An Investigation into the Genetic Diversity of the Foodborne Pathogen *Campylobacter jejuni* using DNA Microarrays

Thesis submitted for the degree of Doctor of Philosophy

By

Olivia L. Champion B.Sc. (Hons) M.Sc.

Department of Infectious and Tropical Diseases,

London School of Hygiene and Tropical Medicine, University of London

March 2005

Abstract

Despite being the principal bacterial cause of gastroenteritis world wide, the epidemiology of *Campylobacter jejuni* is poorly understood. This is largely because the proportion of human disease caused by different sources of infection is unknown.

In this study a diverse strain collection was selected comprising of 91 *C. jejuni* isolates from diverse animal and environmental ecological niches as well as clinical isolates from patients representing a range of disease outcomes. Whole genome comparisons were performed by DNA microarray analysis with the dual aim of identifying genetic markers specific to strains from different ecological niches and identifying novel virulence determinants.

A new phylogenomic technique for the analysis of DNA microarray data was developed combining Bayesian algorithms to model phylogeny based on whole genome data with parsimony based algorithms to identify the key genes contributing to the clade formations. This method revealed a previously undetected *C. jejuni* population structure comprising two main clades, a "chicken clade" and a "non-chicken clade". These statistically supported clades differentiated strains from distinct ecological niches with 94% of strains isolated from chickens and 41% of clinical isolates contained within the "chicken clade". *C. jejuni* isolates from ovine, bovine and sand isolates also formed distinct clades within the "non-chicken clade". Key genes contributing to the distinction of strains from one ecological niche from another were identified. In particular a putative glycosylation islet *cj1321-cj1326* was found to be present in strains within the "chicken clade" but absent or divergent from strains in the "non-chicken clade', a result validated by PCR screening. This locus represents a potential genetic marker of *C. jejuni* strains of avian origin.

The DNA microarray data analysis method described in this study may be used to study other bacterial pathogens facilitating the identification of bacterial phylogenies and genetic markers associated with specific phenotypes.

Contents

Acknowledgements

With thanks to Professor Brendan Wren for his excellent supervision, much needed support and invaluable guidance throughout the duration of my PhD.

My thanks go to the rest of the Wren group, some now departed but especially to Dr Nick Dorrell for his helpful advice, microarray expertise and good humour; Dr Dennis Linton and Dr Andrey Karlyshev for constructive criticism and invaluable training; Ozan Gundogdu and Manu Davis for unfailing technical and administrative support and Dr Stewart Hinchliffe, Jon Cuccui, Dr George Joshua, Dr Richard Stabler, Gill Thacker, Dr Rebecca Langdon, Pippa Strong and Collette Guthrie for stimulating laboratory debates. I am indebted to the members of the BμG@S group, in particular Dr Jason Hinds, Dr Adam Witney, Dr Sally Hussain and Gemma Marsden without whom the *C. jejuni* microarray would not exist. With thanks also to Stephanie Sanos, Eric Tongren, Sarah Howard, Ross Cummings, Chandrabala Shah and Daniel Korbel, my esteemed colleagues, dining companions and friends. With thanks to my family for their endless enthusiasm and my father's 'A star' grade in botany many years ago (was there such a grade as 'A star' in those days?). How could I fail to become a scientist with such remarkable genes? A final word of thanks to my husband, Al, for his artistic input, unfailing support and for helping me see the funny side of things.

This study was funded by a Medical Research Council Studentship.

Declaration

The conclusions reached in this thesis are my own based on the experiments reported herein and published work. Experiments were performed in the Department of Infectious and Tropical Diseases at the London School of Hygiene and Tropical Medicine (LSHTM).

Microarray construction was possible only through the valuable collaboration with the Bacterial Microarray Group at St Georges.
LOS outer core structural analysis was carried out at National Research Canada (NRC), Ottawa, Canada.
Dr Michael Gaunt calculated clade credibility values of phylogenetic trees.

Abbreviations

AFLP	amplified fragment length polymorphism
BGC	Bacillus Calmette-Guerin
BLAST	basic local alignment search tool
BSA	bovine serum albumin
BμG@S	bacterial microarray group at St Georges
CDS	coding sequence
CO_2	carbon dioxide
CPS	capsular polysaccharide
CRU	Campylobacter Reference Unit
DA	direct agglutination
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
EPP	estimated probability of presence
FSA	Food Standards Agency
gDNA	genomic deoxyribonucleic acid
GACK	genomotyping analysis Charlie Kim
GAPDH	glycerol phosphate dehydrogenase
GBS	Guillain Barré syndrome
HMW	high molecular weight
HS	heat stable
IID	Infectious Intestinal Disease Study
Kb	kilo base
LOS	lipooligosaccharide
MAGE-ML	microarray gene expression markup language
MAMP	microbe associated molecular pattern
MLEE	multi locus enzyme electrophoresis
MLST	multi locus sequence typing
N_2	nitrogen
NANA	neuraminic acid (sialic acid)
NCTC	national collection of type cultures
NT	not typable

Abbreviations

AFLP	amplified fragment length polymorphism
BGC	Bacillus Calmette-Guerin
BLAST	basic local alignment search tool
BSA	bovine serum albumin
BμG@S	bacterial microarray group at St Georges
CDS	coding sequence
CO_2	carbon dioxide
CPS	capsular polysaccharide
CRU	Campylobacter Reference Unit
DA	direct agglutination
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
EPP	estimated probability of presence
FSA	Food Standards Agency
gDNA	genomic deoxyribonucleic acid
GACK	genomotyping analysis Charlie Kim
GAPDH	glycerol phosphate dehydrogenase
GBS	Guillain Barré syndrome
HMW	high molecular weight
HS	heat stable
IID	Infectious Intestinal Disease Study
Kb	kilo base
LOS	lipooligosaccharide
MAGE-ML	microarray gene expression markup language
MAMP	microbe associated molecular pattern
MLEE	multi locus enzyme electrophoresis
MLST	multi locus sequence typing
N_2	nitrogen
NANA	neuraminic acid (sialic acid)
NCTC	national collection of type cultures
NT	not typable

O_2	oxygen
OMP	outer membrane protein
ORF	open reading frame
PAUP	phylogenetic analysis using parsimony
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PFGE	pulsed field gel electrophoresis
PR	plasticity region
PRR	pattern recognition receptor
PT	phage type
RAPD	randomly amplified polymorphic DNA
RNA	ribonucleic acid
rRNA	ribosomal ribonucleic acid
RC	reconstructed changes
RDNC	reacts with phage but does not conform to designated type
SDS	sodium dodecyl sulphate
SNP	single nucleotide polymorphism
SSC	saline sodium citrate
ST	serotype
TLR	toll like receptor

List of Tables **page**

List of Figures **page**

1.0 Introduction

1.1 The organism

Campylobacter cells are slender Gram-negative rods approximately 1.5-6 μm long and 0.2-0.5 μm wide, with characteristic spiral morphology. The *Campylobacter* genus contains sixteen species that are members of the delta-epsilon group of proteobacteria, of which *C. jejuni* is most commonly associated with human disease. *C. jejuni* is a microaerobic bacterium requiring an oxygen concentration in the range of 3-15% and a carbon dioxide concentration of 3-5% for growth. The optimal growth temperature of *C. jejuni* is 42°C, the gut temperature of avians, in which the bacterium forms part of the commensal flora. The cells are flagellate with a characteristic corkscrew-like motility mediated usually via a single, polar, unsheathed flagellum (Guerry *et al*, 1992). *C. jejuni* is oxidase positive and is unable to ferment or oxidise sugars. The inability to ferment sugars means many routine biochemical tests used in diagnostic microbiology laboratories to identify pathogens are redundant for the identification of *Campylobacters*. Many of the characteristics of *Campylobacters*, including the microaerobic growth and spiral morphology with polar flagella mediating corkscrew motility, are shared with a second member of the delta-epsilon group of proteobacteria, the genus *Helicobacter* of which *Helicobacter pylori* is the type species (Owen, 1998).

1.2 The disease

Human *Campylobacter* strains were first documented in the 1940's and 1950's as a thermophilic group of 'related vibrios' associated with human diarrhoeal disease (King, 1957; Wheeler and Borchers, 1961). Although first cultured from blood samples the organism was suspected to also be present in the human gut. However, failure to successfully culture the organisms from stool samples due to overgrowth by coliforms prevented this from being confirmed. The Greek word Campylobacter, meaning a curved rod, was proposed by Sebald and Veron in 1963 to distinguish microaerobic vibrios from classical cholera and halophilic groups in the genus *Vibrio* (Sebald and Veron, 1963). Then, in the 1970's, following the development of a selective culturing technique and successful cultures of the organism from human stools (Butzler *et al.*, 1973; Dekeyser *et al.*, 1972) the status of *Campylobacter* as a "new" human pathogen was confirmed (Skirrow, 1977). Since its identification in the

1970's, *C. jejuni* has become the most commonly isolated human gastrointestinal pathogen in the developed world with 44,832 *Campylobacter* cases (provisionally) reported in England and Wales in 2003 (www.hpa.org.uk/infections/topics_az/campy/data_ew.htm). This figure is thought to be an underestimate with the actual number of cases per annum estimated to be approximately half a million (Agency, 2000).

A feature of human *C. jejuni* infection (campylobacteriosis) is the wide spectrum of clinical disease presentation and outcome. This ranges from asymptomatic carriage to the most common presentation of self-limiting diarrhoea and rarely sequelae including septicaemia and neuropathy such as Guillain-Barrè syndrome (GBS).

1.2.1 Asymptomatic carriage

Asymptomatic carriage of *C. jejuni* in humans has rarely been documented. However, several cases were identified by the Infectious Intestinal Diseases (IID) study carried out by the Foods Standards Agency (Agency, 2000). This study was undertaken to estimate the number of cases of gastroenteritis occurring in England and Wales. Seventy general practice surgeries were recruited into the study and groups of people from each practice were asked to report on a weekly basis for six months whether they had suffered from gastroenteritis. All patients who presented to their GPs from 34 of the general practices over a twelve-month period were matched with controls and asked to provide stool specimens. Information was also obtained from both cases and controls regarding risk factors and personal characteristics. The definition of a control in this study was '...persons who have been free of loose stools or significant vomiting for three weeks prior to the onset of illness in the case.' In total over 6000 cases who developed gastroenteritis were matched with controls within the cohorts and both provided stool specimens that underwent microbiological investigations. Interestingly, some of the stool specimens from the control population tested positive for *C. jejuni*. These control patients were the first well-characterised asymptomatic human carriers of *C. jejuni*.

1.2.2 Diarrhoeal disease

The most common presentation of campylobacteriosis is gastroenteritis following an incubation period of 24-72 hours. The illness is characterised by fever, abdominal cramps, headache and loose to bloody diarrhoea, lasting between 5 and 10 days and

usually self –limiting (Blaser, 1997). The acute diarrhoeal illness has clinical manifestations similar to salmonellosis or shigellosis and clinically cannot be distinguished. Definitive diagnosis of campylobacter associated human disease is made by the isolation of the organism from patients' faeces on selective media containing vancomycin, polymyxin B and trimethoprim (Kaijser, 1992). The *Campylobacter* Sentinel Surveillance Scheme, launched in 2000, has collected *C. jejuni* epidemiological and microbiological typing data from twenty-two health authorities in England and Wales (covering a population of about 12 million people). Data from the first two years of the study has shown that diarrhoea (98%), abdominal pain (92%) and fever (85%) were the most commonly reported symptoms associated with *C. jejuni* human infection. However, vomiting (39%) and bloody diarrhoea (33%), although reported less frequently, were still evident in a significant proportion of the cases.

1.2.3 Post *C. jejuni* infection sequelae

Following *C. jejuni* infection, sequelae occasionally occur. These include bacteraemias (occurring in around 1.5/1000 intestinal infections), endocarditis, reactive arthritis, meningitis, post-dysenteric irritable bowel syndrome (PD-IBS), GBS and the rare variant of GBS, Miller Fisher syndrome (MFS) (Kuroki *et al.*, 1993; Skirrow, 1991). GBS is an autoimmune disorder of the peripheral nervous system characterised by weakness of limbs, respiratory muscles and loss of reflexes. Recovery usually takes place over weeks or months but up to 20% of patients may require mechanical ventilation and 15-20% are left with severe neurological defects (Nachamkin *et al.*, 1998).

1.3 Pathogenesis

The human host possesses a number of innate mechanisms, including a mucus layer, an epithelial barrier, peristalsis and acidic pH, through which to protect itself from gastrointestinal pathogens. Following ingestion, *C. jejuni* must first circumvent these physical barriers. Gastric acid rapidly kills *C. jejuni* and ingestion of the pathogen with food aids its survival in the stomach (Blaser *et al.*, 1980; Walker *et al.*, 1986). Human volunteer studies have demonstrated an infectious dose as low as 500 organisms (Black *et al.*, 1988) but variation in host susceptibility and/ or virulence of strains exists.

Following ingestion, successful pathogens must penetrate the mucus layer and adhere to the intestinal epithelium of the small intestine and occasionally of the colon. Most campylobacters are found within deep caecal crypts, thus avoiding clearance by peristalsis. The spiral shape and corkscrew-like motility of the bacterium is important in its ability to colonise the mucus layer of the intestinal tract (Morooka *et al.*, 1985) and *C. jejuni* cells have been shown to adhere to and invade Caco-2 cells more effectively in the presence of a chemical mimicking the viscosity of intestinal mucus (Szymanski *et al.*, 1995). Moreover, pre-treatment of Hep-2 cells with mucin resulted in enhanced invasion of *C. jejuni* in four clinical strains (de Melo and Pechere, 1988). Thus, intestinal mucus appears to promote cell adhesion and invasion of *C. jejuni*. The mechanism of adherence to the intestinal epithelium used by *C. jejuni* is unknown but several adhesins have been identified. CadF, an outer membrane protein (OMP) of *C. jejuni*, promotes the binding of the pathogen to intestinal epithelial cells (Konkel *et al.*, 1999). Similarly, a *jlpA* mutant showed a marked decrease in adhesion to Hep-2 cells (Jin *et al.*, 2001). However, the most well characterised virulence factor facilitating colonisation is the flagellum. Experimental evidence implicates flagella and/or motility as requirements for successful colonisation of the chick intestinal tract and to cross polarised epithelial cell monolayers (Doig *et al.*, 1996; Grant *et al.*, 1993; Nachamkin *et al.*, 1993; Wassenaar *et al.*, 1993). Flagella are the locomotory organelles of bacteria. In *C. jejuni*, flagellin is encoded by *flaA* and *flaB* and the flagella filaments are composed predominantly of the FlaA flagellin interspersed with the highly variable FlaB flagellin (Guerry *et al.*, 1991). A defective *flaA* gene leads to immotile bacteria (Wassenaar *et al.*, 1991). However, mutations in gene *flaB* but not *flaA* result in a motile phenotype, although both genes are required for maximum motility (Kinsella *et al.*, 1997). Non-motile mutants lose the ability to adhere to and consequently invade cells *in vitro* and *in vivo* (Nachamkin *et al.*, 1993; Wassenaar *et al.*, 1991). However, invasion was possible when non motile mutants were centrifuged onto tissue culture cells, indicating that motility is not the only factor involved in invasion (Wassenaar *et al.*, 1991). The secretion of invasion proteins into host cells to promote uptake is well documented in gastroenteric pathogens such as *Salmonella*, *Shigella*, *Yersinia* and enteropathogenic *E. coli* (Zaharik *et al.*, 2002). Upon co-cultivation with host cells, *C. jejuni* expresses nine novel proteins against which rabbit antisera has been raised. Treatment of *C. jejuni* with the rabbit antisera reduced cell invasion by 98%, implicating the production of novel proteins in the role

of host cell invasion (Konkel *et al.*, 1993). It has been demonstrated that one of these secreted proteins is a product of the *Campylobacter* invasion antigen B (*ciaB*) gene. CiaB comprises 610 amino acids with has 40.6% to 45.4% sequence similarity to type III secreted proteins associated with *Salmonella* SipB, *Shigella* IpaB and *Yersinia* YopB proteins. Mutants in *ciaB* were non-invasive for INT-407 cells and using immunofluorescence microscopy the CiaB protein has been shown to translocate into INT-407 cells (Konkel *et al.*, 1999). Similarly, FlaC identified in *C. jejuni* strain ATCC43431, is found in the extracellular milieu of wild type cultures as a secreted protein. Purified recombinant FlaC binds to Hep-2 cells and invasion of the cells by *flaC* null mutants was reduced to 14% suggesting a role for FlaC in cell invasion (Song *et al*, 2004). Genes homologous with those encoding classical type III secretory apparatus have not been identified in NCTC11168. However, *Yersinia* secretes virulence factors through the flagellar apparatus and it has been speculated that *C. jejuni* may secrete proteins using the same mechanism. Indeed, mutations in several genes encoding the basal body, hook and filament structural components of the flagellum indicated that CiaB export requires at least one of two structural filament proteins (Konkel *et al*, 2004). Furthermore, mutations in genes essential for flagellum biosynthesis (*flgF* and *flgE*) resulted in a phenotype that did not secrete FlaC (Song *et al*, 2004). *In H. pylori*, the virulence factor CagA is transported directly into host cells via a type IV secretion system (Christie and Vogel, 2000). Genes found on a plasmid (*pVir*) identified in *C. jejuni* strain 81-176 have significant homology to type IV secretion system genes and inactivation of four of these genes resulted in decreased invasion on INT-407 cells (Bacon *et al.*, 2001). However, the sequenced *C. jejuni* strain NCTC11168 which was isolated from a patient with gastroenteritis did not possess any plasmids (Parkhill *et al.*, 2000). Although the mechanism is not fully understood, *C. jejuni* very clearly invade cells in the human intestinal tract as shown *in vivo* in clinical intestinal biopsies (van Spreeuwel *et al.*, 1985) as well as intestinal biopsies of infected primates and other animal models (Babakhani *et al.*, 1993; Newell and Pearson, 1984; Russell *et al.*, 1993). Eukaryotic cell invasion by *C. jejuni* has also been shown *in vitro*, by the demonstration of the invasion of *C. jejuni* into human epithelial cell lines (Ketley, 1997; Konkel *et al.*, 1992). Damage to gut epithelial cells may occur during invasion and/ or by subsequent cytolethal distending toxin activity. Disruption of the normal functioning of epithelial cells results in their inability to absorb water and ions leading to the characteristic diarrhoeal disease

associated with clinical *C. jejuni* infection. Attempts by the host's immune system to combat the pathogen present in infected tissue results in common *C. jejuni* disease symptoms such as fever and inflammation of the gut epithelium.

The mechanism of pathogenesis for the debilitating and often life-threatening complications associated with *C. jejuni* also remains unclear. Following *C. jejuni* infection in humans, increases in enteroendocrine cells (ECs), T lymphocytes and gut permeability may persist for a year or more and are these are thought to be contributory factors to PD-IBS. Furthermore, it has been hypothesised that the peripheral neuropathy GBS may be caused by molecular mimicry between surface components on the *C. jejuni* cell and human gangliosides. The *C. jejuni* surface antigen, lipooligosaccharide (LOS), has a relatively conserved inner core with a terminal outer core region that can vary in structure between strains and can mimic human gangliosides, GM_2 and GM_3 and to a lesser degree, GD_{1b} and GD_2 (Guerry *et al.*, 2002). The generation of antibodies against LOS that cross-react with human gangliosides is thought to be responsible for the nerve damage observed in GBS (Yuki, 1997). Goodyear *et al* added weight to the molecular mimicry hypothesis by raising monoclonal antibodies against *C. jejuni* ganglioside mimicking LOS that were cross reactive with neural gangliosides and caused neurotransmission block in mice (Goodyear *et al.*, 1999). This molecular mimicry is due to the presence of *N*-acetyl neuraminic acid (NANA), also known as sialic acid, in both LOS and human gangliosides. NANA has been identified as a constituent sugar in the LOS of *C. jejuni* and three genes, (*neuB1*, *neuB2*, and *neuB3*), encoding NANA have been identified (Linton et al, 2000). Although compelling, molecular mimicry as the mechanism of GBS pathogenesis, remains unproven.

After 30 years of research into *C. jejuni,* the full mechanism of pathogenesis is still poorly understood. This is largely because there are no suitable animal models. Several animal models have been tested including mice (Newell and Pearson, 1984), colostrum deprived piglets (Babakhani *et al.*, 1993) and monkeys (Russell *et al.*, 1993). A ferret diarrhoeal disease model (Bacon *et al.*, 2001) and a rabbit ileal loop model (Everest *et al.*, 1993) have been used. However each of these animal models are expensive and difficult to use. Chick colonisation models (Medema *et al.*, 1992; Nachamkin *et al.*, 1993) have also been used to test mutants but since *C. jejuni* is a gut commensal in chickens, this is not an appropriate disease model. Consequently,

al., 1997). Birds also act as a reservoir which may account for the isolation of *C. jejuni* from sand samples (Broman *et al.*, 2002). The proportion of human disease attributable to these different sources of infection is unknown. In addition, studies of risk factors for human disease have consistently failed to identify an explanation for a large proportion of cases (Rodrigues *et al.*, 2001). Thus effective control strategies to minimise or eliminate Campylobacter from the food chain are difficult to implement.

1.4.2 Typing Methods

Phenotypic and genotypic methods for differentiating between *C. jejuni* strains were developed to aid surveillance and epidemiological studies. Traditionally Penner (heat-stable) (Penner and Hennessy, 1980) and Lior (heat-labile) serotyping (Lior *et al.*, 1982) have been used to classify *C. jejuni* strains. Such analysis was introduced to facilitate the tracing of routes and sources of *C. jejuni* infection and the identification of serotypes associated with post-infection neuropathies has been particularly noteworthy (Fujimoto *et al.*, 1997; Kuroki *et al.*, 1993).

More recently molecular techniques such as pulsed-field gel electrophoresis (PFGE) and multi locus sequence typing (MLST) have been used to type strains at a genetic level. Using PFGE the genome is cut into large fragments by rare cutting enzymes (*SmaI* and *KpnI*). DNA fragments are separated by coordinated application of pulsed electric fields from different positions in the electrophoresis cell. Thus, DNA fragments are orientated and separated according to size within agarose gel matrix. Profiles are visualised by staining in ethidium bromide. The band profile of isolates can then be compared and contrasted, with clonal strains identified as those possessing identical profiles (On *et al*, 1998).

MLST is a sequencing based technique that discriminates strains based on variations in the nucleotide sequences (including single nucleotide polymorphisms or SNPs), encoding seven housekeeping loci present in *C. jejuni.* These housekeeping loci were chosen by screening the *C. jejuni* genome database (www.sanger.ac.uk/projects/C_jejuni) for loci encoding enzymes responsible for intermediary metabolism. Suitable genes were then chosen based on a number of criteria including chromosomal location (a minimum distance of 70 kb suggests coinheritance in the same recombination event to be unlikely), suitability for primer design and sequence diversity (indicated from pilot studies). PCR products are amplified with primer pairs and amplification products are purified and sequenced.

Allele profiles are identified, described and then grouped into lineages or clonal complexes using the program BURST (Dingle *et al.*, 2001b). BURST is a heuristic algorithm available within the program START (Jolley *et al.*, 2001) used to identify existing sequence types (STs) and automatically assign ST's to their clonal complex. New ST's that do not belong to a previously identified clonal complex are recognised by the fact the software does not automatically assign them. MLST has been used to study the relationship between 814 Campylobacter strains from a variety of sources (Dingle *et al.*, 2002) indicating that *C. jejuni* strains are genetically highly diverse with a weakly clonal population structure (Dingle *et al.*, 2001a; Suerbaum *et al.*, 2001). These data also indicate that the population of *C. jejuni* strains exhibit significant genetic plasticity. An independent study of 184 *C. jejuni* strains found an unexpected association between strains isolated from cattle and humans suggesting a common source of infection (Schouls *et al.*, 2003). However, because of the carriage of strains of multiple types and an extremely high diversity of strains in animals current typing methods for Campylobacter strains is probably not useful for source tracing and global epidemiology. Although many methods have been described for typing *C. jejuni*, none have been shown to distinguish strains with phenotypic characteristics associated with pathogenesis and / or different ecological habitats. Thus the development of typing schemes to monitor human Campylobacter infections through the identification of foodborne sources and routes of transmission has been unsuccessful.

1.5 Socioeconomic factors

The IID study (Agency, 2000) (Chapter 1.2.1) indicated that *Campylobacter* affects half a million people in England and Wales annually, costing the economy of England alone an estimated 69.5 million pounds each year. There are several areas where these costs arise including the use of GP services such as consultations, microbiological tests and treatment. Furthermore, an average of 6.5 days per annum are taken off work by the adult working population resulting in lost revenue. *Campylobacter* disease also results in the inability to conduct normal household duties by carers and other non-working adults. Although hospital admittance due to *Campylobacter* disease is rare costs do arise due to days spent in hospital, hospital outpatient and accident and emergency visits. The estimated economic burden covers just the primary infection and does not take into consideration the impact of sequelae such as

GBS, MFS, bacteraemias and PD-IBS that have all been linked with *C. jejuni* infection (Allos, 1998; Ang *et al.*, 2001; Blaser, 1995; Thornley *et al.*, 2001).

1.6 The *C. jejuni* genome

The *C. jejuni* genome sequence of NCTC11168, an isolate from a patient with severe gastroenteritis, was annotated and published in 2000 (Parkhill *et al.*, 2000). The single circular chromosome, comprising 1654 predicted protein coding sequences (CDSs) (1,641,481 bp) had a GC content of 30.6% (Figure 1). Surprisingly there were very few repeat gene sequences, no plasmids and no insertion sequences or phage associated sequences revealing little about the mechanism of pathogenesis of this poorly understood pathogen. However, many short homopolymeric tracts were found in CDSs likely to encode surface structures. Such hypervariable sequences are thought to contribute toward the high levels of variation in surface structures and may be involved with survival of the organism in different ecological niches. More recently, a second *C. jejuni* genome, strain RM1221, was sequenced (Fouts, 2005). The single circular genome of this strain was larger than that of NCTC11168, at 1,777,831 bp with1884 CDSs and a slightly lower GC content of 30.31%. As with NCTC11168 no plasmids were found to be associated with the genome but a high proportion of predicted CDSs contained homopolymeric tracts. The average protein sequence percentage identity was calculated for all proteins matching the reference NCTC11168 strain. *C. jejuni* strains NCTC11168 and RM1221 had a high level of protein sequences in common with 1468 proteins averaging 98.41% identity. The incomplete genomes of three further Campylobacter species that rarely cause human disease, *C. coli*, *C. lari* and *C. upsaliensis*, were also recently published (Fouts, 2005). *C. coli* is frequently isolated from pigs but has also been isolated from other animals including cattle and dogs (Stephens et al, 1998). *C. coli* is responsible for 8% of *Campylobacter* infections reported to the PHLS each year making it the second most common *Campylobacter* species associated with human gastroenteritis (Tam *et al.*, 2003). *C. coli* showed the highest percentage identity to *C. jejuni* with 1399 protein sequences averaging 85.81% identity. *C. upsaliensis* showed the next highest level of identity to *C. jejuni* (1261 protein sequences averaging 74.72% identity). The natural host of this species is domestic cats and dogs where it may cause diarrhoea, indeed *C. upsaliensis* was first isolated from canine faeces. However, this species has been also associated with clinical disease, including rare outbreaks and septicaemia (Bourke *et*

al., 1998). *C. lari* showed the lowest level of identity to *C. jejuni* (1251 protein sequences with an average of 68.91% identity). This phenotypically, genotypically and ecologically diverse species has been isolated from wild birds, poultry, cattle, shellfish and water samples (Aarestrup *et al.*, 1997; Endtz *et al.*, 1997; Lastovica and le Roux, 2000; Skirrow, 1994).

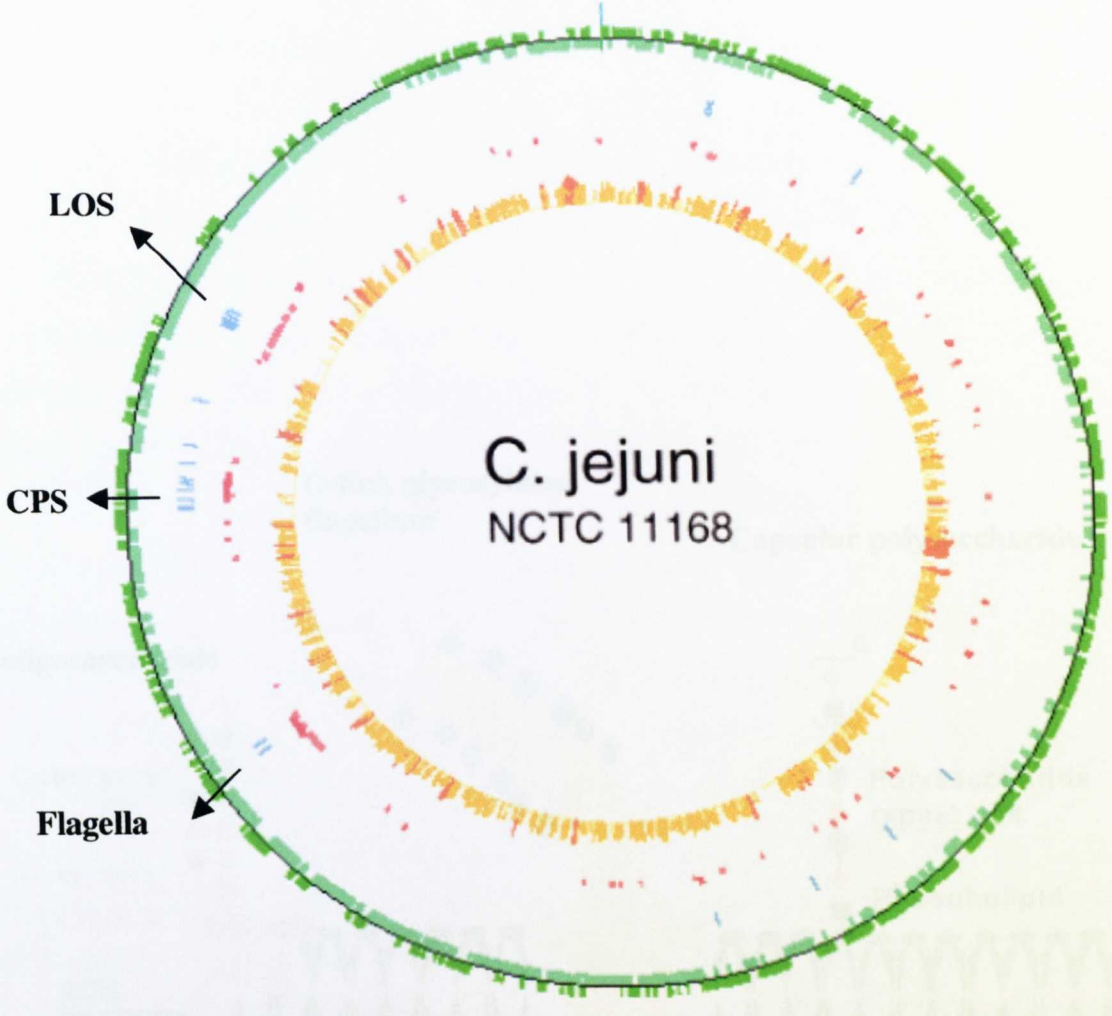

Figure 1. Diagrammatic atlas of the genome of *C. jejuni* strain NCTC11168. From the outside, the first circle (green) shows coding sequences transcribed in a clockwise direction. The next circle (also in green) shows coding sequences transcribed in an anticlockwise direction. The putative origin of replication is represented with a small vertical line at the top. Potentially phase variable genes containing homopolymeric tracts are shown in blue. Genes involved with surface structures transcribed in a clockwise direction are shown in red, those transcribed in an anticlockwise direction in pale red. The innermost histogram shows the similarity of each gene to its *Helicobacter pylori* orthologue, where the height of the bar and the intensity of colour are proportional to the degree of similarity. Adapted from Parkhill *et al*, 2000.

1.7 Virulence determinants

The genome sequence of NCTC11168 revealed distinct genetic loci for three major *C. jejuni* surface antigens; the flagella biosynthesis locus and two major surface-located glycolipids biosynthesised by *C. jejuni*, LOS and capsular polysaccharide (CPS) (Figure 2). In addition, outer membrane proteins such as the major outer membrane protein (MOMP) are important surface antigens allowing exchanges between the bacterium and the environment (Zhang *et al*, 2000).

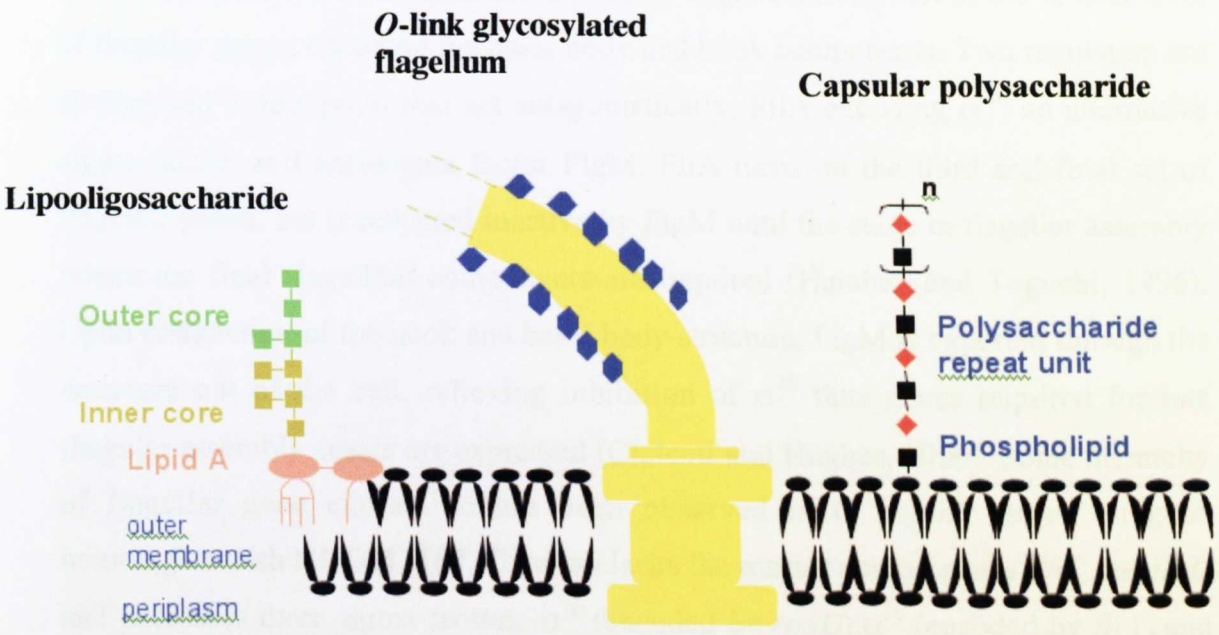

Figure 2. Schematic representation of the three major *C. jejuni* surface antigens, flagellum, LOS and CPS. Adapted from Karlyshev *et al*, 2005 in press.

1.7.1 Flagella

Flagella are the locomotory organelles of bacteria and are probably the best-characterised virulence determinant for *C. jejuni*. Motility facilitates dispersal, allows the cell to move towards and away from different environments and facilitates the penetration of the mucus layer of the intestine. The flagella biosynthesis locus contains over 50 genes (Figure 3) and it is therefore costly (in terms of energy) for the cell to both synthesise components of the flagellum, and to subsequently propel the bacterium. Consequently, flagella biosynthesis must be under tight control to prevent the expenditure of wasted energy by the cell. In *E. coli* and *S. typhimurium* a regulatory cascade determines the order in which flagella biosynthesis genes are expressed (Aizawa, 1996; Harshey and Toguchi, 1996, Chilcott and Hughes, 2000). Motility genes are organised into operons located in four regions of the genome. The master operon comprises regulatory genes *flhC* and *flhD* that are controlled by global cellular signals (cAMP, phosphorylated OmpR and heat shock proteins). The products of *flhC* and *flhD* act as a signal for the cell to begin transcription of the second level of flagellar genes, encoding the basal body and hook components. Two regulators are synthesised at this point that act antagonistically; FliA encoding σ^{28}, an alternative sigma factor, and anti-sigma factor FlgM. FliA turns on the third and final set of flagellar genes, but is rendered inactive by FlgM until the stage in flagellar assembly where the final flagellum components are required (Harshey and Toguchi, 1996). Upon completion of the hook and basal body structure, FlgM is exported through the structure out of the cell, relieving inhibition of σ^{28} thus genes required for late flagellar assembly stages are expressed (Chilcott and Hughes, 2000). Some hierarchy of flagellar gene expression has been observed in *C. jejuni*. Based on gene homologies with NCTC11168, *C. jejuni* lacks the master operon genes *flhC* and *flhD* and possesses three sigma factors, σ^{70} (encoded by *rpoD*),σ^{28} (encoded by *fliA*) and σ^{54} (encoded by *rpoN*) (Parkhill *et al.*, 2000). Early flagellar genes *flgS*, *flgR*, *RpoN* and *FliA* are regulated by σ^{70} (Wosten *et al*, 2004). Insertional inactivation of *fliA* encoding alternative sigma factor σ^{28} resulted in truncated flagella whereas insertional inactivation of *rpoN* encoding σ^{54} resulted in the complete absence of flagella (Jagannathan *et al.*, 2001). A *C. jejuni* homologue of FlgR found in *H. pylori*, part of the NtrC transcriptional activator family associated with σ^{54} promoters, was also insertionally inactivated and these mutants also lacked flagella (Jagannathan *et al.*,

2001). Furthermore, despite the absence of anti sigma factor *flgM* based on gene homologies with NCTC11168, a *flgM* homologue identified in *H. pylori* has since been identified in *C. jejuni* (Colland et al, 2000; Hendrixson and DiRita 2003). The flagellar filament in *C. jejuni* comprises flagellins FlaA and FlaB. Although adjacent to one another *flaA* and *flaB* genes are under the control of different promoters (σ^{28} and σ^{54} promoters respectively) (Wassenaar *et al.*, 1994). The complete lack of flagella in *rpoN* and *flgR* mutants indicates that *flaA*, although under control of a σ^{28} promoter, is no longer transcribed. However, in *H. pylori*, FlgR represses *flaA* transcription and deletion of *flgR* results in the upregulation of *flaA* (Spohn and Scarlato, 1999). In *C. coli*, inactivation of *flhA*, a member of the LcrD/FlbF family whose members are all integral cytoplasmic membrane proteins involved with the regulation of surface proteins results in a failure by the cell to synthesis flagellin (Doig *et al.*, 1996b; Miller *et al.*, 1993). Inactivation of *flhA* can occur naturally via slipped strand mispairing of a short poly T tract in the *flhA* gene leading to high frequency phase variation of gene expression (Park *et al.*, 2000). However, based on gene homologies with NCTC11168, the *flhA* gene in *C. jejuni* does not contain the homopolymeric tract present in *C. coli* thus flagellin gene expression in *C. jejuni* is not controlled by phase variation of *flhA* (Park *et al.*, 2000). *C. jejuni*, therefore, appears to have a different mechanism controlling the expression of flagellar components to both the well-characterised mechanisms in place in *S. typhimurium* and to its close relatives, *C. coli* and *H. pylori*. A two-component signal transduction system, FlgS/FlgR, is at the top of the hierarchy regulating the *fla* regulon in *C. jejuni*. RpoN-dependent genes encoding the flagellar hook-basal body filament complex are activated via the phosphorylation of FlgR (Wosten *et al*, 2004).

In *C. jejuni*, the construction of the flagellum begins with the basal body, containing the flagellar motor and switch. This anchors the flagellum to the cell and is embedded into the inner and outer membranes. The hook (encoded by *flgE*) is then constructed which acts as an axial coupling structure (Kinsella *et al.*, 1997). Transcription of *flgE* is under the control of σ^{54} promoter activity (Luneberg, Glenn-calvo et al, 1998). Finally, once the other components are in place, the filament is constructed. In *C. jejuni* flagellin is encoded by *flaA* and *flaB*, genes that are 95% identical at the nucleotide level (Nuijten *et al.*, 1990). Flagella filaments are composed predominantly

of the FlaA flagellin interspersed with the variable FlaB flagellin (Guerry *et al.*, 1991).

Rotation of the flagellum is driven by the motor in the cell envelope, mediated through the flagella hook. A two component regulatory system, CheA and CheY, regulates chemotaxis in response to attractant and repellent gradients. Phospho-CheY interacts with the flagellar motor to control swimming behaviour (Yao *et al.*, 1997).

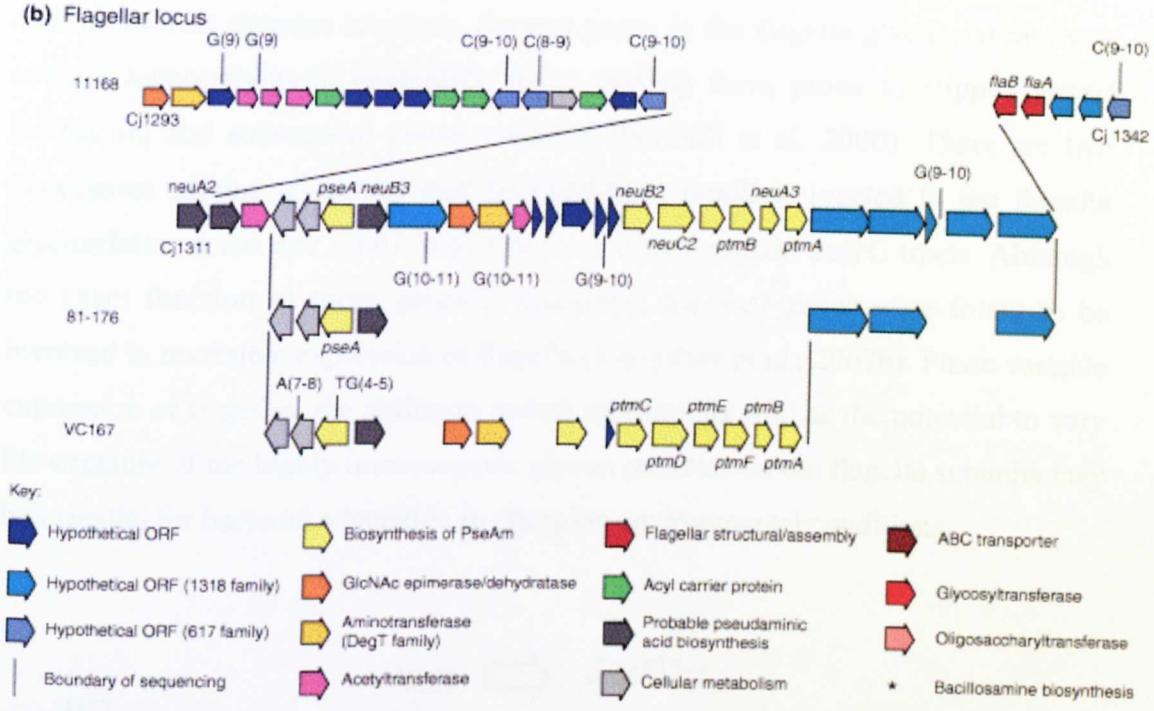

Figure 3. Schematic representation of the flagellar locus including the *O*-linked glycosylation locus in *C. jejuni* strains NCTC11168, 81-176 and *C. coli* strain VC167. Adapted from (Szymanski *et al.*, 2003)

Post-translational modification of flagella through *O*-linked glycosylation has been identified in *C. jejuni* and *C. coli* (Doig *et al.*, 1996b). Pseudaminic acid (Figure 4) is *O*-linked to as many as 19 serine and threonine residues in the central, hydrophilic, surface exposed region of the flagellin protein of *C. jejuni* strain 81 176 (Logan *et al.*, 2002; Thibault *et al.*, 2001). Flagellin proteins are the only known *O*-linked glycosylated molecules in *C. jejuni* and although biological significance for this extensive modification is unknown, the modification appears to be important for flagella assembly, as mutations in some genes involved in pseudaminic acid biosynthesis (e.g. *cj1293*, *pseB*) may result in non-motile and aflagellate cells (Linton *et al.*, 2000). The intracellular accumulation of unmodified flagella subunits in these mutants (Logan *et al.*, 2002), suggests that glycosylation may be required for the

recognition by flagella secretion/assembly apparatus (Thibault *et al.*, 2001). However, these studies were performed in 81-176 that has a truncated glycosylation locus compared to the sequenced strain NCTC11168 (see Figure 3).

The flagellum is a highly antigenic target. Recognition of the flagellin monomers by host cell receptors triggers proinflammatory and adaptive immune responses in enteric pathogens such as *S. typhimurium* and *E. coli* (Berin *et al.*, 2002; Eaves-Pyles *et al.*, 2001). Variation in cell-surface exposed glycan structure may assist in bacterial evasion of host immune response. Several genes in the flagella glycosylation locus contain homopolymeric nucleotide tracts making them prone to slipped strand mispairing and subsequent phase variation (Parkhill et al, 2000). There are two paralogous groups of genes, *maf* (*cj1318*-like) families, located in the flagella glycosylation locus and *cj0617/0618* both of which contain polyG tracts. Although the exact function of these genes is unknown, the *maf* genes were found to be involved in reversible expression of flagella (Karlyshev *et al.*, 2002b). Phase variable expression of flagellin, the ability to switch motility, as well as the potential to vary the structure of the highly immunogenic glycan moieties on the flagella subunits may be essential for bacterial adaptation to changing environmental conditions.

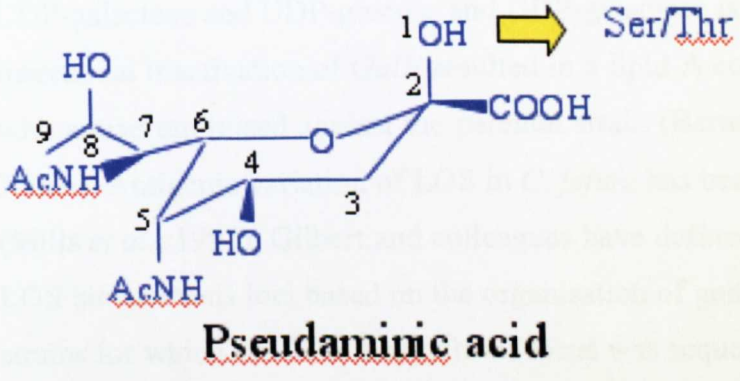

Figure 4. Diagrammatic structure of pseudaminic acid found *O*-linked to serine and threonine residues on *C. jejuni* flagellin. Adapted from Karlyshev *et al*, 2005.

1.7.2 Lipooligosaccharide

Lipopolysaccharide (LPS) is the predominant glycolipid found in the cell wall of Gram-negative bacteria. This predominant microbial antigen is a major bacterial virulence factor, exhibiting endotoxic properties as well as roles in adhesion and antigenic variation. In Enterobacteriacae, LPS comprises three distinct regions;

anchored in the outer membrane is the endotoxic lipid A moiety. Linked to the lipid A moiety through 2-keto-3-deoxyoctulosonic acid (KDO) is a core consisting of an inner and outer region. Finally, attached to the outer core is a long chain of covalently linked oligosaccharides. In *C. jejuni*, the major glycolipid molecule in the cell wall is lipooligosaccharide (LOS) the structure of which is analogous to LPS but with a truncated oligosaccharide chain, limited to only10 saccharide units (Preston *et al.*, 1996) (Figure 5). LOS, like LPS exhibits endotoxic and adhesive properties (McSweegan and Walker, 1986; Naess and Hofstad, 1984). The genetic locus responsible for the biosynthesis of LOS in *C. jejuni* has been identified, *cj1120* to *cj1152*. A gene cluster involved with a general *N*-linked protein glycosylation system in *C. jejuni* directly flanks the LOS biosynthesis locus at the 5' end (Figure 6). These genes were originally thought to be associated with LOS biosynthesis. However, mutagenesis in genes from this locus in *C. jejuni* strain 81-176 resulted in no change to LOS but altered the reactivity of multiple *C. jejuni* proteins to human and rabbit sera (Szymanski *et al.*, 1999; Szymanski *et al.*, 2002). The metabolic pathways and enzymes required to synthesise the antigen are not fully characterised. However *cj1131*(*galE*) has been shown to have an essential role in LOS biosynthesis, encoding UDP-galactose-4-epimerase. The product of *galE* catalyses the interconversion of UDP-galactose and UDP-glucose and UDP-galactose is used in the synthesis of LOS. Insertional inactivation of *GalE* resulted in a lipid A core molecule that did not react with antiserum raised against the parental strain (Bernatchez *et al.*, 2005; Fry *et al.*, 2000a). Antigenic variation of LOS in *C. jejuni* has been documented for many years (Mills *et al.*, 1992). Gilbert and colleagues have defined three classes, A, B and C, of LOS biosynthesis loci based on the organisation of genes in eleven different *C. jejuni* strains for which the LOS biosynthesis locus was sequenced. Class B appears to be an evolutionary intermediary between classes A and C (Gilbert *et al.*, 2002). The sequencing of this locus has revealed further genetic mechanisms for LOS variation including different gene complements, gene inactivation by deletion or insertion, a single mutation leading to inactivation of a glycosyltransferase and single or multiple mutations leading to allelic glycosyltransferases with different acceptor specificities (Gilbert *et al.*, 2002). Furthermore, the genome sequence of NCTC11168 revealed a mechanism for phase variation due to homopolymeric tracts in genes in the LOS biosynthesis locus (Parkhill *et al.*, 2000). Cell surface structures such as LOS are recognised as antigens by the host and lipid A is responsible for many host immune

responses such as platelet aggregation and cytokine activation (Preston *et al.*, 1996). It is therefore advantageous to the microbe to modulate its surface coat in order to evade the host immune system (Gilbert *et al.*, 2002).

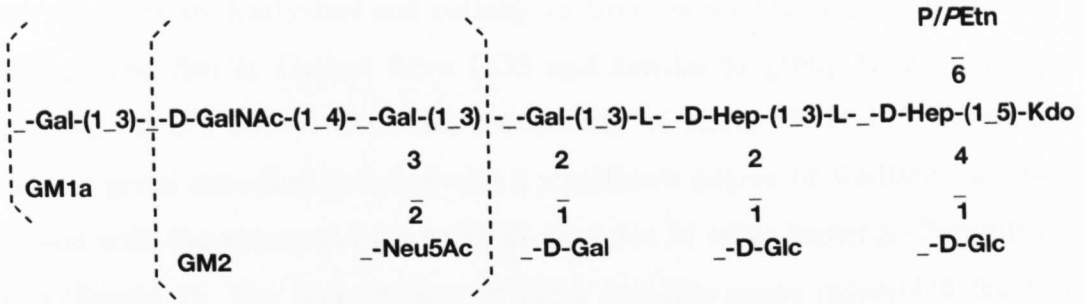

Figure 5. The complete LOS core structure of NCTC11168. Structures analogous to gangliosides GM1 and GM2 are shown in brackets. KDO, 2-keto-3-deoxyoctulosonic acid, Hep, L-glycero-D-mannoheptose, Glc, glucose, Gal, galactose, Neu5Ac, N-acetylneuraminic acid, GalNAc, N-acetyl-D-galactosamine. (from Karlyshev et al, 2005)

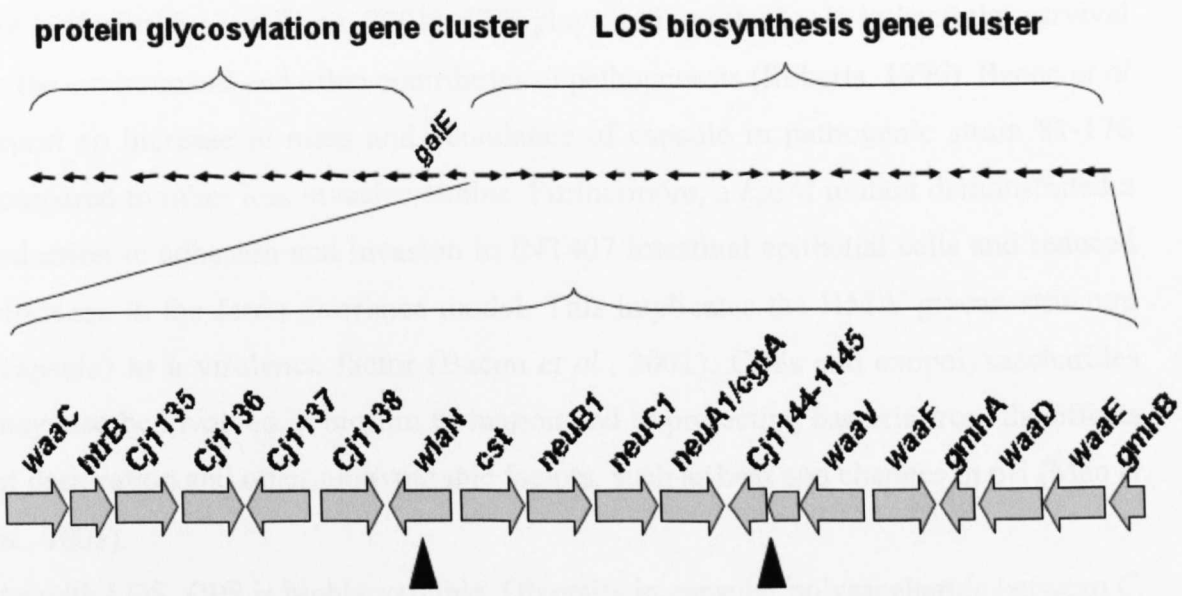

Figure 6. Schematic of the LOS biosynthesis locus and flanking genes associated with *N*-linked protein glycosylation in NCTC11168. The ORF names are written above the arrows denoting the direction of transcription. ORFs containing homopolymeric tracts are indicated with black arrows (from Karlyshev et al, 2005).

1.7.3 Capsule

Another surface structure important in *C. jejuni* virulence and survival is capsular polysaccharide. It was thought that all *C. jejuni* strains produced LOS with a third producing high molecular weight (HMW) glycan structure, known as O-antigen. However, work by Karlyshev and colleagues have shown that all strains possess the HMW glycan that is distinct from LOS and similar to group II and III capsule (Karlyshev *et al.*, 2000). Genomic sequencing of strain NCTC11168 revealed a cluster of genes encoding proteins with a significant degree of similarity to proteins involved with the transport of type II/III capsules in other bacteria (Parkhill *et al.*, 2000) (Figure 7). The arrangement of these *kps*-like genes resembled the *E. coli* paradigm; conserved *kps*-like genes associated with transport of the molecule flank a central, variable 34 kb region, involved in the biosynthesis of diverse polysaccharide structures (Karlyshev *et al.*, 2000). Site-specific insertional mutagenesis of *kpsM*, *kpsS* or *kpsC* in several strains resulted in the loss of a HMW glycan. Karlyshev and colleagues went on to use electron microscopy (EM) to demonstrate that *C. jejuni* produces a polysaccharide capsule that surrounds the cell surface (Karlyshev *et al.*, 2001; Karlyshev and Wren, 2001). CPS plays an important role in bacterial survival in the environment and often contributes to pathogenesis (Roberts, 1996). Bacon *et al* report an increase in mass and abundance of capsule in pathogenic strain 81-176 compared to other less invasive strains. Furthermore, a *kpsM* mutant demonstrated a reduction in adhesion and invasion in INT407 intestinal epithelial cells and reduced virulence in the ferret diarrhoea model. This implicates the HMW glycan structure (capsule) as a virulence factor (Bacon *et al.*, 2001). CPSs and exopolysaccharides may also be involved in biofilm formation and in protecting bacteria from the effects of desiccation and other unfavourable factors, such as heat and changes in pH (Mao *et al.*, 2001).

As with LOS, CPS is highly variable. Diversity in capsular polysaccharide between *C. jejuni* strains occurs by multiple mechanisms including exchange of capsular genes and gene clusters by horizontal transfer, gene duplication, deletion, fusion and contingency gene variation (Karlyshev *et al.*, 2005). Furthermore, *kpsM*, *kpsS* and several genes in the central biosynthetic region are likely to be phase variable as they possess intragenic homopolymeric tracts (Parkhill *et al.*, 2000). *C. jejuni* virulence determinants that have been identified to date are summarised in Table 1.

Figure 7. CPS biosynthesis loci in five *C. jejuni* strains, NCTC11168, HS:19, G1, 81-176 and HS:41. Polysaccharide transport or assembly genes are coloured red. Sugar transferases are shown in blue, genes associated with heptose biosynthesis are shown in pink. Other sugar biosynthesis related genes are shown in green. Genes with no obvious link to sugar biosynthesis are coloured yellow and genes with no similarity or similarity to hypothetical genes are shown in brown (Karlyshev *et al.*, 2005)

Table 1. Summary of putative virulence factors identified in *C. jejuni*

Putative virulence factor	Associated CDS	Reference
Capsular polysaccharide	*cj1410* to *cj1449*	(Karlyshev *et al.*, 2000)
Lipooligosaccharide	*cj1120* to *cj1152*	(Preston and Penner, 1987)
Flagella and *O*-linked glycosylation	*cj1293 to cj1342*	(Newell *et al.*, 1984) (Thibault *et al.*, 2001)
Cytolethal distending toxin	*cj0077, cj0078, cj0079*	(Eyigor *et al.*, 1999)
Putative virulence plasmid encoding type IV secretion system	*pVir*	(Bacon *et al.*, 2000)
Adhesion to fibronectin	*cadF*	(Konkel *et al.*, 1997)
Adhesion to INT 407 cells	*galE*	(Fry *et al.*, 2000)
Adhesion to Hep-2 cells	*jlpA*	(Jin *et al.*, 2001)

1.8 DNA microarray technology

The availability of the genome sequence for the *C. jejuni* strain NCTC11168 (Parkhill *et al.*, 2000) paved the way for the construction of *C. jejuni* DNA microarrays to investigate the relative genome content of different strains. Whole genome microarrays have the potential to demonstrate the link between genetic characteristics and the ability of a strain to inhabit a preferred ecological niche and /or its ability to cause disease. DNA microarrays have been used to compare the genome content of several bacterial pathogens. The application of microarrays for comparative genomics ranges from bacterial pathogens exhibiting wide diversity, genome plasticity to the pathological outcome of infections. This technique, allied with the increasing availability of genome sequence data, is set to revolutionise our ability to distinguish bacteria. DNA microarray analyses have revealed a vast genetic diversity both between genera and within species. Deciphering the mechanisms behind this variability will help us to understand the phylogeny, physiology, ecology and evolution of bacteria. Comparing the genomes of pathogens and non-pathogens within a species can be particularly useful for identifying determinants that are important in virulence, transmission and host specificity.

1.8.1 DNA microarray terminology

In molecular biology terms a "probe" has traditionally been used to describe the labelled sample in solution that hybridises to a bound target. However the term "probe" with respect to microarray experiments has often been used not to describe the labelled sample hybridised to the array, but the element attached to the array. This confusion highlighted the need for a standard terminology that has been implemented through development of MAGE-ML (MicroArray Gene Expression Markup Language; http://www.mged.org). MAGE-ML avoids the use of the term "probe" altogether through the use of the term "reporter" to refer to the elements bound to the microarray surface and "sample" as the labelled nucleic acid in the hybridisation solution that is derived from the DNA or RNA preparation. The "core genes" or "minimal gene set" of a species refers to the number of genes that are present in the genome of each strain hybridised with the microarray. These core genes are necessary to maintain the living cell and are involved with metabolic, biosynthetic, cellular and regulatory processes. As more strains are hybridised so a clearer picture is gained as those genes that are absent or variable from strains exhibiting particular traits can be eliminated from the core genes (Koonin 2003). Conversely, the "variable genes" or "accessory genes" of a species comprise those genes identified as absent or divergent in one or more of the strains hybridised with the microarray. Variable genes are therefore not required for survival of the species but may correlate with specific phenotypic traits. The term "flagged genes" refers to those reporter elements on the microarray that did not produce data of sufficient quality to analyse. This may be due to problems with the actual reporter element printed on the microarray e.g. a reporter element printed on the microarray from small PCR product that produced only a weak band may yield hybridisation data with low signal intensities. Alternatively flagged genes may arise due poor experimental technique such as insufficient or slow washing and rinsing of the microarray slides.

Membrane arrays are usually referred to as "macroarrays" due to the larger format used and the low density of reporter elements on the array. Reporter elements printed on nylon membranes have included both unknown sequences derived from cDNA libraries and sequence defined PCR products. The labelled samples are hybridised to the reporter elements on the membranes using protocols similar to Southern blots. For comparison between different samples, the hybridisation signals from one membrane

must be compared to signals from another membrane, which can lead to problems with normalising the different signal intensities from membranes hybridised at different times.

Glass slide arrays are referred to as "microarrays" as the spot density of the reporter elements is much higher (> 20 fold) than can be achieved using macroarrays. A complete genome sequence enables the design of whole genome arrays, whereby at least one reporter element is produced for every gene in the genome, but may also include intergenic regions, any other additional genomic features or genes from other organisms. The reporter elements are first synthesised, with double-stranded DNA reporter elements often in the form of PCR products, generated from gene-specific PCR or amplified clone inserts of DNA/cDNA libraries, or oligonucleotides. These reporter elements are then spotted onto specially coated glass microscope slides (Hinds, 2002). Direct comparison of two samples (both test and reference conditions) can be achieved on a single array. Competitive hybridisation is one of the major advantages of microarray-based experiments. The differences in signal intensity produced by hybridisation of the two labels can be detected using a dual laser confocal scanner. Another benefit of the microarray approach is the flexibility that allows the arrays to be easily customised to suit particular applications.

Oligonucleotide arrays do not rely on PCR amplification of a gene but utilise oligonucleotides as reporter elements. The target DNA sequence must be known to facilitate oligonucleotide synthesis, which occurs *in situ*, on glass or silicon wafers using combinational chemistry and photolithography (Pease *et al.*, 1994). Oligonucleotide arrays have several advantages over more traditional PCR products; tens of thousands of reporter elements comprising custom designed 20-70 mers promises increased specificity whilst maintaining the sensitivity of PCR products and allowing better discrimination between gene families and overlapping genes. Oligonucleotide reporter elements are more uniform in quality than PCR products, However, cross-hybridisation between homologous DNA sequences is still a problem when analysing array data using double-stranded DNA reporter elements, particularly when performing gene expression studies. Microarrays printed with PCR products may not allow detection of differential expression patterns for highly homologous or overlapping genes. In addition, the PCR amplification process involved in generating the individual reporter elements is labour intensive and expensive, and can have a

high failure rate (typically 5–10%). However, oligonucleotide arrays are often more expensive to produce than PCR based arrays.

Oligonucleotide arrays are available from a number of companies including Affymetrix and Qiagen. The GeneChip® *E. coli* Genome Array was the first Affymetrix microarray product for a prokaryotic organism. *Campylobacter* oligo arrays are also commercially available. For example, one manufactured by MWG Biotech in association with Affymetrix (www.mwg-biotech.com) and another by Qiagen (www.qiagen.com).

1.8.2 Selected examples of the application of microarrays for comparative genomics

DNA microarrays have been used to investigate bacterial pathogens exhibiting wide diversity, genome plasticity, differences in pathological outcome of infections and to study genomes of related species. Dobrindt *et al* 2001 proposed that bacterial genomes consist of a core of genetic material that are conserved in most strains, a minimal gene set shared by the vast majority of bacteria and a flexible pool of strain specific genes allowing the organism to adapt to its environment (Table 2). These principals are borne out in whole genome comparative microarray studies on *Yersinia pestis* (Zhou et al., 2004). Forty-three *Y. pestis* strains of biovars antiqua, orientalis, mediaevalis and microtus were selected to represent the diversity associated with the adaptive evolution of plague. These strains were hybridised with *Y. pestis* microarray containing 4005 ORFs from *Y. pestis* strains CO92 and those unique to strain 91001. Microarray data in this study indicates that the acquisition of genomic islands and plasmids in *Y. pestis* induced the organism's rapid evolution from its close relative, *Y. pseudotuberculosis*. DNA microarrays have also been used to study the evolutionary genomics of 22 strains of *Yersinia pestis* and 10 strains of *Y. pseudotuberculosis*. The loss of eleven DNA loci in *Y. pseudotuberculosis* is thought to have led to the rapid emergence of the life threatening *Y. pestis* species (Hinchliffe *et al.*, 2003). This rapid evolution is believed to have occurred only 1,500 to 20,000 years ago (Achtman *et al.*, 1999). Horizontal gene transfer is also responsible for the diversity between and segregation of the distinct *Y. pestis* biovars. However, genome reduction was also important in the parallel microevolution of *Y. pestis* leading to distinct *Y. pestis* biovars. Biovar Orientalis, associated with modern plague, was responsible for the

mid 19th century plague pandemic that originated in China and continued to spread globally affecting over 60 countries. Microarray data from this study suggests that biovar Orientalis arose from biovar Antiqua, associated with the Justinian plague through the acquisition of a genomic region (DFR13). The authors propose that *Y. pestis* may evolve through genome reduction into a unique ecological niche (natural environment, reservoirs and vector(s)) causing the progeny to inhabit a host niche that does not overlap with its progenitor (Zhou *et al.*, 2004).

Whole genome microarray experiments have been applied to assess the genome plasticity of the laboratory *Mycobacterium tuberculosis* strain H37Rv with the closely related species *Mycobacterium bovis* and passage derivatives of *M. bovis* BCG vaccine strain (Behr *et al.*, 1999). The original BCG vaccine strain was passaged *in vitro* at least 230 times resulting in a strain that maintained immunogenicity but had reduced virulence. Since then, this strain has continued to be passaged resulting in a collection of daughter strains with different phenotypes. The genomic composition of *M. bovis*, *M. tuberculosis* and the BCG daughter strains were compared in order to explain why BCG efficacy varies in human trials. The differences observed indicate that since original derivation, the BCG strains have evolved which may have resulted in loss of protective efficacy. Compared to *M. tuberculosis* 11 regions of 91 ORFs were deleted from one or more pathogenic *M. bovis* strains. For the attenuated BCG vaccine strains, an additional five regions of 38 ORFs had been lost. Transcriptional regulators were proportionally the most common class of genes lost from BCG strains. These results demonstrate that microarray analysis can be used to reconstruct the genealogy of related strains at the genome level.

Table 2. Examples of the application of DNA microarrays for comparative genomics of pathogenic bacteria

Arrayed strain	Number of samples	% variability of genome	Reference
Escherichia coli K12 MG1655	4	10	(Ochman and Jones, 2000)
Mycobacterium tuberculosis H37Rv	19	0.3	(Kato-Maeda *et al.*, 2001)
Mycobacterium tuberculosis H37Rv	13	N/a	(Behr *et al.*, 1999)
Helicobacter pylori J99	42	3	(Israel *et al.*, 2001)
Helicobacter pylori J99 & NCTC 26695	15	22	(Salama *et al.*, 2000)
Campylobacter jejuni NCTC11168	11	21	(Dorrell *et al.*, 2001)
Campylobacter jejuni NCTC11168	18	16	(Pearson *et al.*, 2003)
Campylobacter jejuni NCTC11168	51	36	(Taboada *et al.*, 2004)
Staphylococcus aureus COL	36	22	(Fitzgerald *et al.*, 2001)
El Tor 01 *Vibrio cholerae* N 16961	9	1	(Dziejman *et al.*, 2002)
Yersinia pestis and *Yersinia pseudotuberculosis*	22 *Y. pestis* 10 *Y. pseudotuberculosis*	4 (*Y. pestis*) 16 (*Y. pseudotuberculosis*)	(Hinchliffe *et al.*, 2003)
Yersinia pestis	46 *Y. pestis*	N/a	(Zhou *et al.*, 2004)

It has been hypothesised that the *V. cholerae* strains associated with the current seventh cholera pandemic have acquired genes that may have facilitated the displacement of the pre-existing classical *Vibrio cholerae* strains (Dziejman *et al.*, 2002). The first six recorded cholera pandemics occurred between 1817 and 1923 and were caused by the classical biotype of *V. cholerae*. However in 1961 the El Tor biotype emerged to cause the seventh cholera pandemic and resulted in the eventual global replacement of the classical biotype strains as a cause of disease. A *V. cholerae* microarray based on the El Tor O1 strain N16961, was used to analyse a collection of nine strains of diverse global origin isolated between 1910 and 1992 to investigate this hypothesis. It was possible to differentiate classical biotype strains from El Tor biotype strains and two putative chromosomal islands (VSP-I and VSP-II) with a deviant G+C content were identified in El Tor biotypes. Genes associated with the VSP-1 and VSP-II islands may encode key properties that led to the global success of the El Tor biotype and studies are in progress to determine their potential role both in human infection and in promoting the fitness of *V. cholerae* in environmental ecosystems.

Similarly, the pathological outcome of *Neisseria* species infections has been investigated using microarrays (Perrin *et al.*, 2002). *Neisseria meningitis* (a causative agent of meningitis) is very closely related to *N. gonorrhoeae* (the causative agent of gonorrhoea) and *N. lactamica,* (a harmless commensal of the nasopharynx), yet their disease profiles are very different. This study revealed a series of relatively small sequences scattered throughout the genome which were either specific to *N. meningitis* or shared with *N. gonorrhoeae*, but absent from *N. lactamica*. This study confirmed that the capsule biosynthesis loci and the RTX toxin family were meningococcal specific.

Microarray analysis can allow the rapid identification of genes in the genome of an unsequenced bacterium using a microarray that has been constructed for a closely related species. A microarray was constructed consisting of 96 putative virulence factor genes from an orally toxic *Photorhabdus luminescens* strain W14 (Marokhazi *et al.*, 2003). *Photorhabdus* species are found in the guts of nematodes that invade insects. Some strains are orally toxic and kill the insect host, whilst other strains are non-toxic. The aim of this study was to identify the minimal subset of toxin complex (tc) genes required for oral toxicity by hybridising genomic DNA from both orally toxic and non-toxic strains to this array. A striking split was found in the distribution

of tc genes between orally toxic and non-toxic strains. Orally toxic strains were found to carry all three genes in the *tca* operon (*tcaABC*), whereas those lacking oral toxicity lack *tcaA* and *tcaB*. This study clearly demonstrates that even a very limited microarray can be an extremely useful tool for correlating phenotype and genotype, with particular reference to virulence.

The above are only a small sample of recent microarray-based comparative genomic studies of different bacterial species that were selected to highlight the range of applications for comparative genomic studies that can be undertaken using microarrays.

1.8.3 *C. jejuni* comparative genomics studies

An initial comparative genomics study of 11 human *C. jejuni* isolates using a *C. jejuni* clone array revealed extensive genetic diversity between these *C. jejuni* strains (Dorrell *et al.*, 2001). Many of the strain-variable genes are associated with the biosynthesis of surface structures, including flagella, lipo-oligosaccharide and capsule. This suggests that variation of these determinants may be important in survival, transmission and pathogenesis, indicating that selective pressures have driven profound evolutionary changes to create a diverse *Campylobacter* species. Comparison of the capsule biosynthesis locus reveals conservation of all the genes in this region in strains with the same Penner serotype as strain NCTC11168. By contrast, between five and seventeen of the NCTC11168 genes in this region are either absent or highly divergent in strains of a different serotype to the sequenced strain, providing further evidence that the capsule accounts for Penner serotype specificity (Dorrell *et al.*, 2001). In all at least 21% of the genes present in the sequenced strain appear dispensable, as they are absent or highly divergent in one or more of the isolates tested, defining 1300 out of 1654 predicted CDSs as *C. jejuni* species-specific. These genes mainly encode housekeeping functions such as metabolic, biosynthetic, cellular and regulatory processes. However, many virulence determinants are also conserved, indicating that they are indispensable for *C. jejuni* to cause disease in humans. These include the cytolethal distending toxin, the flagellar structural proteins, the PEB antigenic surface proteins and the general protein glycosylation locus (Dorrell *et al.*, 2001).

Four more *C. jejuni* comparative genomics studies have been published since this original study and the commencement of this thesis using gene specific DNA

microarrays. A comparison of random amplified polymorphic DNA (RAPD) and Penner serotyping with DNA microarray analyses of clinical isolates associated with five independent clusters of infection indicated that DNA microarrays provide a highly specific epidemiological typing tool for analysis of *C. jejuni* isolates (Leonard *et al.*, 2003). In another study the genomic diversity of 18 *C. jejuni* strains from diverse sources was investigated (Pearson *et al.*, 2003). Seven hypervariable plasticity regions (PR) were identified in the genome (PR1 to PR7). PR1 contains genes important in the utilisation of alternative electron acceptors for respiration and may confer a selective advantage to strains in restricted oxygen environments. PR2, 3 and 7 contain many outer membrane and periplasmic proteins and hypothetical proteins of unknown function that might be linked to phenotypic variation and adaptation to different ecological niches. PR4, 5 and 6 contain genes involved in the production and modification of antigenic surface structures including the flagellin glycosylation locus. In this study algorithms were used that selected a dynamic boundary between the conserved and variable genes similar to the Genomotyping Analysis Charlie Kim (GACK) algorithm (Kim *et al.*, 2002). More recently genomic comparisons of 51 strains isolated from food and clinical sources have been integrated with data from three previous *C. jejuni* DNA microarray studies to perform a meta-analysis that included 97 strains from the four separate data sets (Taboada *et al.*, 2004). In this study a large proportion of the variable genes were found to be absent or divergent in single strains only and these uniquely variable genes could be mapped to previously defined variable loci. Thus the authors propose large regions of the *C. jejuni* genome are genetically stable. Of the highly divergent genes that were identified 117 of 122 genes had divergent neighbours and showed high levels of intraspecies variability (Taboada *et al.*, 2004).

Microarrays have been used to investigate genetic markers capable of differentiating strains that cause GBS from those that cause uncomplicated enteritis. Significant genomic heterogeneity among the isolates was revealed but no specific GBS genes or regions were identified (Leonard *et al.*, 2004). DNA microarrays have also been developed to identify *C. jejuni* directly from faecal cloacal swabs (Keramas *et al.*, 2003) in an effort to comply with consumer demands for food safety. Universal bacterial sequences and specific Campylobacter sequences were amplified using multiplex-PCR. This DNA-microarray facilitated the detection of two closely related Campylobacter species, *C. jejuni* and *C. coli* directly from chicken faeces. A non-

NCTC11168 microarray based on *C. jejuni* strain ATCC 43431 has also been developed. A shotgun DNA microarray to identify genes specific to this ATCC 43431 was constructed by spotting 9,000 DNA fragments from a *C. jejuni* ATCC43431 genomic library onto glass slides and hybridising with strain NCTC11168 (Poly *et al.*, 2004). Gene fragments unique to strain ATCC 43431 were sequenced and putative functions were assigned. However, none of the aforementioned studies investigated strains from diverse disease outcome or different ecological niches. Furthermore the sample size used in these studies and from studies on other bacteria (see table 2) was relatively small and the phylogenetic relationships of the strains were not thoroughly examined.

1.9 Phylogenomics

Phylogenomics constitutes the inferral of phylogenies using comparative genomic data from DNA micorarray comparisons or by comparing genome sequences. Such 'whole genome' trees allow evolutionary analysis to be introduced to comparative genomics studies. This method facilitates the identification of clonal populations with similar phenotypic traits. Moreover, genomic similarities and differences of phylogenetically related strains may provide genetic markers specific to strains from common ecological niches or strains causing a particular disease outcome. Such genetic markers may then be further investigated for their biological significance.

Many molecular phylogenetic trees utilise highly conserved and ubiquitous genes such as rRNA that may be amplified and sequenced to infer phylogenies. However, rRNAs are not representative of the microbial genome demonstrated by the discrepancy between phylogenies constructed from protein encoding gene sequences and those from rRNA sequences. The noise and bias of single gene analysis may be reduced through the inferral of phylogeny based on whole genome comparisons (Charlebois *et al.*, 2003). However, such analyses have only been possible due to the completion of genome sequences and the development of high-density microarrays. Current genotyping techniques infer relationships of strains based on a single gene or several gene sequences. Consequently relationships based on genes other than those used for the typing method may be missed.

The initial aim of this project was to develop an improved *C. jejuni* molecular typing system based on genetic markers identified through DNA microarray studies to further the understanding of *C. jejuni* epidemiology. The CPS locus was identified as

the genetic basis for Penner serotyping (Karlyshev *et al.*, 2000) and DNA microarray studies confimed the correlation between variation in Penner serotype and variation in CPS biosynthesis genes (Dorrell *et al.*, 2001). Subsequently, the CPS locus was selected for further investigation to identify potential genetic markers on which to base a molecular epidemiological typing system. We sequenced the full CPS locus of five *C. jejuni* strains representing five different serotypes (HS1, HS19, HS41, HS23 and HS23/36). Structural analyses of the CPS was also performed revealing a good correlation between gene sequence and structure (Karlyshev *et al.*, 2005) (Appendix 7).

Although the detailed genetic study of CPS identified potential genetic markers, we reasoned that clonal populations with common phenotypes would be identified using a comparative phylogenomics approach, through the combination of microarray hybridisation data and robust phylogenetic algorithms. When performed on a relatively large and diverse defined *C. jejuni* strain collection, such analyses may provide novel information on genes relating to ecological niche or disease severity, highlighting more pertinent genetic markers on which to base a molecular epidemiological typing system. Therefore, in this study, strains from the United Kingdom, isolated from a range of ecological niches and clinical outcomes of infection, were selected for detailed comparative phylogenomic analyses.

1.10 Aims of this study

Although several comparative genomics studies of *C. jejuni* have been published, to date none have used phylogenomic analyses to investigate the mechanism of pathogenesis of *C. jejuni.* Furthermore, this method has not been used to identify genetic differences in strains isolated from different ecological niches and thus different sources of human infection. Therefore, the aims of this study were;

- To use the gene specific, composite *C. jejuni* DNA micorarray to compare different methods of analysis to ascertain the optimal method for assessing absent or divergent genes.
- To identify potentially non-pathogenic *C. jejuni* strains that may be used for whole genome comparisons through DNA microarray hybridisations to facilitate the identification of putative virulence determinants.

- To undertake whole genome comparisons on strains from a range of clinical outcomes to identify potential genetic markers characteristic of disease outcome.
- To develop a phylogenomics method to model *C. jejuni* phylogeny and allow the identification of clade specific genetic markers that may be widely applicable for studying other bacterial pathogens.
- To apply comparative phylogenomics to investigate *C. jejuni* strains from a range of hosts and environmental sources facilitating the identification of clones associated with different hosts.
- To identify genetic markers that may be informative for epidemiological studies.

2.0 Materials and Methods

2.1 Microbiology

2.1.1 Bacterial growth conditions

Campylobacter jejuni strains (Appendix 1) were streaked on Columbia blood agar plates (Fluka Biochemika, Steinheim, Switzerland), containing 5% (v/v) horse blood in a Variable Atmosphere Incubator (Don Whitley Scientific, Shipley, UK) for 48 hours. A microaerobic atmosphere of O_2 5%, N_2 85%, CO_2 10% was maintained at 37°C.

2.1.2 Storage of *C. jejuni* strains

Strains were stored in 50% Mueller Hinton (MH) Broth (Oxoid, Basingstoke, UK) and 50% glycerol at -80°C.

2.1.3 Motility plates

Motility of *C. jejuni* strains was tested on 0.4% bacto™ agar plates (Scientific Laboratory Supplies Ltd., Nottingham, UK) in MH broth. Plates were inoculated at their centre, 2 mm beneath the agar surface, inverted and incubated for 48 hours in a Variable Atmosphere Incubator in a microaerobic atmosphere of O_2 5%, N_2 85%, CO_2 10% at 37°C.

2.2 Molecular Microbiology

2.2.1 Isolation of genomic DNA

DNA isolations were performed using a Wizard genomic DNA purification kit (Promega Ltd, Madison, USA). Briefly, 48 hour plate cultures of *C. jejuni* grown on Columbia blood agar plates were suspended in 300 µl of distilled water. 900 µl of Cell Lysis Solution were added and mixed by inversion five times. The cells were incubated at room temperature for 10 minutes. Samples were then spun for 20 seconds at 10,000 g and the supernatant decanted leaving approximately 20 µl for the resuspension of the cells. The cells were resuspended by brief vortexing and 300 µl of Nuclei Lysis Solution was then added, mixed by pipetting five times, and incubated at 37°C for an hour. 1.5 µl of RNase Solution was added, mixed by inversion 25 times and incubated at 37°C for 15 minutes. The suspension was allowed to cool to room temperature before 100 µl of Protein Precipitation Solution was added. After

vortexing for 20 seconds the suspension was spun at 10,000 x g for 10 minutes. The clear supernatant was then decanted into 300 µl of absolute isopropanol and mixed thoroughly. Following overnight incubation at – 20°C, the samples were spun for 10 minutes at 10,000 x g and the supernatant was decanted. 300 µl of 70% ethanol was added to the genomic DNA and mixed gently by inversion. The samples were then spun for a further minute at 10,000 x g and the supernatant was removed. The samples were allowed to air dry for ten minutes before 100 µl of DNA Rehydration Solution was added. Samples were incubated at 65°C for one hour and left at room temperature overnight. The isolated genomic DNA was quantified using a Genequant spectrophotometer and all DNA was stored at 4°C to minimise shearing due to freeze thawing.

2.2.2 Primer design

Oligonucleotide primers for the amplification of nucleotide sequences were designed using Primer3 (Rozen and Skaletsky, 1998; www-genome.wi.mit.edu/genome_software/other/primer3.html) or manually and synthesised by Sigma Genosys Ltd (Haverhill, UK). Primers used in this study are shown in Table 3.

Table 3. Primers used in this study

CDS	Name	Sequence	Length	GC%	Tm°C
Cj0059c;fliY	CjNCTC11168-0059c_f	TACAACCTTAGGAGCCCAAAAA	22	40.9	59.9
Cj0059c;fliY	CjNCTC11168-0059c_r	TTAAAATTTCAAGCGGATCGTT	22	31.8	59.9
Cj0105;atpA	CjNCTC11168-0105_f	AGCCAATTGATGCTAAAGGTGT	22	40.9	60.0
Cj0105;atpA	CjNCTC11168-0105_r	CTTGTGTTTCAATTATCGGCAA	22	36.4	60.0
Cj0145	CjNCTC11168-0145_f	AAAATCATCGACATGAATGCAG	22	36.4	60.0
Cj0145	CjNCTC11168-0145_r	TTTCATCAAATTTTTCCCATCC	22	31.8	60.0
Cj0266c	CjNCTC11168-0266c_f	TCAATCAATTCACCACTCAAGC	22	40.9	60.1
Cj0266c	CjNCTC11168-0266c_r	ATTACCCGGGAGTGTGTCTATG	22	50	60.1
Cj0818	CjNCTC11168-0818_f	CTTGCACTGTACTAAACGCAGC	22	50	60.1
Cj0818	CjNCTC11168-0818_r	AGCACCTACTACCCCCAATCTT	22	50	60.3
Cj0887c;flaD	CjNCTC11168-0887c_f	GGCGGTCAAAGTGCTTTATATC	22	45.5	59.9
Cj0887c;flaD	CjNCTC11168-0887c_r	AAAGTATCTTCGGGTTTGACGA	22	40.9	60.0
Cj1321	CjNCTC11168-1321_f	AAAATGTCATCATCATAGGAGCG	23	39.1	60.3
Cj1321	CjNCTC11168-1321_r	TCTAAGTTTACGCAAGGCAACA	22	40.9	59.9
Cj1322	CjNCTC11168-1322_f	GACTTTGGTTTAATGGGTAAGCA	23	39.1	59.4
Cj1322	CjNCTC11168-1322_r	TTCCGGCGTTAAAATTAGAAAA	22	31.8	59.9
Cj1323	CjNCTC11168-1323_f	AGAACGATTTACCCCATTGAAA	22	36.4	59.7
Cj1323	CjNCTC11168-1323_r	ATTTGCTAAAGCTCCTCGATTG	22	40.9	59.8
Cj1324	CjNCTC11168-1324_f	TGCCGTAAGTGGAGGTAAAGAT	22	45.5	60.0
Cj1324	CjNCTC11168-1324_r	TCTGCACACATTGTTCTATCCC	22	45.5	60.0
Cj1325	CjNCTC11168-1325_f	ACGGATTACTTTTTCCAGATGGT	23	39.1	60.1
Cj1325	CjNCTC11168-1325_r	TTTGCTTTGAAAATACGCTGAA	22	31.8	59.8
Cj1326	CjNCTC11168-1326_f	TACATTTCATCGATAAAGCCGA	22	36.4	59.5
Cj1326	CjNCTC11168-1326_r	AAATATAATGGTGTGCCGATCC	22	40.9	59.9
Cj1365	CjNCTC11168-1365_f	TATGGGGCAAAATTTTATGGAG	22	36.4	60.0
Cj1365	CjNCTC11168-1365_r	CTTCTATCCCAGGTGGATCTTG	22	50	59.9
Cj1376	CjNCTC11168-1376_f	TACTCGATGGAAATGCCTTTTT	22	36.4	59.9
Cj1376	CjNCTC11168-1376_r	TCGCTAAGTTTTTGAGCATTGA	22	36.4	60.0
Cj1545	CjNCTC11168-1545_f	CTGTGATCTATCAAATGCCAGC	22	45.5	59.7
Cj1545	CjNCTC11168-1545_r	ATGCAAATGCCAATACACCATA	22	36.4	60.1
Cj1678	CjNCTC11168-1678_f	GGAGAAATGATTTTATCTGGCG	22	40.9	59.9
Cj1678	CjNCTC11168-1678_r	TATGGTTGAGCCTTGTGAATTG	22	40.9	60.0
Cj1686c;topA	CjNCTC11168-1686c_f	GCAAATTTTGGCAAGGACTATC	22	40.9	59.9
Cj1686c;topA	CjNCTC11168-1686c_r	TTGCACTTTTAAATTTTGGCCT	22	31.8	60.0
16S rRNA	16s rRNA f	ATGGAGAGTTTGATC	15	40	43.0
16S rRNA	16s rRNA r	TGATCCAACCGCAGG	15	60	60.3
23S rRNA	23s rRNA f	ACTAAGAGCGAATGG	15	46.7	47.8

23S rRNA	23s rRNA r	ATTAGTACTGGTCAC	15	40	36.5

2.2.3 Polymerase Chain Reaction (PCR)

PCR reactions were performed in a reaction volume of 50 µl. They consisted of 0.2 mM of each deoxynucleotide triphosphate (dNTP) (Promega, Southampton, UK), 1 Unit of *Taq* DNA polymerase (Promega, Southampton, UK), 0.1 nM of both the forward and reverse primers and 1-100 ng of template DNA. All reactions were carried out using an Omn-E thermal cycler using cycling parameters; denaturation of template gDNA at 94°C for 1 minutes, 94°C for 15 seconds, annealing of primers to single stranded DNA at X°C for 15 seconds, extension of the product at 72°C for Y min (steps 2-4 were cycled 30 times), 72°C for 5 minutes (where X = temperature varied depending on the annealing temperature of the primers (approximately 5°C lower than the melting temperature) and Y = 1 minute per kb expected product).

PCR products were separated in a horizontal 0.7% agarose gel prepared with 1x TAE (40 mM Tris Acetate, 1 mM EDTA (pH8)) by means of an electric current. The gel contained 0.5% ethidium bromide (Promega, Madison, USA). Using 2 µl of loading buffer (0.25% bromophenol blue, 40% (w/v) sucrose) per sample, 5 µl of PCR product was loaded into the gel. A 1 kb ladder was used (Invitrogen, Paisley, UK) to size fragments. An electric charge of 100 V was applied across the gel for approximately 1 hour and PCR products were viewed using an ultraviolet light (Gene Genius Bio Imaging System).

2.3 Construction of *C. jejuni* gene specific composite DNA microarray

2.3.1 Collaboration with the Bacterial Microarray Group at St Georges Hospital Medical School, London (BµG@S)

Steps A-C in the construction of the gene specific *C. jejuni* microarray were all carried out by staff at (BµG@S) (The full protocols for the steps are available at BµG@S www.bugs.sghms.ac.uk).

A. Primer design

Oligonucleotide primers for the amplification of nucleotide sequences were designed using Primer3 (Rozen and Skaletsky, 1998).

B. PCR amplifications and purification

Amplification of gDNA from strain NCTC11168 and non-NCTC11168 strains and downstream liquid handling were performed using a RoboAmp 4200 (MWG Biotech, London, UK). Amplicons were evaluated for size and concentration on agarose gels and 5% of amplicons were sequenced to ensure the expected gene had been amplified. PCR products were purified by precipitation carried out in 96-well plates. Of the total PCR product 50 µl was precipitated and 50 µl was stored as an archive at –20 °C. 4.5 ml 3M sodium acetate and 0.5 ml 100% glycogen were pre-mixed. Precipitation was performed by adding 40 µl isopropanol and 5.0 µl of the sodium acetate and glycogen mixture to 50 µl of the PCR product. This was spun at 3,000 x g for 1 hour at 4°C and the supernatant was removed. A wash was performed using 150 µl of 70% ethanol followed by a 15 minute spin at 3,000 x g at 4°C. The supernatant was removed and the precipitated PCR products were air dried and resuspended in 15 µl 50% DMSO.

C. Printing the microarray

A BioRobotics MicroGridII arrayer (Genomic Solutions, Huntingdon, UK) with a 4x4 split pin configuration was used to print PCR products representing each of the 1654 NCTC11168 predicted CDSs, the 73 non-NCTC11168 genes and controls onto Poly-L-lysine coated slides made in-house at the BµG@S facility and GAPSII-coated slides (Corning Life Sciences, Koolhovenlaan, The Netherlands). Two replicate spots for each predicted CDS were printed on the array. These were not printed adjacently but spaced within a sub-grid to avoid any artefacts affecting both replicates.

D. PCR amplification of positive controls

Positive control reporter elements, 16S rRNA and 23S rRNA, were amplified through PCR reactions carried out manually using the reagents and conditions (Chapter 2.2.3). Negative controls included on the microarray were human B-actin and glycerol phosphate dehydrogenase (GAPDH). Negative controls such as spotting buffer 50% DMSO printed after the top concentration of the 16S or 23S rRNA positive controls were included to check for carry-over between samples when printing. In addition 5-amino-propargyl-2'-deoxycytidine 5- triphosphate coupled to Cy3 fluorescent and

Cy5 fluorescent dyes (herein referred to as Cy3 and Cy5) were printed on the microarrays for orientation of glass slides.

E. Post print processing of microarray slides

Post print processing of slides was performed to acetylate free amino group on the coated glass slides. A 2 L beaker containing 1 L of boiling distilled H_2O was placed on a hot plate whilst a slide drying bench was heated to 100 °C. Wash A and Wash B were prepared as follows;

Wash A; 1 x SSC and 0.05% SDS made up to 400 ml H_2O. Wash B; 0.06 x SSC made up to 400 ml H_2O.

In addition, 400 ml of 95% ethanol was placed in a trough at –20 °C and 6.0 g of succinic anhydride was added to 335 ml 1-methyl-2-pyrolidinone in a 500 ml bottle. Slides printed with the microarray were held array side down in steam rising from the beaker of boiling distilled H_2O for approximately 5 s. The formation of large beads of condensation at this point must be avoided to prevent the swelling of reporter elements. The slides were then placed immediately on the slide drying bench to dry for 10-20 s with the printed microarray facing upwards. Slides were then placed into a UV crosslinker in a plastic tray with the printed microarray facing upwards. Reporter elements spotted on the microarray were crosslinked using 200 mJ/cm^2 (Energy= 2000) of UV radiation. Succinic anhydride was fully dissolved in 1-methyl-2-pyrolidinone and the blocking solution was buffered with 15 ml of 1 M sodium borate (pH 8.0) forming a clear colourless solution. This was poured into a staining trough. Slides were placed in a staining rack and washed in wash A with vigorous agitation for 1 min. The slide rack was then drained and transferred to wash B. Washing with vigorous agitation was then carried out for a further minute before blotting the slide rack once more and transferring it to the blocking solution. Slides were vigorously agitated in blocking solution for 2 min after which time the slides were allowed to soak in blocking solution for a further 20 min.

The slides were then removed from the blocking solution, allowed to drain and then plunged into boiling water and agitated vigorously for 2 min after which the slides were then plunged into the ice cold 95% ethanol and washed for 1 min. Again the slide rack was allowed to drain and slides were dried by centrifugation for 5 min at 1500 rpm.

Slides were then stored in a dust free slide box. Poly L-lysine coated glass slides were used within three months or amino silane coated GAPSII slides within 1 year.

2.4 Comparative genomics

2.4.1 Competitive hybridisation

DNA microarray hybridisations were performed by comparing a chosen test strain against the sequenced (NCTC11168) strain as control. Competitive hybridisation reactions were carried out manually as follows;

A pre-hybridisation solution of 8.75 ml of 20x SSC, 0.25 ml of 20% SDS, 5.0 ml of BSA (100 mg/ ml) made up to 50 ml with distilled water was set up in a Coplin jar and incubated at 65°C during the labelling incubation period to equilibriate.

Control and test DNA were labelled as follows;

Reagent	**Control**	**Test**
Genomic DNA	1 µg	1 µg
Random Hexamers (3µg/µl) (Invitrogen, Paisley UK)	1 µl	1 µl
Distilled H_2O	to 41.5 µl	to 41.5 µl

Heated at 95°C for 5 minutes, snap cooled on ice and briefly centrifuged.

10x buffer (Invitrogen, Paisley UK)	5 µl	5 µl
dNTP's (Promega, Southampton, UK)	1 µl	1 µl
Cy-labelled dCTP (Amersham Biosciences, Amersham, UK)	1.5 µl (Cy3)	1.5 µl (Cy5)
Klenow fragment (10U/µl) (Amersham Biosciences, Amersham, UK)	1 µl	1 µl

Incubated at 37 °C for 90 minutes in the dark.

Microarray slides were incubated in the Pre-hybridisation solution for 20 minutes at 65°C, beginning just before the end of the labelling reactions.

Test and control reactions were combined and purified using the Qiagen MinElute PCR Purification kit, using a two step wash stage of 500 µl and 250 µl volumes of PE and eluting the labelled cDNA from the MinElute column with 21 µl H_2O. Columns retain 1 µl.

The pre-hybridised microarray slides were rinsed in distilled water for 1 minute, then in isopropanol for 1 minute followed by centrifugation briefly at 1,000 x g to dry the slides. Microarray slides were stored in a dust free, covered slide box.

The hybridisation solution was prepared as follows; 20 µl sample, 6.4 µl filter sterilised 20x SSC and 4.6 µl filter sterilised 2% SDS. This was heated at 95°C for 2 minutes and allowed to cool slowly to room temperature.

Hybridisation chambers (made to order, BµG@S) were taken, one per microarray slide, and 2 x 20 µl distilled water added to the corners. Slides were placed into the chambers and Lifter slips (Erie Scientific Company, Portsmouth, US) positioned over the microarray grid under which the hybridisation solution was applied, fully covering the microarray grid. The chamber was then sealed and incubated in a water bath at 65 °C overnight.

2.4.2 High and low stringency washes

Wash solutions were prepared as follows: High stringency wash A; 20x SSC 20 ml, 20% SDS 1 ml, H_2O to 400 ml incubated at 65°C overnight. Low stringency wash B; 20x SSC 2.4 ml, H_2O to 800 ml. Wash solutions were dispensed into three glass washing dishes. Microarray slides were removed from hybridisation chambers and coverslips were removed by rinsing in wash A. Microarray slides were immediately placed into a rack and rinsed with vigorous agitation for 5 minutes. Slides were transferred to a clean rack and rinsed with agitation in wash B. Slides were then washed for a further 2 minutes in fresh Buffer B. Microarray slides were dried by brief centrifugation at 1,000 x g and stored in a dust free, covered slide box for no longer than 2 hours.

2.5 Data Analysis

2.5.1 Scanning of microarray slides

Hybridised microarray slides were stored in a dust free slide box for up to 2 hours. Direct comparison of two samples (both test and reference conditions) was achieved through detection of differences in signal intensity produced by hybridisation of the two labels. This was achieved using a dual laser confocal GMS 418 Arrayscanner (Affymetrix, Santa Clara, US). During a preliminary scan of each microarray slide the gain was adjusted to optimise the intensity of the fluorescence in each channel without allowing reporter elements to 'bleach out', thus preventing the correct calculation of the ratio of fluorescence. The full scan was then performed at the appropriate gain and the image for each channel saved as a TIFF file.

2.5.2 Initial calculation of ratio of fluorescence

TIFF files generated through the scanning process for the test and control channels were overlaid and visualized in ImaGene v5.5 (BioDiscovery, El Segundo, US) software. Following the automated alignment of the test and control channels, a grid was fitted onto data points. The grid was created by entering detailed information about the layout of the microarray. Initially each reporter element must be matched back to its gene identity. To do this, the number of rows and columns in each grid must be entered as well as the minimum and maximum diameter of the reporter elements. This is calculated using a ruler tool available within ImaGene5.5. Subgrids within the metagrid may also be defined. Once defined a grid template may be saved for use on other microarrays from the same print run. However, updated templates may need to be defined if subsequent microarrays differ slightly from an older version. By overlaying the grid onto the reporter elements, gene ID files allow information about reporter elements to be tracked and saved with the quantified signal intensity values.

ImaGene5.5 software uses an automated process to identify reporter elements unsuitable for analysis. These flagged genes included empty spots, negative spots and poor spots. To enable ImaGene5.5 to recognise empty spots a sensitivity threshold was set up of 2.0. For each spot the following ratio is compared to the threshold:

$$R = \frac{\text{signal mean} - \text{background mean}}{\text{background standard deviation}}$$

If this ratio is lower than the threshold value the spot is flagged.

Any spot with a signal mean value lower than the background was flagged as negative. Low quality spots are detected based on a number of criteria:

1. Genes were flagged for background contamination whereby the background of each individual spot is compared to the local background level distribution over the image through a t-test. The resulting P-value is subtracted from 1 and is called 'confidence in contamination presence'. This confidence threshold was set at 1.0 and genes were flagged if the contamination confidence level went over this threshold.
2. Genes were also flagged for a high-ignored percentage. Since some pixels may be assigned to neither signal nor background in the presence of high local contamination it is important that such contamination is excluded from the measurements. Such exclusions are designated 'ignored regions'. If the ignored region was higher than the pre-set threshold of 25% the spot was flagged for low quality.
3. Low quality genes were flagged for a high open perimeter percentage. This is designed to eliminate spots with a specific type of deformation. Imagene computes the percentage of signal perimeter that touches the border of a rectangle. Such a deformation may appear if the spot is significantly shifted from its expected position or if its shape has an abnormal form. Whenever the percentage crossed the pre set threshold of 25% the spot was flagged for low quality.
4. Genes were flagged for having abnormal shape regularity. An algorithm computed a shape regularity measure that characterised the closeness of a spot border to a circle. The first step of this algorithm inscribed the signal area of a spot into a circle. The number of non-signal pixels that fall within this circle was computed and divided by the circle area. This ratio was subtracted from 1 and was called the 'shape regularity'. This variable ranges from 0 (highly non circular) to 1 (a perfect circle). If this ratio fell below the preset threshold of 0.65 the spot was flagged for low quality.
5. Finally, genes were flagged for significant offset from their expected position. The expected position of a spot in a grid is computed by fitting least square lines to the grid's columns and rows. Whenever the offset became larger than a preset threshold of 60% of the distance between grid and nodes, the spot was flagged for low quality.

The ratio of fluorescence for present and marginal data points was automatically calculated. The extracted data was captured, quantified and stored automatically in a

tab-delimited or XML file that may be exported into data analysis software tools including GeneSpring6.1 (Silicon Genetics, Redwood City, US) and Genomotyping Analysis Charlie Kim (GACK) (http://falkow.stanford.edu/whatwedo/software/software.html).

2.5.3 Normalisation of DNA microarray hybridization data

Data output from ImaGene5.5 is recognised by GeneSpring6.1 and replicate hybridisations were loaded into single experiments together. Raw microarray data input from ImaGene5.5 was standardised by normalisation allowing biological variations to be differentiated from variations due to the measurement process. Each gene's measured intensity was divided by its control channel value in each sample; if the control channel was below 10 then this value was used instead. If the control channel and the signal channel were both below 10 then no data was reported. Variations in intensity caused by inconsistencies in sample preparation or washing were also accounted for by normalisations thus each measurement was divided by the 50.0th percentile of all measurements in that sample. Lowess was used as an alternative DNA microarray normalisation. Using this method, a Lowess curve was fitted to the log-intensity versus log-ratio plot. 20% of the data was used to calculate the Lowess fit at each point. This curve was used to adjust the control value for each measurement. If the control value was lower than 10 then this value was used instead.

2.5.4 Comparative genomics analysis

Within GeneSpring6.1 software, hybridsation data was filtered to remove data points unsuitable for further analysis. Empty, negative or poor reporter elements flagged using ImaGene5.5 software (Chapter 2.5.2) were removed, leaving only data points that fluoresced at a 1:1 ratio or above (present) and those between 0.5 and 1 (marginal). A constant cut-off of 0.5 was selected to identify absent or divergent genes in the test strain. Genes identified as absent or divergent were filtered on confidence using a t-test cut off of 0.05.

GACK software was also used to analyse DNA microarray hybridisation data. Prior to analysis with GACK, microarray hybridisation data was normalised using Lowess normalisation. Normalised comparative genomics data was then entered into a data

matrix for input into GACK. This matrix contained the genomic content of each strain represented in binary an example of which is shown in Table 4.

Table 4. A sample of GACK comparative genomics input data matrix

UNIQID	NAME	GWEIGHT	1099E	11848	11856	11919	11818	11973	11974	12196	12241	12450
EWEIGHT		1	1	1	1	1	1	1	1	1	1	1
1	Cj11168-0001	(1A1)	1		1	1	1	1	1	1		1
2	Cj11168-0002	(1B1)	1	1	1	1	1	1		1		1
3	Cj11168-0003	(1C1)	1	1	1	1	1	1	1	1	1	1
4	Cj11168-0004c	(1D1)	1	1	1	1	1	1	1	1	1	1
5	Cj11168-0005c	(1E1)	1	1	1	1	1	1	1	1	1	1
6	Cj11168-0006	(1F1)	1	1	1	1	1	1	1	1	1	1
7	Cj11168-0007	(1G1)	1	1	1	1	1	1	1	1	1	1
8	Cj11168-0008	(1H1)	1		1	1	1	1	1	0	1	
9	Cj11168-0009	(1A2)	1	1	1	1	1	1	1	1	1	1
10	Cj11168-0010c	(1B2)	1	1	1	1	1	1	1	1	1	1
11	Cj11168-0011c	(1C2)	1		1	1	1	1		1		1
12	Cj11168-0012c	(1D2)	1	1	1	1	1	1	1	1	1	1
13	Cj11168-0013	(1E2)	1	1	1	1	1	1	1	1	1	1
14	Cj11168-0014c	(1F2)	1	1	1	1	0	1	1	1	1	1
15	Cj11168-0015c	(1G2)	1	1	1	1	1	1	1	1	1	1
16	Cj11168-0016	(1H2)	1	1	1	1	1	1	1	1	1	1
17	Cj11168-0017c	(1A3)	1	1	1	1	1	1	1	1	1	1
18	Cj11168-0018c	(1B3)	1	1	1	1	1	1	1	1	1	1
19	Cj11168-0019c	(1C3)	1	1	1	1	1	1	1	1	1	1
20	Cj11168-0020c	(1D3)	1	1	1	1	1	1	1	1	1	1
21	Cj11168-0021c	(1E3)	1		1	1	1	1	1	1		1
22	Cj11168-0022c	(1F3)	1	1	1	1	1	1	1	0	1	1
23	Cj11168-0023	(1G3)	1	1	1	1	1	1	1	1	1	1
24	Cj11168-0024	(1H3)	1	1	1	1	1	1	1	1		1
25	Cj11168-0025c	(1A4)	1	1	1	1	1	1	1	1	1	1
26	Cj11168-0026c	(1B4)	1	1	1	1	1	1	1	1	1	1
27	Cj11168-0027	(1C4)	1	1	1	1	1	1	1	1	1	1
28	Cj11168-0028	(1D4)	1	1	1	1	1	1	1	1	1	1
29	Cj11168-0029	(1E4)	1	1	1	1	1	1	1	1	1	1
30	Cj11168-0030	(1F4)	1		1	1	1	1	1	1		1
31	Cj11168-0031	(1G4)	1	1	1	1	0	1	1	1	1	1
32	Cj11168-0032	(1H4)	1	1	1	1	0	1	1	1	0	0
33	Cj11168-0033	(1A5)	1	0	0	0	0	0	0	0	0	0

Using GACK a dynamic cut-off value for the identification of absent or divergent genes was calculated. A normal probability density function was fitted to the major peak and the ratio values at half the peaks maximum height on both sides. The estimated probability or presence (EPP) of a gene was calculated as follows;

EPP= 100 x (expected normal value / observed value)

The mapped normal curve value (which is expected for a distribution in which all spots are present on the array) was divided by the observed data distribution value. Genes were then categorised as absent with an EPP of 0%, present with an EPP of 100% and into a transition region if the EPP was anything other than 0% or 100%.

2.6 Comparative phylogenomics

2.6.1 Data transformation

Using NCTC11168 reporter elements only raw microarray data was transformed into binary using a Perl script (provided by Adam Witney, BμG@S). Using data derived from GeneSpring genes falling below a 0.5 ratio were designated 0 and genes above 0.5 were designated 1. With data derived from GACK, genes that were present or marginal were assigned a '1' and those that were absent were assigned '0'. In both data sets the genes that were flagged using ImaGene5.5 software were denoted 'F'. Strains in which over 5% of the total genes were flagged were manually removed from both data sets as were genes flagged in over 10% of all strains. The remaining flagged genes were denoted with a question mark. The data set was then cross checked back against the raw microarray data, transposed so genes were on the X axis and transformed into Nexus format (Maddison *et al.*, 1997) (Figure 8). This is a modular file format that is easily processed by many bioinformatics programs such as Mr Bayes3.0 (Ronquist and Huelsenbeck, 2003) and MacClade4.0 (Maddison, 2001).

```
#NEXUS

Begin data;

      Dimensions ntax=102 nchar=1627;
      Format datatype=standard missing=? matchchar=.;
      Matrix
1099_E
1111111111111111?11111111110111111111111111111111111111111111111111111111111111111
11111111111111111111111?11111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
1111101?11111111111111111111111111111111111111111111111111?111111111111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
10000000000000000001111111111111111111111111111111111111111111111111111010011111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
1111111111111111111111111111111111111111111111111111111111111111111
11818
11111111111111111111111111011111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111101111111111111111111111111111111111111111111
11111111111111111111111111111111111111111110111111111111111110111111111111111111
11111111111111111111101111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111110011111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
11111110111111111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111110111111111111111111111111111111111111111111111
11111111111111111111111111111111111101111111111111001111111111111111111111111111
11111111111111111111111111111111111111111111111111101111111111111110000000111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111111111011000000000001001
11111111111111111111011111111111111111111111111111111111111111111111111111111111
11011111111111111111111111111111111111111111111111111111111111111111111111110111
11111111011111101011101111111111111111111111111111111111111111111111111110011111
11111111111111111100001000000000000000011100111111111110111111111111111111111111
11111111111111111111111111111110111111111111111111111110000000100011111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
1111111111111111111111111111111111111111111111111111111110001111111
11848
11111111011111111111111111011111111111111111111111111111111111111111111111111111
11111111111111111111111111111111110011111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
11111111100011111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111110111111111111111111111110000000111011111111111111
111111111111111111111111111111111111111111111111111111111111111111111111?1111111
1111111111111111000111111111111111111111111111111111111111110000111111
      ;
End;

begin mrbayes;
      log start filename=Olivia.log replace;
      lset coding=noabsence rate=gamma ngammacat=16 nst=6;
            mcmc ngen=2000000 nchains=4 temp=0.8  printfreq=40 samplefreq=40
Startingtree=Random burnin=100000 savebrlens=yes;
end;
quit;
```

Figure 8. A sample of comparative genomic data in Nexus format with input code for MrBayes3.0 software. Input format for the phylogenetic program Mr Bayes. All microarray hybridisation ratios for strains 1099_E, 11818 and 11848 were transformed into a modular binary file format (Nexus). In this format "1" represents a CDS that is present in the test strain, "0" a CDS that is absent and "?" a CDS that was flagged. The code for running the phylogenetic program Mr Bayes is also shown.

2.6.1 Inferral of phylogeny using Bayesian based algorithms

MrBayes v3.0B4 was used to generate Bayesian trees with a gamma distribution using 16 discrete rates categories to model hybridisation rate heterogeneity between loci. Four incrementally heated Metropolis coupling Markov Chain Monte Carlos were run for 1000000 generations with a 10,0000 generation "burn-in". Phylogenies were sampled every 40 generations.

Statistical support was calculated within PAUP* (Sinauer Associates, Sunderland, Massachusetts) by Michael Gaunt (LSHTM) using Bayesian strict and 95% majority rule consensus phylogeny based on 25,000 phylogenies of *C. jejuni* DNA-DNA microarray binary data.

2.6.2 Graphical representation of phylogenetic tree topology

The topology of phylogenetic trees inferred using Bayesian based algorithms were viewed using TreeView software (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html).

2.6.3 Identification of key CDSs contributing to clade formation

MacClade4: Analysis of Phylogeny and Character Evolution was used for CDS analysis to identify genomic regions contributing to different clades within phylogenetic trees. Data was entered into this programme in Nexus format and the corresponding phylogenetic tree was also loaded. MacClade4 allows the user to visualise the distribution of each CDS in the genome as it occurs throughout the phylogenetic tree as designated by the microarray hybridisations. Strains in which a CDS of interest is absent or divergent are coloured blue and strains in which the same CDS is present are coloured yellow (Chapter 6.2.2). Through visualising the distribution of each CDS in the genome, specific CDSs for a particular clade may be identified.

3.0 Results

Construction of a gene specific composite *Campylobacter jejuni* DNA microarray

3.1 Introduction

3.1.1 Aims

The availability of the genome sequence for the *C. jejuni* strain NCTC11168 (Parkhill *et al.*, 2000) has allowed the construction of DNA microarrays with reporter elements representing each one of the 1654 predicted protein coding sequences (CDSs) for this strain. This chapter describes the development and construction of a gene specific *C. jejuni* DNA microarray containing reporter elements representing NCTC11168 CDSs and novel CDSs not found in the sequenced genome. This type of DNA microarray is more representative of the species and is described as composite.

3.1.2 Historical development of *C. jejuni* DNA microarrays

The first whole genome *C. jejuni* DNA microarray was produced by BμG@S (http://bugs.sghms.ac.uk/index.php) in January 2000. This early *C. jejuni* array was an amplified clone array from an optimal set of plasmid clones from the NCTC11168 genome sequencing project. In 2000, most of the cost in producing PCR product-based DNA microarrays was due to the synthesis of oligonucleotide primer pairs required for the PCR amplification of gene fragments that make up the reporter elements present on the array. This cost was dramatically reduced by utilising clone libraries as the template for amplifying the PCR products. Using this method the same two vector primers are utilised for the amplification of every clone. For example, ordered pUC clone libraries that are a by-product of most bacterial genome sequencing projects can be used by selecting an optimum clone to represent every gene in the genome (Figure 9).

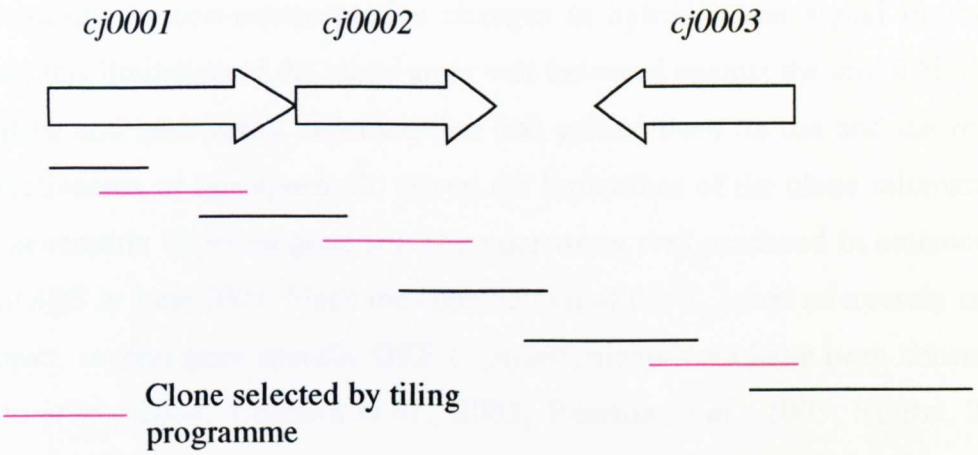

Figure 9. Selection of clones by tiling programme (Dorrell *et al.*, 2001) for DNA microarray. Predicted protein CDSs for NCTC11168 *cj0001*, *cj0002* and *cj0003* are diagrammatically represented with arrows. Overlapping clones from the genome sequencing project for each CDS are coloured black. Clones selected as optimal (coloured pink) using a tiling programme were selected to represent a CDS on the clone array.

The ordered *C. jejuni* NCTC11168 plasmid library generated by the Sanger Institute was used as the template and a single pair of primers based on the pUC19 vector sequence were used to amplify gene fragments representative of the initial 1731 predicted CDSs from strain NCTC11168, to produce a low-cost representative whole genome microarray. The predicted CDSs were based on the preliminary annotation of the NCTC11168 genome sequence (Julian Parkhill, unpublished data), however following completion of this annotation, the number of predicted genes was reduced to 1654 (Parkhill *et al.*, 2000). Therefore the resulting 77 "non-genes" were each flagged in the gene lists used by the analysis software as a "not annotated gene" (Dorrell *et al.*, 2001).

A drawback of this type of microarray is that many of the optimum clones available were not gene-specific and contained adjacent genes or gene fragments. Only 34.5% of the selected clones contained a single gene fragment, with 35.4% containing an adjacent gene fragment and the remaining 30.1% containing more than one gene

fragment adjacent to the selected gene. The presence of these adjacent gene fragments could result in cross-hybridisation when the gene sequence present in the hybridisation sample is present in more than one reporter element present on the array, resulting in non-representative changes in hybridisation signal intensities. However, this limitation of the clone array was balanced against the initial biological information and microarray expertise that was gained from its use and the relative cost effectiveness of this approach. Given the limitations of the clone microarray, a second generation *C. jejuni* gene specific microarray was produced in collaboration with BμG@S in June 2002. Since the construction of the *C. jejuni* microarray used in this project, several gene specific ORF *C. jejuni* microarrays have been constructed (Carrillo *et al.*, 2004; Leonard *et al.*, 2003; Pearson *et al.*, 2003; Stintzi, 2003). However, none of these other microarrays represent non-NCTC11168 CDSs as well as NCTC11168 CDSs.

3.2 Results – non-NCTC11168 CDSs

3.2.1 Identification of non-NCTC11168 CDSs

To make the microarray more representative of the *C. jejuni* species, non-NCTC11168 CDSs were added to the microarray (Tables 5 and 6). These non-NCTC11168 CDSs were obtained by contacting the *Campylobacter* research community and by searching the literature. The percentage identity of these non-NCTC11168 CDSs to the NCTC11168 sequenced strain was calculated using a Basic Local Alignment Search Tool (BLAST) (http://www.sanger.ac.uk/cgi-bin/blast/submitblast/c_jejuni). BLASTn was used to compare the nucleotide sequence against that of NCTC11168. BLASTx was used to compare the six-frame conceptual translation products of the non-NCTC11168 query sequence against a protein sequence database for all sequenced genomes. Sequences with less than 80% identity in the BLASTn search were selected for inclusion on the composite array. MIAME compliant information for all reporter elements included on the gene specific, composite *C. jejuni* microarray are available in the ArrayExpress public repository at;

http://www.ebi.ac.uk/arrayexpress/query/result?queryFor=PhysicalArrayDesign&aAccession=A-BUGS-8

Table 5. Non-NCTC11168 CDSs included on the composite gene specific *C. jejuni* microarray

CDS ID	Accession No.	BLASTn to NCTC11168	BLASTx to NCTC11168	Predicted protein similarity
11351-01	AF334961	4.10E-28	0.19 38/124 (30%)	LOS region sequences; ClpA
11828-01	AF343914	1.40E-07	1.4e-36 62/271 (22%)	LOS region sequences sugar transferase
11828-02	AF343914	2.50E-06 411/705 (58%)	0.77 20/60 (33%)	LOS region sequences ABC transport system
11828-03	AF343914	1.20E-28 743/1258 (59%)	2.0e-9 28/81 (34%)	LOS region sequences CysE
11828-04	AF343914	9.60E-08 246/410 (60%)	0.75 37/114 (32%)	LOS region probable membrane protein
11828-05	AF343914	28/81 (34%)	6.3e-18 13/33 (39%)	LOS region sequences galactosyl transferase
11828-06	AF343914	1.20E-11 460/790 (58%)	3.1e-14 40/111 (36%)	LOS region sequences
11828-07	AF343914	4.70E-28 616/1035 (59%)	3.9e-55 82/251 (32%)	LOS region sequences aminotransferase
11828-08	AF343914	5.40E-17 617/1089 (56%)	0.9997 23/61 (37%)	LOS region sequences Cj0505c
11828-10	AF343914	4.00E-10 424/725 (58%)	no matching pairs	LOS region sequences Cj1137
11828-11	AF343914	3.30E-47 574/904 (63%)	4.2e-44 33/95 (34%)	LOS region sequences transport protein
11828-12	AF343914	3.50E-17 559/955 (58%)	0.775 20/59 (33%)	LOS region sequences Cj1333
2523/90-01	AF334762	4.10E-28 69/106 (56%)	0.021 20/66 (30%)	LOS region; possible
43431-01	AF411225	4.1e-20 395/650 (60%)	2.4e-34 87/241 (36%)	LOS region probable two domain glycosyltransferase
43431-02	AF411225	0.0026 335/588 (56%)	9.7e-14 60/233 (25%)	LOS region Cj1423c possible sugar phosp nucleotidytransferase
43431-03	AF411225	7.0e-08 336/580 (57%)	1.0e-17 59/160 (36%)	LOS region Cj1319 probable nucleotide

				sugar dehydratase
43432-01	AF215659	1.9e-06 440/779 (56%)	0.015 33/131 (25%)	LOS region sequences
43438-01	AF400048	4.8e-59 580/869 (66%)	1.0e-72 147/287 (65%)	LOS region sequences
43438-02	AF400048	8.3e-06 285/478 (59%)	0.0012 32/131 (24%)	LOS region sequences CysE
43449-01	AF401529	1.0e-56 578/869 (66%)	7.1e-73 144/287 (50%)	LOS region sequences
43449-02	AF401529	1.10E-05 438/779 (56%)	0.0010 22/55 (40%)	LOS region sequences CysE
4344c-01	AF167344	2.4e-05 461/825 (55%)	0.0010 22/55 (40%)	LOS region sequences CysE
43456-01	AF401528	1.0e-56 574/865 (66%)	1.5e-79 153/287 (53%)	LOS region Cj1140 unknown function
43456-02	AF401528	1.3e-0.5 462/826 (55%)	0.001 22/55 (40%)	LOS region sequences CysE
480-01		8.1e-0.6 246/414 (59%)	3.60E-12 51/171 (29%)	Kfid
480-02		0.00097 98/152 (64%)	4.60E-08 29/85 (34%)	Kfid
480-03		9.00E-41 412/728 (57%)	1.30E-70 58/177 (32%)	GalE
480-04		1.60E-36 505/887 (56%)	3.20E-63 40/137 (29%)	Glf
480-05		2.00E-14 174/299 (65%)	1.30E-08 26/59 (44%)	Sugar transferase
480-06		0.0058 200/340 (58%)	3.30E-17 78/238 (32%)	DMSO reductase
81176-01		0.00057 181/313 (57%)	0.0057 181/313 (57%)	VirB8 protein from *B. suis*
81176-02		0.055 425/819 (51%)	0.9997 36/139 (25%)	LvhB9 of *L. pneumophila*
81176-03		7.60E-06 172/292 (58%)	No high scoring segment pairs	TrbI of plasmid RK2
81176-04		0.0046 136/226 (60%)	0.0013 57/184 (30%)	Type IV secretion protein
8F169-01	AF334378	1.40E-07 440/759 (57%)	0.11 22/72 (30%)	LOS region Cj1042c transcriptional regulatory protein
P19-01		5.6e-83 610/873 (69%)	1.9e-100 131/309 (42%)	Capsule region

P19-02		9.3e-112 1297/2047 (63%)	7.3e-199 97/174 (55%)	Capsule region
X-1		1.2e-84 675/1019 (66%)	3.5e-108 169/344 (49%)	Capsule region
X-2		1.3e-82 707/1042 (67%)	2.3e-100 155/344 (45%)	Capsule region
X-3		6.3e-10 863/1524 (56%)	2.5e-10 37/110 (33%)	Capsule region
X-4		9.8e-14 648/1148 (56%)	1.7e-14 42/99 (42%)	Capsule region
X-5		2.1e-42 2333/4167 (55%)	7.4e-32 124/428 (28%)	Capsule region
X-6		4.1e-12 475/815 (58%)	no high scoring pairs	Capsule region

Of the 69 non-NCTC11168 CDSs included on the array, 23 were identified from strain 81116 using subtractive hybridisation (Ahmed *et al.*, 2002). Similarity searches had already been carried out for these CDSs and the data published thus BLAST searches were not repeated and these sequences were automatically included on the microarray.

Table 6. Additional 23 non-NCTC11168 sequences included on the microarray identified using subtractive hybridisation (Ahmed *et al.*, 2002)

Insert	Size (bp)	Predicted protein similarity/species	Predicted amino acid identity (%)	Blast score	Accession no.
30	259	No similarity*	–	–	AJ315005
31	232	Arsenite export protein (*arsB*)*; *Sinorhizobium* sp. As4	64	2×10^{21}	AJ315006
46	384	Putative oxidoreductase chain; *Campylobacter jejuni*	61	1×10^{38}	AJ315007
50	124	No similarity*	–	–	AJ315008
60	170	Type I R–M protein (S subunit); *Haemophilus influenzae* Rd	51	2×10^{10}	AJ315009
65	147	No similarity*	–	–	AJ315010
68	443	Type I R–M protein (M subunit); *Salmonella enterica*	32	1×10^{6}	AJ315011
117	280	No similarity	–	–	AJ315012
121	321	No similarity	–	–	AJ315013
135	644	Sty SPI type I R–M (M subunit)*; *Salmonella enterica*	32	2×10^{6}	AJ315014
136	308	Putative type I R–M protein (S subunit); *Streptococcus thermophilus*	39	2×10^{10}	AJ315015
141	301	Putative glycosyltransferase; *C. jejuni*	40	1×10^{11}	AJ131360.1 (CAC01393)
149	208	Hypothetical protein (Cj1337); *C. jejuni*	47	2×10^{8}	AJ315016
183	195	Hypothetical protein (Rv0235c)*; *Mycobacterium tuberculosis*	35	0·033	AJ315017
186	196	Arsenate reductase (*arsC*); *Aquifex aeolicus*	35	8×10^{6}	AJ315018
188	163	Cytochrome *c* oxidase III*; *Trypanosoma brucei*	38	3·4	AJ315019
202	309	Type II R–M protein; *Helicobacter pylori* J99	68	5×10^{32}	AJ315020
205	313	dTDP-glucose 4,6-dehydratase (*rmlB*); *Legionella pneumophila*	51	7×10^{25}	AJ131360.1 (CAC01395)
209	248	Hypothetical protein (Cj1333); *C. jejuni*	48	2×10^{11}	AJ315021
212	194	Probable membrane protein (Cj1560)*; *C. jejuni*	56	5×10^{11}	AJ315022
236	369	γ-Glutamyl transpeptidase; *Helicobacter pylori* 26695	72	5×10^{40}	AJ315023
243	150	No similarity*	–	–	AJ315024
246	297	Abortive phage-resistance protein; *Lactococcus lactis*	34	0·033	AJ315025

* Predicted membrane-spanning domain.

3.2.2 Acquisition of non-NCTC11168 strains

The *C. jejuni* strains from which non-NCTC11168 CDSs were identified were requested from the research groups working with them (Table 7). These strains were cultured and gDNA was isolated (Chapter 2.1.1 and 2.2.1).

Table 7. *C. jejuni* strains from which novel CDSs were identified and the research groups associated with them

***C. jejuni* strain**	**Donated by;**
NCTC11351	Julian Ketley, University of Leicester, UK.
NCTC11828	Julian Ketley, University of Leicester, UK.
2523/90	Julian Ketley, University of Leicester, UK.
ATCC43431	Julian Ketley, University of Leicester, UK.
ATCC43432	Julian Ketley, University of Leicester, UK.
ATCC43438	Julian Ketley, University of Leicester, UK.
ATCC43449	Julian Ketley, University of Leicester, UK.
ATCC4344c	Julian Ketley, University of Leicester, UK.
ATCC43456	Julian Ketley, University of Leicester, UK.
480	Neal Golden, New England Medical Centre, USA
81116	If Ahmed, Veterinary Laboratory Association, UK
81-176	Erin Gaynor, Stanford University, USA
8F 169	Julian Ketley, Birmingham University, UK.
P19	Andrey Karlyshev, London School of Hygiene and Tropical Medicine, UK
X	Andrey Karlyshev, London School of Hygiene and Tropical Medicine, UK

3.3 Results - Construction of the microarray

3.3.1 Design of gene specific primers

An optimised pair of oligonucleotide primers were designed for the 69 additional non-NCTC11168 CDSs and the 1654 NCTC11168 predicted CDSs using Primer3 software (Chapter 2.3.1). The PCR products for each CDS for this second generation *C. jejuni* microarray were designed to have minimal cross-hybridisation to other CDSs in the genome. Primers ranging from 22-25 bases in size were designed to have closely matched melting temperature to aid high-throughput PCR amplification of products ranging in size from 70 - 800 bases with an average length of 388 bp.

3.3.2 Controls and orientation of the microarray slides

PCR products for the 16S and 23S rRNA genes were included on the microarray as positive controls. These genes were amplified manually (Chapter 2.2.3). PCR products of the expected size (1.55 kb (16s) and 3.0 kb (23s)) were observed (Figure 10).

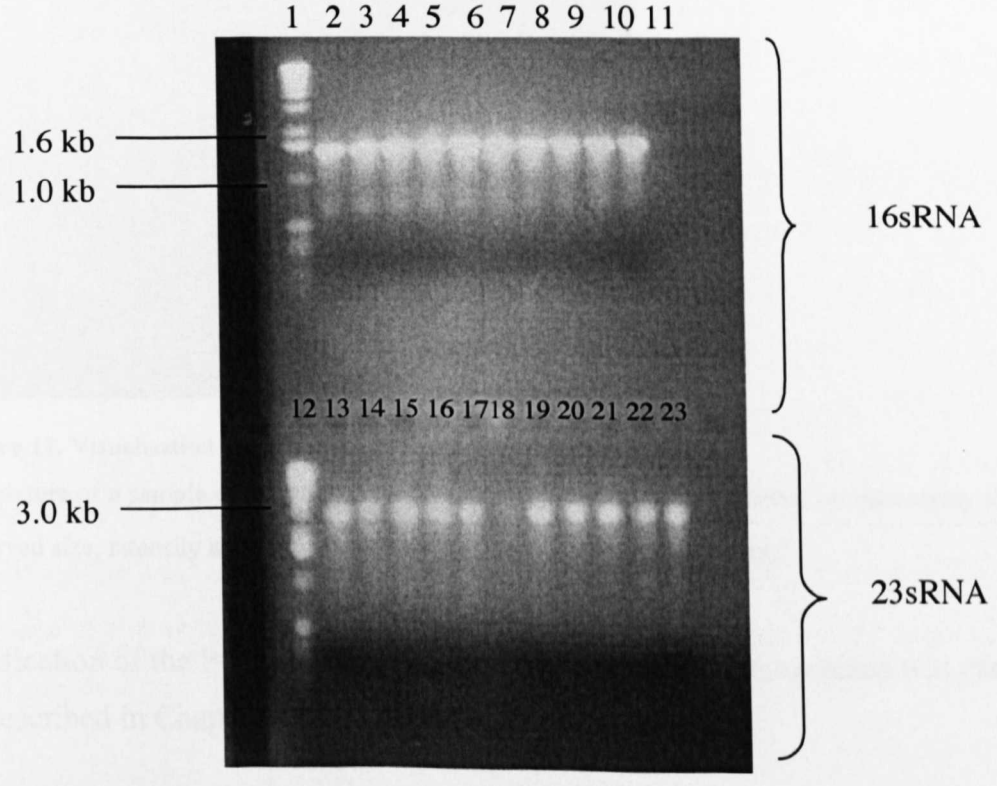

Figure 10. Visualisation of 16S and 23S rRNA PCR products

Lanes 1 and 12 contain the molecular weight marker. Lanes 2-11 contain PCR products of 16S rRNA. Lanes 13-23 contain PCR products for 23S rRNA.

Negative controls and 'landing lights' for orientation of glass slides during scanning were also printed on the arrays (Chapter 2.3.1.D).

3.3.3 Amplification and verification of NCTC11168 coding sequences and non NCTC11168 sequences

Amplification of NCTC11168 CDSs and the novel sequences selected for inclusion on the microarray was carried out using a liquid handling robot RoboAmp 4200 (Chapter 2.3.1). Of the CDSs selected for inclusion on the array, five CDSs were unsuccessfully amplified despite re-ordering the primers used for PCR. These CDSs included one from the sequenced strain (*cj11168-0776c*) and four of the additional CDSs (*cj43456-03, cj466-01, cj466-02, cj81116-09*). PCR products were run on agarose gels to ensure a single product of the expected size and of suitable intensity was observed (Figure 11). In addition 5% of the PCR products were sequenced to verify the expected target CDS had been amplified.

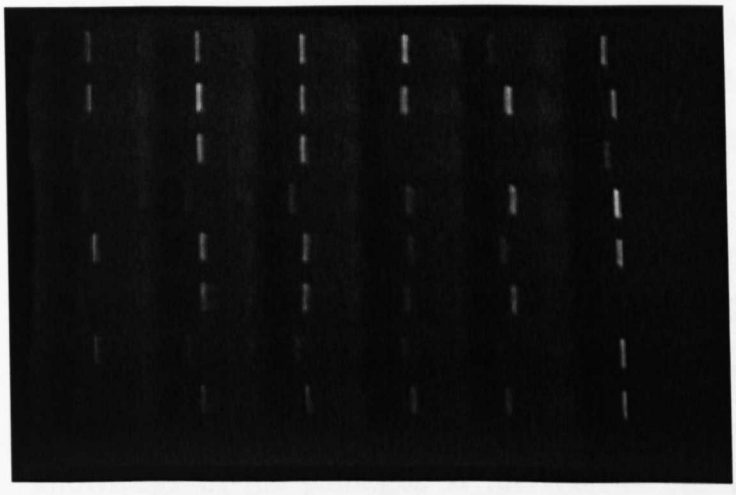

Figure 11. Visualisation of PCR product microarray reporter elements
Gel picture of a sample of *C. jejuni* gene specific PCR products to be printed on microarray slides. The observed size, intensity and presence of single product only was confirmed.

Purification of the PCR products and printing of the microarray slides was carried out as described in Chapter 2.3.1.

3.4 Discussion

Even with gene-specific microarrays, for some microbial species potential problems will still occur with cross hybridisation between reporter elements representing paralogous genes, meaning that genetic rearrangements, insertions, inversions and duplications are difficult to detect. For example, differentiating between separate genes in the *C. jejuni* NCTC11168 *cj0617* or *cj1318* paralogous gene families using such whole gene PCR products as reporter elements would be difficult due to cross hybridisation. This makes it difficult to distinguish whether a particular gene in a paralogous gene family is really absent in a test strain. Using a more selective primer design strategy (Hinds *et al*, 2002), the PCR products for each gene for this second generation *C. jejuni* microarray were designed to have minimal cross-hybridisation to other genes in the genome. This gene specific microarray represented 1654 annotated NCTC11168 predicted CDSs as well as 69 additional CDSs, many of which have been donated from the *Campylobacter* research community. Such non-NCTC11168 CDSs have been identified from the capsule and lipooligosaccharide biosynthesis regions and a putative virulence plasmid, *pVir* from strain 81-176. In addition 23 non-NCTC11168 CDSs were also identified from strain 81116 using subtractive hybridisation. Non-NCTC11168 CDSs were also identified from the literature, such as AF411225, AF215659, AF400048, AF167344, AF401529, AF401528 and AJ131360.

However, since construction of the array, many more non-NCTC11168 CDSs have been identified, including several from strain RM1221 that has been sequenced at The Institute for Genome Research (TIGR). The genome of this *C. jejuni* isolate is larger than that of strain NCTC11168. Thus, the current *C. jejuni* microarray reflects our knowledge of the pathogen at the time of construction but will be continually modified and updated with the addition of novel RM1221 CDSs, newly identified CDSs, as well as the CDSs absent due to PCR failure. All microarray analyses are limited by the genetic information on the array, however, this second generation array moves away from a strain-specific array to an array more representative of the *C. jejuni* species.

4.0 Comparison of DNA microarray data analysis techniques for comparative genomic hybridisations

4.1 Introduction

4.1.1 Aims

DNA microarrays facilitate genome composition analysis based on signal intensity ratios of thousands of reporter elements. Nucleotide variation between test and control strains leads to fluctuations in signal intensity that can be used to identify absent or divergent genes in non-sequenced strains. Strains exhibiting specific phenotypes may be genomotyped by categorising test strain genes as present or absent / divergent based on signal cut-off values. This chapter describes the comparison of two DNA microarray data analysis methods that use different criteria to select the signal cut-off value and therefore designate genes as absent or divergent. A collection of 103 *C. jejuni* strains (Appendix 1) was selected to carry out a comprehensive comparison of fixed and dynamic cut-off values (Chapter 2.5.4) to identify absent or divergent genes from the same data set. Furthermore, by using fixed and dynamic cut-off values to compare the same DNA microarray hybridisation data from two sequenced *C. jejuni* genomes, RM1221 and NCTC11168, the sensitivity of each method could be quantified.

4.1.2 Methods available to analyse DNA microarray data

In order to process, analyse and determine the biological significance of genomic comparisons, the vast data accumulated from each hybridisation experiment must first be normalised allowing biological variations to be differentiated from variations due to the experimental procedures and measurement process. To differentiate genes that are present from genes that are absent or divergent in a test strain, a signal intensity cut-off value must be calculated. There are two main ways used to calculate this cut-off value. First, by selecting a constant cut-off value that applies across all hybridisation experiments and second, by calculating a dynamic cut-off value calculated for each hybridisation experiment. Using a constant cut-off method, a pre-determined ratio of fluorescence between the test and control strain is used to designate presence or absence / divergence for each reporter element on the array. Using the software programme GeneSpring6.1, a constant cut-off value of 0.5 was selected to determine genes that were present or absent/ divergent in each microarray

experiment. This value was selected based on initial hybridisation experiments carried out at BμG@S. Hybridisation data indicates that the signal intensity ratio is 1 in the test and control channel for the majority of genes. Thus the majority of genes are present in both test and control strains. When the test channel fluoresces at half the level or less of the control channel (<0.5) the gene is deemed to be absent or divergent in the test strain (Figure 12).

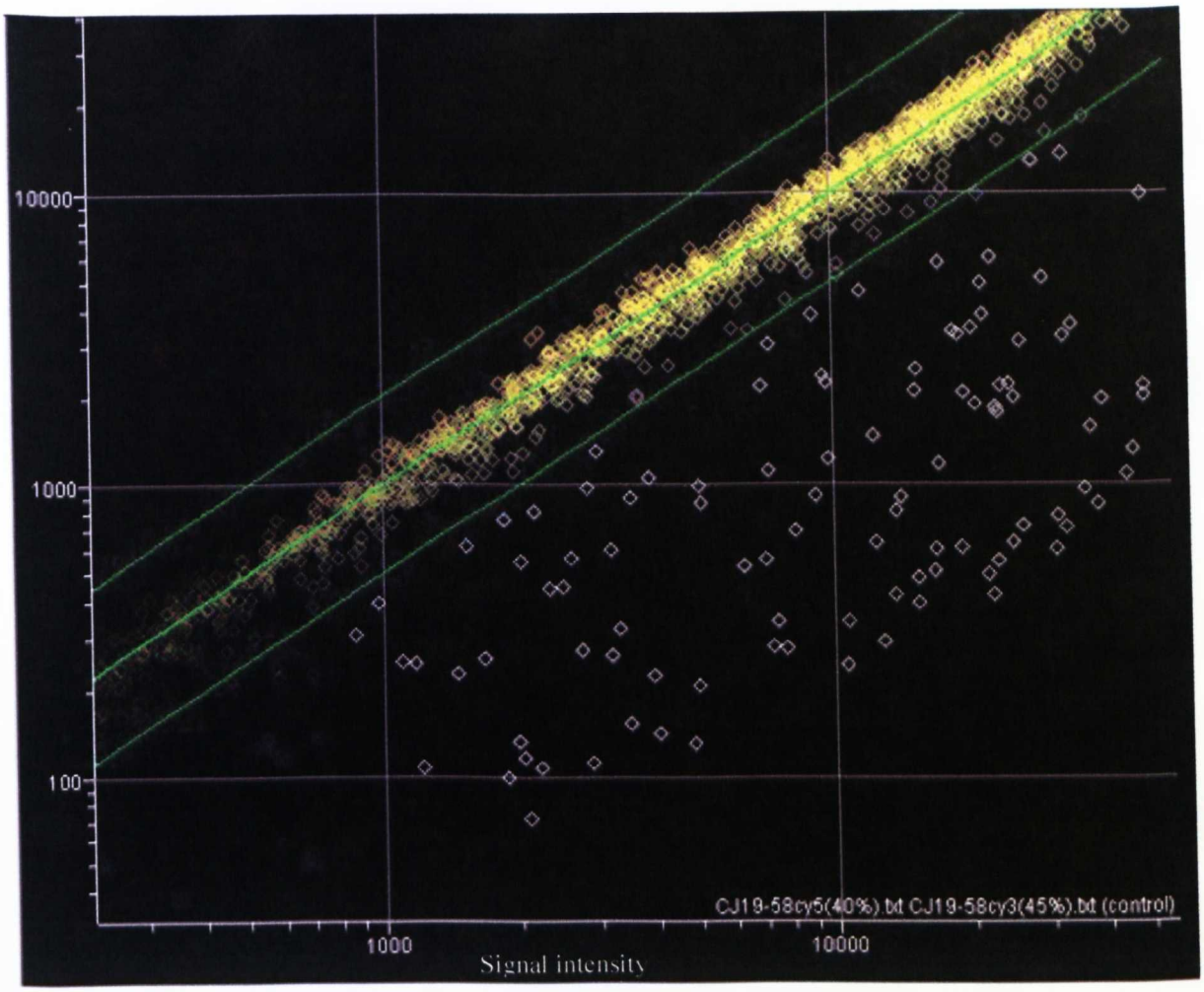

Figure 12. Scatterplot showing the distribution of signal ratios for strain 31485 against NCTC11168 using GeneSpring6.1 software. CDSs that hybridised at a 1:1 ratio are distributed along the central green line and are coloured yellow. The lower line represents the 0.5 cut-off. Test strain CDSs with a ratio of 0.5 or less of the control strain CDS found below the lower green line are coloured white. These CDSs are designated absent or divergent in the test strain 31485.

The second method used to determine which reporter elements are absent or divergent in the test strain utilises a dynamic cut-off value, calculated for each hybridisation experiment. The expected distribution of the reporter elements is compared to the observed distribution for an individual hybridisation and the cut-off value is calculated based on the observed microarray hybridisation data (Chapter 2.5.4). GACK has a singular comparative genomics function to determine genes that are absent or divergent using an algorithm that selects a specific cut-off value for each hybridisation. GACK utilises an algorithm that estimates the main ratio peak, the first step of which is finding the location and height of the major peak and the ratio values at half the peaks maximum height on both sides. A normal probability density

function is fitted using these three parameters. The peak represents the genes that are present in both the test and control strains. The tail to the left (Figure 13) represents the absent or divergent genes. The estimated probability of presence (EPP) of a gene is calculated by dividing the mapped normal curve value which is expected for a distribution in which all spots are present on the array, by the observed data distribution value (Chapter 2.5.4).

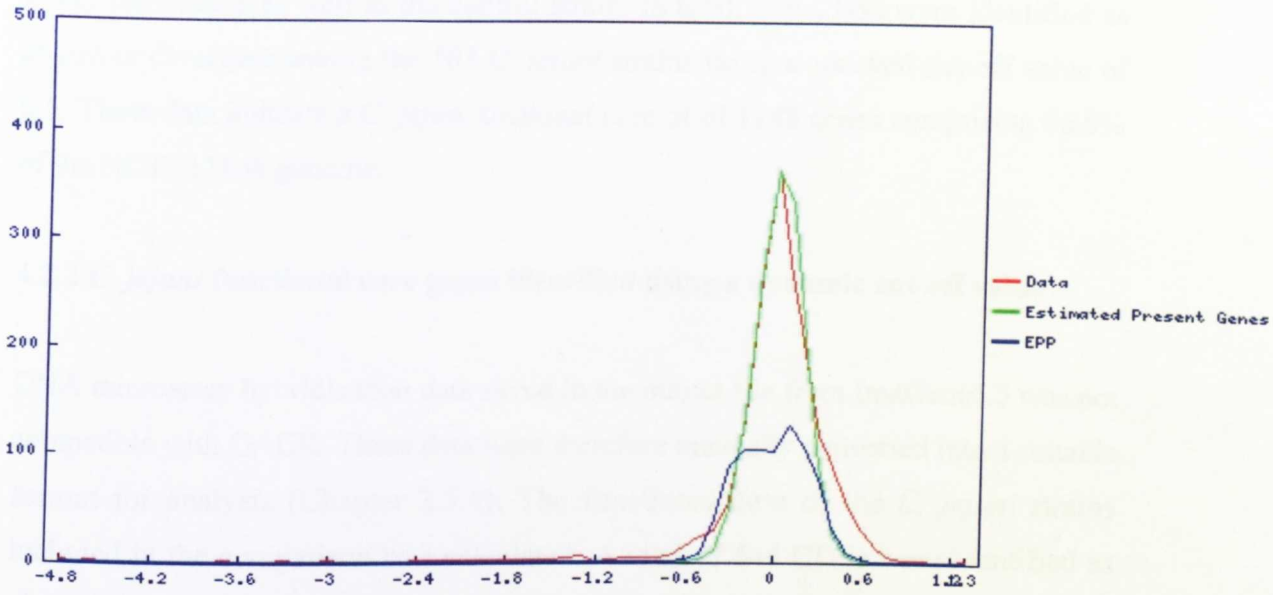

Figure 13. GACK histogram showing the distribution of signal ratios for strain 31485
The dynamic cut-off value for the competitive hybridisation of test strain 31485 against NCTC11168 is calculated within GACK by plotting the normal curve value that is expected for a distribution in which all spots are present on the array (shown in green). The expected distribution is divided by the observed data distribution (shown in red) giving the estimated probability of presence (EPP) (shown in blue).

4.2 Results – Identification of *C. jejuni* functional core genes using a constant and dynamic cut-off value

4.2.1 *C. jejuni* functional core genes identified using a constant cut-off value

Microarray slides were scanned with a GMS 418 Array Scanner (Chapter 2.5.1) and the ratio of fluorescence between the test and control strains was quantified using ImaGene 5.5 software (Chapter 2.5.2). ImaGene5.5 output text files were loaded directly into the GeneSpring6.1 programme with replicate hybridisations loaded into a single experiment. Raw microarray data imported from Imagene5.5 was normalised

(Chapter 2.5.3) allowing biological variations to be differentiated from variations due to experimental procedures.

A constant cut-off value of 0.5 was selected to identify absent or divergent CDSs in each test strain i.e. reporter elements with signal intensity ratios of 0.5 or less compared to the control channel were deemed absent or divergent in the test strain. The species-specific functional core of the 103 *C. jejuni* included in the comparison was determined by calculating the number of CDSs that were present in each of the *C. jejuni* test strains as well as the control strain. In total, 579 CDSs were identified as absent or divergent among the 103 *C. jejuni* strains using a constant cut-off value of 0.5. These data indicate a *C. jejuni* fuctional core of of 1148 genes comprising 66.0% of the NCTC11168 genome.

4.2.2 *C. jejuni* functional core genes identified using a dynamic cut-off value

DNA microarray hybridisation data saved in the output file from ImaGene5.5 was not compatible with GACK. These data were therefore manually converted into a suitable format for analysis (Chapter 2.5.4). The functional core of the *C. jejuni* strains included in the comparison was calculated. A total of 544 CDSs were identified as absent or divergent in 103 *C. jejuni* strains using a dynamic cut-off value. Using dynamic cut-off value a *C. jejuni* functional core of 1183 CDSs was therefore identified, comprising 69.0% of the genome.

4.2.3 Comparison of absent or divergent CDSs identified in *C. jejuni* strains using constant and dynamic cut-off values

Using a constant cut-off value of 0.5, a total of 579 CDSs were identified as absent or divergent in one or more of the *C. jejuni* strains tested, compared with 544 identified using a dynamic cut-off value. This corresponds to a functional core of 1148 CDSs and 1183 CDSs respectively. The difference of 35 CDSs between the two results suggested a variability of only 2% between the two methods for *C. jejuni* species hybridisations. Overall, analysing DNA microarray hybridisation data using a constant cut-off value resulted the designation of a higher proportion of the genome as variable and therefore a smaller functional core. Absent or divergent CDSs identified in *C. jejuni* using the two methods were compared using the Venn diagram function

available within GeneSpring6.1 sotware (Figure 14) to determine whether the same CDSs were identified in both.

Comparison of the CDSs identified as absent or divergent using a constant and dynamic cut-off value highlighted a discrepancy between the CDSs identified with each method. Of the 579 CDSs identified as absent or divergent using a constant cut-off value of 0.5, 479 CDSs were also identified using a dynamic cut-off value. However, 100 CDSs were uniquely identified as absent or divergent using a constant cut-off value. Similarly, of the 544 CDSs identified as absent or divergent using a dynamic cut-off value, 479 CDSs were also identified using a constant cut-off value, however, 65 CDSs were identified uniquely to this method.

A. B.

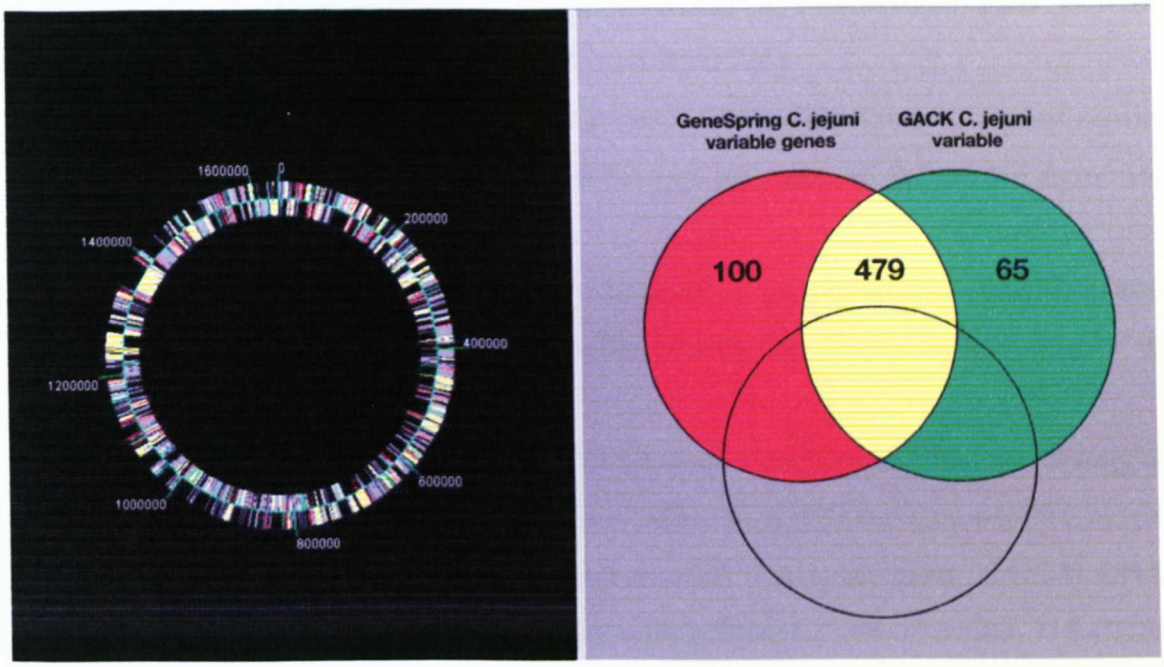

Figure 14. Venn Diagram showing CDSs that were designated absent or divergent in *C. jejuni* strains using a constant cut-off of 0.5 (red circle) and a dynamic cut-off value (green circle). The distribution of *C. jejuni* CDSs in the genome can be visualised in section (A). CDSs present in both lists are shown in yellow can be clearly seen distributed throughout the genome. CDSs identified as absent or divergent using a constant cut-off value of 0.5 are shown in the red circle. CDSs identified as absent or divergent using a dynamic cut-off value are shown in the green circle. CDSs identified as absent or divergent using both methods are shown in the yellow section. Section (B) shows a Venn diagram with the 579 CDSs (red and yellow) identified as absent or divergent using a constant cut-off value of 0.5. In the green and yellow circles 544 identified as absent or divergent using a dynamic cut-off value are shown. CDSs in the yellow overlapping segment are CDSs identified as absent or divergent using both cut-off values.

4.3 Results – Comparison of DNA microarray hybridisation data of two sequenced *C. jejuni* strains (NCTC11168 and RM1221)

4.3.1 Comparison of DNA microarray hybridisation data from *C. jejuni* strain NCTC11168 and RM1221 using a constant cut-off value of 0.5 and a dynamic cut-off value

The DNA of two fully sequenced and annotated *C. jejuni* strains (RM1221 and NCTC11168) were competitively hybridised with the *C. jejuni* microarray in duplicate.

Using a constant cut-off value of 0.5, 159 CDSs were identified as absent or divergent from strain RM1221. Of these 159 variable CDSs, 31 CDSs were non-NCTC11168 CDSs for which direct comparisons were not possible. Using the same RM1221 DNA microarray hybridization data but analyzing with a dynamic cut-off value, 118 CDSs were identified as absent or divergent of which 17 CDSs were non-NCTC11168 CDSs (Appendix 2). The total number of CDSs identified as absent or divergent using a dynamic cut-off value (118 CDSs) was also identified using the constant cut-off of 0.5. However, using a constant cut-off value an additional 41 CDSs were identified as absent or divergent (Appendix 3).

4.3.2 Comparison of DNA microarray data from RM1221 with sequence data

NCTC11168 and RM1221 genome sequences were compared using BLASTn at BμG@S. RM1221 CDSs against NCTC11168 CDS and *C. jejuni* microarray reporter elements against RM1221 CDSs BLASTn results were supplied (Adam Witney, personal communication). BLASTn results indicated whether each of the NCTC11168 CDS had 'no match' to any RM1221 CDSs, or, if a match was found, the nucleotide identity of the two CDSs.

CDSs identified as absent or divergent using constant and dynamic cut-off values were compared back to the RM1221 BLASTn data to determine whether each CDS was absent, divergent or present. Of the 159 CDSs identified as absent or divergent using a constant cut-off value of 0.5, 128 were from NCTC11168 and 31 were non-NCTC11168 CDSs. Thus, 128 of a potential 159 CDSs were compared with the NCTC11168 against RM1221 BLASTn data. No match was found in the genome sequence of RM1221 for 61 CDSs of the 128 NCTC11168 CDSs identified as absent or divergent in RM1221 using a constant cut-off value of 0.5, that is 61 CDSs were correctly identified as absent from RM1221 using a constant cut-off value of 0.5

(Appendix 4). Additionally, 57 of the 128 NCTC11168 CDSs were identified as absent or divergent from RM1221, using a constant cut-off value of 0.5. A total of 10 CDSs were incorrectly identified as absent or divergent using a constant cut-off value of 0.5 as these 10 CDSs were found to have 100% nucleotide identity in both NCTC11168 and RM1221 (Appendix 5).

Analysing RM1221 DNA microarray data with a dynamic cut-off value method identified 118 CDSs as absent or divergent of which 101 CDSs were from NCTC11168 and 17 were non-NCTC11168 CDSs. Comparing the 101 NCTC11168 CDSs that were identified as absent or divergent using a dynamic cut-off value with the BLASTn sequence comparisons revealed that 57 CDSs were correctly identified as absent as these CDSs had no match to any CDSs in the RM1221 genome (Appendix 4). In addition 35 of the 101 absent or divergent CDSs were correctly identified as divergent from the RM1221 genome. However, 9 CDSs were incorrectly identified as absent or divergent as these 9 CDSs has 100% nucleotide identity in both NCTC11168 and RM1221 (Appendix 5). These data are summarized in Table 8.

4.3 Results – Comparison of DNA microarray hybridisation data of two sequenced *C. jejuni* strains (NCTC11168 and RM1221)

4.3.1 Comparison of DNA microarray hybridisation data from *C. jejuni* strain NCTC11168 and RM1221 using a constant cut-off value of 0.5 and a dynamic cut-off value

The DNA of two fully sequenced and annotated *C. jejuni* strains (RM1221 and NCTC11168) were competitively hybridised with the *C. jejuni* microarray in duplicate.

Using a constant cut-off value of 0.5, 159 CDSs were identified as absent or divergent from strain RM1221. Of these 159 variable CDSs, 31 CDSs were non-NCTC11168 CDSs for which direct comparisons were not possible. Using the same RM1221 DNA microarray hybridization data but analyzing with a dynamic cut-off value, 118 CDSs were identified as absent or divergent of which 17 CDSs were non-NCTC11168 CDSs (Appendix 2). The total number of CDSs identified as absent or divergent using a dynamic cut-off value (118 CDSs) was also identified using the constant cut-off of 0.5. However, using a constant cut-off value an additional 41 CDSs were identified as absent or divergent (Appendix 3).

4.3.2 Comparison of DNA microarray data from RM1221 with sequence data

NCTC11168 and RM1221 genome sequences were compared using BLASTn at BμG@S. RM1221 CDSs against NCTC11168 CDS and *C. jejuni* microarray reporter elements against RM1221 CDSs BLASTn results were supplied (Adam Witney, personal communication). BLASTn results indicated whether each of the NCTC11168 CDS had 'no match' to any RM1221 CDSs, or, if a match was found, the nucleotide identity of the two CDSs.

CDSs identified as absent or divergent using constant and dynamic cut-off values were compared back to the RM1221 BLASTn data to determine whether each CDS was absent, divergent or present. Of the 159 CDSs identified as absent or divergent using a constant cut-off value of 0.5, 128 were from NCTC11168 and 31 were non-NCTC11168 CDSs. Thus, 128 of a potential 159 CDSs were compared with the NCTC11168 against RM1221 BLASTn data. No match was found in the genome sequence of RM1221 for 61 CDSs of the 128 NCTC11168 CDSs identified as absent or divergent in RM1221 using a constant cut-off value of 0.5, that is 61 CDSs were correctly identified as absent from RM1221 using a constant cut-off value of 0.5

Table 8. Comparison of data analysis methods for identifying absent or divergent CDSs from the sequenced *C. jejuni* strain RM1221 compared to NCTC11168

Cut-off Value	Constant (0.5)	Dynamic
Total number of CDSs identified as absent or divergent	159	118
Non-NCTC11168 CDS identified as absent or divergent	31	17
NCTC11168 CDSs identified as absent or divergent	128	101
CDSs that are absent from RM1221	61	57
CDSs that are divergent in RM1221	57	35
CDSs identified as absent or divergent from RM1221 that are 100% identical (false negatives)	10	9

4.4 Discussion

4.4.1 Calculation of the *C. jejuni* functional core using two different analysis methods

One of the objectives of this chapter was to compare two DNA microarray data analysis methods on the same collection of *C. jejuni* strains. These data would indicate how the use of different cut-off criteria influences the identification of absent or divergent genes.

Whole genome comparisons were carried out using GeneSpring6.1 software and GACK. This comparison allowed the difference between a constant cut-off value of

0.5 to denote absence or divergence used in GeneSpring6.1 to be compared with a dynamically selected cut-off value used in GACK. Within GeneSpring6.1 it is assumed that most genes will be present in the test and control channel thus ratios are normalised to represent a linear ratio of 1. Using GeneSpring6.1 a constant ratio cut-off value is selected to categorise absent or divergent genes from those that are present in the test strain. In this study genes were categorised as present with a ratio of 1, marginal when this ratio fell below 1 but above 0.5 and absent or divergent with a ratio of 0.5 or below. In bacteria where more than one sequenced strain is available this cut-off value may be chosen empirically by comparing CDSs in which nucleotide variation is known. However, for *C. jejuni* only one annotated genome was available until February 2005, thus the 0.5 cut-off value was not empirically determined but was selected based on hybridisation data from the clone array.

By contrast GACK calculates an independent cut-off ratio for each hybridisation. An EPP is calculated by dividing the expected distribution (if all reporter elements were present in the test strain) by the observed distribution. GACK categorises genes either in binary form as either absent or divergent or trinary (used in this study) where genes are categorised as absent with an EPP of 0%, present with an EPP of 100% and into a transition region if the EPP is anything other than 0% or 100%.

A comparison of these two DNA microarray data analysis methods using the same data set of 103 *C. jejuni* strains identified 579 CDSs (constant cut-off value of 0.5) compared with 544 CDSs (dynamic cut-off) as absent or divergent of which 479 were common to both methods. This corresponded to a functional core of 1148 CDSs (constant cut-off) and 1183 (dynamic cut-off) CDSs respectively. A discrepancy of 100 CDSs identified as absent or divergent only with the constant cut-off value and 65 CDSs were identified as absent or divergent only with a dynamic cut-off value.

4.4.2 Comparison of RM1221 and NCTC11168

The second objective of this chapter was to determine whether a DNA microarray data analysis method utilising a constant or dynamic cut-off value was more sensitive at identifying absent or divergent CDSs.

The publication of a second *C. jejuni* genome sequence, RM1221, in 2005 facilitated the BLASTn comparison of the two sequenced *C. jejuni* strains. BLASTn comparisons of RM1221 CDSs against NCTC11168 CDSs, as well as RM1221 CDSs

against reporter elements were carried out by Adam Witney (BμG@S). BLASTn data provided a quantitative method to demonstrate whether a constant cut-off or a dynamic cut-off value was more sensitive at identifying absent or divergent CDSs. Direct comparisons of two sequenced *C. jejuni* strains, RM1221 and NCTC11168, on the *C. jejuni* DNA microarray facilitated the validation of DNA microarray hybridization data through direct comparisons against the sequence.

RM1221 DNA microarray hybridisation data analysed with a constant cut-off value of 0.5 identified 159 CDSs as absent or divergent of which 128 were NCTC11168 CDSs. By comparing the genome sequences of NCTC11168 and RM1221 the total number of CDSs that had no match with RM1221, and were therefore correctly identified as absent from RM1221 could be determined. Of the 128 CDSs, 61 were correctly identified as absent from RM1221 and a further 57 were divergent. Using a dynamic cut-off value to analyze RM1221 DNA microarray hybridization data 118 CDSs were identified as absent or divergent, including 101 NCTC11168 CDSs of which 57 CDSs were correctly identified as absent and 35 CDSs correctly as divergent. All of which were also identified using the constant cut-off value of 0.5. Thus, the data analysis method utilising constant cut-off value of 0.5 identified all the absent or divergent CDSs identified with a dynamic cut-off, with an additional four absent CDSs and 22 divergent CDSs. However, ten CDSs and nine CDSs labeled as absent or divergent using a constant cut-off and dynamic cut-off value respectively were, in fact, present in RM1221 with 100% nucleotide identity. It is uncertain why these CDSs were identified as absent or divergent in RM1221. However, nine CDSs were mistakenly identified as absent or divergent using both methods (with an extra one CDS identified using the constant cut-off value). It is possible that if multiple paralogous CDSs are present in NCTC11168 for a single RM1221 gene, cross hybridization may occur leading to lower signal intensity in the test channel. This would result in the output data for that reporter element indicating that the CDSs was absent or divergent in the test strain, explaining the identification of false negatives with both analysis methods.

Overall, these data indicate that the use of a DNA microarray data analysis method that utilises a constant cut-off value compared with a dynamic cut-off value facilitates the identification of more of the genuinely absent (61CDSs compared with 57 CDSs) and divergent CDSs (57 CDSs compared with 35 CDSs) in a test strain.

5.0 Comparative genomics of potentially non-pathogenic strains and human *C. jejuni* strains from patients with different clinical presentations

5.1 Introduction

5.1.1 Aims

C. jejuni strain diversity combined with variable host responses results in a complex spectrum of disease outcomes, ranging from asymptomatic colonisation to severe inflammatory diarrhoea and occasionally sequelae such as GBS. Previous studies have shown that remarkable differences also exist between *C. jejuni* strains with respect to phenotypic properties such as cell invasiveness, rates of translocation across cell monolayers, toxin production and colonisation of chickens (Hu and Kopecko, 1999), but the genetic differences between *C. jejuni* strains exhibiting these observed differences are unknown. Precise strain comparisons from well-characterised strains of diverse origins may allow correlates of pathogenesis to be determined and the subsequent identification of potential virulence determinants. In this chapter, strains representing the spectrum of clinical presentations and outcomes of human *C. jejuni* infection have been identified for comparative genomics analysis. The overall aim of this chapter was to identify putative virulence determinants and genetic markers of clinical outcome through the comparative hybridisation of DNA from strains representing a range of clinical outcomes with the *C. jejuni* composite DNA microarray. The strains were divided into four clinical sets to facilitate data comparison; potentially non-pathogenic, gastroenteritis, septicaemia and GBS.

5.1.2 Strains included in the study

Potentially non-pathogenic strains from asymptomatic carriers were identified from the IID study (Chapter 1.2.1). Phenotypic information about these strains is shown in Table 9.

Table 9. Phenotypic information for *C. jejuni* strains isolated from asymptomatic carriers identified from the IID study

*UT= untypeable

**RDNC = reacts with phage but does not conform to designated type

Strain	Serotype	Phage type	Source
33084	HS35	1	Human
33106	HS4	1	Human
31481	HS37	76	Human
31485	UT*	2	Human
31467	HS18	RDNC**	Human
32787	HS18	RDNC**	Human
32799	HS50	5	Human

A second set of potentially non-pathogenic strains, identified by MLST, were included in the study. Two clonal complexes associated with the sand of bathing beaches, and not with human disease cases, livestock or chickens have been identified using MLST (K.E. Dingle, Personal Communication). It is thought that the natural host of these isolates may be wild birds. In Sweden, *C. jejuni* has been identified in black-headed gulls (*Larus ridibundus*) and it is thought that wild birds may act as an important reservoir for *C. jejuni* (Broman *et al.*, 2002). We can, therefore, hypothesise that faecal contamination of the beaches by birds may have preceded the isolation of these strains which may lack genes with potentially important roles in human virulence or chicken colonisation. Six such isolates were included in the study for which phenotypic information is provided in Table 10.

Table 10. Phenotypic information for potentially non-pathogenic *C. jejuni* strains from sand samples

*UT= untypeable

Strain	Penner serotype	Source
1771 (79309)	UT*	Beach isolate
1772 (79260)	HS 55	Beach isolate
1773 (79196)	UT*	Beach isolate
1791 (79207)	HS 2	Beach isolate
1792 (79046)	No data	Beach isolate
1793 (79044)	HS 5	Beach isolate

Having established two potential sources of non-pathogenic strains, strains from different clinical presentations were identified through the Sentinel study (Chapter 1.2.2). Strains representing three cohorts of patients presenting with distinct clinical symptoms; diarrhoea, bloody diarrhoea and vomiting (Table 11) were selected for comparative genomics investigations.

Table 11. Phenotypic information for *C. jejuni* strains with different clinical presentations

*UT= untypeable

Strain	Serotype	Phage type	Diarrhoea	Bloody stool	Vomiting
34555	HS 5	34	1	0	0
35424	UT*	1	1	0	0
35535	UT*	1	1	1	0
35799	UT*	1	0	0	0
36069	HS 5	1	1	0	1
36439	HS12	1	1	0	0
36860	HS21	44	1	0	0
36952	UT*	36	1	0	0

37537	UT*	1	1	1	0
38353	HS 5	1	1	1	0
38556	HS13	No data	1	0	1
38576	UT*	33	1	1	0
38762	HS18	No data	1	1	0
38857	HS23	1	1	0	1
39182	HS13	1	1	0	1
39640	UT*	1	1	1	0
39828	HS42	1	1	0	1
40917	HS21	8	1	1	0
41651	HS16	5	1	0	1
42724	UT*	1	1	1	0
43157	HS13	2	1	0	1
43205	HS 2	33	1	1	0
44464	HS37	44	1	0	0
44811	HS 2	36	1	0	1
44933	HS13	1	1	0	0
44958	HS50	34	1	0	1
45557	HS60	No data	1	1	0
45631	HS 4	1	1	0	0
48612	HS 2	36	1	1	0
50097	HS13	39	1	1	0
52331	HS50	34	1	0	1
52368	UT*	44	1	1	0
53259	HS 2	1	1	0	1
54489	UT*	33	1	0	0
55320	HS13	2	1	1	0
55703	HS13	1	1	0	1
56281	HS50	5	1	1	0
56282	HS50	5	1	0	0
56519	HS12	2	1	1	0
56832	HS50	1	1	0	1

58473	HS 2	62	1	0	0
59161	HS 2	1	1	0	1
59214	UT*	33	1	0	0
59364	HS31	2	1	0	0
59424	HS31	2	1	0	1
59627	UT*	19	1	0	0
62567	UT*	1	1	0	0
62914	UT*	1	1	0	1
63326	HS31	1	1	0	0
64555	HS31	2	1	1	0

The symptoms most commonly associated with campylobacteriosis are fever, abdominal cramps and diarrhoea. However, sequelae can occur, including endocarditis, bacteraemia, meningitis and GBS (Blaser, 1995) (Chapter 1.2.3). GBS is a sequela to infection and *C. jejuni* is the most commonly associated organism that triggers this neuropathy. We may hypothesise that strains associated with sequelae may differ at a genomic level to strain NCTC11168 that was isolated from a patient with gastroenteritis who apparently did not develop GBS.

To investigate a potentially hyperinvasive group of *C. jejuni* strains, seven isolates from the blood samples of patients with *C. jejuni* septicaemia and two *C. jejuni* strains isolated from patients who subsequently developed GBS were selected (Table 12).

Table 12. Phenotypic information on *C. jejuni* strains associated with sequelae

*UT= untypeable

Strain	Source	Serotype (HS)	Phage type	Clinical data
46979	Human blood	50	6	Immunocompromised – Chronic leukaemia
43983	Human blood	50	6	Infected in USA
44119	Human blood	18	2	Immunocompromised – Non-Hodgkins lymphoma
34007	Human blood	18	2	Immune system intact
52472	Human blood	UT*	1	Immune system intact
53250	Human blood	60	2	Immunocompromised – Post operation sternaortic biopsy
47439	Human blood	67	1	Immune system intact
15168	GBS	19	2	No data
18836	GBS	19	2	No data

5.2 Results – DNA microarray hybridisation data

5.2.1 Asymptomatic carriage and other potentially non-pathogenic strains

Microarray hybrisidisations were carried out in duplicate on two experimental replicates for each strain using the method described in Chapter 2.4. Variation between seven *C. jejuni* strains isolated from asymptomatic patients and strain NCTC11168 was observed with between 92 CDSs and 172 CDSs absent or divergent. CDSs that were designated absent or divergent in each individual strain were compared using the Venn diagram function in GeneSpring6.1 producing a gene list containing only genes that were absent or divergent in each of the seven strains. Twenty-two CDSs were consistently absent or divergent in the seven strains (Table 13). This diversity was found largely in the CPS biosynthesis locus in which CDSs

cj1421 and *cj1422* were consistently absent or divergent from all seven isolates. *Cfr*A, a probable iron uptake protein (ferric receptor) was also absent or divergent in each of the seven strains from asymptomatic carriers. In addition, one of the novel non-NCTC11168 CDSs, *cj11828-05* involved with LOS biosynthesis, was also absent or divergent in each of the potentially non-pathogenic strains isolated from asymptomatic patients.

Table 13. CDSs absent or divergent in seven potentially non-pathogenic strains isolated from asymptomatic carriers

CDS	Genbank Annotation
cj0755	CfrA, probable iron uptake protein (ferric siderophore receptor)
cj0139	possible endonuclease
cj1421c	capsule polysaccharide biosynthesis locus
cj1422c	capsule polysaccharide biosynthesis locus
cj1426c	capsule polysaccharide biosynthesis locus
cj1428c	capsule polysaccharide biosynthesis locus
cj1429c	capsule polysaccharide biosynthesis locus
cj1430c	capsule polysaccharide biosynthesis locus
cj1431c	capsule polysaccharide biosynthesis locus
cj1432c	capsule polysaccharide biosynthesis locus
cj1433c	capsule polysaccharide biosynthesis locus
cj1434c	capsule polysaccharide biosynthesis locus
cj1435c	capsule polysaccharide biosynthesis locus
cj1436c	capsule polysaccharide biosynthesis locus
cj1437c	capsule polysaccharide biosynthesis locus
cj1438c	capsule polysaccharide biosynthesis locus
cj1439c	capsule polysaccharide biosynthesis locus
cj1440c	capsule polysaccharide biosynthesis locus
cj1441c	capsule polysaccharide biosynthesis locus
cj1447c	kpsT, probable capsule polysaccharide export ATP-binding protein
cj1448c	kpsM, probable capsule polysaccharide export system inner membrane protein
cj11828-05	Lipo oligosaccharide sequence from non-NCTC11168 strain 11828

Six potentially non-pathogenic strains isolated from bathing beaches were also hybridised with the microarray. Higher levels of variation throughout the genome were observed in the beach isolates than for the potentially non-pathogenic strains from asymptomatic patients. Between 127 CDSs and 225 CDSs were absent or divergent in the beach isolates compared to the control strain, NCTC11168. CDSs

that were designated absent or divergent in each individual strain were compared using Venn diagrams to produce a gene list indicating only CDSs absent or divergent to all the strains included in the study. Forty-eight CDSs were consistently absent or divergent in the six strains (Appendix 6). This diversity was found largely in the LOS biosynthesis locus with five CDSs from this locus, *cj1138*, *cj1139*, *cj1142*, *cj1143* and *cj1144*, divergent in all six isolates. Interestingly, a probable secreted serine protease (*cj1365*) as well as three CDSs associated with iron uptake (*cj0178*, *cj0179* and *cj0181*) were also absent or divergent in each of the six strains. Only two CDSs identified as absent or divergent in strains isolated from asymptomatic carriers were absent or divergent in the beach isolates, *cj0755* and *cj0139* encoding a probable ferric receptor and possible endonuclease respectively. Over 15% of the genes that were absent or divergent in the beach isolated were of unknown function. Table 14 shows a selection of CDSs identified as absent or divergent from each of the potentially non-pathogenic beach isolates.

Table 14. Selected CDSs that were absent or divergent in all potentially non-pathogenic beach isolates (full list in Appendix 6)

CDS	Genbank Annotation
cj0755	CfrA, probable iron uptake protein (ferric siderophore receptor)
cj0139	possible endonuclease
cj1138	LOS biosynthesis
cj1139c	LOS biosynthesis
cj1142	LOS biosynthesis
cj1143	LOS biosynthesis
cj1144	LOS biosynthesis
cj1365c	probable secreted serine protease
cj0178	possible outer membrane siderophore receptor
cj0179	ExbB1, biopolymer transport protein
cj0181	TonB1, possible tonB transport protein

5.2.2 Gastroenteritis; diarrhoea, bloody diarrhoea and vomiting

In total, 14 *C. jejuni* isolates from patients presenting with diarrhoea, 15 with bloody diarrhoea and 15 with vomiting as the predominant symptom of campylobacterioisis were competitively hybridised with NCTC11168 (a clinical isolate from a patient with diarrhoea). Analysis of the hybridisation data revealed no CDSs that were absent or divergent in all of the 14 *C. jejuni* strains from patients presenting with diarrhoea as the predominant symptom compared to NCTC11168. Furthermore, of the 15 isolates from patients with bloody diarrhoea and 15 isolates from patients with vomiting as the predominant symptom, no CDSs were identified as absent or divergent from strains associated with either clinical presentation.

5.2.3 Septicaemia

Seven strains isolated from the blood samples of patients with *C. jejuni* septicaemia were hybridised with the composite DNA micorarray. Between 13 and 130 CDSs were absent or divergent in the strains with a median of 117 CDSs that were absent or divergent. Twenty CDSs were absent/divergent from five of the strains (Table 15), with

six of these CDSs common to six strains and just one, *Cj1677* encoding a probable lipoprotein, absent or divergent in all seven strains.

Table 15. CDSs absent or divergent in five septicaemia strains 52472, 47939, 44119, 43983 and 34007

CDS	Genbank annotation
cj0033	probable integral membrane protein
cj0815	unknown
cj0816	unknown
cj0818	probable lipoprotein
cj1051c	probable restriction modification enzyme
cj1136	LOS biosynthesis locus
cj1137c	LOS biosynthesis locus
cj1138	LOS biosynthesis locus
cj1139c	LOS biosynthesis locus
cj1141	*neuB1* LOS biosynthesis locus
cj1142	*neuC1* LOS biosynthesis locus
cj1143	*neuA1* LOS biosynthesis locus
cj1144c	LOS biosynthesis locus
cj1145c	LOS biosynthesis locus
cj1395	possible pseudogene
cj1677	probable lipoprotein
cj1721c	possible outer membrane protein
cj1722c	unknown
cj1723c	probable periplasmic protein

5.2.4 Microarray analysis of strains associated with GBS sequelae

A total of 131 CDSs were absent or divergent to both GBS strains included in the study, including *cj1686*, or *Top*A encoding DNA topoisomerase I.

5.3 Results – Comparative phylogenomics of human *C. jejuni* strains from the spectrum of disease outcome

5.3.1 Phylogenetic relationships

DNA microarray hybridisation data for 91 *C. jejuni* strains was collated. The strains were all isolated in the UK and included 63 isolates from the spectrum of clinical outcomes described above and 28 non-clinical *C. jejuni* strains (including six potentially non-pathogenic strains isolated from sand from beaches) (Appendix 1). The ratio of fluorescence for each NCTC11168 reporter element was converted into binary form manually and later using a Perl script (provided by Adam Witney). Zero was taken to represent a gene that was absent or divergent (with a ratio of 0.5 or below) and '1' represented a gene that was present in both the control and test strain (with a ratio of above 0.5) (Chapter 2.6.1). This was repeated for 91 *C. jejuni* strains for which whole genome comparisons were carried out.

Clinical strains 39640 and 53259 in which over 5% of the total genes were flagged were manually removed from both data sets as were genes flagged in over 10% of all strains (Table 16). Remaining flagged genes were denoted with a question mark.

Table 16. Flagged genes removed from data sets

Gene removed from data set	Percentage of strains in which gene is flagged
cj0001	18%
cj0002	11%
cj0008	18%
cj0011	22%
cj0030	23%
cj0036	14%
cj0118	66%
cj0121	53%
cj0124	37%
cj0128	26%
cj0143	11%
cj0185	12%
cj0395	22%
cj0735	16%
cj0747	24%
cj0748	21%
cj0776	17%
cj0782	10%
cj0797	10%
cj0801	50%
cj0896	63%
cj1024	63%
cj1063	55%
cj1072	67%
cj1073	66%
cj1106	73%
cj1694	23%

The data set was then cross checked back against the raw microarray data to ensure that genes were correctly named '1', '0' and '?' according to the original gene lists in

GeneSpring6.1. These data were then transformed into Nexus format (Chapter 2.6.1). The Nexus format matrix was used to determine the relationship of strains using Bayesian based algorithms implemented through Mr Bayes v.3.0 software. With samples and saves from every fortieth tree, two million generations of Markov Chain Monte Carlo (MCMC) were completed and this process was optimised by adjusting parameters such as annealing temperature. The first one million trees were then discarded for each optimisation and statistical stability was calculated to determine the robustness of the phylogeny of the last 750,000 trees. Majority rule consensus trees and clade credibility values were obtained for each tree using Phylogenetic Analysis Using Parsimony (PAUP*). Tree topology was viewed using TreeView. The optimal tree was selected based on statistical support for the phylogeny. The tree showed a topology comprising two major clades (two or more strains with a common ancestor), herein referred to as clade A and B. Clade A contained 26/63 (41%) of the total number of clinical isolates including 1/9 (11%) of those that were associated with sequelae (Figure 15). Clade B contained the remaining 37/63 (59%) of clinical isolates including 8/9 (89%) of those that were involved with sequelae i.e. post infection neuropathy and septicaemia.

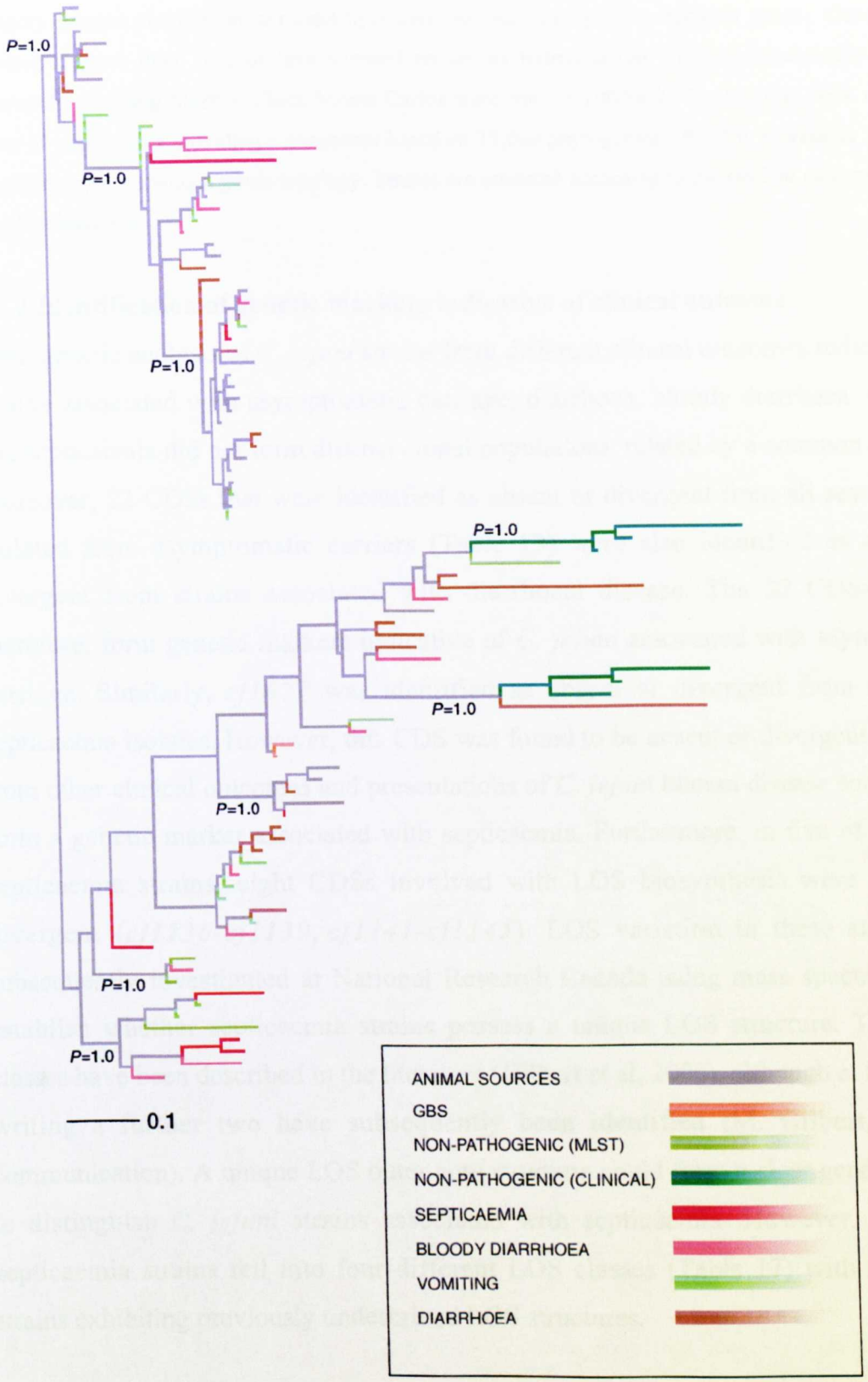
P=1.0
P=1.0
P=1.0
P=1.0
P=1.0
P=1.0
P=1.0
0.1
ANIMAL SOURCES
GBS
NON-PATHOGENIC (MLST)
NON-PATHOGENIC (CLINICAL)
SEPTICAEMIA
BLOODY DIARRHOEA
VOMITING
DIARRHOEA

Figure 15. Phylogenetic relationship of strains associated with different clinical outcomes. Bayesian strict and 95% majority rule consensus phylogeny of *C. jejuni* DNA-DNA microarray binary data (hybridisiation threshold 0.5) with associated branch lengths. The phylogeny incorporates a 16 category gamma distribution to model hybridisation rate heterogeneity between genes. Genes were omitted if more than 10% of taxa showed erroneous hybridisation. Four incrementally heated Metropolis coupling Markov Chain Monte Carlos were run for 1000000. Phylogenies were sampled every 40 generations providing a consensus based on 25,000 phylogenies. $P = 1.0$, represents 100% of all phylogenies showing a given topology. Strains are coloured according to the clinical outcome of the *C. jejuni* infection.

5.3.2 Identification of genetic markers indicative of clinical outcome

Phylogenetic analysis of *C. jejuni* strains from different clinical outcomes indicated that strains associated with asymptomatic carriage, diarrhoea, bloody diarrhoea, vomiting and septicaemia did not form distinct clonal populations, related by a common ancestor. Moreover, 22 CDSs that were identified as absent or divergent from all seven strains isolated from asymptomatic carriers (Table 13) were also identified as absent or divergent from strains associated with diarrhoeal disease. The 22 CDSs do not, therefore, form genetic markers indicative of *C. jejuni* associated with asymptomatic carriage. Similarly, *cj1677* was identified as absent or divergent from all seven septicaemia isolates. However, this CDS was found to be absent or divergent in strains from other clinical outcomes and presentations of *C. jejuni* human disease and does not form a genetic marker associated with septicaemia. Furthermore, in five of the seven septicaemia strains, eight CDSs involved with LOS biosynthesis were absent or divergent (*cj1136-cj1139*, *cj1141-cj1145*). LOS variation in these strains was subsequently investigated at National Research Canada using mass spectrometry to establish whether septicaemia strains possess a unique LOS structure. Three LOS classes have been described in the literature (Gilbert et al, 2000), although at the time of writing a further two have subsequently been identified (M. Gilbert, personal communication). A unique LOS outer core structure could form a clear genetic marker to distinguish *C. jejuni* strains associated with septicaemia. However, six of the septicaemia strains fell into four different LOS classes (Table 17) with two of the strains exhibiting previously undescribed LOS structures.

Table 17. Summary of mass spectrometry analysis of LOS outer core structures from seven *C. jejuni* septicaemia strains

Septicaemia strain	LOS class
34007	D
43983	A
44119	D
47439	Unknown*
47886	C
52472	Unknown*
53250	E

*The unknown LOS outer cores appear to be different from each other so there may be six different outer LOS cores expressed by seven different septicaemia strains

It is clear from these structural analyses that the *C. jejuni* strains associated with septicaemia in this study do not possess a unique LOS structure. Moreover, a high level of structural diversity in the LOS outer core exists between these isolates. Therefore, the LOS outer core structure does not form a genetic marker indicative of *C. jejuni* strains that cause septicaemia.

Unlike strains from all other clinical outcomes, *C. jejuni* strains associated with GBS and the 6 potentially non-pathogenic strains from beaches did form distinct clonal populations. The two GBS isolates also formed a unique clade, demonstrating that these two strains possessed a similar genomic make up distinct from strains associated with other clinical outcomes. Of the 131 CDSs that were identified as absent or divergent from both GBS strains using DNA microarray hybridization data, *cj1686*, or *top*A encoding DNA topoisomerase I, was identified as present in the remaining 89 isolates included in the study.

5.3.3 Identification of genetic markers associated with potentially non-pathogenic strains isolated from sand

Potentially non-pathogenic *C. jejuni* strains isolated from beaches also formed distinct clonal populations (Figure 15). Three beach isolates of MLST ST 177 formed a clonal population unique and distinct from a clonal population formed by three MLST ST 179 beach isolates. Analysis of the DNA microarray hybridisation data indicated that

CDS *cj0059*, putatively encoding FliY, the flagellar motor protein switch, was absent or divergent from beach strains of MLST ST 179 but present in every other strain in the study. PCR amplification of *cj0059* in strains 1791, 1792 and 1793 revealed weak products for strains 1791 and 1793. However, a strong band was observed for 1792 and the control (NCTC11168) (Figure 16).

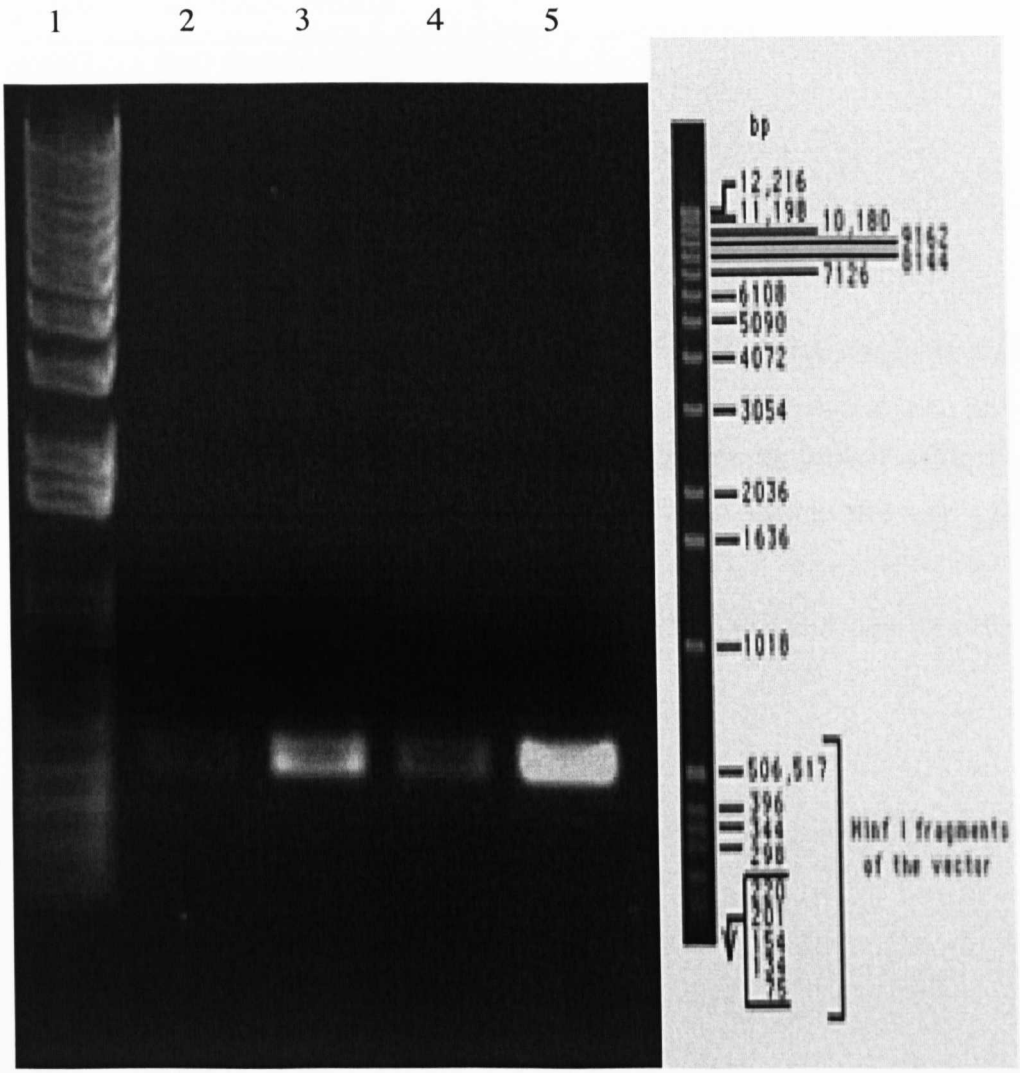

Figure 16. PCR analysis of *cj0059* in beach isolates. Lane 1 shows molecular weight marker. Lane 2 shows product of 1791, lane 3 shows product of 1792, lane 4 shows product of 1793 and lane 5 shows product of control strain NCTC11168.

Five CDSs were identified as uniquely absent or divergent in 3 beach strains of MLST ST 177 (Table 18).

Table 18. CDSs uniquely absent or divergent to beach isolates of MSLT ST 177

CDS	Annotated function
cj0145	unknown
cj0266	probable integral membrane protein
cj0887	*flaD* possible flagellin
cj1545	*m d a B* protein homologue
cj1674	unknown

PCR analysis of the five CDSs in the MLST ST 177 beach isolates indicated that *cj0145*, *cj0887* and *cj1545* were absent from each of the beach strains and present in the control strain NCTC11168 (Figures 17 and 18). However, weak products for *cj0266* were observed in all three strains (Figure 18) and *cj1674* was present in the three beach isolates and the control strain (Figure 19).

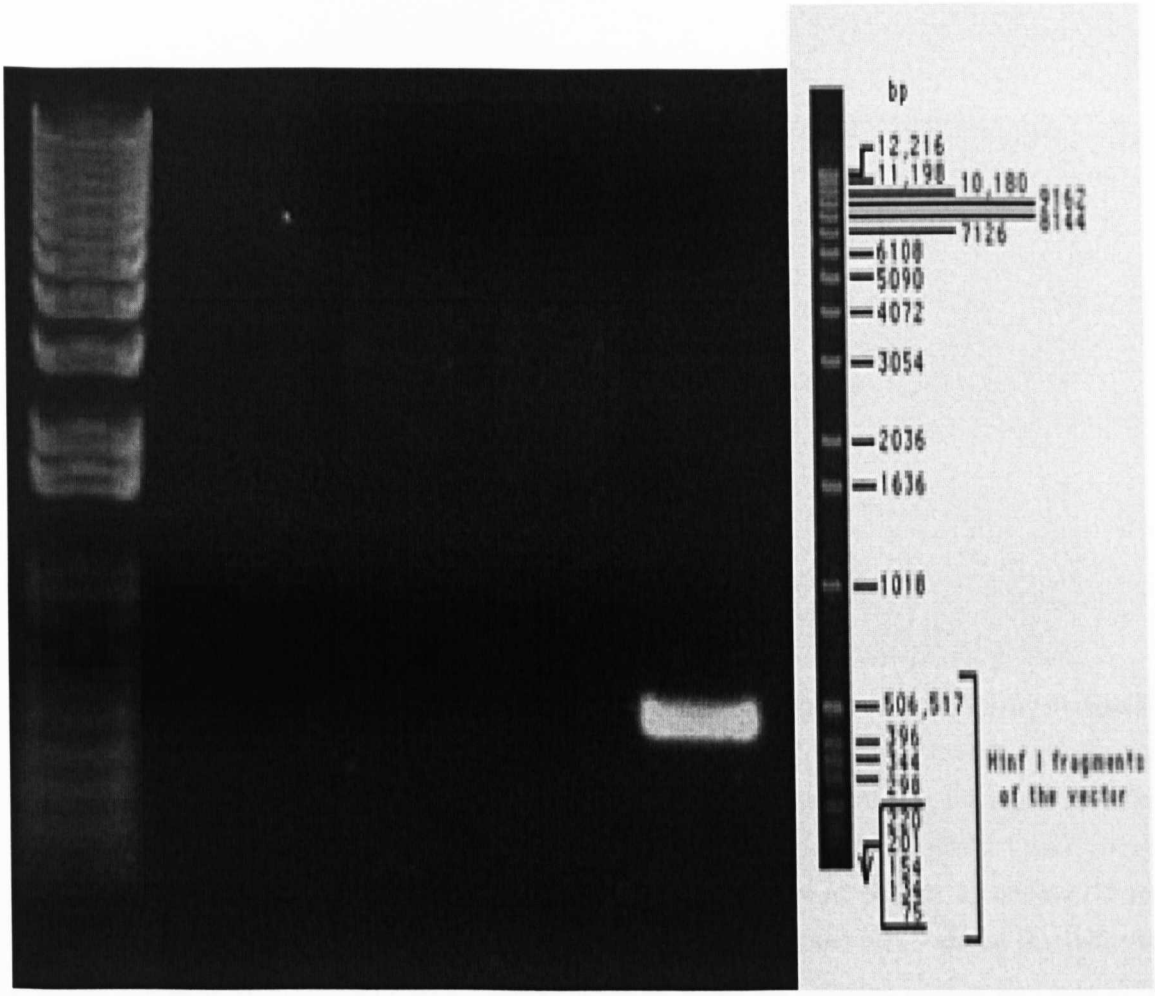

Figure 17. PCR analysis of *cj0145* in MLST ST 179 beach isolates. Lane 1 shows the molecular weight marker. Lane 2 shows strain 1791, lane 3 shows strain 1792, lane 4 shows strain 1793 and lane 5 shows control strain NCTC11168.

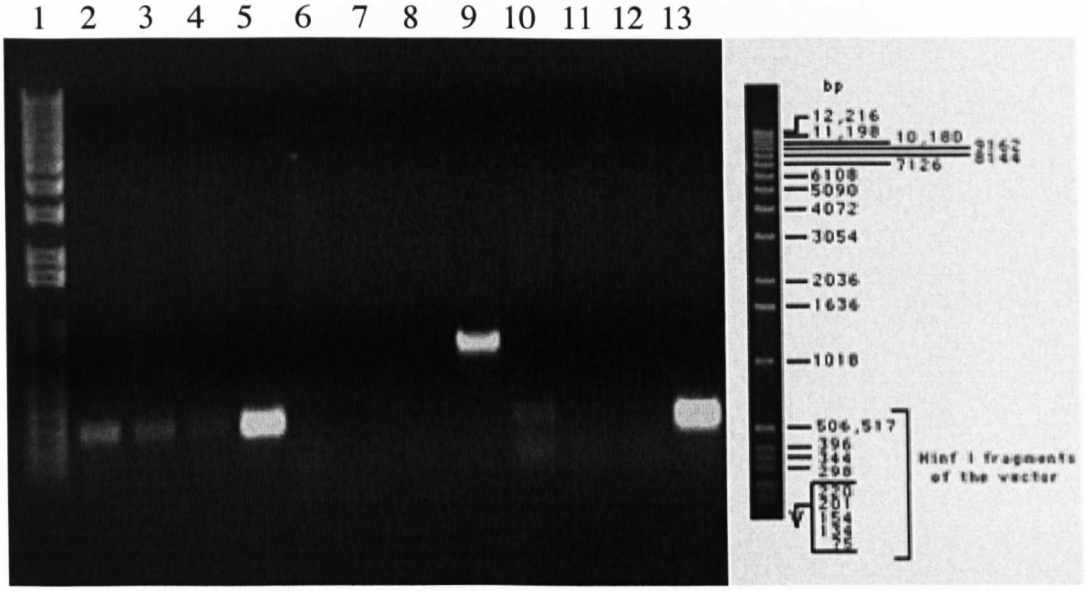

Figure 18. PCR analysis of *cj0266*, *cj0887* and *cj1545* in beach isolates. Lane 1 shows the molecular weight marker, lane 2 shows the observed product for strain 1771 for *cj0266*, lane 3 shows the observed product for strain 1772 for *cj0266,* lane 4 shows the observed product for strain 1773 for *cj0266* and lane 5 shows the observed product for control strain NCTC11168 for *cj0266*. Lane 6 shows the observed product for strain 1771 for *cj0887,* lane 7 shows the observed product for strain 1772 for *cj0887,* lane 8 shows the observed product for strain 1773 for *cj0887* and lane 9 shows the observed product for control strain NCTC11168 for *cj0887*. Lane 10 shows the observed product for strain 1771 for *cj1545,* lane 11 shows the observed product for strain 1772 for *cj1545,* lane 12 shows the observed product for strain 1773 for *cj1545* and lane 13 shows the observed product for control strain NCTC11168 for *cj1545*.

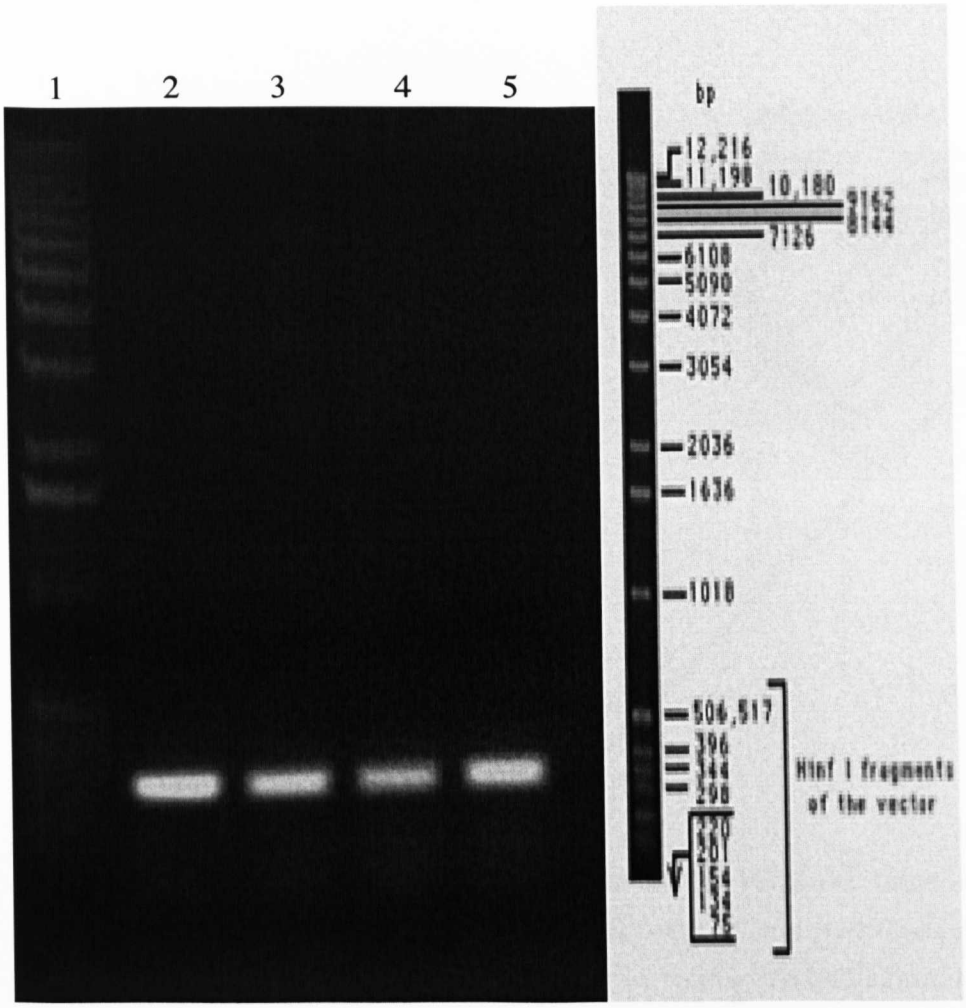

Figure 19. PCR analysis of *cj1674* in beach isolates. Lane 1 shows molecular weight marker, lane 2 shows strain 1771, lane 3 shows strain 1772, lane 4 shows strain 1773 and lane 5 shows control strain NCTC11168.

CDS *cj1365*, encoding a putative secreted serine protease, was identified as absent or divergent from all potentially non-pathogenic beach isolates, both those of MLST ST 177 and MLST ST 179. Although this CDS was identified as absent or divergent in several clinical isolates, *cj1365* was present in 88% of the clinical isolates included in the study (Figure 20). PCR analysis of beach isolates 1771, 1772, 1773, 1791, 1792 and 1793 indicated that *cj1365* is absent from all six environmental isolates (Figure 21). Further analysis of the DNA microarray hybridisation data for these isolates indicated that the CDSs flanking *cj1365* were present in each strain. Subsequent sequencing of this region in the six beach isolates clearly demonstrated that *cj1365* was absent whilst the flanking CDSs were present (Andrey Karlyshev, personal communication), validating the microarray data for this CDS.

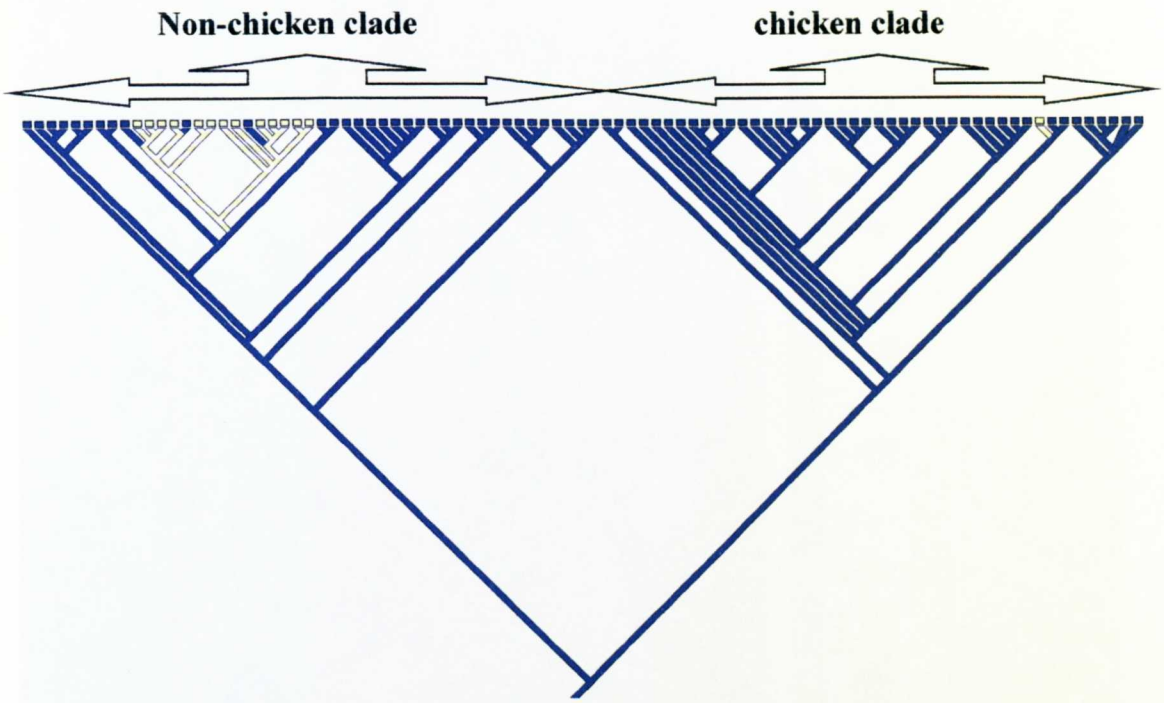

Figure 20. Distribution of *cj1365* among *C. jejuni* strains. Parsimony based gene analysis for determining the distribution of individual CDS *cj1365* throughout the phylogenetic tree. Strains in which *cj1365* are absent are coloured yellow. Strains in which *cj1365* are present are coloured blue. *Cj1365* is absent from beach isolates and 8 clinical isolates. In all other isolates *cj1365* is present.

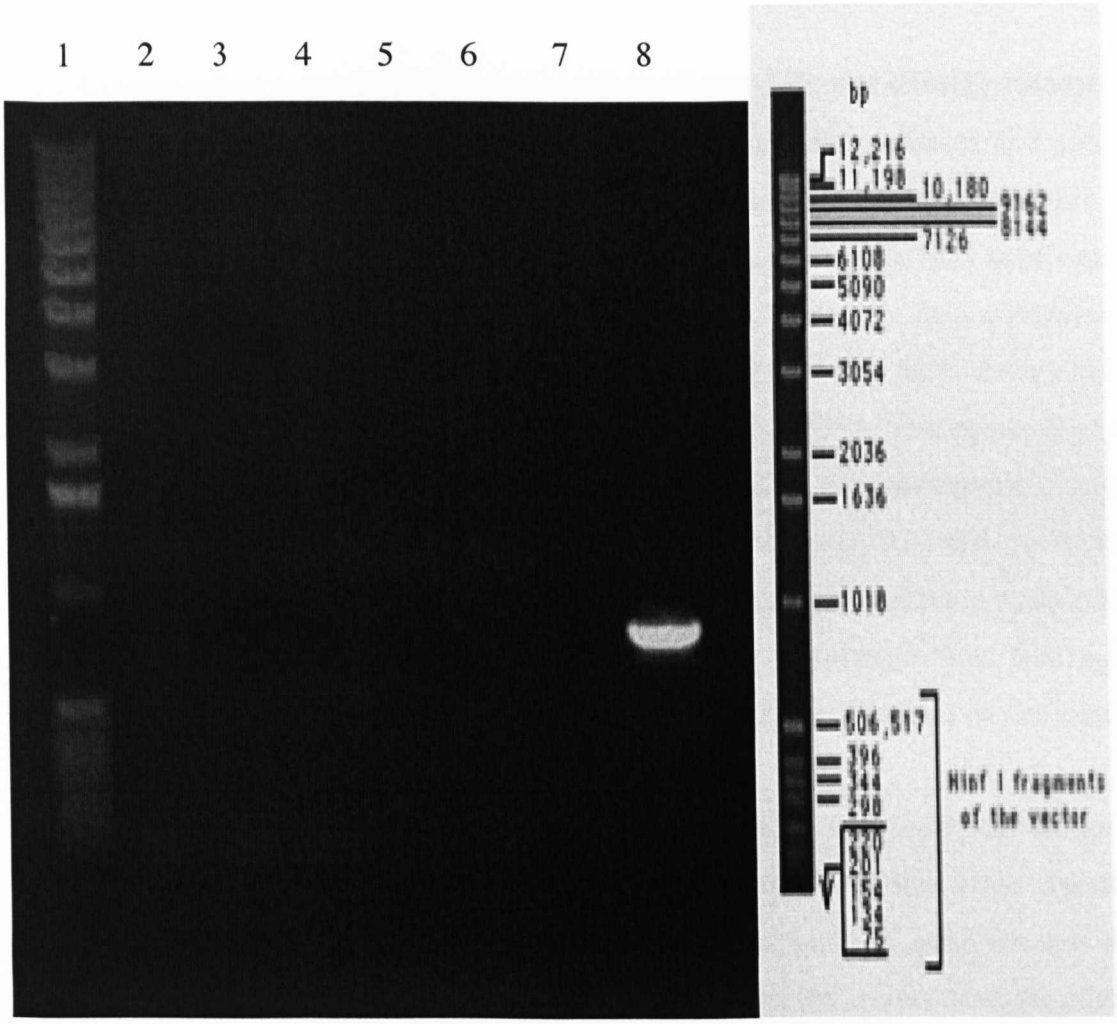

Figure 21. PCR analysis of *cj1365* in beach isolates. Lane 1 shows the molecular weight marker. Lane 2 shows 1771, lane 3 shows 1772, lane 4 shows 1773, lane 5 shows 1791, lane 6 shows 1792, lane 7 shows 1793 and lane 8 shows the control strain NCTC11168.

5.4 Discussion

5.4.1 Comparative genomics of *C. jejuni* strains from different clinical outcomes

The aim of this study was to identify potential virulence determinants and genetic markers associated with strains correlating to different clinical presentations through comparative genomics of a strain collection representative of the full spectrum of campylobacter disease outcome. A major reason for developing new comparative genomics studies for *C. jejuni* is the absence of an animal model that can be used to verify virulence determinants or genetic markers of disease. A consequence of the lack of animal infection model is that, to date, no definitive non-pathogenic *C. jejuni* strains have been characterised. Therefore, in this chapter potentially non-pathogenic *C. jejuni* isolates as well as isolates from different clinical outcomes were examined at the genome level in detail. This approach has the advantage over traditional approaches that use animal models because the isolates are studied in the natural model of infection, the human host.

In order to detect differences between the control strain NCTC11168, and the group of test strains representing a particular clinical outcome, CDSs that were absent or divergent in each strain were compared using Venn diagrams. A gene list was then generated comprising only genes absent or divergent in all the strains from the clinical outcome of interest. This list was then compared back against every other strain hybridised with the array to determine whether any absent or divergent CDSs were unique to that group of strains that may account for the observed differences in clinical outcome of *C. jejuni* infection. CDSs that were absent or divergent from all strains associated with distinct specific clinical outcome were identified, with the exception of strains associated with gastroenteritis. The strain used as a control in each of the competitive hybridisations was NCTC11168, a clinical strain originally isolated in 1977 from a patient with gastroenteritis in Worcester. It may therefore be difficult to tease out subtle genetic differences between strains associated with gastroenteritis using this control strain. The CDSs that were identified as absent or divergent in other clinical outcomes were mainly from the loci encoding surface antigens, in particular the CPS and LOS biosynthesis loci. CPSs and LOS are highly antigenic surface antigens and it is hypothesised that structural variation of these surface glycolipids may play a role in evasion of host immune responses and high levels of variability in these loci has been previously documented in both clinical and animal isolates (Dorrell *et al.*, 2001; Pearson *et al.*, 2003). None of the CDSs

encoding surface antigens formed unique genetic markers for strains associated with different clinical outcomes. However, the LOS biosynthesis locus was identified as a highly variable region in six out of seven septicaemia strains, with CDSs *cj1136-cj1139* and *cj1141-cj1149* identified as absent or divergent in five of the strains. Although this locus was highly variable amongst *C. jejuni* strains from other clinical outcomes the consistency of variable genes found in the majority of strains from this unusual clinical outcome was interesting. The LOS loci of the septicaemia strains were further investigated by mass spectrometry analysis to establish whether the consistent variation in LOS genes found in these strains correlated with a unique LOS outer core. LOS structural analysis revealed that the seven strains possessed four different LOS outer core structures that had been identified in strains from previous studies (M. Gilbert, personal communication). A further two LOS outer core structures that were previously unidentified were also discovered. Thus, of the seven strains analysed six different LOS outer cores were expressed. This study has shown that strains associated with septicaemia do not have a unique LOS outer core but instead display highly variable outer core structures. These data also indicate that there may be many more LOS classes as yet undescribed.

Only strains associated with the post *C. jejuni* infection sequela, GBS, demonstrated a potential genetic marker. *Cj1686* or *top*A encoding DNA topoisomerase 1 was uniquely absent or divergent in the two strains associated with GBS by microarray analysis and no product was observed for *cj1686* in either GBS strain using PCR. However, microarray analysis indicated that *cj1686* was present in the remaining 89 isolates included in the study that weren't associated with GBS. Topoisomerases are enzymes that introduce additional turns in the DNA alpha helix causing the DNA to supercoil making it more compact. The relevance of this single gene in the development of GBS is difficult to speculate on and further GBS associated strains must be investigated. One previous DNA microarray study investigating the genomic composition of twentysix *C. jejuni* strains compared GBS isolates with strains associated with enteritis and did not identify any regions of genetic diversity specific to GBS strains (Leonard *et al.*, 2004). It would be interesting to analyse the DNA microarray hybridisation data from this study using the analysis method employed in this project to determine whether *cj1686* was absent or divergent from GBS associated strains, or more simply to screen the GBS isolates by PCR to determine whether *cj1686* was present. Similarly, Engberg *et al* investigated a worldwide selected population of non-HS19 *C. jejuni* strains associated

with GBS and uncomplicated gastroenteritis (Engberg *et al.*, 2001). The authors were looking for an epidemiologic marker for GBS associated strains. None was found. Such a collection of strains would be ideal to investigate with the *C. jejuni* DNA microarray to determine a genetic basis for their clonality, as well as a potential genetic marker for GBS associated strains.

In conclusion, the analysis of strains from different clinical outcomes using DNA microarrays is limited by the reporter elements represented on the microarray. CDSs that are absent or divergent from a particular group of strains in relation to the control strain may be identified. However, the presence of additional CDSs that may correlate with a particular disease outcome cannot be ruled out. For instance, strains associated with septicaemia may possess "hypervirulent" additional genes not present in NCTC11168. Furthermore, host factors rather than genomic differences between *C. jejuni* strains may be responsible for the differences in the predominant clinical presentations observed in patients.

5.4.2 Identification of CDSs and a potential virulence determinant, *cj1365*, absent from potentially non-pathogenic sand isolates

Six potentially non-pathogenic beach strains isolated from sand on Blackpool beach were included in this study with the aim of identifying putative virulence genes. These six strains each possess MLST types previously undetected in human and chicken isolates and so it was hypothesised that these strains may be potentially non-pathogenic. The beach isolates also formed distinct clades following phylogenomic analysis, indicating that they are distinct from other clinical strains included in the study.

Microarray hybridisation data identified only one CDS, *cj0059* encoding FliY, the putative flagellar motor switch protein, as uniquely absent or divergent in beach isolates of MLST ST179. Using PCR, weak products of the expected size were observed in strains 1791 and 1793 and a strong product was observed in strain 1792. Unfortunately it was not possible to check the motility of these strains as DNA only was provided for these studies. Five genes were identified as uniquely absent or divergent in beach isolates of MLST ST179 (*cj0145*, *cj0266*, *cj0887*, *cj1545* and *cj1674*). Using PCR, weak products were observed for *cj0266*, encoding a probable integral membrane protein, and products were observed for *cj1674* encoding a hypothetical protein. Microarray data is based on the intensity of signal for a

particular CDS in both the control and the test strain. The signal intensity in the test strain may be lower than the control if the portion of the CDS represented on the array is variable or the size of the fragment is small compared to many other reporter elements. Data from such reporter elements may result in the categorisation of a CDS as absent or divergent when it is in fact present in the test strain. However, nucleotide variation in the CDS in the test strain would also result in lower signal intensity in the test channel than that in the control channel. The use of PCR screening to verify microarray hybridisation data for key CDSs is therefore important.

Cj0145 encoding a hypothetical protein, *cj0887* encoding FlaD and *cj1545* encoding a MdaB protein homologue were absent in 1771, 1772 and 1773 by PCR. MdaB was recently characterized as a novel potential antioxidant protein in *H. pylori*. Phenotypic assays of the *H. pylori mdaB* mutant indicated that it was more sensitive to H_2O_2, organic hydroperoxides, and the superoxide-generating agent paraquat and 10% oxygen for growth than the wild type. Exposure of the mutant strain to air for 8 h resulted in the recovery of no viable cells whereas the wild-type strain survived more than 10 h of air exposure. The oxidative stress sensitivity of the *mdaB* mutant resulted in a deficiency in the ability of the mutant to colonize mouse stomachs (Wang and Maier, 2004). The absence of *cj1545 (mdaB)* suggests that these beach strains may be less able to survive in the human gut than strains expressing the potential antioxidant.

A probable secreted serine protease, encoded by CDS *cj1365* was also identified as absent or divergent in all six of the potentially non-pathogenic 'beach' isolates. PCR amplification of *cj1365* indicated that this gene was absent in the six isolates and subsequent sequencing of this region revealed that *cj1365* was absent from each of the beach isolates whilst the flanking genes were present in each (Andrey karlyshev, personal communication). It is unknown whether this gene was 'lost' in beach isolates or gained in other strains. Secreted serine proteases are well- documented virulence factors in pathogens, serving multiple functions including the cleavage of human factor V that aggravates the haemorrhagic colitis characteristic of EHEC infections (Dutta *et al.*, 2002). The cleavage of factor V is widespread amongst bacterial serine proteases secreted by pathogens that cause bloody diarrhoea (Dutta *et al.*, 2002). *C. jejuni* strains isolated from humans have been shown to produce cytotoxic effects include the rounding of CHO and HeLA cells indicating the presence of a toxin that inhibits actin filament formation (Lee *et al.*, 2000). EHEC secretes a serine protease, EspP, which is thought to act as a cytotoxin, disrupting the actin network when

applied to Vero cells (Brunder *et al.*, 1997). Host functional proteins as well as proteins involved in defence are often targeted by bacterial serine proteases. For example, Enteropathogenic *E. coli* (EPEC) produces EspC (Stein *et al.*, 1996). This cleaves both pepsin and human coagulation factor V, exacerbating haemorrhagic colitis. EspP is secreted by enterohaemorrhagic *E. coli* (EHEC). EspP may act as a cytotoxin, disrupting the actin network when applied to Vero cells (Brunder *et al.*, 1997). SepA has a role in intestinal inflammation and tissue invasion in Shigellosis (Benjelloun-Touimi *et al.*, 1995) and Hbp cleaves haemaglobin, playing a role in abscess formation by *Bacteroides fragilis* (Otto *et al.*, 2002). Many pathogenic bacteria secrete proteases but no common role in virulence has been determined. Cytotoxin production by some human *C. jejuni* isolates has been demonstrated and although the nature of the toxin is unknown, detection in human isolates may suggest relevance to clinical disease. However, the secretion pathway utilised by the putative serine protease is unknown and requires further investigation. The absence of *cj1365* from the potentially non-pathogenic strains is noteworthy as this CDS is present in 88% of the clinical isolates hybridised with the microarray. Thus the putative secreted serine protease may form part of the arsenal of virulence factors found in pathogenic *C. jejuni*. Over 15% of the CDSs that were absent or divergent in the six potentially non-pathogenic beach isolates were of unknown function. CDSs of unknown function that have little or no homology to CDSs found in other pathogens are interesting candidates for further investigation as the mechanism of pathogenesis of *C. jejuni* bears little in common to that of well characterized enteropathogens such as *E. coli* and *Salmonella*. We can speculate that the potentially non-pathogenic beach strains may be poorly adapted to survive and cause disease in the human host but this genetic make up may confer a selective advantage for survival in this environmental niche.

6.0 Comparative phylogenomics of *C. jejuni* strains from different ecological niches

6.1 Introduction

6.1.1 Aims

This chapter describes the development of a robust and improved method for the analysis of *C. jejuni* phylogeny to establish whether strains from the same ecological niche are related. Furthermore, the aim was to use this method to identify genetic markers that may distinguish strains from different ecological niches. We have investigated the relationships of *C. jejuni* strains through comparative phylogenomics. This method combines the power of DNA microarrays and robust statistical algorithms to model phylogeny facilitating the inferral of relationships between strains from specific sources. Using this novel approach we can determine whether two or more strains with a common ancestor (herein referred to as a clade) were isolated from the same host and identify the genomic relatedness of strains isolated from different hosts. An understanding of genetic differences between *C. jejuni* strains from different ecological niches should allow the identification of improved epidemiological markers and in the long term, the development of rational approaches to reduce *C. jejuni* in the food chain.

6.1.2 Sources of human *C. jejuni* disease

The consumption of undercooked poultry is the most commonly documented source of human *C. jejuni* infection. This is partly a result of the citation of chicken as the principal risk factor in large case-control studies and the fact that *C. jejuni* is a gut commensal of avians and is therefore easily transferred onto the skin of chicken carcasses during food processing (Friedman *et al.*, 2004; Rodrigues *et al.*, 2001). The presence of *C. jejuni* on raw chicken may result in human disease if the chicken is undercooked. Furthermore, cross-contamination from the chicken to other foods for human consumption may also lead to human *C. jejuni* infection. However, a variety of other sources including cattle, water, milk and wild birds also act as a reservoir (Broman *et al.*, 2002; Engberg *et al.*, 1998; Inglis *et al.*, 2004; Wood *et al.*, 1992). This may account for the isolation of *C. jejuni* from diverse environmental samples. Little is known regarding the contribution of non-chicken sources to the burden of human *C. jejuni* infection and this, in turn, has hindered effective control strategies to

reduce *Campylobacter* in the food chain. Traditional methods for typing were introduced to facilitate epidemiological investigations, in particular the tracing of routes and sources of *C. jejuni* human infection. Phenotypic typing methods such as Penner serotyping are largely based on variation in the surface expressed capsular polysaccharide (CPS) antigen (Karlyshev *et al.*, 2000). However, variation in surface antigens such as CPS and LOS do not correlate with clinical outcome, source or pathogenicity. This is partly because *C. jejuni* strains are genetically diverse (Dingle *et al.*, 2001a), with loci encoding surface antigens exhibiting particularly high levels of variation (Dorrell *et al.*, 2001; Gilbert *et al.*, 2002; Karlyshev *et al.*, 2005). Thus, since current typing methods have generally been unable to identify strains with phenotypic characteristics associated with different ecological habitats, the proportion of human disease attributable to different sources of infection is unknown.

6.2 Results – Comparative phylogenomics

6.2.1 Comparative genomics of *C. jejuni* from different animal hosts

C. jejuni strains were selected for comparative phylogenomics based on the source from which the strains were isolated. Partial funding for this study was provided by the Food Standards Agency (FSA) and so UK isolates only (91) were included represented by 17 *C. jejuni* chicken isolates (19%) and five ovine and bovine isolates (5%). In addition, *C. jejuni* 6 environmental strains (7%) isolated from sand on Blackpool beach (Chapter 5.1.2) were also chosen with the 63 *C. jejuni* clinical isolates mentioned in the previous chapter. The chicken isolates were collected over a ten year period (1991-2001) from three different points in the food chain; broiler flocks, abattoirs and supermarket chicken portions. Using traditional phenotypic typing methods the 17 chicken isolates were represented by five different serotypes (HS2, HS5, HS27, HS44 and HS50) and 11 different phage types (PT1, PT2, PT5, PT15, PT19, PT25, PT34, PT35, PT36, PT44 and PT59). Three of the ovine/ bovine strains were isolated from animals (all HS50) and two from supermarket ox liver portions. All isolates were competitively hybridised in duplicate with the microarray using the sequenced strain NCTC11168 as a control as described in the previous chapter.

The distribution of the strains from different ecological niches within the tree comprised 94% (16/17) of strains from chicken sources falling within the "chicken clade" (a single isolate was not found in this clade) (Figure 22). Strains from ovine/

bovine and sand (environmental, beach) formed distinct clades that were unequivocally supported within the "non-chicken clade". Both clades contained clinical isolates of which 41% (26/63) were found in the "chicken clade" and 59% (37/63) in the "non-chicken clade".

6.2.2 Identification of CDSs differenting the "chicken clade" from the "non chicken clade"

CDSs differentiating the strains in the "chicken clade" from those strains in the "non-chicken clade" were identified using parsimony-based methods implemented through MacClade 4 software (Chapter 2.6.3). Similarly, CDSs contributing to ovine, bovine and environmental sub-clades within clade A were identified. Table 19 shows the CDSs contributing to the formation of the "chicken clade" and table 20 shows the CDSs contributing to the strains forming the "non-chicken clade". These tables indicate whether the CDSs are unique to the clade or if the genotype is shared with strains from other clades and indicate whether these CDSs are randomly dispersed or are associated with other flanking CDSs. Additional data including the number of reconstructed evolutionary changes (RC) of a CDS in a particular branch (this is the value from which the length of the branch is calculated). For example, a value of 1.0 indicates that a CDS is unique to a clade whereas a value of 0.0 indicates that a CDS does not contribute to the formation of a clade. Whether the CDS is absent or divergent and the annotated function of the CDSs are also shown.

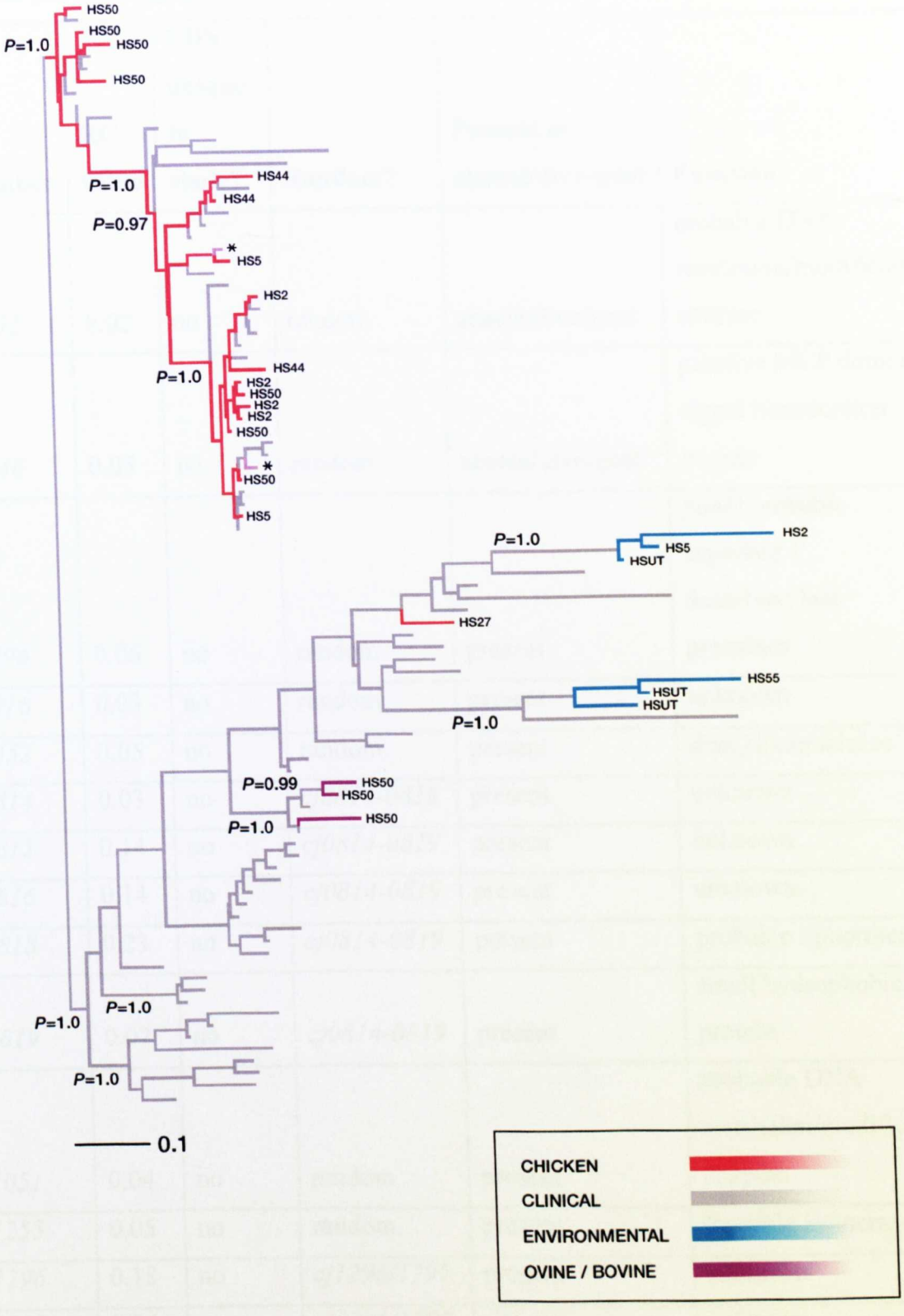

Figure 22. As Figure 15 but strains labelled according to animal/environmental source.

HS = serotype * = ovine / bovine strain isolated from supermarket meat portion

Table 19. CDSs specific to "chicken clade" strains

cj number	RC value	CDS unique to clade?	Random?	Present or absent/divergent	Function
0032	0.02	no	random	absent/divergent	probable DNA restriction/modification enzyme
0246	0.05	no	random	absent/divergent	putative MCP domain signal transduction protein
0296	0.06	no	random	present	*panD* probable aspartate 1 decarboxylase precursor
0416	0.03	no	random	present	unknown
0452	0.05	no	random	present	*dnaQ* exonuclease
0814	0.03	no	*cj0814-0819*	present	unknown
0815	0.14	no	*cj0814-0819*	present	unknown
0816	0.14	no	*cj0814-0819*	present	unknown
0818	0.23	no	*cj0814-0819*	present	probable lipoprotein
0819	0.07	no	*cj0814-0819*	present	small hydrophobic protein
1051	0.04	no	random	present	probable DNA restriction/modification enzyme
1255	0.05	no	random	present	Possible isomerase
1296	0.18	no	*cj1296/1297*	present	unknown
1297	0.14	no	*cj1296/1297*	present	unknown
1321	0.18	no	*cj1321-1339*	present	putative acetyl transferase
1322	0.15	no	*cj1321-1339*	present	probable hydroxyacyl

					dehydrogenases
1323	0.06	no	*cj1321-1339*	present	probable hydroxyacyl dehydrogenases
1324	0.18	no	*cj1321-1339*	present	similarity to *wbpG*, associated with LPS biosynthesis
1325	0.08	no	*cj1321-1339*	present	similar to *cj1330* involved in the synthesis of pseudaminic acid
1338	0.1	no	*cj1321-1339*	present	*flaB*
1339	0.07	no	*cj1321-1339*	present	*flaA*
1376	0.12	no	random	present	putative periplasmic protein
1442	0.08	no	random	present	unknown
1561	0.07	no	*cj1561-1562*	absent/divergent	putative transcriptional regulator
1562	0.03	no	*cj1561-1562*	absent/divergent	unknown
1677	0.11	no	*cj1677/1679*	present	probable lipoprotein
1679	0.12	no	*cj1677/1679*	present	unknown

Table 20. CDSs specific to "non-chicken clade" strains

cj number	RC value	CDS unique to clade?	Random?	Present or absent/divergent	Function
0296	0.06	no	random	absent/divergent	*panD* probable aspartate 1 decarboxylase precursor
0815	0.14	no	*cj0815-0819*	absent/divergent	unknown
0816	0.14	no	*cj0815-0819*	absent/divergent	unknown
0818	0.23	no	*cj0815-0819*	absent/divergent	probable lipoprotein
0819	0.07	no	*cj0815-0819*	absent/divergent	small hydrophobic protein
1321	0.18	no	*cj1321-1339*	absent/divergent	putative acetyl transferase
1322	0.15	no	*cj1321-1339*	absent/divergent	unknown
1323	0.06	no	*cj1321-1339*	absent/divergent	unknown
1324	0.18	no	*cj1321-1339*	absent/divergent	unknown
1325	0.08	no	*cj1321-1339*	absent/divergent	unknown
1338	0.1	no	*cj1321-1339*	absent/divergent	*flaB*
1339	0.07	no	*cj1321-1339*	absent/divergent	*flaA*
1442	0.08	no	random	absent/divergent	unknown
1561	0.07	no	*cj1561-1562*	absent/divergent	putative transcriptional regulator
1562	0.03	no	*cj1561-1562*	absent/divergent	unknown
1677	0.11	no	*cj1677-1679*	present	probable lipoprotein
1679	0.12	no	*cj1677-1679*	present	unknown

These data indicate that the most striking gene clusters absent in the "non-chicken clade" are CDSs *cj1321-cj1325* and *cj1338* and *cj1339*. The genes are present in strains in the "chicken clade". Thus the loss or divergence of this locus is a major genetic region distinguishing the majority of strains from chicken sources from those of other animal and environmental sources. *Cj1326* is closely associated with this genetic island and so this CDS was investigated. The distribution of the differential CDS *cj1321* is shown in Figure 23. This graphical representation clearly shows how *cj1321* is distributed in the two clades.

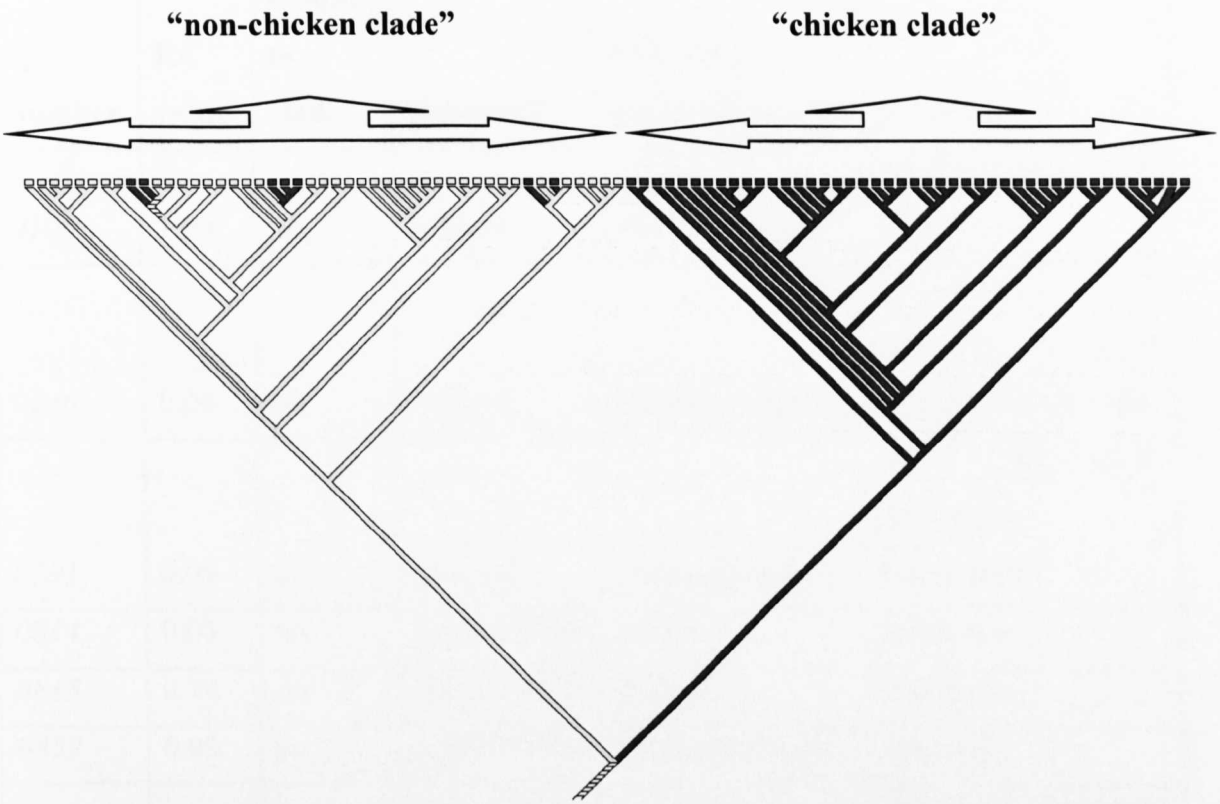

Figure 23. Distribution of *cj1321* among *C. jejuni* strains. Parsimony based gene analysis for determining the distribution of individual CDS *cj1321* throughout the phylogenetic tree. Strains in which *cj1321* are absent are coloured yellow, strains in which *cj1321* are present are coloured blue. Strains in the "chicken clade" all contain *cj1321*. *Cj1321* is absent in strains in the "non chicken" clade with 7 exceptions (5 beach isolates and 2 clinical isolates).

6.2.3 Identification of CDSs distinguishing strains from ovine and bovine animal sources

Three ovine and bovine strains formed a clade with two clinical isolates. The remaining two bovine strains that were isolated from ox liver portions from a supermarket did not fall in this clade. CDSs that were specific to this clade comprising ovine/bovine strains isolated from animals were identified and are shown in Table 21.

Table 21. CDSs specific to ovine/bovine clade

cj number	RC value	CDS unique to clade?	Random?	Present or absent/divergent	Function
0105	1.00	no	random	absent/divergent	*atpA* probable ATP synthase
0246	0.04	no	random	absent/divergent	putative MCP domain signal transduction protein
0291	0.09	no	random	absent/divergent	*glpT* glycerol-3-phosphate transporter
0814	0.03	no	*cj0814-0819*	present	unknown
0815	0.14	no	*cj0814-0819*	present	unknown
0859	0.05	no	*cj0859-0860*	absent/divergent	unknown
0860	0.05	no	*cj0859-0860*	absent/divergent	putative integral membrane protein
0969	0.04	no	*cj0969-0972*	present	pseudogene
0970	0.13	no	*cj0969-0972*	present	unknown
0972	0.05	no	*cj0969-0972*	present	unknown
0987	0.18	no	random	absent/divergent	unknown
1159	0.05	no	*cj1159-1164*	absent/divergent	small hydrophobic protein
1160	0.04	no	*cj1159-1164*	absent/divergent	small hydrophobic

					protein
1164	0.06	no	*cj1159-1164*	absent/divergent	unknown
1296	0.18	no	*cj1296/1301*	present	unknown
1297	0.14	no	*cj1296/1301*	present	unknown
1300	0.08	no	*cj1296/1301*	present	unknown
1301	0.08	no	*cj1296/1301*	present	unknown
1376	0.12	no	random	lost	putative periplasmic protein
1427	0.1	no	random	present	putative sugar nucleotide epomerase/ dehydratase
1442	0.08	no	random	present	unknown
1549	0.12	no	*cj1549-1560*	present	putative type 1 restriction enzyme R protein
1550	0.14	no	*cj1549-1560*	present	putative ATP/GTP binding protein
1551	0.14	no	*cj1549-1560*	present	putative type 1 restriction enzyme S protein
1552	0.18	no	*cj1549-1560*	present	unknown
1553	0.12	no	*cj1549-1560*	present	putative type 1 restriction enzyme M protein
1555	0.12	no	*cj1549-1560*	present	unknown
1556	0.1	no	*cj1549-1560*	present	unknown
1558	0.05	no	*cj1549-1560*	present	putative membrane protein
1560	0.09	no	*cj1549-1560*	present	putative membrane protein
1585	0.17	no	random	present	possible oxidoreductase

Nine contiguous CDSs, *cj1549-cj1553*, *cj1555-cj1556* and *cj1558-cj1560* encoding probable restriction enzymes, putative periplasmic proteins and proteins of unknown function form a large proportion of the CDSs responsible for the differentiation of the ovine and bovine strains in this clade from other strains. However, these CDSs are also present in strains from other sources. CDS *cj0105* (*atpA*), encoding a probable ATP synthase, was identified as absent or divergent from strains found in the ovine/ bovine clade yet this CDS was present in every other isolate (Figure 24).

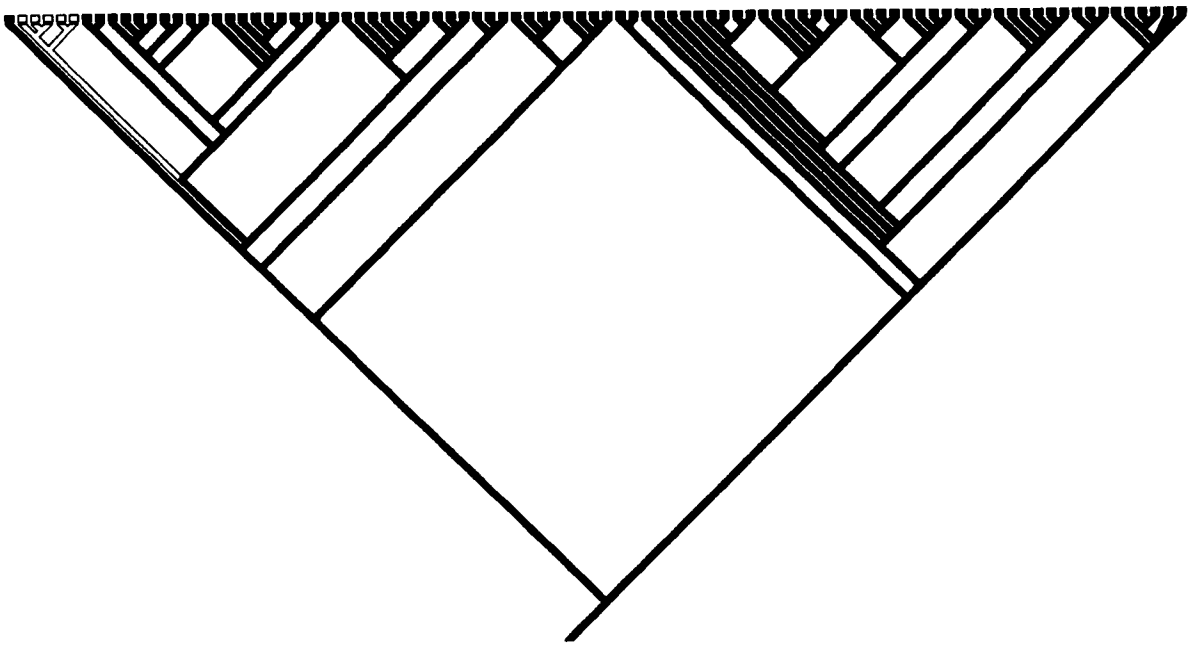

Figure 24. Distribution of *cj0105* among *C. jejuni* strains. Parsimony based gene analysis for determining the distribution of individual CDS *cj0105* throughout the phylogenetic tree. Strains in which *cj0105* are absent are coloured yellow, strains in which *cj0105* are present are coloured blue. *Cj0105* is absent from strains in the "ovine / bovine clade". All other isolates contain *cj0105*.

6.3 Results – Validation of microarray data

6.3.1 PCR analysis of CDSs differentiating "chicken clade" strains from "non chicken clade" strains

CDSs *cj1321-cj1326* were screened in strain NCTC11168 as a positive control as well as chicken strains from the "chicken clade" (Figures 25 and 26). In addition uncharacterised chicken isolates not previously investigated by microarray analysis were screened for *cj1321-cj1326* (Figures 27 and 28) as were clinical strains, ovine/bovine and beach strain from the "non chicken clade" (Figures 29 and 30). Moreover, the presence of *cj1321-cj1326* in the single chicken isolate found in the "non chicken clade" was screened for by PCR. Tables 22 to 26 contain the information for each gel lane.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

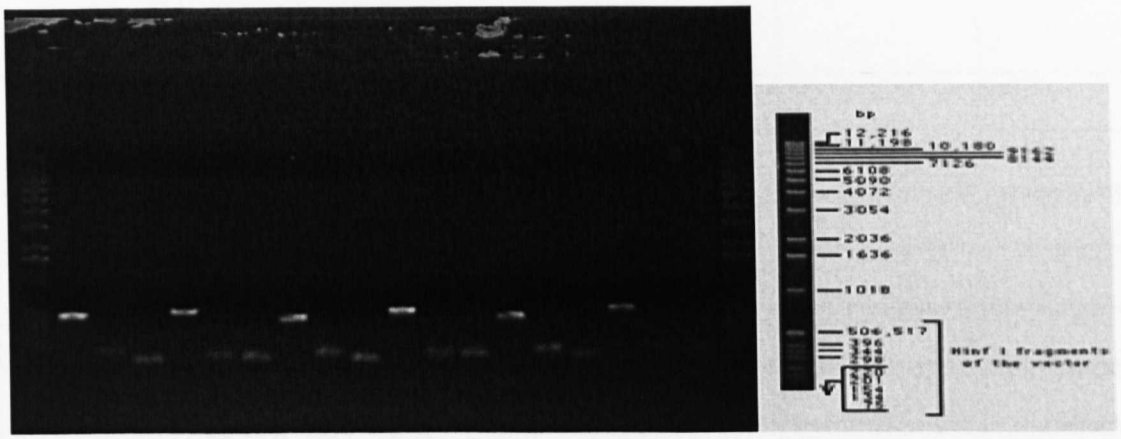

Figure 25. Detection of *cj1321-cj1326* in strain NCTC11168 and chicken isolates 11919 and 11818. Lane 1 contains molecular weight marker. Lanes 2 to 7 show the observed PCR products from DNA isolated from NCTC11168 amplified with primers for *cj1321* to *cj1326* inclusive. Lanes 8 to 13 show the observed PCR products from DNA isolated from chicken isolate 11919 amplified with primers for *cj1321* to *cj1326* inclusive. Lanes 14 to 19 show the observed PCR products from DNA isolated from chicken isolate 11818 amplified with primers for *cj1321- cj1326* inclusive. Table 22 shows the information for each gel lane.

Table 22. Key to Figure 25

Lane	Strain	Source	CDS	Expected	Observed
2	NCTC11168	sequenced strain	*cj1321*	+	+
3	NCTC11168	sequenced strain	*cj1322*	+	+
4	NCTC11168	sequenced strain	*cj1323*	+	+
5	NCTC11168	sequenced strain	*cj1324*	+	+
6	NCTC11168	sequenced strain	*cj1325*	+	+
7	NCTC11168	sequenced strain	*cj1326*	+	+
8	11919	chicken	*cj1321*	+	+
9	11919	chicken	*cj1322*	+	+
10	11919	chicken	*cj1323*	+	+
11	11919	chicken	*cj1324*	+	+
12	11919	chicken	*cj1325*	+	+
13	11919	chicken	*cj1326*	+	+
14	11818	chicken	*cj1321*	+	+
15	11818	chicken	*cj1322*	+	+
16	11818	chicken	*cj1323*	+	+
17	11818	chicken	*cj1324*	+	+
18	11818	chicken	*cj1325*	+	+
19	11818	chicken	*cj1326*	+	+

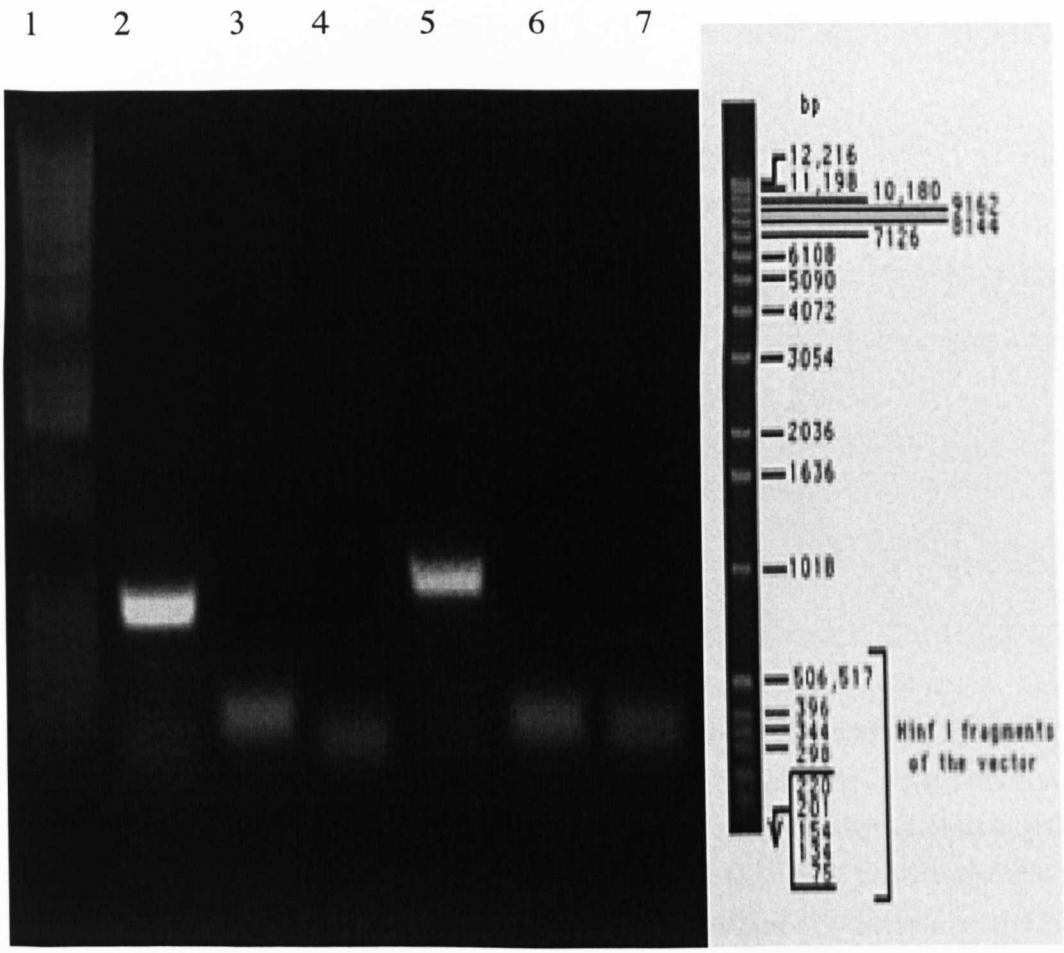

Figure 26. Detection of *cj1321-cj1326* in chicken strain 13411. Lane 1 = molecular weight marker, lane 2 = PCR product for *cj1321*, lane 3 = PCR product for *cj1322*, lane 4 = PCR product for *cj1323*, lane 5 = PCR product for *cj1324*, lane 6 = PCR product for *cj1325*, lane 7 = PCR product for *cj1326*.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

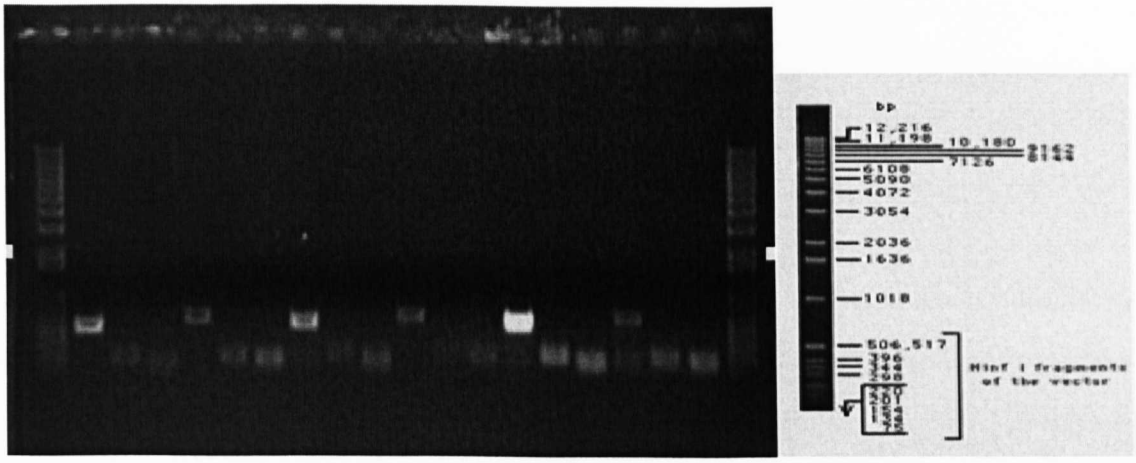

Figure 27. Detection of *cj1321-cj1326* in uncharacterised chicken strains B, 17M and A. Lanes 1 and 20 contain molecular weight marker. Lanes 2 to 7 show the observed PCR products from DNA isolated from uncharacterised chicken isolate B amplified with primers for *cj1321* to *cj1326* inclusive. Lanes 8 to 13 show the observed PCR products from DNA isolated from uncharacterised chicken isolate 17M amplified with primers for *cj1321* to *cj1326* inclusive. Lanes 14 to 19 show the observed PCR products from DNA isolated from uncharacterised chicken isolate A amplified with primers for *cj1321- cj1326* inclusive. Table 23 contains the information for each gel lane.

Table 23. Key for Figure 27

Lane	Strain	Source	CDS	Expected	Observed
2	B	chicken	*cj1321*	+	+
3	B	chicken	*cj1322*	+	+
4	B	chicken	*cj1323*	+	+
5	B	chicken	*cj1324*	+	+
6	B	chicken	*cj1325*	+	+
7	B	chicken	*cj1326*	+	+
8	17M	chicken	*cj1321*	+	+
9	17M	chicken	*cj1322*	+	+
10	17M	chicken	*cj1323*	+	+
11	17M	chicken	*cj1324*	+	+
12	17M	chicken	*cj1325*	+	+
13	17M	chicken	*cj1326*	+	+
14	A	chicken	*cj1321*	+	+
15	A	chicken	*cj1322*	+	+
16	A	chicken	*cj1323*	+	+
17	A	chicken	*cj1324*	+	+
18	A	chicken	*cj1325*	+	+
19	A	chicken	*cj1326*	+	+

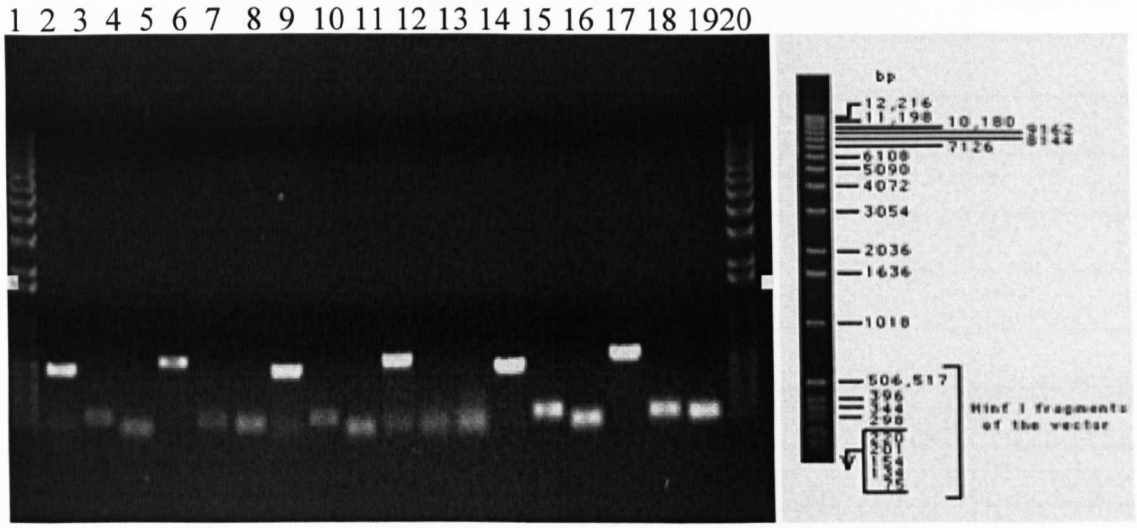

Figure 28. Detection of *cj1321-cj1326* in uncharacterised chicken strains 3852, C and D.

Lanes 1 and 20 contain molecular weight marker. Lanes 2 to 7 show the observed PCR products from DNA isolated from uncharacterised chicken isolate 3852 amplified with primers for *cj1321* to *cj1326* inclusive. Lanes 8 to 13 show the observed PCR products from DNA isolated from uncharacterised chicken isolate C amplified with primers for *cj1321* to *cj1326* inclusive. Lanes 14 to 19 show the observed PCR products from DNA isolated from uncharacterised chicken isolate D amplified with primers for *cj1321- cj1326* inclusive. Table 24 contains the information for each gel lane.

Table 24. Key for Figure 28

Lane	Strain	Source	CDS	Expected	Observed
1	3852	chicken	*cj1321*	+	+
2	3852	chicken	*cj1322*	+	+
3	3852	chicken	*cj1323*	+	+
4	3852	chicken	*cj1324*	+	+
5	3852	chicken	*cj1325*	+	+
6	3852	chicken	*cj1326*	+	+
7	C	chicken	*cj1321*	+	+
8	C	chicken	*cj1322*	+	+
9	C	chicken	*cj1323*	+	+
10	C	chicken	*cj1324*	+	+
11	C	chicken	*cj1325*	+	+
12	C	chicken	*cj1326*	+	+
13	D	chicken	*cj1321*	+	+
14	D	chicken	*cj1322*	+	+
15	D	chicken	*cj1323*	+	+
16	D	chicken	*cj1324*	+	+
17	D	chicken	*cj1325*	+	+
18	D	chicken	*cj1326*	+	+

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

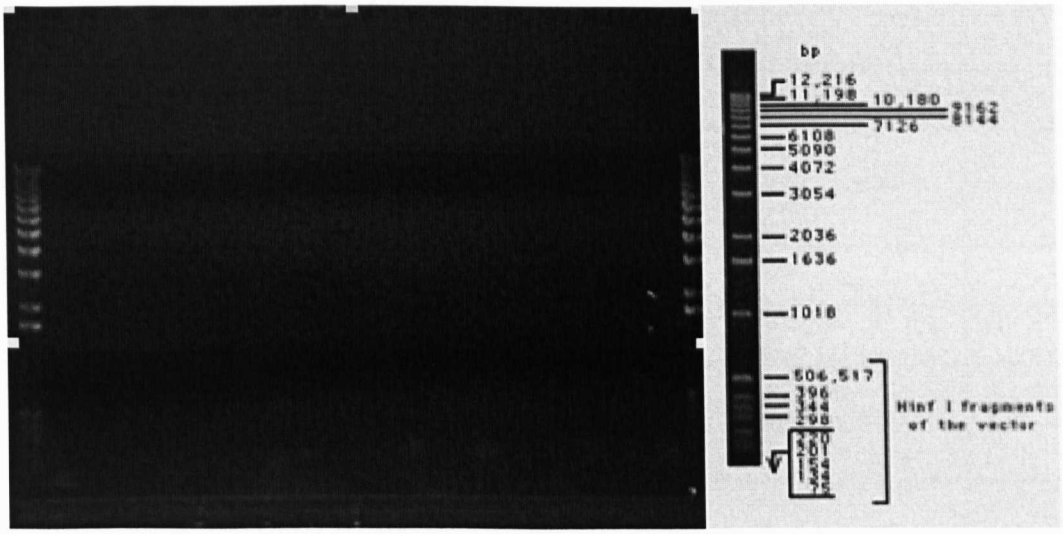

Figure 29. Detection of *cj1321-cj1326* in strains 47693, 1771 and 15168 from the "non chicken clade". All PCR reactions were carried out at the same time as those in figure 25, thus lanes 2 to 7 of figure 25 are the positive control for this gel. Lanes 1 and 20 contain molecular weight marker. Lanes 2 to 7 show the observed PCR products from DNA isolated from the outlier chicken isolate found in the "non chicken" clade amplified with primers for *cj1321* to *cj1326* inclusive. Lanes 8 to 13 show the observed PCR products from DNA isolated from beach isolate 1771 amplified with primers for *cj1321* to *cj1326* inclusive. Lanes 14 to 19 show the observed PCR products from DNA isolated from clinical isolate 15168 found in the "non chicken" clade amplified with primers for *cj1321- cj1326* inclusive. Table 25 contains the information for each gel lane.

Table 25. Key to figure 29

Lane	Strain	Source	CDS	Expected	Observed
2	47693	chicken	*cj1321*	-	-
3	47693	chicken	*cj1322*	-	-
4	47693	chicken	*cj1323*	-	-
5	47693	chicken	*cj1324*	-	-
6	47693	chicken	*cj1325*	-	-
7	47693	chicken	*cj1326*	-	-
8	1771	beach	*cj1321*	-	-
9	1771	beach	*cj1322*	-	-
10	1771	beach	*cj1323*	-	-
11	1771	beach	*cj1324*	-	-
12	1771	beach	*cj1325*	-	-
13	1771	beach	*cj1326*	-	-
14	15168	clinical (A)	*cj1321*	-	-
15	15168	clinical (A)	*cj1322*	-	-
16	15168	clinical (A)	*cj1323*	-	-
17	15168	clinical (A)	*cj1324*	-	-
18	15168	clinical (A)	*cj1325*	-	-
19	15168	clinical (A)	*cj1326*	-	-

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

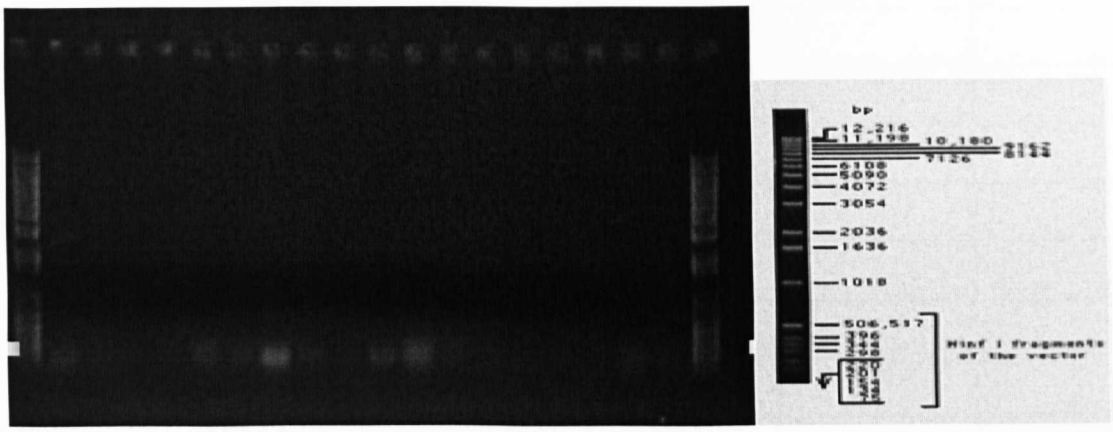

Figure 30. Detection of *cj1321-cj1326* in strains 34007, 18836 and 12241 from "non chicken clade". All PCR reactions were carried out at the same time as those in figure 25, thus lanes 2 to 7 of figure 25 are the positive control for this gel. Lanes 1 and 20 contain molecular weight marker. Lanes 2 to 7 show the observed PCR products from DNA isolated from clinical isolate 34007 found in the "non chicken" clade amplified with primers for *cj1321* to *cj1326* inclusive. Lanes 8 to 13 show the observed PCR products from DNA isolated from clinical isolate 18836 found in the "non chicken" clade amplified with primers for *cj1321* to *cj1326* inclusive. Lanes 14 to 19 show the observed PCR products from DNA isolated from ovine isolate 12241 found in the "non chicken" clade amplified with primers for *cj1321- cj1326* inclusive. Table 26 contains the information for each gel lane.

Table 26. Key for Figure 30

Lane	Strain	Source	CDS	Expected	Observed
2	34007	clinical (A)	*cj1321*	-	-
3	34007	clinical (A)	*cj1322*	-	-
4	34007	clinical (A)	*cj1323*	-	-
5	34007	clinical (A)	*cj1324*	-	-
6	34007	clinical (A)	*cj1325*	-	-
7	34007	clinical (A)	*cj1326*	-	-
8	18836	clinical (A)	*cj1321*	-	-
9	18836	clinical (A)	*cj1322*	-	-
10	18836	clinical (A)	*cj1323*	-	-
11	18836	clinical (A)	*cj1324*	-	-
12	18836	clinical (A)	*cj1325*	-	-
13	18836	clinical (A)	*cj1326*	-	-
14	12241	ovine/bovine	*cj1321*	-	-
15	12241	ovine/bovine	*cj1322*	-	-
16	12241	ovine/bovine	*cj1323*	-	-
17	12241	ovine/bovine	*cj1324*	-	-
18	12241	ovine/bovine	*cj1325*	-	-
19	12241	ovine/bovine	*cj1326*	-	-

6.3.2 Confirmation of motility in selected strains from the "chicken clade" and "non chicken clade"

Due to the location of the *cj1321-cj1326* in the flagellin modification locus, the motility of strains from both the chicken and non-chicken clade was tested to determine whether the absence or presence of these CDSs affected motility. One chicken strain (11818) from the "chicken clade" in which *cj1321-1326* were confirmed as present both by microarray data and PCR was tested for motility. In addition, one clinical isolate (18836) from the "non chicken clade", in which *cj1321-1326* were identified as absent both by microarray analysis and PCR was tested for motility. The motile sequenced strain NCTC11168 (hypermotile variant) was used as a positive control in motility tests (Figure 31). The motility of chicken strain 11818 (Figure 32) and clinical strain, 18836 (Figure 33) was checked by growing the isolates on motility agar, in triplicate (Chapter 2.1.3).

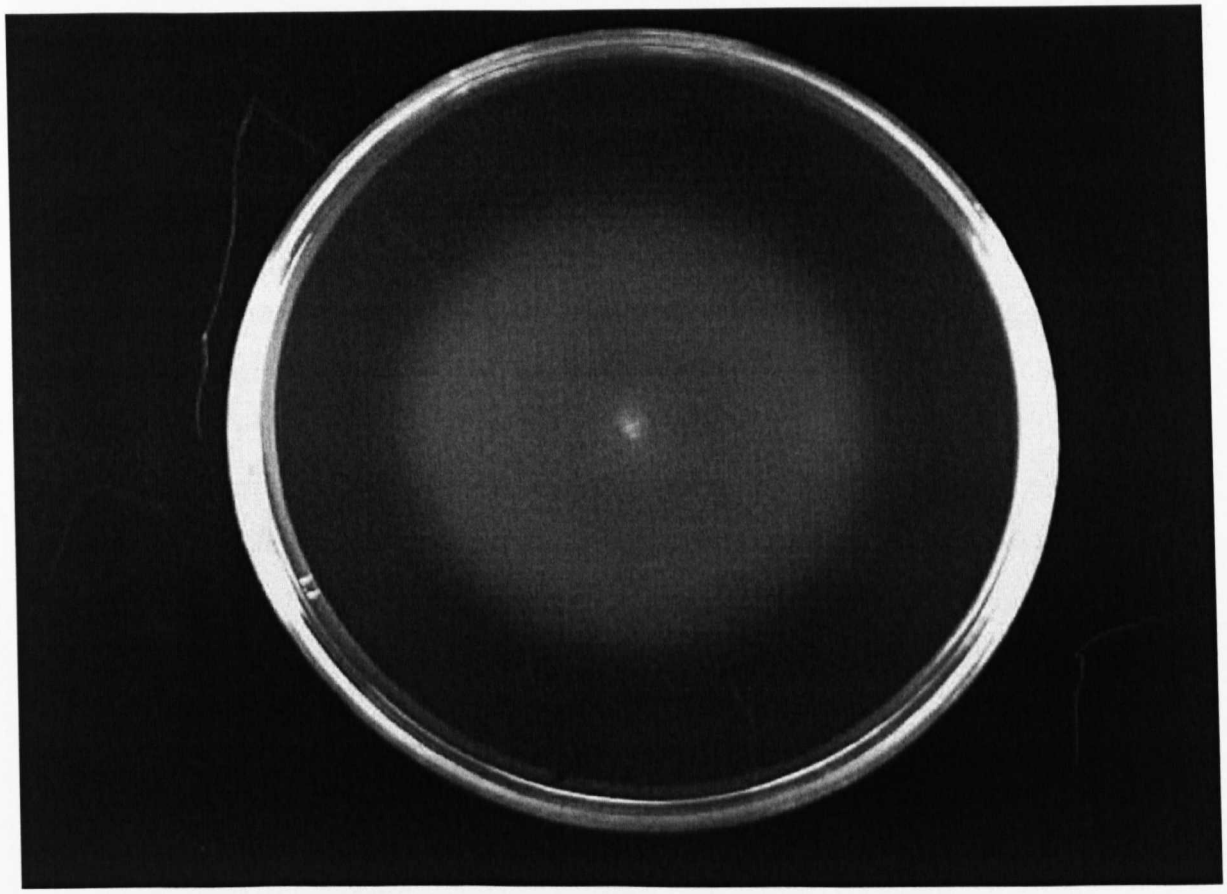

Figure 31. Motility after 48 hours of strain NCTC11168 – positive control. Hypermotile NCTC11168 positive control for motility phenotype. Innoculum placed at centre of motility agar plate and incubated for 48 hours. High levels of motility demonstrated by growth out from centre of plate towards the edge of the petri dish.

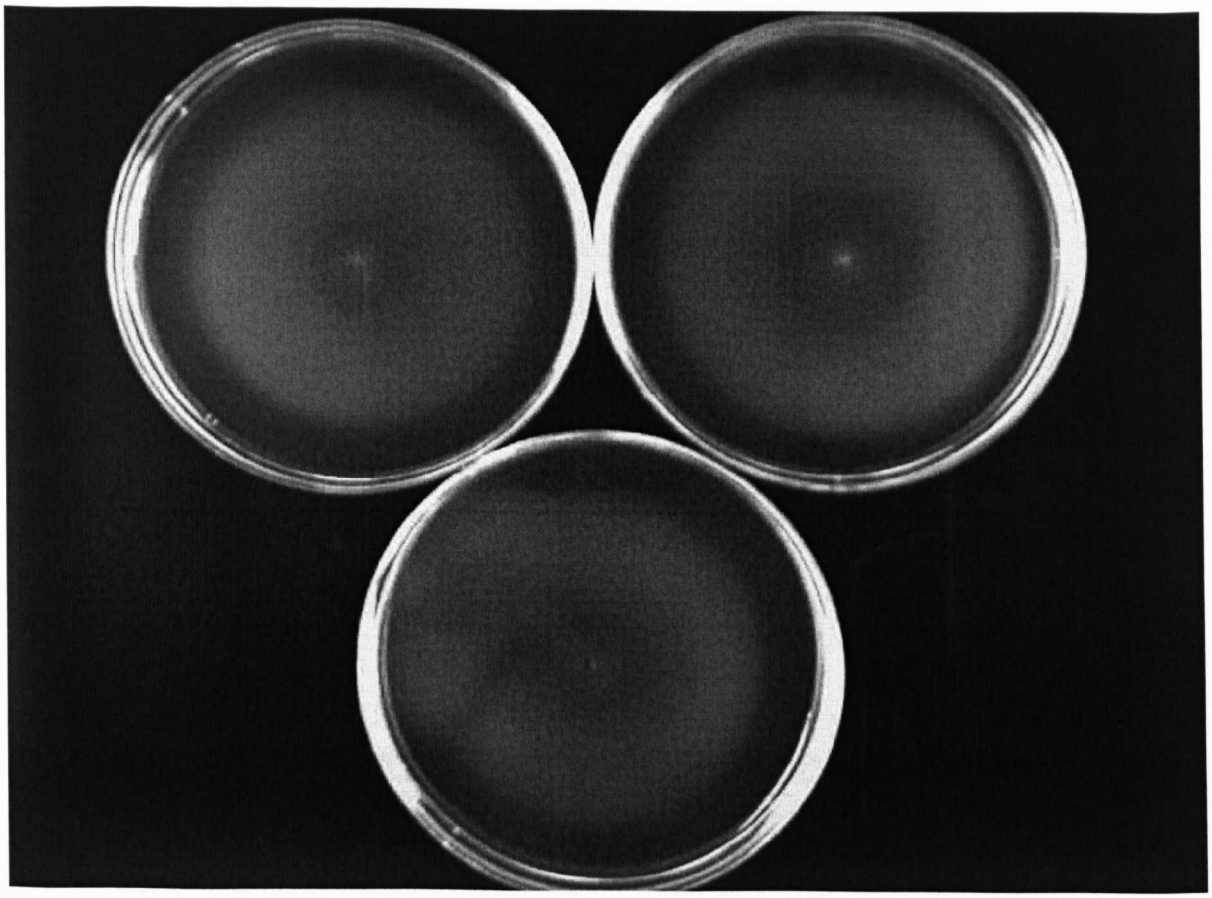

Figure 32. Motility after 48 hours of strain 11818 – chicken isolate from "chicken clade". Three replicates of chicken isolate 11818 tested for motility phenotype. Innoculum placed at centre of motility agar plate and incubated for 48 hours. High levels of motility demonstrated by growth out from centre of plate towards the edge of the petri dish.

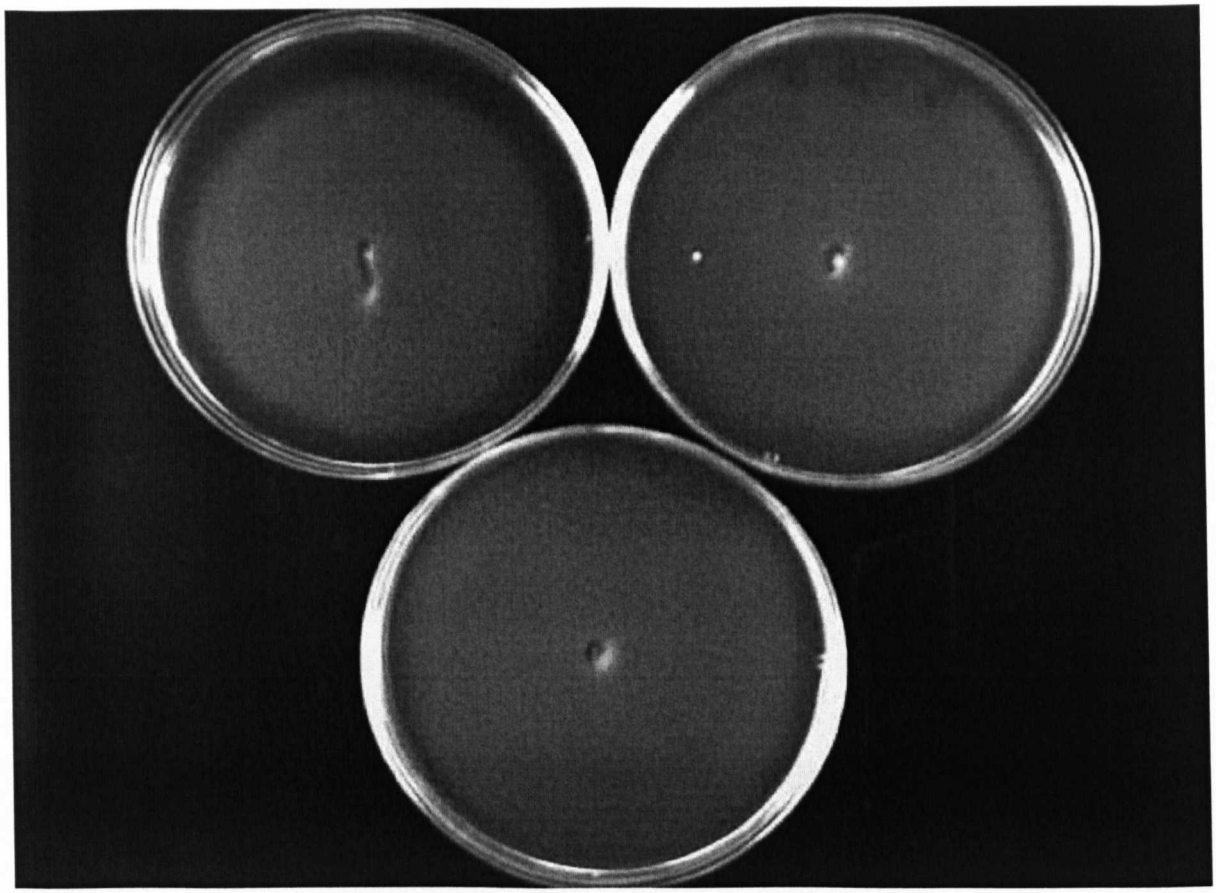

Figure 33. Motility after 48 hours of strain 18836 – clinical isolate from "non chicken clade". Three replicates of clinical isolate 18836 tested for motility phenotype. Innoculum placed at centre of motility agar plate and incubated for 48 hours. High levels of motility demonstrated by growth out from centre of plate towards the edge of the petri dish.

6.4 Discussion

6.4.1 Comparative phylogenomics

In this study phylogenomics utilising comparative genomic data from 91 characterised *C. jejuni* strains from diverse sources combined with Bayesian based algorithms were used to determine the phylogeny of the strains. The aim was to establish whether strains from the same source could be distinguished from strains isolated from other hosts, facilitating the identification of host specific genetic markers that could potentially be used to predict the source of human *C. jejuni* infection.

Bayesian based algorithms implemented through MrBayes3.0 software were used to infer robust hypotheses of phylogenetic relationships using whole genome data from microarray analyses. Relationships were tested using the majority rule and consensus trees generated using the computer programme Phylogenetic Analysis Using Parsimony (PAUP*). Phylogenetic trees indicated that strains isolated from the same host formed distinctive clades with a few exceptions. The "non chicken clade" comprised 52% (47/91) of the total number of strains including 59% (37/63) clinical isolates. Ovine, bovine and environmental source strains formed distinct sub clades within the "non chicken clade". Statistically these were all unequivocally supported (P=1.0). Several CDSs distinguished strains of ovine and bovine origin from other *C. jejuni* strains in the study. However, CDS *cj0105*, encoding a probable ATP synthase, was uniquely absent or divergent in strains isolated from cows and sheep and the two clinical isolates present in this clade. Further research is required to substantiate *cj0105* as a *C. jejuni*-specific ovine/bovine genetic marker.

The "chicken clade", so called as 94% (16/17) of the chicken isolates were all found in this clade, comprised 48% (44/91) of the total number of strains including 41% (26/63) clinical isolates. The chicken isolates included in this study were highly diverse by traditional typing methods, comprising five different serotypes (HS2, HS5, HS27, HS44 and HS50) and eleven different phage types (PT1, PT2, PT5, PT15, PT19, PT25, PT34, PT35, PT36, PT44 and PT59). In addition, chicken strains were isolated from both supermarket poultry portions, abattoirs and flocks between 1991 and 2001 from different geographical locations throughout the UK. Based on the phenotypic and geographic diversity and time frame of isolation of these strains it is highly unlikely that this phylogeny was due to the presence of a clonal *C. jejuni* population found in poultry. Furthermore, traditional typing methods such as

serotyping and phage typing have failed to identify any relationship between these isolates.

Some exceptions to the general host specific trends were seen. One chicken strain was found in the "non-chicken clade"and two ovine bovine strains that were isolated from supermarket ox liver portions, were found in the "chicken clade". We hypothesise that *C. jejuni* adapts to live in specific ecological niches and therefore strains have a preferred host, yet they are able to survive in other hosts. Furthermore, processing of animal carcasses post slaughter may lead to cross contamination of a *C. jejuni* strain to a source other than its preferred host from where it could be isolated. Moreover, this population structure may be specific to UK chicken strains. A representative sample of strains from different geographic locations worldwide would be required for further investigation to establish whether non-UK chicken isolates were also distinguishable from strains isolated from different ecological niches. Other ecological niches that could be investigated include pigs, wild birds, water, milk and pets.

6.4.2 Identification of an avian genetic marker

Using MacClade 4 software, CDS evolution within the hypothesised phylogenetic tree was traced. Tracing shows the most parsimonious hypothesis of ancestral states. That is, the simplest placing of presence and absence or divergence of genes in the common ancestor of a clade. This indicates how the presence and absence or divergence of certain genes in an ancestral strain has led to the formation of a new clade. Conserved regions may be indicative of the selective pressures that caused a CDS to be lost or to diverge or to remain stable. CDSs that were present and absent or divergent at the basal node of the major clades were identified. This technique facilitates the identification of CDS patterns that have been inherited from a common ancestor. Strains in a clade may exhibit common phenotypic traits and so genetic markers correlating to phenotype may be identified. Many of the regions of the genome causing these strains to cluster together into clades were shared with strains in other clades. For example, a CDS may be absent in all strains in one clade, but also absent from strains found in other clades. Likewise, all strains in a clade may possess a CDS with the exception of one. However, some CDSs were clade specific. Microarray analyses were validated by PCR in (16/91) 18% of the strains through the amplification of CDSs that differentiated "chicken" and "non chicken" clades or were

identified as absent or divergent unique to specific clades. These results validated microarray data.

A significant finding was that an "islet" of CDSs, *cj1321-cj1326,* as well as *cj1338* and *cj1339,* were absent or divergent in the common ancestor of the "non-chicken clade". These CDSs form part of the putative flagellin glycosylation locus of *C. jejuni* (Figure 3). Genes *cj1338* and *cj1339* encode flagellin structural proteins FlaA and FlaB which are known to be highly diverse between *C. jejuni* isolates from previous studies. Variability of *C. jejuni* flagellar genes *flaA* and *flaB* has been well documented in both the pre-microarray and microarray eras. In *Pseudomonas aeruginosa* there is a correlation between flagellin sequence variants and the inheritance of a genetic island essential for the glycosylation of flagellin (Arora *et al.*, 2004). This relationship has not been investigated in *C. jejuni.* However, variation in *flaA* between isolates forms the basis of the Fla typing scheme (Meinersmann *et al.*, 1997) and more recently microarray studies carried out by Pearson and colleagues also found low levels of hybridisation of genes *cj1338* and *cj1339* with reporter elements on their microarray (Pearson *et al.*, 2003). Thus the absence or divergence of genes *cj1338* and *cj1339* in this study was not unexpected and may well be indicative of strain-strain variation in these genes.

However, CDSs *cj1321-cj1326* were present in the majority of chicken isolates and clinical strains in the "chicken clade". No studies have been reported so far on CDSs *cj1321* to *cj1326.* From the predicted amino acid sequences, Cj1321 is similar to acetyl transferases from many bacterial species, Cj1322 and Cj1323 have similarity to hydroxyacyl dehydrogenases, Cj1324 has similarity to WbpG, a LPS biosynthesis protein found in several bacteria, Cj1325 is similar to Cj1330 involved in the synthesis of pseudaminic acid and Cj1326 has no obvious similarity to any known protein. Given the similarity of most of the *cj1321* to *cj1326* CDSs to genes involved in carbohydrate biosynthesis or sugar modifications, plus the CDSs' location within the *O*-linked glycosylation locus, this strongly suggests that these CDSs are involved in carbohydrate modification of the flagellum. The presence of this potential glycosylation islet was confirmed by PCR in all chicken isolates included in the study. Additionally, the presence of *cj1321-cj1326* was screened for by PCR in six chicken isolates not included in the comparative phylogenomics study. PCR products were observed for *cj1321-cj1326* in all chicken isolates screened. Furthermore, *cj1321-cj1326* were not observed following PCR screening in ovine/ bovine, beach

isolates and clinical isolates from the "non chicken clade". Thus the loss or divergence of *cj1321-cj1326* from "non chicken clade" strains appears to form the basis of differentiation between the "chicken" and "non-chicken" clades.

In *C. jejuni*, genetic organisation of flagella glycosylation regions varies dramatically not only between different species, but also between the strains. For example, the 81-176 strain of *C. jejuni* lacks a large contiguous region corresponding to CDSs *cj1318-cj1332* present in *C. jejuni* strain NCTC11168 (Figure 3). The deletion may be a result of the presence of several highly similar genes of the *maf* family in the latter strain, which may have created a recombinational hot spot (Karlyshev *et al.*, 2002a). In particular, *maf* genes 1 and 4 are identical at the nucleotide level (Karlyshev *et al.*, 2002a). Microarray studies by Pearson and colleagues also found this locus to be highly variable between strains (Pearson *et al.*, 2003). To date structural studies on the flagellin modification system have focused on *C. jejuni* 81-176 and the *C. coli* strain VC167. No structural data has been reported for NCTC11168 but given this strain contains several additional genes it is likely that other sugar modifications of the flagellin (e.g. modification of pseudaminic acid) are yet to be identified. Glycosylation of flagellin is increasingly being recognised in a number of Gram-negative pathogenic bacteria, including *Pseudomonas aeruginosa, Helicobacter pylori, Vibrio parahaemolyticus* and *Aeromonas* spp (Castric *et al.*, 2001; Gavin *et al.*, 2002; Schirm *et al.*, 2003). The modifications increase the hydrophilicity of flagellin, and often influence the cells' immunogenicity and their interaction with eukaryotic cells (Castric *et al.*, 2001; Gavin *et al.*, 2002; Logan *et al.*, 2002; Schirm *et al.*, 2003). The biological significance of the presence of *cj1321 – cj1326* in chicken isolates in unknown. However, we hypothesise that it may play a role in the colonization of chicken caeca.

Given the unequivocal clade separation of chicken and non-chicken strains it is possible that clinical isolates found in the "chicken clade" were infected from *C. jejuni* contaminated chicken. Conversely, the source of infection of human strains in the "non chicken clade" may have been from other animal and possibly environmental sources. Genetic markers distinguishing *C. jejuni* strains from different ecological niches may allow the source of human infection to be predicted.

7.0 Overall discussion and conclusions

7.1 Background to this project

At the outset of this project the mechanism of pathogenesis of *C. jejuni* was poorly understood. Moreover, it was unknown whether differences in human *C. jejuni* disease outcomes, ranging from asymptomatic colonisation to severe inflammatory diarrhoea and rare sequelae, were due to genetic differences in *C. jejuni* strains and/ or other factors. The epidemiology of *C. jejuni* was also poorly characterised. Traditional typing methods such as serotyping, phage typing and PFGE had been unsuccessful at distinguishing *C. jejuni* strains from different ecological niches. Therefore, the proportion of human disease attributable to different sources of *C. jejuni* infection was unknown, although the predominant source had frequently been cited as poultry (Friedman *et al.*, 2004). Comparative genomics studies into the genetic diversity of *C. jejuni* carried out before and during this project have identified the functional core of various *C. jejuni* strain collections and have attempted to identify genetic markers distinguishing *C. jejuni* strains associated with GBS (Dorrell *et al.*, 2001; Leonard *et al.*, 2004; Pearson *et al.*, 2003; Taboada *et al.*, 2004). However, none of these studies had used detailed phylogenetic analysis to compare large and diverse strain collections.

7.2 Aims of project

The application of DNA microarray analysis of *C. jejuni* strains of diverse origin provides information about which CDSs are present or absent/ divergent in the genomes of individual isolates. Through the development of a gene specific *C. jejuni* DNA microarray and improved data analyses methods the objective of this study was to distinguish *C. jejuni* from well-characterised UK strains of diverse origins, identifying determinants important in virulence, transmission and host specificity.

7.3 Results and conclusions

This project began with the design and construction of a gene specific composite *C. jejuni* DNA microarray following the success of the proof of principal *C. jejuni* clone array (Chapter 3). Reporter elements incorporating PCR products from each CDS in the NCTC11168 genome were amplified using primers designed to amplify NCTC11168 CDSs for minimal cross hybridisation. Finally, the microarray also

included 69 CDSs absent from the NCTC11168 genome. Thus, this second generation array begins the move away from a strain-specific array to a more representative array of the *C. jejuni* species. Additional CDSs included on the microarray were identified, largely from loci encoding surface antigens, and were highlighted in some hybridisation experiments.

When considering the microarray approach to comparative genomics, the limitations of this technology include its inability to detect novel genes absent from the array, so while microarray analysis will reveal genes absent or divergent from the genes represented by the reporter elements, no information on any extra genes present in the test strain will be obtained. The latest *C. jejuni* microarrays produced by BμG@S contain reporter elements representing 69 *C. jejuni* genes that are absent in the sequenced strain NCTC11168. As more *C. jejuni* strains are sequenced and our knowledge of the CDSs they contain increases, microarrays will become less strain-specific and more representative of the species. Another limitation of microarrays is that they cannot detect gene rearrangements. Arrays that use PCR products as reporter elements will also be unable to detect point mutations and small deletions or insertions. Microarrays do not usually include intergenic regions, so variation in promoter sequences will not be analysed. The use of oligonucleotide microarrays or Affymetrix GeneChip™ technology should reduce these limitations considerably. Despite this, microarray-based comparative genomics have already provided important information on the genetic diversity of *C. jejuni* isolates and a clearer picture of the role of genetic diversity in the pathogenesis of this organism has been shown through this study.

Prior to carrying out detailed analysis of DNA microarray hybridisation data, two methods of analysis were compared. The definition of absent or divergent CDSs using a constant cut-off of 0.5, implemented through GeneSpring6.1 was compared with a dynamic cut-off calculated using GACK. Raw microarray data are initially processed and saved using ImaGene5.5 in a format that is recognised by GeneSpring6.1. The data format used by ImaGene5.5 was not compatible with GACK therefore manual data formatting was required. This is labour intensive with room for human error and there is scope for improvement by automating the process. The binary or trinary output of GACK is easily manipulated into formats such as Nexus required for phylogenetic simulations using search methods such as Bayesian based algorithms

that consider each gene directly. However, genome composition data acquired through GeneSpring6.1 must be manually converted into binary format. This process has been semi automated through the development of a perl script by Adam Witney and used in this study. Some discrepancies were found between the number of CDSs identified as absent or divergent from the same strains using these two techniques leading to a direct comparison of two sequenced *C. jejuni* strains, RM1221 and NCTC11168, to determine which data analysis method was more sensitive. This direct comparison of the same hybridisation data using the two methods of analysis indicated that, overall, the use of a constant cut-off of 0.5 through GeneSpring6.1 correctly identified slightly more absent and divergent CDSs. This validated the use of a constant cut-off value of 0.5 for DNA microarray hybridisation data analysis.

Precise strain comparisons of well-characterised UK strains from diverse clinical outcomes were carried out to gain a better understanding the genetic differences between *C. jejuni* strains from the whole spectrum of disease outcome. One aim was to identify correlates of pathogenesis and epidemiological markers for hyperinvasive strains associated with serious sequelae, septicaemia and GBS, that are occasionally associated with *C. jejuni* infection. High levels of LOS variation between strains associated with septicaemia were noted and these strains were subsequently tested at the National Research Canada, Ottawa, for LOS outer core structural analysis to determine whether strains associated with septicaemia possessed a unique LOS outer core structure. Rather than a conserved LOS structure, analysis of these strains identified high levels of variation with the seven strains possessing six different LOS outer cores. This result indicates that the septicaemia isolates included in this study do not lack key CDSs present in NCTC11168 through which the mechanism of pathogenesis may be determined or septicaemia isolates distinguished from less invasive *C. jejuni* strains. However, DNA microarray analysis is limited by the CDSs represented on the microarray. It is possible that *C. jejuni* isolates that cause septicaemia possess additional CDSs that facilitate this hyperinvasive phenotype absent in NCTC11168 or represented in the non-NCTC11168 reporter elements. Such additional CDSs could be identified using methods such as subtractive hybridisation (Ahmed *et al.*, 2002). One of the septicaemia strains from this study, 52471, has been sent to Emily Kay at the Sanger Institute, Cambridge, UK for subtractive hybridisation to determine whether additional CDSs are present. This analysis was underway at the time of writing.

TopA, encoding DNA topoisomeraseI, was absent from two *C. jejuni* strains associated with GBS but present in every other isolate in the study. Although interesting, a previous study carried out to identify genetic markers of *C. jejuni* strains associated with GBS found no specific genes or loci associated with GBS. It would therefore be necessary to screen for *topA* in a larger collection of GBS associated strains to determine whether this single gene is a genetic marker for GBS and to attempt to underpin its biological significance.

It was hypothesised that genome comparisons between pathogenic and non-pathogenic strains would highlight CDSs with a role in causing human diarrhoeal disease. Potentially non-pathogenic strains were identified both from the environment and from human asymptomatic carriers (Chapter 5). Comparison of the genomes of these potentially non-pathogenic *C. jejuni* strains using DNA microarrays revealed detailed information about which CDSs were present and absent or divergent in the genomes of individual isolates. Regions of the genome that were absent or divergent in each of these potentially non-pathogenic strains were highlighted. Common regions of variability included highly divergent loci encoding the surface antigens capsular polysaccharide and LOS. However, a probable secreted serine protease, *cj1365*, was identified as absent or divergent in all six potentially non-pathogenic beach isolates, a result confirmed by PCR analysis and subsequent sequencing of the region. Moreover, *cj1365* was identified as present in 88% of the clinical isolates included in the study by microarray analysis. Secreted serine proteases have a role in virulence in enteric pathogens such as enteropathogenic and enterohaemorrhagic *E. coli* (Brunder *et al.*, 1997; Stein *et al.*, 1996). CDS *cj1365* represents a putative *C. jejuni* virulence determinant and consequently a defined *cj1365* mutant has been constructed in our laboratory for further investigation. Furthermore, 15% of the CDSs identified as absent or divergent in the potentially non-pathogenic beach isolates showed no amino acid similarity to proteins from other bacteria. To date, the beach strains represent the most well characterised potentially non-pathogenic strain collection available. *C. jejuni* appears to have a unique mechanism of pathogenesis, the function unknown CDS are interesting candidates for further investigation into their role in the mechanism of pathogenesis of *C. jejuni*. Consequently, defined mutants in these function unknown CDSs are under construction in our laboratory and phenotypic assays will be used to ascertain their biological function.

In addition to investigating human clinical isolates of diverse origins, differences in genomic content of *C. jejuni* strains from a wider variety of sources were investigated. Whole genome comparisons through microarray hybridisations were coupled with phylogenetic analysis identifying regions within the *C. jejuni* genome that show a high degree of variation between strains. The study of these regions has highlighted genetic factors that may be linked to phenotypic variation and adaptation to different ecological niches. DNA microarrays represent an efficient technology for whole genome comparisons, allowing a bird's eye view of the absence and presence of genes in a given genome compared to the reference genome on the microarray and 'whole genome' trees allow evolutionary analysis to be introduced to comparative genomics studies (Eisen and Fraser, 2003). Cluster algorithms available within software programs such as GeneSpring6.1 facilitate the rapid identification of phylogenetic relationships of strains, identifying patterns of genome content with relation to different parameters including source and serotype. However, the robustness of these phylogenetic relationships cannot be tested. This is because cluster methods lose much of the information present in the data matrix by converting the aligned binary whole genome data into a pairwise distance matrix. The distance matrix is then entered into a tree building method. Search methods such as the Bayesian based algorithms used in this study consider each site directly. Such methods employ greater processing power and consequently analyses take significantly longer than those methods using a distance matrix. However, if trees generated using methods with a distance matrix and those with discrete characters revealed identical phylogeny, the latter tree would provide additional information of the sites in the genome that contribute to each branch. This information is lost when data is converted into a distance matrix. Therefore trees produced using clustering algorithms do not relate back to the original data thus the identification of genes specific to particular groups of strains is not possible. Moreover, clade credibility values cannot be calculated for phylogeny inferred using clustering algorithms.

Bayesian inference of phylogeny has only been proposed relatively recently and has several advantages over other methods including easy interpretation of results, the ability to incorporate prior information if available. Phylogenetic relationships simulated using Bayesian based algorithms require the conversion of microarray data into Nexus format. This is more labour intensive than using clustering algorithms available through GeneSpring6.1, simulations require large amount of processing

power and time and the results must be viewed using a second program, TreeView. Furthermore, during the development of the comparative phylogenomics approach non-NCTC11168 reporter elements included on the composite *C. jejuni* microarray were excluded from the analysis. This was due to difficulties standardising the cut-off level at which CDSs were designated absent/divergent or present in non-NCTC11168 reporter elements for which no control DNA was present. The comparative phylogenomics technique developed in this study may be refined through the inclusion of a greater number of non-NCTC11168 CDSs on the microarray and the inclusion of gDNA from non-NCTC11168 strains in competitive hybridisations. This would facilitate the devlopment of a system where strains were not compared with the genome content of a single strain.

However, this is the first study using this robust comparative phylogenomics approach to investigate bacterial population structures and more specifically, the phylogeny of *C. jejuni*. Relationships between strains with specific phenotypes at a whole genome level were clearly demonstrated and correlation between strains of a particular genomotype and host were apparent. Furthermore, genetic markers associated with specific phenotypes were identified. We have demonstrated a population structure comprising two robust clades; a "chicken clade" containing strains from chicken and clinical sources and a "non-chicken clade" containing strains from non-chicken and clinical sources. A single chicken isolate was also found in the non-chicken clade. Using majority rule consensus trees unequivocal clade credibility values were obtained for the non-chicken and chicken clades, a clade containing strains isolated from ovine and bovine animal sources and clades comprising isolates from beaches. These data clearly show that *C. jejuni* strains in this study, from different animal and environmental sources, can be distinguished. Unexpectedly, 59% of clinical isolates included shared a common ancestor with strains from non-chicken sources. We could hypothesise that clinical isolates found in the "non chicken clade" were transmitted to humans from non-chicken sources. Likewise, strains in the "chicken clade" may have been transmitted from chickens. The results of this study suggest that ovine, bovine and environmental sources contribute substantially to the burden of UK human *C. jejuni* infection. This is supported by the findings of a study by Brown et al that the risk of human exposure to *Campylobacter spp*, and in particular *C. jejuni* from the environment is very high (Brown *et al.*, 2004). If predictions of the source of human *C. jejuni* infection were robust, several important

C. jejuni public health questions could potentially be answered through the implementation of large-scale comparative phylogenomics studies on representative *C. jejuni* isolates for which full epidemiological information is available. Such questions include;

- The proportion of human *Campylobacter* disease caused by infected chicken;
- The proportion of disease caused by infected chicken due to direct consumption compared with the proportion caused by cross contamination;
- The cost of human *Campylobacter* disease caused by infected chicken;
- Trends in the incidence of *Campylobacter* could be retrospectively investigated. For example to determine whether disease caused by infected chicken decreased as fast as disease not caused by infected chicken.

The genomic regions differentiating strains in the "chicken clade" (comprising the 94% of strains from poultry sources and 41% of clinical isolates) from strains in the "non-chicken clade" (comprising strains from ovine and bovine animal sources, beach samples as well as 59% of clinical isolates) were identified. CDSs *cj1321-cj1326* were identified as a key locus differentiating the two clades, with the majority of strains in the "chicken clade" possessing the CDSs whilst they were absent from strains in the "non chicken clade". Little is known about the biological function of genes *cj1321-cj1326.* However, from their location (*cj1314-cj1338*), some genes in this region are involved with *O*-linked glycosylation of flagellin proteins (Szymanski *et al.*, 2003) and *cj1321-cj1326* may have a related function. Therefore we have termed this region a "glycosylation islet". *Cj1321-cj1326* are flanked by *maf1* and *maf4*, genes that are 100% homologous at a nucleotide level (Karlyshev *et al.*, 2002a). This suggests a possible mechanism through which this "glycosylation islet" may be lost, through homologous recombination between *maf1* and *maf4*. Presence of this "glycosylation islet" in the majority of strains isolated from poultry may correlate with post- translational modifications on flagellin. Through *O*-linked glycosylation the flagellin may be masked with sialic acid preventing a proinflammatory response in poultry therefore increasing pathogen survival within the host. The biological role of *cj1321-cj1326* with relation to host specificity is therefore a prime candidate for further research. Research has begun into this genetic locus and preliminary PCR screening studies show that *cj1321-cj1326* are present in UK chicken isolates

included in the comparative genomics study as well as six further UK chicken isolates that were not part of the bank of strains used in the comparative phylogenomic study. In addition strains tested, in which *cj1321-cj1326* were absent, were still fully motile. This was demonstrated using motility agar plates for strains that both possess and lack these CDSs revealing high levels of motility. In addition, strain 81-176 lacks *cj1321-cj1326* and yet has been shown to be highly motile.

7.4 Future studies

We hypothesise that the *cj1321-26* islet enables *C. jejuni* strains that possess it to persist in poultry, by encoding a variant flagellin glycoform. This may play a key role in the bacterium's interaction with its avian host, relating to colonisation, survival and/or immune response. Future studies to investigate this hypothesis may include a thorough investigation of the flagella locus of chicken and non-chicken strains through DNA sequencing of the region. The ability of the different strains to colonise and survive in chickens should then be tested, as well as histological and avian immune response studies. In addition, the structure of the flagellin glycan in the parent NCTC11168 strain and in defined *cj1321-26* mutants could be determined by NMR and mass spectroscopy.

However, to fulfil the promise of these studies, a larger collection of diverse strains of known origin must be investigated. The remit of this study was to investigate isolates from the UK only (due to partial funding of the study by the FSA). Further studies on a large and well-characterised strain collection from non-UK sources would allow the robustness of these phylogenetic relationships to be determined and the main factors contributing to human *C. jejuni* infection could be unequivocally calculated. This may contribute toward the development of an improved molecular typing method such as PCR and would improve our understanding of *C. jejuni* epidemiology to a point where targeted control strategies to prevent this enteropathogen entering and passing through the food chain could be introduced.

8.0 appendices

Appendix 1. Strains used in *C. jejuni* comparative genomics study. Strains in blue were included to calculate the *C. jejuni* functional core genes.

Strain	Source	Origin	HS serotype
33106	Clinical asymptomatic	Campylobacter Reference Lab, UK	4
33084	Clinical asymptomatic	Campylobacter Reference Lab, UK	35
32799	Clinical asymptomatic	Campylobacter Reference Lab, UK	50
31485	Clinical asymptomatic	Campylobacter Reference Lab, UK	untypeable
32787	Clinical asymptomatic	Campylobacter Reference Lab, UK	18
31481	Clinical asymptomatic	Campylobacter Reference Lab, UK	37
31467	Clinical asymptomatic	Campylobacter Reference Lab, UK	18
43983	Clinical septicaemia	Campylobacter Reference Lab, UK	50
44119	Clinical septicaemia	Campylobacter Reference Lab, UK	18
34007	Clinical septicaemia	Campylobacter Reference Lab, UK	18
52471	Clinical septicaemia	Campylobacter Reference Lab, UK	untypeable
53250	Clinical septicaemia	Campylobacter Reference Lab, UK	60
47886	Clinical	Campylobacter	untypeable

	septicaemia	Reference Lab, UK	
47939	Clinical septicaemia	Campylobacter Reference Lab, UK	67
15168	Clinical GBS	Campylobacter Reference Lab, UK	19
18836	Clinical GBS	Campylobacter Reference Lab, UK	19
M1	Clinical	Veterinary Laboratory Agency, UK	21
81116	Clinical	Veterinary Laboratory Agency, UK	6
40671	Clinical	Campylobacter Reference Lab, UK	50
30280	Clinical	Campylobacter Reference Lab, UK	16
30328	Clinical	Campylobacter Reference Lab, UK	16
34555	Clinical diarrhoea	Campylobacter Reference Lab, UK	5
35335	Clinical diarrhoea	Campylobacter Reference Lab, UK	untypeable
35424	Clinical bloody diarrhoea	Campylobacter Reference Lab, UK	untypeable
36069	Clinical vomiting	Campylobacter Reference Lab, UK	5
36439	Clinical diarrhoea	Campylobacter Reference Lab, UK	12
36860	Clinical	Campylobacter	21

	diarrhoea	Reference Lab, UK	
36952	Clinical diarrhoea	Campylobacter Reference Lab, UK	untypeable
37537	Clinical bloody diarrhoea	Campylobacter Reference Lab, UK	untypeable
35799	Clinical diarrhoea	Campylobacter Reference Lab, UK	untypeable
38353	Clinical bloody diarrhoea	Campylobacter Reference Lab, UK	5
38556	Clinical vomiting	Campylobacter Reference Lab, UK	13
38762	Clinical bloody diarrhoea	Campylobacter Reference Lab, UK	18
38857	Clinical vomiting	Campylobacter Reference Lab, UK	23
39182	Clinical vomiting	Campylobacter Reference Lab, UK	13
39828	Clinical vomiting	Campylobacter Reference Lab, UK	42
40917	Clinical bloody diarrhoea	Campylobacter Reference Lab, UK	21
41651	Clinical vomiting	Campylobacter Reference Lab, UK	16
42724	Clinical bloody diarrhoea	Campylobacter Reference Lab, UK	untypeable
43205	Clinical bloody diarrhoea	Campylobacter Reference Lab, UK	2
44811	Clinical vomiting	Campylobacter Reference Lab, UK	2
44933	Clinical	Campylobacter	13

	diarrhoea	Reference Lab, UK	
44958	Clinical vomiting	Campylobacter Reference Lab, UK	50
45557	Clinical bloody diarrhoea	Campylobacter Reference Lab, UK	60
45631	Clinical diarrhoea	Campylobacter Reference Lab, UK	13
48612	Clinical bloody diarrhoea	Campylobacter Reference Lab, UK	2
52331	Clinical vomiting	Campylobacter Reference Lab, UK	50
52368	Clinical bloody diarrhoea	Campylobacter Reference Lab, UK	untypeable
55320	Clinical bloody diarrhoea	Campylobacter Reference Lab, UK	13
55703	Clinical vomiting	Campylobacter Reference Lab, UK	13
56281	Clinical bloody diarrhoea	Campylobacter Reference Lab, UK	50
56282	Clinical diarrhoea	Campylobacter Reference Lab, UK	50
56519	Clinical bloody diarrhoea	Campylobacter Reference Lab, UK	12
56832	Clinical vomiting	Campylobacter Reference Lab, UK	50
58473	Clinical diarrhoea	Campylobacter Reference Lab, UK	2
59161	Clinical vomiting	Campylobacter Reference Lab, UK	2
59214	Clinical	Campylobacter	untypeable

	diarrhoea	Reference Lab, UK	
59364	Clinical diarrhoea	Campylobacter Reference Lab, UK	31
59424	Clinical vomiting	Campylobacter Reference Lab, UK	31
62567	Clinical diarrhoea	Campylobacter Reference Lab, UK	untypeable
62914	Clinical vomiting	Campylobacter Reference Lab, UK	untypeable
63326	Clinical diarrhoea	Campylobacter Reference Lab, UK	31
64555	Clinical bloody diarrhoea	Campylobacter Reference Lab, UK	31
12241	ovine	Campylobacter Reference Lab, UK	50
12481	ovine	Campylobacter Reference Lab, UK	50
13305	bovine	Campylobacter Reference Lab, UK	50
13713	Ox liver portion	Campylobacter Reference Lab, UK	2
12912	Ox liver portion	Campylobacter Reference Lab, UK	50
11818	chicken	Campylobacter Reference Lab, UK	50
11848	chicken	Campylobacter Reference Lab, UK	2
11974	chicken	Campylobacter Reference Lab, UK	44
12196	chicken	Campylobacter	50

		Reference Lab, UK	
12450	chicken	Campylobacter Reference Lab, UK	50
12487	chicken	Campylobacter Reference Lab, UK	50
12567	chicken	Campylobacter Reference Lab, UK	2
13082	chicken	Campylobacter Reference Lab, UK	50
13249	chicken	Campylobacter Reference Lab, UK	44
40217	chicken	Campylobacter Reference Lab, UK	5
13040	chicken	Campylobacter Reference Lab, UK	50
13411	chicken	Campylobacter Reference Lab, UK	44
47693	chicken	Campylobacter Reference Lab, UK	27
11919	chicken	Campylobacter Reference Lab, UK	2
11856	chicken	Campylobacter Reference Lab, UK	50
40209	chicken	Campylobacter Reference Lab, UK	5
11973	chicken	Campylobacter Reference Lab, UK	2
79260 (1771)	beach	Preston Public Health Laboratory	55
79309 (1772)	beach	Preston Public	untypeable

		Health Laboratory	
79196 (1773)	beach	Preston Public Health Laboratory	untypeable
79207 (1791)	beach	Preston Public Health Laboratory	2
79046 (1792)	beach	Preston Public Health Laboratory	-
79044 (1793)	beach	Preston Public Health Laboratory	5
B60	Clinical	Brazil	untypeable
K1	Clinical Watery diarrhoea	Aga Khan University, Pakistan	-
K4	Clinical Watery diarrhoea	Aga Khan University, Pakistan	-
K5	Clinical Watery diarrhoea	Aga Khan University, Pakistan	-
K6	Clinical Watery diarrhoea	Aga Khan University, Pakistan	52
K8	Clinical Watery diarrhoea	Aga Khan University, Pakistan	-
1099 (E)	chicken	Denmark	
922 (F)	chicken	Denmark	
1425 (G)	chicken	Denmark	
835-770 (H)	chicken	Denmark	
RM1221	chicken	USA	
44464	Clinical diarrhoea	Campylobacter Reference Lab, UK	37

Appendix 2. CDSs identified as absent or divergent in *C. jejuni* strain RM1221 using both a dynamic and constant cut-off value of 0.5.

CDS	Genbank annotation
cj0033	probable integral membrane protein
cj0300c	*modC*, probable molybdenum transport ATP-binding protein
cj0819	small hydrophobic protein
cj1122c	*wlaJ*, possible integral membrane protein
cj1138	probable galactosyltransferase
cj1418c	unknown
cj1427c	probable sugar-nucleotide epimerase/dehydratease
cj1432c	possible sugar transferase
cj1440c	probable sugar transferase
cj0815	unknown
cj0970	unknown
cj1143	*neuA1*, probable acylneuraminate cytidylyltransferase (CMP-N-acetylneuraminic acid synthetase)
cj1142	*neuC1*, probable N-acetylglucosamine-6-phosphate 2-epimerase/N-acetylglucosamine-6-phosphatase
cj1679	unknown
cj0139	possible endonuclease
cj1324	unknown
cj1434c	probable sugar transferase
cj1438c	probable sugar transferase
cj1333	unknown
cj1338c	*flaB*, flagellin B
cj1144c	unknown
cj1395	pseudogene
cj1296	unknown

cj1430c	probable nucleotide-sugar epimerase/dehydratase
cj1423c	possible sugar-phosphate nucleotidyltransferase
cj1421c	possible sugar transferase
cj1422c	possible sugar transferase
cj0055c	unknown
cj1308	*acpP4*, possible acyl carrier protein
cj1145c	unknown
cj1300	unknown
cj1431c	unknown
cj0816	unknown
cj0306c	*bioF*, probable 8-amino-7-oxononanoate synthase
cj0304c	*bioC*, possible biotin synthesis protein
cj0303c	*modA*, probable molybdate-binding lipoprotein
cj1259	*porA*, major outer membrane protein (MOMP)
cj1721c	possible outer membrane protein
cj0301c	*modB*, probable molybdenum transport system permease protein
cj1139c	probable galactosyltransferase
cj0302c	unknown
cj0305c	unknown
cj0259	*pyrC*, probable dihydroorotase
cj1136	probable galactosyltransferase
cj0056c	unknown
cj0170	unknown
cj0171	unknown
cj0260c	unknown
cj0264c	probable molybdopterin-containing oxidoreductase
cj0265c	probable cytochrome C-type haem-binding periplasmic protein
cj0295	possible acetyltransferase

cj0423	probable integral membrane protein
cj0424	probable acidic periplasmic protein
cj0425	probable periplasmic protein
cj0565	pseudogene
cj0566	unknown
cj0567	unknown
cj0568	unknown
cj0569	unknown
cj0628	probable lipoprotein
cj0629	possible lipoprotein
cj0818	probable lipoprotein
cj1055c	probable integral membrane protein
cj1137c	unknown
cj1140	unknown
cj1141	*neuB1*, probable N-acetylneuraminic acid synthetase
cj1255	possible isomerase
cj1297	unknown
cj1301	unknown
cj1309c	unknown
cj1321	probable transferase
cj1322	unknown
cj1323	unknown
cj1325	unknown
cj1326	unknown
cj1376	probable periplasmic protein
cj1415c	*cysC*, possible adenylylsulfate kinase
cj1416c	probable sugar nucleotidyltransferase
cj1417c	unknown
cj1419c	possible methyltransferase
cj1420c	unknown

cj1426c	unknown
cj1428c	*fcl*, probable fucose synthetase
cj1429c	unknown
cj1433c	unknown
cj1435c	unknown
cj1436c	probable aminotransferase
cj1437c	probable aminotransferase
cj1439c	*glf*, probable UDP-galactopyranose mutase
cj1441c	*kfiD*, probable UDP-glucose 6-dehydrogenase
cj1442c	unknown
cj1549c	probable type I restriction enzyme R protein
cj1550c	probable ATP/GTP-binding protein
cj1551c	probable type I restriction enzyme S protein
cj1552c	unknown
cj1553c	probable type I restriction enzyme M
cj1556	unknown
cj1677	probable lipoprotein
cj1678	possible lipoprotein
cj1722c	unknown
cj1723c	probable periplasmic protein
cj11828-03	LOS
cj11828-04	LOS
cj11828-06	LOS
cj11828-10	LOS
cj11828-11	LOS
cj43431-01	LOS region sequences
cj43438-01	LOS region sequences
cj460-03	Neal Golden ORF
cj460-04	Neal Golden ORF
cj81116-14	Subtractive hybridisation sequence from strain 81116

cj81176-02	virulence plasmid genes
cjP19-01	capsule region
cjX-01	capsule region
cjX-02	capsule region
cjX-03	capsule region
cjX-04	capsule region
cjX-05	capsule region

Appendix 3. Additional 41 CDSs identified as absent or divergent in *C. jejuni* strain RM1221 using a constant cut-off value of 0.5.

CDS	**Genbank annotation**
cj0008	unknown
cj0109	*exbB3*, probable exbB/tolQ family transport protein
cj0246c	probable MCP-domain signal transduction protein
cj0261c	unknown
cj 0294	unknown
cj0296c	*panD*, probable aspartate 1-decarboxylase precursor
cj0416	unknown
cj0417	unknown
cj0494	small hydrophobic protein
cj0755	*cfrA*, probable iron uptake protein (ferric receptor)
cj0799c	*ruvA*, probable Holliday junction DNA helicase
cj0814	unknown

cj0969	pseudogene
cj0987c	probable integral membrane protein
cj1051c	probable restriction modification enzyme
cj1299	acpP2, probable acyl carrier protein
cj1339c	*flaA*, flagellin A
cj1340c	unknown
cj1341c	Unknown
cj1394	probable fumarate lyase
cj1424c	*gmhA2*, probable phosphoheptose isomerase
cj1425c	possible sugar kinase
cj1448c	*kpsM*, probable capsule polysaccharide export system inner membrane protein
cj1520	unknown
cj1555c	unknown
cj1562	unknown
cj1724c	unknown
cj11828-05	LOS
cj11828-08	LOS
cj43431-02	LOS region sequences
cj43431-03	LOS region sequences
cj43438-02	LOS region sequences
cj43449-02	LOS region sequences
cj81116-03	Subtractive hybridisation sequence
cj81116-09	Subtractive hybridisation

	sequence
cj81116-23	Subtractive hybridisation sequence
cj81116-25	LOS region sequences
cj81116-26	LOS region sequences
cj81176-01	Plasmid ComB1virulence plasmid gene
cj81176-03	Plasmid ComB3 virulence plasmid gene
cjP19-02	capsule region

Appendix 4. CDSs correctly identified as absent in *C. jejuni* strain RM1221 using two data analysis methods. CDS shown in red were not identified using a dynamic cut-off value.

CDS	**Genbank annotation**
cj0056c	unknown
cj0170	unknown
cj0171	unknown
cj0260c	unknown
cj0264c	probable molybdopterin-containing oxidoreductase
cj0265c	probable cytochrome C-type haem-binding periplasmic protein
cj0295	possible acetyltransferase
cj0423	probable integral membrane protein
cj0424	probable acidic periplasmic protein
cj0425	probable periplasmic protein
cj0565	pseudogene

cj0566	unknown
cj0567	unknown
cj0568	unknown
cj0569	unknown
cj0628	probable lipoprotein
cj0629	possible lipoprotein
cj0818	probable lipoprotein
cj1055c	probable integral membrane protein
cj1137c	unknown
cj1140	unknown
cj1141	*neuB1*, probable N-acetylneuraminic acid synthetase
cj1255	possible isomerase
cj1297	unknown
cj1301	unknown
cj1309c	unknown
cj1321	probable transferase
cj1322	unknown
cj1323	unknown
cj1325	unknown
cj1326	unknown
cj1376	probable periplasmic protein
Cj1415c	*cysC*, possible adenylylsulfate kinase
cj1416c	probable sugar nucleotidyltransferase
cj1417c	unknown
cj1419c	possible methyltransferase
cj1420c	unknown
cj1426c	unknown
cj1428c	*fcl*, probable fucose synthetase
cj1429c	unknown

cj1433c	unknown
cj1435c	unknown
cj1436c	probable aminotransferase
cj1437c	probable aminotransferase
cj1439c	*glf*, probable UDP-galactopyranose mutase
cj1441c	*kfiD*, probable UDP-glucose 6-dehydrogenase
cj1442c	unknown
cj1549c	probable type I restriction enzyme R protein
cj1550c	probable ATP/GTP-binding protein
cj1551c	probable type I restriction enzyme S protein
cj1552c	unknown
cj1553c	probable type I restriction enzyme M protein
cj1556	unknown
cj1677	probable lipoprotein
cj1678	possible lipoprotein
cj1722c	unknown
cj1723c	probable periplasmic protein
cj0008	unknown
cj0417	unknown
cj1520	unknown
cj1555c	unknown

Appendix 5. CDSs incorrectly identified as absent in *C. jejuni* strain RM1221 using two data analysis methods. CDS shown in red was not identified using a dynamic cut-off value.

CDS	Genbank annotation
cj0033	probable integral membrane protein
cj0300	*modC*, probable molybdenum transport ATP-binding protein
cj0819	small hydrophobic protein
cj1122c	*wlaJ*, possible integral membrane protein
cj1138	probable galactosyltransferase
cj1418c	unknown
cj1427c	probable sugar-nucleotide epimerase/dehydratease
cj1432c	possible sugar transferase
cj1440c	probable sugar transferase
cj1340c	unknown

Appendix 6. CDSs that are absent or divergent in all putatively non-pathogenic beach isolates

CDS	Genbank annotation
Cj1122c	*wlaJ*, possible integral membrane protein
Cj0033	probable integral membrane protein
Cj0139	possible endonuclease
Cj0177	probable lipoprotein
Cj0178	possible outer membrane siderophore receptor

Cj0179	*exbB1*, biopolymer transport protein
Cj0181	*tonB1*, possible tonB transport protein
Cj0259	*pyrC*, probable dihydroorotase
Cj0261c	unknown
Cj0295	possible acetyltransferase
Cj0297c	*panC*, probable pantoate--beta-alanine ligase
Cj0298c	*panB*, probable 3-methyl-2-oxobutanoate hydroxymethyltransferase
Cj0299	possible periplasmic beta-lactamase
Cj0300c	*modC*, probable molybdenum transport ATP-binding protein
Cj0380c	unknown
Cj0424	probable acidic periplasmic protein
Cj0425	probable periplasmic protein
Cj0501	pseudogene
Cj0565	pseudogene
Cj0569	unknown
Cj0617	unknown
Cj0628	probable lipoprotein
Cj0629	possible lipoprotein
Cj0737	probable periplasmic protein
Cj0755	*cfrA*, probable iron uptake protein (ferric receptor)
Cj0765c	hisS, probable histidyl-tRNA synthetase
Cj0818	probable lipoprotein
Cj0970	unknown
Cj1138	probable galactosyltransferase
Cj1139c	probable galactosyltransferase
Cj1142	*neuC1*, probable N-acetylglucosamine-6-phosphate 2-epimerase/N-acetylglucosamine-6-

	phosphatase
Cj1143	*neuA1*, probable acylneuraminate cytidylyltransferase (CMP-N-acetylneuraminic acid synthetase)
Cj1144c	unknown
Cj1365c	probable secreted serine protease
Cj1376	probable periplasmic protein
Cj1395	pseudogene
Cj1550c	probable ATP/GTP-binding protein
Cj1551c	probable type I restriction enzyme S protein
Cj1555c	unknown
Cj1560	probable membrane protein
Cj1585c	probable oxidoreductase
Cj1677	probable lipoprotein
Cj1721c	possible outer membrane protein
Cj1722c	unknown
Cj1723c	probable periplasmic protein
Cj1725	probable periplasmic protein
Cj1726c	*metA*, probable homoserine O-succinyltransferase
Cj1727c	*metY*, possible O-acetylhomoserine (thiol)-lyase

Molecular Microbiology (2005) 55(1), 90–103 doi:10.1111/j.1365-2958.2004.04374.x

Analysis of *Campylobacter jejuni* capsular loci reveals multiple mechanisms for the generation of structural diversity and the ability to form complex heptoses

Andrey V. Karlyshev,[1] Olivia L. Champion,[1] Carol Churcher,[2] Jean-Robert Brisson,[3] Harold C. Jarrell,[3] Michel Gilbert,[3] Denis Brochu,[3] Frank St Michael,[3] Jianjun Li,[3] Warren W. Wakarchuk,[3] Ian Goodhead,[2] Mandy Sanders,[2] Kim Stevens,[2] Brian White,[2] Julian Parkhill,[2] Brendan W. Wren[1]* and Christine M. Szymanski[3]*

[1]*Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London WCIE 7HT, UK.*

[2]*The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.*

[3]*Institute for Biological Sciences, National Research Council of Canada, Ottawa, Canada, K1A 0R6.*

Summary

We recently demonstrated that *Campylobacter jejuni* produces a capsular polysaccharide (CPS) that is the major antigenic component of the classical Penner serotyping system distinguishing *Campylobacter* into >60 groups. Although the wide variety of *C. jejuni* serotypes are suggestive of structural differences in CPS, the genetic mechanisms of such differences are unknown. In this study we sequenced biosynthetic *cps* regions, ranging in size from 15 to 34 kb, from selected *C. jejuni* strains of HS:1, HS:19, HS:23, HS:36, HS:23/36 and HS:41 serotypes. Comparison of the determined *cps* sequences of the HS:1, HS:19 and HS:41 strains with the sequenced strain, NCTC11168 (HS:2), provides evidence for multiple mechanisms of structural variation including exchange of capsular genes and entire clusters by horizontal transfer, gene duplication, deletion, fusion and contingency gene variation. In contrast, the HS:23, HS:36 and HS:23/36 *cps* sequences were highly conserved. We report the first detailed structural analysis of 81-176 (HS:23/36) and G1 (HS:1) and refine the previous structural interpretations of the HS:19, HS:23, HS:36 and HS:41 serostrains. For the first time, we demonstrate the commonality and function of a second heptose biosynthetic pathway for *Campylobacter* CPS independent of the pathway for lipooligosaccharide (LOS) biosynthesis and identify a novel heptosyltransferase utilized by this alternate pathway. Furthermore, we show the retention of two functional heptose isomerases in *Campylobacter* and the sharing of a phosphatase for both LOS and CPS heptose biosynthesis.

Accepted 23 August, 2004. *For correspondence. E-mail christine.szymanski@nrc-cnrc.gc.ca; Tel. (+1) 613 990 1569; Fax (+1) 613 952 9092; and E-mail brendan.wren@lshtm.ac.uk; Tel. (44) 0 207 927 2288; Fax (44) 0 207 637 4314.

Introduction

Capsular polysaccharides (CPSs) are found on the surface of a large number of bacterial species. CPSs are known to play an important role in bacterial survival and persistence in the environment and often contribute to pathogenesis (Roberts, 1996). In addition, through structural variation, the potential to mimic host cell antigens, and the ability to resist innate mechanisms such as phagocytosis and complement-mediated killing, bacterial CPSs play a role in evasion of host immune responses.

Assembly of these surface polysaccharides is remarkably conserved in bacteria (reviewed in Whitfield and Roberts, 1999). Nucleotide diphosphate sugars are synthesized in the cytoplasm and sequentially added by glycosyltransferases to an undecaprenyl pyrophosphate carrier anchored in the membrane. Many Gram-negative bacteria flip the assembled polysaccharide across the membrane using an ABC transporter consisting of the transmembrane channel, KpsM, and the ATPase, KpsT. These transporters form a complex with four to five additional Kps proteins to ensure proper translocation of the polysaccharide to the bacterial surface. The genetic organization of the capsule gene clusters is also conserved in bacteria with *kps* transporter genes flanking polysaccharide biosynthesis genes, an organization conducive to genetic recombination and reorganization.

Production of CPSs by the enteric pathogen, *Campylobacter jejuni*, remained unnoticed until we initiated sequencing of its genome in 1998. Identification of *kps* genes potentially involved in capsule biosynthesis during sample sequencing of the shot-gun library of NCTC11168 (Karlyshev *et al.*, 1999) prompted a systematic genetic analysis of the corresponding locus and resulted in identification of CPSs in a number of strains of *C. jejuni* (Karlyshev *et al.*, 2000a). These molecules were found to be the

major antigens in the Penner serotyping scheme (Karlyshev *et al.*, 2000a). Similar experiments performed on *C. jejuni* 81-176 confirmed these findings and demonstrated a role for the capsule in serum resistance, epithelial cell invasion and diarrhoeal disease (Bacon *et al.*, 2001). Subsequent characterization of the CPSs by Alcian blue staining (Karlyshev and Wren, 2001) led to the visualization of capsule by electron microscopy (Karlyshev *et al.*, 2001). These experiments suggested that the previously described high-molecular-weight 'lipopolysaccharides' (HMW LPSs) of *C. jejuni* are in fact CPSs.

Recently, the CPS structure of NCTC11168 was determined to contain 6-O-methyl-D-*glycero*-α-L-*gluco*-heptose, β-D-glucuronic acid modified with 2-amino-2-deoxyglycerol, β-D-GalfNAc and β-D-ribose (St Michael *et al.*, 2002). There are several notable features encoded by the *cps* locus of NCTC11168 (St Michael *et al.*, 2002) that correlate well with the published structure: homologues of the GDP-D-*glycero*-D-*manno*heptose pathway (GmhA2, HddA and HddC) (Valvano *et al.*, 2002); homologue of the UDP-glucose dehydrogenase, Udg, involved in the formation of UDP-glucuronic acid (Sieberth *et al.*, 1995); and a homologue of the UDP-pyranose mutase, Glf, predicted to catalyse the reversible conversion of pyranoses to furanoses (Nassau *et al.*, 1996) and shown to cause loss of CPS when mutated in NCTC11168 (St Michael *et al.*, 2002).

In a series of earlier publications on the structural analysis of HMW LPSs (now realized as CPSs) of *C. jejuni*, it was shown that these molecules are highly heterogeneous (Moran *et al.*, 2000). Microarray hybridization analysis also demonstrated some differences in the CPS-related genes between the strains of various serotypes (Dorrell *et al.*, 2001). However, hybridization analysis does not allow detailed investigation of gene content. Sequencing of the *C. jejuni* NCTC11168 genome revealed that the guanosine cytosine (GC) content of the *cps* locus (*cj1415–cj1442*) is lower (26.5%) in comparison to that for the entire genome (30.6%) suggesting that this locus was acquired through horizontal gene transfer (Parkhill *et al.*, 2000). In addition, the biosynthetic region of the *cps* locus is also prone to phase variation because of the presence of six genes with homopolymeric tracts (Parkhill *et al.*, 2000). It was subsequently shown that CPS from 81–176 undergoes antigenic variation at high frequency (Bacon *et al.*, 2001) and that CPS from NCTC11168 can vary in structure (Szymanski *et al.*, 2003). However, the genetic mechanisms underlying the structural heterogeneity and antigenic variation remain unknown.

In the current study, we determined the full nucleotide sequence of *cps* regions from six *C. jejuni* strains of similar and different serotypes and compared these with the sequenced strain NCTC11168 (HS:2). The results demonstrate heterogeneity in the biosynthetic *cps* genes and suggest widespread genetic exchange. Extensive structural studies, including high-resolution magic angle spinning (HR-MAS) nuclear magnetic resonance (NMR) spectroscopy, demonstrated polysaccharide heterogeneity in *Campylobacter* CPS and solved two new structures for G1 (HS:1) and 81-176 (HS:23/36). Structural and sequencing analysis also demonstrated the presence of additional CPS modifications and a large abundance of genes potentially involved in heptose biosynthesis in these clusters. This report describes the first comparative study of *C. jejuni cps* loci with structure and demonstrates the commonality and function of these complex heptose biosynthetic genes in this organism.

Results

Strains and sequencing strategy

Although *C. jejuni* is the major bacterial cause of gastrointestinal disease in developed countries, infection can also lead to the development of neuropathies such as Guillain–Barré syndrome (GBS). Two of the type strains selected in this study, HS:19 (Aspinall *et al.*, 1994a) and HS:41 (Lastovica *et al.*, 1997), have been isolated from patients with gastroenteritis, but represent a serotype commonly associated with GBS, while G1 is a human GBS isolate belonging to the HS:1 serogroup (Karlyshev *et al.*, 2000a). Strain 81-176 is a well-characterized strain isolated from a human outbreak and is highly virulent in human challenge studies (Black *et al.*, 1988). The analysis of CPS-related genes in the latter strain was of particular interest as it reacts with both HS:23- and HS:36-specific anti-sera, and the origin for this dual reactivity remained unclear. The HS:23 and HS:36 serotype reference strains were also investigated for comparison and these were all compared with the genome sequenced strain NCTC11168 (HS:2; Parkhill *et al.*, 2000).

The strategy used for sequencing the variable *cps* loci from the different strains is described in *Experimental procedures*. Briefly, conserved internal *cps* genes in a particular strain were identified by polymerase chain reaction (PCR) with primers specific to NCTC11168 *cps* genes. These primers were then used in long-range PCR in combination with primers specific to the flanking *kps* genes. In some cases, this resulted in overlapping PCR products, which, after sequencing using a standard shotgun procedure, produced a sequence of an entire biosynthetic *cps* region. When necessary, the gaps were closed using PCR with primers derived from the sequences of the long PCR products. Several overlapping reads were analysed before generating a consensus sequence. The diagram shown in Fig. S1 demonstrates the general strategy for amplification of *cps* clusters using long-range PCR. The overall summary of the *cps* sequencing results

Table 1. Biosynthetic *cps* regions of various strains of *C. jejuni*.

Strain	Serotype	Accession No. of nucleotide sequence	Size	% GC	No. of genes	Contingency genes	Sugar phosphate nucleotidyltransferases	Putative glycosyltransferases	No. of different sugar residues found[a]
NCTC11168	HS:2	AL139078	34180	26.5	28	6	2	8	4
NCTC12517	HS:19	BX545860	16727	26.1	13	2	1	4	2[b]
G1	HS:1	BX545859	15180	26.8	11	2	2	2	1[b]
81-176	HS:23/36	BX545858	24625	27.1	21	5	2	7	6
CCUG 10954	HS:23	AY332625	24627	27.0	21	6	2	7	6
ATCC 43456	HS:36	AY332624	24625	26.9	21	6	2	7	6
176.83	HS:41	BX545857	34118	27.2	30	5	2	8	4

a. According to structural analysis.
b. Additional uncharacterized labile substituent present.

is presented in Table 1. A schematic of the *cps* loci compared in this study is shown in Fig. 1. The predicted function of *cps* genes from strain NCTC11168 (HS:2) based on the published genome sequence (Parkhill *et al.*, 2000) and the recently published CPS structure (St Michael *et al.*, 2002) are presented in Table S1. CPS structures of NCTC11168 and the representative serostrains used in this study are shown in Fig. 2.

NCTC12517 (HS:19) CPS

NMR analysis of the HS:19 serostrain used in this study confirmed that the CPS contained the published disaccharide (Fig. 2; data not shown) (Aspinall *et al.*, 1994b) in addition to two acid-labile functional groups that were not reported previously. Both the phosphoramide modification recently described for NCTC11168 (Szymanski *et al.*, 2003) and an unknown labile sugar were observed during the analysis and are currently being characterized. As expected by the lack of heptose in the structure, the *cps* region of the HS:19 serostrain did not contain homologues of the heptose pathway, but did have the *udg* homologue (Table S2) correlating well with the presence of β-D-glucuronic acid which is further substituted with 2-amino-2-deoxyglycerol similar to NCTC11168. The genes responsible for glycerol modification are currently unknown.

G1 (HS:1) CPS

In contrast to NCTC11168 (HS:2) and the HS:19 serostrain, the biosynthetic region of strain G1 (HS:1) is the smallest (15 kb) and contains only 11 genes. Organization of the genes from *cj1415* to *cj1421* in this strain is similar to that of serostrain HS:19 and NCTC11168. However, the remaining genes have no counterparts in the corresponding regions of these strains (Fig. 1; Table S3). In addition, the *cps* locus of G1 does not encode homologues of UDP-glucose 6-dehydrogenase or the heptose pathway

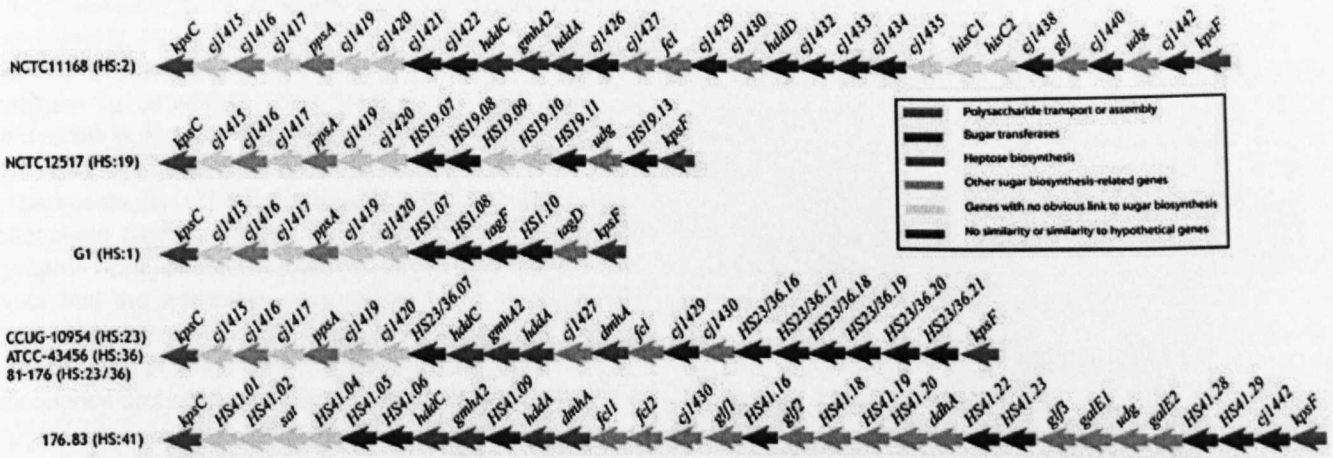

Fig. 1. Graphical representation of the sequenced *cps* biosynthetic regions. In cases with high level of similarity between putative gene products (usually with BLAST *E*-values below 1e-30), the genes were given names of counterparts found in other bacteria. When no such similarity was found, the genes were assigned the names of respective genes from strain NCTC11168. The genes with no similarity to either NCTC11168 or other bacteria are given strain-specific systematic names. The *cps* clusters of serostrains HS:23 and HS:36 are almost identical to that of strain 81-176 (see text).

HS:1 [*P*-4)-α-D-Gal-1-3-Gro-1-]$_n$

NCTC11168 (HS:2) {6-O-Me}-D-*glycero*-α-L-*glc*-Hep-1↓3
[-2)-β-D-Rib*f*-1-5-β-D-Gal*f*NAc-1-4-α-D-GlcA6{NGro}-1-]$_n$
3↑
{OP=O(NH$_2$)OMe}

HS:19 [-4)-β-D-GlcA6(NGro)-1-3-β-D-GlcNAc-1-]$_n$

HS:23/HS:36 [-3)-β-D-GlcNAc-1-3-α-D-Gal-1-2-6d-α-D-*altro*-Hep-1-]$_n$
-6d-3Me-α-D-*altro*-Hep-1-]$_n$
-D-*glycero*-α-D-*altro*-Hep-1-]$_n$
-3Me-D-*glycero*-α-D-*altro*-Hep-1-]$_n$

HS:41 [-2)-β-L-Ara*f*-1-2-β-D-6d-*altro*-Hep*f*-1-2-β-L-6d-Alt*f*-1-]$_n$
-α-D-Fuc*f*-1-]$_n$

Fig. 2. Summary of the representative *C. jejuni* capsular polysaccharide structures described in this study. The CPS structures of the heat-stable (HS) Penner type strains HS:1, HS:19, HS:23 and HS:36 have been reviewed by Moran *et al.* (2000). The structures of NCTC11168 (St Michael *et al.*, 2002) and HS:41 (Hanniffy *et al.*, 1999) CPS have recently been described. NCTC11168 modifications that have been demonstrated to be phase-variable are indicated in curved brackets. Sugars are shown in pyranose configurations unless otherwise noted. *P*, phosphate; Gal, galactose; Gro, glycerol; Me, methyl; Hep, heptose; Rib, ribose; GalNAc, *N*-acetylgalactosamine; GlcA6, glucuronic acid; NGro, aminoglycerol; OP=O(NH$_2$)OMe, phosphoramide; GlcNAc, *N*-acetylglucosamine; Ara, arabinose; Alt, altrose; Fuc, fucose.

(Table S3) and thus the strain may not have the ability to synthesize glucuronic acid or heptose unless the genes are located elsewhere on the chromosome. In contrast, G1 contains a potential *tagD* homologue encoding a glycerol-3-phosphate cytidylyltransferase necessary for the formation of CDP-glycerol (Table S3) (Pooley *et al.*, 1991). The strain also encodes a TagF homologue, which transfers glycerol-phosphate residues from CDP-glycerol (Schertzer and Brown, 2003). Therefore, genetic analysis suggests that the repeating unit of this CPS may contain glycerol-phosphate residues.

Indeed, the HS:1 serostrain was reported to contain glycerol-1-phosphate residues alternating with galactose in the repeating unit (Fig. 2; MacDonald, 1993). The NMR spectra of G1 (Fig. S2) revealed that the structure of this CPS is consistent with the HS:1 structure. However, extensive NMR analysis by COSY, TOCSY, NOESY and HMQC detected galactose and two additional acid-labile groups similar to those observed for HS:19 in both G1 and HS:1 strains. Again, the phosphoramide was confirmed to be one of these modifications by the strong correlation between the OP=O(NH$_2$)OMe ^{1}H resonance at 3.8 p.p.m. (indicated in Fig. S2) and the phosphoramide ^{31}P resonance at 14 p.p.m., in the ^{31}P HMQC spectra. ^{31}P NMR experiments for G1 also confirmed the presence of a phosphate diester linkage, consistent with the reported glycerol-1-phosphate structure in HS:1 (data not shown). Note that the additional anomeric proton resonance at 5.44 p.p.m. in the HR-MAS spectrum of G1 was absent in the partially purified CPS sample suggesting that this resonance probably came from the medium used.

CCUG 10954 (HS:23), ATCC 43456 (HS:36) and 81-176 (HS:23/36) CPS

The *cps* loci of the HS:23 and HS:36 serostrains and of strain 81-176 (which reacts with both HS:23 and HS:36 anti-sera) all have the same gene content and colinear representation (Fig. 1 and Table 1). Pair-wise alignments of the CPS biosynthetic regions (24.6 kb) of these strains show that the HS:23 and HS:36 serostrains share 97.6% DNA sequence identity while strain 81-176 shares 97.6% and 98.9% identity with HS:23 and HS:36 respectively. All pair-wise comparisons showed >93% protein sequence identity except for HS23.08 (HddC) which shared 87.9% and 86.6% identity with the corresponding gene products in HS:36 and 81-176 respectively. Analysis of the predicted products, encoded by the *cps* regions of serostrains HS:23 and HS:36 and of strain 81-176 (Table S4), demonstrate a potential for deoxyheptose biosynthesis because of the presence of genes *cj1423 (hddC)*, *cj1424 (gmhA2)*, *cj1425 (hddA)* and the new gene, *dmhA*, inserted between genes *cj1427* and *fcl* (Fig. 1). The DmhA homologue is suggested to be involved in conversion of heptose to deoxyheptose in *Yersinia pseudotuberculosis* (Pacinelli *et al.*, 2002).

The published CPS structures of HS:23 and HS:36 were found to contain repeating units of α-D-galactose, β-D-GlcNAc and D-*glycero*-D-*altro*-heptose or deoxyheptose variants with and without methyl groups (Fig. 2; Aspinall *et al.*, 1992). However, it was reported that the D-*glycero*-D-*altro*-heptose variant was not detected in the HS:23 serostrain (Aspinall *et al.*, 1992). In this study, HR-MAS spectra of cells and NMR spectra of the partially purified CPS from strain 81-176 demonstrated similar sugar resonances with the HS:23 and HS:36 serostrains (Fig. 3). In all the spectra, the characteristic OMe signal at 3.5 p.p.m. and NAc resonance at 2.05 p.p.m. were observed. In the ^{1}H NMR spectra of the CPS, the anomeric region (4.7–5.5 p.p.m.) of the HS:23 serostrain was the simplest with three anomeric resonances (Fig. 3A). The anomeric region for HS:36 and 81-176 was more complex (Fig. 3B and C) with the spectrum of 81-176 being the most complex.

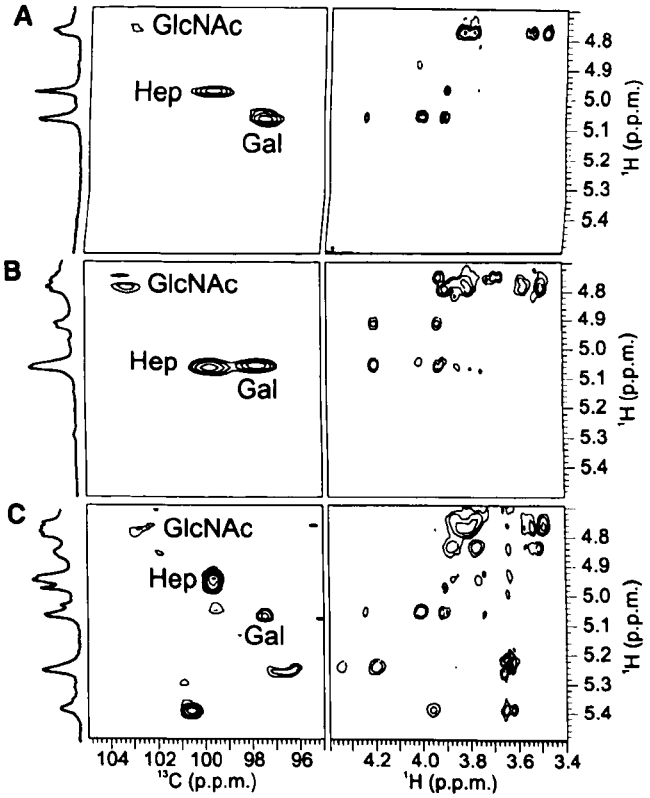

Fig. 3. NMR spectra of *C. jejuni* strains CCUG 10954 (HS:23 serostrain), ATCC 43456 (HS:36 serostrain) and 81-176 (HS:23/HS:36). NMR spectrum and HMQC and TOCSY spectra for CCUG 10954 (A), for ATCC 43456 (B) and for 81-176 (C) respectively. NMR spectra were acquired at 40°C using partially purified CPS. For the HMQC spectra, the anomeric region is shown, along with the ^{1}H NMR spectrum. For the TOCSY, the mixing time was 90 ms. Cross-peaks between signals in the anomeric region (4.7–5.5 p.p.m.) and the sugar ring region (3.4–4.4 p.p.m.) are shown.

2D-NMR experiments were performed to identify the sugar resonances (Fig. 3). The HMQC and TOCSY spectra for the HS:23 serostrain (Fig. 3A) confirmed that the anomeric resonances at 5.06, 4.97 and 4.77 p.p.m. corresponded to Gal, Hep and GlcNAc respectively (Aspinall *et al.*, 1992). In the HMQC spectrum, the C-6 cross-peaks of the 6-deoxy-heptose were observed at 34.8 p.p.m. (^{13}C) and 2.06 and 1.71 p.p.m. (^{1}H).

For the HS:36 serostrain, three anomeric resonances were also observed by HMQC and TOCSY experiments on the CPS (Fig. 3B). The ^{1}H resonance at 4.92 p.p.m. was confirmed to be a non-anomeric resonance using HMQC. While the anomeric carbon resonances had similar chemical shifts, the proton anomeric resonance of the heptose residue was different, probably because of different structural motifs on the heptose residue. In the TOCSY spectrum, the anomeric resonances at 4.76 and 5.06 p.p.m. exhibited connectivities that were similar to those observed for the HS:23 serostrain, indicating the presence of similar sugars in both serostrains. In the HMQC spectrum, resonances characteristic of a 6-deoxy-heptose could not be observed, indicating that for this serostrain this modification was not predominant (see contingency gene analysis below).

The HMQC and TOCSY spectra for strain 81-176 (Fig. 3C) showed correlation patterns similar to those observed for the HS:23 serostrain for the Gal, Hep and *GlcNAc anomeric resonances*, again indicating similar *sugar structures to those of HS:23 and HS:36. Comparison* of the NOESY spectra (data not shown) established that serostrains HS:23 and HS:36 and strain 81-176 exhibited similar NOE patterns for the Gal, Hep and GlcNAc residues, a result that is consistent with the conclusions arrived at from the analysis of the TOCSY experiments. These observations are in agreement with the conservation of the *cps* genes for these three strains. However, structural analysis of 81-176 also demonstrated the presence of additional resonances indicating the presence of a more complex repeating unit or the presence of another polysaccharide structure as was previously suggested (Bacon *et al.*, 2001). The phosphoramide modification observed for NCTC11168 was also observed for strain 81-176 and serostrain HS:36, but not for serostrain HS:23.

176.83 (HS:41) CPS

The major and minor components of CPS isolated from the HS:41 serostrain were described to contain β-L-arabinose, 6-deoxy-β-D-*altro*heptose, 6-deoxy-β-L-altrose and α-D-fucose all in the furanose form (Fig. 2, Hanniffy *et al.*, 1999; Szymanski *et al.*, 2003). NMR analysis demonstrated that the CPS of the sequenced strain used in this study is consistent with the published structure (data not shown). Interestingly, sequencing results from this strain (Table S5) show the *cps* region to be quite outstanding in that it lacks the *cj1415–cj1420* genes conserved in the other strains. However, three heptose-related genes in the middle of the *cps* locus (*gmhA2*, *hddA* and *hddC*) are almost identical to those in NCTC11168 (Fig. 1; Table S5). The mosaic patterns of similarity and divergence indicate that these *cps* regions have a diverse recent ancestry, suggesting that recombination between different *cps* clusters has occurred. Sequencing also demonstrated three UDP-pyranose mutase (*glf*) homologues, consistent with having three of the CPS sugars in the furanose form (note that arabinose is a pentose and therefore is naturally in the furanose configuration). The presence of genes *gmhA2*, *hddA*, *hddC* and *dmhA* (Fig. 1) is consistent with the presence of deoxyheptose in the CPS (Fig. 2). Additional sugar dehydratases will be required for the biosynthesis of fucose and deoxyaltrose and putative homologues are observed in Table S5.

Variation in the contingency genes

The biosynthetic *cps* locus of *C. jejuni* NCTC11168 was

Table 2. Comparison of the contingency genes in the biosynthetic *cps* region of serostrains HS:23, HS:36 and strain 81-176.

	Gene					
Strain	*cj1420*	*HS23/36.07*	*dmhA*	*cj1429*	*HS23/36.17*	*HS23/36.20*
81-176 (HS:23/HS:36)	**G9**	G9/**G10**	**(on)**	G9/**G10**	**G9**	G8/**G9**
CCUG 10954 (HS:23)	G8/**G9**	**G8**	**G9**	G8/**G9**	**G9**	**G8**
ATCC 43456 (HS:36)	G8/**G9**	G8/**G9**	G9/G10	G9/G10	**G9**	G9/**G10**
Gene 'ON'	G9	G9	G9	G9	G9	G9

The number in bold indicates the most frequent variant. For strains CCUG 10954 (HS:23) and ATCC 43456 (HS:36) the 'most frequent variant' is defined as the one corresponding to the strongest signal on a DNA sequencing electrophoregram when we sequenced a PCR product using chromosomal DNA isolated from a confluent plate. For strain 81-176 (HS:23/HS:36) the most frequent variant is determined from sequence analysis of multiple clones. Gene *dmhA* is not phase-variable in strain 81-176.

found to contain six genes with homopolymeric G tracts potentially prone to phase variation (Parkhill *et al.*, 2000). We surveyed the 'ON' and 'OFF' states of these gene homologues in the other strains and found that most of the genes tested are predominantly in the 'ON' state, although many are demonstrated to vary (data not shown). It is possible that such modulation may explain the presence of variant structures in the HS:41 serostrain and in serostrains HS:23 and HS:36 compared with strain 81-176 (HS:23/36). We therefore examined the latter three strains in more detail because they share the same gene content (see above), yet produce capsules with slight differences in CPS structure (Fig. 3). As these three strains share >95% gene identity in their *cps* biosynthetic regions, phase-variable genes could be responsible for the differential expression of deoxyheptose and phosphoramide observed in this study, i.e. HS:23 (Gal, GlcNAc, Hep, deoxyhep), HS:36 (Gal, GlcNAc, Hep, phosphoramide) and 81-176 (Gal, GlcNAc, Hep, deoxyhep, phosphoramide). There are six contingency genes in the *cps* cluster of the HS:23 and HS:36 serostrains but only five in 81-176 as the *dmhA* homologue (ORF#12) is not phase-variable (Table 2). DmhA has been shown to be involved in deoxyheptose synthesis in *Yersinia* and interestingly, in this study, the *dmhA* homologue is functional in 81-176 and HS:23, but variable in HS:36. This may correspond to the detection of deoxyheptose in 81-176 and HS:23 and the difficulty in detecting this heptose variant in HS:36. In the HS:23 serostrain, two 'OFF' genes (*HS:23.07* and *HS:23.20*) show high sequence similarity with the putative glycosyltransferase (*cj1422c*) from NCTC11168 and may play a role in adding the missing phosphoramide. However, function of these contingency genes must be proven experimentally.

Analysis, mutagenesis and complementation of conserved heptose genes from NCTC11168

The *C. jejuni gmhA2* (sedoheptulose-7-phosphate isomerase), *hddA* (D,D-heptose-7-phosphate kinase) and *hddC* (D,D-heptose-1-phosphate guanosyltransferase) gene homologues are conserved in all strains containing heptose in their CPS and also share similarity and colinearity with the respective genes involved in heptose biosynthesis in other bacteria (Fig. 4) (DeShazer *et al.*, 1998; Reckseidler *et al.*, 2001; Pacinelli *et al.*, 2002; Valvano

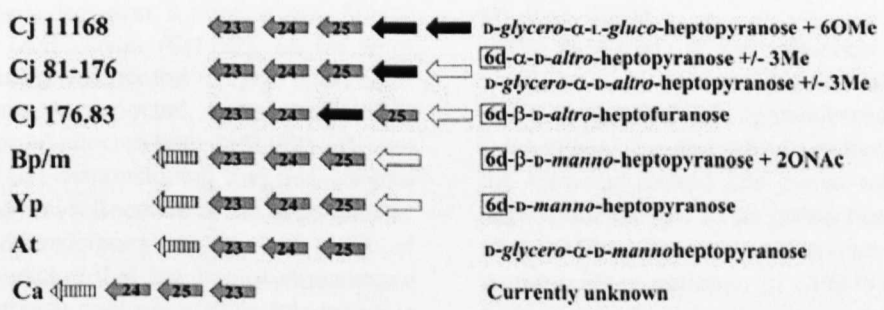

Fig. 4. Evidence for horizontal gene transfer of putative GDP-heptose pathways among bacteria. The top four matches for BlastX analysis of 23: Cj1423 (HddC, nucleotidyltransferase); 24: Cj1424 (GmhA2, isomerase); and 25: Cj1425 (HddA, kinase) were to: Bp/m: *Burkholderia pseudomallei/mallei*; Yp: *Yersinia pseudotuberculosis*; At: *Aneurinibacillus theromoaerophilus*; and Ca: *Clostridium acetobutylicum*. A schematic of the respective heptose gene clusters, shown by grey arrows, adjacent to the configuration of the heptose(s) found in the bacterial polysaccharide structures is shown. An additional gene shown by open arrows (DmhA, dehydratase) was not in the *C. jejuni* NCTC11168 genome but was found in the *cps* loci of strains 81-176 and 176.83 and is suggested to be involved in deoxyheptose formation (Pacinelli *et al.*, 2002), which correlates with the observed deoxy configurations shown by open boxes. Black arrows represent additional gene insertions while hatched arrows represent the Cj1152c phosphatase homologue that is located in the LOS heptose cluster in *C. jejuni* NCTC11168. Note that *hddC*, *gmhA2*, *hddA* and *dmhA* correspond to *wcbM*, *gmhA2*, *wcbL* and *wcbK* in the *Burkholderia* species.

et al., 2002). We therefore mutated these genes in NCTC11168 to examine their role in CPS heptose biosynthesis. Genome sequencing of NCTC11168 identified a second heptose gene cluster (*cj1148–cj1152*) involved in lipooligosaccharide (LOS) biosynthesis (Parkhill *et al.*, 2000; C.M. Szymanski, unpubl. obs.). Thus, for completeness, both heptose gene clusters were compared and both LOS and CPS were examined for structural changes. NCTC11168 has two copies of the heptose isomerase gene, *gmhA* (*cj1149* located in the LOS gene cluster) and *gmhA2* (*cj1424* located in the CPS gene cluster). Mutation of either gene in *C. jejuni* did not have an effect on LOS or CPS (data not shown) suggesting that both products were functional and compensated for each other. To test this hypothesis, we complemented *Escherichia coli* χ_{711} lacking *gmhA* with *C. jejuni gmhA* or *gmhA2*. The mass of the truncated *E. coli* LOS core [1394 and 1516 amu (+PEtn)] was restored back to wild type [2970 and 3093 amu (+PEtn); Fig. S3 and Table S6] in the presence of either *C. jejuni* isomerase and coincided with a slower LOS migration on deoxycholate-PAGE (Fig. S3). Mutation of both *gmhA* and *gmhA2* in *C. jejuni* resulted in the loss of CPS 6-O-methylheptose (Fig. 5C) demonstrated by the disappearance of the heptose anomeric 'a' and methyl resonance 'x' and the truncation of the LOS core (Fig. 5D; Table S7) compared with wild-type CPS (Fig. 5A) and LOS (Fig. 5B; Table S7). These results confirm that both isomerases are functional and involved in both the CPS and LOS heptose pathways.

In contrast, individual mutations of the *C. jejuni* NCTC11168 kinase (*hddA*) and nucleotidyltransferase (*hddC*) homologues resulted in loss of the CPS 6-O-methylheptose (Fig. 5E and I) without altering the LOS mass (Fig. 5F and J; Table S7) demonstrating their role solely in CPS heptose biosynthesis. Unlike other bacteria that contain this alternate heptose pathway, *C. jejuni* lacks a phosphatase homologue (GmhB) in its CPS cluster (Fig. 4, hatched arrows). However, a phosphatase homologue exists in the LOS cluster (Cj1152), so we were interested in determining whether this enzyme could function for both pathways. As expected, mutagenesis of *C. jejuni* NCTC11168 *gmhB* affected both CPS (Fig. 5G) and LOS (Fig. 5H; Table S7) demonstrating that this enzyme is shared by both pathways. Because of the large number of putative glycosyltransferases in the *cps* locus of NCTC11168, we predicted that the heptosyltransferase for this alternate pathway may exist within this locus. In order to identify the NCTC11168 heptosyltransferase, it was necessary to sequentially inactivate all of the putative glycosyltransferases in the *cps* locus (Table S1) as homologues of this enzyme have not been previously described. The NCTC11168 glycosyltransferase mutants were examined by HR-MAS NMR for loss of the heptose anomeric (data not shown). Figure 5K demonstrates that mutagenesis of the putative glycosyltransferase, *cj1431c*, results in loss of CPS 6-O-methylheptose with no change to LOS (Fig. 5L; Table S7). From this analysis, we demonstrate that *cj1431c* is the CPS heptosyltransferase which we have designated *hddD*. A summary of the initial steps of heptose biosynthesis in *C. jejuni* NCTC11168 leading to formation of D-*glycero*-α-L-*gluco*heptose is shown in Fig. 6.

Discussion

In this study, we compared the sequences of capsule biosynthetic loci from six *C. jejuni* strains using a PCR amplification procedure dependent on the presence of highly conserved genes in this region. Comparison of the biosynthetic loci with CPS structural analyses demonstrated a good correlation between gene sequence and structure. The structural studies also identified additional CPS modifications, including the recently identified phosphoramide, which coincide with the presence of additional genes in the *cps* regions potentially involved in biosynthesis of these structures.

As CPSs from strains belonging to serotypes HS:23 and/or HS:36 are closely related in structure, comparative analysis of the corresponding gene clusters was of particular interest. Striking similarity between the *cps* regions of both the HS:23 and HS:36 serostrains and that of strain 81-176, which is of mixed HS:23/HS:36 serotype (Fig. 2), indicates their *cps* loci originated from a common ancestor. The variation in the heptoses and phosphoramide in the respective CPS structures may be attributed to the presence of phase-variable genes whose expression was demonstrated to differ between the three strains. Indeed, phase-variable expression of methyl, ethanolamine, aminoglycerol and phosphoramide groups on the CPS of strain NCTC11168 has recently been described and also shown to result in changes in antibody reactivity (Szymanski *et al.*, 2003).

The availability of the sequence data allowed further comparison of the *cps* regions with that of *C. jejuni* NCTC11168. Both highly conserved and variable genes were found. The biosynthetic genes that are proximal to the transport-related *kps* genes were more conserved, particularly the five to six genes near *kpsC*. Similarly, the study of *Streptococcus pneumoniae cps* loci also revealed a non-random variation of CPS-related genes, with the highest difference for those genes closest to the central region (Jiang *et al.*, 2001). However, the biosynthetic *cps* genes of strains HS:41 and NCTC11168, with the exception of *gmhA2*, *hddA* and *hddC*, were found to have very low levels of similarity. This finding indicates that the *cps* regions of these two strains are the most distantly related. An independent analysis of the *C. jejuni* NCTC11168 genome performed elsewhere (http://www.fut.es/~debb/

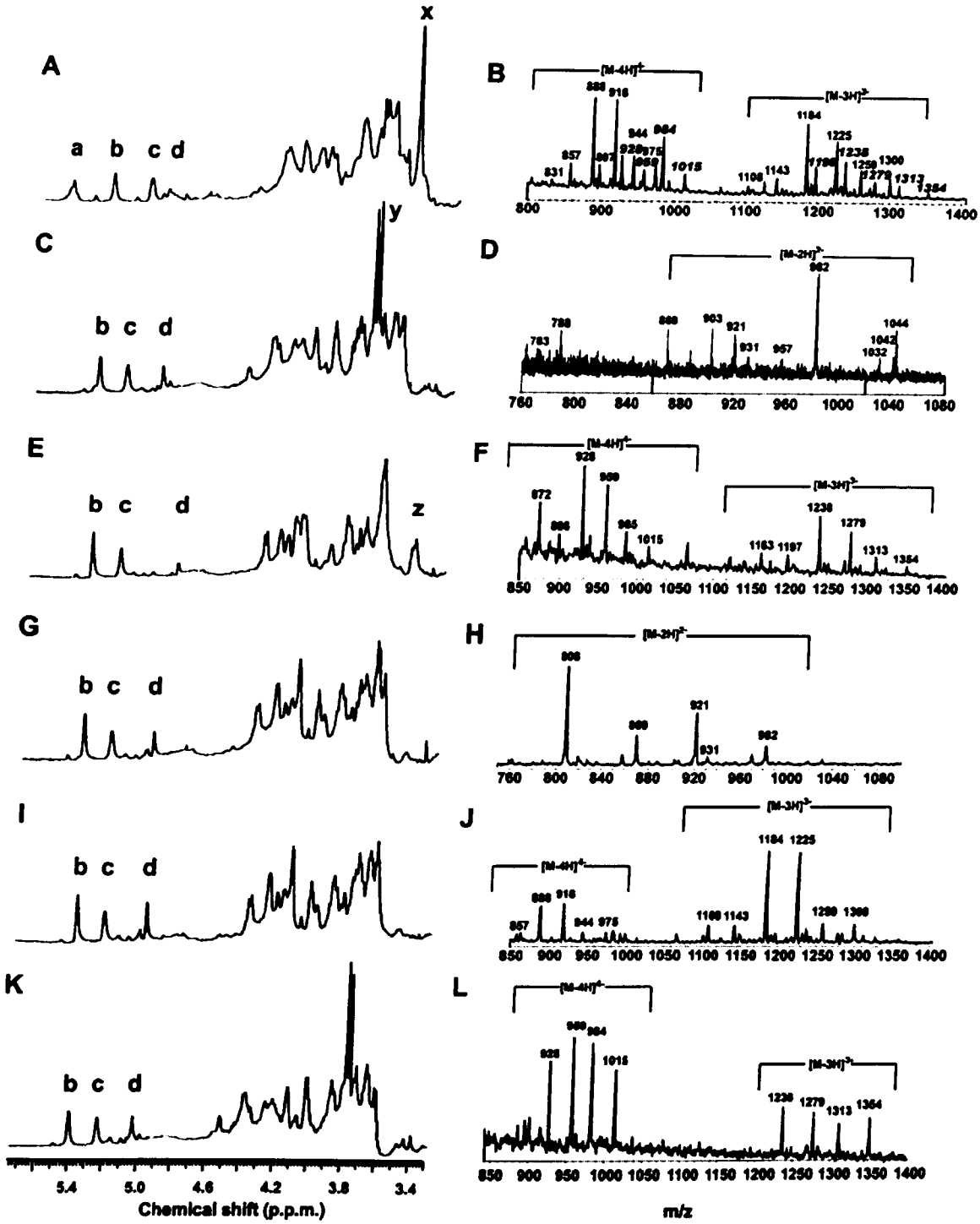

Fig. 5. HR-MAS NMR analysis of capsule (A, C, E, G, I and K) and CE-MS analysis of LOS (B, D, F, H, J and L) of *C. jejuni* NCTC11168 and heptose mutants. (A and B) *C. jejuni* NCTC11168; (C and D) *gmhA/gmhA2* mutant; (E and F) *hddA* mutant; (G and H) *gmhB* mutant; (I and J) *hddC* mutant and (K and L) *hddD* mutant. The anomeric protons of the capsular polysaccharide correspond to: (a) 6-O-methyl-D-*glycero*-α-L-*gluco*heptose; (b) β-D-ribose; (c) α-D-glucuronic acid amidated predominantly with 2-amino-2-deoxyglycerol and (d) β-D-GalfNAc. Other proton resonances of note include: (x) methyl group linked to heptose, (y) the phase-variable phosphoramide and (z) variable ethanolamine replacing 2-amino-2-deoxyglycerol on glucuronic acid. Note that the anomeric resonance 'd' corresponding to GalfNAc shifts due to the relative proportions of the phase-variable substitutions, aminodeoxyglycerol and ethanolamine, on glucuronic acid.

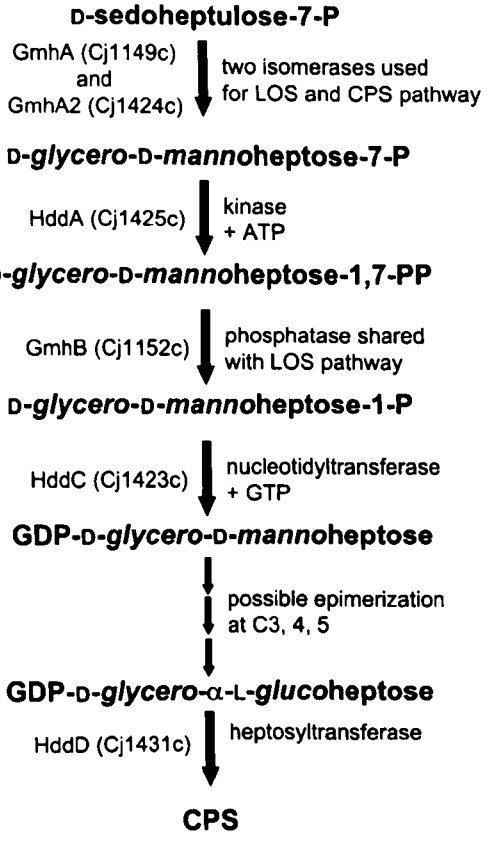

Fig. 6. Summary of the capsular heptose biosynthetic pathway in *C. jejuni* NCTC11168.

HGT/) also predicted that a number of genes have been acquired via horizontal gene transfer, including many from the biosynthetic *cps* region (e.g. *cj1432–cj1442*). These data suggest that exchange of the genes present in the biosynthetic *cps* regions of different *C. jejuni* may provide an efficient mechanism of structural variation in this pathogen. Evidence for this was recently shown in the *C. jejuni* GBS isolate GB11 (HS:2) which is similar to NCTC11168 (HS:2), but has horizontally acquired the entire LOS locus from the HS:19 serostrain (Gilbert *et al.*, 2004). Similar mechanisms of variation in CPS structure via genetic exchange were suggested for *Neisseria meningitidis* (Swartley *et al.*, 1997) and *S. pneumoniae* where more than 90 different capsular serotypes have been described in the latter (Dillard and Yother, 1994; Dillard *et al.*, 1995; Kamerling, 1999).

In addition to the mechanisms of variation attributed to homopolymeric tracts and horizontal gene transfer, extensive intragenomic variation in the *cps* regions can be observed. Some genes, e.g. *cj1421* and *cj1422* in NCTC11168 share long regions of identity, which may have resulted from gene duplication while in other strains only one copy of either of these genes is present. Other genes may have arisen from deletions resulting in the formation of hybrid genes. For example, the N-terminal region of gene H19.11 from serostrain HS:19 is similar to many *C. jejuni* glycosyltransferases, with the first 169 aa residues being almost identical to the N-terminal residues of the Cj1440 protein of NCTC11168. However, the C-terminus of HS19.11 has no similarity to the corresponding region of Cj1440, and resembles instead the Cj1438 glycosyltransferase. The finding supports the possibility of intracistron recombinations between genes performing a similar function (e.g. encoding glycosyltransferases). It remains to be determined whether mosaic gene sequences encode enzymes with altered substrate specificity and contribute to structural variation of the CPS. It is interesting to note that there is an extensive duplication of glycosyltransferase genes in the *cps* loci resulting in approximately double the number of transferases predicted by the structures (Table 1 and Fig. 2).

Strains NCTC11168, HS:23, HS:36, 81-176 and HS:41 were found to contain the following gene homologues: *gmhA2* (sedoheptulose-7-phosphate isomerase), *hddA* (D,D-heptose-7-phosphate kinase) and *hddC* (D,D-heptose-1-phosphate guanosyltransferase) which have enzymatically been shown to be involved in the alternate D-*glycero*-D-*manno*-heptose biosynthetic pathway in *Aneurinibacillus thermoaerophilus* proceeding through GDP- rather than ADP-linked intermediates (Kneidinger *et al.*, 2001). The presence of these genes correlates with the presence of heptose in the respective CPSs demonstrated by structural studies. In addition, the *C. jejuni gmhA2*, *hddA* and *hddC* genes are nearly identical in all loci, suggesting that recent exchange of these genes occurred between strains with different *cps* clusters. The *gmhA2*, *hddA* and *hddC* genes found in *C. jejuni* share similarity with and show colinearity to the respective genes involved in heptose biosynthesis in other bacteria (Fig. 4) (DeShazer *et al.*, 1998; Reckseidler *et al.*, 2001; Pacinelli *et al.*, 2002; Valvano *et al.*, 2002), thus supporting the possibility of being acquired via horizontal gene transfer.

In this study, we have demonstrated that the enzymes encoded by these homologues are also involved in heptose biosynthesis in *C. jejuni*. Furthermore, we have shown that *C. jejuni* has retained two heptose isomerase enzymes (GmhA and GmhA2; Fig. 6) that are both functional in *E. coli* and are used for the first step of heptose biosynthesis in both *C. jejuni* LOS and CPS. Interestingly, the genes required for *Campylobacter* LOS (*cj1148–cj1152*) and CPS (*cj1423–cj1425*) heptose biosynthesis are both clustered on the genome (Parkhill *et al.*, 2000). This is in contrast to most organisms that contain genes involved in ADP-activated heptose biosynthesis scattered throughout their genomes. Also, unlike bacteria containing the alternate GDP-activated pathway, all of the *Campylobacter* CPS clusters lack the phosphatase homologue necessary for the conversion of D,D-heptose-1,7-bisphos-

176.83 (HS:41 serostrain, enteritis isolate; Hanniffy *et al.*, 1999), NCTC12517 (HS:19 serostrain, enteritis isolate), G1 (HS:1, GBS isolate; Karlyshev *et al.*, 2000a), 81-176 (HS:23/36, enteritis isolate used in human challenge studies; Black *et al.*, 1988), CCUG 10954 (HS:23 serostrain, enteritis isolate) and ATCC 43456 (HS:36 serostrain, enteritis isolate) (Table 1). *C. jejuni* strains were grown in microaerophilic conditions at 37°C on 7% blood agar plates for 2 days. The *E. coli* XL2 Blue MRF′ strain (Stratagene), used in cloning experiments, was grown overnight at 37°C on LB agar plates supplemented with 100 μg ml^{-1} ampicillin when necessary. The *E. coli* χ_{705} parent and χ_{711} *gmhA* variant were previously described (Brooke and Valvano, 1996).

Construction and characterization of C. jejuni *NCTC11168 heptose mutants*

C. jejuni NCTC11168 insertional mutants were created and verified as previously described (St Michael *et al.*, 2002). For construction of the *cj1424c* (*gmhA2*) mutant, genes *cj1423c* to *cj1425c* were cloned using the primers: Cj1423cF651 and Cj1425cR41 (see *Supplementary material*; Table S8). The kanamycin resistance cassette (*kan*′) from pILL600 (Labigne-Roussel *et al.*, 1988) was inserted into the *Afl*II restriction site in a non-polar orientation generating pCSc24Km. The chloramphenicol resistance cassette (*cam*′) from pRY109 (Yao *et al.*, 1993) was inserted into the same site in a non-polar orientation to generate pCSc24Cm. For the *cj1425c* (*hddA*) mutant, genes *cj1423c* to *cj1425c* were cloned using the primers described above and *kan*′ was inserted into the *Bcl*I restriction site generating pCSc25. Although sequencing results demonstrated that *kan*′ insertion into *hddA* was in the polar orientation, the resulting phenotype did not result from polar effects on adjacent genes as mutagenesis of *gmhA2* (this study) or *cj1426c* (M. Szymanski, unpubl. obs.) does not alter the CPS heptose. For the *cj1152c* (*gmhB*) mutant, genes *cj1148* to *cj1152c* were cloned using waaFF9 and Cj1152c_R113 and *kan*′ was inserted into the *Sph*I restriction site in a non-polar orientation generating pCSI52. For the *cj1423c* (*hddC*) mutant, genes *cj1422c* to *cj1426c* were cloned using Cj1422cF1869 and Cj1426c_R30 and *kan*′ was inserted into the *Bgl*II restriction site in a non-polar orientation. The resultant construct, pCSc23, was truncated by ≈2 kb. Sequencing demonstrated that the 3′ end of pCSc23 lacked *cj1425c–cj1426c* and the first 84 base pairs of *cj1424c*. For the *cj1149c* (*gmhA*) mutant, genes *cj1148* to *cj1152c* were cloned using the primers described above and the mutant was generated by transposon mutagenesis to create pCSI49 using the EZ::TN™<KAN-2> insertion kit (Epicentre) according to the instructions of the manufacturer. Sequencing with the forward and reverse primers provided with the kit demonstrated that *kan*′ was inserted in a non-polar orientation after 420 bp in *gmhA*. The double *gmhA*::*kan*′/*gmhA2*::*cam*′ mutant was constructed by naturally transforming *gmhA2* with *gmhA* chromosomal DNA (Guerry *et al.*, 1994) and confirmed by PCR that both antibiotic cassettes were retained in the correct genes. The *cj1431* (*hddD*) mutant was made using plasmid cam109e8 constructed as part of the *C. jejuni* NCTC11168 genome sequencing project (Parkhill *et al.*, 2000). The plasmid contains a 1.7 kb fragment of NCTC11168 genomic DNA cloned into pUC18. The *kan*′ cassette from plasmid pJMK30 (van Vliet *et al.*, 1998) was inserted in a non-polar orientation into the *Swa*I site of *cj1431* contained within the fragment.

Complementation of E. coli χ_{711} *with* C. jejuni *isomerase homologues*

Campylobacter jejuni gmhA was amplified using primers constructed with forward *Nde*I and reverse *Sal*I restriction sites (FgmhA and BgmhA) while *gmhA2* was amplified using FgmhA2 and BgmhA2 and inserted in frame into the *Nde*I and *Sal*I sites of cloning vector pCW (Karwaski *et al.*, 2002) to generate CjgmhA and CjgmhA2. The plasmids were then inserted into electrocompetent *E. coli* χ_{711} lacking *gmhA* (Brooke and Valvano, 1996).

The strategy of amplification and sequencing of C. jejuni *CPS regions*

Short sequences of *kpsC* and *kpsF* genes from different strains were derived using the single-primer PCR procedure (Karlyshev *et al.*, 2000b). Primer pairs ak176/ak177 and ak173/ak174 were used for sequencing of the *kpsC* and *kpsF* genes respectively. Alignment of the derived sequences allowed the design of universal primers suitable for long-range PCR in various strains: ak186 for *kpsF* gene and ak188 for *kpsC* (see below). For serostrain HS:19 a more optimal *kpsC* primer ak187 was designed. Long-range PCR with *kpsF* and *kpsC* primers alone failed to produce any product with the reference strain NCTC11168. This could result from a relatively large size of the amplicon (over 36 kb). However, it was possible to amplify the entire *cps* region as two long-range PCR products when *kpsC* and *kpsF* primers were combined with primers derived from the internal biosynthetic genes.

The same strategy was used for amplification of *cps* regions from other strains. The sequencing of the *cps* regions of serostrains HS:23 and HS:36 was performed after the sequencing of the *cps* region of strain 81-176 was complete. The identical gene content and high sequence identity between these three strains allowed us to completely sequence the *cps* regions of serostrains HS:23 and HS:36 using custom-made oligonucleotides.

The biosynthetic *cps* region of strain NCTC11168 contains 28 genes (Karlyshev *et al.*, 2000a). The conserved genes present in the biosynthetic regions of other strains were identified using PCR amplification with primers derived from the sequence of NCTC11168 genome. Primer pairs specific to the biosynthetic *cps* genes of strain NCTC11168 are indicated in the *Supplementary material* (Table S9). The long-range PCR with the flanking *kpsC*- and *kpsF*-specific primers combined with the primers derived from conserved internal genes resulted in overlapping products suitable for generation of complete sequences of the internal biosynthetic regions. Long-range PCR was performed using the Expand 20 kb^{PLUS} PCR System (Roche) using conditions described by the manufacturer.

The long-range PCR products were treated with polynucleotide kinase, sonicated, and blunt-ended with T4 DNA polymerase. Then, 1–2 kb fragments were gel-extracted and

cloned into alkaline phosphatase-treated pUC18 (Promega) before sequencing. For closing gaps, primers corresponding to the ends of the contigs were designed and the regions were amplified and sequenced either directly or after cloning into pGEM-T-Easy vector (Promega) using the automatic sequencer. DNA sequencing was performed on ABI 377 or ABI 3700 automatic sequencers using an ABI PRISM BigDye Terminator Cycle Sequencing Kit (Perkin-Elmer). The sequences generated via shot-gun sequencing were assembled and edited using GAP4 (Bonfield *et al.*, 1995) or GeneTool software, and were deposited at the EMBL database with the accession numbers listed in Table 1.

Multiple sequence alignment was performed using the CLUSTALW program (http://www2.ebi.ac.uk/clustalw/). The *cps* sequences were analysed using Artemis software (Rutherford *et al.*, 2000) and the extracted amino acid sequences were analysed by similarity searches with the BLASTp program against NCTC11168 at http://www.sanger.ac.uk/Projects/C_jejuni/ and a non-redundant protein database at http://www.blast.genome.ad.jp/. The entire *cps* regions were compared with the *cps* region of NCTC11168 using BLASTn and tBLASTx programs (http://www.hgmp.mrc.ac.uk/) followed by the analysis using MSPCRUNCH (http://bioweb.pasteur.fr/seqanal/interfaces/mspcrunch.html) and ACT programs (http://www.sanger.ac.uk).

Isolation and purification of CPS

The CPS was isolated from dried cell mass (≈1 g) by the hot water/phenol method (Westphal and Jann, 1965). The aqueous phase was dialysed against water and lyophilized. The dried sample was then dissolved in water to a 1% solution (w/v) and subjected to ultracentrifugation to yield a gel-like pellet containing LOS and supernatant containing the CPS.

Analytical methods

Sugars were determined by examining their alditol acetate derivatives (Sawardeker *et al.*, 1965) by GLC-MS. Samples were hydrolysed for 4 h using 4 M trifluoroacetic acid at 100°C. The sample was reduced in $NaBD_4$ overnight in H_2O and acetylated with acetic anhydride at 100°C for 2 h using residual sodium acetate as the catalyst. The GLC-MS was equipped with a 30 M DB-17 capillary column (180°C to 260°C at 3.5°C per minute) and MS was performed in the electron impact mode on a Varian Saturn II mass spectrometer.

HR-MAS NMR allows the screening of small amounts of bacterial cells directly without having to purify surface carbohydrates (Szymanski *et al.*, 2003). HR-MAS experiments were performed on a Varian Inova 500 and 600 MHz spectrometer using a gradient 4 mm indirect detection HR-MAS nano-NMR probe (Varian) with a broadband decoupling coil as previously described (St Michael *et al.*, 2002; Young *et al.*, 2002). Proton spectra of cells were acquired with the Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence [90-(τ-180-τ)$_n$-acquisition] to remove broad lines arising from lipids and solid-like material. The total duration of the CPMG pulse ($n2\tau$) was 10 ms with τ set to (1/spin rate). ^{1}H chemical shifts were referenced relative to that of sodium 3-(trimethylsilyl)propri-onate-2,2,3,3-d_4 at 0.0 p.p.m.

High-resolution NMR experiments on the partially purified CPS were acquired using a Varian Inova 500 MHz spectrometer equipped with a Z-gradient 3 mm triple resonance (^{1}H, ^{13}C, ^{31}P) probe. The experiments were performed at 40°C with suppression of the water resonance. The methyl resonance of acetone was used as an internal reference at δ_H 2.225 p.p.m. and δ_C 31.07 p.p.m. Standard pulse sequences from Varian, COSY, TOCSY, NOESY, HMQC and ^{31}P HMQC were used.

Preparation of O-deacylated LOS and CE-MS analysis

LOS preparations and CE-MS analyses were performed as previously described (Szymanski *et al.*, 2003). Briefly, a Prince CE system (Prince Technologies) was coupled to a Q-Star quadrupole/time-of-flight mass spectrometer or an API 3000 mass spectrometer (Applied Biosystems/MDS Sciex) via a microIonspray interface.

Deoxycholate-PAGE of polysaccharides

Proteinase K-treated whole-cell digests of *E. coli* χ_{705}, χ_{711} and χ_{711} complements were prepared and analysed on 16.5% deoxycholate-PAGE as previously described (St Michael *et al.*, 2002).

Acknowledgements

We thank Dr P. Guerry for the kind gift of strain 81-176 and Dr M. Valvano for *E. coli* χ_{705} and χ_{711}. We also thank Dr N.M. Young for support, Dr B. Kneidinger for helpful discussions, Dr J. Nash for mutant primer design and Anna Cunningham and Sonia Leclerc for DNA sequencing at the NRC. We especially thank Laura Fiori and Marc Lamoureux for excellent technical assistance. The work was supported by the Leverhulme Trust, the BBSRC and an MRC Studentship to O.L.C. and the NRC Genomics and Health Initiative.

Supplementary material

The following material is available from
http://www.blackwellpublishing.com/products/journals/suppmat/mmi/mmi4374/mmi4374sm.htm

Fig. S1. Strategy for amplification of the *cps* regions.
Fig. S2. Proton NMR spectra of NCTC12500 (HS:1 serostrain) and G1 (HS:1).
Fig. S3. CE-MS analysis of *E. coli* K12 O-deacylated cells.
Table S1. Strain NCTC11168 (HS:2).
Table S2. Strain NCTC12517 (HS:19).
Table S3. Strain G1 (HS:1).
Table S4. Strains 81-176 (HS:23/HS:36), CCUG 10954 (HS:23) and ATCC 43456 (HS:36).
Table S5. Strain 176.83 (HS:41).
Table S6. Mass assignment of the deacylated LOS from *E. coli* wild type (χ_{705}), *E. coli* isomerase mutant (χ_{711}) and *E. coli* χ_{711} complements from Fig. S3.

Table S7. Mass assignment of the deacylated LOS from *C. jejuni* wild type and mutants from Fig. 5.
Table S8. Primers mentioned in *Experimental procedures*
Table S9. Gene specific PCR primers derived from *C. jejuni* NCTC11168 genome sequence.

References

Aspinall, G.O., McDonald, A.G., and Pang, H. (1992) Structures of the O chains from lipopolysaccharides of *Campylobacter jejuni* serotypes O:23 and O:36. *Carbohydr Res* **231:** 13–30.

Aspinall, G.O., Fujimoto, S., McDonald, A.G., Pang, H., Kurjanczyk, L.A., and Penner, J.L. (1994a) Lipopolysaccharides from *Campylobacter jejuni* associated with Guillain–Barré syndrome patients mimic human gangliosides in structure. *Infect Immun* **62:** 2122–2125.

Aspinall, G.O., McDonald, A.G., and Pang, H. (1994b) Lipopolysaccharides of *Campylobacter jejuni* serotype O:19: structures of O antigen chains from the serostrain and two bacterial isolates from patients with the Guillain–Barré syndrome. *Biochemistry* **33:** 250–255.

Bacon, D.J., Szymanski, C.M., Burr, D.H., Silver, R.P., Alm, R.A., and Guerry, P. (2001) A phase-variable capsule is involved in virulence of *Campylobacter jejuni* 81-176. *Mol Microbiol* **40:** 769–777.

Benz, I., and Schmidt, M.A. (2001) Glycosylation with heptose residues mediated by the *aah* gene product is essential for adherence of the AIDA-I adhesin. *Mol Microbiol* **40:** 1403–1413.

Black, R.E., Levine, M.M., Clements, M.L., Hughes, T.P., and Blaser, M.J. (1988) Experimental *Campylobacter jejuni* infection in humans. *J Infect Dis* **157:** 472–479.

Bonfield, J.K., Smith, K., and Staden, R. (1995) A new DNA sequence assembly program. *Nucleic Acids Res* **23:** 4992–4999.

Brooke, J.S., and Valvano, M.A. (1996) Biosynthesis of inner core lipopolysaccharide in enteric bacteria identification and characterization of a conserved phosphoheptose isomerase. *J Biol Chem* **271:** 3608–3614.

DeShazer, D., Brett, P.J., and Woods, D.E. (1998) The type II O-antigenic polysaccharide moiety of *Burkholderia pseudomallei* lipopolysaccharide is required for serum resistance and virulence. *Mol Microbiol* **30:** 1081–1100.

Dillard, J.P., and Yother, J. (1994) Genetic and molecular characterization of capsular polysaccharide biosynthesis in *Streptococcus pneumoniae* type 3. *Mol Microbiol* **12:** 959–972.

Dillard, J.P., Vandersea, M.W., and Yother, J. (1995) Characterization of the cassette containing genes for type 3 capsular polysaccharide biosynthesis in *Streptococcus pneumoniae*. *J Exp Med* **181:** 973–983.

Dorrell, N., Mangan, J.A., Laing, K.G., Hinds, J., Linton, D., Al-Ghusein, H., *et al.* (2001) Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity. *Genome Res* **11:** 1706–1715.

Gibson, B.W., Campagnari, A.A., Melaugh, W., Phillips, N.J., Apicella, M.A., Grass, S., *et al.* (1997) Characterization of a transposon Tn*916*-generated mutant of *Haemophilus ducreyi* 35000 defective in lipooligosaccharide biosynthesis. *J Bacteriol* **179:** 5062–5071.

Gilbert, M., Godschalk, P.C., Karwaski, M.F., Ang, C.W., van Belkum, A., Li, J., *et al.* (2004) Evidence for acquisition of the lipooligosaccharide biosynthesis locus in *Campylobacter jejuni* GB11, a strain isolated from a patient with Guillain–Barre syndrome, by horizontal exchange. *Infect Immun* **72:** 1162–1165.

Graninger, M., Kneidinger, B., Bruno, K., Scheberl, A., and Messner, P. (2002) Homologs of the Rml enzymes from *Salmonella enterica* are responsible for dTDP-beta-L-rhamnose biosynthesis in the gram-positive thermophile *Aneurinibacillus thermoaerophilus* DSM 10155. *Appl Environ Microbiol* **68:** 3708–3715.

Gronow, S., Oertelt, C., Ervela, E., Zamyatina, A., Kosma, P., Skurnik, M., and Holst, O. (2001) Characterization of the physiological substrate for lipopolysaccharide heptosyltransferases I and II. *J Endotoxin Res* **7:** 263–270.

Guerry, P., Yao, R., Alm, R.A., Burr, D.H., and Trust, T.J. (1994) Systems of experimental genetics for *Campylobacter* species. *Method Enzymol* **235:** 474–481.

Hanniffy, O.M., Shashkov, A.S., Moran, A.P., Prendergast, M.M., Senchenkova, S.N., Knirel, Y.A., and Savage, A.V. (1999) Chemical structure of a polysaccharide from *Campylobacter jejuni* 176.83 (serotype O:41) containing only furanose sugars. *Carbohydr Res* **319:** 124–132.

Jiang, S.M., Wang, L., and Reeves, P.R. (2001) Molecular characterization of *Streptococcus pneumoniae* type 4, 6B, 8, and 18C capsular polysaccharide gene clusters. *Infect Immun* **69:** 1244–1255.

Kamerling, J.P. (1999) Pneumococcal polysaccharides: a chemical view. In *Streptococcus pneumoniae: Molecular Biology and Mechanisms of Disease*. Tomasz, A. (ed.). New York: Mary Ann Liebert, pp. 81–114.

Karlyshev, A.V., and Wren, B.W. (2001) Detection and initial characterization of novel capsular polysaccharide among diverse *Campylobacter jejuni* strains using alcian blue dye. *J Clin Microbiol* **39:** 279–284.

Karlyshev, A.V., Henderson, J., Ketley, J.M., and Wren, B.W. (1999) Procedure for the investigation of bacterial genomes: random shot-gun cloning, sample sequencing and mutagenesis of *Campylobacter jejuni*. *Biotechniques* **26:** 50–52,54,56.

Karlyshev, A.V., Linton, D., Gregson, N.A., Lastovica, A.J., and Wren, B.W. (2000a) Genetic and biochemical evidence of a *Campylobacter jejuni* capsular polysaccharide that accounts for Penner serotype specificity. *Mol Microbiol* **35:** 529–541.

Karlyshev, A.V., Pallen, M.J., and Wren, B.W. (2000b) Single-primer PCR procedure for rapid identification of transposon insertion sites. *Biotechniques* **28:** 1078–1080, 1082.

Karlyshev, A.V., McCrossan, M.V., and Wren, B.W. (2001) Demonstration of polysaccharide capsule in *Campylobacter jejuni* using electron microscopy. *Infect Immun* **69:** 5921–5924.

Karwaski, M.F., Wakarchuk, W.W., and Gilbert, M. (2002) High-level expression of recombinant Neisseria CMP-sialic acid synthetase in *Escherichia coli*. *Protein Expr Purif* **25:** 237–240.

Kneidinger, B., Graninger, M., Puchberger, M., Kosma, P.,

and Messner, P. (2001) Biosynthesis of nucleotide-activated D-glycero-D-manno-heptose. *J Biol Chem* **276:** 20935–20944.

Kneidinger, B., O'Riordan, K., Li, J., Brisson, J.R., Lee, J.C., and Lam, J.S. (2003) Three highly conserved proteins catalyze the conversion of UDP-*N*-acetyl-D-glucosamine to precursors for the biosynthesis of O antigen in *Pseudomonas aeruginosa* O11 and capsule in *Staphylococcus aureus* type 5. Implications for the UDP-*N*-acetyl-L-fucosamine biosynthetic pathway. *J Biol Chem* **278:** 3615–3627.

Labigne-Roussel, A., Courcoux, P., and Tompkins, L. (1988) Gene disruption and replacement as a feasible approach for mutagenesis of *Campylobacter jejuni*. *J Bacteriol* **170:** 1704–1708.

Lastovica, A.J., Goddard, E.A., and Argent, A.C. (1997) Guillain–Barre syndrome in South Africa associated with *Campylobacter jejuni* O:41 strains. *J Infect Dis* **176** (Suppl. 2): S139–S143.

MacDonald, A.G. (1993) *Lipopolysaccharides from* Campylobacter. PhD Thesis. Toronto, Canada: York University.

Moormann, C., Benz, I., and Schmidt, M.A. (2002) Functional substitution of the TibC protein of enterotoxigenic *Escherichia coli* strains for the autotransporter adhesin heptosyltransferase of the AIDA system. *Infect Immun* **70:** 2264–2270.

Moran, A.P., Penner, J.L., and Aspinall, G.O. (2000) *Campylobacter* polysaccharides. In *Campylobacter*. Nachamkin, I., and Blaser, M.J. (eds). Washington, DC: American Society for Microbiology Press, pp. 241–257.

Nassau, P.M., Martin, S.L., Brown, R.E., Weston, A., Monsey, D., McNeil, M.R., and Duncan, K. (1996) Galactofuranose biosynthesis in *Escherichia coli* K-12: identification and cloning of UDP-galactopyranose mutase. *J Bacteriol* **178:** 1047–1052.

Pacinelli, E., Wang, L., and Reeves, P.R. (2002) Relationship of *Yersinia pseudotuberculosis* O antigens IA, IIA, and IVB: the IIA gene cluster was derived from that of IVB. *Infect Immun* **70:** 3271–3276.

Parkhill, J., Wren, B.W., Mungall, K., Ketley, J.M., Churcher, C., Basham, D., *et al.* (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403:** 665–668.

Pooley, H.M., Abellan, F.X., and Karamata, D. (1991) A conditional-lethal mutant of *Bacillus subtilis* 168 with a thermosensitive glycerol-3-phosphate cytidylyltransferase, an enzyme specific for the synthesis of the major wall teichoic acid. *J Gen Microbiol* **137:** 921–928.

Reckseidler, S.L., DeShazer, D., Sokol, P.A., and Woods, D.E. (2001) Detection of bacterial virulence genes by subtractive hybridization: identification of capsular polysaccharide of *Burkholderia pseudomallei* as a major virulence determinant. *Infect Immun* **69:** 34–44.

Roberts, I.S. (1996) The biochemistry and genetics of capsular polysaccharide production in bacteria. *Annu Rev Microbiol* **50:** 285–315.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* **16:** 944–945.

Sawardeker, D.G., Sloneker, J.H., and Jeanes, A. (1965) Quantitative determination of monosaccharides as their alditol acetates by gas liquid chromatography. *Anal Chem* **37:** 1602–1604.

Schertzer, J.W., and Brown, E.D. (2003) Purified, recombinant TagF protein from *Bacillus subtilis* 168 catalyzes the polymerization of glycerol phosphate onto a membrane acceptor *in vitro*. *J Biol Chem* **278:** 18002–18007.

Sieberth, V., Rigg, G.P., Roberts, I.S., and Jann, K. (1995) Expression and characterization of UDPGlc dehydrogenase (KfiD), which is encoded in the type-specific region 2 of the *Escherichia coli* K5 capsule genes. *J Bacteriol* **177:** 4562–4565.

St Michael, F., Szymanski, C.M., Li, J., Chan, K.H., Khieu, N.H., Larocque, S., *et al.* (2002) The structures of the lipooligosaccharide and capsule polysaccharide of *Campylobacter jejuni* genome sequenced strain NCTC 11168. *Eur J Biochem* **269:** 5119–5136.

Swartley, J.S., Marfin, A.A., Edupuganti, S., Liu, L.J., Cieslak, P., Perkins, B., *et al.* (1997) Capsule switching of *Neisseria meningitidis*. *Proc Natl Acad Sci USA* **94:** 271–276.

Szymanski, C.M., St Michael, F., Jarrell, H.C., Li, J., Gilbert, M., Larocque, S., *et al.* (2003) Detection of conserved N-linked glycans and phase variable lipo-oligosaccharides and capsules from *Campylobacter* cells by mass spectrometry and high resolution magic angle spinning NMR spectroscopy. *J Biol Chem* **278:** 24509–24520.

Tullius, M.V., Phillips, N.J., Scheffler, N.K., Samuels, N.M., Munson, R.S., Jr, Hansen, E.J., *et al.* (2002) The *lbgAB* gene cluster of *Haemophilus ducreyi* encodes a beta-1,4-galactosyltransferase and an alpha-1,6-DD-heptosyltransferase involved in lipooligosaccharide biosynthesis. *Infect Immun* **70:** 2853–2861.

Valvano, M.A., Messner, P., and Kosma, P. (2002) Novel pathways for biosynthesis of nucleotide-activated glycero-manno-heptose precursors of bacterial glycoproteins and cell surface polysaccharides. *Microbiology* **148:** 1979–1989.

van Vliet, A.H., Wooldridge, K.G., and Ketley, J.M. (1998) Iron-responsive gene regulation in a *Campylobacter jejuni* fur mutant. *J Bacteriol* **180:** 5291–5298.

Westphal, O., and Jann, K. (1965) Bacterial lipopolysaccharide. Extraction with phenol-water and further applications of the procedure. *Methods Carbohydr Chem* **5:** 88–91.

Whitfield, C., and Roberts, I.S. (1999) Structure, assembly and regulation of expression of capsules in *Escherichia coli*. *Mol Microbiol* **31:** 1307–1319.

Yao, R., Alm, R.A., Trust, T.J., and Guerry, P. (1993) Construction of new *Campylobacter* cloning vectors and a new mutational cat cassette. *Gene* **130:** 127–130.

Young, N.M., Brisson, J.R., Kelly, J., Watson, D.C., Tessier, L., Lanthier, P.H., *et al.* (2002) Structure of the N-linked glycan present on multiple glycoproteins in the Gram-negative bacterium, *Campylobacter jejuni*. *J Biol Chem* **277:** 42530–42539.

9.0 References

Aarestrup, F.M., Nielsen, E.M., Madsen, M., and Engberg, J. (1997) Antimicrobial susceptibility patterns of thermophilic *Campylobacter* spp. from humans, pigs, cattle, and broilers in Denmark. *Antimicrob Agents Chemother* **41**: 2244-2250.

Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A., and Carniel, E. (1999) *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci U S A* **96**: 14043-14048.

Agency, F.S. (2000) A Report of the Study of Infectious Intestinal Disease in England. London: Food Standards Agency.

Ahmed, I.H., Manning, G., Wassenaar, T.M., Cawthraw, S., and Newell, D.G. (2002) Identification of genetic differences between two *Campylobacter jejuni* strains with different colonization potentials. *Microbiology* **148**: 1203-1212.

Aizawa, S.I. (1996) Flagellar assembly in *Salmonella typhimurium*. *Mol Microbiol* **19**: 1-5.

Allos, B.M. (1998) *Campylobacter jejuni* infection as a cause of the Guillain-Barre syndrome. *Infect Dis Clin North Am* **12**: 173-184.

Ang, C.W., De Klerk, M.A., Endtz, H.P., Jacobs, B.C., Laman, J.D., Van Der Meche, F.G.A., and van Doorn, P.A. (2001) Guillain-Barre syndrome- and Miller Fisher Syndrome- Associated *Campylobacter jejuni* Lipopolysaccharides Induce Anti-GM1 annd Anti-GQ1b Antibodies in Rabbits. *Infection and Immunity* **69**: 2462-2472.

Arora, S.K., Wolfgang, M.C., Lory, S., and Ramphal, R. (2004) Sequence polymorphism in the glycosylation island and flagellins of *Pseudomonas aeruginosa*. *J Bacteriol* **186**: 2115-2122.

Babakhani, F.K., Bradley, G.A., and Joens, L.A. (1993) Newborn piglet model for campylobacteriosis. *Infect Immun* **61**: 3466-3475.

Bacon, D.J., Szymanski, C.M., Burr, D.H., Silver, R.P., Alm, R.A., and Guerry, p. (2001) A phase-variable capsule is involved in virulence of *Campylobacter jejuni* 81-176. *Molecular Microbiology* **40**: 769-777.

Behr, M.A., Wilson, M.A., Gill, W.P., Salamon, H., Schoolnik, G.K., Rane, S., and Small, P.M. (1999) Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* **284**: 1520-1523.

Benjelloun-Touimi, Z., Sansonetti, P.J., and Parsot, C. (1995) SepA, the major extracellular protein of *Shigella flexneri*: autonomous secretion and involvement in tissue invasion. *Mol Microbiol* **17**: 123-135.

Berin, M.C., Darfeuille-Michaud, A., Egan, L.J., Miyamoto, Y., and Kagnoff, M.F. (2002) Role of EHEC O157:H7 virulence factors in the activation of intestinal epithelial cell NF-kappaB and MAP kinase pathways and the upregulated expression of interleukin 8. *Cell Microbiol* **4**: 635-648.

Bernatchez, S., Szymanski, C.M., Ishiyama, N., Li, J., Jarrell, H.C., Lau, P.C., Berghuis, A.M., Young, N.M., and Wakarchuk, W.W. (2005) A single bifunctional UDP-GlcNAc/Glc 4-epimerase supports the synthesis of three cell surface glycoconjugates in *Campylobacter jejuni*. *J Biol Chem* **280**: 4792-4802.

Black, R.E., Levine, M.M., Clements, M.L., Hughes, T.P., and Blaser, M.J. (1988) Experimental *Campylobacter jejuni* infection in humans. *J Infect Dis* **157**: 472-479.

Blaser, M.J., Hardesty, H.L., Powers, B., and Wang, W.L. (1980) Survival of *Campylobacter fetus* subsp. *jejuni* in biological milieus. *J Clin Microbiol* **11**: 309-313.

Blaser, M.J. (1995) Campylobacter and Related species. In *Principles and Practice of Infectious Diseases*. Mandell, G.L., Bennett, J.E., Dolin, R. (ed). New York: Churchill Livingstone Inc., pp. 1948-1956.

Blaser, M.J. (1997) Epidemiologic and clinical features of *Campylobacter jejuni* infections. *J Infect Dis* **176 Suppl 2**: S103-105.

Bourke, B., Chan, V.L., and Sherman, P. (1998) *Campylobacter upsaliensis*: waiting in the wings. *Clin Microbiol Rev* **11**: 440-449.

Broman, T., Palmgren, H., Bergstrom, S., Sellin, M., Waldenstrom, J., Danielsson-Tham, M.L., and Olsen, B. (2002) *Campylobacter jejuni* in black-headed gulls (Larus ridibundus): prevalence, genotypes, and influence on *C. jejuni* epidemiology. *J Clin Microbiol* **40**: 4594-4602.

Brown, P.E., Christensen, O.F., Clough, H.E., Diggle, P.J., Hart, C.A., Hazel, S., Kemp, R., Leatherbarrow, A.J., Moore, A., Sutherst, J., Turner, J., Williams, N.J., Wright, E.J., and French, N.P. (2004) Frequency and spatial distribution of environmental *Campylobacter spp. Appl Environ Microbiol* **70**: 6501-6511.

Brunder, W., Schmidt, H., and Karch, H. (1997) EspP, a novel extracellular serine protease of enterohaemorrhagic *Escherichia coli* O157:H7 cleaves human coagulation factor V. *Mol Microbiol* **24**: 767-778.

Butzler, J.P., Dekeyser, P., Detrain, M., and Dehaen, F. (1973) Related vibrio in stools. *J Pediatr* **82**: 493-495.

Carrillo, C.D., Taboada, E., Nash, J.H., Lanthier, P., Kelly, J., Lau, P.C., Verhulp, R., Mykytczuk, O., Sy, J., Findlay, W.A., Amoako, K., Gomis, S., Willson, P., Austin, J.W., Potter, A., Babiuk, L., Allan, B., and Szymanski, C.M. (2004) Genome-wide expression analyses of *Campylobacter jejuni* NCTC11168 reveals coordinate regulation of motility and virulence by *flhA*. *J Biol Chem* **279**: 20327-20338.

Castric, P., Cassels, F.J., and Carlson, R.W. (2001) Structural characterization of the *Pseudomonas aeruginosa* 1244 pilin glycan. *J Biol Chem* **276**: 26479-26485.

Charlebois, R.L., Beiko, R.G., and Ragan, M.A. (2003) Microbial phylogenomics: Branching out. *Nature* **421**: 217.

Chilcott, G.S., Hughes, K.T. (2000). Coupling of flagellar gene experssion to flagellar gene assembly in *Salmonella enterica* serovar *typhimurium* and *Escherichia coli*. *Microbiol Mol Biol Rev* **64**: 694-708

Christie, P.J., and Vogel, J.P. (2000) Bacterial type IV secretion: conjugation systems adapted to deliver effector molecules to host cells. *Trends Microbiol* **8**: 354-360.

Colland, F., Rain, J-C., Gounon, P., Labigne, A., Legrain, P., DeReuse, H. (2001) Identification of the *Helicobacter pylori* anti-sigma28 factor. *Mol Microbiol* **41**:477

de Melo, M.A., and Pechere, J.C. (1988) Effect of mucin on *Campylobacter jejuni* association and invasion on HEp-2 cells. *Microb Pathog* **5**: 71-76.

Dekeyser, P., Gossuin-Detrain, M., Butzler, J.P., and Sternon, J. (1972) Acute enteritis due to related vibrio: first positive stool cultures. *J Infect Dis* **125**: 390-392.

Dingle, K.E., Colles, F.M., Wareing, D.R.A., Ure, R., Fox, A.J., Bolton, F.E., Bootsma, H.J., Willems, R.J.L., Urwin, R., and Maiden, M.C.J. (2001a) Multilocus Sequence Typing System for *Campylobacter jejuni*. *Journal of Clinical Microbiology* **39**: 14-23.

Dingle, K.E., Van Den Braak, N., Colles, F.M., Price, L.J., Woodward, D.L., Rodgers, F.G., Endtz, H.P., Van Belkum, A., and Maiden, M.C. (2001b) Sequence Typing Confirms that *Campylobacter jejuni* Strains Associated with Guillain-Barre and Miller-Fisher Syndromes Are of Diverse Genetic Lineage, Serotype, and Flagella Type. *J Clin Microbiol* **39**: 3346-3349.

Dingle, K.E., Colles, F.M., Ure, R., Wagenaar, J.A., Duim, B., Bolton, F.J., Fox, A.J., Wareing, D.R., and Maiden, M.C. (2002) Molecular characterization of *Campylobacter jejuni* clones: a basis for epidemiologic investigation. *Emerg Infect Dis* **8**: 949-955.

Doig, P., Kinsella, N., Guerry, P., and Trust, T.J. (1996a) Characterization of a post-translational modificationn of *Campylobacter* flagellin: identification of a sero-specific glycosyl moiety. *Molecular Microbiology* **19**: 379-387.

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., and Wren, B.W. (2001) Whole Genome Comparison of *Campylobacter jejuni* Human Isolates Using a Low-Cost Microarray Reveals Extensive Genetic Diversity. *Genome Res* **11**: 1706-1715.

Dutta, P.R., Cappello, R., Navarro-Garcia, F., and Nataro, J.P. (2002) Functional comparison of serine protease autotransporters of enterobacteriaceae. *Infect Immun* **70**: 7105-7113.

Dziejman, M., Balon, E., Boyd, D., Fraser, C.M., Heidelberg, J.F., and Mekalanos, J.J. (2002) Comparative genomic analysis of *Vibrio cholerae*: Genes that correlate with cholera endemic and pandemic disease. *Proc Natl Acad Sci U S A* **99**: 1556-1561.

Eaves-Pyles, T., Murthy, K., Liaudet, L., Virag, L., Ross, G., Soriano, F.G., Szabo, C., and Salzman, A.L. (2001) Flagellin, a novel mediator of Salmonella-induced epithelial activation and systemic inflammation: I kappa B alpha degradation,

induction of nitric oxide synthase, induction of proinflammatory mediators, and cardiovascular dysfunction. *J Immunol* **166**: 1248-1260.

Eisen, J.A., and Fraser, C.M. (2003) Phylogenomics: intersection of evolution and genomics. *Science* **300**: 1706-1707.

Endtz, H.P., Vliegenthart, J.S., Vandamme, P., Weverink, H.W., van den Braak, N.P., Verbrugh, H.A., and van Belkum, A. (1997) Genotypic diversity of *Campylobacter lari* isolated from mussels and oysters in The Netherlands. *Int J Food Microbiol* **34**: 79-88.

Engberg, J., Gerner-Smidt, P., Scheutz, F., Moller Nielsen, E., On, S.L., and Molbak, K. (1998) Water-borne *Campylobacter jejuni* infection in a Danish town---a 6-week continuous source outbreak. *Clin Microbiol Infect* **4**: 648-656.

Engberg, J., Nachamkin, I., Fussing, V., McKhann, G.M., Griffin, J.W., Piffaretti, J.C., Nielsen, E.M., and Gerner-Smidt, P. (2001) Absence of clonality of *Campylobacter jejuni* in serotypes other than HS:19 associated with Guillain-Barre syndrome and gastroenteritis. *J Infect Dis* **184**: 215-220.

Everest, P.H., Goossens, H., Sibbons, P., Lloyd, D.R., Knutton, S., Leece, R., Ketley, J.M., and Williams, P.H. (1993) Pathological changes in the rabbit ileal loop model caused by *Campylobacter jejuni* from human colitis. *J Med Microbiol* **38**: 316-321.

Eyigor, A., Dawson, K.A., Langlois, B.E., and Pickett, C.L. (1999) Cytolethal Distending Toxin genes in *Campylobacter jejuni* and *Campylobacter coli* Isolates: Detection and Analysis by PCR. *Journal of Clinical Microbiology* **37**: 1646-1650.

Fitzgerald, J.R., Sturdevant, D.E., Mackie, S.M., Gill, S.R., and Musser, J.M. (2001) Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. *Proc Natl Acad Sci U S A* **98**: 8821-8826.

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. (2005) Major Structural Differences

and Novel Potential Virulence mechanisms from the Genomes of Multiple *Campylobacter* Species. *PLoS Biol* **3**: 1-14.

Friedman, C.R., Hoekstra, R.M., Samuel, M., Marcus, R., Bender, J., Shiferaw, B., Reddy, S., Ahuja, S.D., Helfrick, D.L., Hardnett, F., Carter, M., Anderson, B., and Tauxe, R.V. (2004) Risk factors for sporadic *Campylobacter* infection in the United States: A case-control study in FoodNet sites. *Clin Infect Dis* **38 Suppl 3**: S285-296.

Fry, B.N., Feng, S., Chen, Y.Y., Newell, D.G., Coloe, P.J., and Korolik, V. (2000a) The galE gene of *Campylobacter jejuni* is involved in lipopolysaccharide synthesis and virulence. *Infect Immun* **68**: 2594-2601.

Fujimoto, S., Allos, B.M., Misawa, N., Patton, C.M., and Blaser, M.J. (1997) Restriction fragment length polymorphism analysis and random amplified polymorphic DNA analysis of *Campylobacter jejuni* strains isolated from patients with Guillain-Barre syndrome. *J Infect Dis* **176**: 1105-1108.

Gavin, R., Rabaan, A.A., Merino, S., Tomas, J.M., Gryllos, I., and Shaw, J.G. (2002) Lateral flagella of *Aeromonas* species are essential for epithelial cell adherence and biofilm formation. *Mol Microbiol* **43**: 383-397.

Gilbert, M., Karwaski, M.F., Bernatchez, S., Young, N.M., Taboada, E., Michniewicz, J., Cunningham, A.M., and Wakarchuk, W.W. (2002) The genetic bases for the variation in the lipo-oligosaccharide of the mucosal pathogen, *Campylobacter jejuni*. Biosynthesis of sialylated ganglioside mimics in the core oligosaccharide. *J Biol Chem* **277**: 327-337.

Goodyear, C.S., O'Hanlon, G.M., Plomp, J.J., Wagner, E.R., Morrison, I., Veitch, J., Cochrane, L., Bullens, R.W., Molenaar, P.C., Conner, J., and Willison, H.J. (1999) Monoclonal antibodies raised against Guillain-Barre syndrome-associated *Campylobacter jejuni* lipopolysaccharides react with neuronal gangliosides and paralyze muscle-nerve preparations. *J Clin Invest* **104**: 697-708.

Grant, C.C., Konkel, M.E., Cieplak, W., Jr., and Tompkins, L.S. (1993) Role of flagella in adherence, internalization, and translocation of *Campylobacter jejuni* in nonpolarized and polarized epithelial cell cultures. *Infect Immun* **61**: 1764-1771.

Guerry, P., Alm, R.A., Power, M.E., Logan, S.M., and Trust, T.J. (1991) Role of two flagellin genes in *Campylobacter* motility. *J Bacteriol* **173**: 4757-4764.

Guerry, P., Szymanski, C.M., Prendergast, M.M., Hickey, T.E., Ewing, C.P., Pattarini, D.L., and Moran, A.P. (2002) Phase variation of *Campylobacter jejuni* 81-176 lipooligosaccharide affects ganglioside mimicry and invasiveness in vitro. *Infect Immun* **70**: 787-793.

Guerry, P.A., RA; Power, ME; Trust, TJ (1992) Molecular and Structural Analysis of *Campylobacter* flagellin. In *Campylobacter jejuni Current Status and Future Trends*. Nachamkin, I.B., M and Tompkins, LS (ed): ASM, pp. 267-281.

Harshey, R.M., and Toguchi, A. (1996) Spinning tails: homologies among bacterial flagellar systems. *Trends Microbiol* **4**: 226-231.

Hendrixson, D.R., DiRita, V.J. (2003) Transcription of sigma54-dependent but not sigma28-dependent flagellar genes in *Campylobacter jejuni* is associated with formation of the flagellar secretory apparatus. *Mol Microbiol* **50**:687-702.

Hinchliffe, S.J., Isherwood, K.E., Stabler, R.A., Prentice, M.B., Rakin, A., Nichols, R.A., Oyston, P.C., Hinds, J., Titball, R.W., and Wren, B.W. (2003) Application of DNA microarrays to study the evolutionary genomics of Yersinia pestis and *Yersinia pseudotuberculosis*. *Genome Res* **13**: 2018-2029.

Hinds, J., K.G. Laing, J.A. Mangan, and P.D. Butcher (2002) Glass slide microarrays for bacterial genomes, in Functional Microbial Genomics. *in Functional Microbial Genomics, B.W. Wren and N. Dorrell, Editors. Elsevier Science: London, UK. 83-99.*

Hu, L., and Kopecko, D.J. (1999) *Campylobacter jejuni* 81-176 associates with microtubules and dynein during invasion of human intestinal cells. *Infect Immun* **67**: 4171-4182.

Inglis, G.D., Kalischuk, L.D., and Busz, H.W. (2004) Chronic shedding of *Campylobacter* species in beef cattle. *J Appl Microbiol* **97**: 410-420.

Israel, D.A., Salama, N., Krishna, U., Rieger, U.M., Atherton, J.C., Falkow, S., and Peek, R.M., Jr. (2001) *Helicobacter pylori* genetic diversity within the gastric niche of a single human host. *Proc Natl Acad Sci U S A* **98**: 14625-14630.

Jagannathan, A., Constantinidou, C., and Penn, C.W. (2001) Roles of *rpoN*, *fliA* and *flgR* in Expression of Flagella in *Campylobacter jejuni*. *Journnal of Bacteriology* **183**: 2937-2942.

Jin, S., Joe, A., Lynett, J., Hani, E.K., Sherman, P., and Chan, V.L. (2001) JlpA, a novel surface-exposed lipoprotein specific to *Campylobacter jejuni*, mediates adherence to host epithelial cells. *Mol Microbiol* **39**: 1225-1236.

Jolley, K.A., Feil, E.J., Chan, M.S., and Maiden, M.C. (2001) Sequence type analysis and recombinational tests (START). *Bioinformatics* **17**: 1230-1231.

Kaijser, B.a.M., F. (1992) Diagnosis of *Campylobacter* Infections. In *Campylobacter jejuni Current Status and Future Trends*. Nachamkin, I.B., M and Tompkins, LS (ed): ASM, pp. 89-92.

Karlyshev, A.V., Linton, D., Gregson, N.A., Lastovica, A.J., and Wren, B.W. (2000) Genetic and biochemical evidence of a *Campylobacter jejuni* capsular polysaccharide that accounts for Penner serotype specificity. *Mol Microbiol* **35**: 529-541.

Karlyshev, A.V., McCrossan, M.V., and Wren, B.W. (2001) Demonstration of polysaccharide capsule in *Campylobacter jejuni* using electron microscopy. *Infect Immun* **69**: 5921-5924.

Karlyshev, A.V., and Wren, B.W. (2001) Detection and initial characterization of novel capsular polysaccharide among diverse *Campylobacter jejuni* strains using alcian blue dye. *J Clin Microbiol* **39**: 279-284.

Karlyshev, A.V., Linton, D., Gregson, N.A., and Wren, B.W. (2002) A novel paralogous gene family involved in phase-variable flagella- mediated motility in *Campylobacter jejuni*. *Microbiology* **148**: 473-480.

Karlyshev, A.V., Champion, O.L., Churcher, C., Brisson, J.R., Jarrell, H.C., Gilbert, M., Brochu, D., St Michael, F., Li, J., Wakarchuk, W.W., Goodhead, I., Sanders, M., Stevens, K., White, B., Parkhill, J., Wren, B.W., and Szymanski, C.M. (2005) Analysis of *Campylobacter jejuni* capsular loci reveals multiple mechanisms for the generation of structural diversity and the ability to form complex heptoses. *Mol Microbiol* **55**: 90-103.

Kato-Maeda, M., Rhee, J.T., Gingeras, T.R., Salamon, H., Drenkow, J., Smittipat, N., and Small, P.M. (2001) Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Res* **11**: 547-554.

Keramas, G., Bang, D.D., Lund, M., Madsen, M., Rasmussen, S.E., Bunkenborg, H., Telleman, P., and Christensen, C.B. (2003) Development of a sensitive DNA microarray suitable for rapid detection of *Campylobacter spp. Mol Cell Probes* **17**: 187-196.

Ketley, J.M. (1997) Pathogenesis of enteric infection by *Campylobacter. Microbiology* **143**: 5-21.

Kim, C.C., Joyce, E.A., Chan, K., and Falkow, S. (2002) Improved analytical methods for microarray-based genome-composition analysis. *Genome Biol* **3**: 0065 1-17.

King, E.O. (1957) Human infections with *Vibrio fetus* and a closely related vibrio. *J Infect Dis* **101**: 119-128.

Kinsella, N., Guerry, P., Cooney, J., and Trust, T.J. (1997) The *flgE* gene of *Campylobacter coli* is under the control of the alternative sigma factor sigma54. *J Bacteriol* **179**: 4647-4653.

Konkel, M.E., Mead, D.J., Hayes, S.F., and Cieplak, W., Jr. (1992) Translocation of *Campylobacter jejuni* across human polarized epithelial cell monolayer cultures. *J Infect Dis* **166**: 308-315.

Konkel, M.E., Mead, D.J., and Cieplak, W., Jr. (1993) Kinetic and antigenic characterization of altered protein synthesis by *Campylobacter jejuni* during cultivation with human epithelial cells. *J Infect Dis* **168**: 948-954.

Konkel, M.E., Klena, J.D., Rivera-Amill, V., Monteville, M.R., Biswas, D., Raphael, B., Michelson, J. (2004) Secretion of virulence proteins from *Campylobacter jejuni* is dependent on a functional flagellar export apparatus. *J Bacteriol* **186**: 3296-3303.

Konkel, M.E., Kim, B.J., Rivera-Amill, V., and Garvis, S.G. (1999) Identification of proteins required for the internalization of *Campylobacter jejuni* into cultured mammalian cells. *Adv Exp Med Biol* **473**: 215-224.

Koonin, E.V. (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol* **2**:127-136.

Kuroki, S., Saida, T., Nukina, M., Haruta, T., Yoshioka, M., Kobayashi, Y., and Nakanishi, H. (1993) *Campylobacter jejuni* strains from patients with Guillain-Barre syndrome belong mostly to Penner serogroup 19 and contain beta-N-acetylglucosamine residues. *Ann Neurol* **33**: 243-247.

Lastovica, A.J., and le Roux, E. (2000) Efficient isolation of campylobacteria from stools. *J Clin Microbiol* **38**: 2798-2799.

Lee, A., Smith, S.C., and Coloe, P.J. (2000) Detection of a novel campylobacter cytotoxin. *J Appl Microbiol* **89**: 719-725.

Leonard, E.E., 2nd, Takata, T., Blaser, M.J., Falkow, S., Tompkins, L.S., and Gaynor, E.C. (2003) Use of an open-reading frame-specific *Campylobacter jejuni* DNA microarray as a new genotyping tool for studying epidemiologically related isolates. *J Infect Dis* **187**: 691-694.

Leonard, E.E., 2nd, Tompkins, L.S., Falkow, S., and Nachamkin, I. (2004) Comparison of *Campylobacter jejuni* isolates implicated in Guillain-Barre syndrome and strains that cause enteritis by a DNA microarray. *Infect Immun* **72**: 1199-1203.

Linton, D., Karlyshev, A.V., Hitchen, P.G., Morris, H.R., Dell, A., Gregson, N.A., and Wren, B.W. (2000) Multiple N-acetyl neuraminic acid synthetase (*neuB*) genes in *Campylobacter jejuni*: identification and characterization of the gene involved in sialylation of lipo-oligosaccharide. *Mol Microbiol* **35**: 1120-1134.

Lior, H., Woodward, D.L., Edgar, J.A., Laroche, L.J., and Gill, P. (1982) Serotyping of *Campylobacter jejuni* by slide agglutination based on heat- labile antigenic factors. *J Clin Microbiol* **15**: 761-768.

Logan, S.M., Kelly, J.F., Thibault, P., Ewing, C.P., and Guerry, P. (2002) Structural heterogeneity of carbohydrate modifications affects serospecificity of *Campylobacter* flagellins. *Mol Microbiol* **46**: 587-597.

Lorenz, E., Lastovica, A., and Owen, R.J. (1998) Subtyping of *Campylobacter jejuni* Penner serotypes 9, 38 and 63 from human infections, animals and water by pulsed field gel electrophoresis and flagellin gene analysis. *Lett Appl Microbiol* **26**: 179-182.

Maddison, D.R., Swofford, D.L., and Maddison, W.P. (1997) NEXUS: an extensible file format for systematic information. *Syst Biol* **46**: 590-621.

Maddison, D.R.a.M., W. P. (2001) MacClade 4: Analysis of Phylogeny and Character Evolution. Version 4.03. Sunderland, Massachusetts: Sinauer Associates.

Mao, Y., Doyle, M.P., and Chen, J. (2001) Insertion mutagenesis of wca reduces acid and heat tolerance of enterohemorrhagic *Escherichia coli* O157:H7. *J Bacteriol* **183**: 3811-3815.

Marokhazi, J., Waterfield, N., LeGoff, G., Feil, E., Stabler, R., Hinds, J., Fodor, A., and Ffrench-Constant, R.H. (2003) Using a DNA Microarray To Investigate the Distribution of Insect Virulence Factors in Strains of *Photorhabdus* Bacteria. *J Bacteriol* **185**: 4648-4656.

McSweegan, E., and Walker, R.I. (1986) Identification and characterization of two *Campylobacter jejuni* adhesins for cellular and mucous substrates. *Infect Immun* **53**: 141-148.

Medema, G.J., Schets, F.M., van de Giessen, A.W., and Havelaar, A.H. (1992) Lack of colonization of 1 day old chicks by viable, non-culturable *Campylobacter jejuni*. *J Appl Bacteriol* **72**: 512-516.

Meinersmann, R.J., Helsel, L.O., Fields, P.I., and Hiett, K.L. (1997) Discrimination of *Campylobacter jejuni* isolates by fla gene sequencing. *J Clin Microbiol* **35**: 2810-2814.

Miller, S., Pesci, E.C., and Pickett, C.L. (1993) A *Campylobacter jejuni* homolog of the LcrD/FlbF family of proteins is necessary for flagellar biogenesis. *Infect Immun* **61**: 2930-2936.

Mills, S.D., Kuzniar, B., Shames, B., Kurjanczyk, L.A., and Penner, J.L. (1992) Variation of the O antigen of *Campylobacter jejuni* in vivo. *J Med Microbiol* **36**: 215-219.

Morooka, T., Umeda, A., and Amako, K. (1985) Motility as an intestinal colonization factor for *Campylobacter jejuni*. *J Gen Microbiol* **131**: 1973-1980.

Nachamkin, I., Yang, X.H., and Stern, N.J. (1993) Role of *Campylobacter jejuni* flagella as colonization factors for three-day-old chicks: analysis with flagellar mutants. *Appl Environ Microbiol* **59**: 1269-1273.

Nachamkin, I., Allos, B.M., and Ho, T. (1998) *Campylobacter* species and Guillain-Barre syndrome. *Clin Microbiol Rev* **11**: 555-567.

Naess, V., and Hofstad, T. (1984) Chemical composition and biological activity of lipopolysaccharides prepared from type strains of *Campylobacter jejuni* and Campolybacter coli. *Acta Pathol Microbiol Immunol Scand [B]* **92**: 217-222.

Newell, D.G., McBride, H., and Pearson, A.D. (1984) The identification of outer membrane proteins and flagella of *Campylobacter jejuni*. *J Gen Microbiol* **130**: 1201-1208.

Newell, D.G., and Pearson, A. (1984) The invasion of epithelial cell lines and the intestinal epithelium of infant mice by *Campylobacter jejuni/coli*. *J Diarrhoeal Dis Res* **2**: 19-26.

Nuijten, P.J., van Asten, F.J., Gaastra, W., and van der Zeijst, B.A. (1990) Structural and functional analysis of two *Campylobacter jejuni* flagellin genes. *J Biol Chem* **265**: 17798-17804.

Ochman, H., and Jones, I.B. (2000) Evolutionary dynamics of full genome content in Escherichia coli. *Embo J* **19**: 6637-6643.

On, S.L., Nielsen, E.M., Engberg, J., and Madsen, M. (1998) Validity of SmaI-defined genotypes of *Campylobacter jejuni* examined by *SalI*, *KpnI*, and *BamHI* polymorphisms: evidence of identical clones infecting humans, poultry, and cattle. *Epidemiol Infect* **120**: 231-237.

Otto, B.R., van Dooren, S.J., Dozois, C.M., Luirink, J., and Oudega, B. (2002) *Escherichia coli* hemoglobin protease autotransporter contributes to synergistic abscess formation and heme-dependent growth of *Bacteroides fragilis*. *Infect Immun* **70**: 5-10.

Owen, R.J. (1998) Helicobacter--species classification and identification. *Br Med Bull* **54**: 17-30.

Park, S.F., Purdy, D., and Leach, S. (2000) Localized reversible frameshift mutation in the *flhA* gene confers phase variability to flagellin gene expression in *Campylobacter coli*. *J Bacteriol* **182**: 207-210.

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., Quail, M.A., Rajandream, M.-A., Rutherford, K.M., van Vliet, A.H.M., Whitehead, S., and Barrell, B.G. (2000) The

genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403**: 665-668.

Pearson, B.M., Pin, C., Wright, J., I'Anson, K., Humphrey, T., and Wells, J.M. (2003) Comparative genome analysis of *Campylobacter jejuni* using whole genome DNA microarrays. *FEBS Lett* **554**: 224-230.

Pease, A.C., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P., and Fodor, S.P. (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci U S A* **91**: 5022-5026.

Pebody, R.G., Ryan, M.J., and Wall, P.G. (1997) Outbreaks of campylobacter infection: rare events for a common pathogen. *Commun Dis Rep CDR Rev* **7**: R33-37.

Penner, J.L., and Hennessy, J.N. (1980) Passive hemagglutination technique for serotyping *Campylobacter fetus* subsp. *jejuni* on the basis of soluble heat-stable antigens. *J Clin Microbiol* **12**: 732-737.

Perrin, A., Bonacorsi, S., Carbonnelle, E., Talibi, D., Dessen, P., Nassif, X., and Tinsley, C. (2002) Comparative genomics identifies the genetic islands that distinguish *Neisseria meningitidis*, the agent of cerebrospinal meningitis, from other Neisseria species. *Infect Immun* **70**: 7063-7072.

Poly, F., Threadgill, D., and Stintzi, A. (2004) Identification of *Campylobacter jejuni* ATCC 43431-specific genes by whole microbial genome comparisons. *J Bacteriol* **186**: 4781-4795.

Preston, A., Mandrell, R.E., Gibson, B.W., and Apicella, M.A. (1996) The lipooligosaccharides of pathogenic gram-negative bacteria. *Crit Rev Microbiol* **22**: 139-180.

Preston, M.A., and Penner, J.L. (1987) Structural and antigenic properties of lipopolysaccharides from serotype reference strains of *Campylobacter jejuni*. *Infect Immun* **55**: 1806-1812.

Rivera-Amill, V., Kim, B.J., Seshu, J., and Konkel, M.E. (2001) Secretion of the virulence-associated Campylobacter invasion antigens from *Campylobacter jejuni* requires a stimulatory signal. *J Infect Dis* **183**: 1607-1616.

Roberts, I.S. (1996) The biochemistry and genetics of capsular polysaccharide production in bacteria. *Annu Rev Microbiol* **50**: 285-315.

Rodrigues, L.C., Cowden, J.M., Wheeler, J.G., Sethi, D., Wall, P.G., Cumberland, P., Tompkins, D.S., Hudson, M.J., Roberts, J.A., and Roderick, P.J. (2001) The study of infectious intestinal disease in England: risk factors for cases of infectious intestinal disease with *Campylobacter jejuni* infection. *Epidemiol Infect* **127**: 185-193.

Ronquist, F., and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572-1574.

Russell, R.G., O'Donnoghue, M., Blake, D.C., Jr., Zulty, J., and DeTolla, L.J. (1993) Early colonic damage and invasion of *Campylobacter jejuni* in experimentally challenged infant Macaca mulatta. *J Infect Dis* **168**: 210-215.

Salama, N., Guillemin, K., McDaniel, T.K., Sherlock, G., Tompkins, L., and Falkow, S. (2000) A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc Natl Acad Sci U S A* **97**: 14668-14673.

Schirm, M., Soo, E.C., Aubry, A.J., Austin, J., Thibault, P., and Logan, S.M. (2003) Structural, genetic and functional characterization of the flagellin glycosylation process in *Helicobacter pylori*. *Mol Microbiol* **48**: 1579-1592.

Schouls, L.M., Reulen, S., Duim, B., Wagenaar, J.A., Willems, R.J., Dingle, K.E., Colles, F.M., and Van Embden, J.D. (2003) Comparative genotyping of *Campylobacter jejuni* by amplified fragment length polymorphism, multilocus sequence typing, and short repeat sequencing: strain diversity, host range, and recombination. *J Clin Microbiol* **41**: 15-26.

Sebald, M., and Veron, M. (1963) [Base DNA Content and Classification of Vibrios.]. *Ann Inst Pasteur (Paris)* **105**: 897-910.

Skirrow, M.B. (1977) *Campylobacter* enteritis: a "new" disease. *Br Med J* **2**: 9-11.

Skirrow, M.B. (1991) Epidemiology of *Campylobacter* enteritis. *Int J Food Microbiol* **12**: 9-16.

Skirrow, M.B. (1994) Diseases due to *Campylobacter*, *Helicobacter* and related bacteria. *J Comp Pathol* **111**: 113-149.

Song, Y.C., Jin, S., Louie, H., Ng, D., Lau, R., Zhang, Y., Weerasekera, R., Al Rashid, S., Ward, L.A., Der, S.D., Chan, V.L. (2004) FlaC, a protein of campylobacter

jejuni TGH9011 (ATCC43431) secreted through the flagellar apparatus, binds epithelial cells and influences cell invasion. *Mol Microbiol* **53**:541-553

Spohn, G., and Scarlato, V. (1999) Motility of *Helicobacter pylori* is coordinately regulated by the transcriptional activator FlgR, an NtrC homolog. *J Bacteriol* **181**: 593-599.

Stein, M., Kenny, B., Stein, M.A., and Finlay, B.B. (1996) Characterization of EspC, a 110-kilodalton protein secreted by enteropathogenic *Escherichia coli* which is homologous to members of the immunoglobulin A protease-like family of secreted proteins. *J Bacteriol* **178**: 6546-6554.

Stintzi, A. (2003) Gene expression profile of *Campylobacter jejuni* in response to growth temperature variation. *J Bacteriol* **185**: 2009-2016.

Suerbaum, S., Lohrengel, M., Sonnevend, A., Ruberg, F., and Kist, M. (2001) Allelic diversity and recombination in *Campylobacter jejuni*. *J Bacteriol* **183**: 2553-2559.

Szymanski, C.M., King, M., Haardt, M., and Armstrong, G.D. (1995) *Campylobacter jejuni* motility and invasion of Caco-2 cells. *Infect Immun* **63**: 4295-4300.

Szymanski, C.M., Yao, R., Ewing, C.P., Trust, T.J., and Guerry, P. (1999) Evidence for a system of general protein glycosylation in *Campylobacter jejuni*. *Mol Microbiol* **32**: 1022-1030.

Szymanski, C.M., Burr, D.H., and Guerry, P. (2002) *Campylobacter* protein glycosylation affects host cell interactions. *Infect Immun* **70**: 2242-2244.

Szymanski, C.M., Logan, S.M., Linton, D., and Wren, B.W. (2003) *Campylobacter*--a tale of two protein glycosylation systems. *Trends Microbiol* **11**: 233-238.

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., and Nash, J.H. (2004) Large-scale comparative genomics meta-analysis of *Campylobacter jejuni* isolates reveals low level of genome plasticity. *J Clin Microbiol* **42**: 4566-4576.

Tam, C.C., O'Brien, S.J., Adak, G.K., Meakins, S.M., and Frost, J.A. (2003) *Campylobacter coli* - an important foodborne pathogen. *J Infect* **47**: 28-32.

Thibault, P., Logan, S.M., Kelly, J.F., Brisson, J.R., Ewing, C.P., Trust, T.J., and Guerry, P. (2001) Identification of the carbohydrate moieties and glycosylation motifs in *Campylobacter jejuni* flagellin. *J Biol Chem* **276**: 34862-34870.

Thornley, J.P., Jenkins, D., Neal, K., Wright, T., Brough, J., and Spiller, R.C. (2001) Relationship of *Campylobacter* toxigenicity in vitro to the development of postinfectious irritable bowel syndrome. *J Infect Dis* **184**: 606-609.

van Spreeuwel, J.P., Duursma, G.C., Meijer, C.J., Bax, R., Rosekrans, P.C., and Lindeman, J. (1985) *Campylobacter* colitis: histological immunohistochemical and ultrastructural findings. *Gut* **26**: 945-951.

Walker, R.I., Caldwell, M.B., Lee, E.C., Guerry, P., Trust, T.J., and Ruiz-Palacios, G.M. (1986) Pathophysiology of *Campylobacter* enteritis. *Microbiol Rev* **50**: 81-94.

Wang, G., and Maier, R.J. (2004) An NADPH quinone reductase of *Helicobacter pylori* plays an important role in oxidative stress resistance and host colonization. *Infect Immun* **72**: 1391-1396.

Wassenaar, T.M., Bleumink-Pluym, N.M., and van der Zeijst, B.A. (1991) Inactivation of *Campylobacter jejuni* flagellin genes by homologous recombination demonstrates that *flaA* but not *flaB* is required for invasion. *Embo J* **10**: 2055-2061.

Wassenaar, T.M., van der Zeijst, B.A., Ayling, R., and Newell, D.G. (1993) Colonization of chicks by motility mutants of *Campylobacter jejuni* demonstrates the importance of flagellin A expression. *J Gen Microbiol* **139 (Pt 6)**: 1171-1175.

Wassenaar, T.M., Bleumink-Pluym, N.M., Newell, D.G., Nuijten, P.J., and van der Zeijst, B.A. (1994) Differential flagellin expression in a *flaA flaB*+ mutant of *Campylobacter jejuni*. *Infect Immun* **62**: 3901-3906.

Wheeler, W.E., and Borchers, J. (1961) Vibrionic enteritis in infants. *Am J Dis Child* **101**: 60-66.

Wood, R.C., MacDonald, K.L., and Osterholm, M.T. (1992) *Campylobacter* enteritis outbreaks associated with drinking raw milk during youth activities. A 10-year review of outbreaks in the United States. *Jama* **268**: 3228-3230.

Wosten, M.M., Wagenaar, J.A., van Putten, J.P. (2004) The Flgs/FlgR two component signal transduction system regulates the *fla* regulon in *Campylobacter jejuni*. *J. Biol Chem* **279**:16214-22.

Yao, R., Burr, D.H., and Guerry, P. (1997) CheY-mediated modulation of *Campylobacter jejuni* virulence. *Mol Microbiol* **23**: 1021-1031.

Yuki, N. (1997) Molecular mimicry between gangliosides and lipopolysaccharides of *Campylobacter jejuni* isolated from patients with Guillain-Barre syndrome and Miller Fisher syndrome. *J Infect Dis* **176 Suppl 2**: S150-153.

Zaharik, M.L., Gruenheid, S., Perrin, A.J., and Finlay, B.B. (2002) Delivery of dangerous goods: type III secretion in enteric pathogens. *Int J Med Microbiol* **291**: 593-603.

Zhang, Q., Meitzler, J.C., Huang, S., Morishita, T. (2000). Sequence polymorphisms, predicted secondary structures, and surface-exposed conformational epitopes of Campylobacter major outer membrane protein. *Infect Immun* **68**: 5679-5689

Zhou, D., Han, Y., Song, Y., Tong, Z., Wang, J., Guo, Z., Pei, D., Pang, X., Zhai, J., Li, M., Cui, B., Qi, Z., Jin, L., Dai, R., Du, Z., Bao, J., Zhang, X., Yu, J., Huang, P., and Yang, R. (2004) DNA microarray analysis of genome dynamics in *Yersinia pestis*: insights into bacterial genome microevolution and niche adaptation. *J Bacteriol* **186**: 5138-5146.