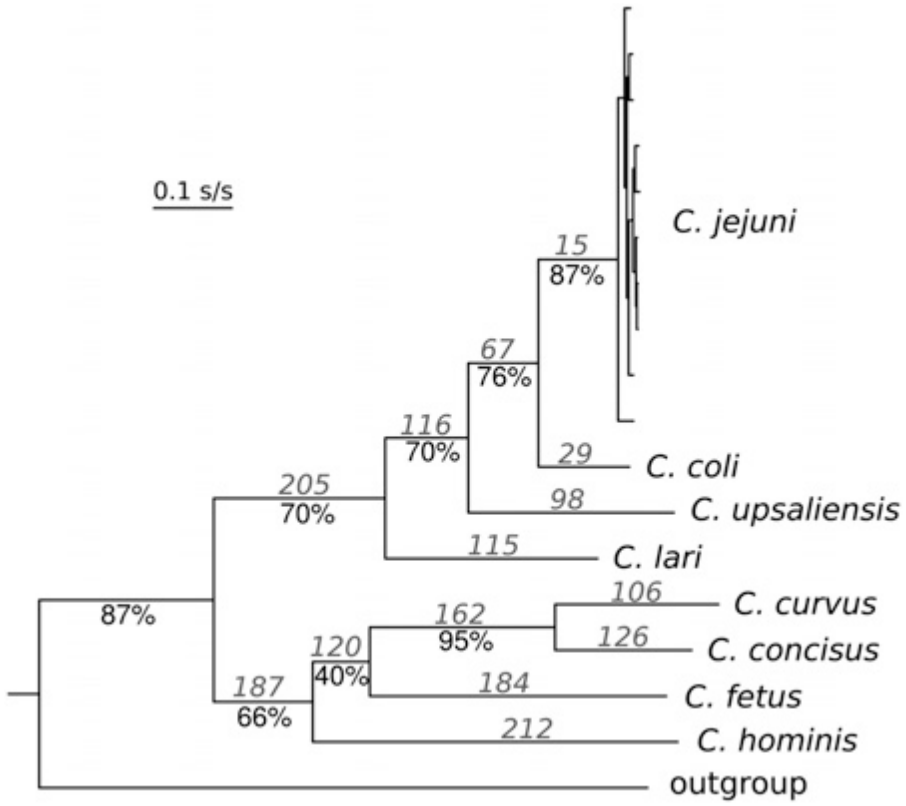

Figure 3.1: Maximum likelihood *Campylobacter* Phylogenetic Species Tree. Taken from Lefébure & Stanhope (2009).

The phylogenetic tree produced by Lefébure & Stanhope (2009) shows clear separation between thermotolerant and non-thermotolerant isolates. Branch labels indicate percentage of gene tree support (below branches) and number of genes determined to be under positive selection (above) as determined in the original study.

This tree was produced based upon an alignment of the core genome after discarding recombinant genes with homogeneous phylogenetic signal. The resulting 511 genes created the tree above through nonparametric bootstrapping to give maximum likelihood tree shown above.

Figure 3.2, overleaf, was produced using MUGSY (Angiuoli & Salzberg, 2010), RAxML (Rokas, 2011), and FigTree (Rambaut, 2007) based upon concatenated MLST sequence data taken from the PubMLST database (Jolley & Maiden, 2010; http://pubmlst.org/campylobacter/) and reveals that various ST403CC *C.*

jejuni isolates group together within this basis of comparison and are distinct from other Clonal Complexes; this is similar for most other included MLST Clonal Compexes (including ST42CC and ST354CC), whereas other Clonal Complex isolates

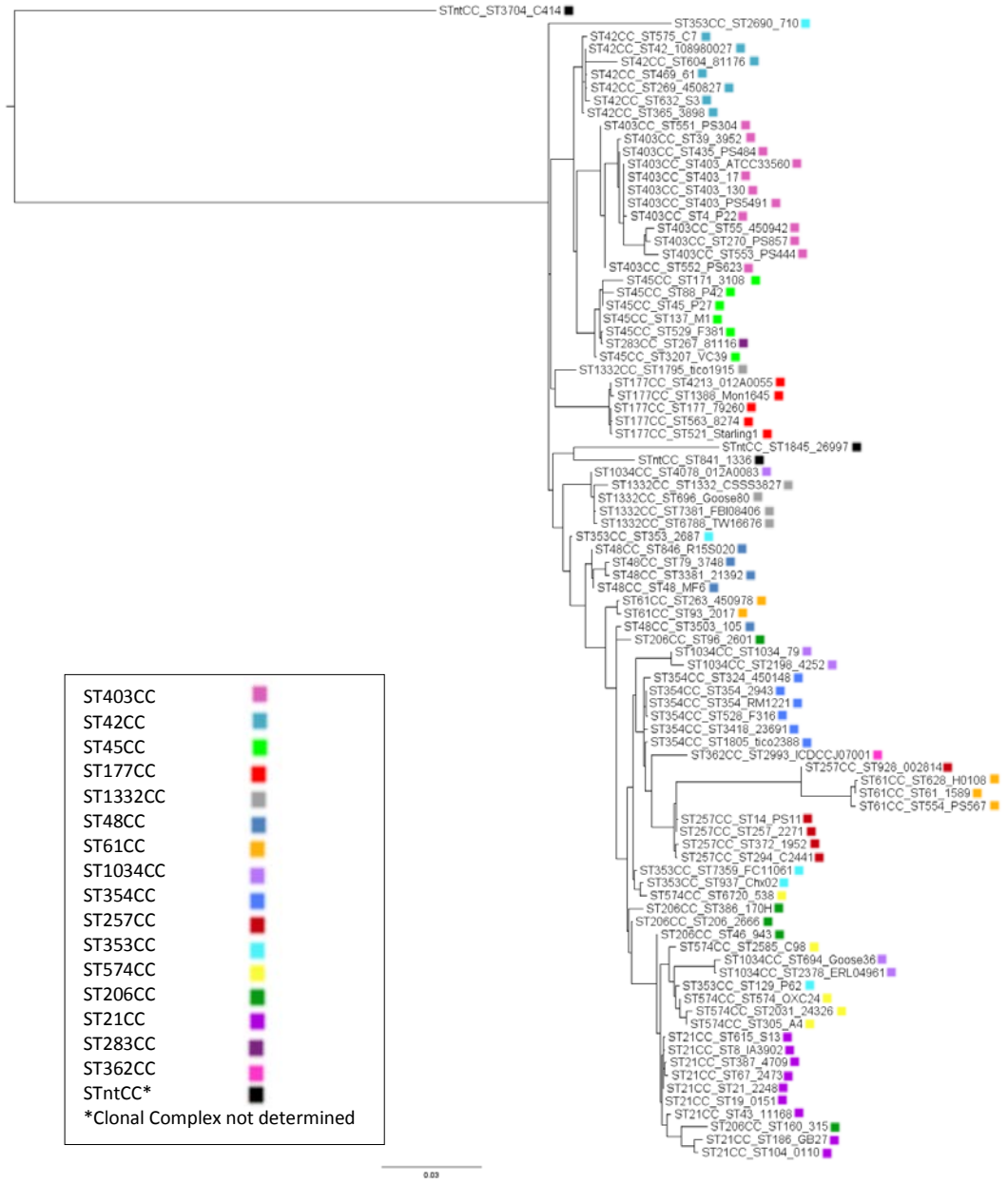

Figure 3.2: MLST Phylogenetic Tree for *C. jejuni* Isolates

Phylogeny tree based upon sequence alignment of ninety *C. jejuni* isolates concatenated MLST sequences.

group together with isolates from other Clonal Complexes (such as the ST21 and ST206 Clonal Complexes). Other Clonal Complexes, notably the host generalist ST61CC displayed visibily distinct groupings based upon concatenated MLST sequence data.

Species boundaries in eukaryotes are often clearly defined and demonstrated, however the distinguishing of species in the study of bacteria is a considerably more controversial area, particularly with the recent developments in 16S rRNA and whole genome sequencing approaches (Lan & Reeves, 2000; Sheppard *et al*, 2008). The issue of the bacterial species concept and flow has been particularly relevant in modern studies of thermo-tolerant *Campylobacter*. In 2008, Sheppard *et al* reported that *C. jejuni* and *C. coli*, due to a 'recent' change in gene flow are converging progressively and 'despeciating'; already they share 86.5% nucleotide similarity, a proportion similar to that between *E. coli* and *Salmonella*; which are closely related species sharing a large common gene pool, but with the maintenance of distinct species boundaries, and are thought to have diverged around 120 million years ago (Gordienko, Kazanov & Gelfand, 2013).

In contrast, Caro-Quintero *et al* (2009) reanalysed the same MLST data used by Sheppard *et al* (2008) and concluded that although recombination is indeed occurring between species, it was not at a rate to suggest species convergence, and asserted that a clear boundary was maintained between *C. jejuni* and *C. coli*. In fact, Caro-Quintero *et al* (2009) finish that the convergence of any two bacterial species has yet to be proven. In addition, Lefébure *et al* (2010) sequenced 42 *C. coli* and 43 *C. jejuni* isolates using Illumina GAII and concluded that the pan genomes of these species are finite, and that the two species have similar pan genome sizes, although the core genome of *C. coli* is larger than the core genome of *C. jejuni*. Lefébure *et al* (2010) also observed that each species had a set of species specific core genes, and they concluded that *C. coli* and *C. jejuni* were 'sympatric' – they share the same hosts and environments but do not 'interbreed'. Lefébure *et al* (2010) consider this evidence of species specific genes and low observed levels of horizontal gene transfer to directly contradict the conclusions drawn by Sheppard *et al* (2008); they suggest that the two cannot be converging ('despeciating') due to the maintenance of these boundaries. Both Caro-Quintero *et al* (2009) and Lefébure *et al* (2010) propose that the study by Sheppard *et al* (2008) was limited by the use of MLST, focusing on only seven house-keeping genes. However more recent work by Sheppard *et al* (2011) expanded upon the original study using a 'word searching'

method and whole genome sequences and confirmed the results originally presented by Sheppard *et al* (2008).

3.1.3 Phylogenetics & Host Adaptation

MLST does not provide explicit information on the grander characteristics of an isolate, however considerable evidence has shown that certain STs or ST complexes can be associated with specific hosts or host types. Research which has used MLST to study the origin of isolates and population structure includes McCarthy *et al* (2007) and Sheppard *et al* (2009): in each case these used concatenated sequence from the seven MLST alleles as an 'ST genotype' for analysis. McCarthy *et al* (2007) included isolates from cattle (n=245), sheep (n=159) and chicken (n=309), and using STRUCTURE – a clustering method for inference of population structure (Pritchard, Stephens & Donnelly, 2000) – were able to predict the source of isolates. McCarthy *et al* (2007) found that predictions made using STRUCTURE based upon the 'MLST genotype' were able to predict isolates as being chicken or bovid more accurately than using just MLST or MLSTCC profile data. The method also showed some success at distinguishing between sheep and cow isolates, but with less accuracy. Sheppard *et al* (2009) analysed a large dataset (n=9093) using both STRUCTURE and an asymmetric island model using the concatenated sequence of the MLST alleles, and demonstrated that the majority of clinical isolates for both *C. jejuni* and *C. coli* were attributed by both methods to chicken origins, with lower numbers associated with ruminants, and fewer isolates for other sources such as wild bird, 'environment', swine and turkey.

Existing literature has shown that some sequence types or complexes can be associated with specific hosts, or host types, including the pig association previously observed for the ST403CC *C. jejuni* isolates (Manning *et al*, 2003). For example, in addition to attributing sources of human clinical isolates, Sheppard *et al* (2009) also observed that strains isolated from farmed poultry and ruminants were substantially different; and even greater variation was found between farmed chickens and wild birds at the same site. Also observed were associations between ST61 with cattle, and ST257 with chickens.

3.1.4 Aims

It was previously observed that the isolates selected for this research (six ST403CC *C. jejuni* isolates) were closely related through MLST, sharing four or more identical alleles with the clonal complex sequence type ST403. This chapter aimed to investigate how closely related the six ST403CC *C. jejuni* isolates are, based upon whole genome sequence rather than solely upon the seven MLST genes. This chapter also investigated how the ST403CC *C. jejuni* isolates fit within other *C. jejuni* and *C. coli* isolates.

Whole genome alignment has become a critical tool in understanding diversity and evolution in bacteria (Angiuoli & Salzberg, 2010) and this project utilised this modern approach to improve understanding of evolution and population diversity in *Campylobacter*.

The specific aims of this chapter were as follows:

- To produce high quality genomic DNA for sequencing for six ST403CC *C. jejuni* isolates, and four *C. coli* isolates.
- To acquire genome sequence data for the ten isolates, courtesy of the University of Exeter.
- To utilise *in silico* MLST methods to confirm MLST results for the six ST403CC *C. jejuni* isolates, and to determine MLST for the four *C. coli* isolates.
- To determine the phylogeny of the ST403CC *C. jejuni* isolates amongst a range of reference *C. jejuni* and *C. coli* sequences.

3.2 Methods

In order to address the phylogeny of the ST403CC *C. jejuni* isolates, quality draft genomes were required.

3.2.1 Obtaining Genomic DNA

The initial step in the process was to produce genomic DNA for sequencing. Genomic DNA was extracted using a traditional phenol-chloroform extraction process.

Bacteria were recovered from frozen stocks (-80°C storage in Mueller-Hinton broth (Oxoid (Thermo Scientific) Item CM0405) containing 20% v/v glycerol (Fisher Scientific Item G/0600/17)) on MCCDA plates (Oxoid (Thermo Scientific) Item CM0739) for 48 hours at 37°C, with an appropriate gas environment produced using CampyGen sachets (Fisher Scientific Item CN025A; developing a gas mixture of 5%O_2, 10%CO_2 and 85% N_2) in gas jars. Bacterial growth was then swabbed from plates and suspended into 1ml sterile PBS, mixed thoroughly to break down any clumps and create a homogenised suspension.

Bacterial cells were then harvested by centrifugation at 4000xg for ten minutes, and resuspended into a lysis solution (2ml TE buffer (Tris-EDTA buffer; Sigma-Aldrich Item 93283) with 150µl of 10% SDS (Sodium Dodecyl Sulphate; Sigma-Aldrich Item L3771) and 150µl of 20mg/ml proteinase K (Sigma Aldrich Item P6556) and mixed thoroughly by inversion and then incubated for sixty minutes at 65°C. The resulting lysate solution was subsequently separated through a series of centrifugation steps. To the lysate solution was added 2ml of phenol-chloroform-isoamyl alcohol (25:24:1; Sigma-Aldrich Item P2069) and the solution was mixed carefully by inversion five times before being transferred to a Phase-Lock tube (5Prime Phase-Lock Gel Light 15ml tubes, SLS Item 2302840) and centrifuged for ten minutes at 1500xg. The upper layer of supernatant was then carefully removed and transferred into 2ml of phenol-chloroform-isoamyl alcohol, mixed by inversion three times, then transferred to a new phase-lock tube and again centrifuged for ten minutes at 1500xg. The previous step was repeated a second time so that the suspension had competed three Phase-Lock separation steps, after which the upper layer of the supernatant was again carefully removed, and then transferred into a fresh tube containing 3ml of room temperature 100% ethanol, and precipitated by gentle inversion. The DNA was then removed using a sterile hook and swirled in ice cold 70% ethanol before being left to dry for up to thirty minutes and subsequently dissolved without vigorous mixing in sterile nuclease free water (Ambion Item AM9938: an initial volume of 200µl of water was added, however where necessary this was increased up to a maximum of 500µl in order to fully dissolve the DNA).

Quality and concentration were checked using a microvolume spectrophotometer (NanoDrop 2000) with samples accepted if they had a 260/280 of 1.8 and 260/230 of 2.0-2.2. Genomic DNA samples were assessed for fragmentation using gel electrophoresis: a 0.5% agarose gel was produced (TAE buffer (Tris-Acetate-EDTA; Fisher Scientific Item ELR-328-010v)) with 0.5% w/v agarose (Fisher Scientific Item BPE1356) with 1% SybrSafe (Life Technologies Item S33102) added prior to pouring. 3µl of sample was mixed with 2µl of loading dye (Biolabs blue loading dye Item B70215) and loaded into wells. The gel was run for 25 minutes at 95v, using 5µl of Norgen HighRanger Plus 100bp ladder (EurolabSupplier Item NO-12015-5), subsequently visualised at 200ms exposure.

If the sample passed quality checks it was subsequently diluted as required to a concentration of approximately 10µg in 100µl and stored at -20°C before being submitted for sequencing.

3.2.2 Sequencing Genomic DNA – Performed by The University of Exeter

Once suitable genomic DNA was prepared, sequencing was kindly carried out at The University of Exeter, using an Illumina HiSeq 2000 machine, creating 36bp paired-end reads, which provided high quality short-reads of sequence, in FastQ format. Briefly, the stages of Illumina sequencing are as follows; firstly, a DNA library is prepared – DNA is fragmented into small pieces and the sheared ends are repaired and adenylated before adapter oligos are joined to each end of the fragments. Once the fragment library is complete, clusters are generated – the DNA fragments are applied to a flow cell which holds a lawn of oligonucleotide primers which bind to the adaptors on the fragments. Enzyme and nucleotides are then added to begin 'bridge amplification', leading to both ends of the fragment being attached to the flow cell, and double stranded. The double stranded fragments are then denatured to leave large numbers of single stranded templates available for binding; amplification leads to the formation of dense clusters of dsDNA. Base-by-base sequencing then takes place using fluorescently labelled reversibly terminated nucleotides which bind competitively (to provide higher accuracy than 'one-at-a-time' nucleotide exposure). Laser excitation is used to identify which base is added

(once again aiding the accuracy of the sequence data, as the cyclic nature of the sequencing reduced the likelihood of homopolymeric tracts confounding the results); this process results in large numbers of accurately sequenced short-reads.

Once sequence data was complete, the information was provided by the University of Exeter, for subsequent assembly and analysis to be carried out by the author.

3.2.3 Assembly of Genomic Sequences

Upon receipt of the raw genomic sequence data, genomes were assembled to provide draft quality genome sequences for subsequent analysis. Both forward and reverse raw FastQ files were trimmed to 16 million lines, if necessary, before being 'shuffled' to provide a consensus sequence for assembly using VELVET (Zerbino & Birney, 2008), with the settings; 31kmer, exp_cov 75 cov_cutoff 8 which were adjusted where necessary (exp_cov 35, 50; cov_cutoff 4, 5, 6) in order to provide average contig lengths of tens of thousands of basepairs. Subsequently the genome sequence was improved using ABACAS (Assefa *et al*, 2009) and IMAGE (Tsai, Otto & Berriman, 2010) using *C. jejuni* RM1221 and *C. coli* RM2228 as reference isolates for each species.

3.2.4 Confirming MLST *in silico*

The newly acquired sequence data were used to confirm the MLST findings for the ST403CC *C. jejuni* isolates described by Manning *et al* (2003), and to establish the MLSTs of the four *C. coli* isolates. Sequence Types were determined *in silico* for the ten sequenced strains, using the PubMLST *Campylobacter* database (available via http://pubmlst.org/campylobacter/), and using Short Read Sequence Typing (SRST, Inouye *et al*, 2012). The PubMLST database inferred the sequence type and clonal complex of each sequence, achieved by uploading the FastA sequence file for each genome to the sequence query page. The pubMLST query was used to query the MLST database, producing output describing the allele version of each loci, including length and position, as well as informing the resulting Sequence Type and Clonal Complex. SRST was used as an additional confirmation method. SRST is a software tool built to assess MLST from short-read sequencing data (Inouye *et al*,

2012). There were two processes required in SRST analysis; firstly a database was created using all known alleles for each locus (as available from PubMLST, downloaded March 2013).

A locally stored database of MLST alleles was created in Ubuntu Linux, using the definitions from the pubMLST database (Jolley & Maiden, 2010: http://pubmlst.org/campylobacter/) and created using BLAST (Altschul *et al*, 1990; Madden, 2002; Camacho *et al*, 2008).

The database was then subsequently queried using nucleotide BLAST (Altschul *et al*, 1990; Madden, 2002; Camacho *et al*, 2008) for matches for each isolate using the raw FastQ sequence files, producing output files describing matching alleles and resulting Sequence Types and Clonal Complexes. SRST also requires model flanking regions for each allele, which were provided from *C. jejuni* 81116.

3.2.5 Core Genome Alignment & Phylogenetic Trees

Phylogenetic analysis is the inference of evolutionary relationships, the output of which is typically displayed as a phylogenetic tree. Phylogenetic trees portray the evolutionary relationships between genetic sequences, in contrast to determining sequence similarity (Rokas, 2011). Phylogenetic trees may be rooted or unrooted: rooted phylogenetic trees include a root at the earliest common ancestor, displaying the direction of evolution; whereas unrooted trees demonstrate distance between sequences but not the direction of evolutionary progress (Rokas, 2011).

Phylogenetic trees were created using 33 *Campylobacter* genomes, including the newly sequenced six porcine *C. jejuni* and four porcine *C. coli*, and a variety of publicly available genomes (Table 3.1). FastA file sequences were aligned using MUGSY (Angiuoli & Salzberg, 2010) to rapidly produce multiple genome alignments without the need for a reference genome. MUGSY is a tool which incorporates four steps and utilises some pre-existing programs within a simple user friendly process, combining: Nucmer pairwise alignment (Kurtz *et al*, 2004); construction of alignment graphs and subsequent refinement (Rausch *et al*, 2008; Doring *et al*,

2008); detection of local colinear blocks (LCBs) and calculation of alignment for LCBs (SeqAn::TCoffee, Rausch *et al*, 2008), within a single wrapper script.

Following completion of MUGSY alignment, the resulting output file was then post-processed to convert the multiple alignment file (.maf) into a single concatenated FastA (.fa) file, which was stripped of non-alphabetic characters ('-'; '.' Schloss *et al*, 2009), before being converted to an appropriate file type - phylip (.phy) - for tree generation using Randomised Accelerated Maximum Likelihood software (RAxML, Rokas, 2011).

RAxML produces a 'most likely' phylogeny by encompassing different search strategies in turn; first by building a starting tree by iteratively selecting random sequences from the query set and determining location. This random selection means that the outcome can be different between repeats on the same sample, therefore the option (-x) in the software was used to dictate the start point for each analysis to minimise the effect of the random selection between trees of the same sequences (-x 12345).

Strain ID	Species	Country of Isolation	Host	Date of Isolation	Sequence Type	Sequence Type Complex
1336	*C. jejuni jejuni*	-	Water	-	841*	~
414	*C. jejuni jejuni*	UK	Water Vole	2000	3704	~
81-176	*C. jejuni jejuni*	USA	Milk	1981	604*	42*
S3	*C. jejuni jejuni*	UK	Sheep	1999	632	42
IA3902	*C. jejuni jejuni*	USA	Sheep	2006	8	21
269.97	*C. jejuni doylei*	South Africa	Human	1997	1845	~
ICDCCJ07001	*C. jejuni jejuni*	China	Human	2007	2993	362
81116	*C. jejuni jejuni*	UK	Human	1981	267	283
M1	*C. jejuni jejuni*	UK	Human	1999	137	45
11168	*C. jejuni jejuni*	UK	Human	1977	43	21
01/10	*C. jejuni jejuni*	UK	Human	-	104*	21*
01/51	*C. jejuni jejuni*	UK	Human	-	19*	21*
RM1221	*C. jejuni jejuni*	USA	Chicken	1997	354	354
857	*C. jejuni jejuni*	UK	Pig	2000	270	403
549.1	*C. jejuni jejuni*	UK	Pig	1999	403	403
623	*C. jejuni jejuni*	UK	Pig	1999	552	403
304	*C. jejuni jejuni*	UK	Pig	1999	551	403
484	*C. jejuni jejuni*	UK	Pig	1999	435	403
444	*C. jejuni jejuni*	UK	Pig	1999	553	403
RM2228	*C. coli*	USA	Chicken	1998	1063	828
JV20	*C. coli*	-	Human	-	860*	828*
111-3	*C. coli*	USA	Pig	2001	1467*	828*
132-6	*C. coli*	USA	Pig	2001	3861*	~
151-9	*C. coli*	USA	Pig	2001	1102*	~
59-2	*C. coli*	USA	Pig	2000	890*	828*
67-8	*C. coli*	USA	Pig	2000	1061*	828*
7—1	*C. coli*	USA	Pig	2001	3860*	~
84-2	*C. coli*	USA	Pig	2000	113*	828*
90-3	*C. coli*	USA	Pig	2001	3862*	~
99/321	*C. coli*	Denmark	Pig	1999	1153*	828*
03/121	*C. coli*	UK	Pig	2003	887*	828*
03/103	*C. coli*	UK	Pig	2003	2732*	828*
03/317	*C. coli*	UK	Pig	2003	1145*	828*

Table 3.1: Isolates included in Phylogenetic Analysis

Details of isolates whose genome sequences were used for phylogenetic analysis.
*determined *in silico* in lieu of formal MLST data.
- data not available or unclear.
~ Sequence type either not able to be assigned to a current MLST complex, or may represent a novel clonal complex.

Additionally, during RAxML, the trees are subjected to 'Lazy Subtree Rearrangement' (LSR), whereby subtrees are removed and reinserted at every possible point on the tree to find the most likely correct position; LSR is applied to the original starting tree and to each best tree determined, until no better tree can

be determined. RAxML was run through 100 bootstrap repetitions, with a defined random seed, using the GTRGAMMA model for determining rate of nucleotide heterogeneity. Following the RAxML process, best trees were subsequently visualised and edited for presentation using FigTree software (Rambaut, 2007: http://tree.bio.ed.ac.uk/software/figtree/).

3.2.6 BLAST Ring Image Generator (BRIG)

In order to visualise regions which were consistent across the *C. jejuni* ST403CC isolates, or additional in the ST403CC isolates compared to reference *C. jejuni* the BLAST Ring Image Generator (BRIG; Alikhan *et al*, 2011) was used. BRIG uses BLAST searches and CGView (Stothard & Wishart, 2005) to produce images demonstrating similarity between a reference sequence and query sequences as a series of concentric rings. BRIG was used to create a series of images; comparing the pig *C. jejuni* against each other, against the pig *C. coli* and against human and chicken *C. jejuni;* comparing pig *C. coli* against each other, against the pig *C. jejuni*, and against human and chicken *C. coli*.

BRIG was run using the Graphical User Interface (GUI), using the default BLAST settings in each case, the first two rings (working from the centre outwards) represent the G:C content and G:C Skew, respectively. Each subsequent ring represents the genome of an isolate, as indicated by the colour coded legend included in the output image. In each output image, three gradients of colour may be seen per ring; the darkest shade represents 100% similarity with the lighter shade indicating the highest cut off value, and the lightest shade the lower cut-off value, whilst white space indicates highly disparate or missing regions. Cut-off values used were 90% and 70% for same species comparisons, and 70% and 50% for across species comparisons.

3.2.7 Single Nucleotide Polymorphisms (SNPs)

An additional method used to investigate the relatedness of the ST403CC *C. jejuni* isolates to each other, and to other *C. jejuni* isolates was a measure of Single Nucelotide Polymorphisms (SNPs) using SMALT (Ponstigl,

2010; https://www.sanger.ac.uk/resources/software/smalt/) and SAMtools (Li *et al*, 2009). SMALT (Ponsitgl, 2009) produces alignments of a query sequence (paired fastQ) against a reference sequence (fastA) by producing a hash table for the reference sequence and subsequently mapping the query reads against it.

SMALT was run in Linux using using a word length (-k) of 17, and a sampling step size (-s; the distance between successive words) of 2 so that every second word was hashed (the default –s is the same as the –k). A higher value for –s reduces the memory demand on the machine; given the power of the machine and size of genome the –s was reduced to 2 for greater detail.

SamTools (Li *et al*, 2009) was used to assess SNP variation between the reference and query sequence based upon the mapping produced using SMALT. SamTools was run in Linux.

3.3 Results

The relatedness of a group of potentially host-adapted *C. jejuni* isolates was investigated by a range of genome sequence based means. The isolates were previously demonstrated to represent different STs within an MLST Clonal Complex (Manning *et al,* 2003), and were earlier shown to be potentially human pathogenic strains (Chapter 2). Genomic DNA was prepared, and subsequently sequenced at The University of Exeter, and sequence data was considered in a variety of manners.

3.3.1 Genomic DNA Extraction

Genomic DNA was extracted from each isolate to provide DNA of suitable quality and concentration for sequencing. Quality control was carried out quantifiably using spectrophotometry, and qualitatively by running through an agarose gel.

Figure 3.3 shows the resulting samples which were subsequently submitted for sequencing. Each of these samples were within acceptable ranges for quality and diluted to a suitable concentration of approximately 10µg or more per 100µl of buffer. The bands shown in Figure 3.3 illustrate varying strengths of solution, but all were within acceptable boundaries. Additionally some shearing can be seen in some lanes (particularly lane 9, *C. jejuni* 304), however due to the high concentration and otherwise good quality these were deemed acceptable for submission. The faint bands visible at the bottom of the gel are RNA left over from

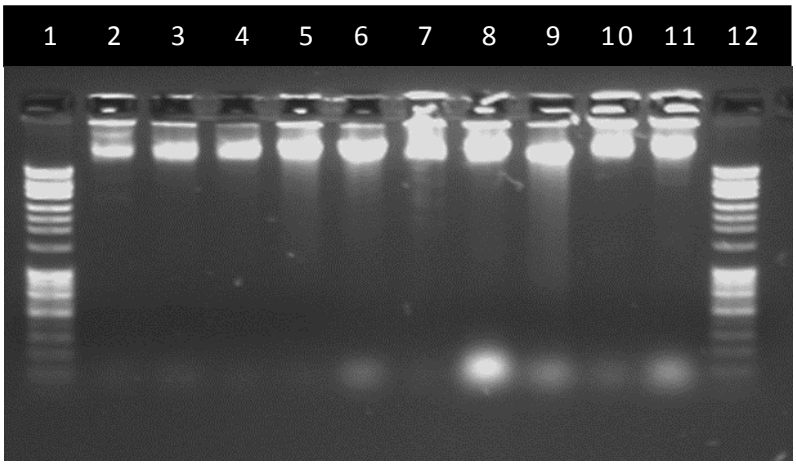

Figure 3.3: Genomic DNA Extracts
The above agarose gel image shows DNA extracts submitted for sequencing. It was produced using a 0.5% agarose gel with Norgen HighRanger Plus 100bp ladder, ran for 25 minutes at 95v, and visualised at 200ms exposure.
Lane 1 Norgen HighRanger Plus 100bp ladder
Lane 2 *C. coli* 03/103
Lane 3 *C. coli* 03/317
Lane 4 *C. jejuni* 549.1
Lane 5 *C. jejuni* 857
Lane 6 *C. jejuni* 444
Lane 7 *C. coli* 03/121
Lane 8 *C. jejuni* 484
Lane 9 *C. jejuni* 304
Lane 10 *C. coli* 99/321
Lane 11 *C. jejuni* 623
Lane 12 Norgen HighRanger Plus 100bp ladder

the extraction process. This could have been prevented by RNAase treatment during the extraction process; however, its presence does not affect the subsequent sequencing process.

3.3.2 MLST

MultiLocus Sequence Types were recorded, where available, from the pubMLST database, and were also observed for each newly sequenced isolate using whole genome sequence based techniques. Sequence types were determined using SRST (Inouye *et al,* 2012) from the raw FastQ sequence data, and also using the pubMLST database, based upon the assembled FastA sequences.

As shown in Table 3.2, some discrepancies were observed between the original MLST done for the ST403CC *C. jejuni* isolates, and the subsequent *in silico* results, however agreement was consistent for all isolates across the two *in silico* methods. Of the six *C. jejuni* ST403CC isolates, four (857, 549.1, 623, 484) were consistent across all three methods, while *C. jejuni* 304 varied at four loci between the original MLST results and the two *in silico* measures, and 444 had one discrepancy between the MLST result and *in silico* methods. Figure 3.4 shows visual representation of the nucleotide variation between the predicted alleles.

As described in Table 3.2, the MLST profile of *C. jejuni* 304 differed at four loci between the previously carried out MLST (Manning *et al*, 2003), and the newly observed *in silico* techniques. The four divergent loci were *glnA*, *gltA*, *glyA* and *tkt*. The allele determined for *glnA* by MLST (allele 27) differed from the allele determined by genome sequence based methods (allele 54) by a single base variation, an A-G substitution at position 139. The differing alleles for *gltA* (25 and 59) vary by two A-G substitutions at positions 12 and 348. More variation existed between the divergent alleles for *glyA* (alleles 19 and 4), with a total of 8 differences spread across the length of the locus, whilst the fourth allele variation, *tkt*, was a single base change at position 28, a substitution between T (allele 22) and C (allele 5). As a result of these differences, *C. jejuni* 304 was designated as Sequence Type 2676 by *in silico* methods, as opposed to ST551 as was initially described; however, both of these STs remain within the ST403 Clonal Complex.

The second isolate with variation in observed MLST results was *C. jejuni* 444, which differed only at one loci; *tkt*, and showed the same discrepancy as *C. jejuni* 304 at this locus – a single base change at location 28 (T-C substitution) leading to

designation as allele 5 instead of allele 22 as described by MLST. This single nucleotide difference for *C. jejuni* 444 means that it would be a novel ST, rather than part of ST553, although once again it would remain part of the ST403CC.

Strain	Method	aspA	glnA	gltA	glyA	pgm	tkt	uncA	ST	ST-Complex
857	MLST	10	27	43	19	6	18	7	270	403
	in silico	10	27	43	19	6	18	7	270	403
	SRST	10	27	43	19	6	18	7	270	403
549.1	MLST	10	27	16	19	10	5	7	403	403
	in silico	10	27	16	19	10	5	7	403	403
	SRST	10	27	16	19	10	5	7	403	403
623	MLST	10	27	59	4	10	5	7	552	403
	in silico	10	27	59	4	10	5	7	552	403
	SRST	10	27	59	4	10	5	7	552	403
304	MLST	10	27♦	25♦	4♦	10	22♦	7	551	403
	in silico	10	54♦	59♦	19♦	10	5♦	7	2676	403
	SRST	10	54♦	59♦	19♦	10	5♦	7	2676	403
484	MLST	10	27	43	19	10	5	7	435	403
	in silico	10	27	43	19	10	5	7	435	403
	SRST	10	27	43	19	10	5	7	435	403
444	*MLST*	10	27	16	4	6	22♦	7	553	403
	in silico	10	27	16	4	6	5♦	7	ND**	403
	SRST	10	27	16	4	6	5♦	7	6601***	403
99/321	MLST	~	~	~	~	~	~	~	~	~
	in silico	33	39	30	82	118	35	17	1153	828
	SRST	33	39	30	82	118	35	17	1153	828
03/121	MLST	~	~	~	~	~	~	~	~	~
	in silico	33	38	30	82	104	85	68	887	828
	SRST	33	38	30	82	104	85	68	887	828
03/317	MLST	~	~	~	~	~	~	~	~	~
	in silico	33	39	30	82	104	44	17	1145	828
	SRST	33	39	30	82	104	44	17	1145	828
03/103	MLST	~	~	~	~	~	~	~	~	~
	in silico	53	38	30	82	104	43	36	5345	828
	SRST	53	38	30	82	104	43	36	5345	828

Table 3.2: MLST Results for Newly Sequenced *C. jejuni* & *C. coli* Isolates

Results for MLST by traditional MLST sequencing, *in silico* determination using pubMLST and using the SRST software.
~ Test not completed.
♦ Discrepancy observed between MLST and *in silico* methods.
**ND - Sequence type not defined by test method.
*** Novel sequence type by test method.

Figure 3.4 shows the relevant segments of alignments where discrepancies occurred between Sequence Typing methods. MLST results for the four *C. coli* isolates were entirely consistent between the web-based PubMLST and the SRST (Inouye *et al*,

2012) attribution methods. Each of the four isolates was attributed to a different, pre-existing Sequence Type, however all four were within the ST828 Clonal Complex. Additionally, all four *C. coli* isolates shared the same alleles at *gltA* (30) and *glyA* (82).

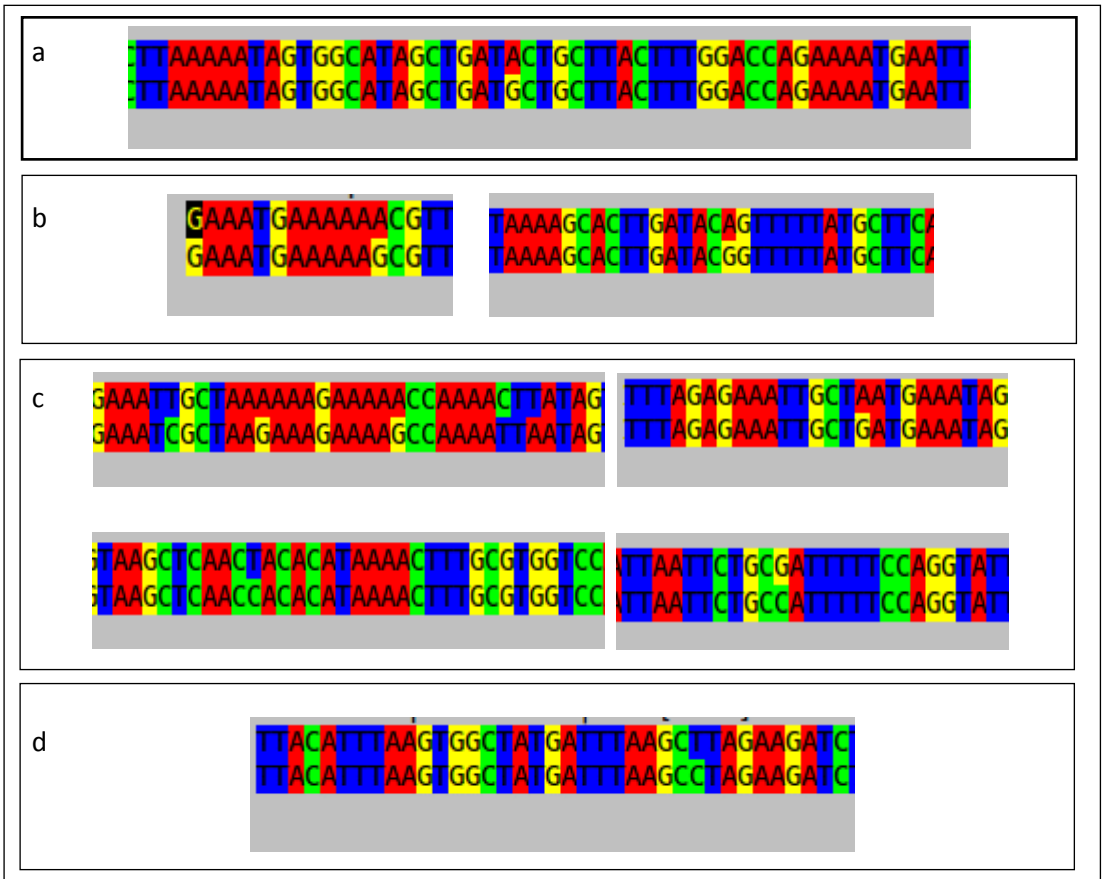

Figure 3.4: Alignments for Discrepant MLST Alleles

In each image, the top row represents the allele determined by the original MLST, and the bottom row represents the discrepant allele observed by *in silico* methods.
a) *glnA* top line allele 27, bottom line allele 54. A-G substitution at position 139.
b) *gltA* top row allele 25, bottom row allele 59. A-G substitutions at positions 12 and 348.
c) *glyA* top row allele four, bottom row allele 19. Eight substitutions in total; T-C at position 114, A-G position 120; A-G position 129; C-T at position 136; T-A at position 138; in second image, A-G substitution at position 202; third image shows the T-C substitution at position 309; fourth image shows the G-C substitution at position 390.
d) *tkt* top row allele 22, bottom row allele five. One T-C substitution at position 28.

3.3.3 Core Genome Alignment & Phylogenetic Trees

Figure 3.5 shows the resulting output from the MUGSY (Angiuoli & Salzberg, 2010) and RAxML (Rokas, 2011) pipeline, following visualisation using FigTree (Rambaut, 2007) and after additional annotation, for the thirty three included genomes. Each

genome was labelled at the end of the branch, and colour coded according to source. The scale bar segment visible in Figure 3.5 illustrates the length of branch representing an amount of genetic change of 0.04; the units of this measure of genetic change are the number of nucleotide changers per site (number of substitutions ÷ length of sequence).

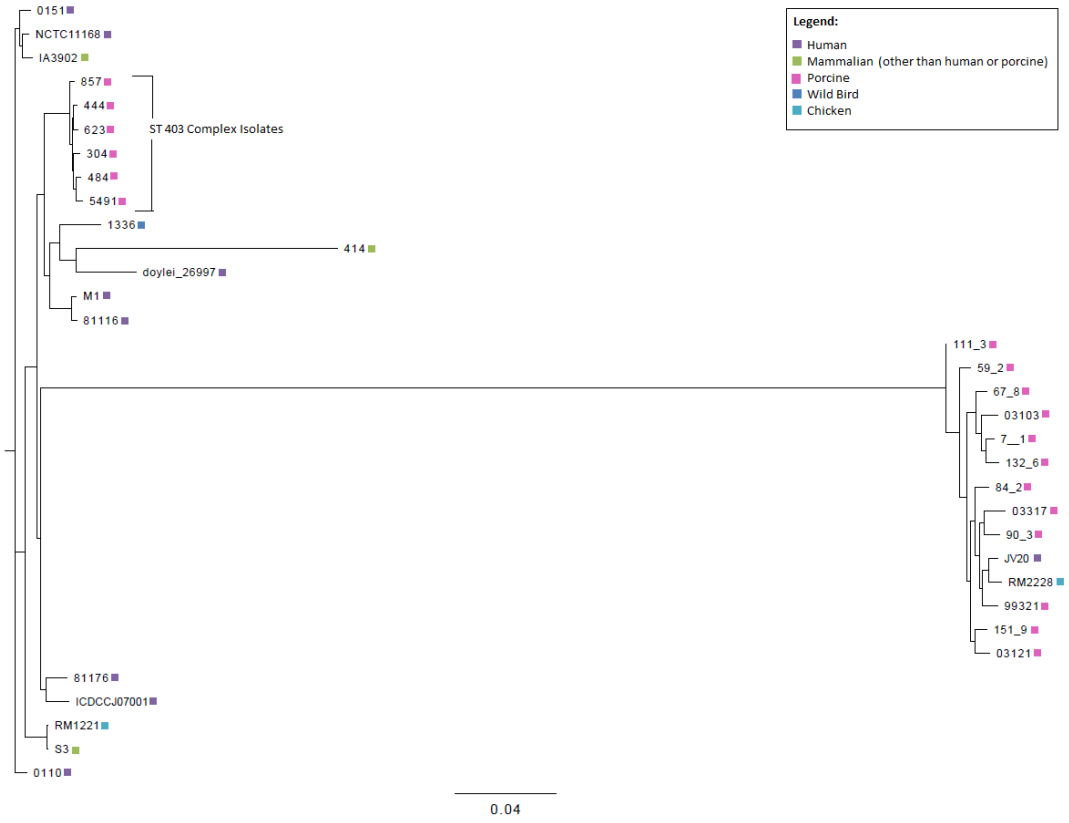

Figure 3.5: Phylogenetic Tree for *C. jejuni* & *C. coli* Isolates

Phylogeny tree based upon core genome alignment of thirty-three *C. coli* and *C. jejuni* whole genome sequences.

The phylogeny tree (Figure 3.5) shows a clear species divide between *C. jejuni* and *C. coli* isolates. The phylogeny of *C. coli* isolates, although limited by the small range of dates (1998-2003), countries (Denmark, UK, USA) and sources (human, chicken, pig), did not appear to be affected by these factors, with observed grouping not reflecting country or date of isolation, or source. A possible exception to this was observed in that the closest relative to the chicken isolate RM2228 was the human microbiome project isolate JV20 (AEER01000000), however, the nearest ancestor to these two isolates was 99/321, the Danish porcine isolate.

The phylogeny of *C. jejuni* isolates did not show a clear association of source, date or location, with the exception of the ST403CC *C. jejuni* isolates, which were shown to be closely related to each other, and formed a distinct cluster within *C. jejuni* phylogeny. The same techniques were used to produce a second phylogeny tree, using just the *C. jejuni* genomes in order to provide a clearer image of the relatedness of the nineteen *C. jejuni* isolates (Figure 3.6). The most distant isolate was *C. jejuni* 414, the bank vole niche specialist, followed by *C. jejuni doylei* 269.97 and the wild bird isolate 1336. The ST403CC *C. jejuni* isolates were closely grouped together within the phylogeny tree. *C. jejuni* 857 was revealed to be the most ancestral of the sequenced ST403CC isolates, with the remainder branching off from it, the most distant of which being 484 and 549.1. The ST403CC *C. jejuni* isolates were seen to represent a distinct clonal group within the *C. jejuni* genomes. Also included in Figure 3.6 is MLST information for the *C. jejuni* isolates (as previously described in Table 3.1), Clonal Complex was used where one has been described, for those which have not been assigned a clonal complex, the ST was used. In general, the location and date of isolation did not affect phylogeny, with distant isolates such as 81176 (USA, 1981) and ICDCCJ07001 (China, 2007) being closely related. Clonal Complex however did concur with whole genome phylogeny, as evidenced both by the grouping together of the ST403CC porcine isolates, and also by the grouping of ST21CC isolates, over a range of dates, and being from both UK and USA and from both human and sheep isolates.

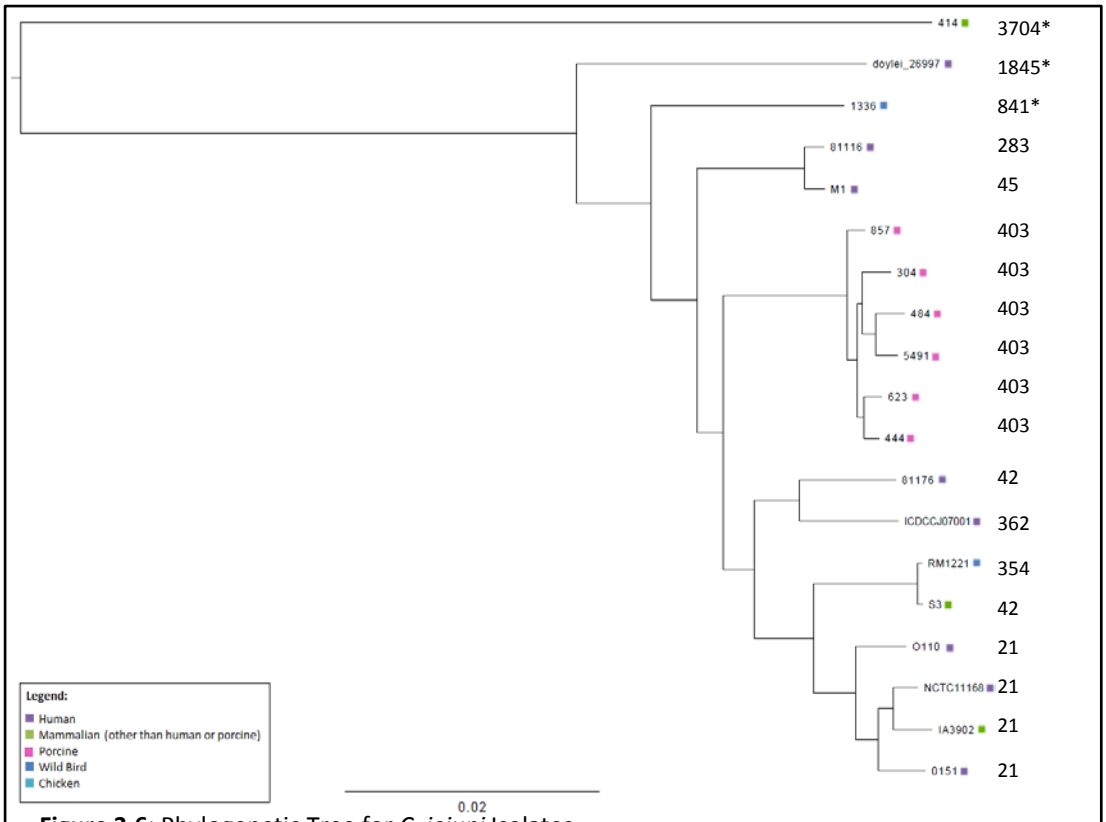

Figure 3.6: Phylogenetic Tree for *C. jejuni* Isolates
Maximum likelihood tree based upon core genome alignment for nineteen *C. jejuni* isolates. MLST Clonal Complex was labelled alongside the isolate name and host legend
* Sequencce Type labelled as Clonal Complex not confirmed.

3.3.4 BLAST Ring Image Generator (BRIG)

BRIG (Alikhan *et al*, 2011) was used to provide a first look at the variation between the ST403CC *C. jejuni* isolates and specific reference isolates, and across the ST403CC isolates. As described above, BRIG produces circular diagrams illustrating BLAST search matches between a reference sequence and query sequences, indicating similarity and discrepancy between the reference and each query sequence.

BRIG was used to show the similarities between reference genome *C. jejuni* 81116, a well-studied historical human isolate, and the ST403CC *C. jejuni* isolates (Figure 3.7). Several regions were observed at which all six of the ST403CC *C. jejuni* isolates were highly divergent from the genome of *C. jejuni* 81116, including some particularly large regions around 50kb, 270-290kb, 650kb, 1080-110kb, 1250-1270kb, 1340-1370kb, 1470kb, 1505kb and 1600kb approximately. Interestingly, around 990-1010kb some regions of discrepancy were observed between *C. jejuni*

81116 and only some of the ST403CC isolates, particularly one clear region which was discrepant in 857, 623 and 444. However, other than this region, the highly divergent regions between *C. jejuni* 81116 and the ST403CC isolates were consistent across the group – little evidence was revealed by this method of individual ST403CC isolates differing from the reference genome where the remaining ST403CC isolates did not.

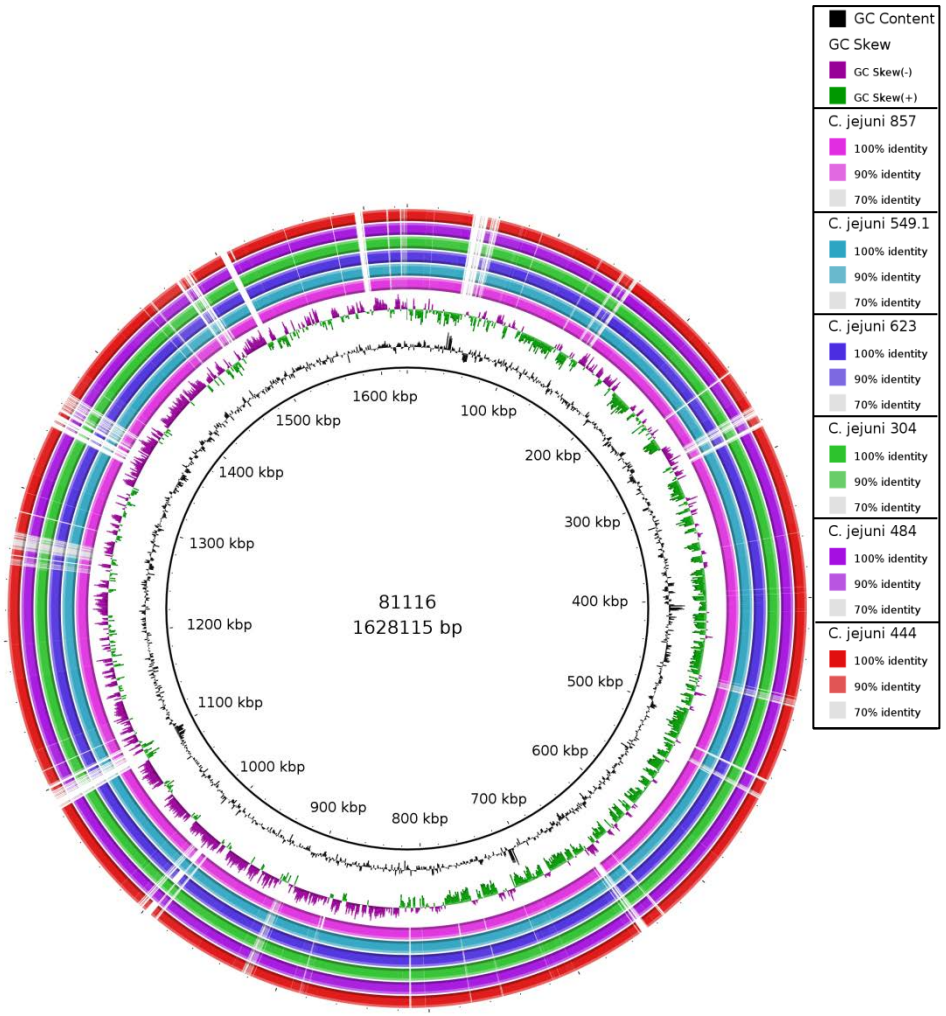

Figure 3.7: BRIG Output for *C. jejuni* 81116 as Reference Genome against ST403CC *C. jejuni* Isolates

Image created with BLAST Ring Image Generator using the human *C. jejuni* isolate 81116 as a backbone for comparison against 6 pig *C. jejuni* sequences (inside to out 857, 549.1, 623, 304, 484, 444). Solid colour indicates 100% similarity, with the fainter colours indicating 90% and 70% identity; areas which are colourless suggest zero similarity.

Highly divergent regions between the six ST403CC isolates and the reference genome 81116 were observed at several points, including a substantial region around 50kb, smaller regions around 160kb and 240kb, several segments around 270-90kb, another region around 470kb, and others around 520kb, 535kb, 635kb and 650kb. There were several small regions around 720kb, 750kb, 770kb, 800kb, 815kb, 885kb and 920kb. Some inconsistent regions of variation were present at around 990-1010kb. Another larger region of discrepancy was observed around 1080-1100kb and another at 1250-1270kb, followed by a smaller region around 1280kb and another larger region around 1340-1370kb. Two more small regions of discrepancy were observed around 1440kb, followed by a group around 1470-1480kb and a larger solid region of discrepancy around 1505kb and some small regions around 1550kb and 1580kb. Another large region was observed around 1600kb, and some final small regions of variation were observed around 1620kb.

A similar comparison was also carried out with another well-known isolate; *C. jejuni* 11168 (another historical human isolate from the UK), as shown in Figure 3.8.

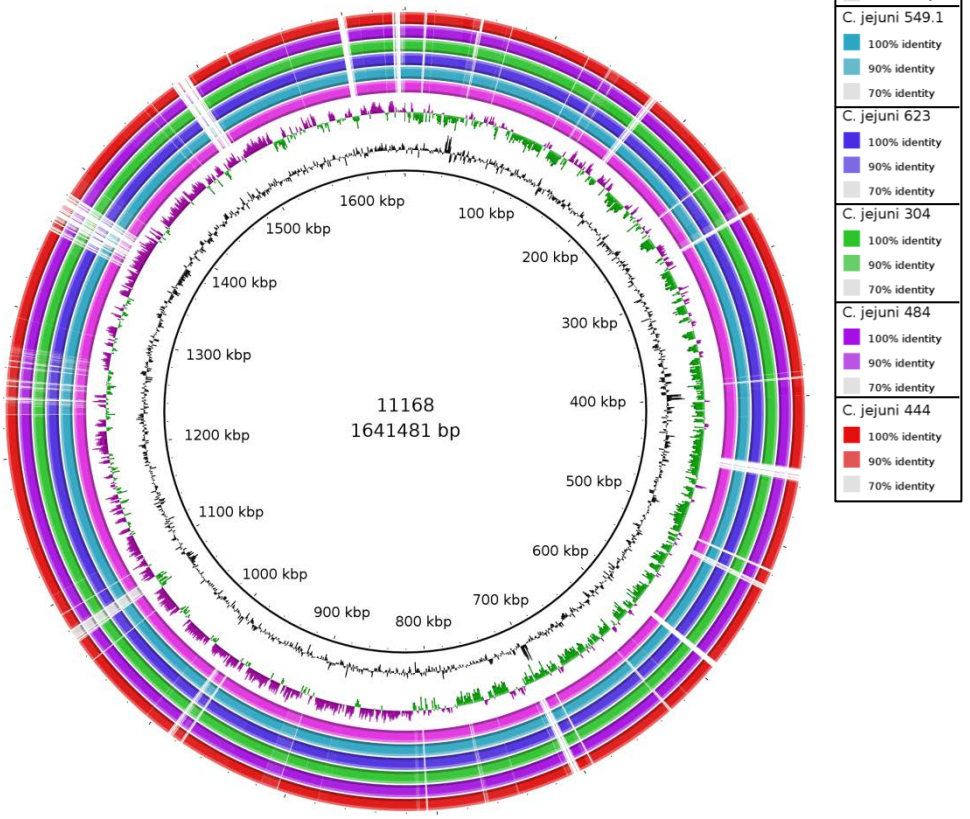

Figure 3.8: BRIG Output for *C. jejuni* 11168 as Reference Genome against ST403CC *C. jejuni* Isolates

Image created with BLAST Ring Image Generator using the human *C. jejuni* isolate 11168 as a backbone for comparison against 6 pig *C. jejuni* sequences (inside to out 857, 549.1, 623, 304, 484, 444). Solid colour indicates 100% similarity, with the fainter colours indicating 90% and 70% identity; areas which are colourless suggest zero similarity.

Highly divergent regions between the six ST403CC isolates and the reference genome 81116 were observed at several points, including around 10kb, 60kb, 80kb, 125kb, 140kb, 150kb, a larger region around 170-180kb, 240kb, 260kb, 380kb. Another larger region of discrepancy was observed between 440-460kb, with other regions around 520kb, 530kb, 590kb, 620kb, 630kb, and a group around 690-710kb. Smaller regions were observed around 740kb, 760kb, 810kb, 900kb, followed by a region around 970-990kb, another small divergent area around 1055kb, and another larger region between 1070-1080kb. A region of small divergent segments was observed around 1220-1280kb, followed by another large region around 1350-1380, a large region around 1480-1500kb, small region around 1560kb and two large regions around 1580-1605kb and 1630-1640kb.

Once again, several regions were observed where all six ST403CC isolates were noticeably and consistently divergent from the reference genome *C. jejuni* 11168, including large regions at approximately 170-180kb, 440-460kb, 1070-1080kb and 1480-1500kb. There also remained little evidence of individual ST403CC isolates being highly divergent from the reference genome, with the exception of a region in *C. jejuni* 444 around 1520kb which was highly conserved amongst the other isolates.

In addition, each of the ST403CC *C. jejuni* isolates were used for reference comparison against the other ST403CC isolates, as shown in Figure 3.9. This overview demonstrated that, despite the close phylogenetic relationship of the ST403CC *C. jejuni* isolates, evidence of strain-strain variation existed across the genomes, although large amounts of 90-100% identity were also observed. Few instances were observed where five of the six strains differed significantly from the reference genome; when *C. jejuni* 623 was used as the reference genome two examples were observed. At approximately 1420kb, 623 shared high similarity with 857 but showed high divergence in 304, 484, 444 and 549.1, whilst at approximately 1680kb, high similarity existed between 623 and 444, with high divergence in the other ST403CC isolates. Other instances of this were observed when *C. jejuni* 304 was used as the reference genome, including at approximately 340kb where high similarity was shared with 549.1 and high divergence from the others, and 1450kb where the region was conserved in 857, and distinct in the other isolates.

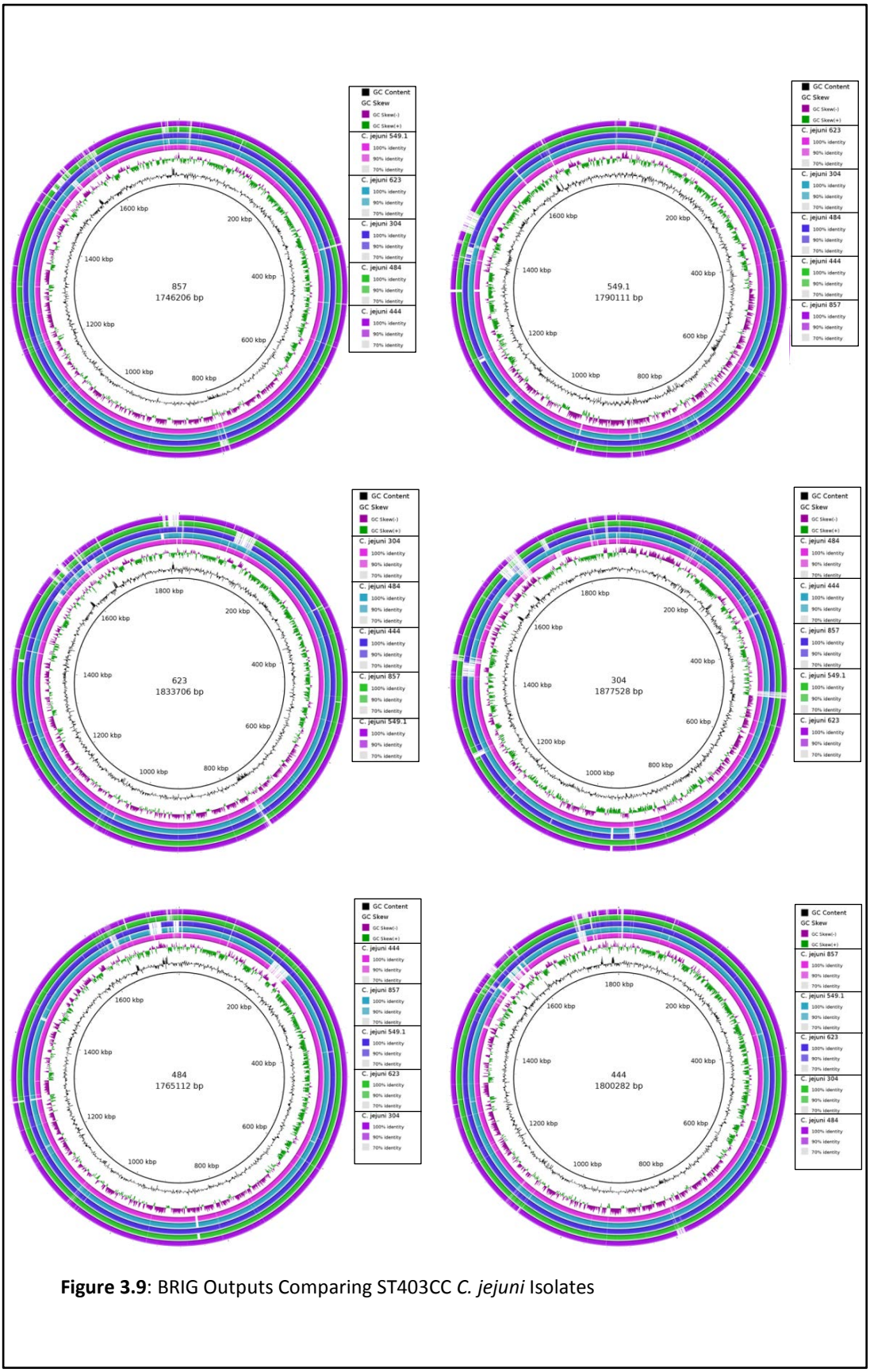

Figure 3.9: BRIG Outputs Comparing ST403CC *C. jejuni* Isolates

3.3.5 Single Nucleotide Polymorphisms

SMALT (Ponstigl, 2010) and SAMtools (Li *et al*, 2009) were used to assess the number of SNPs between isolates. SNP analysis gave an overview of the total number of single nucleotide changes between two isolates, as a measure of variation between isolates.

Reference Strain	Query Strain	No. of SNPs
C. jejuni 81116	*C. jejuni* 484	1110
C. jejuni 857	*C. jejuni* 484	372
C. jejuni 549.1	*C. jejuni* 484	256
C. jejuni 623	*C. jejuni* 484	2831*
C. jejuni 304	*C. jejuni* 484	662
C. jejuni 444	*C. jejuni* 484	413
C. jejuni 81116	*C. jejuni* 623	9846
C. jejuni 857	*C. jejuni* 623	2341
C. jejuni 549.1	*C. jejuni* 623	2736
C. jejuni 304	*C. jejuni* 623	2651
C. jejuni 484	*C. jejuni* 623	2831
C. jejuni 444	*C. jejuni* 623	1734
C. jejuni 81116	*C. jejuni* 857	1166
C. jejuni 81116	*C. jejuni* 01/51	1459
C. jejuni 81116	*C. coli* 03/317	52080
C. coli 99/321	*C. coli* 03/317	2374
C. coli 03/121	*C. coli* 03/317	1791
C. coli 03/103	*C. coli* 03/317	3745

Table 3.3: SNP Results

The table shows the number of single nucleotide polymorphisms calculated between the reference genome and query genome in each case. *623ref484 is approximate – it is taken from the result for 484ref623. As part of the investigation I looked at differences caused by switching reference and query, and found that for 857 vs 484 the variation was 26 calls (372 vs 398) so although it is likely the actual number is different, it is unlikely to vary by more than approximately 7-10% so the number still provides a useful indicator of disparity between strains.

SNPs were determined from *C. jejuni* 81116 for the closest relative amongst the ST403CC *C. jejuni* isolates (857), one of the most distant (484) and one from the middle branch (623), as well as for a *C. coli* isolate (03/317), and a hyper invasive *C. jejuni* isolate (01/51). The number of SNPs observed between the low invasive, human isolate *C. jejuni* 81116 and ST403CC *C. jejuni* 857 (1166) was comparable to

the number of SNPs between 81116 and the hyper invasive isolate 01/51 (1459), as well as being consistent with the number between 81116 and the more distant ST403CC isolate 484 (1110), however a much higher SNP difference was observed between 81116 and 623. The number of SNPs between reference *C. jejuni* genome 81116 and a representative *C. coli* sequence was also calculated for comparison, and, expectedly was much larger even than was observed for *C. jejuni* 623. SNPs across porcine *C. coli* isolates were also assessed and were observed at a similar level to those between *C. jejuni* 623 and the other ST403CC isolates.

SNP comparisons were also completed against the raw FastQ sequence of both the distant isolate 484 and the intermediary isolate 623 for each of the ST403CC isolates. *C. jejuni* 623 was considerably more SNPs away from the other ST403CC isolates, than was observed for *C. jejuni* 484. Fewer SNPs exist between the closely related 623 and 444 than between 623 and other isolates. SNP analysis of *C. jejuni* 484 suggested that the ST403CC isolates were closely related. It may be suggested that 623 (and 444) may have additional genetic content leading to this larger variation from the rest of the isolates.

SNP analysis was also run comparing the 623 FastA reference file against the 623 FastQ query sequences in order to assess the accuracy of the method – 127 SNPs were observed between the two, this appears to indicate inaccuracy in the data, however 127 bases across a genome size of 1.8mb is a relatively low error rate, and if factored into the other SNP results would not significantly alter the interpretation.

3.4 Discussion

This chapter set out to produce genomic sequence data for six potentially porcine host adapted ST403CC *C. jejuni* isolates, and to perform initial analysis of the evolutionary relationship of these isolates to other non-ST403CC *C. jejuni* and *C. coli* isolates - these aims were achieved as described above, and discussion is included below. Genomic DNA extracts were produced, and were submitted for sequencing by the University of Exeter. Genome sequences were produced by the University of Exeter and subsequently returned to the researcher for assembly and analysis.

Whole genome sequence based methods were used to assess the relatedness of ST403CC *C. jejuni* isolates to each other, and to other *C. jejuni* and *C. coli* isolates. The ST403CC *C. jejuni* isolates had previously been shown to be related by MLST (Manning *et al*, 2003), and were previously demonstrated (Chapter Two) to be capable of potentially causing disease in humans. Phylogenetic studies demonstrated that the ST403CC *C. jejuni* isolates were closely related at the whole genome level, as well as based solely upon the seven MLST genes, and indicated some regions where ST403CC *C. jejuni* isolates were uniformly different from human *C. jejuni* isolates. Phylogenetic investigations also determined that variation was observed between the ST403CC isolates, between individual strains and potential 'sub-groups' within the ST403CC isolates.

3.4.1 Sequencing & Assembly

The sequencing approach used produces high quality short-read sequences, which are accurate but rely upon being 'put together' correctly. The assembly process used in this project was implemented to provide genomes which were assembled, arranged and adjusted to reduce gaps, although base error checking was not carried out, the sequences were completed to a greater degree than much of what is publicly available, as it is common to publish raw short-read sequences; whereas these sequences have been assembled and been treated to improve quality of sequence data.

3.4.2 MLST

MLST was carried out using two *in silico* methods using whole genome sequence data. This served as a 'second opinion' for the ST403CC *C. jejuni* isolates, to confirm that they had be correctly identified and ensure inferences were accurate; and also as an 'indicator' method for the newly sequenced *C. coli* isolates which had not previously been studied by traditional MLST.

The results observed for the ST403CC *C. jejuni* isolates were consistent across the two sequence based methods, however they were in disagreement with the original typing data in some instances, as described previously. *C. jejuni* 304 had differing

alleles between the original MLST and the sequenced based methods at four loci, which were attributed to a total of twelve substitutions, whilst *C. jejuni* 444 was different by just one nucleotide in one loci.

Considering the discrepant alleles – 444 was determined to be a novel ST by the *in silico* methods, whereas *C. jejuni* 304 would be altered from ST551 to ST2676. A search for ST2676 on the PubMLST database (29/07/14) gave only one match for this ST – a UK (Aberdeen) cattle isolate (*C. jejuni* 611) from 2006, submitted by Sam Sheppard.

As the *C. coli* isolates had not previously been typed by traditional MLST, the typing done *in silico* is the only data available for them; however, given the high degree of agreement between the *in silico* methods and traditional MLST data for the *C. jejuni* isolates this may be presumed to be representative. Additionally, the accuracy of these methods has been cited (Inouye *et al*, 2012) so a degree of confidence can be placed in these results. Each of the *C. coli* isolates was placed within a currently described Sequence Type, with at least one entry in PubMLST. *C. coli* 99/321 was determined as ST1153 – currently (29/07/14) there are four entries for ST1153 on PubMLST, all representing pig isolates; three from the USA in 2003 and one from Germany in 2007: two of which were included in the paper on generalist lineages by Gripp *et al* (2011), and another from Miller *et al* (2006) in which it was associated with swine in both recorded ST1153 isolates. *C. coli* 03/121 was identified as being part of ST887, for which there were ten recorded isolates in the PubMLST database; one associated with chicken meat or offal (UK, no date), one associated with human stool sample from gastroenteritis (UK, 2007), one recoded simply as 'other animal' (UK) and seven from pigs (three UK, 2004; one USA no date; three Germany (one in 2007 and two in 2008)). *C. coli* 03/317 was found to be in ST1145, which had eight isolates on PubMLST: two USA pig isolates (no date), a pig isolate from The Netherlands (2004), four UK human gastroenteritis stool sample isolates (one 2004 and three from 2011) and a 'human unspecified' isolate (Switzerland, 2009). Lastly, *C. coli* 03/103 was identified as an ST5345 isolate, a ST which had only one recorded isolate on PubMLST, a UK isolate lacking other information. All four of the *C. coli* isolates were deemed to be part of the ST828 Clonal Complex. This is unsurprising

as the ST828CC is known to be one of the most prevalent clonal complexes within *C. coli*, constituting a large part of the 'clade one' livestock and human isolates as described by Sheppard *et al* (2010).

As described by Inouye *et al* (2012), the main previous alternative to acquiring MLST data from whole genome sequences was web-based (Larsen *et al*, 2012), requiring sequences to be uploaded and relying upon a database which is updated once monthly from the PubMLST databases, whereas with SRST the user builds their own databases prior to use and so can be updated when required and as fresh as possible when being used. Comparing the results of using the PubMLST website to assess the MLST of a genome with that of SRST, both methods were equally successful, however SRST does not require completion of assembled genomes prior to analysis so it can provide this information much more quickly and with the same degree of accuracy, however once the genome is completed it is easier to use the browser based PubMLST service.

It was confirmed that the ST403CC porcine isolates were closely related by MLST, although some discrepancies in typing were observed. MLSTs were also determined for the newly sequenced *C. coli* isolates.

3.4.3 Phylogenetic Trees of Relatedness

Trees visualising the phylogeny of *C. jejuni* and *C. coli* isolates were produced based upon whole genome sequence alignment. Genome sequence data for thirty three isolates (nineteen *C. jejuni* and fourteen *C. coli*) were included and an alignment was produced based upon the shared information between the isolates, and variation within it – core genome alignment. This was interpreted as a tree by predicting the distance of each genome from the other isolates.

As described above, phylogenetic grouping did not seem to be largely affected by country or date of isolation, and although some grouping by host type was observed, this was not clear cut – some, but not all, human isolates grouped together, and often grouped with or near to isolates from other sources such as sheep or chicken. MLST Clonal Complex was reflected by whole genome phylogeny

in the cases of ST403CC and ST21CC, however the two ST42CC isolates were not closely related.

Miller *et al* (2006) investigated host associated alleles in *C. coli* from food animals, including 488 isolates from cattle, chickens, pigs and turkeys from USA locations over a six year period. Miller *et al* (2006) revealed host associated MLST alleles and ST profiles in *C. coli*, and also identified 'common' MLST alleles at each locus which were found across all four sources. Considering the *C. coli* isolates newly sequenced in this work, three had 'common *aspA* allele 33', however 03/103 has the swine associated *aspA*53 allele. Two of the isolates had common allele *glnA*39, whilst two had swine associated *glnA*38 (strains 03/121 and 03/103 had *glnA*38, of which 95% of examples were observed in swine by Miller *et al* (2006)). All four strains had *gltA* 30, a 'common' allele, and all four also had common allele *glyA* 82. For *pgm* three of the four isolates had allele 104, a 'common' allele, one (99/321) had swine associated allele 118 (97% in swine, one instance (3%) of it in chicken). Regarding *tkt*, 99/321 had the swine associated allele *tkt*35 (94% swine, 1% cattle, 5% chicken); 03/103 had the 'common' allele *tkt*43; 03/317 had swine associated allele *tkt*44 (93% swine, 4% chicken, 4% turkey); and the 03/121 allele *tkt*85 showed a 'mild' swine association (72% swine, 22% chicken, 6% cattle). Finally for *uncA*; two isolates shared the *uncA*17 'common allele'; *uncA*36 swine associated allele in 03/103 (96% swine, 3% chicken, 1% turkey); and *uncA*68 swine associated allele was found in 03/121 (95% swine, 3% cattle, 3% chicken). Miller *et al* (2006) state that some host-associated alleles have been found in *C. jejuni* however these are much less common than in *C. coli* and typically associated with avian hosts; examples include *gltA*22, *glyA*48, *pgm*17, *uncA*23, each of which are absent from the six ST403CC swine isolates.

Sheppard *et al* (2010) showed that the effect of host association over-rides geographic association in *Campylobacter*, based upon MLST 'genotypes'. Gripp *et al* (2011) ST21 *C. jejuni* isolates were demonstrated to have phage related genes and high levels of recombination, and therefore high levels of genomic diversity (based upon whole genome sequence analysis), however for this ST no features were identified which linked with host type (in terms of genetic information or chicken

colonisation) and Gripp *et al* (2011) therefore concluded that ST21 is characterised by diversity and flexibility allowing a generalist lifestyle, as opposed to the specialism seen in other species and in other STs/isolates of *C. jejuni*.

A strong link between MLST and host association has been described (Sheppard *et al*, 2010) although some STs do not show this pattern and rather exhibit a generalist genotype (Gripp *et al*, 2011). The work here appears to support the findings of Sheppard *et al* (2010) at the whole genome level; although some isolates do group with isolates from different sources, certainly geographical and temporal associations seem less important. The results from whole genome phylogeny also concur with the conclusions of Gripp *et al* (2011); the ST21 isolates were demonstrated to be closely related at the whole genome level, despite host source.

3.4.4 Variation Indicated by BRIG

BRIG is a simple comparison tool used for quick overviews and to give indication of differences and similarities between one reference strain and one or more query isolates. In this project BRIG was used as an early stage test to visualise the variation between well characterised *C. jejuni* isolates and the newly sequenced ST403CC *C. jejuni* isolates, and to check for homology within the ST403CC *C. jejuni* group.

Region (approx. kb)	Predicted content
40-50	putative membrane proteins, hypothetical proteins
270-290	transferases
650	hypothetical proteins, putative phage repressor protein C8J_0649
1080-1100	LOS genes including *waaC* and *waaF*
1250-1270	LOS/LPS genes including *neuA* and *neuB*, flagella genes *flaA* and *flaB*
1340-1379	Capsule region (*kpsS*, *kpsF*, *kpsD*, *kpsE*, *kpsT*, *kpsM*)
1470	R/M modification subunit *rloB*
1505	Putative transporter proteins
1600	Ribosomal proteins

Table 3.4: Contents of Variable Regions Identified by BRIG

As described above, comparisons between the ST403CC *C. jejuni* isolates revealed large degrees of highly conserved sequence, but also demonstrated that considerable strain-strain variation was present between the closely related isolates. This fits with the literature as *C. jejuni* isolates are known to be highly variable even when closely related. At this stage conclusions could not yet be drawn about the content of the variable regions between the isolates due to the lack of annotation of genes present; this will be addressed in later chapters. It is likely that at least some of this variation, as for the two reference isolates, may be linked to known variable regions such as the capsule and LOS regions, however there remains a possibility that these regions may be shared by ST403CC *C. jejuni* isolates.

BRIG images demonstrated that ST403CC *C. jejuni* isolates share highly similar sequence with reference *C. jejuni* isolates but also display considerable and in some cases consistent variation. Some of these regions of variation appear to be linked to known variable regions, whilst others are not well understood from this method of analysis. Likewise ST403CC *c. jejuni* isolates demonstrate high similarity with each other based upon BRIG investigation, however variation between isolates was also observed, the content of which are considered later.

3.4.5 SNPs

SNP analysis revealed an overview of the extent of variation between *C. jejuni* and *C. coli* sequences. Larger numbers of SNPs were observed between the ST403CC 623 genome and the other ST403CC *C. jejuni* isolates, than across the rest of the ST403CC group. This may indicate that *C. jejuni* 623 has undergone considerable horizontal genetic transfer not present in the other ST403CC isolates.

Determination of SNPs has proved useful in rapid determination of common clonal complexes in *C. jejuni* and *C. coli*. In 2004, Best *et al*, determined three SNPS in two MLST alleles which were common across the ST21 Clonal Complex, and were able to use PCR amplification, hybridisation to reference probe, and analysis of the subsequent melting point temperature to rapidly determine whether isolates formed part of the ST21CC. Best *et al* (2007) further expanded this to provide a

scheme which could rapidly identify isolates directly from faecal samples as being part of the six most common Clonal Complexes (ST21CC, ST45CC, ST48CC, ST61CC, ST206CC, ST257CC) as accurately as by carrying out full MultiLocus Sequence Typing in a fraction of the time. However, this SNP based method was of course limited in that it could only ascribe an isolate to one of these six Clonal Complexes, without determining the actual Sequence Type of the isolate, or recognising novel STs.

SNP comparisons are a useful tool for reducing the large volumes of data of full genome sequences into a smaller, informative dataset (Méric *et al*, 2014), determining the number of SNPs between two sequences of interest, or in closer detail to illustrate the location of SNPs in coding sequence. SNP analysis may be of less value in *Campylobacter* due to the high sequence variability in these genomes, being at risk not only of identifying SNPs caused by sequencing error rather than true SNPs, but also with the risk of being skewed by high volumes of closely packed SNPs caused by divergent regions acquired by HGT in the reference sequence which are absent or highly divergent in the query sequence. Another issue caused by the reference based nature of SNP comparison is that it is unable to identify additional genetic content in the query sequence. The method for determining the presence of SNPs used here assessed the number of SNPs across the whole length of reference and query genome, so values are affected by additional/lost accessory genome content. An alternative method would have been to use the core genome alignment to analyse SNPs however the method used required raw FastQ sequence rather than multiple alignment sequence data, additionally although this would have allowed consideration of the SNPs present in the shared portion of the genomes, it would not provide as true an indication of the variance between the isolates. SNP analysis suggested that considerable variation, potentially due to horizontal gene transfer, may exist between the ST403CC *C. jejuni* isolates. At this stage it cannot be predicted what this sequence variation may be, however the content of the genomes will be considered in later chapters, following the annotation of the ST403CC *C. jejuni* genome sequences in Chapter Four.

3.4.6 Summary & Conclusions

ST403CC *C. jejuni* isolates were demonstrated to be closely related not only by MLST, but also at the whole genome level. Regions of variation were identified where ST403CC *C. jejuni* isolates varied from well-studied human *C. jejuni* isolates, and variation was also observed across the ST403CC *C. jejuni* group. It was shown that the ST403CC *C. jejuni* isolates represented a distinct group within *C. jejuni* but are not a 'fuzzy' intermediary group between *C. jejuni* and *C. coli* species. Fuzzy species have been shown to exist, particularly in *Neisseria* (Hanage, Fraser & Spratt, 2005; Corander et al, 2012) however evidence of this phenomenon was not observed for the ST403CC *C. jejuni* isolates, which do not appear to represent an intermediary group with increased recombination with both *C. jejuni* and *C. coli*.

The results observed in this chapter further supported the idea that these ST403CC *C. jejuni* isolates may be an adapted group of isolates based upon whole genome phylogeny, and suggested that they may be lacking or highly variant in some regions compared to other *C. jejuni* genomes. This chapter also provided genome sequences which will be utilised throughout the rest of the thesis. Having confirmed initial sequence based tests, the next chapter considers the annotation of the genomes and more thorough investigation of the genome content and variation.

Chapter Four: Genome Annotation & Investigating Genome Content of ST403 Clonal Complex *C. jejuni* Isolates

4.1 Introduction

Six ST403CC *C. jejuni* isolates were identified as potentially host adapted isolates (Manning *et al*, 2003), and have been demonstrated to be closely related at the whole genome level (Chapter Three), and exhibit pathogenic potential *in vivo* (Chapter Two). Initial genome comparison investigation indicated that certain regions observed in the ST403CC *C. jejuni* isolates were consistently different from other, well studied, *C. jejuni* genomes. It was observed that some of these regions may represent known variable regions including LOS and capsule, however it was not feasible to draw detailed conclusions on the content of these regions, the likelihood that the same divergent event was conserved across the group, or the possible implications this variance may have on the isolates.

The aim of the current chapter therefore was to follow on from the phylogenetic studies of Chapter Three, via the completion of annotation of the ST403CC *C. jejuni* isolate genomes, and subsequent analysis of these in comparison to reference genomes. A number of specific goals are approached in this chapter: annotation of newly sequenced ST403CC *C. jejuni* isolates; to ascertain the core and pan genome of the ST403CC *C. jejuni* isolates within the context of a library of included genome sequences; to consider the presence and potential function of specific genes of interest which may be related to the regions of variation identified previously (Chapter Three), or which may be revealed by pan genome analysis and inform on potential adaptation in ST403CC *C. jejuni* isolates.

4.1.1 Genome Annotation

Acquisition of genome sequence allows for useful comparisons of relatedness in isolates, as demonstrated in Chapter Three. However in order to better understand the genotype of an isolate, the identification and annotation of coding regions is an important step. In the early days of bacterial genome sequencing, producing complete sequences was costly and time consuming enough that manual

annotation of genomes, or manual curation of automated annotations was practical, however with the boom of next generation sequencing techniques this is not typically fast enough to match the production of new sequences (Richardson & Watson, 2012). The caveat with fully automated annotation however is that it can lead to the propagation of problematic annotations which are inaccurate or lacking consistency (Richardson & Watson, 2012). An additional issue with this is transferring annotation from one sequence to a newly sequenced relative, leading to the omission or poor annotation of additional genomic content in newly sequenced isolates (Richardson & Watson, 2012).

Annotation transfer can be a useful tool in rapidly applying annotation to newly sequenced genomes, although it relies on the quality of the reference genome annotation, and can overlook additional genome content (Otto *et al*, 2011; Richardson & Watson, 2012). Otto *et al* (2011) developed a tool (RATT; Rapid Annotation Transfer Tool) to rapidly annotate new genome sequences based upon a reference genome, at a level which is more accurate than existing automated annotation processes.

The quality of a transferred annotation relies on the quality of the reference annotation being utilised (Otto *et al*, 2011; Richardson & Watson, 2012). Various methods exist to create annotated genomes, including approaches such as PROKKA (Prokaryotic Annotation tool) (Seemann, 2014) which combines numerous existing processes to produce rapid, reliable annotation of bacterial genome sequences.

4.1.2 Sequence Similarity

Assessing the similarity to pre-existing sequences often provides the first step in studying new nucleotide or amino acid sequence (Madden, 2013). Traditionally, the major method for determining sequence similarity was DNA-DNA hybridisation – comparing two genomes using direct hybridisation (Luo, Lin & Xu, in Xu (Ed), 2010). The advance of genome sequencing provided a key opportunity to develop computer-based multiple genome similarity comparisons. Similarity searches are highly valuable in investigating nucleotide or protein sequences, indicating function and relatedness of the sequence, based upon the 'first fact' of biomolecular

sequence analysis as described by Gusfield (Gusfield, 1997, in Mushegian, 2007); the concept that a high degree of similarity between sequences implies substantial structural or functional similarity, and additionally the idea that sequence similarity often also indicates an evolutionary relationship, however it is also important to remember that this link does not work in reverse; proteins may share functional or structural similarity despite having dissimilar sequence and being evolutionarily distinct (Gusfield, 1997 in Mushegian, 2007; Madden, 2013).

BLAST (Basic Local Alignment Search Tool) was first introduced by Altschul *et al* in 1990 and further improved by Altschul *et al* (1997), whilst BLAST+ effectively superseded previous versions in 2009 (Camacho *et al*, 2009) and is currently the predominant method in determining sequence similarity (Madden, 2013). BLAST uses heuristic searching (also referred to as informed searching) as opposed to an uninformed search strategy; uninformed searches simply differentiate between states, whereas a heuristic search strategy is provided with additional information in order to predict which state is more 'favourable' (Russell & Norvig, 2013); this strategy is faster and more computationally efficient as it can discard lower scoring results part way through a search if a better scoring match has already been determined. Heuristic searching is commonly used in biological sequence analysis as it is not generally computationally possible to use other, potentially more accurate methods such as the dynamic programming algorithm, which produces a highly accurate alignment of two sequences but is highly demanding (Altschul *et al*, 1990). BLAST uses informed searching with well-defined mutation scores which allow a faster and less computationally demanding calculation of results which are highly similar to those that would be produced via dynamic programming algorithms (Altschul *et al*, 1990). BLAST combines this informed search strategy with a calculated 'E value' predicting the likely accuracy of the result determined by approximating how many matches would occur at the given score by chance. Additionally, BLAST executes 'local' alignments which, rather than requiring the alignment of the whole sequence (global alignment), is suited to finding conserved/functional domains within proteins/CDS. As a result, BLAST identifies small regions ('words') with high similarity and then 'extends' along the sequence to

identify the boundaries of the region of homology (Altschul *et al*, 1990). In contrast, global searching attempts to align the entire lengths of the sequences (Madden, 2013). BLAST therefore is uniquely able to identify conserved regions and domains or motifs compared to global search methods.

As described by Madden (2013) there now exist numerous versions of BLAST, each based upon the same principle but optimised for different sequence types or for speed in sequences known to be closely related. The BLAST versions include nucleotide query against nucleotide database (blastn, megablast), protein query against protein database (blastp, DELTA-BLAST, PSI-BLAST, PHI-BLAST), translated nucleotide query against protein database (blastx), protein query against translated nucleotide database (tblastn) or translated nucleotide query against translated nucleotide database (tblastx). As well as the various types of BLAST query there are numerous ways to utilise BLAST, primarily based upon either using the web-based tool using the online databases at NCBI, or using stand-alone BLAST on an individual machine, which can either utilise the NCBI databases or create custom locally stored datasets.

This chapter utilises BLAST based similarity searching to confirm previous phenotyping results (Chapter Two; Manning *et al*, 2003), to inform of the presence or absence of specific genes associated with virulence, and to investigate any coding sequences of interest revealed through pan genome content analysis.

4.1.3 Hippurate Hydrolysis

As described previously (Chapter One; Chapter Two), hippurate hydrolysis was commonly used as a discriminatory test to distinguish *C. jejuni* from the other thermo-tolerant *Campylobacter* species (Nicholson & Patton, 1995). Differentiation was based upon the expectation that *C. jejuni* isolates would be positive for hippurate hydrolysis, whilst *C. coli* and *C. lari* would be negative for hippurate hydrolysis. However, it has more recently been established that some *C. jejuni* strains are phenotypically hippurate negative, although the *HipO* gene is retained, as determined by PCR (Amri *et al*, 2007; Caner *et al*, 2008).

Previous work (Stephen On, cited by Manning et al, 2003) demonstrated that the ST403CC *C. jejuni* isolates used in this research were phenotypically negative for hippurate hydrolysis. In this chapter it will be investigated whether this negative result is associated with a lack of the *HipO* gene, as is observed in *C. coli*, or confirm the presence of the *HipO* gene despite the negative phenotype, as described elsewhere (Amri *et al*, 2007; Caner *et al*, 2008).

4.1.4 Surface Polysaccharides & Capsule in *C. jejuni*

It was initially thought that *C. jejuni* produced only LOS (lipooligosaccharide); LPS (lipopolysaccharide) lacking O-antigen repeats and with low molecular weight (Logan & Trust, 1984). However, it was later found that around one-third of *C. jejuni* reference strains for serotyping were also able to produce O-antigen; a high molecular weight LPS molecule (Preston & Penner, 1987). Following on from this, as a consequence of the genome sequencing of NCTC11168 (Parkhill *et al*, 2000) it was discovered that *C. jejuni* possessed a number of genes with homology to capsule polysaccharide biosynthesis proteins in other bacteria, despite the fact that capsule had never been detected in laboratory culture or study of *C. jejuni*. This led to the conclusion that the previously assumed O-antigen polysaccharide was in fact capsular polysaccharide (Karlyshev *et al*, 2000). Following this discovery, Bacon *et al* (2001) demonstrated the role of capsule in virulence of *C. jejuni* 81-176 by revealing that, when compared to wild type, a *kpsM* mutant, rendered unable to produce the high molecular weight glycan, was significantly impaired in both adherence and invasion *in vitro* using INT407 cells and was also impaired in the animal model (ferret). The mutant was also rendered untypeable using HS23 and HS36 antisera - effectively confirming Karlyshev *et al*'s conclusions (2000). In 2001, Karlyshev, McCrossan & Wren were able to visualise capsule in an isolate associated with Guillain-Barré syndrome (*C. jejuni* G1), and for a gastroenteritis associated isolate (*C. jejuni* X) but were not able to demonstrate capsule production in the sequenced gastroenteritis isolate *C. jejuni* 11168, despite the presence of the appropriate gene cluster. Capsule is thought to be potentially synthesised by the majority of *C. jejuni* isolates (Karlyshev, McCrossan & Wren, 2001) and has a role in

evading the host immune response, and has been demonstrated to be specifically involved with serum resistance (Keo *et al*, 2011).

BLAST similarity searching will be used in this chapter to confirm the presence of capsule biosynthesis genes. As the ST403CC *C. jejuni* isolates represent a potentially pig adapted group it is of interest to determine whether they have the potential ability to produce capsule, given its role in surviving the human host response; it has been demonstrated that these isolates have the potential to internalise into human colonic cells and therefore cause disease, however if they are lacking means to survive within the host this colonisation may be unlikely to occur.

4.1.5 Virulence Associated Genes in *Campylobacter jejuni*

Numerous virulence associated genes have been identified in *C. jejuni*, some of which show evidence of source bias (Datta, Niwa & Itoh, 2003). The most well recognised virulence genes in *C. jejuni* are *cadF*, *cdtA*, *cdtB*, *cdtC*, *ciaB*, *dnaJ*, *flaA*, *iamB*, *pldA*, *racR*, *virB11* and *wlaN*, each playing a variety of roles. These include genes associated with adherence and colonisation: *flaA* involved in motility and colonisation (Nuijten *et al*, 2000), *cadF* and *dnaJ* mutations in which have been shown to have a significant effect on colonising ability (*dnaJ* is involved in response to environmental stress encountered within the host, whilst *cadF* is suspected to have a role in binding to host cells (Ziprin *et al*, 2001) and racR which is involved in temperature response during chicken colonisation (Brás *et al*, 1999). Genes associated with invasion ability; *ciaB* and *pldA* mutants are poor invaders (Konkel *et al*, 1999; Ziprin *et al*, 2001) and *virB11* has been shown to confer an advantage in invasion *in vitro* and in pathogenicity in the ferret model (Bacon *et al*, 2000). The genes required to produce cytolethal distending toxin; *cdtA,B* and *C* (Purdy *et al*, 2000), and finally a gene which has been linked to GBS - *wlaN* (Linton *et al*, 2000).

The presence of these virulence associated genes in the ST403CC *C. jejuni* isolates will be assessed in order to relate to the literature, with particular reference to host association as described by Datta, Niwa & Itoh (2003).

4.1.6 Defining the Core & Pan Genome

Typically, in Eukaryote genomics, the genome of a single individual provides the substantial majority of the genetic information for its species (Lan & Reeves, 2000). In Prokaryotes, and particularly in bacteria however, the situation is tremendously different. Bacteria undergo constant, considerable, intraspecies genomic exchange and modification (Lan & Reeves, 2000), ranging from the acquisition or loss of plasmids and/or lysogenic phages and the presence or absence of gene groups such as pathogenicity islands (PAIs), to variation in possession of single genes (particularly those associated with metabolism) and to smaller changes including variation of gene coding sequence, or in copy number, all the way down to single nucleotide polymorphisms (SNPs). As such, a single isolate is unlikely to inform upon the genome of a species; unless perhaps in the case of a species with highly reduced genetic recombination such as *Mycoplasma spp.*, which represents a substantially restricted genome. Liu *et al* (2012) observed a core genome of just 196 genes based upon 20 *Mycoplasma* genomes representing 17 different species from a range of highly divergent hosts. To establish a bacterial 'species genome' would therefore typically require numerous diverse individuals to be sequenced before a moderately confident assessment could be drawn. The species genome may also be referred to as the pan genome (Medini *et al*, 2005) and may be loosely defined in this context as 'all genes found in a species', or more generally as 'the collection of genes which are present in all known genomes of a defined group of organisms' (Lapierre & Gogarten, 2009), however even when numerous isolates have been sequenced for a species, it remains common for new genes to be found in each additional strain sequenced. At the time of writing, Lan & Reeves (2000) stated that it had been demonstrated that up to 20% of one strain's DNA might be missing from another individual of the same species, and mathematical predictions (Medini *et al*, 2005; Tettelin *et al*, 2005) suggest that further new genes will continue to be revealed even with up to hundreds of individual sequences for a species. However, the concept of the pan genome increasing to 'infinity' with additional new genes identified with each additional sequenced isolate is likely a weakness of the mathematical prediction model, as suggested by Snipen, Almøy &

Ussery (2009), although it remains unrealistic to expect that the pan genome of a species can be fully ascribed. The concept of characterising a bacterial species' full genome content is therefore rather like chasing the edge of an ever expanding universe, as there will always be additional unsequenced examples of a species; however, each additional genome sequenced can be used to expand our knowledge of the species' pan genome.

Lan & Reeves (2000) explained that the pan genome should be considered as two elements: firstly, the 'core' set of genes found in most, or all, sequenced individuals; these include those that determine the characteristic properties of that species, including the genes essential for function & survival. It is worth noting that this is not the same as the 'minimal' geneset defined by Hutchinson *et al* (1999) which described the minimum essential genome for laboratory growth of *Mycoplasma genitalium*. And secondly, the 'auxiliary genes' (also commonly referred to as the accessory, or dispensable genome), those which may be present in one or a group of individuals and are not essential for the species (Lan & Reeves, 2000), but may, for example, convey an advantage to certain isolates in a given environment or host (Lefébure & Stanhope, 2009).

4.1.7 Core & Pan Genome in *Campylobacter*

The recent advances in sequencing and annotating multiple isolates within a species has created a new opportunity to study the progression of adaptation through assessing differences in the core genome of bacterial species and sub-species groups (Lefébure & Stanhope, 2009) and expanded the understanding of the pan genome (Lapierre & Gogarten, 2009). In 2009 more than 18 genomes were available for *Campylobacter*, representing eight different species (Duong & Konkel), compared to just one in 2000 (Parkhill *et al*, 2000), and four *C. jejuni* genomes in 2007 (Pearson *et al*, 2007). Today these numbers are exponentially higher, with around 352 of identified individual isolates across 27 species, including 173 *C. jejuni* and 98 *C. coli* available publicly as nucleotide sequences, with varying levels of 'completeness' (based upon a search of ncbi taxonomy browser: http://tinyurl.com/mcplzjd in May 2014).

Various attempts to characterise the core and pan genome of *C. jejuni* have been made, with differing results: for a sample of five genomes, Snipen, Almøy & Ussery (2009) observed a *C. jejuni* core genome of 847 and pan genome of 3221 CDS; whilst in 2010 Friis *et al* estimated *C. jejuni* pan and core genome using the 13 'complete' genomes available at the time, consisting 2427 and 1295, respectively; in 2011 Biggs *et al* recorded a *C. jejuni* core genome of 1001 across 13 genomes; and in 2014, Méric *et al* estimated pan and core genome across two *C. coli* and five *C. jejuni* of 3933 and 1035 respectively, they subsequently expanded their study to a total of 130 *C. jejuni* genomes with a core genome of 947 and pan genome of 3648 CDS.

Previous pan and core genome based research has obtained varying degrees of success in investigating host adaptation in *Campylobacter*. In 2010, Lefébure *et al* were unable to ascertain a link between specific genes and host species in *C. jejuni* but did find some evidence of host type associated genes in *C. coli*, whereas in 2005 Champion *et al* were able to identify distinct livestock and non-livestock clades of *C. jejuni*, including a cluster of livestock associated genes potentially associated with flagella glycosylation.

Host adaptation typically involves the gain of additional genes which confer a survival advantage in the given host (Dale & Park, 2004; Bliven & Maurelli, 2012), or the loss or decay of genes which are no longer necessary in the new environment (Bliven & Maurelli, 2012). Gene loss events are typically associated with pathogenic bacteria, an example being the minimalistic genome of obligate intracellular pathogens such as *Yersinia pestis* (Bliven & Maurelli, 2012). Assessment of the pan genome of a group of isolates allows comparative analysis of the accessory genome which may reveal these markers of adaptation.

An assessment of the pan and core genome will be produced using a total of 33 isolates, as detailed in Chapter Three (Table 3.1; Appendix 9.1) in order to identify genes which may have been gained or lost in the ST403CC *C. jejuni* isolates as a result of host adaptation.

4.1.8 Determining the Core & Pan Genome

The basis for determining which genes are shared across a group of genomes relies traditionally upon BLAST (Altschul *et al*, 1990) or FASTA (Lipman & Pearson, 1985; Pearson & Lipman, 1988) program based searching procedures, which require one-by-one searching of each individual CDS against a database to determine its prevalence against a custom built database such as the collective available genomes for a species in order to construct banks of core or accessory genes across the group. Scripts were developed to automate the input of each CDS for querying however this remained a limited and inefficient method which was particularly affected by the low number of genomes within a species which were historically available (Tettelin *et al*, 2005; Pearson, in Misener & Krawetz (Eds), 2000). Also, in the earlier years of expanding genomic sequencing, the general trend in genome analysis tools previously had been designed with the main goal of comparing *between* species rather than *within* species - such as GeConT, xBASE and others, as reviewed by Blom *et al* (2009). Therefore the modern boom in multiple isolates within species created a need for different analysis tools. The primary manner in which the problem of within species comparisons was addressed was through the creation of databases, notably the Microbial Genome Database (MBGD; Uchiyama, 2003) and the Comprehensive Microbial Resource (CMR; Davidsen *et al*, 2010; no longer active), however these each have limitations in their capabilities. As a result, Blom *et al* (2009) introduced EDGAR (Efficient Database frameworks for Comparative Genome Analysis using BLAST Score Ratios), a browser based solution to provide both CDS orthology information, and analysis of pan, core and accessory genome as well as phylogenomics.

EDGAR calculates core genome via an iterative pairwise comparison of the set of genomes. A single genome must be chosen as the reference (A), with 'any' number allowed for comparison (B, C, D and so on). Every gene in genome A which does not have an orthologue in genome B is removed from the data set. The remaining set is then iteratively compared against genomes C, D and so on, until a remaining group of coding sequences is created with orthologues present in all genomes across the group. EDGAR identifies orthologues using Bi-directional Best Hits (BBHs) – to

identify genes with conserved function – via BLAST (blastp with blosum62 similarity matrix). Rather than using absolute BLAST scores, EDGAR relies on BLAST Score Ratio Values (SRVs) – comparing BLAST bit score to maximum bit score (maximum score resulting from aligning the gene against itself). To discriminate orthologous genes from nonspecific hits, a sliding LSW – lowest scoring window – is used to identify the SRV with the lowest hit count within the LSW: this filters out low quality BLAST hits. An SQL (Structured Query Language; a form of programming language for storing and interacting with relational databases) database stores the BBH BLAST information for the all-against-all comparison of the genomes.

The calculation of the pan genome is similar; every gene in genome B that is not in genome A is added to the dataset, and this runs iteratively across the group of genomes producing a total database of all coding sequences present in the group.

As with any genome comparison technique EDGAR has some potential for error, however it is comparable to other techniques and based upon methods which have been demonstrated to be amongst the most effective: the use of BBHs and BLAST SRVs have been demonstrated to be effective and appropriate measures (Altenhoff & Dessimoz 2009; Lerat, Daubin & Moran, 2003). Due to its use of BBHs, EDGAR can overlook instances where a target genome has multiple copies of an orthologous gene - the additional copies will be identified as separate matches, and requires some manual curation to assess these duplicated matches and condense these to a single CDS match in the core or pan genome. Additionally, the use of BLAST SRVs can be less effective in highly divergent genera such as *Corynebacterium* (Blom *et al*, 2009); however EDGAR remains a suitable method for investigation in organisms such as *Campylobacter*. Also, a very small degree of bias (<1; Blom *et al*, 2009), and the size of the resulting calculated core or pan genome, can occur depending on the reference genome selected; in the work conducted here the largest resulting outputs were used in each case.

EDGAR was used to produce predicted core and pan genome tables for the 33 included *C. jejuni* and *C. coli* genomes (Table 3.1; Appendix 9.1) which will be investigated for evidence of increased homology between the porcine ST403CC *C.*

jejuni isolates and the porcine *C. coli* isolates, as well as for any additional or lost *C. jejuni* associated genes in the ST403CC *C. jejuni* isolates.

4.1.9 Aims

The first goal of this chapter is to produce annotated genomes for the newly sequenced, potentially pig host adapted ST403CC *C. jejuni* isolates in order to study genome content. In this chapter, local database BLAST is used to assess the presence of the *HipO* gene, virulence genes and capsule genes in the ST403CC *C. jejuni* isolates. This chapter will also produce information on the core and pan genome of the group of included *C. jejuni* and *C. coli* genomes, which will be assessed to identify genes gained or lost in ST403CC *C. jejuni* isolates. Additional means including BLAST similarity searching will also be applied to CDS of interest which may be revealed by accessory genome analysis.

As has been briefly described here, various approaches have been taken to address the question of the core and pan genome of *C. jejuni* and to use these groups to inform on adaptation of isolates, however due to the nature of the question, and the reliance upon the genomes included in analysis different results for the size of the core and pan genome have been reported. Nonetheless investigating the pan and core genome of species and groups within species provide a useful tool to potentially identify host, environmental or pathogenic associated genes. This chapter attempts to utilise study of the core and pan genome in order to assess the genome content of the six newly sequenced ST403CC *C. jejuni* isolates and identify any genes gained or missing in this group compared to reference non-ST403CC *C. jejuni* and *C. coli* isolates.

This chapter attempts to investigate evidence for potential host adaptation in *C. jejuni* ST403CC isolates. As mentioned above, host adaptation can be characterised by gene gain or loss; consideration of the core and pan genome of the group of isolates is used to assess the presence of additional genes in ST403CC *C. jejuni* isolates, and investigate any genes which have been lost by the ST403CC isolates. Specifically, the major goal for this chapter was to assess evidence for loss of 'poultry associated' CDS and gain of potential 'porcine' or 'mammalian associated'

CDS. This chapter also sets out to consider whether the ST403CC *C. jejuni* isolates, being recovered from a host typically associated with *C. coli*, show any evidence for increased homology with *C. coli* isolates compared to that found in 'typical' *C. jejuni* isolates.

4.2 Methods

In order to assess the shared and variant coding genome content in ST403CC *C. jejuni* isolates compared to other *C. jejuni* and *C. coli* isolates, firstly annotated genomes were required and subsequently the establishment of the core and pan genome for included isolates was necessary.

4.2.1 Annotation

Genome sequences were produced as described in Chapter Three, and were subsequently annotated as follows. One *C. jejuni* ST403CC isolate (*C. jejuni* 857) and one *C. coli* isolate (*C. coli* 03/121) were selected for primary detailed annotation. *C. jejuni* 857 is the probable most 'ancestral' of the sequenced ST403CC *C. jejuni* isolates, as it is closest relative to non-ST403CC isolates by whole genome phylogeny (Chapter Three), whilst *C. coli* 03/121 was selected as it effectively sits in the 'middle ground' of the *C. coli* isolates, also due to the higher clonality in *C. coli* it was less of a concern to pick a 'representative' from the four isolates. Annotation was carried out using PROKKA (Seemann, 2014) for the two selected isolates, and subsequently transferred to the remaining isolates using RATT (Rapid Annotation Transfer Tool; Otto *et al*, 2011), a tool which takes a query and reference genome, and finds all matches for CDS on the reference genome in the query genome, applying the annotation for each. PROKKA annotates assembled genomes without relying on a single reference genome, rather it utilises a database of numerous reference genomes, and is therefore able to find most likely annotation for each CDS individually, and requires little subsequent manual completion.

4.2.2 BLAST Sequence Similarity Searching

Both web-based NCBI BLAST searches and BLAST searches based upon locally produced databases were used to investigate the probable function and prevalence of specific coding sequences.

4.2.2.1 Web-Based BLAST Sequence Similarity Searching

Web-based NCBI BLAST searches were run directly from Artemis sequence viewer (Rutherford *et al*, 2000) using the 'run' option. The primary method used was BLASTp, although BLASTn searches were also included. Results were obtained in the browser output, and the top three scoring matches were recorded for interpretation. Artemis viewer (Rutherford *et al*, 2000) was also used to investigate the neighbouring regions of CDS.

4.2.2.2 Local Database BLAST Sequence Similarity Searching

CDS of interest were queried against locally built genome databases. In order to achieve this, the CDS of interest was selected whilst viewing in Artemis (Rutherford *et al*, 2000) and extracted as FastA sequence which was then used to search for matches in the database. Custom nucleotide databases were created using a locally installed version of BLAST (Camacho *et al*, 2008; 2013), through creating a FastA file containing all of the required genomes; in this case 34 included genomes, the 14 *C. coli*, 13 *C. jejuni*, 6 ST403CC *C. jejuni*, and the additional historical ST403CC *C. jejuni* isolate ATCC33560. BLASTn was employed as the primary search method, where no matches were found tBLASTx was conducted subsequently.

Output from local database BLAST searching was obtained as a plain text file, exported to table format, providing details on which isolates recorded matches, the BLAST similarity score and the E value for each.

4.2.3 Hippurate Hydrolysis Gene *HipO*

The presence or absence of the *HipO* gene was queried against the 34 genome local database, using the *HipO* gene from *C. jejuni* NCTC11168 (*Cj0985c* coding sequence at loci c919903-918406, genome accession AL111168) as the query sequence.

Matches were compared to the score observed for the 'self-hit' – the BLAST score recorded for the *C. jejuni* 11168 *HipO* gene against the *C. jejuni* 11168 genome.

4.2.4 Capsule Production Genes

The presence of a capsule region was investigated using *C. jejuni* 11168 as the reference sequence. The capsule region was queried against the local database using seven individual coding sequences (*kpsC*, *kpsD*, *kpsE*, *kpsF*, *kpsM*, *kpsS*, *kpsT*). Once again the resulting scores were compared to the 'self-match' score for the reference sequence provided.

4.2.5 Virulence Associated Genes

The presence of virulence associated genes were also investigated using local database BLAST similarity searching. Virulence genes used were all taken from the genome sequence of *C. jejuni* 11168 with the exception of *virB11*, which was taken from the *C. jejuni* 81-176 pVir plasmid (Bacon *et al*, 2000: accession; chromosome CP000538,; plasmid pVir AF226280).

4.2.6 Investigating the Core & Pan Genome

The core and pan genome sizes and content were determined, and subsequently analysed to identify coding sequences of interest.

4.2.6.1 Determining the Core & Pan Genome

Following the annotation process, EDGAR (Blom *et al*, 2009) was used to assess the core and pan genome for the 33 included isolates (Chapter 3, Table 3.1; Appendix 9.1).

The pan genome was determined for the 33 included genomes, using each genome in turn as the reference, in order to establish the largest possible pan genome, being determined at 5329 CDS (reference genome *C. coli* 03/103). Core genome of 591 CDS (reference genome *C. jejuni* 81116).

4.2.6.2 Filtering the Pan Genome

The resulting pan genome outputs were then subjected to a 'post-processing pipeline' in order to identify CDS of potential interest for comparative analysis.

The resulting table of CDS was subsequently filtered to remove essential and species specific coding sequences, as illustrated in Figure 4.1. Firstly the core genes across the group were removed - those with orthologues present in 100% of the 33 included genomes; secondly, in order to focus on the ST403CC isolates as a group, all CDS were removed which were not either 100% present or 100% absent across the six ST403CC *C. jejuni* genomes; finally, in order to slim down further to potential genes of interest, the remaining 1220 CDS were further filtered by removing CDS present in 100% of the ST403CC isolates and equal to or greater than 80% of the reference genomes and those absent from 100% of ST403CC isolates and present in equal to or less than 20% of the reference genomes.

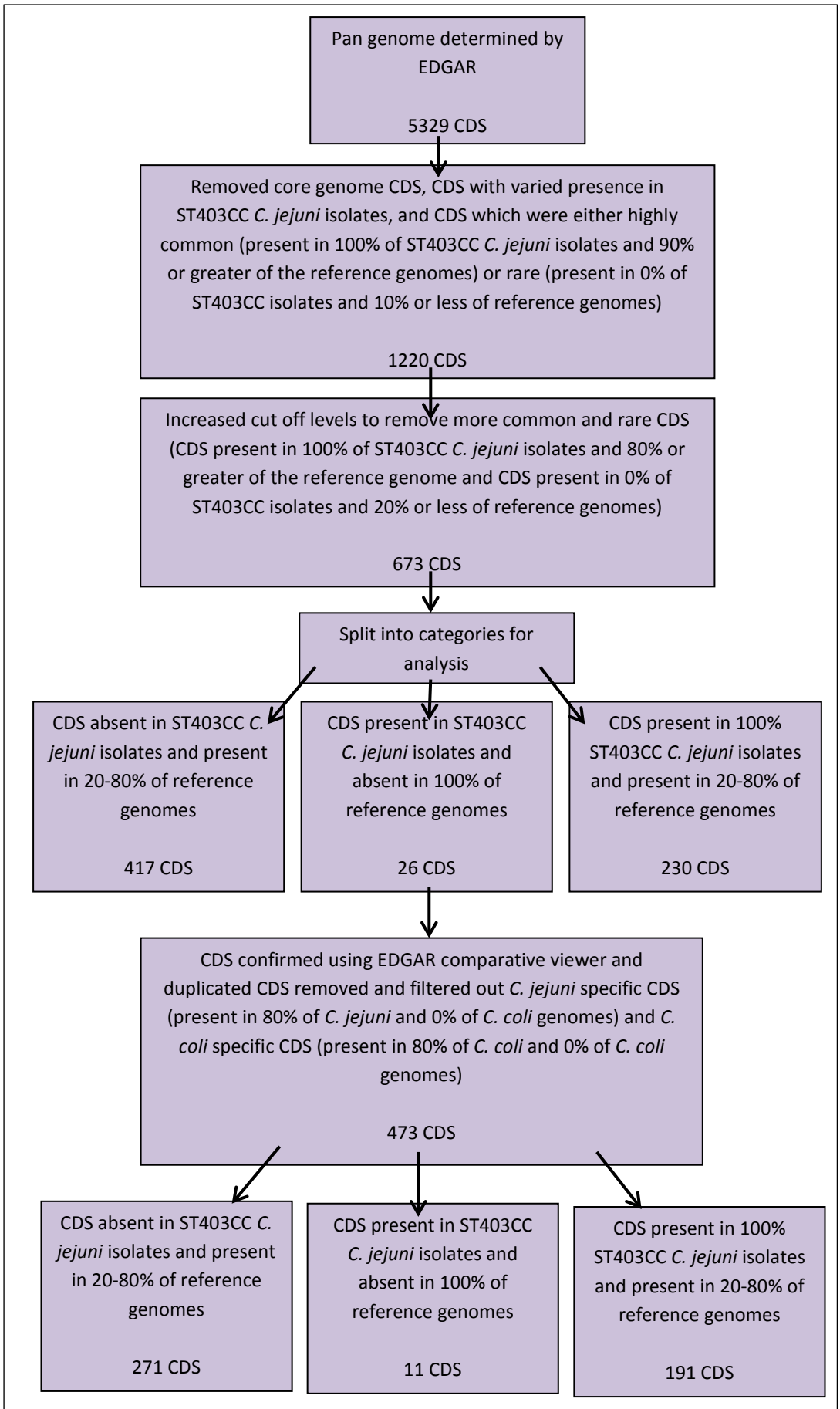

Figure 4.1: Flow Chart Illustrating the Filtering Process for CDS following EDGAR

4.3 Results

In order to assess the genome content of a group of potentially host adapted, closely related ST403CC *C. jejuni* isolates, annotated genomes were produced using genome sequences produced as described in Chapter Three. The ST403CC *C. jejuni* isolates had previously been demonstrated to be potentially capable of causing human illness (Chapter Two), and were shown to be closely related based upon MLST (Manning *et al*, 2003) and whole genome sequence based phylogeny (Chapter Three). Annotated genomes were produced, and the core and pan genome of the included isolates was determined and investigated.

4.3.1 Annotation

Annotated genomes were produced for the newly sequenced six *C. jejuni* ST403CC and four *C. coli* isolates. A single genome for *C. jejuni* (857) and *C. coli* (03/121) were annotated using PROKKA (Seemann, 2014) and transferred onto the remaining *C. jejuni* and *C. coli* sequences respectively, using RATT (Otto *et al*, 2011).

4.3.2 Presence of the *HipO* Gene

Figure 4.2 shows the percentage presence of the *HipO* gene for all *C. jejuni* isolates, all *C. coli* isolates, and for ST403CC *C. jejuni* isolates, as determined by local BLASTn query. As described previously, the ST403CC *C. jejuni* isolates were previously identified as phenotypically negative for hippurate hydrolysis (Manning *et al*, 2003), however, significant matches with *C. jejuni* 11168 *HipO* were observed in all included *C. jejuni* isolates, including the ST403CC *C. jejuni* isolates. No homology was observed in any of the *C. coli* isolates. The 'self-match' score for *C. jejuni* 11168 was 2767, with a mean across remaining the remaining 19 *C. jejuni* genomes of 2600 (with a range of 2025-2678), and for the ST403CC *C. jejuni* isolates the mean observed score was 2642 (with a range of 2628-2645), indicating high conservation of the gene across the included *C. jejuni* isolates, including the phenotypically negative ST403CC isolates.

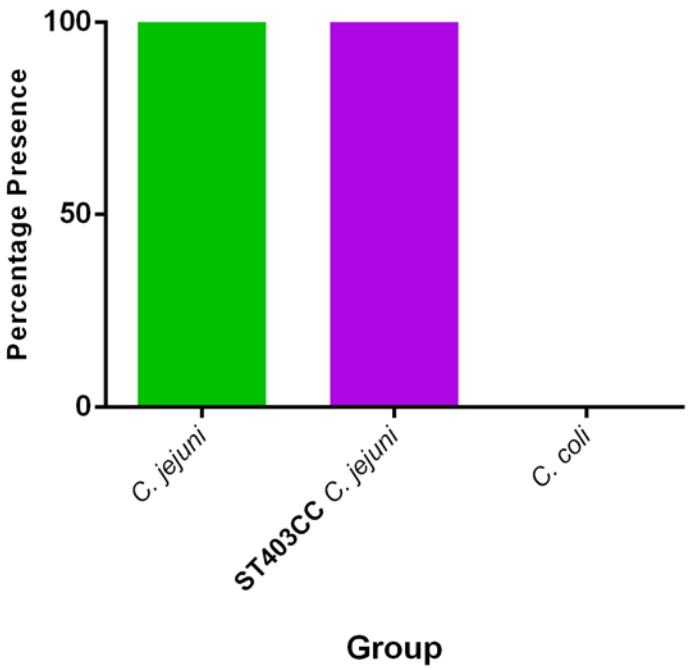

Figure 4.2: Presence of the *HipO* Gene in *C. jejuni* & *C. coli* Isolates

Green colour indicates all *C. jejuni* isolates combined (ST403CC and non-ST403CC). ST403CC *C. jejuni* isolates are indicated by purple colour, *C. coli* is indicated by blue bar, however none is visible due to the zero observed presence of the *HipO* gene in this group.

4.3.3 Capsule Gene Presence

Figure 4.3 combines the results concerning the prevalence of the capsule genes following local database BLAST searches. Individual percentages were as follows: *kpsC* matches were present in 100% of *C. jejuni* isolates and 14% of *C. coli* isolates; *kpsD* matches were present in 100% of included isolates; *kpsE* matches were observed in 100% of *C. jejuni* isolates and 86% of the *C. coli* isolates; *kpsF* had homologues in 100% of *C. jejuni* isolates, and 71% of *C. coli* isolates; *kpsM* was present in 70% of *C. jejuni* isolates, including 100% of the ST403CC *C. jejuni* isolates, and 7% of *C. coli* isolates; *kpsS* had significant matches in all of the included isolates, from both species; and *kpsT* had matches in 100% of *C. jejuni* isolates and 14% of *C. coli* isolates.

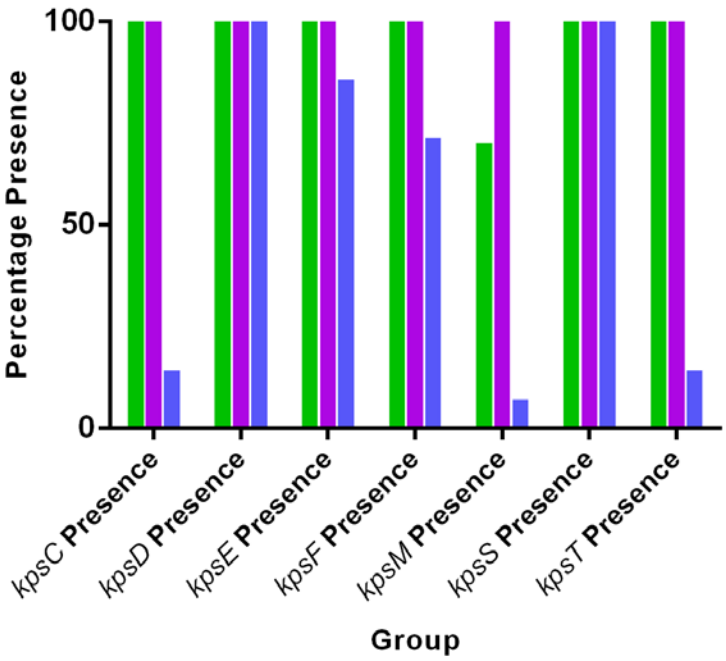

Figure 4.3: Presence of Capsule Associated Genes in *C. jejuni* & *C. coli* Isolates

Green colour is combined *C. jejuni* (ST403CC and non-ST403CC). ST403CC *C. jejuni* isolates are indicated by purple colour and *C. coli* isolates by blue.

4.3.4 Presence of Virulence Genes

Figure 4.4 provides an overview of the presence of virulence genes across the 34 genomes. Five of the eleven virulence associated genes (*cadF*; *ciaB*; *dnaJ*; *iamB* and *racR*) were found to have homologues in 100% of *C. jejuni* isolates and none of the *C. coli* isolates. Four more had no orthologues in *C. coli*, and were present in 100% of the ST403CC *C. jejuni* isolates but did not have complete prevalence across all *C. jejuni* genomes included (*cdtA* and *cdtB* were present in 80% of *C. jejuni* isolates, whilst *cdtC* homologues were found in 85% of the *C. jejuni* isolates and *pldA* had homologues in 90% of the included *C. jejuni* genomes). Finally, two of the investigated virulence associated genes did not have homologues in the ST403CC *C. jejuni* isolates, these CDS also had varied prevalence across the rest of the genomes; *flaA* was found in 30% of *C. jejuni* and 7% of *C. coli*, whilst *wlaN* was present in just 10% of the included *C. jejuni* isolates, and had no homology in *C. coli*.

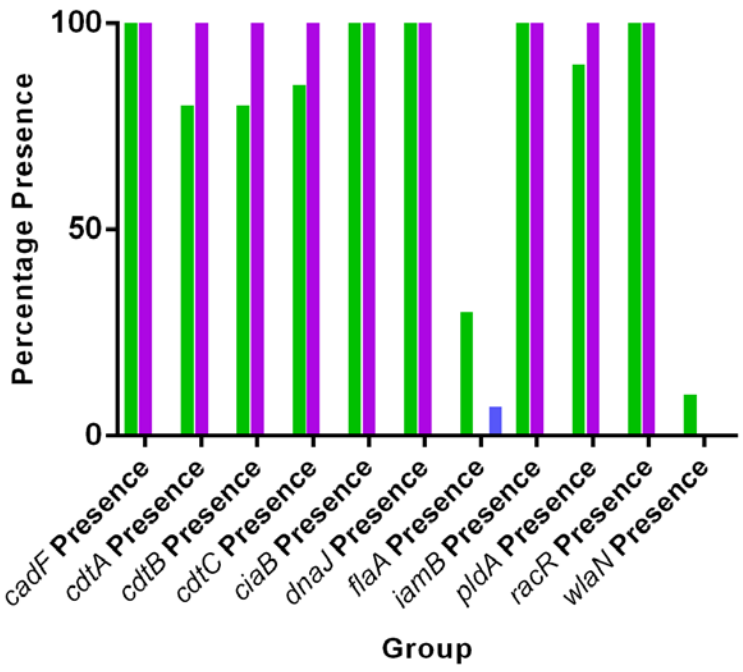

Figure 4.4: Presence of Virulence Associated Genes in *C. jejuni* & *C. coli* Isolates

Green colour indicates combined *C. jejuni* (ST403CC and non-ST403CC). ST403CC *C. jejuni* isolates are indicated by purple colour and *C. coli* isolates by blue.

4.3.5 Core & Pan Genome Analysis

The pan genome was determined for the 33 included genomes, using each genome in turn as the reference, in order to establish the largest possible pan genome, being determined at 5329 CDS (reference genome *C. coli* 03/103). The core genome for the 33 included isolates was 591 CDS (reference genome *C. jejuni* 81116).

The core and pan genome for the ST403CC *C. jejuni* isolates alone were 1444-1450 CDS (using *C. jejuni* 444 and 623 as reference sequences, respectively) and 2216-2220 (using *C. jejuni* 623 and 857 as references, respectively). The pan genome for all included *C. jejuni* isolates was 3887 CDS (with *C. jejuni* 1336 as the reference genome) and the core genome for *C. jejuni* isolates was found to contain between 883 CDS (based upon *C. jejuni* 1336) and 966 CDS (based upon *C. jejuni* 304).

4.3.5.1 Investigating Pan Genome Content

As described above, EDGAR (Blom *et al*, 2012) analysis was used to determine the pan genome for the 33 isolates, and the contents of the pan genome were investigated to identify potentially relevant CDS differentially found in ST403CC *C. jejuni* genomes compared to the reference group. This genome content analysis revealed a total of 379 CDS of interest, which were separated into three categories: CDS which were present in 20-80% of reference genomes and absent from 100% *C. jejuni* ST403CC isolates; CDS present in 20-80% of reference genomes and present in 100% *C. jejuni* ST403CC isolates; CDS absent from 100% of reference genomes and present in 100% *C. jejuni* ST403CC isolates - each of which contained a range of different potentially functional proteins, and the contents of which are discussed in greater detail below.

Figure 4.5 provides an overview of the remaining CDS following EDGAR analysis and subsequent filtering. The majority of the CDS remaining after the filtering steps were those present in a mixture of the reference genomes and absent from the ST403CC *C. jejuni* isolates, with fewer CDS present in all of the ST403CC isolates and a variety of the reference genomes, and a much smaller proportion found only in the ST403CC *C. jejuni* isolates and lacking from the rest of the included genomes.

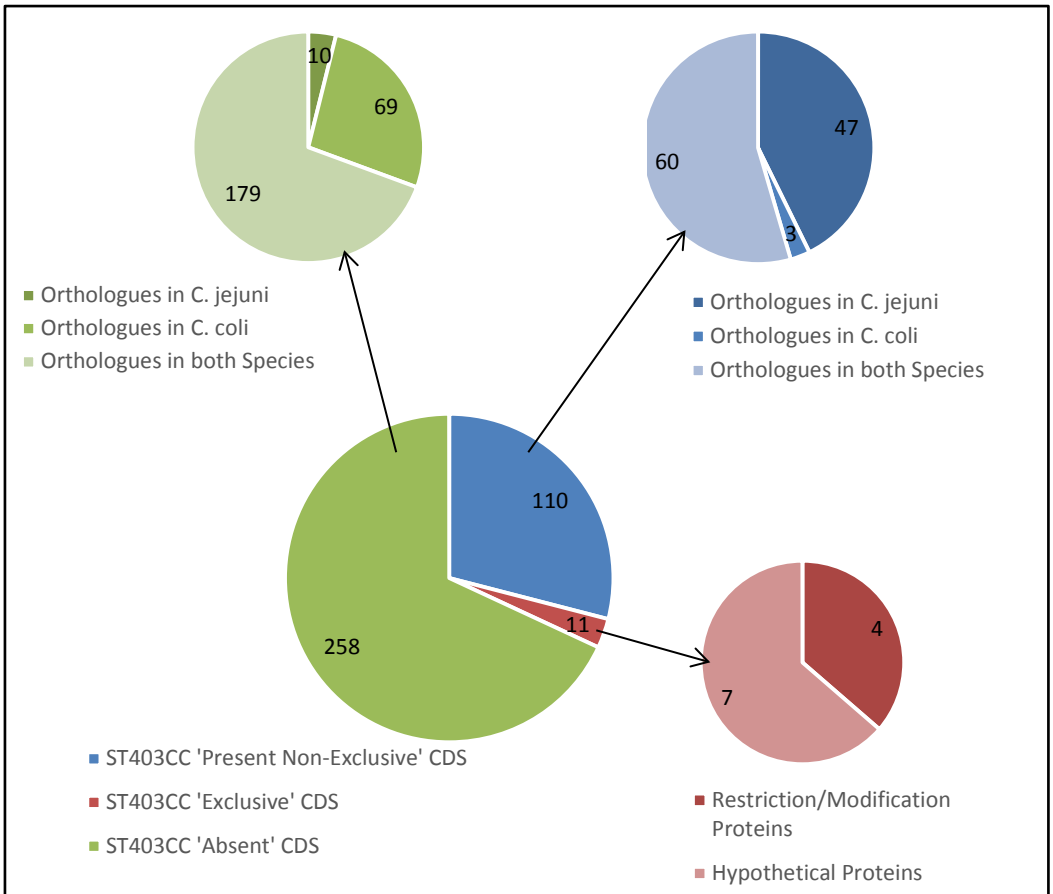

Figure 4.5: CDS of Interest Identified Following EDGAR Analysis

From the pan genome of 5329 CDS, after filtering steps, 379 CDS of interest remain, as represented in the pie charts: 258 CDS 'absent' from ST403CC *C. jejuni* isolates; 110 CDS present in ST403CC *C. jejuni* isolates and between 20 and 80% of reference genomes; and 11 CDS present 'exclusively' in the ST403CC *C. jejuni* isolates based upon the 33 included genomes.

The first of the three categories, referred to as the 'Absent' category, consists of 258 CDS which were found to have orthologues in 20-80% of the included reference genomes, but had no observed orthologues in the six ST403CC *C. jejuni* isolates. This was further sub-categorised by reference genome species; 69 of the CDS had orthologues in *C. coli* references only, ten had orthologues only in *C. jejuni*, and the remaining 179 CDS with no orthologues in ST403CC *C. jejuni* isolates had orthologues in a mixture of *C. jejuni* and *C. coli* reference genomes (Appendix 9.3.1).

The predicted functions of the CDS in the absent group are represented in Figure 4.6. The majority of the CDS which were present in 20-80% of the reference genomes and lacking homologues in the ST403CC *C. jejuni* isolates were hypothetical proteins for which function has not been determined, and the second largest proportion were also proteins of unknown function or other 'miscellaneous' proteins. However a number of functional proteins were found to be lacking or highly divergent in the ST403CC *C. jejuni* isolates, including cell membrane associated proteins, metabolism and biosynthesis associated proteins, and coding sequences associated with prophages, and the conjugal transfer of plasmids.

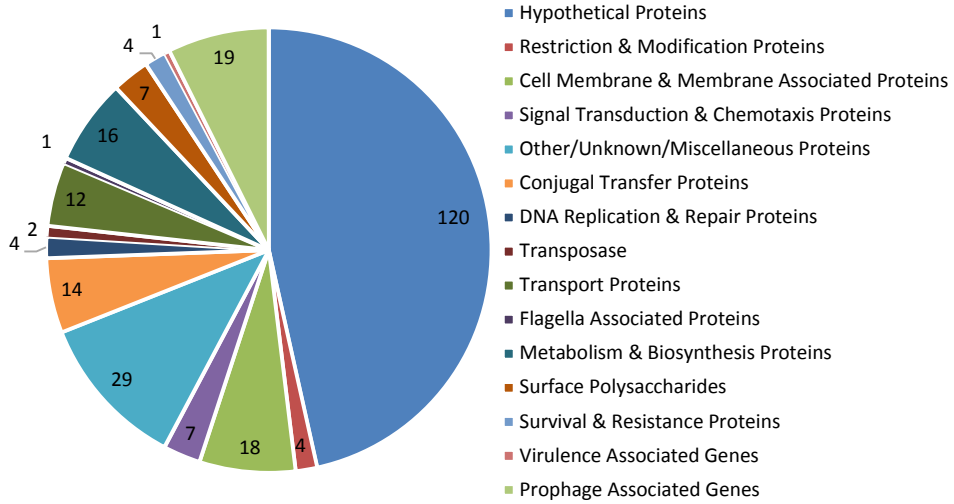

Figure 4.6: Distribution of Protein Function in CDS 'Absent' in ST403CC *C. jejuni* Isolates & with Homologues in Reference Genomes

CDS which were absent in 100% of the ST403CC *C. jejuni* isolates and with homologues present in between 20% and 80% of the reference genomes.

The functional assignment of the ST403CC 'absent' CDS are considered in relation to the three sub-categories; CDS in *C. jejuni* isolates only, CDS in *C. coli* isolates only, and CDS in a combination of *C. jejuni* and *C. coli* isolates, in Figure 4.7.

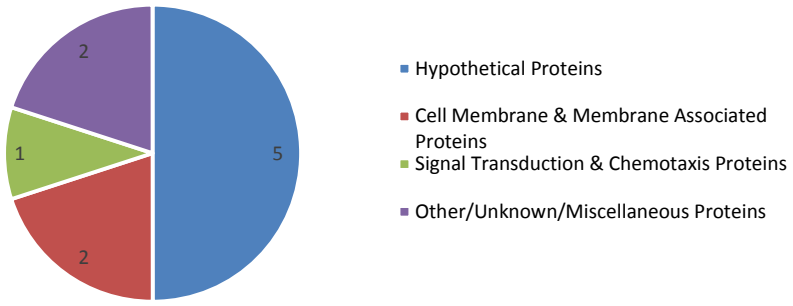

Figure 4.7a

Distribution of protein function in CDS absent in 100% of the ST403CC *C. jejuni* isolates and with homologues present in between 20% and 80% of *C. jejuni* only reference genomes.

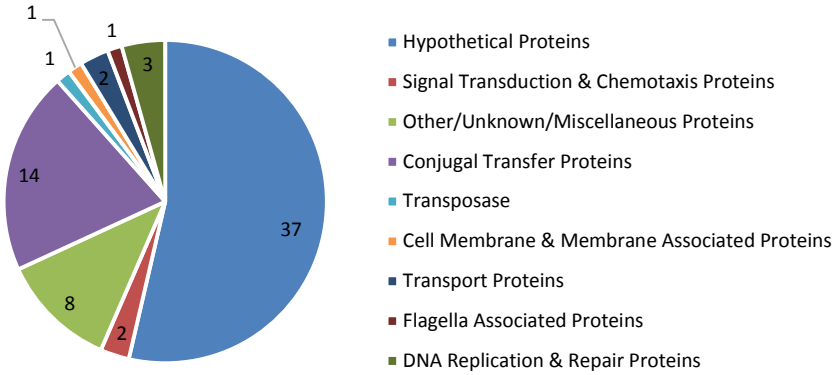

Figure 4.7b

Distribution of protein function in CDS absent in 100% of the ST403CC *C. jejuni* isolates and with homologues present in between 20% and 80% of *C. coli* only reference genomes.

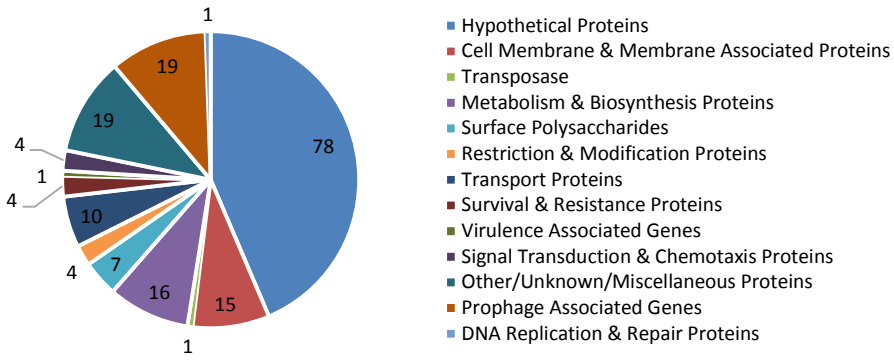

Figure 4.7c

Distribution of protein function in CDS absent in 100% of the ST403CC *C. jejuni* isolates and with homologues present in between 20% and 80% of a mixture of *C. jejuni* and *C. coli* reference genomes.

The majority of the *C. jejuni* associated CDS lacking homologues in the ST403CC *C. jejuni* isolates were hypothetical or miscellaneous proteins, however also observed were two cell membrane associated proteins, and one gene associated with chemotaxis signalling (Figure 4.7a).

A considerable proportion of the *C. coli* associated CDS lacking homologues in the ST403CC *C. jejuni* isolates were also proteins without known predicted function. There was considerable evidence of plasmid acquisition (conjugal transfer) content and DNA replication and repair proteins which were linked with *C. coli* genomes and lacking in ST403CC *C. jejuni* (Figure 4.7b).

Those CDS present in a mixture of the reference *C. jejuni* and *C. coli* genomes lacking homologues in ST403CC *C. jejuni* isolates were also dominated by proteins of unknown or hypothetical status, although also observed were genes associated with cell membrane, metabolism, biosynthesis, transport and prophage content (Figure 4.7c).

As stated above, the second major category resulting from EDGAR analysis was that containing CDS with orthologues present in 20-80% of reference genomes and present in 100% *C. jejuni* ST403CC isolates (n=110). This was separated into three sub-categories, similarly to the 'Absent' group: CDS with orthologues in *C. coli* only (n=3); CDS with orthologues in *C. jejuni* only (n=47); CDS with orthologues in a mixture of both species (n=60). A summary of the predicted function of these CDS is provided in Figure 4.8.

The functional content of the 'non-exclusive' ST403CC CDS are separated by category (*C. jejuni* associated; *C. coli* associated; shared with *C. jejuni* and *C. coli*) in Figure 4.9.

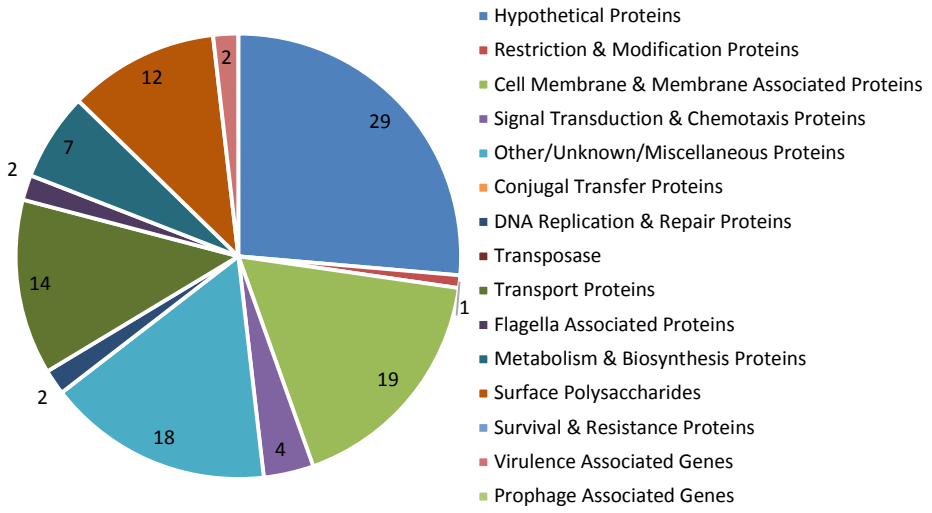

Figure 4.8

Distribution of protein function in CDS present in 100% of the ST403CC *C. jejuni* isolates and with homologues present in between 20% and 80% of the reference genomes.

As depicted in Figure 4.9a, the majority of CDS which were present in ST403CC *C. jejuni* isolates and with homologues in 20-80% of included *C. jejuni* isolates were proteins associated with the cell membrane. A considerable proportion were also associated with transport, although there were also the same number of hypothetical proteins. Additionally, five of the forty-seven CDS in this category were associated with metabolism and biosynthesis.

The three CDS shared between the ST403CC *C. jejuni* isolates and *C. coli* isolates only were all hypothetical proteins (Figure 4.9b).

A range of functions were predicted within the CDS which were shared between the ST403CC *C. jejuni* isolates and a combination of reference *C. jejuni* and *C. coli* genomes. The majority again were hypothetical or unknown proteins, however a considerable number (eleven of sixty) were predicted surface polysaccharides, and smaller proportions represented transport proteins and cell membrane associated CDS (Figure 4.9c).

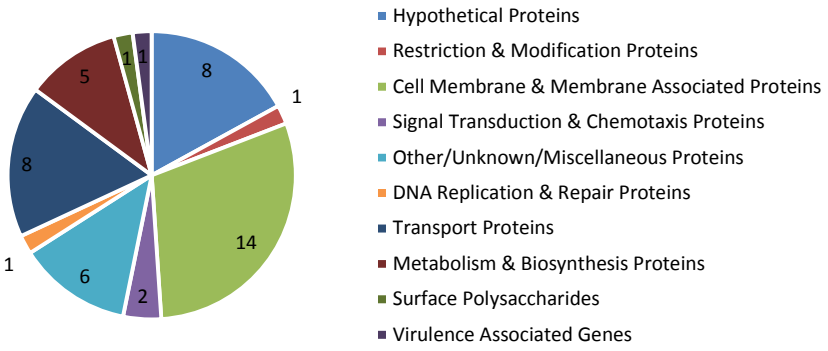

Figure 4.9a

Distribution of protein function in CDS present in 100% of the ST403CC *C. jejuni* isolates and with homologues present in between 20% and 80% of *C. jejuni* only reference genomes.

Figure 4.9b

Distribution of protein function in CDS present in 100% of the ST403CC *C. jejuni* isolates and with homologues present in between 20% and 80% of *C. coli* only reference genomes.

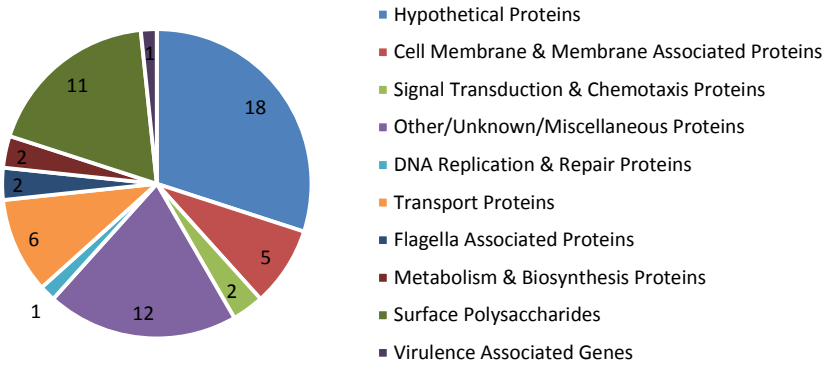

Figure 4.9c

Distribution of protein function in CDS present in 100% of the ST403CC *C. jejuni* isolates and with homologues present in between 20% and 80% of a mixture of *C. jejuni* and *C. coli* reference genomes.

The third group, referred to as the 'exclusive' group, were CDS which were present in 100% of the ST403CC *C. jejuni* isolates and with homologues not found across the reference *C. jejuni* and *C. coli* genomes. This group consisted of seven (64%) hypothetical proteins, and four (36%) restriction/modification system associated proteins (Figure 4.10).

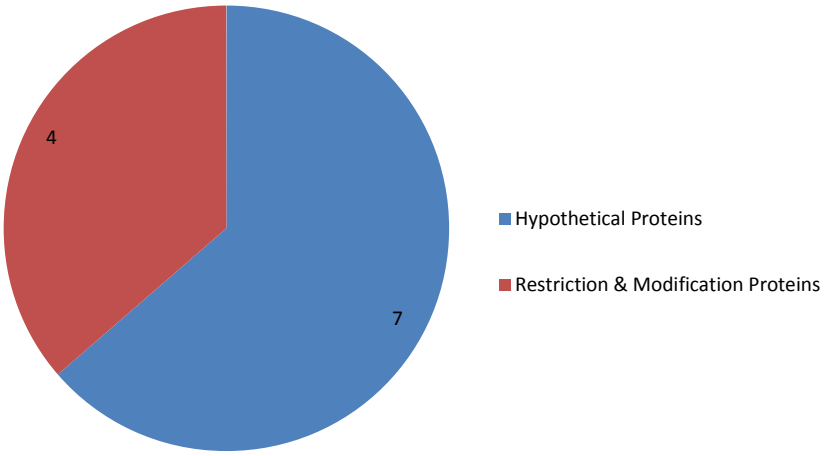

Figure 4.10: Protein Function in CDS Present in ST403CC *C. jejuni* & without Homologues in Reference Genomes

Protein distribution in CDS present in 100% of the ST403CC *C. jejuni* isolates and without homologues in the reference genomes.

Table 4.1 shows the breakdown of the numbers of CDS of interest by group and by species. The majority of the CDS present in reference genomes and not found in the ST403CC *C. jejuni* isolates had homologues in both *C. jejuni* and *C. coli* genomes, with a smaller proportion only having homologues observed in *C. coli* isolates and just 4% of the ST403CC 'absent' CDS found only in reference *C. jejuni* isolates. Of those CDS which were found to be present in all included ST403CC *C. jejuni* isolates and also in a number of the reference genomes, the majority again had homologues in both *C. jejuni* and *C. coli* isolates, although the margin was smaller in this instance, with slightly fewer CDS having orthologues only in *C. jejuni* and a much smaller number having orthologues only in *C. coli* (3%).

Category of CDS	Orthologues in *C. coli*	Orthologues in *C. jejuni*	Orthologues in both species	Total
ST403CC Absent	69	10	179	258
ST403CC Present Non-Exclusive	3	47	60	110
ST403CC 'Exclusive'	-	-	-	11

Table 4.1: Overview of CDS Identified following EDGAR Analysis

A total of 379 coding sequences of potential interest remained after filtering the content of the pan genome. These CDS were subdivided into three categories as shown above.

Further analysis of the ST403CC absent group identified some CDS which were present in the majority of the included *C. jejuni* isolates and lacked homologues in ST403CC *C. jejuni* isolates, including a putative integral membrane protein and a protein hydrolase, present in each of the included *C. jejuni* genomes with the exception of one of the human isolates; the *C. jejuni doylei* isolate 269.97.

The present non-exclusive group did not reveal any CDS with apparent links to host type; some CDS showed some bias towards being in human *C. jejuni* isolates, others with predominantly porcine origin in *C. coli* isolates, however this may be attributed simply to the bias induced by the high numbers of isolates from these sources.

4.3.6 ST403CC *C. jejuni* Associated Coding Sequences

As described above a small number of CDS were identified which were common to the six ST403CC *C. jejuni* isolates, but lacked homology across any of the other genomes included in the pan genome analysis. The presence of these genes which were 'exclusive' to the ST403CC *C. jejuni* isolates by EDGAR analysis were subjected to further consideration; both to confirm their 'exclusivity' in the locally constructed database, and to consider the potential implications of these CDS, including further investigation of their potential role and studying the neighbouring regions for additional information.

4.3.6.1 Predicting Function of ST403CC *C. jejuni* Associated CDS

Following the identification of ST403CC 'exclusive' CDS via EDGAR analysis, web-based NCBI BLAST was utilised to assess the true 'exclusivity' and prevalence of these CDS outside the included 33 genomes, and to further investigate their potential functions.

Table 4.2 gives details of the CDS which were found across the ST403CC *C. jejuni* isolates that did not record homologues in the reference genomes via EDGAR analysis. In order to attempt to determine potential function of the hypothetical proteins, and to assess the 'exclusivity' of these CDS across a larger platform, the eleven CDS were investigated using BLAST searching as described above.

BLAST querying confirmed the likely functions of the predicted proteins, however no additional functional predictions were made for the seven hypothetical proteins, with the exception that one of three matches for hypothetical protein CJ857_01361 was associated with a possible metabolic enzyme. Matches were observed for all eleven of the ST403CC 'exclusive' CDS as determined by EDGAR based analysis, demonstrating that homologues do exist in a larger database.

CDS	Predicted Function (PROKKA)	Blastp Hits		Percentage Coverage	Percentage Identity
		Description	Isolate		
CJ857_00074	Hypothetical protein	Hypothetical protein	*C. jejuni* ATCC33560	*100*	*99*
CJ857_00075	Hypothetical protein	Hypothetical protein	*C. jejuni* ATCC33560	*78*	*99*
CJ857_00839	Hypothetical protein	Hypothetical protein	*C. coli* K3	*100*	*95*
		Partial hypothetical protein	*C. jejuni* ATCC33560	*88*	*99*
CJ857_00896	R.HinP1 restriction endonuclease	Hypothetical protein	*C. jejuni* ATCC33560	*100*	*100*
		Hypothetical protein	*H. cinaedi* ATCC BAA-847	*96*	*74*
		Type II R-M system restriction endonuclease	*H. cinaedi* PAGU611	*95*	*74*
CJ857_00897	Modification methylase HhaI	DNA methyltransferase	*C. jejuni* ATCC33560	*100*	*99*
CJ857_01361	Hypothetical protein	Hypothetical protein	*C. jejuni* 2008-979	*100*	*100*
		2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase	*C. jejuni* CG8421	*100*	*99*
		Hypothetical protein	*C. jejuni* LMG 23263	*100*	*97*
CJ857_01649	Hypothetical protein	Hypothetical protein	*C. jejuni* LMG23223	*100*	*99*
			C. jejuni LMG23210	*100*	*99*
			C. jejuni 1997-10	*100*	*99*
			C. jejuni 2008-979	*100*	*99*
CJ857_01723	Hypothetical protein	Hypothetical protein	*C. coli* LMG23336	*100*	*100*
CJ857_01724	R.Pab1 restriction endonuclease	Hypothetical protein	*C. coli* H56	*100*	*100*
			C. coli LMG23336	*100*	*99*
CJ857_01734	Hypothetical protein	Hypothetical protein	*C. jejuni* 129-258	*100*	*100*
CJ857_01735	Recombination protein F	Putative ATPase	*C. jejuni* 129-258	*100*	*100*

Table 4.2: CDS Present in ST403CC *C. jejuni* Isolates without Homologues in Reference Genomes

This table describes the BLASTp NCBI search results for the eleven CDS which were only found amongst ST403CC *C. jejuni* isolates following EDGAR analysis.

4.3.6.2 Further Investigation of ST403CC *C. jejuni* Associated CDS

Following on from web-based BLAST searching, the ST403CC *C. jejuni* isolate 'exclusive' CDS as determined by EDGAR analysis were also further investigated using local database BLAST, and consideration of the surrounding coding regions. Local BLAST queries were used as described previously, using the sequences extracted from *C. jejuni* 857 for database queries against the 34 genome local database; including the 33 genomes included in pan genome analysis, and the additional ancestral ST403CC isolate *C. jejuni* ATCC33560, as well as an additional local database containing additional *C. coli* genomes. Local database BLAST searching confirmed presence in 100% of ST403CC isolates and zero of reference isolates for CJ857_00074, CJ857_00075, CJ857_00839, CJ857_00896, CJ857_00897, CJ857_01734 and CJ857_01735. CJ857_00839 did record some low scoring or high error matches in other isolates in the database (Appendix 9.4.1), however these were not considered to be significant matches. CJ857_01361 recorded several low scoring matches which were not significant, however in addition to the ST403CC *C. jejuni* isolates, an additional significant match was observed in *C. jejuni* 01/51. The three remaining EDGAR 'exclusive' CDS (CJ857_01649, CJ857_01723 and CJ857_01724) each showed significant matches in *C. coli* 03/317 in addition to the *C. jejuni* ST403CC isolates. Both CJ857_01723 and CJ857_01724 showed further significant matches when queried against the *C. coli* database in *C. coli* H56 and LMG23336 (Appendix 9.4.2).

Investigation using Artemis (Rutherford *et al*, 2000) was used in order to give further consideration to the size and position of the CDS, including the possibility of the CDS being pseudogenes; rather than being truly 'exclusive' CDS, these may represent *C. jejuni* genes which have become 'pseudogenised' – function and part of the coding sequence may be lost or degraded to the extent that it no longer appears related to the original CDS as seen in other isolates.

Figure 4.11 illustrates the region on the *C. jejuni* 857 genome containing CJ857_00074 and CJ857_00075. The first point to notice is the potential that CJ857_00074 and CJ857_00075 may have previously been a single coding sequence, indicated by the shortness of CJ857_00074, and the observable stop codons in all

three reading frames. In the *C. jejuni* 857 genome the pair of CDS are located in between *proB* and *fmt*, which as Figure 4.11b shows, were observed to be neighbouring CDS in *C. jejuni* 11168. In both genomes *proB* (Glutamate 5-kinase; Cj0097; 11168) is 756bp in length, so there was no evidence of this being fragmented; similarly *fmt* (methionyl-tRNA formyltransferase; Cj0098; 11168) has the same length (918bp) in 11168 and 857 genomes and therefore was not considered to be degraded. BLASTx searching of the region recorded low scoring partial matches in hypothetical proteins in the genome of *C. jejuni* ATCC33560.

Figure 4.11: Visualisation of ST403CC Associated CDS *CJ857_00074* & *CJ857_00075* in *C. jejuni* Isolate 857

a) Artemis viewer screen for *C. jejuni* 857 showing the region around CJ857_00074 and CJ857_00075.
b) Artemis viewer screen for the *C. jejuni* 11168 *proB* region.

It was deemed to be probable that CJ857_00074 and CJ857_00075 are pseudogenes. They do not appear to be products of the degradation of neighbouring genes *proB* or *fmt* however, and it remains unclear whether this is frameshift based, genuine stop codon insertion mutation, or sequencing error.

Figure 4.12, overleaf**,** shows the region of *C. jejuni* 857 containing the 'exclusive' CDS CJ857_000839, which is one of a number of consecutive short CDS (120-246bp) found upstream of *xerD_2* - an additional second copy of *xer_D* , which is downstream of *intA* in *C. jejuni* 857. *IntA* is a potential prophage intergrase, and may suggest that this region was acquired as part of a prophage**.**

Both CJ857_00896 and CJ857_00897 (hhaIM) can be seen in Figure 4.13**.** In *C. jejuni* 11168, *thiH* and *thiG* are direct neighbours on the chromosome. These two CDS may represent an insertion.

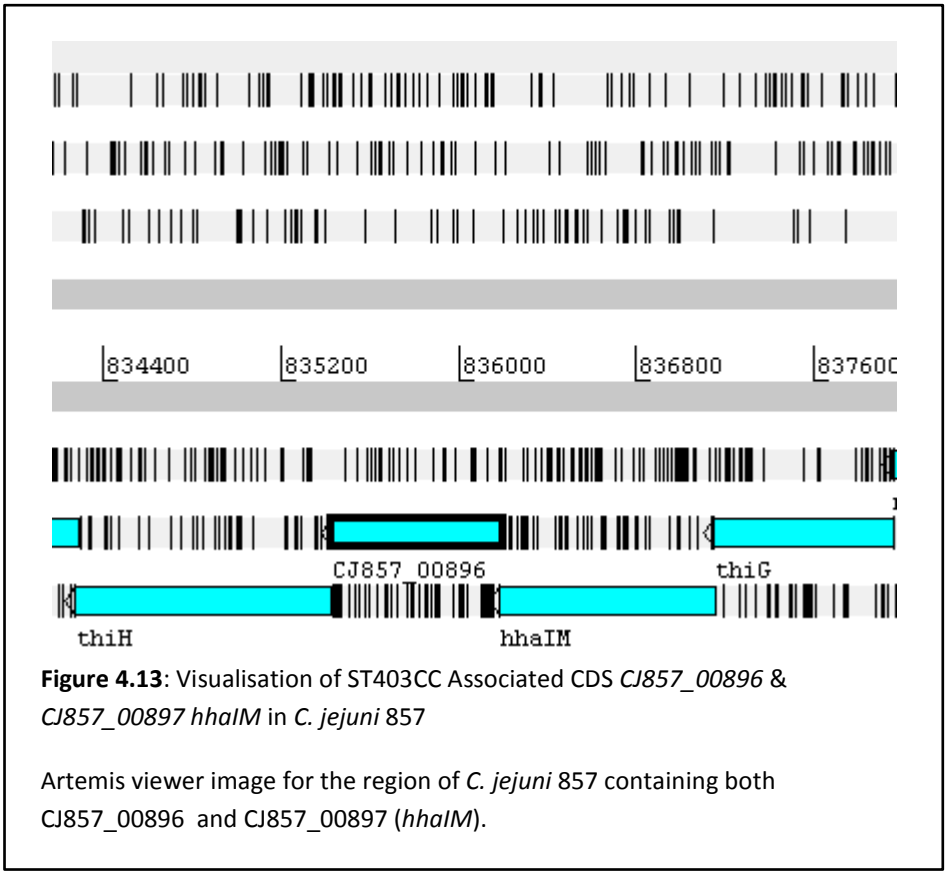

Figure 4.13: Visualisation of ST403CC Associated CDS *CJ857_00896* & *CJ857_00897 hhaIM* in *C. jejuni* 857

Artemis viewer image for the region of *C. jejuni* 857 containing both CJ857_00896 and CJ857_00897 (*hhaIM*).

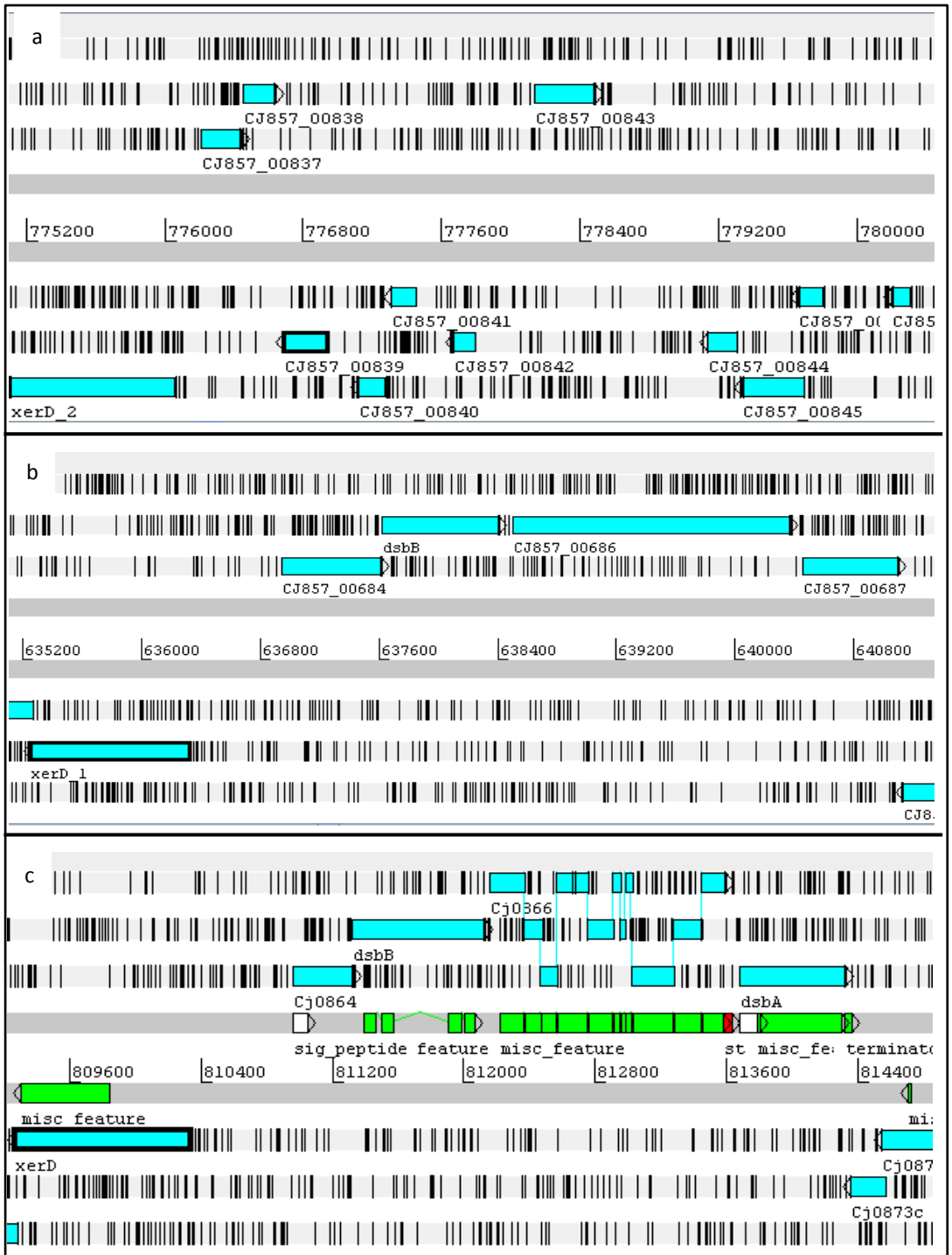

Figure 4.12: Visualisation of ST403CC Associated CDS *CJ857_00839* in *C. jejuni* 857

Artemis viewer output for the *C. jejuni* 857 CJ857_00839 region, including the second copy or *XerD*, plus output for the first copy of *XerD* in *C. jejuni* 857, and the related region in *C. jejuni* 11168.

a) CJ857_00839 region in *C. jejuni* 857
b) *C. jejuni* 857 region containing *xerD_1*
c) *xerD* region in *C. jejuni* 11168 genome

As shown in Figure 4.14, the ST403CC *C. jejuni* 'exclusive' CDS CJ857_01361 neighbours *purM* (Phosphoribosylformylglycinamidine cyclo-ligase) and *dcuD*_4 in *C. jejuni* 857. The *purM* sequence was the same size in the genomes of both *C. jejuni* 857 and 11168, and the region was similar upstream to in 11168 (*purM*, *coaE*, *dapF*), however the downstream area was different. In *C. jejuni* 857, *dcuD* copies 1,2 and 3 are neighbouring and in fact are probably one pseudogeneised copy of *dcuD*, however *dcuD_4* was located remote from these, it was also truncated

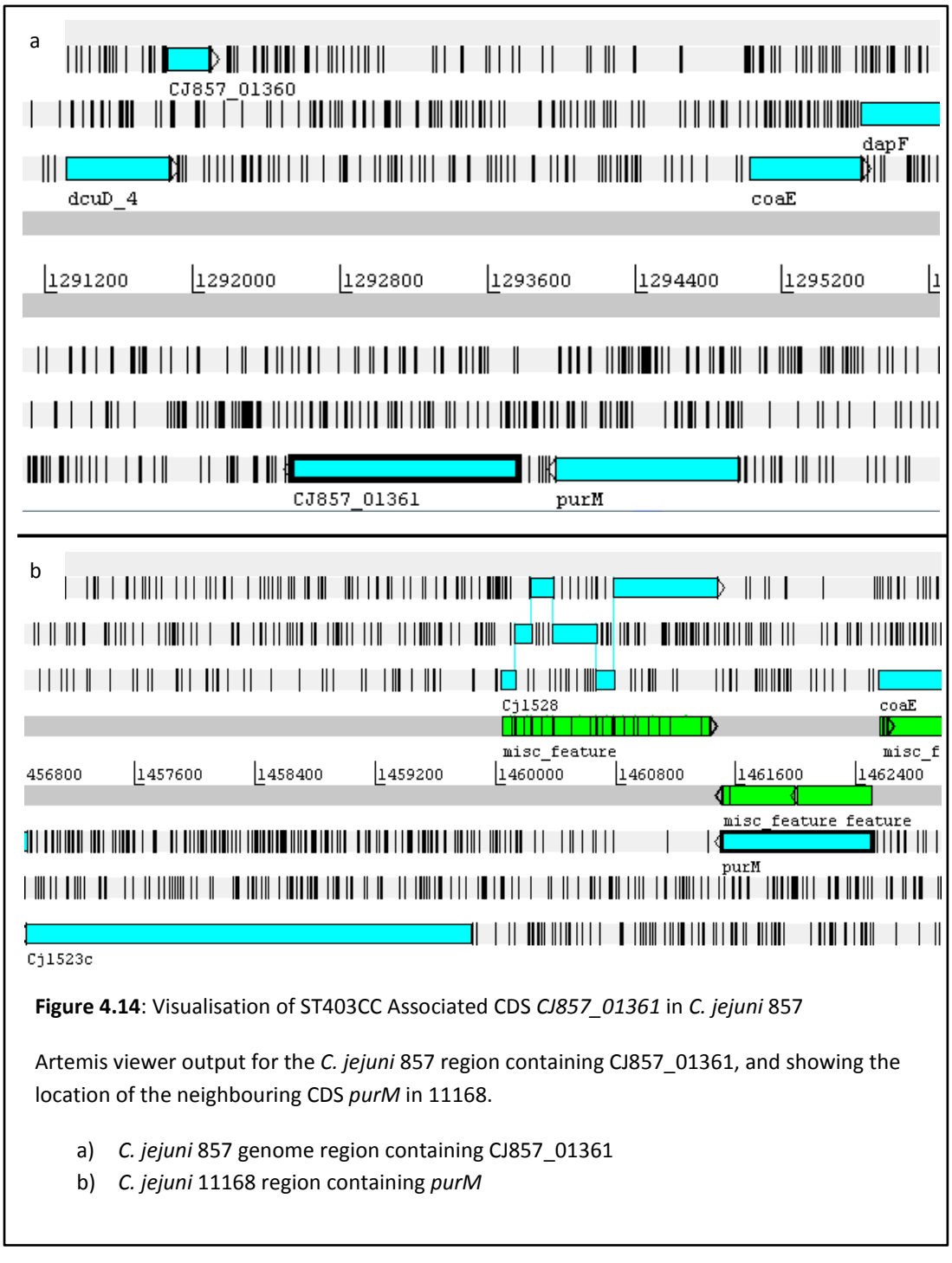

Figure 4.14: Visualisation of ST403CC Associated CDS *CJ857_01361* in *C. jejuni* 857

Artemis viewer output for the *C. jejuni* 857 region containing CJ857_01361, and showing the location of the neighbouring CDS *purM* in 11168.

a) *C. jejuni* 857 genome region containing CJ857_01361
b) *C. jejuni* 11168 region containing *purM*

(555bp) compared to the full *dcuD* (UniProtKB:P45428; 1368bp; Putative cryptic C4-dicarboxylate transporter) and is quite possibly pseudogeneised with CJ857_01360. Interestingly the CDS neighbouring *purM* in *C. jejuni* 11168 (*Cj1528*) was also described as a putative C4-dicarboxylate anaerobic carrier pseudogene - therefore it may be likely that all three (*dcuD_4*, *CJ857_01360*, *CJ857_01361*) are an equivalent pseudogene in ST403CC *C. jejuni*.

Figure 4.15 shows the region of *C. jejuni* 857 containing 'exclusive' CDS CJ857_01649. This region contained numerous hypothetical CDS, many of which were small and likely to be degraded fragments. The nearest downstream neighbours are *fcl* and *rfbC*; these two CDS are also close to each other in *C. jejuni* 11168 however the upstream region is the capsule region, which is not the case in *C. jejuni* 857. CJ857_01649 may be a 'true' hypothetical protein (861bp) however it is directly downstream of 01650 (1542bp) so it is possible the two were once one, or that they are jointly expressed.

Figure 4.15: Visualisation of ST403CC Associated CDS *CJ857_01649* in *C. jejuni* 857

Artemis viewer output for the *C. jejuni* 857 region containing the 'exclusive' CDS CJ857_01649.

Figure 4.16 shows the region of the *C. jejuni* 857 genome region containing CJ857_01723 and CJ857_01724. In *C. jejuni* 857 the 'exclusive' hypothetical protein CJ857_01723 had two neighbouring hypothetical proteins, and lies upstream of *carB* and downstream of *tal*. *C. jejuni* 11168 had a single hypothetical coding region

in between *carB* and *tal*. For both genomes, both *carB* and *tal* were found to be the same length. It appears possible that CJ857_01723 (hypothetical protein, 1068bp), CJ857_01724 (restriction endonuclease, 714bp), and CJ857_01725 (hypothetical protein, 423bp) may represent a pseudogenised gene as they are direct neighbours in different reading frames.

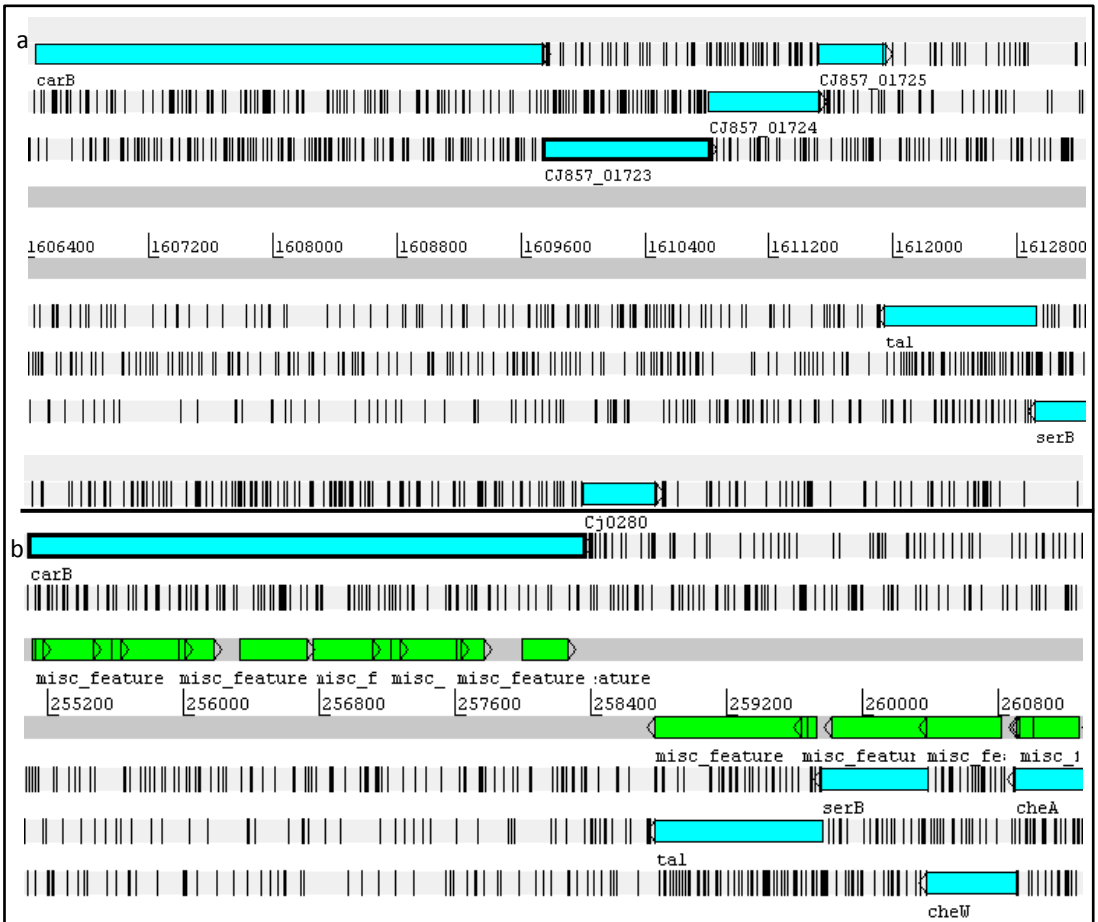

Figure 4.16: Visualisation of ST403CC Associated CDS *CJ857_01723* & *CJ857_01724* in *C. jejuni* 857

Artemis view showing the *C. jejuni* 857 CDS CJ857_01723 and CJ857_01724, neighbouring *carB*, and the *carB* region as found in *C. jejuni* 11168.

a) In *C. jejuni* 857 showing CJ857_01723 and CJ857_01724 with neighbour *carB.*
b) In *C. jejuni* 11168 the *carB* region similar but has only one hypothetical protein, in different reading frame

Figure 4.17 shows the Artemis view containing CJ857_01734 and CJ857_01735 and the downstream neighbours *lpxB* (1095bp), *greA* (486bp). Both of these CDS are present and neighbouring in 11168, with same lengths as observed in *C. jejuni* 857. Upstream of *lpxB* in 11168 is *peb3* (periplasmic protein), *glpT* (putative glycerol-3-

phosphate transporter) pseudogene. The *glpT* pseudogene is also present in *C. jejuni* 857, upstream near *surE*.

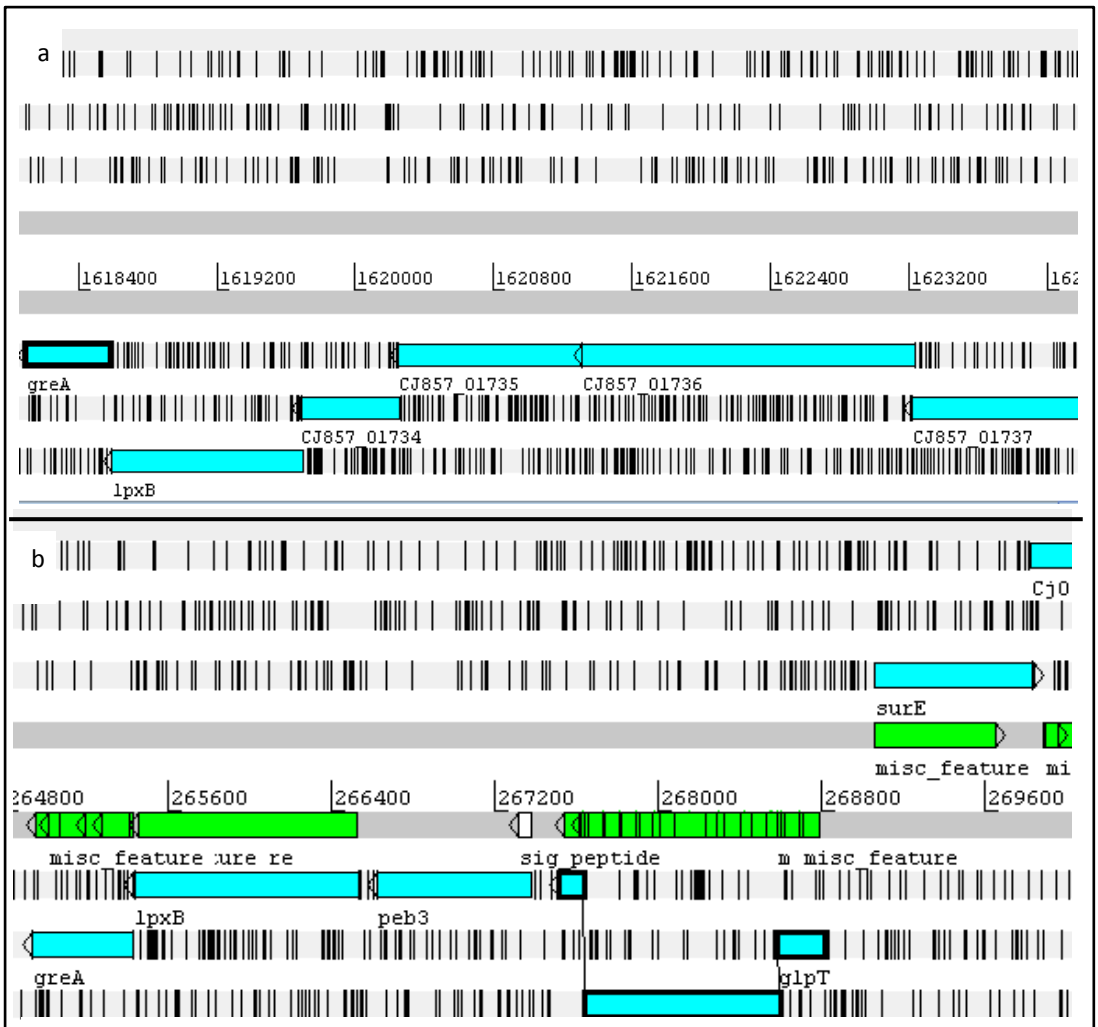

Figure 4.17: Visualisation of ST403CC Associated CDS *CJ857_01734* & *CJ857_01735* in *C. jejuni* 857

Artemis output for the region containing CJ857_01734 and CJ857_01735 in *C. jejuni* 857, and region relating to neighbouring CDS in *C. jejuni* 11168.

a) CJ857_01734 in ST403CC *C. jejuni* 857 upstream of *greA* and *lpxB*
b) Region in *C. jejuni* 11168 *greA* and *lpxB* next to *peb3* protein.

It was also noted (Figure 4.17) that the next neighbour gene in *C. jejuni* 11168 was *surE*. This is also upstream in 857, however more potential CDS exist in between, including CJ857_01736, which is not an ST403CC 'exclusive' CDS, but is a putative methyltransferase, and CJ857_01737 which is a type III restriction-modification system enzyme. CJ857_01734 may be a pseudogene or hypothetical protein; the function is not determined, and it marks the beginning of an apparent inserted region in *C. jejuni* 857 compared to *C. jejuni* 11168.

4.4 Discussion

The work undertaken in this chapter provided well annotated genomes for the newly sequenced *C. jejuni* ST403CC isolates, and used them to assess the core and pan genome of the selected group of isolates, in order to establish CDS potentially involved in host adaptation of the ST403CC group. In general terms, the apparent association of ST403CC *C. jejuni* isolates with the porcine host, and the uncommon lack of hippurate hydrolysis in phenotypic testing (Manning *et al*, 2003) had been thought to indicate potential host adaptation in this clonal group; it might be logical to predict therefore that evidence of host adaptation could be evidenced in part by the loss of 'poultry associated' CDS, the sharing of genes with porcine or mammal associated isolates, or increased sharing of CDS typically found in *C. coli*.

The results produced in this chapter have determined the presence of specific genes of interest in the ST403CC *C. jejuni* isolates, and provided further insight into the prevalence of ST403CC 'exclusive' CDS as elucidated from pan genome analysis. This section will consider the implications of the determined gene presence, whilst considering the potential roles of the variation observed.

4.4.1 Annotation

RATT (Otto *et al*, 2011) provides rapid annotation for newly sequenced genomes, however it relies on the presence and quality of a pre-existing reference genome annotation; if there is no currently annotated closely related genome, RATT cannot successfully be used, and RATT is limited by the quality of the existing annotation, and by the content of the annotated reference genome; any mistakes in the reference annotation will be propagated to the new sequences, and any additional coding content will not be annotated. RATT, therefore, is a useful tool for rapidly providing an 'approximate' annotation, but requires extra work to 'complete' the annotation.

PROKKA (Seemann, 2014), unlike RATT (Otto *et al*, 2011), has the benefit of not requiring a reference genome; instead PROKKA uses a range of tools to identify coding regions and subsequently query those regions against a database to determine most probable annotations. PROKKA has the benefit of not being limited

to a single reference genome, and therefore can produce improved accuracy of annotation, however due to the nature of automated annotation using databases post annotation 'sense-checking' may be required if an annotation seems unlikely. PROKKA also encounters the same potential pitfall as RATT in relying on the accuracy of existing annotations, however due to the inclusion of large protein databases this effect is minimised.

Both RATT (Otto *et al*, 2011) and PROKKA (Seemann, 2014) provide rapid and accurate annotations, which may not be as accurate as full manual annotation of a genome but are highly useful, and significantly less time and labour intensive.

As described by Richardson & Watson (2012) the rapid production of genome sequences has meant that manual annotation of genomes simply cannot keep pace with the production of new information, and automated annotation therefore has become hugely important. Various inconsistencies in annotation cause problems in automated annotation, including inconsistent annotation of homologous proteins, and even spelling mistakes in some annotations, as well as instances where the same gene names have been assigned to different sequences with different product (Richardson & Watson, 2012). Hypothetical proteins remain a significant burden in annotation databases, and a limitation of transferring and automating annotation is that these hypothetical proteins remain and propagate without function being determined (Richardson & Watson, 2012). As described by Richardson & Watson (2012) the most useful approach at current is to apply an automated method, and include manual curation to improve accuracy; this is the method that was applied here, wherein genes in regions of interest were manually investigated to attempt to confirm likely products.

4.4.2 Presence of the *HipO* Gene in ST403CC *Campylobacter jejuni*

As described, *HipO* presence was confirmed for all included *C. jejuni* isolates, including the six ST403CC *C. jejuni* isolates which were previously characterised as phenotypically negative for hippurate reduction. It can therefore be concluded that these are new examples of *HipO* positive, hippurate reduction negative (gene positive, phenotype negative) isolates, as have been described previously (Caner *et*

al, 2008). The HipO gene from *C. jejuni* 11168 was selected as the gene for comparison as this is one of the most well studied *C. jejuni* isolates, this was also deemed an appropriate reference due to the modular search methods employed by BLAST, as described above. The use of this *HipO* gene was confirmed to be suitable by the results obtained - as expected matches scoring high similarity with the 'self-match' score were observed across the *C. jejuni* isolates, with no significant matches observed across the *C. coli* isolates.

The hippuricase gene was initially sequenced from *C. jejuni* TGH9011 due to its strong positive hippurate hydrolysis reaction, and was determined to consist of 1152 nucleotides, coding for 383 amino acids (Hani & Chan, 1995). Brief analysis revealed that the *hipO* gene was of equal length in the original sequenced isolate (*C. jejuni* TGH9011), the well characterised hippurate hydrolysis positive isolates NCTC11168 and 81116, and the phenotypically hippuricase negative ST403CC *C. jejuni* isolates. There was no evidence of truncation or variation in the *hipO* coding sequence between known hippurate hydrolysis positive isolates and the hippurate hydrolysis negative ST403CC *C. jejuni* isolates. The gene appears to remain potentially functional in these isolates.

It is thought that hippurate negative *C. jejuni* isolates are relatively uncommon (Totten *et al*, 1987), so it remains an interesting characteristic to be shared across the group. Inaccurate phenotype testing can be an issue in determining the hippurate hydrolysis status of an isolate, although false negative results are less frequent than false positives (Amri *et al*, 2007). It could therefore be argued that false-negative results were observed for the tested ST403CC isolates, however this is unlikely as it was consistent across all isolates, and were carried out by a reputable lab with appropriate control measures (Manning *et al*, 2003). The results recorded for *hipO* and hippurate hydrolysis for ST403CC *C. jejuni* isolates also serves as additional evidence that for speciation, PCR based methods are more valuable than phenotypic testing in this instance, as is often seen in the literature (Denis *et al*, 1999; Wainø *et al*, 2003; Caner *et al*, 2008).

4.4.3 Capsule Synthesis Gene Presence in ST403CC *Campylobacter jejuni*

The capsule has been demonstrated to play a role in colonisation and invasion into host cells in *C. jejuni* 81-176 (Bacon *et al*, 2001); therefore it is interesting to consider capsule homology in ST403CC *C. jejuni* isolates. A previous study by Karlyshev *et al* (2000) first identified the seven capsule associated CDS in *C. jejuni*, including three (*kpsM*; *kpsS*; *kpsC*) which were conserved across all the included isolates (nine strains including 11168 and 81116). For the twenty *C. jejuni* isolates included in this study, homologues were observed for six of the seven capsule genes, whilst *kpsM* was present in only 70% of *C. jejuni* isolates, but was found in all ST403CC *C. jejuni* genomes. Considerably lower capsule gene prevalence was observed for the *C. coli* isolates; providing another example where the ST403CC *C. jejuni* isolates have much more in common with 'normal' *C. jejuni* isolates. Only structural genes were used in this research, further investigation would be required to identify the specific Penner serotype of the isolates; as cited previously, it was determined that both ST403CC *C. jejuni* isolates 484 and 444 were HS:23 serotype whilst the remaining four sequenced isolates were deemed non typeable (Manning *et al*, 2003; Frost *et al*, 1998). These results suggest that these ST403CC isolates are all potentially capable of producing capsule but cannot inform upon serotype.

4.4.4 Presence of Virulence Associated Genes in ST403CC *Campylobacter jejuni*

Monteville *et al* (2003) investigated the role of *cadF* (Campylobacter adhesion to fibronectin) in the process of invasion, stating that *cadF* has been repeatedly exhibited to be conserved among *C. jejuni* and *C. coli* (Konkel *et al*, 1999; Amri *et al*, 2007; Datta *et al*, 2003; Rozynek *et al*, 2005; Nayak, Stewart & Nawaz, 2005). It is surprising therefore to find two isolates which do not appear to have *cadF* homologues, however it may be that it is degraded in these porcine associated *C. coli* isolates, or that they are divergent enough that they did not score BLAST matches; invasion assays have not been completed on these isolates so it is also possible that they are non-pathogenic strains. It could also be a fault in sequencing or assembly in these two genomes.

cdtA,B,C were present in most (89%) of the studied *C. jejuni,* with no matches in the *C. coli* isolates included in the database. The two *C. jejuni* isolates lacking *cdtABC* were *C. jejuni doylei* 269.97 and *C. jejuni jejuni* IA3902. These two are potentially 'divergent' strains, although the lack in 269.97 is surprising as it is a human pathogenic isolate. This coincides with previous study be Rozynek *et al* (2005), who observed *cdtA,B,C* presence in almost all of their tested *C. jejuni* isolates, and in less than 6% of included *C. coli* isolates in work by. Also, Datta, Niwa & Itoh (2003) found 100% prevalence of *cdtABC* in their PCR study on 111 *C. jejuni* isolates from various sources.

The virulence associated genes *ciaB* and *dnaJ* were observed as being present in 100% of the included *C. jejuni* isolates and absent from all included *C. coli*. This correlates with research by Datta, Niwa & Itoh (2003) who observed high levels of prevalence of *ciaB* in *C. jejuni* with little variation by source (100% of poultry meat isolates, 100% of broiler faeces isolates, 98% of human clinical isolates, 85% of bovine faeces isolates) and for *dnaJ* (98% in human clinical samples, 91% poultry meat samples, 100% broiler faeces samples, 100% bovine faeces samples). Both *iamb* and *racR* were also found in all tested *C. jejuni* isolates and none of the included *C. coli* strains, whereas Datta, Niwa & Itoh (2003), found that *racR* was less prevalent than many of the virulence genes they studied (98% in human clinical isolates, 91% of poultry meat isolates, 86% of broiler faecal isolates, 77% of bovine faecal isolates). Due to the nature of short read sequencing as used in this project, and the subsequent necessity of reference based assembly, the issue of potential mis-assembly should be considered particularly in the case of a potential false-positive result for gene presence. Brás *et al* (1999) originally identified the RacR-RacS type two component regulatory system in *C. jejuni* 81116 and determined that mutation of *racR* reduced colonisation ability in the chicken model. The reference used for genome assembly of the ST403CC isolates was *C. jejuni* RM1221; which was found (as were all *C. jejuni* in this study) to be positive for *racR* by BLAST searching. The annotated genome of RM1221 (accession CP000025.1) contains identified *racR* - which is the same length and orientation as in 81116 (accession CP000814.1).

The possibility that the 100% coverage of *racR* observed in the included *C. jejuni* isolates was due to mis-assembly or sequence inaccuracy was considered however the presence of *racR* in ST403CC *C. jejuni* isolates is unlikely to be due to the sequence assembly against RM1221 as it is present at the same length not as a short CDS fragment or pseudogene giving a false positive result. Additionally, *racR* was observed in all included *C. jejuni* isolates which were independently assembled and annotated by various laboratory groups; therefore the likelihood of each of these positive results being false seems slim. Additionally, the source group with the lowest prevalence of *racR* in Datta, Niwa & Itoh's (2003) study was bovine faecal associated isolates, of which the only included isolate in this study was ATCC33560. This isolate was positive for *racR* according to BLAST searching; however this does not contradict the fact that other bovine isolates may be negative for *racR*. Finally, the average contig lengths for the ST403CC *C. jejuni* sequences before alignment against the *C. jejuni* RM1221 reference genome were in the thousands of base pairs so it unlikely that a short fragment was accidentally aligned with the *racR* CDS.

In Datta, Niwa & Itoh's study (2003) *flaA* was found to be present across 100% of *C. jejuni* isolates included in their study regardless of source, however here presence was determined for only 30% of *C. jejuni* and 7% of *C. coli* isolates, with no homologues observed in ST403CC *C. jejuni* isolates. Flagella are not essential for colonisation, however mutant studies have shown that mutants without *flaA* are impaired in colonisation compared to wild type strains (Wassenaar *et al*, 1993). Despite this, 'normal' levels of motility were previously observed for the ST403CC *C. jejuni* isolates (Chapter Two); it is probable that this apparent absence is in fact due to low homology resulting in BLAST similarity not being observed. The *flaA* gene is known to have highly variable regions, with highly conserved end regions (Chapter One; Harrington, Thomson-Carter & Carter, 1997; Meinersmann *et al*, 1997), it is possible that the variation was such that the conserved signature was not detected by this method.

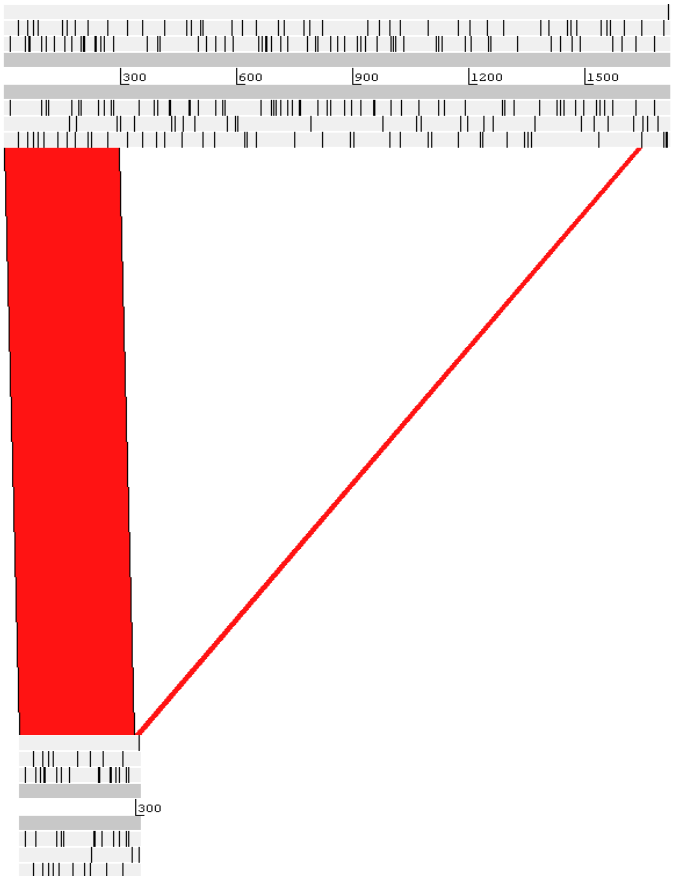

Figure 4.18: Artemis Comparison Tool View of *flaA* in *C. jejuni* 11168 & 857.

The Artemis Comparison Tool (ACT; Carver *et al*, 2005) output showing the *flaA* coding region in *C. jejuni* 11168 (top) against the *flaA* coding region in ST403CC *C. jejuni* isolate 857.

Using ACT (Artemis Comparison Tool; Carver *et al*, 2005), Figure 4.18 was produced visualising the differences in sequence between the *flaA* region in *C. jejuni* 11168 and ST403CC *C. jejuni* isolate 857. Figure 4.18 indicates that the ends of the *flaA* region in *C. jejuni* 11168 have high similarity with the *flaA* region in *C. jejuni* 857, however it is clear that the highly variable middle region appears to be missing from the ST403CC isolate - this may indicate that the *flaA* gene has been degraded in *C. jejuni* 857; however this may also potentially be due to mis-assembly during reference based assembly of the short-read genome sequence, as previously discussed regarding *racR*. Investigation of the *flaA* region in *C. jejuni* 857 using Artemis (Rutherford *et al*, 2000) revealed that the larger *flaA* CDS area expected appears pseudogenised due to stop codons in the area, leading to the tagged *flaA* coding sequence being shorter (1718bp in 11168, 312bp in 857) as shown in ACT Figure 4.18. As well as the expected neighbouring *flaB* coding region, *C. jejuni* 857

also contained a duplicate predicted *flaA* region in a different location on the chromosome, which is longer (2253bp) and may explain the motile phenotype observed in Chapter Two. As described by Meinersmann & Hiett (2000) *flaA* is found as part of a duplicate region with *flaB* which allows flagellation alongside antigenic phase variation .

Datta, Niwa & Itoh (2003) observed had high prevalence of *pldA* in *C. jejuni* isolates included (human clinical isolates 91%, poultry meat and broiler faeces isolates 100%, bovine faeces 85%), as was found here.

Finally, *wlaN* was observed in just 10% of included *C. jejuni* isolates, with no homology in ST403CC strains, and no homology in *C. coli* strains. This is consistent with literature, as *wlaN* was also one of the least common virulence genes as tested by Datta, Niwa & Itoh (2003) with the highest amount being in the human clinical isolates (25%), followed by poultry meat isolates (24%), then bovine faecal samples (8%) and broiler faeces (5%). Linton *et al* characterised *wlaN* in 2000, and observed it in three of the six isolates included in their study. It may be concluded that *wlaN* was absent from the database of isolates, with the exception of the 'self-hit' in *C. jejuni* 11168 and the high scoring match (1677) in *C. jejuni* 01/51; a known hyper invasive isolate (Javed *et al*, 2010), despite being absent from the other included hyper invasive isolate (*C. jejuni* 01/51).

4.4.5 Coding Sequences Variably Present in ST403CC and Non-ST403CC *Campylobacter* Genomes

The CDS which were determined to be missing from the *C. jejuni* ST403CC genomes may indicate loss of genes no longer required by ST403CC isolates which are typically needed by *C. jejuni*, or by *Campylobacters* associated with specific host types. It was observed that CDS lacking in ST403CC isolates frequently had homologues in both reference *C. jejuni* and *C. coli* genomes and represented a range of protein functions, particularly membrane associated proteins, transporters and potential metabolic proteins. This may suggest differential metabolic processes being used in the ST403CC *C. jejuni* isolates, however a clear pattern was not observed to support the idea that poultry host associated metabolic processes were

lost in the potentially porcine adapted ST403CC *C. jejuni* isolates; the ST403CC absent CDS were spread across reference genomes across the different sources and host types.

CDS present in ST403CC and also found in reference genomes may inform upon what remains common between these 'unusual' isolates and *C. jejuni* from various sources, they may also provide useful information on similarity between the ST403CC isolates and references from the porcine or mammalian host, and particularly demonstrate commonality between ST403CC *C. jejuni* isolates and *C. coli* genomes. The CDS shared between ST403CC *C. jejuni* and non-ST403CC *C. jejuni* did not exhibit bias towards host source, and were typically spread across a variety of diverse origins. Three CDS were observed which, by EDGAR analysis, were present in all ST403CC *C. jejuni* isolates, with homologues only in *C. coli*, two of which were in only 43% of included *C. coli* isolates, and were not restricted to a single source, and one present in 100% of *C. coli* genomes. The presence of these hypothetical proteins indicates some similarity between ST403CC *C. jejuni* and *C. coli* isolates; however it does not confirm increased sharing between ST403CC *C. jejuni* and *C. coli* as a result of surviving in the typically *C. coli* associated porcine host. The 'present non-exclusive' CDS group did not reveal any CDS with apparent links to host type; some CDS showed some bias towards being in human *C. jejuni* isolates, others with predominantly porcine origin in *C. coli* isolates, however this may be attributed simply to the bias induced by the high numbers of isolates from these sources.

The *C. jejuni* ST403CC 'exclusive' CDS; found in all of the ST403CC isolates but lacking homologues in the rest of the included genomes may inform of genes gained by ST403CC isolates which may potentially aid in, or be related to, host adaptation, or may serve as genetic markers for this group of isolates. BLAST searches carried out on the 'exclusive' CDS were unable to assign additional functional information on the hypothetical proteins, but confirmed the likelihood of the restriction modification genes. Homology for the various CDS was determined in eight *C. jejuni* strains, three *C. coli* strains, one *C. upsaliensis* and two *H. cinaedi* isolates, the majority of which had homology with only one or two of the CDS in question, and

ranging from a number of sources (human clinical isolates, including a GBS associated *C. upsaliensis* isolate and a *H. cinaedi* strain associated with bacteraemia, chicken meat and faeces, and bovine faeces), however one isolate - *C. jejuni* ATCC33560 - had significant homology with five of the eleven CDS, including two of the restriction/modification enzymes. Further investigation revealed that this isolate was in fact also an ST403CC isolate; a newly available sequence for a historical strain isolated from bovine faeces in Belgium prior to 1970, which was found to be closely related to the porcine source *C. jejuni* ST403CC isolates at the whole genome level, despite the differences in host, time and location.

These 'exclusive' CDS represent only a minimal proportion of the pan genome for the 33 isolates, and consisting largely of hypothetical proteins appear relatively uninformative, however the evaluation of the role of these CDS leads to two potentially important conclusions: the studied ST403CC *C. jejuni* isolates have each acquired a set of restriction modification associated enzymes, which do not have homologous genes present in included EDGAR genomes, and following BLAST analysis have few high scoring matches in the NCBI database; and the identification of a temporally and geographically distinct *C. jejuni* ST403CC isolate sharing some of these 'exclusive' CDS.

4.4.6 Coding Sequences Associated with ST403CC *Campylobacter jejuni*

Although it appears unexpected that additional homologues to the ST403CC *C. jejuni* 'exclusive' CDS were observed following local database BLAST queries, this can be explained by considering the methodology implemented by EDGAR; specifically, it iteratively judges each annotated CDS region, recording each CDS with significant homology to it. As stated previously, EDGAR relies upon the tagged, annotated CDS, whereas standard BLAST searching as applied in this chapter studies homology in raw sequence data, and is not affected by annotation as it uses raw fasta format sequence data. Therefore the additional homologues can be described as CDS not annotated in the genbank files used for EDGAR analysis. This is evidenced as the additional homologues were observed in genomes with transferred annotations; the annotation applied to *C. jejuni* 01/51 was transferred

from the annotation of *C. jejuni* 01/10, and the homologue to CJ857_01361 was not annotated as it is not present in *C. jejuni* 01/10. By the same principle, the annotation used to *C. coli* 03/317 was transferred from the annotated genome of *C. coli* 03/121 - again the CDS (*CJ857_01649; CJ857_01723; CJ857_01724*) which are present on *C. coli* 03/317 genome were not found in the annotated *C. coli* 03/121.

The more detailed analysis carried out here on the ST403CC 'exclusive' CDS following EDGAR analysis revealed that several of these are potential pseudogenes, lacking function, possible evidence of genomic degradation; for example it was determined that *CJ857_01361* may represent a pseudogenised dicarboxylate transporter. However, there are challenges in ascribing CDS or pseudogenes; given the numerous definitions and debate around bacterial pseudogenes it can be difficult to assign pseudogenes (Karro *et al*, 2007), this is particularly difficult for hypothetical proteins with unknown function - typically a gene has become a pseudogene if it is altered such that it is no longer a functional protein, however this can be harder to recognise when the function is unknown – particularly, for example, if two short hypothetical proteins are commonly found neighbouring to each other, do they represent two parts of a common pseudogene, or are they two separate unknown hypothetical proteins?

Analysis of the ST403CC 'exclusive' CDS and their neighbouring regions also revealed the possibility that the ST403CC isolates may have gained prophage DNA at some stage; as indicated by the presence of *intA* - a prophage integrase - in the region containing *CJ857_00839*, and also through the region surrounding the *CJ857_01734* and *CJ857_01735* CDS which contains a number of short potential coding regions. This information bears further examination, and the following chapter (Chapter Five) will attempt to uncover whether phage may play a role in what makes ST403CC *C. jejuni* isolates different from other *C. jejuni*.

The observation of these ST403CC *C. jejuni* 'exclusive' CDS and their possible roles as restriction endonucleases, modification methylase and recombination protein also warrants further consideration. The results of this and the previous chapter provide evidence that the ST403CC *C. jejuni* isolates are not undergoing increased

genetic exchange with *C. coli*, however the presence of these specific 'exclusive' CDS may provide a new explanation for this; *CJ857_00896, CJ857_00897, CJ857_01724* and *CJ857_01735* are homologues of genes associated with restriction/modification systems. R/M systems are known to play a role in reducing genomic inflow, specifically, they 'protect' the genome by preventing the integration of new sequence information (Wilson, 1991). This will be investigated subsequently (Chapter Six).

4.4.7 Summary & Overall Discussion

Gripp *et al* (2011) combined virulence studies and genome analysis and were unable to find genes related to specific hosts, or demonstrate fixed host adaptation. Therefore they concluded that some *C. jejuni* isolates, and specifically those within Sequence Type 21 exhibited the behaviour of a generalist rather than a specialist, that is, the variation in host isolates was thought to be linked to variations in expression rather than to the loss or gain of coding regions. It is possible that this phenomenon is also occurring in these ST403CC *C. jejuni* isolates; they may not have explicitly altered genome content in order to specialise to the porcine or mammalian host, rather they may be capable of generalist survival in a potentially non-optimal host.

The information provided by pan genome analysis has provided new information on the genes specific to, and shared by, the ST403CC *C. jejuni* isolates, however it must be acknowledged that the pan genome concept does have its limitations. It is important to consider the fact that, by its nature, the calculated pan genome is always dependent on the sample of genomes included; restricted by the number and origin of isolates included. As such, the work presented here is limited by the range of reference and target genomes included, however the selected genomes were included to represent range of hosts and sources, including well-characterised isolates, as well as some automated pipeline output sequences where these represented the best available sequences for the host type in question. It is important to acknowledge the potential limitations of the database used, however Méric *et al* (2014), in their work developing new approaches to the problem of the

pan genome found that a relatively small number of genomes were able to represent a significant portion of the total sample: They estimated a *C. jejuni* pan genome based upon 130 isolates at 3648 CDS, having identified 99% of this total using their initial comparison of just 75 *C. jejuni* genomes, similarly in *C. coli* the pan genome was estimated as 3520 CDS for 62 genomes, 99% of which was established by the comparison of just 40 genomes. It is therefore reasonable to expect that although it cannot provide a complete picture, the selection of genomes included can provide some useful insight into the characteristics of the query isolates. In order to address the limitations of the pan genome references included, BLASTp searching was implemented to consider the prevalence of the 'exlcusive' CDS in the public NCBI database. Significant scoring matches were found for each of the eleven CDS, however several of these were not in *Campylobacter*, and each only had a small number of matches. They were CDS with few homologues even across the much larger database so considering them 'exclusive' or rather 'associated' was a reasonable step. BLASTp search results for the eleven 'exclusive' CDS confirmed the likely function of the four restriction/modification/recombination proteins, however the method was unable to provide functional information for the seven hypothetical proteins.

It has also been suggested that the approach of viewing the genome in terms of accessory and core genes may be too simple to demonstrate the full picture of how genes are shared across a population. For example, for a sample of five genomes, Snipen, Almøy & Ussery (2009) observed a *C. jejuni* core genome of 847 and pan genome of 3221 CDS, whilst their mixture model based predictions estimated a likely 470 and 6587 CDS respectively. Using their multiple component model they found that most of their included species had an optimal model with 3 components, whilst the remainder had even more; this could indicate that the standard split of core and dispensable genes is limited as none of the mixture models were most effective with just two components, however it is suggested that this implied weakness may be due to skewed data: if several of the included individuals are 'related' in some way; such as pathogenic potential or epidemiology than this may create an extra sub-group indicating genes shared across this sub-population.

In addition to the caveats surrounding the pan genome concept, there are also some limitations to each method used to calculate the pan or core genome. In this instance, EDGAR is reliable in accurately predicting homologues and provides a useful picture of the pan or core genome, however it is limited by its requirement for annotated genomes meaning that un-annotated genomes cannot be included for study, and that the pan genome determined is also restricted by the completeness and accuracy of the annotations included, as it runs its iterative searching process based on tagged coding regions, and therefore does not consider additional non coding sequence.

In order to support the theory that the ST403CC isolates undergo increased genetic exchange with *C. coli,* it may be expected that a number of CDS would be found to be common between the ST403CC isolates and the *C. coli* reference genomes. It may also be predicted that the ST403CC isolates may share less information with the non-porcine *C. jejuni* isolates. However, the EDGAR based genome content analysis failed to identify significant sharing of homologous CDS between ST403CC isolates and *C. coli*, and few ST403CC absent CDS which are *C. jejuni* specific.

In the ST403CC *C. jejuni* genomes, there is not a clear pattern of the gain of mammalian host associated genes, or loss of poultry host associated genes. Additionally, the ST403CC *C. jejuni* isolates do not exhibit a large number of homologues shared with *C. coli* isolates.

In summary, this chapter has provided estimated core and pan genome information for thirty-three *C. jejuni* and *C. coli* isolates; this allowed investigation into CDS which were absent from, shared in, or exclusive to ST403CC *C. jejuni* isolates within the context of the included genomes. It was revealed that ST403CC *C. jejuni* isolates, despite sharing a host commonly associated with *C. coli* isolates, do not exhibit evidence of increased *C. coli* associated content; however the analysis did reveal a small number of ST403CC *C. jejuni* isolates which may play a role in recombination of this group of isolates, as well as raising the question of the role of potential prophage content in ST403CC *C. jejuni* isolates.

Chapter Five: Prevalence & Influence of Integrated Genomic Regions in the ST403CC *Campylobacter jejuni* Chromosome

5.1 Introduction

The uptake of new genomic content from external sources is a key event in evolution and adaptation of bacterial strains. Genetic acquisition occurs by a number of means, including horizontal gene transfer between bacterial isolates, uptake of phage DNA and integration of plasmids. The information transferred can range from short sequences and individual coding sequences, up to large regions such as genomic islands and in some cases can confer a fitness advantage to the bacteria. This chapter uses genome sequence based analysis to investigate the imported genetic content of the ST403CC *C. jejuni* isolates in order to investigate the role of 'foreign' DNA in these strains.

The identification of a possible phage integrase gene in ST403CC *C. jejuni* 857 in the neighbouring region of a possible ST403CC group associated CDS highlighted the potential role that phage acquisition may have played in separating the ST4303CC group from other *C. jejuni* genomes. This chapter therefore will focus on phage and other foreign DNA in the ST403CC *C. jejuni* isolates; beginning with assessing the prevalence of the integrase identified in *C. jejuni* 857, and considering the overall potential phage or other integrated genomic regions in the ST403CC *C. jejuni* genomes.

5.1.1 Mechanisms of Horizontal Genetic Transfer

As mentioned previously, the forms of integrated genetic elements vary in size, type and origin. There exist three major mechanisms for horizontal exchange of DNA in bacteria; conjugation, natural transformation and transduction. Conjugation is the direct, active exchange of DNA from a donor to a recipient cell. Conjugation is typically associated with plasmid DNA but also occurs with non-plasmid regions known as Integrating Conjugative Elements (ICEs). Transformation occurs in certain bacteria which are able to uptake and integrate external DNA from their surrounding environment. Transduction is the process by which non-viral DNA is

shared via viral particles (Burrus & Waldor, 2004; Dale & Park, 2004). Examples of transferrable elements include insertion sequences, transposons, genomic islands (GIs), bacteriophages and plasmids (Vernikos & Parkhill, 2006; Hacker & Kaper, 2000; Dale & Park, 2004). Insertion sequences (IS) are the most basic mobile genetic element; containing a transposase - for the movement of the element, and with inverted repeats at either end (one of the targets for identifying inserted regions). ISs are usually approximately 1300-1500bp in length but can be longer or shorter, and may be present as numerous copies on a genome; for example *E. coli* can hold six copies of *IS1* (768bp) as well as numerous copies of other IS elements (Dale & Park, 2004). Insertion sequences can have deleterious effects on genes but do not contain additional coding sequence, whereas transposons are essentially IS elements which also contain additional genetic content, such as metabolic or resistance genes (Dale & Park, 2004). Genomic islands (GIs) are large elements (typically 5-500kbp in length; Langille, Hsiao & Brinkman, 2008), which are transferred through HGT, and harbour a group of functionally linked genes, whilst pathogenicity islands (PAIs) are a GIs which confer virulence associated genes (Vernikos & Parkhill, 2006).

It is well established that pathogenicity factors can be carried and shared via mobile genetic elements, including bacteriophage (once internalised into a bacterium it may then be shared to other bacteria via HGT) , transposons, plasmids and more recently identified PAIs (Hacker & Kaper, 2000).

5.1.2 Bacteriophages & Prophages

Due to the identification of a potential phage gene in ST403CC C. jejuni isolate 857, consideration will be given to the prevalence of the intA gene specifically, and to establish the presence and influence of phage DNA in the ST403CC C. jejuni isolates. In order to consider the role of phage content, these must first be considered in further detail.

5.1.2.1 Bacteriophages

Viruses exist which infect all kinds of living cells; bacteriophages, also referred to as phages, are virus particles which infect bacterial cells (Dale & Park, 2004). Bacteriophages are similar to other viruses, being comprised of a protein coat encasing either DNA or RNA sequence, and replicating within a host cell (Dale & Park, 2004). Bacteriophages can be separated into two major groups – lytic phages and temperate phages (Zhou *et al*, 2011). Temperate phages may also be referred to as latent phages or lysogenic phages.

The main process of the viral life cycle involves lysis of the host cell. The virus particle first attaches to and enters the cell, then following entry, the production of 'early' genes begins, utilising the host's enzymes; next the viral nucleic acids are copied repeatedly at maximal speed and eventually the 'late' genes are expressed – those involved in producing new phage particles, before finally the progeny phages are built and released through the lysis of the host cell (Birge, 1994; Dale & Park, 2004). This typical progression of viral infection can also be described as the lytic cycle, and is the most common outcome of bacteriophage infection, however the temperate phages are able to enter a lysogenic phase within the host cell, at which time the viral DNA is preserved as a prophage – the viral nucleotide sequence is integrated into the DNA sequence of the host cell and is replicated at the same rate as the host cell (rather than at the unrestricted rate seen during lytic growth) and is passed onto each bacterial daughter cell as part of the bacterial chromosome, thus being maintained within the bacterial strain (Birge, 1994; Campbell, in Neidhardt *et al* (Eds), 1996; Dale & Park, 2004).

5.1.2.2 Prophages: Integrated Bacteriophages

A phage which has been integrated into the host genome is referred to as a prophage, whilst a prophage which has been inactivated and is no longer capable of entering the lytic infectious cycle is known as a cryptic prophage (Birge, 1994; Campbell, in Neidhardt *et al* (Eds), 1996; Casjens, 2003). Once a temperate phage has become a prophage within a host cell, the prophage is able to preserve its own presence within the host DNA by producing a protein repressor in order to prevent

the transcription of the phage 'late' genes, including the structural proteins. When the lysogenic phase has been entered, the lytic cycle does not resume unless triggered by some outside factor (Birge, 1994; Campbell, in Neidhardt *et al* (Eds), 1996), as a result prophages can persist within the bacterial genome for significant periods of time. Whilst they are behaving as part of the host sequence however, prophage genes can be lost in the same way as any other coding sequence – they are subject to the same mutation, degradation and recombination pressures as any other part of the host DNA (Birge, 1994). Cryptic prophages can still play a significant role in the host bacteria as they may still harbour functional bacterial genes (Zhou *et al*, 2011).

5.1.3 Identifying Prophages & other HGT Elements

Traditionally, identifying prophages relied upon inducing lysis by damaging the bacterial host cell; however this was experimentally demanding, and limited to prophages which remain fully intact and functional, also not all prophages can be induced by the same conditions (Zhou *et al*, 2011, Campbell, in Neidhardt *et al* (Eds), 1996). With the advance of sequencing techniques, genomic methods were developed to circumvent the restrictions of laboratory identification of prophages; typically these methods involved the examination of known potential attachment sites, or the identification of atypical nucleotide content which may represent prophage DNA, however each of these approaches also suffer weaknesses, due to the fact that prophages do not always use consistent integration sites, and may not always indicate abnormal sequence. In order to investigate possible prophage content without encountering these difficulties, this project makes use of PHAST (PHAge Search Tool; Zhou *et al*, 2011), which attempts to address these previous limitations and provides rapid results at a degree of accuracy comparable or superior to other current methods (Zhou *et al*, 2011).

As described previously however, prophages are just one means by which 'foreign' DNA may be internalised into the bacterial genome. Alienhunter (Vernikos & Parkhill, 2006) uses IVOM (Interpolated Variable Order Motifs) methods to investigate integrated DNA, with particular focus on genomic islands. Many genes

acquired through horizontal transfer HGT have negative effects and will cause bacteria to be lost from the population, some can appear neutral, conferring no advantage or disadvantage therefore not coming under selective pressure, the persistence of these depends largely on luck. Finally there are those HGT events that confer an advantage to the recipient isolate, or those which are able to elicit their own maintenance and propagation – these can be shared quickly across a population (Thomas & Nielsen, 2005).

Transferrable elements, particularly GIs, often have clear motifs, such as direct repeats or inverted repeats, and various methods exist to identify these regions and predict 'foreign' regions of genomic content. Vernikos & Parkhill (2006) developed a modern approach which does not reply upon high quality annotation already being completed for a genome, and utilises new techniques to improve the prediction of 'alien' sequence.

5.1.4 Bacteriophages in *Campylobacter*

Having described briefly the general characteristics of integrated genomic DNA it is now important to consider integrated DNA in *Campylobacter* as described in the literature. The majority of known *Campylobacter* bacteriophages are lytic phages (Sails *et al*, 1998), however it has long been established that *Campylobacter* may also harbour lysogenic phage. In 1968 Firehammer & Border demonstrated an inducible lysogenic prophage from a cattle origin '*Vibrio fetus*' (now classified as *Campylobacter fetus*, as described in Clark & Ng, 2008).

Campylobacters regularly encounter lytic bacteriophages during colonisation of the avian intestinal system. Scott *et al* (2007) investigated phage survival and the response to phage 'attack' and demonstrated that *C. jejuni* isolates which colonised chickens represented a distinct phage resistant type which undergo genomic rearrangements in response to phage. These isolates undergo intra-genomic inversions in sections of sequence between integrated Mu-like phage sequences within their chromosomes, causing the inversion of large segments of the genome. These modified isolates then exhibit a clear phenotype; they are resistant to phage infection, are poor colonisers of the chicken gut, and produce infectious virions of

the phage CampMu. Scott *et al* (2007) also demonstrated that this was a reversible event, with the original phage sensitive type being restored by passage through the chicken gut.

Bacteriophages have also been considered as potential therapeutic agents for *Campylobacter* infection in the face of rising antimicrobial resistance (Connerton, Timms & Connerton, 2011) and form the basis of a typing system (Sails *et al*, 2008).

The presence of temperate phage in *Campylobacter* was not well established until the availability of genome sequence data in recent years revealed that prophages were present in some but not all strains (Fouts *et al*, 2005; Connerton, Timms & Connerton, 2011). The *Myoviridae* are the most common group of bacteriophages found in *Campylobacter,* these can be split into four groups; three groups of lytic virions (described by Sails *et al*, 1998 in typing scheme) and the latent group (Connerton, Timms & Connerton, 2011).

In recent years various studies have investigated the presence and content of integrated elements in *Campylobacter* genomes, including notable works by Fouts *et al* (2005), Parker *et al* (2006) and Clark & Ng (2008).

Fouts *et al* (2005) investigated the genome content of four *Campylobacter* isolates from different species (*C. coli* strain RM2228; *C. jejuni* RM1221; *C. lari* strain RM2100 and *C. upsaliensis* strain RM3195) with a focus on identifying the major sequence differences between them. Amongst their discoveries was the recognition of four integrated elements within the genome of *C. jejuni* RM1221, termed CJIE (*Campylobacter jejuni* Integrated Elements) 1-4, the content of which was described in detail (Fouts *et al*, 2005; Parker *et al*, 2006). CJIE1 was found to be a Mu-like prophage region, designated CMLP1 (*Camyplobacter* Mu-like phage one). CJIE2 may represent either a cryptic prophage, or a non prophage genomic island, as it did not display any recognised capsid, scaffold or portal phage proteins. CJIE3 likely represents a genomic island or possibly an integrated plasmid region as it does not contain any prophage proteins. CJIE4 contained potential phage related proteins including methylases, endonucleases and repressor proteins, however the

region lacked most prophage structural proteins. CJIE4 also showed some homology with *C. lari* RM2100 prophage element.

Parker *et al* (2006) followed on from the work conducted by Fouts *et al* in 2005, investigating the prevalence of the *C. jejuni* RM1221 CJIE regions across a group of 67 *C. jejuni* and 12 *C. coli* isolates using PCR, as well as a DNA microarray based upon both *C. jejuni* 11168 and RM1221 sequences. Parker *et al* (2006) demonstrated that, of their study group, 27% of *C. jejuni* isolates contained two or more RM1221 CJIEs and 55% of *C. jejuni* isolates contained at least one RM1221 CJIE-like region. Additionally, 58% of the included *C. coli* were positive (by PCR) for one of the four RM1221 CJIEs – always either CJIE1 or CJIE3. Additionally they showed that in all *C. jejuni* isolates containing CJIE/CMLP1, it occupied a different chromosomal location; demonstrating that this phage (and potentially others) can be located at varying sites, adding to the potential for resulting variation.

Clark & Ng (2008) used southern blotting techniques to investigate the prevalence of the seven individual prophage genes present in *C. jejuni* RM1221 CMLP1; demonstrating that within their population of 35 *C. jejuni* isolates 77% with one or more gene, and 43% with five or more, leading them to conclude that CMLP1 (CJIE1) represents an individual example within a family of phages which exist within *C. jejuni*.

Barton *et al* (2007) conducted a study based upon the Mu-like phage described by Fouts *et al* (2005) investigating via PFGE the content of *C. jejuni* isolates from a water-borne outbreak, and demonstrated that the difference in PFGE patterns for four water-borne outbreak isolates was caused by the differential presence of phages and showed that prophages may have a role in adaptive *C. jejuni* biology in addition to natural transformation.

5.1.5 Aims

Integrated genomic DNA from bacterial and viral sources has been demonstrated to play a role in adaptation and variation in bacteria, and is known to occur within *C. jejuni*. Data revealed in Chapter Four indicated the possible presence of prophage

DNA within the ST403CC *C. jejuni* isolates. As such the goal of this chapter is to utilise sequence based techniques to investigate the prevalence and potential influence of prophages or other integrated genomic elements in the evolutionary separation of the ST403CC *C. jejuni* isolates from other *C. jejuni*.

5.2 Methods

In this chapter a range of sequence based methods are combined to investigate the presence of 'foreign' DNA content within the genomes of potentially host adapted ST403CC *C. jejuni* isolates.

This process will consider the presence and content of any integrated phage sequence, and investigate other inserted elements such as IS, transposons or genomic islands.

5.2.1 Investigating the *intA* Gene

The predicted prophage integrase gene *intA* was observed in *C. jejuni* 857 during investigation of regions surrounding ST403CC 'exclusive' genes - CDS identified during pan genome analysis which were common to all six included ST403CC *C. jejuni* isolates and with homologues not present within the included *C. jejuni* and *C. coli* reference genomes. Firstly, in order to determine the prevalence of *intA* across the six ST403CC *C. jejuni* genomes, and to investigate the content of the neighbouring sequences, regions were examined using Artemis viewer (Rutherford *et al*, 2000) as described previously. The annotated genomes were searched using the 'go to' feature built into Artemis (Goto Feature with Gene Name) to quickly locate *intA*, and the surrounding area viewed for other CDS of potential relevance. This was done using annotations transferred from the PROKKA annotation (Seemann, 2014; using RATT; Otto *et al*, 2011) of ST403CC *C. jejuni* isolate 857, in order to maintain the same CDS tags and annotation.

Local database similarity searches were carried also carried out, using BLASTn, as introduced in Chapter Four (Altschul *et al*, 1990). Local BLAST database searches were conducted using nucleotide BLAST, and the *intA* sequence was used as taken from *C. jejuni* strain 857 and queried against the 34 genome local database

containing genome sequence for the 33 genomes included in phylogenetic and pan genome analysis (Table 3.1; Appendix 9.1), as well as the additional historical ST403CC *C. jejuni* isolate ATCC33560.

5.2.2 Identifying Prophage Content

To investigate prophage content in the ST403CC *C. jejuni* isolates, PHAST (Zhou *et* al, 2011) was run using the browser based service at http://phast.wishartlab.com/index.html submitting each sequence individually.

5.2.2.1 Improved Annotation for PHAST Analysis

Each of the six ST403CC *C. jejuni* genomes was annotated using PROKKA (Seemann, 2014) to produce the genbank format annotated files used for PHAST investigation.

5.2.2.2 PHAST Analysis of Prophage Content

To investigate the consistency of the method, both 'raw' FastA and annotated genbank format files were used and received consistent results; subsequently, results obtained using the genbank format files were used for analysis and discussion in order to relate to the annotated genomes. For *C. jejuni* isolate ATCC33560 files were used by submitting the unannotated genome via genbank accession 380625642.

5.2.3 Identifying Integrated Genomic Content using AlienHunter

Alienhunter (Vernikos & Parkhill, 2006) was run with command line using fasta files, and the output was read into the annotated genome sequence in Artemis (Rutherford *et al*, 2000).

5.3 Results

The determined foreign DNA content for *C. jejuni* ST403CC isolates is described in this section, beginning with a summary of the *intA* related content using Artemis (Rutherford *et al*, 2000) and local database BLAST searches (Altschul *et al*, 1990) and followed by detailed information provided by PHAST (Zhou *et al*, 2011) and AlienHunter (Vernikos & Parkhill, 2006) analysis.

5.3.1 Prevalence of *intA*

As described previously, the presence of *intA* detected in ST403CC *C. jejuni* isolate 857 in Chapter Four lead to interest in prophage regions in ST403CC genomes. The first step taken was to confirm whether *intA* homologues existed in all sequenced ST403CC *C. jejuni* isolates, and to investigate the surrounding regions in the genomes for additional prophage associated content, as well as considering the prevalence of this gene amongst non-ST403CC *C. jejuni* or *C. coli* genomes.

As shown in Table 5.1, a homologue of *C. jejuni* 857 *intA* gene was observed in each included ST403CC *C. jejuni* isolate. The *intA* coding sequence was found to be a consistent length (627bp, 208aa) in the six newly sequenced ST403CC *C. jejuni* isolates, however some variation was observed in its location and the neighbouring CDS. In *C. jejuni* strain 857, *intA* is located between CJ857_00834 (downstream) and xerD_2 (upstream). This same organisation is also observed in 549.1, 623, 304 and 444 (although the neighbouring region is more heavily pseudogenised in 444). The same pattern is not observed for strain 484 however, where *intA* is found at a distinct and distant site.

In the case of *C. jejuni* 484, *CJ857_00839* (an ST403CC 'exclusive' gene as described in Chapter Four) and *xerD_2* are still found located close to each other, and *intA* is still a direct neighbour of *CJ857_00834*, however it is also linked to possible pseudogene *CJ857_01602* on the complementary strand (Figure 5.1). *CJ857_01602* is also in the region in *C. jejuni* 444 however it is upstream of *intA* in this case and remains frameshift pseudogeneised but with *CJ857_00839* overlapping.

It was noted that in each case the region surrounding *intA* tended to be made up of hypothetical proteins likely to be pseudogenes or fragments, including one of the previously identified 'ST403CC exclusive' CDS, however no additional prophage linked CDS were observed.

In Figure 5.1 additional potential ORFs can be seen - these were not annotated due to the transference of annotation from *C. jejuni* 857 to *C. jejuni* 484 which simply transferred the initial annotation, and cannot ascribe new ORFs; however, NCBI

nucleotide BLAST searches were carried out and these ORFs would only represent additional hypothetical protein regions, and therefore do not significantly affect the investigation.

Isolate	Size (bp)	Location	Downstream Neighbour	Upstream Neighbour
C. jejuni 857	627	774069..774695	CJ857_00834	xerD_2
C. jejuni 549.1	627	581511..582137	CJ857_00834	xerD_2
C. jejuni 623	627	731890..732516	CJ857_00834	xerD_2
C. jejuni 304	627	1061331..1061957	CJ857_00834	xerD_2
C. jejuni 484	627	805772..806398	CJ857_00834	CJ857_01602
C. jejuni 444	627	776779..777405	CJ857_00834	xerD_2

Table 5.1: Presence of the *intA* Gene in ST403CC *C. jejuni* Isolates

Comparing the sequence length, location and neighbouring CDS of the predicted *intA* gene in ST403CC *C. jejuni* isolates

Nucleotide BLAST queries were run against the local databases to investigate whether *intA* was specific to the ST403CC *C. jejuni* isolates (full output for local database BLAST queries are included in Appendix 9.4).

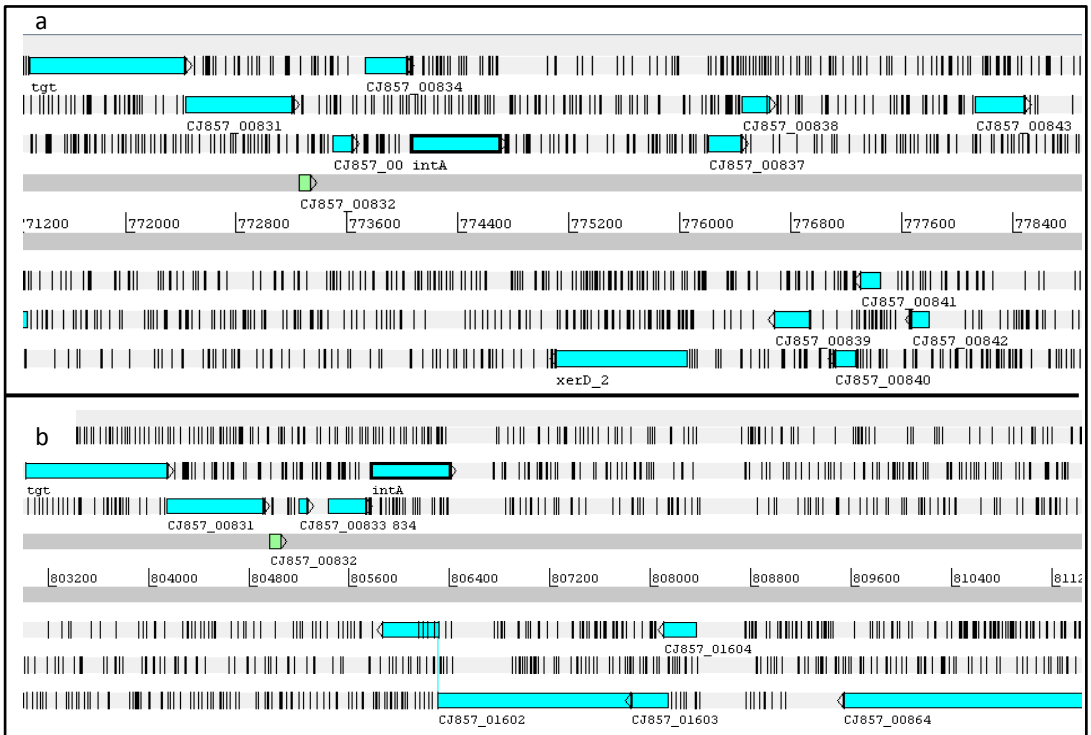

Figure 5.1: The *intA* Gene in ST403CC *C. jejuni* Isolates 857 & 484

intA coding sequence in a) 857 and b) 484. Other ST403CC *C. jejuni* isolates also possess *intA* homologue in highly similar location and configuration as 857.

The 'self-match' score for *C. jejuni* 857 *intA* gene (1158) was closely matched by all other ST403CC isolates (range 1153-1158, mean 1156). Additional matches were observed however, one in *C. jejuni* 1336 (score 957) and in *C. coli* 67-8 (score 1079), 2685 (score 1024), 80352 (score 1079), LMG9860 (score 1068; range 1024-1079, mean 1063). The presence of *intA* was shown to be common to all sequenced ST403CC *C. jejuni* isolates, but also occurs in other non-ST403CC *C. jejuni* and in *C. coli*. These isolates represent diverse host types; *C. jejuni* 1336 is a wildlife associated isolate (French *et al*, 2005) whilst *C. coli* 67-8 is porcine, *C. coli* 2685 is from turkey, *C. coli* 80352 is from chicken, and *C. coli* LMG9860 is a human isolate (Lefébure *et al*, 2010).

Although the *intA* gene did not appear to be associated with additional phage DNA in ST403CC *C. jejuni* isolates, it was still considered worth-while to conduct analysis of possible prophage content in the ST403CC isolate genomes. First using PHAST to look for specifically for the presence of prophage and subsequently using AlienHunter to confirm results from PHAST and consider other kinds of foreign DNA including integrated elements and plasmid associated content.

5.3.2 PHAST Analysis & Prophage Content

PHAST provided overview images for each isolate; showing a circular genome diagram indicating the presence, location and status of prophage content (Figure 5.2). As Figure 5.2 shows, at least one prophage region was predicted for each of the porcine origin ST403CC isolates, however no prophage region was identified in the historical, bovine ST403CC isolate ATCC33560. As shown in the legend, PHAST indicates the possible status of a prophage region as intact (red colour), incomplete (grey) or questionable (green); each of the ST403CC porcine *C. jejuni* isolates contained at least one potential intact prophage region, whilst some strains contained additional intact or incomplete prophage regions.

Considering each isolate individually, *C. jejuni* 857 had only one phage region, which was determined to be an intact region of approximately 8kbp, whereas *C. jejuni* strain 549.1 contained three prophage regions; regions one (28118bp) and three (6792bp) were designated as intact prophage, whilst region two (18799) was

considered incomplete. Strain 623 was found to carry two intact predicted phage regions, measuring approximately 31 and 5kbp.

Figure 5.2: Overview of Prophage Content in ST403CC *C. jejuni* Isolates as Determined by PHAST

PHAST output images showing circular genome with identified prophage regions.
a. *C. jejuni* 857
b. *C. jejuni* 549.1
c. *C. jejuni* 623
d. *C. jejuni* 304
e. *C. jejuni* 484
f. *C. jejuni* 444
g. *C. jejuni* ATCC33560

Strain 304 contained three observed phage regions of approximately 13, 42 and 5kbp in length respectively, the first of which was incomplete, with the remaining two being considered intact. Strain 484 had two phage regions, both intact, at sizes of 49 and 19kbp respectively. Finally, strain 444 contained only one predicted phage region, with a length of around 5kbp and predicted to be an intact prophage region.

In summary, two of the porcine ST403CC *C. jejuni* isolates (857 and 444) were thought to have only one (intact) prophage region, two (623 and 484) were observed as having two (intact) prophage regions, whilst the final two strains (549.1 and 304) each contained three predicted prophage regions; two intact and one incomplete.

The circular genome diagrams produced by PHAST provided a useful overview of the potential phage content of ST403CC *C. jejuni* isolates; below these regions are discussed in greater detail.

As stated previously, each of the porcine ST403CC *C. jejuni* isolates contained at least one intact prophage region; further investigation showed that this represented a common prophage: with matching sequence in Enterobacteriaceae phage F1 sensu lato, a group IV RNA phage (Figure 5.3). The PHAST predicted F1 prophage regions range from 5389-19215 bp in length, with between 6 and 12 predicted coding sequences. The length is skewed somewhat by strain 484 - which registered a distant phage attachment site as part of the same phage region - if this were discarded from consideration the predicted phage region would be approximately 9kbp. Table 5.2 summarises the size and predicted content of the type *Enterobacteriaceae* F1 phage compared to that F1 prophage region predicted in each ST403CC isolate.

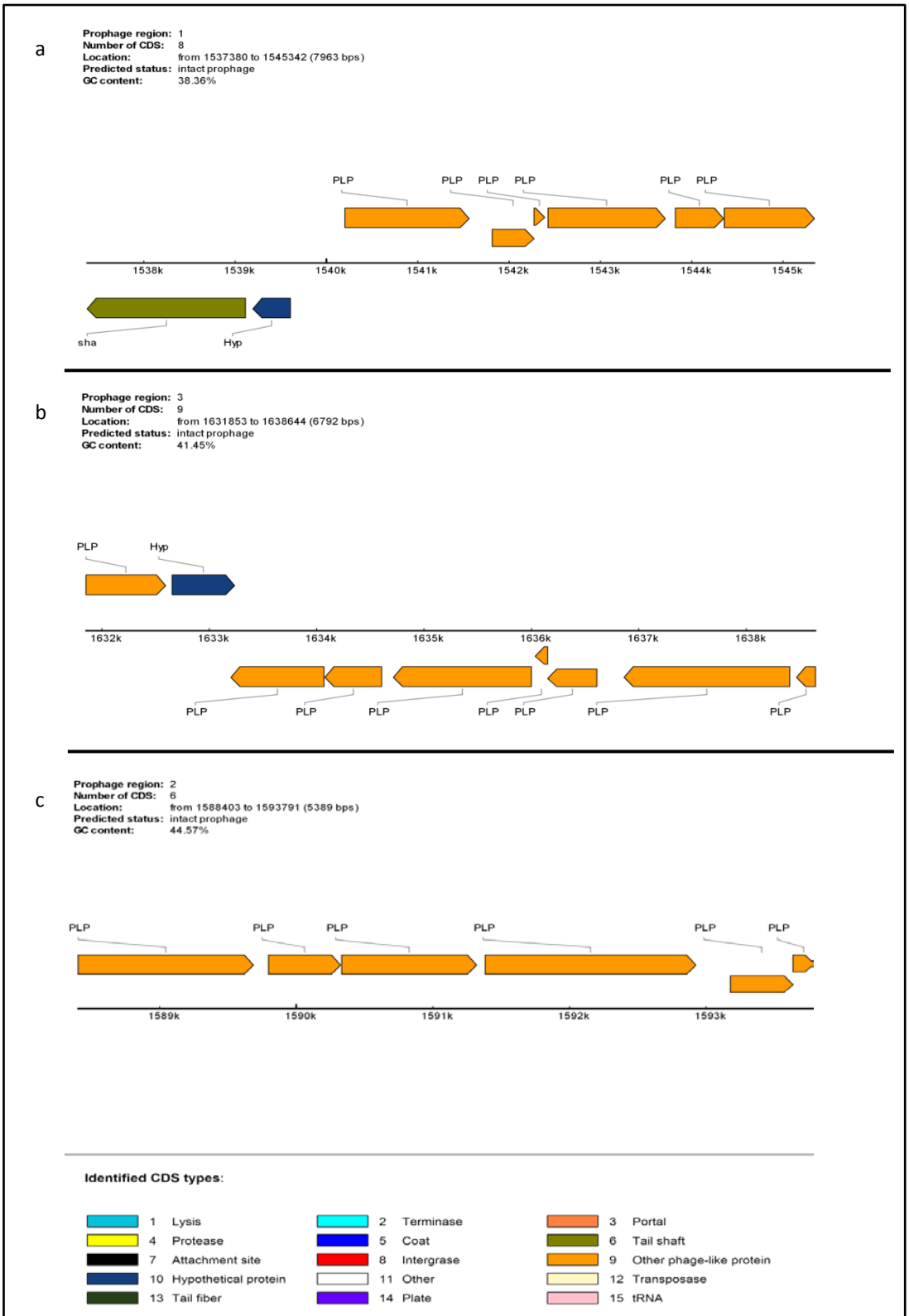

Figure 5.3a: Content of Common Prophage Region in ST403CC *C. jejuni* Isolates

PHAST regions identified by similarity with Enterobacteriaceae phage F1 sensu lato for isolates a) *C. jejuni* 857, b) *C. jejuni* 549.1, c) *C. jejuni* 623

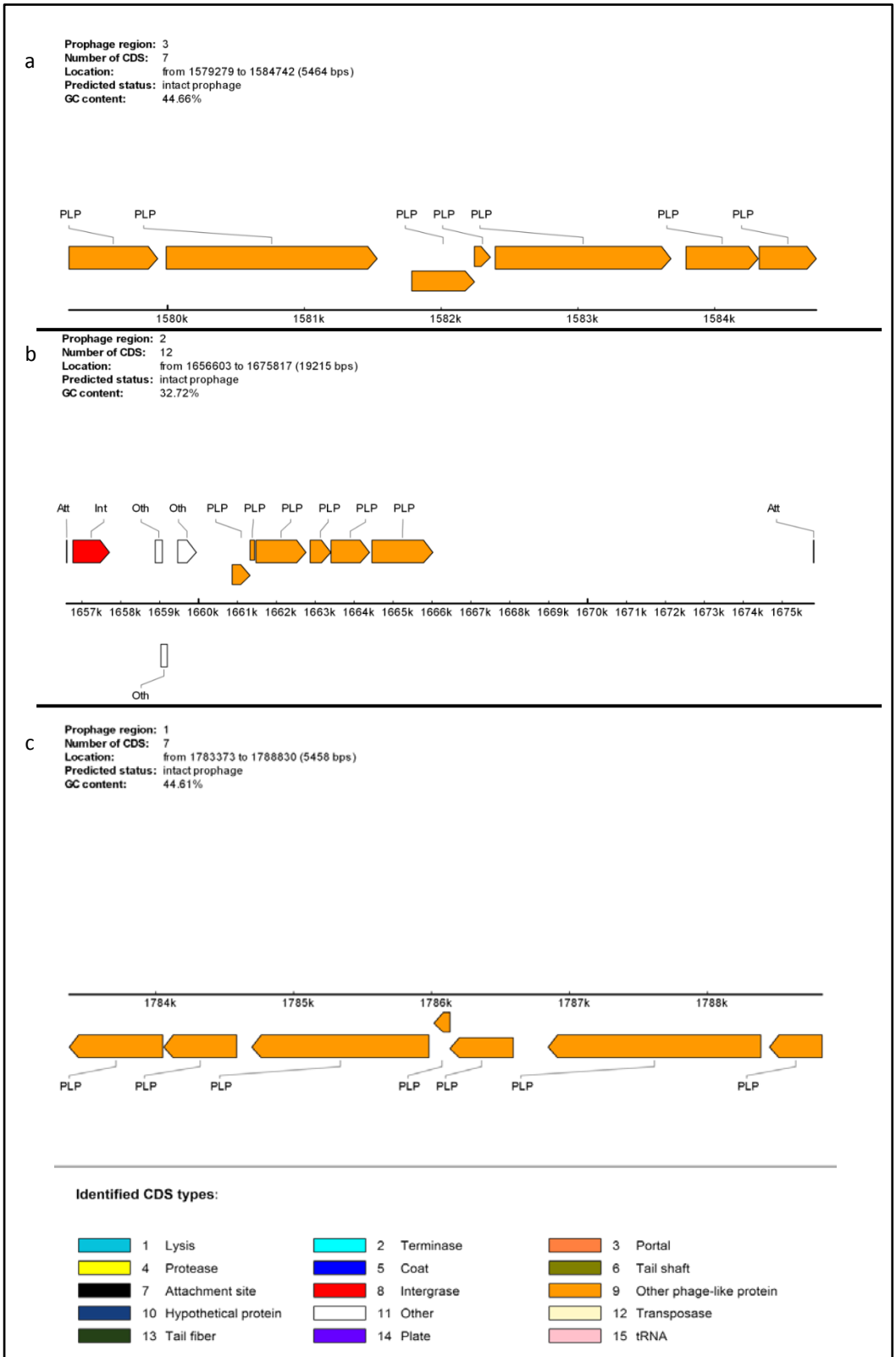

Figure 5.3b: Content of Common Prophage Region in ST403CC *C. jejuni* Isolates

PHAST regions identified by similarity with Enterobacteriaceae phage F1 sensu lato for isolates a) *C. jejuni* 304, b)*C. jejuni* 484, c) *C. jejuni* 444

The reference genome for *Enterobacteriaceae* phage F1 sensu lato is 4276bp in length, with three coding regions - a maturation protein, a phase-variable coat protein, and the replication protein (Inokuchi *et al*, 1988), whereas the ST403CC regions were tagged with considerably more potential coding regions. The coding content is described in further detail later in the section.

Figures 5.3a and 5.3b provide detailed PHAST output on the *Enterobacteriaceae* F1 phage region identified in each of the porcine ST403CC *C. jejuni* isolates. The majority of the coding regions tagged by PHAST in the *Enterobacteriaceae* phage F1 were categorised as 'other phage like protein', some of these have more detailed descriptions in the comprehensive PHAST output, however in order to give a more detailed picture of the regions, PROKKA annotations were combined in the analysis of the region.

Sequence	F1 sensu lato	857	549.1	623	304	484	444
Length (bp)	4276	7963	6792	5389	5464	19215	5458
#CDS	3	8	9	6	7	12	7

Table 5.2: Prophage Region Common to ST403CC *C. jejuni* Isolates

Comparing the sequence length and content of the predicted phage region shared by the ST403CC *C. jejuni* isolates against the original phage identified as a potential match.

The identified *Enterobacteriaceae* prophage F1 region observed in the porcine ST403CC *C. jejuni* isolates contained between 6 (strain 623) and 12 (strain 484) CDS according to PHAST study (mean number of CDS; 8). Considering the PROKKA annotation of these regions, the number of CDS observed remained consistent for strains 857, 549.1, 623, 304 and 444, however the region described by PHAST in strain 484 contained a number of additional predicted coding regions within the large empty space observed by PHAST, where PHAST identified 12 potential prophage coding regions, PROKKA identified a total of 24 potential coding regions, including xerD_2, two flagellin proteins, glycosyltransferase family two protein, carbon starvation protein A and general stress protein A, guanylate kinase *gmk*,

twin arginine translocase protein A, arginine tRNA ligase, and two spermidine export proteins *mdtI* and *mdtJ*, as well as 8 hypothetical proteins.

Considerable similarity was observed for the F1 prophage region across the isolates; homologues of 'bacteriophage scaffolding protein D', 'microvirus J protein', 'capsid protein (F protein)', 'major spike protein (G protein)' and 'microvirus H protein (pilot protein)' were found in all six of the ST403CC isolates. Additionally 'bacteriophage replication gene A protein GPA' was found in five of the six porcine ST403CC isolates, lacking a homologue only in *C. jejuni* 484. Additional content included hypothetical proteins in *C. jejuni* strains 857, 549.1 and 484, a 'phosphate import ATP-binding protein *pstB'* in strain 549.1 and numerous additional isolates were observed in strain 484 as described above.

Whereas ST403CC *C. jejuni* strains 857 and 444 contained only the *Enterobacteriaceae* phage F1, the remaining four isolates contained either two or three prophages in total. Two prophages were identified which were each shared by two of the ST403CC porcine *C. jejuni* isolates. The first of these was *Enterobacteriaceae* prophage Fels_2.

Enterobacteriaceae phage Fels_2 matching regions were identified in *C. jejuni* 549.1 and 623, as shown in Figure 5.4 with lengths of 28118bp and 31087bp and with 44 and 49 coding sequences respectively identified by PHAST; the 44 coding sequences are identical between the two, with 623 having an additional five coding sequences at the upstream end of the phage region, each of which are considered hypothetical proteins by PHAST.

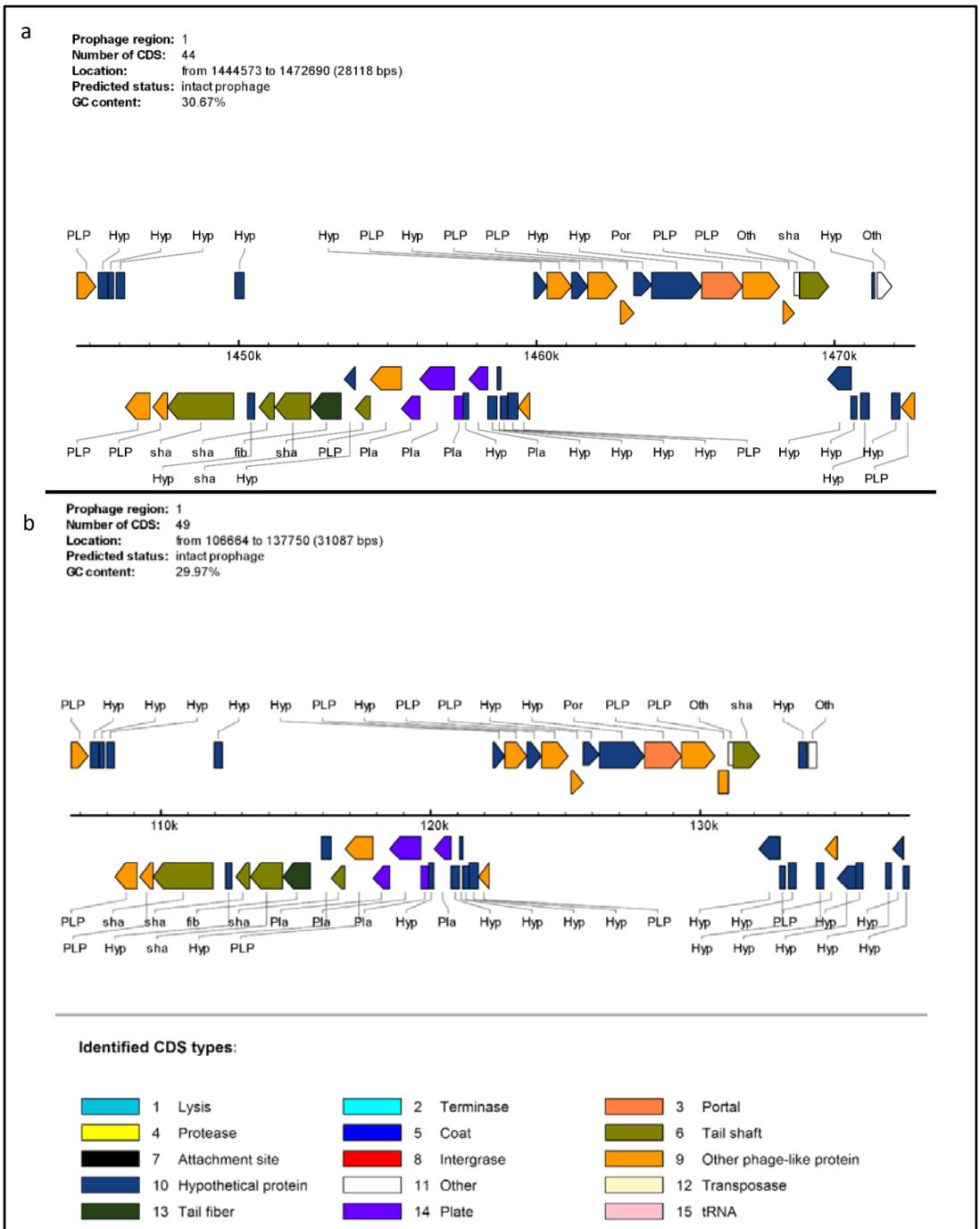

Figure 5.4: Content of Prophage Region Observed in ST403CC *C. jejuni* Isolates 549.1 & 623

Regions identified as related to *Enterobacteriaceae* prophage Fels_2 in *C. jejuni* strains a) 549.1 and b)623.

The second prophage which was determined to be common to two of the six porcine ST403CC *C. jejuni* isolates was the prophage *Haemophilus* phage SuMu. PHAST analysis identified regions matching Haemophilus phage SuMu in both *C. jejuni* 549.1 and 484, as shown in Figure 5.5. The region is approximately 18kbp in

549.1 and 49kbp in 484, with the regions having 15 and 58 coding sequences, respectively.

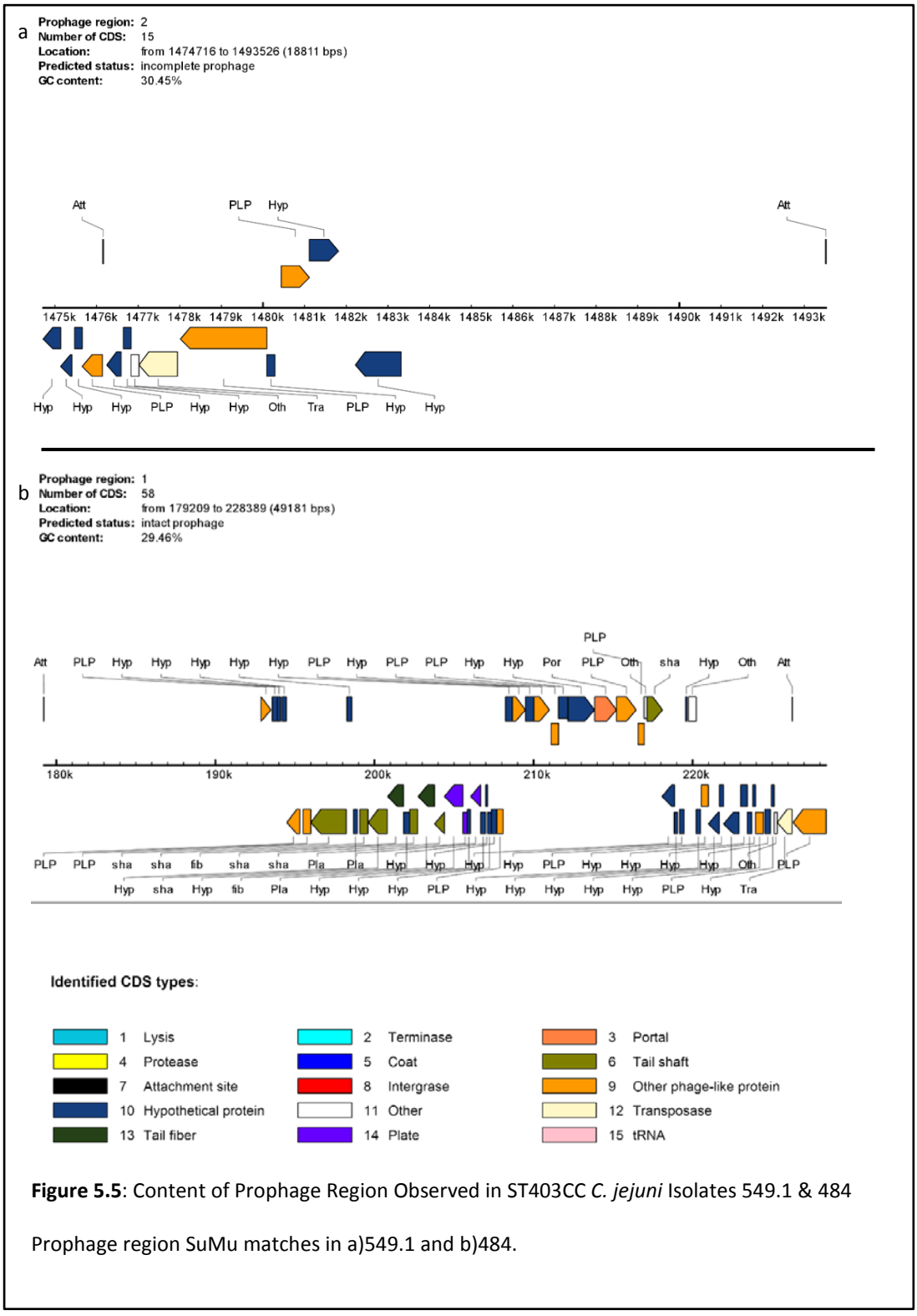

Figure 5.5: Content of Prophage Region Observed in ST403CC *C. jejuni* Isolates 549.1 & 484

Prophage region SuMu matches in a)549.1 and b)484.

PHAST tagged 58 CDS for the 484 version of SuMu, whereas PROKKA annotation of the same area tagged 72 CDS. Two of the PHAST CDS (CDS one and fifty eight, the Att proteins) do not have corresponding annotation in PROKKA annotation, however the PROKKA annotation has 16 additional CDS not tagged by PHAST at the downstream end (between where the Att 'should be' and the second PHAST CDS (PLP) which corresponds to CJ484_00217). Ten PHAST CDS are shared between the 549.1 and 484 SuMu prophage regions; CDS one to ten in 549.1 match CDS 49-58 in 484. Additionally, the main 'body' of the SuMu prophage region in 484 is strikingly similar to the Fels_2 region previously described; PHAST CDS 1-46 of strain 623 Fels_2 region correspond (by prokka annotation) to CDS 2-47 of SuMu in strain 484.

Finally there were also two phage types which were observed only in one of the ST403CC isolates, both of which were observed within strain 304. These included *Mannheimia* phage phiMHaA1 (Figure 5.6) and *Synechococcus* phage S-SKS1 (Figure 5.7).

In ST403CC *C. jejuni* isolate 304 the first additional prophage matched with *Mannheimia* phage phiMHaA1, and was 13493bp in length, with 16 coding sequences described by both PHAST and PROKKA.

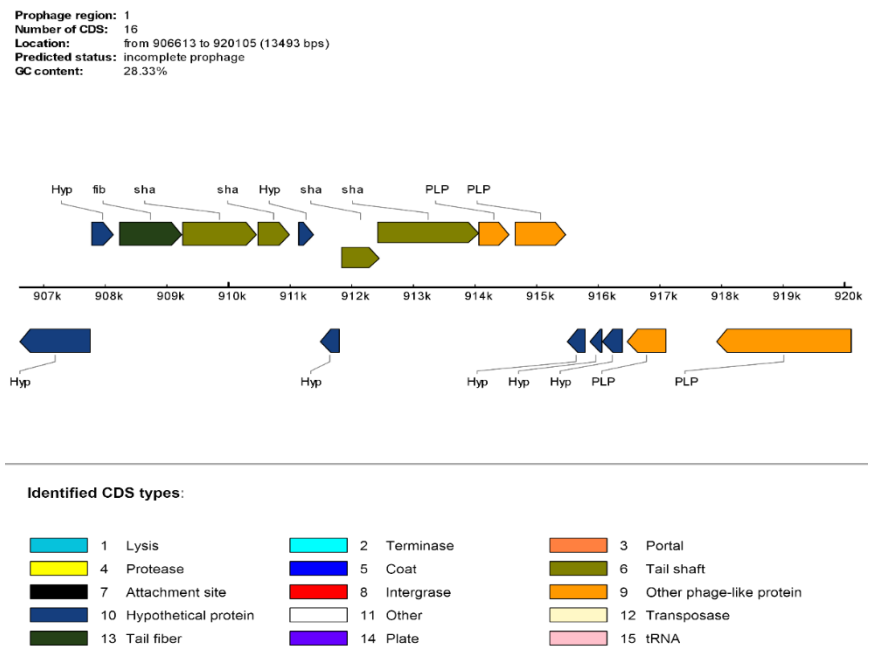

Figure 5.6: Content of First Prophage Region Observed in ST403CC *C. jejuni* Isolate 304

The second prophage region associated only with ST403CC *C. jejuni* isolate 304 recorded similarity with *Synechoccus* phage S-SKSI, is 42706bp in length and has 64 coding sequences identified by both PHAST and PROKKA. In PHAST analysis the S-SKSI region was found to be bookended by a pair of *Att* proteins, however these were not identified by PROKKA, in fact they lay within larger CDS by this means of annotation (a hypothetical protein at PHAST CDS one and a putative transcriptional regulatory protein at CDS 64).

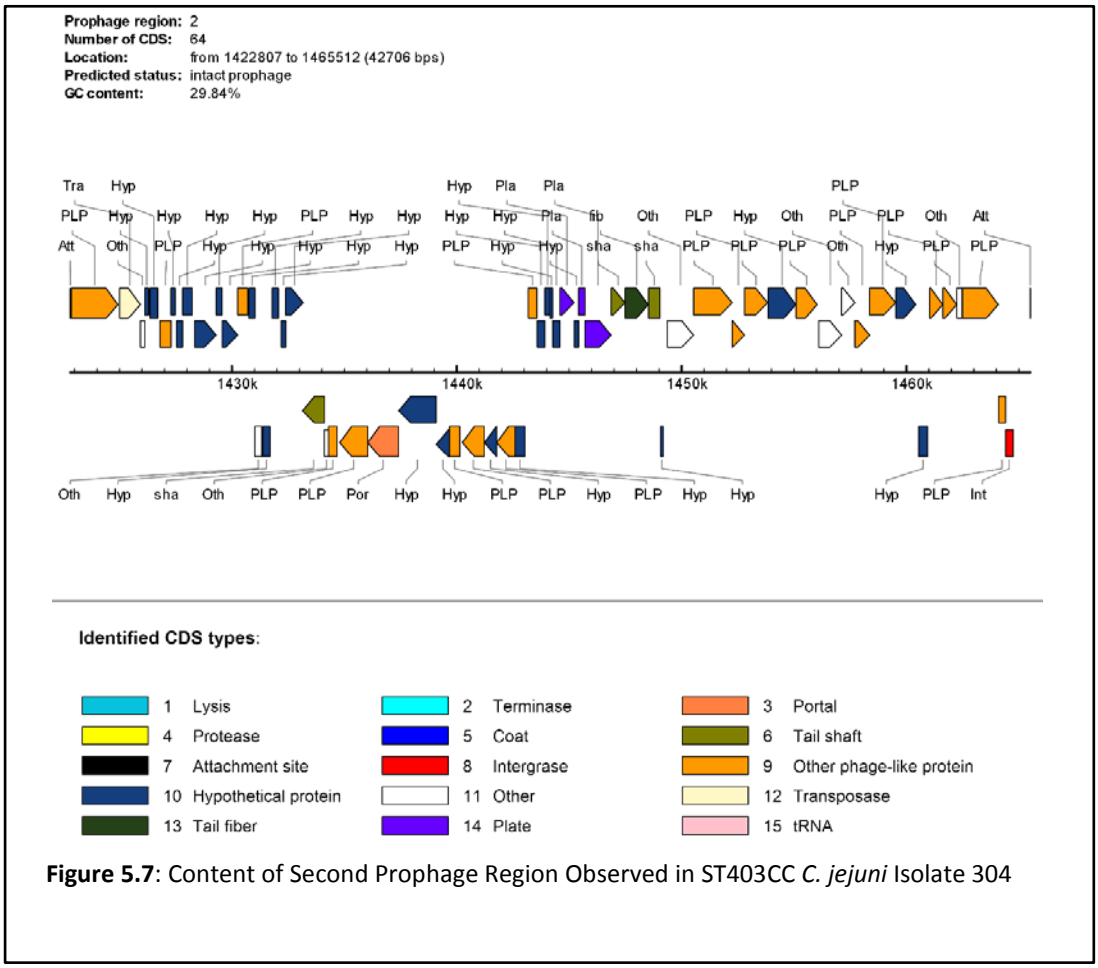

Figure 5.7: Content of Second Prophage Region Observed in ST403CC *C. jejuni* Isolate 304

Table 5.3 provides an overview of the predicted prophage content recorded in ST403CC *C. jejuni* isolates, as predicted by PHAST (Zhou *et al*, 2011).

Strain	#CDS	Region Length	Phage Similarity	GC Percentage	Genome GC Percentage	Phage State
857	8	7963	Enterobacteria Fl sensu lato	38.36%	30.06	Intact
549.1	44	28118	Enterobacteria Fels 2	30.67%	30.08	Intact
	13	18799	Haemophilus SuMu	30.45%		Incomplete
	9	6792	Enterobacteria Fl sensu lato	41.45%		Intact
623	49	31087	Enterobacteria Fels 2	29.97%	30.02	Intact
	6	5389	Enterobacteria Fl sensu lato	44.57%		Intact
304	16	13493	Mannheimia phiMHaA1	28.33%	29.97	Incomplete
	62	42693	Synechococcus S-SKS1	29.84%		Intact
	7	5464	Enterobacteria Fl sensu lato	44.66%		Intact
484	56	49181	Haemophilus SuMu	29.46%	30.23	Intact
	10	19203	Enterobacteria Fl sensu lato	32.72%		Intact
444	7	5458	Enterobacteria Fl sensu lato	44.61%	30.01	Intact
ATCC33560	0	0	0	0	-	None

Table 5.3: Summary of Prophage Regions Identified by PHAST Analysis in ST403CC *C. jejuni* Isolates
Displaying the number, size, status and content of each predicted prophage region in investigated ST403CC *C. jejuni* isolates.

5.3.3 Integrated Genomic Content & AlienHunter

Having considered results from initial impressions using Artemis and BLAST, and following the more detailed, prophage focused, approach of PHAST the final method employed was AlienHunter, which as stated previously looks for all sections of nucleotide sequence of potentially 'outside' sources, this would include prophage but also genomic islands and other inserted regions. Results from AlienHunter are summarised in Table 5.4.

As previously described, AlienHunter (Vernikos & Parkhill, 2006) produces output which can be read into Artemis (Rutherford *et al*, 2000) over annotated sequence files, to aid further investigation into location and content of potential 'foreign' regions.

Isolate	Number of AlienHunter Regions
C. jejuni 857	81
C. jejuni 549.1	82
C. jejuni 623	64
C. jejuni 304	81
C. jejuni 484	64
C. jejuni 444	45
C. jejuni ATCC33560	66

Table 5.4: Integrated Regions in ST403CC *C. jejuni* Isolates Predicted by Alien Hunter Analysis

Displayed are the number of predicted 'alien' regions determined by the AlienHunter software in each of the studied ST403CC *C. jejuni* isolates.

Visual AlienHunter results were colour coded from white through pink to red, with darker regions (red) representing higher scores indicating more divergent regions and paler colours indicating lower divergence, higher similarity to the host genome.

As shown in Figure 5.8, overleaf, AlienHunter found a considerable number of potential inserted regions across each of the ST403CC *C. jejuni* isolates including ATCC33560 (range 45-82, mean 69).

5.3.4 Relating PHAST & AlienHunter Prophage Results

In order to relate the results of AlienHunter with those from PHAST, in this section the overlap between PHAST and AH described regions are investigated. Firstly considering *C. jejuni* ST403CC isolate 857 as stated previously contained one identified prophage region, with homology to *Enterobacteria* phage F1, located between 1537380 and 1545342 on the chromosome. This region maps almost exactly with an identified 'alien' region from AH results; AH region 61, existing at bases 1537500 to 1550000.

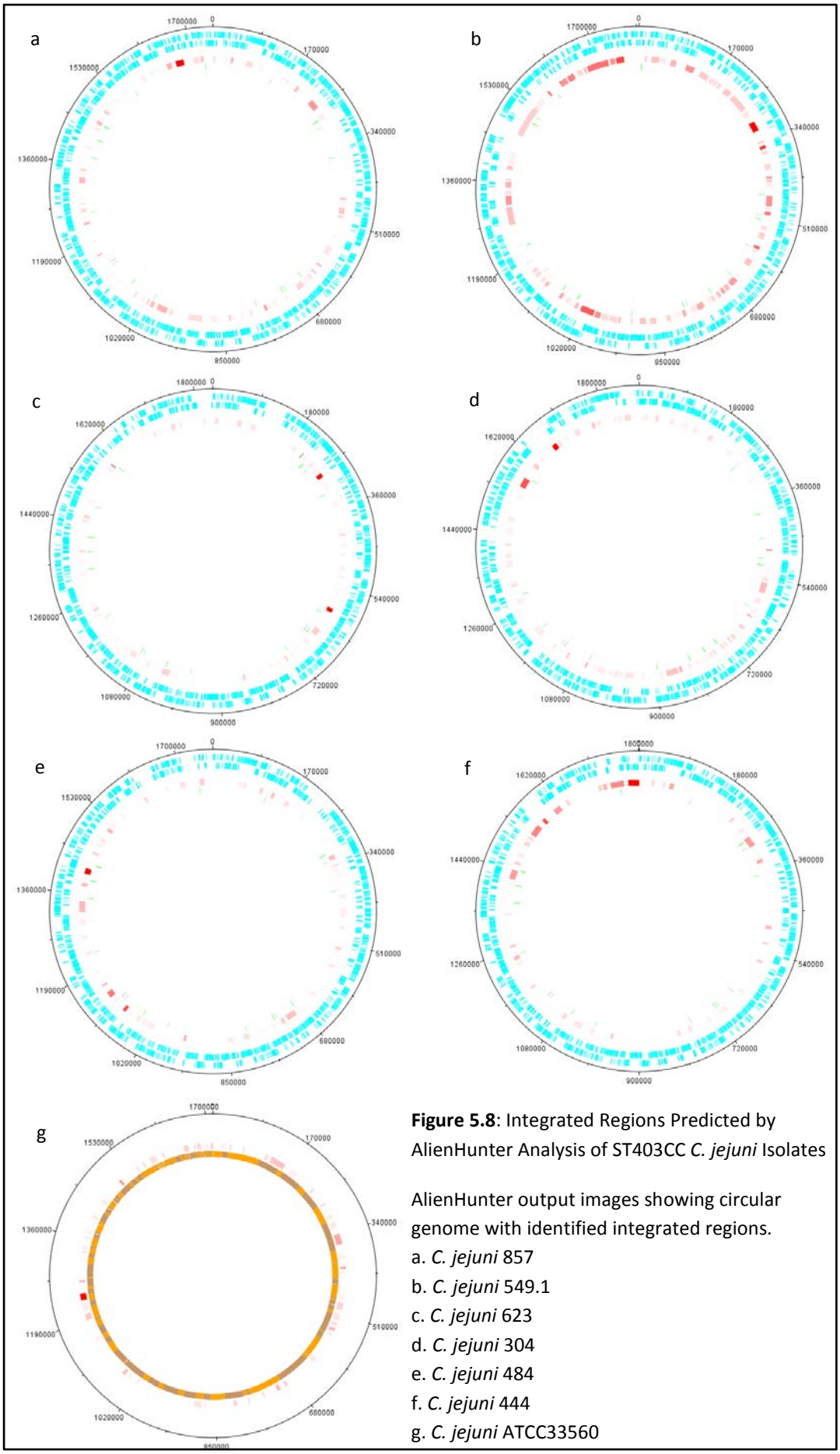

Figure 5.8: Integrated Regions Predicted by AlienHunter Analysis of ST403CC *C. jejuni* Isolates

AlienHunter output images showing circular genome with identified integrated regions.

a. *C. jejuni* 857
b. *C. jejuni* 549.1
c. *C. jejuni* 623
d. *C. jejuni* 304
e. *C. jejuni* 484
f. *C. jejuni* 444
g. *C. jejuni* ATCC33560

Strain 549.1 also demonstrated some agreement between PHAST and AlienHunter; the F1 phage region (1631853..1638644bp) is contained within AlienHunter region 76 (1627500..1642500), whilst both the Fels_2 and SuMu regions were found to be associated with AH region 71; Fels_2 (1444573..1472690) displayed partial overlap with AH region 71, whilst SuMu (1474716..1493526) was contained within the AH71 region (1465000..1532500). For strain 623 however no such association was observed, for both F1 (1588403..1593791) and Fels_2 (106664..137750) no corresponding region was observed by AlienHunter investigation. Strain 304 had mixed results; as with 623 there was no corresponding AH region for 304's F1 region (1579279..1584742), however both the Mannheimia phage region (906613..920105) and Synechococcus region (1422807..1465512) were associated with AlienHunter regions; the Mannheimia phage contained AH region 40 (907500..912500) whilst the Synechococcus region contained both AH region 64 (1432500..1437500) and region 65 (1440000..1450000). Isolate 484 recorded Alienhunter region matches for both PHAST prophage regions; F1 (1656603..1675817) contained region AH 62 (1662500..1667500) and SuMu region contained AH 8 (210000..215000). Finally, strain 444 also contained agreement between PHAST and AlienHunter, with its F1 prophage region (1783373..1788830) contained within Alienhunter region 45 (1777500..1800282). It was also noted that *intA* was not located within the AlienHunter identified 'foreign' regions.

5.3.5 Summary of Results

In summary, the phage associated gene *intA* was confirmed present in each of theST403CC *C. jejuni* isolates, however it was also found in one other *C. jejuni* isolate included in the study, and four *C. coli* isolates. The location of *intA* showed some variation in location and surrounding CDS within the ST40CC *C. jejuni* isolates, however it was commonly associated with a second copy of xerD_2 and CJ857_00839.

Potential prophage regions were observed in each of the porcine ST403CC *C. jejuni* isolates, however no prophage regions were observed in bovine ST403CC *C. jejuni* isolate ATCC33560. Each of the six ST403CC isolates shared one common called

region, with some strains having just this one region and others containing upto three potential prophage regions; further consideration of the origin and impact of these regions is discussed below.

Additional predicted integrated regions were identified by study with AlienHunter, which showed some agreement with PHAST prophage results, but did not indicate the presence of an integrated region containing the *intA* homologue.

5.4 Discussion

It was established in this chapter that ST403CC *C. jejuni* isolates contain predicted integrated genomic regions, including but not limited to potential prophage regions. The contents of these regions were described, however further discussion is required to establish potential implications of these regions and their contents in order to draw conclusions.

5.4.1 Prevalence of *intA*

The *intA* homologue observed in ST403CC *C. jejuni* strain 857 led to the investigation of the possibility of prophage content within the ST403CC *C. jejuni* isolates. The gene *intA* is a prophage integrase, also referred to as *intX* or *slpA*, described by Blattner *et al* (1997) in their annotation of the genome sequence of *E. coli* strain K12, located on the prophage CP4-57 (as described by first observed by Kirby, Trempy & Gottesman (1994) who identified CP4-57 (cryptic P4-like prophage at minute 57). A potential homologue of *intA* (*CJ857_00835*) was present in each (porcine) ST403CC isolate (local blast also confirmed presence in ATCC33560), although the region surrounding it varied between strains (for *C. jejuni* 857, 623, 549.1, 304 and 444 *intA* was between CJ857_00834 and xerD_2, whilst in 484 it was at a divergent region, and still neighboured with CJ857_00834, though xerD_2 and rest of region were still at the 'normal' location as in the other five). In each case the CDS near *intA* are largely hypothetical, and probably represent pseudogenes. Local database BLAST searches showed that *intA* homologues were present in one of the thirteen included non-ST403CC *C. jejuni* isolates, and four of forty-eight *C. coli* genomes. The presence of *intA* was not associated only with ST403CC *C. jejuni*

isolates, and was also not associated with a specific host type, with the *intA* positive strains coming from a variety of sources; *C. jejuni* 1336 was first isolated by French *et al* (2005) during a UK MLST study investigating association of specific sequence types with certain environments, and was subsequently investigated further by Hepworth *et al* (2011) using comparative genomic hybridisation, and genome sequencing, and was described as a representative isolate from a group of water and wildlife isolates; whereas *C. coli* isolates 67-8, 2685, 80352 and LMG9860 were each sequenced as part of a large sequencing project by Lefébure *et al* (2010) investigating the genome dynamics of *C. coli* and *C. jejuni*, each representing different source types (swine, turkey, chicken, human).

The neighbouring region surrounding *intA* consisted largely of short potential coding sequences, mostly described as hypothetical proteins and also including in each case a second copy of xerD, a tyrosine recombinase. In five of the six porcine ST403CC isolates, the neighbouring region also contained one of the 'ST403CC exclusive EDGAR CDS' as described in Chapter Four, however this, as most of the CDS in the region are very short regions which may represent degraded genes.

Results observed using both PHAST and AlienHunter demonstrated that the potentially prophage associated gene *intA* observed in ST403CC *C. jejuni* isolates was not located within a predicted integrated prophage region. There may be several reasons for this seemingly surprising result; firstly, genome annotation was carried out using PROKKA (Seemann, 2014), which uses a large database to find the most likely matches for each CDS, however it does not 'sense check' observed matches, so the process can give occasionally produce some inaccurate answers. An NCBI BLASTp query for *CJ857_00835 intA* recorded matches for 30S ribosomal protein S15 in various *C. jejuni* and *C. coli* genomes (100% coverage, 99% identity). Secondly, the assembly of the short read sequence data may have placed the gene in an inaccurate location; assembly was carried out using a reference to build against so each area is lined up against that, this can mean that additional or divergent areas can be tagged on towards the 'end' of the genome, however it is unlikely that a single CDS would be put in the 'wrong' place. Thirdly, is the possibility that the *intA* homologue in fact be within a prophage region which was

overlooked by both PHAST and AlienHunter, however this is considered highly unlikely, as it would be very surprising for a genuine prophage region to be missed by both methods. Finally it may be possible that this *intA* coding sequence was genuinely an integration protein for a prophage, which was under neutral or negative selection pressure and was degraded so as to be unrecognisable to the detection methods utilised.

5.4.2 PHAST Prophage Content

Although the potential prophage gene *intA* did not appear to be associated with integrated prophage regions in ST403CC *C. jejuni* isolates, as described earlier, potential prophage content was observed for each of the six porcine ST403CC *C. jejuni* isolates. Two strains had only one potential prophage region, two had two and two had three.

Potential prophage regions were initially determined for each of the six porcine ST403CC *C. jejuni* isolates with a single common predicted prophage shared across the six strains. However, during subsequent analysis and following an update to the PHAST database the evidence indicated that the shared predicted prophage region tagged by PHAST analysis as a potential relative of *Enterobacteria* phage F1 was in fact potentially associated with phage phix174. Carrying out repeat PHAST runs using the recently updated viral database gives matches previously identified as the 'f1 region' being designated 'phix174' regions. This indicates that rather than being a group specific prophage this may be a remnant of phix174 control spikes from the Illumina sequence run. This phage is commonly used as a control during Illumina sequencing procedures, such as those used in this project. Although efforts can be made to strip remnants of phiX174 data from the sequence, and targeted reference based assembly makes an extra step to help remove these sequences, fragments can be left behind in the final sequence. It was therefore considered likely that this potential common region is remnant phix174 sequence from the Illumina sequencing procedure. *Enterobacteria* phage F1 is a relative of phiX174, both being from the family Microvirus, there remains some possibility that this region may represent a novel phage associated with porcine ST403CC *C. jejuni* isolates however

this is highly unlikely due to the similarity to the phiX174 phage, given knowledge of the Illumina sequencing process, and the different type of prophage normally observed in *Campylobacter*.

This potentially problematic revelation however does not cause significant changes to the overall results or conclusions regarding prophage content in ST403CC *C. jejuni* isolates, as the potential F1/phiX174 region contains only phage associated CDS and would have no effect on the 'behaviour' of the isolates. Additionally the prophage region would not have served as a marker for ST403CC isolates as it was not observed in the non-porcine ST403CC *C. jejuni* isolate. Also it is established that the issue of contaminating phiX174 sequence is highly prevalent in published sequence data.

PHAST analysis initially identified potential prophage regions in each of the six porcine *C. jejuni* ST403CC isolates, however as described above further analysis revealed that the shared prophage region is most likely a remnant from the sequencing procedure. This leaves the remaining results wherein two isolates contain two potential prophage regions, two contain one potential prophage region, and two have no prophage content. As such it can be said that as in other *Campylobacter* isolates, prophage content may have a role in determining individual characteristics however it does not provide a mechanism for specialism or host adaptation in ST403CC *C. jejuni* isolates.

The initial prophages *Enterobacteria* phage Fels_2, *Haemophilus* phage SuMu, *Mannheimia* phage phiMHaA1, *Synechococcus* phage S-SKS1 each belong within the order Caudovirales, an order of single stranded DNA viruses containing three families; *Myoviridae*, *Siphoviridae*, and *Podoviridae*. Repeated PHAST runs using the updated database were carried out during the analysis of phiX174 content, from which the same regions were determined as being prophage but recorded new higher scoring matches; each of the prophage regions had homology with *Campylobacter* phages NCTC12673 or CP30A. Phage CP30A is not currently published in literature but was sequenced by Connerton, Siringan & Cummings in 2012 as part of a study into recombination and diversity of bacteriophages in

Campylobacter from chickens, whilst phage NCTC12673 is one of the virulent 'typing' bacteriophage as described in Sails *et al* (1998), as this is a type III virulent *Campylobacter* phage it is not likely that this phage is integrated into the genome of the ST403CC *C. jejuni* isolates, rather it serves as confirmation that these are Caudovirales bacteriophages, probably of the family *Myoviridae*.

PHAST searches against a phage database, and although it is the most up to date and comprehensive phage database currently available this is still somewhat of a limitation; it cannot find additional prophage content which has not already been described and added to the database, and when it does describe a match it can only provide the most similar relative which it currently has. This means that although a prophage match was described as '*Mannheimia* phiMHaA1' for example, this does not confirm the presence of this phage, rather it serves as an indicator of homology with this phage. It may be more beneficial to consider the order or family of phage observed rather than the individual phages described.

The 'real' prophages were all of order *Caudovirales* (old matches and new matches) so it is acceptable to state that four of six porcine ST403CC contained one or two potential Caudovirales order prophages. This is 'acceptable' as these are known to be the most common *Campylobacter* lysogens (Connerton, Timms & Connerton, 2011; Sails *et al*, 1998). Additionally this relates to the observed overlaps and similarity between the different phage regions observed; as described previously regions of the 'SuMu' phage were almost identical to regions of the 'Fels_2' phage in other isolates, and so on, again this implies homology and relatedness of prophage regions.

To provide comparisons for PHAST reliability, control runs were carried out for *C. jejuni* isolates 11168 and RM1221; these showed concurrence with literature such as Fouts *et al* (2005) in that zero prophage regions were observed for 11168 and four regions were observed for RM1221, however only three of the four RM1221 integrated regions are thought to be prophage related so this highlights some weakness in PHAST.

5.4.3 AlienHunter Integrated Genomic Content

Having considered the information from PHAST analysis the final dataset to consider was that produced from AlienHunter (Vernikos & Parkhill, 2006) analysis. AlienHunter searches for integrated genomic elements and recorded considerable numbers of predicted integrated regions across the ST403CC *C. jejuni* genomes (45-82 regions). Due to the lack of published data using AlienHunter for studying *Campylobacter* comparisons were provided using *C. jejuni* RM1221 (52 regions), 11168 (24 regions) and 81116 (29 regions); the number of regions predicted by AlienHunter for ST403CC isolates (45-82) were comparable to, albeit higher than, those for well-known *C. jejuni* isolates RM1221, 11168 and 81116.

Considering the association between PHAST and Alienhunter results, as expected, most phage regions identified by PHAST were also recognised by AlienHunter, along with a number of additional non-prophage integrated regions. As PHAST searches only for prophages with matches in its database, any novel prophages which might be overlooked by PHAST might well be picked up by AlienHunter, plus the presence of other types of integrated regions including insertion sequences and genomic islands. It was noted that some prophage regions predicted by PHAST were not identified as integrated regions by Alienhunter. As AlienHunter searches for specific patterns and for divergence from the host genome whereas PHAST directly compares sequence to a database of virus sequences, AlienHunter can overlook integrated regions which have prophage homology if they do not appear significantly distinct from the host chromosome. This can occur when an integrated region has been within the host sequence for an extended period of time, as it can begin to look more like that host genome, and degradation may disrupt target sequences so that they can no longer be recognised (Vernikos & Parkhill, 2006; Lawrence & Ochman, 1997).

In order to provide perspective for the observed results data from the literature must be considered. When the AlienHunter software was first developed and published it was used on *Salmonella* Typhi genomes, and was able to accurately predict all of the known genomic islands and integrated prophages, as well as some

previously unidentified island regions – it was successful, and more accurate than previous methods. They observed 17 *Salmonella* pathogenicity islands (three of which were newly discovered by this method), and five bacteriophages; these numbers of integrated regions were considerably lower than the numbers observed here in ST403CC C*. jejuni*, however *C. jejuni* is known to be highly competent and readily undertake the uptake of DNA from the environment through natural transformation (Gaasbeek *et al*, 2009), whereas *Salmonella* is not naturally competent (Ferguson, Heinemann & Kennedy, 2002) therefore it is logical to encounter considerably more integrated regions in these isolates.

5.4.4 Integrated Genomic Content in *Campylobacter*

As described previously, the major work published regarding *C. jejuni* integrated genomic regions focuses on comparisons between and against isolates 11168 and RM1221; such as that by Parker *et al* (2006) which investigated the prevalence of integrated elements and genomic diversity across a total of 79 *C. jejuni* and *C. coli* isolates, however they compared solely against the four integrated elements found in *C. jejuni* RM1221 , so provided no information on other additional integrated genomic regions, however they were able to demonstrate that integrated regions with some homology to RM1221 integrated elements were common across *C. jejuni* and *C. coli*, and not distinct to a single isolate or group of isolates. Parker *et al* (2006) demonstrated that genes carried by lysogenic bacteriophages play a role in genomic diversity in *Campylobacters*.

Other studies on the effects of integrated regions in *C. jejuni* have also been based upon the integrated elements in *C. jejuni* RM1221, including work which suggested strains carrying CJIE1 may be slightly more virulent *in vitro* than those without (Clark *et al*, 2012) and studies considering the impact of integrated elements on further genomic variation. It has been established that the uptake of exogenous DNA (transformation) is a method used by some bacterial species to increase genetic diversity. *C. jejuni* is naturally competent for transformation, however it has been demonstrated that some strains lack this ability (Gaasbeek *et al*, 2009). In 2009, Gaasbeek *et al* demonstrated using knockout and complementation that the gene

dns, a periplasmic DNase found in CJIE1, inhibits natural transformation in *C. jejuni,* and also identified other CDS contained within CJIE which were associated with non-naturally competent *C. jejuni* strains. Further to this, in 2010 Gaasbeek *et al* demonstrated similar roles for DNases encoded on CJIE2 and CJIE4. This is of interest as a homologue of *dns* was observed within prophage regions of the four ST403CC *C. jejuni* isolates carrying predicted prophages, suggesting that these isolates may lack natural competence. It is possible that secreted DNAase enzymes are protecting the ST403CC *C. jejuni* isolates from the uptake of integrated genomic regions. A standard test to confirm the action of deoxyribonulease activity could be carried out using DNAse agar (such as Sigma Aldrich item number D2560), however the production of DNAase would not prevent all uptake of new genetic information.

5.4.5 General Discussion & Conclusions

Some general conclusions can be drawn from this chapter. Firstly, it was observed that the apparent presence of *intA* did not appear to show any association with *Campylobacter* species, multilocus sequence type, or host source. Despite being described as a prophage integrase, the *intA* homologue observed in ST403CC *C. jejuni* isolates was not found to be located near a prophage or other integrated region. As stated previously (section 5.4.1), it is most likely that this CDS was annotated inaccurately due to the limitations of annotation relying on the accuracy of previous annotations, and may in fact be a 30S ribosomal protein, however there remains a slight possibility that this region represents a historical cryptic prophage.

Although the presence of *intA* (*CJ857_00835*) was not an indicator of integrated genomic content, considerable potential integrated regions were observed across the ST403CC *C. jejuni* genomes. Foreign DNA was found to be present in ST403CC *C. jejuni* isolates, and like other strains, may play a role in adaptation and evolution, however no evidence was observed of prophage regions being implicated in specialisation of the ST403CC group.

Analysis of prophage content demonstrated that ST403CC *C. jejuni* isolates may, but do not always, contain Caudovirales prophages. These are the most common form of integrated phages observed in Campylobacters and as such suggest that in this

respect, the ST403CC *C. jejuni* isolates reflect the published data for non-ST403CC isolates. ST403CC *C. jejuni* are capable of gaining and losing prophage similarly to other *C. jejuni* isolates and may gain some advantage individually as with any other strain, however they do not share a common prophage which indicates niche adaptation. It may be stated that whatever pressure led to this group appearing distinct by MLST and whole genome core phylogeny, was not linked to prophages or genomic islands. Relating prophage and integrated genomic content to whole genome based core phylogeny; no significant links were observed, there was no association between branch/ancestry of the ST403CC isolates and the number or 'type' of phage, nor the number of AlienHunter regions.

The overall conclusion regarding the prevalence and influence of integrated regions on the genomes of ST403CC *C. jejuni* isolates therefore is that although they may contain integrated content, there is no evidence of a specific adaptive event, adding further evidence that these isolates are generalist strains, rather than being specifically host adapted. This chapter has suggested that the integration of 'foreign' DNA may potentially play a role in the adaptation of ST403CC *C. jejuni* isolates, as it may in other non-ST403CC *C. jejuni* isolates; however the results observed provided no evidence for specific integrated content leading to specialisation in these isolates. It has also raised the possibility that some ST403CC isolates may be non-naturally competent isolates, which may have an impact in their ability to share genomic information and may go some way to explain their apparent isolation at the whole genome phylogeny level. The next stage in the project will partly carry on this idea of reduced genomic exchange in ST403CC *C. jejuni* isolates, by returning to the outcomes of previous chapters and considering the role and effect of potential new restriction modification systems in ST403CC *C. jejuni* strains.

Chapter Six: Recombination Events in ST403CC *Campylobacter jejuni* Isolates

6.1 Introduction

Previous work identified a group of isolates which were closely related by MLST, and were thought to potentially represent a host-adapted group of C. jejuni isolates (Manning *et al*, 2003). Work presented in this thesis has demonstrated that a selection of these isolates are potentially pathogenic (Chapter Two) and are closely related at the whole genome level (Chapter Three). Investigations were carried out to attempt to discover evidence for host adaptation based upon genome content. Despite the feature of the porcine host being more commonly associated with *C. coli*, the ST403CC *C. jejuni* did not show evidence of higher sharing of genome content with *C. coli* compared to *C. jejuni* isolates from more typical sources such as chicken and human infections. Identified during analysis of genome content were a group of potential restriction/modification system enzymes associated with ST403CC isolates which did not have homologues across the included *C. jejuni* and *C. coli* reference genomes. Given the role of r/m systems and the apparent lack of sharing with *C. coli* determined by investigating whole genes; this chapter will consider the rate of (homologous) recombination within ST403CC *C. jejuni* isolates compared to other *C. coli* and *C. jejuni* isolates.

6.1.1 Uptake of Genetic Content

Three means exist which allow bacteria to acquire and assimilate new genetic content: transduction, conjugation, and transformation (Gilbreath *et al*, 2011). Transduction and conjugation both rely on an 'infectious' donor providing DNA to the recipient cell, the donor being phage or plasmid, respectively (Redfield, 2001), whereas transformation is the integration of loose genetic content available in the environment (Krüger & Stingl, 2011) which results in exchanges of sequence content within a localised environment (Feil & Spratt, 2001) due to the recipient cell acquiring parts of the DNA of other cells within a closed space which have been broken open.

6.1.2 Homologous Recombination

Homologous recombination plays a significant role in bacterial adaptation, with the gain of new DNA providing one of the major forces driving bacterial evolution (Feil & Spratt, 2001; Gilbreath *et al*, 2011; Tang *et al*, 2009; Krüger & Stingl, 2011) and is particularly significant in species which experience selection pressures due to exposure to antimicrobials or the need to evade the host immune system (Feil *et al*, 2000; Suerbaum *et al*, 2001). Neutral variations often occur via recombination (Spratt & Maiden, 1999; Feil *et al*, 2000); that is, variation in sequence within conserved genes such as those necessary for essential metabolic processes, as such, MLST scheme 'housekeeping' genes are a useful target for recombination studies and have been the basis for numerous studies.

6.1.3 Recombination in *Campylobacter*

The extent to which recombination occurs in *C. jejuni* is debated (Fearnhead *et al*, 2005; Suerbaum *et al*, 2001; Wilson *et al*, 2009; Yu *et al*, 2012), but is considered to be between 2-8 times more than the levels of point mutation. Wilson *et al*, 2009 estimated that recombination was occurring at twice the rate of mutation (2.67 times more), but that despite this the changes are subjected to a 'purifying selection' which 'purges 60% of novel variation'. Yu *et al* (2012) observed much higher proportion of recombination than mutation (between 2.97 and 8.91 times), although both papers use MLST sequences for recombination analysis, Yu *et al* (2012) studied only single locus variants (SLVs - MLSTs which share six out of seven alleles). Fearnhead and colleagues (2005) however estimated that recombination was occurring at a similar rate to mutation using MLST data for farm and wildlife isolates of *C. jejuni* using an approximate likelihood model. They also concluded that small recombination fragments are much more likely than rare instances of complete genes being transferred by recombination. Suerbaum *et al* (2001) also studied MLST alleles, but used the HOMOPLASY method (as described by Maynard Smith & Smith, 1998), and concluded that *C. jejuni* frequently undergoes intraspecies recombination - with homoplasy ratios between 0.36 and 0.48 for five assessed fragments (where 0 signifies entirely clonal and 1 signifies free

recombination). Each of these examples uses MLST allele sequence data as the basis for recombination analysis, however here it was possible to use whole genome sequence to investigate recombination frequency in potentially host adapted ST403CC *C. jejuni* isolates.

6.1.4 Methods for Revealing Recombination

BAPS (Bayesian Analysis of Population Structure) is a statistical method for predicting structure of a given population of sequences, first introduced by Corander, Waldmann & Sillanpää (2003) and subsequently revised (Corander *et al,* 2008) to have improved computational efficiency and allow for analysis of larger datasets, and again to allow more complex analysis including MLST classification (Cheng *et al*, 2011). Observing recombination and population structure is particularly challenging for genomes which undergo large amount of Horizontal Gene Transfer. The most recent advance in the BAPS software is hierarchical BAPS, which uncovers nested populations structures from multiple sequence alignments (Cheng *et al*, 2013). BRATNextGen (Bayesian Recombination Tracker) (Marttinen *et al*, 2008) is part of the BAPS software package, which determines recombination events for large datasets of heterogeneous populations within closely related sequences.

Recent publications to make use of BAPS and BRAT include De Been *et al*'s study of *Enterococcus faecium* (2013), McNally *et al*'s work with *E. coli* (2013), Tang *et al*'s 2009 *Neisseria* study and Willems *et al*'s *Enterococcus faecium* paper (2012). As an example, De Been *et al* (2013) used BRATNextGen as described by Marttinen *et al* (2012) to investigate the role of recombination in the evolution of *Enterococcus faecium*. The Enterococci are gut commensal organisms, but have played an increasing role in opportunistic disease in recent decades. Initially opportunistic *Enterococcus* infection was mostly associated with *E. faecalis*, but in the last twenty years *E. faecium* cases have increased in prevalence and is now considered as important as *E. faecalis*; this rise in *E. faecium* infection is associated with its ability to readily acquire new genetic elements, particular antibiotic resistance genes. BRAT analysis revealed that recombination was playing a significant role in the

evolution of *E. faecium*; specifically they observed an increase in recombination within an important hospital associated sub population compared to commensal isolates, but also saw that 'modern' clinical isolates show a reduction in recombination. This concurs with McNally *et al* (2013) and Willems *et al* (2012) in the idea that new clinical strains can arise through acquisition of advantageous genes, and subsequently become a more clonal group with less recombination as a result of new characteristics.

6.1.5 Aims

The aims of this chapter were

- To investigate the level of recombination in ST403CC *C. jejuni* isolates compared to that in other *C. jejuni* and *C. coli* isolates.
- To determine whether recombination events are shared between ST403CC *C. jejuni* from pigs, and *C. coli* porcine isolates: although Chapter Four didn't identify significant sharing between *C. coli* isolates and the porcine ST403CC *C. jejuni* isolates based upon coding regions/whole genes it was of interest whether homologous recombination events may exist which represent a marker for host adaptation and admixture with the typical porcine occupier, *C. coli*.
- To consider recombination rates in ST403CC *C. jejuni* isolates in relation to the whole genome phylogeny determined in Chapter Three, and also to consider the content of any observed recombinant regions, with reference to the integrated sequence elements revealed in Chapter Five.

In this chapter, frequency of recombination events and genetic admixture is determined for whole genome sequences of ST403CC *C. jejuni* isolates using BAPS 6.2 software package, and compared to non ST403CC *C. jejuni* and *C. coli* strains, as described in previous chapters.

6.2 Methods

The recombinant status of ST403CC *C. jejuni* isolates was determined, and compared to other *C. jejuni* and *C. coli* isolates, using the isolates previously

described for use in genetic analysis (Table 3.1; Appendix 9.1), and the additional ST403CC isolate ATCC33560.

6.2.1 Revealing Recombination

Recombination events and genetic admixture were investigated to reveal concealed population structure not exposed by other methods.

6.2.1.1 Core Genome Alignment for Recombination Analysis

Core genome alignment was produced for thirty-four *C. jejuni* and *C. coli* isolates, using Mugsy (Angiuoli & Salzberg, 2010) as described previously (Chapter Three).

6.2.1.2 Heirarchical BAPS Completed by Jukka Corander & Colleagues at The University of Helsinki

Hierarchical BAPS was used with settings as described in Corander *et al* (2012) to create clustering results for multiple sequence alignment.

6.2.1.3 BRATNextGen Analysis Completed by Jukka Corander & Colleagues at The University of Helsinki

BRATNextGen was used with default settings as put forward by Marttinen and colleagues (2012) to estimate recombination events based on evolutionary lineage clusters determined using Bayesian analysis.

6.2.1.4 Phylogenetic Tree Based upon the Non-Recombinant Core Genome Completed by Jukka Corander & Colleagues at The University of Helsinki

A phylogenetic tree was produced, showing the relatedness of isolates based upon the remaining core genome alignment upon discarding the recombinant regions.

6.2.2 Investigating Content of Recombinant Regions

Further investigation was carried out to consider the content of recombinant regions in ST403CC *C. jejuni* isolates, and to investigate the presence of shared recombinant regions in ST403CC *C. jejuni* isolates, as well as quantifying

recombination in ST403CC *C. jejuni* compared to non-ST403CC *C. jejuni* and *C. coli* isolates.

6.2.2.1 Converting BRATNextGen Output Carried out by Simon Harris at the Sanger Centre

Post-processing was kindly carried out by Simon Harris at the Sanger Centre, producing BRAT recombination output file which was readable in Artemis, to allow visualisation of the location and content of recombinant regions.

6.2.2.2 Applying Statistics to Quantified Recombination Results

The frequency of recombinant events in *C. jejuni* and *C. coli* isolates were analysed; in order to determine the statistical significance of variation in recombination between groups - ST403CC *C. jejuni*, non-ST403CC *C. jejuni*, and *C. coli* using T-test to determine whether variation between groups was statistically significant, or likely to be due to chance. Statistical analysis was used to test the Null Hypothesis: there will be no significant difference in recombinant frequency between groups. Specifically no significant difference in the number of recombination events will be observed between ST403CC *C. jejuni* and non-ST403CC *C. jejuni*.

The standard T-test used previously throughout this thesis (section 2.2.7) was not appropriate for the recombination frequency data due to the assumption of Normal distribution in this method. The values in the samples (recombination frequency per group) do not indicate normal distribution, and do not have distributions of the same shape, therefore the Mann-Whitney T-test was selected as an appropriate measure, using Log values.

6.2.2.3 Recombinant Region Content

Once the Artemis readable file was kindly produced by Simon Harris (Sanger Centre), Artemis sequence viewer (Rutherford *et al*, 2000) was used to investigate the recombinant regions of the ST403CC *C. jejuni* isolates. The output of the core genome alignment was a multiple genome fastA file; from this the feature representing the core aligned genome sequence of ST403CC *C. jejuni* 857 was selected and saved as a new fastA file; onto which the complete annotation of *C.*

jejuni 857 was transferred using RATT (Otto *et al*, 2011), producing an annotated core genome region for the isolate; into this was read the recombinant region annotation produced by Simon Harris, providing an annotated core genome revealing recombinant regions in ST403CC *C. jejuni* isolate 857. This was then used to investigate regions identified by BRATNextGen analysis which were unique or otherwise interesting in ST403CC *C. jejuni* isolates, and identify coding regions in these areas.

6.2.2.4 Investigating the Function of CDS in Recombinant Regions

BLAST searching was implemented in order to ascertain potential function and relatedness of coding sequences. As described in Chapter Four, both the built-in NCBI BLAST function in Artemis was used to identify related sequences in the global database, and local database BLAST searching was applied to investigate sequence homology across the specific thirty-four included genomes.

6.3 Results

The recombination and genetic admixture of *C. jejuni* and *C. coli* isolates was assessed, with the aim to reveal the recombinant profile of ST403CC *C. jejuni* isolates; a group of closely related and potentially host adapted strains. This section comprises three sub-sections: the first summarising Hierarchical BAPS and BRATNextGen output, followed by the quantification of recombination events, and subsequently content analysis of recombination regions.

6.3.1 Genetic Admixture & Recombination in ST403CC *Campylobacter jejuni*

The degree of genetic admixture was determined for thirty-four *C. jejuni* and *C. coli* isolates, based upon their core genome alignment. Hierarchical BAPS grouped isolates into clusters of related sequences, with dashes representing regions of variation between related isolates. The number and location of recombination events were assessed using BRATNextGen for the same group of thirty-four isolates; with assessment of the source of the recombinant content. Figures 6.1 and 6.2 were produced by Jukka Corander & colleagues at the University of Helsinki, with annotation added by the author.

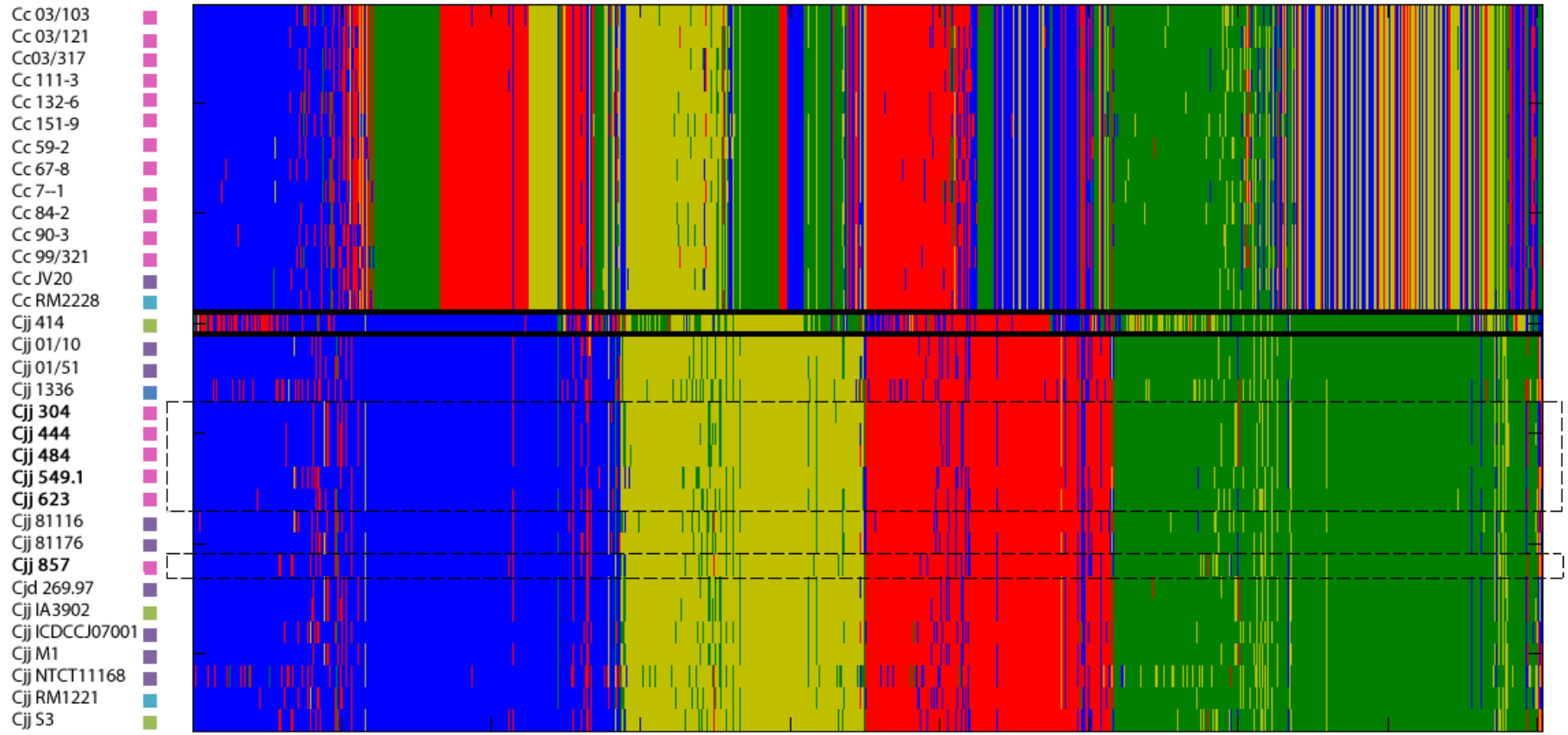

Figure 6.1: Admixture of *C. jejuni* & *C. coli* Isolates Revealed by Hierarchical BAPS

Heirarchical BAPS was run by J. Corander & Team at the University of Helsinki, producing this output image. Annotation was added by L. Morley at the Nottingham Trent University

Each row represents an isolate: isolates are clustered according to level 1 BAPS, separated by black horizontal lines. (BAPS level 1 = *C. coli*; level 2 = *C. jejuni jejuni* 414; level 3 = *C. jejuni jejuni*/*C. jejuni doylei* Hierarchical BAPS produces vertical lines to indicate 'SNPs', however, due to the size of the alignment each pixel in fact corresponds to numerous SNPs.

Figure 6.1 shows hierarchical BAPS output. Each row represents an isolate and isolates are clustered by level 1 BAPS, with clusters separated by black horizontal lines. Hierarchical BAPS analysis grouped all *C. coli* isolates into one cluster, *C. jejuni jejuni* 414 as a separate cluster of its own, and the remaining *C. jejuni jejuni* and *C. jejuni doylei* strains, including the ST403CC isolates, as a third cluster. Hierarchical BAPS produces vertical lines to indicate 'SNPs', however, due to the size of the alignment each pixel in fact corresponds to numerous SNPs.

High levels of admixture were observed in some *C. jejuni* isolates, including 11168 and 1336 in particular. The overview produced by BAPS analysis also suggests that, in fact, lower levels of admixture seem to be observed in ST403CC *C. jejuni* isolates compared to the other included *C. jejuni* genomes

Figure 6.2, overleaf, shows visual output from BRATNextGen analysis. No significant recombination events were observed for the included *C. coli* isolates, or for *C. jejuni* isolate 414 (the highly niche adapted bank vole isolate); although *C. jejuni* otherwise shows recombination events. ST403CC strains showed a reduced number of significant recombination events compared to other *C. jejuni* isolates, and exhibited a distinct 'recombination profile' including three recombination events consistent across all 7 isolates. Recombination events which originate from *C. coli* appear red in Figure 6.2, whilst blue segments are from *C. jejuni* sources, and turquoise are predicted to be from other external sources. It was observed that *C. coli* sequence is incorporated into all recombining *C. jejuni* isolates, however there was no evidence of increased acquisition of *C. coli* sequence within the ST403CC group

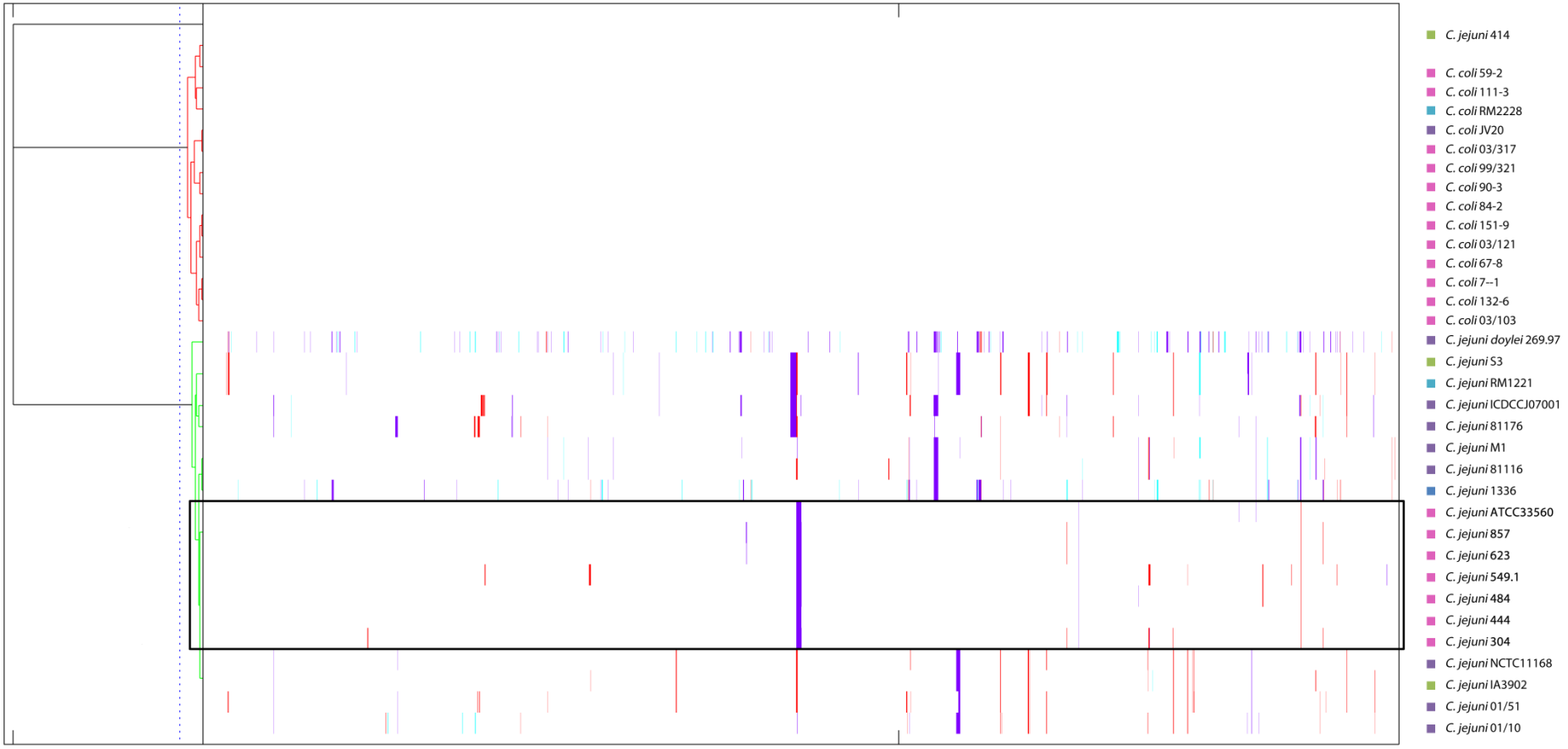

Figure 6.2: Recombination Events in *C. jejuni* & *C. coli* Isolates Revealed by BRATNextGen Analysis

BRATNextGen was run by J. Corander & Team at the University of Helsinki, producing this output image. Annotation was added by L. Morley at the Nottingham Trent University.

Each row represents an isolate, as indicated on the right. Each vertical line represents a recombination event; red lines indicate import from *C. coli*; blue indicates import from *C. jejuni*. ST403CC isolates are highlighted with bold text, and enclosed in black box.

Note: C. jejuni strains are *C. jejuni jejuni* unless specified otherwise.

Figure 6.3 depicts a phylogenetic tree based upon the remaining core genome alignment data, following removal of recombinant regions. This tree is similar in appearance to the previously shown phylogenetic trees, the overall population structure is not noticably altered.

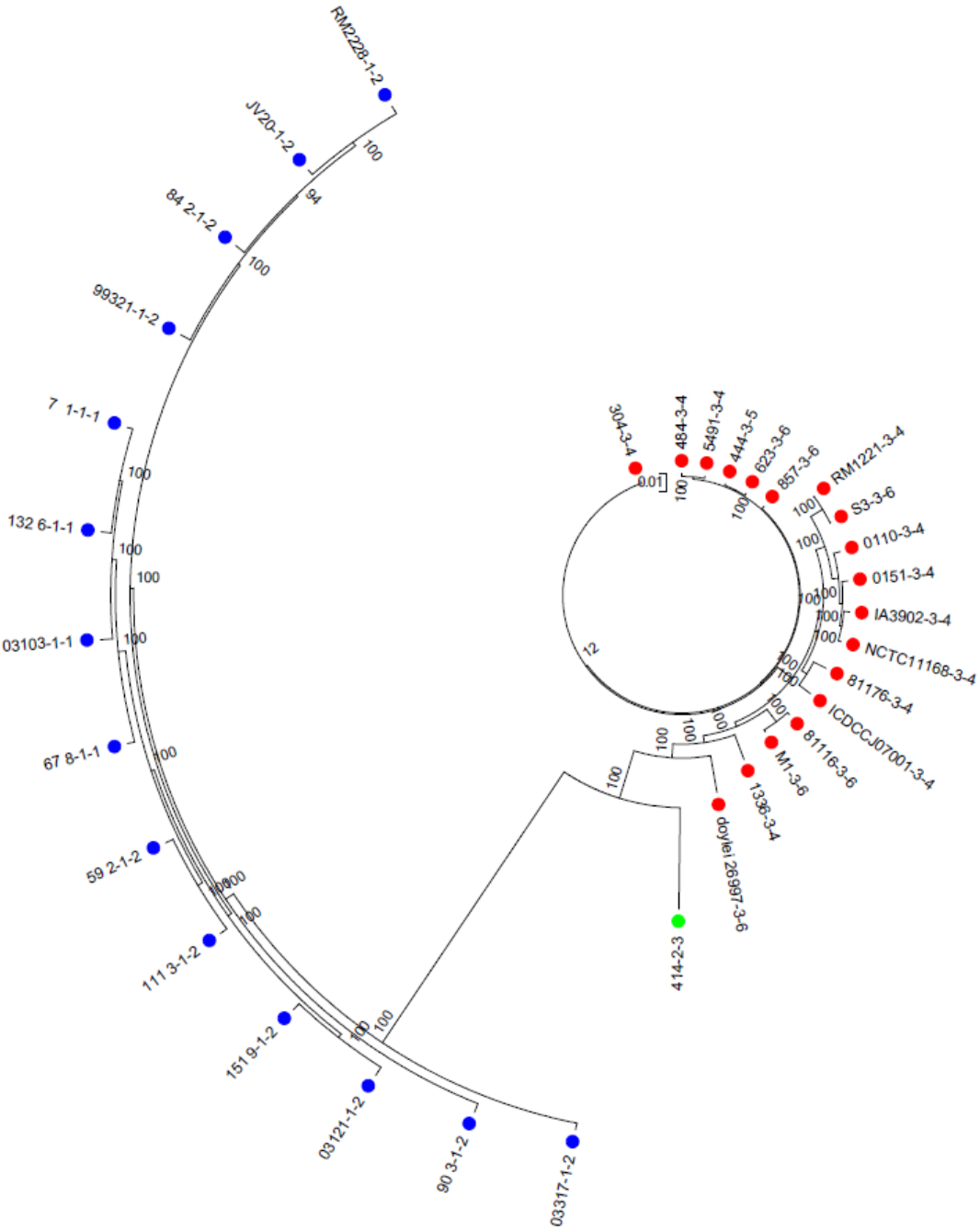

Figure 6.3: Phylogenetic Tree of *C. jejuni* & *C. coli* Isolates based upon Non-Recombinant Core Genome Regions

Phylogenetic tree based upon core genome alignment following removal of recombinant regions, including colour coding relating to BAPS categories. This tree was produced by J. Corander at the University of Helsinki

Figure 6.3 was produced by Jukka Corander & colleagues at the University of Helsinki.

Initial assessment of BAPS and BRATNextGen results revealed that ST403CC *C. jejuni* isolates group within other non-ST403CC *C. jejuni* isolates based upon genetic admixture, and confirmed that the ST403CC *C. jejuni* isolates do not exhibit an increase in the uptake of *C. coli* associated genetic content, as revealed based upon coding sequence analysis in Chapter Four. It was visible that the ST403CC *C. jejuni* isolates contained fewer recombinant regions than their non-ST403CC *C. jejuni* relatives, but more than was observed for the included *C. coli* isolates. To draw further information from these results, quantification was carried out to reveal the differences in recombination in ST403CC and non-ST403CC *C. jejuni* and *C. coli*.

6.3.2 Quantification of Recombination Events

Results are presented on the size and frequency of recombinant regions in ST403CC and non-ST403CC *C. jejuni* and *C. coli* isolates.

The percentage of the core genome made up of recombination fragments was calculated as shown in Table 6.1. The percentage of the core genome made up of recombination fragments in non-ST403CC *C. jejuni* (0.017) was more than double that of the ST403CC isolates (0.007). This means that, generally, each recombinant region in the studied ST403CC *C. jejuni* isolates was shorter in nucleotide length (mean 5923 nucleotides) than the recombinant regions observed in non-ST403CC *C. jejuni* (mean 14372 nucleotides) isolates, and as such, represent a smaller proportion of the core genome. The standard deviation observed for non-ST403CC *C. jejuni* isolates was also higher than that of the ST403CC *C. jejuni*, suggesting that the smaller size of recombinant fragments was comparatively consistent in ST403CC *C. jejuni* isolates in contrast to higher degree of variation in length of recombinant regions in non-ST403CC *C. jejuni* genomes. This may suggest a lesser degree of genetic diversity in ST403CC *C. jejuni* than other *C. jejuni* isolates, although still higher than that observed in *C. coli*.

Group	Mean Length of Recombination	Standard Deviation Length of Recombination	Mean Percentage of Recombination	Standard Deviation Percentage of Recombination
C. coli	0	0	0	0
ST403CC *C. jejuni*	5923	1639.2	0.007	0.002
Non-ST403CC *C. jejuni*	14372	8860.4	0.017	0.0103

Table 6.1: Percentage of Core Genome Comprised of Recombinant Fragments

The percentage of the core genome which consisted of recombinant fragments was calculated from the mean scores for each group; *C. coli*, ST403CC *C. jejuni* and non-ST403CC *C. jejuni*

The percentage of (the total number of) SNPs (across the whole genome) which occur within recombinant regions calculated as shown in Table 6.2. The percentage of SNPs across the whole genome which occur within recombination fragments was very similar for the ST403CC *C. jejuni* and non-ST403CC *C. jejuni*, however, the raw values show a considerable reduction of SNP frequency in ST403CC *C. jejuni* compared to non-ST403CC *C. jejuni*.

Group	Total Number of SNPs in Recombination Fragments (R)	Total Number of SNPs in Non-Recombinant Core Genome (M)	Percentage of SNPs in Core Genome contained within Recombinant Regions
C. coli	0	32993	0.00
ST403CC *C. jejuni*	1104	6161	17.92
Non-ST403CC *C. jejuni*	14421	81633	17.67
All Strains	19115	162325	11.78

Table 6.2: Percentage of SNPs Observed in Recombinant Regions

The percentage of the SNPs across the core genome which were contained within recombinant regions was calculated for *C. coli* isolates, ST403CC *C. jejuni* isolates and for non-ST403CC *C. jejuni* isolates, based upon the mean number of SNPs for each group.

Range and mean values of recombinant event frequencies are shown in Table 6.3. No significant recombination events were observed for any *C. coli* isolate, or for *C. jejuni* 414. The non-ST403CC *C. jejuni* recombination frequency ranged from 0-200 events (mean 48.5), however, this includes two notable outliers – *C. jejuni jejuni* 414 and *C. jejuni doylei* - with zero and 200 recombination events respectively. Discarding these two outlier values, the range for recombination frequencies in non-ST403CC *C. jejuni* was 27-95 (mean 39.1), compared to the range for the ST403CC isolates of 4-15 events (mean 9.9).

Group	Range	Mean	Standard Deviation
C. coli	0	0	0
Non-ST403CC *C. jejuni*	0-200	49	50
Non-ST403CC *C.jejuni**	27-95	39	21
ST403CC *C. jejuni*	4-15	10	4

Table 6.3: Quantification of Recombinant Events in *C. jejuni* & *C. coli* Genomes

The range and mean of the number of recombination events observed in each group of isolates studied; *C. coli*, ST403CC *C. jejuni* and non-ST403CC *C. jejuni*.

*Non-ST403CC *C. jejuni* isolates excluding extraneous strains *C. jejuni jejuni* 414 and *C. jejuni doylei 269*.97

Figure 6.4 provides a visual representation of the values in Table 6.3; using the mean values of recombination frequencies as determined by BRATNextGen analysis, by category, with standard deviation.

To confirm the statistical significance of the difference in recombination event frequency between groups, the Mann-Whitney T-test was used, with Log values - however this cannot be applied to zero values, and was not therefore suitable for comparison of the *C. coli* results, or the outlier *C. jejuni* isolate 414. The high recombination outlier *C. jejuni doylei* 269.97 was also excluded from statistical analysis as described in Table 6.3. The Mann-Whitney T-test (two-tailed, unpaired, nonparametric means, based upon Log values) confirmed that the difference in

recombination event frequency between ST403CC *C. jejuni* and non-ST403CC *C. jejuni* (excluding outlier values *C. jejuni* 414 and *C. jejuni doylei* 269.97) was statistically significant (P<0.0001).

These results indicate that the Null hypothesis can be rejected - significant variation was observed between ST403CC *C. jejuni* and non-ST403CC *C. jejuni*, with ST403CC isolates displaying statistically significantly fewer recombination events

As stated previously, zero significant recombination events were observed in *C. coli*. As such, this data was unsuitable for analysis based upon Log values, therefore the Mann-Whitney T-test was run on the raw values to provide an indication of significance. Significant difference in variation was observed between *C. coli* and ST403CC *C. jejuni* (P<0.0001), as well as between *C. coli* and non-ST403CC *C. jejuni* (P <0.0001), however these results were based upon raw values and are therefore less dependable; although the difference between these groups was clearly visible as shown in Figure 6.4 and Figure 6.2.

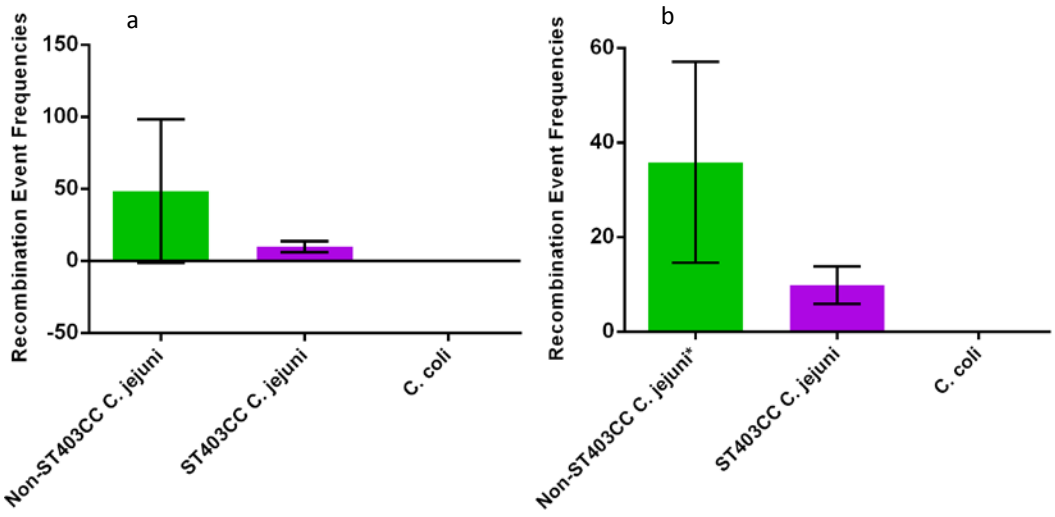

Figure 6.4: Mean Frequences of Recombination Events in *C. jejuni* & *C. coli* Isolates
6.4a shows the mean values and standard deviation for recombination events for *C. coli*, non-ST403CC *C. jejuni* and ST403CC *C. jejuni*. Large standard deviation gave rise to big error bars due to two 'unusual' *C. jejuni* strains; *C. jejuni* 414, which showed no significant recombination events; and *C. jejuni* 269.97, which is a different subspecies (*C. jejuni doylei*) and showed a very high rate of recombination.
6.4b shows the mean values and standard deviation for recombination events for *C. coli*, non-ST403CC *C. jejuni* and ST403CC *C. jejuni* after discarding the values for the *C. jejuni* 414 and 269.97.

6.3.3 Content of Recombinant Regions in ST403CC *Campylobacter jejuni*

BRATNextGen analysis led to the recognition of three recombinant regions which are common and identical or almost identical across the seven ST403CC *C. jejuni* isolates (Table 6.4). One of which occurs at a site with no neighbouring recombination events in non-ST403CC *C. jejuni* isolates, whilst the other two are located nearby to dissimilar recombination events in other *C. jejuni* isolates. These may represent 'recombination hotspots' in the genome, where recombination occurs readily, as recombination occurs near this site in the majority of studied *C. jejuni* isolates, although the recombination event observed in the ST403CC *C. jejuni* isolates is distinct from those taking place in non-ST403CC isolates and is uniform across the ST403CC isolate group.

Proceeding from the beginning of the alignment, the first of the three ST403CC common recombination events occurs at 426229-429980bp (426229-429957bp in *C. jejuni* 444 and ATCC33560; 426226-429980 in *C. jejuni* 623). A neighbouring recombination event exists in *C. jejuni* isolates 81-176, RM1221 and S3 (426035-427187bp), however this originates from a different source; origin 1 in these non-ST403CC isolates, origin 4 in the ST403CC isolates. Prokka annotation for this region indicated the presence of *bioA* (Adenosylmethionine-8-amino-7-oxononanoate aminotransferase) and *carB* (carbamoyl phosphate synthase large subunit), as shown in Figure 6.5**.** BLASTx search for this complete 3752bp region confirmed similarity with *bioA* and *carB* genes. BLASTx searching also confirmed matches for these genes in the non-ST403CC isolates, revealing that this recombination event does not signify gain of different coding sequence. Additionally, local BLASTn searches were carried out as described in Chapter Four to confirm sequence similarity between the ST403CC *C. jejuni* isolates. The UniProt database (http://www.uniprot.org/) entry for *bioA* is made up of 455 amino acids, observed lengths in ST403CC isolates 250(304)-436(857) and for *carB* 1073bp, against 1079 in 857 and 817 in 304.

Region	Start	End	Origin	Isolate
Region 1	426229	429980	4	304
	426229	429957	4	444
	426229	429980	4	484
	426229	429980	4	5491
	426226	429980	4	623
	426229	429980	4	857
	426229	429957	4	ATCC33560
Region 2	629005	629206	4	304
	629005	629206	4	444
	629005	629206	4	484
	629005	629206	4	5491
	629005	629206	4	623
	629005	629206	4	857
	629005	629214	4	ATCC33560
Region 3	788994	789264	1	304
	788994	789264	1	444
	788994	789264	1	484
	789000	789264	1	5491
	789000	789264	1	623
	789000	789264	1	857
	789000	789264	1	ATCC33560

Table 6.4

Recombination Events Common across ST403CC *C. jejuni*

The second ST403CC common recombination event was located 629005-629206bp (629005-629214 in ATCC33560) and only 202-210 base pairs in length. PROKKA (Seemann, 2014) annotation did not reveal any coding sequence within this region,

however investigation via Artemis (Figure 6.6) revealed that this region is in fact contained within a larger coding sequence (628938-630212c), which was revealed by BLASTx search to encode a hydrolase, with database matches in *C. jejuni* 87459 and *C. coli* H8, 1417 and others.

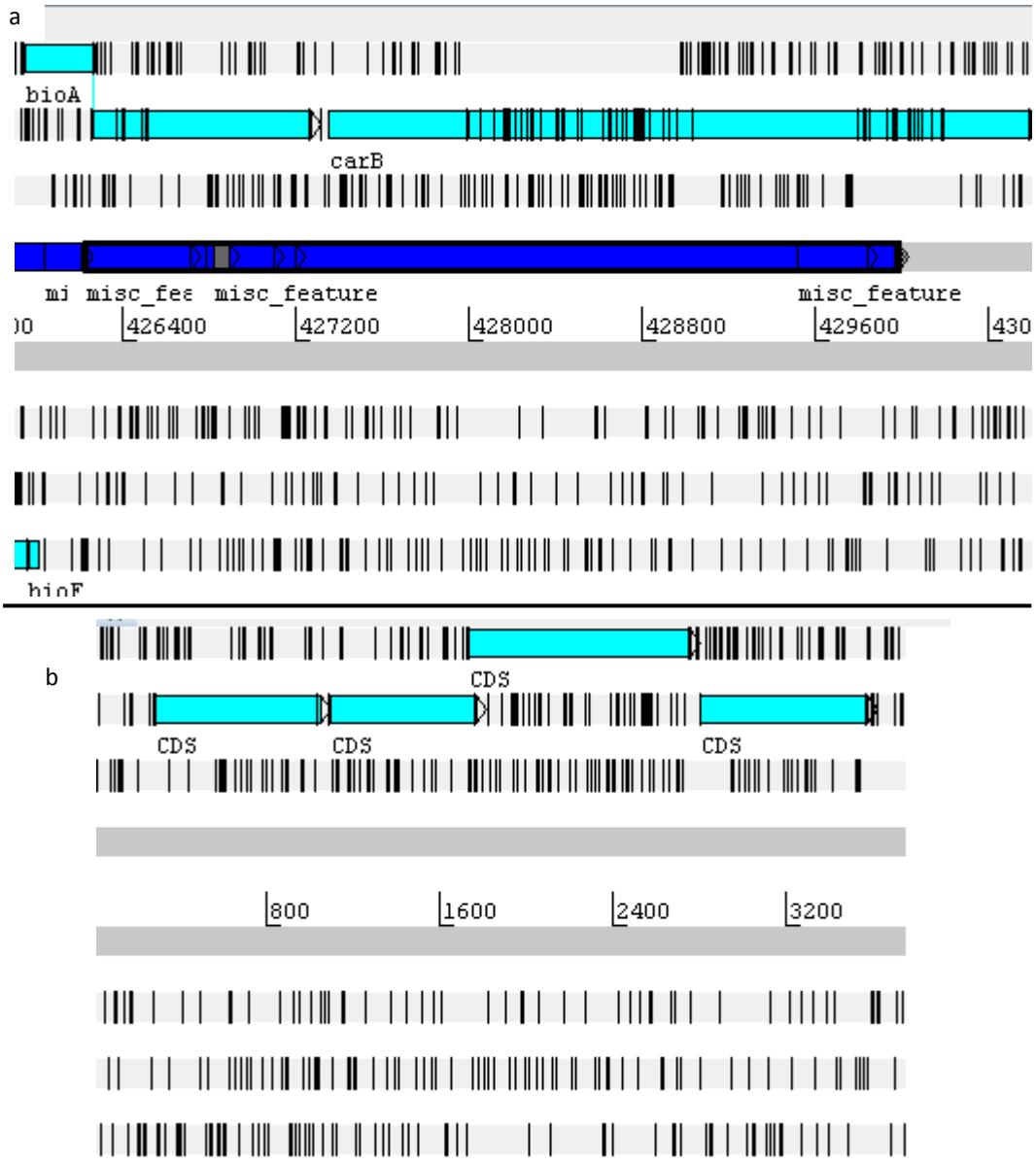

Figure 6.5: Visualisation of the First Recombinant Region Common to the ST403CC *C. jejuni* Isolates

6.5a: as Annotated for strain 857
6.5b: as seen in strain 304

Local BLASTn searches of the recombination fragment revealed homology in all *C. coli* included in study, and no matches in non-ST403CC *C. jejuni*, however the same

search queried with the full coding sequence (1275bp in *C. jejuni* 857) revealed homology in all included reference genomes.

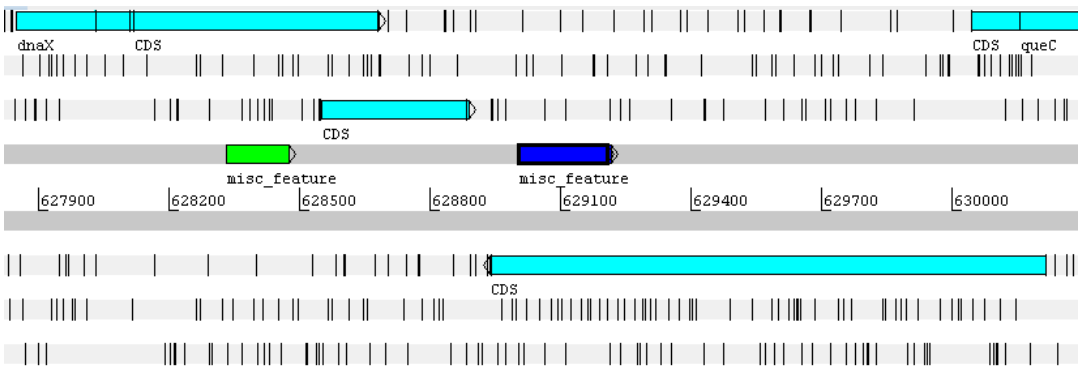

Figure 6.6: Visualisation of the Second Recombinant Region Common to theST403CC *C. jejuni* Isolates The blue 'misc feature' shown in this image represents the recombinant event in ST403CC *C. jejuni*, whilst on thecomplementary strand at this location is a larger coding sequence.

The third ST403CC common recombination event (789000-789264; 788994-789264 in 304, 444, 484), as with the first, is directly neighbouring other recombination events in non-ST403CC *C. jejuni* isolates from different origins, however in this instance the ST403CC fragment comes from origin 1 and is neighboured on both sides by non-ST403CC recombination events. The non-ST403CC isolate recombinant regions downstream are from origin 4, however the upstream recombinant region,

Figure 6.7: Visualisation of the Third Recombinant Region Common to the ST403CC *C. jejuni* Isolates The highlighted grey 'misc feature' includes the ST403CC recombination event, but also overlaps with a recombination event in 81116 and M1.

present in *C. jejuni* 81116 and M1 are from origin 1. The 265-271bp recombinant region is contained within coding sequence CJ857_01161 (788484-789326; 843bp), as identified via Prokka annotation.

BLASTx search for the recombinant region showed homology with coding sequences identified as nucleotidases (in *C. coli* H56, 1948, 1091), flagellar protein (in *C. coli* LMG 9860) or PseG (pseudominic acid biosynthesis associated protein G in *C. coli* JV20). Local BLASTn searches revealed that the both recombinant region and the full coding sequence CJ857_01161 had homology in all reference genomes.

6.4 Discussion

ST403CC *C. jejuni* isolates have previously been demonstrated to be closely related based upon MLST and whole genome sequence alignment. The six sequenced ST403CC isolates were associated with the porcine host, and have been demonstrated to be closely related to a historical, geographical and host distinct additional ST403CC *C. jejuni* isolate ATCC33560. The six studied porcine ST403CC *C. jejuni* isolates have also been demonstrated to be potentially capable of causing disease in humans, and have been demonstrated based upon study of the pan genome not to have acquired considerable *C. coli* associated content, despite colonising a host typically associated with *C. coli*. The work carried out in this chapter revealed that ST403CC *C. jejuni* isolates do not contain increased admixture with *C. coli* within the core genome, but do exhibit a distinct profile of reduced recombination compared to non-ST403CC *C. jejuni* isolates.

6.4.1 Heirarchical BAPS & BRATNextGen Investigation of Admixture & Recombination

Hierarchical BAPS results confirm that again the ST403CC *C. jejuni* isolates group within the 'normal' *C. jejuni* isolates. The degree of admixture varies across the six isolates, but they appear middle to low in terms of 'rank', compared to high admixture in NCTC11168, 1336. Hierarchical BAPS shows a large degree of uniform admixture within the *C. coli* cluster. This reflects previous work (Kinana *et al*, 2007; Sheppard *et al,* 2008) as the included *C. coli* isolates are all either confirmed or

suspected Clade 1 *C. coli,* which are known to share sequence both with *C. jejuni* and *C. coli*, yet are also a highly clonal population.

BRATNextGen is conservative in estimating recombination events, so the results observed may not represent every recombination event which has taken place (De Been *et al*, 2013) but do provide a useful overview of significant recombination events. Also, although the previously identified restriction/modification system associated proteins found in the ST403CC *C. jejuni* isolates (Chapter Four) were not located within identified recombinant regions, it could be that these are part of small recombinant regions representing a false negative in BRATNextGen results.

6.4.2 Recombination Frequency

Recombination frequency in ST403CC *C. jejuni* isolates was found to be significantly lower than in non-ST403CC *C. jejuni*, and significantly higher than in *C. coli*. The lack of significant recombination events observed for *C. coli* is in line with previous research which observed lower levels of recombination in *C. coli* than *C. jejuni* (Yu *et al*, 2012; Sheppard *et al* 2008); Yu *et al* (2012) observed that recombination gave rise to around seven times more SLV generation than does mutation in *C. jejuni*, whereas in *C. coli* both mutation and recombination play a similar level. The level of recombination observed for non-ST403CC *C. jejuni* is as could be expected – frequent evidence of recombination events, with considerable variation between strains. *C. jejuni* 414 was unusual in that it showed no evidence of significant recombination events, however this is justifiable, as it is a specialist niche isolate which survives in an isolated host (*Myodes glareolus*, the bank vole), it is known to have lost regions which are common to more 'typical' *C. jejuni* isolates and although it is also described as having gained new genetic information to survive in the bank vole host (Hepworth *et al*, 2011), these will form part of the accessory genome and therefore would not be observed in the BRAT analysis.

6.4.3 Content of Recombinant Regions in ST403CC *Campylobacter jejuni*

Although 'region 1' showed that both ST403CC and non-ST403CC isolates possessed *bioA* and *carB* genes, the ST403CC isolates are lacking several other '*bio'* genes, or they are so disparate that they weren't identified as part of the core genome

alignment. ST403CC isolates contain both *bioF* and *bioA*, although whether *bioA* is likely to be functional may be uncertain, as it shows possible frameshift. *carB* likewise may have a frameshift – annotation using PROKKA identified it as being entirely in the second reading frame for 857, despite lots of potential stop codons, but examining empty reading frames such as in 304 region 1 suggest it may involve a frameshift to frame 1. It seems possible that the recombination event may not have a functional change (although considering the potential frameshift, it may), but it may serve as a 'marker' as it is consistent across the ST403CC isolates. It was also observed that when searching the public BLAST database the closest matches to these genes in the ST403CC *C. jejuni* were seen in *C. coli* isolates; suggesting that these genes are more like the *C. coli bioA* and *carB* genes.

The second recombinant region common to the ST403CC *C. jejuni* isolates was the small recombination segment (202-210bp) was confirmed by local database BLAST searching as present in these *C.* jejuni and not in the included non-ST403CC *C. jejuni*, although lower scoring matches were observed in all included *C. coli*. The ORF which the recombination fragment is revealed to be contained within however (length in 857 of 1275bp) has homologues in all 34 isolates. This suggests that this potential hydrolase is present in 'all' thermotolerant *Campylobacter*, but that a different 'fingerprint' is observed in *C. coli* and in ST403CC *C. jejuni*.

The third recombinant region, similar to the first also appeared to be a distinct recombination event in ST403CC *C. jejuni* isolates which does not appear to have a significant effect on the coding sequence.

6.4.4 Overall Discussion & Conclusions

ST403CC *C. jejuni* isolates show a significant reduction in the number of recombination events compared to non ST403CC *C. jejuni* isolates, and have little import of *C. coli* associated DNA. This is accompanied by the acquisition of four restriction/modification enzymes (as described in Chapter Four). R/M systems 'protect' the strain from the uptake of new genetic information and may, theoretically, explain the lack of recombination in the ST403CC *C. jejuni* isolates.

However, it is worth noting that the lack of observed recombination in the *C. coli* isolates appears to be in contradiction with what has been observed in the literature. Sheppard *et al* (2013) have observed genome wide introgression in *C. coli* isolates from ST828CC; taking up large amounts of *C. jejuni* DNA. Lefébure *et al* (2010) observed no recombination or evidence of introgression in *C. coli*, however this was due to their included sample containing only clade one isolates (no non-agricultural isolates) and their applied method being affected by this bias; as all of the included isolates may have contained the same introgression signature, the analysis wouldn't necessarily have been able to 'filter' the event and identify it. It may be possible that the included BRATNextGen analysis suffered false negatives in not identifying recombination events in the *C. coli* isolates, however it is more probable that the recombination occurring would be part of the accessory genome and therefore not included in the analysis based upon the core genome alignment used in this technique.

The observation of significantly reduced recombination in ST403CC *C. jejuni* isolates disproves the initial supposition that these isolates may exhibit increased uptake of *C. coli* content, however this in itself represents a way in which the ST403CC *C. jejuni* isolates emulate the behaviour of *C. coli* via the reduction in recombination.

Chapter Seven: Discussion

7.1 Summary

Previous work identified, by MLST, a clonal complex group of *C. jejuni* isolates which were considered a potentially host adapted group of strains (Manning *et al*, 2003). For the purpose of this thesis, a selection of these isolates representing six different sequence types within the clonal complex (ST403CC) were studied for evidence of this potential host adaptation.

It was hypothesised that these ST403CC *C. jejuni* isolates, being associated with the porcine host more commonly associated with *C. coli*, may exhibit increased genetic recombination with *C. coli*, and at an extreme may represent an intermediary group or 'fuzzy species' within the boundary between *C. jejuni* and *C. coli*. It was also suggested that the ST403CC *C. jejuni* isolates might demonstrate a distinct phenotype as a result of adapting to survive the porcine host, such as exhibiting differential ability to survive environmental stressors or internalise into human cells in cell culture experiments. As such, the work carried out in this thesis aimed to investigate evidence of niche adaptation in these isolates through assessment of phenotypic characteristics; whole genome phylogeny; core and pan genome content; genetic variation; and genomic recombination.

The initial ST403CC isolates which inspired this research were recovered by Manning *et al* (2003) and were associated primarily with pigs and with cows to a lesser extent; with no ST403CC isolates associated with human infection in the UK during that study. At a similar timepoint, Duim *et al* (2003) found an association with ST403CC *C. jejuni* and human gastroenteritis in Curaçao. A recent overview of the recorded isolates of ST403CC *C. jejuni* on the PubMLST database (Jolley & Maiden, 2010: http://pubmlst.org/campylobacter/) is shown in Figure 7.1 overleaf: the ST403 sequence type, and the ST403CC clonal complex have been isolated from food mammals, dogs, and frequently from human cases, but have not been recovered from poultry.

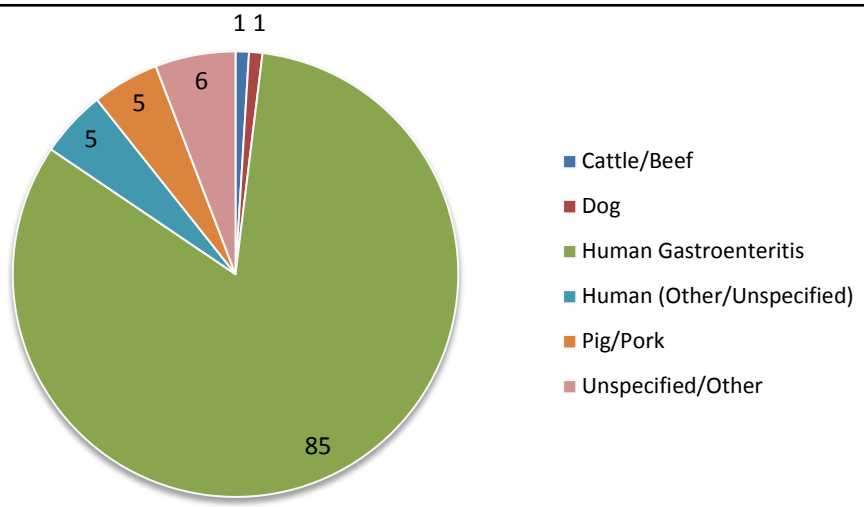

Figure 7.1a

Sources of Sequence Type ST403 isolates as reported on PubMLST (total 103 isolates recorded).

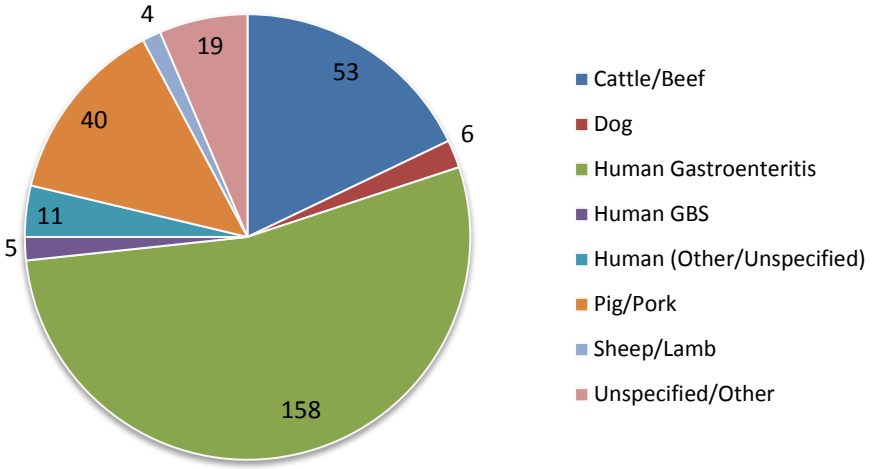

Figure 7.1b

Sources of all ST403 Clonal Complex isolates (sharing four or more of seven loci with ST403) as reported on PubMLST (Total 296 isolates recorded).

There exist many STs within the ST403CC which are represented by only one isolate, and therefore only one reported source. Of those with multiple isolates in a Sequence Type, some such as ST270 and ST403 represent a variety of source types, whilst others seem specific to a single host type or source. ST933 represents the latter with the majority of isolates associated with cattle and beef products, and a small number of human gastroenteritis cases. Others such as ST5408 have only

been isolated from a single source type (in this case pigs and pork products) and not yet linked with human cases.

The work presented in this thesis demonstrated that the selected ST403CC *C. jejuni* isolates did not display a distinct pattern of phenotypic behaviour when compared to other *C. jejuni* and *C. coli* isolates, rather giving equivalent results to previously characterised *C. jejuni*; with the typical strain-strain variation observed within these species, however it was demonstrated that these isolates were capable of motility and host cell invasion *in vitro* – adhering to and internalising into human colonic Caco-2 cells – indicating that these isolates, given the opportunity would potentially be capable of causing human illness. It was also confirmed that the isolates are closely related based upon whole genome phylogeny, but do exhibit between strain variation in genetic content. The ST403CC *C. jejuni* isolates do not exhibit increased sharing of CDS with *C. coli* compared to non-ST403CC *C. jejuni* isolates, and contain also only a small number of CDS shared with none of the included *C. jejuni* and *C. coli* isolates in the group-pan genome analysis carried out. The coding sequence content of ST403CC *C. jejuni* isolates was largely 'normal' for *C. jejuni*, however some potential prophage content was found which was common to all six included ST403CC *C. jejuni* isolates. Predicted prophage content was also observed which represented part of the variation between isolates in the ST403 Clonal Complex.

As described in Chapter Six, the ST403CC *C. jejuni* isolates were demonstrated to exhibit a significant reduction in the rate of recombination occurring, compared to non-ST403CC *C. jejuni* isolates, which was accompanied by and possibly explained by, the acquisition of four restriction modification enzymes which are not common in sequenced *C. jejuni* - in the local database searches using 34 *C. jejuni* and *C. coli* genomes, three of the four CDS had homologues only in the seven ST403CC *C. jejuni* isolates, whilst the fourth had an additional homologue in a porcine *C. coli* isolate - and may 'protect' the isolates from the uptake of new genetic information. Although the ST403CC *C. jejuni* isolates did not show an increased homology with *C. coli* isolates, the reduced recombination profile observed reflects a genetic similarity to *C. coli*. *C. coli* isolates undergo lower recombination frequencies than *C. jejuni*, as demonstrated by BRATNextGen in Chapter Six and also in the literature

(Sheppard *et al* 2008; Yu *et al*, 2012). It may be postulated that the ST403CC *C. jejuni* isolates are in fact 'behaving' more like *C. coli* than non-ST403CC *C. jejuni* isolates due to this significant reduction in recombination.

7.2 Related ST403CC Isolate *Campylobacter jejuni* ATCC33560

As described in Chapter Four NCBI blast searching coding regions from the ST403CC *C. jejuni* isolates revealed frequent high scoring matches with a *C. jejuni* isolate ATCC33560 – which was revealed to be a geographically and temporally distinct ST403CC isolate (ST403) from a bovine faecal sample. This isolate was therefore included in subsequent analyses, to investigate whether the 'effects' of ST403CC over-ride the effects of host, time and country. The additional ST403CC isolate was found to be largely similar to the six porcine ST403CC *C. jejuni* isolates, with the exception of its lack of prophage content as revealed in Chapter Five. It is noted however that the lack of prophage content in ATCC33560 does not necessarily signify a difference in ability to acquire prophage; it may be that due to the different time, location or host of isolation the isolate may not have encountered bacteriophage to integrate. It is also possible that there is some difference in ST403CC isolates associated with cattle not encountering or not acquiring phage content, however further research would be required on additional isolates in order to reveal this. As shown below (Figure 7.2) *C. jejuni* ATCC33560 clusters with the other ST403CC *C. jejuni* isolates based upon whole genome phylogeny, but is visibly distinct.

C. jejuni ATCC33560 is the quality control isolate for *C. jejuni* antimicrobial sensitivity testing, and was used to demonstrate the influence of the CmeR regulator by Hyytiäinen & Hänninen (2012).

In addition, as described in Chapter Six, *C. jejuni* ATCC33560 also fits the reduced recombination profile of the ST403CC *C. jejuni* isolates, suggesting that the unusual characteristics of ST403CC isolates are not specific to porcine isolates but related to some other linking factor within the clonal complex.

7.3 ST403CC *Campylobacter jejuni* Isolates in the Literature

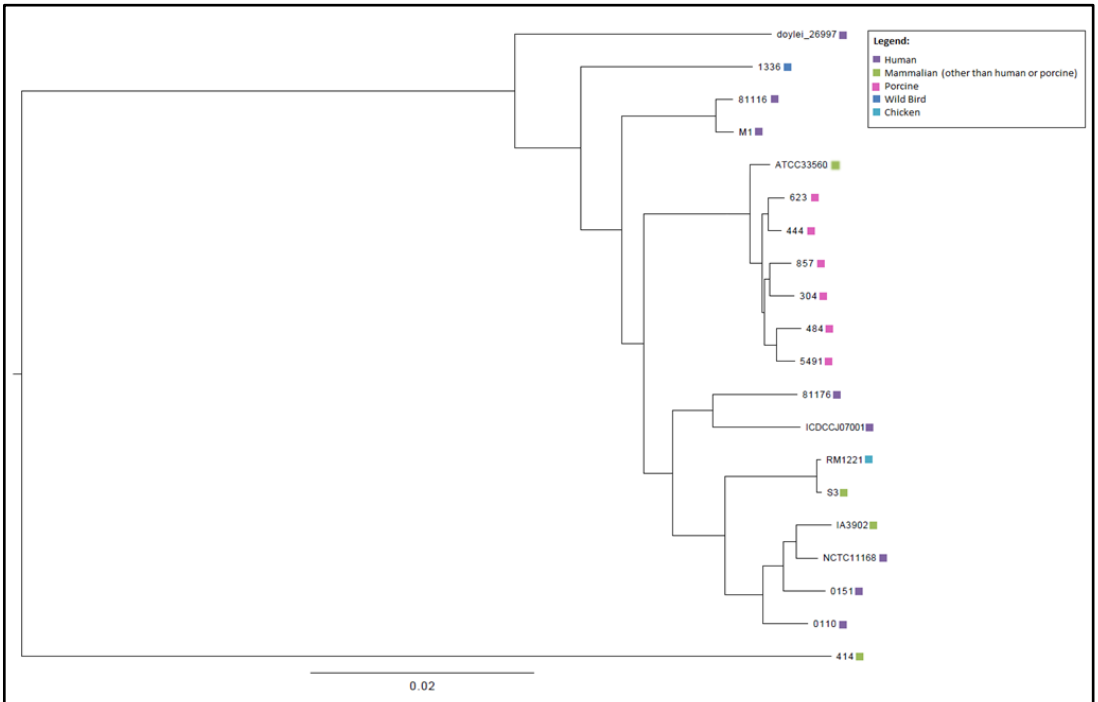

Figure 7.2

Whole genome alignment phylogeny tree for 20 *C. jejuni* isolates, including the additional ST403CC *C. jejuni* isolate ATCC33560

The ST403 clonal complex was primarily associated with pigs, and also with cattle to a seemingly lesser extent (Manning *et al*, 2003) however isolates of this clonal complex have increasingly been identified from a range of sources, and have also been associated with human clinical infections.

In an investigation into an increased burden of *Campylobacteriosis* and GBS in Curaçao, Duim *et al* (2003) identified 41 isolates (21% of the isolates fully characterised in the study) within the ST403 clonal complex, which were associated with a temporal association with November-February (although this is the wet season there and has the hightest levels of *Campylobacter* infection in general) and were all Sequence Type 403. Duim *et al* (2003) also noted difficulty in AFLP typing ST403CC isolates, possibly linked to four adenine specific methyltransferases known to be produced in *C. jejuni* (11168; Parkhill *et al*, 2000), as methylation can block restriction by *Hind*III – one of the AFLP enzymes – however there was no explanation of whether this was responsible, or why this effect might be particularly

pronounced in these isolates. This may potentially be associated with results observed in Chapter Four; additional methyltransferases were identified in the six sequenced ST403CC isolates, perhaps these have a role in preventing restriction by *Hind*III – although this is just postulation which has not been demonstrated it represents a possible explanation.

ST403CC isolates were also recovered from dogs during a study by Parsons *et al* (2009) where they demonstrated that a relatively low proportion of domestic dogs may carry *C. jejuni* and may in fact acquire *C. jejuni* infection from the same sources as humans, they still represent a risk for zoonotic *C. jejuni* infection.

Additionally, ST403CC *C. jejuni* isolates have recently been associated with raw bulk milk in Italian dairy herds (Bianchini *et al*, 2014). Bianchini *et al* (2014) also recorded the presence of ST21CC and ST48CC isolates which are known 'generalist' *C. jejuni* lineages (ST21, Gripp *et al*, 2011; ST48, Sheppard *et al*, 2009). ST21CC may be described as a 'true generalist' *C. jejuni* lineage. The research by Gripp *et al* (2011) demonstrated high genomic diversity linked with large scale recombination and acquisition of prophage content. The ST48CC was associated with a range of host types in a survey by Sheppard *et al* (2009).

Grove-White *et al* (2011) surveyed cattle and sheep in Lancashire and found some association between ST21CC and both cattle (16.7%) and sheep (13.2%), ST48 linked with sheep (13.2% and only in 2.9% of cattle isolates) and ST403CC with cattle (16.7%, compared to 2.8% in sheep).

ST403CC isolates were associated with a winter peak in a six year study by Cody *et al* (2012) and were generally ciprofloxacin sensitive. Cody *et al* (2012) described ST403CC isolates as being generally associated with a wide range of hosts and typically not with poultry. The winter peak of ST403CC recovery observed by Cody *et al* (2012) also correlates with the wet season peak observed by Duim *et al* (2003), whilst contradicting the usual summer seasonal peak for *C. jejuni* in the UK. It may be proposed that there could be an unrevealed link between certain clonal complexes of *C. jejuni* and the wetter seasons, however this may not be linked with mammalian hosts, and the cause of the association is not yet clear. Both the

ST403CC and ST353CC had winter (wet) peaks during a 6 year longitudinal study in Oxfordshire in the UK; whereas the ST45, ST283 and ST42 Clonal complexes exhibited the typical summer peak of *C. jejuni* in the UK (Cody *et al*, 2012). Both the ST45CC and ST283CC are chicken associated clonal complexes, however ST42CC is commonly associated with ruminants; and although ST403CC is associated with a range of hosts excluding poultry, the ST353CC is associated with chicken - thereby there is little suggestion that this link is associated with a difference between poultry and non-poultry isolates (Cody *et al*, 2012). Interestingly, the ST353CC identified as a wet season associated complex by Cody *et al* (2012), was also one of the ten major clonal complexes identified by Duim *et al* (2003) in Curaçao alongside the ST403CC, where it also matched the rainy season peak.

Sheppard *et al* (2009) revealed an association with ST403CC and cattle and sheep. Also, although they found the majority of clinical isolates included were related to chicken associated genotypes, they did observe a small proportion of clinical isolates from the ST403CC.

ST403CC *C. jejuni* isolates have even been associated with GBS cases (Islam *et al*, 2009) during a survey in Bangladesh, representing five of the ten GBS related isolates included in their study. Islam *et al* (2009) observed an association between ST403CC and Penner serotype HS:23 in their study on GBS and enteritis isolates. Of the six sequenced ST403CC isolates used in this research, in previous study, four were deemed untypeable, however isolates 484 (PS484) and 444 (PS444) were designated HS:23 (Manning *et al*, 2003; Frost et al 1998). Across the sixteen porcine ST403CC isolates in total included in Manning *et al* (2003), 9 were untypeable, and a total of four serotypes were observed across the remainder (23 (n=4), 35 (n=1), 29 (n=1), 22 (n=1)).

It may be that, rather than representing a host adapted isolate as first predicted, the ST403 clonal complex may represent a more generalist lineage, with a bias towards mammalian hosts. Works published in the literature also link ST403CC isolates with enteric infection and even with GBS so it is revealed that these may represent a threat lurking within the meat production industry.

7.4 Gene Content of ST403CC *Campylobacter jejuni* Isolates

The annotated genomes produced were subsequently used to return to the regions of variance identified in Chapter Three (Figure 3.9) which were found to contain largely hypothetical proteins, unfortunately investigation of these regions was not highly informative. Although one of the divergent regions in 623 contained *flgE*, *dnaA*, *dnaN*, *gyrB* – however it is difficult to accurately identify exact locations from BRIG so this may be directly upstream of the variant region.

Raphael *et al* (2005) described the gene *cbrR* (*Campylobacter* bile resistance regulator) in *C. jejuni* 11168 (coding region *Cj0643*). This gene was shown to provide resistance to sodium deoxycholate; mutants were highly sensitive and the complemented mutant had restored resistance. *cbrR* mutants were also less able to colonise in the chick model. In the thirty four genomes included in analysis, a *cbrR* homologue was identified in all included *C. jejuni* strains and none of the *C. coli* isolates. The match for the niche adapted isolate *C. jejuni* 414 was much lower however suggesting the gene may be degraded. All of the ST403CC *C. jejuni* contained a *cbrR* homologue, which supports the level of sodium deoxycholate sensitivity observed in Chapter Two, as they are seemingly able to respond to and survive sodium deoxycholate stress.

Karlyshev *et al* (2014) characterised a previously unstudied peptidase in *C. jejuni* 11168, *Cj0511*. They confirmed that this was a serine peptidase and demonstrated that mutants in *Cj0511* were poor colonisers in the chicken model. A BLASTn search of the local database containing thirty four *C. jejuni* and *C. coli* genomes, as described in Chapter Four revealed that a high-scoring *Cj0511* homologue was present in all included *C. jejuni* isolates, with approximately 50% similarity to homologues also found in each included *C. coli* isolate. This supports their conclusion that this may play a role in colonisation, as all tested *C. jejuni* strains contained a *Cj0511* homologue and were able to invade host cells in cell culture experiments. However this cannot directly be compared as chicken colonisation experiments were not done in this work.

Champion *et al* (2005) identified a group of CDS which were seemingly linked with livestock strains of *C. jejuni* (*Cj1321-Cj1326* in *C. jejuni* 11168). Subsequently, during the re-annotation of *C. jejuni* 11168 (Gundogdu *et al*, 2007) the CDS Cj1325 and Cj1326 were joined to represent a single gene split across two reading frames (designated as *Cj1325*). Homologues of *Cj1321-1323* from *C. jejuni* 11168 were observed in local database BLASTn searching in the two hyper invasive human *C. jejuni* isolates (01/10 and 01/51), the human microbiome isolate *C. coli* JV20 and seven porcine *C. coli* isolates, whereas *Cj1324* and *Cj1325* had homologues in the same ten previously described as well as in the bank vole adapted *C. jejuni* isolate 414, the human blood isolate *C. jejuni doylei* 269.97, the sheep isolate *C. jejuni* IA3902 and the seven ST403CC *C. jejuni* isolates. These results largely fit with those observed by Champion *et al* (2005), as these isolates are mostly livestock associated, the notable exceptions being human clinical or commensal isolates; these may likely be explained by the fact that the probable source of the infections originated from a livestock source. In the case of *C. coli* JV20 having been associated with the human gut microbiome, this seemingly suggests that this is a pathogenic isolate however it could possibly be suggested that this is a potential pathogen which had either not reached a high enough dose level to create symptoms, or was post-illness and not fully cleared, or represents an example of adaptive immunity.

7.5 Host Adaptation & Host Generalists

Bronowski, James & Winstanley (2014) describe that *C. jejuni* must survive oxygen stress in different environments - and that *Campylobacter* uses catalase rather than SOD (superoxide dismutase) as in most bacteria. The *katA* gene is the major catalase in *C. jejuni* however another has been recently identified - *Cj1386* downstream of katA. A brief check using local database blast searching revealed that a *katA* homologue was observed in all thirty four included genomes, with greater homology in *C. jejuni* than *C. coli*. Homology matches were also observed for all thirty four genomes for *Cj1386* however scores were much lower (and discarded) for all *C. coli* isolates (approximately 50% homology score, and higher E. values).

As described in the introductory chapter (1.7 Host Association) in 2005, French *et al* noticed a prevalence of *uncA*17 allele in ST61CC cattle isolates. In this study *uncA*17 was found in two of the four newly sequenced porcine *C. coli* isolates. All six newly sequenced porcine ST403CC *C. jejuni* isolates contained *uncA*7; this allele is also found in some ST61CC isolates.

Sheppard *et al* (2014) demonstrated that recombination does not occur in the natural host environment between two of the major generalist *C. jejuni* lineages, despite considerable opportunity in common hosts. Sheppard *et al* (2014) propose that the barrier to recombination between these lineages is not a physical prevention - as they can recombine easily in the lab - there may be some unclear ecological niche reason preventing recombination between these two lineages.

In this thesis, it has been attempted to discern evidence of host adaptation within a potential niche adapted group of *C. jejuni*; this has been evidenced in some lineages, but is not always a discernable feature (Méric *et al*, 2014; Sheppard *et al*, 2014). It may be postulated that seemingly generalist lineages simply have not had a long enough time period to adapt to a given host, and although this may be the case in some examples, this has also been refuted in studies such as that by Sheppard *et al* (2014).

Méric *et al* (2014) used a similar method to the approach taken here in Chapter Four, of filtering through the pan genome content to identify genes associated with species or sub-group of isolates, however Méric *et al* (2014) have developed a new technique to identify these genes.

Based upon seven *C. jejuni* and *C. coli* isolates Méric *et al* (2014) observed a pan genome of 3933 CDS, with a core genome of 1035 and accessory genome of 2792 CDS. This aligns with the results observed in this research; the core genome determined for thirty three included genomes was 591 CDS with a pan genome across the thirty three isolates of 5329 CDS. The core represented approximately 11% of the total pan genome, and the accessory genome (calculated simply as the size of the pan genome size minus the core genome size) was 4738 CDS. Whilst for the six ST403CC isolates alone, the core genome was 1444 CDS, with a pan genome

of 2220 CDS. The core represented approximately 65% of the pan genome, with an accessory genome of 776 CDS.

Méric *et al* (2014) used this seven genome reference pan genome to investigate genetic variation both between the two species and within specific lineages within the species. They did not identify specific CDS linked with specific lineages (clonal complexes) as they tended to also be present in some other smaller lineage. However they did observe some 'generalist' genes which were associated with the multiple host associated ST21CC and ST45CC, plus others which were related to host restricted groups (ST353CC and ST61CC). This falls in line with the results observed here also; few CDS were observed which were distinct to ST403CC isolates, however this is not surprising due to the close relationship between *C. jejuni* clonal complexes. Méric *et al* (2014) also observed gene flow between the generalist lineage ST21CC and *C. coli* clade one isolates, as has been previously described (Sheppard *et al* 2013; Sheppard *et al* 2008).

Yahara *et al* (2014) produced a new approach for identifying 'hotspots' of recombination in bacterial genomes. Using this they demonstrated three recombinant 'hotspots' in *C. jejuni* which were largely associated with membrane associated proteins. The work presented here supports the findings of Yahara *et al* (2014) in that the regions where the ST403CC *C. jejuni* isolates tended to vary from other *C. jejuni* were observed to be largely consisting of membrane associated proteins.

7.6 Temperature Effects

The normal body temperature range of the pig is 38.7-39.8°C, and in cattle 36.7-39.3°C (beef cattle 36.7-39.1°C; dairy cattle 38.0-39.3°C) compared to 40.6-43°C in chickens (as described in the Merck veterinary manual; http://www.merckmanuals.com/vet/appendixes/reference_guides/normal_rectal_temperature_ranges.html) and 37°C in humans. In the course of the phenotypic experiments carried out it was demonstrated that the ST403CC *C. jejuni* isolates grow successfully at 37°C and 42°C, however no specific experimentation was carried out in this region. It may be of future interest to carry out a

temperature growth curve experiment, to ascertain the optimum growth temperature of ST403CC *C. jejuni* isolates and determine whether it reflects the 42°C optimum typically associated with *C. jejuni*, and similar to the chicken host, or whether perhaps the optimum of these isolates may be closer to the normal body temperature of pigs or mammals in general.

7.7 Benefits & Limitations of Genomic Investigation

It is important to remember in this era of ever increasing genomic study, that genome alone cannot inform us of how the bacteria may be behaving in the host/ecosystem/even in the lab. A simple example that illustrates this is the issue of hippurate hydrolysis. As described early on (Chapter Two), hippurate hydrolysis was a traditional discriminator between *C. jejuni* and *C. coli* - wherein *C. coli* are negative and *C. jejuni* are positive, however in newer studies it was shown that some *C. jejuni*, including the isolates which provided the basis for this work, are phenotypically hippurate negative. Although as shown in Chapter Four these isolates do still maintain the *hipO* gene - for the isolates included in the analysis here it was shown that presence or absence of the *hipO* gene can separate the two species, however it does not inform upon which are able to express the enzyme. This is an issue which may be relevant in many genome based studies - it is important to remember that it may give us clues and associations but cannot give the whole picture.

7.8 Further Work, Future Perspectives & Conclusions

The work produced in this thesis has revealed that the potential porcine associated ST403CC *C. jejuni* isolates previously identified (Manning *et al*, 2003) may represent a mammalian niche associated lineage, and that this grouping appears to transcend temporal and geographical factors.

The ST403 Clonal Complex is associated with a significantly reduced rate of genomic recombination compared to other *C. jejuni* isolates from a range of sources and clonal complexes. However the low recombination rate is mirrored in *C. coli*, which undergo less genomic recombination than *C. jejuni*, and also in the highly niche

adapted *C. jejuni* isolate 414 which did not exhibit recombination; although it may be difficult to affirm whether this is due to genetic or environmental barriers.

If it were a feasible prospect, a future goal would be to acquire sequence data for additional ST403CC isolates, including human clinical isolates and a range of additional sources, in order to investigate whether the reduced recombination profile is common across the remainder of the group or whether it differs between human and 'mammalian' hosts. This would also provide a basis to further consider the potential role of prophages in this group of isolates.

The work presented here has provided new information on the effect of host adaptation on evolution and recombination in *C. jejuni*. It has been demonstrated that a group of ST403CC *C. jejuni* isolates exhibit a profile of reduced recombination in isolates from both pigs and cattle and may represent a mammalian associated specialist lineage of *C. jejuni*.

8: References

Alikhan, N.F.; Petty, N.K.; Zakour, N,L,B.; & Beatson, S.A. BLAST Ring Image Generator (BRIG): Simple Prokaryote Genome Comparisons. *BMC Genomics*, 2011, **12**; 402-412

Allos, B.M. *Campylobacter jejuni* Infections: Update on Emerging Issues & Trends. *Clinical Infectious Diseases*, 2001, **32**; 1201-1206

Altenhoff, A.M. & Dessimoz, C. Phylogenetic & Functional Assessment of Orthologs Inference Projects & Methods. *PLoS Computational Biology*, 2009, **5**; e1000262

Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W. & Lipman, D.J. Basic Local Alignment Tool. *Journal of Molecular Biology*, 1990, **215**; 403-410

Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W. & Lipman, D.J. Gapped BLAST & PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Research*, 1997, **25**; 3389-3402

Amri, A.A.; Senok, A.C.; Ismaeel, A.Y.; Al-Mahmeed, A.E. & Botta, G.A. Multiplex PCR for Direct Identification of *Campylobacter* spp. in Human & Chicken Stools. *Journal of Medical Microbiology*, 2007, **56**; 1350-1355

Angiuoli, S.V. & Salzberg, S.L. Mugsy: Fast Multiple Alignment of Closely Related Whole Genomes. Bioinformatics, 2010, **27**; 334-342

Aroori, S.V.; Cogen, T.A. & Humphrey, T.J. The Effect of Growth Temperature on the Pathogenicity of *Campylobacter*. *Current Topics in Microbiology & Immunology*, 2013, **67**; 333-340

Assefa, S.; Keane, T.M.; Otto, T.D.; Newbold, C. & Berriman, M. ABACAS: Algorithm-Based Automatic Contiguation of Assembled Sequences. *Bioinformatics*, 2009, **25**; 1968-1969

Atabay, H.I. & Corry, J.E.L. The Isolation & Prevalence of Campylobacters from Dairy Cattle using a Variety of Methods. *Journal of Applied Microbiology*, 1998, **84**; 733-740

Babakhani, F.K.; Bradley, G.A. & Joens, L.A. Newborn Piglet Model for Campylobacteriosis. *Infection & Immunity*, 1993, **61**; 3466-3475

Bacon, D.J.; Alm, R.A.; Burr, D.H.; Hu, L.; Kopecko, D.J.; Ewing, C.P.; Trust, T.J. & Guerry, P. Involvement of a Plasmid in Virulence of *Campylobacter jejuni* 81-176. *Infection & Immunity*, 2000, **68**; 4384-4390

Bacon, D.J.; Szymanski, C.M.; Burr, D.H.; Silver, R.P.; Alm, R.A. & Guerry, P. A Phase-Variable Capsule is Involved in Virulence of *Campylobacter jejuni* 81-176. *Molecular Microbiology*, 2001, **40**; 769-777

Baillon, M-L.A.; van Vliet, A.H.M.; Ketley, J.M.; Constantinidou, C. & Penn, C.W. An Iron-Regulated Alkyl Hydroperoxide Reductase (AhpC) Confers Aerotolerance & Oxidative Stress Resistance to the Microaerophilic Pathogen *Campylobacter jejuni*. *Journal of Bacteriology*, 1999, **181**; 4798-4804

Barton, C.; Ng, L-K.; Tyler, S.D. & Clask, C.G. Temperate Bacteriophages Affect Pulsed-Field Gel Electrophoresis Patterns of *Campylobacter jejuni*. *Journal of Clinical Microbiology*, 2007, **45**; 386-391

Bereswill, S.; Fischer, A.; Plickert, R.; Haag, L-M.; Otto, B.; Kühl, A.A.; Dashti, J.I.; Zautner, A.E.; Muñoz, M.; Loddenkemper, C.; Groß, U.; Göbel, U.B. & Heimesaat, M.M. Novel Murine

Infection Models Provide Deep Insights into the "Ménage à Trois" of *Campylobacter jejuni*, Microbiota & Host Innate Immunity. *PLoS One*, 2011, **6**; e20953

Best, E.L.; Fox, A.J.; Frost, J.A. & Bolton, F.J. Identification of *Campylobacter jejuni* Multilocus Sequence Type ST-21 Clonal Complex by Single-Nucleotide Polymorphism Analysis. *Journal of Clinical Microbiology*, 2004, **42**; 2836-2839

Best, E.L.; Fox, A.J.; Owen, R.J.; Cheesbrough, J. & Bolton, F.J. Specific Detection of *Campylobacter jejuni* from Faeces using Single Nucleotide Polymorphisms. *Epidemiology 7 Infection*, 2007, **135**; 839-846

Bianchini, V.; Borella, L.; Benedetti, V.; Parisi, A.; Miccolupo, A.; Santoro, E.; Recordati, C. & Luini, M. Prevalence in Bulk Tank Milk & Epidemiology of *Campylobacter jejuni* in Dairy Herds in Northern Italy. *Applied & Environmental Microbiology*, 2012, **80**; 1832-1837

Biggs, P.J.; Fearnhead, P.; Hotter, G.; Mohan, V.; Collins-Emerson, J.; Kwan, E.; Besser, T.E.; Cookson, A.; Carter, P.E. & French, N.P. Whole-Genome Comparison of Two *Campylobacter jejuni* Isolates of the Same Sequence Type Reveals Multiple Loci of Different Ancestral Lineage. *PLoS One*, 2011, **6**; e27121

Birge, E.A. *Bacterial & Bacteriophage Genetics*, 3rd Edition. Springer-Verlag

Blattner, F.R.; Plunkettm G.III.; Bloch, C.A.; Perna, N.T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J.D.; Rode, C.K.; Mayhew, G.F.; Gregor, J.; Davis, N.W.; Kirkpatrick, H.A.; Goeden, M.A.; Rose, D.J.; Mau, B. & Shao, Y. The Complete Genome Sequence of *Escherichia coli* K-12. *Science*, 1997, **277**; 1453-1462

Bliven, K.A. & Maurelli, A.T. Antivirulence Genes: Insights into Pathogen Evolution through Gene Loss. *Infection & Immunity*, 2012, **80**; 4061-4070

Blom, J.; Albaum, S.P.; Doppmeier, D.; Pühler, A.; Vorhölter, F-J.; Zakrzewski, M. & Goesmann, A. EDGAR: A Software Framework for the Comparative Analysis of Prokaryotic Genomes. *BMC Bioinformatics*, 2009, **10**; 154

Brás, A.M.; Chatterjee, S.; Wren, B.W.; Newell, D.G. & Ketley, J.M. A Novel *Campylobacter jejuni* Two-Component Regulatory System Important for Temperature-Dependent Growth & Colonisation. *Journal of Bacteriology*, 1999, **181**; 3298-3302

Bronowski, C.; James, C.E. & Winstanley, C. Role of Environmental Survival in Transmission of *Campylobacter jejuni*. *FEMS Microbiology Letters*, 2014, **356**; 8-19

Burrus, V. & Waldor, M.K. Shaping Bacterial Genomes with Integrative & Conjugative Elements. *Research in Microbiology*, 2004, **155**; 376-386

Butzler, J.P. *Campylobacter*, from Obscurity to Celebrity. *Clinical Microbiology & Infection*, 2004, **10**; 868-876

Butzler, J.P.; Dekeyser, P.; Detrain, M. & Dehaen, F. Related *Vibrio* in Stools. *The Journal of Pediatrics*, 1973, **82**; 493-495

Cabiscol, E.; Tamarit, J. & Ros, J. Oxidative Stress in Bacteria & Protein Damage by Reactive Oxygen Species. *International Microbiology*, 2000, **3**; 3-8

Camacho, C.; Madden, T.; Coulouris, G.; Ma, N.; Tao, T.; Agarwala, R. & Morgulis, A. BLAST Command Line Applications User Manual, 2008

Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K. & Madden, T.L. BLAST+: Architecture & Applications. *BMC Bioinformatics*, 2009, **10**; 421

Campbell, A.M. 'Bacteriophages' in Neidhardt, F. C. (Ed. in Chief), R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, & H. E. Umbarger (Eds). 1996. *Escherichia coli & Salmonella: Cellular & Molecular Biology*. American Society for Microbiology

Caner, V.; Cokal, Y.; Cetin, C.; Sen, A. & Karagenc, N. The Detection of *hipO* Gene by Real-Time PCR in Thermophilic *Campylobacter* spp. with Very Weak & Negative Reaction of Hippurate Hydrolysis. *Antonie van Leeuwenhoek*, 2008, **94**; 527-532

Carey, J.B.Jr. Chapter 3: 'Bile Salt Metabolism in Man' in Nair, P.P. & Kritchevsky, D. (Eds) *The Bile Acids: Volume 2; Physiology & Metabolism*, 1973. Plenum Press

Caro-Quintero, A.; Rodriguez-Castaño, G.P.; Konstantinidis, K.T. Genomic Insights into the Convergence & Pathogenicity Factors of *Campylobacter jejuni* & *Campylobacter coli* Species. *Journal of Bacteriology*, 2009, **191**; 5824-5831

Carver, T.J.; Rutherford, K.M.; Berriman, M.; Rajandream, M-A.; Barrell, B.G. & Parkhill, J. ACT: the Artemis Comparison Tool. *Bioinformatics*, 2005, **21**; 3422-3423

Carver, T.; Harris, S.R.; Berriman, M.; Parkhill, J. & McQuillan, J.A. Artemis: an Integrated Platform for Visualisation & Analysis of High-Throughput Sequence-Based Experimental Data. *Bioinformatics*, 2012, **28**; 464-469

Casjens, S. Prophages & Bacterial Genomics: What have we Learned So Far? *Molecular Microbiology*, 2003, **49**; 277–300

Champion, O.L.; Gaunt, M.W.; Gundogdu, O.; Elmi, A.; Witney, A.A.; Hinds, J.; Dorrell, N. & Wren, B.W. Comparative Phylogenomics of the Food-Borne Pathogen *Campylobacter jejuni* Reveals Genetic Markers Predictive of Infection Source. *PNAS*, 2005, **102**; 16043-16048

Champion, O.L.; Cooper, I.A.M.; James, S.L.; Ford, D.; Karlyshev, A.; Wren, B.W.; Duffield, M.; Oyston, P.C.F. & Titball, R.W. *Galleria mellonella* as an Alternative Infection Model for *Yersinia pseudotuberculosis*. *Microbiology*, 2009, **155**; 1516-1522

Champion, O.J.; Karlyshev, A.V.; Senior, N.J.; Woodward, M. ; La Ragione, R.; Howard, S.L.; Wren, B.W. & Titball, R.W. Insect Infection Model for Campylobacter jejuni reveals that O-methyl Phosphoramidate has Insecticidal Activity. *Journal of Infectious Diseases*, 2010, **201**; 776-782

Chang, N. & Taylor, D.E. Use of Pulsed-Field Agarose Gel Electrophoresis to Size Genomes of *Campylobacter* Species & to Construct a *Sal*I Map of *Campylobacter jejuni* UA580. *Journal of Bacteriology*, 1990, **172**; 5211-5217

Cheng, L.; Connor, T.R.; Aanensen, D.M.; Spratt, D.G. & Corander, J. Bayesian Semi-Supervised Classification of Bacterial Samples using MLST Databases. *BMC Bioinformatics*, 2011, **12**; 302

Cheng, L.; Connor, T.R.; Sirén, J.; Aanensen, D.M. & Corander, J. Hierarchical & Spatially Explicit Clustering of DNA Sequences with BAPS Software. *Molecular Biology & Evolution*, 2013, **30**; 1224-1228

Clark, C.G. & Ng, L-K. Sequence Variability of *Campylobacter* Temperate Bacteriophages. *BMC Microbiology*, 2008, **8**; 49

Clark, C.G.; Grant, C.C.R.; Pollari, F.; Marshall, B.; Moses, J.; Tracz, D.M. & Gilmour, M.W. Effects of the *Campylobacter jejuni* CJIE1 Prophage Homologues on Adherence & Invasion in Culture, Patient Symptoms, & Source of Infection. *BMC Microbiology*, 2012, **12**; 269

Cody, A.J.; McCarthy, N.M.; Wimalarathna, H.L.; Colles, F.M.; Clark, L.; Bowler, I.C.J.W.; Maiden, M.C.J. & Dingle, K.E. A Longitudinal 6-Year Study of the Molecular Epidemiology of Clinical *Campylobacter* Isolates in Oxfordshire, United Kingdom. *Journal of Clinical Microbiology*, 2012, **50**; 3193-3201

Colles, F.M.; Jones, K.; Harding, R.M. & Maiden, M.C.J. Genetic Diversity of *Campylobacter jejuni* Isolates from Farm Animals & the Farm Environment. *Applied & Environmental Microbiology*, 2003, **69**; 7409-7413

Connerton, P.L.; Timms, A.R. & Connerton, I.F. *Campylobacter* Bacteriophages & Bacteriophage Therapy. *Journal of Applied Microbiology*, 2011, **111**; 255-265

Corander, J.; Waldmann, P. & Sillanpää, M.J. Bayesian Analysis of Genetic Differentiation between Populations. *Genetics*, 2003, **163**; 367-374

Corander, J.; Marttinen, P.; Sirén, J. & Tang, J. Enhanced Bayesian Modelling in BAPS Software for Learning Genetic Structures of Populations. *BMC Bioinformatics*, 2008, **9**; 539

Corander, J.; Connor, T.R.; O'Dwyer, C.A.; Kroll, J.S. & Hanage, W.P. Population Structure in the *Neisseria* & the Biological Significance of Fuzzy Species. *Journal of the Royal Society Interface*, 2012, **9**; 1208-1215

Croucher, N.J. & Thomson, N.R. Studying Bacterial Transcriptomes using RNA-seq. *Current Opinion in Microbiology,* 2010, **13**; 619-624

Crushell, E.; Harty, S.; Sharif, F. & Bourke, B. Enteric *Campylobacter*: Purging its Secrets? *Pediatric Research*, 2004, **55**; 3-12

Dale, J.W. & Park, S.F. 2004. *Molecular Genetics of Bacteria*, 4th Edition. John Wiley & Sons, Ltd

Datta, S.; Niwa, H. & Itoh, K. Prevalence of 11 Pathogenic Genes of *Campylobacter jejuni* by PCR in Strains Isolated from Humans, Poultry Meat & Broiler & Bovine Faeces. *Journal of Medical Microbiology*, 2003, **52**; 345-348

Davidsen, T.; Beck, E.; Ganapathy, A.; Montgomery, R.; Zafar, N.; Yang, Q.; Madupu, R.; Goetz, P.; Galinsky, K.; White, O. & Sutton, G. The Comprehensive Microbial Resource. *Nucleic Acids Research*, 2010, **38**; D340-D345

Day, W.A.Jr.; Sajecki, J.L.; Pitts, T.M. & Joens, L.A. Role of Catalase in *Campylobacter jejuni* Intracellular Survival. *Infection & Immunity*, 2000, **68**; 6337-6345

De Been, M.; van Schaik, W.; Cheng, L.; Corander, J. & Willems, R.J. Recent Recombination Events in the Core Genome are Associated with Adaptive Evolution in *Enterococcus faecium*. *Genome Biology & Evolution*, 2013, **5**; 1524-35

De Boer, P.; Duim, B.; Rigter, A.; van der Plas, J.; Jacobs-Reitsma, W.F. & Wagenaar, J.A. Computer-Assisted Analysis & Epidemiological Value of Genotyping Methods for *Campylobacter jejuni* & *Campylobacter coli*. *Journal of Clinical Microbiology*, 2000, **38**; 1940-1946

Debruyne, L.; Gevers, D. & Vandamme, P. Chapter One: 'Taxonomy of the Family Campylobacteriaceae' in Nachamkin, I.; Szymanski, C.M. & Blaser, M.J. (Eds) *Campylobacter*, 3rd Edition, 2008. American Society for Microbiology Press

Denis, M.; Rivoal, S.K.; Blivet, E.D.; Salvat, G. & Colin, P. Development of a m-PCR Assay for Simultaneous Identification of *Campylobacter jejuni* & *C. coli*. *Letters in Applied Microbiology*, 1999, **29**; 406-410

Dewey, C.N. & Pachter, L. Evolution at the Nucleotide Level: the Problem of Multiple Whole Genome Alignment. *Human Molecular Genetics*, 2006, **15**; r51-r56

Dingle, K.E.; Colles, F.M.; Wareing, D.R.A.; Ure, R.; Fox, A.J.; Boton, F.E.; Bootsma, H.J.; Willems, R.J.L.; Urwin, R. & Maiden, M.C.J. Multilocus Sequence Typing System for *Campylobacter jejuni*. *Journal of Clinical Microbiology*, 2001, **39**; 14-23

Dingle, K.E.; Colles, F.M.; Falush, D. & Maiden, M.C.J. Sequence Typing & Comparison of Population Biology of *Campylobacter coli* & *Campylobacter jejuni*. *Journal of Clinical Microbiology*, 2005, **43**; 340-347

Dingle, K.E.; McCarthy, N.D.; Cody, A.J.; Peto, T.E.A. & Maiden, M.C.J. Extended Sequence Typing of *Campylobacter* spp., United Kingdom. *Emerging Infectious Diseases*, 2008, **14**; 1620-1622

Dorrell, N.; Mangan, J.A.; Laing, K.G.; Hinds, J.; Linton, D.; Al-Ghusein, H.; Barrell, B.G.; Parkhill, J.; Stoker, N.G.; Karlyshev, A.V.; Butcher, P.D. & Wren, B.W. Whole Genome Comparison of *Campylobacter jejuni* Human Isolates using a Low-Cost Microarray Reveals Extensive Genetic Diversity. *Genome Research*, 2001, **11**; 1706-1715

Döring, A.; Weese, D.; Rausch, T. & Reinert, K. SeqAn An Efficeint, Generic C++ Library for Sequence Analysis. *BMC Bioinformatics*, 2008, **9**; 11

Doyle, L.P. The Etiology of Swine Dysentery. *American Journal of Veterinary Research*, 1948, **9**; 50-51

Duim, B.; Wassenaar, T.M.; Rigter, A. & Wagenaar, J. High-Resolution Genotyping of *Campylobacter* strains Isolated from Poultry & Humans with Amplified Fragment Length Polymorphism Fingerprinting. *Applied & Environmental Microbiology*, 1999, **65**; 2369-2375

Duim, B.; Godschalk, P.C.R.; van den Braak, N.; Dingle, K.E.; Dijkstra, J.R.; Leyde, E.; van der Plas, J.; Colles, F.M.; Endtz, H.P.; Wagenaar, J.A.; Maiden, M.C.J. & van Belkum, A. Molecular Evidence for Dissemination of Unique *Campylobacter jejuni* Clones in Curaçao, Netherlands Antilles. *Journal of Clinical Microbiology,* 2003, **41**; 5593-5597

Duong, T. & Konkel, M.E. Comparative Studies of *Campylobacter jejuni* Genomic Diversity reveal the Importance of Core & Dispensable Genes in the Biology of this Enigmatic Food-Borne Pathogen. *Current Opinion in Biotechnology*, 2009, **20**; 158-165

Ellerbroek, L.I.; Lienau, J-A. & Klein, G. *Campylobacter* spp. in Broiler Flocks at Farm Level & the Potential for Cross-Contamination During Slaughter. *Zoonoses & Public Health*, 2010, **57**; e81-e88

Elliott, W.H. Chapter 11: 'Metabolism of Bile Acids in Liver & Extrahepatic Tissues' in Danielsson, H. & Sjövall, J. (Eds) *Sterols & Bile Acids*, 1985, Elsevier

Endtz, H.P.; van West, H.; Godschalk, P.C.R.; de Haan, L.; Halabi, Y.; van den Braak, N.; Kesztyüs, B.I.; Leyde, E.; Ott, A.; Verkooyen, R.; Price, L.J.; Woodward, D.L.; Rodgers, F.G.; Ang, C.W.; van Koningsveld, R.; van Belkum, A. & Gerstenbluth, I. Risk Factors Associated with *Campylobacter jejuni* Infections in Curaçao, Netherlands Antilles. *Journal of Clinical Microbiology*, 2003, **41**; 5588-5592

Ethelburg, S.; Simonsen, J.; Gerner-Smidt, P.; Olsen, K.E.P. & Mølbak, K. Spatial Distribution & Registry-Based Case-Control Analysis of *Campylobacter* Infections in Denmark, 1991-2001. *American Journal of Epidemiology*, 2005, **162**; 1008-1015

Fearnhead, P.; Smith, N.G.C.; Barrigas, M.; Fox, A. & French, N. Analysis of Recombination in *Campylobacter jejuni* from MLST Population Data. *Journal of Molecular Evolution*, 2005, **61**; 333-340

Fearnley, C.; Manning, G.; Bagnall, M.; Javed, M.A.; Wasenaar, T.M. & Newell, D.G. Identification of Hyperinvasive *Campylobacter jejuni* Strains Isolated from Poultry & Human Clinical Sources. *Journal of Medical Microbiology*, 2008, **57**; 570-580

Feil, E.J.; Maynard Smith, J.; Enright, M.C. & Spratt, B.G. Estimating Recombinational Parameters in *Streptococcus pneumonia* from Multilocus Sequence Typing Data. *Genetics*, 2000, **154**; 1439-1450

Feil, E.J. & Spratt, B.G. Recombination & the Population Structures of Bacterial Pathogens. *Annual Review of Microbiology*, 2001, **55**; 561-590

Ferguson, G.C.; Heinemann, J.A. & Kennedy, M.A. Gene Transfer between *Salmonella enterica* Serovar Typhimurium inside Epithelial Cells. *Journal of Bacteriology*, 2002, **184**; 2235-2242

Firehammer, B.D. & Border, M. Isolation of Temperate Bacteriophages from *Vibrio fetus***.** *American Journal of Veterinary Research,* 1968, **29;** 2229-2235

Fitzgerald, C.; Owen, R.J. & Stanley, J. Comprehensive Ribotyping Scheme for Heat-Stable Serotypes of *Campylobacter jejuni*. *Journal of Clinical Microbiology*, 1996, **34**; 265-269

Florent, A. Les Deux Vibrioses Génitales: a Vibriose due à *Vibrio fetus venerealis* et al Vibriose d'Origine Intestinale due à *V. fetus intestinalis*. *Meded. Veeartsenijsch. Rijksuniv. Gent.* 1959, **3**; 1-60

Fouts, D.E.; Mongodin, E.F.; Mandrell, R.E.; Miller, W.G.; Rasko, D.A.; Ravel, J.; Brinkac, L.M.; DeBoy, R.T.; Parker, C.T.; Daugherty, S.C.; Dodson, R.J.; Durkin, A.S.; Madupu, R.; Sullivan, S.A.; Shetty, J.U.; Ayodeji, M.A.; Shvartsbeyn, A.; Schatz, M.C.; Badger, J.H.; Fraser, C.M. & Nelson, K.E. Major Structural Differences & Novel Potential Virulence Mechanisms from the Genomes of Multiple *Campylobacter* Species. *PLoS Biology*, 2005, **3**:e15

Fox, E.M.; Raftery, M.; Goodchild, A. & Mendzm G.L. *Campylobacter jejuni* Response to Ox-Bile Stress. *FEMS Immunology & Medical Microbiology*, 2007, **49**; 165-172

Franco, D.A. *Campylobacter* Species: Considerations for Controlling a Food-Borne Pathogen. *Journal of Food Protection*, 1988, **51**; 145-153

French, N.; Barrigas, M.; Brown, P.; Ribiero, P.; Williams, N.; Leatherbarrow, H.; Birtles, R.; Bolton, E.; Fearnhead, P. & Fox, A. Spatial Epidemiology & Natural Population Structure of *Campylobacter jejuni* Colonising a Farmland Ecosystem. *Environmental Microbiology*, 2005, **7**; 1116-1126

Friis, C.; Wassenaar, T.M.; Javed, M.A.; Snipen, L.; Lagesen, K.; Hallin, P.F.; Newell, D.G.; Toszeghy, M.; Ridley, A.; Manning, G. & Ussery, D.W. Genomic Characterisation of *Campylobacter jejuni* Strain M1. PLoS One, 2010, **5**; e12253

Frost, J.A.; Oza, A.N.; Thwaites, R.T. & Rowe, B. Serotyping Scheme for *Campylobacter jejuni* & *Campylobacter coli* Based on Direct Agglutination of Heat-Stable Antigens. *Journal of Clinical Microbiology*, 1998, **36**; 335-339

Garret, R.H. & Grisham, C.M. 2009. *Biochemistry*, 5th Edition. Brooks/Cole

Gaasbeek, E.J.; Wagenaar, J.A.; Guilhabert, M.R.; Wösten, M.M.S.M.; van Putten, J.P.M.; van der Graaf-van Bloois, L.; Parker, C.T. & van der Wal, F.J. A DNase Encoded by Integrated Element

CJIE1 Inhibits Natural Transformation of *Campylobacter jejuni*. *Journal of Bacteriology*, 2009, **191**; 2296-2306

Gaunt, P.N. & Piddock, L.J.V. Ciprofloxacin Resistant *Campylobacter* spp. in Humans: an Epidemiological & Laboratory Study. *Journal of Antimicrobial Chemotherapy*, 1996, **37**; 747-757

Gaynor, E.C. & Szymanski, C.M. The 30th Anniversary of *Campylobacter, Helicobacter*, & Related Organisms Workshops - What Have We Learned in Three Decades? *Frontiers in Cellular & Infection Microbiology*, 2012, **2**; 1-5

Gibson, J.; Lorenz, E. & Owen, R.J. Lineages within *Campylobacter jejuni* defined by Numerical Analysis of Pulsed-Field Gel Electrophoretic DNA Profiles. *Journal Medical Microbiology*, 1997, **46**; 157-163

Gilbreath, J.J.; Cody, W.L.; Merrell, D.S. & Hendrixson, D.R. Change is Good: Variations in Common Biological Mechanisms in the Epsilonproteobacterial Genera *Campylobacter* & *Helicobacter*. *Microbiology & Molecular Biology Reviews*, 2011, **75**; 84-132

Gillespie, I.A.; O'Brien, S.J.; Frost, J.A.; Adak, G.K.; Horby, P.; Swan, A.V.; Painter, M.J. & Neal, K.R. A Case-Case Comparison of *Campylobacter coli* & *Campylobacter jejuni* Infection: a Tool for Generating Hypotheses. *Emerging Infectious Diseases*, 2002, **8**; 937-942

Golden, N.J. & Acheson, D.W.K. Identification of Motility & Autoagglutination *Campylobacter jejuni* Mutants by Random Transposon Mutagenesis. *Infection & Immunity*, 2002, **70**; 1761-1771

Gordienko, E.N.; Kazanov, M.D. & Gelfand, M.S. Evolution of Pan-Genomes of *Escherichia coli*, *Shigella* spp., & *Salmonella enterica*. *Journal of Bacteriology*, 2013, **195**; 2786-2792

Grajewski, B.A.; Kusek, J.W. & Gelfand, H.M. Development of a Bacteriophage Typing System for *Campylobacter jejuni* & *Campylobacter coli*. *Journal of Clinical Microbiology*, 1985, **22**; 13-18

Grant, A.J.; Woodward, J. & Maskell, D.J. Development of an *ex vivo* Organ Culture Model using Human Gastro-Intestinal Tissue & *Campylobacter jejuni*. *FEMS Microbiology Letters*, 2006, **263**; 240-243

Grant, C.C.R.; Konkel, M.E.; Cieplak, W. & Tompkins, L.S. Role of Flagella in Adherence, Internalisation, & Translocation of *Campylobacter jejuni* in Nonpolarised & Polarised Epithelial Cell Cultures. *Infection & Immunity*, 1993, **61**; 1764-1771

Grant, K.A. & Park, S.F. Molecular Characterisation of *katA* from *Campylobacter jejuni* & Generation of a Catalase-Deficient Mutant of *Campylobacter coli* by Interspecific Allelic Exchange. *Microbiology*, 1995, **141**; 1369-1376

Green, C.G.; Krause, D.O. & Wylie, J.L. Spatial Analysis of *Campylobacter* Infection in the Canadian Province of Manitoba. *International Journal of Health Geographics*, 2006, **5**:2

Griekspoor, P.; Colles, F.M.; McCarthy, N.D.; Hansbro, P.M.; Ashhurst-Smith, C.; Olsen, B.; Hasselquist, D.; Maiden, M.C.J. & Waldenstrøm, J. Marked Host Specificity & Lack of Phylogeographic Population Structure of *Campylobacter jejuni* in Wild Birds. *Molecular Ecology*, 2013, **22**; 1463-1472

Gripp, E.; Hlahla, D.; Didelot, X.; Kops, F.; Maurischat, S.; Tedin, K.; Alter, K.; Ellerbroek, L.; Schrieber, K.; Schomburg, D.; Janssen, T.; Bartholomaus, P.; Hofreuter, D.; Woltemate, S.; Uhr, M.; Brenneke, B.; Gruening, P.; Gerlach, G.; Wieler, L.; Suerbaum, S. & Josenhans, C. Closely Related *Campylobacter jejuni* Strains from Different Sources Reveal a Generalist rather than a Specialist Lifestyle. *BMC Genomics*, 2011, **12**; 584

Grove-White, D.H.; Leatherbarrow, A.J.H.; Cripps, P.J.; Diggle, P.J. & French, N.P. Molecular Epidemiology & Genetic Diversity of *Campylobacter jejuni* in Ruminants. *Epidemiology & Infection*, 2011, **139**; 1661-1671

Gundogdu, O.; Bentley, S.D.; Holden, M.T.; Parkhill, J.; Dorrell, N. & Wren, B.W. Re-Annotation & Re-Analysis of *Campylobacter jejuni* NCTC11168 Genome Sequence. *BMC Genomics*¸2007, **8**; 162-169

Gunn, J.S. Mechanisms of Bacterial Resistance & Response to Bile. *Microbes & Infection*, 2000, **2**; 907-913

Gusfield, D. *Algorithms on Strings, Trees & Sequences: Computer Science & Computational Biology*. Cambridge University Press

Habib, I.; Uyttendaele, M. & De Zutter, L. Survival of Poultry-Derived *Campylobacter jejuni* of Multilocus Sequence Type Clonal Complexes 21 & 45 under Freeze, Chill, Oxidative, Acid & Heat Stresses. *Food Microbiology*, 2010, **27**; 829-834

Hacker, J. & Kaper, J.B. Pathogenicity Islands & the Evolution of Microbes. *Annual Review of Microbiology*, 2000, **54**; 641-679

Hald, B. & Madsen, M. Healthy Puppies & Kittens as Carriers of *Campylobacter* spp., with Special Reference to *Campylobacter upsaliensis*. *Journal of Clinical Microbiology*, 1997, **35**; 3351-3352

Hanage, W.P.; Fraser, C. & Spratt, B.G. Fuzzy Species among Recombinogenic Bacteria. *BMC Biology*, 2005, **3**; 6

Harrington, C.S.; Thomson-Carter, F.M. & Carter, P.E. Evidence for Recombination in the Flagellin Locus of *Campylobacter jejuni*: Implications for the Flagellin Gene Typing Scheme. *Journal of Clinical Microbiology*, 1997, **35**; 2386-2392

Harvey, P. & Leach, S. Analysis of Coccal Cell Formation by *Campylobacter jejuni* using Continuous Culture Techniques, & the Importance of Oxidative Stress. *Journal of Applied Microbiology*, 1998, **85**; 398-404

Haughton, P.N.; Lyng, J.G.; Cronin, D.A.; Morgan, D.J.; Fanning, S. & Whyte, P. Efficacy of UV Light Treatment for the Microbiological Decontamination of Chicken, Associated Packaging, & Contact Surfaces. *Journal of Food Protection*, 2011, **8**; 109-117

Hepworth, P.J.; Ashelford, K.E.; Hinds, J.; Gould, K.A.; Witney, A.A.; Williams, N.J.; Leatherbarrow, H.; French, N.P.; Birtles, R.J.; Mendonca, C.; Dorrell, N.; Wren, B.W.; Wigley, P.; Hall, N. & Winstanley, C. Genomic Variations Define Divergence of Water/Wildlife-Associated *Campylobacter jejuni* Niche Specialists from Common Clonal Complexes. *Environmental Microbiology*, 2011, **13**; 1549-1560

Herrmann, B. & Burman, L.G. Chemoattracting effect of human urine on motile *Escherichia coli*. *FEMS Microbiology Letters*, 1983, **20**; 411-415

Hu, L. & Kopecko, D.J. *Campylobacter jejuni* 81-176 Associates with Microtubules & Dynein during Invasion of Human Intestinal Cells. *Infection & Immunity*, 1999; **67**; 4171-4182

Hu, L. & Kopecko, D.J. Chapter 19: 'Invasion' in Ketley, J.M. & Konkel, M.E. (Eds) *Campylobacter Molecular & Cellular Biology*, 2008. Horizon Bioscience

Hugdahl, M.B.; Beery, J.T. & Doyle, M.P. Chemotactic Behaviour of *Campylobacter jejuni*. *Infection & Immunity*, 1988, **56**; 1560-1566

Hutchinson, C.A.III.; Peterson, S.N.; Gill, S.R.; Cline, R.T.; White, O.; Fraser, C.M.; Smith, H.O. & Venter, J.C. Global Transposon Mutagenesis & a Minimal *Mycoplasma* Genome. *Science*, 1999, **286**; 2165-2169

Hyytiäinen, H. & Hänninen, M-J. Quality Control Strain *Campylobacter jejuni* ATCC 33560 contains a Frameshift Mutation in the CmeR Regulator. *Antimicrobial Agents & Chemotherapy*, 2012, **12**; 1148

Inokuchi, Y.; Jacobson, A.B.; Hirose, T.; Inayama, S. & Hirashima, A. Analysis of the Complete Nucleotide Sequence of the Group IV RNA Coliphage SP. *Nucleic Acisd Research*, 1988, **16**; 6205-6221

Inouye, M.; Conway, T.C.; Zobel, J. & Holt, K.E. Short Read Sequence Typing (SRST): Multi-Locus Sequence Types from Short Reads. *BMC Genomics*, 2012, **13**; 338

Islam, Z.; van belkum, A.; Wagenaar, J.A.; Cody, A.J.; de Boer, A.G.; Tabor, H.; Jacobs, B.C.; Talukder, K.A. & Endtz, H.P. Comparative Genotyping of *Campylobacter jejuni* Strains from Patients with Guillain-Barré Syndrome in Bangladesh. *PLoS One*, 2009, **4**; e7257

Janssen, R.; Krogfelt, K.A.; Cawthraw, S.A.; van Pelt, W.; Wagenaar, J.A. & Owen, R.J. Host-Pathogen Interactions in *Campylobacter* Infections: the Host Perspective. *Clinical Microbiology Reviews*, 2008, **21**; 505-518

Javed, M.A.; Grant, A.J.; Bagnall, M.C.; Maskell, D.J.; Newell, D.G. & Manning, G. Transposon Mutagenesis in a Hyper-Invasive Clinical Isolate of *Campylobacter jejuni* Reveals a Number of Genes with Potential Roles in Invasion. *Microbiology*, 2010, **156**; 1134-1143

Jolley, K.A. & Maiden, M.C.J. BIGSdb: Scalable Analysis of Bacterial Genome Variation at the Population Level. *BMC Bioinformatics*, 2010, **11**; 595

Jones, F.S.; Orcutt, M. & Little, R.B. *Vibrios* (*Vibrio jejuni*, N.SP.) associated with Intestinal Disorders of Cows & Calves. *Journal of Experimental Medicine*, 1931, **53**; 853-863

Jore, S.; Viljugreinm H.; Brun, E.; Heier, B.T.; Borck, B.; Ethelburg, S.; Hakkinen, M.; Kuusi, M.; Reiersen, J.; Hansson, I.; Engvall, E.O.; Løfdahl, M.; Wagenaar, J.A.; van Pelt, W. & Hofshagen, M. Trends in *Campylobacter* Incidence in Broilers & Humans in Six European Countries, 1997-2007. *Preventative Veterinary Medicine*, 2010, **93**; 33-41

Karlyshev, A.V.; Linton, D.; Gregson, N.A.; Lastovica, A.J. & Wren, B.W. Genetic & Biochemical Evidence of a *Campylobacter jejuni* Capsular Polysaccharide that Accounts for Penner Serotype Specificity. *Molecular Microbiology*, 2000, **35**; 529-541

Karlyshev, A.V.; McCrossan, M.V. & Wren, B.W. Demonstration of Polysaccharide Capsule in *Campylobacter jejuni* using Electron Microscopy. *Infection & Immunity*, 2001, **69**; 5921-5924

Karlyshev, A.V.; Thacker, G.; Jones, M.A.; Clements, M.O. & Wren, B.W. *Campylobacter jejuni* Gene *cj0511* Encodes a Serine Peptidase Essential for Colonisation. *FEBS Open Bio*, 2014, **9**; 468-72

Karro, J.E.; Yan, Y.; Zheng, D.; Zhang, Z.; Carriero, N.; Cayting, P.; Harrison, P. & Gerstein, M. Pseudogene.org: a Comprehensive Database & Comparison Platform for Pseudogene Annotation. *Nucleic Acids Research*, 2007, **35**; D55-D60

Keo, T.; Collins, J.; Kunwar, P.; Blaser, M.J. & Iovine, N.M. *Campylobacter* Capsule & Lipooligosaccharide Confer Resistance to Serum & Cationic Antimicrobials. *Virulence*, 2011, **2**; 30-40

Ketley, J.M. Pathogenesis of Enteric Infection by *Campylobacter*. *Microbiology*, 1997; **143**; 5-21

Killiny, N. & Almeida, R.P.P. Gene Regulation Mediates Host Specificity of a Bacterial Pathogen. *Environmental Microbiology Reports*, 2011, **3**; 791-797

Kinana, A.D.; Cardinale, E.; Bahsoun, I.; Tall, F.; Sire, J-M.; Breurec, S.; Garin, B.; Boye, C. S-B. & Perrier-Gros-Claude, J-D. *Campylobacter coli* Isolates derived from Chickens in Senegal: Diversity, Genetic Exchange with *Campylobacter jejuni* & Quinolone Resistance. *Research in Microbiology*, 2007, **158**; 138-142

King, E.O. Human Infections with *Vibrio fetus* & a Closely Related *Vibrio*. *The Journal of Infectious Diseases*, 1957, **101**; 119-128

Kirby, J.E.; Trempy, J.E. & Gottesman, S. Excision of a P4-Like Cryptic Prophage Leads to Alp Protease Expression in *Escherichia coli*. *Journal of Bacteriology*, 1994, **176**; 2068-2081

Konkel, M.E.; Kim, B.J.; Rivera-Amill, V. & Garvis, S.G. Bacterial Secreted Proteins are required for the Internalisation of *Campylobacter jejuni* into Cultured Mammalian Cells. *Molecular Microbiology*, 1999, **32**; 691-701

Krause-Gruszczynksa, M.; Rohde, M.; Hartig, R.; Genth, H.; Schmidt, G.; Keo, T.; König, W.; Miller, W.G.; Konkel, M.E. & Backert, S. Role of the Small Rho GTPases Rac1 & Cdc42 in Host Cell Invasion of *Campylobacter jejuni*. *Cellular Microbiology*, 2007, **9**; 2431-2444

Krüger, N-J. & Stingl, K. Two Steps away from Novelty - Principles of Bacterial DNA Uptake. *Molecular Microbiology*, 2011, **80**; 860-867

Kurtz, S.; Phillippy, A.; Delcher, A.L.; Smoot, M.; Shumway, M.; Antonescu, C. & Salzberg, S.L. Versatile & Open Software for Comparing Large Genomes. *Genome Biology*, 2004, **5**; Epub R12

Kwiatek, K.; Wolton, B. & Stern, N.J. Prevalence & Distribution of *Campylobacter* spp. On Poultry & Selected Red Meat Carcasses in Poland. *Journal of Food Protection*, 1990, **53**; 127-130

Lacroix, M. Persistant Use of "False" Cell Lines. *International Journal of Cancer*, 2008, **122**; 1-4

Lan, R. & Reeves, P.R. Intraspecies Variation in Bacterial Genomes: The Need for a Species Genome Concept. *Trends in Microbiology*, 2000, **8**; 396-401

Langille, M.G.I.; Hsaio, W.W.L. & Brinkman, F.S.L. Evaluation of Genomic Island Predictors using a Comparative Genomics Approach. *BMC Bioinformatics*, 2008, **9**; 329

Langdon, R.H.; Cuccui, J. & Wren, B.W. *N*-linked Glycosylation in Bacteria: an Unexpected Application. *Future Microbiology*, 2009, **4**; 401-412

LaPierre, P. & Gogarten, J.P. Estimating the Size of the Bacterial Pan Genome. *Trends in Genetics*, 2009, **25**; 107-110

Larsen, M.V.; Cosentino, S.; Rasmussen, S.; Friis, C.; Hasman, H.; Marvig, R.L.; Jelsbak, L.; Sicherltz-Pontén, T.; Ussery, D.W.; Aarestrup, F.M. & Lund, O. Multilocus Sequence Typing of Total-Genome-Sequenced Bacteria. *Journal of Clinical Microbiology*, 2012, **50**; 1355-1361

Lastovica, A.J. & Allos, B.M. Chapter Seven: 'Clinical Significance of *Campylobacter* & Related Species, Other Than *Campylobacter jejuni* & *Campylobacter coli*' in Nachamkin, I.; Szymanski, C.M. & Blaser, M.J. (Eds) *Campylobacter*, 3rd Edition, 2008. American Society for Microbiology Press

Lawrence, J.G. & Ochman, H. Amelioration of Bacterial Genomes: Rates of Change & Exchange. *Journal of Molecular Evolution*, 1997, **44**; 383-397

Layton, S.L.; Morgan, M.J.; Cole, K.; Kwon, Y.M.; Donoghue, D.J.; Hargis, B.M. & Pumford, N.R. Evaluation of *Salmonella*-Vectored *Campylobacter* Peptide Epitopes for Reduction of *Campylobacter jejuni* in Broiler Chickens. *Clinical & Vaccine Immunology*, 2011, **18**; 449-454

Lefébure, T. & Stanhope, M.J. Pervasive, Genome-Wide Positive Selection, Leading to Functional Divergence in the Bacterial Genus *Campylobacter*. *Genome Research*, 2009, **19**; 1224-1232

Lefébure, T.; Pavinski Bitar, P.D.; Suzuki, H. & Stanhope, M.J. Evolutionary Dynamics of Complete *Campylobacter* Pan-Genomes & the Bacterial Species Concept. *Genome Biology & Evolution*, 2010, **2**; 646-655

Leonard, E.E.II.; Takata, T.; Blaser, M.J.; Falkow, S.; Tompkins, L.S. & Gaynor, E.C. Use of an Open-Reading Frame-Specific *Campylobacter jejuni* DNA Microarray as a New Genotyping Tool for Studying Epidemiologically Related Isolates. *Journal of Infectious Diseases*, 2003, **187**; 691-694

Lerat, E.; Daubin, V. & Moran, N.A. From Gene Trees to Organismal Phylogeny in Prokaryotes: The Case of the γ-Proteobacteria. *PLoS Biology*, 2003, **1**; 101-109

Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homwe, N.; Marth, G.; Abecasis, G. & Durbin, R. The Sequence Alignment/Map Format & SAMTools. *Bioinformatics*, 2009, **25**; 2078-2079

Lin, J.; Sahin, O.; Overbye Michel, L. & Zhang, Q. Critical Role of Multidrug Efflux Pump CmeABC in Bile Resistance & *in vivo* Colonisation of *Campylobacter jejuni*. *Infection & Immunity*, 2003, **71**; 4250-4259

Lin, J. Novel Approaches for *Campylobacter* Control in Poultry. *Foodborne Pathogens & Disease*, 2009, **6**; 755-765

Linton, D.; Gilbert, M.; Hitchen, P.G.; Dell, A.; Morris, H.R.; Wakarchuk, W.W.; Gregson, N.A. & Wren, B.W. Phase Variation of a ß-1,3 Galactosyltransferase involved in Generation of the Ganglioside GM_1-Like Lipo-Oligosaccharide of *Campylobacter jejuni*. *Molecular Microbiology*, 2000, **37**; 501-514

Linton, D.; Dorrell, D.; Hitchen, P.G.; Amber, S.; Karlyshev, A.V.; Morris, H.R.; Dell, A.; Valvano, M.A.; Aebl, M. & Wren, B.W. Functional Analysis of the *Campylobacter jejuni* N-Linked Protein Glycosylation Pathway. *Molecular Microbiology*, 2005, **55**; 1695-1703

Lior, H.; Woodward, D.L.; Edgar, J.A.; Laroche, L.J. & Gill, P. Serotyping of *Campylobacter jejuni* by Slide Agglutination Based on Heat-Labile Antigenic Factors. *Journal of Clinical Microbiology*, 1982, **15**; 761-768

Lipman, D.J. & Pearson, W.R. Rapid & Sensitive Protein Similarity Searches. *Science*, 1985, **227**; 1435-1441

Liu, W.; Fang, L.; Li, M.; Li, S.; Guo, S.; Luo, R.; Feng, Z.; Li, B.; Zhou, Z.; Shao, G.; Chen, H. & Xiao, S. Comparative Genomics of Mycoplasma: Analysis of Conserved Essential Genes & Diversity of the Pan-Genome. *PLoS One*, 2012, **7**; e35698

Logan, S.M. & Trust, T.J. Structural & Antigenic Heterogeneity of Lipopolysaccharides of *Campylobacter jejuni* & *Campylobacter coli*. *Infection & Immunity*, 1984, **45**; 210-216

Luber, P. Cross-Contamination versus Undercooking of Poultry Meat or Eggs - which Risks Need to be Managed First? *International Journal of Food Microbiology*, 2009, **134**; 21-28

Ludwig, W. & Klenk, H-P. 'Overview: A Phylogenetic Backbone & Taxonomic Framework for Prokaryotic Systematics' in Garrity, G.M. (Ed) *Bergey's Manual of Systematic Bacteriology*, 2nd Edition, Volume 4, 2001. Springer

Luo, Y.; Lin, K. & Xu, J. Chapter 2: 'Comparative Microbial Genomics: Analytical Tools, Population Genetic Patterns & Evolutionary Implications' in Xu, J. (Ed) *Microbial Population Genetics*, 2008. Caister Academic Press

Madden, T. Chapter 16: The BLAST Sequence Analysis Tool, in McEntyre, J. & Ostell, J. (Eds) *The NCBI Handbook*, 2002/2013. National Centre for Biotechnology Information

Madigan, M.T. & Martinko, J.M. *Brock Biology of Microorganisms*, International Edition, 11th Edition, 2006. Pearson Benjamin Cummings

Maiden, M.C.J. & Dingle, K.E. Chapter Two: 'Population Biology of Campylobacter jejuni & Related Organisms' in Nachamkin, I.; Szymanski, C.M. & Blaser, M.J. (Eds) *Campylobacter*, 3rd Edition, 2008. American Society for Microbiology Press

Malik-Kale, P.; Parker, C.T. & Konkel, M.E. Culture of *Campylobacter jejuni* with Sodium Deoxycholate Induces Virulence Gene Expression. *Journal of Bacteriology*, 2008, **190**; 2286-2297

Manning, G.; Dowson, C.G.; Bagnall, M.C.; Ahmed, I.H.; West, M. & Newell, D.G. Multilocus Sequence Typing for Comparison of Veterinary & Human Isolates of *Campylobacter jejuni*. *Applied & Environmental Microbiology*, 2003, **69**; 6370-6379

Marttinen, P.; Baldwin, A.; Hanage, W.P.; Dowson, C.; Mahenthiralingam, E. & Corander, J. Bayesian Modelling of Recombination Events in Bacterial Populations. *BMC Bioinformatics*, 2008, **9**; 421

Marttinen, P.; Hanage, W.P.; Croucher, N.J.; Connor, T.R.; Harris, S.R.; Bentley, S.D. & Corander, J. Detection of Recombination Events in Bacterial Genomes from Large Population Samples. *Nucleic Acids Research*, 2012, **40**; e6

Masanta, W.O.; Heimesaat, M.M.; Bereswill, S.; Tareen, A.M.; Lugert, R.; Groβ, U. & Zautner, A.E. Modification of Intestinal Microbiota & its Consequences for Innate Immune Response in the Pathogenesis of Campylobacteriosis. *Clinical & Developmental Immunology*, 2013, 526860

Maynard Smith, J. & N.H. Smith. Detecting Recombination from Gene Trees. *Molecular Biology & Evolution*, 1998, **15**; 590-599

McCarthy, N.D.; Colles, F.M.; Dingle, K.E.; Bagnall, M.C.; Manning, G.; Maiden, M.C.J. & Falush, D. Host Associated Genetic Import in *Campylobacter jejuni*. *Emerging Infectious Diseases*, 2007, **13**; 267-272

McNally, A.; Cheng, L.; Harris, S.R. & Corander, J. The Evolutionary Path to Extraintestinal Pathogenic, Drug-Resistant *Escherichia coli* is marked by Drastic Reduction in Detectable Recombination within the Core Genome. *Genome Biology & Evolution*, 2013, **5**; 699-710

McFadyean, J. & Stockman, S. *Report of the Departmental Committee appointed by the Board of Agriculture & Fisheries to Inquire into Epizootic Abortion, Part III, Abortion in Sheep.* London, HMSO, 1913

Medini, D.; Donati, C.; Tettelin, H.; Masignani, V. & Rappuoli, R. The Microbial Pan-Genome. *Current Opinion in Genetics & Development*, 2005, **15**; 589-594

Meinersmann, R.J.; Helsel, L.O.; Fields, P.I. & Hiett, K.L. Discrimination of *Campylobacter jejuni* Isolates by *fla* Gene Sequencing. *Journal of Clinical Microbiology*, 1997, **35**; 2810-2814

Meinersmann, R.J. & Hiett, K.L. Concerted Evolution of Duplicate *fla* Genes in *Campylobacter*. *Microbiology*, 2000, **146**; 2283-2290

Méric, G.; Yahara, K.; Mageiros, L.; Pascoe, B.; Maiden, M.C.J.; Jolley, K. & Sheppard, S.K. A Reference Pan-Genome Approach to Comparative Bacterial Genomics: Identification of Novel Epidemiological Markers in Pathogenic *Campylobacter*. *PLoS One*, 2014, **9**; e92798

Miller, W.G.; Englen, M.D.; Kathariou, S.; Wesley, I.V.; Wang, G.; Pittenger-Alley, L.; Siletz, R.M.; Muraoka, W.; Fedorka-Cray, P.J. & Mandrell, R.E. Identification of Host-Associated Alleles by Multilocus Sequence Typing of *Campylobacter coli* Strains from Food Animals. *Microbiology*, 2006, **152**; 245-255

Misawa, N. & Blaser, M.J. Detection & Characterisation of Autoagglutination Activity by *Campylobacter jejuni*. *Infection & Immunity*, 2000, **68**; 6168-6175

Monteville, M.R.; Yoon, J.E. & Konkel, M.E. Maximal Adherence & Invasion of INT 407 Cells by *Campylobacter jejuni* requires the CadF Outer-Membrane Protein & Microfilament Reorganisation. *Microbiology*, 2003; **149**; 153-165

Moore, J.E.; Lanser, J.; Heuzenroeder, M.; Ratcliff, R.M.; Millar, B.C. & Madden, R.H. Molecular Diversity of *Campylobacter coli* & *Campylobacter jejuni* Isolated from Pigs at Slaughter by *fla*A-RFLP Analysis & Ribotyping. *Journal of Veterinary Medicine B*, 2002, **49**; 388-393

Moore, J.E.; Corcoran, D.; Dooley, J.S.G.; Fanning, S.; Lucey, B.; Matsuda, M.; McDowell, D.A.; Mégraud, F.; Millar, B.C.; O'Mahony, R.; O'Riordan, L.; O'Rourke, M.; Rao, J.R.; Rooney, P.J.; Sails, A. & Whyte, P. *Campylobacter*. *Veterinary Research*, 2005, **36**; 351-382

Murphy, C.; Carroll, C. & Jordan, K.N. Environmental Survival Mechanisms of the Foodborne Pathogen *Campylobacter jejuni*. *Journal of Applied Microbiology*, 2006, **100**; 623-632

Mushegian, A.R. *Foundations of Comparative Genomics,* 2007. Elsevier

Nachamkin, I.; Ung, H. & Li, M. Increasing Fluoroquinolone Resistance in *Campylobacter jejuni*, Pennsylvania, USA, 1982-2001. *Emerging Infectious Diseases*, 2002, **8**; 1501-1503

Nayak, R.; Stewart, T.M. & Nawaz, M.S. PCR Identification of *Campylobacter coli* & *Campylobacter jejuni* by Partial Sequencing of Virulence Genes. *Molecular & Cellular Probes*, 2005, **19**; 187-193

Neal-McKinney, J.M. & Konkel, M.E. The *Campylobacter jejuni* CiaC Virulence Protein is Secreted from the Flagellum & delivered to the Cytosol of Host Cells. *Frontiers in Cellular & Infection Microbiology*, 2012, **2**; 31

Nesbakken, T.; Eckner, K.; Høidal, H.K. & Røtterud, O-J. Occurrence of *Yersinia enterocolitica* & *Campylobacter* spp. in Slaughter Pigs & Consequences for Meat Inspection, Slaughtering, & Dressing Procedures. *International Journal of Food Microbiology*, 2003, **80**; 231-240

Newell, D.G.; Koopmans, M.; Verhoef, L.; Duizer, E.; Aidara-Kane, A.; Sprong, H.; Opsteegh, M.; Langelaar, M.; Threfall, J.; Scheutz, F.; van der Giessen, J. & Kruse, H. Food-Borne Diseases - The Challenges of 20 Years Ago still Persist while New Ones Continue to Emerge. *International Journal of Food Microbiology*, 2010, **139**; s3-s15

Newell, D.G.; Elvers, K.T.; Dopfer, D.; Hansson, I.; Jones, P.; James, S.; Gittins, J.; Stern, N.J.; Davies, R.; Connerton, I.; Pearson, D.; Salvat, G. & Allen, V.M. Biosecurity-Based Interventions &

Strategies to Reduce *Campylobacter* spp. on Poultry Farms. *Applied & Environmental Microbiology*, 2011, **77**; 8605-8614

Nicholson, M.A. & Patton, C.M. Evaluation of Disc Method for Hippurate Hydrolysis by *Campylobacter* Species. *Journal of Clinical Microbiology*, 1995, **33**; 1341-1343

Novik, V.; Hofreuter, D. & Galán, J.E. Identification of *Campylobacter jejuni* Genes Involved in Its Interaction with Epithelial Cells. *Infection & Immunity*, 2010, **78**; 3540-3553

Nuijten, P.J.M.; van den Berg, A.J.G.; Formentini, I.; van der Zeijst, B.A.M. & Jacobs, A.A.C. DNA Rearrangements in the Flagellin Locus of an *flaA* Mutant of *Campylobacter jejuni* during Colonisation of Chicken Ceca. *Infection & Immunity*, 2000, **68**; 7137-7140

Otto, T.D.; Dillon, G.P.; Degrave, W.S. & Berriman, M. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Research*, 2011, **39**; e57

Palyada, K.; Sun, Y-Q.; Flint, A.; Butcher, J.; Naikare, H. & Stintzi, A. Characterisation of the Oxidative Stress Stimulon & PerR Regulon of *Campylobacter jejuni*. *BMC Genomics*, 2009, **10**; 481

Parham, P. *The Immune System*, 3rd Edition, 2009. Garland Science, Taylor & Francis Group.

Parker, C.T.; Quiñones, B.; Miller, W.G.; Horn, S.T. & Marshall, R.E. Comparative Genomic Analysis of *Campylobacter jejuni* Strains Reveals Diversity due to Genomic Elements Similar to those Present in *C. jejuni* Strain RM1221. *Journal of Clinical Microbiology*, 2006, **44**; 4125-4135

Parker, C.T.; Miller, W.G.; Horn, S.T. & Lastovica, A.J. Common Genomic Features of *Campylobacter jejuni* subsp. *doylei* Strains Distinguish them from *C. jejuni* subsp. *jejuni*. *BMC Microbiology*, 2007, **7**; 50

Parkhill, J.; Wren, B.W.; Mungall, K.; Ketley, J.M.; Churcher, C.; Basham, D.; Chillingworth, T.; Davies, R.M.; Feltwell, T.; Holroyd, S.; Jagels, K.; Karlyshev, A.V.; Moule, S.; Pallen, M.J.; Penn, C.W.; Quall, M.A.; Rajandream, M-A.; Rutherford, K.M.; van Vliet, A.H.M.; Whitehead, S. & Barrell, B.G. The Genome Sequence of the Food-Borne Pathogen *Campylobacter jejuni* Reveals Hypervariable Sequences. *Nature*, 2000, **403**; 665-668

Parsons, B.N.; Cody, A.J.; Porter, C.J.; Stavisky, J.H.; Smith, J.L.; Williams, N.J.; Leatherbarrow, A.J.H.; Hart, C.A.; Gaskell, R.M.; Dingle, K.E. & Dawson, S. Typing of *Campylobacter jejuni* Isolates from Dogs by Use of Multilocus Sequence Typing & Pulsed-Field Gel Electrophoresis. *Journal of Clinical Microbiology*, 2009, **47**; 3466-3471

Patrick, M.E.; Mahon, B.E.; Zansky, S.M.; Hurd, S. & Scallan, E. Riding in Shopping Carts & Exposure to Raw Meat & Poultry Products: Prevalence of, & Risk Factors Associated with, this Risk Factor for *Salmonella* & *Campylobacter* infection in Children Younger than 3 Years. *Journal of Food Protection*, 2010, **6**; 1097-1100

Pearson, B.M.; Gaskin, D.J.H.; Segers, R.P.A.M.; Wells, J.M.; Nuijten, P.J.M. & van Vliet, A.H.M. The Complete Genome Sequence of *Campylobacter jejuni* 81116 (NCTC11828). *Journal of Bacteriology*, 2007, **189**; 8402-8403

Pearson, W.R. & Lipman, D.J. Improved Tools for Biological Sequence Comparison. *PNAS*, 1988, **85**; 2444-2448

Pearson, W.R. Chapter 10: 'Flexible Sequence Similarity Searching with the FASTA3 Program Package'in Misener, S. & Krawetz, S.A. (Eds), *Methods in Molecular Biology, volume 132: Bioinformatics Methods & Protocols*, 2000. Humana Press

Penner, J.L. & Hennessy, J.N. Passive Haemagglutination Technique for Serotyping *Campylobacter fetus* subsp. *jejuni* on the Basis of Soluble Heat-Stable Antigens. *Journal of Clinical Microbiology*, 1980, **12**; 732-737

Penner , J.L. The Genus *Campylobacter*: A Decade of Progress. *Clinical Microbiology Reviews*, 1988, **1**; 157-172

Pezzotti, G.; Serafin, A.; Luzzi, I.; Mioni, R.; Milan, M. & Perin, R. Occurrence & Resistance to Antibiotics of *Campylobacter jejuni* & *Campylobacter coli* in Animals & Meat in Northeastern Italy. *International Journal of Food Microbiology*, 2003, **82**; 281-287

Pollock, K.; Locking, M.; Browning, L.; Smith-Palmer, A.; Brownlie, S. Gastro-Intestinal & Foodborne Infections: Laboratory Reports for Common Bacterial, Protozoal & Viral Infections 2009. *Health Protection Scotland Weekly Report*, 2010; **44;** 38–41

Ponstigl, SMALT, 2010, https://www.sanger.ac.uk/resources/software/smalt/

Pritchard, J.K.; Stephens, M. & Donnelly, P. Inference of Population Structure using Multilocus Genotype Data. *Genetics*, 2000, **155**; 945-959

Public Health Servies Report: Population & Environmental Health Group, Ministry of Health. 2009. *Notifiable & other Diseases in New Zealand 2008 Annual Surveillance Report*.

Preston, M.A. & Penner, J.L. Structural & Antigenic Properties of Lipopolysaccharides from Serotype Reference Strains of *Campylobacter jejuni*. *Infection& Immunity*, 1987, **55**; 1806-1812

Purdy, D.; Buswell, C.M.; Hodgson, A.E.; McAlpine, K.; Henderson, I. & Leach, S.A. Characterisation of Cytolethal Distending Toxin (CDT) Mutants of *Campylobacter jejuni*. *Journal of Medical Microbiology*, 2000, **49**; 473-479

Rajkovic, A.; Tomic, N.; Smigic, N.; Uyttendaele, M.; Ragaert, P. & Devlieghere, F. Survival of *Campylobacter jejuni* on Raw Chicken Legs Packed in High-Oxygen or High-Carbon Dioxide Atmosphere after the Decontamination with Lactic Acid-Sodium Lactate Buffer. *International Journal of Food Microbiology*, 2010, **140**; 201-206

Rambaut, 2007: http://tree.bio.ed.ac.uk/software/figtree/

Raphael, B.H.; Periera, S.; Flom, G.A.; Zhang, Q.; Ketley, J.M. & Konkel, M.E. The *Campylobacter jejuni* Response Regulator, CbrR, Modulates Sodium Deoxycholate Resistance & Chicken Colonisation. *Journal of Bacteriology*, 2005, **187**; 3662-3670

Rausch, T.; Emde, A-K.; Weese, D.; Döring, A.; Notredame, C. & Reinert, K. Segment-Based Multiple Sequence Alignment. *Bioinformatics*, 2008, **24**; i187-i192

Redfield, R.J. Do Bacteria have Sex? *Nature Reviews Genetics,* 2001, **2**; 634-639

Richardson, E.J. & Watson, M. The Automatic Annotation of Bacterial Genomes. *Briefings in Bioinformatics*, 2012, **14**; 1-12

Rokas, A. Phylogenetic Analysis of Protein Sequence Data using the Randomised Accelerated Maximum Likelihood (RAXML) Program. *Current Protocols in Molecular Biology*, 2011, **Supplement 19**; 19.11.1-19.11.14

Romaniuk, P.J.; Zoltowska, B,; Trust, T.J.; Lane, D.J.; Olsen, G.J.; Pace, N.R. & Stahl, D.A. *Campylobacter pylori*, the Spiral Bacterium Associated with Human Gastritis, is Not a True *Campylobacter* sp. *Journal of Bacteriology*, 1987, **169**; 2137-2141

Rozynek, E.; Dzierzanowska-Fangrat, K.; Jozwiak, P.; Popowski, J.; Korsak, D. & Dzierzanowska, D. Prevalence of Potential Virulence Markers in Polish *Campylobacter jejuni* & *Campylobacter coli* Isolates obtained from Hospitalized Children & from Chicken Carcasses. *Journal of Medical Microbiology*, 2005, **54**; 615-619

Russell, R.G. & Blake, D.C. Cell Association & Invasion of Caco-2 Cells by *Campylobacter jejuni*. *Infection & Immunity*, 1994; **62**; 3773-3779

Russell, S.J. & Norvig, P. *Artificial Intelligence: A Modern Approach*, 2010. Prentice Hall

Rutherford, K.; Parkhill, J.; Crook, J.; Horsnell, T.; Rice, P.; Rajandream, M.A. & Barrell, B. Artemis: Sequence Visualization & Annotation. Bioinformatics, 2000; **16**; 944-9455

Sails, A.D.; Wareing, D.R.A.; Bolton, F.J.; Fox, A.J. & Curry, A. Characterisation of 16 *Campylobacter jejuni* & *C. coli* Typing Bacteriophages. *Journal of Medical Microbiology*, 1998, **47**; 123-128

Samuel, M.C.; Vugia, D.J.; Shallow, S.; Marcus, R.; Segler, S.; McGivern, T.; Kassenborg, H.; Reilly, K.; Kennedy, M.; Angulo, F. & Tauxe, R.V. Epidemiology of Sporadic *Campylobacter* Infection in the United States & Declining Trend in Incidence, FoodNet 1996-1999. *Clinical Infectious Diseases*, 2004, **38**; s165-s174

Schloss, P.D.; Westcott, S.L.; Ryabin, T.; Hall, J.R.; Hartmann, M.; Hollister, E.B.; Lesniewski, R.A.; Oakley, B.B.; Parks, D.H.; Robinson, C.J.; Sahl, J.W.; Stres, B.; Thallinger, G.G.; van Horn, D.J. & Weber, C.F. Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing & Comparing Microbial Communities. *Applied & Environmental Microbiology*, 2009, **75**; 7537-7541

Scott, A.E.; Timms, A.R.; Connerton, P.L.; Carrillo, C.L.; Radzum, K.A. & Connerton, I.F. Genome Dynamics of *Campylobacter jejuni* in Response to Bacteriophage Predation. *PLoS Pathogens*, 2007, **3**; 1142-1151

Sebald, M. & Véron, M. Teneur en Bases de l'ADN et Classification des *Vibrions*. *Annales de l'Institut Pasteur*, 1963, **105**; 897-910

Seemann, T. Prokka: Rapid Prokaryotic Genome Annotation. *Bioinformatics*, 2014, **30**; 2068-2069

Senior, N.J.; Bagnall, M.C.; Champion, O.L.; Reynolds, S.E.; La Ragione, R.M.; Woodward, M.J.; Salguero, F.J. & Titball, R.W. *Galleria mellonella* as an Infection Model for *Campylobacter jejuni* Virulence. *Journal of Medical Microbiology*, 2011, **60**; 661-669

Sheppard, S.K.; McCarthy, N.D.; Falush, D. & Maiden, M.C.J. Convergence of *Campylobacter* Species: Implications for Bacterial Evolution. *Science*, 2008; **320**; 237-239

Sheppard, S.K.; Dallas, J.F.; Strachan, N.J.C.; MacRae, M.; McCarthy, N.D.; Wilson, D.J.; Gormley, F.J.; Falush, D.; Ogden, I.D.; Maiden, M.C.J. & Forbes, K.J. *Campylobacter* Genotyping to Determine the Source of Human Infection. *Clinical Infectious Diseases*, 2009, **48**; 1072-1078

Sheppard, S.K.; Colles, F.; Richardson, J.; Cody, A.J.; Elson, R.; Lawson, A.; Brick, G.; Meldrum, R.I.; Little, C.L.; Owen, R.J.; Maiden, M.C.J. & McCarthy, N.D. Host Association of *Campylobacter* Genotypes Transcends Geographic Variation. *Applied & Environmental Microbiology*, 2010, **76**; 5269-5277

Sheppard, S.K.; McCarthy, N.D.; Jolley, K.A. & Maiden, M.C.J. Introgression in the Genus *Campylobacter*: Generation & Spread of Mosaic Alleles. *Microbiology*, 2011(a), **157**; 1066-1074

Sheppard, S.K.; Colles, F.M.; McCarthy, N.D.; Strachan, N.J.C.; Ogden, I.D.; Forbes, K.J.; Dallas, J.F. & Maiden, M.C.J. Niche Segregation & Genetic Structure of *Campylobacter jejuni* Populations from Wild & Agricultural Host Species. *Molecular Ecology*, 2011(b), **20**; 3484-3490

Sheppard, S.K.; Didelot, X.; Jolley, K.A.; Darling, A.E.; Pascoe, B.; Meric, G.; Kelly, D.J.; Cody, A.; Colles, F.M.; Carter, P.; Miller, W.G.; McCarthy, N.D.; Owen, R.; Litrup, E.; Egholm, M.; Arrourtit, J.P.; Bentley, S.D.; Parkhill, J.; Maiden, M.C.J. & Falush, D. Progressive Genome-Wide Introgression in Agricultural *Campylobacter coli*. *Molecular Ecology*, 2013, **22**; 1051-1064

Sheppard, S.K.; Cheng, L.; Méric, G.; de Haan, C.P.A.; Llarena, A-K.; Marttinen, P.; Vidal, A.; Ridley, A.; Clifton-Hadley, F.; Connor, T.R.; Strachan, N.J.C.; Forbes, K.; Colles, F.M.; Jolley, K.A.; Bentley, S.D.; Maiden, M.C.J.; Hänninen, M-L.; Parkhill, J.; Hanage, W.P. & Corander, J. Cryptic Ecology among Host Generalist *Campylobacter jejuni* in Domestic Animals. *Molecular Ecology*, 2014, **23**; 2442-2451

Siemer, B.L.; Nielsen, E.M. & On, S.L.W. Identification & Molecular Epidemiology of *Campylobacter coli* Isolates from Human Gastroenteritis, Food & Animal Sources by Amplified Fragment Length Polymorphism Analysis & Penner Serotyping. *Applied & Environmental Microbiology*, 2005, **71**; 1953-1958

Skirrow, M.B. *Campylobacter* Enteritis: a 'New' Disease. *British Medical Journal*, 1977, **2**; 9-11

Skirrow, M.B. & Benjamin, J. '1001' Campylobacters: Cultural Characteristics of Intestinal Campylobacters from Man & Animals. *Journal of Hygiene*, 1980, **85**; 427-442

Smith, K.E.; Besser, J.M.; Hedberg, C.W.; Leano, F.T.; Bender, J.B.; Wicklund, J.H.; Johnson, B.P.; Moore, K.A. & Osterholm, M.T. Quinolone-Resistant *Campylobacter jejuni* Infections in Minnesota, 1992-1998. *The New England Journal of Medicine*, 1999, **340**; 1525-1532

Smith, T. Spirilla associated with Disease of the Fetal Membranes in Cattle (Infectious Abortion). *Journal of Experimental Medicine*, 1918, **28**; 701-721

Smith, T. & Orcutt, M.L. *Vibrios* from Calves & their Serological Relation to *Vibrio fetus*. *Journal of Experimental Medicine*, 1927, **45**; 391-397

Smith, T. & Taylor, M.S. Some Morphological & Biochemical Characters of the Spirilla (*Vibrio fetus* n. sp.) associated with Disease of the Fetal Membranes in Cattle. *Journal of Experimental Medicine*, 1919, **30**; 299-312

Snipen, L.; Almøy, T. & Ussery, D.W. Microbial Comparative Pan-Genomics using Binomial Mixture Models. *BMC Genomics*, 2009, **10**; 385

Song, Y.C.; Jin, S.; Louie, H.; Ng, D.; Lau, R.; Zhang, Y.; Weerasekera, R.; Al Rashid, S.; Ward, L.A.; Der, D. & Chan, V.L. FlaC, a Protein of *Campylobacter jejuni* TGH9011 (ATCC43431) Secreted through the Flagellar Apparatus, Binds Epithelial Cells & Influences Cell Invasion. *Molecular Microbiology*, 2004, **53**; 541-553

Spratt, B.G. & Maiden, M.C.J. Bacterial Population Genetics, Evolution & Epidemiology. *Philosophical Transactions of the Royal Society of London*, 1999, **354**; 701-710

Sproston, E.L.; Ogden, I.D.; MacRae, M.; Forbes, K.J.; Dallas, J.F.; Sheppard, S.K.; Cody, A.; Colles, F.; Wilson, M.J. & Strachan, N.J.C. Multi-Locus Sequence Types of *Campylobacter* Carried by Flies & Slugs Acquired from Local Ruminant Faeces. *Journal of Applied Microbiology*, 2010, **109**; 829-838

Stanley, K.; Cunningham, R. & Jones, K. Isolation of *Campylobacter jejuni* from Groundwater. *Journal of Applied Microbiology*, 1998, **85**; 189-191

Stothard, P. & Wishart, D.S. Circular Genome Visualisation & Exploration using CGView. *Bioinformatics*, 2005, **21**; 537-539

Strachan, N.J.C.; Gormley, F.J.; Rotariu, O.; Ogden, I.D.; Miller, G.; Dunn, G.M.; Sheppard, S.K.; Dallas, J.F.; Reid, T.M.S.; Howie, H.; Maiden, M.C.J. & Forbes, K. Attribution of *Campylobacter* Infections in Northeast Scotland to Specific Sources by Use of Multilocus Sequence Typing. *Journal Infectious Diseases*, 2009, **199**; 1205-1208

Strachan, N.J.C. & Forbes, K.J. The Growing UK Epidemic of Human Campylobacteriosis. *The Lancet*, 2010, **376**; 665-667

Strachan, N.J.C.; Rotariu, O.; Smith-Palmer, A.; Cowden, J.; Sheppard, S.K.; O'Brien, S.J.; Maiden, M.C.J.; Macrae, M.; Bessell, P.R.; Matthews, L.; Reid, S.W.J.; Innocent, G.T.; Ogden, I.D. & Forbes, K.J. Identifying the Seasonal Origins of Human Campylobacteriosis. *Epidemiology & Infection*, 2013(a), **141**; 1267-1275

Strachan, N.J.C.; Rotariu, O.; MacRae, M.; Sheppard, S.K.; Smith-Palmer, A.; Cowden, J.; Maiden, M.C.J. & Forbes, K.J. Operationalising Factors that Explain the Emergence of Infectious Diseases: A Case Study of the Human Campylobacteriosis Epidemic. *PLOS One*, 2013(b), **8**; e79331

Suerbaum, S.; Lohrengel, M.; Sonnevend, A.; Ruberg, F. & Kist, M. Allelic Diversity & Recombination in *Campylobacter jejuni*. *Journal of Bacteriology*, 2001, **183**; 2553-2559

Svedhem, A. & Kaijser, B. Isolation of *Campylobacter jejuni* from Domestic Animals & Pets: Probable Origin of Human Infection. *Journal of Infection*, 1981, **3**; 37-40

Szewzyk, U.; Szewzyk, R.; Manz, W. &Schleifer, K.-H.Microbiological Safety of Drinking Water.Annual Review of Microbiology, 2000, **54**; 81-127

Taboada, E.N.; Acedillo, R.R.; Carrillo, C.D.; Findlay, W.A.; Medeiros, D.T.; Mykytczuk, O.L.; Roberts, M.J.; Valencia, C.A.; Farber, J.M. & Nash, J.H.E. Large Scale Comparative Genomics Meta-Analysis of *Campylobacter jejuni* Isolates Reveals Low Level of Genome Plasticity. *Journal of Clinical Microbiology*, 2004, **42**; 4566-4576

Takata, T.; Fujimoto, S. & Amako, K. Isolation of Nonchemotactic Mutants of *Campylobacter jejuni* & their Colonisation of the Mouse Intestinal Tract. *Infection & Immunity*, 1992, **60**; 3596-3600

Tang, J.; Hanage, W.P.; Fraser, C. & Corander, J. Identifying Currents in the Gene Pool for Bacterial Populations using an Integrative Approach. *PLoS Computational Biology*, 2009, **5**; e1000455

Terra, V.S.; Mills, D.C.; Yates, L.E.; Abouelhadid, S.; Cuccui, J. & Wren, B.W. Recent Developments in Bacterial Protein Glycan Coupling Technology & Glycoconjugate Vaccine Design. *Journal of Medical Microbiology*, 2012, **61**; 919-926

Tettelin, H.; Masignani, V.; Cieslewicz, M.J.; Donati, C.; Medini, D.; Ward, N.L.; Angiuoli, S.V.; Crabtree, J.; Jones, A.L.; Durkin, A.S.; DeBoy, R.T.; Davidsen, T.M.; Mora, M.; Scarselli, M.; y Ros, I.M.; Peterson, J.D.; Hauser, C.R.; Sundaram, J.P.; Nelson, W.C.; Madupu, R.; Brinkac, L.M.; Dodson, R.J.; Rosovitz, M.J.; Sullivan, S.A.; Daugherty, S.C.; Haft, D.H.; Selengut, J.; Gwinn, M.L.; Zhou, L.; Zafar, N.; Khouri, H.; Radune, D.; Dimitrov, G.; Watkins, K.; O'Connor, K.J.B.; Smith, S.; Utterback, T.R.; White, O.; Rubens, C.E.; Grandi, G.; Madoff, L.C.; Kasper, D.L.; Telford, J.L.; Wessels, M.R.; Rappuoli, R. & Fraser, C. Genome Analysis of Multiple Pathogenic Isolates of *Streptococcus agalactiae*: Implications for the Microbial 'Pan-Genome'. *PNAS*, 2005, **102**; 13950-13955

Thomas, C.M. & Nielsen, K.M. Mechanisms of, & Barriers to, Horizontal Gene Transfer between Bacteria. *Nature Reviews Microbiology*, 2005, **3**; 711-721

Totten, P.A.; Patton, C.M.; Tenover, F.C.; Barrett, T.J.; Stamm, W.E.; Steigerwalt, A.G.; Lin, J.Y.; Holmes, K.K. & Brenner, D.J. Prevalence & Characterisation of Hippurate-Negative *Campylobacter jejuni* in King County, Washington. *Journal of Clinical Microbiology*, 1987, **25**; 1747-1752

Tsai, I.J.; Otto, T.D. & Berriman, M. Improving Draft Assemblies by Iterative Mapping & Assembly of Short Reads to Eliminate Gaps. *Genome Biology*, 2010, **11**; r41

Uchiyama, I. MBGD: Microbial Genome Database for Comparative Analysis. *Nucleic Acids Research*, 2003, **31**; 58-62

Unicomb, L.; Ferguson, J.; Riley, T.V. & Collignon, P. Fluoroquinolone Resistance in *Campylobacter* Absent from Isolates, Australia. *Emerging Infectious Diseases*, 2003, **9**; 1482-1483

van Vliet, A.H.M.; Baillon, M-L.A.; Penn, C.W. & Ketley, J.M. *Campylobacter jejuni* Contains Two Fur Homologs: Characterisation of Iron-Responsive Regulation of Peroxide Stress Defence Genes by the PerR Repressor. *Journal of Bacteriology*, 1999, **181**; 6371-6376

Vandamme, P. & De Ley, J. Proposal for a New Family, *Campylobacteriaceae*. *International Journal of Systematic Biology*, 1991, **41**; 451-455

Vandamme, P.; Falsen, E.; Rossau, R.; Hoste, B.; Segers, P.; Tytgat, R. & De Ley, J. Revision of *Campylobacter*, *Helicobacter*, & *Wolinella* Taxonomy: Emendation of Generic Descriptions & Proposal of *Arcobacter* gen. nov. *International Journal of Systematic Bacteriology*, 1991, **41**; 88-103

Vandamme, P. & Goossens, H. Taxonomy of *Campylobacter*, *Arcobacter*, & *Helicobacter*: A Review. *Zentrallblatt für Bakteriologie*, 1992, **276**; 447-472

Vandamme, P.; Daneshvar, M.I.; Dewhirst, F.E.; Paster, B.J.; Kersters, K.; Goossens, H. & Moss, C.W. Chemotaxonomic Analyses of *Bacteriodes gracilis* & *Bacteroides ureolyticus* & Reclassification of *B. gracilis* as *Campylobacter gracilis* comb. nov. *International Journal of Systematic Bacteriology*, 1995, **45**; 145-152

Vellinga, A. & Van Loock, F. The Dioxin Crisis as Experiment to Determine Poultry-Related *Campylobacter* Enteritis. *Emerging Infectious Diseases*, 2002, **8**; 19-22

Vernikos, G.S. & Parkhill, J. Interpolated Variable Order Motifs for Identification of Horizontally Acquired DNA: Revisiting the *Salmonella* Pathogenicity Islands. *Bioinformatics*, 2006, **22**; 2196-2203

Véron, M. & Chatelain, R. Taxonomic Study of the Genus *Campylobacter* Sebald & Véron & Designation of the Neotype Strain for the Type Species, *Campylobacter fetus* (Smith & Taylor) Sebald & Véron. *International Journal of Systemic Bacteriology*, 1973, **23**; 122-134

Wainø, M.; Bang, D.D.; Lund, M.; Nordentoft, S.; Andersen, J.S.; Pedersen, K. & Madsen, M. Identification of Campylobacteria Isolated from Danish Broilers by Phenotypic Tests & Species-Specific PCR Assays. *Journal of Applied Microbiology*, 2003, **95**; 649-655

Wassenaar, T.M. Following an Imaginary *Campylobacter* Population from Farm to Fork & Beyond: a Bacterial Perspective. *Letters in Applied Microbiology*, 2011, **53**; 253-263

Wassenaar, T.M.; Bleumink-Pluym, N.M.C. & van der Zeijst, B.A.M. Inactivation of *Campylobacter jejuni* Flagellin Genes by Homologous Recombination Demonstrates that *flaA* but not *flaB* is required for Invasion. *The EMBO Journal*, 1991; **10**; 2055-2061

Wassenaar, T.M.; van der Zeijst, B.A.M.; Ayling, R. & Newell, D.G. Colonisation of Chicks by Motility Mutants of *Campylobacter jejuni* Demonstrates the Importance of Flagellin A Expression. *Journal of General Microbiology*, 1993, **139**; 1171-1175

Wheeler, J.G.; Sethi, D.; Cowden, J.M.; Wall, P.G.; Rodrigues L.C.; Tompkins, D.S.; Hudson, M.J. & Roderick, M.J. Study of Infectious Intestinal Disease in England: Rates in the Community, Presenting to General Practice, & Reported to National Surveillance. *British Medical Journal*, 1999, **318**; 1046-1050

Williams, N.J.; Jones, T.R.; Leatherbarrow, H.J.; Birtles, R.J.; Lahuerta-Marin, A.; Bennett, M. & Winstanley, C. Isolation of a Novel *Campylobacter jejuni* Clone Associated with the Bank Vole, *Myodes glareolus*. *Applied & Environmental Microbiology*, 2010, **76**; 7318-7321

Willems, R.J.L.; Top, J.; van Shalk, W.; Leavis, H.; Bonten, M.; Sirén, J.; Hanage, W.P. & Corander, J. Restricted Gene Flow among Hospital Subpopulations of *Enterococcus faecium*. *mBio*, 2012, **3**; e00151-12

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

Wilson, D.J.; Gabriel, E.; Leatherbarrow, A.J.H.; Cheesbrough, J.; Gee, S.; Bolton, E.; Fox, A.; Hart, C.A.; Diggle, P.J. & Fearnhead, P. Rapid Evolution & the Importance of Recombination to the Gastroenteric Pathogen *Campylobacter jejuni*. *Molecular Biology & Evolution*, 2009, **26**; 385-397

Wilson, G.G. Organisation of Restriction-Modification Systems. *Nucleic Acids Research*, 1991, **19**; 2539-2566

Wilson, I.G. & Moore, J.E. Presence of *Salmonella* spp. & *Campylobacter* spp. in Shellfish. *Epidemiology & Infection*, 1996, **116**; 147-153

Wilson, R.J.L.; Top, J.; van Schalk, W.; Leavis, H.; Bonten, M.; Sirén, J.; Hanage, W.P. & Corander, J. Restricted Gene Flow among Hospital Subpopulations of *Enterococcus faecium*. *MBio*, 2012, **3**; e00151-12

Winstanley, C. Spot the Difference: Applications of Subtractive Hybridisation to the Study of Bacterial Pathogens. *Journal of Medical Microbiology*, 2002, **51**; 459-467

Wren, B.W. Microbial Genome Analysis: Insights into Virulence, Host Adaptation & Evolution. *Nature Reviews Genetics*, 2000, **1**; 30-39

Yahara, K.; Didelot, X.; Ansari, M.A.; Sheppard, S.K. & Falush, D. Efficient Inference of Recombination Hot Regions in Bacterial Genomes. *Molecular Biology & Evolution*, 2014, **31**; 1593-1605

Yan, W.; Chang, N. & Taylor, D.E. Pulsed-Field Gel Electrophoresis of *Campylobacter jejuni* & *Campylobacter coli* Genomic DNA & its Epidemiologic Application. *Journal of Infectious Diseases*, 1991, **163**; 1068-1072

Young, K.T.; Davis, L.M. & DiRita, V.J. *Campylobacter jejuni*: Molecular Biology & Pathogenesis. *Nature Reviews Microbiology*, 2007, **5**; 665-679

Yu, S.; Fearnhead, P.; Holland, B.R.; Biggs, P.; Maiden, M. & French, N. Estimating the Relative Roles of Recombination & Point Mutation in the Generation of Single Locus Variants in *Campylobacter jejuni* & *C. coli*. *Journal of Molecular Evolution*, 2012, **74**; 273-280

Yuki, N. Infectious Origins of, & Molecular Mimicry in, Guillain-Barré & Fisher Syndromes. *The Lancet Infectious Diseases*, 2001, **1**; 29-37

Zerbino, D.R. & Birney, E. Velvet: Algorithms for De Novo Short Read Assembly using de Bruijn Graphs. *Genome Research*, 2008, **1**; 821-829

Zhou, Y.; Liang, Y.; Lynch, K.H.; Dennis, J.J. & Wishart, D.S. PHAST: A Fast Phage Search Tool. *Nucleic Acids Research*, 2011, **39**; W347-W352

Ziprin, R.L.; Young, C.R.; Byrd, J.A.; Stanker, L.H.; Hume, M.E.; Gray, S.A.; Kim, B.J. & Konkel, M.E. Role of *Campylobacter jejuni* Potential Virulence Genes in Cecal Colonization. *Avian Diseases*, 2001, **45**; 549-557

Zilbauer, M.; Dorrell, N.; Wren, B.W. & Bajaj-Elliot, M. *Campylobacter jejuni*-Mediated Disease Pathogenesis: An Update. *Transactions of the Royal Society of Tropical Medicine & Hygiene*, 2008, **102**; 123-129

Zweifel, C.; Zychowska, M.A. & Stephan, R. Prevalence & Characteristics of Shiga Toxin-Producing *Escherichia coli*, *Salmonella* spp. & *Campylobacter* spp. Isolated from Slaughtered Sheep in Switzerland. *International Journal of Food Microbiology*, 2004, **92**; 45-53

Appendix 9.1 Isolate Information Table

Species	Isolate	Date	Country	Source	Accession	ST	ST CC	CDS Prefix	Included in				
									Chapter 2: Phenotyping	Chapter 3: Phylogeny	Chapter 4: Genome Content	Chapter 5: Prophage	Chapter Six: Recombination
C. jejuni jejuni	857	2000	UK	Pig	N/A	270	403	CJ857	✓	✓	✓	✓	✓
C. jejuni jejuni	549.1	1999	UK	Pig	N/A	403	403	CJ857	✓	✓	✓	✓	✓
C. jejuni jejuni	623	1999	UK	Pig	N/A	552	403	CJ857	✓	✓	✓	✓	✓
C. jejuni jejuni	304	1999	UK	Pig	N/A	551	403	CJ857	✓	✓	✓	✓	✓
C. jejuni jejuni	484	1999	UK	Pig	N/A	435	403	CJ857	✓	✓	✓	✓	✓
C. jejuni jejuni	444	1999	UK	Pig	N/A	553	403	CJ857	✓	✓	✓	✓	✓
C. jejuni jejuni	ATCC33560	1970	Belgium	Cow	380625642	403	403	N/A				✓	✓
C. jejuni jejuni	81116	1981	UK	Human	CP000814.1	267	283	C8J_	✓	✓	✓		✓
C. jejuni jejuni	11168	1977	UK	Human	AL111168.1	43	21	Cj	✓	✓	✓		✓
C. jejuni jejuni	1336	-	-	Water	CM000854.1	841*	~	C1336_		✓	✓		✓
C. jejuni jejuni	414	2000	UK	Water Vole	CM000855.1	3704	~	C414_		✓	✓		✓
C. jejuni jejuni	81-176	1981	USA	Milk	CP000538.1	604*	42*	CJJ81176_		✓	✓		✓
C. jejuni jejuni	S3	1999	UK	Sheep	CP001960.1	632	42	CJS3_		✓	✓		✓
C. jejuni jejuni	IA3902	2006	USA	Sheep	CP001876.1	8	21	CJSA_		✓	✓		✓
C. jejuni jejuni	ICDCCJ07001	2007	China	Human	CP002029.1	2993	362	ICDCCJ07001_		✓	✓		✓
C. jejuni jejuni	M1	1999	UK	Human	CP001900.1	137	45	CJM1_		✓	✓		✓
C. jejuni jejuni	01/10	-	UK	Human	-	104*	21*	CJ0110_		✓	✓		✓
C. jejuni jejuni	01/51	-	UK	Human	-	19*	21*	CJ0110_		✓	✓		✓
C. jejuni jejuni	RM1221	1997	USA	Chicken	CP000025.1	354	354	CJE		✓	✓		✓
C. jejuni doylei	269.97	1997	South Africa	Human	CP000768.1	1845	~	JJD26997_		✓	✓		✓
C. coli	RM2228	1998	USA	Chicken	AAFL00000000	1063	828	CCO		✓	✓		✓
C. coli	JV20	-	-	Human	AEER00000000	860*	828*	HMPREF9399_		✓	✓		✓
C. coli	111-3	2001	USA	Pig	AIMI00000000	1467*	828*	cco1_		✓	✓		✓
C. coli	132-6	2001	USA	Pig	AINA00000000	3861*	~	cco5_		✓	✓		✓
C. coli	151-9	2001	USA	Pig	AINQ00000000	1102*	~	cco8_		✓	✓		✓
C. coli	59-2	2000	USA	Pig	AIND00000000	890*	828*	cco6_		✓	✓		✓
C. coli	67-8	2000	USA	Pig	AINI00000000	1061*	828*	cco7_		✓	✓		✓
C. coli	7—1	2001	USA	Pig	AIMZ00000000	3860*	~	cco4_		✓	✓		✓
C. coli	84-2	2000	USA	Pig	AIMS00000000	113*	828*	cco12_		✓	✓		✓
C. coli	90-3	2001	USA	Pig	AIMJ00000000	3862*	~	cco10_		✓	✓		✓
C. coli	99/321	1999	Denmark	Pig	N/A	1153*	828*	CC03121_	✓	✓	✓	✓	✓
C. coli	03/121	2003	UK	Pig	N/A	887*	828*	CC03121_	✓	✓	✓	✓	✓
C. coli	03/103	2003	UK	Pig	N/A	2732*	828*	CC03121_	✓	✓	✓	✓	✓
C. coli	03/317	2003	UK	Pig	N/A	1145*	828*	CC03121_	✓	✓	✓	✓	✓

Details of isolates used in this study. - information not available. ♦MLST/CC determined *in silico* . ~Clonal Complex not assigned

		C. jejuni 81116	*C. jejuni* 11168	*C. jejuni* 857	*C. jejuni* 549.1	*C. jejuni* 623	*C. jejuni* 304	*C. jejuni* 484	*C. jejuni* 444	*C. coli* 99/321	*C. coli* 03/121	*C. coli* 03/103	*C. coli* 03/317
Motility Diameter (mm)	Minimum	19	21	18	6	3	29	13	9	3	3	6	9
	Maximum	19	21	24	6	3	31	15	17	4	3	18	13
	Mean	19	21	21.67	6	3	30	14	13	3.333	3	11.67	11
	Std. Deviation	0	0	3.215	0	0	1.414	1.414	4	0.5774	0	6.028	2.828
Autoagglutination (change in O600nm)	Minimum	0.245	0.229	0.6	0.716	0.7	0.289	0.805	0.567	0.67	0.482	0.648	0.816
	Maximum	0.285	0.444	0.667	0.717	0.818	0.496	1.028	0.656	0.731	0.502	0.726	0.993
	Mean	0.2607	0.3543	0.643	0.7163	0.7677	0.362	0.896	0.6133	0.7053	0.4887	0.674	0.9157
	Std. Deviation	0.02136	0.1118	0.03732	0.0005773	0.06088	0.1162	0.117	0.04461	0.03163	0.01155	0.04503	0.09059
Percentage Adhesion Efficiency	Minimum	2.08	0.3567	1.151	1.1	0.1931	2	1.25	1.152	0.6255	1.333	1.35	2.186
	Maximum	2.83	0.7833	1.395	1.166	0.2863	4.667	1.415	1.461	2.498	2.333	2.392	3.098
	Mean	2.455	0.57	1.273	1.133	0.2397	3.333	1.333	1.306	1.562	1.833	1.871	2.642
	Std. Deviation	0.5303	0.3017	0.1723	0.04667	0.06586	1.886	0.1167	0.2183	1.324	0.7071	0.7366	0.6453
Percentage Invasion Efficiency	Minimum	0.003415	0.161	0.001304	0.00466	0.00279	0.01333	0.02085	0.002456	0.009363	0.0004113	0.003003	0.02005
	Maximum	0.004835	0.2557	0.001364	0.005	0.003721	0.01653	0.0225	0.002535	0.009363	0.0004333	0.003473	0.02552
	Mean	0.004125	0.2083	0.001334	0.00483	0.003255	0.01493	0.02168	0.002495	0.009363	0.0004223	0.003238	0.02279
	Std. Deviation	0.001004	0.06694	4.24E-05	0.0002404	0.0006585	0.002263	0.001167	5.54E-05	0	1.56E-05	0.0003323	0.003864
Zone of Inhibition (mm) 3% Hydrogen Peroxide	Minimum	24	29	25	22	30	15	24	26	28	16	25	20
	Maximum	26	35	27	24	34	24	28	28	28	20	28	25
	Mean	25	31.67	25.67	23	31.67	20	26.33	27	28	18	26.33	22.67
	Std. Deviation	1	3.055	1.155	1	2.082	4.583	2.082	1	0	2	1.528	2.517
Zone of Inhibition (mm) 3% Pyrogallol	Minimum	35	30	30	32	38	29	31	36	30	30	29	35
	Maximum	41	36	34	55	50	44	42	40	36	34	43	39
	Mean	37.6	32.4	31.8	43.2	43.8	35.6	35.8	37.4	32	32.2	38.4	36.8
	Std. Deviation	2.408	2.881	1.789	8.701	5.02	5.771	5.263	1.949	2.55	1.483	5.727	1.483
Zone of Inhibition (mm) Cholic Acid	Minimum	7	8	7	7	8	10	8	6	6	6	5	8
	Maximum	8	9	9	8	9	13	10	7	8	7	7	8
	Mean	7.667	8.333	7.667	7.667	8.333	12	9	6.667	7	6.667	6	8
	Std. Deviation	0.5774	0.5774	1.155	0.5774	0.5774	1.732	1	0.5774	1	0.5774	1	0
Zone of Inhibition (mm) Deoxycholic Acid	Minimum	5	6	7	7	6	7	7	7	6	6	6	6
	Maximum	7	8	8	8	7	8	7	10	8	7	8	7
	Mean	6.333	7	7.667	7.667	6.667	7.667	7	8	7	6.667	7	6.667
	Std. Deviation	1.155	1	0.5774	0.5774	0.5774	0.5774	0	1.732	1	0.5774	1	0.5774
Zone of Inhibition (mm) Chenodeoxy cholic Acid	Minimum	6	6	5	6	5	7	6	5	5	6	7	8
	Maximum	6	7	6	7	6	8	7	8	6	7	8	10
	Mean	6	6.333	5.333	6.333	5.333	7.333	6.667	6.667	5.667	6.333	7.667	9
	Std. Deviation	0	0.5774	0.5774	0.5774	0.5774	0.5774	0.5774	1.528	0.5774	0.5774	0.5774	1

Overview of phenotyping results

CDS	Cjj 1336 Wild Bird	Cjj 414 Water Vole	Cjj 52 Sheep	Cjj IA3902 Sheep	Cjd 26997 Human Blood	Cjj 81-176 Human (Milk)	Cjj ICDCC10 7001 Human GBS Stool	Cjj 81116 Human (Water)	Cjj M1 Human (Chicken)	Cjj 1116B (Human)	Cjj 0130 Human	Cjj 0151 Human	Cc JV20 Human (Microbiome)	Cjj RM1221 Chicken	Cc RM2228 Chicken	Cc 111-3 Pig	Cc 132-6 Pig	Cc 151-9 Pig	Cc 59-2 Pig	Cc 67-8 Pig	Cc 7--1 Pig	Cc 84-2 Pig	Cc 90-3 Pig	Cc 99321 Pig	Cc 03121 Pig	Cc 03103 Pig	Cc 03312 Pig	Cjj 857 Pig	Cjj 5491 Pig	Cjj 623 Pig	Cjj 304 Pig	Cjj 484 Pig	Cjj 444 Pig
CJJ0672																																	
putative periplasmic protein	✔	-	✔	✔	-	✔	-	✔	-	✔	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CJJ0246c																																	
putative MCP-domain signal transduction protein	-	-	✔	✔	-	✔	-	-	-	✔	✔	✔	-	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CJJ0344																																	
hypothetical protein CJJ0344	✔	-	✔	✔	✔	✔	-	✔	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CJE1384																																	
hypothetical protein	✔	-	✔	✔	✔	✔	-	-	-	-	-	-	-	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CJJ0556																																	
putative amidohydrolase family protein	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	✔		✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CJJ0555																																	
putative dicarboxylate carrier protein MatC	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	✔		✔	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-
CJJ 0033																																	
hypothetical protein	✔	✔	-	-	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CJJ0417																																	
hypothetical protein CJJ0417	✔	-	-	✔	-	✔	-	✔	✔	✔	✔	✔		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CJJ158c																																	
small hydrophobic protein	-	-	✔	-	-	✔	✔	✔	✔	✔	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CJJ 1559																																	
hypothetical protein	✔	-	-	-	-	✔	-	✔	✔	-	✔	✔		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HMPREF9399_0599																																	
conserved hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	✔	-	✔	✔	-	-	-	✔	✔	✔	✔	✔	-	-	-	-	-	-
HMPREF9399_0256																																	
conserved hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	-	✔	-	-	-	✔	✔	-	✔	✔	-	-	-	-	-	-	-
CCO1311																																	
hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	-	✔	-	-	-	✔	-	-	✔	✔	-	-	-	-	-	-	-
cco4_04499																																	
hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	-	✔	✔	✔	-	-	✔	✔	✔	-	-	-	-	-	-
cco4_04519																																	
hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	-	✔	✔	✔	-	-	✔	✔	-	-	-	-	-	-	-
cco4_04524																																	
hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	-	✔	✔	✔	-	-	✔	✔	-	-	-	-	-	-	-
CCO0280																																	
methyl-accepting chemotaxis protein (tlpA)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	-	-	✔	✔	-	-	-	-	-	-	-
CCO1740																																	
Hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	-	✔	✔	✔	-	-	-	-	✔	✔	-	-	-	-	-	-	-
CCO1688															✔		✔	-	-	-	-	✔	-	✔	-	✔	✔	-	-	-	-	-	-	-
hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-																				
CCO0184																																	
Hemolysin, putative	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-
cco1_04927																																	
hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	-	-	-	-	-	-	-	-
CCO1041																																	
conserved hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	-	-	✔	-	-	-	✔	✔	✔	✔	✔	-	-	-	-	-	-	-
CCO1040																																	
conserved hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	-	-	✔	-	-	-	✔	✔	✔	✔	✔	-	-	-	-	-	-	-
cco1_08394																																	
hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	-	✔	-	-	-	✔	✔	-	-	-	-	-	-	-	-
HMPREF9399_0022																																	
conjugal transfer protein TrbD	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	-	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-
HMPREF9399_0023																																	
conjugal transfer pilus assembly protein TrbC	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-
HMPREF9399_0024																																	
conjugative transfer protein TrbB	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-
HMPREF9399_0013																																	
conserved hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-
HMPREF9399_0012																																	
conjugal transfer protein TrbI	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-
HMPREF9399_0011																																	
conjugal transfer protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-
cco5_00320																																	
hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	✔	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-
HMPREF9399_0010																																	
single-strand binding protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-
HMPREF9399_0009																																	
conjugal transfer protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-
HMPREF9399_0008																																	
DNA primase TraC	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-
HMPREF9399_0007																																	
conjugal transfer protein TraG	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-
HMPREF9399_0006																																	
conjugal transfer pilus assembly protein TraL	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-
cco1_07549																																	
hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	✔	✔	✔	✔	-	-	✔	✔	✔	-	-	-	-	-	-	-
CCO1639																																	
ISCco1, transposase orfA	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	✔	-	-	✔	-	✔	✔	-	-	-	-	-	-	-	-
CCO0036																																	
Chain B, Stru	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	-	-	-	-	-	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-
HMPREF9399_0018																																	
P-type conjugative transfer protein TrbJ	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	-	✔	✔	-	-	-	-	-	-	-	-
HMPREF9399_0031																																	
TraJ protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	-	✔	✔	-	-	-	-	-	-	-	-
HMPREF9399_0032																																	
relaxase/mobilization nuclease domain protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	-	✔	✔	-	-	-	-	-	-	-	-
HMPREF9399_0033																																	
hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-
HMPREF9399_0034																																	
plasmid partition ParA protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-
HMPREF9399_0035																																	
conjugal transfer protein TraM	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-
HMPREF9399_0036																																	
plasmid transfer protein TraF	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-
CCO0533																																	
membrane protein , putative	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	-	-	-	-	-	-	-
cco4_06586																																	
ferrous iron transport protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	-	✔	-	-	-	✔	✔	-	-	-	-	-	-	-	-
cco4_03101																																	
hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	✔	✔	✔	-	✔	✔	✔	✔	-	-	-	-	-	-	-
HMPREF9399_1768																																	
flagellin	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-
HMPREF9399_1770																																	
motility accessory factor	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	✔	-	✔	✔	-	-	✔	✔	-	-	-	-	-	-	-	-	-	-	-
cco1_08991																																	
methyl-accepting chemotaxis signal transduction protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	-	✔	-	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-
CCO1784																																	
hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-
cco1_08404																✔	✔	✔	✔	✔	✔													
hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-							-	-	-	-	-	-	-	-	-	-	-	-	-
CCO1173																																	
hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-

CDS	Cj 1136 Wild Bird	Cj 414 Water Vole	Cj 53 Sheep	Cj IA3902 Sheep	Cjst 26997 Human Blood	Cj 81-176 Human (Milk)	Cj ICDCC07 001 Human GBS Stool	Cj 81116 Human (Water)	Cj M1 Human (Chicken)	Cj 11168 (human)	Cj 0110 Human	Cj 0151 Human	Cc JV20 Human (Microbiome)	Cj RM1221 Chicken	Cc RM2228 Chicken	Cc 111-3 Pig	Cc 132-6 Pig	Cc 151-9 Pig	Cc 59-2 Pig	Cc 67-8 Pig	Cc 7-1 Pig	Cc 84-2 Pig	Cc 90-3 Pig	Cc 99321 Pig	Cc 03121 Pig	Cc 03131 Pig	Cc 03317 Pig	Cj 857 Pig	Cj 5491 Pig	Cj 623 Pig	Cj 364 Pig	Cj 484 Pig	Cj 444 Pig
CCO1076 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-
CCO1054 conserved hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	-	✔	✔	✔	✔	-	-	-	✔	-	✔	-	-	-	-	-	-
CCO1183 thioredoxin, putative	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	-	✔	✔	-	✔	-	✔	✔	✔	-	-	-	-	-	-	-	-
CCO0066 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-
CCO1105 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	-	-	✔	-	✔	✔	✔	✔	-	✔	-	-	-	-	-	-
CCO1249 conserved hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-
cjo1_06747 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-
CCO1587 DnaJ-related protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	-	✔	✔	✔	-	✔	-	-	-	-	-	-	-	-
cjo1_04490 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	-	-	✔	✔	-	✔	-	✔	-	-	-	-	-	-
cjo1_04020 DNA adenine methylase	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	-	✔	✔	✔	✔	-	-	✔	-	-	-	-	-	-	-	-
cjo1_04029 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	-	✔	✔	✔	✔	-	-	✔	-	-	-	-	-	-	-	-
cjo1_04030 DNA methyltransferase	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	-	✔	✔	✔	✔	-	-	✔	-	-	-	-	-	-	-	-
CCO1671 catalase-like protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	-	✔	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
cjo1_05605 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	✔	✔	✔	✔	-	✔	-	-	✔	-	-	-	-	-	-
CCO0968 conserved hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	-	✔	-	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-
HMPREF0099_0017 conserved hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	✔	-	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-
cjo1_03616 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	-	-	✔	✔	✔	-	-	✔	-	-	-	-	-	-
CCO1514 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	✔	-	-	-	✔	✔	✔	-	-	-	-	-	-	-	-	-	-
CCO1191 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-
CCO0017 Chain B, Sru	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	-	-	✔	-	-	-	-	-	-	-	-
cjo1_00200 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-
CCO1307 conserved hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	-	✔	✔	✔	-	-	✔	✔	-	-	-	-	-	-	-	-	-	-
HMPREF0099_0018 conserved hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	✔	✔	-	✔	-	✔	✔	-	✔	-	-	-	-	-	-	-	-	-
CCO0935 Tat (twin-arginine translocation) pathway signal sequence domain	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	-	✔	-	-	-	-	-	-	-	-	-	-	-
CJ0960 putative integral membrane protein	-	-	✔	-	✔	✔	-	✔	✔	✔	-	-	✔	✔	✔	-	-	✔	✔	-	-	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CJ0818 putative lipoprotein	-	-	-	✔	-	-	-	-	-	✔	✔	✔	✔	-	✔	✔	-	✔	✔	-	-	✔	✔	✔	✔	✔	-	-	-	-	-	-	-
CCO1339 ISCon1, transposase	-	-	✔	-	-	-	✔	-	-	-	✔	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CJ0 0705 hypothetical protein	-	-	-	-	-	✔	-	✔	✔	-	✔	-	✔	-	✔	-	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CCO1269 hypothetical protein	-	✔	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-
CCO1281 cobalamin synthesis protein/P47K family protein	-	-	-	-	✔	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CCO1280 conserved hypothetical protein	-	-	-	-	✔	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CCO1279 polysaccharide deacetylase family protein	-	-	-	-	✔	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-
CCO1278 hydrolase, carbon-nitrogen family	-	-	-	-	✔	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CJ1270c putative integral membrane protein	✔	✔	✔	✔	-	✔	-	-	-	✔	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CJ1 1076 hypothetical protein	-	-	-	-	-	-	-	✔	✔	-	-	-	-	-	-	-	-	✔	✔	-	✔	✔	✔	-	-	✔	✔	-	-	-	-	-	-
CJ1 1077 c	-	-	-	-	-	-	-	✔	✔	-	-	-	-	-	-	-	-	✔	✔	-	✔	✔	✔	-	-	✔	✔	-	-	-	-	-	-
CJ1 1078 hypothetical protein	-	-	-	-	-	-	-	✔	✔	-	-	-	-	-	-	-	-	✔	✔	-	✔	✔	✔	-	-	✔	✔	-	-	-	-	-	-
CJ1 1080 hypothetical protein	-	-	-	-	-	-	-	✔	-	-	-	-	-	-	-	-	-	✔	✔	-	✔	✔	✔	-	-	✔	✔	-	-	-	-	-	-
CJ1 1082 hypothetical protein	-	-	-	-	-	-	-	✔	✔	-	-	-	-	-	-	-	-	✔	✔	-	✔	✔	✔	-	-	✔	✔	✔	-	-	-	-	-
CJ1 1084 hypothetical protein	-	-	-	-	-	-	-	✔	✔	-	✔	-	-	-	-	-	-	✔	✔	-	✔	✔	✔	-	-	✔	✔	✔	-	-	-	-	-
CJ1 1090 hypothetical protein	-	-	-	-	-	-	-	✔	✔	-	-	-	✔	-	-	✔	✔	✔	-	✔	✔	✔	-	-	-	✔	✔	-	-	-	-	-	-
CJ1375 putative glycosyltransferase	-	-	-	✔	-	-	-	✔	✔	✔	-	-	✔	-	-	✔	✔	✔	-	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-	-
CJE0051 hypothetical protein	-	-	✔	-	-	-	-	-	-	-	-	-	✔	✔	✔	-	-	-	✔	✔	✔	-	-	-	✔	✔	✔	-	-	-	-	-	-
CJ0826c restriction modification enzyme	-	-	✔	✔	-	-	✔	✔	✔	✔	-	-	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CJE179 biopolymer transport protein	-	-	✔	✔	-	-	✔	-	-	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CJE181 TonB transport protein	-	-	✔	✔	-	-	✔	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CJ1 1459 arsenical-resistance protein, putative	-	-	✔	-	-	-	-	✔	-	-	✔	-	✔	✔	✔	-	✔	✔	-	-	-	-	✔	✔	✔	✔	✔	-	-	-	-	-	-
CJ1 1458 arsenate reductase	-	-	✔	-	-	-	-	✔	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
cjo5_04288 hypothetical protein	-	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	-	✔	-	✔	✔	✔	✔	-	-	-	-	-	-
CCO0182 hemagglutinin/hemolysin-related protein	-	✔	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CCO0137 transcriptional regulator, Crp family, putative	-	-	-	-	-	-	✔	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CJ1 0988 hypothetical protein	-	✔	-	-	✔	-	-	✔	✔	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
cjo1_06184 CHT protein	-	-	-	-	✔	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔	-	-	-	✔	✔	✔	-	-	-	-	-	-
cjo1_06409 GTP-binding protein HflX	-	-	-	-	✔	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	-	-	-	✔	✔	-	-	-	-	-	-	-
CCO0168 conserved hypothetical protein	-	✔	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	-	-	-	✔	✔	✔	-	-	-	-	-	-
CCO0171 conserved hypothetical protein	-	✔	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	-	-	-	✔	✔	✔	-	-	-	-	-	-
CCO0017 hypothetical protein	-	✔	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	-	-	-	✔	✔	✔	-	-	-	-	-	-
HMPREF0099_0027 conserved hypothetical protein	-	✔	-	-	✔	-	-	-	-	-	-	-	✔	-	-	-	-	✔	-	✔	✔	✔	✔	-	✔	✔	-	-	-	-	-	-	-
CCO0048 hypothetical protein	-	-	-	-	✔	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	-	-	-	✔	✔	-	-	-	-	-	-	-
CJE1103 conserved hypothetical protein	-	-	-	-	✔	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	✔	✔	-	-	-	-	-	-	-
CJE0189 membrane protein, putative	-	-	✔	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CJE0188 membrane protein, putative	-	-	✔	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CJE0187 membrane protein, putative	-	-	✔	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CCO0413 tricarboxylate transport protein TctA, putative	-	-	-	-	✔	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CCO0412 tricarboxylate transport protein TctB, putative	-	-	-	-	✔	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CCO0411 tricarboxylate transport protein TctA, putative	-	-	-	-	✔	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CJE1127 hypothetical protein	-	✔	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	-	-	✔	-	✔	-	-	-	✔	✔	✔	-	-	-	-	-	-
CJE1126 hypothetical protein	-	✔	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	-	-	✔	-	✔	-	-	-	✔	✔	✔	-	-	-	-	-	-
CJE1884 lipopolysaccharide core biosynthesis protein, putative	-	✔	✔	-	-	-	-	-	-	-	-	-	✔	✔	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	-	-	-	-	-	-
CJ1723c putative periplasmic protein	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CJ1725 putative periplasmic protein	-	-	✔	-	✔	✔	✔	-	-	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CJ1726c homoserine O-succinyltransferase	-	-	✔	-	-	✔	✔	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-
CCO0026 YeaF/YedE family protein	-	-	-	-	✔	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-	-

CDS	Cjj 1336 Wild Bird	Cjj 414 Water Vole	Cjj S3 Sheep	Cjj IA3902 Sheep	Cjd 26997 Human Blood	Cjj 81-176 Human (Milk)	Cjj ICDCCJ07001 Human GBS Stool	Cjj 81116 Human (Water)	Cjj M1 Human (Chicken)	Cjj 11168 (Human)	Cjj 0110 Human	Cjj 0151 Human	Cc JV20 Human (Microbiome)	Cjj RM1221 Chicken	Cc RM2228 Chicken	Cc 111-3 Pig	Cc 132-6 Pig	Cc 151-9 Pig	Cc 59-2 Pig	Cc 67-8 Pig	Cc 7--1 Pig	Cc 84-2 Pig	Cc 90-3 Pig	Cc 99321 Pig	Cc 03121 Pig	Cc 03103 Pig	Cc 03317 Pig	Cjj 857 Pig	Cjj 5491 Pig	Cjj 623 Pig	Cjj 304 Pig	Cjj 484 Pig	Cjj 444 Pig
Cj1727c putative O-acetylhomoserine (thiol)-lyase	-	-	✔	-	✔	✔	✔	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-
CCO0039 sodium:solute symporter family protein	-	-	-	-	✔	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CCO0040 Protein of unknown function, DUF485 superfamily	-	-	-	-	✔	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
Cj0486 putative sugar transporter	✔	-	✔	✔	✔	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-
Cj0488 conserved hypothetical protein Cj0488	✔	-	✔	✔	✔	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-
Cj0487 putative amidohydrolase	✔	-	✔	✔	✔	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-
Cj0485 putative oxidoreductase	✔	-	✔	✔	✔	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-
Cj0484 putative MFS (Major Facilitator Superfamily) transport protein	✔	-	✔	✔	-	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-
Cj0483 putative altronate hydrolase C-terminus	✔	-	✔	✔	✔	-	-	-	-	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-
Cj0482 putative altronate hydrolase N-terminus	✔	-	✔	✔	-	-	-	-	-	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-
Cj0481 putative dihydrodipicolinate synthase	✔	-	✔	✔	✔	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-
Cj0480c putative transcriptional regulator	✔	-	✔	✔	✔	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-
cco5_00550 mannose-1-phosphate guanylyltransferase/mannose-6-phosphate isomerase	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	-	-	-	-	✔	✔	✔	-	-	-	-	-	-	-
cco5_00545 GDP-D-mannose dehydratase	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	-	-	-	-	✔	✔	✔	-	-	-	-	-	-	-
cco1_04120 WbdK	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-	✔	✔	-	✔	-	-	-	-	✔	✔	✔	-	-	-	-	-	-	-
Cj0122 hypothetical protein Cj0122	-	-	✔	✔	-	-	-	-	-	✔	-	✔	✔	✔	✔	-	-	-	-	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-	-
CJE0140 methyl-accepting chemotaxis protein	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	-	-	-	✔	✔	-	-	-	-	-	-
cco5_04298 helicase, putative	-	-	-	-	-	-	✔	-	-	-	-	-	-	-	-	-	✔	-	-	-	-	✔	-	✔	✔	✔	✔	-	-	-	-	-	-
cco5_04303 hypothetical protein	✔	-	-	-	-	-	✔	-	-	-	-	-	-	-	-	-	✔	-	-	-	-	✔	-	✔	✔	✔	✔	-	-	-	-	-	-
CCO1658 major facilitator family transporter	-	✔	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
CJE1429 conserved hypothetical protein	-	✔	✔	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	-	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CCOA0077 hypothetical protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-
CJE0555 hypothetical protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0554 hypothetical protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0553 hypothetical protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0550 site-specific DNA-methyltransferase, putative	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	✔	-	✔	-	-	-	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
Cj1321 putative transferase	-	-	-	-	-	-	-	-	-	✔	✔	-	✔	-	-	✔	✔	✔	✔	-	-	✔	-	✔	✔	-	-	-	-	-	-	-	-
Cj1322 hypothetical protein Cj1322	-	-	-	-	-	-	-	-	-	✔	✔	-	✔	-	-	✔	✔	✔	✔	-	-	✔	-	✔	✔	-	-	-	-	-	-	-	-
Cj1323 hypothetical protein Cj1323	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	-	-	✔	✔	✔	✔	-	-	✔	-	✔	✔	-	-	-	-	-	-	-	-
CJE0587 hypothetical protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0588 hypothetical protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0589 hypothetical protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE1103 conserved hypothetical protein	-	-	-	-	✔	-	-	-	-	-	-	-	-	✔	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-
Cj0070c hypothetical protein Cj0070c	-	-	-	-	✔	-	-	✔	-	✔	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-
Cj1309c hypothetical protein Cj1309c	-	✔	-	✔	✔	-	-	-	-	✔	✔	✔	✔	-	✔	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CJE1445 hypothetical protein	✔	✔	✔	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0598 hypothetical protein	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0599 hypothetical protein	✔	✔	-	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0600 hypothetical protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0601 hypothetical protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
Cj0299 putative periplasmic beta-lactamase	✔	-	✔	✔	-	-	-	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-
CJE0732 type III restriction-modification enzyme	-	-	✔	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	-	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-
CJE0731 type III restriction/modification enzyme, methylase subunit	-	-	✔	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	-	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-
CJE0574 hypothetical protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0575 hypothetical protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0576 conserved domain protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0577 terminase B protein, putative	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0578 conserved domain protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0579 hypothetical protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0580 hypothetical protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0581 hypothetical protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0582 hypothetical protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	-	✔	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
Cj1591 50S ribosomal protein L36	-	-	✔	✔	-	✔	-	✔	-	✔	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-
CCO1298 sodium/pantothenate symporter , putative	-	✔	-	-	✔	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-
Cj1208 putative 5-formyltetrahydrofolate cyclo-ligase family protein	✔	✔	✔	✔	-	-	-	✔	✔	✔	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-
CJE1611 GDP-mannose 4,6-dehydratase	-	-	✔	-	✔	✔	✔	-	-	-	✔	-	✔	✔	-	✔	-	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CCOA0206 tetracycline resistance protein	-	-	-	✔	-	-	-	-	-	-	✔	-	-	-	✔	✔	✔	✔	✔	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-

CDS	Cj 1336 Wild Bird	Cj 414 Water Vole	Cj 53 Sheep	Cj IA3902 Sheep	Cjd 28997 Human Blood	Cj 81-176 Human (Milk)	Cji ICDCC07 001 Human GBS Stool	Cj 81116 Human (Water)	Cj M1 Human (Chicken)	Cj 11168 (Human)	Cj 0130 Human	Cj 0151 Human	Cc JV20 Human (Microbiome)	Cji RM1221 Chicken	Cc RM3228 Chicken	Cc 111-3 Pig	Cc 132-6 Pig	Cc 151-9 Pig	Cc 59-2 Pig	Cc 67-8 Pig	Cc 7--1 Pig	Cc 84-2 Pig	Cc 90-3 Pig	Cc 99321 Pig	Cc 03121 Pig	Cc 03101 Pig	Cc 03137 Pig	Cj 857 Pig	Cj 5491 Pig	Cj 623 Pig	Cj 304 Pig	Cj 484 Pig	Cj 444 Pig
CJ1560 putative permease	-	✔	✔	✔	-	✔	-	✔	-	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	-	-	-	-	-	-	-	-
CJ0490 putative aldehyde dehydrogenase C-terminus	✔	-	✔	✔	✔	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	-	✔	-	-	-	-	-	-
CJ0819 small hydrophobic protein	-	-	-	✔	✔	-	-	-	-	✔	-	-	✔	-	✔	✔	-	✔	✔	-	-	✔	✔	-	-	-	-	-	-	-	-	-	-
CJ1043c putative transcriptional regulatory protein	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	✔	-	✔	✔	✔	✔	-	✔	-	-	-	-	-	-
CJ0177 putative iron transport protein	-	-	✔	✔	-	-	✔	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	-	-	-	-	-	-
CJ0178 putative TonB-dependent outer membrane	-	-	✔	✔	-	-	✔	-	-	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	-	✔	-	-	-	-	-	-
CCO1689 hypothetical protein	✔	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	-	-	-	-	-	-
CJ0768 hypothetical protein CJ0768	-	-	✔	✔	-	-	-	-	-	✔	-	-	-	-	✔	✔	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-	-	-	-	-
HMPREF9099_0919 antirepressor	-	-	-	-	✔	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	-	✔	-	-	-	-	-	-	-	-
HMPREF9099_0926 conserved hypothetical protein	-	✔	-	-	✔	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	✔	-	✔	-	-	-	-	-	-	-	-
CCOA0076 hypothetical protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	✔	-	✔	-	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-
cjc4_08036 hypothetical protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-
CJ0967 putative periplasmic protein	-	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	✔	-	✔	-	-	✔	-	-	-	✔	✔	-	✔	-	-	-	-	-	-
CJ1 0987 AlkA	-	✔	-	-	✔	-	-	✔	✔	-	-	-	✔	-	-	-	✔	-	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-
CJ1 0986 hypothetical protein	-	✔	-	-	✔	-	-	✔	✔	-	-	-	✔	-	-	-	✔	-	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-
CJ1 0985 hypothetical protein	-	✔	-	-	-	-	-	✔	✔	-	-	-	✔	-	-	-	✔	-	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-
CJ0 1 1376 NAD-dependent epimerase/dehydratase	-	-	-	-	-	-	✔	-	✔	-	-	-	✔	-	-	-	-	✔	-	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-
CJ0 1 1380 Glucose-1-phosphate cytidylyltransferase	✔	-	-	-	-	-	✔	-	✔	-	-	-	✔	-	-	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-
CJ0144 putative methyl-accepting chemotaxis signal transduction protein	-	-	✔	-	-	✔	✔	✔	✔	✔	-	-	✔	✔	-	-	-	✔	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJ1 1349 hypothetical protein	-	✔	-	-	✔	-	-	✔	✔	-	-	-	-	-	✔	-	-	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CJ0811N 0645 hypothetical protein	-	-	-	-	✔	✔	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-
CJE0473 hypothetical protein	✔	-	✔	-	✔	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	-	-	-	-	✔	-	-	-	-	-	-	-	-
CJE1101 hypothetical protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	✔	-	-	✔	-	-	-	-	-	-	✔	-	-	-	-	-	-	-	-
CJ1 0928 hypothetical protein	-	-	-	-	-	✔	✔	✔	✔	-	-	-	✔	-	-	-	-	✔	-	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-
CJE0466 hypothetical protein	✔	-	✔	-	✔	-	✔	-	-	-	-	-	✔	✔	✔	-	-	-	✔	-	-	-	-	-	-	-	✔	-	-	-	-	-	-
CJ0424 putative acidic galactoside protease	-	-	-	✔	-	-	✔	-	-	✔	✔	✔	-	-	-	-	-	-	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CJ0425 putative periplasmic protein	-	-	-	✔	-	-	✔	-	-	✔	✔	✔	-	-	-	-	-	-	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CCO1213 glycosyl transferase, group 1 family protein	✔	✔	-	-	✔	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJ0 0644 hypothetical protein	-	✔	-	-	-	-	-	✔	✔	-	-	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0235 bacteriophage DNA transposition protein A, putative	-	✔	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0269 bacteriophage DNA transposition protein B, putative	-	✔	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0268 conserved domain protein	-	✔	✔	-	-	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0265 host-nuclease inhibitor protein Gam, putative	-	✔	✔	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0231 tail fiber protein H, putative	-	✔	✔	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0232 phage tail protein, putative	-	✔	✔	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0233 baseplate assembly protein I, putative	-	✔	✔	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0234 baseplate assembly protein W, putative	-	✔	✔	-	-	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0236 baseplate assembly protein V, putative	-	✔	✔	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0237 conserved hypothetical protein	-	✔	✔	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0238 hypothetical protein	-	✔	✔	-	-	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0239 lysoprotein, putative	-	✔	✔	-	-	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0240 hypothetical protein	-	✔	✔	-	-	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0241 conserved hypothetical protein	-	✔	✔	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0243 hypothetical protein	-	✔	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0244 Mu-like prophage I protein, putative	-	✔	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0245 hypothetical protein	-	✔	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0246 conserved hypothetical protein	-	✔	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0247 conserved domain protein	-	✔	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0248 hypothetical protein	-	✔	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0249 phage uncharacterized protein	-	✔	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0250 hypothetical protein	-	✔	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0251 prophage MuSo1, F protein, putative	-	✔	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0252 phage tail protein, putative	-	✔	✔	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0253 tail protein X, putative	-	✔	✔	-	-	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0254 tail protein D, putative	-	✔	✔	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0258 conserved domain protein	-	✔	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0261 hypothetical protein	-	✔	-	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0229 conserved hypothetical protein	-	✔	✔	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0228 hypothetical protein	-	-	✔	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0227 major tail sheath protein	-	✔	✔	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0226 phage major tail tube protein, putative	-	✔	✔	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0220 DNA adenine methylase	-	-	✔	-	-	-	✔	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJ0 0648 hypothetical protein	-	✔	✔	-	-	-	-	✔	✔	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJ0 0649 phage repressor protein, putative	-	✔	✔	-	-	-	✔	✔	✔	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJ1341c motility accessory factor (function unknown)	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	✔	✔	-	-	-	-	-	✔	-	-	-	-	-	-	-	-	-	-	-	-
CJE1607 hypothetical protein	✔	-	✔	-	-	-	-	-	-	-	✔	-	-	✔	✔	-	-	-	-	-	-	✔	-	-	-	-	-	-	-	-	-	-	-
CJE1606 haloacid dehalogenase-like hydrolase domain/phosphoribulokinase domain protein	✔	-	✔	-	-	-	-	-	-	-	✔	-	-	✔	✔	-	-	-	-	-	-	✔	✔	-	-	-	-	-	-	-	-	-	-
CJE1605 capsular polysaccharide biosynthesis protein, putative	✔	-	✔	-	-	-	-	-	-	-	✔	-	-	✔	✔	-	-	-	-	-	-	✔	✔	-	-	-	-	-	-	-	-	-	-
CJE1604 capsular polysaccharide biosynthesis protein, putative	✔	-	✔	-	-	-	-	-	-	-	✔	-	-	✔	✔	-	-	-	-	-	-	✔	✔	-	-	-	-	-	-	-	-	-	-
CJE0924c putative methyl-accepting chemotaxis signal transduction protein	-	-	-	-	✔	✔	✔	✔	-	✔	✔	-	✔	-	-	-	-	-	-	-	-	-	✔	-	-	-	-	-	-	-	-	-	-
CJE0556c hypothetical protein CJE0556c	-	-	-	✔	-	✔	-	-	-	✔	✔	✔	-	-	✔	-	-	-	-	-	-	-	-	✔	-	-	-	-	-	-	-	-	-
CJE1136 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	-	✔	✔	-	-	-	✔	✔	✔	✔	✔	-	-	-	-	-	-
CJ1139c beta-1,3 galactosyltransferase	-	-	✔	✔	✔	✔	✔	-	-	✔	✔	-	-	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

CDS	Cjj 1336 Water	Cjj 414 Water Vole	Cjj 53 Sheep	Cjj IA3902 Sheep	Cjj 81-176 Milk	Cjd 26997 Human Blood	Cjj ICDCCJ 07001 Human GBS Stool	Cjj 81116 Human (Water)	Cjj M1 Human (Chicken)	Cjj 11168 (Human)	Cjj 0110 Human	Cjj 0151 Human	Cc JV20 Human (Microbiome)	Cjj RM1221 Chicken	Cc RM2228 Chicken	Cc 111-3 Pig	Cc 132-6 Pig	Cc 151-9 Pig	Cc 59-2 Pig	Cc 67-8 Pig	Cc 7--1 Pig	Cc 84-2 Pig	Cc 90-3 Pig	Cc 99321 Pig	Cc 03121 Pig	Cc 03103 Pig	Cc 03317 Pig	Cjj 857 Pig	Cjj 5491 Pig	Cji 623 Pig	Cji 304 Pig	Cji 484 Pig	Cji 444 Pig
CJ857_00074 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ857_00075 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ857_00839 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ857_00896 R.HinP1I restriction endonuclease	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ857_00897 Modification methylase HhaI	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ857_01361 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ857_01649 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ857_01723 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ857_01724 R.Pab1 restriction endonuclease	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ857_01734 hypothetical protein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ857_01735 recombination protein F	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔

CDS	Cjj 1336 Water	Cjj 414 Water Vole	Cjj S3 Sheep	Cjj IA3902 Sheep	Cjd 26997 Human Blood	Cjj 81-176 Milk	Cjj ICDCCJ 07001 Human GBS Stool	Cjj 81116 Human Water)	Cjj M1 Human (Chicken)	Cjj 11168 (Human)	Cjj 0110 Human	Cjj 0151 Human	Cc JV20 Human (Microbiome)	Cjj RM1221 Chicken	Cc RM2228 Chicken	Cc 111-3 Pig	Cc 132-6 Pig	Cc 151-9 Pig	Cc 59-2 Pig	Cc 67-8 Pig	Cc 7--1 Pig	Cc 84-2 Pig	Cc 90-3 Pig	Cc 99321 Pig	Cc 03121 Pig	Cc 03103 Pig	Cc 03317 Pig	Cjj 857 Pig	Cjj 5491 Pig	Cjj 623 Pig	Cjj 304 Pig	Cjj 484 Pig	Cjj 444 Pig
CJ0778																																	
major antigenic peptide PEB2	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	✔	-	-	-	-	-	-	-	✔	-	✔	✔	✔	✔	✔	✔
CJ1376																																	
putative periplasmic protein	-	✔	-	✔	-	-	✔	-	-	✔	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
CJ1192																																	
putative C4-dicarboxylate transport protein	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	✔	-	-	-	✔	-	-	✔	✔	-	✔	✔	✔	✔	✔	✔
CJ1069																																	
conserved hypothetical protein CJ1069	✔	✔	✔	✔	✔	✔	-	✔	-	✔	✔	✔	✔	✔	-	-	✔	-	-	✔	✔	✔	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔
CBU_0076																																	
hypothetical protein	-	-	✔	-	-	-	-	✔	✔	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
CJ0908																																	
putative periplasmic protein	-	-	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
CJ1255																																	
putative	-	✔	-	✔	✔	✔	-	-	-	✔	✔	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
CJ1301																																	
hypothetical protein CJ1301	✔	✔	-	✔	✔	-	✔	-	-	✔	✔	✔	-	-	-	✔	✔	-	✔	✔	✔	✔	-	✔	✔	✔	-	✔	✔	✔	✔	✔	✔
CBU_0398																																	
hypothetical protein	-	-	✔	-	-	-	-	✔	✔	-	-	-	✔	✔	✔	✔	✔	-	-	✔	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔
CJ0380c																																	
hypothetical protein CJ0380c	-	-	✔	✔	-	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
CJ0295																																	
putative acetyltransferase	-	-	-	✔	-	-	-	✔	-	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
CBU_1620																																	
hypothetical protein	✔	-	-	-	-	✔	-	✔	✔	-	-	-	✔	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
CJ1547																																	
Blc protein homolog	-	-	-	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔
CBU_1415																																	
hypothetical protein	-	-	✔	-	-	✔	✔	✔	✔	-	✔	✔	✔	-	-	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔
CJ1485c																																	
putative periplasmic protein	-	-	✔	✔	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
CJE1098																																	
hypothetical protein	-	-	-	-	✔	-	-	-	-	-	-	-	-	✔	✔	✔	✔	-	✔	✔	✔	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CCOA0075																																	
transcriptional regulator, Crp/Fnr family	✔	✔	-	-	✔	-	-	-	-	-	-	-	-	-	✔	✔	-	-	✔	-	✔	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ1328																																	
putative UDP-N-acetylglucosamine 2-epimerase	-	✔	✔	✔	✔	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	✔	-	✔	✔	-	-	✔	✔	✔	✔	✔	✔
CJ1329																																	
putative sugar-phosphate nucleotide transferase	-	✔	✔	✔	✔	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	✔	✔	✔	✔	✔	✔
CJ1330																																	
hypothetical protein CJ1330	-	✔	✔	✔	✔	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	✔	✔	✔	✔	-	-	✔	✔	✔	✔	✔	✔
CJ1331																																	
acylneuraminate cytidylyltransferase (flagellin modification)	-	✔	✔	✔	✔	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	✔	✔	✔	✔	-	-	✔	✔	✔	✔	✔	✔
CJ1332																																	
putative oxidoreductase (flagellin modification)	-	✔	✔	✔	✔	-	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	✔	✔	✔	✔	-	-	✔	✔	✔	✔	✔	✔
CJ0916c																																	
conserved hypothetical protein CJ0916c	✔	✔	✔	✔	✔	✔	-	✔	-	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	-	-	✔	✔	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ1415c																																	
putative adenylylsulfate kinase	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	✔	✔	✔	✔	✔	-	✔	✔	-	✔	-	✔	✔	✔	✔	✔	✔	✔
CJ0685c																																	
Invasion protein CiaA	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	-	✔	-	✔	✔	✔	-	-	-	✔	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔
CJ1419c																																	
putative methyltransferase	-	✔	-	✔	✔	✔	-	✔	✔	✔	✔	✔	-	-	-	✔	✔	✔	✔	✔	-	✔	✔	-	✔	-	✔	✔	✔	✔	✔	✔	✔
CJ1418c																																	
putative transferase	-	✔	-	✔	✔	✔	-	✔	✔	✔	✔	✔	-	-	-	✔	✔	✔	✔	✔	-	✔	✔	-	✔	-	✔	✔	✔	✔	✔	✔	✔
CJ1417c																																	
putative amidotransferase	-	✔	-	✔	✔	✔	-	✔	✔	✔	✔	✔	-	-	-	✔	✔	✔	✔	✔	-	✔	✔	-	✔	-	✔	✔	✔	✔	✔	✔	✔
CJ1416c																																	
putative sugar nucleotidyltransferase	-	✔	-	✔	✔	✔	-	✔	✔	✔	✔	✔	-	-	-	✔	✔	✔	✔	✔	-	✔	✔	-	✔	-	✔	✔	✔	✔	✔	✔	✔
CJ1602																																	
conserved hypothetical protein CJ1602	-	✔	✔	✔	-	✔	✔	-	-	✔	✔	✔	-	✔	-	✔	✔	-	✔	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ0571																																	
putative transcriptional regulator	-	-	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	-	-	✔	✔	✔	✔	✔	✔
CBU_1426																																	
hypothetical protein	✔	✔	-	-	-	-	✔	✔	✔	-	✔	✔	✔	-	✔	✔	✔	✔	✔	-	✔	✔	✔	-	✔	-	✔	✔	✔	✔	✔	✔	✔
CBU_0399																																	
lipoprotein, putative	-	-	-	-	-	-	-	✔	✔	-	-	-	✔	-	✔	✔	✔	-	-	✔	✔	✔	✔	-	-	-	-	✔	✔	✔	✔	✔	✔

CDS	Cjj 1336 Water	Cjj 414 Water Vole	Cjj 57 Sheep	Cjj IA3902 Sheep	Cjd 26997 Human Blood	Cjj 81-176 Milk	Cjj ICDCCJ 07001 Human GBS Stool	Cjj 81116 Human (Water)	Cjj M1 Human (Chicken)	Cjj 11168 (Human)	Cjj 0110 Human	Cjj 0151 Human	Cc JV20 Human (Microbiome)	Cjj RM1221 Chicken	Cc RM2228 Chicken	Cc 111-3 Pig	Cc 132-6 Pig	Cc 151-9 Pig	Cc 59-2 Pig	Cc 67-8 Pig	Cc 7--1 Pig	Cc 84-2 Pig	Cc 90-3 Pig	Cc 99321 Pig	Cc 03121 Pig	Cc 03103 Pig	Cc 03317 Pig	Cjj 857 Pig	Cjj 5491 Pig	Cjj 623 Pig	Cjj 304 Pig	Cjj 484 Pig	Cjj 444 Pig
CBJ_0400 hypothetical protein	-	-	-	-	-	-	-	✔	✔	-	-	-	✔	-	✔	✔	✔	-	-	✔	✔	✔	✔	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ1198 ferrous iron transport protein	✔	-	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	-	✔	✔	✔	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔
CJ1335 pseudogene (putative MmgE/PrpD family protein)	✔	-	✔	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	-	-	✔	✔	✔	✔	✔	✔
CJ1365c putative secreted serine protease	-	✔	✔	✔	-	✔	✔	-	-	✔	✔	✔	-	✔	-	✔	-	-	✔	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ1055c putative sulfatase family protein	-	✔	-	✔	-	✔	✔	-	-	✔	✔	✔	-	-	-	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	-	✔	✔	✔	✔	✔	✔	✔
CJE1096 conserved domain protein	✔	-	-	-	✔	-	-	-	-	-	-	-	-	✔	✔	-	✔	-	✔	✔	✔	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CJE1723	✔	-	✔	-	✔	-	-	-	-	-	✔	-	✔	✔	✔	-	-	✔	-	-	-	-	✔	-	-	-	-	✔	✔	✔	✔	✔	✔
MisA protein, putative	✔	-	✔	-	✔	-	-	-	-	-	✔	-	✔	✔	✔	-	-	✔	-	-	-	-	✔	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ0265c putative cytochrome C-type haem-binding periplasmic protein	✔	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	-	-	✔	-	-	-	-	✔	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ0727 putative periplasmic solute-binding protein	-	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	✔	✔	✔	-	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔	✔
CJ1141 sialic acid synthase (N-acetylneuraminic acid synthetase)	-	-	-	✔	-	✔	✔	✔	-	✔	✔	-	-	-	-	-	-	-	✔	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ1142 putative UDP-N-acetylglucosamine 2-epimerase	-	-	-	✔	-	✔	-	✔	-	✔	✔	-	-	-	-	-	-	-	✔	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CBJ_0588 hypothetical protein	-	✔	-	✔	-	✔	-	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	✔	-	-	✔	✔	✔	✔	✔	✔
CJ1325 putative methyltransferase	✔	✔	-	✔	✔	-	-	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	✔	✔	-	-	✔	✔	✔	✔	✔	✔
CJE1284 lipooligosaccharide biosynthesis glycosyltransferase, putative	✔	✔	✔	-	✔	✔	-	-	-	-	-	-	-	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ1506c	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
putative MCP-type signal transduction protein	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ1354 enterochelin uptake ATP-binding protein	-	✔	✔	-	✔	-	-	✔	✔	✔	-	✔	-	-	✔	-	-	✔	✔	-	-	-	✔	✔	-	-	-	✔	✔	✔	✔	✔	✔
CJ1063 putative acetyltransferase	✔	-	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	-	✔	-	-	✔	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CJ81176_0207 conserved hypothetical protein	-	✔	-	✔	✔	✔	-	-	-	-	✔	✔	-	-	-	-	-	-	-	-	-	-	-	✔	-	-	-	✔	✔	✔	✔	✔	✔
CJ0861c para-aminobenzoate synthase glutamine amidotransferase component II	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	✔	-	-	-	-	-	-	-	✔	-	-	-	-	✔	✔	✔	✔	✔	✔	✔
CJ0079c cytolethal distending toxin A	-	-	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	✔	-	✔	✔	-	-	✔	✔	✔	✔	✔	✔
CBJ_1278 hypothetical protein	-	-	-	-	-	-	✔	✔	✔	-	-	-	-	-	✔	-	-	✔	-	-	-	-	-	✔	-	-	✔	✔	✔	✔	✔	✔	✔
CJ1638 DNA primase	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔	✔	✔	✔
CBJ_0064 hypothetical protein	-	-	-	-	-	-	✔	✔	✔	-	-	-	✔	-	✔	-	-	-	-	✔	-	✔	-	✔	✔	-	✔	✔	✔	✔	✔	✔	✔
CCO0265 CAAX amino terminal protease family protein	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
CCO0375 Transporter protein	-	-	-	-	-	-	-	-	✔	-	-	-	✔	-	✔	✔	✔	-	✔	✔	✔	-	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
CCO1142 conserved hypothetical protein	✔	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	✔	-	✔	✔	✔	-	✔	-	✔	✔	-	✔	✔	✔	✔	✔	✔	✔	✔
CCO1143 transporter, MFS superfamily	✔	-	-	-	-	-	-	-	-	-	-	-	✔	-	✔	-	-	-	-	-	-	-	✔	-	✔	-	✔	✔	✔	✔	✔	✔	✔	✔

Appendix 9.4.1 Local Database BLAST Results for HipO Gene, and Virulence, Capsule, and ST403CC Associated Genes in C. jejuni and C. coli

Species		*C. jejuni jejuni*		*C. jejuni jejuni*		*C. jejuni jejuni*		*C. jejuni jejuni*		*C. jejuni jejuni*		*C. jejuni jejuni*		*C. jejuni jejuni*		*C. jejuni jejuni*		*C. jejuni jejuni*	
Isolate		857		549.1		623		304		484		444		ATCC33560		81116		11168	
CDS		Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value
Typing CDS	*HipO*	2645	0	2645	0	2645	0	2639	0	2645	0	2645	0	2628	0	2678	0	2767	0
Capsule Associated CDS	*kpsC*	3109	0	3109	0	3109	0	3109	0	3109	0	3109	0	3109	0	1711	0	3823	0
	kpsD	2798	0	2837	0	2837	0	2798	0	2832	0	2832	0	2832	0	3059	0	3064	0
	kpsE	1884	0	1879	0	1879	0	1884	0	1879	0	1884	0	1879	0	2067	0	2067	0
	kpsF	1483	0	1511	0	1522	0	1483	0	1489	0	1489	0	1489	0	1552	0	1751	0
	kpsM	1048	0	1048	0	1003	0	1048	0	1003	0	1048	0	998	0	1447	0	1447	0
	kpsS	2015	9	2032	0	2032	0	2032	0	2032	0	2032	0	2037	0	2012	0	2189	0
	kpsT	976	0	998	0	1026	0	998	0	1020	0	992	0	1026	0	1225	0	1225	0
Virulence Associated CDS	*cadF*	1707	0	1707	0	1707	0	1707	0	1773	0	1707	0	1707	0	1724	0	1773	0
	cdtA	1413	0	1413	0	1413	0	1413	0	1413	0	1413	0	1413	0	1469	0	1491	0
	cdtB	1240	0	1240	0	1240	0	1240	0	1240	0	1251	0	1240	0	1430	0	1474	0
	cdtC	987	0	987	0	987	0	987	0	981	0	987	0	987	0	1026	0	1053	0
	ciaB	3253	0	3258	0	3253	0	3314	0	3258	0	3269	0	3258	0	3314	0	3386	0
	dnaJ	1953	0	1941	0	1953	0	1941	0	1953	0	1953	0	1941	0	1892	0	2073	0
	flaA	636	0	933	0	623	2E-177	769	0	885	0	627	2E-178	881	0	909	0	3175	0
	iamB	1901	0	1884	0	1890	0	1884	0	1879	0	1890	0	1890	0	1895	0	2050	0
	pldA	1779	0	1729	0	1779	0	1735	0	1790	0	1779	0	1768	0	1735	0	1829	0
	racR	1203	0	1203	0	1203	0	1203	0	1203	0	1203	0	1197	0	1125	0	1242	0
	virB11	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	wlaN	99	8.00E-20	292	4.00E-78	99	8.00E-20	296	3.00E-79	296	3.00E-79	292	4.00E-78	292	4.00E-78	NM	NM	1685	0
ST403CC Associated CDS	*CJ857_00074*	466	5E-131	466	5E-131	466	5E-131	466	5E-131	466	5E-131	466	5E-131	466	5E-131	NM	NM	NM	NM
	CJ857_00075	2438	0	2431	0	2438	0	2431	0	2438	0	2438	0	2416	0	NM	NM	NM	NM
	CJ857_00839	455	1E-127	455	1E-127	449	5E-126	449	5E-126	455	1E-127	449	5E-126	394	2E-109	NM	NM	NM	NM
	CJ857_00896	1419	0	1419	0	1419	0	1419	0	1419	0	1419	0	1419	0	NM	NM	NM	NM
	CJ857_00897	1779	0	1779	0	972	0	1779	0	1779	0	1779	0	1779	0	NM	NM	NM	NM
	CJ857_01361	2250	0	2211	0	2204	0	2204	0	2211	0	2204	0	2242	0	NM	NM	NM	NM
	CJ857_01649	1591	0	1585	0	1580	0	1585	0	1585	0	1585	0	1580	0	NM	NM	NM	NM
	CJ857_01723	1973	0	1973	0	1973	0	1973	0	1973	0	1973	0	1973	0	NM	NM	NM	NM
	CJ857_01724	1319	0	1319	0	1319	0	1308	0	1319	0	1314	0	1319	0	NM	NM	NM	NM
	CJ857_01734	1037	0	1037	0	1037	0	1037	0	1037	0	1037	0	1037	0	NM	NM	NM	NM
	CJ857_01735	1951	0	1951	0	1951	0	1951	0	1951	0	1951	0	1951	0	NM	NM	NM	NM

BLAST Results for capsule, virulence, and ST403CC associated CDS. Grey results were considered 'low scoring matches'. Results in red were discarded scores due to low homology or high error values. Yellow highlighted scores indicate the 'self-match' of the CDS used to investigate homology

Species		*C. jejuni jejuni*		*C. jejuni jejuni*		*C. jejuni jejuni*		*C. jejuni jejuni*		*C. jejuni jejuni*		*C. jejuni jejuni*		*C. jejuni jejuni*		*C. jejuni jejuni*		*C. jejuni jejuni*	
Isolate		1336		414		81-176		S3		IA3902		ICDCCJ07001		M1		01/10		01/51	
CDS		Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value
Typing CDS	*HipO*	2628	0	2025	0	2617	0	2656	0	2545	0	2628	0	2684	0	2678	0	2579	0
Capsule Associated CDS	*kpsC*	3149	0	1517	0	3458	0	3109	0	3086	0	3096	0	1810	0	3452	0	3325	0
	kpsD	2771	0	2433	0	2959	0	2942	0	2638	0	2809	0	2798	0	2926	0	2987	0
	kpsE	1768	0	1613	0	1912	0	1906	0	1435	0	1978	0	1945	0	1873	0	1884	0
	kpsF	1530	0	1352	0	1461	0	1434	0	1428	0	1530	0	1483	0	1485	0	1461	0
	kpsM	965	0	1013	0	NM	NM	NM	NM	NM	NM	NM	NM	1042	0	1070	0	1064	0
	kpsS	2039	0	1683	0	2050	0	2089	0	1786	0	2032	0	2039	0	2150	0	1971	0
	kpsT	998	0	1009	0	1003	0	1003	0	815	0	1020	0	987	0	1003	0	1003	0
Virulence Associated CDS	*cadF*	1735	0	1480	0	1480	0	1762	0	1618	0	1762	0	1724	0	1773	0	1773	0
	cdtA	294	9E-79	NM	NM	1485	0	1463	0	NM	NM	1463	0	1469	0	1491	0	1483	0
	cdtB	492	3E-138	NM	NM	1474	0	1458	0	NM	NM	1447	0	1430	0	1474	0	1474	0
	cdtC	878	0	NM	NM	1053	0	1048	0	NM	NM	1042	0	1026	0	1053	0	1053	0
	ciaB	3153	0	2580	0	3319	0	3330	0	3236	0	3341	0	3308	0	3380	0	3341	0
	dnaJ	1853	0	1692	0	1958	0	2039	0	1825	0	2002	0	1892	0	1953	0	2073	0
	flaA	809	0	741	0	909	0	2396	0	1762	0	765	0	909	0	942	0	1977	0
	iamB	1929	0	1574	0	1945	0	1945	0	1845	0	1945	0	1895	0	2006	0	2050	0
	pldA	1424	0	1356	0	1724	0	1773	0	252	7E-66	1718	0	1735	0	1757	0	1731	0
	racR	1031	0	1020	0	1242	0	1242	0	1064	0	1230	0	1125	0	1242	0	1242	0
	virB11	NM	NM	NM	NM	1762	0	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	wlaN	NM	NM	NM	NM	NM	NM	267	2.00E-70	NM	NM	NM	NM	NM	NM	379	3.00E-104	1677	0
ST403CC Associated CDS	*1857_0007*	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	1857_0007	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	1857_0083	130	7E-30	NM	NM	NM	NM	NM	NM	316	5E-86	NM	NM	NM	NM	NM	NM	NM	NM
	1857_0089	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	1857_0089	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	1857_0136	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	2239	0
	1857_0164	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	1857_0172	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	1857_0172	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	1857_0173	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	1857_0173	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM

BLAST Results for capsule, virulence, and ST403CC associated CDS. Grey results were considered 'low scoring matches'. Results in red were discarded scores due to low homology or high error values. Yellow highlighted scores indicate the 'self-match' of the CDS used to investigate homology

Appendix 9.4.1 Local Database BLAST Results for HipO Gene, and Virulence, Capsule, and ST403CC Associated Genes in C. jejuni and C. coli

Species		*C. jejuni jejuni*		*C. jejuni doylei*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*	
Isolate		RM1221		269.97		RM2228		JV20		111-3		132-6		151-9		59-2		67-8	
CDS		Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value
Typing CDS	*HipO*	2651	0	2545	0	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
Capsule Associated CDS	*kpsC*	3109	0	3086	0	NM	NM	603	3.00 E-171	3009	0	NM	NM	597	2.00 E-169	1855	0	569	3.00 E-161
	kpsD	2942	0	2638	0	1325	0	1297	0	3009	0	1301	0	1314	0	2848	0	1277	0
	kpsE	1906	0	1435	0	662	0	656	0	1851	0	684	0	667	0	1940	0	756	0
	kpsF	1434	0	1428	0	NM	NM	924	0	1417	0	939	0	929	0	1417	0	NM	NM
	kpsM	NM	NM	NM	NM	NM	NM	NM	NM	1048	0	592	3 E-168	NM	NM	NM	NM	592	3 E-168
	kpsS	2089	0	1786	0	1116	0	1116	0	2100	0	1127	0	1155	0	2111	0	1133	0
	kpsT	1003	0	815	0	682	0	682	0	1009	0	693	0	688	0	1020	0	704	0
Virulence Associated CDS	*cadF*	1762	0	1618	0	601	6E-171	595	3E-169	601	6E-171	601	6E-171	NM	NM	601	6E-171	NM	NM
	cdtA	1463	0	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	cdtB	1458	0	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	cdtC	1048	0	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	ciaB	3330	0	3236	0	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	dnaJ	2039	0	1825	0	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	flaA	2379	0	1762	0	2361	0	848	0	710	0	837	0	793	0	817	0	499	4E-140
	iamB	1945	0	1845	0	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	pldA	1773	0	252	7E-66	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	racR	1242	0	1064	0	782	0	787	0	793	0	793	0	793	0	793	0	793	0
	virB11	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	1219	0	NM	NM	NM	NM	NM	NM
	wlaN	285	6.00E-76	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
ST403CC Associated CDS	*CJ857_00074*	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	CJ857_00075	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	CJ857_00839	128	3E-29	316	5E-86	NM	NM	NM	NM	128	3E-29	NM	NM	NM	NM	268	2E-71	NM	NM
	CJ857_00896	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	CJ857_00897	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	CJ857_01361	NM	NM	NM	NM	682	0	682	0	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	CJ857_01649	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	CJ857_01723	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	CJ857_01724	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	CJ857_01734	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	CJ857_01735	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM

BLAST Results for capsule, virulence, and ST403CC associated CDS. Grey results were considered 'low scoring matches'. Results in red were discarded scores due to low homology or high error values. Yellow highlighted scores indicate the 'self-match' of the CDS used to investigate homology

Appendix 9.4.1 Local Database BLAST Results for HipO Gene, and Virulence, Capsule, and ST403CC Associated Genes in C. jejuni and C. coli

Species		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*	
Isolate		7—1		84-2		90-3		99/321		03/121		03/103		03/317	
CDS		Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value
Typing CDS	*HipO*	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
Capsule Associated CDS	*kpsC*	575	7.00 E-163	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	kpsD	1282	0	1321	0	1319	0	1319	0	1303	0	1330	0	1301	0
	kpsE	756	0	673	0	NM	NM	667	0	662	0	NM	NM	684	0
	kpsF	NM	NM	898	0	NM	NM	929	0	933	0	926	0	939	0
	kpsM	592	3 E-168	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	kpsS	1127	0	1171	0	1133	0	1166	0	1155	0	1122	0	1144	0
	kpsT	704	0	643	0	649	0	688	0	671	0	649	0	704	0
Virulence Associated CDS	*cadF*	601	6E-171	601	6E-171	601	6E-171	601	6E-171	601	6E-171	601	6E-171	601	6E-171
	cdtA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	cdtB	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	cdtC	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	ciaB	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	dnaJ	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	flaA	414	1E-114	715	0	893	0	525	6E-148	623	2E-177	893	0	300	4E-80
	iamB	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	pldA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	racR	793	0	787	0	793	0	793	0	793	0	787	0	787	0
	virB11	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	wlaN	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
ST403CC Associated CDS	*CJ857_00074*	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	CJ857_00075	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	CJ857_00839	265	2E-70	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	322	1E-87
	CJ857_00896	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	CJ857_00897	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	CJ857_01361	NM	NM	682	0	675	0	675	0	676	0	676	0	NM	NM
	CJ857_01649	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	1555	0
	CJ857_01723	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	1973	0
	CJ857_01724	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	1319	0
	CJ857_01734	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
	CJ857_01735	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM

BLAST Results for capsule, virulence, and ST403CC associated CDS. Grey results were considered 'low scoring matches'. Results in red were discarded scores due to low homology or high error values. Yellow highlighted scores indicate the 'self-match' of the CDS used to investigate homology

Appendix 9.4.2: Local Database BLAST Results for ST403CC Associated CDS in an Extended *C. coli* Database

Species	C. coli		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*	
Isolate	37-05		202-04		317-04		1091		1098		1148		1417		1891		1909		1948		1957		1961	
CDS	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value
CJ857_00839	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
CJ857_01361	682	0	676	0	689	0	682	0	NM	NM	682	0	676	0	682	0	NM	NM	NM	NM	NM	NM	NM	NM
CJ857_01723	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
CJ857_01724	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM

Species	*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*	
Isolate	2548		2553		2680		2685		2688		2692		2698		80352		86119		H6		H8	
CDS	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value
CJ857_00839	333	7E-91	128	4E-29	NM	NM	NM	NM	NM	NM	NM	NM	246	1E-64	248	3E-65	NM	NM	NM	NM	267	7E-71
CJ857_01361	682	0	671	0	682	0	667	0	682	0	NM	NM	NM	NM	NM	NM	682	0	671	0	688	0
CJ857_01723	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
CJ857_01724	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM

Species	*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*		*C. coli*	
Isolate	H9		H56		LMG9853		LMG9854		LMG9860		LMG23336		LMG23341		LMG23342		LMG23344		Z156		Z163	
CDS	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value	Score	E value
CJ857_00839	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	265	3E-70	NM	NM	NM	NM
CJ857_01361	671	0	NM	NM	660	0	682	0	676	0	671	0	667	0	673	0	671	0	682	0	682	0
CJ857_01723	NM	NM	1973	0	NM	NM	NM	NM	NM	NM	1973	0	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
CJ857_01724	NM	NM	1319	0	NM	NM	NM	NM	NM	NM	1314	0	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM

BLAST Results for capsule, virulence, and ST403CC associated CDS. Grey results were considered 'low scoring matches'. Results in red were discarded scores due to low homology or high error values.