Identification of Genetic Differences Between Strains of *Campylobacter jejuni*

Emily J. Kay

University of Cambridge

Darwin College

2005

This dissertation is submitted for the degree of Doctor of Philosophy.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

Emily Kay

1/8/05

Acknowledgements

First of all I would like to thank my supervisor, Julian for all his help and support over the past (nearly) four years. I would also like to thank my emergency back-up supervisor, Al Ivens for all his sage advice on the use of radioactivity. There are so many people that I have variously pestered and stolen reagents from over the past years that I will probably forget to acknowledge all of them.

There are so many people at the Sanger without whom this work would not have been possible; including all of Mike Quail's team, for providing helpful advice on cloning,
the people of the YAC lab/ clone resources for use of hoods and the gridding machine, and Andy Mungall for letting us use his hotroom.

Then there are all the collaborators from places near and far who have helped me no end, including Brendan Wren's group at the LSHTM for supplying me with copious quantities of DNA, especially Olivia who is a star; Duncan Maskell's group at the Vet labs in Cambridge for letting me do some experiments with real live Campy, especially Chris who was very patient and helpful despite me not having the faintest clue what I was doing half the time, and Diane Taylor's group in Canada, especially Ameera, again for sending large quantities of DNA

Special thanks have to go to the people who shared an office with me over the last few months, including Helena, keeper of the teabags and master of the kettle, Ali the random gossip factory and John the only guy in the office, I know I haven't been the easiest person to put up with lately. I thank you all for your witty office banter, arguments and the constant supply of caffeinated beverages. I must apologise especially to Helena (and Ali) for subjecting you to rough drafts of my thesis.

Finally to my housemates; thanks for your patience, and to my parents; thanks as always for your support.

Abstract

The bacterial pathogen *Campylobacter jejuni* is known to cause a range of diseases from inflammatory diarrhoea to the autoimmune Guillain-Barré syndrome, although in some individuals infection with *C. jejuni* can be asymptomatic. The difference in outcome of the infection is likely to be the result of a number of factors including genetic differences between the infecting strains and susceptibility of the infected individual. As *C. jejuni* is known to be genetically variable, this project has involved the comparison of a number of unsequenced strains of *C. jejuni* against the sequenced strain 11168, in order to discover novel chromosomal sequences that may be responsible for the different phenotypes of these strains.

Four strains, representing a range of clinical outcomes and survival in different environmental niches were compared against the sequenced strain 11168 using a nylon macro-array based technique. This has resulted in the identification of 483 Kb of sequence containing 595 novel predicted genes within small-insert genomic libraries. Many of the novel predicted genes are associated with surface polysaccharide, flagellar biosynthesis and modification in addition to hypothetical genes. Also a number of genes identified were associated with restriction modification, metabolism and respiration. A few predicted proteins showed homology to genes associated with transposons, plasmid conjugation, chemotaxis and adhesion. Using this data 31 larger-insert BAC clones containing predicted genes of interest were identified and sequenced in order to determine the extent of these chromosomal islands. Within these sequences predicted genes have been identified that might be implicated in the distinct phenotypes of these strains.

A chromosomal tetracycline resistance determinant was discovered amongst remnants of transposon associated genes. A similar insert was found in two of eight tetracycline resistant clinical isolates studied. This presents the possibility that a transposon may be responsible for disseminating tetracycline resistance in some strains of *C. jejuni.*

Two plasmids from one of the strains used in this study were sequenced. In one of the plasmids an inverting region was discovered and analysed, and the possibility that this is responsible for variable expression of a type IV secretion system was investigated.

Glossary of terms

aa	amino acid
ACT	Artemis Comparison Tool
AFLP	amplified fragment length polymorphism
AIDS	Aquired Immunodeficiency Syndrome
BAC	Bacterial Artificial Chromosome
BAP	Bacterial Alkaline Phosphatase
Bp	Base pair(s)
BSA	Bovine Serum Albumin
C	Cytosine
CDS	Coding Sequence
CDSC	Communicable Disease Surveillance Centre
CDT	Cytolethal Distending Toxin
dATP	2′-Deoxyadenosine 5′-triphosphate
dCTP	2′-Deoxycytidine 5′-triphosphate
DDW	double distilled water
dGTP	2′-Deoxyguanosine 5′-triphosphate
DMSO	dimethyl sulphoxide
dTTP	2′-Deoxythiamine 5′-triphosphate
EDTA	ethylene diamine tetra-acetic acid
FSA	Food Standards Agency
G	Guanine
GBS	Guillain-Barré syndrome
GGT	gamma-glutamyl transpeptidase
HLA	Human Lymphocyte Antigen
HR-MAS NMR	High-resolution magic angle spinning nuclear magnetic resonance
Id	Identity
IgG	immunoglobulin G
IPA	Invasion Protein Antigen
IPTG	isopropylthio-β-D-galactoside
IVS	Intervening sequence
Kb	kilobase(s)
LB	Luria-Bertani broth
LMP	Low Melting Point
LOS	Lipooligosaccharide

LPS	Lipopolysaccharide
LSHTM	The London School of Hygiene and Tropical Medicine
MFS	Miller Fisher Syndrome
MLST	Multilocus Sequence Typing
MSP	Maximal Segment Pairs
MW	Molecular Weight
PCR	Polymerase Chain Reaction
PEG	polyethylene glycol
PFGE	pulsed-field gel electrophoresis
PMSF	phenylmethylsulfonyl fluoride
RAPD	Randomly amplified polymorphic DNA
RFLP	restriction fragment length polymorphism
RM	restriction-modification
SDS	Sodium Dodecyl Sulphate
SOD	Superoxide Dismutase
SSD	Strain Specific DNA
ST	Sequence Type
T	Thymine
TE	Tris- EDTA
TLP	Transducer-like protein
TMAO	trimethylamine-*N*-oxide
TPS	Two Partner Secretion
TYE	Tryptone Yeast Extract
UV	Ultra Violet
WU-BLAST	Washington University-Basic Local Alignment Sequence Tool
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactoside
YOP	*Yersinia* outer protein

Table of Contents

1. Introduction

1.1 Background

Campylobacter jejuni is the most common cause of bacterial diarrhoeal disease worldwide. Little is known about the ability of this organism to cause disease. A wide range of phenotypic and genotypic diversity has been reported for this species along with a range of disease outcomes.

1.1.1 Classification

C. jejuni belongs to the delta-epsilon group of proteobacteria within the family *Campylobactereaceae,* which also includes the genera *Helicobacter* and *Arcobacter*. The delta-epsilon group is significantly divergent from the gamma subgroup which contains many human enteropathogens such as the salmonellae, *Escherichia coli* and *Shigella* [1]. The genus *Campylobacter* now includes about twenty species and subspecies, eight of which are known to cause human gastrointestinal disease [2]. *Campylobacter* was previously thought to be purely a pathogen of animals until the 1970s when it was discovered that *Campylobacter* caused diarrhoea in man [3].

Campylobacter are Gram-negative with a low G+C content chromosome (30%). They are non-spore forming, spiral rod shaped bacteria 0.2-0.8 µm wide and 0.5-5 µm long. Cells are typically motile and move in a corkscrew-like motion propelled by a single polar flagellum. They require a microaerobic environment for growth and are thermophilic with an optimum growth temperature of 42-43°C [1].

1.1.2 Physiology and metabolism

1.1.2.1 Growth

Many aspects of the physiology and metabolism of these organisms remain poorly understood due in part to difficulties in cultivating members of the *Campylobacter* genus. The majority of *Campylobacter* strains need to be cultured in a microaerobic environment consisting of 5-10% (v/v) oxygen and 5-10% (v/v) carbon dioxide in complex growth media with additional supplements [4]. *C. jejuni* is susceptible to a wide variety of antimicrobial treatments and food processing methods such as drying, freezing and salting. *C. jejuni* is also sensitive to osmotic stress and osmotic pressure, oxygen concentrations above 5% and has not been reported to grow at temperatures below 30°C [2].

After several days of *in vitro* culture, cells have been noted to change from spiral to coccoid forms accompanied by a loss of culturability. It remains controversial as to whether this represents a survival mechanism or is a degenerative form but it has been reported that the change to coccoid forms does not require *de novo* protein synthesis. This suggests that the change to coccoid forms is not actively controlled and therefore may represent cell injury [5]. *C. jejuni* is not thought to mount a stationary-phase response to limited nutrient availability, or the build up of toxic waste products, which is characterized in many other bacteria by increased resistance to environmental stress. Resuscitation of aged cultures has been demonstrated but is more likely to represent growth in numbers of residual viable cells rather than reversal of a viable but not culturable state [6]. It is possible that a subpopulation within stationary phase cultures of *Campylobacter* may be better able to cope with injury [7].

1.1.2.2 Transport and iron uptake

The genome sequences of *C. jejuni* show that it has a limited capacity for biosynthesis and therefore many transport systems exist for the acquisition of essential amino acids, other nutrients and ions from the external environment [4;8;9].

Iron is essential for all organisms as it is a cofactor of many enzymes e.g. peroxidases and cytochromes; in addition it is used in electron transport and redox reactions [10]. Iron-uptake systems are often considered to be virulence factors as iron availability is limited in mammalian host tissues. In response to low iron availability bacteria may produce low-molecular weight iron chelators called siderophores [10]. It has been suggested that certain strains of *C. jejuni* may produce siderophores; however, no siderophores have been characterized from *C. jejuni* [11]. *C. jejuni* may be able to scavenge siderophores produced by other bacteria in the intestinal tract [12] as several systems for the uptake of iron complexed to siderophores have been discovered. These uptake systems include the *ceuBCDE* operon for the uptake of enterochelin and *cfrA* which has been proposed as a ferric enterobactin receptor [13] and which is only present in some strains [14]. Also cj0178 may be the receptor of an as yet unidentified iron source [10]. In addition a haemin/haemoglobin uptake system exists encoded by *chuABCD* with a *chuA* mutant being unable to use haemoglobin or haemin as an iron source [13]. Haem compounds may be released by the host at the site of inflammation [12] and therefore be accessible to *C. jejuni* once it leaves the intestinal tract.

Excess iron can be toxic to cells causing oxidative stress therefore iron uptake is tightly regulated. In *C. jejuni* there are two iron-response regulators, the ferric uptake regulator Fur and PerR, which regulates peroxide stress defence proteins AhpC and KatA [4], underlining the link between iron uptake and oxidative stress resistance. *C. jejuni* also possesses the ability to store iron for iron-limited environments as ferritin (*cft*) [15].

1.1.2.3 Carbon metabolism

C. jejuni has no phosphofructokinase, a glycolysis enzyme, and therefore cannot significantly metabolize externally supplied sugars. *C. jejuni* has been proposed to obtain carbon and energy requirements from tricarboxylic acid (TCA) cycle intermediates, some organic acids and amino acids; in particular *C. jejuni* has been shown to grow in culture using serine, glutamate, aspartate, asparagine, glutamine and proline as the sole carbon source [16]. A recent study has shown that *Campylobacter* strains fall into three distinct metabolic groups: 91% of *C. jejuni* strains tested were able to oxidize α–ketoglutarate, succinate, fumarate and aspartic acid; 7% of *C. jejuni* strains were unable to metabolize α-ketoglutatarate and 2% of *C. jejuni* strains were unable to oxidize succinate, fumarate and aspartic acid [17]. This highlights the fact that there is metabolic diversity between different strains of *C. jejuni*. Proteases might be important to the nutrition of the organism under carbon-limiting conditions by breaking down proteins into constituent parts that can be fed into TCA cycle [16].

1.1.2.4 Electron transport

C. fetus has been shown to grow anaerobically by respiring formate and fumarate in a similar way to *Wolinella succinogenes* along with some other members of the *Campylobacter* genus. *C. jejuni* in contrast has been reported not to be able to grow anaerobically [18] even though the genome sequence of strain NCTC 11168 revealed the presence of genes for fumarate reductase and other genes known to be involved in anaerobic electron transport pathways from other bacteria [4]. Electron acceptors other than oxygen may be important for growth in the avian gut and also the mammalian gut where oxygen is limited. The respiratory chain in *C. jejuni* appears to be highly branched and complex with many cytochromes [4] suggesting an ability to adapt to different environmental conditions.

1.1.3 *Campylobacter* infection

1.1.3.1 Epidemiology

About 90% of human *Campylobacter* isolates in England and Wales are *C. jejuni* with most of the remainder being *C. coli* [19]. Reports of campylobacteriosis are not normally distinguished at the species level so the peak of 58,059 cases in 1998 were reported to the Communicable Disease Surveillance Centre (CDSC) simply as *Campylobacter* [20]. More recent figures would tend to suggest a decrease in cases over the past few years, with the Health Protection Agency (HPA) receiving 47,597 laboratory reports of *Campylobacter* in faecal isolates during 2002, 7,317 less than the previous year [21]; this trend is also apparent in the USA [22].

In the UK there is a seasonal variation in incidence, with a peak in late spring, a lesser peak in autumn and a winter low. Regional variation also occurs with a greater incidence in rural rather than urban populations [16]. The incidence of campylobacteriosis is highest in males under 1 year old with a second peak occurring in adults aged 25-34. Incidence is higher in males than females for all age groups [20].

Even though cases of acute gastroenteritis caused by *Campylobacter* now outnumber those caused by *Salmonella,* outbreaks of campylobacteriosis are rarer than outbreaks of salmonellosis; only 12 general outbreaks of *Campylobacter*, affecting 239 people were reported to the Communicable Disease Surveillance Centre (CDSC) between 1995 and 1996, compared to 233 outbreaks of salmonellosis involving 4,946 people [20]. Part of the reason for this could be that *Campylobacter* does not multiply in foods; however only a low dose is required to cause disease: 50-100 cells if not lower, depending on the infecting strain [2].

C. jejuni is found naturally in the gastrointestinal tract of birds (particularly poultry), cattle and domestic pets, where it rarely causes disease. Transmission to humans has been reported from a variety of sources including raw or undercooked meat, especially poultry

[2;20]. The Food Standards Agency (FSA) has quoted that an average of 50% of retail chickens in the United Kingdom are contaminated by *Campylobacter* [23] although reports of contamination vary dramatically with location, sampling season and different producers of raw retail chickens [24;25].

Other sources of infection are unpasteurised milk, bird-pecked milk on doorsteps and untreated water: *Campylobacter* may be shed into surface water by birds and can survive for many weeks at low temperatures. However most infections remain unexplained by recognised risk factors [20].

1.1.3.2 Disease outcomes

C. jejuni can cause a spectrum of disease ranging from asymptomatic colonisation to severe inflammatory diarrhoea and has also been associated with bacteraemia, endocarditis, meningitis, urinary tract infection and other extraintestinal diseases including Guillain-Barré syndrome (GBS) and Miller Fisher syndrome (MFS) [26]. Post infective complications are rare with about 1% developing reactive arthritis and about 0.1% developing GBS [16]. GBS is an autoimmune-mediated disorder of the peripheral nervous system resulting in paralysis [27] and MFS is considered a variant of GBS that causes paralysis of ocular muscles, ataxia and a loss of tendon reflex [28]. GBS has been associated with *C. jejuni* serotypes 0:41[29] and 0:19[30]; with the latter serotype (0:19) the risk of developing GBS may be as high as 1 in 150 [29]. Amplified fragment length polymorphism (AFLP) analysis has shown that strains associated with GBS or MFS do not belong to a distinct genetic group [31] suggesting that host factors are a major determining factor in the onset of these disorders.

In its uncomplicated form campylobacteriosis is characterised by fever, abdominal cramping and diarrhoea (with or without faecal leukocytes). The incubation period is normally 1-7 days after which profuse diarrhoea frequently lasts 2-3 days accompanied by acute abdominal pains [26]. The average duration of illness calculated from nine outbreaks was 4.6 days although one third of patients from these outbreaks were ill for more than seven days [32]. The disease is usually self-limiting and in the majority of cases people recover without the aid of antibiotics; the relapse rate being 5-10% [26]. The proportion of patients admitted to hospital varies but is generally cited as between 5-10% and fatalities are rare, usually only occurring in the elderly or the immunocompromised [32]. In circumstances where antibiotic treatment is necessary, for example in prolonged or systemic infection, erythromycin is used as the drug of choice but fluoroquinolones and tetracycline may also be used [26;32].

There is a marked discrepancy between disease outcomes in developed and developing countries. In developed countries campylobacteriosis can be quite severe with bloody diarrhoea a common feature of infection whereas in developing countries diarrhoea is more likely to be watery. In developing countries the disease predominantly affects young children possibly relating to developed immunity resulting in subsequent asymptomatic infection [33].

It has been reported that travellers abroad are more likely to develop disease symptoms similar to those they would develop if they contracted the disease in their own country. This suggests that host susceptibility is an important factor and that differences in disease outcome are unlikely to be solely due to strain differences in geographically separate areas. Indeed in human volunteer studies the same strain can cause different severity of illness in different people [16].

Immunocompromised people are more at risk of developing disease, with one report suggesting that in Acquired Immunodeficiency Syndrome (AIDS) patients in Los Angeles during the period 1983-1987 the incidence of disease was 39 times higher than in the general population. This may however be an overestimate as AIDS patients are considered to be more likely to report to health services on development of symptoms [34]. It has also been suggested that immunocompromised people often develop a more severe form of disease [26]. There is also a strong association between reactive arthritis, a rare complication, and people who have the human lymphocyte antigen HLA-B27 [34]. These facts highlight the role of the immune system in severity of disease.

The reasons for variable host response are not clear but may depend on a combination of the virulence of the infecting strain, the challenge dose, and the susceptibility of the patient [32].

1.1.4 Pathogenesis

1.1.4.1 Models of infection

Although *C. jejuni* causes a large number of infections each year the bacterium is nutritionally fastidious and supposedly extremely susceptible to environmental stresses (section 1.1.2.1). This apparent paradox highlights the fact that *C. jejuni* remains a poorly understood pathogen.

Part of the reason for the poor understanding of this pathogen is due to a lack of suitable animal models to assess virulence [35]. *Campylobacter* has the ability to colonize the intestinal tract of many animals including humans, pigs, cattle and birds but in most hosts *Campylobacter* behaves as a commensal gut organism. Only primates and possibly ferrets show disease outcomes similar to those in humans with other animal models e.g. chickens and mice being used for colonization studies. So far factors which may explain why only a restricted number of species succumb to disease have been elusive [16].

Human volunteer challenges have been used in the past to study pathogenicity. In these reports large oral inocula were needed to cause illness [36] whereas low doses have been implicated in waterborne outbreaks [29] so even this may be a poor reflection of how *C. jejuni* causes infection naturally.

Despite the above limitations to experimental determination of *C. jejuni* pathogenicity certain factors have been shown to be important for disease progression as described below.

1.1.4.2 Motility

The primary stage of infection or colonization involves *Campylobacter* moving towards cell surfaces. The distinctive type of corkscrew-like movement displayed by *Campylobacter* is thought to be an adaption allowing penetration of the mucus overlying the intestinal epithelium [12]. This movement is mediated by flagella. There are two genes encoding flagellin in *C. jejuni: flaA* and *flaB*, and these genes are arranged in tandem on the chromosome and show a high degree of sequence identity to each other. The flagella are constructed from multimers of flagellin; FlaA flagellin protein is the major component with a small amount of FlaB flagellin protein. These flagellin proteins are attached by a hook protein to a basal structure which is embedded in the membrane and, along with the stator units (MotA and MotB plus FliMNG), acts as a motor for rotation [29].

The flagella are post-translationally modified by phosphorylation and glycosylation [37] and they exhibit phase and antigenic variation [15]. The *flaA* and *flaB* genes are independently transcribed by different types of promoter, σ^{28}- and σ^{54}-dependent respectively. The expression of *flaB* seems to be environmentally regulated by temperature and pH [16]. It has also been demonstrated that the flagella are able to secrete proteins into the extracellular milieu [38].

Motility is directed by chemotaxis which allows the organism to locate and move towards the mucus layer of the gut [16]. Mucin, L-serine and L-fucose all act as chemoattractants for *C. jejuni* and bile acids act as chemorepellants [16]. Ten proteins containing methyl-accepting chemotaxis domains have been identified within the genome sequence of *C. jejuni* strain NCTC 11168 [8;12].

1.1.4.3 Adherence

Several adhesins have been described in *Campylobacter* including flagellin, lipopolysaccharide (LPS) and a number of membrane proteins [16]. In the case of flagella it is difficult to separate adhesion from motility functions. Binding to host cells is proposed to be mediated by proteins synthesized constitutively as heat killed bacteria are still able to bind [39]. Different strains have been noted to vary in their ability to adhere to epithelial cells *in vitro* [40].

A mutant of PEB1, a homologue of cluster 3 binding proteins of bacterial ABC transporters, showed a 50-100 fold decrease in adherence and 15-fold decrease in invasion of epithelial cells in culture [41]. An outer membrane protein CadF (Campylobacter adhesion to fibronectin) is required for adherence and may stimulate invasion upon binding to fibronectin [42]. A lipoprotein, JlpA, has also been shown to be an adhesin [43].

CheY, which affects the rate of flagellar motor switching, has been shown to have an effect on both adherence and invasion. *C. jejuni* mutants containing two copies of *cheY* are non-adherent and non-invasive and *cheY*$^{-}$ strains are hyperadherent and hyperinvasive *in vitro* [44].

1.1.4.4 Cellular Invasion

Strains of *C. jejuni* differ in their ability to invade human cell lines *in vitro* [16;40]. Once *C. jejuni* has attached to a gut epithelial surface a subpopulation goes on to invade the epithelial cells; this invasion has been correlated with inflammatory disease [12]. At least two mechanisms of invasion have been proposed. The first mechanism involves actin reorganization and accumulation within the mammalian cell, beneath the site of bacterial attachment, followed by microfilament-mediated uptake. The second mechanism is microfilament independent and instead utilises a microtubule mediated uptake system involving coated pit formation. In addition both clathrin-coated pits and clathrin-independent caveolae have been implicated in endocytosis of the pathogen [45]. This data suggests that there may be several different mechanisms at work.

C. jejuni invasion is dependent on both *de novo* synthesized bacterial proteins and host cell signal transduction [12]: at least 8 bacterial proteins are produced and secreted upon co-cultivation of *C. jejuni* and the human embryonic intestinal cell line, INT-407. One of these proteins is CiaB which has been identified as necessary for internalization into cells [46].

In addition to intracellular invasion, translocation across cell monolayers has been observed and has been shown to be inhibited by chloramphenicol, suggesting that *de novo* bacterial protein synthesis is required. Some strains are able to translocate across confluent cell monolayers despite being classed as non-invasive and indeed electron microscopy studies have shown bacteria passing through and between host cells suggesting that both paracellular and transcellular routes are used [15].

1.1.4.5 Intracellular survival

After invasion of the intestinal epithelial cells, *C. jejuni* appear to be largely confined within endosomal vacuoles [45]. Bacterial numbers in INT407 cells have been shown to decrease after phagosome-lysosome fusion after which most bacteria have adopted a coccoid morphology and there is little evidence of intracellular multiplication [45]. *Campylobacter* have also been observed free in the cytoplasm of cells *in vitro* and *in vivo* [45].

Campylobacter may translocate across cells allowing the bacteria to reach the bloodstream and deeper tissues [29]. *Campylobacter* are sensitive to complement-mediated lysis and as such are thought to be rapidly killed upon traversing the epithelium [29]. The host inflammatory response leads to polymorphonuclear leucocytes and monocytes infiltrating intestinal epithelium [15]. If *Campylobacter* do survive early immune responses and circulate in the bloodstream they will eventually be taken up by macrophages where the bacteria may survive for up to seven days [29]. Strain differences in serum resistance [47] and phagocyte-mediated killing have been reported [16]. Mechanisms of intracellular survival are unknown although serum resistance has been linked to sialylation of lipooligosaccharide [48].

1.1.4.6 Toxins

As the levels of invading *Campylobacter* are thought to make up less than 1% of applied bacteria on a monolayer of cells in culture, the action of toxins has been proposed as a pathogenesis factor [49], indeed certain aspects of *Campylobacter* disease would be consistent with the action of toxins.

There are two major classes of proteinaceous toxins which would be relevant to the observed pathogenicity of *Campylobacter*; enterotoxins and cytotoxins. Enterotoxins bind to cellular receptors, enter the cell and elevate intracellular cAMP levels causing excess secretion of fluid resulting in watery diarrhoea. Cytotoxins kill target cells by inhibition of cellular protein synthesis or inhibition of actin filament formation [2] which may be consistent with diarrhoea containing blood and inflammatory cells [16].

At least six different types of toxin have been proposed to be encoded by *Campylobacter* strains including cholera-like toxin and various cytotoxins [29]. Reports of enterotoxin production vary widely between isolates with some isolates producing enterotoxin and some not. No correlation was found between enterotoxin production and prevalent Lior or Penner serotypes (see section 1.1.5.1 for explanation of serotype schemes). It has been suggested that enterotoxin production results in watery type diarrhoea as opposed to inflammatory bloody diarrhoea. However this could be determined by host factors (section 1.1.3.2) [49].

Several different cytotoxins have been proposed to be encoded by *Campylobacter* including cytolethal distending toxin (CDT). The early effects of CDT are similar to those of an enterotoxin but after several days cells show distension and death [49]. CDT appears to be able to cause some diarrhoeal symptoms in the rat-ileal-loop assay [50] but the mechanisms of action of this toxin are unknown. CDT is encoded by *cdtABC* which have been found in all *C. jejuni* strains studied so far [29;51]. The CDT locus is found in other *Campylobacter* species although it shows high sequence divergence between species [50]. The amount of toxin produced by isolates varies even though all strains seem to possess the genes that encode the toxin [49].

Haemolytic toxins have also been described, with 92% of *C. jejuni* strains tested showing haemolysis on blood agar. Hepatotoxin and shiga-like toxin have also been proposed but studies on these toxins are contradictory [49].

With the exception of CDT, genes encoding proposed toxins have not been isolated or identified from sequencing projects [29]. The importance of these proposed toxins in disease remains unclear.

1.1.4.7 Surface polysaccharide structures

All strains produce lipooligosaccharide (LOS); a lipid A molecule joined to core oligosaccharide. LOS with a structure mimicking human ganglioside GM1, thought to be produced to evade the immune system, has been postulated to be an important factor in the development of GBS [52]. In addition it was thought that about one third of strains also produced a high molecular-weight lipopolysaccharide (LPS) which contained an O-chain consisting of repeating oligosaccharide. This has since been shown to be a capsular polysaccharide. Interestingly, capsule has been shown to be present even in strains previously not thought to produce LPS [53]. The capsule has also been proved to be the basis of the Penner typing scheme (section 1.1.5.1), and it has been proposed to aid surface spreading, contribute to serum resistance, phagocytic killing and cell toxicity [15]. Capsular polysaccharide is poorly immunogenic which may aid in resistance to host-specific immune response [35] and may protect the cells from desiccation when in the environment.

1.1.4.8 Environmental survival

Anything which aids growth or survival in either the host or general environment could be considered a virulence factor if it aids transmission. Oxygen stress defences are used to deal with toxic oxygen metabolites produced during normal metabolism, during transmission or when in contact with host immune defences [12]. Superoxide stress defence is mediated by the superoxide dismutase (SOD) SodB whilst peroxide stress defence is mediated by catalase and alkyl hydroperoxide reductase (AhpC) [12].

The majority of *Campylobacters* are phenotypically catalase-positive (KatA). Catalase reduces oxygen stress by detoxifying hydrogen peroxide to oxygen and water [10]. Prior exposure to oxidative stress has been shown to increase rates of invasion, and catalase has been shown to contribute to intramacrophage survival. However it does not play a role in intraepithelial cell survival [54].

SodB is a superoxide dismutase which catalyzes the breakdown of superoxides into hydrogen peroxide and oxygen; mutants in *sodB* are attenuated in intracellular survival and colonization [10]. AhpC alkyl hydroperoxide reductase converts reactive hydroperoxides to the corresponding alcohols [10].

The ability of *Campylobacter* to persist outside the host environment may be an important factor in transmission and therefore a determinant of which strains may infect individuals as *C. jejuni* is thought to be susceptible to a wide range of environmental stresses and does not grow at ambient temperatures. *Campylobacter* lose culturability at different rates when transferred to water [16]. *Campylobacter* are found in natural water sources throughout the year but appear to survive better when the water is cold [34]. Some strains are able to survive for up to four weeks at 4-10°C: this persistence was increased when bacteria were introduced with biofilms of indigenous water flora [16]. Factors determining strain variation in persistence are currently poorly understood but different isolates may vary in the virulence determinants they carry [16].

1.1.4.9 Plasmids

Recently the role of plasmids in *Campylobacter* virulence has been studied [55;56]. However, not all highly invasive strains have plasmids; the prevalence of plasmids in *Campylobacter* has been estimated by several sources as 19-53% [55]. In a survey of 688 isolates from diverse sources 32% were found to harbour plasmid DNA of size ranging 2-162 Kb with 16% harbouring multiple plasmids. No plasmid type was common to all *Campylobacter* [57].

Although plasmids may be involved in virulence there are a large number of clinical isolates that possess no plasmids so chromosomal determinants must also be of importance in virulence. Plasmids have also been implicated in the spread of antibiotic resistance. Tetracycline resistance is largely associated with plasmids [57] but other resistances may be chromosomally mediated [58]. Two *C. jejuni* plasmids both containing type IV secretion systems have been identified in strain 81-176; pVir and pTet [55]. These plasmids will be discussed further in chapter 3.

1.1.4.9 Chicken colonization

The colonization of chickens by *Campylobacter* is thought to be an important consideration in disease causation by the organism as consumption of poultry products has been implicated in transmission of the bacteria to humans, although the source of most infections remains unidentified [20]. Levels of chicken colonization are high with up to 10^9 *Campylobacter* being recovered from a single chicken [59]. Different strains of *Campylobacter* are known to differ in their ability to colonize chickens [60].

1.1.4.10 Pathogenesis Summary

Phenotypic differences have been observed for traits implicated in virulence such as adherence, invasive properties, toxin production, serum resistance, chicken colonization potential, aerotolerance and temperature tolerance [29]. These phenotypic differences may correlate with differences in clinical outcome of disease, survival of the bacterium in the environment and transmission of the bacterium between hosts. Phenotypic difference may reflect underlying genotypic diversity.

1.1.5 Subtyping and diversity

Subtyping methods have been developed for *Campylobacter* and are frequently used in surveillance and epidemiological studies. In order to trace sources of infection, discrimination between different strains is necessary.

1.1.5.1 Serotyping

Two well established subtyping techniques are serotyping schemes. The Lior scheme is based on heat-labile antigens and a bacterial agglutination method [61] which now recognizes over 100 serotypes of *C. jejuni* [29]. The Penner scheme is based on heat-stable antigens using a passive hemagglutination technique [62]. The Penner scheme has recently been shown to be based on capsular polysaccharide [53] and currently identifies more than 60 serotypes of *C. jejuni* [29]. Phage typing has been used to give finer discrimination, and there are currently 336 serotype-phage type combinations. However there are a number of strains which remain untypable using traditional methods (19% of human isolates) [19]. Serotyping has been used to monitor *Campylobacter* on contaminated foodstuffs. A survey of poultry food products in Denmark found that 85% of *Campylobacter* isolates were *C. jejuni* with certain Penner serotypes being more common than others [63].

Serotyping techniques are labour intensive and as such are largely limited to reference laboratories [34]. Molecular techniques already in use for other bacteria have been adapted for *Campylobacter* in order to provide more accessible typing procedures [64].

1.1.5.2 Molecular subtyping

Flagellin typing (*fla* typing) utilises restriction fragment length polymorphism (RFLP) in the PCR-amplified flagellin locus [29]. Due to variations in procedure, results from different laboratories cannot be directly compared which limits the usefulness of this assay for tracking infections [64]. In addition, due to recombination in the flagellin locus, the long term applicability of this assay for subtyping has been called into question as it does not accurately represent the entire genome [65]. However, *fla* typing has been used to demonstrate not only diversity between environmental isolates of *C. jejuni* but also to link certain *fla* types from environmental isolates to those from cases of human campylobacteriosis in the same geographical area [66].

Pulsed-field gel electrophoresis (PFGE) also shows diversity between isolates of *C. jejuni* [29;66]. The method is based on digestion of the bacterial chromosome by restriction enzymes that cleave the DNA infrequently. A major problem with this is that differences in electrophoretic conditions can lead to apparent differences in the profiles obtained even for the same DNA preparation which may make the comparison of different PFGE patterns unreliable [64].

Ribotyping involves gel electrophoresis of digested genomic DNA followed by Southern blot hybridization with a probe specific for rRNA genes. There is limited discriminatory power for this typing method due to there being only 3 rRNA gene copies present in the genomes of *Campylobacter* species which means isolates can not be reliably identified at the subspecies level [64].

Randomly amplified polymorphic DNA (RAPD) analysis uses short non-specific primers to arbitrarily amplify DNA products under low-stringency PCR conditions. Band patterns consist of both weak and strong amplicons which can complicate interpretation. In addition up to 14% of strains examined may be untypeable due to DNase activity and there is poor reproducibility [64]. RAPD analysis has been used to distinguish between invasive and non-invasive isolates based on band differences; a distinct RAPD profile was found in 63% of invasive strains but was also found in 16% of non-invasive strains [67]. RAPD analysis has also shown genetic diversity between *C. jejuni* isolates from human faeces, seawater and poultry products [68].

Amplified fragment length polymorphism (AFLP) analysis involves the complete digestion of chromosomal DNA with two restriction enzymes, one with a 4 bp recognition site and the other with a 6 bp recognition site, followed by PCR amplification based on restriction sites [64]. The bacterial subtypes recognized by one technique often do not correlate with the subtypes determined by other typing techniques [29]. However AFLP analysis and multilocus sequence typing (MLST) have been shown to give similar genetic groupings when performed on the same isolates [69]. In one study AFLP analysis identified more than 100 different profiles amongst human, chicken and cattle isolates of *C. jejuni* and showed that isolates from human and cattle were more likely to show similar banding patterns than those from chickens [69].

A major limiting factor in some subtyping schemes is the reproducibility and comparison of results between laboratories as well as the fact that a number of strains are untypable using certain techniques. In order to provide a standardised test that is easy to perform and compare between laboratories, an MLST scheme for *Campylobacter* has been set up based on seven housekeeping loci [70]. The results of initial typing using 194 strains indicate that *C. jejuni* is genetically diverse with a weakly clonal population structure. Using this technique 155 sequence types (STs) were observed with 26% being unique. Some STs were consistent with Penner serotype however some displayed a high level of diversity within serotypes [70].

1.1.5.3 Subtyping Summary

These subtyping techniques demonstrate a wide range of phenotypic and genotypic diversity between *Campylobacter* isolates in different environmental and clinical settings, although it is difficult to compare between techniques and even between laboratories. It is difficult, based on these techniques, to know the full extent of genotypic diversity within the species *C. jejuni.*

1.2 Genomic studies

In some bacteria an increase in observed pathogenicity has been attributed to the uptake and incorporation of virulence genes which in some cases cluster in regions known as pathogenicity islands [71], for example the *cagA* pathogenicity island of *Helicobacter pylori* [72]. The *cag* pathogenicity island is a 40 Kb island flanked by direct repeats; this island encodes 31 CDSs including the CagA cytotoxin and a type IV secretion system [73]. Pathogenicity islands are characterized by different G+C content to the core chromosome G+C content, instability, integration at specific loci e.g. tRNA genes, presence of mobility elements and the presence of direct repeats [71]. Pathogenic *Yersina* can be divided into low-pathogenicity strains, which induce mild intestinal infection in humans, and high-pathogenicity strains, which induce severe systemic infection in humans [74]. Several genes responsible for the high-pathogenicity phenotype are clustered on a genomic island termed the high-pathogenicity island [75]. This high-pathogenicity island present in *Yersinia* spp. is particularly unstable and can be lost at frequencies of up to 10^{-5} per generation [76]. Virulence genes may also be located on transmissible genetic elements such as transposons [71], plasmids [77;78] or bacteriophages [79;80]. However, in *C. jejuni* the degree to which genetic differences contribute to variations in disease outcome and epidemiological characteristics is as yet unclear [29]. In the current genomics age large scale methods have been adopted to explore bacterial strain diversity.

1.2.1 Genome sequencing

1.2.1.1 *C. jejuni* strain NCTC 11168

C. jejuni strain NCTC 11168 was isolated from a case of human campylobacteriosis in 1977 and is a commonly used laboratory strain. In 2000 the sequence of *C. jejuni* strain NCTC 11168 was published [8]. The genome of strain NCTC 11168 is 1641481 bp with 94.3% predicted to code for proteins. Out of 1654 predicted Coding Sequences (CDSs) approximately 22% of *C. jejuni* genes had no matches to previously identified genes with known function. Only 55.4% of *C. jejuni* CDSs had orthologues in the closely related bacterium *Helicobacter pylori*. The majority of predicted CDSs did not appear to be organized into operons or clusters. Exceptions to this include the lipooligosaccharide (LOS) and capsular polysaccharide biosynthesis clusters. Interestingly these polysaccharide biosynthesis clusters have a lower G+C content than the rest of the chromosome. The sequence data was largely unable to elucidate novel candidate genes for the production of toxins, adhesins, invasins and other classical virulence determinants with the exception of components of sialylation pathways and the cytolethal distending toxin genes (*cdtABC*). There were also a lack of bacteriophage, inserted sequence (IS) elements and obvious pathogenicity islands [8].

One striking discovery from the genome sequence was the identification of 24 regions of sequence polymorphism which mainly consisted of poly G/C tracts which alter in length due to slipped-strand mispairing. Slipped-strand mispairing involves denaturation and displacement of the strands of DNA in a duplex followed by mispairing of complementary bases within a short repeat, e.g. a homopolymeric tract. When slipped-strand mispairing is followed by replication or repair this can lead to the insertion or deletion of one or more bases within the homopolymeric tract [81]. If the number of bases inserted or deleted is not a multiple of three this in turn can alter the expression of specific proteins by shifting their translational reading frame. Tract length variation resulting in translational frameshifting has been shown in other bacteria to be responsible for phase variation whereby bacteria randomly vary surface properties or antigenicity [82]. The phase-variable genes in *C. jejuni* predominantly cluster in the lipo-oligosaccharide, capsule and flagellar biosynthesis regions indicating that *C. jejuni* is also using this mechanism to alter surface properties.

1.2.1.2 *C. jejuni* strain RM1221

More recently the genome sequence of *C. jejuni* strain RM1221, which was isolated from the skin of a retail chicken, has been published [9] and at 1777831 bp was larger than that of strain NCTC 11168. Strain RM1221 was predicted to encode 1884 proteins which was 230 more than in strain NCTC 11168. The LOS and capsular polysaccharide biosynthesis loci were predicted to encode many different CDSs to that of strain NCTC 11168. The genomes of strain NCTC 11168 and strain RM1221 were shown to be syntenic but this synteny was disrupted by four genomic islands in strain RM1221. Three of the islands were phage derived with the fourth likely to be of plasmid origin [9]. Strain RM1221 also appeared to be devoid of functional IS elements.

1.2.2 Comparative genomic studies

1.2.2.1 Microarrays

Microarrays have been popular in recent years, with many bacterial species being studied, and have proved useful for comparing diversity with respect to sequenced strains. Dorrell *et al.* [51] have used a microarray approach to reveal extensive genetic diversity within *C. jejuni*: 21% of genes in strain NCTC 11168 were absent or highly divergent in one or more of the 11 *C. jejuni* strains tested. Genes for virulence determinants hypothesised to be necessary for *C. jejuni* to cause disease in humans including the cytolethal distending toxin, flagellar structural proteins, phospholipase A, the PEB antigenic surface proteins, and proteins potentially involved in host pathogen interactions such as CiaB, CadF and CheY were conserved in all the strains tested [51]. With regard to strain diversity this approach can only show genes that are missing or significantly divergent compared to strain NCTC 11168 and not replacements or insertions that may exist within the genome of these different strains. There have also been two other comparative papers using strain NCTC 11168 microarrays and various other strains [83;84]. Pearson *et al.* [83] showed that between 2.6% and 10.2% of the NCTC 11168 CDSs were absent or divergent in the 18 *C. jejuni* strains tested. Variable CDSs were located in seven plasticity regions on the genome of strain NCTC 11168 which included the LOS and capsular polysaccharide biosynthesis loci as well as flagellin biosynthesis and post translational modification loci. Taboada *et al.* [84] compared 51 *C. jejuni* strains to NCTC 11168 showing that 20% of NCTC 11168 CDSs were divergent in at least one of the test strains. Most of these variable genes were located in 16 plasticity regions including surface polysaccharide biosynthesis and modification loci and restriction modification loci as well as regions containing hypothetical CDSs.

C. jejuni strain ATCC 43431 has been analysed using a shotgun microarray to identify genes unique to this strain in comparison to strain NCTC 11168. This method identified 130 complete and incomplete CDSs [85]. Many LOS and capsule associated genes were discovered along with restriction modification genes, integrases and hypothetical genes. However, these CDSs were only fragments and were not expanded to give genomic context.

1.2.2.2 Subtractive hybridization

Subtractive hybridization has been used for some time as a method for identifying genes expressed in one cell type but not another by hybridizing cDNA to RNA [86;87]. The method has since been adapted for identifying DNA differences between bacterial strains [88;89]. Subtractive hybridization was evaluated as a method of comparing genomes using the two sequenced strains of *H. pylori* where it was shown to identify 95% of CDSs unique to one strain compared to the other [90].

Ahmed *et al.* [91] have used subtractive hybridization as a technique to identify gene fragments in strain 81116 that were not present in strain NCTC 11168. Strain 81116 is also a human campylobacteriosis isolate but has been proposed to show greater colonization potential of chickens than the strain NCTC 11168. In strain 81116, 24 fragments that were unique to this strain (less than 75% identity at the nucleotide level to NCTC 11168) were identified and used to hybridize to genomic DNA from 9 other strains: one insert was unique to 81116, one was present in all 9 tested strains and the rest showed variable distributions [91]. Gene fragments identified included those with similarity to restriction-modification enzymes, arsenic-resistance genes and cytochrome C oxidase III genes [91]. However these fragments were not characterized further to obtain entire genes and to assess their distribution across the genome. The method of subtractive hybridization has drawbacks, including limited coverage of the genome, and the production of only small fragments of novel DNA which must then be cloned after manipulation which can introduce biases.

1.2.2.3 Differential hybridization in other organisms

Others have addressed the problem of identifying novel sequences in one bacterial strain compared to another. Liang *et al.* used a differential hybridization approach to identify differences between *Pseudomonas aeruginosa* strain X24509 and the sequenced strain PA01 [92]. They developed a method of differential hybridization using arrayed libraries of cloned DNA fragments and found a genomic island (PAGI-1) that was present in 85% of pathogenic isolates. PAGI-1 was sequenced and CDSs within it predicted. Several CDSs showed sequence similarity to known genes including dehydrogenase genes, genes coding for proteins implicated in detoxification of reactive oxygen species and transcriptional regulators. The role of PAGI-1 in *Pseudomonas aeruginosa* is unknown but it may encode genes that have a role in protecting the cell against oxidative damage and the transcriptional regulators may control the expression of chromosomal genes providing a selective advantage for strains that have acquired PAGI-1. This study showed that differential hybridization is a valid approach for identifying virulence factors in conjunction with sequence data as Liang *et al*. were able to identify the insertion site of PAGI-1 [92]. This method has several advantages: i) there are no cloning steps after manipulation, ii) a breadth of coverage can be achieved by generating large libraries, and iii) the length of sequences studied can be modified by altering the insert sizes of the clone libraries.

1.3 Aims of this thesis

C. jejuni has been demonstrated to be both genotypically and phenotypically diverse. Differences in phenotype between strains can often be due to novel genes or islands present in one strain compared to another. These genotypic differences may relate to different clinical outcomes, epidemiological characteristics or environmental persistence. In order to explore this possibility:

i) Genomic DNA arrays of *C. jejuni* strains with different phenotypic characteristics will be created.

ii) DNA present in the strains to be tested, which is not present in the sequenced strain NCTC 11168, will be identified using a differential genomic DNA hybridization approach with small-insert libraries.

iii) Strain specific DNA will be characterized by sequencing and annotation in order to identify potential virulence or survival factors.

iv) The extent and context of novel regions containing these factors will be determined relative to the chromosome of strain NCTC 11168 by sequencing larger-insert libraries.

2. Materials and Methods

2.1 Materials

2.1.1 Bacterial Strains and plasmids

Table 2.1: *C. jejuni* strains and plasmids used in this study

Strain	characteristics	Penner serotype	Reference/Source
NCTC 11168	Sequenced strain, human isolate (Worcester 1977), source: unknown	0:2	NCTC 11168 [A] [3;8]
81-176	Human isolate (Minnesota 1981), source: raw milk	0:23/36	Black *et al.* 1988 [B] [36]
81-176 plasmids	pTet and pVir	-	Bacon *et al.* 2000 [55]
M1	Human isolate (UK 2000), source: poultry	0:9	Unpublished/ VLA, Weybridge[C]
40671	Human outbreak isolate (UK 2000), source: water source suspected, unproven by epidemiology	0:50, phage type 6	Champion [D] [93]
52472	Blood invasive human isolate, source: unknown	untypable, phage type 1	Unpublished/ PHLS, Colindale[D]

[A] National Collection of Type Cultures, Colindale, London, UK

[B] Vanderbilt University, Nashville, Tennessee, USA

[C] Veterinary Laboratories Agency (VLA), Weybridge, UK

[D] Public Health Laboratory Service (PHLS), Colindale, London, UK

Purified genomic DNA from all strains and plasmids was provided by Brendan Wren's group at the London School of Hygiene and Tropical Medicine (LSHTM, London, UK).

2.1.2 Reagents

Reagents used in this study were purchased from Fisher Scientific (Loughborough, Leicestershire, UK), BDH Laboratory supplies (VWR International, Dorset, UK) or Sigma (Dorset, UK) unless otherwise stated. Restriction endonucleases were purchased from New England Biolabs (NEB) (Hitchin, Hertfordshire, U.K.).

2.2 General Methods

2.2.1 Growth of *Escherichia coli* clones

Escherichia coli clones were grown for 18-22 hrs at 37°C, agitating at 320 rpm in 2x Luria-Bertani broth (2LB; 20 mg/ml tryptone, 10 mg/ml yeast extract and 10 mg/ml NaCl) containing 0.1 mg/ml Ampicillin for pUC clones, 12.5 μg/ml Chloramphenicol for pBAC clones or other antibiotics as appropriate.

2.2.2 Preparation of DNA

2.2.2.1 Isopropanol preparation for isolation of pUC plasmid DNA

pUC clones were grown (section 2.2.1) in a volume of 1 ml of 2LB per well in 96-well boxes (Beckman). Boxes were spun in a centrifuge (Eppendorf 5810R) at 4000 rpm to pellet the cells. Culture supernatant was decanted and cells were resuspended in 120 μl Glucose-Tris-EDTA (GTE; 20% glucose, 1 M Tris-HCl, pH8.0, 0.1 M EDTA) plus 60 μg/ml RNase A (Q-Biogene, Cambridge, UK). After resuspension 120 μl 0.2 N NaOH/ 1% SDS was added followed by 120 μl 3 M potassium acetate. In order to remove cell debris and precipitate plasmid DNA 140 μl of cell lysate was pipetted into a filter plate (Costar 3504) which was placed above a storage plate (Costar 3365 serocluster) containing 140 μl of isopropanol, and both plates were spun together for 15 mins at 4000 rpm and 4˚C in a centrifuge. The supernatant was discarded and pellets washed with 100 μl 70% ethanol by spinning for 5 mins at 4000 rpm and 4°C. Pellets were dried then resuspended in 60 µl of autoclaved double distilled water (DDW).

2.2.2.2 Vacuum preparation for isolation of BAC DNA

Clones were grown (section 2.2.1) in a volume of 1.5 ml per well in 96-well boxes (Beckman). The boxes were spun in a centrifuge (Eppendorf 5810R) for 3 mins at 4000 rpm to pellet the cells and the culture supernatant was discarded. Cells were resuspended in 100 µl GTE containing 0.1 mg/ml RNaseA (Q-Biogene) on a box vortexer (Luckham) set on speed 8 for 3 mins. After resuspension, 100 µl 0.2 N NaOH/ 1% SDS was added and cell suspensions were mixed again on a box vortexer for 1 min, incubated at room temperature for 2 mins then 100 µl 3 M potassium acetate was added and the cell suspensions mixed using a box vortexer for 2 mins. The cell lysate was then transferred into a filter plate with pore size 0.65 µM (MADVN6550 Millipore) on a vacuum manifold with a second filter plate (MANUBAC50 Millipore) underneath and a vacuum of 10-15 mmHg was applied until the cell lysate had passed through the top filter plate. The bottom filter plate (MANUBAC50) was then transferred to the top of the vacuum manifold and a vacuum of 20-25 mmHg was applied until the contents had passed through the filter plate into a waste receptacle. To wash the DNA, 200 µl of DDW was added to the filter plate and the vacuum of 20-25 mmHg reapplied until the plate was dry. The filter plate was then removed from the vacuum manifold and 35 µl 10 mM Tris-HCl pH8 was added to neutralize and resuspend the DNA. Filter plates were vortexed on box vortexer (Luckham) speed 5 for 10 mins to aid resuspension. The DNA was then pipetted from the filter plate into a 96-well plate for storage (costar serocluster).

2.2.2.3 Low-throughput preparation of DNA

Clones were grown (section 2.2.1) in a volume of 6 ml. Overnight cultures were spun in a centrifuge (Eppendorf 5810R) in 50 ml tubes (Falcon) for 15 mins at 3000 rpm and 4°C. The culture supernatant was discarded and cells pellets resuspended in 200 µl GTE containing 0.1 mg/ml RNaseA (Q-Biogene). The cell mixture was then pipetted into 1.5 ml tubes (eppendorf) containing 400 µl 0.2 N NaOH/ 1% SDS and incubated at room temperature for 5 mins, 300 µl 3 M Potassium acetate was then added and the tubes were spun in a centrifuge (Eppendorf 5415D) for 15 mins at 13000 rpm. To precipitate the DNA, 750 µl of cell lysate was transferred into a fresh 1.5 ml tube (eppendorf) containing 450 µl of isopropanol and the tubes were spun for 15 mins at 13000 rpm in a centrifuge. The supernatant was removed and the pellets were washed with 1 ml 70% ethanol. The pellets were dried and then resuspended in 20-60 µl TE (10 mM Tris: 0.1 mM EDTA) as appropriate to the pellet size.

2.2.3 Gel electrophoresis

DNA fragments were separated by agarose gel electrophoresis using 0.5% or 0.8% w/v agarose or low melting point (LMP) agarose (Invitrogen, Paisley, UK) in either Tris-acetate-EDTA (TAE) or Tris-Borate-EDTA (TBE) buffer [94]. Samples were loaded in ficoll loading dye (1mg/ml Bromophenol blue, 0.1mg/ml Ficoll 400, 1xTBE v/v) diluted ¼ with running buffer .

DNA was visualized by adding 10 µg/ml ethidium bromide to the running buffer and leaving to stain for 10-30 mins before viewing under ultraviolet. Alternatively, DNA was visualized by adding VistraGreen (Amersham, Buckinghamshire, UK) according to the manufacturers instructions, in a volume of DDW sufficient to cover the gel, and staining for 30 mins before viewing on a Dark Reader (Clare Chemical research, www.clarechemical.com). Fragment size was determined by comparison with appropriate DNA ladders including a λ *Hin*dIII digest (NEB) and pBR322 *Bst*NI digest (NEB) mix, 1 Kb ladder (Invitrogen), 100 bp ladder (Invitrogen) or Raoul™ (Q-Biogene).

2.2.4 Polymerase Chain Reaction

All Polymerase Chain Reaction (PCR) amplifications including sequencing were carried out on a Peltier Thermal Cycler (PTC-225 MJ Research, Bio-Rad, Hertfordshire, UK) except for reactions incorporating radio-labelled nucleotides (see section 2.3.5). Oligonucleotide primers were all synthesized in-house.

PCR amplification was performed using *Taq* DNA polymerase (Applied Biosystems, Foster City, CA, USA). Reactions contained 10-100 ng of DNA template, 50 pmoles of each primer (forward and reverse) and dNTPs (Amersham) at a final concentration of 0.25 mM each in a total reaction volume of 50 µl. PCR conditions consisted of 94°C for 30 s to denature the template, followed by 30 cycles of 92°C for 30 s, a variety of temperatures (based on individual primer Tm) for 30 s and 72°C for 1-4 min (depending on expected length of product), unless otherwise stated.

2.2.5 Restriction enzyme digests

Restriction endonuclease digestion was performed over a period of 3-4 hrs at the specified optimum temperature (usually 37°C) using 1 unit of enzyme per µg of DNA unless otherwise specified. Double digests were performed using buffers compatible with both restriction enzymes according to the manufacturer's protocol. Complete digestion was verified using agarose gel electrophoresis (section 2.2.3).

2.2.6 DNA manipulation

2.2.6.1 Gel extraction

2.2.6.1.1 Spin column kit

Fragments were excised from the agarose gel, and purified using MinElute™ gel extraction kit (Qiagen, Sussex, UK) according to the manufacturer's instructions.

2.2.6.1.2 LMP agarace digestion

LMP agarose gel slices were melted in 1.5 ml tubes (eppendorf) in a waterbath at 65°C for 5 mins. The tubes were then transferred to a waterbath at 42°C and allowed to equilibrate before 5 µl of AgarAce® (Promega, Southampton, UK) was added for every 200 µl gel volume. Samples were incubated at 42°C for at least 20 mins. DNA was then extracted by phenol extraction (section 2.2.6.2) and purified by ethanol precipitation (2.2.6.3).

2.2.6.2 Phenol extraction

An equal volume of TE-buffered phenol (Sigma) was added to samples and agitated using a vortex genie (Scientific Industries, New York, USA) for 1 min. Samples were incubated on ice for 5 mins then spun in a centrifuge (Eppendorf 5415D) for 3 mins at 13000 rpm. The aqueous layer was pipetted into a 0.5 ml tube (eppendorf) and incubated on ice for 5 mins before being spun in a centrifuge for 3 mins at 13000 rpm. The aqueous layer was then transferred into a 1.5 ml tube (eppendorf).

2.2.6.3 Ethanol precipitation

DNA was precipitated by adding 1/10th the sample volume of 1 M NaCl, 2.5 volumes of 70% ethanol; 1 µl pellet paint (Novagen, Darmstadt, Germany) was used to aid pellet visualization. Samples were incubated either at -70°C for 1 hr or -20°C overnight before being spun in a centrifuge (eppendorf 5417R) for 30 mins at 14000 rpm and 4°C. Pellets were washed with 1 ml 70% ethanol before centrifugation for 5 mins at 1400 rpm and 4°C, and then dried before resuspending as appropriate.

2.3 Differential hybridization methods

2.3.1 Construction of a pUC19 library of DNA fragments

2.3.1.1 Preparation of DNA

Approximately 10 µg of chromosomal DNA was sheared using a sonicator (xl2020 sonicator, Heat systems Inc., New York, USA) in a final volume of 60 µl to create fragments between 12 kb and 500 bp. The ends of sonicated DNA were repaired using 0.3 µl mung bean nuclease (256 U/µl, Amersham Pharmacia Biotech, Piscataway, NJ, USA) incubated at 30˚C for 10 mins in the presence of mung bean buffer (15 mM Sodium acetate, 25 mM NaCl, 0.5 mM $ZnCl_{2,}$ 2.5% glycerol). The DNA was then ethanol precipitated (section 2.2.6.3) and size fractionated by LMP agarose gel electrophoresis (section 2.2.3). Fragments of appropriate size were recovered and purified (section 2.2.6.1.2).

2.3.1.2 Ligation

Purified DNA fragments were ligated to pUC19 vector in the following reaction mix: 3 µl DNA solution, 0.3 µl pUC19 *Sma*I-Bacterial Alkaline Phosphatase (BAP) (Q-Biogene 40 ng/µl), 0.4 µl ligase buffer, 0.3 µl T4 DNA ligase (5 U/µl Roche, Lewes, East Sussex, UK). The ligation mixture was incubated at 12-14˚C overnight. Ligation was terminated using 1 µl Proteinase K (Roche 14 mg/ml) in a final volume of 50 µl and incubated at 50˚C for 1 hr.

2.3.1.3 Transformation

Transformation was performed using 0.5 µl of ligation mixture, 40 µl electrocompetent *Escherichia coli* DH10B (Invitrogen) and an electroporation device (BioRad, genepulser) set at 1.7 Kv, 200 Ω, 25 μF. Cells were allowed to recover in 500 μl SOC [95] at 37°C for 1 hr and plated on Tryptone Yeast Extract (TYE) plates (15 mg/ml Agar; 8 mg/ml NaCl; 10 mg/ml Bacto Tryptone; 5 mg/ml yeast extract) containing 0.1 mg/ml ampicillin, with 2.5 mg 5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-gal) and 2 mg isopropylthio-β-D-galactoside (IPTG). White colonies were picked after overnight incubation at 37˚C.

2.3.2 Construction of a pBACe3.6 library of DNA fragments

2.3.2.1 Preparation of pBACe3.6 vector

The vector pBACe3.6 [96] was supplied in *E. coli* DH10B cells by Pieter de Jong from the Children's Hospital Oakland Research Institute, USA (http://bacpac.chori.org). The vector pBACe3.6 was prepared for cloning as described in 'Current Protocols in Human Genetics' [97].

Briefly, a caesium chloride gradient was used to remove and purify the vector DNA from the host cell DNA, after total DNA extraction. The pUC stuffer fragment was removed by restriction endonuclease digestion and the resultant "sticky ends" dephosphorylated to prevent vector recircularization. Control ligations with a 16 kb fragment of lambda DNA were performed to check the quality of the vector.

2.3.2.2 Preparation of DNA

DNA fragments of chromosomal DNA were prepared by limited digestion with an optimised dilution of *Sau*3A1 in 10x bovine serum albumin (BSA) for an optimized length of time (determined by previous trial digestions) at 37˚C in a 200 μl volume, to give DNA fragments of between 10-40 Kb as appropriate. The enzyme was inactivated and DNA extracted using phenol (section 2.2.6.2) followed by ethanol precipitation (section 2.2.6.3). DNA fragments were separated using a 0.4% low melting point (LMP) agarose gel (section 2.2.3). The appropriate size fractions were recovered from the gel and purified (section 2.2.6.1.2).

2.3.2.3 Ligation

Approximately 20 ng of the purified DNA size fraction was used for ligation in a 50 µl reaction volume containing 10 ng pBACe3.6 pre-cut with *Bam*HI (section 2.3.2.1), 9 µl 30% polyethylene glycol (PEG) 8000, 1 µl 0.1 M $MgCl_2$, 5 µl ligation buffer and 1 µl 1/10 ligase (diluted in 10x BSA). The ligations were incubated overnight at 16˚C and terminated by adding 2.5 µl of 0.5 M EDTA and 1 µl of 14 mg/ml Proteinase K (Roche), incubated for 1 hr at 37˚C then 1 µl of 100 mM phenylmethylsulfonyl fluoride (PMSF) (sigma) was added. The ligation mixture was then dialysed on a 0.025 µm pore size microdialysis filter (MF-Millipore VSWP, Millipore UK Ltd, Watford, UK) floated on 0.5 x TE v/v in a petri dish and incubated at 4˚C for 3 hrs. The ligation mix was then pipetted from the filter.

2.3.2.4 Transformation

Transformation was performed using 1 µl of each ligation, 20 µl electrocompetent *E. coli* DH10B cells (Invitrogen) and a CellPorator (Life Technologies, Paisley, UK) equipped with a voltage booster set at 4 kΩ, 330 µF, 13 kV/cm, fast charge. Cells were allowed to recover in 500 µl SOC [95] by incubation at 37°C for 1 hr before being plated onto TYE plates containing 20 µg/ml chloramphenicol and 5% sucrose, and incubated overnight at 37˚C. Colonies were picked the following day.

2.3.3 Propagation of library clones

Clones were picked either manually or using the Sanger Institute automated picking facility into media (section 2.2.1) plus 7.5% glycerol for storage. All libraries were routinely tested for phage and *Pseudomonas* contamination by spotting colonies onto agar plates seeded with a DH10B lawn or onto *Pseudomonas* selective agar with C-N supplement (Oxoid, Basingstoke, Hampshire, UK) according to manufacturer's instructions.

2.3.4 Preparation of colony arrays

2.3.4.1 Colony blotting

The clones were arrayed onto 78 x 119mm Nytran N membrane (Schleicher and Schuell, Dassel, Germany), supported on agar plates, using the Sanger Institute automated robotic arraying facilities. pUC plasmid clones were arrayed in duplicate in a 384-pin 4x4 gridding pattern resulting in 6144 clones per filter. BAC clones were arrayed in duplicate in a 96-pin 4x4 gridding pattern resulting in 1536 clones per filter. After colony blotting, the agar plates plus membranes were incubated for 16-18 hrs at 37°C.

2.3.4.2 Lysis of bacterial clones

The membranes were placed colony side up on chromatography paper soaked in 10% SDS for 5 mins. Membranes were then transferred to chromatography paper soaked in denaturing solution (0.5 N NaOH/ 1.5 M NaCl) for 10 mins then allowed to dry for 10-20 mins on dry chromatography paper. After drying the membranes were briefly submerged in neutralizing solution (0.5 M Tris-HCl pH 7.4/ 1.5 M NaCl), followed by a 5 min wash with fresh neutralizing solution on an orbital shaker. A further 5 min wash in neutralizing solution was performed, followed by a 5 min wash in 1/10 neutralizing solution. The membranes were then washed with agitation in 2X SSC/ 0.1% SDS for 5 mins, 2X SSC for 5 mins then washed twice with 50 mM Tris-Hcl pH7.4 for 5 mins. Membranes were air dried DNA side up on chromatography paper for a minimum of 6 hours before UV cross-linking for 2 mins on a transilluminator (254 nm).

2.3.5 Hybridization of membranes

2.3.5.1 Radiolabelled probe generation using random octamers

Random octamer primers were annealed to denatured DNA templates and extended by the Klenow fragment of DNA polymerase I, incorporating one radiolabelled nucleotide and three unlabelled nucleotides, to form a probe [98]. This labelling reaction was carried out using 100 ng of sonicated DNA and components of the BioPrime DNA labelling system (Invitrogen) according to manufacturer's instructions, except for using an in-house dNTP mix (0.6 mM of A, G, T) and 30 µCi [α-^{32}P]dCTP (Amersham). Labelling reactions were incubated at 37˚C for 1 hr then 5 µl of stop buffer (0.5 M EDTA) was added and the probes were purified using a microspin G25 column (Amersham). An equal volume of sonicated human placental DNA (Sigma) was added to each probe, then the mixture was denatured at 99˚C for 5 mins.

2.3.5.2 Generation of radiolabelled probe using PCR

Primer pairs complementary to pUC clone insert sequences were used to amplify regions of DNA from pUC clones (section 2.2.4). Products were checked for size using agarose gel electrophoresis (section 2.2.3) then purified through spin columns (microspin S-400HR, Amersham) according to the manufacturer's instructions. These purified templates were then added to 1 µl of 10x PCR buffer (50 mM KCl, 5 mM Tris pH8.5, 2.5 mM $MgCl_2$), 0.4 µl of 2.5 mM dATP, dTTP and dGTP, 0.5 units *Taq* polymerase and 4 µCi [α-^{32}P]dCTP in a 10 µl reaction volume. PCR amplification was performed on a thermocycler (Perkin and Elmer, Boston, MA, USA) at 96°C for 30 s to denature the template, followed by 30 cycles of 92°C for 30 s, 53°C for 30 s and 72°C for 2 mins. Probes were purified using a microspin G25 column (Amersham) then denatured at 99˚C for 5 mins.

2.3.5.3 Hybridization

Membrane sets were soaked in 2x SSC (1x SSC is 15 mM sodium citrate and 0.15 M NaCl) prior to pre-hybridization. Membrane sets were pre-hybridized in glass tubes, in a hybridization oven at 65˚C in 15 ml Church buffer (1 mM EDTA, 0.5 M $NaHPO_4$ (pH7.2), 7% Sodium Dodecyl Sulphate (SDS) 1% BSA) for 1-3 hrs. Pre-hybridization buffer was decanted and probe was added to the membranes in a fresh volume of Church buffer and incubated at 65°C overnight.

2.3.5.4 Washes

The day following hybridization (section 2.3.4.3) the membranes were washed twice at room temperature with 2x SSC/ 0.1% SDS for 20 mins, then twice at 65˚C with 0.1x SSC/ 0.1% SDS for 20 mins. The membranes were then sealed in Saran wrap, placed in an autoradiograph cassette and exposed to photographic film. It was found that a rinse step using 2x SSC/ 0.1% SDS conducted in the hybridization tubes improved the quality of results for differential hybridization reactions.

2.3.6 Sequencing of library clones

2.3.6.1 Sequencing primers

5` - 3`

M13F: TGTAAAACGACGGCCAGT

pUC18R: GCGGATAACAATTTCACACAGGA

T7: TAATACGACTCACTATAGGG

Sp6: ATTTAGGTGACACTATAG

2.3.6.2 Sequencing reactions

All sequencing was performed using ABI Prism® BigDye® terminator chemistry (Applied Biosystems) and loaded on ABI 3700 capillary sequencing machines at the Sanger Institute sequencing facility according to their protocols.

2.3.6.2.1 pUC clone sequencing

Sequencing reactions were conducted using 3 µl of DNA (20 ng/ul), 0.25 µl BigDye v3.1, 2.5 µl BigDye buffer (400 mM Tris pH9, 10 mM $MgCl_2$), and primer (either M13F or pUC18R) to a final concentration of 3 pM in a total reaction volume of 9 µl. DNA was amplified by thermocycling with the following conditions:-

96˚C 30 s

$$\left.\begin{array}{l}92^{\circ}\text{C } 4\text{ s}\\ 50^{\circ}\text{C } 4\text{ s}\\ 60^{\circ}\text{C } 1\text{ min } 50\text{ s}\end{array}\right\} \text{ 44 cycles}$$

10˚C holding temperature

After thermocycling DNA was precipitated by adding 25 µl of precipitation mix (60 mM sodium acetate, 4 µM EDTA in 96% ethanol) and spinning in a centrifuge for 20 mins at 4000 rpm and 4°C. Pellets were then washed with 30 µl 70% ethanol before spinning in a centrifuge for 5 mins at 4000 rpm and 4°C and dried before being loaded on sequencing machines.

2.3.6.2.2 Sequencing from primers internal to cloning vector

Sequencing reactions were carried out as in section 2.3.6.2.1 except primers designed from the insert sequence were used at a final concentration of 30 pM. Either plasmid or PCR products purified using spin columns (Microspin S-400HR, Amersham) were used as sequencing templates.

2.3.6.2.3 BAC end sequencing

Sequencing reactions were conducted using 10 µl DNA (650-800 ng), 3 µl of BigDye v3.1, 3 µl BigDye buffer (400 mM Tris pH9, 10 mM $MgCl_2$) and 30 pM of primer (either T7 or Sp6) in a reaction volume of 20 µl.

DNA was amplified by thermocycling with the following conditions:-

95˚C 5 mins

95˚C 30 s
51˚C 10 s 75 cycles
60˚C 4 mins

10˚C holding temperature

After thermocycling DNA was precipitated by adding 5 µl 3 M Sodium acetate and 125 µl 96% ethanol, the plates were spun in a centrifuge for 1 hr at 4000 rpm and 4°C. The pellets were washed with 100 µl 70% ethanol and spun in a centrifuge for 15 mins at 4000 rpm and 4°C and then dried before being loaded on sequencing machines.

2.3.7 Analysis of sequence

The trace files were processed using Asp (http://www.sanger.ac.uk/software/sequencing/docs/asp/) and basecalled using Phred [99]. The individual read sequences were compared to the query sequence using WUBLASTN from the Washington University Basic Local Alignment Sequence Tool algorithms (WU-BLAST; http://blast.wustl.edu [100]). MSPcrunch [101] was then used to map query sequences back to the relevant subject sequence. Reads to be further analysed were assembled using Phrap (Green, P., unpublished) into contiguous sequences. Coding sequences (CDSs) were predicted within Artemis [102] and the translated protein sequences were compared to a non-redundant protein database using WUBLASTP [100] and FASTA [103]. Predicted proteins were compared against the Pfam database of protein domain

Hidden Markov models (http://www.sanger.ac.uk/software/Pfam/). The protein sequences were also searched for signal peptides (http://www.cbs.dtu.dk/services/signalP-2.0/), transmembrane helices (http://www.cbs.dtu.dk/services/TMHMM-2.0) and prosite motifs (http://www.expasy.ch/prosite/). Clustal X was used for protein and DNA alignments [104]. Shading was supplied by Boxshade server (http://www.ch.embnet.org/software/BOX_form.html). NJplot [105] was used to visualize guide trees produced by clustal X. EMBOSS applications (http://emboss.sourceforge.net/apps/) "needle" for Needleman-Wunsch [106] global alignments and "water" for Smith-Waterman [107] local alignments were used.

3. Plasmid sequencing and annotation

3.1 Introduction

Bacon *et al.* [55] identified two plasmids in *C. jejuni* strain 81-176 and partially characterized them. One plasmid was implicated in virulence (pVir) where mutation of the genes *comB3* and *virB11*, type IV secretion system homologues, were both found separately to reduce adherence to and invasion of a host cell line. The other plasmid was implicated in tetracycline resistance (pTet) [55]. As at the time of starting this project the plasmid sequences were unavailable, it was decided to sequence these plasmids in order to be able to differentiate which unique 81-176 genes identified in the hybridization experiments (chapter 4) were present on the chromosome and which were present on the plasmids.

3.2 Results

3.2.1 Overview of methods

The plasmids were sequenced using a shotgun strategy: purified 81-176 plasmid DNA was used to construct a library of plasmid DNA fragments of 2-4 Kb in pUC19 (section 2.3.1), plasmid DNA containing cloned fragments was prepared from the *E. coli* host strain (section 2.2.2.1) and enough inserts sequenced using forward and reverse primers to provide 10-fold coverage of the original *Campylobacter* plasmids (1413 reads) (section 2.3.6.2.1). The sequence reads were assembled using Phrap (Green, P., unpublished) into 4 large contiguous regions of 30788 bp, 23049 bp, 10671 bp and 10372 bp. Unfortunately no pUC clones were found containing end-sequences in more than one of these contiguous regions that would bridge the gaps. In order to join these sequences primers were designed to be complementary to the ends of each contiguous region and used in PCR reactions in each possible combination (section 2.2.4). Successfully amplified products were purified using columns (section 2.3.5.2) and then sequenced using the PCR primers in order to close the gaps (section 2.3.6.2.2). The final consensus sequences for each 81-176 plasmid were constructed in all places from at least four reads, with reads in both forward and reverse directions. The sequences were annotated using Artemis [102] to predict coding sequences (CDSs) and FASTA [103] to search protein databases (section 2.3.7).

3.2.2 Identification of replication origins

The circular plasmid genomes were broken, for the purpose of annotation, and each predicted CDS was assigned an identity number from this point. CDSs on pVir were called pVir1-54; CDSs on pTet were called pTet1-50. A sensible place to split a circular sequence would be at the origin of replication. In order to identify the predicted origin several features found at replication origins were searched for. There are several mechanisms for plasmid replication in bacteria including strand displacement, rolling circle and theta replication. A feature of replication origins common to all methods of replication is the presence of directly repeated sequences to which replication proteins bind. In addition there may also be an adjacent A+T rich region containing repeats and one or more *dnaA* boxes where the host DnaA initiator protein binds [108]. The origin of replication is sometimes found adjacent to the gene encoding the replication initiation protein [109]. It is unclear how many repeats are necessary for replication initiation, as an origin from the chromosome of *Coxiella burnetii* was characterized which contains only two *dnaA* boxes and three A+T rich 21-mers [110]. Many characteristic features associated with replication, e.g. *dnaA* boxes, are not found at all replication origins [111].

In pVir the proposed origin has an A+T content of 83% directly preceding a highly repetitive region. This repetitive region is followed by a predicted protein (pVir54c) with similarity to a replication protein (RepA) from *Erysipelothrix rhusiopathiae*. This probably represents the origin of replication for this plasmid.

In pTet there is a potential replication protein (pTet1) which has similarity to a replication protein from the plasmid ps23 of *Selenomonas ruminantium*. There is no high A+T region preceding this putative replication protein but there are two 40 bp repeats located just upstream. It was decided to split the sequence before the putative replication protein as there was no strong candidate for an origin of replication in this plasmid. However, there is a 273 bp region between pTet19 and pTet20 which contains a small cluster of three short repeats 20-30 bp in length as visualized using dotter [112]. This region has an A+T content of 81% and it is possible that this represents an origin of replication, as the replication protein and origin need not be located in the same place [113].

3.2.3 General characteristics

The plasmid pVir was found to be 37473 bp with a G+C content of 26%. A total of 54 CDSs were predicted covering 86% of the plasmid sequence. The plasmid pTet was found to be larger than pVir at 45204 bp and has a G+C content of 29% which is closer to that of the *C. jejuni* genome (strain NCTC 11168 having 30.6% and strain RM1221 30.3%). In total 50 CDSs were predicted in pTet covering 92% of the plasmid sequence; this is fewer than predicted for pVir as the average length of predicted CDSs from pVir is 597 bp compared to 835 bp in pTet.

There are predicted CDSs with similarity to DNA replication and plasmid conjugation proteins on both plasmids as well as many CDSs with similarity to hypothetical proteins from other bacteria and CDSs with no detectable similarity to proteins from other organisms (Appendix 1- pVir and Appendix 2 – pTet).

A circular representation of each plasmid is shown in **Fig 3.1.**

3.2.3.1 Characteristics of pVir

The plasmid pVir contains 37 predicted CDSs which show no detectable homology to proteins from other bacteria. This represents 69% of the total predicted CDSs for this plasmid. There are also 4 predicted CDSs which show similarity to hypothetical proteins from *Helicobacter pylori*. Other CDSs with similarity to proteins from *Helicobacter pylori* include a putative topoisomerase (pVir38), involved in DNA replication, and a putative partition gene (pVir52), involved in segregating low copy number plasmids into daughter cells of the host bacterium [114]. In addition to the topoisomerase and *parA* homologues there are other predicted CDSs that show homology to genes involved in plasmid maintenance e.g. single-stranded binding protein (pVir40) and *repA* (pVir54c). The plasmid is also predicted to encode type IV secretion system homologues. The predicted CDSs pVir26, pVir27, pVir28, pVir29, pVir30, pVir33 correspond to *virB4*, *virB8*, *virB9*, *virB10*, *virB11* and *virD4* of *Agrobacterium tumefaciens*. The *virB8*, *virB9*, *virB10* and *virB11* homologues were previously identified in pVir by Bacon *et al.* [55]. There are also some predicted CDSs with homology to genes from conjugative plasmids that are not involved in the formation of a type IV secretion apparatus. These include a homologue of TrbM from the *Escherichia coli* plasmid RP4 (pVir3) and a conjugal transfer protein homologue of *Rhizobium loti* (pVir37).

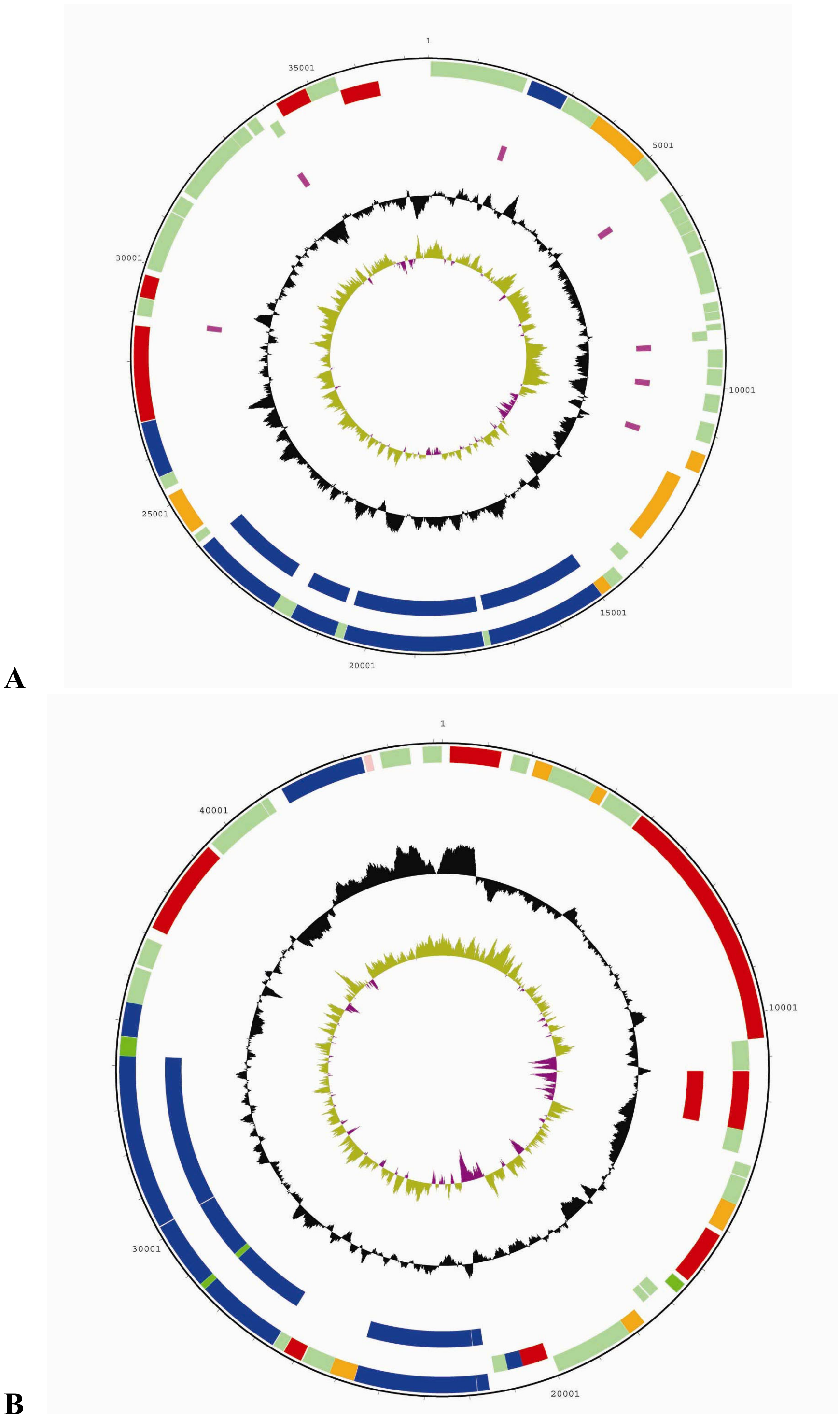


Fig 3.1: A circular representation of the plasmid sequences. A: pVir, B: pTet.

The circles represent the following features, numbering from the outside in: 1, 2, all CDSs (transcribed clockwise and anticlockwise respectively); 3, CDSs predicted to encode type IV secretion system homologues transcribed clockwise; 4, in A only: repeat units, in B only: as 3, transcribed anticlockwise; 5, G+C content; 6, GC deviation ((G-C)/ (G+C)) with a window size of 250 bp and a step size of 10 bp. The 12 o'clock position of each circle represents the predicted origin of replication and CDS colours represent the following putative functions: red, information transfer (transcription/ translation + DNA/ RNA modification); light green, unknown; dark green, surface; orange, conserved hypothetical; blue, pathogenicity/ adaptation; pink, bacteriophage/ IS elements.

From **Fig 3.1A** it can be seen that the CDSs predicted to encode a type IV secretion system have a higher G+C content than the rest of the plasmid. Although the overall G+C content of pVir is 26% the region containing the type IV secretion system homologues is 29.4% G+C which is much closer to that of the *C. jejuni* chromosome and pTet, indicating that, in effect, the rest of the plasmid has a lower G+C content (approximately 24%). The majority of the CDSs are transcribed in one direction, and there are only a few predicted CDSs transcribed in the opposite direction and these correlate with changes in GC deviation, indicating potential recent re-arrangements. It has been noted that several bacterial genomes show a preference for G over C on the leading strand extending from the origin of replication to the termination region [115]. Strand compositional asymmetry may arise due to a combination of factors including replication and repair mechanisms, transcription, and selective constraints affecting amino acid and codon usage [115]. Strand compositional asymmetry may not be as apparent in plasmids between the origin and terminus of replication as it is for bacterial chromosomes [116].

There are many repeats in the sequence of pVir (**Fig 3.1A**). CDSs pVir17 and pVir18 are flanked by a perfect 156 bp direct repeat (rep3 and rep4). This repeat unit is present at 5 other intergenic locations on the plasmid in a partial or imperfect form giving 7 units in total. Repeat 6 and repeat 2 share the highest identity to repeat 3 and repeat 4 although repeat 2 has an 11 bp section that breaks the identity in the middle. Repeat 7, repeat 1 and repeat 5 are less conserved towards the ends showing highest identity in the middle (**Fig 3.2**). These repeat units are spread evenly around the lower G+C portion of the plasmid in intergenic regions and are themselves A+T rich. It is unclear what function these repeats may have in this plasmid.

```
rep3     1 AAAAAAGGGGGAAAATGTTTCGG-TTTGGTGCAAAATGAGTTTAAAAA-AACATATAAAA
rep4     1 AAAAAAGGGGGAAAATGTTTCGG-TTTGGTGCAAAATGAGTTTAAAAA-AACATATAAAA
rep6     1 AAAAAAAGGGGAAAATGTTTCGG-TTTGGTGCAAAATGAGTTTAAAAA-A--ATATAAAT
rep2     1 AAAAAAAGGGGAAAATGTTTTGGGTTTGGTGCAAAATGAGTTTAAAAA-AACATATAAAA
rep7     1 GAAAAAATCTCTTAATTCTAATTTTATTCTACGACACTATATTATAAATAACATATAAAT
rep1     1 GCCAAAAGATGAAAACGGTTTTCCTGTTTTT----ACTTTTTAAAAATTAACATATAAAA
rep5     1 ATTATATCATAAATATTAATAAAAATCAATATTTATCATTAATAAATATAATTATTAAAA


rep3    59 AAAGTATAAATAACATATAAA-----------AAATGTATATATTAAGTATAGATTAAGT
rep4    59 AAAGTATAAATAACATATAAA-----------AAATGTATATATTAAGTATAGATTAAGT
rep6    57 AACCTATAAATAACATATAAA-----------AAAGGTATAGATTAAGTATAGATTAAGT
rep2    60 AAAGTATAAATAACATATAAATAACATATAAAAAAGGTATATATTAAGTATAGATTAAGT
rep7    61 AACATATAAAAAACCTATAAA-----------AAATGTATATATTAAGTATATATTAAGT
rep1    57 AATGTATAAATAACATATAAA-----------AAAGGTATATATTAAGTATATACTAAGT
rep5    61 AACATATAAAAAACATATAAA-----------AAAGGTATAGATTAAGTATATATTAAGT


rep3   108 ATAAAAAAGGTATAATTATAATAACAAAAACAAAAGACAAAGG-CAAAAA
rep4   108 ATAAAAAAGGTATAATTATAATAACAAAAACAAAAGACAAAGG-CAAAAA
rep6   106 ATAAAAAAGGTATAATTATAATAACAAAAACAAAAGACAAAGG-CAAAAA
rep2   120 ATAAAAAAGGTATAATTATAATAACAAAAACAAAAGACAAAGGACAAAAA
rep7   110 ATAAAAAATGTATAATTATAAGA-CAAAA--CAAAGACAAAGGATTAAAG
rep1   106 ATAAAAAAGGTATAATTTTACAA--AAAAGGAGAAATATAGTGAGAAAAA
rep5   110 ATAAAAAATGTATAATTATATTAATTTAATTTTTAAAATAGGAGTAAAAA
```

Fig 3.2: Alignment of the long repeat units of plasmid pVir. Repeat units were numbered sequentially from the predicted origin of replication (see fig 3.1A). Repeats 3 and 4 are perfect 156 bp direct repeats. Repeats 2 and 6 are imperfect repeats with repeat 2 containing an 11 bp interruption in the centre of the conserved region. Repeats 1, 5 and 7 share the most identity in the central portion of the sequence and are less conserved towards the ends.

3.2.3.2 Characteristics of pTet

In the plasmid pTet there are 18 predicted CDSs with no detectable homology to proteins from other bacteria; representing 36% of the total CDSs. There are also a number of predicted CDSs with homology to hypothetical proteins from other organisms. There are more type IV secretion system homologues in pTet than there are in pVir. Many of the type IV secretion system homologues in pTet share highest similarity to proteins from *Actinobacillus actinomycetemcomitans* (pTet27, pTet32, pTet33, pTet37, pTet39 and pTet11). The predicted CDS pTet27 is similar to a predicted ATPase from *Actinobacillus actinomycetemcomitans* which shows homology at the N-terminus to VirB3 and VirB4 in the

C-terminus. The other predicted CDSs are homologous to VirB5, VirB6, VirB10, VirD4, and VirD2 of *Agrobacterium tumefaciens* respectively. *Actinobacillus actinomycetemcomitans* is a human pathogen associated with periodontal disease that encodes a type IV secretion system on the 25 Kb plasmid pVT745 that is also present on the chromosome of another strain [117;118]. There are also CDSs similar to homologues of a type IV secretion system from the partially sequenced plasmid pCjA13; pTet35, pTet36 and pTet38 which are equivalent to VirB8, VirB9 and VirB11 of *Agrobacterium tumefaciens* respectively [56]. Downstream of the type IV secretion system homologues there is a CDS with similarity to the lipoprotein MagB13 from *Actinobacillus actinomycetemcomitans* followed by a CDS with homology to TrbM from *Haemophilus aegyptius* in the same arrangement as in plasmid pVT745 of *Actinobacillus actinomycetemcomitans*. There are some predicted CDSs which show homology to proteins involved in plasmid maintenance e.g. a replication protein (pTet1), a single-stranded binding protein (pTet30), a DNA primase (pTet16) and a topoisomerase (pTet44). There is also a member of the site specific DNA recombinase family (pTet23c) and in the same region there is also a CDS (pTet24c) with homology to VapD2 from *Riemerella anatipestifer* plasmid pCFC1. Proteins containing a VapD N-terminal domain have been implicated in virulence [119].

From **Fig 3.1B** it is apparent that there is a region of high G+C around the tetracycline resistance gene (42%) and a dip in G+C content before the replication protein. As with pVir predicted CDSs on the opposite strand correlate with changes in GC deviation.

In pTet the putative recombinase pTet23 is located on a region that is flanked by imperfect 31 bp inverted repeats (**Fig 3.3**); 25 bp out of the 31 bp are identical. These repeats enclose the region including pTet23-pTet25 which encodes proteins on the opposite strand to the surrounding ones. There is also a further set of imperfect inverted repeats (26 bp out of 34 bp are identical) within the first that surrounds pTet24-pTet25. It is possible that this region is invertible with the recombinase acting on one pair of repeats. Recombinases are known to play a role in plasmid replication by resolving plasmid multimers [120]. Recombinases have also been implicated in variable expression of proteins by inverting regions of DNA containing promoters to switch on and off downstream genes. In several studies this has been implicated in generating bacterial cell surface diversity [121-123]. In the case of pTet there is a predicted promoter present on the invertible region of DNA that is positioned directly before genes predicted to encode type IV secretion system homologues. There are three sigma factors in *C. jejuni*, RpoD (sigma 70), FliA (sigma 28) and RpoN (sigma 54) [124]. It has been suggested that the *C. jejuni rpoD* promoters contain a periodic signal, involving variation in A+T content and T stretches, instead of a conserved -35 box [125], however another group have proposed a consensus sequence for the -35 region [126]. The region upstream of pTet26, representing the start of the type IV secretion system operon, does not contain sequence with strong agreement to the proposed *Campylobacter rpoD* promoter consensus sequences. Using the bacterial promoter prediction program BPROM (http://www.softberry.com) which searches for agreement to the *Escherichia coli* σ^{70} consensus -10 sequences of AACTAAATT and TTTTATAAT, -35 sequences of TTGAAT and TTTAAT were predicted on opposite strands (**Fig 3.4**) suggesting a bidirectional promoter region between the inner and outer inverted repeats present between pTet25 and pTet26. The Neural Network Promoter Prediction program (http://www.fruitfly.org/seq_tools/promoter.html) also identified putative promoters between the two repeat units, IR1 and IR2, on both strands. However, it should be noted that these predictions are based on the *Escherichia coli* paradigm and may not hold for *Campylobacter;* in addition transcription of this operon may be under the control of an alternative sigma factor. The inverted repeat unit is located only 18 bp upstream of the start codon of the predicted CDS pTet26 suggesting that a promoter for this operon would be located within the inverted repeat region. This suggests that control of transcription of the type IV system could be under the control of this putatively variable promoter and this will be discussed further in chapter 7.

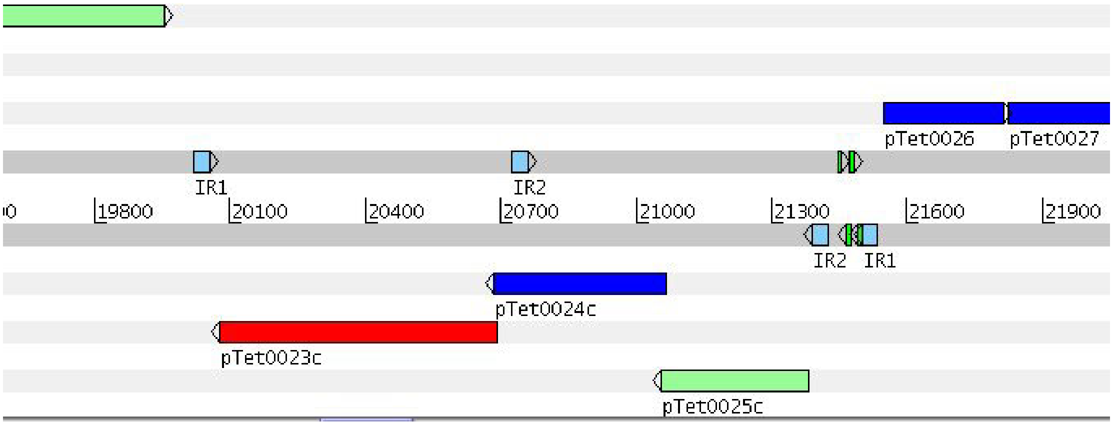


Fig 3.3: Putative invertible region in the plasmid pTet. The region is viewed using Artemis [102], forward and reverse DNA lines are represented by the central dark grey bars. The three forward and three reverse reading frames translated from the DNA sequence are represented by the light grey bars. Open boxes show features: sets of inverted repeats are marked by light blue boxes labelled IR1 and IR2 on the DNA lines. The sites of predicted promoters are labelled by dark green boxes on the forward and reverse DNA lines appropriate to their predicted orientation. CDSs are marked on their reading frames with the following colours to indicate functional categories: light green, unknown; red, information transfer (transcription/ translation + DNA/RNA modification); blue, pathogenicity/ adaptation. pTet23c is a putative site-specific DNA recombinase which may invert the regions between either set of inverted repeats.

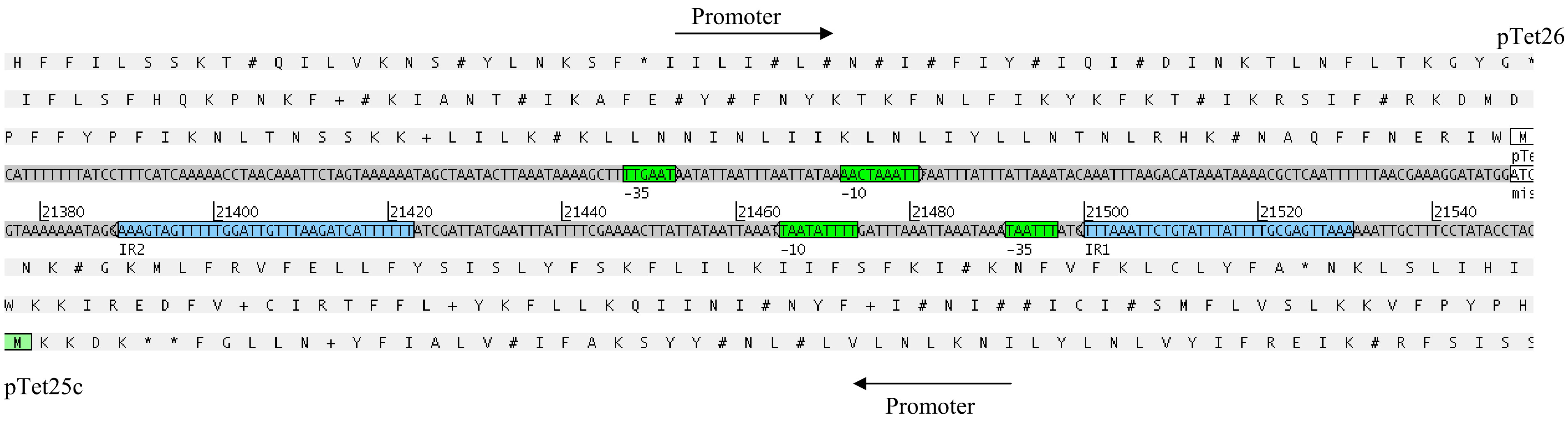


Fig 3.4: Region between pTet25c and pTet26 predicted to contain a promoter. The region is viewed using Artemis [102], forward and reverse DNA lines are represented by the central dark grey bars. The three forward and three reverse reading frames translated from the DNA sequence are represented by the light grey bars. Open boxes show features: the inverted repeats are marked by light blue boxes labelled IR1 and IR2 on the DNA lines. There are predicted promoters on both strands positioned between the two repeat units, positioned upstream of the start of pTet25c and the start of the putative type IV secretion system, pTet26.

3.3 Discussion

3.3.1 Comparison to published sequences

During the course of this study the sequences of pVir [127] and pTet [128] were published. The numbers below refer to the locations of features in the sequence determined in this study.

3.3.1.1 pVir

The sequence of pVir from this study shows good agreement with the published pVir sequence, with a 99% nucleotide match. There are however some small differences between the two sequences with the pVir sequence from this study containing an extra 5 bp; there are 10 bp differences in all. The predicted CDSs of pVir from the Bacon study have been named Cjp and numbered from VirB8/ComB1 [127].

In the pVir sequence from this study there is 1 bp missing before base 25348 which results in a frame shift relative to the Bacon sequence, extending the N-terminus of the hypothetical pVir36 to a total length of 89 aa compared to Cjp09 which is 48 aa (**Fig 3.5**). An extra G at base 33098 leads to a frame shift relative to the Bacon sequence which results in the hypothetical pVir48 having 2 aa less than Cjp22 (**Fig 3.6**).

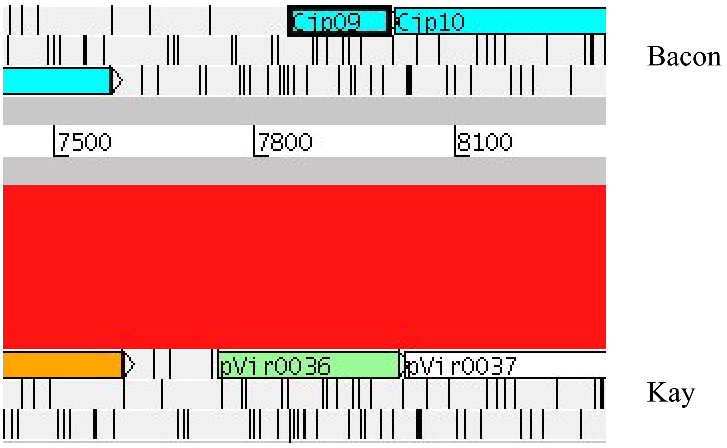


Fig 3.5: A WUBLASTN comparison of the published pVir sequence with the pVir sequence from this study in the region of pVir36. The comparison is viewed using the Artemis Comparison Tool (ACT) where blocks of red indicate sequence homology and the intensity of colour is proportional to the percent identity. However, single base pair changes cannot be accurately represented. The three forward translated reading frames from each sequence are represented by light grey bars and stop codons are indicated by vertical black lines. CDSs are indicated by open boxes and the CDSs from this study are coloured to represent functional categories: orange, conserved hypothetical; light green, unknown; white, pathogenicity/ adaptation. A base pair difference between the two sequences results in a frame shift compared to the Bacon sequence and extends the N-terminus of CDS pVir36. pVir36 is predicted to encode 89 aa whereas Cjp09 is predicted to encode 48 aa.

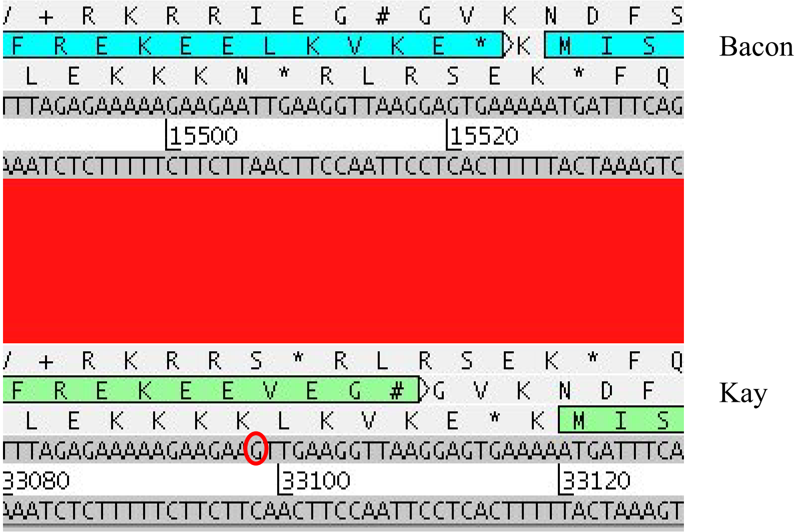
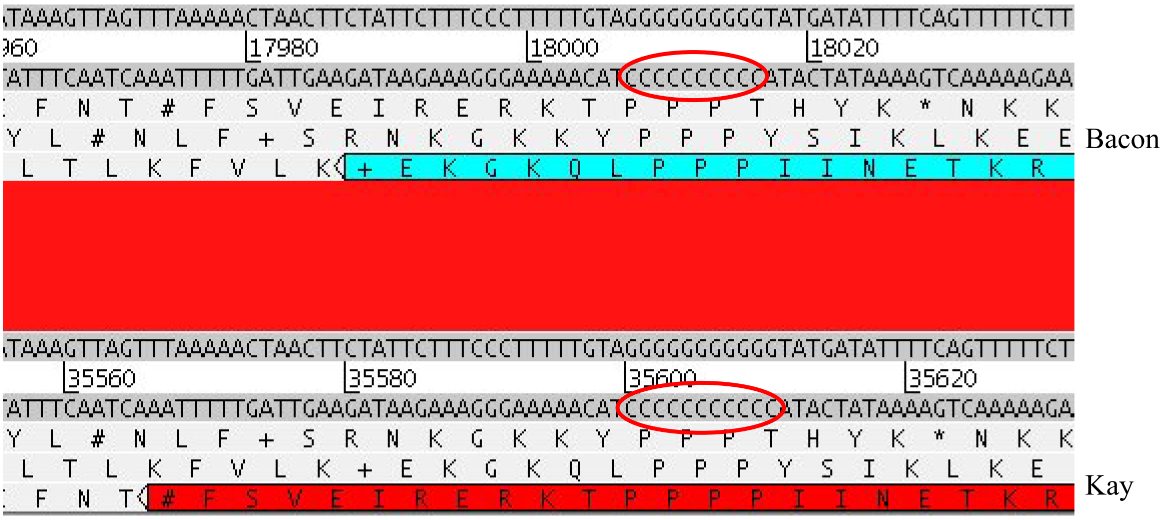

/ + R K R R I E G # G V K N D F S
F R E K E E L K V K E * K M I S
L E K K K N * R L R S E K * F Q
TTTAGAGAAAAAGAAGAATTGAAGGTTAAGGAGTGAAAAATGATTTCAG
15500
15520
AAATCTCTTTTTCTTCTTAACTTCCAATTCCTCACTTTTTACTAAAGTC
Bacon
/ + R K R R S * R L R S E K * F Q
F R E K E E V E G # G V K N D F
L E K K K K L K V K E * K M I S
TTTAGAGAAAAAGAAGAAGTTGAAGGTTAAGGAGTGAAAAATGATTTCA
33080
33100
33120
AAATCTCTTTTTCTTCTTCAACTTCCAATTCCTCACTTTTTACTAAAGT
Kay

Fig 3.6: A WUBLASTN comparison of the published pVir sequence with the pVir sequence from this study in the region of pVir48. The comparison is viewed using ACT where blocks of red indicate sequence homology and the intensity of colour is proportional to the percent identity. However, single base pair changes cannot be accurately represented. Forward and reverse DNA lines are in dark grey, and the three forward translated reading frames from each sequence are represented

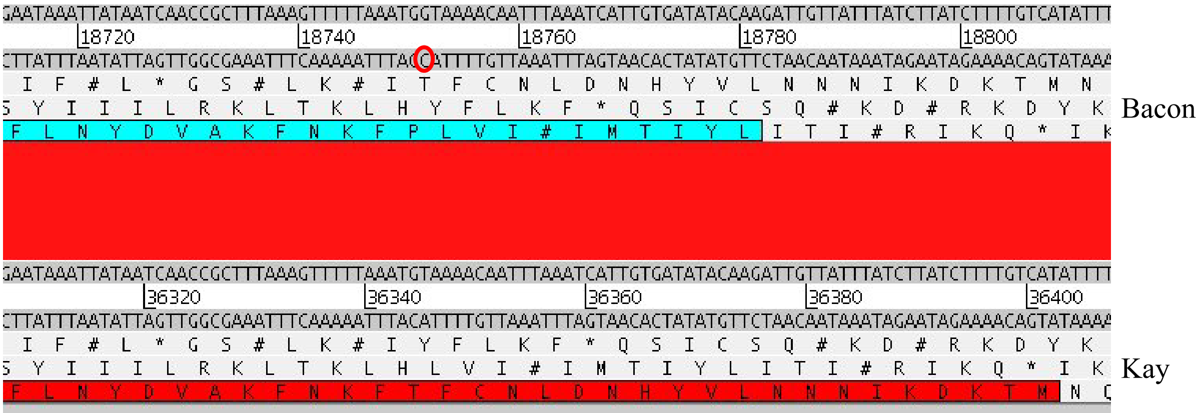
by light grey bars. CDSs are indicated by open boxes and the CDSs from this study are coloured to represent functional categories: light green, unknown. An extra base in the sequence from this study is circled in red; this extra base leads to a frame shift in CDS pVir0048 compared to CDS Cjp22 in the published sequence.

There are 3 bp differences in the putative *repA* gene. There is a homopolymeric tract of guanine residues (G) which appears to vary between G(10-11) before base 35611 (**Fig 3.7**). In the shotgun assembly three reads contained G(10) and seven reads contained G(11) suggesting that this homopolymeric tract varies in length due to slip-strand misspairing. This would vary the final 6-11 aa of the RepA protein, which is unlikely to have a functional consequence. Also in this predicted gene there is a base pair difference at 35898 leading to a predicted amino acid change from K in the Bacon sequence to E in the pVir sequence from this study. Also, importantly, there is a base missing before 36345 which results in a frame shift relative to the Bacon sequence giving an uninterrupted reading frame for *repA* indicating that this is not a pseudogene in this version of the sequence (**Fig 3.8**). RepA is a replication initiator protein which recognizes specific sequences at the origin of replication and is required by most plasmids replicating by the theta mechanism, in addition to the host DnaA protein, to initiate plasmid replication. There are however some examples where the initiation of plasmid replication can occur in the absence of a plasmid encoded initiator protein, the best characterized of which is ColE1 [108]. It would be necessary to conduct further studies to identify whether pVir54c, predicted to encode a RepA protein, is required for plasmid replication, or if it is accessory and therefore may accumulate deleterious mutations without lethal consequence for the plasmid.



```
TAAAGTTAGTTTAAAAACTAACTTCTATTCTTTCCCTTTTTGTAGGGGGGGGGGTATGATATTTTCAGTTTTTCTT
960                   17980                   18000                   18020
ATTTCAATCAAATTTTTGATTGAAGATAAGAAAGGGAAAAACATCCCCCCCCCCATACTATAAAAGTCAAAAAGAA
 F  N  T  #  F  S  V  E  I  R  E  R  K  T  P  P  P  T  H  Y  K  *  N  K  K
Y  L  #  N  L  F  +  S  R  N  K  G  K  K  Y  P  P  P  Y  S  I  K  L  K  E  E     Bacon
 L  T  L  K  F  V  L  K  +  E  K  G  K  Q  L  P  P  P  I  I  N  E  T  K  R

TAAAGTTAGTTTAAAAACTAACTTCTATTCTTTCCCTTTTTGTAGGGGGGGGGGGTATGATATTTTCAGTTTTTCT
    35560                   35580                   35600                   35620
ATTTCAATCAAATTTTTGATTGAAGATAAGAAAGGGAAAAACATCCCCCCCCCCCATACTATAAAAGTCAAAAAGA
Y  L  #  N  L  F  +  S  R  N  K  G  K  K  Y  P  P  P  T  H  Y  K  *  N  K  K
 L  T  L  K  F  V  L  K  +  E  K  G  K  Q  L  P  P  P  Y  S  I  K  L  K  E
 F  N  T  #  F  S  V  E  I  R  E  R  K  T  P  P  P  P  I  I  N  E  T  K  R     Kay
```

Fig 3.7: A WUBLASTN comparison of the published pVir sequence with the pVir sequence from this study showing the C-terminus of *repA*. The comparison is viewed using ACT where blocks of red indicate sequence homology and the intensity of colour is proportional to the percent identity. However, single base pair changes cannot be accurately represented. Forward and reverse DNA lines are in dark grey, and the three reverse translated reading frames from each sequence are represented by light grey bars. CDSs are indicated by open boxes and the CDSs from this study are coloured to represent functional categories: red, information transfer (transcription/ translation + DNA/ RNA modification). The C-terminus of the putative *repA* gene contains a homopolymeric tract which alters the last 11aa of the encoded protein.



```
GAATAAATTATAATCAACCGCTTTAAAGTTTTTAAATGGTAAAACAATTTAAATCATTGTGATATACAAGATTGTTATTTATCTTATCTTTTGTCATATTT
       18720                   18740                   18760                   18780                   18800
CTTATTTAATATTAGTTGGCGAAATTTCAAAAATTTACCATTTTGTTAAATTTAGTAACACTATATGTTCTAACAATAAATAGAATAGAAAACAGTATAAA
 I  F  #  L  *  G  S  #  L  K  #  I  T  F  C  N  L  D  N  H  Y  V  L  N  N  N  I  K  D  K  T  M  N
S  Y  I  I  I  L  R  K  L  T  K  L  H  Y  F  L  K  F  *  Q  S  I  C  S  Q  #  K  D  #  R  K  D  Y  K     Bacon
 F  L  N  Y  D  V  A  K  F  N  K  F  P  L  V  I  #  I  M  T  I  Y  L  I  T  I  #  R  I  K  Q  *  I  K

GAATAAATTATAATCAACCGCTTTAAAGTTTTTAAATGTAAAACAATTTAAATCATTGTGATATACAAGATTGTTATTTATCTTATCTTTTGTCATATTTT
          36320                   36340                   36360                   36380                   36400
CTTATTTAATATTAGTTGGCGAAATTTCAAAAATTTACATTTTGTTAAATTTAGTAACACTATATGTTCTAACAATAAATAGAATAGAAAACAGTATAAAA
 I  F  #  L  *  G  S  #  L  K  #  I  Y  F  L  K  F  *  Q  S  I  C  S  Q  #  K  D  #  R  K  D  Y  K
S  Y  I  I  I  L  R  K  L  T  K  L  H  L  V  I  #  I  M  T  I  Y  L  I  T  I  #  R  I  K  Q  *  I  K     Kay
 F  L  N  Y  D  V  A  K  F  N  K  F  T  F  C  N  L  D  N  H  Y  V  L  N  N  N  I  K  D  K  T  M  N  Q
```

Fig 3.8: A WUBLASTN comparison of the published pVir sequence with the pVir sequence from this study showing the N-terminus of *repA*. The comparison is viewed using ACT where blocks of red indicate sequence homology and the intensity of colour is proportional to the percent identity. However, single base pair changes cannot be accurately represented. Forward and reverse DNA lines are in dark grey, and the three reverse translated reading frames from each sequence are represented by light grey bars. CDSs are indicated by open boxes and the CDSs from this study are coloured to represent functional categories: red, information transfer (transcription/ translation +

DNA/ RNA modification). There is a base missing compared to the published sequence which is circled in red; this leads to a frame shift compared to the published sequence leading to a complete replication gene.

Within the putative origin of replication there is a large homopolymeric tract of adenine residues (A). In the pVir sequence of this study there are A(21) and in the Bacon sequence there are A(20). Also in a non-coding region, there is an extra bp at 5422 between pVir6 and pVir7. In the same region at base 5692 there is an extra base in the repeat region. Further on, in pVir9, there is an extra base at 6604 which leads to a frame shift and extension of pVir9 to 73 aa rather than 39 aa in Cjp37 (**Fig 3.9**). There is an extra base at 14490 leading to a frame shift relative to the Bacon sequence which causes Cjp50 (30 aa) and Cjp51 (45 aa) to fuse in the same reading frame giving pVir24 (101 aa) (**Fig 3.10**).

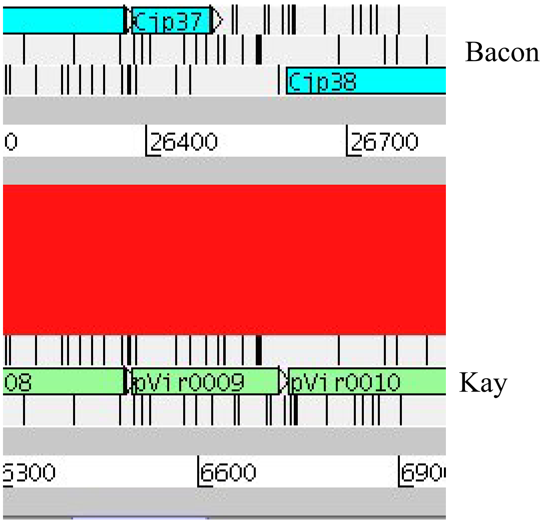


Fig 3.9: A WUBLASTN comparison of the published pVir sequence with the pVir sequence from this study in the region of pVir9. The comparison is viewed using ACT where blocks of red indicate sequence homology and the intensity of colour is proportional to the percent identity. However, single base pair changes cannot be accurately represented. The three forward translated reading frames from each sequence are represented by light grey bars and stop codons are indicated by vertical black lines. CDSs are indicated by open boxes and the CDSs from this study are coloured to represent functional categories: light green, unknown. An extra base in the sequence from this study results in a frame shift relative to the Bacon sequence; pVir0009 is predicted to encode 73 aa whereas the CDS Cjp37 is predicted to encode 39 aa.

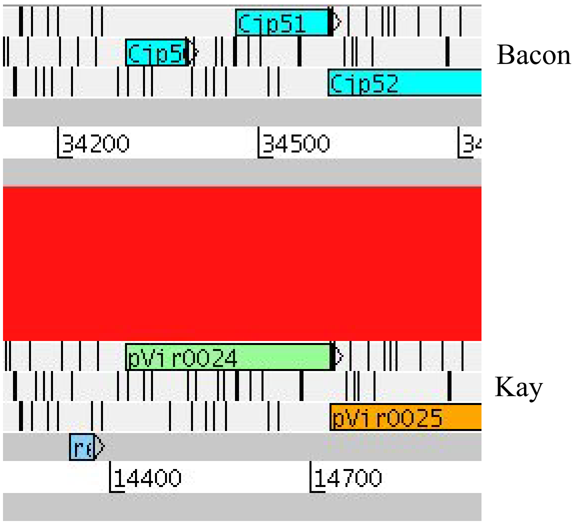


Fig 3.10: A WUBLASTN comparison of the published pVir sequence with the pVir sequence from this study in the region of pVir24. The comparison is viewed using ACT where blocks of red indicate sequence homology and the intensity of colour is proportional to the percent identity. However, single base pair changes cannot be accurately represented. The three forward translated reading frames from each sequence are represented by light grey bars and stop codons are indicated by vertical black lines. CDSs are indicated by open boxes and the CDSs from this study are coloured to represent functional categories: orange, conserved hypothetical; light green, unknown. An extra base in the sequence from this study results in a frame shift relative to the Bacon sequence; the CDS pVir0024 appears to be a fusion of Cjp50 and Cjp51.

With the exception of *repA* these differences occur either in intergenic regions or hypothetical CDSs and may represent variation within the population, as the plasmids used as sequencing templates were isolated separately.

3.3.1.2 pTet

The sequence of pTet in this study shows 99% nucleotide identity across the entire length to the published version. The pTet sequence from this study is 1bp shorter and there are 9 differences in total from the published sequence. In the published sequence the predicted CDSs have been named cpp or cmg for the mating associated genes. At base 21115 there is a synonymous base change of TCC to TCT in the hypothetical CDS pTet25 compared to cpp29. There is an extra base at 24460 which results in a frame shift relative to the published sequence resulting in an extension to the C-terminus of pTet27 (922 aa) compared to cmgB3/4 (883 aa) (**Fig 3.11**) the ATPase from *Actinobacillus actinomycetemcomitans* with which these CDSs share identity is 923 aa. There are 2 bp differences at 24708 and 24709 (GT to AC) and also a base missing before 24715 which leads to a frame shift causing an extension to the N-terminus of pTet28 hypothetical protein (188 aa) compared to cpp32 (162 aa) (**Fig 3.11**). There is 1 less base at 28377 plus an extra base at 28414 leading to an extension at the C-terminus of pTet33 VirB6 homologue (332 aa) compared to cmgB6 (281 aa) (**Fig 3.12**). At base 28863 there is a synonymous base change of CCG to CCT in VirB8. One base is missing before 35080 relative to the published sequence which results in a frame shift extending the reading frame of pTet41 which is predicted to encode a TrbM homologue (254 aa) compared to cpp45 (143 aa) (**Fig 3.13**). The TrbM-like protein of *Haemophilus aegyptius* is 217 aa long.

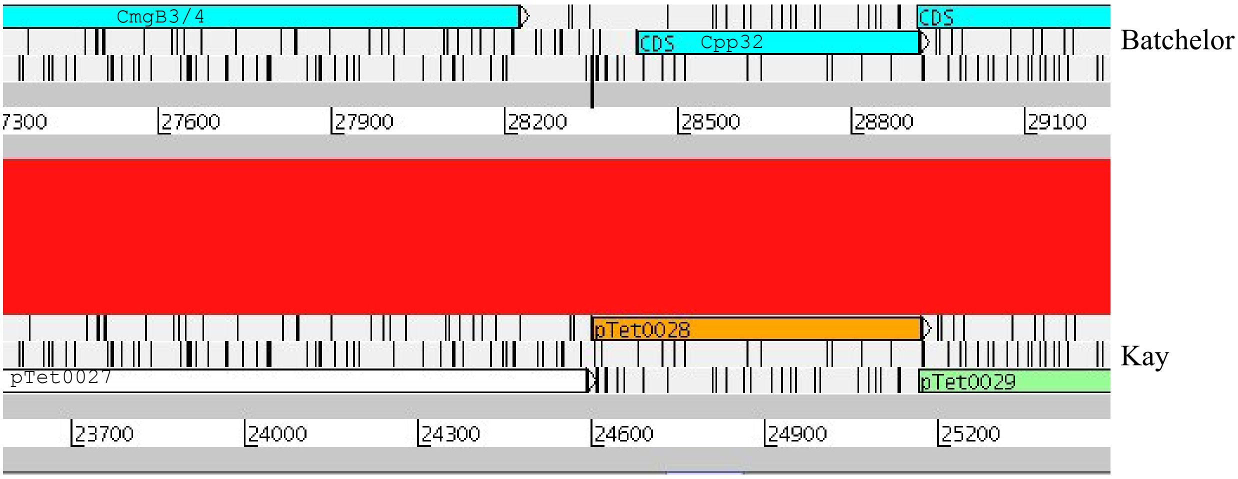


Fig 3.11: A WUBLASTN comparison of the published pTet sequence with the pTet sequence from this study in the region of pTet28. The comparison is viewed using the ACT where blocks of red indicate sequence homology and the intensity of colour is proportional to the percent identity. However, single base pair changes cannot be accurately represented. The three forward translated reading frames from each sequence are represented by light grey bars and stop codons are indicated by vertical black lines. CDSs are indicated by open boxes and the CDSs from this study are coloured to represent functional categories: orange, conserved hypothetical; light green, unknown; white,

pathogenicity/ adaptation. The C-terminus of pTet27 and the N-terminus of pTet28 are extended relative to cmgB3/4 and cpp32 from the published sequence respectively.

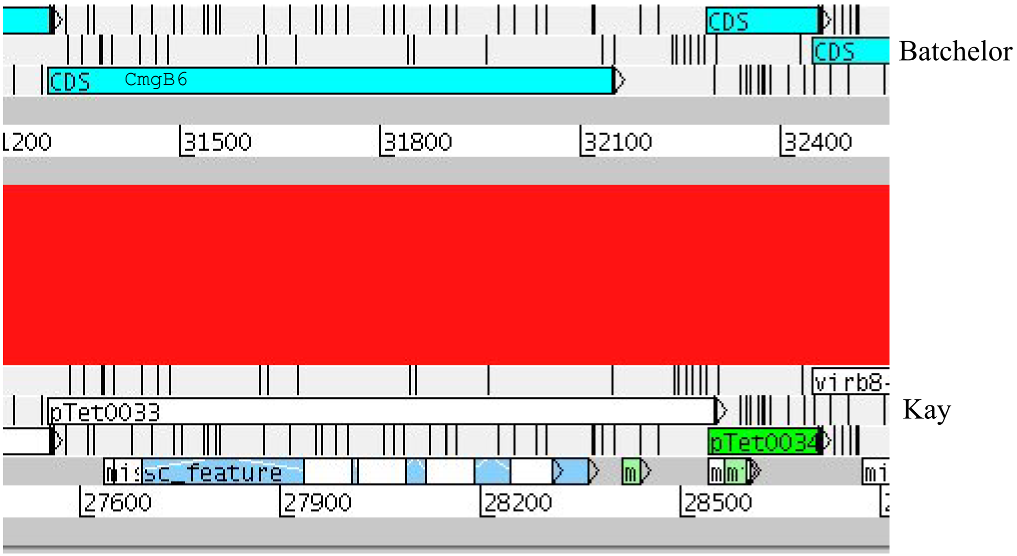


Fig 3.12: A WUBLASTN comparison of the published pTet sequence with the pTet sequence from this study in the region of pTet33. The comparison is viewed using ACT where blocks of red indicate sequence homology and the intensity of colour is proportional to the percent identity. However, single base pair changes cannot be accurately represented. The forward and reverse DNA lines are represented in dark grey, the three forward translated reading frames from each sequence are represented by light grey bars and stop codons are indicated by vertical black lines. Features are indicated by open boxes: pFam (blue), signalP (white), tmhmm (white) and prosite (green) matches are indicated on the DNA lines; CDSs are represented on the reading frame lines and CDSs from this study are coloured to represent functional categories: white, pathogenicity/ adaptation; dark green, surface associated. The C-terminus of CDS pTet33 is extended compared to cmgB6.

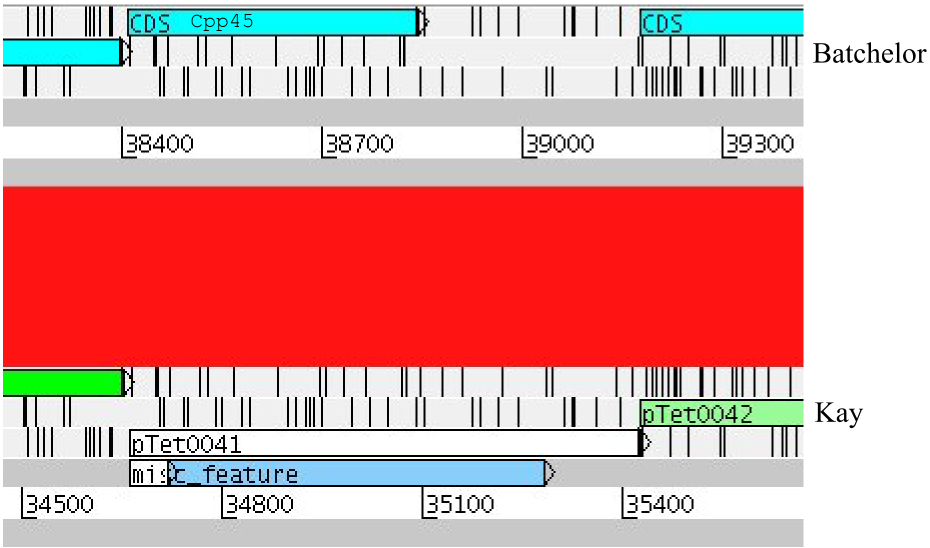


Fig 3.13: A WUBLASTN comparison of the published pTet sequence with the pTet sequence from this study in the region of pTet41. The comparison is viewed using ACT where blocks of red indicate sequence homology and the intensity of colour is proportional to the percent identity. However, single base pair changes cannot be accurately represented. The forward and reverse DNA lines are represented in dark grey, the three forward translated reading frames from each sequence are represented by light grey bars and stop codons are indicated by vertical black lines. Features are indicated by open boxes: pFam (blue) and signalP (white) matches are indicated on the DNA lines; CDSs are represented on the reading frame lines and CDSs from this study are coloured to represent functional categories: white, pathogenicity/ adaptation; dark green, surface associated; light green, unknown. A base missing compared to the published sequence leads to a frame shift extending the C-terminus of CDS pTet0041, a homologue of TrbM, compared to cpp45.

3.3.2 Characteristics of pVir and pTet

Bacon *et al.* suggest that plasmids have been found in 19-53% of *C. jejuni* strains [55], with the best characterized being plasmids encoding antibiotic resistance determinants. Plasmids have been implicated in virulence in other bacteria. The virulence plasmids of *Yersinia* spp encode a type III secretion system involved in the secretion of plasmid encoded *Yersinia* outer proteins (Yops) that block phagocytosis, and induce cytokine expression and apoptosis [78;129]. *Shigella flexneri* contains a virulence plasmid that encodes a type III secretion system which secretes invasion protein antigens (Ipa) that induce uptake of the pathogen into eukaryotic cells, apoptosis, and vacuole membrane lysis [77;129]. However, the role of plasmids in *C. jejuni* virulence has not been well studied to date. More recently several pVir genes have been subjected to mutational analysis [127]. Mutation of the predicted CDSs cjp15 and cjp29 which have no detectable homology to proteins from other bacteria, cjp32, which has similarity to Cj0041 and cjp49, a homologue of *Helicobacter pylori* HP0996, resulted in reduced invasion of a host cell line when compared to wild type levels [127]. pTet conjugation genes have also been subject to mutational analysis showing that cmgB3/4 is required for conjugal transfer [128]. Others have looked at the distribution of plasmids within populations of *C. jejuni* showing that few strains carry pVir. In one study one out of 16 plasmid containing clinical isolates was found to contain pVir, with 8 containing sequences with homology to pTet [56]. Another study found 18 out of 104 clinical isolates contained pVir [130].

In *Helicobacter pylori* there are two separate and functionally independent type IV secretion systems, one for protein translocation (*cag*) and one for natural transformation (*comB*) [131;132]. From the results of Bacon *et al.* it appears that the type IV secretion system of pVir may function in both roles: a mutation in *comB3* reduced adherence, invasion and natural transformation, although mechanisms for these traits are unknown. The *Yersinia* *enterocolitica* flagellum export apparatus has been shown to also secrete several extracellular proteins including YplA [133] showing that some export apparatuses may be multifunctional in contrast to the separate systems of *Helicobacter pylori*.

pTet also contains components of a type IV secretion/ conjugation system. Comparison to other type IV secretion systems (**Fig 3.14**) shows that the systems in pTet and pVir are similar except pTet has VirB2, B3, B5, B6 and B7 homologues. The only VirB homologue that is present in all known conjugation systems and absent from protein secretion systems is the homologue of VirB5 from *Agrobacterium tumefaciens*. Although the function of VirB5 is not known it has been suggested that it may be a minor structural component of the pilus along with VirB2 in *Agrobacterium tumefaciens* [134]. Most importantly pTet also contains a nickase (MagA2) homologue, also known as a relaxase, which would not be expected in a protein secretion system. Relaxases play an essential role in conjugative DNA transfer by nicking the DNA which must then be unwound by a DNA helicase to produce the single strand of DNA transferred to the recipient cell [135]. The MagA2 homologue, like MagA2 of *Actinobacillus actinomycetemcomitans* pVT745, does not contain any nucleotide-binding motifs or a helicase domain which is sometimes present in relaxases [117]. There is however a homologue of *Sinorhizobium meliloti* bacteriophage PBC5 DNA methylase within 800 bp on the opposite strand which contains a helicase domain which may play a role in DNA transfer.

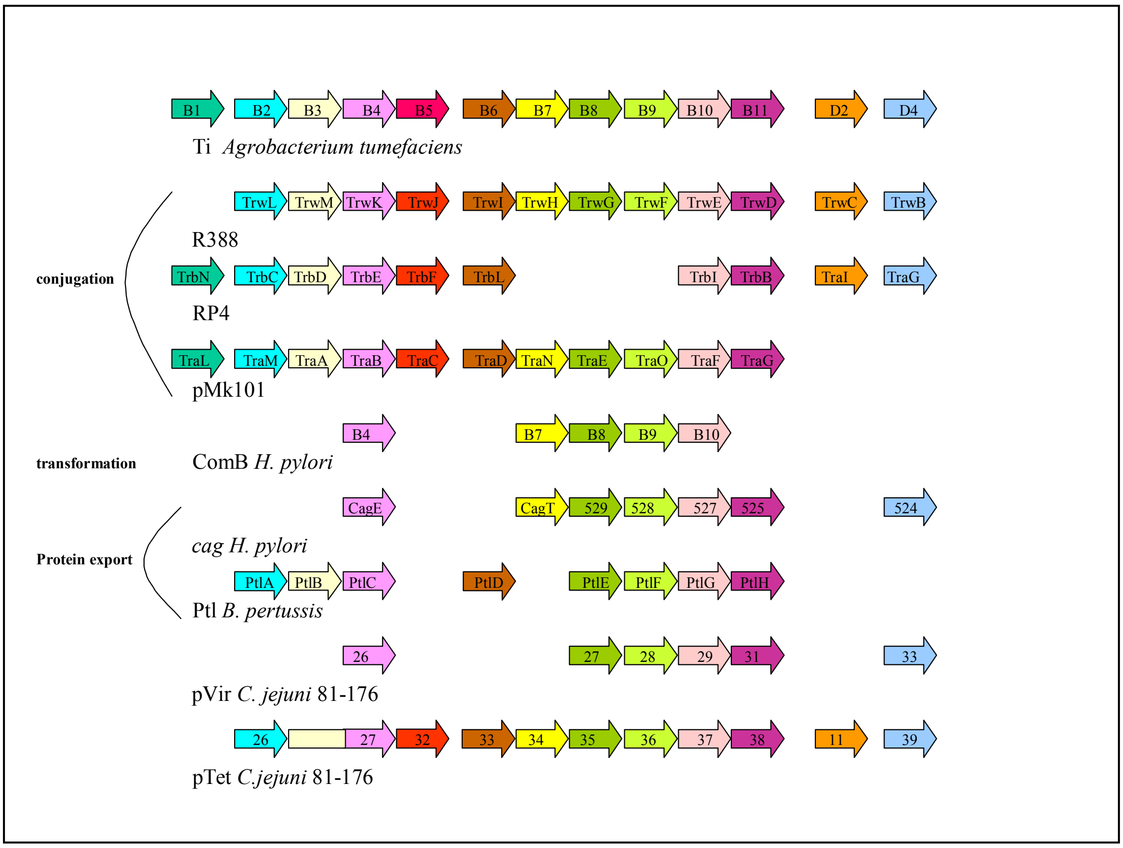


Fig 3.14: Schematic comparison of proteins involved in type IV bacterial secretion systems. Proteins that are homologous are shown with arrows of the same colour. The order of proteins expressed here is not necessarily the order in which the genes appear on respective stretches of DNA with the exception of *Agrobacterium tumefaciens.* Arrow sizes are not representative of gene or protein size. The plasmids R388, RP4 and pMk101 represent gram-negative conjugation schemes. The *Helicobacter pylori comB* region is involved in natural transformation by DNA uptake while the *Helicobacter pylori cag* and *Bordetella pertussis* ptl proteins form toxin secretion systems. Predicted CDSs of pTet and pVir from this study have been added with the locus_id number of the homologous CDS. This figure was adapted from the data of Cossart, P. *et al*. (2000) [129]; Firth, N. *et al*. (1996) [136]; Christie, P. J. (1997) [137] and Novak, K. F. *et al*. (2001) [118].

VirB/VirD4 homologues can mediate transfer of DNA and protein as they encode a transmembrane pilus structure. In *Agrobacterium tumefaciens* a complex of transfer proteins is necessary to chaperone DNA out of the cell. The transfer complex consists of a single VirD2 molecule bound to the 5` end of the DNA which is coated with VirE2, a single-strand DNA binding protein This suggests that these proteins carry a sequence necessary for export [134]. It may be that in the case of pVir, proteins that have roles in virulence are translocated, but this can not be confirmed until the secreted proteins are identified. Other groups are currently working towards this aim [127]. There has also been evidence of pVir genes being transcribed under the control of a σ^{54}-regulated promoter along with the flagellar secretory apparatus [138]. More recently a strong association between campylobacteriosis patients with bloody diarrhoea and the presence of pVir has been found [130]. Blood in patient stools is generally associated with invasion which supports the findings of Bacon *et al.* who suggested that pVir was associated with invasion of intestinal epithelial cells [55;127].

In order to characterize the true origin of replication for both of these plasmids it would be necessary to assess which regions could autonomously replicate. Other studies have attempted to identify origins of replication by cloning suspected regions into antibiotic resistance plasmids lacking an origin of replication [110]. As these plasmids are large it would be intuitive to expect that they are low copy number, however, only pVir contained a homologue of a partition gene. When the predicted partition gene cjp26 was mutated it showed no detectable phenotype and the plasmid appeared to be stably maintained [127]. Another intriguing characteristic in pTet that warrants further investigation is the possibility that the putative type IV secretion system is under the control of a variable promoter, this possibility will be discussed further in chapter 7.

4. Analysis of pUC libraries

4.1 Introduction

In order to screen the *C. jejuni* strains, 81-176, M1, 40671 and 52472 for regions of DNA not present in the sequenced strain NCTC 11168 a method of differential hybridization [92] was used. The strains were selected to show a range of phenotypes. Strain 81-176 is a commonly studied laboratory strain and as such many novel regions have already been identified. In addition strain 81-176 contains two plasmids not present in NCTC 11168 so this strain should provide a good reference to evaluate the method. Strain M1 was contracted by a scientist who developed severe inflammatory gastroenteritis following a visit to a poultry abattoir in the UK. Chickens have been suggested as an important route of transmission to humans, and as different strains are known to differ in their colonization potentials [60] this strain may provide information about colonization factors in addition to being virulent in humans. Strain 40671 is an outbreak strain; outbreaks of *Campylobacter* are rare, in addition this strain has been associated with water which may indicate that this strain is adapted to survival in the wider environment. Strain 52472 was isolated from a patient with septicaemia which may indicate that this strain has invasion factors and is adapted to survive within the blood stream. The differential hybridization step was planned as the initial phase of the project, in order to sample the entire genome. This would give an idea of the extent and variety of genes that are not present in strain NCTC 11168 and identify regions for further, in depth, analysis.

4.2 Results

4.2.1 Identification of DNA present in test strains and absent from NCTC 11168

A library of fragments 0.8-1.2kb was constructed in pUC19 for strains 81-176, M1, 40671 and 52472. Genes of *C. jejuni* NCTC 11168 have an average length of 948 bp [8] so an insert size of approximately 1 kb was selected for gene comparison as larger inserts are more likely to contain flanking DNA present in both strains being compared. Libraries of strains 81-176, 40671 and 52472 each consisted of 8064 clones and the library of strain M1 consisted of 8448 clones, representing roughly 5-fold coverage of the genome assuming a similar genome size to NCTC 11168. The idealised equation $P=1-e^{-x}$, where P = probability of a base being represented and x = raw coverage [139], indicates that with 5-fold coverage the library should represent 99.3% of the genome. These clones were arrayed onto a set of 3 (moderately charged) nylon membranes: one set was hybridized with labelled “self” genomic DNA and the other with labelled NCTC 11168 genomic DNA. Clones that hybridized to “self” genomic DNA, but not to NCTC 11168 genomic DNA were selected for sequencing (**Fig 4.1**). Sequence reads were then compared to the complete NCTC 11168 genome sequence using WUBLASTN and, in the case of 81-176 to the sequences of the plasmids pVir and pTet from this study (chapter 3, Appendix1 and Appendix 2). Sequence reads from the test strains that showed more than 85% nucleotide id (identity) to any of the comparator sequences were eliminated from further analysis. Reads that showed less than 85% nucleotide id to compared sequences were assembled using Phrap (Green, P., unpublished) into contiguous regions, then viewed and annotated using Artemis [102]. At this stage the assembled contiguous regions were again compared to the sequence of strain NCTC 11168. As some of the reads may have been of poor sequence quality the assembled consensus sequence had a higher similarity to the genome of strain NCTC 11168 across the entire length in some instances. In strain 81-176 this accounted for 5 regions, in strain M1 8 regions and in strain 52472 7 regions which were discounted from further analysis unless they occurred next to re-arrangement events compared to the NCTC 11168 chromosome or represented more complete versions of pseudogenes in NCTC 11168.

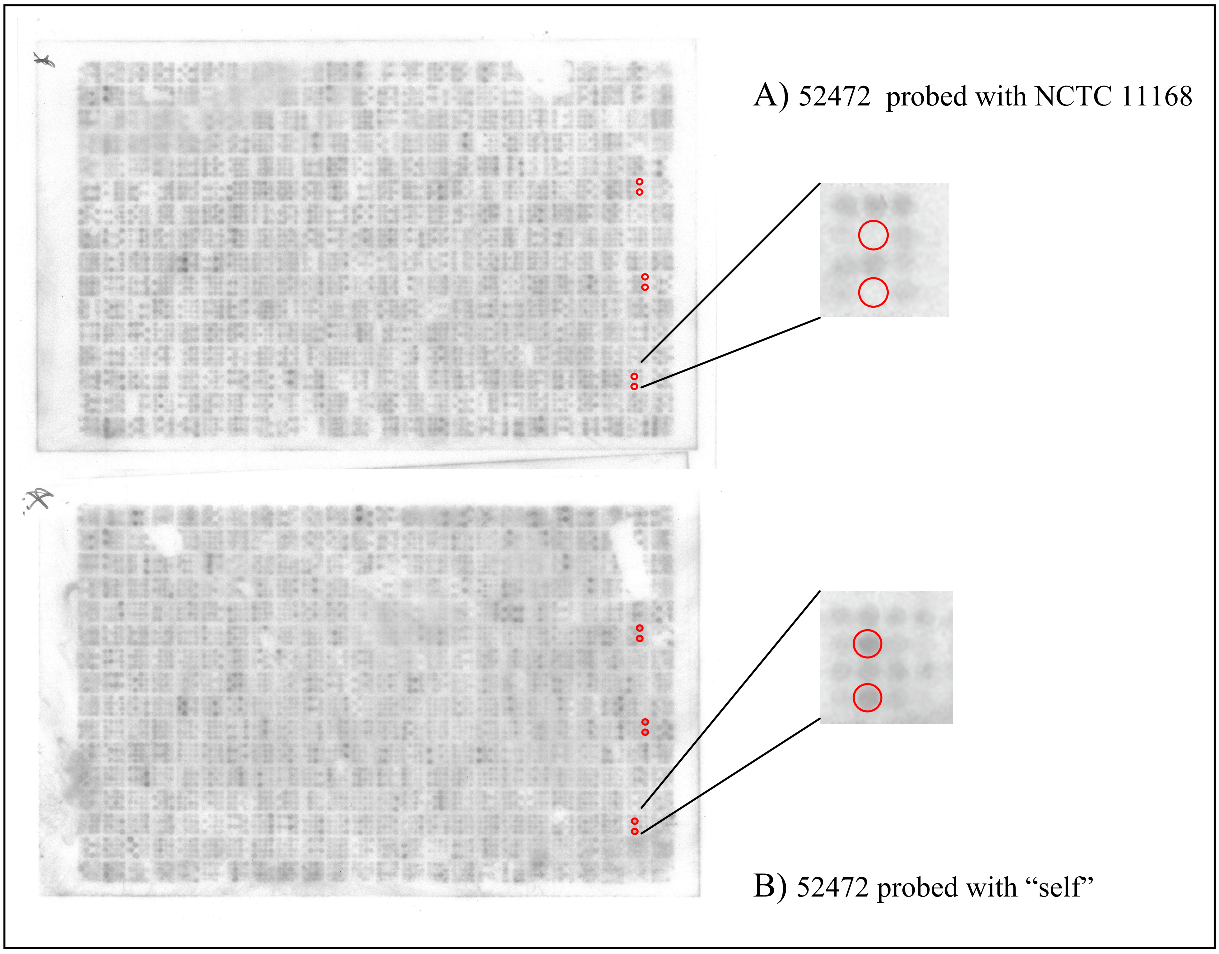


Fig 4.1: Differential genomic DNA hybridization array. A library of DNA fragments from each strain was spotted onto nylon membranes. A) shows an example of a membrane from strain 52472 probed with genomic DNA from NCTC 11168 and B) shows a membrane from strain 52472 probed with its own genomic DNA. The circled spots show representative examples of duplicate clones carrying DNA inserts that show significantly reduced hybridization with the NCTC 11168 probe compared with the "self" probe, and are therefore likely to carry inserts specific to that strain.

4.2.2 False positive and false negative testing

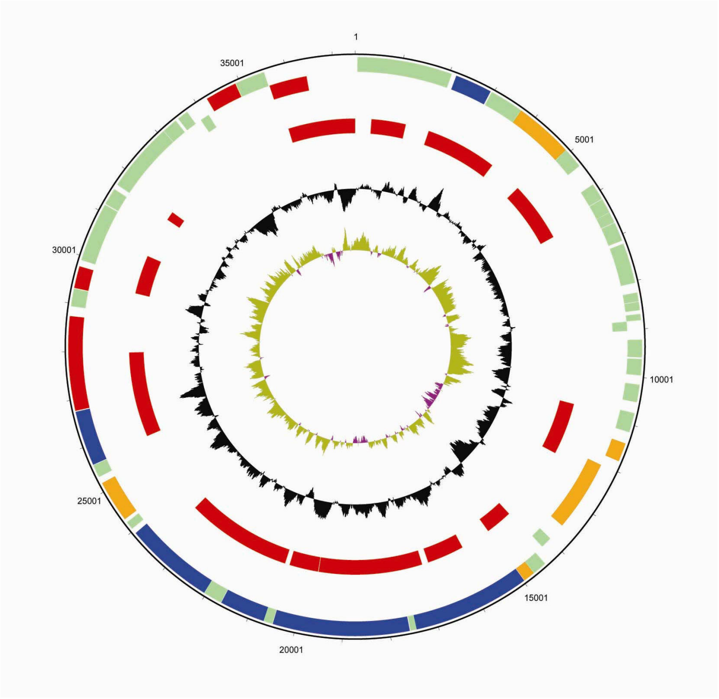
To validate the accuracy of the method, false positive and false negative values were calculated for each of the libraries. False positives were designated as clones that showed differential hybridization patterns and yet had more than 85% sequence similarity to chromosomal DNA from strain NCTC 11168. False negatives were designated as clones that did not show a differential hybridization pattern and yet had less than 85% sequence similarity to chromosomal DNA from strain NCTC 11168. For strain 81-176 135 out of 654 sequence reads (21%), for strain M1 276 out of 807 sequence reads (34%), for strain 40671 140 out of 413 sequence reads (34%) and for strain 52472 98 out of 1439 sequence reads (7%) were false positives. To test false negatives 192 clones were sequenced in both directions from strains 81-176 and M1. For strain 81-176 89% (168 out of 188 successfully end-sequenced) had more than 85% nucleotide id to strain NCTC 11168 whereas 11% were novel, and for strain M1 86% (161 out of 187 successfully end-sequenced) had more than 85% nucleotide id to strain NCTC 11168 whereas 14% were novel.

The sequence reads from pUC clones identified as containing end-sequences with less than 85% nucleotide id to NCTC 11168, from the false negative testing screen, were compared to the differential hybridization results. In strain 81-176 only 3 clones out of the 20 identified as novel (15%) had not been identified in the differential hybridization screen. Of these 3 clones only one contained DNA that had not already been sequenced from other pUC clones within the library. On closer inspection of the sequence from this clone it was apparent that although the overall sequence was 80% similar to *chuA* there were regions of higher similarity within that. This suggests that the method is unsuitable for reliably picking up small variations in sequence similarity and that using a library with 5-fold coverage of the genome partially compensates for variable hybridization of individual clones. For strain M1, 10 clones out of the 26 identified as novel (38%) had not been identified by the hybridization screen. Of these clones 5 contained DNA already sequenced leaving 19% of novel DNA not previously identified by sequencing other library clones.

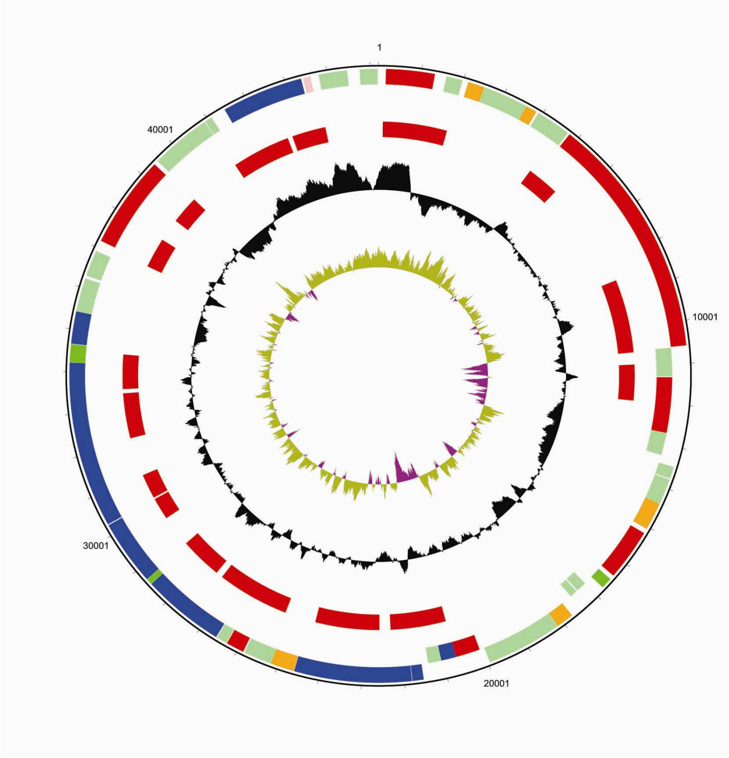
Only 11% of randomly selected library clones from strain 81-176 contained novel sequence whereas 14% of randomly selected library clones from strain M1 contained novel sequence. Using differential hybridization data to select clones for sequencing, 83% of clones from strain 81-176 and 66% of clones from strain M1 contained novel sequence. This technique therefore provides a good enrichment; however, it is estimated that around 20% of novel DNA will be missed using this method alone. Further sequencing of BAC clones encompassing novel regions should identify more sequence in these selected areas.

4.2.3 Strain 81-176 clones with matches to pVir and pTet

Of the 81-176 sequenced clones, 86 reads out of 654 matched to pVir, covering 37473 bp (50%), and 95 reads matched to pTet, covering 20413 bp (45%) of the plasmid sequence using WUBLASTN with an 85% nucleotide id cut-off. There appeared to be some distribution bias with some regions of the plasmid receiving heavier coverage than others and some regions devoid of matches entirely (**Fig 4.2**). The regions that were not covered by the differential hybridization screen were also absent from the initial plasmid shotgun assembly, possibly indicating that these regions are refractory to cloning. This could be for a number of reasons: these regions could contain products that are toxic to the *Escherichia coli* host or contain products that interfere with normal replication. For example, in pVir the regions surrounding the putative partition gene were not sampled. In pTet both the putative origin of replication and the putative origin of transfer were not sampled. Also the region between pVir8-pVir 19 was not sampled. This has low G+C content at 22%, and as the DNA was sonicated prior to cloning highly A+T rich regions may have been lost. It has been shown previously that the initial rate of shearing during sonication is reproducibly more rapid for A+T rich DNA [140].



A



B

Fig 4.2: A circular representation of the plasmid sequences. A: pVir, B: pTet.

The circles represent the following features, numbering from the outside in: 1, 2, all CDSs (transcribed clockwise and anticlockwise respectively); 3, the position of 81-176 pUC clone sequence reads identified from the differential hybridization screen with more than 85% similarity to the plasmid sequence; 4, G+C content; 5, GC deviation ((G-C)/ (G+C)) viewed using a window size of 250 bp, with a step size of 10 bp. The 12 o'clock position of each circle represents the predicted origin of replication and CDS colours represent the following putative functions: red, information transfer (transcription/ translation + DNA/ RNA modification); light green, unknown; dark green, surface; orange, conserved hypothetical; blue, pathogenicity/ adaptation; pink, bacteriophage/ IS elements.

The concentration of plasmid DNA compared to chromosomal DNA in the DNA preparation used to make the library was not known so a direct comparison of the likely coverage of novel regions of chromosomal DNA can not be made based on the coverage of plasmid DNA.

4.2.4 General Features of novel pUC assemblies

For each of the contiguous regions of assembled pUC reads the predicted CDSs were analysed using FASTA to search protein databases and assign putative functions (section 2.3.7) (Appendix 3, 4, 5 and 6). In some cases the contiguous regions contained several novel CDSs and also a region of high identity to strain NCTC 11168, indicating a probable insertion/substitution event compared to the NCTC 11168 genome.

For strain 81-176 the 473 reads were assembled into 58 contiguous regions representing 85,755 bp of sequence containing 108 partial or complete predicted CDSs. For strain M1 the 531 reads were assembled into 81 contigs representing 113,180 bp of sequence containing 156 predicted CDSs. Strain 40671 contains the smallest amount of novel sequence identified, with 273 reads assembled into 59 contigs representing 78,923 bp of sequence containing 100 predicted CDSs. Strain 52472 contains the largest amount of novel sequence identified, with 1341 reads assembled into 101 contigs representing 205,235 bp of sequence containing 279 predicted CDSs.

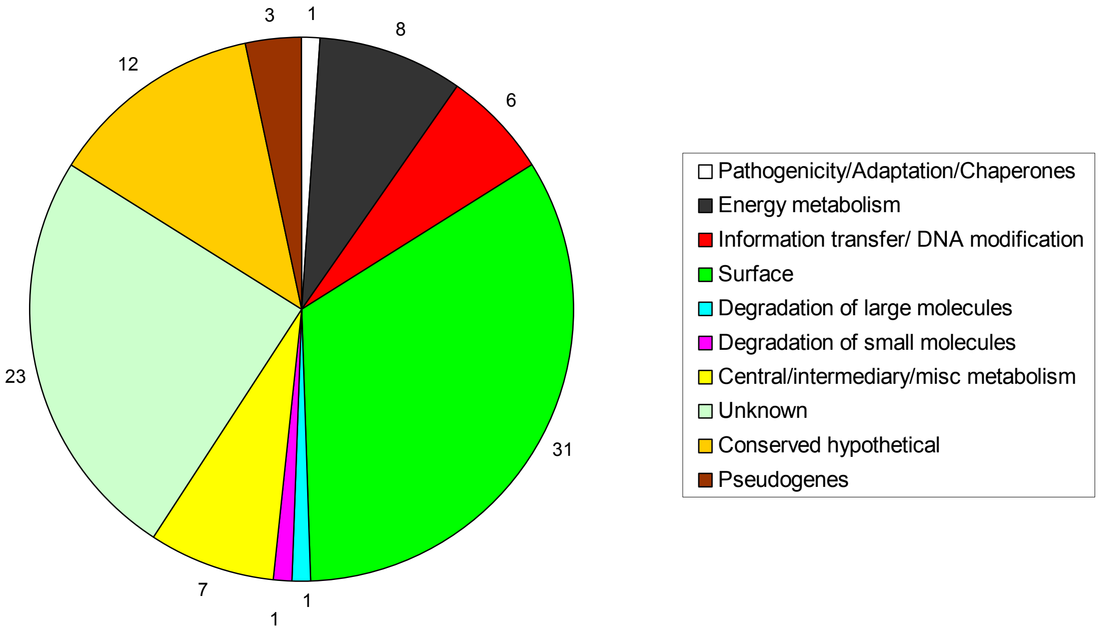
Discounting CDS matches with more than 95% amino acid id to NCTC 11168 across entire length, 93 novel genes were discovered in strain 81-176, 137 in strain M1, 97 in strain 40671 and 268 in strain 52472. The CDSs with more than 95% amino acid id were found to be either towards the ends of contiguous regions containing novel sequence or next to insertion or deletion (indel)/ rearrangement events, which may provide useful positional information.

Strain 52472 has many bacteriophage associated genes which are not present in NCTC 11168. It is possible that many of the hypothetical genes from this strain are also bacteriophage associated as, in addition to genes required for assembly, bacteriophage carry many genes with as yet undetermined function. Without more sequence information from the surrounding regions it is not possible to distinguish between these and chromosomal hypothetical genes. Bacteriophage genes tend to be less conserved and less easy to recognize by similarity searches [141]. Sequence reads assembled into contiguous regions predicted to encode bacteriophage associated proteins have a greater depth of coverage than other contiguous regions. This could indicate that similar bacteriophage are integrated at multiple sites in the genome.

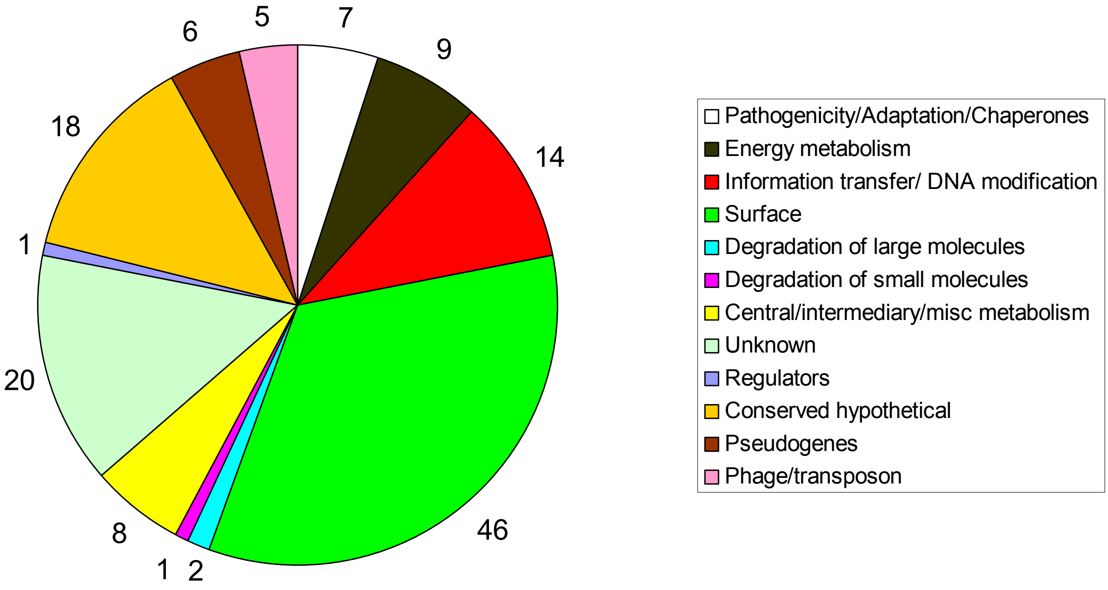
Excluding phage associated CDSs in strain 52472 the largest class of predicted CDSs in all strains tested are hypothetical (**Fig 4.3**). There are 35 hypothetical CDSs in strain 81-176 (38% of novel), 38 in strain M1 (28% of novel), 46 in strain 40671 (47% of novel) and 63 in strain 52472 (24% of novel). In the genome sequences of strains NCTC 11168 and RM1221 22% and 29% of predicted CDSs were classed as hypothetical. In strains 81-176 and 40671 the proportion of novel CDSs classified as hypothetical is greater than that identified for the chromosomal background of the sequenced strains. In strain M1 there are actually more predicted surface associated CDSs (46, 34%) than hypothetical CDSs (38, 28%). Surface associated CDSs probably make up the second largest category overall with 31 in strain 81-176 (33% of novel), 22 in strain 40671 (23% of novel) and 24 in strain 52472 (9% of novel). Another major category for all strains is information transfer/DNA modification which includes restriction-modification (RM) associated CDSs: 6 in strain 81-176, 14 in strain M1, 9 in strain 40671 and 19 in strain 52472. There are also some predicted

CDSs associated with general metabolism: 7 in strain 81-176, 8 in strain M1, 3 in strain 40671 and 13 in strain 52472. The rest of the categories appear to vary according to strain: strains 81-176 and M1 contain several predicted CDSs associated with energy metabolism, and strains 40671 and 52472 contain many CDSs associated with pathogenicity and adaptation.

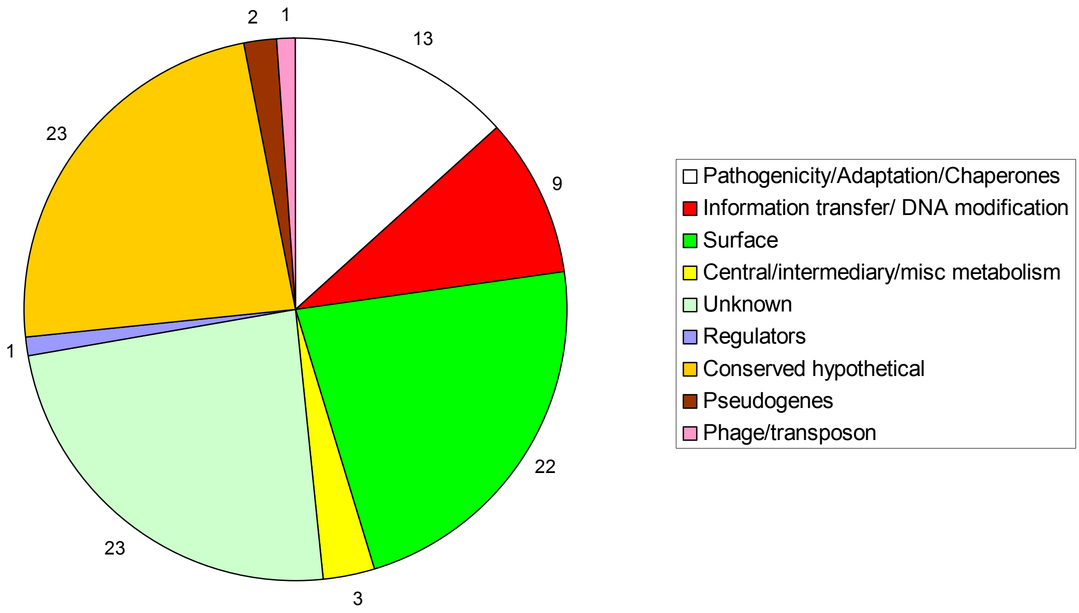
A. Functional categories of predicted CDSs in 81-176 pUC assemblies

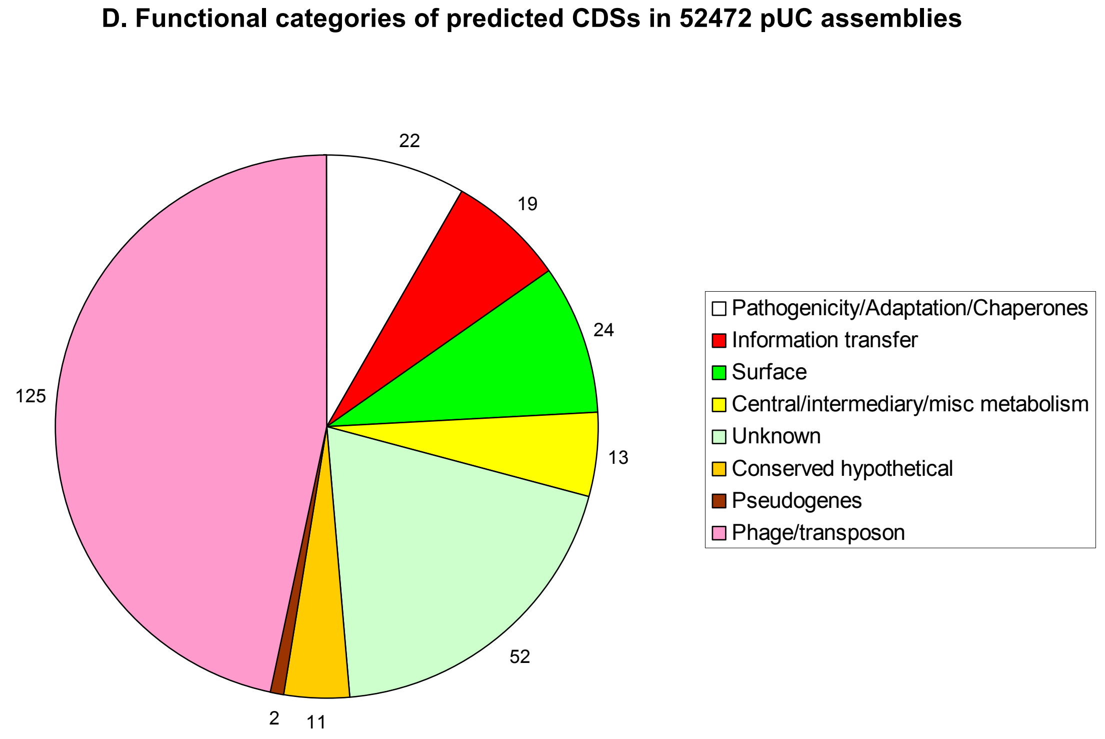


B. Functional categories of predicted CDSs in M1 pUC assemblies



C. Functional categories of predicted CDSs in 40671 pUC assemblies





D. Functional categories of predicted CDSs in 52472 pUC assemblies

Fig 4.3: Pie-chart diagrams depicting functional categories for novel predicted CDSs. Functional categories are described in the keys beside each chart. Certain categories make up a large proportion of the novel CDSs of all strains: DNA modification; Surface; unknown; conserved hypothetical.

From the FASTA analysis it is also apparent that there are a number of CDSs with some identity to CDSs from strain NCTC 11168 rather than being novel genes. This category is made up of genes with between 65 and 95% amino acid id to CDSs in strain NCTC 11168. In 81-176 this category accounts for 30% (28) of the novel CDSs, in M1 20% (26), in 40671 13% (13) and in 52472 9% (24). However, the numbers of CDSs with 65-95% amino acid id to NCTC 11168 in strains 81-176, M1 and 52472 are similar so this actually reflects the fact that strain 81-176 has a smaller proportion of novel CDSs compared to the rest and that strains 40671 and 52472 have a larger proportion of novel CDSs. The vast majority of CDSs with 65-95% amino acid id to CDSs in strain NCTC 11168 are surface associated with some of the rest being associated with metabolism and some hypothetical.

In the strain NCTC 11168 genome sequence CDSs are numbered from cj0001 to cj1731 with CDSs on the complementary strand suffixed with a c. In this study CDSs have been numbered sequentially with a strain identifier of 8 for strain 81-176, M for strain M1, 4 for strain 40671 and 5 for strain 52472, followed by a P for pUC library. Contiguous regions have been named similarly with a strain identifier, a library identifier followed by the read name from the first sequence read of that contiguous region. Data for the predicted novel CDSs are presented in Appendix 3, 4, 5 and 6.

4.2.5 Regions showing limited identity to genes in NCTC 11168

4.2.5.1 Surface associated

4.2.5.1.1 N-linked glycosylation locus cj1119c-cj1130c

The predicted CDSs MP0007 and MP0008 (MP2f03q) show 97% id to WlaI and WlaK showing that WlaJ is missing in this location in strain M1 as it is in strain 81116 [142]and RM1221. This region has been shown to be highly conserved between several strains [37].

4.2.5.1.2 Lipo-oligosaccharide biosynthesis locus cj1131-1152

The lipo-oligosaccharide (LOS) gene cluster of *C. jejuni* is one of the most highly studied regions within this bacterium and has been demonstrated to be highly variable between strains. DNA sequences from the LOS region of 11 *C. jejuni* strains were compared by Gilbert *et al*. [143] and assigned to one of 3 classes, A, B or C.

In strain M1 all the predicted CDSs from the LOS region show highest identity to CDSs from strain 81116, also known as strain NCTC 11828 [144], which does not fall into the A, B or C class system (**Fig 4.4**). The predicted CDS MP0121 (MP2b12p) shows high identity to WlaNA and the partial CDSs MP0120 and MP0029 (MP1f05p) both show high identity to WlaNB. CDS MP0028 (MP1f05p) shows high identity to a transferase, RlmA, and CDS MP0010 (MP5b05p) shows 59% id to a DTPT dehydratase from *H. hepaticus*. Predicted CDS MP0051 (MP3d07q) shows high identity to a hypothetical CDS, MP0052 to an aminotransferase and MP0053 to a membrane protein. Predicted CDSs MP0015 and MP0016 (MP3b05q) show high identity to two glycosyltransferases from strain 81116 with MP0016 showing 63% id to CgtA (Cj1138) from strain NCTC 11168. The predicted CDSs MP0034 (MP2f12q) and MP0019 (MP3d02q) show high identity to an O-acetylation protein and MP0020 (MP3d02q) to a hypothetical protein which are inserted upstream of *gmhA* (MP0021). This arrangement is present in class B1 LOS clusters [59].

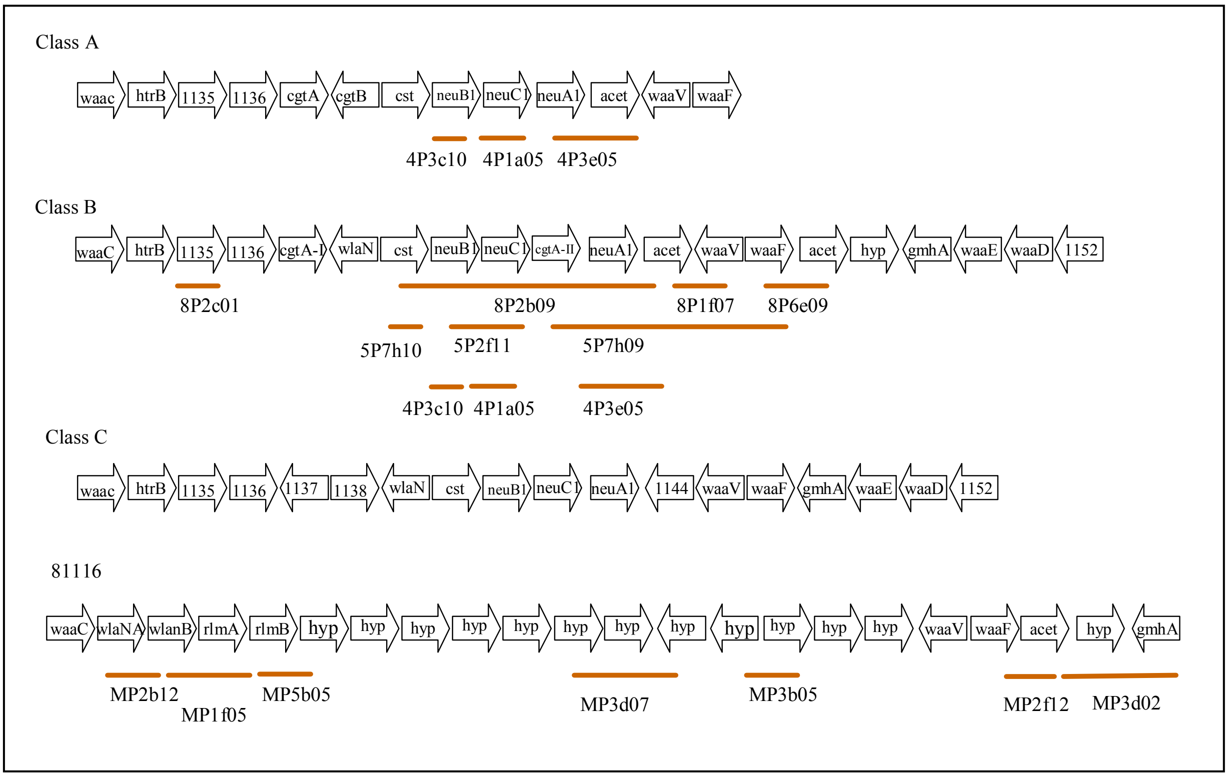


Fig 4.4: **Schematic representation of the genes from different LOS classes** of *C. jejuni* according to Gilbert *et al.* 2002 [143]. Arrow sizes are not representative of gene or protein size. Arrows are labelled with predicted protein products; numbers represent CDS locus id numbers from *C. jejuni* NCTC 11168; hyp= hypothetical protein; acet= acetyltransferase. Contiguous regions from this study are represented by lines underneath the arrows and are labelled with their contig identifiers. Strains ATCC43446, OH4384, OH4382, ATCC43432, ATCC4360 and ATCC43438 belong to class A; strains ATCC43449 and ATCC43456 belong to class B; strains NCTC 11168, ATCC43429 and ATCC43430 belong to class C. Data for the LOS locus of strain 81116 was obtained from NCBI, accession numbers AJ131360 and AF343914 (Oldfield *et al.* 2002) [144].

In strain 81-176 the predicted CDS, 8P0099 (8P2c01q) shows 74% id to glycosyltransferase Cj1135 but 100% id to a glycosyltransferase from strain 43456 (**Fig 4.4**). In strain 52472 the predicted CDS, 5P0268 (5P7h10p) shows 95% id to Cst-II from strain 43432 and in strain 81-176 the predicted CDS, 8P0063 matches to Cst-II previously sequenced from strain 81-176. For strain 81-176 8P0064 NeuB1, 8P0065 NeuC1, 8P0066 CgtA-II, 8P0067 NeuA1 and partial CDSs 8P0068 and 8P0105 acetyltransferase (8P2b09p) show high amino acid id and a similar arrangement of the genes that encode them to strain 43456 which belongs to LOS cluster class B [143;145] (**Fig 4.4**). This arrangement of genes seems to be shared in strain 40671 with CDS 4P0010 (4P3c10p) predicted to encode NeuB1 [146], 4P0092 (4P1a05q) predicted to encode NeuC1, 4P0050 (4P3e05q) predicted to encode NeuA1 and 4P0049 predicted to encode an acetyltransferase [145]. This LOS arrangement also appears to be present in strain 52472: 5P0109 (5P2f11q) is predicted to encode NeuB1, 5P0110 predicted to encode NeuC1, 5P0130 (5P7h09p) predicted to encode CgtA-II, 5P0131 predicted to encode NeuA1 and 5P0132 predicted to encode an acetyltransferase.

In strain 52472 the predicted CDS 5P0133 shows high amino acid id to WaaV from strain lio87 and CDS 5P0134 shows high identity to WaaF from strain 81116; both predicted CDSs also show high identity to WaaF from RM1221. In strain 81-176 the predicted CDS 8P0104 (8P1f07q) shows high identity to WaaV from strain 43456 [143], 8P0006 (8P6e09q) shows high identity to WaaF and 8P0005 high identity to a hypothetical CDS from regions previously sequenced in strain 81-176 [147].

4.2.5.1.3 Flagellar associated genes

In strains 81-176 and 40671 the homologues of cj0043 encoding the flagellar hook protein FlgE appear to be variable. The predicted CDS 8P0054 (8P4a03p) shows high identity to FlgE previously sequenced from strain 81-176 and predicted CDS 4P0012 (4P1c07p) shows homology to FlgE from strain lio7 and but these two predicted CDSs, 8P0054 and 4P0012, do not share high identity with each other [138].

In strain 81-176 the predicted CDS 8P0010 (8P7e10q) shows 89% amino acid id to the putative aminotransferase Cj1294. In strain M1 MP0135 (MP4h07p) shows 89% id to the hypothetical protein Cj1295 and MP0136 appears to be a fusion of the genes predicted to encode aminoglycoside N3`-acetyltransferases Cj1296 and Cj1297 with 79% and 56% id to each respectively although MP0136 appears to be more similar to RM1221 CJE1488. In strain NCTC 11168 these proteins have a homopolymeric tract between them so they can be translated as a single gene if a frame shift occurs due to slip-strand mispairing. In strain 40671 4P0098 (4P1g08p) shows 76% id to the hypothetical protein Cj1305. In strain 81-176 8P0041 (8P3e08q) shows high identity (97%) to NeuA2, involved in biosynthesis of glycosyl moieties [148], but 8P0040 only shows 62% id to the hypothetical protein Cj1310.

Parts of the flagellar cluster have previously been sequenced in strain 81-176. The following CDSs match to these previously sequenced genes; 8P0015 (8P7d11q) matches to an orthologue of Cj1333 and 8P0045 (8P1c09q) matches to an orthologue of Cj1337 [148]. Both strains M1 and 40671 appear more similar to 81-176 than NCTC 11168 in this region. For example, in strain M1 MP0040 (MP3b03q) shows 59% id to Cj1334 from strain NCTC 11168 and 76% to an orthologue of Cj1334 from strain 81-176. In strain 40671 4P0062 (4P3f10p) shows 73% id to Cj1334 from strain NCTC 11168 and 95% id to an orthologue of Cj1334 from strain 81-176 [148]. In strain M1 partial CDS MP0024 (MP3e04p) shows 57% id to Cj1337 from strain NCTC 11168 and 99.8% id to an orthologue of Cj1337 from strain 81-176. In strain 40671 (4P1g09q) 4P0070 shows 61% id to Cj1337 from strain NCTC 11168. In strain 81-176 partial CDSs 8P0044 (8P1c09q) and 8P0069 (8P6a11p), and in strain M1 MP0064 (MP1b10q) show high id to FlaB from 81116 [148-150]. In strain 40671 4P0069 shows 91% amino acid id to FlaB from *C. coli*.

In strain 81-176 8P0027 (8P8b05p) shows 100% id to FlaA from strain d2677 and in strain M1 MP0018 (MP4a03q) shows 100% id to FlaA from strain 81116. In strain 40671 4P0025 (4P1f06p) shows 71.5 % id to FlaA from strain NCTC 11168. 8P0026 (8P3h05p) shows 34% id to hypothetical protein Cj1340 [149;150] and 8P0096 shows 78% id to hypothetical protein Cj1342.

The only CDSs in strain 52472 that match to this region are 5P0087 (5P5a07q) which shows 94% id to hypothetical protein Cj1341 and 5P0086 which shows 60% id to hypothetical protein Cj1342. This suggests that the flagellar region of this strain is much more similar to that of strain NCTC 11168 than the other strains.

4.2.5.1.4 Capsule locus cj1413-1448

It has recently been demonstrated that the capsule region is highly variable with many genes being acquired by horizontal transfer along with gene duplications, deletions and fusions [151]. Due to the extensive variation it is likely that some of the identified novel surface associated genes discovered here may be part of the capsule locus but without further sequence information it is not possible to identify where exactly they belong on the chromosome. Relatively few predicted genes can be linked to the capsule. Where this is possible, most of the CDSs are from strain 81-176 as this has already been sequenced in its entirety [151]. The capsule sequence became available after the annotation of the novel 81-176 regions in this study therefore a comparison using WUBLASTN revealed more matches to this area (**Fig 4.5**). 8P0036 shows 63% id to Cj1442 and 8P0037 shows 96% id to KpsF of strain NCTC 11168 (8P5a10q). Other predicted CDSs match to this region: 8P0043 (8P1e08q) which shows 78% id to DmhA of *Yersinia pseudotuberculosis*, predicted to be involved in the conversion of heptose to deoxyheptose, 8P0001 (8P5c06p) which shows 56% id to Fcl of NCTC 11168 and 8P0028 which shows 41% id to Cst-I from strain oh4384, which has been associated with GBS [146], and 8P0029 (8P2d02p) which shows 41% id to Cj1431 from strain NCTC 11168 [151].

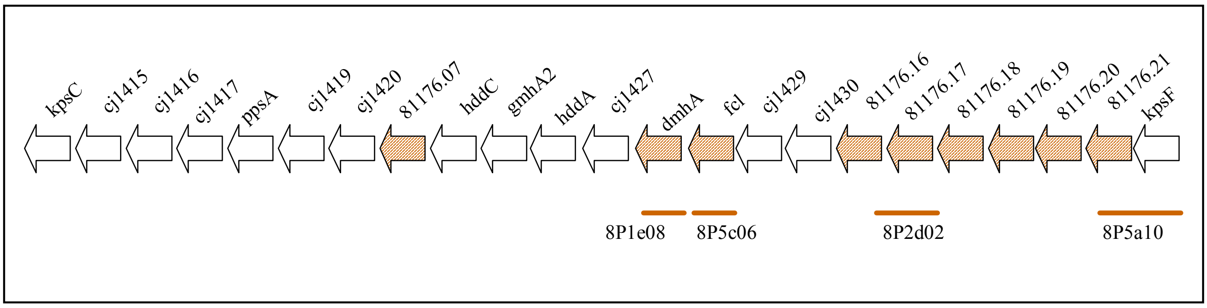


Fig 4.5: Capsule locus of *C. jejuni* strain 81-176. CDSs from the capsule locus of strain 81-176 determined by Karlyshev *et al.* 2005 [151] (accession number BX545858) are depicted by arrows. The size of the arrows is not representative of gene or protein size. Contiguous regions from this study are represented by lines underneath the arrows and are labelled with their contig identifiers. Stripped arrows represent genes that are novel/ significantly divergent from strain NCTC 11168.

Strain 40671 also has some predicted CDSs which may be associated with the capsule region. 4P0063 matches 8P0043 and also shows 78% id to DmhA from *Y. pseudotuberculosis* (4P1a10p). Also present on this contiguous region are the predicted CDSs 4P0064 which shows 59% id to Fcl from strain NCTC 11168, 4P0065 which shows 81% id to the sugar epimerase Cj1430 and 4P0066 which shows 37% id to the sugar transferase Cj1421.

Also potentially located in the capsular region are predicted CDSs 4P0058 (4P1d02q) which shows 69% id to Cj1421c from strain NCTC 11168, 4P0059 which shows 58% id to Cst-I 58% from strain oh4384 and 4P0007 (4P3f04p) which shows 50% id to the sugar transferase Cj1440 and 4P0008 which shows 84% id to Cj1421.

4.2.5.1.5 Miscellaneous

There are a number of predicted CDSs which are similar at the amino acid level to NCTC 11168 proteins but are only based on single read coverage which may mean that sequencing errors are the cause of any observed variation rather than true variation. This applies to strain 52472 predicted CDS 5P0269 which shows 90% amino acid id to the periplasmic protein Cj0168 (5P5h05p) although 5P0269 seems to be more similar to RM1221 CJE0163. In strain M1 MP0134 shows 86% amino acid id to membrane protein Cj0692c (MP3e02q) and MP0129 shows 88% amino acid id to membrane protein Cj1049 (MP2e10p). In strain 40671 the predicted CDS 4P0091 shows 88% amino acid id to CfrA (Cj0755) (4P1a06p).

There are also a number of predicted CDSs which are similar at the amino acid level to CDSs from strain NCTC 11168 but have higher sequence coverage than those discussed above. In strain 52472 the predicted CDS 5P0074 shows 76% amino acid id to lipoprotein Cj0629 with a large gap in the centre of the match (5P6f05q). In strain M1 the predicted CDS MP0013 shows 69% amino acid id to the membrane protein PorA (Cj1259) (MP4e02q) and a higher identity to strain X7199 at 88% amino acid id. CDS MP0035 shows 91% id to ChuA (Cj1614) (MP3e06q). In strain 81-176 the predicted CDS 8P0035 (8P6e04q) shows 92% id to the membrane associated protein Cj0835.

In strains M1, 81-176 and 40671 the predicted CDS MP0139 (MP5h05p) shows 65% and the predicted CDSs 8P0021 (8P5a05p) and 4P0016 (4P1d05p) show 64% id to the membrane protein Cj1721 although all are similar to each other and also to RM1221 CJE1891.

4.2.5.2 Metabolism

4.2.5.2.1 Molybdate transport region cj0294-cj0310

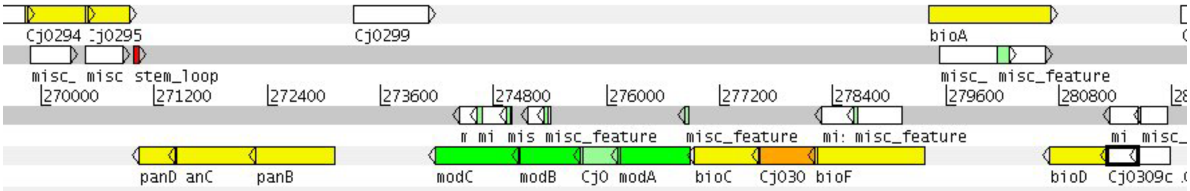


Fig 4.6: Molybdate transport region of *C. jejuni* strain NCTC 11168. The region from cj0294-cj0310 is viewed using Artemis. The forward and reverse DNA lines are represented by the central dark grey lines. The light grey lines represent all three forward and reverse translated reading frames respectively. Open boxes represent features: pFam and prosite features are shown on the DNA lines; CDSs are shown on the frame lines. CDSs are coloured to indicate functional category: white, pathogenicity/ adaptation/ chaperones; yellow, central/ intermediary/ miscellaneous metabolism; dark green, surface; light green, unknown; orange, conserved hypothetical.

cj0294-cj0310 is known to be a region with limited identity between strains [83-85]. In strain 81-176 the contiguous region 8P7b08p contains predicted CDSs 8P0051, 8P0052 and 8P0053 with high amino acid identity to cj0294, cj0296 and cj0297 respectively showing that a homologue of the predicted acetyltransferase cj0295 is missing in this location in strain 81-176. ModA-C (**Fig 4.6**) are variable in both strains 81-176 and 52472 with 8P0060 and 5P0075 showing 82% amino acid id to ModA, 8P0058 and 5P0077 showing 85% amino acid id to ModB, and 8P0057 and 5P0078 showing 76% and 78% amino acid id to ModC from strain NCTC 11168 respectively. The predicted CDSs 8P0059 and 5P0076 appear to be the most divergent showing only 65% amino acid id to the hypothetical protein Cj0302 (8P4e04p, 5P4e07q). 5P0122 appears to be variable showing 74% id to BioC, 8P0046 and 5P0123 show 67% and 68% id to Cj0305 respectively and 8P0047 and 5P0124 both show 76% id to BioF (8P3b10q, 5P6g02q). The entire region shows high identity between the strains 81-176, 52472 and also RM1221 possibly suggesting that the whole region in NCTC 11168 has been acquired by homologous recombination from a more divergent source.

4.2.5.2.2 Region cj0807-cj0813

In strain 52472 the region homologous to cj0807-cj0813 appears to be highly variable. The predicted CDS 5P0006 (5P2g09p) shows 77% amino acid id to the hypothetical protein Cj0808 and the partial CDSs 5P0007 (5P2g09p) and 5P0089 (5P6c03q) show 90% and 78% amino acid id respectively to the hydrolase Cj0809. The predicted CDS 5P0090 (5P6c03q) shows 74.1% amino acid id to the NH(3)-dependent NAD(+) synthetase NadE (Cj0810) and the partial CDSs 5P0091 (5P6c03q) and 5P0029 (5P3a03q) show 82% and 84% id to the tetraacyldisaccharide 4`-kinase LpxK (Cj0811). The partial CDSs 5P0028 (5P3a03q) and 5P0048 (5P7a07p) show 78% and 75% amino acid id respectively to threonine synthase ThrC (Cj0812) and 5P0047 (5P7a07p) shows 83% amino acid id to KdsB (Cj0813).

4.2.5.2.3 Miscellaneous

In strain 52472 there are predicted CDSs that show limited amino acid id to the proteins Cj0021-Cj0023 from strain NCTC 11168. On contiguous region 5P7d08q predicted CDS 5P0117c shows 86% amino acid id to the hypothetical protein Cj0021 of strain NCTC 11168, 5P0118c shows 82% amino acid id to the ribosomal pseudouridine synthase protein Cj0022 and 5P0119c shows 94% amino acid id to PurB (Cj0023). This contiguous region is constructed from 12 reads across 2.3 Kb giving a good depth of coverage so poor sequence quality is unlikely to account for the amino acid differences.

In both strains 81-176 and M1 there are predicted CDSs which show limited amino acid id to a cytoplasmic L-asparaginase, AnsA. CDS 8P0103 shows 83.46% amino acid id and MP0110c shows 86.13% amino acid id to AnsA from NCTC 11168.

In strain 81-176 a predicted CDS, 8P0092, with homology to *purU* (cj0789) from strain NCTC 11168 appears to be shorter than in NCTC 11168. The predicted CDS 8P0092 is 158 aa long compared to Cj0789 which is 274 aa in NCTC 11168. It is not possible to say whether this gene would still be functional or whether a duplication event may have occurred and a full length copy of the gene is present elsewhere on the chromosome. The predicted CDS 8P0092 is located on contiguous region 8P7g11p and a rearrangement compared to NCTC 11168 seems to have occurred with a predicted CDS with high identity to the iron uptake transporter, cj0173c, occurring upstream of the *purU* homologue.

4.2.5.3 Hypothetical genes

There are many examples of hypothetical proteins that vary between the strains being studied that have already been discussed in the context of other regions. However, there are some examples of hypothetical genes varying at other locations on the chromosome that have been identified in this study. In strain 81-176 the hypothetical protein Cj0403 appears longer than in NCTC 11168 with 8P0004 (8P8b03p) being 232 aa long and Cj0403 being only 181 aa. In strain M1 the predicted CDS MP0027 (MP1g01q) shows 91% id to the hypothetical protein Cj1178.

4.2.5.4 Pseudogenes

One striking feature apparent from the pUC assemblies is the variability among predicted pseudogenes and their surrounding genes from those in strain NCTC 11168. In strain 81-176 the predicted CDS 8P0013 appears to be a fusion of the genes encoding the small hypothetical proteins Cj1158-Cj1160 (8P7g05p) and on the same contiguous region 8P0012 shows 83% id to the membrane protein Cj1161. Also in strain 81-176 the intact CDS 8P0108 (8P7e09p) shows similarity at the nucleotide level to the pseudogene Cj0742 and shows 32% id to an afimbrial adhesin from *Escherichia coli*.

In both strains 81-176 and M1 there is variation around the arylsulfatase pseudogene Cj0866. The CDSs 8P0107 (8P6a06p) and MP0036 (MP4f03q) show high identity to the previously characterized arylsulfatase protein from 81-176 [152]. In addition several of the CDSs surrounding the arylsulfatase pseudogene in NCTC 11168 appear to vary in strains 81-

176 and M1. The periplasmic protein Cj0864 appears variable in strain M1 with MP0154 (MP4d12p) showing 92% id across part to the protein in strain NCTC 11168 but appearing more similar to RM1221 CJE0951. In strain 81-176 predicted CDS 8P0011c shows 48% identity to DsbA across the entire length but also shows 100% id to parts of Cj0864 with a 101 aa insert between aa 43 and 44. Although there is little overlap between MP0154 and 8P0011 both show high identity to RM1221 CJE0951.

Also in both strains 81-176 and M1 the region between cj0967-cj0975 appears variable. In strain 81-176 8P0042 appears to be a fusion of the genes encoding hypothetical proteins Cj0970-973 (8P2e09q). MP0141 (MP4c04p) is a pseudogene showing 96% id across part of the full length periplasmic protein Cj0967. Also present on the same contiguous region is MP0142 which shows 36% id to a hemagglutinin-related protein from the 2.1 Mb mega-plasmid of *Ralstonia solanacearum*. MP0143 (MP2g07q) appears to be a fusion of the genes encoding Cj0970-Cj0973 showing between 56% and 95% id to each individual protein and MP0144 shows 97% id across part to putative secretion protein Cj0975.

In strain M1 MP0132 (MP3b01p) shows 43% id to the secreted protease EspC from *Escherichia coli*. This may represent an intact version of the pseudogene cj0223 as this predicted CDS is present downstream of *argC* and shows 96% nucleotide id to strain NCTC 11168.

In both strains M1 and 52472 MP0155 (MP4e06p) and 5P0277 (5P5g10q) show 33% and 40% id respectively to a PrpD protein homologue from *Bradyrhizobium japonicum* that may be required for propionate catabolism. These predicted CDSs are located downstream of cj1394 and may represent a functional version of the pseudogene cj1395. It also appears that this gene is complete in RM1221 (CJE1583).

In strain 52472 5P0052 (5P5c07q) shows 51% id to a glycerol-3-phosphate transporter from *Escherichia coli*. As this is present next to *surE* it may be that the N-terminus of this transporter, which is present in a different reading frame in strain NCTC 11168 (cj0191), is present in the same frame in this strain.

4.2.6 Predicted CDSs shared between test strains but absent from NCTC 11168

It was decided to investigate how many of the novel genes were present in multiple test strains. Using a combination of WUBLASTN and reciprocal FASTA it was possible to assess the distribution of the CDSs and partial CDSs identified so far. As some of the predicted CDSs are only partial it is possible that more CDSs are shared between strains but the overlap between them is not large enough to be able to ascertain with confidence whether these genes are present in more than one strain. The results are presented in **Fig 4.7**.

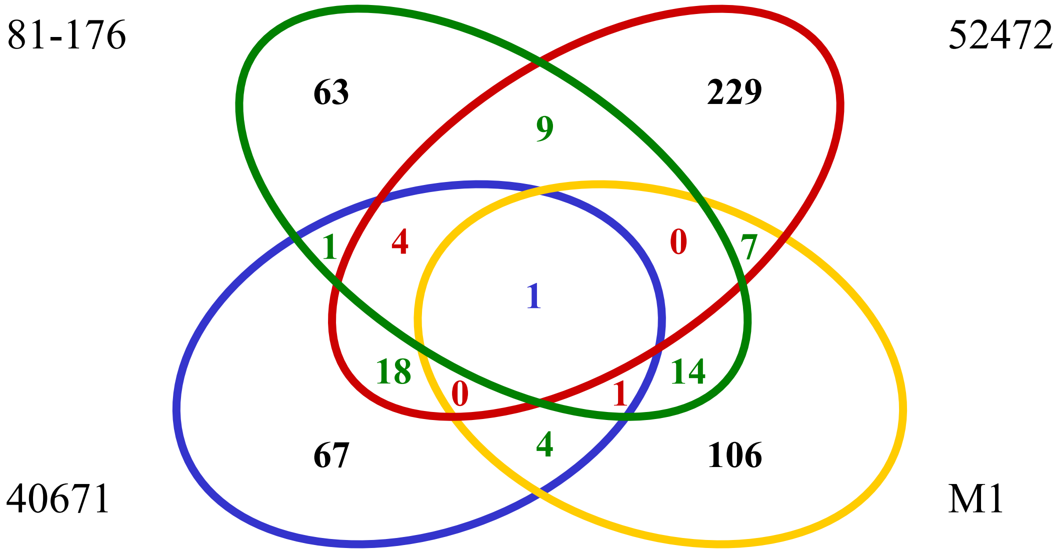


Fig 4.7: Venn diagram of predicted novel CDSs shared between strains in this study. The green ellipse represents novel CDSs identified in strain 81-176; the red ellipse represents novel CDSs identified in strain 52472; the yellow ellipse represents novel CDSs identified in strain M1 and the blue ellipse represents novel CDSs identified in strain 40671. Numbers in black represent predicted

CDSs unique to each strain; numbers in green represent CDSs shared between two strains; numbers in red represent CDSs shared between three strains and the number in blue represents a CDS shared between all strains. From this initial screen it appears that the majority of novel CDSs are unique to each strain.

4.2.6.1 CDSs present in all test strains

There is one CDS that had homologues present in all the test strains and also in RM1221 (CJE0757), and which putatively encodes a di-/tripeptide transporter. In strains 81-176, M1, 40671 and 52472 the predicted CDSs 8P0055 (8P6g02p), MP0038 (MP1a12p), 4P0078 (4P3e03p), and the partial CDSs 5P0055 (5P5f02p) and 5P0002 (5P6e05q) show 48%, 47%, 53%, 50% and 45% id to a di-tripeptide ABC transporter from *Photorhabdus luminescens* respectively. The CDSs 8P0055, MP0038, 4P0078 and 5P0055 all show over 90% id to each other.

There appear to be two transporters located on the same contiguous regions of DNA in strains 81-176, M1 and 52472. In strain 81-176 8P0056 shows 48% id to an ABC transporter from *Photorhabdus luminescens*, in strains M1 MP0039 and 52472 5P0054 show 44% and 48% id to a transporter from *Y. pseudotuberculosis* respectively. The sequence of strain 40671 does not extend this far so it is not possible to tell whether a homologous CDS is present in this strain or not.

4.2.6.2 CDSs present in three test strains

4.2.6.2.1 Shared between 81-176, M1 and 40671

The variable membrane protein Cj1721 had homologues in strains 81-176 (8P0021), M1 (MP0139), 40671 (4P0016) and also RM1221 (section 4.2.5.1.5).

4.2.6.2.2 Shared between 81-176, 40671, 52472

There are four predicted CDSs that had homologues present in strains 81-176, 40671 and 52472. These are all associated with the LOS biosynthesis cluster and encode the proteins NeuB1 (8P0064, 4P0010, 5P0109), NeuC1 (8P0065, 4P0092, 5P0110), NeuA1 (8P0067, 4P0050, 5P0131) and an acetyltransferase (8P0068, 4P0049, 5P0132) as discussed in section 4.2.5.1.2.

4.2.6.3 CDSs present in two test strains

4.2.6.3.1 Shared between 81-176 and M1

There are 14 CDSs shared between strains 81-176 and M1. As previously mentioned in section 4.2.4.4 8P0107 and MP0036 are predicted to encode an arylsulfatase production protein previously identified in 81-176 [152]. Also already mentioned in section 4.2.5.2.3, are 8P0103 and MP0110 encoding a variable AnsA protein, and mentioned in section 4.2.5.1.3 are 8P0069 (8P6a11p) and MP0064 (MP1b10q) which are predicted to encode FlaB.

Half of the shared genes are associated with respiratory chains, some of which are inserted before a cj0031 homologue in both strains, relative to strain NCTC 11168. The predicted CDSs 8P0083 (8P7d05p) and MP0108 (MP4d08p) show 62% and 54% id to Cj0031 respectively. Also present on the same contiguous regions of DNA are the predicted CDSs 8P0082 and MP0107 which show 67% and 68% id to a gamma-glutamyltranspeptidase (GGT) from *Helicobacter pylori*, and the predicted CDSs 8P0081 and MP0106 which show 59% and 54% id to a cytochrome C biogenesis protein from *Wolinella succinogenes*. More predicted CDSs with homology to cytochrome C biogenesis proteins are present with the predicted CDS 8P0084 (8P4b02p) and the predicted partial CDSs MP0050 (MP1h01q) and MP0089 (MP4e08q) showing 55%, 54% and 49%

respectively to a cytochrome c protein from *Shewanella oneidensis*. The predicted CDS 8P0085 (8P4b02p) and the predicted partial CDS MP0090 (MP4e08q) show 36% and 39% id to a cytochrome c protein from *Geobacter sulfurreducens* and *Shewanella oneidensis* respectively and MP0118 (MP2e03p) shows 28% id to a formate dehydrogenase protein from *Vibrio cholerae*. The predicted CDSs 8P0086 (8P4b02p) and MP0117 (MP2e03p) both show 39% id to a hypothetical protein from *W. succinogenes*. Also present on the same contiguous region in 81-176 is the predicted CDSs 8P0087 which shows 37% id to a cytochrome C protein from *Helicobacter hepaticus*.

There appear to be more respiratory associated genes located downstream of the cj1584 homologue in strain M1. The predicted CDSs MP0103 (MP1g06p) and 8P0078 (8P6g03q) both show 62% id to DmsA from *W. succinogenes*. On the same contiguous region in 81-176 there are further oxidoreductase homologues with 8P0079 and MP0069 (MP5c01p) showing 62% and 63% id to FdhB, a putative oxidoreductase, and 8P0080 and MP0068 showing 47% and 43% id to MraY, a hypothetical protein from *W. succinogenes* respectively. Also in strain M1 MP0067 (MP5c01p) shows 38% id to a hypothetical protein from *W. succinogenes* although this is not present in the 81-176 assembly. None of these putative respiratory chain associated proteins have homologues in strain RM1221.

In addition to respiratory associated proteins there are also some hypothetical proteins that are shared, for example MP116 (MP2d03p) shows 51% id to a hypothetical protein from *Helicobacter hepaticus* but also matches 8P0098c (8P1a12p) which shows 41% id to LpsA from *V. parahaemolyticus* and matches RM1221 CJE1884. The predicted CDSs MP0066 (MP3b09p) and 8P0039 (8P2h05p) show 24% and 23% id to a hypothetical protein from *Fusobacterium nucleatum* respectively.

4.2.6.3.2 Shared between 81-176 and 40671

There is only one predicted CDS present in both strains 81-176 (8P0043) and 40671 (4P0063) which is located in the capsule region and shows homology to the DmhA protein from *Y. pseudotuberculosis* as discussed in section 4.2.5.1.4.

4.2.6.3.3 Shared between 81-176 and 52472

There are nine predicted CDSs that are shared between strains 81-176 and 52472, seven of which are located in the molybdate transport region. The proteins ModC (8P0057, 5P0078), ModB (8P0058, 5P0077), Cj0302 (8P0059, 5P0076), ModA (8P0060, 5P0075), Cj0305 (8P0046, 5P0123), BioF (8P0047, 5P0124) and BioA (8P0048, 5P0125) are highly similar in both strains as well as RM1221 as discussed in section 4.2.5.2.1.

Another predicted CDS with high identity between the two strains is 8P0060 and 5P0130 encoding CgtA-II in the LOS biosynthesis cluster discussed in section 4.2.5.1.2. Also showing high identity between the two strains is 8P0050 (8P1b01p) and 5P0083 (5P8g09p) which show homology to HsdM from *V. cholerae*. In strain 52472 this appears to be inserted next to DnaK but in 81-176 there is no positional information although there is an adjacent predicted CDS, 8P0049, present on the same contiguous region which shows 37% id to a type I restriction modification protein from *Methanosarcina mazei*.

4.2.6.3.4 Shared between M1 and 40671

There are four predicted CDSs that are shared between strains M1 and 40671 which are all associated with restriction modification (RM) systems. In strain 40671 there appears to be a novel region inserted downstream of cj1047. Predicted CDS 4P0046 (4P1g12p) shows 39% id to a type I RM protein from *Archaeoglobus fulgidus* which appears to be present in two pieces in M1; MP0048 (MP2g01p) which shows 46% id to a type I RM protein from *Archaeoglobus fulgidus* and MP0049 which shows 33% id to a type I RM protein from *W.*

succinogenes. Also present on these respective contiguous regions are the predicted CDSs 4P0045 and MP0047 which both show 27% id to a hypothetical protein from *Shewanella oneidensis*. There is also an additional predicted CDS in strain 40671, 4P0044 that shows 47% id to a hypothetical protein from *Bacteroides thetaiotaomicron*.

There is another contiguous region associated with RM that is shared between the two strains. The predicted CDSs 4P0067 (4P3b11p) and MP0081 (MP4g01p) show 71% and 70% id to a type I RM protein from *W. succinogenes* respectively and the predicted CDSs 4P0067 and MP0080 show 46% and 45% id to a hypothetical protein and RlfA from bacteriophage P1 respectively.

4.2.6.3.5 Shared between M1 and 52472

There are seven predicted CDSs that are shared between strains M1 and 52472. A shared region that has already been discussed in section 4.2.4.4 contains MP0155 (MP4e06p) and 5P0277 (5P5g10q) which are predicted to encode proteins involved in catabolism of propionates, which are short-chain fatty acids found in the intestinal lumen [153].

Not discussed before are the predicted CDSs MP0101 (MP1d11p) and 5P0165 (5P5d09p) which are both predicted to encode TetO, and MP0100 and 5P0025 (5P4d12p) which show homology to a small hypothetical protein from a transposon. This hypothetical CDS is found adjacent to *tetO* on pTet. However, in M1 the adjacent hypothetical predicted CDS MP0099 does not show homology to pTet. This potential tetracycline resistance locus is investigated further in chapter 5.

There is also a putative phage repressor protein that appears to be shared between the two strains encoded by MP0062 (MP3c05q) and 5P0248 (5P3b03q) and also RM1221. However, the surrounding predicted CDSs are not shared: in the case of strain 52472 this CDS is part of a 15 Kb region containing many phage associated genes.

There are a few predicted CDSs associated with restriction modification systems that are shared. The predicted CDSs MP0087 (MP2c11p) and 5P0024 (5P5e08p) show high identity to the HsdM from strain RM2227 and RM1170 respectively and also to RM1221 as do the predicted hypothetical proteins MP0087 and 5P0104 (5P2e10p) [154]. Also in strain M1 on this contiguous region MP0088 shows high identity to HsdS from strain RM1163 and MP0086 shows 42% id to the decarboxylase PcaC from *M. acetivorans*. Also associated with the RM locus of RM1221 are the predicted CDSs 5P0105 (5P2e10p) and MP0017 (MP2h08p) which show identity to CJE1727 and CJE1728.

4.2.6.3.6 Shared between 40671 and 52472

There are 18 predicted CDSs which are shared between strains 40671 and 52472 which are all homologous to proteins encoded on the plasmid pTet. It is possible that these strains possess a conjugative plasmid similar to pTet as none of these regions show homology to any known chromosomally located genes. None of these show homology to proteins present in strain RM1221.

There are other matches on these contiguous regions and on other contiguous regions that are not shared between the two strains. These include 4P3a12p from strain 40671, and 5P4d02q, 5P6a01q, 5P6b02q, 5P5d09p and 5P4d12p from strain 52472. These may not have been sequenced in the respective strains or possibly different versions of the plasmid are present in the different strains.

Strain 40671 contains 18387 bp of sequence that matches to pTet representing 41% of the plasmid but contains most of the type IV secretion system genes with the exception of the *virB5* and *virD2* homologues. Strain 52472 contains 33034bp of sequence that matches to pTet representing 73% of the plasmid and contains homologues of all the type IV secretion system genes. It may also be possible that part of the plasmid is inserted on the chromosome in a similar way to the plasmid derived island of RM1221. Further work would be needed to explore this possibility.

4.2.6.4 Predicted novel CDSs present in RM1221

The complete genome sequence of *C. jejuni* strain RM1221 has recently been published [9]. Table 4.1 below shows the number of genes identified in the pUC libraries that are present in the sequence of RM1221.

Table 4.1: CDSs identified in the pUC libraries that are present in RM1221.

81-176	M1	40671	52472
21	30	4	108

4.2.6.4.1 Shared between 81-176 and RM1221

There is only one predicted CDS in strain 81-176 that shares identity with RM1221 but with none of the other strains including NCTC 11168. This predicted hypothetical CDS 8P0031 (8P8h11p) shows identity to RM1221 CJE0905 and appears to be inserted downstream of the hypothetical gene cj0121; interestingly there appears to be a hypothetical gene inserted in the same place in strain M1 although these hypothetical genes are not similar.

4.2.6.4.2 Shared between M1 and RM1221

The predicted CDS MP0145 (MP4f07p) appears to be inserted downstream of the M1 homologue of the uptake permease *ceuB* (cj1352) and shows 80% id to the hypothetical protein Cj0970. This arrangement seems to be conserved between strains M1 and RM1221.

Also conserved between the two strains is RM1221 CJE0312 and MP0030 (MP2h03q) which shows 55% id to Cj0262, a methyl-accepting chemotaxis signal transduction protein. There is also a conserved CDS associated with restriction-modification systems, RM1221 pseudogene CJE1720 and MP0071 (MP2g03q) which shows high identity to HsdR from strain 81116 and appears to be inserted before a homologue of the dehydrogenase cj1548 [154].

4.2.6.4.3 Shared between 40671 and RM1221

The hypothetical predicted CDS 4P0085c (4P1a12q) appears to be conserved between strains 40671 and RM1221 CJE0388.

4.2.6.4.4 Shared between 52472 and RM1221

Most of the predicted CDSs in strain 52472 that show homology to strain RM1221 are bacteriophage associated or hypothetical proteins. In total there are 84 bacteriophage associated CDSs that share high identity to strain RM1221 leaving 41 bacteriophage associated CDSs that are novel to strain 52472. These novel bacteriophage associated CDSs are interspersed with the matches to RM1221. There are also 23 hypothetical CDSs that show homology to strain RM1221.

There are some DNA modification associated CDSs that appear to be shared between the two strains. The predicted CDS 5P0035 (5P6a05p) shows high identity to MloA, Methylase-linked ORF, from strain 1852 and RM1221, and the predicted CDS 5P0065 (5P8c04p) shows 53% id to a type III RM protein from *Helicobacter pylori* and to RM1221. Also on contiguous region 5P8g05q there are four predicted CDSs with high identity to RM1221 CDSs CJE0255-CJE0258, including 2 hypothetical proteins (5P0069, 5P0070), an extracellular deoxyribonuclease (5P0068) and a DNA binding protein (5P0067). There is also a methyltransferase 5P0136 (5P3e03p) inserted downstream of cj0259 *pyrC* which shows identity to RM1221 CJE0310.

There are some plasmid associated genes shared between the two strains with 5P0108 (5P6a01q) putatively encoding a TraC-like protein and showing identity to RM1221 pseudogene CJE1121, and 5P0121 (5P1d01q) showing identity to a hypothetical protein and RM1221. Both 5P0108 and 5P0121 show homology to pTet. RM1221 is known to have a large insert of novel DNA predicted to be of plasmid origin [9].

The predicted CDS 5P0066 (5P5h03q) shows some homology to the autotransporter domain of VacA from *Helicobacter pylori* although this may be a pseudogene in this strain as there is a stop codon in the middle of the CDS. This region shows similarity to RM1221 at the nucleotide level although the reading frame is disrupted by several frame shifts in RM1221. In strain M1 there is also a predicted CDS MP0023 (MP2g06p) that shows homology to the autotransporter domain of VacA which appears to be inserted after cj1359 (*ppK*). However, this partial CDS is apparently intact and does not show high identity to 5P0066 possibly as they match to different areas of VacA.

4.2.7 Predicted CDSs unique to each test strain

4.2.7.1 Strain 81-176

4.2.7.1.1 Restriction modification

MP0062 (8P6d08p) shows 45% id to a type I RM system protein from *M. mazei*.

4.2.7.1.2 Hypothetical

The hypothetical CDS 8P0024 (8P3a07q) appears to be inserted upstream of an orthologue of *cj1658*, predicted to encode a membrane protein. There are also two hypothetical CDSs inserted downstream of *secY* (*cj1688*), 8P0076 and 8P0077 (8P7f11p) which show 35% and 38% id to hypothetical proteins from *Clostridium perfringens* and *Rhizobium loti* respectively. There are also many predicted CDSs that show no detectable homology to previously sequenced genes; 8P0094 (8P2h12p), 8P0008, 8P0009 (8P2a01p), 8P0101 (8P4c05q), 8P0095 (8P5e04q) and 8P0097 (8P3d09q).

4.2.7.1.3 Surface (transport)

The predicted CDS 8P0023 (8P6a01p) shows 21% id to the secretion associated protein HxuB from *Haemophilus influenzae* which appears to be inserted upstream of cj0976. 8P0002 (8P6a02q) shows 30% id to a putative adhesin from *Chromobacterium violaceum*. 8P0007 (8P6h01q) shows 39% id to a C4-dicarboxylate transporter from *V. vulnificus* this contiguous region shows 95% nucleotide id to the NCTC 11168 genome in the region of pseudogene cj1389. 8P0016-8P0019 (8P1b02p) appear to be fragments of genes predicted to encode membrane associated proteins inserted upstream of an orthologue of cj1308, including an acetyltransferase and 33% id across part of WbkC from *Brucella melitensis*.

4.2.7.1.4 Miscellaneous

8P0089 (8P6d10q) putatively encodes a novel secreted serine protease that shows 40% id to Cj1365 and is inserted between orthologues of cj1368 and cj1369. 8P0070 (8P7f02p) shows 42% id to part of TraN from *Sphingomonas aromaticivorans* and 8P0071 shows 20% id to part of TraG from *E. coli*. This region shows some homology to the TraG pseudogene identified in strain M1 (section 4.2.6.2.4) although the arrangement of open reading frames appears to be different.

4.2.7.2 Strain M1

4.2.7.2.1 Restriction modification

The predicted CDS MP0070 (MP2g03q) shows high identity to RloA from *C. jejuni* strain 1551 and the predicted CDSs MP0112 and MP0113 (MP1f03p) show high identity to HsdS and RloB from strain 81116 [154]. MP0002 (MP5d06p) shows 92% id to a type I RM protein from strain p37. In addition MP0014 (MP3f12p) shows 53% id to the endonuclease Cj0139.

4.2.7.2.2 Hypothetical

There are many unique hypothetical CDSs in strain M1. The hypothetical CDS MP0074 (MP4h06p) is inserted downstream of cj0123 and MP0056 and MP0057 (MP5h04p), which show 34% and 36% id to a hypothetical protein from *Helicobacter hepaticus,* appear to be inserted downstream of cj1223. The predicted CDSs MP0109 (MP1c08p) and MP0122 (MP5b01p) show 39% and 57% id to a hypothetical protein from *Helicobacter hepaticus* and to Cj1305 respectively. There are also seven predicted CDSs that show no significant homology to previously sequenced genes.

4.2.7.2.3 Surface

There are many predicted CDSs in M1 that show homology to surface associated proteins, far more than in any other category. There are a large number of predicted CDSs associated with sugar modification. MP0031 (MP1b09q) shows 40% id to a phosphodiesterase from *Bradyrhizobium japonicum*, MP0032 shows 28% id to a hydrolase from *Caulobacter crescentus* and MP0033 shows 36% id to an ABC transporter from *Brucella suis*. MP0041 (MP5c06p) shows 44% id to an O-antigen biosynthesis protein, WbyH and MP0042 shows 33% id to a reductase, AscF from *Y. pseudotuberculosis*. MP0043 (MP1h04q) shows 56% to an epimerase, EpsS from *Methylobacillus* and MP0044 shows 53% id to a galactopyranose mutase from *Helicobacter hepaticus*. MP0078 (MP3e01p) shows 34% id to a glucose epimerase from *Pyrococcus furiosus* and MP0079 shows 38% id to a glucose dehydrogenase from *Pyrococcus abyssi*. MP0058 (MP5d03p) shows 49% id to a glucose dehydrogenase, UgdH from *Agrobacterium tumefaciens* and MP0059 shows 68% id to UDP-glucose 4-epimerase from *F. nucleatum* 68%. MP0095 (MP3d08p) shows 28% id to the hypothetical protein Cj1431, MP0096 shows 59% id to DdhA, glucose-1-phosphate cytidylyltransferase, from *Y. enterocolitica* and MP0097 shows 60% id to a glucose dehydratase from *F. nucleatum*.

There are also other surface associated CDSs that are not predicted to be involved in sugar modification. MP0046 (MP3d04q) shows 25% id to a hypothetical protein putatively involved in adhesion from *Chromobacterium violaceum*. The rest are associated with transport systems. MP0114 and MP0115 (MP1b04q) show 36% and 57% id to two proteins associated with an ABC transporter from *Rhizobium loti*. There also appears to be a transporter inserted downstream of cj1523 with MP0151 (MP2b05p) showing 36% id to a dicarboxylate transporter from *V. vulnificus* suggesting that this may be a more complete version of the pseudogene cj1528. This contig shows 65% id at amino acid level to the pseudogene cj1528. Downstream of cj1687 there are three predicted CDSs MP0091 (MP3h01q) showing 44% id to a transport system permease from *Rhodopseudomonas palustris* and MP0092 and MP0093 showing 49% and 45% to ABC transport proteins from *Rhizobium loti* and *Agrobacterium tumefaciens* respectively.

4.2.7.2.4 Miscellaneous

There are several plasmid remnants with MP0076 (MP3a05q) showing 40% id to part of a replication protein from *Treponema denticola*, located adjacent to MP0077 which shows 46% id to TnpV, a hypothetical protein from a transposon of *Clostridium difficile*; these predicted CDSs appear to be inserted part way through a homologue of cj0770, which is predicted to encode a membrane protein, possibly denoting transposon activity. The genome of NCTC 11168 is unusual in the fact that it does not contain any bacteriophage remnants. The predicted CDS MP0119 (MP3a03p) shows 31% id to a hypothetical protein from a bacteriophage of *Salmonella enterica* Typhimurium. In addition there appears to be a pseudogene MP0104 (MP4e01q) with 21% id to TraG from *V. vulnificus* inserted upstream of cj0937, which is predicted to encode a membrane protein.

There are also various protease matches with MP0001 (MP2d02q) showing 37% id to the serine protease SigA from *Shigella flexneri*, MP0148 (MP2f07q) showing 46% id to a haemoglobin protease from *Escherichia coli* (this contiguous region shows 92% nucleotide id to cj0223) and MP0054 (MP3e11p) showing 24% id to a haemolysin from *Xanthomonas axonopodis*.

The CDSs MP0082 and MP0083 (MP1g05q) show 44% and 57% id to the oxidoreductases Cj0414 and Cj0415 respectively.

4.2.7.3 Strain 40671

4.2.7.3.1 Restriction modification

The other strains in this study appear to have many unique restriction modification associated proteins but in strain 40671 there is only 4P0090 (4P1b05p) which shows 60% id to Cj0032 a RM enzyme.

4.2.7.3.2 Hypothetical

There are, in contrast, many hypothetical predicted CDSs. There are two hypothetical proteins (4P0004 and 4P0005 4P3d01p) inserted downstream of a homologue of cj0138. 4P0031 (4P2d09p) is a pseudogene inserted upstream of a homologue of cj0121. 4P0035 (4P1b12q) is also a pseudogene showing 51% id to a hypothetical protein from *Chromobacterium violaceum*. There are 13 predicted CDSs that do not show detectable homology to any known proteins from other bacteria.

There are also some examples of predicted CDSs that show homology to hypothetical proteins from other bacteria e.g. 4P0018 (4P1h08q) shows 35% id to *Helicobacter hepaticus*, 4P0061 (4P3a10q) shows 30% id to *Helicobacter pylori* J99, 4P0020 and 4P0021 (4P1d03p) show 32% and 47% id to *W. succinogenes* and *Helicobacter pylori* J99 respectively. 4P0081 (4P1c06p) shows 50% id to *W. succinogenes*, 4P0083 (4P2c10p) shows 26% id to *Clostridium perfringens*, 4P0033 (4P3c01q) shows 53% id to *Actinobacillus suis* and 4P0034 shows 29% id to a C-methyltransferase from *Bordetella bronchiseptica*.

4.2.7.3.3 Surface

There are several matches to hypothetical proteins associated with capsule clusters from other bacteria. These include 4P0019 (4P3d10p) which shows 49% id to Cj1341, 4P0030 (4P3g02p) which shows 26% id to a hypothetical protein from the capsular gene cluster of *Actinobacillus suis* and 4P0022 (4P3g08p) which shows 28% id to Cj1431. 4P0026 and 4P0027 (4P1b06q) show 59% and 40% id to a hypothetical and LPS biosynthesis protein from *Pseudomonas syringae* and 4P0028 and 4P0029 show 58% and 41% id to two hypothetical proteins from *Actinobacillus suis*. In addition, 4P0095 (4P2a08p) shows 39% id to an acetyltransferase from strain 43446.

4.2.7.3.4 Miscellaneous

4P0039 (4P2b07p) shows 45% id to a pyridine nucleotide-oxidoreductase from *Bacteroides thetaiotaomicron* inserted upstream of an orthologue of cj1069. There are many members of the pyridine nucleotide-oxidoreductase family including glutathione reductases, lipoamide reductases, mercuric reductases, trypanothione reductases and thioredoxin reductases many of which are associated with metabolic pathways or stress responses [155]. However, on closer analysis this putative oxidoreductase did not appear to cluster strongly with any of the above members of the same family.

4P0006 (4P1d01p) shows only 40% id to an MCP-type chemotaxis protein from strain NCTC 11168 possibly suggesting a novel chemotaxis receptor protein. 4P0051 and 4P0052 (4P1e06p) show 62% and 41% to a hydrolase and a hypothetical protein from *Pseudomonas syringae* and 4P0053 shows 25% id to a C-methyltransferase from *Leptospira interrogans*.

4.2.7.4 Strain 52472

4.2.7.4.1 Restriction modification

In strain 52472 there are many predicted CDSs that show homology to RM associated proteins. The predicted CDSs 5P0060, 5P0061 (5P7e11p) and 5P0036 (5P6a05p) show high identity to HsdR, RloF and HsdS respectively from strain RM1170 [154]. CDS 5P0013 (5P3d03p) shows 73% id to a type II RM protein from RM1221 and on the same contiguous region 5P0014 shows 57% id to a hypothetical protein from *Helicobacter pylori*. 5P0003 (5P2f05p) shows 36% id to a type I RM protein from *Staphylococcus aureus*. 5P0037 and 5P0038 (5P5d07p) show 57% and 62% id to a type III RM protein and DNA methyltransferase from *Helicobacter pylori* respectively. Also homologous to the same *Helicobacter pylori* type III RM protein is 5P0065 (5P8c04p) which shows 53% id. There are also fragments of CDSs that match to RM proteins. 5P0073 (5P1e12q) shows 46% id to a type I RM protein from *M. mazei* and 5P0279 (5P7e10q) shows 34% id to a type I RM protein from an uncultured Archaeon, although these may be pseudogenes as the CDSs are much shorter than the genes they share identity with.

4.2.7.4.2 Hypothetical CDSs

There are a number of unique hypothetical CDSs. 5P0040 (5P5g12q) shows 33% id to a hypothetical protein from *Nitrosomonas europea* and appears to be inserted next to *aspB* (cj0762c). Inserted next to *panB* are two hypothetical CDSs, 5P0080 (5P5e04p) which shows 44% id to a hypothetical protein from *Helicobacter hepaticus* and 5P0081.

Other hypothetical proteins without any information as to where they may located on the chromosome are 5P0059 (5P8b01p), 5P0071 (5P5h08p) which may be phage associated as it shows 41% id across part to a hypothetical protein from *Salmonella enterica* Typhi, located adjacent to phage genes, and 5P0141 (5P3c11q), 5P0142 which shows 40% id to a hypothetical protein from *Helicobacter hepaticus* and 5P0143.

4.2.7.4.3 Surface

There are relatively few unique predicted surface proteins with 5P0044 (5P8h04p) showing 38% id to the periplasmic protein Cj0737 and 5P0053 (5P6b10q) showing 52% id to a transport permease from *Escherichia coli*.

4.2.7.4.4 Miscellaneous

The predicted CDS 5P0087 (5P4h09p) shows 46% id to a DNA methyltransferase from *Helicobacter pylori* and 5P0088 shows 31% id to serine/threonine protein kinase from the yeast *Debaryomyces hansenii* which will be discussed further in chapter 5.

4.3 Discussion

4.3.1 Surface structures

There are many surface associated genes which have been identified in this study, a large portion of which show limited identity to NCTC 11168 rather than being completely novel. Also, surface associated genes make up a large portion of the genes that are shared between the different strains in this study. There is a great deal of low level variation within predicted surface associated proteins as demonstrated by predicted CDSs that show 65-95% aa id to CDSs from strain NCTC 11168. It is likely that surface proteins show a lower level of similarity across the backbone than housekeeping proteins as surface proteins are exposed to the external environment and therefore may be antigenic, stimulating an immune response upon infection of a host and leading to diversifying selection of the genes. As is the case for FlgE, the flagellar hook protein, the central portion has been demonstrated to be highly variable whilst the remainder is relatively conserved as the central portion is surface exposed [156].

Dorrell *et al.* [51] found many of the NCTC 11168 genes absent or highly divergent in one or more of the 11 test strains used in their microarray study were associated with the biosynthesis of surface structures. These surface structures included flagella, lipo-oligosaccharide and the capsular polysaccharide biosynthesis regions [51]. *C. jejuni* synthesises both a low molecular weight lipopolysaccharide (LPS) lacking O-antigen repeats, termed LOS, and also a high molecular weight polysaccharide responsible for Penner serotype and thought to be a capsular polysaccharide [53]. Variation in surface structures may be important in evading the immune response of the infected host.

The LOS regions have been highly studied and it appears that strains 81-176 and 52472 belong to class B [143] (**Fig 4.4**). Strain M1 appears to have high identity to the LOS region of 81116. Few novel predicted CDSs identified in this study could be matched to the LOS region for strain 40671; those that are putatively associated with the LOS locus may belong to either class A or class B. A recent study of the LOS loci of 123 *C. jejuni* strains has suggested an extra 3 classes of LOS loci [157] in addition to the 3 already proposed by Gilbert *et al.* [143]. Parker *et al.* have assigned strain 81-176 to class B1 and 81116 to a new class, E [157]. Strain M1 can therefore be putatively assigned to class E based on homology to strain 81116. *Campylobacter* varies LOS structure by altering gene content as well as through recombination within genes and homopolymeric tract variation. The result is that the host is presented with constantly varying surface antigens [59].

Capsular polysaccharides contain negatively charged molecules which increase resistance to phagocytosis and, because they are highly hydrated molecules, they may protect bacteria from desiccation [35]. The capsular polysaccharide is therefore functional for *C. jejuni* survival within a host and in the wider environment. As the capsule locus has been shown to be highly divergent between strains, ranging between 15-34 Kb [151] and containing many horizontally acquired genes, it may be that many of the novel genes identified in this study are located in this region. Without further positional information it is not possible to assign novel CDSs to a particular chromosomal region. For example, there are many sugar modification proteins that are homologous to those from capsule regions in other bacteria in strain M1 e.g. MP0041 (44% id to WbyH O-antigen of *Y. pseudotuberculosis* [158]), along with various epimerases, mutases and glucose dehydrogenases. There are also many hypothetical proteins matching to the capsule clusters of other bacteria in strain 40671, including those from *Actinobacillus suis* and *Pseudomonas syringae*. However, there are many similar classes of genes present at the LOS, flagellar and capsule loci and it has recently been shown that some genes can be shared between capsule and LOS [151]. Therefore, although these predicted CDSs match to proteins associated with the capsule from other bacteria it is not possible to be confident that they are involved in capsular polysaccharide production in these strains.

In strains 81-176, M1 and 40671 there are many genes associated with the flagellar locus that vary. Provisional data from Dorrell *et al.* suggested that the flagella locus of 81-176 is missing large sections of DNA (or may be highly divergent) compared with NCTC 11168 [51]. This has been confirmed by Thibault *et al.* [148] who found that in strain 81-176 orthologues of cj1318-cj1332 are missing, as are orthologues of cj1335 and cj1336. The gene encoding flagellin is present on the *C. jejuni* chromosome in two copies and intragenomic and intergenomic recombination between *flaA* and *flaB* genes of *C. jejuni* has been demonstrated to generate antigenic diversity [65]. The surface exposed portions of flagellins are modified with several monosaccharide units of pseudaminic acid [37;148;159]. Flagella have been shown to have adhesive properties which are an important virulence determinant as, prior to invasion, the bacteria must attach to the epithelial cells [129].

4.3.2 Transport

Campylobacter has been shown to have a large number of transporters and in this study a number of ABC transporters were identified. ABC transporters use ATP hydrolysis to power the uptake and efflux of solutes across the cell membrane. These transporters play a major role in nutrient uptake and may be involved in secretion of toxins and antimicrobial agents. Currently there are 22 subfamilies of ABC importers and 24 subfamilies of ABC exporters [160]. Many found in this study are putatively associated with di-tripeptide transport suggesting a role in nutrient uptake.

Strains 81-176 and M1 contain homologues of DcuC: a dicarboxylate transporter. C4-dicarboxylates like succinate, fumarate and malate can be metabolized by bacteria under both aerobic and anaerobic conditions [161]. In NCTC 11168 there are *dcuA* and *dcuB* homologues but no functional *dcuC* homologue although there are two pseudogenes, cj1528 and cj1389, with homology to *dcuC*. In strain 81-176 it appears likely that the *dcuC* homologue is a more complete version of pseudogene cj1389 whereas in strain M1 the *dcuC* homologue shows some similarity to the pseudogene cj1528. DcuAB are used for electroneutral fumarate:succinate antiport which is required in anaerobic fumarate respiration. DcuC can replace DcuA or DcuB in catalyzing fumarate-succinate exchange and fumarate uptake but usually DcuC carriers function in succinate efflux during fermentation [161].

4.3.3 Restriction modification

Many restriction modification system associated genes were identified in this study, several of which appear to be shared between strains. Restriction modification (RM) systems are comprised of pairs of endonucleases and DNA-methyltransferases that recognise the same DNA sequences. The endonucleases catalyze double-strand cleavage of DNA and methyltransferases catalyze the addition of a methyl group to one nucleotide in each strand of the recognition sequence that results in prevention of cleavage of self DNA. There are four types of RM systems classified by subunit composition, cofactor requirements and position of DNA cleavage site [162].

This study has shown variation in RM genes, e.g. cj0032, as well as novel genes with similarity to RM genes of other species including those in *V. cholerae* (47% id), *Caulobacter crescentus* (40% id), *M. mazei* (45%) and *Archaeoglobus fulgidus* (39%). It has been suggested by others that there are multiple R-M systems in *C. jejuni* [91]. Ahmed *et al.* found that within 24 fragments novel to strain 81116, 6 were similar to RM enzymes [91] and provisional microarray data from Dorrell *et al.* suggests that the restriction modification /methylase genes are a particularly variable between strains when compared to NCTC 11168 [51]. RM genes have also been identified in other comparative studies [83-85].

A recent paper has studied the diversity within the type I RM locus [154]. Based on these data it would appear that M1 has an RM locus equivalent to that of strain 81116 and 52472 has a locus equivalent to strain RM1170. In contrast, strain 40671 has many type I RM proteins homologous to those in other bacteria and does not appear to fall within this typing scheme.

RM systems may function in protecting the cell from bacteriophage infection or invasion by foreign plasmid or genomic DNA, as foreign DNA is unlikely to possess the methylation pattern characteristic of the host cell DNA and will therefore be susceptible to cleavage. *C. jejuni* is a naturally competent organism so it may not be beneficial to cleave all foreign DNA. It has been suggested that restriction of homologous DNA taken up by the cell may aid recombination by generating double-stranded breaks in the DNA [163]. Certain *C. jejuni* RM systems have been shown to be phase variable [51], possibly allowing RM properties of the cell to vary in order to facilitate recombination of foreign DNA or to provide a higher degree of protection against infection by bacteriophage.

4.3.4 Metabolism

In strains 81-176 and 52472 the entire molybdate transport region was found to be variable at the amino acid level from FASTA results (ModA 82% id, Cj0302c 65% id, ModB 85% and ModC 76% and 78% id). This region was also identified by Dorrell *et al.* who listed *modC*, cj0300c as absent/highly divergent from some of the test strains compared to strain NCTC 11168 (http://www.sghms.ac.uk/depts/medmicro/bugs/GR-1858). Molybdate plays a key role in anaerobic respiration by incorporation into molybdoenzymes including DmsABC and formate dehydrogenase, all of which are involved in the reduction of alternative electron acceptors to nitrate [164].

A WUBLASTN comparison of the molybdate transport region, including bioF identified in 81-176 and 52472, to strain RM1221 showed 99% similarity. BioF is involved in the biosynthesis of biotin which is an essential prosthetic group for carboxylase enzymes which each catalyse an essential metabolic reaction [164]. It is possible that this region in NCTC 11168 is under diversifying selection or that it has been horizontally transferred from another source, although the reasons why this region may vary are unclear. The natural competence of *C. jejuni* and high recombination rate are thought to be involved in generating diversity at the cell surface [51] but may also be involved in generating diversity elsewhere in the genome.

There is a homologue of a PrpD family protein in strains M1 and 52472, and there is also a homologue in NCTC 11168, but this is a pseudogene. PrpD is required for propionate catabolism *via* the 2-methylcitric acid cycle [153]. Catabolism of propionate could provide an abundant carbon source for these bacteria as propionate is a short chain fatty acid found in the intestinal lumen [153]. A number of oxidoreductases were identified in this study that are either novel or show limited identity to those from NCTC 11168. Oxidoreductases play a role in many aspects of metabolism, and it is difficult to ascribe a specific function to most of those found.

4.3.5 Respiration

Several reductases potentially involved in respiration were found among the CDSs predicted on novel 81-176 and M1 contiguous regions: these included homologues of *W. succinogenes* DmsA, a dimethyl sulfoxide reductase (62% id), FdhB, an oxidoreductase (47% id), and MraY, a hyothetical protein with similarity to dimethyl sulfoxide reductase (43% id). Also potentially involved in respiration are the CDSs with similarity to the cytochrome C biogenesis proteins of *W. succinogenes, Shewanella oneidensis and Geobacter sulfurreducens*. *C. jejuni* has a complex and highly branched respiratory chain and many cytochromes as well as the possibility of anaerobic growth with fumarate as the terminal electron acceptor [4]. This diversity in respiratory associated proteins may aid survival in the different ecological niches to which *C. jejuni* is exposed: for example, the avian and mammalian gut which is essentially anaerobic [4].

Strains 81-176 and M1 each contain CDSs with similarity to *Helicobacter pylori* γ-glutamyl transpeptidase (GGT) (66% id). GGTs have a major role in glutathione metabolism which in turn has a role in protection of the bacterial cell against oxidative stress [165]. GGTs may also play a role in transport of amino acids across cell membranes in bacteria [166]. Both a cytochrome C oxidase III and a GGT specific to 81116 were found by Ahmed *et al.* [91] again underlining strain variation in respiratory and oxidative stress associated genes.

4.3.6 Chemotaxis

In this study a chemotaxis receptor protein has been found in M1 that is also present in RM1221 but not in strain NCTC 11168. There is also a different novel chemotaxis protein in 40671. It has been noted that in the NCTC 11168 genome the carboxy-terminal portion of the methyl-accepting proteins representing the signalling domains is highly conserved. This portion is proposed to be highly conserved in order to interact with CheW which is part of the signal transduction complex. However, the receptor domains may be highly variable representing specificity for different substrates [167]. This will be discussed further in chapter 5.

4.3.7 Pseudogenes

Although this study was not designed to investigate pseudogenes, 8 pseudogenes from NCTC 11168 have been identified through the differential hybridization screen and appear to vary. It should be borne in mind that the depth of coverage in this pUC screen will not give definitive results with regard to pseudogenes in all cases. In addition there are a number of novel genes that are probably pseudogenes. This is perhaps not unexpected as genes that are not shared between strains are likely to be accessory and perhaps only required under a subset of ecological conditions which the bacterium may encounter and therefore more likely to pick up deleterious mutations.

4.3.8 Characteristics of each strain

4.3.8.1 Strain 81-176

Strain 81-176 is a highly studied strain with two plasmids not found in strain NCTC 11168, as such this strain was selected as a way of testing the differential hybridization method. This strain has the highest proportion of CDSs with 65-95% amino acid id to CDSs of strain NCTC 11168. In addition to the plasmid sequences this study also identified novel respiration associated genes and relatively few RM associated genes when compared to the other strains used in this study. Amino acids are a useful nutrient source and in strain 81-176 a novel serine protease was identified, which contains an autotransporter domain and a subtilase family domain. Proteases have also been implicated in virulence in bacteria, for example the IgA protease of *Neisseria* and *Haemophilus* [168;169]. In addition a putative adhesin, a di-tripeptide transporter, and a CDS with partial homology to TraG were identified. These may be good candidates for further investigation.

A recent study by Poly *et al.* has used a microarray to identify CDSs present in 81-176 that are absent in strain NCTC 11168 [170]. Poly *et al.* identified 58 contiguous regions constituting 63 Kb of novel DNA sequence predicted to encode 86 CDSs. Of these 58 regions identified by Poly *et al.*, 37 have been identified in this study; these 37 regions correlate to 24 out of the 58 regions identified in this study (several of the novel regions identified by Poly *et al.* mapped to single contigs in this study). This means that out of the regions identified by Poly *et al.* 36% have been missed in this study; most of the regions missed are from the LOS region or capsule region and have a low G+C content [170]. Of the regions identified in this study 59% were missed by Poly *et al.*

4.3.8.2 Strain M1

Strain M1 has the most diversity of all the strains. There are many CDSs predicted to be associated with surface structures, sugar modification, RM, respiratory chain, as well as some putative adhesins (**Fig 4.3**). Of the novel CDSs from this strain the putative autotransporter warrants further investigation as this may have virulence functions, and the adhesins may be important for chicken colonization or virulence in humans. There is also a predicted CDS with high identity to *tetO* which appears to be in a distinct context compared to that in pTet.

4.3.8.3 Strain 40671

Strain 40671 has the highest proportion of hypothetical CDSs of all the strains with nearly half of the novel CDSs categorized as hypothetical. Many of these hypothetical CDSs are unique to this strain. There are many CDSs which show homology to genes located in capsule biosynthesis loci of other bacteria which may suggest that this strain has diversity in the polysaccharide biosynthesis regions although this possibility will need to be explored further. There are some homologues of CDSs from pTet including homologues of many type IV secretion system genes possibly indicating that this strain has a plasmid. There is also a novel chemotaxis associated CDS.

4.3.8.4 Strain 52472

Strain 52472 has a large number of bacteriophage associated CDSs. Bacteriophage are known to be highly mosaic in structure with a high rate of horizontal exchange between bacteriophage occupying similar ecological niches [171]. Bacteriophage may pick up virulence determinants when they excise although none have been found in this study from strain 52472 or in RM1221 [9].

In this strain there are also a number of regions which show only limited identity to NCTC 11168, e.g. the molybdate transport region. Many homologues of pTet CDSs have also been identified in this strain including homologues of all the type IV secretion system genes indicating the possible presence of a plasmid. Strain 52472 has the highest number of predicted CDSs associated with RM of all strains in this study (**Fig 4.3**).

4.3.9 Summary

The pUC screen has identified a large amount of variation between the test strains confirming that *C. jejuni* is a highly variable species. However, it is difficult to tell without the context of the surrounding genes what systems may be functional and which may be pseudogenes. It is also very likely that there is redundancy within the identified CDSs, with several partial CDSs actually belonging to the same genes. In order to explore some of these regions further it was decided to sequence BAC library clones that contain some of the more interesting genes (chapter 5).

5. Analysis of BAC libraries

5.1 Introduction

The method of differential hybridization with small insert pUC libraries has identified a range of interesting novel predicted CDSs as discussed in chapter 4. It is not possible to gain an accurate prediction of function from fragments of predicted CDSs so to study regions of interest in more detail larger insert BAC libraries were used. As some of the consensus sequences from the pUC contigs are based on few reads, sequencing BAC clones to a higher depth of coverage will provide more accurate sequence as well as providing context for the CDSs and identifying the insertion point relative to the chromosome of strain NCTC 11168.

5.2 Results

5.2.1 Overview of methods

Two small-insert BAC libraries were constructed for each strain under investigation with 15-20 kb and 20-40 kb inserts, each representing 5-fold coverage of the genome (section 2.3.2). Each library was arrayed in duplicate onto membranes (section 2.3.4). Genes of interest from pUC assemblies were selected for further analysis representing a selection of CDSs shared between strains, CDSs unique to each strain, CDSs which may be involved in virulence and CDSs which may be functional homologues of pseudogenes in strain NCTC 11168. Oligonucleotide primers were designed from regions of best coverage from the contiguous regions containing the gene of interest and used to generate radiolabelled probes to screen the BAC libraries (section 2.3.5). For strain 81-176, 7 probes were designed, strain M1, 11 probes were designed, strain 40671, 7 probes and strain 52472, 7 probes were designed (**Table 5.1**).

Table 5.1 A: strain 81-176 initial probes

Probe id	CDS	contig	match	Organism with match
8P6a02	8P0002c	8P6a02q	Putative adhesin	*Chromobacterium violaceum*
8P2e09	8P0042	8P2c09q	hypothetical	*Campylobacter jejuni*
8P4d10	8P0055c	8P6g02p	DTPT transporter	*Photorhabdus luminescens*
8P3c06	8P0070	8P7f02p	TraG fragment	*Escherichia coli*
8P1d09	8P0076	8P7f11p	hypothetical	*Clostridium perfringens*
8P5c02	8P0078	8P6g03q	DmsA	*Wolinella succinogenes*
8Pf01	8P0081	8P0081	Cytochrome C biogenesis	*Wolinella succinogenes*

Table 5.1 B: strain M1 initial probes

Probe id	CDS	contig	match	Organism with match
5P1h08	MP0023c	MP2g06p	autotransporter	*Helicobacter pylori*
8P4d10	MP0038c	MP1a12p	DTPT transporter	*Photorhabdus luminescens*
MP3d04	MP0046c	MP3d04q	Putative adhesin	*Chromobacterium violaceum*
MP3e11	MP0054	MP3e11p	Putative haemolysin	*Xanthomonas axonopodis*
8P1b12	MP0090	MP4e08q	Cytochrome C	*Shewanella oneidensis*
MP1d11	MP0101	MP1d11p	TetO	*Campylobacter jejuni*
8P5c02	MP0103	MP1g06p	DmsA	*Wolinella succinogenes*
8P3c06	MP0104c	MP4e01q	TraG pseudogene	*Vibrio vulnificus*
MP3b01	MP0133	MP3b01p	EspC	*Escherichia coli*
MP4c04	MP0141	MP4c04p	Haemagglutinin-related protein	*Ralstonia solanacearum*
MP2f07	MP0149	MP2f07q	Haemoglobin protease	*Escherichia coli*

Table 5.1 C: strain 40671 initial probes

Probe id	CDS	contig	match	Organism with match
4P1d01	4P0006	4P1d01p	MCP signal transduction	*Campylobacter jejuni*
4P1f05	4P0035	4P1b12q	hypothetical	*Chromobacterium violaceum*
4P1e10	4P0039	4P2b07p	oxidoreductase	*Bacteroides thetaiotaomicron*
4P1e06	4P0052c	4P1e06p	hypothetical	*Pseudomonas syringae*
4P1h09	4P0060c	4P3a10q	hypothetical	*Helicobacter pylori*
4P1a10	4P0063c	4P1a10p	DmhA	*Yersinia pseudotuberculosis*
4P1a12	4P0085c	4P1a12q	hypothetical	*Leishmania tarentolae*

Table 5.1 D: strain 52472 initial probes

Probe id	CDS	contig	match	Organism with match
5P5a12	5P0044	5P8h04p	Periplasmic protein	*Campylobacter jejuni*
5P1h08	5P0066	5P5h03q	autotransporter	*Helicobacter pylori*
5P4h09	5P0088	5P4h09p	Serine-threonine protein kinase	*Debaryomyces hansenii*
5P5a06	5P0116	5P5a06p	VirB4	*Campylobacter coli*
5P3h01	5P0196	5P7b11p	hypothetical	Bacteriophage D3112
5P5g10	5P0277c	5P5g10q	PrpD family protein	*Bradyrhizobium japonicum*
5P3e01	5P0080c	5P5e04p	hypothetical	*Helicobacter hepaticus*

Clones that hybridized to each probe were end-sequenced (section 2.3.6.2.3) and compared against the strain NCTC 11168 genome sequence using WUBLASTN (section 2.3.7). Clones with one or both ends containing sequence complementary to strain NCTC 11168 were selected for subcloning and sequenced using a shotgun strategy. The inserts from BAC clones were released by digesting with the restriction enzyme *Not*I (section 2.2.5) and separated from the vector backbone using agarose gel electrophoresis (section 2.2.3). DNA was extracted from the gel (section 2.2.6.1.2) then fragmented using sonication and cloned into pUC19 vector (section 2.3.1.1). *Escherichia coli* colonies containing the subclones were propagated (section 2.2.1) then the subclone DNA was prepared (section 2.2.2.1) and sequenced (section 2.3.5.2.1).

If sequence complementary to the strain NCTC 11168 genome sequence was not found at both ends of the insert in the initial end-sequence screening, BAC clones with one matching end were sequenced using a shotgun strategy then more primers were designed further along the novel region and the process was repeated until the extent of the novel region was found. Novel sequence was finished to a depth of at least 4 reads, with reads in both directions and a consensus base quality of at least 30.

BAC sequences were named in the following way: first the strain designator character, 8 for strain 81-176, M for strain M1, 4 for strain 40671 and 5 for strain 52472; next the library designator character, B for BAC library; then the library plate number followed by the well reference. Thus the sequence 8B4F10 would be generated from the BAC clone of strain 81-176 located in well F10 of plate number 4. The BAC sequences were arranged in the same orientation as the NCTC 11168 chromosome with CDSs encoded on the complementary strand labelled with a ‘c’.

5.2.2 Respiration

From the pUC assemblies it became apparent that there are a number of respiratory associated CDSs that are shared between strains 81-176 and M1. A probe (8Pf01) for a predicted CDS with homology to the cytochrome C biogenesis protein from *W. succinogenes* (**Table 5.1A**) was used to identify a novel region in 81-176. The initial shotgun sequence did not cover the entire novel insert so in order to expand this region to find the extent of the novel insert a second probe was designed (8P1b12) for a CDS with homology to a cytochrome C protein from *Shewanella oneidensis*. This probe was also used to identify the corresponding region in strain M1 (**Table 5.1B**). BAC 8B4F10 shows that strain 81-176 contains an insert between the rDNA and a homologue of cj0033. This novel insert replaces cj0030 relative to the NCTC 11168 chromosome (**Fig 5.1** and **Table 5.2**).

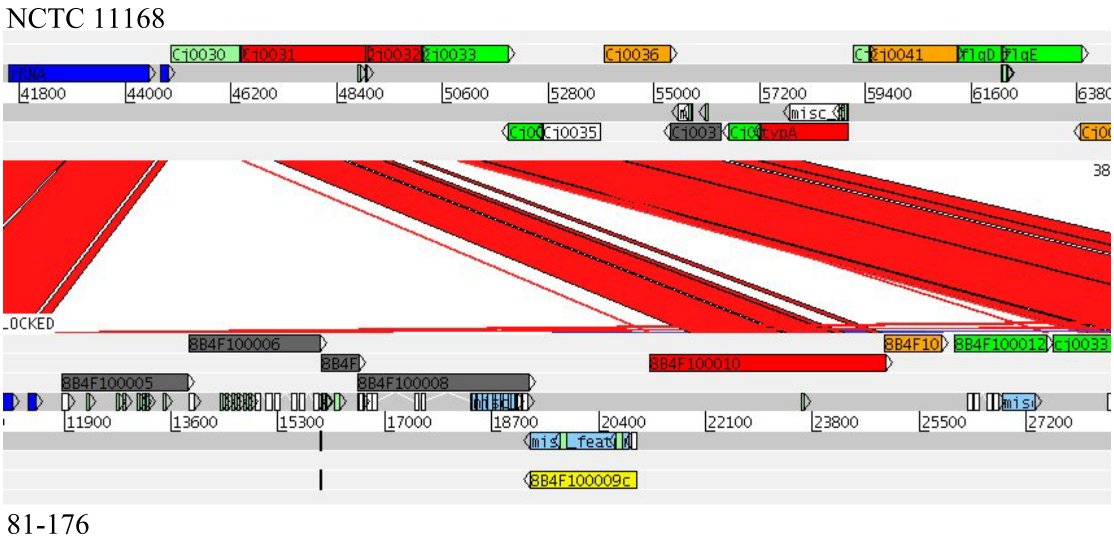
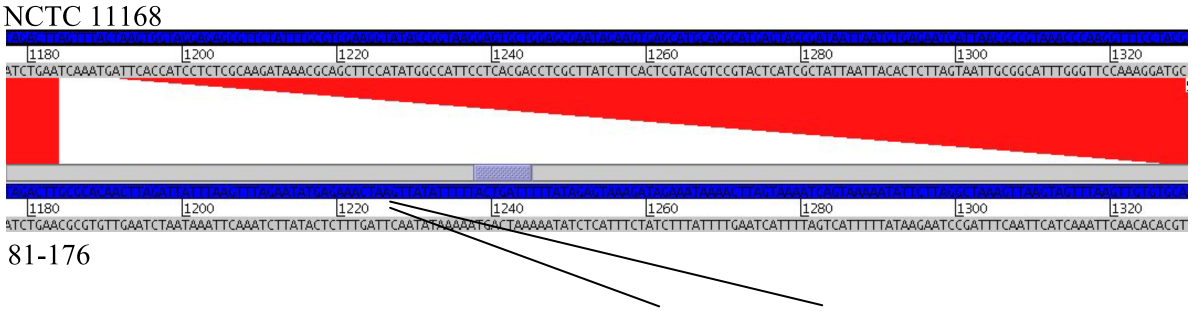


Fig 5.1: Blastn comparison of strain NCTC 11168 and strain 81-176 BAC clone 8B4F10. The comparison is viewed using the Artemis Comparison Tool (ACT) (Rutherford, K., unpublished). Blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark grey lines; the three-frame forward and reverse DNA translations are represented by light grey lines. Features are represented by open boxes: pFam (light blue), tmhmm (white), signalP (white) and prosite (green) matches, and rDNA (dark blue), are indicated on the DNA lines. CDSs are marked on the frame lines; in NCTC 11168 the CDSs are all on one frame line irrespective of reading frame. CDSs are coloured according to functional category: grey, energy metabolism; yellow, central/ intermediary/ miscellaneous metabolism; red, information transfer/ DNA modification; orange, conserved hypothetical; dark green, surface; light green, unknown; white, pathogenicity/ adaptation/ chaperones. In strain 81-176 there are 5 CDSs (8B4F10_5-8B4F10_9c) between rDNA and cj0031 and 2 CDSs (8B4F10_11-8B4F10_12) between cj0031 and cj0033. 8B4F10_10 has regions of similarity to cj0031 but the N- and C-terminus are novel. As the rDNA is present in 3 copies on the chromosome, the red lines from the rDNA of strain 81-176 indicate matches to those other copies.

Table 5.2: Predicted novel CDSs identified from BAC clone 8B4F10.

Locus_id	Putative function	Organism with match	SWALL	E-value	%id
8B4F10_5	Cytochrome C	*Shewanella oneidensis*	Q8EJI6	2.6e-135	55.24
8B4F10_6	Hypothetical	*Shewanella oneidensis*	Q8EJI5	2.7e-12	39.43
8B4F10_7	Thiol:disulfide interchange protein	*Helicobacter pylori*	Q9ZKD5	3.3e-13	36.31
8B4F10_8	Cytochrome C biogenesis	*Wolinella succinogenes*	Q9S1E4	2.5e-124	41.13
8B4F10_9c	Gamma-glutamyl transferase	*Helicobacter pylori*	Q9ZK95	5.5e-135	67.2
8B4F10_10	Type II RM enzyme	*Campylobacter jejuni*	Q9PJ80	0	85.25
8B4F10_11	hypothetical	*Enterococcus faecalis*	AA081633	9.3e-4	29
8B4F10_12	Membrane carboxypeptidase	*Clostridium acetobutylicum*	Q97GR5	6.9e-05	33.33

The corresponding region in strain M1 was deduced from BAC MB2B4. In strain M1 the BAC MB2B4 only contained the region between *recJ* (cj0028) and cj0031 relative to the chromosome of strain NCTC 11168. The BAC clone sequences of 8B4F10 and MB2B4 show 99% nucleotide identity although they contain a different intervening sequence (IVS) in the 23s rDNA [172] (**Fig 5.2**).



```
M1_23SrRNA      1201 tttaagtttagaatatgagaaactaagttatgttagttatatttttact   1250
                     ||||||||||||||||||||||||          |||||||||||||||
81176_23SrRNA   1201 tttaagtttagaatatgagaaacta----------agttatatttttact   1240
```

Fig 5.2: Blastn comparison of strain NCTC 11168 and strain 81-176 BAC clone 8B4F10 23S rDNA sequence. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. The sequence of the 23S rDNA is marked by the dark blue boxes. In strain 81-176 there is a 145 bp IVS which replaces 8 bp relative to the sequence of NCTC 11168. Below the ACT comparison is an alignment of the region of difference between the IVS of strains M1 and 81-176. There are an extra 10 bp in the IVS of M1 compared to the IVS of strain 81-176.

As the sequences in strain 81-176 and M1 are so similar, only the CDSs from strain 81-176 will be discussed further. Downstream of the rDNA there are four predicted cytochrome C associated genes, in the first cytochrome C homologue 8B4F10_5 there are 6 prosite cytochrome c family heme-binding site signatures as well as a signal peptide. There are also 6 cytochrome c family heme-binding site signatures and a signal peptide in the second novel CDS 8B4F10_6 as well as 6 transmembrane helices. There are 14 transmembrane helices in 8B4F10_8, an NrfI, cytochrome C biogenesis protein homologue. This BAC also contains homologues of a gamma glutamyl transpeptidase and a RM protein as discussed in chapter 4. The strain M1 BAC clone sequence of MB2B4 does not extend to the CDS predicted to encode the RM protein. Downstream of the CDS predicted to encode an RM protein in strain 81-176 are a hypothetical CDS and a CDS predicted to encode a membrane carboxypeptidase.

The BAC sequence 8B4F10 contains the 81-176 contiguous pUC regions 4b02p, 7d05p, 1a07p and 8e07p which cover 74% of the novel DNA. In strain M1 the BAC MB2B4 contains contiguous pUC regions 1h01q, 4e08q, 2e03p, 4d08p and 2g10p which cover 85% of the novel DNA.

In other contigs within the pUC assemblies a number of dimethyl sulfoxide reductase homologues were found and are shared between strains 81-176 and M1. In order to investigate this region further a probe (8P5c02) was designed from 8P0078 a homologue of *dmsA* from *W. succinogenes* (**Table 5.1**). The same probe was used for both 81-176 and M1 libraries.

In strains 81-176 and M1, this insert is between cj1584c and cj1586 replacing the oxidoreductase cj1585 and is located on BACs 8B1E5 and MB5G8 (**Fig 5.3** and **Table 5.3**). This region appears to encode a *dmsABC* operon homologous to *W. succinogenes* with conserved gene order (**Fig 5.4**).

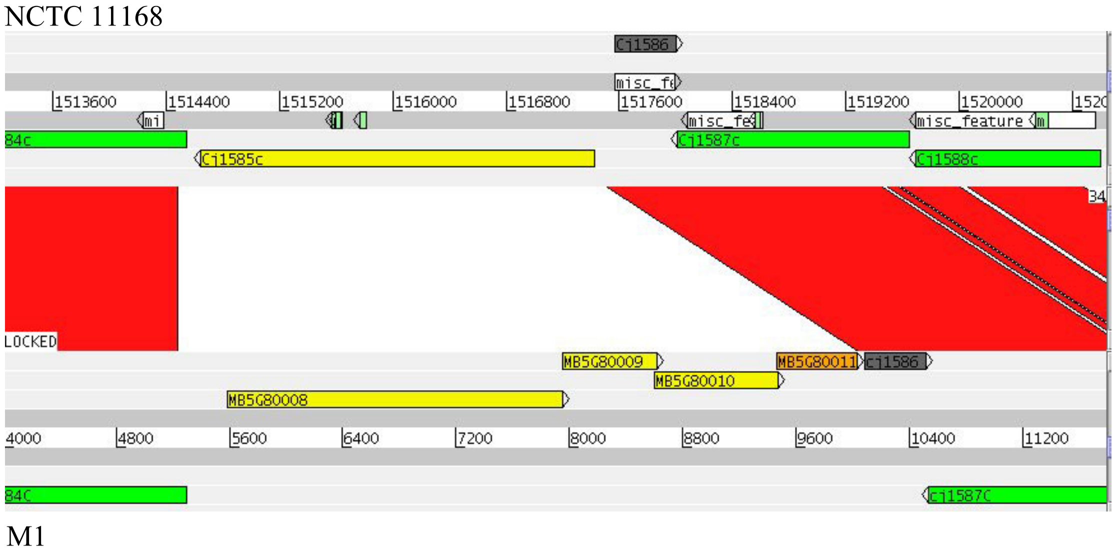


Fig 5.3: Blastn comparison of strain NCTC 11168 and strain M1 BAC clone MB5G8. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark grey lines; the three-frame forward and reverse DNA translations are represented by light grey lines. Features are indicated by open boxes: for strain NCTC 11168 pFam (white) and prosite (green) matches are marked on the DNA lines; CDSs are marked on the translated reading frame lines. CDSs are coloured to indicate functional category: dark green, surface; yellow, central/ intermediary/ miscellaneous metabolism; orange, conserved hypothetical; grey, energy metabolism. In strain M1 an operon of oxidoreductases replaces the oxidoreductase cj1585c compared to strain NCTC 11168.

Table 5.3: Predicted novel CDSs identified from BAC clone MB5G8

locus_id	Putative function	Organism with match	SWALL	E-value	% id
MB5G8_8	Dimethyl sulfoxide reductase	*Wolinella succinogenes*	Q7MRE1	9.9e-198	60.67
MB5G8_9	Oxidoreductase	*Wolinella succinogenes*	Q7M8T2	1.8e-55	63.13
MB5G8_10	Hypothetical	*Wolinella succinogenes*	Q7MRE0	8.8e-40	42.5
MB5G8_11	Hypothetical	*Wolinella succinogenes*	Q7MRD9	6.1e-14	31.72

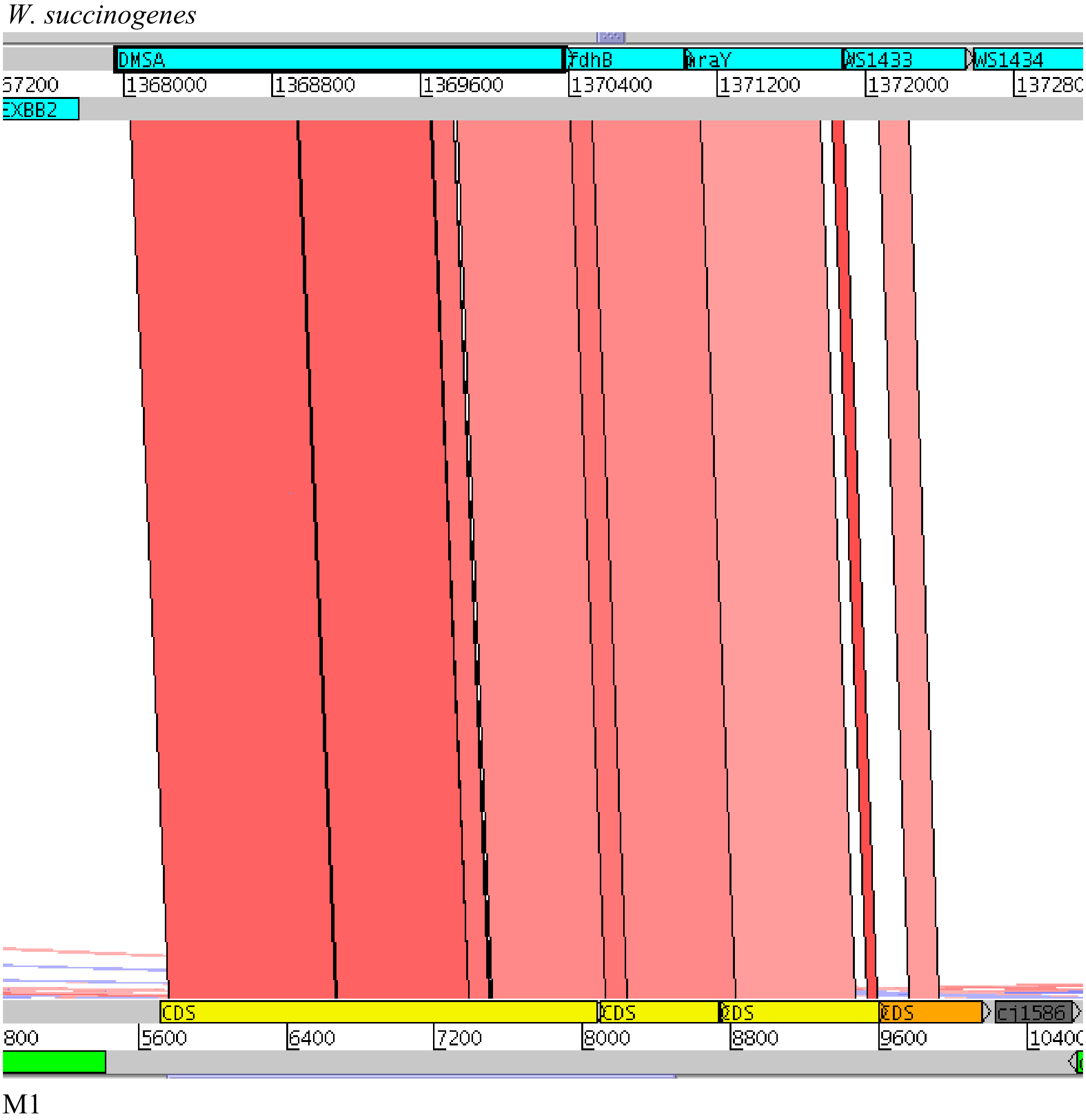
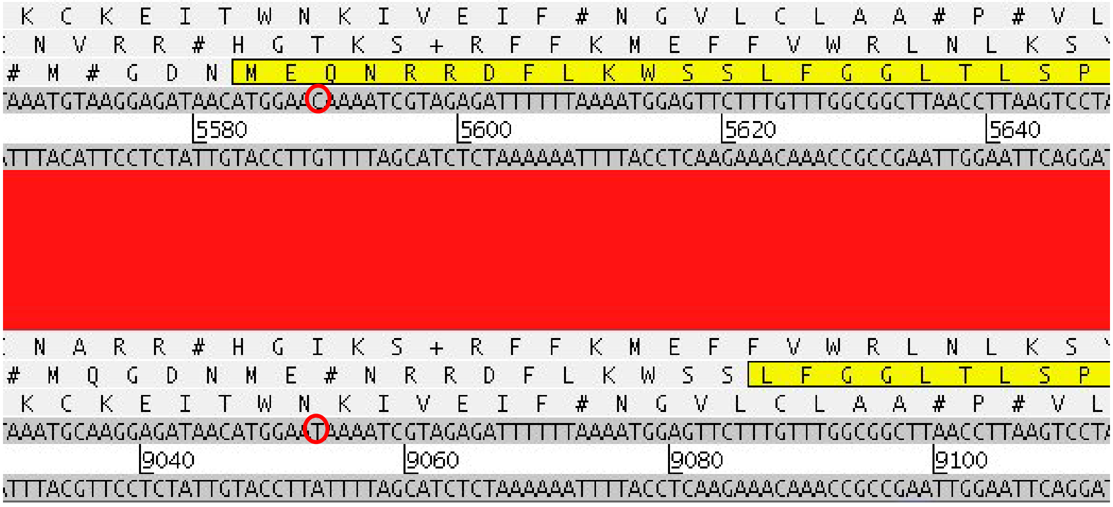


Fig 5.4: **tblastx comparison of *W. succinogenes* and the strain M1 BAC clone sequence of MB5G8.** The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. CDSs are indicated by open boxes. CDSs from *W. succinogenes* (accession number BX571656, Baar *et al.* [173]) are coloured blue. CDSs from strain M1 (MB5G8_7-MB5G8_12) are coloured according to functional category: dark green, surface; yellow, central/ intermediary/ miscellaneous metabolism; orange, conserved hypothetical; grey, energy metabolism.

In strains 81-176 and M1 this region shares 98% identity at the nucleotide level. The main difference between M1 and 81-176 is that the DmsA homologue is predicted to be 787 aa in M1 and only 774 aa in 81-176 as the predicted start site of this protein is located 13 aa downstream of that in strain M1 due to a stop codon being generated in this reading frame by a base pair change giving TAA instead of CAA (**Fig 5.5**).

M1



```
 K  C  K  E  I  T  W  N  K  I  V  E  I  F  #  N  G  V  L  C  L  A  A  #  P  #  V  L
:  N  V  R  R  #  H  G  T  K  S  +  R  F  F  K  M  E  F  F  V  W  R  L  N  L  K  S  Y
#  M  #  G  D  N  M  E  Q  N  R  R  D  F  L  K  W  S  S  L  F  G  G  L  T  L  S  P
TAAATGTAAGGAGATAACATGGAACAAAATCGTAGAGATTTTTTAAAATGGAGTTCTTTGTTTGGCGGCTTAACCTTAAGTCCTA
                  5580                5600                5620                5640
ATTTACATTCCTCTATTGTACCTTGTTTTAGCATCTCTAAAAAATTTTACCTCAAGAAACAAACCGCCGAATTGGAATTCAGGAT

:  N  A  R  R  #  H  G  I  K  S  +  R  F  F  K  M  E  F  F  V  W  R  L  N  L  K  S  Y
#  M  Q  G  D  N  M  E  #  N  R  R  D  F  L  K  W  S  S  L  F  G  G  L  T  L  S  P
 K  C  K  E  I  T  W  N  K  I  V  E  I  F  #  N  G  V  L  C  L  A  A  #  P  #  V  L
TAAATGCAAGGAGATAACATGGAATAAAATCGTAGAGATTTTTTAAAATGGAGTTCTTTGTTTGGCGGCTTAACCTTAAGTCCTA
              9040                9060                9080                9100
ATTTACGTTCCTCTATTGTACCTTATTTTAGCATCTCTAAAAAATTTTACCTCAAGAAACAAACCGCCGAATTGGAATTCAGGAT
```

81-176

Fig 5.5: Blastn comparison of the sequence from strain M1 and strain 81-176 BAC clones MB5G8 and 8B1E5. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match, however, single base pair differences cannot be accurately represented. Forward and reverse DNA sequences are represented by dark grey lines; the three-frame forward DNA translations are represented by light grey lines. The CDSs MB5G8_8 and 8B1E5_12 are represented by open yellow boxes. These homologues of *dmsA* have a different predicted start site due to a base change, circled in red, generating a stop codon in strain 81-176.

The BAC 8B1E5 contains strain 81-176 contiguous pUC region 6g03q covering 70% of the novel DNA. The BAC MB5G8 contains strain M1 contiguous pUC region 1g06p and 5c01p covering 79% of the novel DNA.

5.2.3 Transport

5.2.3.1 di-tripeptide transporters

From the pUC assemblies described in chapter 4 it was apparent that there was a di-tripeptide transporter shared between all the strains in the study. It was decided to explore this region in two of the strains to see if there was any low level variation between them. The probe 8P4d10 was designed from a predicted CDS encoding a homologue of a di-tripeptide transporter from *Photorhabdus luminescens* (**Table 5.1**). This probe identified the BACs 8B2F5 in strain 81-176 and MB3F5 in strain M1.

The BAC sequences of 8B2F5 and MB3F5 share 98% nucleotide identity and both contain two di-tripeptide transporters inserted between cj0653c and cj0659c (**Fig 5.6** and **Table 5.4**). In NCTC 11168 there is pseudogene cj0654c which shows homology to all but the C-terminal portion of the right hand transporter (MB3F5_12c). The left hand transporter (MB3F5_11c, 8B3F5_8c) in both strains contains a frame shift but at different locations: 81-176 has A(7) at 428bp while M1 has A(8) extending the reading frame in this strain. At 454bp there is an extra GT compared to strain 81-176 leading to a frame shift (**Fig 5.7**). The right hand transporter (MB3F5_12c) is complete in strain M1 but there is a frame shift in 81-176 (8B2F5_9c); strain M1 has A(8) in the homopolymeric tract but strain 81-176 has A(7) (**Fig 5.8**). Comparison of this region to strain RM1221 shows that the left hand transporter (CJE0757) is complete but there is no CDS equivalent to the right hand transporter as there are multiple stop codons interrupting the reading frame (pseudogene CJE0758).

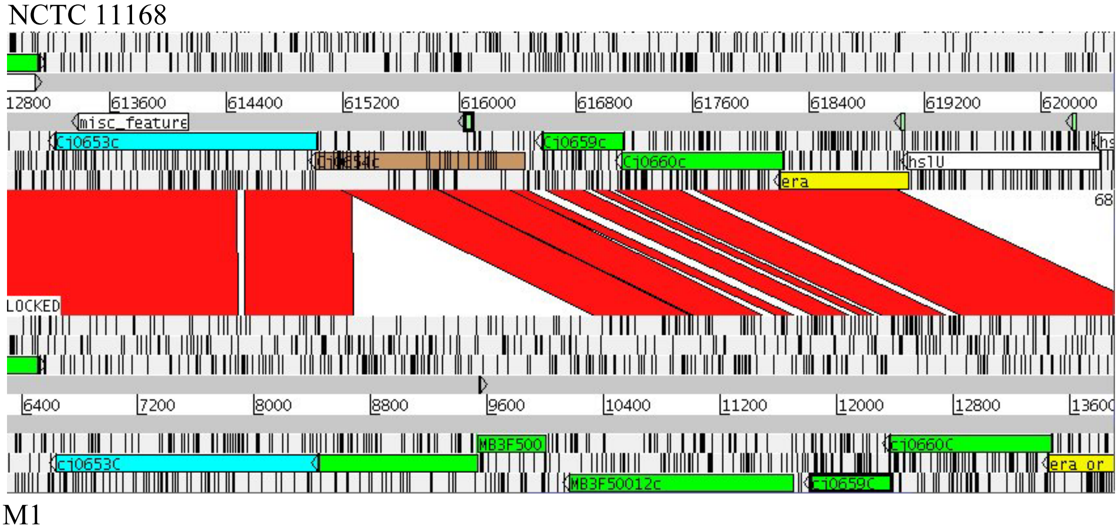
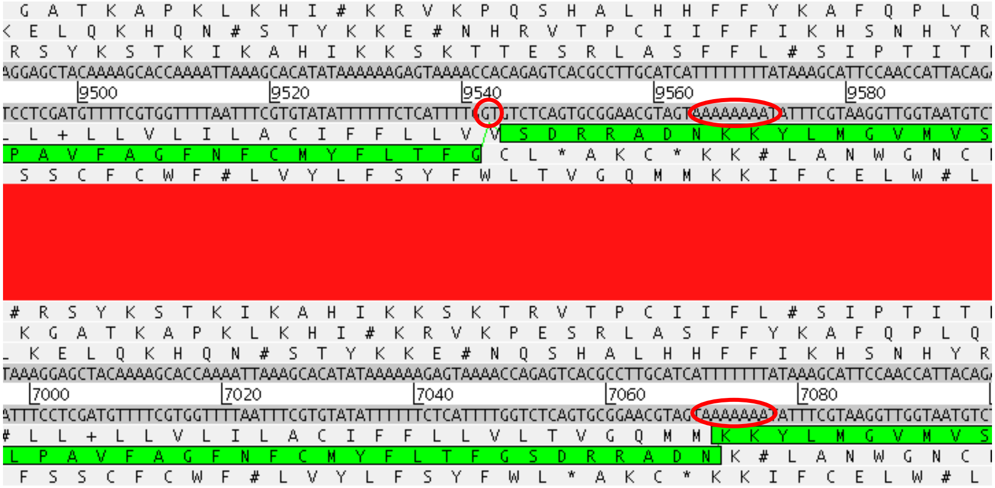


Fig 5.6: Blastn comparison of strain NCTC 11168 and strain M1 BAC clone MB3F5 sequence. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark grey lines; the three-frame forward and reverse DNA translations are represented by light grey lines. Stop codons are indicated by vertical black lines. Features are represented by open boxes: pFam (white) and prosite (green) matches are marked on the DNA lines; CDSs are marked on the translated reading frame lines. CDSs are coloured according to functional category: dark green, surface; blue, degradation of large molecules; brown, pseudogenes; yellow, central/ intermediary/ miscellaneous metabolism; white, pathogenicity/ adaptation/ chaperones. The pseudogene cj0654c shows homology to the N-terminus of MB3F5_12c and the C-terminus of MB3F5_11c, possibly indicating a deletion event.

Table 5.4: Predicted novel CDSs identified from BAC clone MB3F5

locus_id	putative function	organism with match	SWALL	E-value	%id
MB3F5_11c	di-/tripeptide transporter	*Photorhabdus luminescens*	Q7N5W6	1.2e-79	46.43
MB3F5_12c	di-/tripeptide transporter	*Lactococcus lactis*	P36574	5.9e-55	34.57

M1 MB3F5_11c

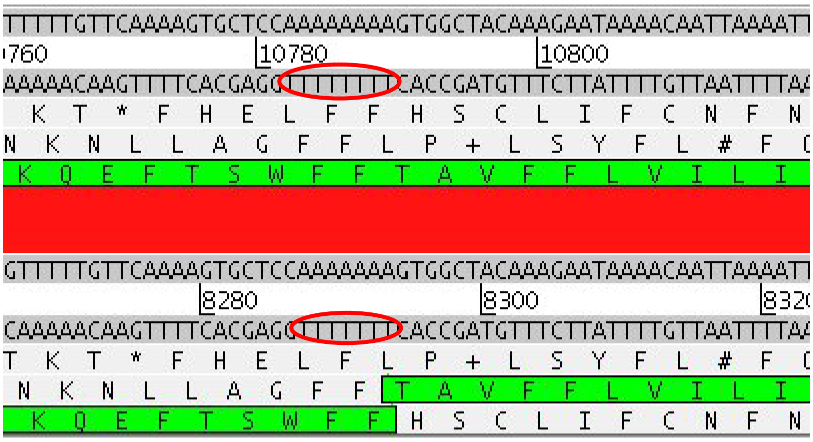


```
 G A T K A P K L K H I # K R V K P Q S H A L H H F F Y K A F Q P L Q
< E L Q K H Q N # S T Y K K E # N H R V T P C I I F F I K H S N H Y R
R S Y K S T K I K A H I K K S K T T E S R L A S F F L # S I P T I T
AGGAGCTACAAAAGCACCAAAATTAAAGCACATATAAAAAAGAGTAAAACCACAGAGTCACGCCTTGCATCATTTTTTTTATAAAGCATTCCAACCATTACAG
      9500          9520          9540          9560          9580
TCCTCGATGTTTTCGTGGTTTTAATTTCGTGTATATTTTTTCTCATTTTGTGTCTCAGTGCGGAACGTAGTAAAAAAAATATTTCGTAAGGTTGGTAATGTC
L + L L V L I L A C I F F L L V V S D R R A D N K K Y L M G V M V S
P A V F A G F N F C M Y F L T F G C L * A K C * K K # L A N W G N C
 S S C F C W F # L V Y L F S Y F W L T V G Q M M K K I F C E L W # L

# R S Y K S T K I K A H I K K S K T R V T P C I I F L # S I P T I T
 K G A T K A P K L K H I # K R V K P E S R L A S F F Y K A F Q P L Q
 K E L Q K H Q N # S T Y K K E # N Q S H A L H H F F I K H S N H Y R
TAAAGGAGCTACAAAAGCACCAAAATTAAAGCACATATAAAAAAGAGTAAAACCAGAGTCACGCCTTGCATCATTTTTTTTATAAAGCATTCCAACCATTACAG
  7000          7020          7040          7060          7080
ATTTCCTCGATGTTTTCGTGGTTTTAATTTCGTGTATATTTTTTCTCATTTTGGTCTCAGTGCGGAACGTAGTAAAAAAATATTTCGTAAGGTTGGTAATGTC
# L L + L L V L I L A C I F F L L V L T V G Q M M K K Y L M G V M V S
L P A V F A G F N F C M Y F L T F G S D R R A D N K # L A N W G N C
 F S S C F C W F # L V Y L F S Y F W L * A K C * K K I F C E L W # L
```

81-176 8B2F5_8c

Fig 5.7: Blastn comparison of sequence from strain M1 and strain 81-176 BAC clones MB3F5 and 8B2F5 in the region of MB3F5_11c. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match, however, small regions of difference cannot be accurately represented. Forward and reverse DNA sequences are represented by dark grey lines; the three-frame forward and reverse DNA translations are represented by light grey lines. The CDSs are marked by open boxes on the translated DNA lines. Both predicted CDSs MB3F5_11c and 8B2F5_8c contain frame shifts. Regions of difference are outlined in red.

M1 MB3F5_12c



```
TTTTTGTTCAAAAGTGCTCCAAAAAAAAAGTGGCTACAAAGAATAAAACAATTAAAATT
760             10780              10800
AAAAACAAGTTTTCACGAGGTTTTTTTTCACCGATGTTTCTTATTTTGTTAATTTTAA
  K  T  *  F  H  E  L  F  F  H  S  C  L  I  F  C  N  F  N
N  K  N  L  L  A  G  F  F  L  P  +  L  S  Y  F  L  #  F
 K  Q  E  F  T  S  W  F  F  T  A  V  F  F  L  V  I  L  I

GTTTTTGTTCAAAAGTGCTCCAAAAAAAAGTGGCTACAAAGAATAAAACAATTAAAATT
           8280              8300              832
CAAAAACAAGTTTTCACGAGGTTTTTTTCACCGATGTTTCTTATTTTGTTAATTTTAA
T  K  T  *  F  H  E  L  F  L  P  +  L  S  Y  F  L  #  F
 N  K  N  L  L  A  G  F  F  T  A  V  F  F  L  V  I  L  I
  K  Q  E  F  T  S  W  F  F  H  S  C  L  I  F  C  N  F  N
```

81-176 8B2F5_9c

Fig 5.8: Blastn comparison of sequence from strain M1 and strain 81-176 BAC clones MB3F5 and 8B2F5 in the region of MB3F5_12c. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match, however single bp changes cannot be accurately represented. Forward and reverse DNA sequences are represented by dark grey lines; the three-frame reverse DNA translations are represented by light

grey lines. In strain M1 predicted CDS MB3F5_12c is complete whereas in strain 81-176 CDS 8B2F5_9c has a frame shift predicted to occur at the homopolymeric A tract circled in red.

8B2F5 contains the 81-176 contiguous pUC sequence 6g02p covering 95% of the novel DNA and MB3F5 contains the strain M1 contiguous pUC sequence 1a12p covering the entire novel sequence.

5.2.3.2 Autotransporter

From the pUC assemblies there appeared to be an autotransporter with homology to part of VacA from *Helicobacter pylori* present in strains M1 and 52472, although it was unclear whether this was complete in strain 52472. A probe 5P1h08 was designed to study this region in more depth in strains 52472 and M1 (**Table 5.1 D**). This probe identified the BAC clones MB1B12 and 5B3E12 which contain autotransporter homologues inserted between orthologues of cj1359 and cj1360c compared to the chromosome of NCTC 11168 (**Table 5.5**, **Table 5.6** and **Fig 5.9**). In strain M1 this BAC also contains differences in the downstream region, with hypothetical CDSs MB1B12_3 and MB1B12_4 found between orthologues of *ceuE* and cj1356c compared to the chromosome of strain NCTC 11168 (**Fig 5.10**). The predicted CDS MB1B12_4 shares high identity with a CDS previously identified in strain 81-176 [174].

Table 5.5: Predicted novel CDSs identified from BAC clone MB3E12

locus_tag	putative function	organism with match	SWALL	E-value	% id
MB1B12_3	hypothetical	-			
MB1B12_4	hypothetical	*C. jejuni*	Q6QNL7	7.5e-33	95.69
MB1B12_9c	autotransporter	*Helicobacter pylori*	O25579	1.3e-13	23.14

Table 5.6: Predicted novel CDSs identified from BAC clone 5B3E12

locus_tag	putative function	organism with match	SWALL	E-value	% id
5B3E12_5c	autotransporter pseudogene	*Helicobacter pylori*	Q9ZHT4	4.2e-11	22.05

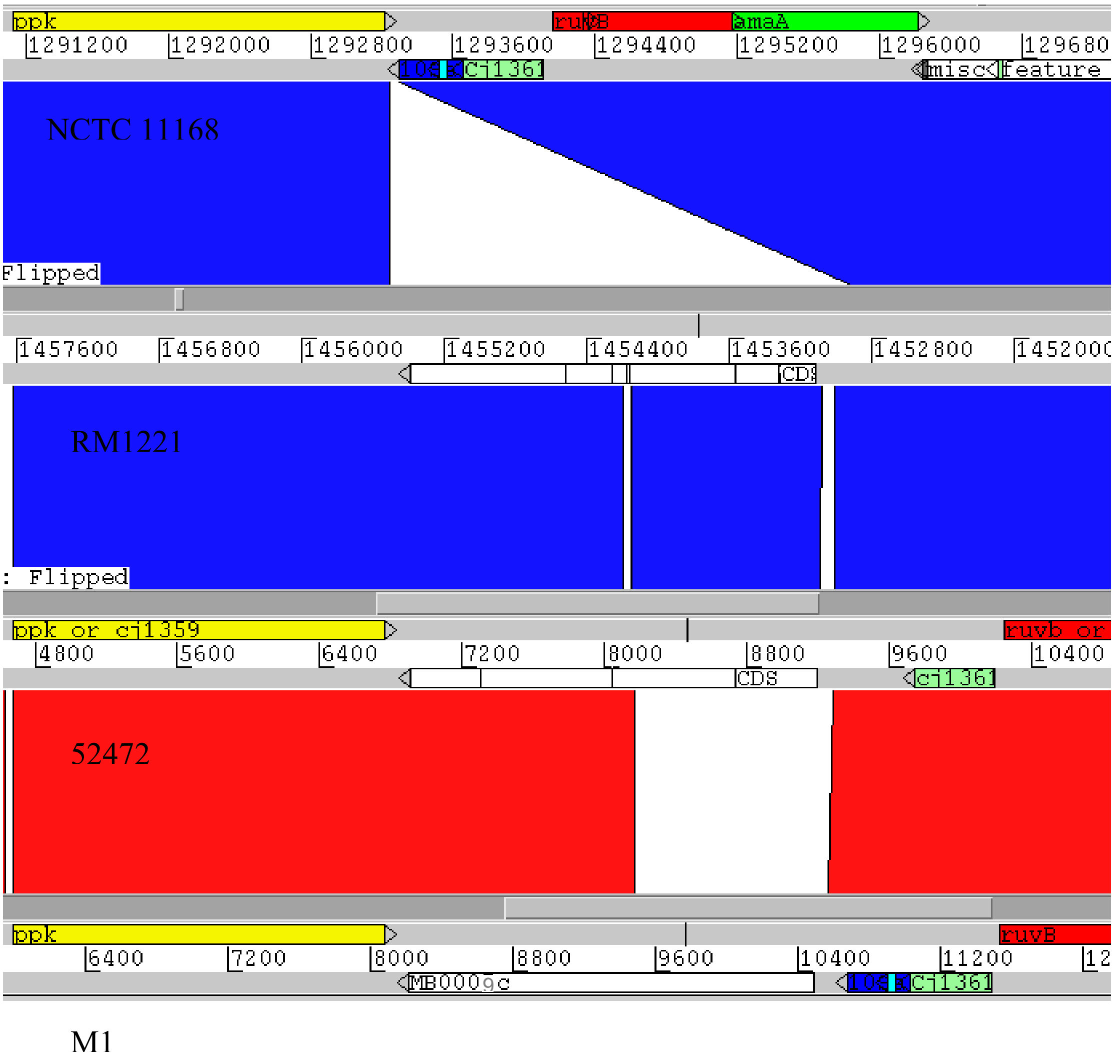


Fig 5.9: Blastn comparison of sequences from strain NCTC 11168, RM1221, 52472 BAC clone 5B3E12 and M1 BAC clone MB1B12. The comparison is viewed using ACT; blocks of red or blue indicate sequence homology with the colour intensity proportional to the percent id of the match. CDSs are represented by open boxes and are coloured according to functional category: yellow, central/ intermediary/ miscellaneous metabolism; white, pathogenicity/ adaptation/ chaperones; blue, stable RNA; light green, unknown; red, information transfer/ DNA modification; dark green, surface. An autotransporter is inserted between *ppk* and cj1361c relative to the sequence of NCTC 11168. In strains RM1221 and 52472 this is a pseudogene as indicated by the vertical black lines showing stop

codons interrupting the reading frame. In strain M1 only the C-terminal half of this CDS shows homology to strains RM1221 and 52472.

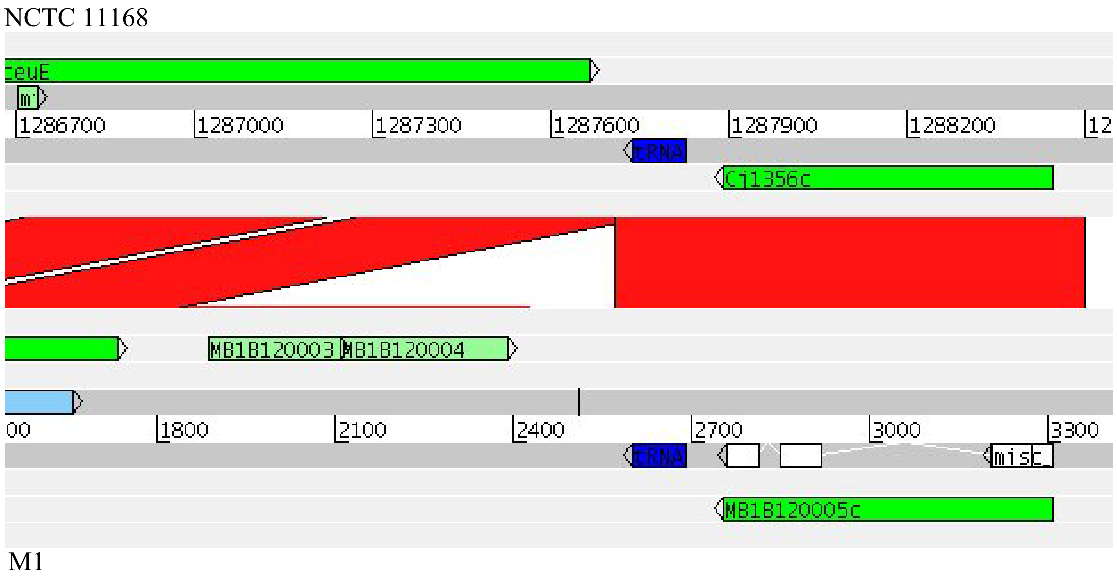


Fig 5.10: Blastn comparison of sequence from strain NCTC 11168 and strain M1 BAC clone MB1B12. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match, however single bp changes cannot be accurately represented. Forward and reverse DNA sequences are represented by dark grey lines; the three-frame forward and reverse DNA translations are represented by light grey lines. Features are represented by open boxes: pFam (light blue), tmhmm (white) and prosite (green) matches are marked on the DNA lines along with tRNA (dark blue). CDSs are marked on the translated reading frame lines and are coloured according to functional category: dark green, surface and light green, unknown. Two novel hypothetical CDSs are inserted between *ceuE* and a tRNA in strain M1.

Autotransporters contain an N-terminal passenger domain followed by a C-terminal autotransporter domain [175]. The putative autotransporters from strains M1 and 52472 share only 70.1% aa id overall with the initial signal peptide and autotransporter domain being strongly conserved but the passenger domain being different (**Fig 5.11**). Only the autotransporter domain shows homology to VacA from *Helicobacter pylori*. This region is also present in strain RM1221 (**Fig 5.9**) although in both strain RM1221 and strain 52472 the autotransporter is a pseudogene (CJE1549-CJE1552, 5B3E12_5), containing multiple stop codons within the reading frame.

```
                                 signal peptide
MB1B12_9c          1 MKKNASSKILLSLGVATLLYSGAFAAEITFNGDSDLDKYFDINEKDNVAT     50
                     ||||.||||||||||||||||||||.||.....||:..||:.|.||  ..
5B3E12_5c          1 MKKNTSSKILLSLGVATLLYSGAFAQEINLTESSDIGNYFEENGKD--IN     48

MB1B12_9c         51 FKN-ENYKNKQDVTFNIS-----TSAFDDAPEDTKINIDLG-NNSLTLKN     93
                     .|| :|||. ||::..:|     ...:|.|  |.:.||.:| ||:|:.|:
5B3E12_5c         49 LKNPDNYKG-QDLSIKMSVWDLPNDDYDSA--DYRFNIGIGKNNTLSFKH     95

MB1B12_9c         94 QMDYQGKTAALVKNFNVDAKDFKTTDIGLSYFNAGIINANFTMEGSGKDF    143
                     .   ..:..|.|.|.|..||:.|||||.|..|...:||.:.||..|   .
5B3E12_5c         96 N---NSEHPAYVTNLNATAKEVKTTDIVLQAFAPSVINGDLTMTSS---L    139

MB1B12_9c        144 DLGNIDKNKASSLLIFNGSRENTNDTVNGSLTVNGDFSTTNSAIVSMKSD    193
                     |....:..|.|.::::|.:.|  ..:.|||||:||:| |.:..:.:...:
5B3E12_5c        140 DEAITEDEKGSGIILYNETVE--GKSANGSLTINGNF-TADKTLFATYGN    186

MB1B12_9c        194 TFKVNGTATLKEAGLGFLSQSYSNLDVNDFIALRAKDIKTDTLNE--DTN    241
                     ..||||.|.|..:..|.:.:||::|:.|:.:.::|||...|.|.|  :.|
5B3E12_5c        187 FVKVNGAANLTNSNFGLMKRSYTDLEANNVVIVQAKDFNKDILEEKSNNN    236

MB1B12_9c        242 AGALILKTASSYINENLLNGDDYAA--YLDVTDDKKYGG---AFVDYKLS    286
                     ||||:||.||.||:.::...|...|  .:|::|:.|||.   ..||||||
5B3E12_5c        237 AGALLLKFASDYISTDVQGKDPLEAGTIIDISDEDKYGDGEKGLVDYKLS    286

MB1B12_9c        287 LKNCGGDKCLVINGGATAAAKNLTNQIAVDLEAITRIIDG-LDNEQ----    331
                     ::||||:||||||||.|||||:...|:.||::.|.::::. .|::|
5B3E12_5c        287 VQNCGGNKCLVINGGVTAAAKDKLVQLQVDIDTIDKLLENEFDSDQDEEW    336

MB1B12_9c        332 --AKKALQEQKTELEKLQQEAMQNGGKIDDEKYIDLVNKNSNLNLSANDK    379
                       ||:||::|||||:.:.:||.:|||||||||||||||||||.||::|||
5B3E12_5c        337 AKAKEALEKQKTELQTMLEEAEKNGGKIDDEKYIDLVNKNSN*NLNSNDK    386

MB1B12_9c        380 ASILVLRSITEQLGSIGADLASREGVKLALQIKKDTDNTGKSVSNFNSAS    429
                     ||||.|||||||||||||||||||||||||.||||||||||||||.||||
5B3E12_5c        387 ASILALRSITEQLGSIGADLASREGVKLALDIKKDTDNTGKSVSNLNSAS    436
                                                          autotransporter domain
MB1B12_9c        430 SAVNTTMNISNDVSIGSRVAMLNNPFGTYASKMNGLKFAALDSDMRPSYV    479
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
5B3E12_5c        437 SAVNTTMNISNDVSIGSRVAMLNNPFGTYASKMNGLKFAALDSDMRPSYV    486

MB1B12_9c        480 NEYTNSVWANAFGGANIIDGDSGAMYGATVGVDKQANDNVLWGAYFTYAN    529
                     |||||||||||||||||||||||||||||:||||||||:||||.||||.|
5B3E12_5c        487 NEYTNSVWANAFGGANIIDGDSGAMYGATIGVDKQANDDVLWGVYFTYTN    536

MB1B12_9c        530 AKIKDNNLEQKSDNFQLGMYSTINIAPQWELNLKAYAQVSPTKQDNVQVD    579
                     ||||||||||||||||||||||||||||||||||||||||||||||||:|
5B3E12_5c        537 AKIKDNNLEQKSDNFQLGMYSTINIAPQWELNLKAYAQVSPTKQDNVQID    586

MB1B12_9c        580 GAYNSDYTSKFLGLSANAGRVFDLSDNTLFIKPFAGVNYYFSYTPSHTEN    629
                     |||||||||||||||||||||||.||||||||||||| ||||||||||||
5B3E12_5c        587 GAYNSDYTSKFLGLSANAGRVFDFSDNTLFIKPFAGV-YYFSYTPSHTEN    635

MB1B12_9c        630 GAIAKDIDSMKNNSVSVEVGAEFRKYMNENSYIFVTPKIEQFVINSGDDY    679
                     ||||||||||||||||:|||||||||||||||||||||||||||||||||
5B3E12_5c        636 GAIAKDIDSMKNNSVSIEVGAEFRKYMNENSYIFVTPKIEQFVINSGDDY    685

MB1B12_9c        680 TANLAVNNAFFTSVEANNKKKTYGQIIVGGNVDFTNQLSMNLGFGAKQIL    729
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
5B3E12_5c        686 TANLAVNNAFFTSVEANNKKKTYGQIIVGGNVDFTNQLSMNLGFGAKQIL    735

MB1B12_9c        730 AGKVDNKNETYLSGQVGLKYKF    751
                     ||||||||||||||||||||||
5B3E12_5c        736 AGKVDNKNETYLSGQVGLKYKF    757
```

Fig 5.11: Alignment of the predicted autotransporters MB1B12_9c from strain M1 and 5B3E12_5c from strain 52472. The protein sequences were aligned using the EMBOSS program ‘water’, which uses the Smith-Waterman algorithm. The signal peptide and autotransporter domains are indicated by red lines above the protein sequence. Similarity between the two predicted CDSs

only occurs in the N-terminal signal peptide domain and the C-terminal autotransporter domain. The passenger domain which determines the function of the autotransporter is not conserved between the two strains.

The BAC sequence MB1B12 contains strain M1 contiguous pUC region 2g06p and also 3e11p which was also used as probe, as this pUC sequence contained a weak match to a putative haemolysin from *X. axonopodis*. The pUC sequences cover 55% of the novel DNA from this BAC. BAC sequence 5B3E12 contains strain 52472 contiguous pUC region 5h03q covering 59% of the novel sequence.

5.2.3.3 Two partner transporter

The probes 8P2e09 (**Table 5.1A**), designed from a hypothetical CDS with homology to *C. jejuni*, MP4c04 (**Table 5.1B**), designed from a haemagglutinin-related protein *Ralstonia solanacearum* and 5P5a12 (**Table 5.1D**), designed from a homologue of a periplasmic protein in *C. jejuni,* identified a region with limited identity to cj0967-cj0975 in strain NCTC 11168. In strains 81-176 and M1 the respective probes also identified this region in an alternative chromosomal location. The BAC sequences 8B1A11 (**Table 5.7**), MB5C4 (**Table 5.8**) and 5B5G5 (**Table 5.9**) contain regions of novel DNA located between cj0967 and cj0975 with respect to the NCTC 11168 chromosome (**Fig 5.12**). In 81-176 8B1A11_9 is a pseudogene with homology to cj0967, this is followed by a putative secreted protein 8B1A11_10 then a putative secretor protein 8B1A11_11 with 91% aa id to Cj0975. This region is similar overall to NCTC 11168 at the nucleotide level. In M1 MB5C4_5, a homologue of cj0967*,* is also a pseudogene, followed by MB5C4_6, a putative secreted protein with a haemagglutinin domain and MB5C4_7, a secretor HxuB homologue. There are also two homologues of an iron binding associated gene cj0241. In 5B5G5 all three genes are present as pseudogenes. These sequences indicate that the previously annotated NCTC 11168 CDSs cj0968-cj0974 are actually fragments of a single pseudogene.

Table 5.7: Predicted novel CDSs identified from BAC clone 8B1A11

locus_tag	putative function	organism with match	SWALL	E-value	% id
8B1A11_9	periplasmic protein pseudogene	*Campylobacter jejuni*	Q9PNW9	0	94.22
8B1A11_10	hypothetical	*Campylobacter jejuni*	Q9PNW7	2.5e-22	83.81
8B1A11_11	outer-membrane protein	*Campylobacter jejuni*	Q7AR82	2.7e-184	91.92

Table 5.8: Predicted novel CDSs identified from BAC clone MB5C4

locus_tag	putative function	organism with match	SWALL	E-value	% id
MB5C4_5	periplasmic protein pseudogene	*Campylobacter jejuni*	Q9PNW9	0	98.55
MB5C4_6	hypothetical	*Campylobacter jejuni*	Q9PNW7	1.9e-33	95.31
MB5C4_7	heme-hemopexin utilization protein	*Haemophilus influenzae*	AAQ10738	5.1e-18	24.19
MB5C4_8	hemerythrin-like protein	*Campylobacter jejuni*	Q9PIQ3	2.2e-09	34.09
MB5C4_9	hemerythrin-like protein	*Campylobacter jejuni*	Q9PIQ3	4.7e-08	36.06

Table 5.9: Predicted novel CDSs identified from BAC clone 5B5G5

locus_tag	putative function	organism with match	SWALL	E-value	% id
5B5G5_9	periplasmic protein pseudogene	*Campylobacter jejuni*	Q9PNW9	0	94.01
5B5G5_10	BpaA pseudogene	*Burkholderia pseudomallei*	AA019442	3.3e-09	25.2
5B5G5_11	heme-hemopexin utilization protein pseudogene	*Haemophilus influenzae*	AAQ10738	3.2e-15	23.91

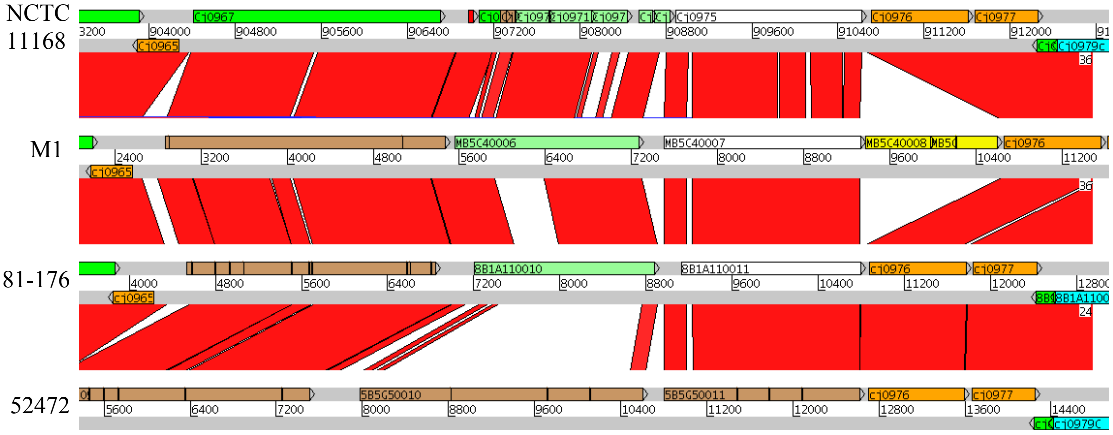


Fig 5.12: Blastn comparison of sequence from strain NCTC 11168, strain M1 BAC clone MB5C4, strain 81-176 BAC clone 8B1A11 and strain 52472 BAC clone 5B5G5. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. CDSs are represented by open boxes which are coloured according to functional category: Dark green, surface; orange, conserved hypothetical; brown, pseudogenes; light green, unknown; white, pathogenicity/ adaptation/ chaperones; yellow, central/ intermediary/ miscellaneous metabolism; blue, degradation of large molecules. This region is predicted to contain a two partner transport (TPS) system with secreted partner MB5C4_6, 8B1A11_10 and 5B5G5_10, and secretor partner MB5C4_7, 8B1A11_11 and 5B5G5_11.

The BAC clone sequence of 8B1A11 contains strain 81-176 contiguous pUC regions 2e09q, 4c05q, 6a01p and 6h03q. BAC clone sequence MB5C4 contains strain M1 contiguous pUC regions 4c04p, 2g07q and 1c08p. In BAC MB5C4 55% of the novel sequence is covered by pUC assemblies. The BAC sequence 5B5G5 contains strain 52472 contiguous pUC regions 8h04p and 8b01p covering 62% of the novel sequence.

This region appears to be duplicated in strains 81-176 and M1 only, being found between cj0500 and *hemH* on BAC sequences 8B1D8 and MB6A1. The BAC sequences 8B1D8 (**Table 5.10**) and MB6A1 (**Table 5.11**) possibly represent a recent duplication event. The BAC clone 8B1D8 contains the same predicted CDSs as 8B1A11 and the BAC clone MB6A1 contains the same predicted CDSs as MB5C4 but inserted between cj0500 and cj0503c replacing the pseudogene cj0501 without paralogues of the two iron binding associated genes in strain M1 (**Fig 5.13** and **Fig 5.14**). BAC sequence 8B1D8 contains strain

81-176 contiguous pUC regions 2e09q and 4c05q and MB6A1 contains strain M1 contiguous pUC sequences 4c04p and 2g07q.

Table 5.10: Predicted novel CDSs identified from BAC clone 8B1D8

locus_tag	putative function	organism with match	SWALL	E-value	% id
8B1D8_5	periplasmic protein pseudogene	*Campylobacter jejuni*	Q9PNW9	0	93.85
8B1D8_6	hypothetical	*Campylobacter jejuni*	Q9PNW7	2.6e-22	83.81
8B1D8_7	outer-membrane protein	*Campylobacter jejuni*	Q7AR82	3.2e-184	91.92

Table 5.11: Predicted novel CDSs identified from BAC clone MB6A1

locus_tag	putative function	organism with match	SWALL	E-value	% id
MB6A1_9	periplasmic protein pseudogene	*Campylobacter jejuni*	Q9PNW9	0	98.55
MB6A1_10	hypothetical	*Campylobacter jejuni*	Q9PNW7	1.7e-33	95.31
MB6A1_11	heme hemopexin utilization protein	*Haemophilus influenzae*	P45356	5.2e-18	24.76

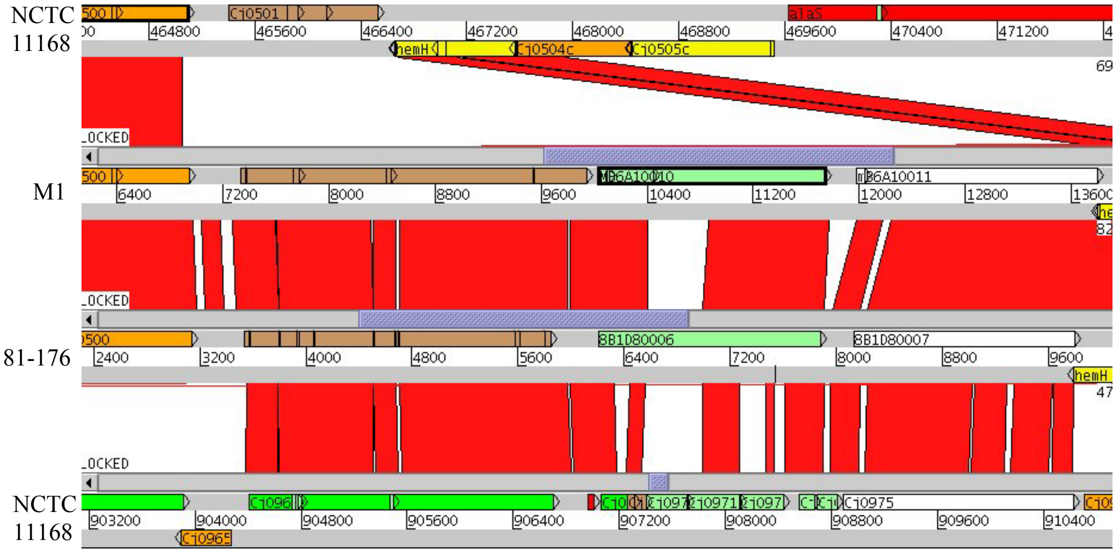


Fig 5.13: Blastn comparison of sequence from strain NCTC 11168, strain M1 BAC clone MB6A1 and strain 81-176 BAC clone 8B1D8. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. CDSs are represented by open boxes and are coloured according to functional category: orange, conserved hypothetical; brown, pseudogenes; dark green, surface; light green, unknown; white,

pathogenicity/ adaptation/ chaperones; yellow, central/ intermediary/ miscellaneous metabolism; red, information transfer/ DNA modification. The three central CDSs from M1 and 81-176 show homology to cj0967-cj0975 from NCTC 11168 and are inserted between cj0500 and *hemH* relative to NCTC 11168.

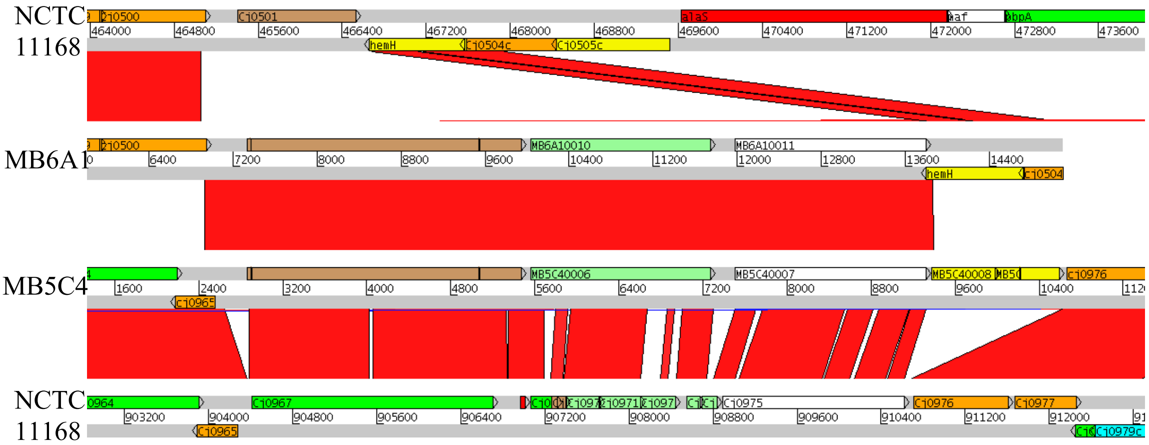


Fig 5.14: Blastn comparison of sequence from strain NCTC 11168 and strain M1 BAC clones MB6A1 and MB5C4. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. CDSs are represented by open boxes and are coloured according to functional category: dark green, surface; orange, conserved hypothetical; brown, pseudogenes; light green, unknown; white, pathogenicity/ adaptation/ chaperones; yellow, central/ intermediary/ miscellaneous metabolism; red, information transfer/ DNA modification; blue, degradation of large molecules. The three central CDSs from strain M1, which include a predicted TPS system, have been duplicated and are present at two sites relative to the chromosome of NCTC 11168.

The probes 8P6a02 (**Table 5.1A**) and MP3d04 (**Table 5.1B**), both designed from CDSs with homology to a putative adhesin from *Chromobacterium violaceum,* identified a similar region in another chromosomal location between cj0737 and cj0742 in BAC sequences 8B2A11 (**Table 5.12**) and MB5B1 (**Table 5.13**). This region is 81.7% similar at the nucleotide level between the two strains (**Fig 5.15**). There appears to be a larger portion of difference between the putative secreted proteins 8B2A11_3 and MB5B1_2 which share 89.5% aa id, MB5B1_2 is 1049 aa compared to 8B2A11_3 which is 615 aa. It is possible that 8B2A11_3 and 8B2A11_4 may have been a single CDS at one stage as they both show homology to MB5B1_2. There is also a frame shift in 8B2A11_3 which may denote that this CDS is no longer functional. Again these sequences indicate that NCTC 11168 CDSs cj0737-cj0741 probably represent fragments of a pseudogene. In strain NCTC 11168 cj0742 is a pseudogene but in strains 81-176 and M1 homologues of this gene are complete. In strain 81-176 BAC 8B2A11 downstream of the rDNA there is a replacement event; *cfrA* is missing which is consistent with previous findings (**Fig 5.16**) [14].

Table 5.12: Predicted novel CDSs identified from BAC clone 8B2A11

locus_tag	putative function	organism with match	SWALL	E-value	% id
8B2A11_3	periplasmic protein	*Campylobacter jejuni*	Q7AR90	1.3e-54	69.69
8B2A11_4	hypothetical	*Campylobacter jejuni*	Q9PPG7	5.7e-85	91.4
8B2A11_5	outer-membrane protein	*Campylobacter jejuni*	Q7AR82	1.6e-86	48
8B2A11_6	hypothetical	-			

Table 5.13: Predicted novel CDSs identified from BAC clone MB5B1

locus_tag	putative function	organism with match	SWALL	E-value	% id
MB5B1_2	adhesin	*Haemophilus influenzae*	Q48028	3.2e-06	23.89
MB5B1_3	hypothetical	*Campylobacter jejuni*	Q9PPG7	1.8e-3	95.83
MB5B1_4	outer membrane protein	*Campylobacter jejuni*	Q7AR82	5e-98	47.58
MB5B1_5	hypothetical	-			

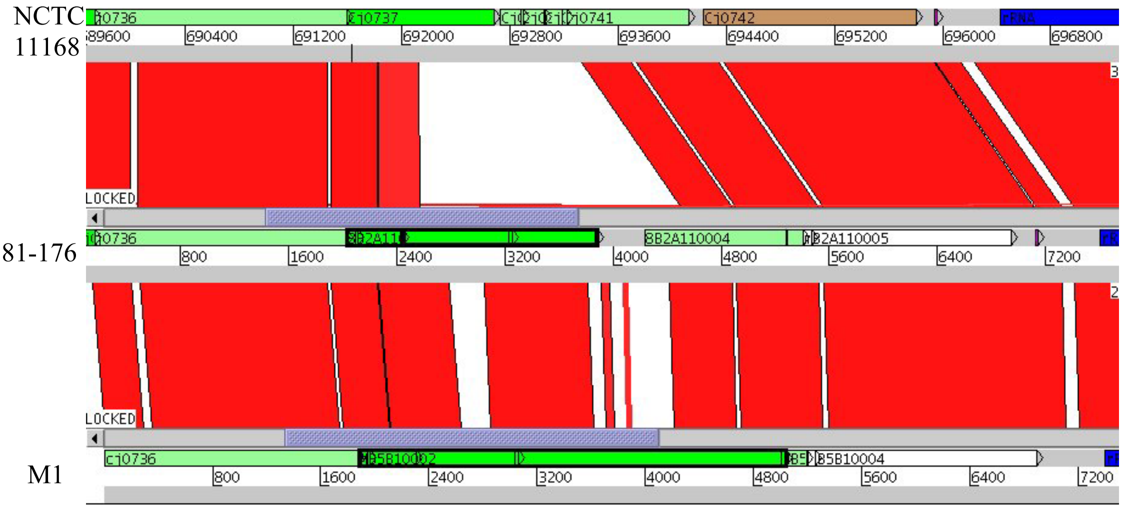


Fig 5.15: Blastn comparison of sequence from strain NCTC 11168, strain 81-176 BAC clone 8B2A11 and strain M1 BAC clone MB5B1. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. CDSs are represented by open boxes and are coloured according to functional category: light green, unknown; dark green, surface; brown, pseudogenes; white, pathogenicity/ adaptation/ chaperones; blue, rDNA. This region is predicted to contain a TPS system; 8B2A11_3 and MB5B1_2 are predicted to be secreted proteins, and 8B2A11_5 and MB5B1_4 are predicted to be secretor proteins. The N- and C- terminus of MB5B1_2 show homology to 81-176 and NCTC 11168 possibly suggesting that in 81-176 and NCTC 11168 the TPS is degrading.

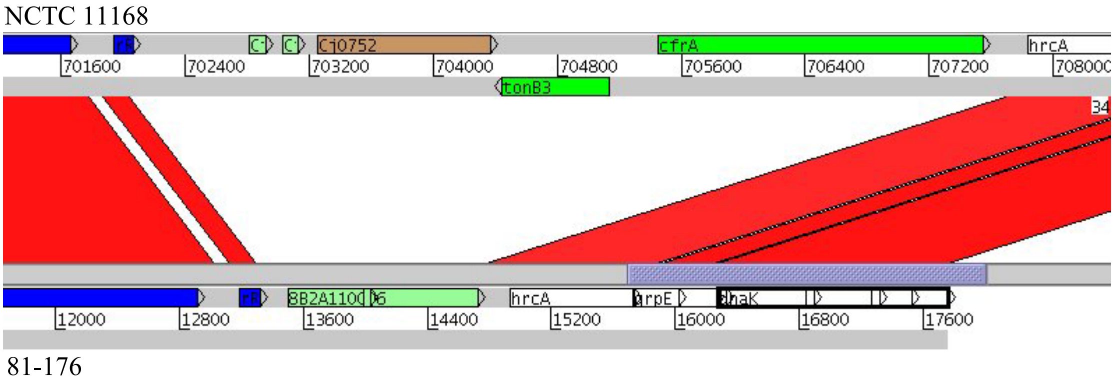


Fig 5.16: Blastn comparison of sequence from strain NCTC 11168 and strain 81-176 BAC clone 8B2A11. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. CDSs are represented by open boxes and are coloured according to functional category: blue, rDNA; light green, unknown; brown, pseudogenes; dark green, surface; white, pathogenicity/ adaptation/ chaperones. The region between rDNA and *hrcA* has been replaced by a hypothetical CDS in strain 81-176.

The BAC sequence 8B2A11 contains strain 81-176 contiguous pUC regions 7e09p, 2a01p and 6a02q covering 84% of novel sequence. The BAC sequence MB5B1 contains strain M1 contiguous pUC region 3d04q covering 63% of novel sequence.

5.2.4 Plasmid

Early data suggested that in strain 52472 the homologues of CDSs from the plasmid pTet identified in the pUC screen might be located on the chromosome. On closer examination this turned out not to be the case but had arisen on account of chimeric BAC sequences being generated incorporating phage DNA, chromosomal DNA and plasmid DNA. A probe, 5P5a06, was designed from a contiguous region from the pUC assemblies containing a homologue of VirB4, and identified the BAC sequence 5B4B1 which contains part of the contiguous pUC region 3c07q and all of the pUC regions 5e02q, 6b02q, 6c04p, 5a06p, 5h08p and 6a01q of strain 52472 which cover 54% of the novel sequence.

There is no evidence that BAC 5B4B1 is chromosomally located as the ends of this BAC clone sequence are complementary to pTet (**Fig 5.17**). There is an insert containing bacteriophage genes between pTet17 and pTet20, replacing half of pTet17 and all of pTet18 and pTet19. More work would need to be done to examine whether this represents a plasmid or whether it is chromosomally located.

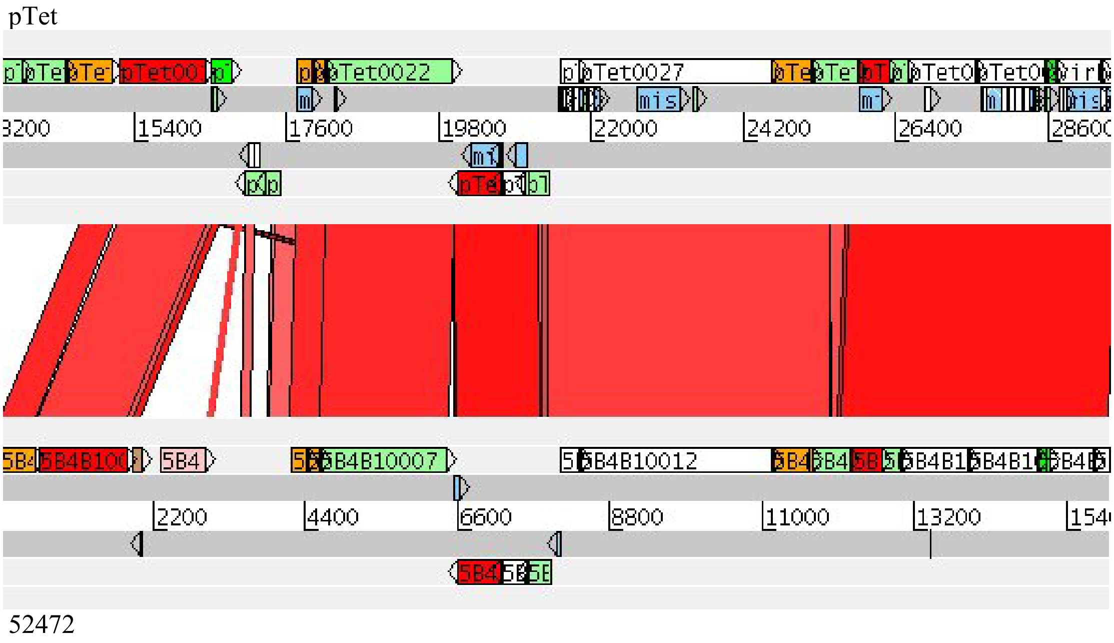


Fig 5.17: tblastx comparison of sequence from pTet and strain 52472 BAC clone 5B4B1. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark grey lines; DNA translations are represented by light grey lines. Open boxes represent features: pFam (blue), tmhmm (white), signalP (white) and prosite (green) matches are indicated on the DNA lines. CDSs are marked on one frame line irrespective of translational reading frame and are coloured according to functional category: light green, unknown; orange, conserved hypothetical; red, information transfer/ DNA modification; dark green, surface; brown, pseudogenes; pink, bacteriophage/ IS elements; white, pathogenicity/ adaptation/ chaperones. Strain 52472 contains sequence homologous to that of pTet with the exception of a small region containing a bacteriophage associated gene.

Other studies have identified partial CDSs with homology to the plasmid conjugation associated protein TraG, in strain 81116 [91] and in strain 43431 [85]. In the pUC assemblies it appeared that a TraG like protein might be present in strains M1 and 81-176. The probe 8P3c06 was designed to investigate this region further and identified the BACs 8B2B11 (**Table 5.14**) and MB2F11 (**Table 5.15**).

Table 5.14: Predicted novel CDSs identified from BAC clone 8B2B11

Locus_id	Putative function	Organism with match	SWALL	E-value	% id
8B2B11_15	hypothetical	-			
8B2B11_16	hypothetical	*Caulobacter crescentus*	Q9A4G5	6.7e-3	39.34
8B2B11_17	hypothetical	-			
8B2B11_18c	hypothetical	-			
8B2B11_19c	hypothetical	-			
8B2B11_20c	TraG fragment	*Escherichia coli*	P33790	1.5e-4	20.44
8B2B11_21c	TraN fragment	*Sphingomonas aromaticivorans*	O85935	2.3e-17	42

Table 5.15: Predicted novel CDSs identified from BAC clone MB2F11

Locus_id	Putative function	Organism with match	SWALL	E-value	% id
MB2F11_3	hypothetical	*Caulobacter crescentus*	Q9A4G5	7.3e-3	39.34
MB2F11_4	hypothetical	-			
MB2F11_5c	TraG pseudogene	*Escherichia coli*	P33790	1.1e-11	21.4
MB2F11_6c	TraN fragment	*Sphingomonas aromaticivorans*	O85935	1e-16	43.7

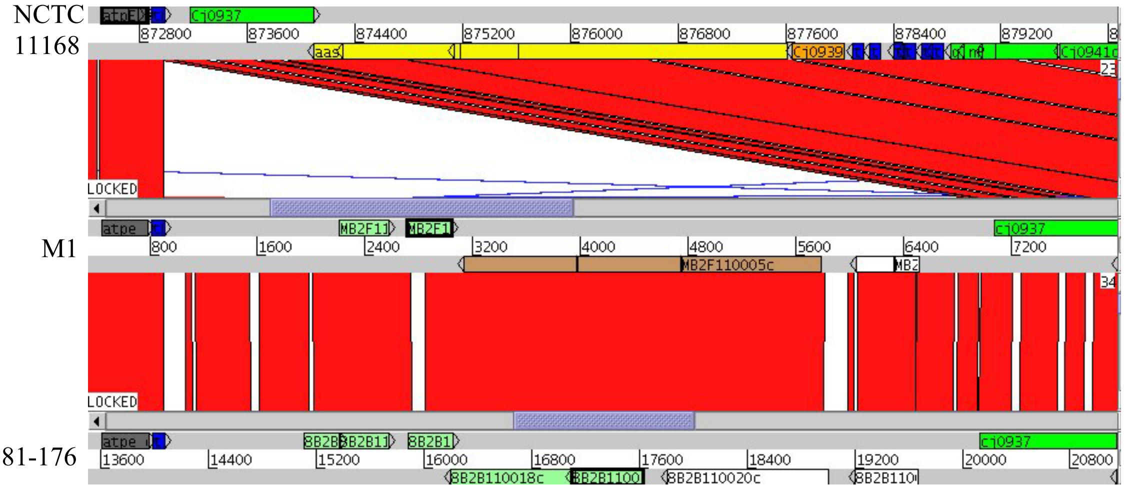


Fig 5.18: Blastn comparison of sequence from strain NCTC 11168, strain M1 BAC clone MB2F11 and strain 81-176 BAC clone 8B2B11. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. CDSs are represented by open boxes and are coloured according to functional category: grey, energy metabolism; dark blue, stable RNA; light green, unknown; brown, pseudogenes; white, pathogenicity/ adaptation/ chaperones; dark green, surface; yellow, central/ intermediary/ miscellaneous metabolism; orange, conserved hypothetical. In both strain M1 and strain 81-176 a

similar region containing a TraG homologue (MB2F11_5c and 8B2B11_20c) is inserted between tRNA and cj0937 relative to the chromosome of NCTC 11168.

In both strains 81-176 and M1 there is an insert between the tRNA-Leu, after cj0936 (*atpE*), and cj0937 relative to the chromosome of strain NCTC 11168 (**Fig 5.18**). These inserts are 97% similar with 182 bp differences which appear to affect the position of reading frames. There appear to be several hypothetical CDSs as well as CDSs with partial homology to TraN and TraG. TraN from *Sphingomonas aromaticivorans* is 704 aa but in strain 81-176 and M1 the matching CDSs are 150 and 153 aa long. TraG from *Escherichia coli* is 938 aa but in strain 81-176 and M1 the matching CDSs are 396 and 881 aa long, and in M1 the *traG* homologue is predicted to be a pseudogene. In strain RM1221 there is a TraG-like protein (CJE1107) of 529 aa located on a chromosomal island predicted to be of plasmid origin [9]. The *traG* fragments in strain 81-176 and strain M1 show 85% nucleotide identity to the gene predicted to encode a TraG-like protein in strain RM1221.

8B2B11 contains the 81-176 contiguous pUC regions 7f02p and 4a04q covering 76% of the novel sequence. MB2F11 contains the M1 contiguous pUC region 4e01q covering 66% of the novel sequence.

5.2.5 Chemotaxis

In strain 40671 a novel MCP-type chemotaxis protein was identified in the pUC assemblies. The probe 4P1d01 was designed and identified BAC 4B1D7. The novel CDS 4B1D7_13c, predicted to encode an MCP-type chemotaxis protein, was identified as being adjacent to an orthologue of cj0261c in 4B1D7 (**Table 5.16** and **Fig 5.19**). However, the N-terminal region of this protein was not identified in any of the BAC clones from the 40671 library. This predicted CDS shows high identity to the repeated C-terminal region of MCP-type chemotaxis proteins Cj0262c, Cj0144 and Cj1564, which includes the MCP signal domain.

The predicted protein shows 70% aa id to Cj0262c although this reflects the high identity of the signal transduction domain (**Fig 5.20**).

Table 5.16: Predicted novel CDSs identified from BAC clone 4B1D7

Locus_id	Putative function	Organism with match	SWALL	E-value	% id
4B1D7_13c	MCP transduction protein	*Campylobacter jejuni*	Q9PIN3	8.6e-104	70.85

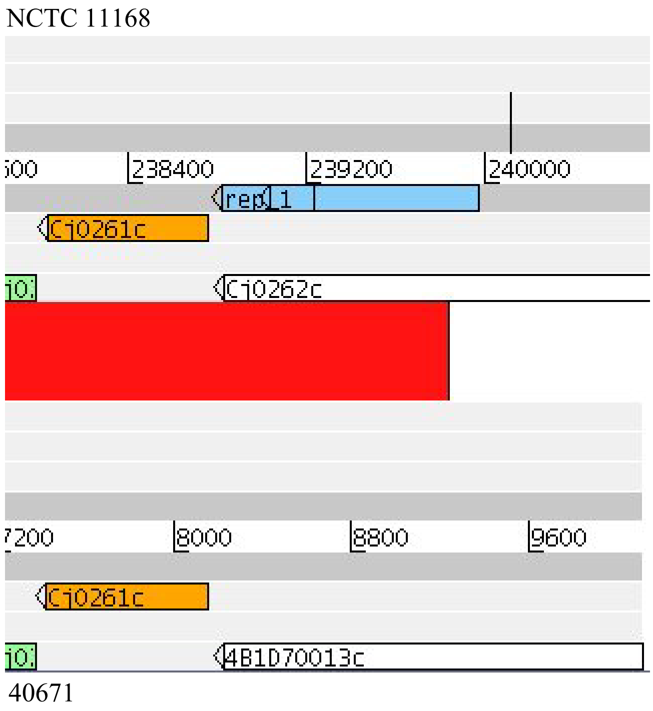


Fig 5.19: Blastn comparison of sequence from strain NCTC 11168 and strain 40671 BAC clone 4B1D7. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark grey lines; the three-frame forward and reverse DNA translations are represented by light grey lines. Features are represented by open boxes: a repeat unit (blue) is marked on the DNA line; CDSs are marked on the translated DNA lines and are coloured according to functional category: light green, unknown; orange, conserved hypothetical; white, pathogenicity/ adaptation/ chaperones. The CDS 4B1D7_13c is predicted to encode an MCP transduction protein. Repeat units are present in three MCP transduction proteins: cj0262c, cj0144 and cj1564. These repeat units contain the signal transduction domain of these chemotaxis proteins suggesting that the receptor portion of this protein is novel.

```
Cj0262c           37 FDSRN---NTYELLKDTQLKT-MQDVDAFFKSYAMSKRNGIQILANELTN     82
                     |||..   ..:..|:|:.:.| ::.:|..||:  ...::...|..:|.|.
4B1D70013c        14 FDSSGYAAYAFLYLQDSSILTHVESLDKNFKN--SDGKSVTMIFFDETTG     61

Cj0262c           83 R----------PDMSDEELINLIKVIKKVNDYDLVYVGFDNTGKNYQSDD    122
                     :          .:.|...:|..||...:..|.|.:::| ..:..||...:
4B1D70013c        62 KAGGIKSIHAPSNFSQLPIIEKIKKNARYGDLDTIFLG-SPSRLNYDGTE    110

Cj0262c          123 QILDLSKGYDTKNRPWYKAAKEAKKLIVTEPYKSAASGEVGLTYAAPFYD    172
                     .:                                      |:....|.::
4B1D70013c       111 FL--------------------------------------GINLGMPLFN    122

Cj0262c          173 RNGNFRGVVGGDYDLANFSTNV----LTVGKSDNTFTEVLDSEGTILFND    218
                     :.|.|.|:||..:|....|..:    |...|.|..|  ::..:|.|:.:.
4B1D70013c       123 KEGKFIGIVGFTFDFLEISETILDPKLDFYKDDLRF--LITDQGVIVIHK    170

Cj0262c          219 EVAKILTKTELSIN-------IANAIKANPALIDPRNQD-----------    250
                     ....|| ||...||       |.:|:|.:..||.....|
4B1D70013c       171 NKDAIL-KTLPEINQDASVQLIIDAVKNHKDLIIDNYVDLSGNLSYAGVA    219

Cj0262c          251 TLFTAKD--HQGVDYAIMCNSAFNPLFRICTITENKVYTEAVNSILMKQV    298
                     :..|..|  |..:.......|.|.||:.:             |.||   :
4B1D70013c       220 SFSTLGDSSHWSMVVTAPKKSIFAPLYEL-------------NFIL---I    253

Cj0262c          299 IVGIIAIIIALILIRFLISRSL-SPLAAIQTGLTSFFDFINYKTKNVSTI    347
                     .:.||.:|..||::.|.:...: |.|..|...|.:||||||:||||||||
4B1D70013c       254 SIAIIVLIAILIILYFCVKNIVGSKLPIIVNSLQNFFDFINHKTKNVSTI    303

Cj0262c          348 EVKSNDEFGQISNAINENILATKRGLEQDNQAVKESVQTVSVVEGGNLTA    397
                     |||||||.||:...|||||||||||||||||||||||:||.|||||||||
4B1D70013c       304 EVKSNDELGQMGKIINENILATKRGLEQDNQAVKESVETVHVVEGGNLTA    353

Cj0262c          398 RITANPRNPQLIELKNVLNKLLDVLQARVGSDMNAIHKIFEEYKSLDFRN    447
                     ||||||||||||||||||||||||||.|||||||||||||||||||||||
4B1D70013c       354 RITANPRNPQLIELKNVLNKLLDVLQVRVGSDMNAIHKIFEEYKSLDFRN    403

Cj0262c          448 KLENASGSVELTTNALGDEIVKMLKQSSDFANALANESGKLQTAVQSLTT    497
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
4B1D70013c       404 KLENASGSVELTTNALGDEIVKMLKQSSDFANALANESGKLQTAVQSLTT    453

Cj0262c          498 SSNSQAQSLEETAAALEEITSSMQNVSVKTSDVITQSEEIKNVTGIIGDI    547
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
4B1D70013c       454 SSNSQAQSLEETAAALEEITSSMQNVSVKTSDVITQSEEIKNVTGIIGDI    503

Cj0262c          548 ADQINLLALNAAIEAARAGEHGRGFAVVADEVRKLAERTQKSLSEIEANT    597
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
4B1D70013c       504 ADQINLLALNAAIEAARAGEHGRGFAVVADEVRKLAERTQKSLSEIEANT    553

Cj0262c          598 NLLVQSINDMAESIKEQTAGITQINDSVAQIDQTTKDNVEIANESAIISS    647
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
4B1D70013c       554 NLLVQSINDMAESIKEQTAGITQINDSVAQIDQTTKDNVEIANESAIISS    603

Cj0262c          648 TVSDIANNILEDVKKKRF    665
                     ||||||||||||||||||
4B1D70013c       604 TVSDIANNILEDVKKKRF    621
```

Fig 5.20: Alignment of the predicted MCP-type chemotaxis proteins Cj0262c and 4B1D7_13c. The EMBOSS program 'water' was used to align the sequences using the Smith-Waterman algorithm. The first red mark signifies the beginning of the repeat domain in NCTC 11168 and the second red mark signifies the beginning of the signal domain indicating that the entire repeat domain is not conserved but the entire signal transduction domain is. The receptor region of these proteins is not conserved suggesting that they may respond to different environmental signals.

5.2.6 Tetracycline resistance

In strain M1 the pUC assemblies identified a *tetO* gene. However, there was no similarity to pTet in any of the surrounding DNA and no other homologues of pTet CDSs were identified in the pUC screen. Probe MP1d11 was designed to locate the tetracycline resistance determinant, *tetO*. The probe identified BAC MB2G11 which contains the strain M1 pUC sequences 1d11p and 3a05q which cover 97% of the novel sequence. In MB2G11 there is a *tetO* determinant located in the middle of a gene cj0770c which is predicted to encode putative periplasmic protein (**Table 5.17** and **Fig 5.21**).

Table 5.17: Predicted novel CDSs identified from BAC clone MB2G11

Locus_id	Putative function	Organism with match	SWALL	E-value	% id
MB2G11_13c	hypothetical	-			
MB2G11_14c	hypothetical	*Enterococcus faecalis* tn916	Q56396	4.4e-14	66.66
MB2G11_15c	TetO	*Campylobacter jejuni*	Q84FM6	0	99.53
MB2G11_16c	TnpV fragment	*Clostridium difficile*	O05416	7.2e-6	46.42
MB2G11_17c	hypothetical	-			
MB2G11_18c	Rep fragment	*Treponema denticola*	Q9AQF2	2.6e-15	39.5

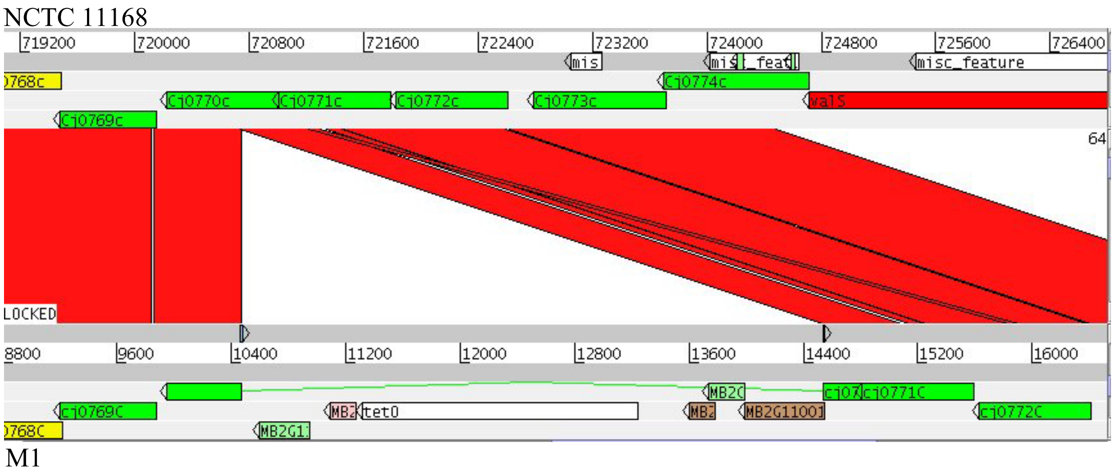


Fig 5.21: Blastn comparison of sequence from strain NCTC 11168 and strain M1 BAC clone MB2G11. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark grey lines; the three-frame reverse DNA translations are represented by light grey lines. Features are represented by open boxes: pFam (white) and prosite (green) matches are marked on the DNA lines; CDSs are marked on the translated frame lines and are coloured according to functional category: yellow, central/ intermediary/ miscellaneous metabolism; dark green, surface; light green, unknown; pink, bacteriophage/ IS elements; white, pathogenicity/ adaptation/ chaperones; brown, pseudogenes and partial genes; red, information transfer/ DNA modification. In strain M1 6 CDSs are inserted within cj0770c relative to strain NCTC 11168.

Tetracycline resistance determinants are often found on plasmids in *C. jejuni* [176], but this determinant is chromosomally located. The *tetO* gene is surrounded by fragments of genes normally found on plasmids for example MB2G11_18 showing 40% id to the central portion of a replication protein (Rep) from plasmid pTS1 of *T. denticola*. The fact that the insert is located in the centre of a gene is reminiscent of a transposon insertion although there are no transposase genes and there are also no inverted repeats which might be expected to be present in a functional transposon. There are however CDSs that show homology to CDSs present on transposons although none of these are predicted to encode a transposase. The predicted CDS MB2G11_16c shows 46% aa id to the C-terminal portion of TnpV located on a chloramphenicol-resistance transposon from *Clostridium perfringens* [177]. Interestingly only the central portion of this inserted region shows homology to pTet; this includes the *tetO* gene and also a small CDS downstream which shows 67% aa id to a hypothetical protein from the conjugative transposon tn916 from *Enterococcus faecalis* which carries a *tetM* determinant. This will be discussed further in chapter 6.

5.2.7 Hypothetical genes

There were a number of hypothetical genes identified in the pUC assemblies. Some of these were chosen to explore in more depth to see where they were inserted and if they were associated with other as yet unidentified genes or whether expanding these regions would give a functional context.

In strain 81-176 the probe 8P1d09 was used to identify a homologue of a hypothetical gene from *Clostridium perfringens.* The BAC sequence 8B1H2 contains the 81-176 pUC sequence 7f11p which covers the entire novel region so the BAC sequence added depth of coverage and positional information but no more novel sequence. 8B1H2 contains two CDSs that show homology to hypothetical proteins from other bacteria, one of which is a pseudogene. These CDSs are located between cj1687 and *secY* relative to the chromosome of NCTC 11168 (**Table 5.18** and **Fig 5.22**). The predicted pseudogene 8B1H2_4 has a sugar transport domain between aa residues 7-400 and an MFS_1 domain between aa residues 12-369. The MFS_1 domain is present in the major facilitator superfamily, a class of transporters capable of transporting small solutes in response to chemiosmotic ion gradients [178]. In addition this predicted CDS is predicted to have 12 transmembrane helices and a lipoprotein attachment site. This arrangement of domains is very similar to those predicted for cj1687, which encodes a putative efflux protein. CDS 8B1H2_3c shows homology to a hypothetical protein from *Rhizobium loti* and contains a pfam domain PF02129, x-pro dipeptidyl-peptidase, between aa residues 24-558. This domain is found in peptidases which perform a range of functions.

Table 5.18: Predicted novel CDSs identified from BAC clone 8B1H2

Locus_id	Putative function	Organism with match	SWALL	E-value	% id
8B1H2_3c	hypothetical	*Rhizobium loti*	Q98CJ2	4.5e-94	39.13
8B1H2_4c	hypothetical pseudogene	*Clostridium perfringens*	Q8XNB6	6.2e-42	33.49

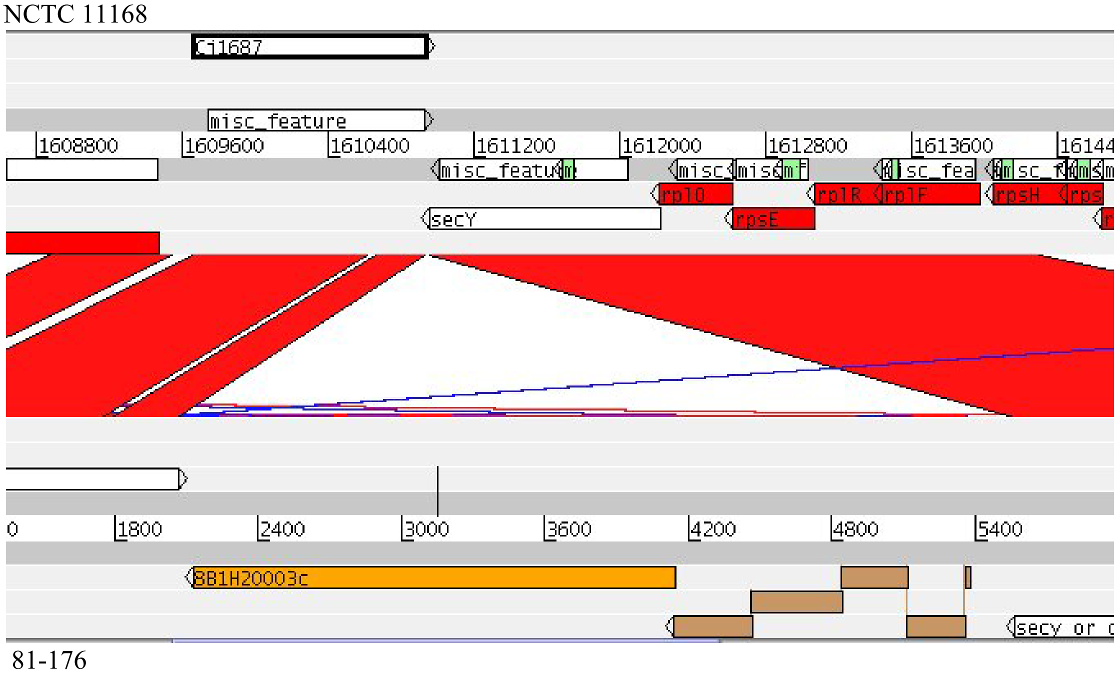


Fig 5.22: Blastn comparison of sequence from strain NCTC 11168 and strain 81-176 BAC clone 8B1H2. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark grey lines; the three-frame forward and reverse DNA translations are represented by light grey lines. Features are represented by open boxes: pFam (white) and prosite (green) matches are marked on the DNA lines; CDSs are marked on the translated frame lines and are coloured according to functional category: white, pathogenicity/ adaptation/ chaperones; orange, conserved hypothetical; brown, pseudogenes; red, information transfer/ DNA modification. Two CDSs are inserted between cj1687 and *secY* relative to strain NCTC 11168, one of which is a pseudogene.

Probe 4P1a12 was used to identify the BAC 4B2B1. This BAC contains pUC sequence 4P1a12q which covers 41% of the novel sequence. There are two hypothetical CDSs located between cj0341c and *uvrA* which are located in a region of very low G+C content, 23% (**Table 5.19** and **Fig 5.23**). In 4B2B1_5c there is a frame shift possibly suggesting that this is a pseudogene, or it actually might be two separate proteins as there is a plausible start site located within the second frame. There are 12 transmembrane helices for CDS 4B2B1_5c and 6 for 4B2B1_6c suggesting that these putative proteins may be membrane associated. There are no pfam family A matches in these two CDSs. This region is also present in RM1221 where it is annotated as two separate genes CJE0387 and CJE0388.

Table 5.19: Predicted novel CDSs identified from BAC clone 4B2B1

Locus_id	Putative function	Organism with match	SWALL	E-value	% id
4B2B1_5c	hypothetical	*Plasmodium falciparum*	Q8IBJ6	1e-08	31.5
4B2B1_6c	hypothetical	*Leishmania tarentolae*	Q34937	7e-4	27.11

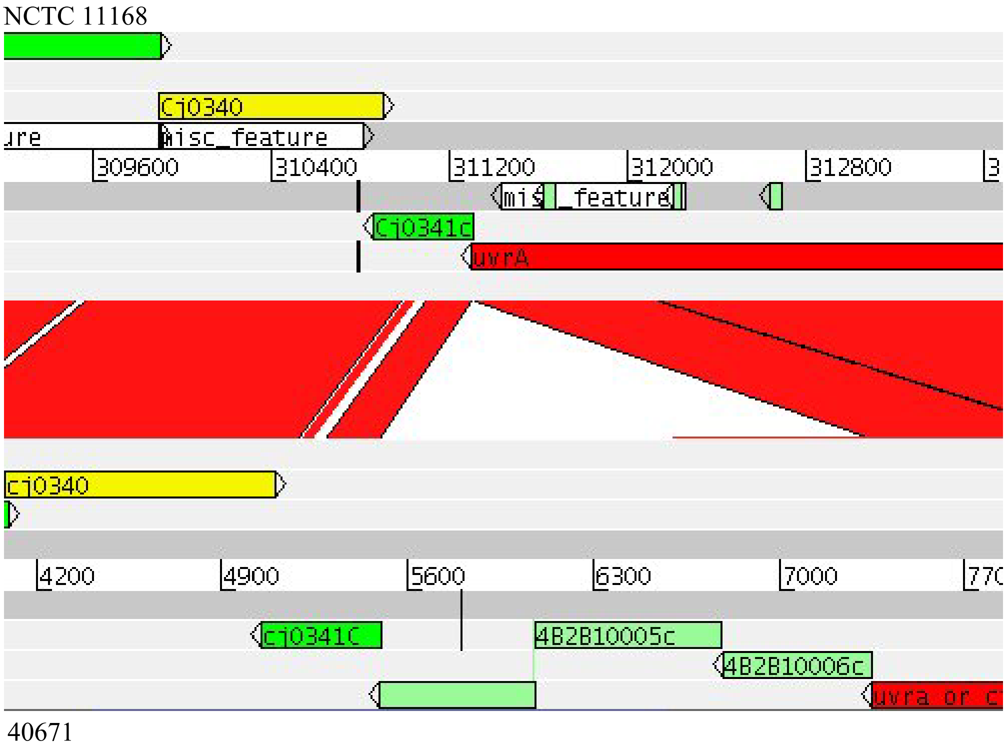


Fig 5.23: Blastn comparison of sequence from strain NCTC 11168 and strain 40671 BAC clone 4B2B1. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark grey lines; the three-frame forward and reverse DNA translations are represented by light grey lines. Features are represented by open boxes: pFam (white) and prosite (green) matches are marked on the DNA lines; CDSs are marked on the translated frame lines and are coloured according to functional category: yellow, central/ intermediary/ miscellaneous metabolism; dark green, surface; light green, unknown; red, information transfer/ DNA modification. Two hypothetical CDSs in strain 40671 are inserted between cj0341c and *uvrA* relative to strain NCTC 11168, one of which is predicted to contain a frame shift within the coding sequence.

Probe 4P1f05 was used to identify the BAC sequence 4B3G11. This BAC contains pUC sequence 4P1b12q which covers the entire novel region. There is one hypothetical CDS between *dnaX* and cj1161 relative to the chromosome of NCTC 11168 (**Table 5.20** and **Fig 5.24**). In NCTC 11168 there are a number of small hypothetical CDSs on the opposite strand not present in strain 40671. 4B3G11_12 contains a lipoprotein lipid attachment site and also a DAO domain at residues 78-118 containing Pyr_redox_2 at residues 78-161 and an amino oxidase domain at residues 532-632. These domains are found in FAD dependent oxidoreductases. Amine oxidases provide source of ammonium and can be involved in catabolism of polyamines. This CDS only matches to hypothetical proteins from other bacteria and not to characterized oxidoreductases. In RM1221 a pseudogene CJE1294 shows identity to 4B3G11_ 12 although the pseudogene CJE1294 is much shorter (519 bp compared to 1902 bp).

Table 5.20: Predicted novel CDSs identified from BAC clone 4B3G11

Locus_id	Putative function	Organism with match	SWALL	E-value	% id
4B3G11_12	hypothetical	*Chromobacterium violaceum*	Q7NTJ9	3.3e-126	52.36

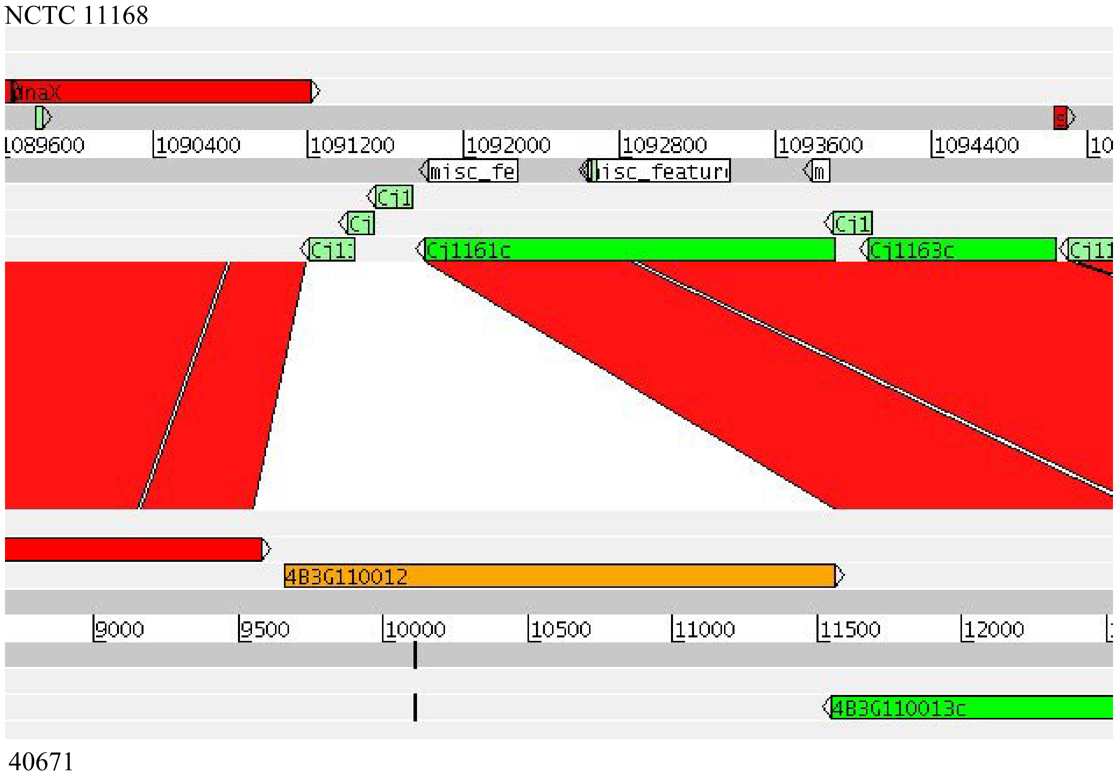


Fig 5.24: Blastn comparison of sequence from strain NCTC 11168 and strain 40671 BAC clone 4B3G11. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark grey lines; the three-frame forward and reverse DNA translations are represented by light grey lines. Features are represented by open boxes: pFam (white) and prosite (green) matches are marked on the DNA lines; CDSs are marked on the translated frame lines and are coloured according to functional category: red, information transfer/ DNA modification; light green, unknown; orange, conserved hypothetical; dark green, surface. In strain 40671 a conserved

hypothetical CDS replaces three hypothetical CDSs between *dnaX* and cj1161c relative to strain NCTC 11168.

5.2.8 Restriction Modification

The importance of RM systems has been discussed in chapter 4 (section 4.3.3). In its simplest form a restriction modification system consists of a restriction enzyme and a methylase protein with the same substrate specificity. Many predicted RM associated CDSs were identified in the pUC screen. The probe 4P1h09, designed from a CDS with homology to a hypothetical protein from *Helicobacter pylori* was used to identify BAC 4B3G8. The probe 5P4h09 was used to identify the location of the homologue of a serine-threonine protein kinase from *D. hansenii* (this is discussed later). These probes identified putative novel RM systems in strain 40671 and strain 52472 that are inserted in a similar location although the inserts are not the same (**Table 5.21**, **Table 5.22** and **Fig 5.25**).

Table 5.21: Predicted novel CDSs identified from BAC clone 4B3G8

Locus_id	Putative function	Organism with match	SWALL	E-value	% id
4B3G8_2c	type III RM r protein	*Helicobacter pylori*	O25923	4.1e-55	31.42
4B3G8_3c	hypothetical	-			
4B3G8_4c	type II RM methyltransferase	*Helicobacter pylori*	Q9ZJM2	3.6e-34	36.53

Table 5.22: Predicted novel CDSs identified from BAC clone 5B3G4

Locus_id	Putative function	Organism with match	SWALL	E-value	% id
5B3G4_6c	serine-threonine protein kinase	*Geobacillus kaustophilus*	Q8WQH7	1.9e-11	37.17
5B3G4_7c	methyltransferase	*Helicobacter pylori*	O25315	3e-46	48.31
5B3G4_8c	type III RM r protein	*Helicobacter pylori*	O25314	6e-79	55.34

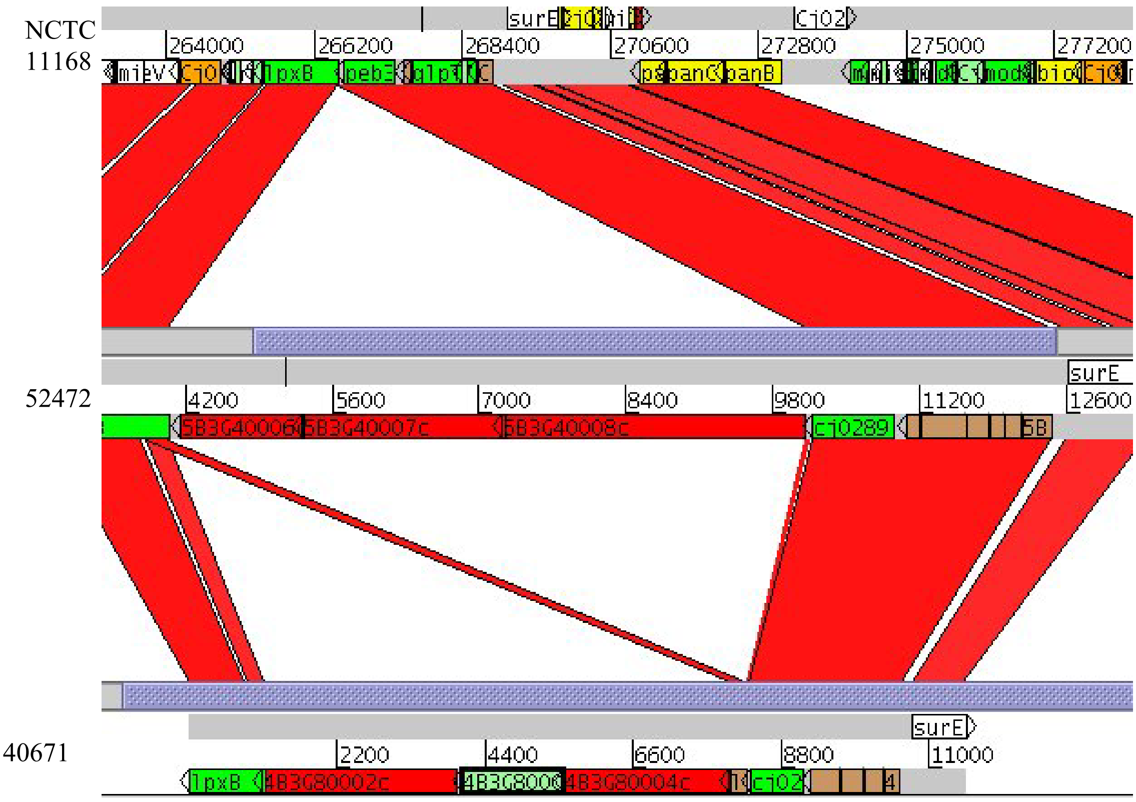


Fig 5.25: Blastn comparison of sequence from strain NCTC 11168, strain 52472 BAC clone 5B3G4 and strain 40671 BAC clone 4B3G8. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. CDSs are represented by open boxes and are coloured according to functional category: white, pathogenicity/ adaptation/ chaperones; orange, conserved hypothetical; dark green, surface; light green, unknown; red, information transfer/ DNA modification; brown, pseudogenes and partial genes; yellow, central/ intermediary/ miscellaneous metabolism. In strain 52472 and 40671 novel restriction modification loci are inserted between cj0289 and *lpxB*. In strain 40671 *lpxB* is interrupted and the N-terminus duplicated, this is indicated by the diagonal red block extending across the novel insert.

4B3G8 contains three novel predicted CDSs replacing the C-terminus of *lpxB,* although a complete copy of *lpxB* is present downstream of the novel insert. *lpxB* encodes a lipid-A-disaccharide synthase, a major component of the cell wall, and if disrupted is likely to be lethal to the bacterium. The novel predicted CDSs are predicted to encode a homologue of a methylase protein from *Helicobacter pylori*, a hypothetical protein with peptidase domain (peptidase_c14 residues 3-224) and a homologue of a restriction protein from *Helicobacter pylori*. BAC 4B3G8 contains the pUC sequence 4P3a10q which covers 55% of novel sequence.

5B3G4 contains three novel predicted CDSs inserted between *peb3* and *lpxB*. The novel CDSs are predicted to encode a homologue of a restriction protein from *Helicobacter pylori*, a homologue of an adenine specific methyltransferase from *Helicobacter pylori* followed by a putative protein kinase with a tyrosine protein kinase specific active-site signature. This predicted CDS (5B3G4_6c) carries a protein kinase pfam domain at aa residues 14-300. The arrangement of a protein kinase associated with RM genes has been found in the phage growth limitation system of *Streptomyces coelicolor* [179] and will be discussed in section 5.3.5. The BAC 5B3G4 includes the strain 52472 pUC sequences 4h09p, 5d07p, 8c04p and 5c07q which cover 96% of the novel DNA.

5.2.9 Capsule

In strain 40671 there were a number of CDSs with homology to hypothetical proteins from other bacteria which are located within polysaccharide biosynthesis loci of these bacteria. The probe 4P1a10 was designed to identify a CDS predicted to encode a homologue of DmhA from *Y. pseudotuberculosis.* The probe 4P1e06 was designed to identify a CDS with homology to a hypothetical protein from *Pseudomonas syringae.* These probes both identified the capsule locus in this strain. Two BAC clones were sequenced to span the extent of this novel region. The BACs 4B1B2 and 4B3H2 contained the pUC sequences 4P1a10p, 4P1e06p, 4P3f04p, 4P1b06q, 4P3g02p, 4P3c01q, 4P3g08p and 4P2e08p which cover 62% of the novel sequence.

The capsule region is large, containing 33838 bp which runs from the N-terminal portion of strain NCTC 11168 cj1418c to the C-terminal portion of *kpsD* (**Table 5.23**, **Fig 5.26** and **Fig 5.27**). Within this region there is very little homology between the strains

40671 and NCTC 11168. The region between cj1418-cj1420 seems conserved then there is an alternate form of cj1421/cj1422 sugar transferase. Between cj1422 and cj1423 there is an approximately 12 Kb insert of novel CDSs, cap5-cap18. cj1423-cj1425 are conserved, cj1426 is missing and cj1427 is present (**Fig 5.27**). A GDP-*mannoheptose-4,6 dehydratase (dmhA)* is inserted before a divergent *fcl* as in strain 81-176. cj1429 is missing, cj1430 is present, downstream of which there is an approximately 8 kb insert between cj1430 and cj1442 (cap26-28), replacing genes in NCTC 11168 between cj1430 and cj1442. This later half appears more similar to strain 81-176. Interestingly Cap26c contains a glycosyltransferase domain between aa residues 5-241 and an adhesion associated domain between aa residues 532-656; the F5/8 type C domain (PF00754).

Table 5.23: Predicted novel CDSs identified from BAC clones spanning the capsular biosynthesis locus of strain 40671.

Locus_id	Putative function	Organism with match	SWALL	E-value	% id
4Bcap_4c	sugar transferase	*Campylobacter jejuni*	Q9PMN6	1.1e-107	51.12
4Bcap_5c	sugar transferase	*Campylobacter jejuni*	Q5M6U5	3.3e-2	34.23
4Bcap_6c	polysaccharide biosynthesis protein	*Campylobacter jejuni*	Q5HT01	5.9e-34	28.47
4Bcap_7c	sugar transferase	*Campylobacter jejuni*	Q5M6U2	6.3e-05	22.1
4Bcap_8c	hypothetical	*Actinobacillus suis*	Q84CG7	6.9e-42	53.31
4Bcap_9c	hypothetical	*Escherichia coli*	Q8L0V7	2.3e-56	39.50
4Bcap_10c	hypothetical	*Actinobacillus suis*	Q84CG6	2.1e-26	57.94
4Bcap_11c	nucleotidyl transferase	*Yersinia enterocolitica*	Q692L3	2.7e-33	48.55
4Bcap_12c	hypothetical	*Pseudomonas syringae*	Q889N9	1.3e-19	58.76
4Bcap_13c	hypothetical	-			
4Bcap_14c	c-terminus hypothetical	*Yersinia enterocolitica*	Q692L0	0.21	29.41
4Bcap_15c	N-terminus hypothetical	*Yersinia enterocolitica*	Q692L0	1.9e-12	29.38
4Bcap_16c	hydrolase	*Yersinia enterocolitica*	Q692L1	1e-39	63.31

Locus_id	Putative function	Organism with match	SWALL	E-value	% id
4Bcap_17c	hypothetical	*Pseudomonas syringae*	Q889P2	1.4e-28	40.67
4Bcap_18	sugar transferase	*Campylobacter jejuni*	Q5M6U2	1.9e-51	41.48
4Bcap_19c	heptose-1-phosphate guanosyltransferase	*Campylobacter jejuni*	Q5M6R1	5.8e-72	90.95
4Bcap_20c	phosphoheptose isomerase	*Campylobacter jejuni*	Q5M6R0	3.1e-71	97
4Bcap_21c	sugar kinase	*Campylobacter jejuni*	Q5HSZ4	3.8e-127	98.23
4Bcap_22c	UDP-glucose 4-epimerase	*Campylobacter jejuni*	Q6EF85	4.8e-116	99.68
4Bcap_23c	GDP-mannoheptose-4,6 dehydratase	*Campylobacter jejuni*	Q6EF84	1.6e-104	98.24
4Bcap_24c	fucose synthetase	*Campylobacter jejuni*	Q9PMM9	2.4e-74	59.1
4Bcap_25c	nucleotidyl-sugar epimerase	*Campylobacter jejuni*	Q5M6T7	4.6e-71	93.92
4Bcap_26c	sugar transferase	*Campylobacter jejuni*	Q5M6T5	6.4e-27	26.28
4Bcap_27	sugar transferase	*Campylobacter jejuni*	Q5M6M6	2e-107	54.36
4Bcap_28c	hypothetical	*Actinobacillus suis*	Q84CH0	2e-125	42.94
4Bcap_29c	sugar transferase	*Campylobacter jejuni*	Q5M6S1	2.9e-194	92.63

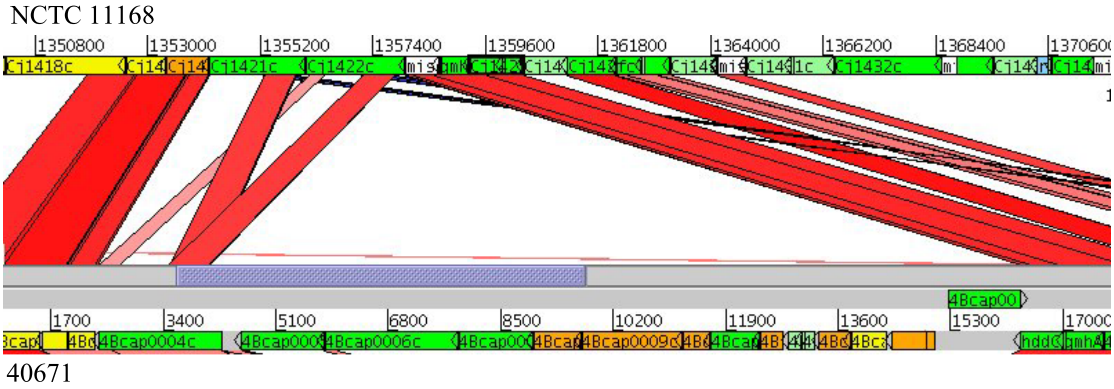


Fig 5.26: tblastx comparison of sequence from strain NCTC 11168 and strain 40671 capsular polysaccharide locus. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. CDSs are represented by open boxes and are coloured according to functional category: yellow, central/ intermediary/

miscellaneous metabolism; orange, conserved hypothetical; dark green, surface; light green, unknown. In strain 40671 there is an insert of approximately 12 Kb between cj1422c and *hddC*.

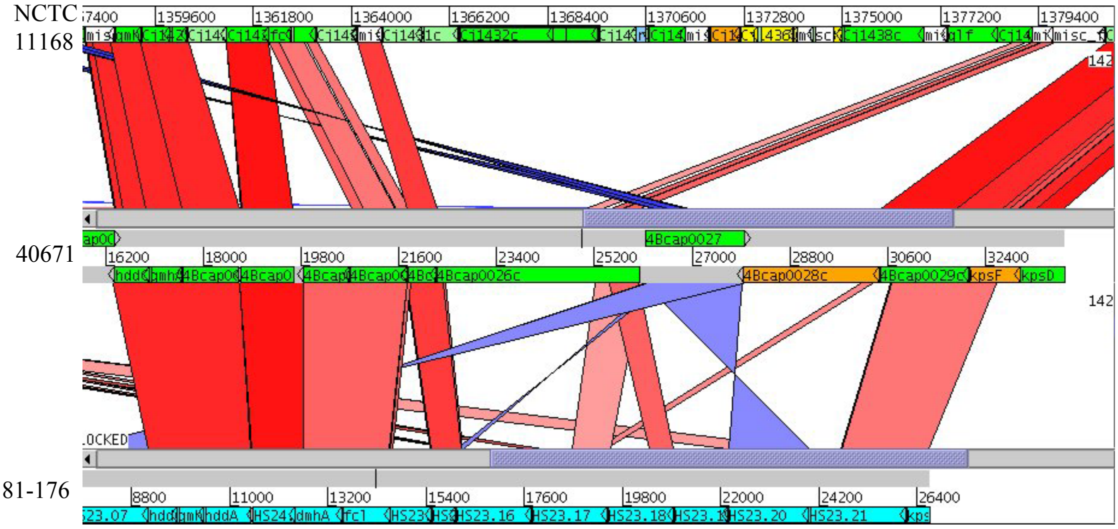


Fig 5.27: tblastx comparison of sequence from strain NCTC 11168, strain 40671 and strain 81-176 capsular polysaccharide locus. The comparison of *hddC* to *kpsF* is viewed using ACT; blocks of red or blue indicate sequence homology with the colour intensity proportional to the percent id of the match. CDSs are indicated by open boxes and, for strain NCTC 11168 and strain 40671, are coloured according to functional category: dark green, surface; light green, unknown; orange, conserved hypothetical; yellow, central/ intermediary/ miscellaneous metabolism. The sequence of the capsular polysaccharide locus of strain 81-176 has been previously determined by Karlyshev *et al.* 2005 [151] (accession number BX545858). CDSs from strain 81-176 are coloured blue irrespective of functional category. This figure illustrates the fact that there are inversions, insertions and deletions within the capsular biosynthesis locus

There are 9 homopolymeric tracts within the capsule region of this strain. The first is G(10) in 4Bcap_3c, cj1420 which is known to be variable in other strains. G(9) in 4Bcap_4c; G(9) in between 4Bcap_4c and 4Bcap_5c, G(9) in 4Bcap_7c; G(10) in 4Bcap_17c; G(10) in 4Bcap_18 which is known to vary in 81-176 and HS:36; G(11) in 4Bcap_23c also known to vary; G(7) in 4Bcap_26c shows some homology to HS23.17 which does not vary; G(9) in 4Bcap_27 shows homology to HS23.20 which does vary in some strains [151]. Phase variation using slipped-strand mispairing at homopolymeric tracts has been shown to be one of the ways *Campylobacter* can vary its surface structures. There are 5 homopolymeric tracts in the capsular region of strain NCTC 11168 that have been shown to vary. No variation is seen here, as these are single BAC subclones from the chromosome, and not subject to the *C. jejuni* cytoplasmic context [8].

5.2.10 Bacteriophage

Many bacteriophage associated CDSs were identified in the pUC screen of strain 52472. The probes 5P3h01, designed to identify a CDS with homology to a hypothetical protein from bacteriophage D3112, and 5P3e01, designed to identify a CDS with homology to a hypothetical protein from *Helicobacter hepaticus,* identified the BAC 5B6C12. Another BAC, 5B6F7, containing bacteriophage sequences was identified possibly due to cross reactivity of the probe. On comparing the pUC assemblies to these BAC sequences, many of the bacteriophage related CDSs showed more that 85% nucleotide id to the BAC sequences. Of the full length matches 5B6C12 contains pUC contigs 5P7h02q, 5P7b11p, 5P5e04p, 5P3g06p, 5P2c11q and 5P2b12q which cover the entire BAC sequence. The BAC 5B6F7 contains pUC contigs 5P2e12p, 5P2c11q, 5P3g06p, 5P4g03q and 5P7h02q which cover 93% of the novel sequence.

In strain 52472 the sequence from the pUC library identified many bacteriophage genes. In the sequence of strain NCTC 11168 there were no phage remnants which is relatively rare for a bacterial genome [8]. In comparison the strain RM1221 has three Mu-like bacteriophage insertions [9]. BAC 5B6F7 contains approximately 24 Kb of phage DNA inserted in the middle of hypothetical CDS cj1305c relative to the chromosome of strain NCTC 11168 (**Table 5.24**). The other end of this BAC does not contain any DNA matching to strain NCTC 11168 so the extent of the insert and insertion point relative to the strain NCTC 11168 chromosome can not be determined. Parts of this phage show similarity to the integrated 37 Kb Mu-like phage of RM1221 in position 207005-244247 (**Fig 5.28**). The

BAC 5B6C12 has a phage insert between *panB* and cj0299 although this region is in three pieces; 10815 bp, 7534 bp and 13655 bp. These two bacteriophage inserts seem very similar to each other at the right hand end where the bacteriophage structural proteins are encoded but are divergent in the hypothetical CDSs at the left hand end of the bacteriophage (**Fig 5.29**).

Table 5.24: Predicted novel CDSs identified from BAC clone 5B6F7

Locus_id	Putative function	Organism with match	SWALL	E-value	% id
5B6F7_1	emm-like protein	RM1221	Q5HTH7	7.8e-13	100
5B6F7_2	site-specific DNA-methyltransferase	RM1221	Q5HTH8	3.8e-102	97.61
5B6F7_3	hypothetical	RM1221	Q5HTH9	9.3e-39	100
5B6F7_4	site-specific recombinase	RM1221	Q5HTI1	1.4e-143	100
5B6F7_5c	hypothetical	-			
5B6F7_6c	hypothetical	Bacteriophage D3112	Q6TM76	1.4e-22	29.48
5B6F7_7c	hypothetical	RM1221	Q5HWS5	9.4e-34	98.08
5B6F7_8c	lipoprotein	RM1221	Q5HWS3	6.6e-27	98.72
5B6F7_9c	hypothetical	RM1221	Q5HWS2	1.1e-36	97.35
5B6F7_10c	hypothetical	RM1221	Q5HWS1	3.2e-52	99.23
5B6F7_11c	hypothetical	-			
5B6F7_12c	hypothetical	-			
5B6F7_13c	hypothetical	-			
5B6F7_14c	major head subunit protein	Bacteriophage D3112	Q6TM67	1.5e-14	35.59
5B6F7_15c	hypothetical	-			
5B6F7_16	hypothetical	-			
5B6F7_17	baseplate assembly protein v	RM1221	Q5HWS6	2.4e-74	98.57
5B6F7_18	hypothetical	RM1221	Q5HWS7	1.1e-22	98.41
5B6F7_19	baseplate assembly protein w	*Campylobacter coli*	Q9K5E0	4.7e-35	97.92
5B6F7_20	baseplate assembly protein J	RM1221	Q5HWS9	5.4e-129	98.2
5B6F7_21	tail protein	RM1221	Q5HWT0	2.9e-70	91.26
5B6F7_22	tail fiber protein h	RM1221	Q5HWT1	1.3e-80	75.59
5B6F7_23	hypothetical	RM1221	Q5HWT2	1.3e-55	95.83
5B6F7_24	hypothetical	RM1221	Q5HWT3	5e-52	98.37
5B6F7_25	hypothetical	RM1221	Q5HWT4	1.8e-123	98.52
5B6F7_26	major tail sheath protein	RM1221	Q5HWT5	1.5e-144	96.97

Locus_id	Putative function	Organism with match	SWALL	E-value	% id
5B6F7_27	major tail tube protein	RM1221	Q5HWT6	1.9e-20	41.92
5B6F7_28	hypothetical	-			
5B6F7_29	tail tape measure protein	RM1221	Q5HWU0	5.2e-22	26.06
5B6F7_30	tail protein	RM1221	Q5HWR0	2.7e-25	57.26
5B6F7_31	tail protein d	RM1221	Q5HWQ8	1.8e-49	47.1
5B6F7_32	DNA adenine methylase	RM1221	Q5HWU2	1.5e-103	98.52
5B6F7_33c	hypothetical	RM1221	Q5HWU3	5.5e-18	96.67
5B6F7_34c	hypothetical	RM1221	Q5HWU6	1.1e-31	97.17
5B6F7_35c	repressor protein	RM1221	Q5HWU7	1.1e-78	97.61

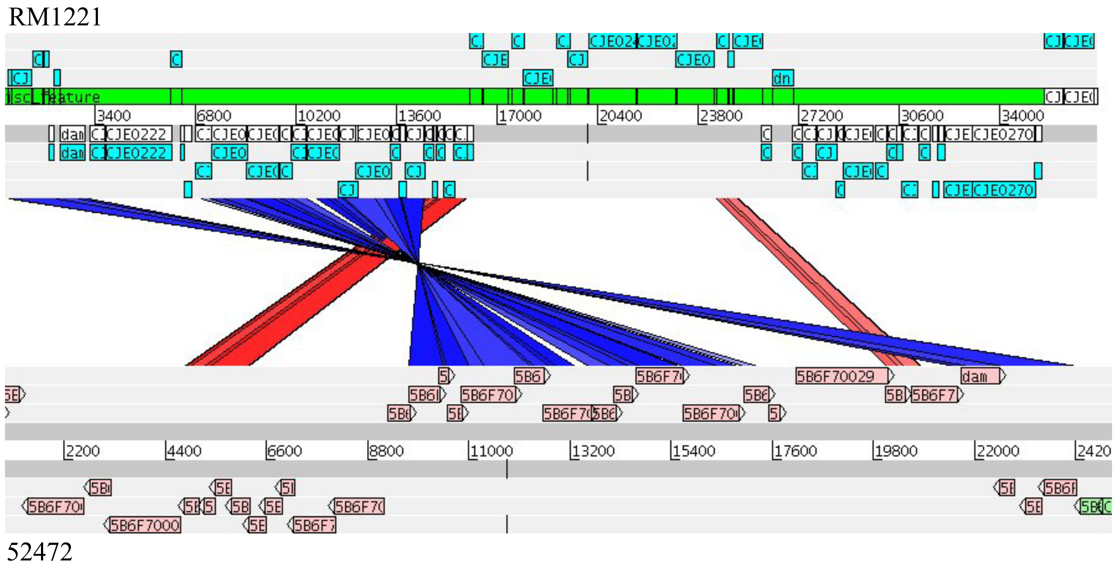


Fig 5.28: tblastx comparison of bacteriophage sequence from strain RM1221 and strain 52472 BAC clone 5B6F7. The comparison is viewed using ACT; blocks of red or blue indicate sequence homology with the colour intensity proportional to the percent id of the match. Blocks of blue indicate that the homologous region is on the opposite strand. Forward and reverse DNA sequences are represented by dark grey lines; the three-frame forward and reverse DNA translations are represented by light grey lines. Features are represented by open boxes: CDSs are marked on the translated frame lines; CDSs from RM1221 are coloured blue and bacteriophage CDSs from strain 52472 are coloured pink. Many of the CDSs from the integrated Mu-like bacteriophage, located between 207005-244247 bp, on the chromosome of strain RM1221 (Fouts *et al.* 2005 [9], accession number CP000025) are conserved in strain 52472.

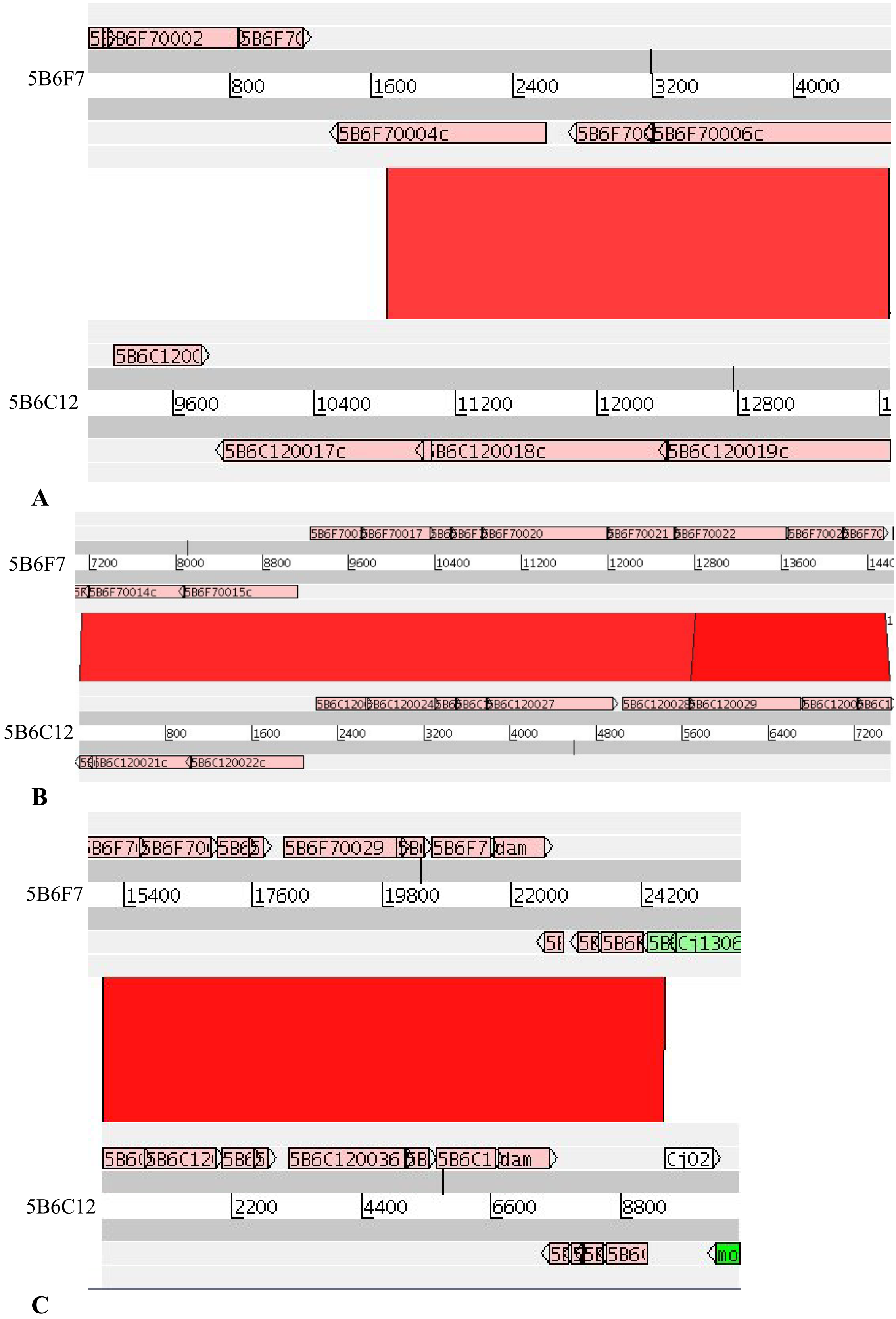


Fig 5.29: WUBLASTN comparison of sequence from strain 52472 BAC clones 5B6F7 and 5B6C12. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are

represented by dark grey lines; DNA translations are represented by light grey lines. CDSs are marked by open boxes on one translated frame line irrespective of reading frame. CDSs are coloured according to functional category: pink, bacteriophage; light green, unknown; dark green, surface; white, pathogenicity/ adaptation/ chaperones. The sequence of BAC 5B6C12 is in three contigs A (13655 bp), B (7534 bp) and C (10815 bp). The CDSs predicted to encode bacteriophage structural proteins are conserved between the two phage inserts in strain 52472 but the hypothetical proteins located at the left hand side of contig A are not conserved between the two.

5.2.11 Metabolism

There were a number of metabolism associated genes identified from the pUC assemblies. Some of these predicted CDSs show homology to genes from strain NCTC 11168 but with some sequence difference, others appeared to be unique to the test strain they were identified within.

A probe 5P5g10 was designed to expand the region around the predicted CDS with homology to a PrpD family protein of *Bradyrhizobium japonicum*. This identified BAC 5B2F2 which contains pUC sequence 5P5g10q. 5B2F2 shows 95% nucleotide identity to strain NCTC 11168 with 265 bp changes over the entire length (**Table 5.25** and **Fig 5.30**). The PrpD homologue, CDS 5B2F2_8, has 91% nucleotide identity to strain NCTC 11168. The CDS with homology to a c4-dicarboxylate transporter from *V. vulnificus* also shows homology to pseudogene cj1389, downstream of this CDS there is a complete *metC* homologue rather than two separate CDSs as in NCTC 11168. Downstream of the CDS with homology to fumarate lyase there is a CDS with homology to a MmgE/PrpD family protein which also shows homology to pseudogene cj1395. Together these appear to represent a functional metabolic operon which is largely defunct in strain NCTC 11168. In strain RM1221 the dicarboxylate transporter homologue is a pseudogene but MetC and PrpD family proteins are complete.

Table 5.25: Predicted novel CDSs identified from BAC clone 5B2F2

Locus_id	Putative function	Organism with match	SWALL	E-value	% id
5B2F2_5	c4-dicarboxylate transporter	*Vibrio vulnificus*	Q7MJB8	1.4e-30	36.84
5B2F2_6	cystathionase beta-lyase	*Bordetella bronchiseptica*	Q7WM51	4.1e-71	48.57
5B2F2_7	fumarate lyase	*Campylobacter jejuni*	Q9PMR1	2.7e-168	96.92
5B2F2_8	MmgE/PrpD family protein	*Bradyrhizobium japonicum*	Q89W77	1.6e-39	32.73

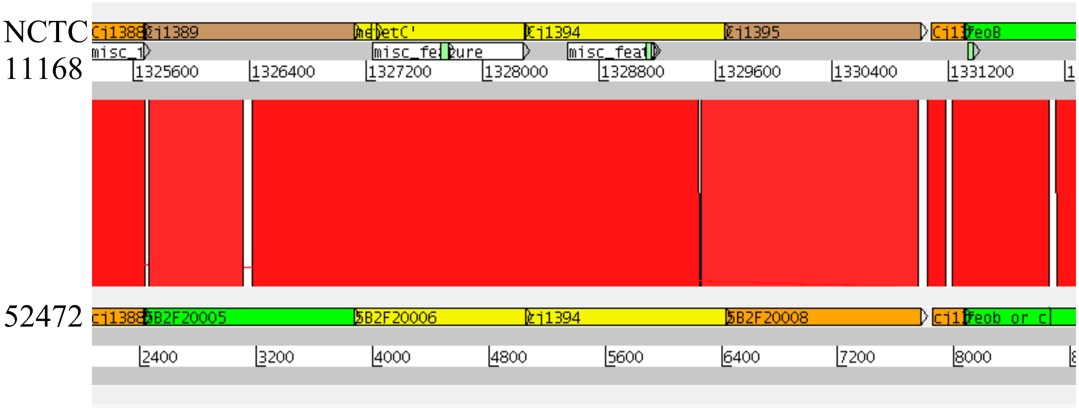


Fig 5.30: Blastn comparison of sequence from strain NCTC 11168 and strain 52472 BAC clone 5B2F2. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark grey lines; DNA translations are represented by light grey lines. Features are represented by open boxes: pFam (white) and prosite (green) matches are marked on the DNA lines; CDSs are marked on one translated frame line irrespective of reading frame and are coloured according to functional category: orange, conserved hypothetical; brown, pseudogenes; dark green, surface; yellow, central/ intermediary/ miscellaneous metabolism. This region shows homology between the two strains, however, the two pseudogenes in strain NCTC 11168 appear to be complete in strain 52472 suggesting that the metabolic operon may be functional in strain 52472.

A homologue of a pyridine nucleotide-disulfide oxidoreductase from *Bacteroides thetaiotaomicron* was identified in the pUC assemblies of strain 40671. A probe was generated for this (4P1e10) and identified the BAC 4B5G11 which contains the pUC sequence 4P2b07p covering 55% of the novel sequence. 4B5G11 contains two CDSs with homology to nitroreductases corresponding to pseudogene cj1064 and *rdxA* (**Table 5.26** and **Fig 5.31**). The first has only 35% aa id to RdxA. There is also an insert of a putative pyridine nucleotide-disulfide oxidoreductase and a CDS similar to a hypothetical protein from *H. hepaticus* inserted between cj1069 and cj1070 relative to the chromosome of strain NCTC 11168.

Table 5.26: Predicted novel CDSs identified from BAC clone 4B5G11

Locus_id	Putative function	Organism with match	SWALL	E-value	% id
4B5G11_7c	pyridine nucleotide-disulfide oxidoreductase	*Bacteroides thetaiotaomicron*	Q8A7I2	6.1e-74	44.68
4B5G11_8c	hypothetical	*Helicobacter hepaticus*	Q7VI88	1.4e-16	44.05

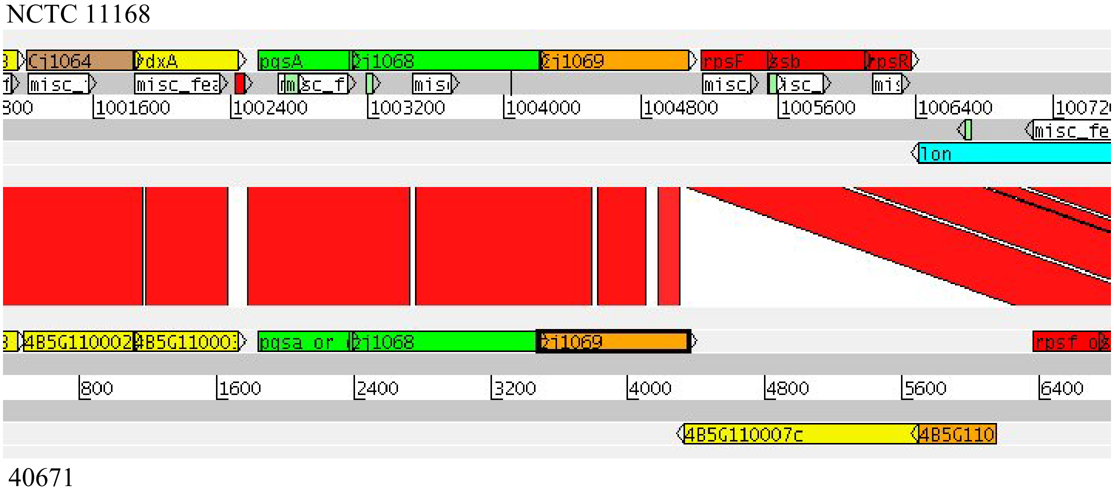


Fig 5.31: Blastn comparison of sequence from strain NCTC 11168 and strain 40671 BAC clone 4B5G11. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark grey lines; DNA translations are represented by light grey lines. Features are represented by open boxes: pFam (white) and prosite (green) matches are marked on the DNA lines; CDSs are marked on one translated frame lines irrespective of reading frame and are coloured according to functional category: yellow, central/ intermediary/ miscellaneous metabolism; brown, pseudogenes; dark green, surface; orange, conserved hypothetical; red, information transfer/ DNA modification; blue, degradation of large molecules. In strain 40671 a putative oxidoreductase and a

conserved hypothetical CDS are inserted between cj1069 and *rpsF* relative to the chromosome of strain NCTC 11168.

5.2.12 Pseudogenes

The probes MP2f07, designed to identify a CDS with homology to a haemoglobin protease from *Escherichia coli*, and MP3b01, designed to identify a CDS with homology to an enterotoxin from *Escherichia coli* identified the BAC MB5D4. It was decided to explore this region further as the pUC assembly data suggested that this region homologous to cj0223 might be intact in strain M1. These pUC regions have high nucleotide similarity to the pseudogene cj0223 of strain NCTC 11168 with 92% and 96% respectively. In strain M1 MB5D4 shows a slightly more complete form of cj0223 enterotoxin. It has 88% nucleotide id to NCTC 11168 across its entire length but is longer in M1 by 117 aa (**Table 5.27** and **Fig 5.32**). The frame shifts in this pseudogene all occur at homopolymeric T tracts although it is unlikely that, as there are three frame shifts, these homopolymeric tract lengths could all vary to give a functional gene. Most variable homopolymeric tracts are G or C in *C. jejuni* [8;9]. The BAC shotgun sequence is at a high enough depth of coverage to be confident about the sequence quality however, as the shotgun sequence is based on a pUC library generated from a single BAC clone it would not be possible to see homopolymeric tract length variation. It is possible that if the gene is not under selective pressure, short homopolymeric tracts represent a point where mutations can easily accumulate. In RM1221 the region homologous to cj0223 is much more degenerate and there is a Mu-like bacteriophage insert between this and *argC*.

Table 5.27: Predicted novel CDSs identified from BAC clone MB5D4

Locus_id	Putative function	Organism with match	SWALL	E-value	% id
MB5D4_3	enterotoxin pseudogene	*Escherichia coli*	Q9EZE7	5.3e-35	27.39

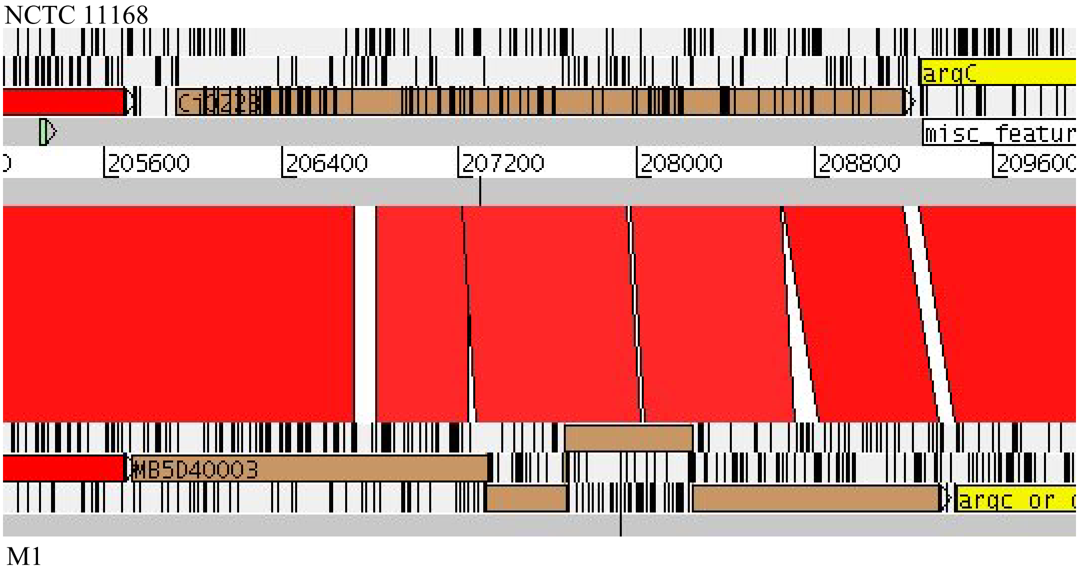


Fig 5.32: Blastn comparison of sequence from strain NCTC 11168 and strain M1 BAC clone MB5D4. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark grey lines; the three-frame forward DNA translations are represented by light grey lines. Stop codons are marked by vertical black lines. Features are represented by open boxes: pFam (white) and prosite (green) matches are marked on the DNA lines; CDSs are marked on the translated frame lines and are coloured according to functional category: red, information transfer/ DNA modification; brown, pseudogenes; yellow, central/ intermediary/ miscellaneous metabolism. In strain M1 only three frame shifts interrupt the reading frame of pseudogene MB5D4_3 compared to the highly interrupted cj0223.

5.3 Discussion

5.3.1 23S rDNA in intervening sequence (IVS)

In the strains 81-176 and M1 an IVS was found in the 23S rDNA, identified from the BAC clone sequences of 8B4F10, MB2B4, 8B2A11 and MB5B1. This is in the same position as a characterised IVS from *C. jejuni* strains F38011, M275 and 78-27 (**Fig 5.33**) replacing the same 8 base pairs and probably allowing the RNA to form a similar stemloop structure [172] (**Fig 5.34**). The IVS has been shown to be excised from the transcribed RNA, cleaving the 23S rRNA into two pieces. Cleaving the 23S rRNA in this way does not appear to hinder ribosomal function [172]. IVSs are also seen in *Salmonella enterica* Typhimurium and *Y. enterocolitica* in approximately the same location. It has been postulated that the IVS may protect the bacterium from bacteriocins that cleave 23S rRNA [172].

```
F38011ivs   1 GCACACAACTTAGATTATTTAAGTTTAGAATATGAGAAACTAAGTTATATGTTTAGTTAT
M1ivs       1 GCACACAACTTAGATTATTTAAGTTTAGAATATGAGAAACTAAGTTAT--GTTTAGTTAT
81-176ivs   1 GCGCACAACTTAGATTATTTAAGTTTAGAATATGAGAAACTAAGTTAT------------

F38011ivs  61 ATTTTTACTGATTTTTATAGAGTAAAGATAGAAATAAAACTTAGTAAAATCAGTAAAAAT
M1ivs      59 ATTTTTACTGATTTTTATAGAGTAAAGATAGAAATAAAACTTAGTAAAATCAGTAAAAAT
81-176ivs  49 ATTTTTACTGATTTTTATAGAGTAAAGATAGAAATAAAACTTAGTAAAATCAGTAAAAAT

F38011ivs 121 ATTCTTAGACTAAAGTTAAGTAGTTTAAGTTGTGTGC
M1ivs     119 ATTCTTAGGCTAAAGTTAAGTAGTTTAAGTTGTGTGC
81-176ivs 109 ATTCTTAGGCTAAAGTTAAGTAGTTTAAGTTGTGTGC
```

Fig 5.33: Alignment of C. jejuni 23S rDNA intervening sequences. DNA sequences were aligned using clustal X. Sequences were taken from strains F38011 (accession number L33972), M1 BAC MB2B4 and 81-176 BAC 8B2A11. The IVS from strain M1 is missing 2 bp compared to the IVS from strain F38011 and the IVS from strain 81-176 is missing 12 bp in the same location as strain M1 compared to the IVS from strain F38011.

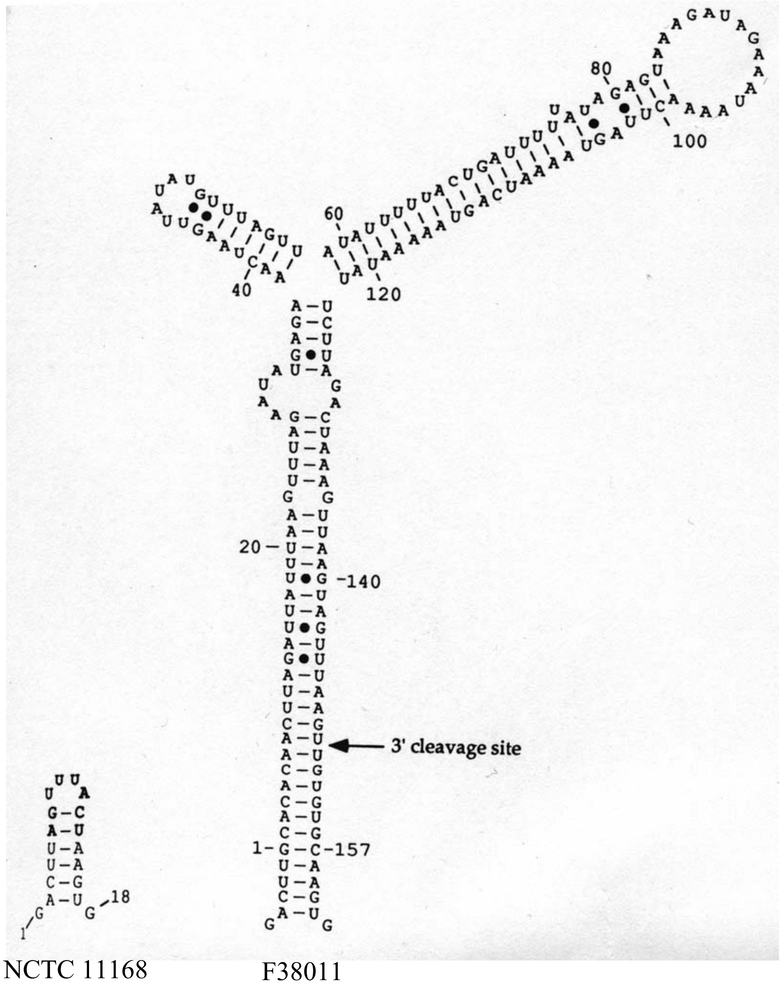


Fig 5.34: Predicted secondary structure of IVS from *C. jejuni* strain F38011 23s rRNA. Figure reproduced from Konkel *et al.* 1994 [172]. The 8 bp in strain NCTC 11168 that are replaced by the IVS are shown in bold on the left hand loop. The right hand loop shows the sequence and predicted secondary structure of the IVS from strain F38011 (accession L33972). The nucleotides of the IVS are numbered from 1 to 157; the nucleotides below this point are identical to the nucleotides from NCTC 11168 shown in the left hand loop that are not indicated in bold. Dots indicate uridine-guanine base pairings.

The differences between the previously characterized IVS and the IVSs from this study occur either in the side loop between base 49 and 60 in strain 81-176 or in regions not required for stem loop structure, base 3 and base 129 (these numbers refer to the previously characterized IVS). It has been noted that if a strain has an IVS that the same one will be present in all three rDNA copies. This seems to be the case in all the rDNA containing BACs that have been sequenced in this study (8B4F10, MB2B4, 8B2A11 and MB5B1).

5.3.2 Respiration

In this study a novel putative cytochrome C biogenesis operon has been identified in strains 81-176 and M1. Cytochrome C is located in the periplasm: either soluble or anchored to the cytoplasmic membrane. Respiratory nitrite reductase, periplasmic TMAO reductase and periplasmic nitrate reductase are all thought to use cytochrome C in electron transfer reactions [180].

The homologues of cytochrome C associated proteins would be consistent with a type II system of cytochrome C biogenesis as seen in *Helicobacter pylori* and Gram positive bacteria [181]. In a type II system an apocytochrome is secreted through the membrane then haem groups are added. CcsA and Ccs1 proteins are postulated to function together in a complex to secrete and attach haem groups. ResA, a thioredoxin, is needed to reduce the apocytochrome in order for the haem groups to be attached [181]. Additionally a fourth protein CcdA is hypothesized to be required for assembly; this is potentially involved in the transfer of reducing equivalents and is required in late stage of cytochrome C maturation in *Bacillus subtilis* [182]. In this study a homologue of NrfI in *W. succinogenes* [183] was identified. NrfI has similarity to both CcsA and Ccs1 from Gram-positive bacteria. In addition a thiol disulphide oxidoreductase homologue was identified which may play the role of ResA. The only homologue from the type II biogenesis scheme missing is CcdA. This is also missing in *W. succinogenes* [183].

A wide range of terminal reductases have been predicted in the genome sequence of strain NCTC 11168 including dimethyl sulfoxide reductases: Cj0264c has been shown to reduce trimethylamine-*N*-oxide (TMAO) and dimethyl sulfoxide (DMSO) [18]. In this study a putative *dmsABC* operon was identified in strains 81-176 and M1. Dimethyl sulfoxide reductases are involved in the reduction of alternative electron acceptors to nitrate and are associated with anaerobic respiration. The reduction of dimethyl sulfoxide requires DmsABC. DmsA is the catalytic subunit containing a molybdopterin cofactor, DmsB is an electron carrier containing four iron-sulfur clusters and DmsC serves as a membrane anchor for the other two subunits [184]. The DmsABC homologues from this study all share highest identity to those from *W. succinogenes* which is strictly anaerobic growing by fumarate respiration with formate or hydrogen as substrate [185]. A *dmsABC* operon in *Actinobacillus pleuropneumoniae* has been linked to acute phase of infection [186]. DMSO is a cryoprotectant produced by some algae and so would be available in aquatic environments [18]. As *C. jejuni* cannot grow anaerobically, these alternative electron acceptors may be used under severely oxygen limited conditions. It has been shown that in strain NCTC 11168 most alternative electron acceptors still require oxygen to function [18]. It is possible that strict anaerobic growth in *C. jejuni* is not possible due to the presence of a I-type ribonucleotide reductase (RNR) which requires oxygen to reduce ribonucleotides to 2`-deoxyribonucleotides which are required for DNA synthesis and repair [18].

5.3.3 Transport

Protein secretion is extremely important for bacteria, and is used in many aspects of pathogenicity. There are 5 major secretion pathways. Type I secretion requires 3 accessory proteins, the type example of secretion being the haemolysin HlyA of *Escherichia coli* [187]. The haemolysin is secreted through a channel spanning both the inner and outer membrane. Type II secretion uses 14 accessory proteins and requires the sec pathway for secretion across the inner membrane. The type example of secretion is PulA *Klebsiella oxytoca* [188]. Type III secretion is very complex and requires a structure that spans inner and outer membrane, an example of type III secreted proteins are the YOPs of *Y. enterocolitica* [189]. Type IV secretion systems are typified by secretion of T-DNA from *Agrobacterium* *tumefaciens* [137] and pertussis toxin of *Bordetella pertussis* [190] and requires at least nine proteins. Type V secretion (autotransport) is the least complicated requiring only 1 protein (see below); an example of an autotransporter is VacA of *Helicobacter pylori* [191]. The two partner secretion (TPS) system is like a type V system but with the passenger and transporter domains of the type V system being encoded by two separate proteins [192].

Autotransporters are known to have various virulence functions including adhesins, toxins and proteases [175;192]. Comparison of the putative autotransporters identified in this study from strains 52472 and M1 shows that although the CDS is present in both, the passenger domain, which will define the function of the protein, is different. This illustrates the fact that horizontal exchange of small sections of DNA within a gene rather than just whole genes may have important functional consequences. Autotransporters all possess: an N-terminal sequence for secretion through the inner membrane using the sec dependent pathway, a secreted mature protein (passenger domain) and a C-terminal beta-domain which forms a pore through which the passenger portion of the protein is secreted through the outer membrane (autotransporter domain) [192]. Once the passenger domain has been secreted through the outer membrane it may either remain attached, e.g. Hsr of *Helicobacter mustelae*, or be cleaved after which it may remain associated with the membrane, e.g. pertactins of *Bordetella* spp., or be released into the extracellular environment, e.g. IgA1 protease from *Neisseria* and *Haemophilus* [175].

In this study a TPS system has also been identified. It has been suggested that this TPS system is defunct in *C. jejuni* strain RM1221 and strain NCTC 11168, and *C. coli* [9]. Only *C. lari* appears to have undisrupted homologues of both required proteins, TpsA and TpsB [9]. There appear to be two TPS systems and in strains 81-176 and M1, one of these TPS systems appears to be duplicated at an alternative chromosomal location. The two original locations contain pseudogenes/ fragmented secreted proteins in strain NCTC 11168.

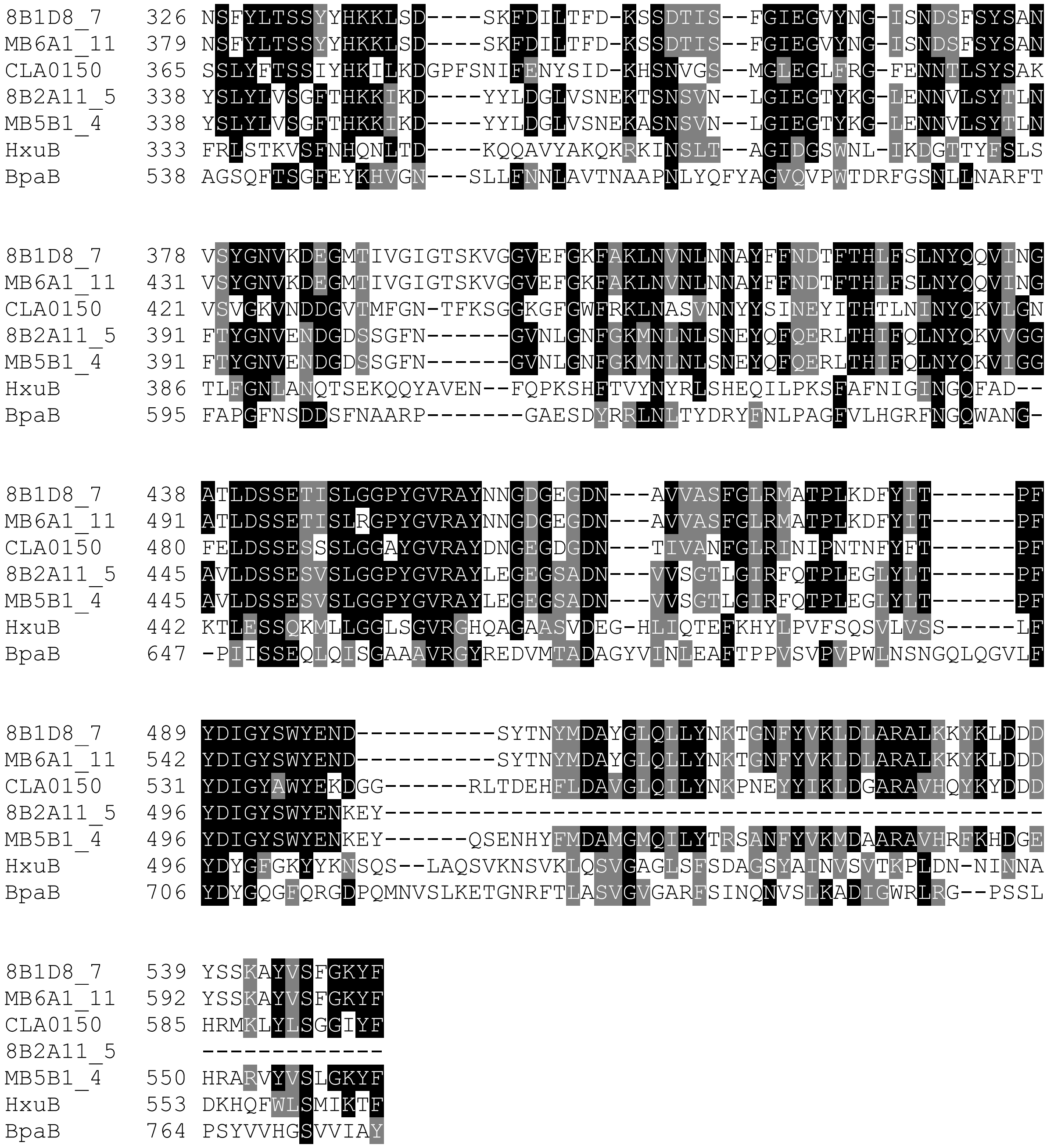
The TPS system located between cj0967 and cj0975 relative to the NCTC 11168 chromosome also appears to have degraded in strain 52472. Most of the secreted proteins from TPS systems of other bacteria are large [193;194]. In 81-176 putative secreted proteins encoded by 8B1A11_10, 8B2A11_3, and 8B1D8_6 are small and in the case of 8B2A11_3 the CDS does not extend to the 8B2A11_5 predicted to encode the TpsB protein, suggesting that a large CDS has accumulated mutations leading to smaller fragmented CDSs; although 8B1A11_10, 8B2A11_3 and 8B1D8_6 may still be secreted. In strain M1, MB5B1_2, which is homologous to an adhesin from *Haemophilus influenzae,* is the largest predicted secreted protein with the reading frame extending from the secretion signal to the CDS predicted to encode a TpsB protein. In the TPS secretion model both partner proteins have sec-dependent signal peptides for translocation through the inner membrane. In addition, TpsA contains an N-terminal signal sequence for transport across the outer membrane mediated by TpsB, this signal region contains the conserved motif, NPNGI and another less conserved motif, NPNL [195], which is not present in all secreted proteins [193]. This extended N-proximal signal sequence is proposed to allow both sec-dependent secretion through the inner membrane then secretion through the transporter protein to occur at the same time. In the putative secreted proteins from *Campylobacter* there is a conserved NPNGI motif (**Fig 5.35**) but in 8B2A11 and MB5B1 there is an M rather than N in conserved signal sequence, which may destroy the signal. There may be other specific signal motifs for secretion of *C. jejuni* proteins *via* this mechanism.

```
MB5C4_6      1 ------------------------MKKLNKLSLSLVVGS---LLFTQSYALPSGGKFTHG
8B1A11_10    1 ------------------------MKKLNKLSLSLVVGS---LLFTQSYALPSGGKFTHG
8B2A11_3     1 -----------------------MKKMSKHIVLSFAVSS---LLFSQAYALPQGGKFTHG
MB5B1_2      1 -----------------------MKKMSKHIVLSFAVSS---LLFSQAYALPQGGKFTHG
CLA0151      1 -----------------------MKKLANHIILSGVTVS---MLFSPLMALPSGGKFTHG
BpaA         1 MKNATARRLYIKKSRMMKSNTLHIKPLVFHVATALAALQGGFLFSSVAWAAPTGSQVVAG
HxuA         1 ------------------------MYKLNVISLIILTTCSG-AAYASTPDFPQHHKTVFG

MB5C4_6     34 TSGSISVSGGTMNISGSKTNSVIQWGGGFNIANGETVNFKGN--GYNYLNIVYGSKSSHI
8B1A11_10   34 TSGSISSSNGTMNISGSKTNSVIQWGGGFNIASGETVNFKGS--GYNYLNIVYGSKSSHI
8B2A11_3    35 TSGTIHTSGNTVTITGKGQNHVIQWGGGFNIAQGESVNFTTS--GKNYLNIAYQKDASKI
MB5B1_2     35 TSGTIHTSGNTVTITGKGQNHVIQWGGGFNIGQNESVNFNGK--NQNYLNIAYQKDASKI
CLA0151     35 TSGTITTNGNNMNISGNGINSVIQWGGGFNIANGEKVNFGGK--DKNYLNIAHGTSKSTI
BpaA        61 -SASIGVSGATTIVNQGSNRAIINWKN-FNVGSGETVRFIAPNTASATLNRVVGSLPSSI
HxuA        36 TVTIEKTTADKMTIKQGSDKAQIDWKS-FDIGQKKEVKFEQPNEHAVAYNRVIGGNASQI

MB5C4_6     92 DGTLEGGTNNIFLINPNGIVVGKDGSINAN-RVFLSASSIGDKEMKEFAKDGK-------
8B1A11_10   92 DGKLEGGSNNIFLINPNGIVVGKGGSINAN-RVYLSTSSVSNEDMQRFANGVS-------
8B2A11_3    93 NGALNGGNNNIFLVNPMGVLIGKTGTITAG-KFVASTTPLNDENVKTFLKQGASF-----
MB5B1_2     93 DGALNGGNNNIFLVNPMGVLIGKTGTITAG-KFVASTTPLSDDNVKTFLEKGASF-----
CLA0151     93 AGILNAGGNNVFLINPNGVIITKTGTINAN-RFVASTSSMSDGDMKAFANLKSFEDGLSF
BpaA       119 NGLVQ-GNGRVFLINPNGILVGQGGAINVQGGFVASTGNISDSAFMQGGAMVLSGDKGQI
HxuA        95 QGKLT-ANGKVYLANPNGVIITQGAEINVA-GLLATTKDLERISENSNSYQFTRR-----
```

Fig 5.35: Alignment of the N-terminal region of TpsA proteins. Protein sequences were aligned using clustal X. Protein sequences of predicted CDSs from this study are shown on the top 4 lines. CLA0151 is a protein from *Campylobacter lari* RM2100 (accession number AEL54392); BpaA is a protein from *Burkholderia pseudomallei* (accession number AA019442); HxuA is a protein from *Haemophilus influenzae* (accession number AAQ10730). The conserved secretion motif NPNGI has been underlined.

The TpsB transporters appear to be highly specific and only secrete one protein, usually one transcribed from the same locus [195]. With this in mind it would appear that each of these transporter locations functions separately so if one member of the pair accumulates deleterious mutations then another system will not be able to compensate and no proteins will be secreted. The TpsB, transporter proteins are more highly conserved than the TpsA, secreted proteins. **Fig 5.36** shows an alignment of the TpsB proteins. It is possible that the proteins secreted in *C. jejuni* act as adhesins, as they contain adhesion associated protein domains, although further work would be needed to assess whether these proteins are secreted and what, if any, function they perform.

```
8B1D8_7      1 ------------------------------------------------------------
MB6A1_11     1 -----------------------MKRIILLSSLAILSLYASDTKDNKKTIQMLEQSPYKE
CLA0150      1 -----------------------MKKLSLCAIALSSLIYANEGGISIAKNDIEKVIELSP
8B2A11_5     1 -----------------------MRKILVVLVLLQVFSHAEELN----NNKIRELIESSP
MB5B1_4      1 -----------------------MRKILVVLVLLQVFSHAEELN----NNKIRELIESSP
HxuB         1 -----------------------MKMRPRYSVIASAVSLGFVLS------------KSVM
BpaB       121 REHGITPVEATNGAALSAGNTNGMAAGAVLAPAAVQDGVPSSTVAAPSATRAARMLPSDL

8B1D8_7      1 ----------------MIIIDHSNTSDDNNSKT--------INTKKNTQ-KDNNNTQKNQ
MB6A1_11    38 DANLKNYNNTLKVKDGVIIIDHSNTSDDNNSKT--------INTKKNTQ-KDNNNTQKNQ
CLA0150     38 DRNLPQNK----------AIKENLKTKDDYIKT--------QEAKKDFEAKKKALKEKLQ
8B2A11_5    34 EANEPQNK----------------------------------NLKNTLK------NQKSP
MB5B1_4     34 EANEPQNK----------------------------------NLKNTLK------NQKSP
HxuB        26 ALDRPDTG----------------------------------SLNRELE--------QRQ
BpaB       181 AVSPPSQRASTIASEPAAASTEQLSASSMPVLAGAREIGSITLASRDITRPQPSSDEAAK

8B1D8_7     36 PNLSNDNTLK-TKTPNSNTPSLKNTSKEESIHKVSFSFHITNKNIN-FKDLGLDEQV-LQ
MB6A1_11    89 PNLSNDNTLK-TKTPNSNTPSLKNTSKEESIHKVSFSFHITNKNIN-FKDLGLDEQV-LQ
CLA0150     80 ENKASEETNS-QTNTNSNN---NTTTTKKVITK--YKFIITNENTS-FKKIGIKEED-LQ
8B2A11_5    54 VNFKEQNTTN-ITNSQTDQ------NEAKVFVR-EYVLHIDNKDLT-FKKLRISEKE-IQ
MB5B1_4     54 VNFKEQNTTN-ITNSQTDQ------NEAKVFVR-EYVLHIDNKDLT-FKKLRISEKE-IQ
HxuB        44 IQSEAKPSGE-LFNQTANS-------PYTAQYKQGLKFPLTQVQILDRNNQEVVTDE-LA
BpaB       241 EAAQEQCGGIGIPSRATPSRPKLPALSSQAVADSYRQSLVQPGNISAEPGIPTTDLEGLE

8B1D8_7     93 EALNDYKKESISVQDLQDIANIISYYVQVSGYPAATAYIPQQELK-DQIQINITLGVLGK
MB6A1_11   146 EALNDYKKESISVQDLQDIANIISYYVQVSGYPAATAYIPQQELK-DQIQINITLGVLGK
CLA0150    132 LLISEFSTKKFSLQDLQDISNIIAYYFQVNGYPAATAYVPQQEFE-DSVQINIALGTLGK
8B2A11_5   104 DAIAEYRNQELSLQNLKDITNIIAYYCQVSGYPSATAYIPPQDLSSNKVQINIAFGTLGK
MB5B1_4    104 DAIAEYRNQELSLQNLKDITNIIAYYCQVSGYPSATAYIPPQDLSSNKVQINIAFGTLGK
HxuB        95 HILKNYVGKEVSLSDLSNLANEISEFYRHNNYLVAKAILPPQEIEQGTVKILLLKGNVGE
BpaB       301 AKLRPFIGQPLDSSLIQKITRVATQYVSAQTDNLVNVYVPPQQLQNGNLVVVFAAAKLGQ

8B1D8_7    152 YVVQNNSSVRDYAIESKLPN--HKGEIITTKLVEDAVYKVNEMYGIQTLASLKAGDNPGE
MB6A1_11   205 YVVQNNSSVRDYAIESKLPN--HKGEIITTKLVEDAVYKVNEMYGIQTLASLKAGDNPGE
CLA0150    191 YIIKNKTTIKDYFVESKLNER-IKGKIISTKLIEDSVYKVNEMYGVQTLAGLQAGENVGE
8B2A11_5   164 VIIKNNSGVRDYALESKLNKN-LKGKVITTKNVENEIYKINEIYGIQTNANLQSGDGYGE
MB5B1_4    164 VIIKNNSGVRDYALESKLNKN-LKGKVITTKNVENEIYKINEIYGIQTNANLQSGDGYGE
HxuB       155 IRLQNHSALSNKFVSRLSNTTVNTSEFILKDELEKFALTINDVPGVNAGLQLSAGKKVGE
BpaB       361 IRTEGQKHISSHDLKCQIRLR--PGDNVDLKTLTDDLTFLNTSPWRQVSSSFTPGAEPGD

8B1D8_7    210 TDVVIETTPSDSFVSVLFYGDNYGIKESGRYRGGASMSFNNIAHQ-GDSLNAYLQRSD-E
MB6A1_11   263 TDVVIETTPSDSFVSVLFYGDNYGIKESGRYRGGASMSFNNIAHQ-GDSLNAYLQRSD-E
CLA0150    250 TDIVIEVEP-DTKANVLLYADNYGIESAGDIRAGISMGFNSLFNM-GDYYNFYLQSSN-E
8B2A11_5   223 SDVIIEVNK-GDSATLTLYSNNYGTKETGRFRAGMSQSLNNIARQ-GDNLNFYLQDSD-E
MB5B1_4    223 SDVIIEVNK-GDSATLTLYSNNYGTKETGRFRAGMSQSLNNIARQ-GDNLNFYLQDSD-E
HxuB       215 ANLLIKIND-AKRFSSYVSVDNQGNKYTGRYRLAAGTKVSNLNGW-GDELKLDLMSSNQA
BpaB       419 ADLVLQTVD-RYPLRVYGSWDNTGTSLTGLNRWRTGVNWGDAFGIVGSRLDYSFAMGNTP

8B1D8_7    268 AQTNYGISYTTFLGNLKITPSYSK--GNYALGGIWREFDFIGTSENLGVDLKYPLWITTY
MB6A1_11   321 AQTNYGISYTTFLGNLKITPSYSK--GNYALGGIWREFDFIGTSENLGIDLKYPLWITTY
CLA0150    307 NQINYGASYTFFLGNLKITPSISQ--GTYSLGGEYKEVGFSGTSRNFGIDFSYPVWINTN
8B2A11_5   280 NQIDYGINYSTFIGNLKITPFATQ--GHYVLGGIYRNLGFYGDSMNVGVNFSYPVFLYTE
MB5B1_4    280 NQIDYGINYSTFIGNLKITPFATQ--GHYVLGGIYRNLGFYGDSMNVGVNFSYPVFLYTE
HxuB       273 NLKNARIDYSSLIDGYSTRFGVTANYLDYKLGGNFKSLQSQGHSHTLGAYLLHPTIRTPN
BpaB       478 RELMEHTLQYTMPTSYRDTLTFTGNYSSSNAAIEDGTFNVKGKNIQASAQWTHLLGGPAA
```



```
8B1D8_7    326 NSFYLTSSYYHKKLSD----SKFDILTFD-KSSDTIS--FGIEGVYNG-ISNDSFSYSAN
MB6A1_11   379 NSFYLTSSYYHKKLSD----SKFDILTFD-KSSDTIS--FGIEGVYNG-ISNDSFSYSAN
CLA0150    365 SSLYFTSSIYHKILKDGPFSNIFENYSID-KHSNVGS--MGLEGLFRG-FENNTLSYSAK
8B2A11_5   338 YSLYLVSGFTHKKIKD----YYLDGLVSNEKTSNSVN--LGIEGTYKG-LENNVLSYTLN
MB5B1_4    338 YSLYLVSGFTHKKIKD----YYLDGLVSNEKASNSVN--LGIEGTYKG-LENNVLSYTLN
HxuB       333 FRLSTKVSFNHQNLTD----KQQAVYAKQKRKINSLT--AGIDGSWNL-IKDGTTYFSLS
BpaB       538 AGSQFTSGFEYKHVGN---SLLFNNLAVTNAAPNLYQFYAGVQVPWTDRFGSNLLNARFT

8B1D8_7    378 VSYGNVKDEGMTIVGIGTSKVGGVEFGKFAKLNVNLNNAYFFNDTFTHLFSLNYQQVING
MB6A1_11   431 VSYGNVKDEGMTIVGIGTSKVGGVEFGKFAKLNVNLNNAYFFNDTFTHLFSLNYQQVING
CLA0150    421 VSVGKVNDDGVTMFGN-TFKSGGKGFGWFRKLNASVNNYYSINEYITHTLNINYQKVLGN
8B2A11_5   391 FTYGNVENDGDSSGFN------GVNLGNFGKMNLNLSNEYQFQERLTHIFQLNYQKVVGG
MB5B1_4    391 FTYGNVENDGDSSGFN------GVNLGNFGKMNLNLSNEYQFQERLTHIFQLNYQKVIGG
HxuB       386 TLFGNLANQTSEKQQYAVEN--FQPKSHFTVYNYRLSHEQILPKSFAFNIGINGQFAD--
BpaB       595 FAPGFNSDDSFNAARP-------GAESDYRRLNLTYDRYFNLPAGFVLHGRFNGQWANG-

8B1D8_7    438 ATLDSSETISLGGPYGVRAYNNGDGEGDN---AVVASFGLRMATPLKDFYIT------PF
MB6A1_11   491 ATLDSSETISLRGPYGVRAYNNGDGEGDN---AVVASFGLRMATPLKDFYIT------PF
CLA0150    480 FELDSSESSSLGGAYGVRAYDNGEGDGDN---TIVANFGLRINIPNTNFYFT------PF
8B2A11_5   445 AVLDSSESVSLGGPYGVRAYLEGEGSADN---VVSGTLGIRFQTPLEGLYLT------PF
MB5B1_4    445 AVLDSSESVSLGGPYGVRAYLEGEGSADN---VVSGTLGIRFQTPLEGLYLT------PF
HxuB       442 KTLESSQKMLLGGLSGVRGHQAGAASVDEG-HLIQTEFKHYLPVFSQSVLVSS-----LF
BpaB       647 -PIISSEQLQISGAAAVRGYREDVMTADAGYVINLEAFTPPVSVPVPWLNSNGQLQGVLF

8B1D8_7    489 YDIGYSWYEND----------SYTNYMDAYGLQLLYNKTGNFYVKLDLARALKKYKLDDD
MB6A1_11   542 YDIGYSWYEND----------SYTNYMDAYGLQLLYNKTGNFYVKLDLARALKKYKLDDD
CLA0150    531 YDIGYAWYEKDGG------RLTDEHFLDAVGLQILYNKPNEYYIKLDGARAVHQYKYDDD
8B2A11_5   496 YDIGYSWYENKEY-----------------------------------------------
MB5B1_4    496 YDIGYSWYENKEY------QSENHYFMDAMGMQILYTRSANFYVKMDAARAVHRFKHDGE
HxuB       496 YDYGFGKYYKNSQS--LAQSVKNSVKLQSVGAGLSFSDAGSYAINVSVTKPLDN-NINNA
BpaB       706 YDYGQGFQRGDPQMNVSLKETGNRFTLASVGVGARFSINQNVSLKADIGWRLRG--PSSL

8B1D8_7    539 YSSKAYVSFGKYF
MB6A1_11   592 YSSKAYVSFGKYF
CLA0150    585 HRMKLYLSGGIYF
8B2A11_5       -------------
MB5B1_4    550 HRARVYVSLGKYF
HxuB       553 DKHQFWLSMIKTF
BpaB       764 PSYVVHGSVVIAY
```

Fig 5.36: Alignment of TpsB secretor proteins. The protein sequences were aligned using clustal X. The CDSs from this study are 8B1D8_7, MB6A1_11, 8B2A11_5 and MB5B1_4. CLA0150 is a protein from *Campylobacter lari* RM2100 (accession number EAL54391); HxuB is a protein from *Haemophilus influenzae* (accession number AAQ10738) and BpaB is a protein from *Burkholderia pseudomallei* (accession number AA019443).

5.3.4 Chemotaxis

Chemotactic responses have been shown in *Campylobacter* and suggested to be important factors in colonization of the intestinal mucosa [196]. Mutants of two MCP-type chemotaxis genes, cj0019c and cj0262c, in strain 81-176 were shown to be deficient in colonization of the chick gastrointestinal tract [197]. In strain 40671 there is a putative MCP-type chemotaxis gene with homology to the C-terminus of Cj0262c.

The genome sequence of strain NCTC 11168 identified 10 chemotaxis receptor proteins. Six of these including Cj0144, Cj0262c and Cj1564 belong to group A of transducer-like proteins (Tlp). These three proteins all contain an identical C-terminus. These group A proteins show a similar structural organization to methyl-accepting proteins of *Escherichia coli* and are proposed to act in a similar way [167]. It has been proposed that group A Tlps sense ligands external to the cell with their extracellular domain. In the *Escherichia coli* paradigm Tlps are proposed to bind to complexes of CheW and CheA, with their intracellular domain, in order to respond to changes in gradients of chemoattractants or chemorepellents. When a chemorepellent binds or there is a lack of chemoattractant binding (depending on the specificity of the extracellular receptor domain), CheA autophosphorylates. The phosphate residue is then transferred to CheY, a soluble response regulator protein, which once phosphorylated, is able to bind to the flagellar switch protein, FliM. This induces a change in the direction of flagellum rotation. Conversely when a Tlp binds a chemoattractant CheA autophosphorylation is inhibited, which in turn decreases phosphorylated CheY so the bacterium continues to move in the same direction [198]. Bacteria respond to changes in gradients of chemoattractants and repellents so there are feedback mechanisms in place proposed to act *via* reversible methylation involving CheB and CheR [167].

The putative MCP chemotaxis CDS identified as novel in 40671 was identified in the place of cj0262c relative to the chromosome of NCTC 11168. It may be that this protein is a novel Tlp however some proteins with homology to chemotaxis receptor proteins have been shown to be involved in other systems and not directly with chemotaxis. For example, a similar domain to the highly conserved signalling domain of MCP-type chemotaxis proteins is found in HlyB of *V. cholerae* implicated in toxin secretion and PilJ of *Pseudomonas aeruginosa* required for the production of type IV fimbriae [199]. If the MCP-like protein in strain 40671 is involved directly in chemotaxis it may have a different function to Cj0262c which is required for wild type levels of colonization of the chick gastrointestinal tract [197]. This may reflect the different environmental niches these strains are best adapted to.

5.3.5 Restriction Modification

From the BAC clone sequences two RM systems were identified in strains 40671 and 52472. In strain 40671 there is a restriction and a methylation homologue. In strain 52472 there is a restriction and a methylation homologue and also a protein kinase homologue, which is unusual. Interestingly, this arrangement of a methyltransferase followed by a protein kinase has been associated with a phage growth limitation system in *Streptomyces coelicolor* [179]. The *pgl* locus of *Streptomyces coelicolor* consists of four genes *pglWXYZ* which most closely resembles a type I RM system. It is proposed to act by targeted modification of bacteriophage or bacteriophage DNA which inhibits bacteriophage growth on reinfection of the same host [179].

5.3.6 Capsule

A complex of conserved Kps proteins is responsible for translocating the assembled polysaccharide across the cell membrane (Karlyshev 2005). These transporter genes flank the polysaccharide biosynthesis genes. In strain 40671 homologues of the GDP-D-glycero-

D-mannoheptose pathway (GmhA2, HddA and HddC) are present in the capsule locus. However, homologues of the UDP-glucose dehydrogenase, Udg and of the UDP-pyranose mutase, Glf are not present. In this respect this capsule more closely resembles that of 81-176. There is also a DmhA homologue which is proposed to convert heptose to deoxyheptose but an HddD heptosyltransferase homologue is lacking [151]. Heptose residues found in some cell surface glycoconjugates are required for adhesion [59]. In order to comprehensively study the structure of the capsule for this strain, a technique such as high-resolution magic angle spinning (HR-MAS) nuclear magnetic resonance (NMR) spectroscopy could be used. HR-MAS NMR has been used to examine glycan modifications of the NCTC 11168 [200;201] capsule and also for 81-176 and G1 [151].

5.3.7 Bacteriophage

In strain 52472 two novel inserts were found containing bacteriophage related CDSs. *Campylobacter* are known to carry phage and indeed phage typing has been used to give finer discrimination between serotypes [19]. Some *Campylobacter* strains are resistant to bacteriophage [202]. The phage inserts in strain 52472 have approximately 30% G+C content, similar to that of chromosome. Bacteriophage Mu from *E. coli* is around 50% G+C and where Mu-like bacteriophage have integrated in other bacteria they often show a disparity of G+C content compared to the chromosome. For example, FluMu of *Haemophilus influenzae* has 50% G+C compared to 38% G+C for the chromosome [203].

Mu-like bacteriophage are known to integrate into nearly random chromosomal locations and also replicate by transposition. During the lytic cycle Mu transposes to several sites around the host genome [203]. This can cause disruption to various host genes which would not occur during the lysogenic stage. Bacteriophage are known to be highly mosaic in nature [171], acquiring DNA by homologous and nonhomologous recombination. Bacteriophage have been cited as a common mechanism for genomic rearrangement and in some cases can enhance the virulence of the infected bacteria. Bacteriophage can encode toxins as in the case of Shiga toxin in *Escherichia coli* 01571:H7 [79] and the serum resistance determinant *bor* of *Escherichia coli* [80]. However, in this instance there are no previously characterized virulence determinants on these putative prophage. It is also not apparent whether these bacteriophage are complete, whether they are inducible or whether they are remnants permanently inserted into the chromosome.

5.3.8 Pseudogenes

Genes present in some strains and not in others may be accessory and therefore subject to reduced selective pressure in many environments. A number of predicted CDSs identified in this study appear to be pseudogenes in one strain or another. For example, the di-tripeptide transporter is a pseudogene in NCTC 11168 and possibly in other strains depending on whether homopolymeric tracts give rise to frame shift by slip-strand mispairing. Phase variation cannot be seen in the shotgun sequencing of BAC clones as the libraries used for sequencing are derived from a single clone. However, in the entire NCTC 11168 genome there is only 1 variable tract out of 22 that is associated with poly A/T, the rest are all G/C. In the novel predicted CDSs most of the frame shifts occur at poly A/T tracts with the exception of the capsule region of strain 40671 which contains G/C homopolymeric tracts. Further investigation would be required to see if phase variation occurs, if the predicted CDSs are pseudogenes or if the fragments can function independently.

The putative autotransporter in 52472, various parts of the TPS system, the TraG-like island, some CDSs within the *tetO* insert and a hypothetical CDS in 81-176 are all likely to be pseudogenes. Also a number of pseudogenes in strain NCTC 11168 are complete in the other strains tested. The region surrounding the CDS predicted to encode a PrpD-family protein in strain 52472 contains pseudogenes in strain NCTC 11168. Not only may novel CDSs be pseudogenes but also CDSs in the region of insertion: for example, in strain M1 a novel region predicted to encode TetO has inserted within an orthologue of cj0770 which may not be functional in this strain due to the insertion. Other examples include a partial *lpxB* in strain 40671 RM insert and *glpT* pseudogene in strains 40671 and 52472. It is impossible to tell at entire genome level how many pseudogenes there are in one strain compared to another but there appear to be many in the regions that differ between strains. However, of the novel predicted CDSs in this study, excluding those predicted within bacteriophage or plasmid DNA, 21% are inactivated in one or more strains compared to the chromosomal background of 1.3% in strain NCTC 11168 and 2.5% in RM1221.

5.3.9 Overview

In strain 81-176 8 regions were sequenced, in strain M1 10 regions were sequenced, in strain 40671 6 regions were sequenced and in strain 52472 7 regions were sequenced. This represents an expansion of 37% of 81-176, 30% of M1, 31% of 40671 and 36% of 52472 over the novel pUC regions. Some of the BACs sequenced contained novel sequence that was not present in the pUC assemblies. In strain 81-176 22%, strain M1 23%, strain 40671 38% and strain 52472, discounting bacteriophage, 35% of the unique BAC sequence had not been previously identified in the pUC assemblies. This shows that sequencing BAC libraries is a useful and complementary technique to the differential hybridization pUC screen to find the context and the entire sequence of novel regions of DNA.

Insertion of pathogenicity islands in bacteria are often associated with tRNA genes and insertion sequence elements at their boundaries. These regions may sometimes be flanked by direct repeats [71]. In this study insertion of novel DNA appears more likely to have occurred by recombination as there are no obvious tRNA pathogenicity islands or insertion sequence elements. The only examples of inserts adjacent to tRNAs are the TraG-like islands of strains 81-176 and M1 and an insert of two hypothetical CDSs in MB1B12. The only possible transposon associated insert is *tetO* in strain M1 and 23-45. There are regions of novel DNA which show homology to other delta epsilon proteobacteria e.g *W. succinogenes*, *Shewanella oneidensis* and *Helicobacter pylori.* This may suggest that recombination is a more common method of incorporation of novel genes into strains of *C. jejuni* than the incorporation of mobile pathogenicity islands. There are indel events near to rDNA which may represent a good place for recombination as rDNA tends to be highly conserved. Out of the 3 copies of rDNA in the *C. jejuni* genome there are indel events adjacent to two of them in strains 81-176 and M1 when compared to strain NCTC 11168. In *C. coli cfrA* is located downstream of an rDNA and is also located downstream of an rDNA in some strains of *C. jejuni* [14]. These findings are consistent with the two published genome sequences of *C. jejuni* not containing any classical pathogenicity islands or IS elements [8;9].

5.3.9.1 Strain 81-176

Strain 81-176 was originally an outbreak strain originating from raw milk. In this part of the study a putative cytochrome C biogenesis operon was discovered and also a putative *dmsABC* operon. This suggests a level of respiratory diversity. Also several transport associated regions were discovered; di-tripeptide transport and three putative two partner secretion systems. This strain also contained an island with a TraG-like homologue and several plasmid associated gene remnants, and an insert of hypothetical CDSs. Respiratory chain divergence could allow this strain to survive under reduced oxygen tensions such as those found in the mammalian and avian gut. Nutrient uptake could allow this strain to survive in different environmental niches to strain NCTC 11168.

5.3.9.2 Strain M1

Strain M1 was isolated from a scientist who developed severe inflammatory gastroenteritis after a visit to a poultry abattoir. This strain, like strain 81-176, also contains a putative cytochrome C biogenesis operon and a putative *dmsABC* operon. Transport associated regions were also found in this strain, DTPT transport, three TPS systems and a putative autotransporter. A TraG-like island, a *tetO* chromosomal insert and an enterotoxin pseudogene were also found. As well as respiratory diversity this strain also has a number of potential adhesins in the form of the TPS systems and possibly the autotransporter. Such adhesins could be a factor involved in chicken colonization.

5.3.9.3 Strain 40671

Strain 40671 is an outbreak strain thought to be associated with water. In this strain two novel regions containing hypothetical CDSs, a RM system, an oxidoreductase, a novel capsule and a novel MCP-type chemotaxis receptor were discovered. As this strain was associated with water the capsule may aid environmental survival. Strain differences in survival in water have been shown, and when *C. jejuni* was incorporated into natural biofilms, then it could survive for weeks [204;205]. Biofilms form by cell-cell adhesion so capsule polysaccharide may be an important factor. However, there are likely to be many more factors involved such as sensing [206]. A putative oxidoreductase and several hypothetical CDSs which may be associated with metabolism were identified, however further work would be needed to identify what function these predicted CDSs have. If metabolic pathways vary between strains they are unlikely to be essential and may represent accessory pathways that are useful to that particular isolate.

5.3.9.4 Strain 52472

Strain 52472 was isolated from a patient with septicaemia. In this strain two inserts of bacteriophage associated DNA, plasmid genes, an RM system, metabolism associated CDSs, and pseudogenes of an autotransporter and TPS system have been identified. The plasmid genes include all the components of a type IV secretion system. Metabolism associated

CDSs include a homologue of a *prpD* family gene. PrpD family proteins are associated with the catabolism of propionate, a short chain fatty acid found in the intestinal lumen, *via* the 2-methylcitric acid cycle [153]. There are however more enzymes required in this pathway and the possibility that strain 52472 could catabolize propionate will need to be explored further.

6. TetO analysis from multiple clinical isolates

6.1 Introduction

Campylobacteriosis is usually self-limiting and treated by replacing fluids and restoring electrolyte balance. Antibiotics are generally only used in severe infections in which case erythromycin is used as the drug of choice [207] but tetracycline may also be used [26]. *C. jejuni* resistance to erythromycin remains low among clinical isolates [208] while resistance to other antibiotics has increased in the last few decades [26]. Tetracycline resistance has increased dramatically over the past 20 years, with resistant isolates in Canada rising from 8% to 50%. The tetracycline resistance of clinical *C. jejuni* isolates has been recorded as 55% in North America and up to 95% in Thailand. In addition tetracycline has been used prophylactically and therapeutically as a feed additive for poultry [209;210]. In the Netherlands tetracycline resistance in *Salmonella* spp. was found to increase in poultry when tetracycline was administered, and decrease after a ban on the use of tetracycline for growth promotion in animal feed, although the link between resistant strains in poultry and clinically resistant isolates remains unproven [210]. Most tetracycline resistance is thought to be plasmid mediated although a chromosomally mediated tetracycline resistance determinant has been reported for *C. coli* [176]. This determinant could be transferred into tetracycline sensitive *C. coli* by natural transformation where it inserted into the same site, however this was deemed to be due to recombination rather than chromosomal insertion mediated for example by a transposon [58].

In strain M1 a chromosomally located *tetO* gene which encodes a tetracycline resistance determinant was located (section 5.2.6) and in order to investigate whether this represents a transposable element, it was decided to look at other strains that might also harbour *tetO* on the chromosome. A recent study showed that in tetracycline resistant clinical isolates from Canada 67% contained plasmids. However, 32 strains contained *tetO* but plasmid DNA could not be isolated from these strains using three separate methods [207].

6.2 Methods

6.2.1 Strains and primers used in this study

Table 6.1: *C. jejuni* strains used in this study

strain	Location of isolation, Canada	Year of isolation	Blood in stool?	Contains pVir?
23-45	Spruce Grove, Alberta	2000	Yes	Yes
16-60	Fort Smith, Alberta	1999	No	No
16-48	Edmonton, Alberta	1999	No	No
24-50	Yellowknife	2001	No	No
24-34	Valemont	2001	Yes	No
16-02	St. Albert, Alberta	1999	No	No
25-69	Edmonton, Alberta	2001	No	No
25-25	Edmonton, Alberta	2001	No	No

Diane Taylor's group (University of Alberta, Canada) supplied DNA from 8 strains of *C. jejuni* for which plasmids had not been identified for this part of the project.

Table 6.2: Primer used in this study

Primer	Sequence 5`-3`	tm	notes
MP1d11L	CTAATAACATCCCTCAAATGC	54.1	*tetO*
MP1d11R	AATTATGGGAAACGATGAAC	54.1	*tetO*
5B3D12vL	ATATGCTAAAGAATCAGGAATG	53	*virB4* region
5B3D12vR	TATTGTCTATGCTCGAAACC	53	*virB4* region

6.2.2 Overview

Libraries of DNA from all the strains were constructed in pBACe3.6 with an average insert size of 10-15 Kb (section 2.3.2). These libraries were arrayed in duplicate onto nylon membranes (section 2.3.3) and hybridized (section 2.3.4) either to a probe generated from the *tetO* sequence of M1 or a probe from the contig containing a homologue of *virB4* from the BAC clone sequence of strain 52472 which both showed high similarity to the pTet sequence. Clones which hybridized to either probe were then selected for end sequencing. The two end sequences were compared to the sequences of both strain NCTC 11168 and pTet using WUBLASTN to give an idea of the location of *tetO*. Strains that appeared to harbour a small chromosomal insertion were selected for BAC shotgun sequencing using the same approach as for sequencing the other BACs in this study (section 5.2.1).

6.3 Results

6.3.1 BAC end sequencing

In order to determine which strains harboured a small chromosomal insert containing the *tetO* gene BAC libraries were screened using radiolabelled probes generated from both *tetO* and a region containing a homologue of *virB4*. The *tetO* probe was used to locate BAC clones containing the tetracycline resistance determinant. The *virB4* probe was used to determine whether more plasmid DNA was present in these strains as the probe was designed to hybridize to a region approximately half way round the plasmid compared to *tetO*. If upon end sequencing the BAC clones only plasmid DNA was discovered there would be a high probability that these strains contained a tetracycline resistance plasmid (possibly chromosomally integrated) rather than a transposon type insert (**Table 6.3**).

Table 6.3: **End sequence data from BAC clones that hybridized to probes**. The ‘well’ column represents the well reference of each positive hybridization clone. The ‘match to 11168’ column represents the location on the NCTC 11168 chromosome in base pairs that the end sequences match to using WUBLASTN. Similarly the ‘match to pTet’ column represents the location of the end sequences on the pTet plasmid (45204bp). The ‘notes’ column shows likely locations of poor sequence quality reads on the pTet plasmid or NCTC 11168 chromosome, and also matches of novel DNA sequences that matched neither pTet nor the NCTC 11168 chromosome.

Strain	probe	well	match to 11168	match to pTet	notes
23-45	*tetO*	E21:1	-209370	-42331 (*tetO*)	
		D21:2	195474-203366		
		I1:2	187660-198835		
		J21:2	1332211-200302		
	virB4	no matches			
16-60	*tetO*	A21:3		-4389	~40000
		B11:3		38920-6960	
		G9:3		38067-4452	
		H21:3		36061-4544	
		A12:4	407763-421374		
		E2:4		-4761	~35342
	virB4	J10:3		15022-28846	
		D7:4		-29666	~19000
		M2:4		17812-30027	
16-48	*tetO*	D17:6		-4512	~40000
		M24:6		38093-6504	
	virB4	A17:5		18603-29929	
		N10:5		17085-29069	
24-50	*tetO*	H5:7		38922-5749	
		L10:7		38280-5604	
		E1:8		38799-4544	
		J5:8		38797-7541	
	virB4	H9:7		-29464	~19120
		P21:7		-26086	~18107
24-34	*tetO*	F13:1	190516-198933		
		I18:1	192590-199546		
		K19:1	192601-200225		
		L9:1	190462-		~198162
		L23:1	191476-198650		
		H20:2	195672-		~206077
		H23:2	192754-198835		
	virB4	no matches			
16-02	*tetO*	H3:4		33768-44445	
		O9:4		38922-5909	
	virB4	A24:3	-111653		
		B14:3	1307047-	-31580	
		H11:4	12644-	-28970	
25-69	*tetO*	D11:5		38295	~4700
		I21:5		38271-5801	
		N13:6		-6568	~40000
	virB4	H1:5	200752-	15022	
25-25	*tetO*	B5:7		36158-2645	
		E20:7		34415-3568	
		H18:7		35242-4387	
		I2:7		38813-5369	
		M15:7		35097-2795	
		O21:7		38156-4240	
		J10:8		35192-4453	
	virB4	C15:8		-26307	protein kinase
		I18:8		-29649	virB5 pCC31
		L12:8		-29524	~15685

Both strains 23-45 and 24-34 hybridized to the *tetO* probe but not the *virB4* probe. Sequences from the ends of BAC clones from these strains are anchored in the NCTC 11168 genome sequence suggesting that they harbour a small chromosomal insertion containing *tetO*.

Strains 16-60, 16-48, 24-50, 16-02, 25-69 and 25-25 seem to contain plasmid DNA as well as *tetO* as these strains also hybridized to the *virB4* probe. In these strains sequences from the ends of BAC clones containing DNA that hybridized to the *tetO* probe contained sequence complementary to pTet. This suggests that a small chromosomal insert is not present in strains 16-60, 16-48, 24-50, 16-02, 25-69 and 25-25. The *virB4* probe also hybridized to DNA from clones of these strains. Strain 25-25 appeared to contain sequence not contained in pTet with the BAC clone C15:8 end sequence encoding a predicted CDS with 37% aa id to a serine-threonine protein kinase from *Streptomyces avermitilis*. Further investigation would be required to explore the possibility that this strain contains novel DNA in the same region as plasmid DNA with homology to pTet. The BACs identified with the *virB4* probe from strains 16-02 and 25-69 have end sequence matches to NCTC 11168 which may indicate that plasmid DNA is inserted into the chromosome while the BAC end sequences from strains 16-48, 24-50 and 25-25 have no matches to strain NCTC 11168 chromosomal DNA.

6.3.2 BAC insert sequencing

DNA inserts from BAC clones I1:2 and I18:1 from strains 23-45 and 24-34 respectively were subcloned into pUC19 and sequenced using a shotgun strategy (section 5.2.1). These clones were selected for sequencing as the end sequence reads were complementary to regions on the strain NCTC 11168 chromosome consistent with the expected insert size of the BAC clones. In addition the end sequences from several other clones from these strains, which hybridized to the tetO probe, were complementary to the same chromosomal region.

Both strains 23-45 and 24-34 contained an approximately 5.5 kb insertion in a CDS orthologous to cj0203 from strain NCTC 11168 and showed 99% sequence similarity to each other with only 19 bp differences (**Table 6.4** and **Fig 6.1**).

Table 6.4: CDSs from strain 23-45 encoded on the *tetO* insert.

Locus_id	Length in aa	Putative function	Database match	Organism with match	SWALL	E-value	% id
2345_5	212	transposase	OrfA	*Helicobacter pylori*	Q9RMU7	4.5e-64	85.84
2345_6	430	unknown	OrfB	*Helicobacter pylori*	Q9RMU6	7e-92	62.88
2345_7c	57	unknown	hypothetical	*Enterococcus faecalis* tn916	Q56396	6e-14	66.66
2345_8c	639	Tetracycline resistance	TetO	*Campylobacter jejuni*	TETO_CAMJE	0	99.84
2345_9c	124	unknown	TnpV	*Clostridium difficile*	O05416	6.7e-15	44.91
2345_10c	53	Gene fragment	DNA relaxation protein	*Fusobacterium nucleatum*	Q9L9V7	9e-09	58.82

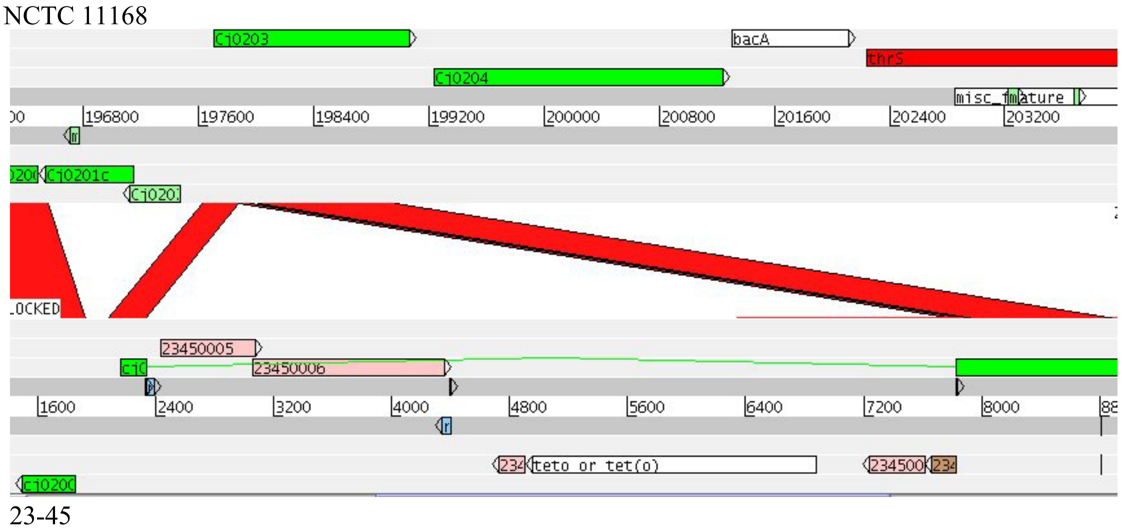


Fig 6.1: **Blastn comparison of strain NCTC 11168 compared to strain 23-45.** The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark grey lines; the three-frame forward and reverse DNA translations are represented by light grey lines. Features are represented by open boxes: inverted repeats (blue) and target site duplications, flanking the putative transposon (2345_5 and 2345_6) and at the edges of the insert, are marked on the DNA lines; CDSs are marked on the translated frame lines. CDSs are coloured according to functional category: dark green, surface; light green, unknown; pink, bacteriophage/ IS elements; white,

pathogenicity/ adaptation/ chaperones; brown, pseudogenes and partial genes; red, information transfer/ DNA modification. In strain 23-45 cj0201c and cj0202c are missing, and 6 novel CDSs including a tetracycline resistance gene are inserted within cj0203, compared to the chromosome of strain NCTC 11168.

As the sequence from strains 23-45 and 24-34 are so similar only the sequence of strain 23-45 will be discussed further. Interestingly, within this insertion there is a putative transposon homologous to IS607 from *Helicobacter pylori* [211]. The predicted CDS 2345_5 shows 86% id to OrfA which has been shown to be necessary for transposition whilst the predicted CDS 2345_6 shows 63% id to OrfB which has no known function. These transposon associated CDSs are flanked by inverted repeats. It has also been noted that IS607 inserts either next to, or between, adjacent GG nucleotides generating a 2-bp or 0-bp target sequence duplication. This would be consistent with the insertion site in cj0203. It is possible that the transposon might have picked up the adjacent DNA from a plasmid similar to pTet containing a *tetO* homologue, although there does not appear to be an inverted repeat at the other side of the insert that could lead to extended excision.

The *tetO* gene and the CDS immediately downstream appear to be highly conserved between pTet, M1 and 23-45/24-34 while other surrounding CDSs differ. Strains 23-45 and 24-34 are both predicted to encode a CDS with homology to TnpV from *Clostridium difficile.* In strain M1 there is also a CDS with homology to TnpV although the CDS only matches to the C-terminus of TnpV suggesting that it is a fragment (**Fig 6.2**).

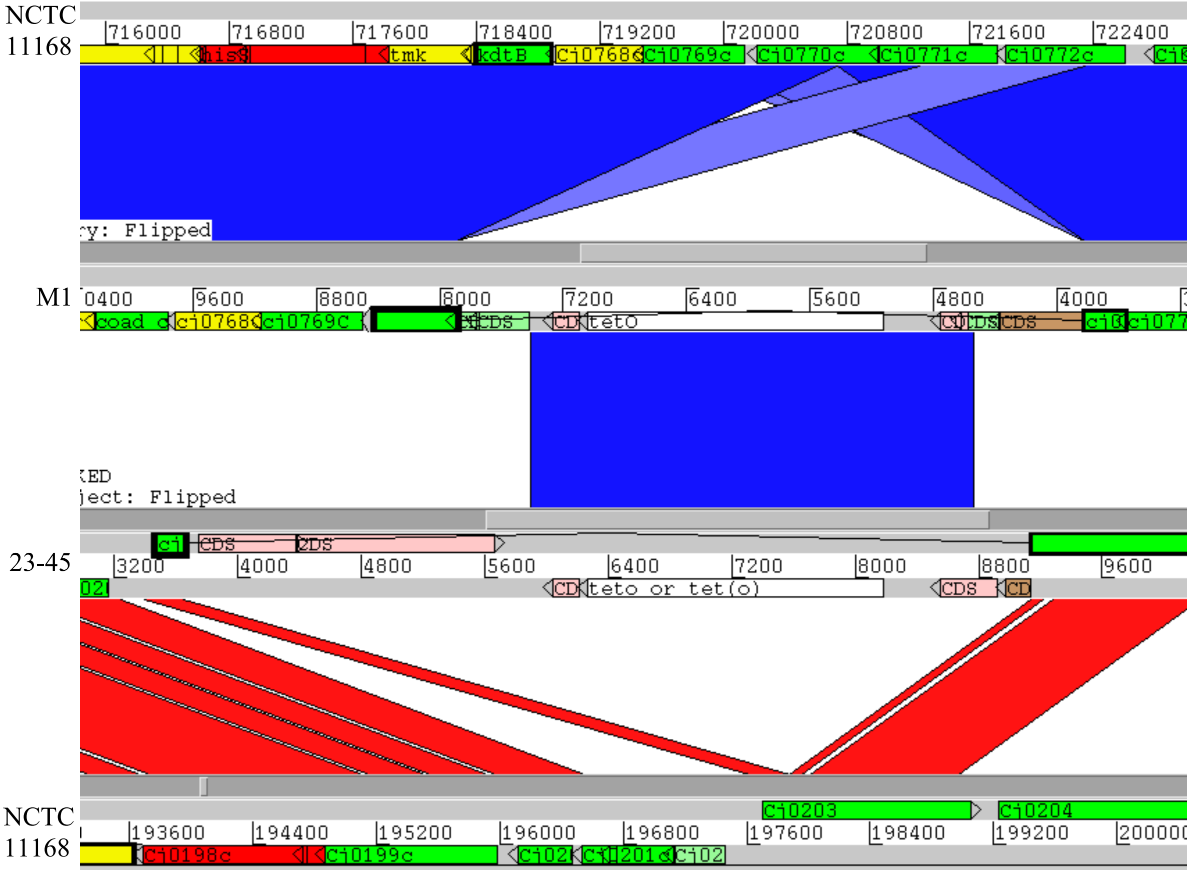


Fig 6.2: **A blastn comparison of strains NCTC 11168, M1 and 23-45**. The comparison is viewed using ACT; blocks of red or blue indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark grey lines. CDSs are represented by open boxes and are coloured according to functional category: yellow, central/ intermediary/ miscellaneous metabolism; red, information transfer/ DNA modification; dark green, surface; light green, unknown; pink, bacteriophage/ IS elements; white, pathogenicity/ adaptation/ chaperones; brown, pseudogenes and partial genes. The pale blue blocks between cj0770c, cj0771c and cj0772c show that these genes have a degree of identity to each other. Both strain M1 and strain 23-45 have a chromosomal *tetO* insertion but these insertions occur at different places relative to the chromosome of strain NCTC 11168. The inserts in strain M1 and strain 23-45 share homology in the central portion with the *tetO* gene, downstream CDS and part of the upstream CDS being conserved between the two.

Unfortunately as both 23-45 and 24-34 have a very similar insert at an identical chromosomal location it is not possible to say whether the insert is mobile, whether a subsection of the insert is mobile or what would constitute a minimum mobile element. It would be interesting to look at other strains to gather more information on insertions of this type.

6.4 Discussion

The gene *tetO* has been found in many *Campylobacter* isolates but usually associated with plasmids. Tetracycline acts by binding to the prokaryotic ribosome and inhibiting the elongation phase of protein synthesis. TetO is a ribosomal protection protein that acts by dislodging tetracycline from the ribosome and allowing aa-tRNA to bind to the A site and thus allows protein synthesis to continue [212]. Both strains 23-45 and 24-34 contain a novel region including a *tetO* gene that has been inserted on the chromosome within a homologue of CDS cj0203 which is predicted to encode a periplasmic protein. For strains 16-60, 16-48, 24-50, 16-02, 25-69 and 25-25 it appears likely that the *tetO* gene is found in conjunction with many other plasmid genes but whether this represents a cryptic plasmid or a large chromosomal insertion of plasmid related DNA remains to be established.

The novel chromosomal insertion in strains 23-45 and 24-34 appears to contain a homologue of the *Helicobacter pylori* transposable element IS607 in addition to *tetO* and other predicted CDSs but it is unclear how these two segments may interact. In addition it appears that in pTet, M1 and 23-45/24-34 the *tetO* gene is always associated with a 172 bp downstream CDS of unknown function that interestingly is present downstream of *tetM* in the transposon tn916 of *Enterococcus faecalis*. This may be required for tetracycline resistance, for transposition, or it may be physically linked to *tetO* without functional consequence. It would be interesting to explore the function of this gene further. These two CDSs are the only conserved CDSs between the insertions in strains 23-45 and 24-34 and that in M1. There are also CDSs with homology to TnpV from *Clostridium perfringens* which is present on a chloramphenicol-resistance transposon. A function for TnpV in

Clostridium perfringens was not found but it was discounted as a transposase [177]. In strain M1 there is only a partial CDS with homology to TnpV (**Fig 6.2** and section 5.2.6). It may be that a transposon is responsible for exchanging these fragments between plasmid and chromosome, picking up extra genes along the way by imperfect excision. Further study of more strains containing chromosomally located *tetO* genes would be needed to identify whether this is indeed a mobile element and if it is mobile what constitutes the minimum mobile element. It would also be interesting to explore the function of the CDS that is located downstream of *tetO* in all of the regions sequenced in this study.

7. Mutagenesis of *C. jejuni* strain 81-176 plasmid pTet

7.1 Introduction

As previously discussed in chapter 3 a predicted CDS with homology to site-specific DNA recombinase genes from other bacteria was found within the plasmid pTet from strain 81-176, flanked by 31 bp inverted repeats. Enclosed within the inverted repeats is a predicted promoter which may drive transcription of the downstream predicted type IV secretion system. It is possible that this putative DNA invertase acts on the repeats to invert the whole region, switching the type IV secretion system on and off as it switches orientation. It was therefore decided to explore the implications of this region further.

7.2 Materials and Methods

7.2.1 Reagents

Reagents for this work were purchased from Oxoid (Hampshire, U.K.), Sigma (Dorset, U.K.) or Roche (Lewes, East Sussex, U.K.) unless otherwise stated.

7.2.2 Strains and plasmids

In addition to strains 81-176 and NCTC 11168 (section 2.1.1) a strain NCTC 11168 *cj0742::cat* mutant (C. Coward and A. Grant; Department of Veterinary Medicine, Cambridge) was used for conjugation experiments (section 7.2.3.5). A kanamycin resistance cassette from plasmid pRY107 and chloramphenicol resistance cassette from plasmid pRY111 were used with permission from P. Guerry (Naval Medical Research Institute, Bethesda, USA) [213].

7.2.3 Methods

7.2.3.1 Primer sequences

Table 7.1: Primer sequences used in this part of the study. Engineered restriction sites are shown in bold.

Name	Sequence 5`-3`	Tm	notes
pTir1	ATTAAGCGTAGTGATCCAATG	54.6	Inverting region
pTir2	GTCAAAGAACTAAAGCAGGAC	53.4	Inverting region
pTir3	TTATCATTCCTGATTCTTTAGC	53.1	Inverting region
pTir3b	AACTCGTCTTTCATTTATTGG	53.7	Inverting region
pTir4	AACTCCACCAAGTGCATAC	53.3	Inverting region
pTir4b	TAAAATGATACATCCCATCG	53.5	Inverting region
Km-3.claI	CC**ATCGAT**TGCGTAAGAACATAGAAAGG	53.29	Km cassette plus ClaI site
Km-P-5.stuI	CT**AGGCCT**AAATGGCTAAAATGAGAATATCAC	55.81	Km cassette plus StuI site
Tet23-r3.claI	CT**ATCGAT**GCTATGATGGTCTATTGTGGTTTATC	59.27	Inverse PCR plus ClaI site
Tet23-r5.stuI	TC**AGGCCT**TCATCGTTCTAAATCTTTTAACTTTTG	58.56	Inverse PCR plus StuI site
Km.verif-5	AGTGGTATGACATTGCCTTC	55.06	Mutant verification
Km.verif-3	GTTCCACATCATAGGTGGTC	55.14	Mutant verification
pT23L	GCTCTAGCATTTTCTAGTCCTG	55.32	DNA invertase for Southern
pT23R	TTTGGATGAAATTTTGGAAG	55.29	DNA invertase for Southern
Ampli.Km-3	TGCGTAAGAACATAGAAAGG	53.29	Km cassette
Ampli.Km-P-5	AAATGGCTAAAATGAGAATATCAC	55.81	Km cassette
Ampli.Km-5	GCTGTTTTCTGGTATTTAAGG	53.53	Km cassette
T27L	CTTGGCAGATGGTAGAAGAC	54.94	Plasmid copy number
T27R	CAAAGGGCTAAGATCAAGAG	54.36	Plasmid copy number
1CTL	AAAGTTGGTTGAATGATTGG	55.05	Chromosomal copy number
1CTR	AGCATACCACAAGAATCCAC	55.06	Chromosomal copy number
1CKL	TTTATCACAGCAGATGCAAG	55.03	Chromosomal copy number
1CKR	GAGAGCGGAGTAACATCAAG	55.11	Chromosomal copy number

7.2.3.1 Growth of *Campylobacter jejuni*

All experiments involving growth of *C. jejuni* were carried out in the laboratory of D. Maskell (Department of Veterinary Medicine, Cambridge). *C. jejuni* was grown under microaerophilic conditions at 42°C in a microaerophilic cabinet, MAC5 VA500 microaerophilic workstation (Don Whitley Scientific, Shipley, U.K.). Microaerophilic conditions consisted of 85% Nitrogen, 5% Oxygen, 5% Hydrogen and 5% Carbon dioxide. *C. jejuni* was grown on Mueller-Hinton (MH) agar (Oxoid) containing Trimethoprim (5 µg/ml). Other supplements included defibrinated horse blood (5% v/v) (TCS Biosciences, Buckingham, U.K.), Tetracycline (20 µg/ml), Chloramphenicol (10 µg/ml) and Kanamycin (25 µg/ml) as appropriate.

7.2.3.2 Chromosomal DNA preparation

7.2.3.2.1 Phenol/Chloroform extraction

C. jejuni was grown for 24-48 hrs on MH agar plus defibrinated horse blood and trimethoprim, cells were harvested in Phosphate Buffered Saline (PBS) (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 2 mM KH_2PO_4 pH7.4) and pelleted by centrifugation for 1 min at 8000 rpm (Eppendorf 5415D) in a 1.5 ml tube (eppendorf). The supernatant was discarded and the cell pellet was resuspended in 1 ml Sucrose Tris EDTA (STE) (TE 10:1 and 25% sucrose) then spun in a centrifuge for 3 mins at 10000 rpm. This step was repeated with pelleting performed for 3 mins at 13000 rpm. The cell pellet was resuspended in 1 ml STE. The cell suspension was placed on ice and 25 µl lysozyme (40 mg/ml in 0.25 M Tris pH8) was added. The sample was incubated on ice for 5 mins then 50 µl of Proteinase K (20 mg/ml) was added and the sample was mixed by inversion. Next 15 µl of RNase (10 mg/ml) was added and the sample incubated on ice for 5mins before 200 µl EDTA (0.5 M) was added. Finally 125 µl Sarkosyl (10% w/v) was added and the sample incubated on ice for 2 hours. The sample was then incubated overnight at 50°C. The following day the sample was transferred to a 50 ml tube (Falcon) and the volume was made up to 5 ml with TE before 5 ml of phenol/chloroform (Sigma) was added. The sample was mixed by inversion then spun in a centrifuge (Eppendorf 5810R) for 40 mins at 4000 rpm and 15°C. The top layer was transferred to a new 50 ml tube (Falcon) and an equal volume of phenol/chloroform added before the sample was spun in a centrifuge for 15 mins at 4000 rpm and 15°C. The previous step was repeated with centrifugation for 10 mins at 4000 rpm and 15°C. The top layer was transferred to a new tube and an equal amount of chloroform (Sigma) was added, then the sample was spun for 1 hour at 4000 rpm. The previous step was repeated with centrifugation for 5 mins at 4000 rpm and 15°C. The top layer was transferred to a new 50 ml tube (Falcon) and 18.5 ml 100% ethanol added, and the sample was spun in a centrifuge for 3 mins at 4000 rpm 4°C. The supernatant was removed and the pellet washed with 10 ml 70% ethanol by spinning in a centrifuge for 3 mins at 4000 rpm and 4°C. The supernatant was removed and the pellet dried by incubating for 10 mins at 37 °C. The dried pellet was resuspended in an appropriate volume of buffer EB (Qiagen).

7.2.3.2.2 Qiagen column prep

Genomic DNA was purified using the Qiagen® Genomic-tip System with a 100/G genomic tip (Qiagen) according to manufacturer's specifications.

7.2.3.3 Quantitative Polymerase Chain Reaction (QPCR)

QPCR was performed using Brilliant SYBR Green QPCR Master Mix (Stratagene, La Jolla, CA, USA) according to the manufacturer's instructions. Reactions were thermocycled using an Mx3000P instrument (Stratagene) with an initial denaturing step of 10 mins 95°C followed by amplification steps of 92°C 30 s, 53°C 30 s and 72°C for 2 mins for 40 cycles. The data generated was analysed with the software provided.

7.2.3.4 Mutagenesis of pTet

7.2.3.4.1 Constructing a suicide vector

PCR of a promoterless kanamycin resistance gene to engineer terminal restriction enzyme sites was conducted as in (section 2.2.4) using the primers Km-3.claI and Km-P-5.stuI (section 7.2.3.1). Inverse PCR was conducted using a pUC clone, 8pT2G1, containing an insert consisting of the DNA invertase region from the pTet shotgun assembly as a template and the primers Tet23-r3.claI and Tet23-r5.stuI. DNA was amplified by thermocycling with the following conditions:-

94°C 4 mins

94°C 10 s
43-55°C 1 min 10 cycles
68°C 5 mins

94°C 10 s
43-55°C 1 min 25 cycles
68°C 5 mins
+10 s/ cycle

72°C 20 mins

10°C holding temperature

PCR products were verified and purified using agarose gel electrophoresis (section 2.2.3) followed by gel extraction (section 2.2.6.1). The purified gel fragments were digested with both ClaI and StuI (section 2.2.5), phenol extracted (section 2.2.6.2) and ethanol precipitated (section 2.2.6.3). The fragments were ligated by mixing 1.5 µl of the inverse PCR product and 1.5 µl of the kanamycin cassette with 0.4 µl of 10x ligase buffer, 0.3 µl of ligase (Roche 5 U/µl) and 0.3 µl of DDW. The ligation mixture was incubated and the ligation reaction terminated (section 2.3.1.2) then the pUC construct was transformed into an

Escherichia coli DH10B host (section 2.3.1.3). Recombinant pUC plasmid DNA was purified and knockout mutation constructs were verified by PCR (section 2.2.4), restriction enzyme digest with *Cla*I and *Stu*I or *Sac*I and *Xba*I (section 2.2.5) and sequencing (section 2.3.6).

7.2.3.4.2 Transformation

Overnight culture plates of *C. jejuni* strain 81-176 were harvested and spotted onto fresh MH agar plates containing defibrinated horse blood and trimethoprim. The plates were incubated under microaerophilic conditions for 3 hours at 42°C, then harvested and spotted onto fresh plates plus 1-10 μg of pUC construct DNA. The transformation mixtures were incubated for 4 hrs under microaerophilic conditions before serially diluting and plating onto selective MH agar plates containing defibrinated horse blood, trimethoprim, tetracycline and kanamycin.

7.2.3.4.3 Colony PCR

In order to verify the incorporation of the kanamycin resistance cassette in the correct location on pTet in *C. jejuni* strain 81-176, PCR was performed (section 2.2.4), with colonies used as templates instead of purified DNA. Colonies were suspended in water and boiled for 10 mins before being added to the reaction mix.

7.2.3.5 Southern blotting

Southern blotting was performed according to established methods: DNA was digested (section 2.2.5), fragments were separated by agarose gel electrophoresis (section 2.2.3), DNA fragments were transferred to Nytran N+ membrane (Amersham, Buckinghamshire, UK) using capillary transfer [214]. The membrane was then hybridized to a radioactively labelled oligonucleotide probe generated by PCR (section 2.3.5.2).

7.2.3.5 Conjugation experiments

Conjugation experiments were performed as previously described [128]. Donor strains 81-176 (pTet) or 81-176 (pTet/pTet23ΔKm), and the recipient strain NCTC 11168 *Cj0742::cat* were grown for approximately 24 hours on MH agar plus defibrinated horse blood, trimethoprim and the appropriate selective antibiotics (section 7.2.3.1). Strain 81-176 (pTet) was grown with tetracycline (20 µg/ml), strain 81-176 (pTet/pTet23ΔKm) was grown with tetracycline (20 µg/ml) and kanamycin (25 µg/ml), and strain NCTC 11168 *Cj0742::cat* was grown with chloramphenicol (10 µg/ml). Cells were harvested in MH broth at an optical density (OD) at 600 nm of 1 giving approximately 10^8 Colony Forming Units (CFU) per ml. Cells from each strain were serially diluted in MH broth and spread in triplicate on MH agar plates containing trimethoprim at this stage to determine CFU per ml. Equal quantities of donor and recipient cells (50 µl) were mixed and DNaseI (10 U/ml, Roche) was added. This conjugation mixture was then spotted onto MH agar plates containing defibrinated horse blood and trimethoprim but without selective antibiotics. Strains were also spotted individually onto MH agar plates containing defibrinated horse blood and trimethoprim as controls. The agar plates were incubated under microaerophilic conditions for 16-18 hours at 42°C. Subsequently the cells were harvested in MH broth, serially diluted and spread in triplicate onto MH agar plates containing trimethoprim with and without tetracycline and chloramphenicol and incubated under microaerophilic conditions for 2-3 days at 42°C.

7.3 Results

7.3.1 Invertible region

The putative CDS pTet23, which is predicted to encode a DNA invertase, is located on a region of the plasmid pTet that is flanked by two sets of inverted repeats (**Fig 7.1A**). In order to determine whether this region is inverting and whether inversion occurs using the outer repeats, the inner repeats, or both, a set of oligonucleotide primers were designed (**Fig 7.1**). PCR amplification using various primer combinations was conducted using prepared pTet plasmid DNA from strain 81-176 in order to identify whether inversion occurs (**Fig 7.2** and **Fig 7.3**). Fig 7.2 shows that products are obtained between primer sets pTir1-pTir2 and pTir3-pTir4 representing the invertible region in the sequenced orientation and between pTir2-pTir4 and pTir1-pTir3 representing the invertible region in the inverted orientation. The obtained product sizes are consistent with the predicted product sizes if only the outer repeats are used for inversion (**Fig 7.1**). As the product sizes for the different orientations are very similar the PCR amplifications were repeated using different primer sets to obtain different sized products to double check the inversion (**Fig 7.3**). These product sizes are again consistent with predicted product sizes if the region inverts using the outer repeats only (**Fig 7.1**).

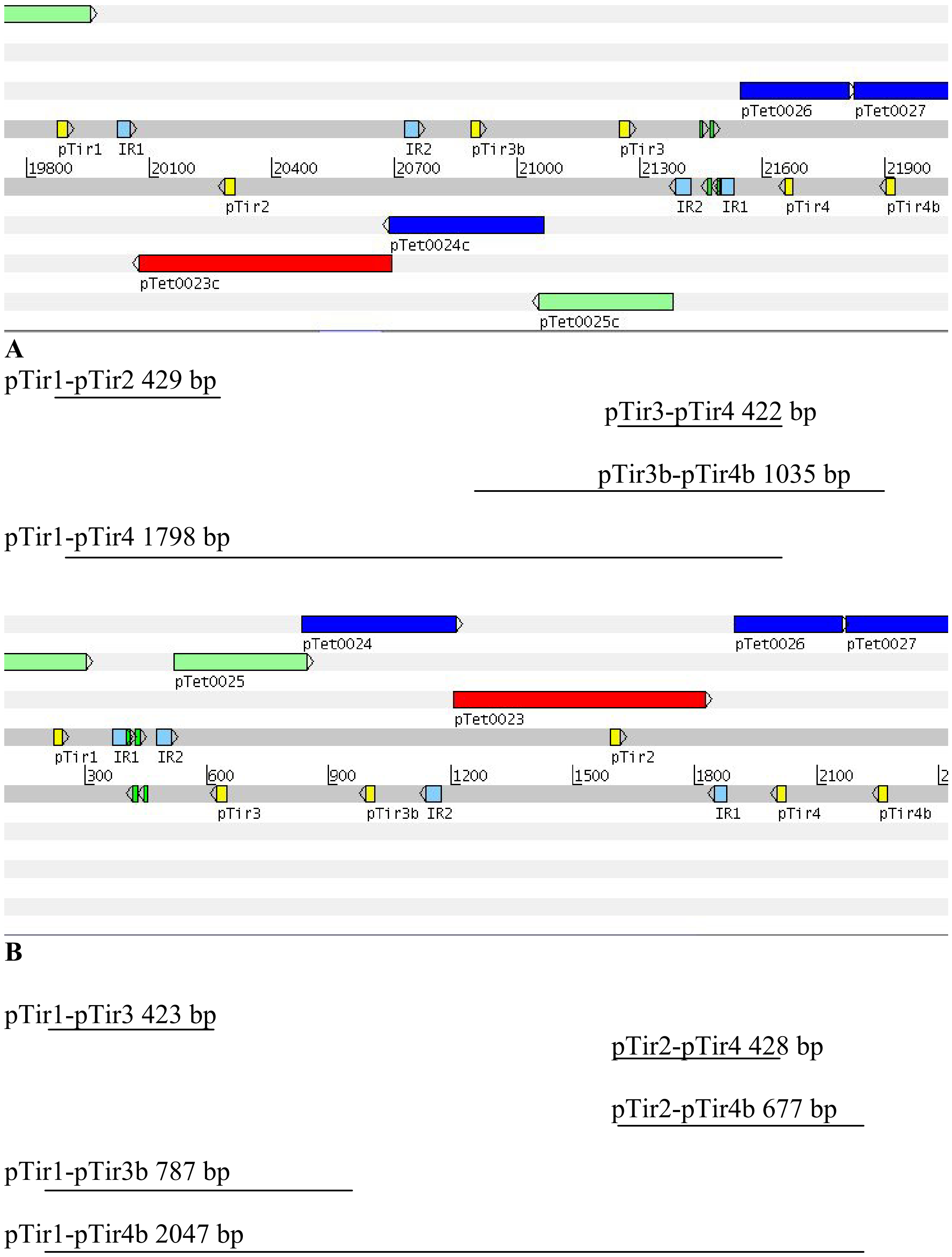


Fig 7.1: *C. jejuni* strain 81-176 plasmid pTet invertible region. A, invertible region in the sequenced orientation; **B**, invertible region in the inverted orientation. Regions are viewed using Artemis [102]. The forward and reverse DNA sequences are represented by dark grey lines. The three-frame forward and reverse DNA translations are represented by light grey lines. Features are represented by open boxes: oligonucleotide primer sequences (yellow), inverted repeats (blue) and promoters (green), are marked on the DNA lines; CDSs are marked on the translated frame lines.

CDSs are coloured according to functional category: light green, unknown; red, information transfer; blue, pathogenicity/ adaptation/ chaperones. Below the diagrams the expected sizes of PCR amplification products are indicated.

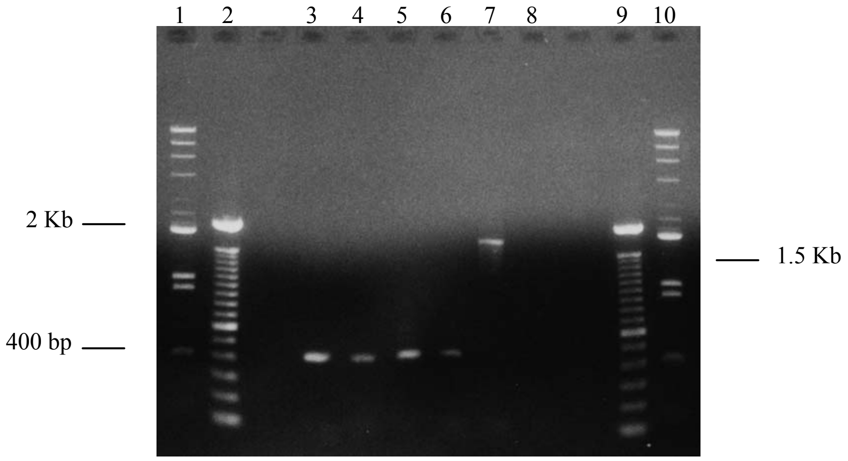


Fig 7.2: PCR of invertible region from extracted *C. jejuni* 81-176 plasmid DNA. Lanes 1 and 10 contain a mixture of λ *Hin*dIII and pBR322 *Bst*NI digested DNA, and lanes 2 and 9 contain 100 bp marker. Strain 81-176 plasmid DNA was amplified with the following primers: lane 3, pTir1-pTir2; lane 4, pTir3-pTir4; lane 5, pTir2-pTir4; lane 6, pTir1-pTir3; lane 7, pTir1-pTir4; lane 8, pTir3-pTir2. These results are consistent with the invertible region inverting between inverted repeat set 1.

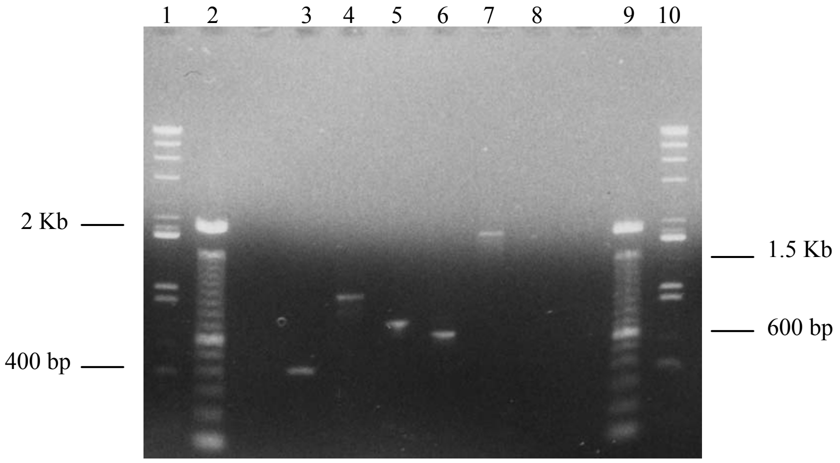


Fig 7.3: PCR of invertible region from extracted *C. jejuni* plasmid DNA. Lanes 1 and 10 contain a mixture of λ *Hin*dIII and pBR322 *Bst*NI digested DNA, and lanes 2 and 9 contain 100 bp marker. Strain 81-176 plasmid DNA was amplified with the following primers; lane 3, pTir1-pTir2; lane 4, pTir3b-pTir4b; lane 5, pTir1-pTir3b; lane 6, pTir2-pTir4b; lane 7, pTir1-pTir4b; lane 8, pTir3b-pTir2. These results suggest that the region is inverting between inverted repeat set 1.

The PCR products between primer sets pTir1-pTir2, pTir3-pTir4, pTir2-pTir4 and pTir1-pTir3 were used as sequencing templates and sequenced using each of the primers used to generate the products. These products were selected for sequencing as they were small enough to be sequenced completely from both directions and would cover the inverted repeats so the exact site of inversion could be verified. The sequences from the inverted orientation (pTir2-pTir4 and pTir1-pTir3) were then aligned using clustal X to the sequence in the 'sequenced' orientation to illustrate the inversion and to confirm that this was occurring at the inverted repeats (**Fig 7.4**). Fig 7.4 shows that the invertible region inverts within the perfect repeat area of IR1.

```
primer1->primer2       AAAATCTTTATTTGATTAACATTTACAATCAAGGAGTGTTGTATGAATGAATGGGAATTA
                       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
primer1->primer3       AAAATCTTTATTTGATTAACATTTACAATCAAGGAGTGTTGTATGAATGAATGGGAATTA
                         :::         :  ::     : :::      :  :   :: :::::        :
primer4->primer3       TCAATACCGCCAGCACCAAATGCAAAAATAGGACTGATTAAAATTAATGATAATAGCCTT

                                                            IR1
                                                 <---------------------------->
primer1->primer2       GAAAGACAAAGACAAATAAGATAGACAAATTGAGCGTTTTATTTTATTACAAAATTTTTT
                       ::::::::::::::::::::::::::::::::::::::::::::   :   :::::: :
primer1->primer3       GAAAGACAAAGACAAATAAGATAGACAAATTGAGCGTTTTATTTATGTCTTAAATTTGTA
                                       :    :  : ::::::::::::::::::::::::::::::::::
primer4->primer3       TTTTTCATCCATATCCTTTCGTTAAAAAATTGAGCGTTTTATTTATGTCTTAAATTTGTA


primer1->primer2       ATATTTTTCTTAAATGAAAATTTAGCTTCCTTTTAATTTTCGTTTTTTGATAAACCACAA
                        :   :   : :: : ::  :     :    :: ::::    :: ::  : :      :
primer1->primer3       TTTAATAAATAAATTAAATTTAGTTTTATAATTAAATTAATATTATTCAAAAGCTTTTAT
                       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
primer4->primer3       TTTAATAAATAAATTAAATTTAGTTTTATAATTAAATTAATATTATTCAAAAGCTTTTAT


primer1->primer2       TAGACCATCATAGCTTTTTCCATTATCTTTGTATAATAACTTCCAAATACTTTTAATAGA
                       :  :  :: :    :::::   :    :::::    :   :    ::    :  :: :
primer1->primer3       TTAAGTATTAGCTATTTTTTACTAGAATTTGTTAGGTTTTTGATGAAAGGATAAAAAAAT
                       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
primer4->primer3       TTAAGTATTAGCTATTTTTTACTAGAATTTGTTAGGTTTTTGATGAAAGGATAAAAAAAT


primer1->primer2       TAAATCTTTATCAAGCAAGGTTTGTATTTTATCTATATCCTTATCATAAA
                          ::    :             : ::  :::          :
primer1->primer3       GGCATACCCACTTTTAGGTACAAGAATAGTATTAGATGAAGAAAAGATTT
                       ::::::::::::::::::::::::::::::::::::::::::::::::::
primer4->primer3       GGCATACCCACTTTTAGGTACAAGAATAGTATTAGATGAAGAAAAGATTT
```

A

```
primer1->primer2        AAAATCTTTATTTGATTAACATTTACAATCAAGGAGTGTTGTATGAATGAATGGGAATTA
                          :::         :  ::     : :::      :  :   :: :::::        :
primer4->primer2        TCAATACCGCCAGCACCAAATGCAAAAATAGGACTGATTAAAATTAATGATAATAGCCTT
                        ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
primer4->primer3        TCAATACCGCCAGCACCAAATGCAAAAATAGGACTGATTAAAATTAATGATAATAGCCTT

                                                             IR1
                                                  <----------------------------->
primer1->primer2        GAAAGACAAAGACAAATAAGATAGACAAATTGAGCGTTTTATTTTATTACAAAATTTTTT
                                        :    :  : ::::::::::::::::::::::::::::::::::
primer4->primer2        TTTTTCATCCATATCCTTTCGTTAAAAAATTGAGCGTTTTATTTTATTACAAAATTTTTT
                        ::::::::::::::::::::::::::::::::::::::::::::   :   :::::: :
primer4->primer3        TTTTTCATCCATATCCTTTCGTTAAAAAATTGAGCGTTTTATTTATGTCTTAAATTTGTA


primer1->primer2        ATATTTTTCTTAAATGAAAATTTAGCTTCCTTTTAATTTTCGTTTTTTGATAAACCACAA
                        ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
primer4->primer2        ATATTTTTCTTAAATGAAAATTTAGCTTCCTTTTAATTTTCGTTTTTTGATAAACCACAA
                         :   :   : :: : ::  :     :    :: ::::    :: ::  : :      :
primer4->primer3        TTTAATAAATAAATTAAATTTAGTTTTATAATTAAATTAATATTATTCAAAAGCTTTTAT


primer1->primer2        TAGACCATCATAGCTTTTTCCATTATCTTTGTATAATAACTTCCAAATACTTTTAATAGA
                        ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
primer4->primer2        TAGACCATCATAGCTTTTTCCATTATCTTTGTATAATAACTTCCAAATACTTTTAATAGA
                        :  :  :: :    :::::   :    :::::    :   :    ::    :  :: :
primer4->primer3        TTAAGTATTAGCTATTTTTTACTAGAATTTGTTAGGTTTTTGATGAAAGGATAAAAAAAT


primer1->primer2        TAAATCTTTATCAAGCAAGGTTTGTATTTTATCTATATCCTTATCATAAA
                        ::::::::::::::::::::::::::::::::::::::::::::::::::
primer4->primer2        TAAATCTTTATCAAGCAAGGTTTGTATTTTATCTATATCCTTATCATAAA
                           ::    :             : ::  :::          :
primer4->primer3        GGCATACCCACTTTTAGGTACAAGAATAGTATTAGATGAAGAAAAGATTT
```

B

```
primer2->primer1       TTTATGATAAGGATATAGATAAAATACAAACCTTGCTTGATAAAGATTTATCTATTAAAA
                              :          :::  :: :             :    ::     : ::  :
primer3->primer1       AAATCTTTTCTTCATCTAATACTATTCTTGTACCTAAAAGTGGGTATGCCATTTTTTTAT
                       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
primer3->primer4       AAATCTTTTCTTCATCTAATACTATTCTTGTACCTAAAAGTGGGTATGCCATTTTTTTAT


primer2->primer1       GTATTTGGAAGTTATTATACAAAGATAATGGAAAAAGCTATGATGGTCTATTGTGGTTTA
                          ::    :   :    :::::    :   :::::    : ::  :  : :      :
primer3->primer1       CCTTTCATCAAAAACCTAACAAATTCTAGTAAAAAATAGCTAATACTTAAATAAAAGCTT
                       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
primer3->primer4       CCTTTCATCAAAAACCTAACAAATTCTAGTAAAAAATAGCTAATACTTAAATAAAAGCTT


                                                                            <------
primer2->primer1       TCAAAAAACGAAAATTAAAAGGAAGCTAAATTTTCATTTAAGAAAAATATAAAAAATTTT
                       :  :: ::    :::: ::    :     :  :: : :: :   :   :  : :::::::
primer3->primer1       TTGAATAATATTAATTTAATTATAAAACTAAATTTAATTTATTTATTAAATACAAATTTA
                       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
primer3->primer4       TTGAATAATATTAATTTAATTATAAAACTAAATTTAATTTATTTATTAAATACAAATTTA

                           IR1
                       ----------------------->
primer2->primer1       GTAATAAAATAAAACGCTCAATTTGTCTATCTTATTTGTCTTTGTCTTTCTAATTCCCAT
                         :   ::::::::::::::::::::::::::::::::::::::::::::::::::::::
primer3->primer1       AGACATAAATAAAACGCTCAATTTGTCTATCTTATTTGTCTTTGTCTTTCTAATTCCCAT
                       :::::::::::::::::::::::: :  :    :                 :
primer3->primer4       AGACATAAATAAAACGCTCAATTTTTTAACGAAAGGATATGGATGAAAAAAAGGCTATTA


primer2->primer1       TCATTCATACAACACTCCTTGATTGTAAATGTTAATCAAATAAAGATTTT
                       ::::::::::::::::::::::::::::::::::::::::::::::::::
primer3->primer1       TCATTCATACAACACTCCTTGATTGTAAATGTTAATCAAATAAAGATTTT
                       ::::: ::   :  :      ::: :     ::  :         :::
primer3->primer4       TCATTAATTTTAATCAGTCCTATTTTTGCATTTGGTGCTGGCGGTATTGA
```

C

```
primer2->primer1       TTTATGATAAGGATATAGATAAAATACAAACCTTGCTTGATAAAGATTTATCTATTAAAA
                       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
primer2->primer4       TTTATGATAAGGATATAGATAAAATACAAACCTTGCTTGATAAAGATTTATCTATTAAAA
                              :          :::  :: :             :    ::     : ::  :
primer3->primer4       AAATCTTTTCTTCATCTAATACTATTCTTGTACCTAAAAGTGGGTATGCCATTTTTTTAT


primer2->primer1       GTATTTGGAAGTTATTATACAAAGATAATGGAAAAAGCTATGATGGTCTATTGTGGTTTA
                       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
primer2->primer4       GTATTTGGAAGTTATTATACAAAGATAATGGAAAAAGCTATGATGGTCTATTGTGGTTTA
                          ::    :   :    :::::    :   :::::    : ::  :  : :      :
primer3->primer4       CCTTTCATCAAAAACCTAACAAATTCTAGTAAAAAATAGCTAATACTTAAATAAAAGCTT

                                                                            <------
primer2->primer1       TCAAAAAACGAAAATTAAAAGGAAGCTAAATTTTCATTTAAGAAAAATATAAAAAATTTT
                       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
primer2->primer4       TCAAAAAACGAAAATTAAAAGGAAGCTAAATTTTCATTTAAGAAAAATATAAAAAATTTT
                       :  :: ::    :::: ::    :     :  :: : :: :   :   :  : ::::::
primer3->primer4       TTGAATAATATTAATTTAATTATAAAACTAAATTTAATTTATTTATTAAATACAAATTTA

                           IR1
                       ----------------------->
primer2->primer1       GTAATAAAATAAAACGCTCAATTTGTCTATCTTATTTGTCTTTGTCTTTCTAATTCCCAT
                       :::::::::::::::::::::::: :  :    :                 :
primer2->primer4       GTAATAAAATAAAACGCTCAATTTTTTAACGAAAGGATATGGATGAAAAAAAGGCTATTA
                         :   ::::::::::::::::::::::::::::::::::::::::::::::::::::::
primer3->primer4       AGACATAAATAAAACGCTCAATTTTTTAACGAAAGGATATGGATGAAAAAAAGGCTATTA


primer2->primer1       TCATTCATACAACACTCCTTGATTGTAAATGTTAATCAAATAAAGATTTT
                       ::::: ::   :  :      ::: :     ::  :         :::
primer2->primer4       TCATTAATTTTAATCAGTCCTATTTTTGCATTTGGTGCTGGCGGTATTGA
                       ::::::::::::::::::::::::::::::::::::::::::::::::::
primer3->primer4       TCATTAATTTTAATCAGTCCTATTTTTGCATTTGGTGCTGGCGGTATTGA
```

D

Fig 7.4: clustal alignments of DNA sequence from inverted orientation PCR products. The middle sequence represents the invertible region sequenced in the inverted orientation between the primers specified to the left of the sequence. The outer sequences represent the invertible region in the sequenced orientation between the primers specified to the left of the sequence. The inverted repeat, IR1 is marked above the alignments; **A** and **B** show the left hand repeat, and **C** and **D** show the right hand repeat. The alignments show that the region is inverting within the perfect repeat area of IR1.

The sequenced orientation of this invertible region appears to be predominant within the purified DNA sample used for these experiments as during the shotgun assembly this region was not identified in the inverted orientation and all attempts to sequence products obtained with primers pTir1 and pTir4 showed the region in the sequenced orientation. However a BAC clone was obtained from 81-176 DNA that contained the region in the inverted orientation. If the plasmid is high copy number then the putative variable promoter may have no observable effect at the bacterial cell level, also if the plasmid is multi-copy then any mutant construct may be compromised by homologous recombination with a wild type plasmid if the strains are *recA* positive. For this reason it was decided to experimentally determine the plasmid copy number.

7.3.2 Determination of plasmid copy number

Usually low copy number plasmids have partition genes which allow the plasmid to be incorporated into daughter cells whereas high copy number plasmids can rely upon random distribution between daughter cells [114]. As no homologues of partition genes were discovered in pTet it was decided to determine the relative plasmid copy number using QPCR to amplify regions from the chromosome and from the plasmid.

The fluorescence of SYBR Green dye increases when bound to double stranded DNA allowing the accumulation of PCR product to be measured. To determine the relative copy number the threshold cycle (Ct) is used which is the lowest cycle of amplification at which the fluorescence is detectable above the background level and therefore product is starting to be formed. The more initial template DNA there is the fewer cycles are necessary to produce fluorescence that is detectable above the background level. The value Ct has been shown to be inversely proportional to the log of the initial copy number of the starting material [215]. Comparing measurements from the initial amplification stage is more accurate than from subsequent cycles when the availability of reaction components can be a limiting factor in product amplification.

QPCR (section 7.2.3.3) was performed using total DNA extracted by a phenol/chloroform method (section 7.2.3.2.1) as a template. Primer sets 1CTL/R (designed from *pheT* of NCTC 11168) and 1CKL/R (designed from *dnaK* of NCTC 11168) were used to amplify regions of the chromosome and primer set T27L/R (designed from pTet27) was used to amplify a region from pTet. Primers were designed from the chromosome in genes not likely to be variable between strains 81-176 and NCTC 11168 as they perform housekeeping functions, at different positions around the chromosome. Primers were designed to give products of approximately equal sizes: 1CTL/R 340 bp; 1CKL/R 341 bp; T27L/R 367 bp. The PCRs were performed in triplicate using 1/10, 1/20 and 1/50 dilutions of the 81-176 total DNA as a template. Amplification plots from the QPCR reaction (**Fig 7.5**) indicate that for each of the dilutions the fluorescence produced using each of the primer pairs begins to exceed the background level at approximately the same amplification cycle and that there is a difference between each of the dilutions. The relative quantities were determined using the amplification product from the primer set 1CKL/R with a 1/50 dilution of the template DNA as a reference to which all other values were compared (**Fig 7.6**). The relative quantities for each primer set show that a 1/20 dilution of template is approximately 2.5x greater than 1/50 dilution of template, and 1/10 dilution of template is approximately 5x greater than 1/50 dilution of template. The relative quantities between primer sets for each dilution are approximately equal which shows that pTet is a low copy number plasmid.

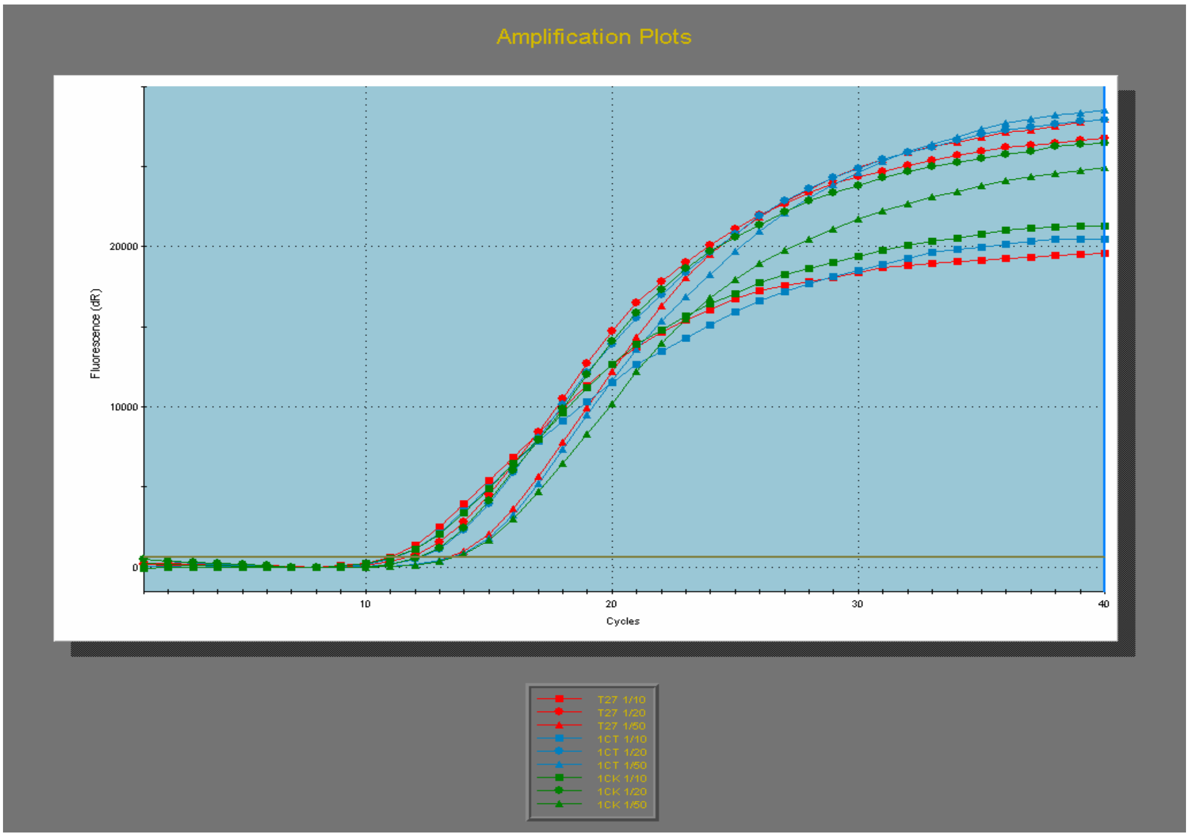


Fig 7.5: QPCR amplification plots for the determination of plasmid copy number. Primer sets (pTet, T27; chromosomal, 1CT and 1CK) and dilutions of template DNA are indicated in the key. The horizontal gold line indicates the threshold fluorescence value. The threshold cycle (Ct), where the fluorescence from an amplification reaction is just detectable above the threshold value, has been shown to be inversely proportional to the log of the initial copy number of the starting material [215]. From this graph it can be seen that fluorescence for each of the primer pairs, at the same dilution of template DNA, begins to exceed the threshold level at approximately the same amplification cycle. There is a visible difference in Ct for each of the template dilutions, with the lower dilutions producing detectable fluorescence after fewer cycles, as expected.

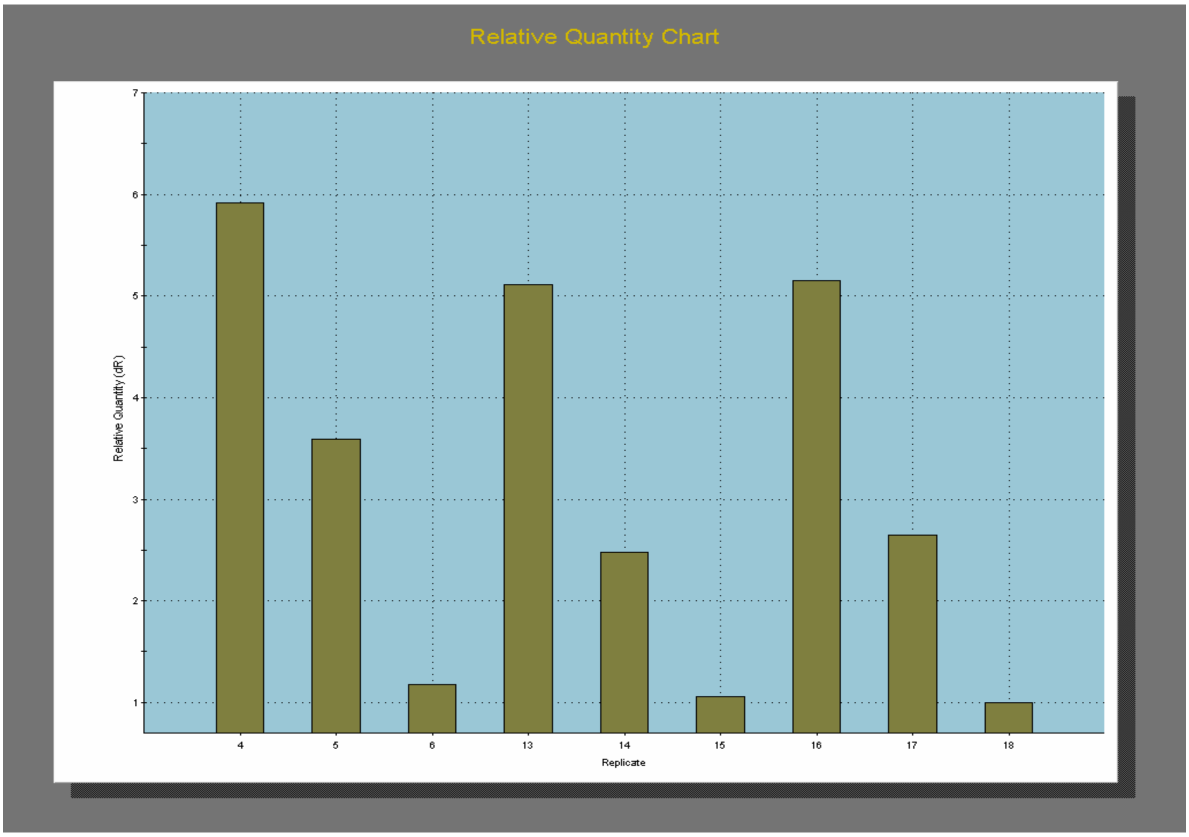


Fig 7.6: QPCR Relative Quantity Chart. Relative quantities are determined from the Ct values of each amplification compared to the Ct value of replicate 18 which is given a value of 1. Primer sets to amplify a region from pTet, Tet27, and regions from the chromosome, 1CT and 1CK, were used at various concentrations of extracted total DNA from strain 81-176. Each of the replicates is based on an average of the results for each primer set at each dilution of template DNA. Relative quantities of pTet DNA, determined from primer set Tet27, are shown in replicates 4, 5 and 6, at a 1/10, 1/20 and 1/50 dilution of template DNA respectively. Relative quantities of chromosomal DNA, determined from primer set 1CT, are shown in replicates 13, 14 and 15, at a 1/10, 1/20 and 1/50 dilution of template DNA respectively. Relative quantities of chromosomal DNA, determined from primer set 1CK, are shown in replicated 16, 17 and 18, at a 1/10, 1/20 and 1/50 dilution of template DNA respectively. The 1/20 dilutions are roughly 2.5 x the 1/50 dilutions, and 1/10 dilutions are roughly 5 x the 1/50 dilutions, within one primer set. The comparison of quantities between primer sets indicates that pTet is a low copy number plasmid.

7.3.3 Knockout mutagenesis of the site specific recombinase pTet0023

A pUC plasmid, 8pT2G1, from the library used for shotgun sequencing (chapter 3) containing the invertible region was used to create an antibiotic cassette knockout of the invertase in order to identify whether this gene is responsible for the inversion seen in the plasmid. The plasmid 8pT2G1 contains an insert of 2615 bp from base 19297-21911 on the pTet plasmid (**Fig 7.7**). Inverse PCR with the primers Tet23-r3.claI and Tet23-r5.stuI was used to amplify 8pT2G1 minus the invertase and to engineer terminal *Cla*I and *Stu*I restriction enzyme sites respectively. A kanamycin resistance gene (*apha-3)* from plasmid pRY107 was amplified by PCR without its promoter using the primers Km-P-5.stuI and Km-3.claI to engineer terminal restriction enzyme sites that would be complementary to the amplified plasmid 8pT2G1. Both of these products were digested with *Stu*I and *Cla*I then ligated together resulting in a pUC plasmid, 8pT2G1ΔKm, with the putative DNA invertase, pTet0023, replaced with a promoterless kanamycin resistance gene (section 7.2.3.4.1) (**Fig 7.8**). A promoterless kanamycin resistance gene was used so as not to affect transcription of downstream genes by introducing a strong promoter that could cause read through transcription. Knockout constructs were verified by PCR, restriction enzyme digests and sequencing. The invertible region was shown by PCR to invert in the plasmid 8pT2G1 within the *Escherichia coli* host (**Fig 7.9**). Once the putative DNA invertase, pTet23c, had been replaced with a kanamycin resistance gene the region no longer inverted within the *Escherichia coli* host suggesting that this region was now fixed in one orientation (**Fig 7.10**). As the wild-type region inverts in *Escherichia coli* some of the 8pT2G1 template used for the inverse PCR would have had the invertible region in the inverted orientation. Attempts were made to isolate an 8pT2G1ΔKm clone containing an insert with the invertible region fixed in the alternative direction. DNA from pools of colonies was analysed by restriction enzyme digests, PCR and sequencing but a mutant plasmid with the invertible region fixed in the inverted orientation could not be purified from the background of clones with the invertible region in the sequenced orientation.

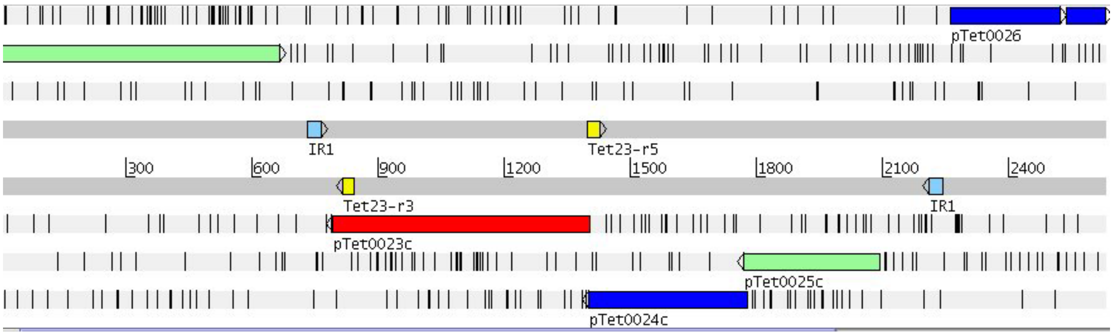


Fig 7.7: Insert sequence of the pUC clone 8pT2G1. The region is viewed using Artemis [102]. The forward and reverse DNA sequences are represented by dark grey lines. The three-frame forward and reverse DNA translations are represented by light grey lines with stop codons indicated by vertical black lines. Features are represented by open boxes: oligonucleotide primer sequences (yellow) and inverted repeats (blue) are marked on the DNA lines; CDSs are marked on the translated frame lines. CDSs are coloured according to functional category: light green, unknown; red, information transfer; blue, pathogenicity/ adaptation/ chaperones. A mutant was constructed using inverse PCR (from the sites indicated) to amplify this region without the invertase, pTet23, and to engineer restriction enzyme sites in order to attach a kanamycin resistance cassette in its place.

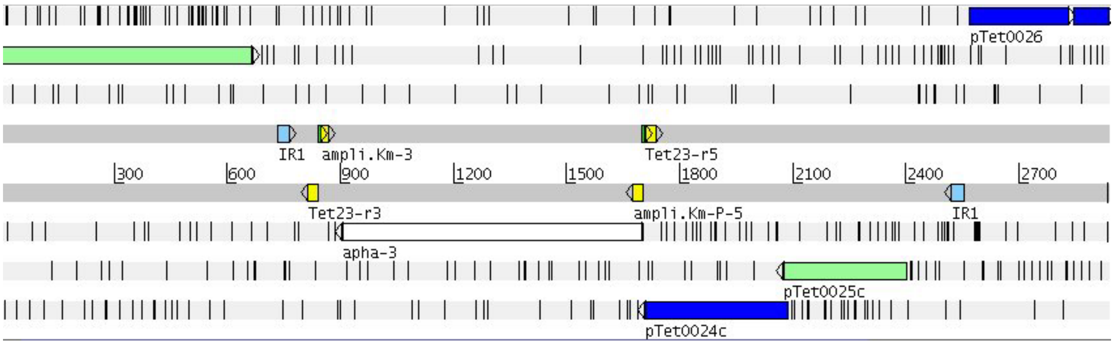


Fig 7.8: Insert sequence of the pUC clone 8pT2G1ΔKm. This region is viewed using Artemis [102]. The forward and reverse DNA sequences are represented by dark grey lines. The three-frame forward and reverse DNA translations are represented by light grey lines and stop codons are marked by vertical black lines. Features are represented by open boxes: oligonucleotide primer sequences (yellow), inverted repeats (blue) and promoters (green), are marked on the DNA lines; CDSs are marked on the translated frame lines. CDSs are coloured according to functional category: light green, unknown; red, information transfer; blue, pathogenicity/ adaptation/ chaperones. The kanamycin resistance cassette is coloured white and replaces the invertase in this mutant construct.

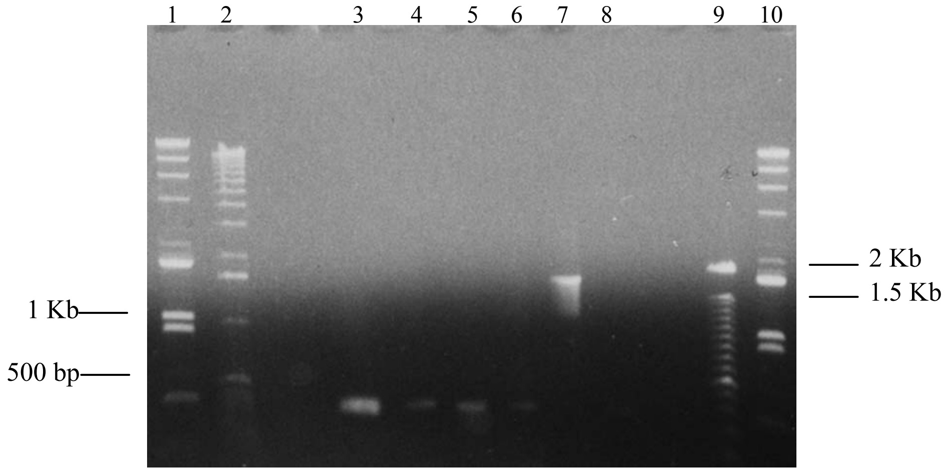


Fig 7.9: PCR of invertible region from extracted 8pT2G1 DNA. Lanes 1 and 10 contain a mixture of λ *Hin*dIII and pBR322 *Bst*NI digested DNA, lane 2 contains 1 Kb marker DNA, and lane 9 contains 100 bp marker DNA. Sample lanes contain 8pT2G1 plasmid DNA amplified with the following primers; lane 3, pTir1-pTir2; lane 4, pTir1-pTir3; lane 5, pTir3-pTir4; lane 6, pTir2-pTir4; lane 7, pTir1-pTir4; lane 8, pTir3-pTir2. The fact that bands are produced from combinations of primers spanning the inverted repeats shows that the region is able to invert within an *Escherichia coli* host.

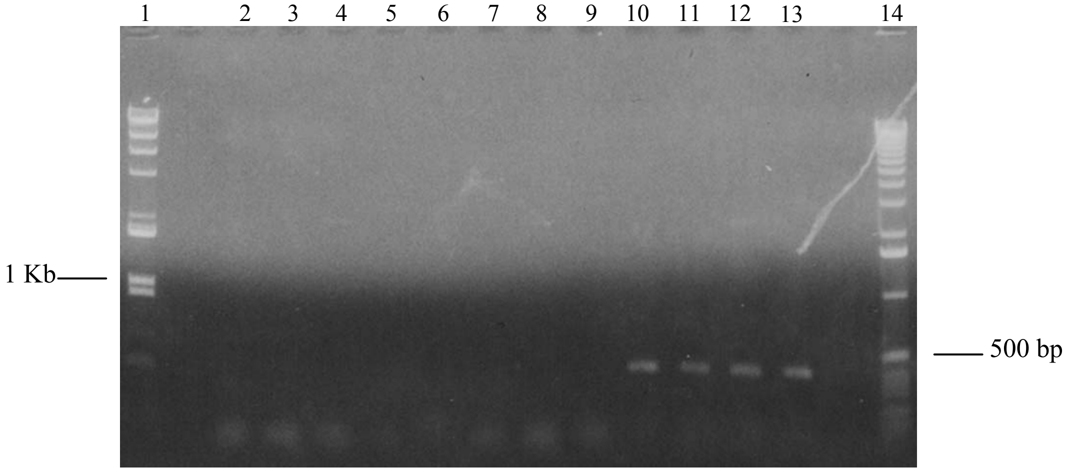


Fig 7.10: PCR amplification between various primers in 8pT2G1ΔKm colonies. Lane 1 contains a mixture of λ *Hin*dIII and pBR322 *Bst*NI digested DNA and lane 14 contains 1 Kb marker DNA. Lanes 2, 3, 4 and 5 contain amplifications products produced using primers pTir1-pTir2 in 8pT2G1ΔKm colonies A11, E2, D3 and G6 respectively; lanes 6, 7, 8 and 9 contain amplification products produced using primers pTir1-pTir3 in 8pT2G1ΔKm colonies A11, E2, D3 and G6 respectively; lanes 10, 11, 12 and 13 contain amplification products produced using primers pTir3-pTir4 in 8pT2G1ΔKm colonies A11, E2, D3 and G6 respectively. Bands are only produced with the primers pTir3-pTir4 showing that the invertase has been deleted and the invertible region has been fixed in the sequenced orientation.

The purified pUC plasmid, 8pT2G1ΔKm, was used as a suicide vector to transform *C. jejuni* strain 81-176 (section 7.2.3.4.2) with the intention that the knockout mutation would be transferred to the indigenous pTet plasmid. Cells were then plated onto MH agar plates containing defibrinated horse blood, trimethoprim, kanamycin and tetracycline to select for recombinant knockout mutants. *C. jejuni* colonies that grew on these selective plates were checked by colony PCR to assess whether the putative DNA invertase gene on the mutagenized pTet had been replaced with the kanamycin resistance gene, by homologous recombination with the suicide vector. Primers flanking the invertible region (pTir1-pTir4) were used to amplify DNA from 12 colonies (**Fig 7.11**).

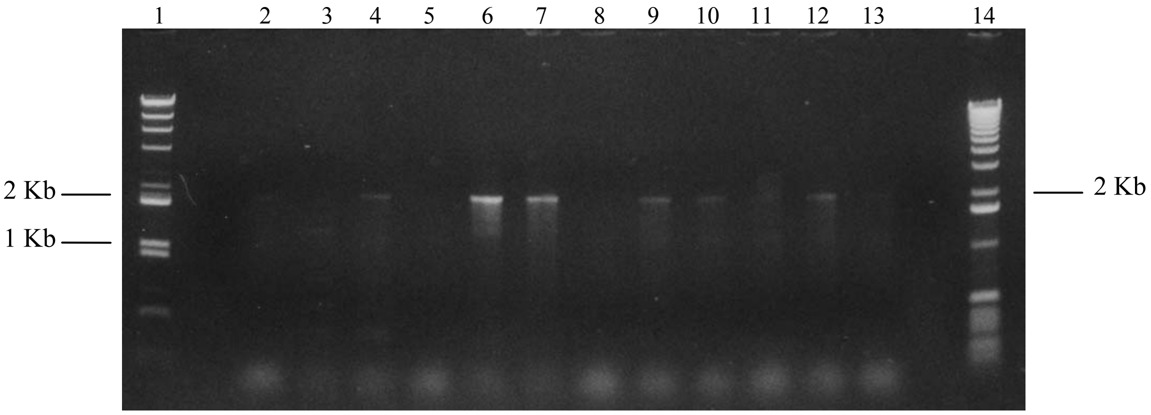


Fig 7.11: Colony PCR from 12 colonies that grew on selective media after transformation with mutant construct using primers pTir1-pTir4. Lane 1 contains a mixture of λ *Hin*dIII and pBR322 *Bst*NI digested DNA and lane 14 contains 1 Kb marker DNA. Lanes 2-13 contain amplification products, from 12 colonies that grew on selective media, produced using primers pTir1-pTir4. Only colonies 5 and 6 (products in lanes 6 and 7) were selected for further analysis as the bands were of the right size (approximately 2 Kb) and lanes did not appear to contain spurious products.

The colonies with amplification products in lanes 6 and 7 (**Fig 7.11**) were selected for further experiments as the amplification products were of the expected size, 2108 bp, compared to wild type, 1799 bp. DNA was extracted using a genomic DNA kit (7.2.3.2.2) and PCR was used to verify the position of the kanamycin resistance gene insert. **Fig 7.12** shows 81-176 (pTet/pTet23 Km) mutant 5 compared to 81-176 (pTet) using primers pTir1-pTir4. There is a larger band for the mutant as the kanamycin resistance gene is larger than the DNA invertase gene it replaces (2108 bp compared to 1799 bp). **Fig 7.13** shows that the kanamycin gene has inserted in the expected orientation and that the kanamycin gene is present without promoter, the band size of 855 bp is consistent with the kanamycin gene having no promoter (**Fig 7.14**). **Fig 7.15** shows that the region no longer inverts within strain 81-176 indicating that the DNA invertase gene is indeed responsible for the inversion in the *C. jejuni* background.

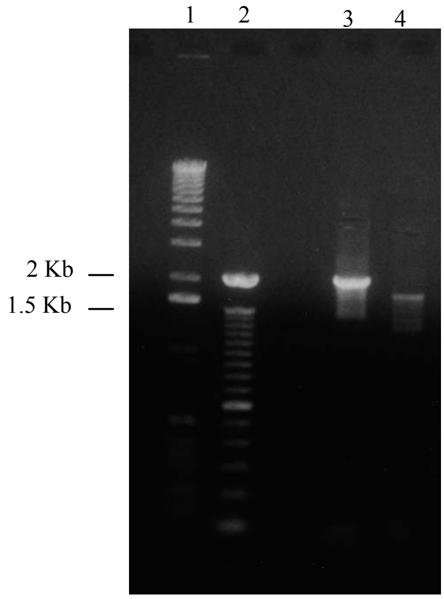


Fig 7.12: PCR amplification of the predicted invertible region from pTet. Lane 1 contains 1 Kb DNA marker and lane 2 contains 100 bp DNA marker. Lanes 3 and 4 contain PCR amplification products produced using primers pTir1-pTir4 from template DNA: lane 3, 81-176 (pTet/pTet23ΔKm) mutant 5; lane 4, wild type 81-176 (pTet). The PCR product from the mutant is larger (approx. 2 Kb) than the wild type (approx 1.8 Kb) which is expected if the kanamycin resistance cassette has replaced the invertase.

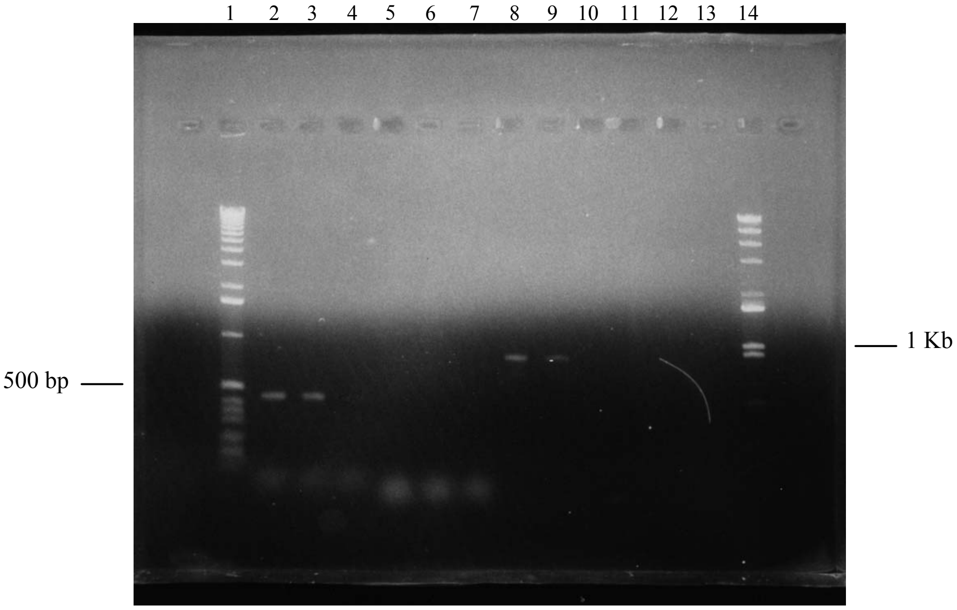


Fig 7.13: PCR amplification of pTet mutants and WT. Lane 1 contains 1 Kb DNA marker and lane 14 contains a mixture of λ *Hin*dIII and pBR322 *Bst*NI digested DNA. Lanes 2, 3 and 4 contain amplifications using primers Km.verif-5-pTir1 in strain 81-176 (pTet/pTet23ΔKm) mutant 5 and 6, and wild type 81-176 (pTet) respectively. There is no band produced with wild type template DNA as there is no kanamycin resistance cassette. Lanes 5, 6 and 7 contain amplifications using primers Km.verif-5-pTir4 in strain 81-176 (pTet/pTet23ΔKm) mutant 5 and 6, and wild type 81-176 (pTet) respectively. There are no bands for wild type or mutant as in the wild type there is no kanamycin resistance cassette and in the mutants the invertible region is fixed in one orientation. Lanes 8, 9 and 10 contain amplifications using primers ampli.Km-3-ampli.Km-P-5 in strain 81-176 (pTet/pTet23ΔKm) and wild type 81-176 (pTet) respectively. Bands are only produced for the mutants as the wild type has no kanamycin resistance cassette. Lanes 11, 12 and 13 contain amplifications using primers ampli.Km-3-ampli.Km-5 in strain 81-176 (pTet/pTet23ΔKm) and wild type 81-176 (pTet) respectively. No bands are produced for the mutants as the inserted kanamycin resistance cassette has no promoter.

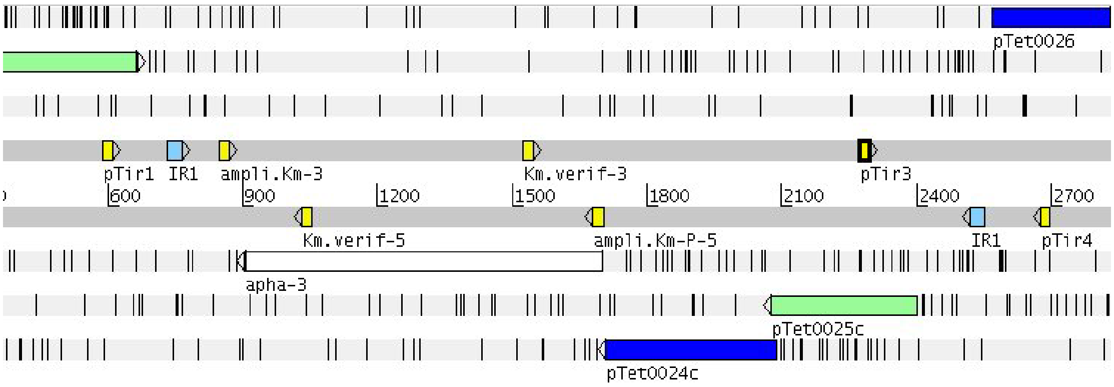


Fig 7.14: Representation of the invertible region of *C. jejuni* strain 81-176 (pTet/pTet23ΔKm). The region is viewed using Artemis [102]. The forward and reverse DNA sequences are represented by dark grey lines. The three-frame forward and reverse DNA translations are represented by light grey lines. Features are represented by open boxes: oligonucleotide primer sequences (yellow) and inverted repeats (blue) are marked on the DNA lines; CDSs are marked on the translated frame lines. CDSs are coloured according to functional category: light green, unknown; blue, pathogenicity/ adaptation/ chaperones. The kanamycin resistance gene *apha-3* is coloured white.

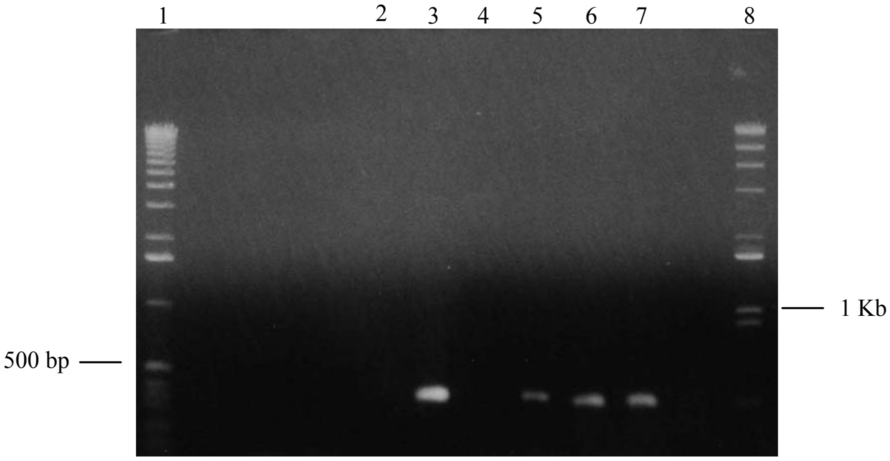


Fig 7.15: PCR amplification between various primers in mutant and wild type 81-176 strains. Lane 1 contains 1 Kb DNA marker and lane 8 contains a mixture of λ *Hin*dIII and pBR322 *Bst*NI digested DNA. Lanes 2 and 3 contain amplifications using primers pTir1-pTir2 in strain 81-176 (pTet/pTet23ΔKm) and wild type 81-176 (pTet) respectively. Lanes 4 and 5 contain amplifications using primers pTir1-pTir3 in strain 81-176 (pTet/pTet23ΔKm) and wild type 81-176 (pTet) respectively. Lanes 6 and 7 contain amplifications using primers pTir3-pTir4 in strain 81-176 (pTet/pTet23ΔKm) and wild type 81-176 (pTet) respectively. This shows that the invertible region no longer inverts in the mutant.

In order to show that the kanamycin resistance gene had inserted in the right location in pTet and that there was only a single copy, purified total DNA from wild type 81-176 (pTet) and 81-176 (pTet/pTet23ΔKm) mutants 5 and 6 were analysed by Southern blotting. DNA was digested with enzymes *Bgl*II and *Hin*dIII. *Bgl*II has been used previously for analysis of pTet [55] and the restriction sites are located outside the invertible region (**Fig 7.16 A**). *Hin*dIII sites are present both within and outside the invertible region resulting in the DNA invertase being located on different size fragments in the wild type pTet depending on which orientation the invertible region is in (**Fig 7.16 A** and **B**). For the pTet/pTet23ΔKm mutant an extra *Hin*dIII site had been introduced downstream of the kanamycin resistance gene (*apha-3*) which means there are two *Hin*dIII sites within the invertible region in the mutant (**Fig 7.16 C**). The undigested and digested DNA was separated by agarose gel electrophoresis in duplicate (**Fig 7.17**). The DNA was then transferred to a Nytran + membrane by capillary transfer (section 7.2.3.5). The membrane was cut in two and each section was hybridized in two parallel reactions to radiolabelled probes generated by PCR amplification from the kanamycin resistance gene, using primers ampli.Km-P-5 and ampli.Km-3, and the putative DNA invertase gene, using primers pT23L and pT23R (**Fig 7.14** and **Fig 7.18**).

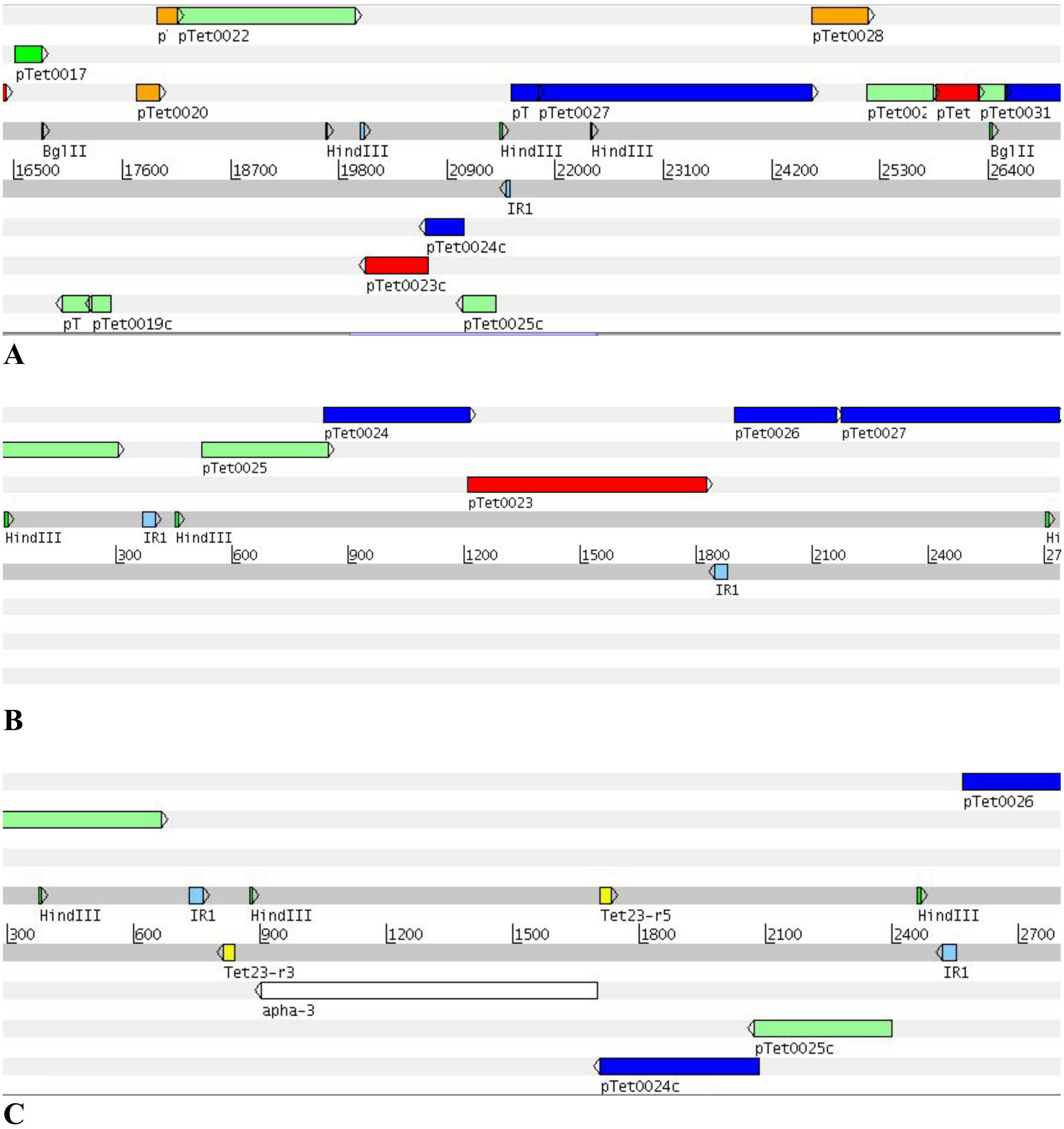


Fig 7.16: Representation of restriction sites within pTet. **A** is 81-176 (pTet) in the 'sequenced' orientation, **B** is 81-176 (pTet) in the 'inverted' orientation, **C** is 81-176 (pTet/pTet23ΔKm). The regions are viewed using Artemis [102]. The forward and reverse DNA sequences are represented by dark grey lines. The three-frame forward and reverse DNA translations are represented by light grey lines. Features are represented by open boxes: oligonucleotide primer sequences (yellow), inverted repeats (blue) and restriction enzyme sites (green), are marked on the DNA lines; CDSs are marked on the translated frame lines. CDSs are coloured according to functional category: light green, unknown; red, information transfer; blue, pathogenicity/ adaptation/ chaperones; orange, conserved hypothetical. The kanamycin resistance gene, *apha-3*, is coloured white. The *Bgl*II restriction sites are located outside the region of inversion, **A**. The *Hin*dIII restriction sites are located such that different size fragments will be produced depending on the orientation of the invertible region, **B**. For the mutant an extra *Hin*dIII restriction site has been introduced downstream of the *apha-3* gene, **C**.

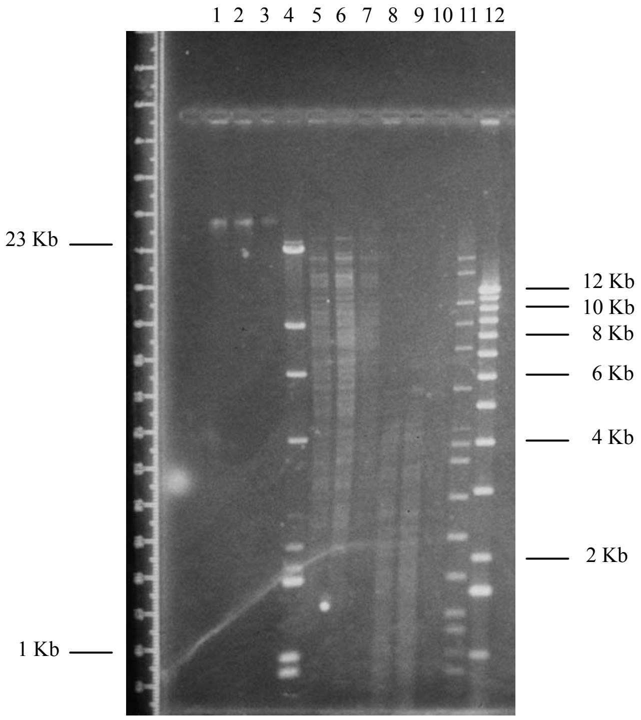


Fig 7.17: Mutant and wild type restriction enzyme digests used for southern transfer. Lanes 1, 2 and 3 contain undigested DNA from 81-176 (pTet/pTet23ΔKm) mutants 5 and 6, and wild type 81-176 (pTet); lane 4 contains marker λ *Hin*dIII/pBR322 *Bst*NI; lanes 5, 6 and 7 contain *Bgl*II digested DNA from 81-176 (pTet/pTet23ΔKm) mutants 5 and 6, and wild type 81-176 (pTet); lanes 8, 9 and 10 contain *Hin*dIII digested DNA from 81-176 (pTet/pTet23ΔKm) mutants 5 and 6, and wild type 81-176 (pTet); lane 11 contains Raoul marker and lane 12 contains 1 Kb marker.

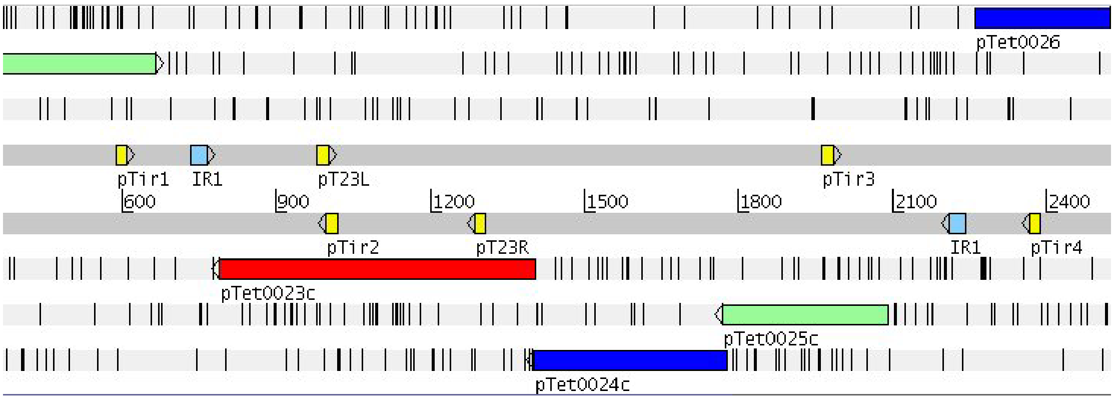


Fig 7.18: Sequence of the wild type *C. jejuni* strain 81-176 pTet with the invertible region in the sequenced orientation viewed using Artemis. The forward and reverse DNA sequences are represented by dark grey lines. The three-frame forward and reverse DNA translations are represented by light grey lines with stop codons indicated by vertical black lines. Features are represented by open boxes: oligonucleotide primer sequences (yellow) and inverted repeats (blue) are marked on the DNA lines; CDSs are marked on the translated frame lines. CDSs are coloured according to functional category: light green, unknown; red, information transfer; blue, pathogenicity/ adaptation/ chaperones. The primers pT23L and pT23R were used to amplify a region from the invertase gene for use as a template to construct a radiolabelled probe for the Southern blot.

Only DNA from 81-176 (pTet/pTet23ΔKm) mutants 5 and 6 hybridized to the kanamycin probe and only the wild type 81-176 (pTet) DNA hybridized to the DNA invertase probe (**Fig 7.19**). The *Bgl*II digested DNA from the wild type has a band size of 9644 bp while the *Bgl*II digested mutant DNA has a band size of 9953 bp. The *Hin*dIII digested wild type DNA has band sizes of 2255 bp for the invertible region in the inverted orientation and 1780 bp for the invertible region in the sequenced orientation. It is apparent that the hybridization signal from the band in the sequenced orientation is much stronger than that for the inverted orientation. The *Hin*dIII digested mutant DNA has a band size of 1588 bp. The observed and expected band sizes correlate.

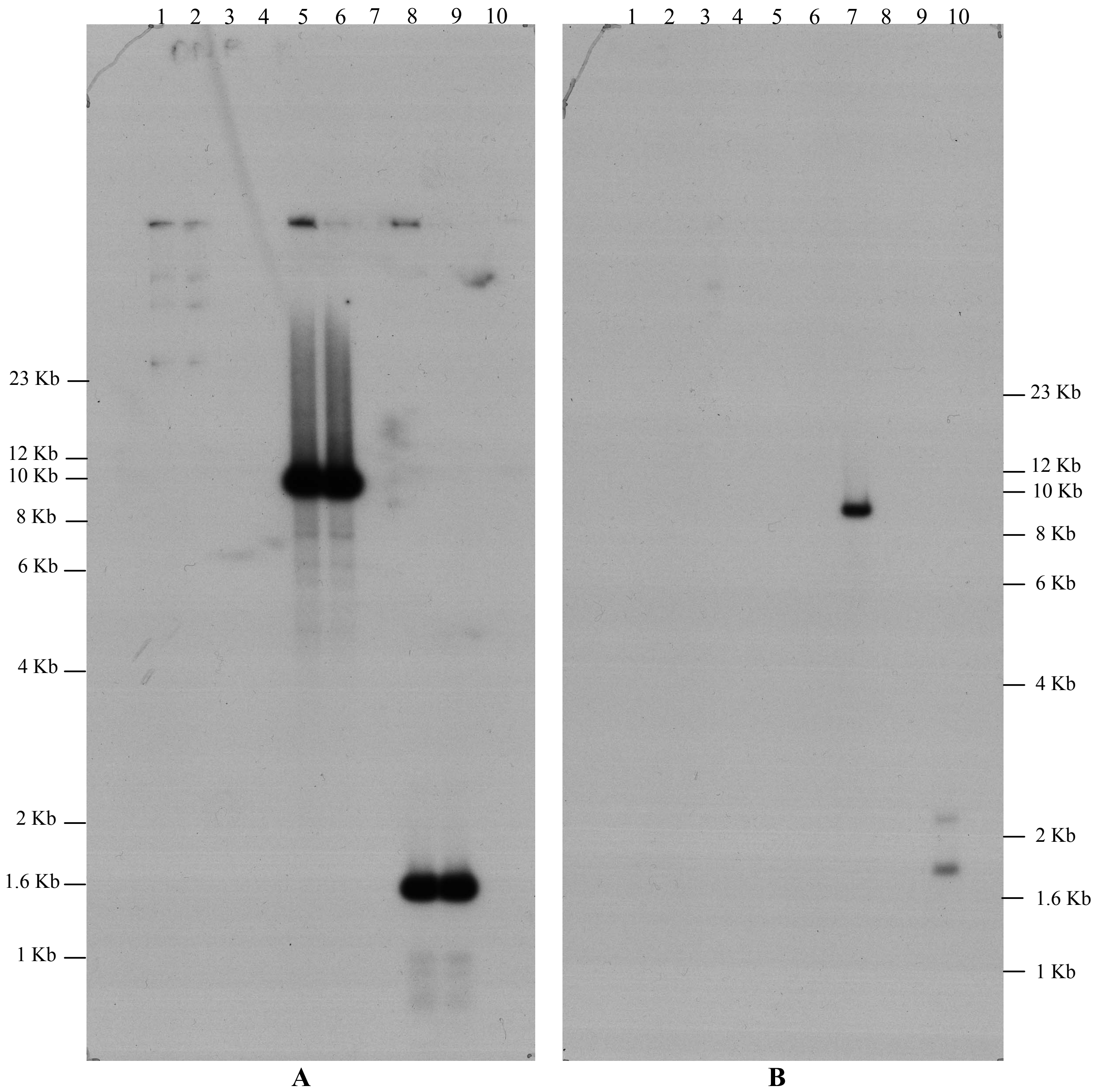


Fig 7.19: Southern blot showing pTet23 knockout mutant. Probing was performed in duplicate. **A** is using kanamycin gene probe. **B** is using invertase gene probe. Lanes 1, 2 and 3 contain undigested DNA from 81-176 (pTet/pTet23ΔKm) mutants 5 and 6, and wild type 81-176 (pTet); lane 4 contains marker λ *Hin*dIII/pBR322 *Bst*NI; lanes 5, 6 and 7 contain *Bgl*II digested DNA from 81-176 (pTet/pTet23ΔKm) mutants 5 and 6, and wild type 81-176 (pTet); lanes 8, 9 and 10 contain *Hin*dIII digested DNA from 81-176 (pTet/pTet23ΔKm) mutants 5 and 6, and wild type 81-176 (pTet). From these blots it can be seen that in the mutant the invertase has been replaced with a kanamycin resistance cassette. Also from the two bands seen in lane 10, where the wild type DNA was digested with *Hin*dIII and hybridized with the invertase gene probe, it appears that the inverted orientation is less common (weaker, upper band).

7.3.4 Conjugation experiments

Strain NCTC 11168 *cj0742::cat* was used as the recipient strain for conjugation experiments, as a strain with an alternative chromosomal antibiotic resistance marker to tetracycline and kanamycin was needed so that conjugation events could be measured by selecting for recipient cells that had received pTet from donor cells (section 7.2.3.5). Both 81-176 (pTet) and strain (pTet/pTet23ΔKm) were used as donor strains. Conjugation reactions were performed in the presence of DNase I to prevent uptake of the plasmid by natural transformation as opposed to transfer by conjugation. Transconjugants were grown on plates containing tetracycline and chloramphenicol to select for NCTC 11168 cells that had received pTet or pTet23ΔKm. In addition single strains were plated on double selective plates (tetracycline and chloramphenicol) to check for spontaneous resistance mutants. The frequency of transconjugation was expressed as the number of NCTC 11168 cells that had received pTet or pTet23ΔKm per donor cell in the initial conjugation mix. All serial dilutions were plated in triplicate and conjugations between donor and recipient were performed in duplicate. The results from conjugation experiments show that there is no significant difference between conjugation frequencies of wild type 81-176 (pTet) and knockout mutant 81-176 (pTet/pTet23ΔKm) mutants 5 and 6 (**Fig 7.20**). This indicates that the putative promoter may be in the "on" orientation and therefore drive transcription of the down stream type IV secretion system homologues when the invertible region is in the 'sequenced' orientation.

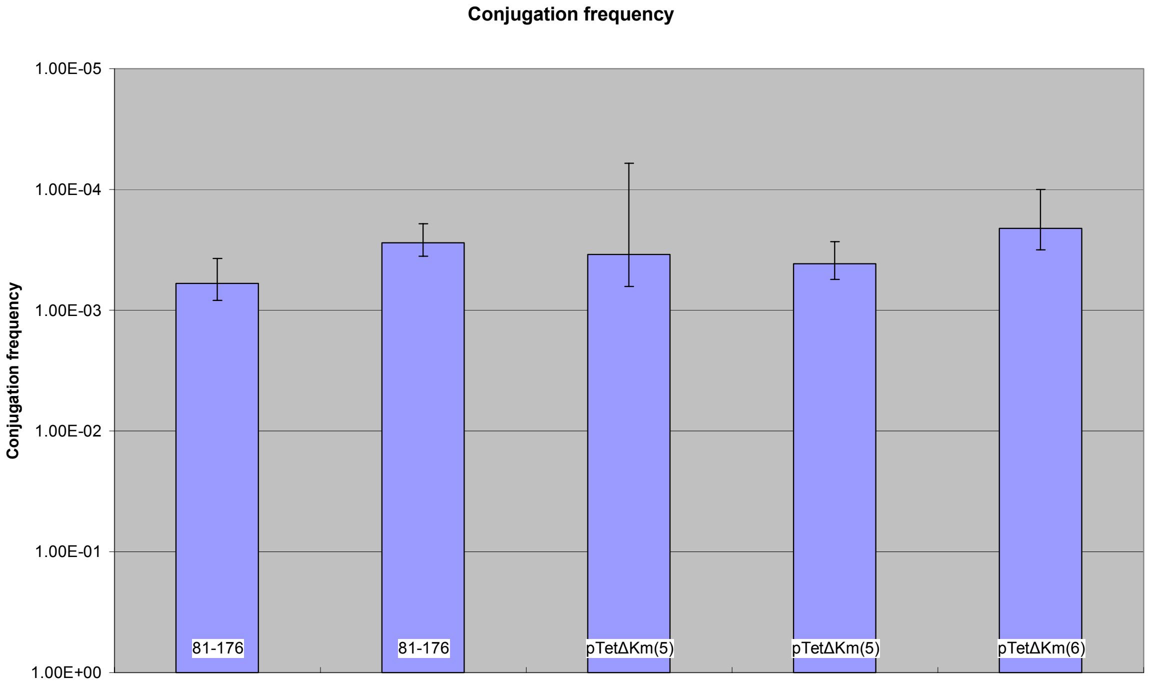


Fig 7.20: Transfer of pTet and pTet23ΔKm to strain NCTC 11168 (*cj0742::cat*). Conjugation frequencies are expressed as the number of transconjugants obtained per donor cell in the initial conjugation culture. Error bars indicate one standard error either side of the mean value. Two biological replicates are shown for wild type and pTet23Km mutant 5, only one for pTet23Km mutant 6. This indicates that there is no significant difference between wild type and mutant conjugation frequencies.

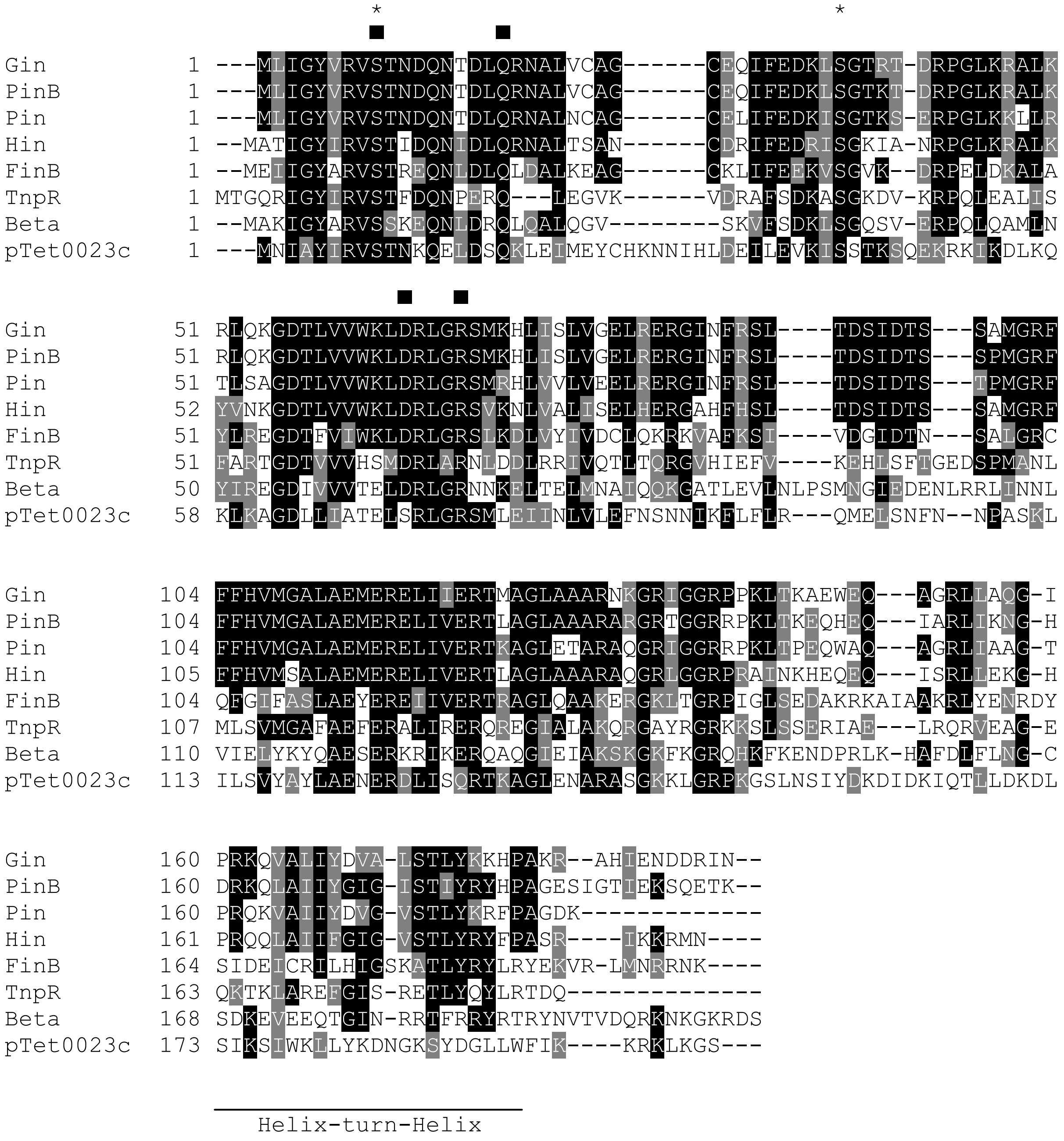
7.4 Discussion

The aims of this section of the project were to assess whether the region in pTet containing a putative DNA invertase, flanked by 31 bp inverted repeats, does invert in the plasmid pTet, whether the DNA invertase is responsible for the inversion and what the implications of this are for the plasmid and host.

Site-specific DNA recombinases can be involved in many biological functions; the family includes DNA invertases, resolvases and integrases. In a multicopy plasmid of *Clostridium perfringens* a resolvase gene has been proposed to act in resolving plasmid multimers into monomers. This is thought to stabilize the plasmid and allow more efficient partitioning [216]. The resolvase of the transposon γδ resolves cointegrate intermediates formed during intermolecular transposition of the parent transposons [217]. Inversion of DNA segments has also been linked to alternation of expression between sets of genes when promoters are located on the ends of invertible regions. In *Salmonella enterica* Typhimurium and other closely related Salmonella spp. a recombinase, Hin, is responsible for inverting a region of DNA that places a promoter upstream of one of two copies of a flagellar gene [218]. This inversion requires a host protein, factor II, and a histone-like protein, and the rate of inversion is increased by an additional enhancer element [219]. This enhancer is located within the N-terminal coding region of the recombinase and has been shown to be cis-acting. The cis-acting sequence is also present in Gin, a recombinase from bacteriophage Mu [219] which controls host range alternation [217]. In bacteriophage Mu two sets of proteins involved in tail fibre production are alternately expressed, thereby altering host range. The rate of inversion is low, presumably to keep one phenotype through one infectious cycle [220]. Expression linked to the G(+) orientation of the invertible region is required for infection of *Escherichia coli* K12 whereas expression linked to the G(-) orientation is required for infection of *Shigella sonnei* and *Escherichia coli* C [221].

Inversion of a region within *Campylobacter fetus* has been proposed to vary surface layer proteins. The *sapA* promoter is located on an inverting stretch of DNA which upon inversion is positioned upstream of two different oppositely orientated gene cassettes [123]. Some systems are more complex with different gene cassettes being inverted upstream of static promoters, such as the Min-like system in R64-related Inc plasmids [123]. The IncIα plasmid R64 contains seven 19bp repeats which upon inversion orientate different C-terminal regions in frame downstream of a fixed N-terminal region to create seven different genes [222].

Site specific recombinases have conserved regions (**Fig 7.21**); the carboxy terminal region is involved in DNA recognition whereas the amino terminal region is responsible for mediating inversion [217]. The amino terminus is highly conserved but the carboxy terminus is more divergent as might be expected with the different target specificities displayed by the recombinases [216]. The mechanism of action of serine recombinases is not clearly understood [223]. The recombinases interact with short DNA repeats, bringing them together, and then all four strands are cleaved and re-ligated. In serine recombinases a staggered break is generated at a 2 bp sequence and transient linkages are formed between the phosphate groups at the recessed 5`-ends of the newly broken DNA and serine residues within the catalytic domain of the recombinase. The strands are exchanged and then re-ligated *via* an unknown mechanism [223].



```
                           *                                *
                           ■        ■
Gin            1 ---MLIGYVRVSTNDQNTDLQRNALVCAG------CEQIFEDKLSGTRT-DRPGLKRALK
PinB           1 ---MLIGYVRVSTNDQNTDLQRNALVCAG------CEQIFEDKLSGTKT-DRPGLKRALK
Pin            1 ---MLIGYVRVSTNDQNTDLQRNALNCAG------CELIFEDKISGTKS-ERPGLKKLLR
Hin            1 --MATIGYIRVSTIDQNIDLQRNALTSAN------CDRIFEDRISGKIA-NRPGLKRALK
FinB           1 --MEIIGYARVSTREQNLDLQLDALKEAG------CKLIFEEKVSGVK--DRPELDKALA
TnpR           1 MTGQRIGYIRVSTFDQNPERQ---LEGVK------VDRAFSDKASGKDV-KRPQLEALIS
Beta           1 --MAKIGYARVSSKEQNLDRQLQALQGV--------SKVFSDKLSGQSV-ERPQLQAMLN
pTet0023c      1 ---MNIAYIRVSTNKQELDSQKLEIMEYCHKNNIHLDEILEVKISSTKSQEKRKIKDLKQ

                              ■   ■
Gin           51 RLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSL----TDSIDTS---SAMGRF
PinB          51 RLQKGDTLVVWKLDRLGRSMKHLISLVGELRERGINFRSL----TDSIDTS---SPMGRF
Pin           51 TLSAGDTLVVWKLDRLGRSMRHLVVLVEELRERGINFRSL----TDSIDTS---TPMGRF
Hin           52 YVNKGDTLVVWKLDRLGRSVKNLVALISELHERGAHFHSL----TDSIDTS---SAMGRF
FinB          51 YLREGDTFVIWKLDRLGRSLKDLVYIVDCLQKRKVAFKSI----VDGIDTN---SALGRC
TnpR          51 FARTGDTVVVHSMDRLARNLDDLRRIVQTLTQRGVHIEFV----KEHLSFTGEDSPMANL
Beta          50 YIREGDIVVVTELDRLGRNNKELTELMNAIQQKGATLEVLNLPSMNGIEDENLRRLINNL
pTet0023c     58 KLKAGDLLIATELSRLGRSMLEIINLVLEFNSNNIKFLFLR---QMELSNFN--NPASKL

Gin          104 FFHVMGALAEMERELIIERTMAGLAAARNKGRIGGRPPKLTKAEWEQ---AGRLLAQG-I
PinB         104 FFHVMGALAEMERELIVERTLAGLAAARARGRTGGRRPKLTKEQHEQ---IARLIKNG-H
Pin          104 FFHVMGALAEMERELIVERTKAGLETARAQGRIGGRRPKLTPEQWAQ---AGRLIAAG-T
Hin          105 FFHVMSALAEMERELIVERTLAGLAAARAQGRLGGRPRAINKHEQEQ---ISRLLEKG-H
FinB         104 QFGIFASLAEYEREIIVERTRAGLQAAKERGKLTGRPIGLSEDAKRKAIAAKRLYENRDY
TnpR         107 MLSVMGAFAEFERALIRERQREGIALAKQRGAYRGRKKSLSSERIAE---LRQRVEAG-E
Beta         110 VIELYKYQAESERKRIKERQAQGIEIAKSKGKFKGRQHKFKENDPRLK-HAFDLFLNG-C
pTet0023c    113 ILSVYAYLAENERDLISQRTKAGLENARASGKKLGRPKGSLNSIYDKDIDKIQTLLDKDL

Gin          160 PRKQVALIYDVA-LSTLYKKHPAKR--AHIENDDRIN--
PinB         160 DRKQLAIIYGIG-ISTIYRYHPAGESIGTIEKSQETK--
Pin          160 PRQKVAIIYDVG-VSTLYKRFPAGDK-------------
Hin          161 PRQQLAIIFGIG-VSTLYRYFPASR----IKKRMN----
FinB         164 SIDEICRILHIGSKATLYRYLRYEKVR-LMNRRNK----
TnpR         163 QKTKLAREFGIS-RETLYQYLRTDQ--------------
Beta         168 SDKEVEEQTGIN-RRTFRRYRTRYNVTVDQRKNKGKRDS
pTet0023c    173 SIKSIWKLLYKDNGKSYDGLLWFIK----KRKLKGS---

                 ______________________
                    Helix-turn-Helix
```

Fig 7.21: Site specific DNA recombinase alignments. The protein sequences were aligned using clustal X. * = serine residues likely to be covalently linked to DNA during recombination. ■ = functionally important as identified by missense mutation analysis of γδ resolvase as identified by Newman and Grindley, 1984 [217]. The amino acid sequence of pTet0023c is compared with the sequences of site-specific recombinases from: Bacteriophage Mu, Gin (accession JWBPU); *Shigella sonnei*, PinB (accession BAA00556); *Escherichia coli*, Pin (accession AAA24391); *Salmonella enterica* Typhimurium LT2, Hin (accession NP_461699); *Bacteroides fragilis*, FinB (accession YP_209695); *Escherichia coli* Tn21, TnpR (accession RPEC21); *Streptococcus pyogenes*, Beta (accession AAR27194).

In this study PCR and sequencing was used to confirm the fact that the pTet0023-pTet0025 region flanked by inverted repeats does invert in *C. jejuni* and in *Escherichia coli*. In addition it was proved that the DNA invertase gene is responsible for this inversion. When the DNA invertase was knocked out inversion could not be demonstrated. The frequency of DNA inversion in the *C. fetus* study was found to be 1.3-5.7 x10^{-4} per generation [123]. The Southern blot showed that there was a much weaker signal from the band corresponding to the invertible region in the inverted orientation compared to in the sequenced orientation which suggests that inversion of this region does not occur very often. This is backed up by the observation that the opposite orientation was not seen in the shotgun sequence assembly of pTet (chapter 3). It may be that when the invertible region is in the inverted orientation that the inverted repeats represent a better substrate for the DNA invertase than the inverted repeats when the invertible region is in the sequenced orientation. As the inverted repeats are imperfect it is possible that the sequence surrounding the repeats in the inverted orientation alters the secondary structure of the DNA making the repeat more accessible to the invertase than when the invertible region is in the sequenced orientation. If this is the case then the forward and reverse inversion reactions may proceed at different rates which would explain why the invertible region has been predominantly found in the sequenced orientation in this study.

There is a predicted promoter located between pTet0025 and pTet0026 which would be upstream of the type IV secretion gene homologues when the invertible region is in the sequenced orientation. It would appear that, as there was no difference in conjugation frequencies when the inverted region was fixed in this orientation, this represents the promoter “on” position and that in the inverted orientation the promoter could be in an “off” position. If the conjugation genes are under the control of a variable promoter then it may be beneficial to the bacterium, and therefore the plasmid, under some circumstances to not have a type IV pilus. This could be because the pilus may present a target for the immune system during infection of a host. A type IV pilus is needed for conjugation and transfer of the plasmid so it is beneficial for plasmid propagation. By controlling its conjugation system with a phase variable promoter, the plasmid could maximise its opportunity for transfer, whilst minimising its risk to the bacterial host. To prove that this is a phase-variable promoter, the conjugation experiments would need to be performed with the invertible region fixed in the opposite orientation in order to see if this has the predicted effect on conjugation frequencies. Unfortunately attempts to isolate a clone fixed in the inverted orientation were unsuccessful; this may be due to the unequal inversion frequencies discussed above.

In addition experiments could be done to determine the exact location of a promoter in the region upstream of the type IV secretion system gene homologues using RNase partial digestion and S1 nuclease digestion assays [224;225]. The promoter could also be used to drive expression of a promoterless reporter gene to prove it is functional and to determine the rate of inversion of the DNA segment [126]. Another possibility would be to extract RNA from wild type 81-176 and mutants with the invertible region fixed in both orientations and use reverse transcriptase PCR (RT-PCR) to see whether the mutant in the opposite orientation has decreased transcription of conjugation genes and to see how far into the type IV secretion system this effect extends.

8. Final discussion

Campylobacter is known to cause disease in humans. Symptoms of campylobacteriosis range from mild to severe with neurological sequelae. Different strains of *C. jejuni* have been shown to be genetically diverse and show a range of phenotypes that may relate to clinical outcome. In order to identify genetic differences between strains a similar differential hybridisation approach, using macroarrays, to that developed by Liang *et al.* [92] was used. This method was chosen above other comparative genomic methods that could be used, for example microarrays, which are relatively difficult to set up or subtractive hybridization, which requires a great deal of sequencing to achieve adequate coverage of novel regions [90] and in addition requires a subcloning step after PCR amplification and subtraction which can introduce biases. The method of subtractive hybridization has also been associated with a high level of false positive sequence identification [91].

The use of macroarrays has been shown to be sensitive for use with expression profiling so it should be a sensitive enough technique to identify clones carrying inserts that differ between strains, provided arraying and hybridization conditions have been optimised [226]. One limitation of the method is that some regions may not be picked out by hybridisation due to labelling difficulties or secondary structure of the DNA. There may be regions of the chromosome not represented in the clone libraries due to clone distribution biases [227] but generating a clone library with 5x coverage of the genome should ensure that as much of the genome as possible is represented. In addition some clones may be lost if they did not grow well or were lost during handling and membrane preparation but the effects of this should again be limited by using 5x coverage of the genome.

The analysis of the efficacy of the method from the differential hybridization of pUC libraries for strains 81-176 and M1 suggested that approximately 20% of novel sequence was missed by the pUC library sequencing. When the pUC contigs were compared to the novel sequences from BAC clones 22% extra sequence was identified in strain 81-176; 23% extra sequence was identified in strain M1; 38% extra sequence was identified in strain 40671; and 35% extra sequence was identified in 52472 discounting bacteriophage associated sequence. Bacteriophage sequence was discounted from the coverage figures as the pUC sequences associated with bacteriophage are all highly similar and therefore true coverage cannot be determined. The extra sequence gained from the BAC clones also shows that the combination of two libraries constructed using different methods (i.e. pUC with sonication and BAC with enzyme digestion) is a powerful tool for identifying novel regions between different strains and may go some way to compensating for biases that occur using either method individually. However it is clear that this approach will never be as comprehensive as a full genome sequence and that there are likely to be regions that have not been identified using this method.

Other methods, such as microarrays, appear to show bias as well [170]. As strain 81-176 has been extensively studied many strain specific genes have already been characterized, some of which have been identified in this study. Using strain 81-176 as a benchmark it is clear that this method is not identifying all the novel regions when compared to strain 11168. This is likely to be due to both bias in the library construction and problems with the hybridization method. The pUC libraries identified 37 contigs out of 58 identified in a recent 81-176 microarray study [170]. A further 3 contigs were discovered in the BAC sequences, leaving 18 out of 58 which have not been identified in this study (31%) of the total number identified by Poly *et al*. [170]. However the microarray method failed to identify many regions that have been expanded in this study: 59% of pUC sequence and 40% of novel BAC sequence was not identified but this may in part be due to different hybridization stringencies and a 2.8x coverage of the genome sequence in the microarray study [170].

Not only do additional genes in the form of plasmids, bacteriophage and pathogenicity islands contribute to virulence but the loss of genes may also have a marked effect. For example *Shigella* spp. and enteroinvasive *Escherichia coli* have a large deletion around the lysine decarboxylase gene, *cadA*, when compared to the *Escherichia coli* laboratory strain K-12. When this gene was introduced to *Shigella,* attenuated virulence was seen and enterotoxin activity was also inhibited [228]. The method of differential hybridization, comparing pUC libraries of test strains to genomic DNA from strain 11168 can not detect regions that are absent from the test strains compared to 11168. Differential hybridization data can be interpreted in conjunction with microarray data comparing these strains against 11168 to give an overall picture of the gene content of these strains. In order to truly evaluate different methods it would be necessary to compare two sequenced strains as has been done for subtractive hybridization with two strains of *Helicobacter pylori* [90].

In this study a number of strains were compared using differential hybridization to represent a range of characteristics present within the *Campylobacter* species. Strain 81-176 represents a highly studied laboratory strain originally isolated from an outbreak thought to be associated with raw milk. Strain M1 represents a strain with a clear transmission link between poultry and humans as disease was developed after a visit to a poultry abattoir. Strain 40671 represents a strain from an outbreak thought to be associated with water. Strain 52472 was isolated from a patient with septicaemia. These strains were all compared to the sequenced strain 11168 using a differential hybridization method. This method led to the identification of 93 CDSs in strain 81-176 some of which were expanded using BAC libraries to give 8 novel regions from BAC clones. In strain M1 137 CDSs were identified some of which were expanded into 10 novel regions; 97 CDSs in 40671 were expanded into 6 novel regions; and 268 CDSs in strain 52472 were expanded into 7 regions.

The strain 81-176 has two plasmids; pVir and pTet. In order to discriminate between plasmid encoded and chromosomally encoded novel CDSs in strain 81-176 these plasmids were sequenced. Plasmids pVir and pTet were originally isolated from strain 81-176 but subsequently very similar plasmids have been isolated from other strains: 17% of clinical isolates from Canada contained pVir [130] and 50% of clinical isolates from another study were resistant to tetracycline, 67% of which harboured tetracycline resistance plasmids [207]. In addition the tetracycline resistance plasmid appears to be highly conserved between species with plasmids pCC31 from *C. coli* and pTet from *C. jejuni* showing 94.3% aa id to each other [128]. Interestingly predicted CDSs with homology to those from pTet have been discovered from the pUC screen in the strains 40671 and 52472. It is unclear whether these regions carrying the predicted CDSs may be integrated into the chromosome or present on plasmids, however these strains appear to contain a full complement of homologues necessary to form a type IV secretion system. In strain 40671 homologues of VirB5 and VirD2 were not identified but these are also missing from some type IV secretion systems involved in protein secretion [72;190].

An interesting feature apparent from the sequencing of pTet was the identification of a putative DNA invertase, flanked by 31bp inverted repeats, upstream of homologues of a type IV secretion system. This region is thought to hold a promoter which is moved upon inversion of the DNA segment to be located upstream of a putative type IV secretion system. The region flanked by inverted repeats has been shown to invert and the DNA invertase was shown to be responsible for this inversion. Further studies would be needed to confirm whether the type IV secretion system is under variable control and what effect this has on conjugation.

Although plasmids represent one method of horizontal transfer of genes into a strain and have been implicated in carrying virulence factors in other bacteria; such as YOPs of

Yersinia enterocolitica [78] and IPA of *Shigella flexneri* [77] as well as in *Campylobacter* [127], chromosomal determinants may also be important. In many other bacteria, pathogenicity islands, phage, IS elements and transposons have been implicated in the dissemination of virulence determinants [71]. Areas adjacent to tRNA genes have been shown to be a common site of insertion for such mobile elements, as is the case for pathogenicity islands of uropathogenic *Escherichia coli* and also for the integration of bacteriophage [71]. However no insertion sequence (IS) elements were found within the genome sequences of *C. jejuni* strains 11168 or RM1221, and only the genome of RM1221 contained genomic islands in the form of integrated phage and plasmid DNA [8;9].

A total of 595 partial and complete predicted CDSs were identified using the differential hybridization approach. There were some common themes between all the test strains with surface structure associated CDSs, transporters, restriction modification CDSs and hypothetical CDSs being identified which have all been shown to vary between strains in other studies [83-85;91]. In addition each of the strains used in this study had a unique profile of predicted CDSs.

Strain 81-176 contained the fewest chromosomal differences to 11168 of all the test strains as identified by the pUC library sequencing. The BAC libraries were used to locate and expand the sequence from a putative cytochrome C biogenesis operon and a putative dimethyl sulfoxide reductase operon which may aid survival under reduced oxygen tensions such as those found in the human and animal gut. Strain 81-176 also contains a novel putative serine protease and di-tripeptide transporters which may aid survival by providing nutrients. This strain also contained three putative TPS systems but further work will be needed to identify whether these are functional, as several other strains contain degenerate forms of these.

Strain M1 also contained a novel putative cytochrome C biogenesis operon and a *dmsABC* operon as identified in strain 81-176. The three putative TPS systems identified in this strain contain different putative secreted proteins to those identified in strain 81-176 and those of strain M1 show homology to adhesins. It is possible that these putative adhesins aid colonization of the chicken. Strain M1 also contains a novel putative autotransporter, and although a function for this cannot be extrapolated based on sequence homology, many autotransporters have been associated with virulence functions from toxins to adhesins [192]. Intriguingly a chromosomal *tetO* insert was discovered in this strain leading to the possibility that it may be associated with a transposable element. This possibility was explored further by identifying the location of *tetO* genes in clinical tetracycline resistant isolates from Canada that were not thought to contain tetracycline resistance plasmids. Out of 8 isolates studied 2 were found to contain a similar chromosomal insertion to that of strain M1 with a conserved *tetO* and downstream hypothetical CDS but different surrounding CDSs. This insertion also contained homologues of the IS607 transposable element of *Helicobacter pylori* [211]. This poses the intriguing possibility that *tetO* may be located on a transposable element. However more work will be needed to explore whether or not this region is mobile and whether it can be transmitted between strains. Previously studied tetracycline resistance determinants from *C. jejuni* have all been carried on plasmids, although not all tetracycline resistant strains appear to contain plasmids [207].

Many predicted CDSs from the pUC screen in strain 40671 were hypothetical with no predicted function. Further investigation identified many of these as being associated with a novel capsule locus. As this strain has been associated with water it may be that this capsule leads to increased survival of 40671 in the environment compared to other strains. A number of hypothetical CDSs that may be associated with metabolism and a novel oxidoreductase were identified. Oxidoreductases are used in many metabolic pathways so further investigation would be needed to elucidate the precise function of these in strain 40671. A novel putative MCP chemotaxis CDS was identified which may also aid survival if this allows the bacterium to respond to environmental conditions. Components of a putative type IV secretion system were identified in strain 40671 but further work would be needed to examine whether these are colocalized with other homologues of pTet CDSs either on a plasmid or integrated into the chromosome.

Strain 52472 also has components of a type IV secretion system. This strain contains many regions of bacteriophage associated DNA some of which were identified as being inserted in the chromosome. There are homologues of an autotransporter and a TPS system, both represented by pseudogenes, suggesting that the intact equivalents performed no function which enhanced survival of this strain in the environmental niche it inhabited. There are many metabolism associated CDSs which showed limited identity to those from strain 11168, as identified by the pUC screen, and also an intact homologue of a PrpD-family protein as present in strain RM1221. These data suggest that strain 52472 may have different metabolic capabilities to strain 11168 although the implications of this for survival would require further study. Strain 52472 appears more similar to strain RM1221 than to strain 11168 with many of the CDSs identified as not being present in strain 11168 being found in strain RM1221. These similarities were mostly due to strains 52472 and RM1221 containing similar bacteriophage DNA. In strain 52472 a putative RM operon that also contained a putative protein kinase was identified. This arrangement is reminiscent of the phage limitation system of *Streptomyces coelicolor* [179]. It would be interesting to explore the implications of this putative operon in strain 52472.

There are many pseudogenes among the novel inserts suggesting that these are accessory genes only beneficial to the bacterium in a small subset of the environmental niches which it inhabits. Of the novel CDSs identified in this study 21% are inactivated in one strain or another. This is compared to the number of pseudogenes in the chromosomal core of 1.3% in strain 11168. It should be noted that comparisons within this study identified many previously unrecognized pseudogenes on the chromosome of strain 11168. This suggests that if DNA present in one strain is found in another, as identified by DNA hybridization, this does not indicate that genes carried on this DNA are functional.

An important feature apparent from this study is that metabolic pathways may be variable and these may play a key role in adaptation and survival in different environments. It is thought that *C. jejuni* is extremely susceptible to a wide range of environmental stresses and does not grow below 30°C [2] but may persist in the environment for several weeks [16]. It appears likely that along with traditional features such as surface polysaccharide (LOS and capsule) that other features associated with accessory metabolic pathways, respiration, uptake of different nutrients and catabolism may be important for differential survival in the environment.

This study has identified many novel regions that could be involved in pathogenicity. Further work could be done to explore this possibility. It may be possible to examine the amount of respiratory divergence between these test strains by comparing growth rates in media supplemented with different respiratory substrates, such as formate, lactate and pyruvate [180]. Growth could also be compared in the presence of the alternative terminal electron acceptors fumarate, nitrate, dimethyl sulphoxide (DMSO) and trimethylamine-*N*-oxide (TMAO), under anaerobic conditions [229]. Strains 81-176 and M1 are predicted to encode a dimethyl sulfoxide reductase operon. In order to test this possibility alternative electron acceptor activity could be measured using methyl or benzyl viologen-linked reductase assays [18]. The role of the putative cytochrome C in strains 81-176 and M1 could be explored in cell-free preparations using physiological electron donor and acceptor systems [185].

A putative autotransporter and putative TPS systems were identified. Protein secretion could be analysed using a similar approach to Konkel *et al.* who used 2D gel electrophoresis and immunoblot analyses to identify proteins expressed on incubation with epithelial cells and then used fluorescence microscopy to visualize secreted proteins upon binding to fluorescently labelled antibodies to bacterial protein [46]. Adhesion assays could also be performed. Other regions of interest that warrant further investigation are the MCP chemotaxis genes from strains M1 and 40671, the novel capsule of strain 40671 and the protein kinase associated with RM system in strain 52472.

Pathogenicity islands are often associated with incorporation of large DNA segments with a different G+C content to the surrounding chromosomal DNA suggesting recent transfer from another organism [71]. With the exception of the capsule region in 40671 (24 Kb) and the bacteriophage (>24 Kb) and plasmid regions in strain 52472 the average insert size for all strains was small (4 Kb) with a range of 865 bp to 14146 bp. Very few multi-gene inserts were found, with the highest number of novel CDSs identified in BAC 8B4F10 (7). There were even examples of domains within a gene being novel, e.g. MCP chemotaxis CDS from strain 40671 and the putative autotransporter which contains a different passenger domain in strains M1 and 52472. There was also no marked difference in G+C content of the novel regions identified from BAC sequencing from the background chromosomal G+C content. Many of the inserts contained DNA with a G+C content +/- 2% of the 30% average of *C. jejuni*. Notable exceptions to this are the *tetO* insert in strain M1 which has a G+C content of 40%; the capsule region in 40671 which has G+C content of 26%; the small insert between *ceuE* and tRNA in strain M1 with a G+C content of 19%; a predicted CDSs downstream of cj0031 in strain 81-176 with a G+C content of 24%; two hypothetical proteins in 40671 4B2B1 with a G+C content of 22% and the RM inserts between *peb3* and *lpxB* in strain 40671 and 52472 which have a G+C content of 25%. In this study relatively few inserts were found adjacent to tRNA genes, the only exceptions being the TraG-like insert in strains 81-176 and M1, and the insert between orthologues of *ceuE* and cj1356c in strain M1. Only the *tetO* insert of strain M1 seems to be associated with a transposon. The remainder of inserts are more likely to have occurred by homologous recombination with exogenously acquired DNA. This could be the case for novel regions located adjacent to rDNA regions as these will be conserved between similar species. A number of the novel inserts show high identity to CDSs from other delta-epsilon proteobacteria e.g. *Helicobacter pylori*, *Shewanella oniedensis* and *W. succinogenes.*

Even when strains have no detectable DNA differences there can still be marked differences in levels of transcription. It has been noted that the sequenced 11168 and the original clinical isolate of the same strain vary in colonization, gene expression and virulence phenotype even though no differences could be detected by multiple high resolution molecular genotyping techniques [230]. Changes in gene expression must be important to allow the bacterium to shift its metabolism and respiration to cope with changing environments. *C. jejuni* has been shown to survive in water and retail meats as well as poultry [230]. Studies have shown that passage under different conditions can also alter virulence phenotypes [16;91], and another study has identified the flagellar regulatory system as important for pathogenesis [231]. Reduced virulence was shown to be attributable to reduced expression of genes with σ^{28} or σ^{54} promoters. *flhA,* a component of the flagellar export apparatus, is important for expression of genes with σ^{28} or σ^{54} promoters and σ^{28} represses expression of σ^{54} [231].

This study has provided a comprehensive survey of differences between four strains with different characteristics when compared to strain 11168 for which the genome sequence is available. A range of novel DNA which may well be involved in virulence or environmental survival of these strains has been identified providing targets for further research. Attempts to identify the global gene pool of *C. jejuni* coupled to transcription studies may help in attempts to elucidate the pathogenicity of this organism.

References

1. **Vandamme, P.** (2000). Taxonomy of the family *Campylobacteraceae*. In *Campylobacter*, p. 3-26, 2nd edition. Nachamkin, I. and Blaser, M.J., Editors. ASM Press: Washington DC.
2. **Solomon, E.B. and Hoover, D.G.** (1999). *Campylobacter Jejuni*: A bacterial paradox. *Journal of Food Safety*. **19**: 121-136.
3. **Skirrow, M.B.** (1977). Campylobacter enteritis: a "new" disease. *British Medical Journal*. **2**(2): 9-11.
4. **Kelly, D.J.** (2001). The physiology and metabolism of *Campylobacter jejuni* and *Helicobacter pylori*. *Journal of Applied Microbiology*. **90**: 16S-24S.
5. **Thomas, C., Hill, D.J., and Mabey, M.** (1999). Morphological changes of synchronized Campylobacter jejuni populations during growth in single phase liquid culture. *Letters in Applied Microbiology*. **28**(3): 194-198.
6. **Bovill, R.A. and MacKey, B.M.** (1997). Resuscitation of 'non-culturable' cells from aged cultures of *Campylobacter jejuni*. *Microbiology*. **143**: 1575-1581.
7. **Kelly, A.F., Park, S.F., Bovill, R.A., and MacKey, B.M.** (2001). Survival of *Campylobacter jejuni* during Stationary Phase: Evidence for the Absence of a Phenotypic Stationary-Phase Response. *Applied and Environmental Microbiology*. **67**(5): 2248-2254.
8. **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.
9. **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., Shavartsbeyn, A., Schatz, M.C., Badger, J.H., Fraser, C.M., and Nelson, K.E.** (2005). Major Structural Differences and Novel Potential Virulence Mechanisms from the Genomes of Multiple *Campylobacter* Species. *PLoS Biology*. **3**(1): 72-85.

10. **van Vliet, A.H.M., Ketley, J.M., Park, S.F., and Penn, C.W.** (2002). The role of iron in *Campylobacter* gene regulation, metabolism and oxidative stress defense. *Fems Microbiology Reviews*. **26**: 173-186.

11. **Field, L.H., Headley, V.L., Payne, S.M., and Berry, L.J.** (1986). Influence of Iron on Growth, Morphology, Outer Membrane Protein Composition, and Synthesis of Siderophores in *Campylobacter jejuni. Infection and Immunity*. **54**(1): 126-132.

12. **van Vliet, A.H.M. and Ketley, J.M.** (2001). Pathogenesis of enteric *Campylobacter* infection. *Journal of Applied Microbiology*. **90**: 45S-56S.

13. **Palyada, K., Threadgill, D., and Stintzi, A.** (2004). Iron Acquisition and Regulation in *Campylobacter jejuni. Journal of Bacteriology*. **186**(14): 4714-4729.

14. **Guerry, P., Perez-Casal, J., Yao, R., McVeigh, A., and Trust, T.J.** (1997). A Genetic Locus Involved in Iron Utilization Unique to Some *Campylobacter* Strains. *Journal of Bacteriology*. **179**(12): 3997-4002.

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

16. **Leach, S.A.** (1997). Growth, survival and pathogenicity of enteric campylobacters. *Reviews in Medical Microbiology*. **8**(3): 113-124.

17. **Mohammed, K.A.S., Miles, R.J., and Halablab, M.A.** (2004). The pattern and kinetics of substrate metabolism of *Campylobacter jejuni* and *Campylobacter coli. Letters in Applied Microbiology*. **39**(3): 261-266.

18. **Sellars, M.J., Hall, S.J., and Kelly, D.J.** (2002). Growth of *Campylobacter jejuni* Supported by Respiration of Fumarate, Nitrate, Nitrite, Trimethylamine-*N*-Oxide, or Dimethyl Sulfoxide Requires Oxygen. *Journal of Bacteriology*. **184**(15): 4187-4196.

19. **Newell, D.G., Frost, J.A., Duim, B., Wagenaar, J.A., Madden, R.H., van der Plas, J., and On, S.L.W.** (2000). New Developments In The Subtyping Of *Campylobacter* Species. In *Campylobacter*, p. 27-44, 2nd edition. Nachamkin, I. and Blaser, M.J., Editors. ASM Press: Washington D. C.

20. **Tam, C.C.** (2001). *Campylobacter* reporting at its peak year of 1998: don't count your chickens yet. *Communicable Disease and Public Health*. **4**(3): 194-199.

21. Health ProtectionAgency, 2005*; Campylobacter spp. Laboratory reports of faecal isolates England & Wales, 1986-2004.* http://www.hpa.org.uk/infections/topics_az/campy/data_ew.htm. Last accessed 1st April 2005.

22. CDC, 2004*; Preliminary FoodNet Data on the Incidence of Infection with Pathogens Transmitted Commonly Through Food --- Selected Sites, United States, 2003*. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5316a2.htm. Last accessed 1st April 2005.

23. **website, F.S.A.** (2001). Salmonella in retail chicken drops to all time low, but the battle with Campylobacter continues. *http://www.foodstandards.gov.uk/wales/pressreleases/lowsalmonellainchicken*.

24. **Wilson, I.G.** (2002). Salmonella and campylobacter contamination of raw retail chickens from different producers: a six year survey. *Epidemiology and Infection*. **129**(3): 635-645.

25. **Meldrum, R.J., Tucker, D., and Edwards, C.** (2004). Baseline Rates of *Campylobacter* and *Salmonella* in Raw Chicken in Wales, United Kingdom, in 2002. *Journal of Food Protection*. **67**(6): 1226-1228.

26. **Nachamkin, I., Engberg, J., and Aarestrup, F.M.** (2000). Diagnosis and Antimicrobial susceptibility of *Campylobacter* species. In *Campylobacter*, p. 45-66, 2nd edition. Nachamkin, I. and Blaser, M.J., Editors. ASM Press: Washington D. C.

27. **Nachamkin, I., Allos, B.M., and Ho, T.W.** (2000). *Campylobacter jejuni* infection and the association with Guillain-Barré syndrome. In *Campylobacter*, p. 155-175, 2nd edition. Nachamkin, I. and Blaser, M.J., Editors. ASM Press: Washington D. C.

28. **Willison, H.J. and O'Hanlon, G.M.** (2000). Antiglycosphingolipid antibodies and Guillain-Barré syndrome. In *Campylobacter*, p. 259-285, 2nd edition. Nachamkin, I. and Blaser, M.J., Editors. ASM Press: Washington, D. C.

29. **Wassenaar, T.M. and Blaser, M.J.** (1999). Pathophysiology of *Campylobacter jejuni* infections of humans. *Microbes and Infection*. **1**: 1023-1033.

30. **Kuroki, S., Saida, T., Nukina, M., Haruta, T., Yoshioka, M., Kobayashi, Y., and Nakanishi, H.** (1993). *Campylobacter jejuni* Strains from Patients with Guillain-Barré Syndrome Belong Mostly to Penner Serogroup 19 and Contain ß-N-Acetylglucosamine Residues. *Annals of Neurology*. **33**(3): 243-247.

31. **Duim, B., Ang, C.W., van Belkum, A., Rigter, A., van Leeuwen, N.W.J., Endtz, H., and Wagenaar, J.A.** (2000). Amplified Fragment Length Polymorphism Analysis of *Campylobacter jejuni* Strains Isolated from Chickens and from Patients with Gastroenteritis or Guillain-Barré or Miller Fisher Syndrome. *Applied and Environmental Microbiology*. **66**(9): 3917-3923.

32. **Skirrow, M.B. and Blaser, M.J.** (2000). Clinical aspects of *Campylobacter* infection. In *Campylobacter*, p. 69-82, 2nd edition. Nachamkin, I. and Blaser, M.J., Editors. ASM Press: Washington, DC.

33. **Oberhelman, R.A. and Taylor, D.N.** (2000). *Campylobacter* infections in developing countries. In *Campylobacter*, p. 139-153, 2nd edition. Nachamkin, I. and Blaser, M.J., Editors. ASM Press: Washington D. C.

34. **Altekruse, S.F., Stern, N.J., Fields, P.I., and Swerdlow, D.L.** (1999). Campylobacter jejuni - An Emerging Foodborne Pathogen. *Emerging Infectious Diseases*. **5**(1): 28-35.

35. **Wren, B.W., Linton, D., Dorrell, N., and Karlyshev, A.V.** (2001). Post genome analysis of *Campylobacter jejuni. Journal of Applied Microbiology*. **90**: 36S-44S.

36. **Black, R.E., Levine, M.M., Clements, M.L., Hughes, T.P., and Blaser, M.J.** (1988). Experimental *Campylobacter jejuni* infections in Humans. *The Journal of Infectious Diseases*. **157**(3): 472-479.

37. **Szymanski, C.M., Logan, S.M., Linton, D., and Wren, B.W.** (2003). *Campylobacter* - a tale of two protein glycosylation systems. *Trends in Microbiology*. **11**(5): 233-238.

38. **Konkel, M.E., Klena, J.D., Rivera-Amill, V., Monteville, M.R., Biswas, D., Raphael, B., and Mickelson, J.** (2004). Secretion of Virulence Proteins from *Campylobacter jejuni* Is Dependent on a Functional Flagellar Export Apparatus. *Journal of Bacteriology*. **186**(11): 3296-3303.

39. **Konkel, M.E. and Cieplak JR., W.** (1992). Altered Synthetic Response of *Campylobacter jejuni* to Cocultivation with Human Epithelial Cells Is Associated with Enhanced Internalization. *Infection and Immunity*. **60**(11): 4945-4949.

40. **Konkel, M.E. and Joens, L.A.** (1989). Adhesion to and Invasion of HEp-2 Cells by *Campylobacter* spp. *Infection and Immunity*. **57**(10): 2984-2990.

41. **Pei, Z., Burucoa, C., Grignon, B., Baqar, S., Huang, X.-Z., Kopecko, D.J., Bourgeois, A.L., Fauchere, J.-L., and Blaser, M.J.** (1998). Mutation in the *peb1A* Locus of *Campylobacter jejuni* Reduces Interactions with Epithelial Cells and Intestinal Colonization of Mice. *Infection and Immunity*. **66**(3): 938-943.

42. **Monteville, M.R., Yoon, J.E., and Konkel, M.E.** (2003). Maximal adherence and invasion of INT 407 cells by *Campylobacter jejuni* requires the CadF outer-membrane protein and microfilament reorganization. *Microbiology*. **149**: 153-165.

43. **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. *Molecular Microbiology*. **39**(5): 1225-1236.

44. **Yao, R., Burr, D.H., and Guerry, P.** (1997). CheY-mediated modulation of *Campylobacter jejuni* virulence. *Molecular Microbiology*. **23**(5): 1021-1031.

45. **Wooldridge, K.G. and Ketley, J.M.** (1997). *Campylobacter*-host cell interactions. *Trends in Microbiology*. **5**(3): 96-102.

46. **Konkel, M.E., Kim, B.J., Rivera-Amill, V., and Garvis, S.G.** (1999). Bacterial secreted proteins are required for the internalization of *Campylobacter jejuni* into cultured mammalian cells. *Molecular Microbiology*. **32**(4): 691-701.

47. **Blaser, M.J., Perez, G.P., Smith, P.F., Patton, C., Tenover, F.C., Lastovica, A.J., and Wang, W.I.** (1986). Extraintestinal *Campylobacter jejuni* and *Campylobacter coli* infections: host factors and strain characteristics. *Journal of Infectious Diseases*. **153**(3): 552-559.

48. **Guerry, P., Ewing, C.P., Hickey, T.E., Prendergast, M.M., and Moran, A.P.** (2000). Sialylation of Lipooligosaccharide Cores Affects Immunogenicity and Serum Resistance of *Campylobacter jejuni. Infection and Immunity*. **68**(12): 6656-6662.

49. **Wassenaar, T.M.** (1997). Toxin Production by *Campylobacter* spp. *Clinical Microbiology Reviews*. **10**(3): 466-476.

50. **Pickett, C.L.** (2000). *Campylobacter* Toxins and Their Role in Pathogenesis. In *Campylobacter*, p. 179-190, 2nd edition. Nachamkin, I. and Blaser, M.J., Editors. ASM Press: Washington, D.C.

51. **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 Research*. **11**(10): 1706-1715.

52. **Yuki, N., Susuki, K., Koga, M., Nishimoto, Y., Odaka, M., Hirata, K., Taguchi, K., Miyatake, T., Furukawa, K., Kobata, T., and Yamada, M.** (2004). Carbohydrate mimicry between human ganglioside GM1 and *Campylobacter jejuni* lipooligosaccharide causes Guillain-Barré syndrome. *Proceedings of The National Academy of Sciences USA*. **101**(31): 11404-11409.

53. **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. *Molecular Microbiology*. **35**(3): 529-541.

54. **Day JR, W.A., Sajecki, J.L., Pitts, T.M., and Joens, L.A.** (2000). Role of Catalase in *Campylobacter jejuni* Intracellular Survival. *Infection and Immunity*. **68**(11): 6337-6345.

55. **Bacon, D.J., Alm, R.A., Burr, D.H., Hu, L., Kopecko, D.J., Ewing, C.P., Trust, T.J., and Guerry, P.** (2000). Involvement of a Plasmid in Virulence of *Campylobacter jejuni* 81-176. *Infection and Immunity*. **68**(8): 4384-4390.

56. **Schmidt-Ott, R., Pohl, S., Burghard, S., Weig, M., and Groß, U.** (2005). Identification and characterization of a major subgroup of conjugative *Campylobacter jejuni* plasmids. *Journal of Infection*. **50**: 12-21.

57. **Tenover, F.C., Williams, S., Gordon, K.P., Nolan, C., and Plorde, J.J.** (1985). Survey of Plasmids and Resistance Factors in *Campylobacter jejuni* and *Campylobacter coli. Antimicrobial Agents and Chemotherapy*. **27**(1): 37-41.

58. **Trieber, C.A. and Taylor, D.E.** (2000). Mechanisms of antibiotic resistance in *Campylobacter*. In *Campylobacter*, p. 441-454, 2nd edition. Nachamkin, I. and Blaser, M.J., Editors. American Society for Microbiology: Washington, D. C.

59. **Karlyshev, A.V., Ketley, J.M., and Wren, B.W.** (2005). The *Campylobacter jejuni* glycome. *Fems Microbiology Reviews*. **29**: 377-390.

60. **Korolik, V., Alderton, M.R., Smith, S.C., Chang, J., and Coloe, P.J.** (1998). Isolation and molecular analysis of colonising and non-colonising strains of *Campylobacter jejuni* and *Campylobacter coli* following experimental infection of young chickens. *Veterinary Microbiology*. **60**: 239-249.

61. **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. *Journal of Clinical Microbiology*. **15**(5): 761-768.

62. **Penner, J.L. and Hennessy, J.N.** (1980). Passive Hemmagglutination Technique for Serotyping *Campylobacter fetus* subsp. *jejuni* on the Basis of Soluble Heat-Stable Antigens. *Journal of Clinical Microbiology*. **12**(6): 732-737.

63. **Nielsen, E.M. and Nielsen, N.L.** (1999). Serotypes and typability of *Campylobacter jejuni* and *Campylobacter coli* isolated from poultry products. *International Journal of Food Microbiology*. **46**: 199-205.

64. **Wassenaar, T.M. and Newell, D.G.** (2000). Genotyping of *Campylobacter* spp. *Applied and Environmental Microbiology*. **66**(1): 1-9.

65. **Harrington, C.S., Thomson-Carter, F.M., and Carter, P.E.** (1997). Evidence for Recombination in the Flagellin Locus of *Campylobacter jejuni*: Implications for the Flagellin Gene Typing Scheme. *Journal of Clinical Microbiology*. **35**(9): 2386-2392.

66. **Fitzgerald, C., Stanley, K., Andrew, S., and Jones, K.** (2001). Use of Pulsed-Field Gel Electrophoresis and Flagellin Gene Typing in Identifying Clonal Groups of *Campylobacter jejuni* and *Campylobacter coli* in Farm and Clinical Environments. *Applied and Environmental Microbiology*. **67**(4): 1429-1436.

67. **Carvalho, A.C.T., Ruiz-Palacios, G.M., Ramos-Cervantes, P., Cervantes, L.-E., Jiang, X., and Pickering, L.K.** (2001). Molecular Characterization of Invasive and Noninvasive *Campylobacter jejuni* and *Campylobacter coli* Isolates. *Journal of Clinical Microbiology*. **39**(4): 1353-1359.

68. **Hernandez, J., Fayos, A., Ferrus, M.A., and Owen, R.J.** (1995). Random amplified polymorphic DNA fingerprinting of *Campylobacter jejuni* and *C. coli* isolated from human faeces, seawater and poultry products. *Research in Microbiology*. **146**: 685-696.

69. **Schouls, L.M., Reulen, S., Duim, B., Wagenaar, J.A., Willems, R.J.L., Dingle, K.E., Colles, F.M., and van Embden, J.D.A.** (2003). Comparative Genotyping of *Campylobacter jejuni* by Amplified Fragment Length Polymorphism, Multilocus Sequence Typing, and Short Repeat Sequencing: Strain Diversity, Host Range, and Recombination. *Journal of Clinical Microbiology*. **41**(1): 15-26.

70. **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.** (2001). Multilocus Sequence Typing System for *Campylobacter jejuni. Journal of Clinical Microbiology*. **39**(1): 14-23.

71. **Hacker, J., Blum-Oehler, G., Mühldorfer, I., and Tschäpe, H.** (1997). Pathogenic islands of virulent bacteria: structure, function and impact on microbial evolution. *Molecular Microbiology*. **23**(6): 1089-1097.

72. **Censini, S., Lange, C., Xiang, Z., Crabtree, J.E., Ghiara, P., Borodovsky, M., Rappuoli, R., and Covacci, A.** (1996). *cagT,* a pathogenicity island of *Helicobacter pylori,* encodes type I-specific and disease-associated virulence factors. *Proceedings of The National Academy of Sciences USA*. **93**: 14648-14653.

73. **Stein, M., Rappuoli, R., and Covacci, A.** (2001). The *cag* Pathogenicity Island. In *Helicobacter pylori: Physiology and Genetics*, p. 345-353Mobley, H.L.T., Mendz, G.L., and Hazell, S.L., Editors. ASM Press: Washington, D.C.

74. **Carniel, E.** (2001). The *Yersinia* high-pathogenicity island: an iron-uptake island. *Microbes and Infection*. **3**: 561-569.

75. **Carniel, E., Guilvout, I., and Prentice, M.** (1996). Characterization of a Large Chromosomal "High-Pathogenicity Island" in Biotype 1B *Yersinia enterocolitica. Journal of Bacteriology*. **178**(23): 6743-6751.

76. **Fetherston, J.D., Schuetze, P., and Perry, R.D.** (1992). Loss of the pigmentation phenotype in Yersinia pestis is due to the spontaneous deletion of 102Kb of chromosomal DNA which is flanked by a repetative element. *Molecular Microbiology*. **6**: 2693-2704.

77. **Venkatesan, M.M., Buysse, J.M., and Kopecko, D.J.** (1988). Characterization of invasion plasmid antigen genes (*ipaBCD*) from *Shigella flexneri. Proceedings of The National Academy of Sciences USA*. **85**: 9317-9321.

78. **Zink, D.L., Feeley, J.C., Wells, J.G., Vanderzant, C., Vickery, J.C., Roof, W.D., and O'Donovan, G.A.** (1980). Plasmid-mediated tissue invasiveness in *Yersinia enterocolitica. Nature*. **283**: 224-226.

79. **Brüssow, H., Canchaya, C., and Hardt, W.-D.** (2004). Phages and the Evolution of Bacterial Pathogens: from Genomic Rearrangements to Lysogenic Conversion. *Microbiology and Molecular Biology Reviews*. **68**(3): 560-602.

80. **Barondess, J.J. and Beckwith, J.** (1990). A bacterial virulence determinant encoded by lysogenic coliphage lambda. *Nature*. **346**: 871-874.

81. **Levinson, G. and Gutman, G.A.** (1987). Slipped-Strand Mispairing: A Major Mechanism for DNA Sequence Evolution. *Molecular Biology and Evolution*. **4**(3): 203-221.

82. **Moxon, E.R., Rainey, P.B., Nowak, M.A., and Lenski, R.E.** (1994). Adaptive evolution of highly mutable loci in pathogenic bacteria. *Current Biology*. **4**(1): 24-33.

83. **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 Letters*. **554**: 224-230.

84. **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.E.** (2004). Large-Scale Comparative Genomics Meta-Analysis of *Campylobacter jejuni* Isolates Reveals Low Level of Genome Plasticity. *Journal of Clinical Microbiology*. **42**(10): 4566-4576.

85. **Poly, F., Threadgill, D., and Stintzi, A.** (2004). Identification of *Campylobacter jejuni* ATCC 43431-Specific Genes by Whole Microbial Genome Comparisons. *Journal of Bacteriology*. **186**(14): 4781-4795.

86. **Jones, K., Shapero, M.H., Chevrette, M., and Fournier, R.E.K.** (1991). Subtractive Hybridization Cloning of a Tissue-Specific Extinguisher: TSE1 Encodes a Regulatory Subunit of Protein Kinase A. *Cell*. **66**: 861-872.

87. **Hedrick, S.M., Cohen, D.I., Nielsen, E.A., and Davis, M.M.** (1984). Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. *Nature*. **308**: 149-153.

88. **DeShazer, D., Waag, D.M., Fritz, D.L., and Woods, D.E.** (2001). Identification of a *Burkholderia mallei* polysaccharide gene cluster by subtractive hybridization and demonstration that the encoded capsule is an essentail virulence determinant. *Microbial Pathogenesis*. **30**: 253-269.

89. **Tinsley, C.R. and Nassif, X.** (1996). Analysis of the genetic differences between *Neisseria meningitidis* and *Neisseria gonorrhoeae*: Two closely related bacteria expressing two different pathogenicities. *Proceedings of The National Academy of Sciences USA*. **93**: 11109-11114.

90. **Agron, P.G., Macht, M., Radnedge, L., Skowronski, E.W., Miller, W., and Andersen, G.L.** (2002). Use of subtractive hybridization for comprehensive surveys of prokaryotic genome differences. *FEMS Microbiology Letters*. **211**: 175-182.

91. **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.

92. **Liang, X., Pham, X.-Q.T., Olson, M.V., and Lory, S.** (2001). Identification of a Genomic Island Present in the Majority of Pathogenic Isolates of *Pseudomonas aeruginosa. Journal of Bacteriology*. **183**(3): 843-853.

93. **Champion, O.L., Best, E.L., and Frost, J.A.** (2002). Comparison of Pulsed-Field Gel Electrophoresis and Amplified Fragment Length Polymorphism Techniques for Investigating Outbreaks of Enteritis Due to Campylobacters. *Journal of Clinical Microbiology*. **40**(6): 2263-2265.

94. **Sambrook, J. and Russell, D.W.** (2001). Agarose gel electrophoresis. In *Molecular Cloning: A laboratory manual*, p. 5.4-5.13, 3rd edition. Argentine, J., Editor. Cold Spring Harbour Laboratory Press: New York.

95. **Sambrook, J. and Russell, D.W.** (2001). Media. In *Molecular Cloning: A Laboratory Manual*, p. A2.1-A2.12, 3rd edition. Argentine, J., Editor. Cold Spring Harbor Laboratory Press: New York.

96. **Frengen, E., Weichenhan, D., Zhao, B., Osoegawa, K., van Geel, M., and de Jong, P.J.** (1999). A Modular, Positive Selection Bacterial Artificial Chromosome Vector with Multiple Cloning Sites. *Genomics*. **58**: 250-253.

97. **Osoegawa, K., de Jong, P.J., Frengen, E., and Ioannou, P.A.** (1999). Support Protocol 1: Preparation of BAC/PAC Vector for Cloning. In *Current Protocols in Human Genetics*, p. UNIT 5.15Dracopoli, N.C., et al., Editors. John Wiley & Sons, Inc.

98. **Sambrook, J. and Russell, D.W.** (2001). Random Priming: Radiolabeling of Purified DNA Fragments by Extension of Random Oligonucleotides. In *Molecular Cloning: A Laboratory Manual*, p. 9.4-9.8, 3rd edition. Argentine, J., Editor. Cold Spring Harbor Laboratory Press: New York.

99. **Ewing, B., Hillier, L., Wendl, M.C., and Green, P.** (1998). Base-Calling of Automated Sequencing Traces Using Phred. I. Accuracy Assessment. *Genome Research*. **8**: 175-185.

100. **Gish, W.** (1996-2002). http://blast.wustl.edu.

101. **Sonnhammer, E.L.L. and Durbin, R.** (1994). A Workbench for large-scale sequence homology analysis. *Computer Applications in the Biosciences*. **10**(3): 301-307.

102. **Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.-A., and Barrell, B.** (2000). Artemis: sequence visualization and annotation. *Bioinformatics*. **16**(10): 944-945.

103. **Pearson, W.R. and Lipman, D.J.** (1988). Improved tools for biological sequence comparison. *Proceedings of The National Academy of Sciences USA*. **85**: 2444-2448.

104. **Thompson, J.D., Higgins, D.G., and Gibson, T.J.** (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*. **22**(22): 4673-4680.

105. **Perriére, G. and Gouy, M.** (1996). WWW-query: An on-line retrieval system for biological sequence banks. *Biochimie*. **78**(5): 364-369.

106. **Needleman, S.B. and Wunsch, C.D.** (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*. **48**(3): 443-453.

107. **Smith, T.F. and Waterman, M.S.** (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*. **147**(1): 195-197.

108. **del Solar, G., Giraldo, R., Ruiz-Echievarria, M.J., Espinosa, M., and Diaz-Orejas, R.** (1998). Replication and Control of Circular Bacterial Plasmids. *Microbiology and Molecular Biology Reviews*. **62**(2): 434-464.

109. **Luo, N. and Zhang, Q.** (2001). Molecular Characterization of a Cryptic Plasmid from *Campylobacter jejuni. Plasmid*. **45**: 127-133.

110. **Suhan, M., Chen, S.-Y., Thompson, H.A., Hoover, T.A., Hill, A., and Williams, J.C.** (1994). Cloning and Characterization of an Autonomous Replication Sequence from *Coxiella burnetii. Journal of Bacteriology*. **176**(17): 5233-5243.

111. **Mackiewicz, P., Zakrzewska-Czerwinska, J., Zawilak, A., Dudek, M.R., and Cebrat, S.** (2004). Where does bacterial replication start? Rules for predicting the *oriC* region. *Nucleic Acids Research*. **32**(13): 3781-3791.

112. **Sonnhammer, E.L.L. and Durbin, R.** (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*. **167**(2): GC1-10.

113. **Yokoyama, E., Matsuzaki, Y., Doi, K., and Ogata, S.** (1998). Gene encoding a replication initiator protein and replication origin of conjugative plasmid pSA1.1 of *Streptomyces cyaneus* ATCC14921. *FEMS Microbiology Letters*. **169**: 103-109.

114. **Bignell, C. and Thomas, C.M.** (2001). The bacterial ParA-ParB partitioning proteins. *Journal of Biotechnology*. **91**: 1-34.

115. **Mrázek, J. and Karlin, S.** (1998). Strand compositional asymmetry in bacterial and large viral genomes. *Proceedings of The National Academy of Sciences USA*. **95**: 3720-3725.

116. **Fraser, C.M., Casjens, S., Huanng, W.M., Sutton, G.G., Clayton, R., Lathigra, R., White, O., Ketchum, K.A., Dodson, R.J., Hickey, E.K., Gwinn, M., Dougherty, B., Tomb, J.-F., Fleischmann, R.D., Richardson, D., Peterson, J.D., Kerlavage, A.R., Quackenbush, J., Salzberg, S.L., Hanson, M., van Vugt, R., Palmer, N., Adams, M.D., Gocayne, J., Weidman, J., Utterback, T., Watthey, L., McDonald, L., Artiach, P., Bowman, C., Garland, S., Fujii, C., Cotton, M.D., Horst, K., Roberts, K., Hatch, B., Smith, H.O., and Venter, J.C.** (1997). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature*. **390**: 580-586.

117. **Galli, D.M., Chen, J., Novak, K.F., and LeBlanc, D.J.** (2001). Nucleotide Sequence and Analysis of Conjugative Plasmid pVT745. *Journal of Bacteriology*. **183**(5): 1585-1594.

118. **Novak, K.F., Dougherty, B., and Peláez, M.** (2001). *Actinobacillus actinomycetemcomitans* harbours type IV secretion system genes on a plasmid and in the chromosome. *Microbiology*. **147**: 3027-3035.

119. **Katz, M.E., Strugnell, R.A., and Rood, J.I.** (1992). Molecular Characterization of a Genomic Region Associated with Virulence in *Dichelobacter nodosus*. *Infection and Immunity*. **60**(11): 4586-4592.

120. **Alonso, J.C., Weise, F., and Rojo, F.** (1995). The *Bacillus subtilis* Histone-like Protein Hbsu Is Required for DNA Resolution and DNA Inversion Mediated by the ß Recombinase of Plasmid pSM19035. *The Journal of Biological Chemistry*. **270**(7): 2938-2945.

121. **Patrick, S., Parkhill, J., McCoy, L.J., Lennard, N., Larkin, M.J., Collins, M., Sczaniecka, M., and Blakely, G.** (2003). Multiple inverted DNA repeats of *Bacteroides fragilis* that control polysaccharide antigenic variation are similar to the hin region inverted repeats of *Salmonella typhimurium*. *Microbiology*. **149**: 915-924.

122. **Krinos, C.M., Coyne, M.J., Weinacht, K.G., Tzianabos, A.O., Kasper, D.L., and Comstock, L.E.** (2001). Extensive surface diversity of a commensal microorganism by multiple DNA inversions. *Nature*. **414**: 555-558.

123. **Dworkin, J. and Blaser, M.J.** (1996). Generation of *Campylobacter fetus* S-layer protein diversity utilizes a single promoter on an invertible DNA segment. *Molecular Microbiology*. **19**(6): 1241-1253.

124. **Park, S.F.** (2000). Environmental Regulatory Genes. In *Campylobacter*, p. 423-440, 2nd edition. Nachamkin, I. and Blaser, M.J., Editors. ASM Press: Washington, D.C.

125. **Peterson, L., Larsen, T.S., Ussery, D.W., On, S.L.W., and Krogh, A.** (2003). RpoD Promoters in *Campylobacter jejuni* Exhibit a Strong Periodic Signal Instead of a -35 Box. *Journal of Molecular Biology*. **326**: 1361-1372.

126. **Wösten, M.M.S.M., Boeve, M., Koot, M.G.A., van Nuenen, A.C., and van der Zeijst, B.A.M.** (1998). Identification of *Campylobacter jejuni* Promoter Sequences. *Journal of Bacteriology*. **180**(3): 594-599.

127. **Bacon, D.J., Alm, R.A., Hu, L., Hickey, T.E., Ewing, C.P., Batchelor, R.A., Trust, T.J., and Guerry, P.** (2002). DNA sequence and mutational analyses of the pVir plasmid of Campylobacter jejuni 81-176. *Infection and Immunity*. **70**(11): 6242-6250.

128. **Batchelor, R.A., Pearson, B.M., Friis, L.M., Guerry, P., and Wells, J.M.** (2004). Nucleotide sequences and comparison of two large conjugative plasmids from different *Campylobacter* species. *Microbiology*. **150**: 3507-3517.

129. **Cossart, P., Boquet, P., Normark, S., and Rappuoli, R.**, *Cellular Microbiology*. 2000, Washington D. C.: ASM Press. 12-13.

130. **Tracz, D.M., Keelan, M., Ahmed-Bentley, J., Gibreel, A., Kowalewska-Grochowska, K., and Taylor, D.E.** (2005). pVir and Bloody Diarrhea in *Campylobacter jejuni* Enteritis. *Emerging Infectious Diseases*. **11**(6): 838-843.

131. **Hofreuter, D., Odenbreit, S., and Haas, R.** (2001). Natural transformation competence in *Helicobacter pylori* is mediated by the basic components of a type IV secretion system. *Molecular Microbiology*. **41**(2): 379-391.

132. **Kuipers, E.J., Israel, D.A., Kusters, J.G., and Blaser, M.J.** (1998). Evidence for a Conjugation-Like Mechanism of DNA Transfer in *Helicobacter pylori. Journal of Bacteriology*. **180**(11): 2901-2905.

133. **Young, G.M., Schmiel, D.H., and Miller, V.L.** (1999). A new pathway for the secretion of virulence factors by bacteria: The flagellar export apparatus functions as a protein-secretion system. *Proceedings of The National Academy of Sciences USA*. **96**: 6456-6461.

134. **Zupan, J.R., Ward, D., and Zambryski, P.** (1998). Assembly of the VirB transport complex for DNA transfer from *Agrobacterium tumefaciens* to plant cells. *Current Opinion in Microbiology*. **1**: 649-655.

135. **Byrd, D.R. and Matson, S.W.** (1997). Nicking by transesterification: the reaction catalysed by a relaxase. *Molecular Microbiology*. **25**(6): 1011-1022.

136. **Firth, N., Ippen-Ihler, K., and Skurray, R.A.** (1996). Structure and Function of the F Factor and Mechanism of Conjugation. In *Escherichia coli and Salmonella: cellular and molecular biology*, p. 2377-2401, 2nd edition. Neidhardt, F.C., et al., Editors. ASM Press: Washington D. C.

137. **Christie, P.J.** (1997). *Agrobacterium tumefaciens* T-Complex Transport Apparatus: a Paradigm for a New Family of Multifunctional Transporters in Eubacteria. *Journal of Bacteriology*. **179**(10): 3085-3094.

138. **Hendrixson, D.R. and DiRita, V.J.** (2003). Transcription of σ54-dependent but not σ28-dependent flagellar genes in Campylobacter jejuni is associated with formation of the flagellar secretory apparatus. *Molecular Microbiology*. **50**(2): 687-702.

139. **Lander, E.S. and Waterman, M.S.** (1988). Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis. *Genomics*. **2**: 231-239.

140. **Deininger, P.L.** (1983). Random Subcloning of Sonicated DNA: Application to Shotgun DNA Sequence Analysis. *Analytical Biochemistry*. **129**: 216-223.

141. **Pedulla, M.L., Ford, M.E., Houtz, J.M., Karthikeyan, T., Wadsworth, C., Lewis, J.A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N.R., Brucker, W., Kumar, V., Kandasamy, J., Keenan, L., Bardarov, S., Kriakov, J., Lawrence, J.G., Jacobs, W.R.J., Hendrix, R.W., and Hatfull, G.F.** (2003). Origins of highly mosaic mycobacteriophage genomes. *Cell*. **113**(2): 171-182.

142. **Fry, B.N., Korolik, V., ten Brinke, J.A., Pennings, M.T.T., Zalm, R., Teunis, B.J.J., Coloe, P.J., and van der Zeijst, B.A.M.** (1998). The lipopolysaccharide biosynthesis locus of *Campylobacter jejuni* 81116. *Microbiology*. **144**: 2049-2061.

143. **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. *The Journal of Biological Chemistry*. **277**(1): 327-337.

144. **Oldfield, N.J., Moran, A.P., Millar, L.A., Prendergast, M.M., and Ketley, J.M.** (2002). Characterization of the *Campylobacter jejuni* Heptosyltransferase II Gene, *waaF*, Provides Genetic Evidence that Extracellular Polysaccharide Is Lipid A Core Independent. *Journal of Bacteriology*. **184**(8): 2100-2107.

145. **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. *Infection and Immunity*. **70**(2): 787-793.

146. **Gilbert, M., Brisson, J.-R., Karwaski, M.-F., Michniewicz, J., Cunningham, A.-M., Wu, Y., Young, N.M., and Wakarchuk, W.** (2000). Biosynthesis of Ganglioside Mimics in *Campylobacter jejuni* OH4384: Identification of the Glycosyltransferase Genes, Enzymatic Synthesis of Model Compounds, and Characterization of Nanomole Amounts by 600-MHz 1H and 13C NMR. *The Journal of Biological Chemistry*. **275**(6): 3896-3906.

147. **Kanipes, M.I., Holder, L.C., Corcoran, A.T., Moran, A.P., and Guerry, P.** (2004). A Deep-Rough Mutant of *Campylobacter jejuni* 81-176 Is Noninvasive for Intestinal Epithelial Cells. *Infection and Immunity*. **72**(4): 2452-2455.

148. **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. *The Journal of Biological Chemistry*. **276**(37): 34862-34870.

149. **Nuijten, P.J.M., van Asten, F.J.A.M., Gaastra, W., and van der Zeijst, B.A.M.** (1990). Structural and Functional Analysis of Two *Campylobacter jejuni* Flagellin Genes. *The Journal of Biological Chemistry*. **265**(29): 17798-17804.

150. **Nuijten, P.J.M., van den Berg, A.J.G., Formentini, I., van der Zeijst, B.A.M., and Jacobs, A.A.** (2000). DNA Rearrangements in the Flagellin Locus of a *flaA* Mutant of *Campylobacter jejuni* during Colonization of Chicken Ceca. *Infection and Immunity*. **68**(12): 7137-7140.

151. **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., 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. *Molecular Microbiology*. **55**(1): 90-103.

152. **Yao, R. and Guerry, P.** (1996). Molecular Cloning and Site-Specific Mutagenesis of a Gene Involved in Arylsulfatase Production in *Campylobacter jejuni. Journal of Bacteriology*. **178**(11): 3335-3338.

153. **Horswill, A.R. and Escalante-Semerena, J.C.** (1999). *Salmonella typhimurium* LT2 Catabolizes Propionate via the 2-Methylcitric Acid Cycle. *Journal of Bacteriology*. **181**(18): 5615-5623.

154. **Miller, W.G., Pearson, B.M., Wells, J.M., Parker, C.T., Kapitonov, V.V., and Mandrell, R.E.** (2005). Diversity within the *Campylobacter jejuni* type I restriction-modification loci. *Microbiology*. **151**: 337-351.

155. **Perry, A.C.F., Ni Bhriain, N., Brown, N.L., and Rouch, D.A.** (1991). Molecular characterization of the *gor* gene encoding glutathione reductase from *Pseudomonas aeruginosa*: determinants of substrate specificity among pyridine nucleotide-disulphide oxidoreductases. *Molecular Microbiology*. **5**(1): 163-171.

156. **Lüneberg, E., Glenn-Calvo, E., Hartmann, M., Bär, W., and Frosch, M.** (1998). The Central, Surface-Exposed Region of the Flagellar Hook Protein FlgE of *Campylobacter jejuni* Shows Hypervariability among Strains. *Journal of Bacteriology*. **180**(14): 3711-3714.

157. **Parker, C.T., Horn, S.T., Gilbert, M., Miller, W.G., Woodward, D.L., and Mandrell, R.E.** (2005). Comparison of *Campylobacter jejuni* Lipooligosaccharide Biosynthesis Loci from a Variety of Sources. *Journal of Clinical Microbiology*. **43**(6): 2771-2781.

158. **Skurnik, M., Peippo, A., and Ervelä, E.** (2000). Characterization of the O-antigen gene clusters of *Yersinia pseudotuberculosis* and the cryptic O-antigen gene cluster of *Yersinia pestis* shows that the plague bacillus is most closely related to and has evolved from *Y. pseudotuberculosis* serotype O:1b. *Molecular Microbiology*. **37**(2): 316-330.

159. **Goon, S., Kelly, J.F., Logan, S.M., Ewing, C.P., and Guerry, P.** (2003). Pseudaminic acid, the major modification on *Campylobacter* flagellin, is synthesized via the Cj1293 gene. *Molecular Microbiology*. **50**(2): 659-671.

160. **Davidson, A.L. and Chen, J.** (2004). ATP-Binding Cassette Transporters in Bacteria. *Annual Review of Biochemistry*. **73**: 241-268.

161. **Janausch, I.G., Zientz, E., Tran, Q.H., Kröger, A., and Unden, G.** (2002). C4-dicarboxylate carriers and sensors in bacteria. *Biochimica et Biophysica Acta*. **1553**: 39-56.

162. **Wilson, G.G. and Murray, N.E.** (1991). Restriction and Modification Systems. *Annual Review of Genetics*. **25**: 585-627.

163. **Mobley, H.L.T., Mendz, G.L., and Hazell, S.L.**, *Helicobacter pylori: physiology and genetics*. 2001, Washington, D. C.: ASM Press.

164. **Neidhardt, F.C., Curtis III, R., Ingraham, J.L., Lin, E.C.C., Low, K.B., Magasanik, B., Reznikoff, W.S., Riley, M., Schaechter, M., and Umbarger, H.C.**, *Escherichia coli and Salmonella: cellular and molecular biology*. 2nd ed. 1996, Washington D. C.: ASM Press.

165. **Carmel-Harel, O. and Storz, G.** (2000). Roles of the glutathione- and thioredoxin-dependent reduction systems in the *Escherichia coli* and *Saccharomyces cerevisiae* responses to oxidative stress. *Annual Review of Microbiology*. **54**: 439-461.

166. **Chevalier, C., Thiberge, J.-M., Ferrero, R.L., and Labigne, A.** (1999). Essential role of *Helicobacter pylori* gamma-glutamyltranspeptidase for the colonization of the gastric mucosa of mice. *Molecular Microbiology*. **31**(5): 1359-1372.

167. **Marchant, J., Wren, B.W., and Ketley, J.M.** (2002). Exploiting genome sequence: predictions for mechanisms of *Campylobacter* chemotaxis. *Trends in Microbiology*. **10**(4): 155-159.

168. **Lomholt, H. and Kilian, M.** (1994). Antigenic relationships among immunoglobulin A1 proteases from Haemophilus, Neisseria, and Streptococcus species. *Infection and Immunity*. **62**(8): 3178-3183.

169. **Vitovski, S., Read, R.C., and Sayers, J.R.** (1999). Invasive isolates of *Neisseria meningitidis* possess enhanced immunoglobulin A1 protease activity compared to colonizing strains. *FASEB Journal*. **13**: 331-337.

170. **Poly, F., Threadgill, D., and Stintzi, A.** (2005). Genomic Diversity in *Campylobacter jejuni*: Identification of *C. jejuni* 81-176-Specific Genes. *Journal of Clinical Microbiology*. **43**(5): 2330-2338.

171. **Hendrix, R.W.** (2003). Bacteriophage genomics. *Current Opinion in Microbiology*. **6**: 506-511.

172. **Konkel, M.E., Marconi, R.T., Mead, D.J., and Cieplak JR., W.** (1994). Identification and characterization of an intervening sequence within the 23S ribosomal RNA genes of *Campylobacter jejuni. Molecular Microbiology*. **14**(2): 235-241.

173. **Baar, C., Eppinger, M., Raddatz, G., Simon, J., Lanz, C., Klimmek, O., Nandakumar, R., Gross, R., Rosinus, A., Keller, H., Jagtap, P., Linke, B., Meyer, F., Lederer, H., and Schuster, S.C.** (2003). Complete genome sequence and analysis of *Wolinella succinogenes. Proceedings of The National Academy of Sciences USA*. **100**(20): 11690-11695.

174. **MacKichan, J.K., Gaynor, E.C., Chang, C., Cawthraw, S., Newell, D.G., Miller, J.F., and Falkow, S.** (2004). The *Campylobacter jejuni dccRS* two-component system is required for optimal *in vivo* colonization but is dispensable for *in vitro* growth. *Molecular Microbiology*. **54**(5): 1269-1286.

175. **Henderson, I.R., Navarro-Garcia, F., and Nataro, J.P.** (1998). The great escape: structure and function of the autotransporter proteins. *Trends in Microbiology*. **6**(9): 370-378.

176. **Ng, L.K., Stiles, M.E., and Taylor, D.E.** (1987). DNA probes for identification of tetracycline resistance genes in Campylobacter species isolated from swine and cattle. *Antimicrobial Agents and Chemotherapy*. **31**(11): 1669-1674.

177. **Bannam, T.L., Crellin, P.K., and Rood, J.I.** (1995). Molecular genetics of the chloramphenicol-resistance transposon Tn4451 from *Clostridium perfringens*: the TnpX site-specific recombinase excises a circular transposon molecule. *Molecular Microbiology*. **16**(3): 535-551.

178. **Pao, S.S., Paulsen, I.T., and Saier Jr., M.H.** (1998). Major Facilitator Superfamily. *Microbiology and Molecular Biology Reviews*. **62**(1): 1-34.

179. **Sumby, P. and Smith, M.C.M.** (2002). Genetics of the phage growth limitation (Pgl) system of *Streptomyces coelicolor* A3(2). *Molecular Microbiology*. **44**(2): 489-500.

180. **Gennis, R.B. and Valley, S.** (1996). Respiration. In *Escherichia coli and Salmonella: cellular and molecular biology*, p. 217-261Neidhardt, F.C., et al., Editors. ASM Press: Washington, D. C.

181. **Kranz, R., Lill, R., Goldman, B., Bonnard, G., and Merchant, S.** (1998). Molecular mechanisms of cytochrome c biogenesis: three distinct systems. *Molecular Microbiology*. **29**(2): 383-396.

182. **Schiött, T., von Wachienfeldt, C., and Hederstedt, L.** (1997). Identification and Characterization of the *ccdA* Gene, Required for Cytochrome C Synthesis in *Bacillus subtilis. Journal of Bacteriology*. **179**(6): 1962-1973.

183. **Simon, J., Gross, R., Einsle, O., Kroneck, P.M.H., Kröger, A., and Klimmek, O.** (2000). A NapC/NirT-type cytochrome c (NrfH) is the mediator between the quinone pool and the cytochrome c nitrite reductase of *Wolinella succinogenes. Molecular Microbiology*. **35**(3): 686-696.

184. **Weiner, J.H., Rothery, R.A., Sambasivarao, D., and Trieber, C.A.** (1992). Molecular analysis of dimethylsulfoxide reductase: a complex iron-sulfur molybdoenzyme of *Escherichia coli. Biochimica et Biophysica Acta*. **1102**: 1-18.

185. **Lascelles, J. and Calder, K.M.** (1985). Participation of Cytochromes in Some Oxidation-Reduction Systems in *Campylobacter fetus. Journal of Bacteriology*. **164**(1): 401-409.

186. **Baltes, N., Hennig-Pauka, I., Jacobsen, I., Gruber, A.D., and Gerlach, G.F.** (2003). Identification of Dimethyl Sulfoxide Reductase in *Actinobacillus pleuropneumoniae* and Its Role in Infection. *Infection and Immunity*. **71**(12): 6784-6792.

187. **Holland, I., Kenny, B., and Blight, M.** (1990). Hemolysin secretion from *E. coli. Biochimie*. **72**: 131-141.

188. **Pugsley, A.P., Kornacker, M.G., and Poquet, I.** (1991). The general protein-export pathway is directly required for extracellular pullulanase secretion in *Escherichia coli* K12. *Molecular Microbiology*. **5**(2): 343-352.

189. **Wilharm, G., Lehmann, V., Neumayer, W., Trcek, J., and Heesemann, J.** (2004). *Yersinia enterocolitica* type III secretion: Evidence for the ability to transport proteins that are folded prior to secretion. *BMC Microbiology*. **4**(1): 27-36.

190. **Weiss, A.A., Johnson, F.D., and Burns, D.L.** (1993). Molecular characterization of an operon required for pertussis toxin secretion. *Proceedings of The National Academy of Sciences USA*. **90**: 2970-2974.

191. **Reyrat, J.-M., Pelicic, V., Papini, E., Montecucco, C., Rappuoli, R., and Telford, J.L.** (1999). Towards deciphering the *Helicobacter pylori* cytotoxin. *Molecular Microbiology*. **34**(2): 197-204.

192. **Henderson, I.R. and Nataro, J.P.** (2001). Virulence Functions of Autotransporter Proteins. *Infection and Immunity*. **69**(3): 1231-1243.

193. **Brown, N.F., Logue, C.A., Boddey, J.A., Scott, R., Hirst, R.G., and Beacham, I.R.** (2004). Identification of a novel two-partner secretion system from *Burkholderia pseudomallei. Molecular Genetics and Genomics*. **272**: 204-215.

194. **Schiebel, E., Schwarz, H., and Braun, V.** (1989). Subcellular Location and Unique Secretion of the Hemolysin of *Serratia marcescens. The Journal of Biological Chemistry*. **264**(27): 16311-16320.

195. **Jacob-Dubuisson, F., Locht, C., and Antoine, R.** (2001). Two-partner secretion in Gram-negative bacteria: a thrifty, specific pathway for large virulence proteins. *Molecular Microbiology*. **40**(2): 306-313.

196. **Hugdahl, M.B., Beery, J.T., and Doyle, M.P.** (1988). Chemotactic Behavior of *Campylobacter jejuni. Infection and Immunity*. **56**(6): 1560-1566.

197. **Hendrixson, D.R. and DiRita, V.J.** (2004). Identification of *Campylobacter jejuni* genes involved in commensal colonization of the chick gastrointestinal tract. *Molecular Microbiology*. **52**(2): 471-484.

198. **Levit, M.N., Liu, Y., and Stock, J.B.** (1998). Stimulus response coupling in bacterial chemotaxis: receptor dimers in signalling arrays. *Molecular Microbiology*. **30**(3): 459-466.

199. **Harkey, C.W., Eveiss, K.D., and Peterson, K.M.** (1994). The *Vibrio cholerae* Toxin-Coregulated-Pilus Gene *tcpI* encodes a Homolog of Methyl-Accepting Chemotaxis Proteins. *Infection and Immunity*. **62**(7): 2669-2678.

200. **St Michael, F., Szymanski, C.M., Li, J., Chan, K.H., Khieu, N.H., Larocque, S., Wakarchuk, W., Brisson, J.-R., and Monteiro, M.A.** (2002). The structures of the lipooligosaccharide and capsule polysaccharide of *Campylobacter jejuni* genome sequenced strain NCTC 11168. *European Journal of Biochemistry*. **269**: 5119-5136.

201. **Szymanski, C.M., St Michael, F., Jarrell, H.C., Li, J., Gilbert, M., Larocque, S., Vinogradov, E., and Brisson, J.-R.** (2003). Detection of Conserved *N*-Linked Glycans and Phase-variable Lipooligosaccharides and Capsules from Campylobacter Cells by Mass Spectrometry and High Resolution Magic Angle Spinning NMR Spectroscopy. *The Journal of Biological Chemistry*. **278**(27): 24509-24520.

202. **Connerton, P.L., Loc Carrillo, C.M., Swift, C., Dillon, E., Scott, A., Rees, C.E.D., Dodd, C.E.R., Frost, J.A., and Connerton, I.F.** (2004). Longitudinal Study of *Campylobacter jejuni* Bacteriophages and Their Hosts from Broiler Chickens. *Applied and Environmental Microbiology*. **70**(7): 3877-3883.

203. **Morgan, G.J., Hatfull, G.F., Casjens, S., and Hendrix, R.W.** (2002). Bacteriophage Mu Genome Sequence: Analysis and Comparison with Mu-like Prophages in *Haemophilus*, *Neisseria* and *Deinococcus. Journal of Molecular Biology*. **317**: 337-359.

204. **Buswell, C.M., Herlihy, Y.M., Lawrence, L.M., McGuiggan, J.T.M., Marsh, P.D., Keevil, C.W., and Leach, S.A.** (1998). Extended Survival and Persistence of *Campylobacter* spp. in Water and Aquatic Biofilms and Their Detection by Immunofluorescent-Antibody and -rRNA Staining. *Applied and Environmental Microbiology*. **64**(2): 733-741.

205. **Pearson, A.D., Greenwood, M., Healing, T.D., Rollins, D., Shahamat, M., Donaldson, J., and Colwell, R.R.** (1993). Colonization of Broiler Chickens by Waterborne *Campylobacter jejuni. Applied and Environmental Microbiology*. **59**(4): 987-996.

206. **Jefferson, K.K.** (2004). What drives bacteria to produce a biofilm? *FEMS Microbiology Letters*. **236**: 163-173.

207. **Gibreel, A., Tracz, D.M., Nonaka, L., Ngo, T.M., Connell, S.R., and Taylor, D.E.** (2004). Incidence of Antibiotic Resistance in *Campylobacter jejuni* Isolated in Alberta, Canada, from 1999 to 2002, with Special Reference to *tet*(O)-Mediated Tetracycline Resistance. *Antimicrobial Agents and Chemotherapy*. **48**(9): 3442-3450.

208. **Nachamkin, I., Ung, H., and Li, M.** (2002). Increasing fluoroquinolone resistance in Campylobacter jejuni, Pennsylvania, USA, 1982-2001. *Emerging Infectious Diseases*. **8**(12): 1501-1503.

209. **Avrain, L., Humbert, F., L'Hospitalier, R., Sanders, P., Vernozy-Rozand, C., and Kempf, I.** (2003). Antimicrobial resistance in *Campylobacter* from broilers: association with production type and antimicrobial use. *Veterinary Microbiology*. **96**: 267-276.

210. **Aarestrup, F.M.** (1999). Association between the consumption of antimicrobial agents in animal husbandry and the occurrence of resistant bacteria among food animals. *International Journal of Antimicrobial Agents*. **12**: 279-285.

211. **Kersulyte, D., Mukhopadhyay, A.K., Shirai, M., Nakazawa, T., and Berg, D.E.** (2000). Functional Organization and Insertion Specificity of IS607, a Chimeric Element of *Helicobacter pylori. Journal of Bacteriology*. **182**(19): 5300-5308.

212. **Connell, S.R., Tracz, D.M., Nierhaus, K.H., and Taylor, D.E.** (2003). Ribosomal Protection Proteins and Their Mechanism of Tetracycline Resistance. *Antimicrobial Agents and Chemotherapy*. **47**(12): 3675-3681.

213. **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**(1): 127-130.

214. **Sambrook, J. and Russell, D.W.** (2001). Southern Blotting: Capillary Transfer of DNA to Membranes. In *Molecular Cloning: A laboratory manual*, p. 6.39-6.46, 3rd edition. Argentine, J., Editor. Cold Spring Harbor Laboratory Press: New York.

215. **Higuchi, R., Fockler, C., Dollinger, G., and Watson, R.** (1993). Kinetic PCR Analysis: Real-time Monitoring of DNA Amplification Reactions. *Nature Biotechnology*. **11**(9): 1026-1030.

216. **Garnier, T., Saurin, W., and Cole, S.T.** (1987). Molecular characterization of the resolvase gene, *res*, carried by a multicopy plasmid from *Clostridium perfringens*: common evolutionary origin for prokaryotic site-specific recombinases. *Molecular Microbiology*. **1**(3): 371-376.

217. **Newman, B.J. and Grindley, N.D.F.** (1984). Mutants of the γδ Resolvase: A Genetic Analysis of the Recombination Function. *Cell*. **38**: 463-469.

218. **Silverman, M. and Simon, M.** (1980). Phase Variation: Genetic Analysis of Switching Mutants. *Cell*. **19**: 845-854.

219. **Johnson, R.C., Bruist, M.F., and Simon, M.I.** (1986). Host Protein Requirements for In Vitro Site-Specific DNA Inversion. *Cell*. **46**: 531-539.

220. **Plasterk, R.H.A., Kanaar, R., and van de Putte, P.** (1984). A genetic switch *in vitro*: DNA inversion by Gin protein of phage Mu. *Proceedings of The National Academy of Sciences USA*. **81**: 2689-2692.

221. **van de Putte, P., Cramer, S., and Giphart-Gassler, M.** (1980). Invertible DNA determines host specificity of bacteriophage Mu. *Nature*. **286**: 218-222.

222. **Komano, T., Kubo, A., and Nisioka, T.** (1987). Shufflon: multi-inversion of four contiguous DNA segments of plasmid R64 creates seven different open reading frames. *Nucleic Acids Research*. **15**(3): 1165-1172.

223. **Smith, M.C.M. and Thorpe, H.M.** (2002). Diversity in the serine recombinases. *Molecular Microbiology*. **44**(2): 299-307.

224. **van den Berg, E., Zwetsloot, J., Noordermeer, I., Pannekoek, H., Dekker, B., Dijkema, R., and van Ormondt, H.** (1981). The structure and function of the regulatory elements of the Escherichia coli uvrB gene. *Nucleic Acids Research*. **9**(21): 5623-5643.

225. **Aiba, H.** (1983). Autoregulation of the Escherichia coli *crp* Gene: CRP Is a Transcriptional Repressor for Its Own Gene. *Cell*. **32**: 141-149.

226. **Bertucci, F., Bernard, K., Loriod, B., Chang, Y.-C., Granjeaud, S., Birnbaum, D., Nguyen, C., Peck, K., and Jordan, B.R.** (1999). Sensitivity issues in DNA array-based expression measurements and performance of nylon microarrays for small samples. *Human Molecular Genetics*. **8**(9): 1715-1722.

227. **Chissoe, S.L., Marra, M.A., Hillier, L., Brinkman, R., Wilson, R.K., and Waterston, R.H.** (1997). Representation of cloned genomic sequences in two sequencing vectors: correlation of DNA sequence and subclone distribution. *Nucleic Acids Research*. **25**(15): 2960-2966.

228. **Maurelli, A.T., Fernández, R.E., Bloch, C.A., and Rode, C.K.** (1998). "Black holes" and bacterial pathogenicity: A large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli. Proceedings of The National Academy of Sciences USA*. **95**: 3943-3948.

229. **Beliaev, A.S. and Saffarini, D.A.** (1998). *Shewanella putrefaciens mtrB* encodes an outer membrane protein required for Fe(III) and Mn(IV) reduction. *Journal of Bacteriology*. **180**(23): 6292-6297.

230. **Gaynor, E.C., Cawthraw, S., Manning, G., MacKichan, J.K., Falkow, S., and Newell, D.G.** (2004). The Genome-Sequenced Variant of *Campylobacter jejuni* NCTC 11168 and the Original Clonal Clinical Isolate Differ Markedly in Colonization, Gene Expression, and Virulence-Associated Phenotypes. *Journal of Bacteriology*. **186**(2): 503-517.

231. **Carrillo, C.D., Taboada, E., Nash, J.H.E., Lanthier, P., Kelly, J.F., Lau, P.C., Verhulp, R., Mykytczuk, O.L., 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*. *The Journal of Biological Chemistry*. **279**(19): 20327-20338.

APPENDIX 1: predicted CDSs and their protein similarities in the *C. jejuni* strain 81-176 plasmid pVir

Locus_id	length	Putative function	Informative database match	Organism with match	SWALL	E-value	id
pVir1	373	Unknown	-	-			
pVir2	292	Unknown	-	-			
pVir3	260	Unknown	TrbM	*E. coli*	Q03537	5.1e-11	31.84
pVir4	239	Unknown	-	-			
pVir5	417	Unknown	Hypothetical Hp0444	*H. pylori*	O25192	4.3e-12	32.54
pVir6	142	Unknown	-	-			
pVir7	114	Unknown	-	-			
pVir8	102	Unknown	-	-			
pVir9	73	Unknown	-	-			
pVir10	131	Unknown	-	-			
pVir11	136	Unknown	-	-			
pVir12	143	Unknown	-	-			
pVir13	61	Unknown	-	-			
pVir14	56	Unknown	-	-			
pVir15	42	Unknown	-	-			
pVir16c	66	Unknown	-	-			
pVir17	121	Unknown	-	-			
pVir18	111	Unknown	-	-			
pVir19	120	Unknown	-	-			
pVir20	134	Unknown	-	-			
pVir21	130	Periplasmic protein	Cj1456c	*C. jejuni*	Q9PMK4	4.6e-31	90.38
pVir22c	523	Unknown	Hypothetical jhp0942	*H. pylori*	Q9ZKJ3	7.3e-13	28.06
pVir23c	82	Unknown	-	-			
pVir24	101	Unknown	-	-			
pVir25	80	Unknown	Hypothetical Hp0042	*H. pylori*	O25190	2.6e-03	35.29
pVir26	822	Type IV secretion system protein	VirB4	*H. pylori*	O25189	1.7e-44	33.87
pVir27	225	Type IV secretion system protein	VirB8/ComB1	*C. jejuni*	Q9KIS2	3.1e-81	100

Locus_id	length	Putative function	Informative database match	Organism with match	SWALL	E-value	id
pVir28	356	Type IV secretion system protein	ComB2	*C. jejuni*	Q9KIS1	1.1e-124	100
pVir29	378	Type IV secretion system protein	ComB3	*C. jejuni*	Q9KIS0	9.2e-122	100
pVir30	66	Unknown	-	-			
pVir31	317	Type IV secretion system protein	VirB11	*C. jejuni*	Q9KIR9	9.8e-115	100
pVir32	135	Unknown	-	-			
pVir33	628	Type IV secretion system protein	VirD4	*E. coli*	Q91UW5	4e-20	24.52
pVir34	56	Unknown	-	-			
pVir35	293	Unknown	Hypothetical jhp0926	*H. pylori*	Q9ZKK9	8e-03	21.56
pVir36	89	Unknown	-	-			
pVir37	382	Conjugal transfer protein	Mlr9255	*Rhizobium loti*	Q981S2	1.3e-03	22.41
pVir38	655	Topoisomerase	TopA2	*H. pylori*	Q9ZKL6	7.9e-37	44.89
pVir39	121	Unknown	-	-			
pVir40	152	Single-stranded DNA-binding protein	Ssb-p1	Bacteriophage P1	Q9XJG4	4.3e-12	30.24
pVir41	57	Unknown	-	-			
pVir42	211	Unknown	-	-			
pVir43	155	Unknown	-	-			
pVir44	117	Unknown	-	-			
pVir45	70	Unknown	-	-			
pVir46	156	Unknown	-	-			
pVir47	137	Unknown	-	-			
pVir48	135	Unknown	-	-			
pVir49	107	Unknown	-	-			
pVir50	77	Unknown	-	-			
pVir51c	67	Unknown	-	-			
pVir52	222	Partition protein	ParA	*H. pylori*	O25646	1.1e-14	38.02
pVir53	209	Unknown	-	-			
pVir54c	278	Replication initiation protein	RepA	*Erysipelothrix rhusiopathiae*	Q9RHE5	1.1e-13	30.73

APPENDIX 2: predicted CDSs and their protein similarities for the *C. jejuni* strain 81-176 plasmid pTet

Locus id	length	Putative function	Informative database match	Organism with match	SWALL	E-value	id
pTet1	382	Replication initiation protein	replication protein	*Selenomonas ruminantium* plasmid ps23	Q55007	1.9e-29	36.48
pTet2	126	Unknown	-	-			
pTet3	132	Unknown	Hypothetical cjp38	*C. jejuni*	Q8GJB7	1.6e-16	40
pTet4	170	Unknown	-	-			
pTet5	185	Unknown	-	-			
pTet6	88	Unknown	Hypothetical rgi82	*Oryza sativa*	Q944E8	5.2e-03	30.3
pTet7	186	Unknown	-	-			
pTet8	88	Unknown	-	-			
pTet9	1932	DNA methylase	Orf23	*Sinorhizobium meliloti* phage PBC5	Q8W6K4	3.7e-135	38.19
pTet10c	234	Unknown	-	-			
pTet11c	462	Nickase	MagA2	*Actinobacillus actinomycetemcomitans*	Q9F276	8.9e-25	32.26
pTet12c	183	unknown	-	-			
pTet13	93	Unknown	-	-			
pTet14	203	Unknown	-	-			
pTet15	217	Unknown	Hypothetical jhp0950	*H. pylori*	Q9ZKI5	1.8e-19	39.63
pTet16	408	DNA primase	TraC	*E. coli*	P27189	4.1e-15	31.56
pTet17	87	Lipoprotein	MagB5	*Actinobacillus actinomycetemcomitans*	Q9F247	1.1e-02	37.7
pTet18c	85	Unknown	-	-			
pTet19c	61	Unknown	-	-			
pTet20	72	Unknown	Hypothetical jhp0960	*H. pylori*	Q9ZKH6	5.6e-10	52.77
pTet21	67	Unknown	Hypothetical jhp0961	*H. pylori*	Q9ZKH5	4.4e-13	68.42
pTet22	597	Unknown	Hypothetical amv156	*Amsacta moorei* entomopoxvirus	Q9EMP3	6.1e-04	22.74
pTet23c	204	Site-specific DNA recombinase	Soao172	*Shewanella oneidensis*	Q8E7Z6	1e-14	33.16

Locus id	length	Putative function	Informative database match	Organism with match	SWALL	E-value	id
pTet24c	125	Virulence-associated protein	Vap2	*Riemerella anatipestifer* pCFC1	O85171	1.9e-04	36.26
pTet25c	107	Unknown	-	-			
pTet26	87	Type IV secretion system protein	VirB2	*E. coli*	Q91UX6	1e-06	35.36
pTet27	922	ATPase	MagB3	*Actinobacillus actinomycetemcomitans*	Q9F245	1e-128	40.67
pTet28	188	Unknown	Hypothetical	*C. jejuni* pCjA13	Q847A4	1.3e-14	44.8
pTet29	221	Unknown	-	-			
pTet30	141	Single-strand DNA binding protein	Ssb-1	*Geobacter sulfurreducens*	AAR35527	5.3e-11	33.58
pTet31	86	Unknown	-	-			
pTet32	323	Unknown	MagB4	*Actinobacillus actinomycetemcomitans*	Q9F246	6e-19	32.66
pTet33	332	Unknown	MagB6	*Actinobacillus actinomycetemcomitans*	Q9F248	1.4e-15	25.93
pTet34	55	Lipoprotein	Cj1074c	*C. jejuni*	Q9PNM0	0.24	44.68
pTet35	220	Type IV secretion system protein	VirB8-like protein	*C. jejuni* pCjA13	Q847A8	1.1e-77	100
pTet36	295	Type IV secretion system protein	VirB9-like protein	*C. jejuni* pCjA13	Q847A7	3.7e-112	97.28
pTet37	391	Type IV secretion system protein	MagB10	*Actinobacillus actinomycetemcomitans*	Q9F252	5.7e-39	39.74
pTet38	330	Type IV secretion system protein	VirB11-like protein	*C. jejuni* pCjA13	Q847A5	4.6e-117	99.69
pTet39	603	Type IV secretion system protein	MagB12	*Actinobacillus actinomycetemcomitans*	Q9F254	4.9e-89	42.64
pTet40	145	Lipoprotein	MagB13	*Actinobacillus actinomycetemcomitans*	Q9F255	4.5e-03	26.57

Locus id	length	Putative function	Informative database match	Organism with match	SWALL	E-value	id
pTet41	254	Unknown	TrbM-like protein	*Haemophilus aegyptius* pF3031	Q8VRC6	4.5e-11	37.17
pTet42	265	Unknown	-	-			
pTet43	206	Unknown	-	-			
pTet44	730	Topoisomerase	TraE	*E. coli*	Q60215	1.8e-80	41.89
pTet45	473	Unknown	Hypothetical	*Plasmodium falciparum*	P21421	2.5e-03	25.39
pTet46	59	Unknown	Hypothetical cjp20	*C. jejuni*	Q8GJD3	3.1e-07	46.42
pTet47	639	Tetracycline resistance	TetO	*C. jejuni*	AAA23033	0	99.84
pTet48	57	Unknown	Hypothetical Orf6	*Enterococcus faecalis* transposon tn916	Q56396	4.3e-14	66.66
pTet49	222	Unknown	-	-			
pTet50	140	unknown	-	-			

APPENDIX 3

Predicted CDSs for sequenced pUC library clones of strain 81-176

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
5c06.p	2	753	8P0001	250	Fcl cj1428c	11168	Q9PMM9	56.4	2.70E-51	250
6a02.q	2	730	8P0002c	243	adhesin	*Chromobacterium violaceum*	Q7NY05	30		
8b03.p	2	815	8P0003	35	GlyA cj0402	11168	P24531	97.14	2.80E-10	35
			8P0004	232	hypothetical cj0403	11168	Q9PIA2	100	1.10E-69	176
6e09.q	2	1055	8P0005	251	hypothetical from LOS cluster	*C.jejuni* strain 11351 81176	Q9ALY2	100	1.60E-99	251
			8P0006c	73	WaaF	*C.jejuni* strain 81176	Q6TDC6	100	7.20E-30	73
6h01.q	2	770	8P0007	256	c4-dicarboxylate transporter	*Vibrio fulnificus*	Q7MJB8	38.93	6.00E-26	244
2a01.p	2	919	8P0008	86	no matches					
			8P0009	60	no matches					
7e10.q	2	532	8P0010c	176	aminotransferase cj1294	11168	Q9PN05	89.2	3.90E-53	176
7e07.q	2	772	8P0011c	189	DsbA cj0872	11168	Q9PP57	48.04	1.20E-28	179
7g05.p	2	1358	8P0012	52	Cj1161	11168	Q9PND4	83.67	2.50E-13	49
			8P0013	173	hydrophobic protein cj1158c	11168	Q9PND7	84.21	6.80E-22	76
			8P0014c	196	DnaX cj1157	11168	Q9PND8	95.91	9.30E-65	196
7d11.q	2	931	8P0015c	310	cj1333 like hypothetical	81-176	Q7X518	100	3.00E-125	309
1b02.p	3	1444	8P0016	165	ribosomal acetyltransferase	*Ureaplasma parvum*	Q9PQI0	29.1	1.60E-02	134
			8P0017	60	no matches					
			8P0018	136	WbkC	*Brucella melitensis*	Q9ZHX0	33.96	1.10E-03	106
			8P0019	74	acyl carrier protein cj1308	11168	Q9PMZ1	93.05	1.80E-21	72
5a05.p	3	967	8P0020	61	cj1724c hypothetical	11168	Q9PLV4	100	3.60E-23	60
			8P0021	199	cj1721c outer membrane protein	11168	Q9PLV7	63.77	5.20E-48	196
6a01.p	3	1000	8P0022c	74	hypothetical cj0976	11168	Q9PNW3	94.59	7.70E-26	74
			8P0023c	226	heme-hemopexin HxuB	*Haemophilus influenzae*	AAQ10738	20.5	2.20E-02	239
3a07.q	3	1446	8P0024	70	no matches	match to 1580383-1580533				
			8P0025	261	membrane protein cj1658	11168	Q9PM19	96.52	1.70E-82	259
8b05.p	3	1693	8P0026	187	hypothetical cj1340c	11168	Q9PMV9	34.44	1.20E-16	180
			8P0027	226	FlaA	*C.jejuni* strain d2677	Q9R953	100	1.80E-74	226
2d02.p	4	1229	8P0028	336	Cst-I	*C. jejuni* strain oh4384	Q9RGF1	41.14	1.20E-31	367
			8P0029	43	hypothetical cj1431c	11168	Q9PMM6	41.02	6.70E-01	39

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
8h11.p	4	1479	8P0030	59	hypothetical cj0121	11168	Q9PJ06	98.27	7.00E-22	58
			8P0031	402	cj0243c hypothetical	11168	Q9PIQ2	21.51	0.005	344
1a07.p	5	1123	8P0032c	94	no matches					
			8P0033c	279	cj0032	11168	Q9PJ79	64.93	4.90E-54	288
6e04.q	5	1000	8P0034	196	AcnB cj0835c aconitate hydratase	11168	Q9PP88	98.46	5.70E-76	196
			8P0035	116	cj0834c periplasmic protein	11168	Q9PP89	92.24	9.30E-39	116
5a10.q	5	1348	8P0036c	225	cj1442c	11168	Q9PML5	63.34	1.50E-50	221
			8P0037c	224	KpsF	11168	Q9PML4	95.92	5.50E-76	221
2h05.p	5	1445	8P0038c	46	no matches					
			8P0039c	433	hypothetical	*Fusobacterium nucleatum*	Q8REK3	23.59	2.70E-03	339
3e08.q	6	1267	8P0040c	330	cj1310c hypothetical	11168	Q9PMY9	62.95	1.40E-78	332
			8P0041	61	NeuA2	11168	Q9PMY8	96.72	2.40E-20	61
2e09.q	6	1438	8P0042	451	cj0971	11168	Q9PNW7	83.81	2.30E-22	105
1e08.q	6	1086	8P0043c	340	DmhA	*Yersinia pseudotuberculosis*	Q8G8E4	78.2	2.30E-99	335
1c09.q	7	1346	8P0044	218	FlaB	81116	Q9RF25	100	3.60E-66	218
			8P0045c	217	cj1337	81-176	Q7X517	100	7.80E-71	217
3b10.q	8	1985	8P0046c	135	cj0305c	11168	Q9PIJ4	66.66	4.00E-33	135
			8P0047c	380	BioF	11168	Q9PIJ3	75.78	5.40E-112	380
			8P0048	124	BioA	11168	Q9PIJ2	94.35	2.50E-47	124
1b01.p	4	2336	8P0049c	72	type I RM mm2978	*Methanosarcina mazei*	Q8PSU8	37.03	6.40E-03	54
			8P0050c	636	rm cc0620	*Caulobacter crescentus*	Q9AAH8	39.62	7.10E-58	641
7b08.q	8	1272	8P0051	130	cj0294 moeb/thif family protein	11168	Q9PIK5	95.38	2.90E-46	130
			8P0052c	126	PanD cj0296c	11168	Q9PIK3	98.41	2.30E-43	126
			8P0053c	137	PanC cj0297c	11168	Q9PIK2	96.35	8.90E-43	137
4a03.p	11	1388	8P0054c	462	FlgE	81-176	Q83WM5	100	1.10E-177	462
6g02.p	11	1765	8P0055c	412	DTPT transporter (disrupted)	*Photorhabdus luminescens*	Q7N5W6	47.99	1.30E-79	398
			8P0056c	117	ABC transporter	*Photorhabdus luminescens*	Q7N5W6	47.66	1.10E-15	107
4e04.p	12	1893	8P0057c	164	ModC	11168	Q9PIJ9	76.22	1.00E-39	164
			8P0058c	222	ModB	11168	Q9PIJ8	85.13	2.40E-70	222
			8P0059c	133	cj0302c	11168	Q9PIJ7	64.61	1.80E-28	130
			8P0060c	109	ModA	11168	Q9PIJ6	81.65	1.40E-30	109
6d08.p	16	2885	8P0061	76	no matches					
			8P0062	879	type I RM mm2976	*Methanosarcina mazei*	Q8PSV0	44.63	6.10E-131	867

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
2b09.p	14	4502	8P0063	126	Cst-iI	81-176	Q9L9Q5	98.4	5.20E-46	125
			8P0064	346	NeuB1	*C.jejuni* strain atcc 43456	Q93D04	100	6.70E-129	346
			8P0065	374	NeuC1	*C.jejuni* atcc 43456	Q93D03	100	5.60E-134	374
			8P0066	315	CgtA-II	*C.jejuni* atcc 43449 and 43456	Q934C5	100	1.20E-126	315
			8P0067	221	NeuA1	*C.j strain* 43456	Q933W2	100	1.40E-82	221
			8P0068	117	acetyltransferase	*C.jejuni* strain atcc 43449	Q93CZ2	100	2.30E-46	117
6a11.p	28	2110	8P0069	576	FlaB	*C.jejuni* 81116	Q9RF25	97.74	3.90E-174	576
7f02.p	31	4287	8P0070	150	TraN	*Sphingomonas aromaticivorans*	O85935	42	2.60E-17	150
			8P0071	396	TraG	*Escherichia coli*	P33790	20.44	1.70E-04	357
			8P0072	174	no matches					
			8P0073	294	no matches					
			8P0074c	93	no matches					
7f11.p	37	3740	8P0075	49	SecY cj1688	11168	Q9PLZ0	100	1.70E-18	49
			8P0076	398	hypothetical	*Clostridium perfringens*	Q8XNB6	34.7	8.00E-43	412
			8P0077	670	hypothetical	*Rhizobium loti*	Q98CJ2	39.13	5.90E-94	672
6g03.q	38	3087	8P0078	740	DmsA	*Wolinella succinogenes*	Q7MRE1	62.01	5.40E-189	745
			8P0079	218	FdhB	*Wolinella succinogenes*	Q7M8T2	62.67	2.00E-55	217
			8P0080	70	MraY hypothetical	*Wolinella succinogenes*	Q7MRE0	47.14	2.40E-07	70
7d05.p	41	4416	8P0081	519	cyt C biogenesis protein	*Wolinella succinogenes*	Q7M7P8	59.45	3.20E-121	518
			8P0082c	556	GGT jhp1046	*H.pylori* j99	Q9ZK95	67.2	2.90E-134	558
			8P0083	306	cj0031	11168	Q9PJ80	61.93	9.80E-63	310
4b02.p	47	5554	8P0084	656	cytochrome C	*Shewanella oneidensis*	Q8EJI6	55.24	1.60E-136	677
			8P0085	689	cytochrome C family protein	*Geobacter sulfurreducens*	AAR33608	36.31	2.50E-59	614
			8P0086	194	hypothetical	*Wolinella succinogenes*	Q7MQN4	38.88	3.10E-23	198
			8P0087	234	cyt C biogenesis protein	*Helicobacter hepaticus*	Q7VHG9	37.97	4.30E-24	237
6d10.q	56	4739	8P0088	273	cj1368	11168	Q9PMT2	89.37	6.70E-97	273
			8P0089	1121	cj1365c serine protease	11168	Q9PMT5	39.66	2.30E-79	1147
			8P0090	147	cj1369 transport	11168	Q9PMT1	81.63	8.20E-45	147
7g11.p	2	1380	8P0091	218	iron uptake ABC transport cj0173c	11168	Q9PIV6	99.08	2.40E-73	218
			8P0092c	158	PurU cj0790	11168	Q9PPC9	100	2.00E-53	146
			8P0093c	61	RNA nucleotidyltransferase cj0789	11168	Q9PPD0	98.21	7.10E-19	56
2h12.p	2	906	8P0094c	286	no matches					
5e04.q	1	396	8P0095c	93	no matches					
3h05.p	1	662	8P0096c	219	cj1342c hypothetical	11168	Q9PMV7	78.53	4.10E-71	219

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
3d09.q	1	176	8P0097c	54	no matches					
1a12.p	1	658	8P0098c	218	LpsA	*Vibrio parahaemolyticus*	Q87T79	41.36	2.40E-32	220
2c01.q	1	596	8P0099c	162	glycosyltransferase	*C.jejuni* strain atcc 43456	Q93D08	100	1.40E-60	162
4a04.q	1	634	8P0100c	88	Atpe cj0936	11168	Q9PNZ7	90.9	2.20E-22	88
4c05.q	1	641	8P0101	157	no matches					
8e07.p	1	880	8P0102	119	exonuclease recj cj0028	11168	Q9PJ83	97.36	4.60E-41	114
			8P0103	127	Ansa cj0029	11168	Q9PJ82	83.46	4.40E-33	127
1f07.q	1	595	8P0104	37	WaaV	*C.jejuni* strain 43456	Q93D01	100	1.50E-12	37
			8P0105c	160	acetyltransferase	*C.jejuni* strain atcc 43456,	Q93D02	98.75	9.40E-59	161
6a06.p	2	1189	8P0106c	213	hypothetical dsba cj0872	11168	Q9PP57	98.12	9.00E-77	213
			8P0107c	141	arylsulfatase AstA	81-176	Q46098	100	2.80E-54	141
7e09.p	2	901	8P0108c	143	afimbrial adhesin	*Escherichia coli*	Q93QU8	32.39	0.00034	142
5g02.p	1	197			N/A	11168				
1a08.p	1	357			N/A					
5b12.q	1	666			N/A					
6h03.q	2	742			N/A	11168				
6h12.q	1	274			N/A	11168				

APPENDIX 4

Predicted CDSs for sequenced pUC library clones of strain M1

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
2d02.q	2	512	MP0001	137	serine protease SigA	*Shigella flexneri* 2a	Q9L8L1	37.4	3.80E-08	139
5d06.p	2	671	MP0002	223	restriction modification protein CjeI	*C. jejuni* strain p37	Q9JN06	91.55	6.80E-73	225
3a07.p	2	907	MP0003	61	cj1058c	11168	Q9PNN3	77.77	1.50E-10	45
			MP0004	57	cj1057c	11168	Q9PNN4	94.73	9.90E-15	57
			MP0005	184	cj1056c (disrupted)	11168	Q9PNN5	76.34	4.80E-50	186
3a10.q	2	555	MP0006c	163	no matches					
2f03.q	2	664	MP0007c	73	WlaK	*C. jejuni* strain 81116	O86158	98.63	1.90E-26	73
			MP0008c	115	WlaI	*C.jejuni* strain 81116	O86157	100	5.30E-43	115
4e10.q	2	457	MP0009c	121	cj1375	11168	Q9PMS5	94.95	1.40E-39	119
5b05.p	2	823	MP0010c	229	DTPT dehydratase	*Helicobacter hepaticus*	Q7VJZ3	59.29	1.80E-47	226
4e04.p	2	752	MP0011	101	cj0032 RM	11168	Q9PJ79	65.34	1.30E-17	101
			MP0012	148	cj0033 membrane	11168	Q9PJ78	39.37	8.10E-06	160
4e02.q	2	624	MP0013	207	PorA membrane	*C. jejuni* Strain x7199	Q9F782	88.37	1.30E-67	215
3f12.p	2	812	MP0014c	233	cj0139 endonuclease	11168	Q9PIY8	53.28	3.70E-28	259
3b05.q	3	1437	MP0015c	185	glycosyltransferase	*C. jejuni* Strain 11828	Q9ALT2	100	8.50E-72	185
			MP0016c	266	glycosyltransferase	*C. jejuni* Strain 11828	Q9ALT1	100	2.30E-87	228
2h08.p	3	746	MP0017c	195	hypothetical	*C. jejuni* Strain rm1221	Q8RN32	97.43	3.00E-70	195
4a03.q	3	1095	MP0018c	365	FlaA	*C. jejuni* Strain 81116	FLA2_CAMJ	100	1.20E-116	365
3d02.q	6	1972	MP0019	57	alginate O-acetylation protein	*C. jejuni* Strain 11828	Q9ALT7	100	5.40E-22	57
			MP0020	371	hypothetical	*C. jejuni* Strain 11828	Q9ALT8	97.99	6.90E-135	349
			MP0021c	186	cj1149c isomerase	11168	LPC1_CAMJ	96.77	1.40E-65	186
2g06.p	3	887	MP0022	94	ppK cj1359	11168	PPK_CAMJE	98.91	4.00E-29	92
			MP0023c	152	VacA	*H. pylori* J99	Q9ZME6	26.41	7.70E-03	159
3e04.p	3	1277	MP0024c	425	cj1337 hypothetical	*C.jejuni* Strain 81-176	Q7X517	99.76	7.90E-159	424
3e08.p	4	1095	MP0025	273	no matches					
2c03.p	4	794	MP0026c	242	no matches					
1g01.q	4	944	MP0027c	314	cj1178c acidic	11168	Q9PNB7	91.42	2.90E-80	315
1f05.p	4	1115	MP0028c	307	RlmA transferase	*C.jejuni* strain 81116	Q9K5D0	98.37	1.40E-110	307
			MP0029c	38	glycosyltransferase wlanB	*C.jejuni* strain 81116	Q9K5D1	100	3.40E-17	38
2h03.q	4	718	MP0030	239	cj0262c chemotaxis	11168	Q9PIN3	55.46	3.60E-45	238

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
1b09.q	4	1934	MP0031c	264	phosphodiesterase	*Bradyrhizobium japonicum*	Q89MQ1	40.9	3.50E-34	264
			MP0032c	202	HAD hydrolase	*Caulobacter crescentus*	Q9Q7S7	28.19	2.30E-06	188
			MP0033c	156	ABC transporter	*Brucella suis*	Q8FUP0	36.36	4.50E-09	165
2f12.q	4	1522	MP0034	461	0-acetylation protein	*C.jejuni* strain 11828	Q9ALT7	100	1.90E-186	459
3e06.q	5	1419	MP0035	469	cj1614 ChuA	11168	Q9PM61	91.19	2.10E-174	477
4f03.q	6	1431	MP0036	317	arylsulfatase	*C.jejuni* strain 81-176	Q46098	99.68	3.60E-129	317
			MP0037	130	cj0872 DsbA	11168	Q9PP57	96.15	6.30E-48	130
1a12.p	6	1908	MP0038c	496	ABC transporter (disrupted)	*Photorhabdus luminescens*	CAE14106	47.58	5.60E-80	496
			MP0039c	81	di-tripeptide transporter	*Yersinia pseudotuberculosis*	Q669J3	44.73	2.20E-09	76
3b03.q	6	1733	MP0040c	552	cj1334 hypothetical	*C.jejuni* strain 81-176	Q7X519	76.71	1.40E-136	481
5c06.p	6	1714	MP0041	428	WbyH (o-antigen)	*Yersinia pseudotuberculosis*	Q9RCB8	43.88	1.10E-65	417
			MP0042c	146	AscF reductase	*Yersinia pseudotuberculosis*	Q57103	32.79	1.10E-07	125
1h04.q	7	2681	MP0043	225	EpsS epimerase	*Methylobacillus*	Q83VQ2	56.05	2.00E-47	223
			MP0044	384	Glf galactopyranose mutase	*Helicobacter hepaticus*	Q7VJP0	53.48	7.50E-74	359
			MP0045	291	hypothetical	*C.jejuni* strain 11828	Q9ALS8	28.04	3.00E-09	296
3d04.q	8	1528	MP0046c	508	adhesin	*Chromobacterium violaceum*	AAQ59146	24.77	5.00E-03	440
2g01.p	8	1953	MP0047	167	hypothetical	*Shewanella oneidensis*	Q8E9K9	26.61	3.60E-05	139
			MP0048	169	type I RM	*Archaeoglobus fulgidus*	O28563	45.94	3.40E-13	111
			MP0049	226	type I RM	*Wolinella succinogenes*	CAE10680	32.57	1.30E-07	221
1h01.q	8	1192	MP0050	381	cytochrome c	*Shewanella oneidensis*	Q8EJI6	54	5.70E-71	400
3d07.q	8	1703	MP0051	116	hypothetical (los locus)	*C.jejuni* strain 11828	Q9ALT0	95.69	7.40E-35	116
			MP0052c	361	aminotransferase	*C.jejuni* strain 11828	Q9ALS9	98.6	4.40E-139	358
			MP0053c	77	membrane protein	*C.jejuni* strain tgh9011	Q6EB21	84.5	2.10E-20	71
3e11.p	8	1247	MP0054	375	weak match to hemolysin	*Xanthomonas axonopodis*	Q8PHP1	23.89	5.30E-02	318
5h04.p	10	1763	MP0055c	69	iron binding protein	11168	Q7AR79	79.7	6.50E-19	69
			MP0056c	220	hypothetical	*Helicobacter hepaticus*	Q7VK87	34.32	6.70E-16	201
			MP0057c	206	hypothetical	*Helicobacter hepaticus*	Q7VK87	36.22	3.10E-19	196
5d03.p	8	1526	MP0058	432	UGDH glucose dehydrogenase	*Agrobacterium tumefaciens*	Q8U8E3	48.84	4.10E-78	434
			MP0059	34	UDP-glucose 4-epimerase	*Fusobacterium nucleatum*	Q8RGC6	67.64	1.50E-05	34
3c05.q	9	1729	MP0060	183	ribosomal protein	*Vibrio vulnificus*	Q8DF32	32.96	1.90E-06	179
			MP0061c	115	no matches					
			MP0062c	209	putative phage repressor protein	Bacteriophage phi ETA	Q9G039	28.89	5.20E-05	180
1b10.q	10	1565	MP0063	45	cj1337 hypothetical	*C.jejuni* strain 81-176	Q7X517	100	2.40E-14	45
			MP0064c	464	FlaB	*C.jejuni* strain 81116	Q9RF25	100	3.90E-144	462

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
3b09.p	10	1924	MP0065c	81	no matches					
			MP0066c	559	hypothetical	*Fusobacterium nucleatum*	Q8REK3	23.71	1.50E-04	485
5c01.p	11	1646	MP0067c	85	hypothetical	*Wolinella succinogenes*	CAE10494	38.09	2.80E-06	84
			MP0068c	288	DmsC type gene (MraY)	*Wolinella succinogenes*	CAE10493	42.5	8.10E-40	287
			MP0069c	179	oxidoreductase FdhB	*Wolinella succinogenes*	CAE10492	62.77	1.50E-43	180
2g03.q	11	2942	MP0070c	66	RloA	*C.jejuni* strain Rm1551 & rm1850	Q8G8E7	100	1.90E-22	66
			MP0071c	769	HsdR	*C.jejuni* strain 81116	Q8RIX1	100	0	769
			MP0072	71	cj1548c dehydrogenase	11168	Q9PMC1	100	3.40E-30	71
4h06.p	13	1792	MP0073	110	cj0123c	11168	Q9PJ04	90.9	3.70E-36	110
			MP0074c	446	hypothetical	*Plasmodium falciparum*	Q8IHQ0	19.2	0.012	453
3a05.q	12	1401	MP0075	39	periplasmic protein cj0770c	11168	Q9PPE9	100	1.50E-05	22
			MP0076	149	hypothetical reP	*Treponema denticola*	Q9AQF2	39.59	8.00E-14	149
			MP0077	60	hypothetical TnpV	*Clostridium difficile*	O05416	46.42	6.00E-06	56
3e01.p	14	1779	MP0078	146	glucose epimerase	*Pyrococcus furiosus*	Q8U170	34.09	1.70E-07	132
			MP0079	376	glucose dehydrogenase	*Pyrococcus abyssi*	Q9UZI8	38.33	1.50E-42	373
4g01.p	15	1955	MP0080c	85	RlfA	Bacteriophage P1	Q71TB8	44.57	2.10E-07	85
			MP0081c	552	type I RM	*Wolinella succinogenes*	CAE10680	70.27	6.60E-149	555
1g05.q	15	2785	MP0082	238	cj0414 oxidoreductase	11168	Q9PI91	44.03	3.20E-34	243
			MP0083	571	cj0415 oxidoreductase (disrupted)	11168	Q9PI90	57.14	7.70E-131	574
2c11.p	15	3856	MP0084	67	hypothetical	*C. jejuni* strain rm1221	Q8RN32	100	8.20E-22	65
			MP0085	149	hypothetical	*C.jejuni* strain rm1221	Q8RN33	97.84	2.20E-51	139
			MP0086	251	decarboxylase pcac	*Methanosarcina acetivorans*	Q8TTM1	42.57	1.30E-37	249
			MP0087c	496	HsdM	*C.jejuni* strain rm2227	Q8RN18	96.77	2.60E-181	496
			MP0088c	198	HsdS	*C.jejuni* strain rm1163 & rm1508	Q8G8A9	99.48	6.40E-74	194
4e08.q	16	1909	MP0089	164	cytochrome C	*Shewanella oneidensis*	Q8EJI6	49.08	1.20E-26	163
			MP0090	457	hpothetical/ possible cyt C	*Shewanella oneidensis*	Q8EJI5	39.43	1.70E-12	142
3h01.q	16	2537	MP0091c	118	permease protein	*Rhodopseudomonas palustris*	Q6NDI1	43.75	7.00E-15	112
			MP0092c	285	ABC transporter permease	*Rhizobium loti*	Q98JZ2	48.54	1.00E-49	274
			MP0093c	372	ABC transporter	*Agrobacterium tumefaciens*	Q8UIA7	45.43	1.10E-48	372
			MP0094c	41	cj1687	11168	Q9PLZ1	100	1.50E-16	41
3d08.p	18	2768	MP0095	153	Cj1431c hypothetical	11168	Q9PMM6	28.32	2.90E-04	173
			MP0096	264	DdhA (los)	*Yersinia enterocolitica*	Q56860	59.47	4.80E-60	264
			MP0097	452	glucose dehydratase	*Fusobacterium nucleatum*	EAA24619	60.67	6.00E-109	445
			MP0098	50	no matches					

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
1d11.p	18	2713	MP0099c	124	no matches					
			MP0100c	57	hypothetical from transposon	*Enterococcus faecalis*	Q56396	66.66	4.10E-14	57
			MP0101c	582	TetO	*C.jejuni* plasmid pCjA13	Q84FM6	99.48	1.40E-202	577
1g06.p	21	2261	MP0102c	55	cj1584c periplasmic	11168	Q9PM91	83.33	9.80E-16	54
			MP0103	600	DmsA	*Wolinella succinogenes*	CAE10491	61.69	6.00E-155	603
4e01.q	35	4308	MP0104c	881	TraG pseudogene	*Vibrio vulnificus*	BAC97743	21.04	4.10E-11	879
			MP0105	51	cj0937 membrane protein	11168	Q9PNZ6	100	8.40E-20	51
4d08.p	42	4924	MP0106	813	cytochrome C	*Wolinella succinogenes*	CAE11153	54.26	5.90E-172	820
			MP0107c	556	GGT	H. pylori J99	Q9ZK95	67.74	9.40E-136	558
			MP0108	182	cj0031 RM	11168	Q9PJ80	53.8	3.70E-30	184
1c08.p	2	715	MP0109	197	hypothetical (disrupted)	*Helicobacter hepaticus*	Q7VK87	39.28	1.40E-19	196
2g10.p	2	1010	MP0110c	202	Ansa cj0029	11168	Q9PJ82	86.13	8.00E-59	202
			MP0111c	87	RecJ cj0028	11168	Q9PJ83	96.55	1.20E-32	87
1f03.p	2	1035	MP0112c	166	HsdS	c.j strain rm1049, rm1861, 81116	Q8RJ16	100	2.70E-64	166
			MP0113c	179	RloB	c.j strain rm1049, rm1861, 81116	Q8RIW9	100	1.70E-66	179
1b04.q	2	760	MP0114c	170	ABC transporter (disrupted)	*Rhizobium loti*	Q98JZ4	36.25	1.20E-13	160
			MP0115c	60	ABC transporter permease	*Rhizobium loti*	Q98JZ3	56.66	1.10E-10	60
2d06.q	2	593			no predicted CDSs					
2d03.p	2	762	MP0116c	206	hypothetical	*Helicobacter hepaticus*	Q7VIF8	51.33	6.50E-33	187
2e03.p	2	824	MP0117c	70	hypothetical	*Wolinella succinogenes*	Q7MQN4	39.34	2.70E-04	61
			MP0118c	187	formate dehydrogenase	*Vibrio cholerae*	Q9KRX2	28	9.00E-04	
3a03.p	1	646	MP0119c	151	hypothetical	*S. typhimurium* phage ST64B	Q8HAA0	30.87	1.30E-06	149
2b12.p	1	591	MP0120c	118	WlanB glycosyltransferase	*C.jejuni* strain 81116	Q9K5D1	100	1.00E-41	118
			MP0121c	78	WlanA (lipid A sysnthesis cluster)	*C.jejuni* strain 81116	Q9K5D2	100	1.00E-32	78
5b01.p	1	585	MP0122	93	cj1305c hypothetical	11168	Q9PMZ4	57.81	8.00E-12	64
2c05.p	1	425	MP0123c	104	no matches					
1e03.q	1	358	MP0124	119	hypothetical	*Pasteurella multocida*	Q9CKR7	39.02	1.10E-03	82
2a08.q	1	471	MP0125c	99	NADH dehydrogenase	*Strongyloides stercoralis*	CAD90562	36.45	3.30E-03	96
4d09.p	1	814	MP0126c	49	no matches					
			MP0127c	222	hypothetical	*Plasmodium yoelii yoelii*	EAA18980	24.27	0.0093	173
2e10.p	1	805	MP0128	20	transferase cj1050c	11168	Q9PNP1	95	6.60E-06	20
			MP0129	199	membrane protein cj1049c	11168	Q9PNP2	87.94	2.90E-65	199
			MP0130	48	Dape or Cj1048c	11168	Q9PNP3	100	1.70E-17	47
2g02.q	2	664	MP0131c	220	pgi cj1535c pseudogene	11168	G6PI_CAMJE	82.27	4.00E-64	220

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
3b01.p	4	1024	MP0132	89	EspC	*Escherichia coli*	P77070	43.18	4.40E-11	88
			MP0133	225	Argc cj0224	11168 & TGH9011	ARGC_CAMJ	98.22	6.80E-86	225
3e02.q	1	733	MP0134	238	membrane protein cj0692c	11168	Q9PPL5	85.71	4.00E-54	238
4h07.p	2	932	MP0135	154	cj1295 hypothetical	11168	Q9PN04	88.88	6.40E-53	153
			MP0136	157	cj1296 & cj1297	11168	Q9PN03	79.04	1.30E-31	105
							Q9PN02	56.75	2.10E-03	37
5h05.p	4	1775	MP0137	53	hydrophobic protein	11168	Q9PLV0	97.5	4.40E-13	40
			MP0138	127	cj1724c hypothetical	11168	Q9PLV4	100	8.80E-51	127
			MP0139	214	cj1721c outer membrane protein	11168	Q9PLV7	65.42	2.10E-55	214
			MP0140c	106	cj1720 hypothetical	11168	Q9PLV8	100	8.80E-38	107
4c04.p	3	1057	MP0141	202	Cj0967 periplasmic protein	11168	Q9PNW9	96.42	1.90E-31	112
			MP0142	115	hemagglutinin-related protein/ adhesin	*Ralstonia solanacearum*	Q8XQ42	36.28	6.00E-05	113
2g07.q	8	1844	MP0143	470	cj0970, cj0971, cj0972, cj0973	11168	Q9PNW7	95.31	4.90E-34	128
							Q9PNW8	85.85	1.20E-21	99
							Q9PNW6	55.78	2.30E-09	95
							Q9PNW5	93.54	1.60E-03	31
			MP0144	65	Cj0975	11168	Q7AR82	97.29	6.90E-09	37
4f07.p	2	730	MP0145c	31	hypothetical	11168	Q9PNW8	80	7.50E-05	30
			MP0146c	151	ceub uptake permease cj1352	11168	Q9PMU7	98.01	1.10E-50	151
			MP0147c	35	pldA	11168	Q9PMU8	97.14	7.90E-14	35
2f07.q	4	1193	MP0148	89	haemoglobin protease	*Escherichia coli*	Q8FKM0	45.97	6.10E-07	87
			MP0149	162	no matches					
			MP0150	89	no matches					
2b05.p	4	1063	MP0151c	255	dicarboxylate transporter	*Vibrio vulnificus*	BAC95008	35.77	1.00E-20	232
			MP0152	31	hypothetical Cj1523c	11168	Q9PME1	96.77	5.00E-11	31
4d12.p	1	788	MP0153c	97	cj0865 oxidoreductase DsbB	11168	DSBI_CAMJE	95.78	2.40E-39	95
			MP0154c	167	Cj0864 periplasmic protein	11168	Q9PP59	91.76	7.30E-23	85
4e06.p	10	2078	MP0155c	423	Bll0816 hypothetical	*Bradyrhizobium japonicum*	Q89W77	33.48	1.40E-39	427
			MP0156c	266	cj1394 fumarate lyase	11168	Q9PMR1	95.11	1.70E-95	266

APPENDIX 5

Predicted CDSs for sequenced pUC library clones of strain 40671

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
3e04.q	2	606	4P0001	86	TraH	*Comamonas acidovorans* pUO1	BAC82023	41.79	1.60E-08	67
			4P0002	110	no matches					
3d01.p	2	1132	4P0003	115	cj0138	11168	Q9PIY9	90.38	2.00E-30	104
			4P0004	95	no matches					
			4P0005	111	no matches					
1d01.p	2	1060	4P0006	345	mcp-type signal transduction	11168	Q9PMF7	40	5.60E-47	345
3f04.p	4	2259	4P0007c	570	cj1440c sugar transferase	11168	Q9PML7	49.64	2.00E-46	423
			4P0008	139	cj1421c sugar transferase	11168	Q9PMN6	84.17	7.10E-43	139
1b09.p	4	988	4P0009c	342	MagB10	*Actinobacillus actinomycetemcomita*	Q9F252	40.35	2.20E-34	342
3c10.p	4	1147	4P0010c	333	sialic acid biosynthesis	*C. jejuni* strain 43446	Q9L9Q4	99.09	4.20E-123	332
3c08.q	4	1341	4P0011	315	no matches					
1c07.p	4	1052	4P0012c	350	FlgE	*C.jejuni* strain lio7	O86148	99.41	6.90E-122	344
3f07.p	4	1011	4P0013c	336	MagB12	*Actinobacillus actinomycetemcomita*	Q9F254	40.95	8.10E-44	337
2d04.p	4	1442	4P0014c	95	no matches					
1d05.p	4	1179	4P0015	101	hypothetical cj1724c	11168	Q9PLV4	100	1.40E-39	101
			4P0016	213	cj1721c membrane protein	11168	Q9PLV7	64.01	5.00E-54	214
1h08.q	4	1079	4P0017	140	hypothetical	*Helicobacter hepaticus*	Q7VGU0	35.43	4.80E-08	127
			4P0018	97	no matches					
3d10.p	4	1236	4P0019	384	hypothetical cj1341c	11168	Q9PMV8	48.55	3.40E-62	381
1d03.p	4	958	4P0020	143	hypothetical	*Wolinella succinogenes*	Q7MQT2	32.37	3.80E-09	139
			4P0021	174	hypothetical jhp0950	*H. pylori* J99	Q9ZKI5	46.7	3.20E-20	167
3g08.p	4	1498	4P0022	494	cj1431c hypothetical	11168	Q9PMM6	28.14	2.20E-22	430
1e07.p	4	1047	4P0023	87	VirB2	*Escherichia coli*	Q91UX6	35.36	9.90E-07	82
			4P0024	147	TriC	*Yersinia enterocolitica*	CAD58564	39.16	1.20E-12	143
1f06.p	4	1038	4P0025	309	FlaA	*C. jejuni* serotype 0:19	Q99QL6	100	2.70E-89	309
1b06.q	6	2049	4P0026	111	hypothetical	*Pseudomonas syringae*	Q889N9	58.76	6.50E-20	97
			4P0027	241	lipopolysaccharide biosynthesis	*Pseudomonas syringae*	Q889P3	40.49	2.40E-23	242
			4P0028	132	hypothetical	*Actinobacillus suis*	Q84CG6	57.93	9.00E-27	126
			4P0029	142	hypothetical	*Actinobacillus suis*	Q84CG5	40.55	8.60E-16	143
3g02.p	6	1643	4P0030c	521	hypothetical	*Actinobacillus suis*	Q84CG8	26.03	2.40E-11	338

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
2d09.p	6	1199	4P0031c	338	no matches					
			4P0032c	41	cj0121	11168	Q9PJ06	95.12	1.40E-14	41
3c01.q	7	1537	4P0033	243	hypothetical	*Actinobacillus suis*	Q84CG7	53.3	4.70E-42	242
			4P0034	265	c-methyltransferase	*Bordetella bronchiseptica*	Q7WR30	29.16	1.40E-06	216
1b12.q	8	2219	4P0035	633	hypothetical	*Chromobacterium violaceum*	Q7NTJ9	51.42	7.70E-120	634
			4P0036c	53	cj1161c ATPase	11168	Q9PND4	70.21	3.80E-11	47
1g01.p	8	1564	4P0037c	122	no matches					
			4P0038c	89	no matches					
2b07.p	8	1572	4P0039	446	oxidoreductase	*Bacteroides thetaiotaomicron*	Q8A7I2	44.61	7.70E-73	455
			4P0040c	82	cj1069	11168	Q9PNM5	83.54	2.60E-22	79
3f05.q	8	2881	4P0041c	295	virB9-like protein	*C.jejuni* plasmid pCjA13	Q847A7	97.28	3.50E-112	295
			4P0042c	220	virB8-like protein	*C.jejuni* plasmid pCjA13	Q847A8	100	1.10E-77	220
			4P0043c	333	magb06	*Actinobacillus actinomycetemcomita*	Q9F248	26.33	1.60E-15	319
1g12.p	9	2582	4P0044	74	hypothetical	*Bacteroides thetaiotaomicron*	Q8A5B1	47.22	5.80E-06	72
			4P0045	167	hypothetical	*Shewanella oneidensis*	Q8E9K9	26.61	2.60E-05	139
			4P0046	480	type I RM	*Archaeoglobus fulgidus*	O28563	38.63	6.70E-16	176
			4P0047	63	cj1047c	11168	Q9PNP4	88.88	1.00E-18	63
			4P0048	33	cj1046c Moeb	11168	Q9PNP5	93.93	3.10E-14	33
3e05.q	9	1813	4P0049c	276	acetyltransferase	*C.jejuni* strain 43432	Q9F0M5	98.91	3.90E-106	277
			4P0050c	221	NeuA1	*C.jejuni* strain 81-176, 43456, 4344	Q933W2	98.64	3.90E-82	221
1e06.p	10	2212	4P0051c	116	hydrolase	*Pseudomonas syringae*	Q889P1	62.28	8.20E-26	114
			4P0052c	211	hypothetical	*Pseudomonas syringae*	Q889P2	40.67	6.10E-29	209
			4P0053	295	c-methyltransferase	*Leptospira interrogans*	Q8F5S5	25	2.00E-09	276
3g05.p	10	3379	4P0054	655	MagB03	*Actinobacillus actinomycetemcomita*	Q9F245	44.82	9.20E-103	647
			4P0055	188	hypothetical	*C.jejuni* plasmid pCjA13	Q847A4	44.8	1.30E-14	183
			4P0056	221	no matches					
			4P0057	45	SSB	*C.jejuni* plasmid pVir	Q8GJE0	48.88	7.60E-06	45
1d02.q	10	1988	4P0058	402	cj1421c sugar transferase	11168	Q9PMN6	69.38	4.10E-86	343
			4P0059	228	Cst-I (disrupted)	*C.jejuni* strain 0h4384	Q9RGF1	57.85	7.60E-39	242
3a10.q	10	3859	4P0060c	830	hypothetical jhp1285	*H. pylori* J99	Q9ZJM1	30.28	5.60E-43	885
			4P0061c	413	no matches					
3f10.p	11	1692	4P0062c	539	Cj1334 hypothetical	*C.jejuni* strain 81-176	Q7X519	94.83	2.00E-168	465
1a10.p	12	3065	4P0063c	331	DmhA	*Yersinia pseudotuberculosis*	Q8G8E4	78.46	9.30E-97	325
			4P0064c	351	Fcl cj1428c	11168	Q9PMM9	59.07	2.90E-75	347
			4P0065c	181	cj1430c sugar epimerase	11168	Q9PMM7	80.66	4.40E-59	181
			4P0066c	126	cj1421c sugar transferase	11168	Q9PMN6	37.39	7.60E-05	115

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
3b11.p	15	2239	4P0067c	60	phage hypothetical	Bacteriophage P1	Q9XJP4	46.42	1.10E-03	56
			4P0068c	683	type I RM	*Wolinella succinogenes*	Q7M8H9	71.04	8.20E-168	632
1g09.q	25	3249	4P0069	501	FlaB	*Campylobacter coli*	P18245	91.18	9.80E-139	499
			4P0070c	569	cj1337 hypothetical	11168	Q9PMW2	61.13	1.90E-136	566
1h01.p	2	1191	4P0071	382	hypothetical jhp0928	*H. pylori* J99	Q9ZKK7	42.96	8.80E-50	384
1b02.q	2	725	4P0072c	152	hypothetical ydaa	*Wolinella succinogenes*	Q7MQX6	39.43	6.30E-14	142
			4P0073c	89	Vap2	*Riemerella anatipestifer*	O85171	35.59	1.30E-01	59
1d06.p	2	1050	4P0074	246	hypothetical	*H. pylori* 26695	O25892	45	4.00E-30	220
3f11.p	2	1346	4P0075c	436	putative DNA methylase	*Sinorhizobium meliloti* phage PBC5	Q8W6K4	44.87	7.20E-39	312
1d10.p	2	1151	4P0076	330	VirB11-like protein (disrupted)	*C.jejuni* plasmid pCjA13	Q847A5	99.69	4.70E-117	329
2c02.p	2	633	4P0077c	199	TraE (virB8)	*Escherichia coli*	Q60215	36.54	6.90E-14	197
3e03.p	2	744	4P0078c	247	ABC transporter	*Photorhabdus luminescens*	Q7N5W6	52.67	1.90E-55	243
1c06.p	2	1352	4P0079c	304	no matches					
			4P0080c	97	no matches					
			4P0081c	39	hypothetical	*Wolinella succinogenes*	Q7MQT0	50	0.00017	36
2b11.p	2	791	4P0082c	262	hypothetical	*H. pylori* J99	Q9ZKK7	48.47	4.00E-43	262
2c10.p	1	800	4P0083c	201	hypothetical	*Clostridium perfringens*	Q93M99	26.15	1.20E-02	195
			4P0084c	78	no matches					
1a12.q	1	772	4P0085c	205	ATPase 6	*Leishmania tarentolae*	Q33561	22.87	1.10E-02	188
3a12.p	1	819	4P0086	32	no matches					32
			4P0087	67	hypothetical	*H. pylori* J99	Q9ZKH5	66.66	1.40E-12	57
			4P0088	131	no matches					131
3a12.q	1	769	4P0089	252	TrbM-like protein	*Haemophilus aegyptius*	Q8VRC6	37.17	4.50E-11	191
1b05.p	1	696	4P0090c	230	type II RM (cj0032)	11168	Q9PJ79	60.08	2.40E-43	228
1a06.p	1	827	4P0091c	225	CfrA cj0755	11168	Q9PPG3	88	3.80E-78	225
1a05.q	1	695	4P0092	231	sialic acid biosynthesis	*C.jejuni* strain atcc43432	Q9F0M7	99.56	5.20E-82	231
1g03.p	1	810	4P0093c	87	no matches					
			4P0094c	194	no matches					
2a08.p	1	229	4P0095	70	acetyltransferase	*C.jejuni* strain 43446	Q9K379	38.57	1	70
2e08.p	1	728	4P0096c	151	no matches					
			4P0097c	90	no matches					
1g08.p	1	847	4P0098c	281	cj1305c hypothetical protein	11168	Q9PMZ4	75.97	2.10E-80	283
3d03.p	1	770	4P0099c	158	hypothetical cj1337	11168	Q9PMW2	70.77	6.30E-39	154
			4P0100c	45	efflux protein cj1174	11168	Q9PNC1	100	1.40E-15	45
1g10.p	1	151			no predicted CDSs					
2c04.p	1	122			no predicted CDSs					
2b06.p	1	90			no predicted CDSs					

APPENDIX 6

Predicted CDSs for sequenced pUC library clones of strain 52472

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
7h02.q	1	515	5P0001c	88	hypothetical	RM1221	Q5HWU3	98.33	1.10E-19	60
6e05.q	2	726	5P0002	241	di-/tripeptide transporter	RM1221	Q5HVB7_CA	96.9	6.30E-84	226
2f05.p	2	1099	5P0003c	316	type I RM	*Staphylococcus aureus*	Q6GD64_ST	35.5	2.90E-31	307
6d05.q	2	671	5P0004	223	cj0929 Pepa	11168	AMPA_CAMJ	97.76	5.40E-80	223
2g09.p	2	1098	5P0005	139	cj0807 oxidoreductase	11168	Q9PPB3_CA	98.51	1.10E-47	135
			5P0006c	104	cj0808c hydrophobic hypothetical	11168	Q9PPB2_CA	77.22	3.70E-30	101
			5P0007c	83	cj0809c hydrolase	11168	Q9PPB1_CA	90.36	4.30E-31	83
5b12.p	2	1017	5P0008	144	hypothetical	RM1221	Q5HTG8_CA	71.05	1.20E-04	38
			5P0009	78	hypothetical	RM1221	Q5HTH0_CA	92.3	2.90E-24	78
6c07.q	2	625	5P0010c	197	hypothetical	RM1221	Q5HW50_CA	98.96	1.50E-63	126
6f10.p	2	918	5P0011	300	cj0765c hiss	11168	SYH_CAMJE	87.29	1.30E-102	299
3g06.p	2	799	5P0012c	265	base plate assembly	RM1221	Q5HWS9_CA	98.11	3.70E-88	265
3d03.p	2	1137	5P0013	235	type II RM	RM1221	Q5HXC7_CA	73.39	2.20E-63	233
			5P0014	144	hypothetical	*H. pylori*	O26049_HEL	56.55	1.20E-21	145
3a04.p	2	763	5P0015	192	TrbM (cpp45)	*C. coli*	Q69BE2_CA	71.74	1.70E-52	184
			5P0016	60	hypothetical cpp46	*C. jejuni* pTet	Q69B91_CA	98.3	4.50E-17	
6a09.p	2	906	5P0017c	71	hypothetical	RM1221	Q5HTH6_CA	97.02	7.30E-22	67
			5P0018c	74	hypothetical	RM1221	Q5HTH5_CA	93.24	6.20E-27	74
			5P0019c	134	hypothetical	RM1221	Q5HTH4_CA	98.51	1.20E-50	134
4e02.q	2	543	5P0020c	179	cj1218c Riba	11168	Q9PN77_CA	95.5	2.40E-60	178
5f10.q	2	675	5P0021	224	cj0411 ATP/GTP binding protein	11168	Q9PI94_CAM	97.3	7.90E-68	223
4e01.p	2	764	5P0022	104	cj0578c Tatc sec-independent transloca	11168	TATC_CAMJ	97.08	1.50E-39	103
			5P0023	146	cj0577c QueA	11168	QUEA_CAMJ	97.26	1.40E-53	146
5e08.p	2	889	5P0024	243	HsdM	*C.jejuni* strain rm 1170	Q8RN38_CA	100	2.90E-90	242
4d12.p	2	974	5P0025	55	hypothetical cpp2	*C. jejuni* pTet	Q69BD4_CA	97.73	4.90E-17	44
			5P0026	117	hypothetical cpp8	*C. jejuni* pTet	Q69BC8_CA	99.14	1.00E-42	116
			5P0027	132	hypothetical cpp9	*C. jejuni* pTet	Q69BC7_CA	100	9.10E-47	132
3a03.q	3	1058	5P0028c	237	cj0812 Thrc	11168	Q9PPA8_CA	78.48	5.70E-70	237
			5P0029c	118	cj0811 Lpxk tetraacyldisaccharide kinas	11168	LPXK_CAMJ	84.21	4.90E-37	114
6c11.p	3	1062	5P0030c	323	hypothetical	RM1221	Q5HXA9_CA	99.69	1.90E-106	323
4d02.q	3	979	5P0031	200	hypothetical cpp46	*C. jejuni* pTet	Q69B91_CA	99	8.40E-67	200
			5P0032	102	hypothetical cpp47	*C. jejuni* pTet	Q69B90_CA	98	1.20E-34	102

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
8a03.p	4	815	5P0033c	80	hypothetical	RM1221	Q5HTH3_CA	98.63	1.10E-24	73
			5P0034c	106	hypothetical	RM1221	Q5HTH2_CA	99.06	7.20E-35	106
6a05.p	4	1794	5P0035c	124	MloA	*C.jejuni* strain rm 1852	Q8RN19_CA	100	1.40E-38	124
			5P0036c	395	HsdS	*C.jejuni* strain rm 1170	Q8RN40_CA	100	6.00E-152	395
5d07.p	4	1131	5P0037	70	type III RM	*H. pylori*	O25314_HEL	56.52	1.30E-10	69
			5P0038	304	DNA methyltransferase	*H. pylori*	O25315_HEL	61.98	8.60E-48	242
5g12.q	4	788	5P0039	74	cj0762c aspb aspartate aminotransferas	11168	Q9PPF7_CA	91.89	2.90E-24	74
			5P0040	185	hypothetical	*Nitrosomonas europea*	Q82T36_NIT	33.15	9.90E-10	184
5h07.q	4	1432	5P0041	324	hypothetical	RM1221	Q5HWQ1_CA	99.68	1.90E-116	320
			5P0042	89	hypothetical	RM1221	Q5HWQ2_CA	100	3.20E-37	89
			5P0043	57	hypothetical	RM1221	Q5HWQ3_CA	98.11	1.00E-16	53
8h04.p	4	997	5P0044	331	periplasmic protein cj0737	11168	Q7AR90_CA	38.02	2.90E-22	334
8c09.p	4	1085	5P0045	102	virion morphogenesis protein	RM1221	Q5HWU1_CA	96.94	1.30E-36	98
			5P0046	212	dam DNA adenine methylase	RM1221	Q5HWU2_CA	96.49	8.00E-63	171
7a07.p	4	1280	5P0047c	239	cj0813 KdsB	11168	Q9PPA7_CA	82.85	8.20E-75	239
			5P0048c	157	cj0812 Thrc	11168	Q9PPA8_CA	75.48	1.00E-41	155
4g03.q	4	993	5P0049c	158	phage tail protein	RM1221	Q5HWTo_CA	96.81	3.00E-57	157
			5P0050c	170	base plate assembly	RM1221	Q5HWS9_CA	98.82	3.10E-55	170
5c07.q	6	1056	5P0051c	206	cj0293 Sure	11168	SURE_CAMJ	93.78	6.60E-67	193
			5P0052	98	transporter	*Escherichia coli*	Q8FAP1_EC	51.06	1.30E-14	94
6b10.q	5	949	5P0053c	298	transport system permease	*Escherichia coli*	Q8X8T6_EC	52.03	9.00E-59	296
5f02.p	5	1246	5P0054	65	di-/tripeptide transporter	RM1221	Q5HVB7_CA	60.66	2.40E-11	61
			5P0055	282	di-/tripeptide transporter	RM1221	Q5HVB7_CA	99.65	5.40E-115	282
5f11.q	5	2192	5P0056c	215	hypothetical					
			5P0057c	224	signal peptidase I	RM1221	Q5HTF9_CA	33.18	1.40E-17	223
			5P0058	177	dna transition protein a	RM1221	Q5HWP2_CA	87.82	6.20E-43	156
8b01.p	6	875	5P0059	221	hypothetical					
7e11.p	6	1591	5P0060	287	HsdR	*C.jejuni* strain rm 1170	Q8RN42_CA	99.29	1.50E-94	283
			5P0061	238	RloF	*C.jejuni* strain rm 1170	Q8RN41_CA	100	7.10E-89	238
5d04.q	6	1564	5P0062	248	hypothetical	RM1221	Q5HWR6_CA	98.79	1.40E-90	248
			5P0063	152	hypothetical	RM1221	Q5HWR5_CA	98.49	2.80E-45	132
			5P0064	83	hypothetical	RM1221	Q5HWR4_CA	100	3.90E-28	82
8c04.p	6	2259	5P0065	752	type III RM r protein	*H. pylori*	O25314_HEL	52.78	7.30E-58	773
5h03.q	6	1528	5P0066	470	VacA autotransporter domain	*H. pylori*	Q9ZHT4_Va	23.05	9.20E-06	192
8g05.q	6	1619	5P0067c	94	DNA binding protein	RM1221	Q5HWQ7_CA	97.87	5.30E-31	94
			5P0068	223	DNS extracellular deoxyribonuclease	RM1221	Q5HWQ6_CA	99.55	2.80E-91	223
			5P0069c	91	hypothetical	RM1221	Q5HWQ5_CA	100	2.30E-31	91
			5P0070c	84	hypothetical	RM1221	Q5HWQ4_CA	100	2.80E-33	84

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
5h08.p	6	1272	5P0071c	207	hypothetical	*Salmonella typhi*	Q8Z3Y2_SAL	41.44	8.10E-19	152
7g06.p	6	1220	5P0072	382	hypothetical	RM1221	Q5HWR2_CA	99.19	6.80E-134	369
1e12.q	7	1723	5P0073c	493	type I RM	*Methanosarcina mazei*	Q8PSV0_ME	45.52	1.80E-75	503
6f05.q	7	1270	5P0074	423	cj0629 possible lipoprotein	11168	Q9PHN8_CA	76.21	5.60E-46	269
4e07.q	7	2026	5P0075	77	cj0303c ModA	11168	Q9PIJ6_CAM	81.81	5.60E-22	77
			5P0076	133	cj0302c	11168	Q9PIJ7_CAM	64.61	9.80E-29	130
			5P0077	222	cj0301c ModB	11168	Q9PIJ8_CAM	84.68	3.40E-70	222
			5P0078	240	cj0300c ModC	11168	Q9PIJ9_CAM	78.33	1.80E-59	240
5e04.p	8	1871	5P0079c	73	cj0298c PanB	11168	PANB_CAMJ	98.59	1.10E-22	71
			5P0080c	236	hypothetical	*Helicobacter hepaticus*	Q7VI60_HEL	44.29	3.30E-33	228
			5P0081c	104	hypothetical					
8g09.p	8	2154	5P0082	93	DnaK	11168	DNAK_CAMJ	94.318	1.10E-20	88
			5P0083	579	HsdM (disrupted)	*Vibrio cholerae*	Q9KR74_VIB	47.03	1.80E-68	608
5a07.q	8	2062	5P0084	122	cj1343c putative periplasmic protein	11168	Q9PMV6_CA	98.36	2.20E-42	122
			5P0085	416	cj1342c hypothetical	11168	Q9PMV7_CA	60.24	7.70E-97	415
			5P0086	144	cj1341c hypothetical	11168	Q9PMV8_CA	94.44	2.10E-49	144
4h09.p	8	1368	5P0087	338	DNA methyltransferase	*H. pylori*	O25315_HEL	46	3.10E-39	313
			5P0088	112	serine-threonine protein kinase	*Debaryomyces hansenii*	Q6BHW6_DE	31.13	7.30E-04	106
6c03.q	8	1506	5P0089c	55	Glx2 putative hydrolase	11168	Q9PPB1_CA	78.182	2.90E-15	55
			5P0090	248	cj0810 Nade	11168	NADE_CAMJ	74.07	1.20E-59	243
			5P0091	164	cj0811 Lpxk tetraacyldisaccharide kinas	11168	LPXK_CAMJ	82.31	5.90E-52	164
4g04.p	8	1892	5P0092c	271	Mu-like prophage I protein	RM1221	Q5HWR8_CA	99.26	3.80E-91	271
			5P0093c	144	hypothetical	RM1221	Q5HWR9_CA	100	6.10E-52	144
			5P0094	131	hypothetical	RM1221	Q5HWS1_CA	96.12	7.20E-51	129
6c04.p	8	1547	5P0095	515	Cmgb3/4	*C. jejuni* pTet	Q69BA6_CA	96.89	9.50E-195	515
2g11.q	8	1062	5P0096c	152	hypothetical	RM1221	Q5HVS2_CA	100	1.80E-58	152
			5P0097c	123	hypothetical	RM1221	Q5HVS4_CA	100	2.10E-49	123
8d12.p	9	2441	5P0098	412	prophage muso1 f protein	RM1221	Q5HWR1_CA	100	1.50E-148	412
			5P0099	124	phage tail protein	RM1221	Q5HWR0_CA	100	9.10E-45	124
			5P0100	140	tail protein D	RM1221	Q5HWQ8_CA	99.28	2.80E-49	140
2c11.q	10	1275	5P0101c	63	hypothetical	RM1221	Q5HWS7_CA	98.41	8.70E-23	63
			5P0102c	210	base plate assembly protein V	RM1221	Q5HS6_CAM	98.57	4.20E-75	210
			5P0103c	86	hypothetical	no matches				
2e10.p	10	1854	5P0104c	149	hypothetical	*C.jejuni* strain rm 1221	Q8RN33_CA	97.84	1.10E-50	139
			5P0105c	391	transporter	*C.jejuni* strain rm 1221	Q5HSN2_CA	98.72	1.20E-139	391
7c10.p	11	1615	5P0106	501	hypothetical phage protein	RM1221	Q5HWR3_CA	100	7.20E-174	442
6a01.q	12	1925	5P0107c	87	cpp23	*C. jejuni* pTet	Q69BB4_CA	97.7	4.30E-30	87
			5P0108c	409	cpp22 (TraC like)	*C. jejuni* pTet	Q69BB5_CA	85.92	2.60E-128	412
2f11.q	12	2036	5P0109	298	sialic acid synthase	*C.jejuni* strain oh4384	Q9LAK2_CA	99.66	2.00E-115	298
			5P0110	374	NeuC1	*C.jejuni* strain atcc43456	Q93D03_CA	98.66	7.20E-132	374

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
5a06.p	12	3535	5P0111c	238	Cmgb3/4 (virB4)	*C. coli*	Q69BF6_CA	93.25	7.50E-87	237
			5P0112c	87	Cmgb2 (VirB2)	*C. jejuni* pTet	Q69BA7_CA	90.8	7.20E-27	87
			5P0113	107	cpp29 hypothetical	*C. jejuni* pTet	Q69BA8_CA	99.07	2.20E-41	107
			5P0114	125	virulence-associated protein d	*C. jejuni* pTet	Q69BA9_CA	98.4	1.90E-45	125
			5P0115	204	site-specific recombinase	*C. jejuni* pTet	Q69BB0_CA	99.02	1.40E-69	204
			5P0116c	286	cpp26 hypothetical	*C. jejuni* pTet	Q69BB1_CA	94.38	1.70E-97	285
7d08.q	12	2314	5P0117c	269	cj0021c hypothetical	11168	Q9PJ90_CAM	85.82	1.20E-90	268
			5P0118c	298	cj0022c ribosomal pseudouridine syntha	11168	Q9PJ89_CAM	82.37	2.60E-94	295
			5P0119	130	cj0023 purb	11168	Q9PJ88_CAM	93.7	4.60E-41	127
1d01.q	13	2884	5P0120c	844	cpp14 hypothetical	*C. jejuni* pTet	Q69BC2_CA	99.39	0.00E+00	824
			5P0121c	88	cpp13 hypothetical	*C. coli*	Q69BH3_CA	100	1.10E-29	88
6g02.q	14	2648	5P0122c	120	cj0304c BioC	11168	Q9PIJ5_CAM	74.16	3.40E-33	120
			5P0123c	203	cj0305c hypothetical	11168	Q9PIJ4_CAM	68.47	1.10E-51	203
			5P0124c	380	cj0306c BioF	11168	Q9PIJ3_CAM	75.78	4.70E-111	380
			5P0125	156	cj0307 BioA	11168	Q9PIJ2_CAM	96.15	1.40E-60	156
2e12.p	14	3074	5P0126	198	site-specific DNA-methyltransferase	RM1221	Q5HVW9_CA	90.91	8.40E-70	198
			5P0127	117	hypothetical	RM1221	Q5HTH9_CA	100	1.17E+02	
			5P0128c	391	site-specific recombinase	RM1221	Q5HTI1_CAM	100	1.60E-143	391
			5P0129c	144	hypothetical	no matches				
7h09.p	15	3530	5P0130	309	Cgta-II (disrupted)	*C. jejuni* strain atcc 43449	Q934C5_CA	99.68	5.90E-125	309
			5P0131	245	NeuA1	*C. jejuni* strain atcc 43438	Q93MP7_CA	97.28	6.80E-81	221
			5P0132	277	acetyltransferase (disrupted)	*C. jejuni* strain atcc 43446	Q9L9Q2_CA	97.83	1.10E-103	277
			5P0133c	270	WaaV	*C. jejuni* lio87	Q6T5A5_CA	95.17	2.70E-102	269
			5P0134	109	WaaF	*C. jejuni* strain nctc 11828	Q6TDC6_CA	97.96	2.90E-34	98
3e03.p	15	2697	5P0135	315	cj0259 Pyrc	11168	Q9PIN6_CAM	77.84	2.40E-97	316
			5P0136	576	DNA methyltransferase	RM1221	Q5HWK5_CA	97.24	6.30E-209	579
3e06.p	16	2317	5P0137	211	hypothetical	RM1221	Q5HTE9_CA	99.05	6.30E-67	211
			5P0138	127	hypothetical	RM1221	Q5HTF0_CA	97.64	1.30E-40	127
			5P0139	124	hypothetical	RM1221	Q5HVS5_CA	99.19	5.20E-52	124
			5P0140	294	hypothetical	RM1221	Q5HTF2_CA	99.66	4.50E-96	294
3c11.q	17	3353	5P0141c	51	hypothetical	no matches				
			5P0142c	704	hypothetical	*Helicobacter hepaticus*	Q7VI58_HEL	40.29	6.20E-81	752
			5P0143c	103	hypothetical	no matches				

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
1c03.q	19	3151	5P0144c	90	hypothetical	RM1221	Q5HWP9_CA	96.66	1.40E-27	90
			5P0145c	161	gam protein (phage)	*C. coli*	Q9K5D6_CA	98.75	8.90E-53	160
			5P0146c	112	hypothetical	*C. coli*	Q9K5D7_CA	94.64	5.80E-37	112
			5P0147c	143	hypothetical	no matches				
			5P0148c	242	hypothetical	*Helicobacter hepaticus*	Q7VI56_HEL	48.73	5.60E-36	236
8g04.q	21	2356	5P0149c	130	phage terminase	RM1221	Q5HTC7_CA	98.46	8.90E-40	130
			5P0150c	113	HNH endonuclease domain protein	RM1221	Q5HTC6_CA	92.04	3.90E-41	113
			5P0151c	96	hypothetical	RM1221	Q5HTC5_CA	100	2.10E-37	96
			5P0152	174	hypothetical	no matches				
5e02.q	23	2304	5P0153	297	Cmgb5 (virB5)	*C. jejuni* pTet	Q69BA1_CA	98.65	1.90E-98	297
			5P0154	332	Cmgb6 (virB6)	*C. coli*	Q69BF0_CA	85.46	3.60E-98	330
			5P0155	55	Cmbg7 (virB7)	*C. jejuni* pTet	Q69B99_CA	100	8.50E-23	55
			5P0156	89	Cmgb8 (virB8)	*C. jejuni* pTet	Q847A8_CA	100	2.30E-31	89
3g09.q	24	4636	5P0157c	110	hypothetical	RM1221	Q5HWQ0_C	100	5.30E-43	110
			5P0158c	90	hypothetical	RM1221	Q5HWP9_CA	100	1.90E-28	90
			5P0159c	161	gam protein	RM1221	Q5HWP7_CA	100	3.70E-53	161
			5P0160c	113	hypothetical	*C. coli* plasmid pBT9810	Q9K5D7_CA	95.57	2.30E-38	113
			5P0161c	307	DNA transposition protein B	RM1221	Q5HWP3_CA	97.07	1.40E-103	307
			5P0162c	419	DNA transposition protein A	RM1221	Q5HWP2_CA	97.85	1.40E-145	419
3b02.p	24	2873	5P0163	730	Cpp49 (VirB8)	*C. coli*	Q69BD8_CA	98.77	0.00E+00	730
			5P0164	141	Cpp50 hypothetical	*C. coli*	Q69BD7_CA	100	5.00E-50	141
5d09.p	28	2775	5P0165c	617	TetO	*C. jejuni* pTet	Q69BD5_CA	99.83	0.00E+00	617
			5P0166c	59	hypothetical Cpp51	*C. coli*	Q69BD6_CA	100	5.90E+01	
			5P0167c	113	hypothetical Cpp50	*C. coli*	Q69BD7_CA	100	2.20E-38	112
2f06.p	39	6771	5P0168	198	Cpp18 hypothetical	*C. coli*	Q69BG8_CA	100	2.40E-54	183
			5P0169	462	(cpp17) nickase MagA2	*C. coli*	Q69BG9_CA	98.92	2.10E-164	462
			5P0170	234	Cpp16 hypothetical	*C. coli*	Q69BH0_CA	100	8.00E-93	234
			5P0171c	242	Cpp15 hypothetical	*C. coli*	Q69BH1_CA	100	1.80E-90	242
			5P0172c	1057	Cpp14 hypothetical	*C. coli*	Q69BH2_CA	100	0.00E+00	1054
3c07.q	41	5026	5P0173	206	Virb9-like protein	*C. jejuni* plasmid pCjA13	Q847A7_CA	100	1.90E-76	206
			5P0174	398	Cmgb10 (VirB10)	*C. jejuni* pTet	Q69B96_CA	100	1.00E-143	398
			5P0175	330	Virb11-like protein	*C. jejuni* pTet	Q69B95_CA	100	4.40E-119	348
			5P0176	603	MagB12 (virD4)	*C. jejuni* pTet	Q69B94_CA	100	0.00E+00	603
			5P0177	145	Cpp44 cag island protein	*C. jejuni* pTet	Q69B93_CA	100	2.70E-54	145

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
7c07.p	43	4998	5P0178c	62	hypothetical	RM1221	Q5HTG6_CA	95.16	4.10E-20	62
			5P0179c	95	hypothetical	RM1221	Q5HTG5_CA	100	5.10E-37	95
			5P0180c	244	dna binding protein Roi	RM1221	Q5HTG4_CA	97.54	6.80E-76	244
			5P0181c	94	hypothetical	RM1221	Q5HTG3_CA	95.75	1.10E-26	94
			5P0182	130	hypothetical	no matches				
			5P0183	105	hypothetical	no matches				
			5P0184	326	hypothetical	*Lactobacillus johnsonii*	Q74HW8_LA	30.2	2.00E-13	308
			5P0185	206	hypothetical	no matches				
			5P0186	71	hypothetical	no matches				
7f04.p	48	6568	5P0187c	740	tail tape measure protein	RM1221	Q5HWU0_CA	98.92	0	738
			5P0188	108	hypothetical	RM1221	Q5HWT9_CA	98.15	1.20E-37	108
			5P0189c	104	hypothetical	RM1221	Q5HWT7_CA	100	4.80E-23	79
			5P0190c	169	major tail tube protein	RM1221	Q5HWt6_CA	100	1.10E-60	169
			5P0191c	397	major tail sheath protein	RM1221	Q5HWT5_CA	98.24	7.30E-147	397
			5P0192c	335	hypothetical protein	RM1221	Q5HWT4_CA	94.93	1.70E-120	335
			5P0193c	128	hypothetical	RM1221	Q5HWT3_CA	90.08	8.70E-47	121
			5P0194c	104	hypothetical	RM1221	Q5HWT2_CA	95.15	5.00E-37	103
7b11.p	55	5186	5P0195	104	hypothetical	RM1221	Q5HWS5_CA	98.08	8.20E-34	104
			5P0196	508	hypothetical	Bacteriophage D3112	Q6TM76_BP	29.48	1.50E-22	502
			5P0197	460	hypothetical	*Shewanella oneidensis*	Q8EDR3_SH	21.27	1.60E-08	470
			5P0198	377	prophage muso1 F protein	RM1221	Q5HWR1_CA	27.67	8.40E-20	365
			5P0199c	167	phage virion morphogenesis protein	RM1221	Q5HWU1_CA	28.74	1.20E-04	167
5g07.q	51	7892	5P0200	81	hypothetical	no matches				
			5P0201c	160	phage virion morphogenesis protein	RM1221	Q5HWU1_CA	28.57	8.50E-05	168
			5P0202c	142	hypothetical	no matches				
			5P0203c	86	hypothetical	no matches				
			5P0204c	128	hypothetical	RM1221	Q5HWQ0_CA	100	4.00E-50	128
			5P0205c	90	hypothetical	RM1221	Q5HWP9_CA	98.89	5.20E-28	90
			5P0206c	161	host-nuclease inhibitor protein gam	RM1221	Q5HWP7_CA	100	3.70E-53	161
			5P0207c	112	hypothetical	*C. coli*	Q9K5D7_CA	94.64	5.80E-37	112
			5P0208c	143	hypothetical	no matches				
			5P0209c	285	transposition protein	*Helicobacter hepaticus*	Q7VI56_HEL	46.02	1.00E-41	289
			5P0210c	705	DNA transposition protein A	RM1221	Q5HWP2_CA	27.14	1.10E-17	689

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
2b12.p	82	9176	5P0211c	113	hypothetical	RM1221	Q5HWS2_CA	97.35	1.10E-36	113
			5P0212c	129	hypothetical	RM1221	Q5HWS1_CA	99.23	3.00E-52	129
			5P0213c	121	hypothetical	no matches				
			5P0214c	117	hypothetical	no matches				
			5P0215c	89	hypothetical	no matches				
			5P0216c	295	major head subunit	Bacteriophage D3112	Q6TM67_BP	35.59	1.50E-14	295
			5P0217c	346	hypothetical	no matches				
			5P0218	154	hypothetical	no matches				
			5P0219	210	baseplate assembly protein V	RM1221	Q5HWS6_CA	99.52	4.60E-76	210
			5P0220	63	hypothetical	RM1221	Q5HWS7_CA	98.41	8.70E-23	63
			5P0221	96	baseplate assembly protein w	*C. coli*	Q9K5E0_CA	97.92	7.10E-34	96
			5P0222	388	baseplate assembly protein J	RM1221	Q5HWS9_CA	99.49	1.50E-129	388
			5P0223	206	phage tail protein	RM1221	Q5HWT0_CA	93.69	7.90E-72	206
			5P0224	343	tail fibre protein H	RM1221	Q5HWT1_CA	75.29	8.90E-80	340
			5P0225	168	hypothetical	RM1221	Q5HWT2_CA	95.83	3.20E-56	168
			5P0226	69	hypothetical	RM1221	Q5HWT3_CA	98.55	9.00E-29	69
4h04.p	102	8165	5P0227c	107	hypothetical	RM1221	Q5HTE8_CA	97.26	2.50E-25	73
			5P0228c	521	hypothetical	RM1221	Q5HTE7_CA	99.62	1.30E-167	521
			5P0229c	210	hypothetical	RM1221	Q5HTE6_CA	99.52	6.20E-67	210
			5P0230c	107	phage head-tail adaptor	RM1221	Q5HTE5_CA	100	6.40E-39	105
			5P0231c	145	hypothetical	RM1221	Q5HTE4_CA	100	1.60E-37	104
			5P0232c	83	hypothetical	RM1221	Q5HTE2_CA	100	7.90E-25	83
			5P0233c	388	major capsid protein, hk97 family	RM1221	Q5HTE1_CA	100	4.00E-136	388
			5P0234c	185	hypothetical	RM1221	Q5HTE0_CA	100	3.50E-64	185
			5P0235c	289	hypothetical	RM1221	Q5HTD9_CA	100	7.20E-119	289
			5P0236c	639	hypothetical	RM1221	Q5HTD8_CA	99.53	5.30E-185	639

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
7e05.p	124	10407	5P0237	67	phage terminase, small subunit	RM1221	Q5HTC7_CA	83.08	4.30E-17	65
			5P0238	541	phage terminase, large subunit	RM1221	Q5HTC8_CA	100	0.00E+00	541
			5P0239	144	toxin-antitoxin protein	RM1221	Q5HTC9_CA	98.61	1.20E-51	144
			5P0240	390	portal protein, hk97 family	RM1221	Q5HTD0_CA	100	1.20E-143	390
			5P0241	188	phage protein, hk97 gp10 family	RM1221	Q5HTD1_CA	100	7.90E-61	180
			5P0242	116	hypothetical	RM1221	Q5HTD2_CA	99.14	7.50E-39	116
			5P0243	326	hypothetical	RM1221	Q5HTD3_CA	99.39	7.20E-114	326
			5P0244	118	hypothetical	RM1221	Q5HTD4_CA	100	2.10E-37	118
			5P0245	71	hypothetical	RM1221	Q5HTD5_CA	100	1.80E-25	71
			5P0246	124	hypothetical	RM1221	Q5HTD7_CA	97.67	7.30E-13	43
			5P0247	1224	hypothetical	RM1221	Q5HTD8_CA	94.2	0	1224
3b03.q	175	15477	5P0248	222	phage repressor protein	RM1221	Q5HWU7_CA	97.61	8.90E-79	209
			5P0249	106	hypothetical protein	RM1221	Q5HWU6_CA	97.17	1.10E-31	106
			5P0250	95	hypothetical	RM1221	Q5HWU3_CA	96.67	5.80E-18	60
			5P0251c	276	dam DNA adenine methylase	RM1221	Q5HWU2_CA	98.52	1.30E-103	271
			5P0252c	322	tail protein d	RM1221	Q5HWQ8_CA	47.1	2.50E-49	327
			5P0253c	124	phage tail protein	RM1221	Q5HWR0_CA	57.26	2.50E-25	124
			5P0254c	654	tail tape measure protein, tp901 family	RM1221	Q5HWU0_CA	26.06	4.70E-22	765
			5P0255c	78	hypothetical	RM1221	Q5HWT7_CA	31.51	1.30E-02	73
			5P0256c	171	major tail tube protein	RM1221	Q5HWt6_CA	41.92	1.90E-20	167
			5P0257c	396	major tail sheath protein	RM1221	Q5HWT5_CA	96.97	2.10E-144	396
			5P0258c	337	hypothetical	RM1221	Q5HWT4_CA	98.52	2.20E-123	337
			5P0259c	123	hypothetical	RM1221	Q5HWT3_CA	98.37	4.80E-52	123
			5P0260c	168	hypothetical	RM1221	Q5HWT2_CA	95.83	1.20E-55	168
			5P0261c	343	tail fiber protein H	RM1221	Q5HWT1_CA	75.59	1.60E-80	340
			5P0262c	206	tail protein	RM1221	Q5HWT0_CA	91.26	3.30E-70	206
			5P0263c	388	baseplate assembly protein J	RM1221	Q5HWS9_CA	98.2	6.90E-129	388
			5P0264c	96	baseplate assembly protein W	*C. coli*	Q9K5E0_CA	97.92	5.10E-35	96
			5P0265c	63	hypothetical	RM1221	Q5HWS7_CA	100	3.70E-23	63
			5P0266c	210	baseplate assembly protein V	RM1221	Q5HWS6_CA	99.05	3.10E-76	210
			5P0267c	104	hypothetical	RM1221	Q5HWS5_CA	100	2.40E-34	104
7h10.p	1	717	5P0268	237	Cst-II, alpha-2,3-sialyltransferase	*C. jejuni* strain 43432	Q9F0M9_CA	95.28	6.20E-91	233
5h05.p	1	781	5P0269	60	cj0168c periplasmic protein	11168	Q9PIW0_CA	90	3.10E-16	60
			5P0270	23	cj0167c integral membrane protein	11168	Y167_CAMJ	95.65	8.50E-10	23
6e09.p	1	510	5P0271	144	cj1624c sdaa L-serine dehydratase	11168	Q9PM51_CA	96.52	7.90E-52	144
7h07.p	1	689	5P0272c	52	hmcd domain protein	RM1221	Q5HXA6_CA	94	1.70E-17	50
			5P0273	126	hypothetical	RM1221	Q5HXA8_CA	96.15	1.60E-14	52
							Q5HXA7_CA	57.38	3.20E-06	61

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
7h08.q	1	438	5P0274c	145	tail protein d	RM1221	Q5HWQ8_CA	97.86	4.70E-49	140
6b02.q	1	678	5P0275c	157	hypothetical cpp32	*C. coli*	Q69BF5_CA	67.68	6.10E-36	164
			5P0276c	64	cmgb3/4	*C. coli*	Q69BF6_CA	93.75	2.60E-18	64
5g10.q	1	752	5P0277c	88	Bll0816 protein (propionate catabolism?	*Bradyrhizobium japonicum*	Q89W77	40	2.10E-05	85
			5P0278c	158	cj1394 fumarate lyase	11168	Q9PMR1_CA	94.93	7.30E-52	158
7e10.q	1	272	5P0279	50	type I RM fragment	uncultured Archaeon	Q64AS4_9A	33.8	6.40E-04	71
5d06.q	1	22			no predicted CDSs					
1b11.p	1	107			no predicted CDSs					
6h07.q	1	32			no predicted CDSs					
1d11.p	1	430			no predicted CDSs					
1f11.p	1	206			no predicted CDSs					
5c01.q	1	47			no predicted CDSs					
6b11.p	1	30			no predicted CDSs					
6d11.p	1	326			no predicted CDSs					
4c05.p	1	302			no predicted CDSs					
5f01.p	1	252			no predicted CDSs					

APPENDIX 1: predicted CDSs and their protein similarities in the *C. jejuni* strain 81-176 plasmid pVir

Locus_id	length	Putative function	Informative database match	Organism with match	SWALL	E-value	id
pVir1	373	Unknown	-	-			
pVir2	292	Unknown	-	-			
pVir3	260	Unknown	TrbM	*E. coli*	Q03537	5.1e-11	31.84
pVir4	239	Unknown	-	-			
pVir5	417	Unknown	Hypothetical Hp0444	*H. pylori*	O25192	4.3e-12	32.54
pVir6	142	Unknown	-	-			
pVir7	114	Unknown	-	-			
pVir8	102	Unknown	-	-			
pVir9	73	Unknown	-	-			
pVir10	131	Unknown	-	-			
pVir11	136	Unknown	-	-			
pVir12	143	Unknown	-	-			
pVir13	61	Unknown	-	-			
pVir14	56	Unknown	-	-			
pVir15	42	Unknown	-	-			
pVir16c	66	Unknown	-	-			
pVir17	121	Unknown	-	-			
pVir18	111	Unknown	-	-			
pVir19	120	Unknown	-	-			
pVir20	134	Unknown	-	-			
pVir21	130	Periplasmic protein	Cj1456c	*C. jejuni*	Q9PMK4	4.6e-31	90.38
pVir22c	523	Unknown	Hypothetical jhp0942	*H. pylori*	Q9ZKJ3	7.3e-13	28.06
pVir23c	82	Unknown	-	-			
pVir24	101	Unknown	-	-			
pVir25	80	Unknown	Hypothetical Hp0042	*H. pylori*	O25190	2.6e-03	35.29
pVir26	822	Type IV secretion system protein	VirB4	*H. pylori*	O25189	1.7e-44	33.87
pVir27	225	Type IV secretion system protein	VirB8/ComB1	*C. jejuni*	Q9KIS2	3.1e-81	100

Locus_id	length	Putative function	Informative database match	Organism with match	SWALL	E-value	id
pVir28	356	Type IV secretion system protein	ComB2	*C. jejuni*	Q9KIS1	1.1e-124	100
pVir29	378	Type IV secretion system protein	ComB3	*C. jejuni*	Q9KIS0	9.2e-122	100
pVir30	66	Unknown	-	-			
pVir31	317	Type IV secretion system protein	VirB11	*C. jejuni*	Q9KIR9	9.8e-115	100
pVir32	135	Unknown	-	-			
pVir33	628	Type IV secretion system protein	VirD4	*E. coli*	Q91UW5	4e-20	24.52
pVir34	56	Unknown	-	-			
pVir35	293	Unknown	Hypothetical jhp0926	*H. pylori*	Q9ZKK9	8e-03	21.56
pVir36	89	Unknown	-	-			
pVir37	382	Conjugal transfer protein	Mlr9255	*Rhizobium loti*	Q981S2	1.3e-03	22.41
pVir38	655	Topoisomerase	TopA2	*H. pylori*	Q9ZKL6	7.9e-37	44.89
pVir39	121	Unknown	-	-			
pVir40	152	Single-stranded DNA-binding protein	Ssb-p1	Bacteriophage P1	Q9XJG4	4.3e-12	30.24
pVir41	57	Unknown	-	-			
pVir42	211	Unknown	-	-			
pVir43	155	Unknown	-	-			
pVir44	117	Unknown	-	-			
pVir45	70	Unknown	-	-			
pVir46	156	Unknown	-	-			
pVir47	137	Unknown	-	-			
pVir48	135	Unknown	-	-			
pVir49	107	Unknown	-	-			
pVir50	77	Unknown	-	-			
pVir51c	67	Unknown	-	-			
pVir52	222	Partition protein	ParA	*H. pylori*	O25646	1.1e-14	38.02
pVir53	209	Unknown	-	-			
pVir54c	278	Replication initiation protein	RepA	*Erysipelothrix rhusiopathiae*	Q9RHE5	1.1e-13	30.73

APPENDIX 2: predicted CDSs and their protein similarities for the *C. jejuni* strain 81-176 plasmid pTet

Locus id	length	Putative function	Informative database match	Organism with match	SWALL	E-value	id
pTet1	382	Replication initiation protein	replication protein	*Selenomonas ruminantium* plasmid ps23	Q55007	1.9e-29	36.48
pTet2	126	Unknown	-	-			
pTet3	132	Unknown	Hypothetical cjp38	*C. jejuni*	Q8GJB7	1.6e-16	40
pTet4	170	Unknown	-	-			
pTet5	185	Unknown	-	-			
pTet6	88	Unknown	Hypothetical rgi82	*Oryza sativa*	Q944E8	5.2e-03	30.3
pTet7	186	Unknown	-	-			
pTet8	88	Unknown	-	-			
pTet9	1932	DNA methylase	Orf23	*Sinorhizobium meliloti* phage PBC5	Q8W6K4	3.7e-135	38.19
pTet10c	234	Unknown	-	-			
pTet11c	462	Nickase	MagA2	*Actinobacillus actinomycetemcomitans*	Q9F276	8.9e-25	32.26
pTet12c	183	unknown	-	-			
pTet13	93	Unknown	-	-			
pTet14	203	Unknown	-	-			
pTet15	217	Unknown	Hypothetical jhp0950	*H. pylori*	Q9ZKI5	1.8e-19	39.63
pTet16	408	DNA primase	TraC	*E. coli*	P27189	4.1e-15	31.56
pTet17	87	Lipoprotein	MagB5	*Actinobacillus actinomycetemcomitans*	Q9F247	1.1e-02	37.7
pTet18c	85	Unknown	-	-			
pTet19c	61	Unknown	-	-			
pTet20	72	Unknown	Hypothetical jhp0960	*H. pylori*	Q9ZKH6	5.6e-10	52.77
pTet21	67	Unknown	Hypothetical jhp0961	*H. pylori*	Q9ZKH5	4.4e-13	68.42
pTet22	597	Unknown	Hypothetical amv156	*Amsacta moorei* entomopoxvirus	Q9EMP3	6.1e-04	22.74
pTet23c	204	Site-specific DNA recombinase	Soao172	*Shewanella oneidensis*	Q8E7Z6	1e-14	33.16

Locus id	length	Putative function	Informative database match	Organism with match	SWALL	E-value	id
pTet24c	125	Virulence-associated protein	Vap2	*Riemerella anatipestifer* pCFC1	O85171	1.9e-04	36.26
pTet25c	107	Unknown	-	-			
pTet26	87	Type IV secretion system protein	VirB2	*E. coli*	Q91UX6	1e-06	35.36
pTet27	922	ATPase	MagB3	*Actinobacillus actinomycetemcomitans*	Q9F245	1e-128	40.67
pTet28	188	Unknown	Hypothetical	*C. jejuni* pCjA13	Q847A4	1.3e-14	44.8
pTet29	221	Unknown	-	-			
pTet30	141	Single-strand DNA binding protein	Ssb-1	*Geobacter sulfurreducens*	AAR3552 7	5.3e-11	33.58
pTet31	86	Unknown	-	-			
pTet32	323	Unknown	MagB4	*Actinobacillus actinomycetemcomitans*	Q9F246	6e-19	32.66
pTet33	332	Unknown	MagB6	*Actinobacillus actinomycetemcomitans*	Q9F248	1.4e-15	25.93
pTet34	55	Lipoprotein	Cj1074c	*C. jejuni*	Q9PNM0	0.24	44.68
pTet35	220	Type IV secretion system protein	VirB8-like protein	*C. jejuni* pCjA13	Q847A8	1.1e-77	100
pTet36	295	Type IV secretion system protein	VirB9-like protein	*C. jejuni* pCjA13	Q847A7	3.7e-112	97.28
pTet37	391	Type IV secretion system protein	MagB10	*Actinobacillus actinomycetemcomitans*	Q9F252	5.7e-39	39.74
pTet38	330	Type IV secretion system protein	VirB11-like protein	*C. jejuni* pCjA13	Q847A5	4.6e-117	99.69
pTet39	603	Type IV secretion system protein	MagB12	*Actinobacillus actinomycetemcomitans*	Q9F254	4.9e-89	42.64
pTet40	145	Lipoprotein	MagB13	*Actinobacillus actinomycetemcomitans*	Q9F255	4.5e-03	26.57

Locus id	length	Putative function	Informative database match	Organism with match	SWALL	E-value	id
pTet41	254	Unknown	TrbM-like protein	*Haemophilus aegyptius* pF3031	Q8VRC6	4.5e-11	37.17
pTet42	265	Unknown	-	-			
pTet43	206	Unknown	-	-			
pTet44	730	Topoisomerase	TraE	*E. coli*	Q60215	1.8e-80	41.89
pTet45	473	Unknown	Hypothetical	*Plasmodium falciparum*	P21421	2.5e-03	25.39
pTet46	59	Unknown	Hypothetical cjp20	*C. jejuni*	Q8GJD3	3.1e-07	46.42
pTet47	639	Tetracycline resistance	TetO	*C. jejuni*	AAA23033	0	99.84
pTet48	57	Unknown	Hypothetical Orf6	*Enterococcus faecalis* transposon tn916	Q56396	4.3e-14	66.66
pTet49	222	Unknown	-	-			
pTet50	140	unknown	-	-			

APPENDIX 3

Predicted CDSs for sequenced pUC library clones of strain 81-176

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
5c06.p	2	753	8P0001	250	Fcl cj1428c	11168	Q9PMM9	56.4	2.70E-51	250
6a02.q	2	730	8P0002c	243	adhesin	*Chromobacterium violaceum*	Q7NY05	30		
8b03.p	2	815	8P0003	35	GlyA cj0402	11168	P24531	97.14	2.80E-10	35
			8P0004	232	hypothetical cj0403	11168	Q9PIA2	100	1.10E-69	176
6e09.q	2	1055	8P0005	251	hypothetical from LOS cluster	*C.jejuni* strain 11351 81176	Q9ALY2	100	1.60E-99	251
			8P0006c	73	WaaF	*C.jejuni* strain 81176	Q6TDC6	100	7.20E-30	73
6h01.q	2	770	8P0007	256	c4-dicarboxylate transporter	*Vibrio fulnificus*	Q7MJB8	38.93	6.00E-26	244
2a01.p	2	919	8P0008	86	no matches					
			8P0009	60	no matches					
7e10.q	2	532	8P0010c	176	aminotransferase cj1294	11168	Q9PN05	89.2	3.90E-53	176
7e07.q	2	772	8P0011c	189	DsbA cj0872	11168	Q9PP57	48.04	1.20E-28	179
7g05.p	2	1358	8P0012	52	Cj1161	11168	Q9PND4	83.67	2.50E-13	49
			8P0013	173	hydrophobic protein cj1158c	11168	Q9PND7	84.21	6.80E-22	76
			8P0014c	196	DnaX cj1157	11168	Q9PND8	95.91	9.30E-65	196
7d11.q	2	931	8P0015c	310	cj1333 like hypothetical	81-176	Q7X518	100	3.00E-125	309
1b02.p	3	1444	8P0016	165	ribosomal acetyltransferase	*Ureaplasma parvum*	Q9PQI0	29.1	1.60E-02	134
			8P0017	60	no matches					
			8P0018	136	WbkC	*Brucella melitensis*	Q9ZHX0	33.96	1.10E-03	106
			8P0019	74	acyl carrier protein cj1308	11168	Q9PMZ1	93.05	1.80E-21	72
5a05.p	3	967	8P0020	61	cj1724c hypothetical	11168	Q9PLV4	100	3.60E-23	60
			8P0021	199	cj1721c outer membrane protein	11168	Q9PLV7	63.77	5.20E-48	196
6a01.p	3	1000	8P0022c	74	hypothetical cj0976	11168	Q9PNW3	94.59	7.70E-26	74
			8P0023c	226	heme-hemopexin HxuB	*Haemophilus influenzae*	AAQ10738	20.5	2.20E-02	239
3a07.q	3	1446	8P0024	70	no matches	match to 1580383-1580533				
			8P0025	261	membrane protein cj1658	11168	Q9PM19	96.52	1.70E-82	259
8b05.p	3	1693	8P0026	187	hypothetical cj1340c	11168	Q9PMV9	34.44	1.20E-16	180
			8P0027	226	FlaA	*C.jejuni* strain d2677	Q9R953	100	1.80E-74	226
2d02.p	4	1229	8P0028	336	Cst-I	*C. jejuni* strain oh4384	Q9RGF1	41.14	1.20E-31	367
			8P0029	43	hypothetical cj1431c	11168	Q9PMM6	41.02	6.70E-01	39

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
8h11.p	4	1479	8P0030	59	hypothetical cj0121	11168	Q9PJ06	98.27	7.00E-22	58
			8P0031	402	cj0243c hypothetical	11168	Q9PIQ2	21.51	0.005	344
1a07.p	5	1123	8P0032c	94	no matches					
			8P0033c	279	cj0032	11168	Q9PJ79	64.93	4.90E-54	288
6e04.q	5	1000	8P0034	196	AcnB cj0835c aconitate hydratase	11168	Q9PP88	98.46	5.70E-76	196
			8P0035	116	cj0834c periplasmic protein	11168	Q9PP89	92.24	9.30E-39	116
5a10.q	5	1348	8P0036c	225	cj1442c	11168	Q9PML5	63.34	1.50E-50	221
			8P0037c	224	KpsF	11168	Q9PML4	95.92	5.50E-76	221
2h05.p	5	1445	8P0038c	46	no matches					
			8P0039c	433	hypothetical	*Fusobacterium nucleatum*	Q8REK3	23.59	2.70E-03	339
3e08.q	6	1267	8P0040c	330	cj1310c hypothetical	11168	Q9PMY9	62.95	1.40E-78	332
			8P0041	61	NeuA2	11168	Q9PMY8	96.72	2.40E-20	61
2e09.q	6	1438	8P0042	451	cj0971	11168	Q9PNW7	83.81	2.30E-22	105
1e08.q	6	1086	8P0043c	340	DmhA	*Yersinia pseudotuberculosis*	Q8G8E4	78.2	2.30E-99	335
1c09.q	7	1346	8P0044	218	FlaB	81116	Q9RF25	100	3.60E-66	218
			8P0045c	217	cj1337	81-176	Q7X517	100	7.80E-71	217
3b10.q	8	1985	8P0046c	135	cj0305c	11168	Q9PIJ4	66.66	4.00E-33	135
			8P0047c	380	BioF	11168	Q9PIJ3	75.78	5.40E-112	380
			8P0048	124	BioA	11168	Q9PIJ2	94.35	2.50E-47	124
1b01.p	4	2336	8P0049c	72	type I RM mm2978	*Methanosarcina mazei*	Q8PSU8	37.03	6.40E-03	54
			8P0050c	636	rm cc0620	*Caulobacter crescentus*	Q9AAH8	39.62	7.10E-58	641
7b08.q	8	1272	8P0051	130	cj0294 moeb/thif family protein	11168	Q9PIK5	95.38	2.90E-46	130
			8P0052c	126	PanD cj0296c	11168	Q9PIK3	98.41	2.30E-43	126
			8P0053c	137	PanC cj0297c	11168	Q9PIK2	96.35	8.90E-43	137
4a03.p	11	1388	8P0054c	462	FlgE	81-176	Q83WM5	100	1.10E-177	462
6g02.p	11	1765	8P0055c	412	DTPT transporter (disrupted)	*Photorhabdus luminescens*	Q7N5W6	47.99	1.30E-79	398
			8P0056c	117	ABC transporter	*Photorhabdus luminescens*	Q7N5W6	47.66	1.10E-15	107
4e04.p	12	1893	8P0057c	164	ModC	11168	Q9PIJ9	76.22	1.00E-39	164
			8P0058c	222	ModB	11168	Q9PIJ8	85.13	2.40E-70	222
			8P0059c	133	cj0302c	11168	Q9PIJ7	64.61	1.80E-28	130
			8P0060c	109	ModA	11168	Q9PIJ6	81.65	1.40E-30	109
6d08.p	16	2885	8P0061	76	no matches					
			8P0062	879	type I RM mm2976	*Methanosarcina mazei*	Q8PSV0	44.63	6.10E-131	867

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
2b09.p	14	4502	8P0063	126	Cst-iI	81-176	Q9L9Q5	98.4	5.20E-46	125
			8P0064	346	NeuB1	*C.jejuni* strain atcc 43456	Q93D04	100	6.70E-129	346
			8P0065	374	NeuC1	*C.jejuni* atcc 43456	Q93D03	100	5.60E-134	374
			8P0066	315	CgtA-II	*C.jejuni* atcc 43449 and 43456	Q934C5	100	1.20E-126	315
			8P0067	221	NeuA1	*C.j strain* 43456	Q933W2	100	1.40E-82	221
			8P0068	117	acetyltransferase	*C.jejuni* strain atcc 43449	Q93CZ2	100	2.30E-46	117
6a11.p	28	2110	8P0069	576	FlaB	*C.jejuni* 81116	Q9RF25	97.74	3.90E-174	576
7f02.p	31	4287	8P0070	150	TraN	*Sphingomonas aromaticivorans*	O85935	42	2.60E-17	150
			8P0071	396	TraG	*Escherichia coli*	P33790	20.44	1.70E-04	357
			8P0072	174	no matches					
			8P0073	294	no matches					
			8P0074c	93	no matches					
7f11.p	37	3740	8P0075	49	SecY cj1688	11168	Q9PLZ0	100	1.70E-18	49
			8P0076	398	hypothetical	*Clostridium perfringens*	Q8XNB6	34.7	8.00E-43	412
			8P0077	670	hypothetical	*Rhizobium loti*	Q98CJ2	39.13	5.90E-94	672
6g03.q	38	3087	8P0078	740	DmsA	*Wolinella succinogenes*	Q7MRE1	62.01	5.40E-189	745
			8P0079	218	FdhB	*Wolinella succinogenes*	Q7M8T2	62.67	2.00E-55	217
			8P0080	70	MraY hypothetical	*Wolinella succinogenes*	Q7MRE0	47.14	2.40E-07	70
7d05.p	41	4416	8P0081	519	cyt C biogenesis protein	*Wolinella succinogenes*	Q7M7P8	59.45	3.20E-121	518
			8P0082c	556	GGT jhp1046	*H.pylori* j99	Q9ZK95	67.2	2.90E-134	558
			8P0083	306	cj0031	11168	Q9PJ80	61.93	9.80E-63	310
4b02.p	47	5554	8P0084	656	cytochrome C	*Shewanella oneidensis*	Q8EJI6	55.24	1.60E-136	677
			8P0085	689	cytochrome C family protein	*Geobacter sulfurreducens*	AAR33608	36.31	2.50E-59	614
			8P0086	194	hypothetical	*Wolinella succinogenes*	Q7MQN4	38.88	3.10E-23	198
			8P0087	234	cyt C biogenesis protein	*Helicobacter hepaticus*	Q7VHG9	37.97	4.30E-24	237
6d10.q	56	4739	8P0088	273	cj1368	11168	Q9PMT2	89.37	6.70E-97	273
			8P0089	1121	cj1365c serine protease	11168	Q9PMT5	39.66	2.30E-79	1147
			8P0090	147	cj1369 transport	11168	Q9PMT1	81.63	8.20E-45	147
7g11.p	2	1380	8P0091	218	iron uptake ABC transport cj0173c	11168	Q9PIV6	99.08	2.40E-73	218
			8P0092c	158	PurU cj0790	11168	Q9PPC9	100	2.00E-53	146
			8P0093c	61	RNA nucleotidyltransferase cj0789	11168	Q9PPD0	98.21	7.10E-19	56
2h12.p	2	906	8P0094c	286	no matches					
5e04.q	1	396	8P0095c	93	no matches					
3h05.p	1	662	8P0096c	219	cj1342c hypothetical	11168	Q9PMV7	78.53	4.10E-71	219

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
3d09.q	1	176	8P0097c	54	no matches					
1a12.p	1	658	8P0098c	218	LpsA	*Vibrio parahaemolyticus*	Q87T79	41.36	2.40E-32	220
2c01.q	1	596	8P0099c	162	glycosyltransferase	*C.jejuni* strain atcc 43456	Q93D08	100	1.40E-60	162
4a04.q	1	634	8P0100c	88	Atpe cj0936	11168	Q9PNZ7	90.9	2.20E-22	88
4c05.q	1	641	8P0101	157	no matches					
8e07.p	1	880	8P0102	119	exonuclease recj cj0028	11168	Q9PJ83	97.36	4.60E-41	114
			8P0103	127	Ansa cj0029	11168	Q9PJ82	83.46	4.40E-33	127
1f07.q	1	595	8P0104	37	WaaV	*C.jejuni* strain 43456	Q93D01	100	1.50E-12	37
			8P0105c	160	acetyltransferase	*C.jejuni* strain atcc 43456,	Q93D02	98.75	9.40E-59	161
6a06.p	2	1189	8P0106c	213	hypothetical dsba cj0872	11168	Q9PP57	98.12	9.00E-77	213
			8P0107c	141	arylsulfatase AstA	81-176	Q46098	100	2.80E-54	141
7e09.p	2	901	8P0108c	143	afimbrial adhesin	*Escherichia coli*	Q93QU8	32.39	0.00034	142
5g02.p	1	197			N/A	11168				
1a08.p	1	357			N/A					
5b12.q	1	666			N/A					
6h03.q	2	742			N/A	11168				
6h12.q	1	274			N/A	11168				

APPENDIX 4

Predicted CDSs for sequenced pUC library clones of strain M1

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
2d02.q	2	512	MP0001	137	serine protease SigA	*Shigella flexneri* 2a	Q9L8L1	37.4	3.80E-08	139
5d06.p	2	671	MP0002	223	restriction modification protein CjeI	*C. jejuni* strain p37	Q9JN06	91.55	6.80E-73	225
3a07.p	2	907	MP0003	61	cj1058c	11168	Q9PNN3	77.77	1.50E-10	45
			MP0004	57	cj1057c	11168	Q9PNN4	94.73	9.90E-15	57
			MP0005	184	cj1056c (disrupted)	11168	Q9PNN5	76.34	4.80E-50	186
3a10.q	2	555	MP0006c	163	no matches					
2f03.q	2	664	MP0007c	73	WlaK	*C. jejuni* strain 81116	O86158	98.63	1.90E-26	73
			MP0008c	115	WlaI	*C.jejuni* strain 81116	O86157	100	5.30E-43	115
4e10.q	2	457	MP0009c	121	cj1375	11168	Q9PMS5	94.95	1.40E-39	119
5b05.p	2	823	MP0010c	229	DTPT dehydratase	*Helicobacter hepaticus*	Q7VJZ3	59.29	1.80E-47	226
4e04.p	2	752	MP0011	101	cj0032 RM	11168	Q9PJ79	65.34	1.30E-17	101
			MP0012	148	cj0033 membrane	11168	Q9PJ78	39.37	8.10E-06	160
4e02.q	2	624	MP0013	207	PorA membrane	*C. jejuni* Strain x7199	Q9F782	88.37	1.30E-67	215
3f12.p	2	812	MP0014c	233	cj0139 endonuclease	11168	Q9PIY8	53.28	3.70E-28	259
3b05.q	3	1437	MP0015c	185	glycosyltransferase	*C. jejuni* Strain 11828	Q9ALT2	100	8.50E-72	185
			MP0016c	266	glycosyltransferase	*C. jejuni* Strain 11828	Q9ALT1	100	2.30E-87	228
2h08.p	3	746	MP0017c	195	hypothetical	*C. jejuni* Strain rm1221	Q8RN32	97.43	3.00E-70	195
4a03.q	3	1095	MP0018c	365	FlaA	*C. jejuni* Strain 81116	FLA2_CAMJI	100	1.20E-116	365
3d02.q	6	1972	MP0019	57	alginate O-acetylation protein	*C. jejuni* Strain 11828	Q9ALT7	100	5.40E-22	57
			MP0020	371	hypothetical	*C. jejuni* Strain 11828	Q9ALT8	97.99	6.90E-135	349
			MP0021c	186	cj1149c isomerase	11168	LPC1_CAMJ	96.77	1.40E-65	186
2g06.p	3	887	MP0022	94	ppK cj1359	11168	PPK_CAMJE	98.91	4.00E-29	92
			MP0023c	152	VacA	*H. pylori* J99	Q9ZME6	26.41	7.70E-03	159
3e04.p	3	1277	MP0024c	425	cj1337 hypothetical	*C.jejuni* Strain 81-176	Q7X517	99.76	7.90E-159	424
3e08.p	4	1095	MP0025	273	no matches					
2c03.p	4	794	MP0026c	242	no matches					
1g01.q	4	944	MP0027c	314	cj1178c acidic	11168	Q9PNB7	91.42	2.90E-80	315
1f05.p	4	1115	MP0028c	307	RlmA transferase	*C.jejuni* strain 81116	Q9K5D0	98.37	1.40E-110	307
			MP0029c	38	glycosyltransferase wlanB	*C.jejuni* strain 81116	Q9K5D1	100	3.40E-17	38
2h03.q	4	718	MP0030	239	cj0262c chemotaxis	11168	Q9PIN3	55.46	3.60E-45	238

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
1b09.q	4	1934	MP0031c	264	phosphodiesterase	*Bradyrhizobium japonicum*	Q89MQ1	40.9	3.50E-34	264
			MP0032c	202	HAD hydrolase	*Caulobacter crescentus*	Q9Q7S7	28.19	2.30E-06	188
			MP0033c	156	ABC transporter	*Brucella suis*	Q8FUP0	36.36	4.50E-09	165
2f12.q	4	1522	MP0034	461	0-acetylation protein	*C.jejuni* strain 11828	Q9ALT7	100	1.90E-186	459
3e06.q	5	1419	MP0035	469	cj1614 ChuA	11168	Q9PM61	91.19	2.10E-174	477
4f03.q	6	1431	MP0036	317	arylsulfatase	*C.jejuni* strain 81-176	Q46098	99.68	3.60E-129	317
			MP0037	130	cj0872 DsbA	11168	Q9PP57	96.15	6.30E-48	130
1a12.p	6	1908	MP0038c	496	ABC transporter (disrupted)	*Photorhabdus luminescens*	CAE14106	47.58	5.60E-80	496
			MP0039c	81	di-tripeptide transporter	*Yersinia pseudotuberculosis*	Q669J3	44.73	2.20E-09	76
3b03.q	6	1733	MP0040c	552	cj1334 hypothetical	*C.jejuni* strain 81-176	Q7X519	76.71	1.40E-136	481
5c06.p	6	1714	MP0041	428	WbyH (o-antigen)	*Yersinia pseudotuberculosis*	Q9RCB8	43.88	1.10E-65	417
			MP0042c	146	AscF reductase	*Yersinia pseudotuberculosis*	Q57103	32.79	1.10E-07	125
1h04.q	7	2681	MP0043	225	EpsS epimerase	*Methylobacillus*	Q83VQ2	56.05	2.00E-47	223
			MP0044	384	Glf galactopyranose mutase	*Helicobacter hepaticus*	Q7VJP0	53.48	7.50E-74	359
			MP0045	291	hypothetical	*C.jejuni* strain 11828	Q9ALS8	28.04	3.00E-09	296
3d04.q	8	1528	MP0046c	508	adhesin	*Chromobacterium violaceum*	AAQ59146	24.77	5.00E-03	440
2g01.p	8	1953	MP0047	167	hypothetical	*Shewanella oneidensis*	Q8E9K9	26.61	3.60E-05	139
			MP0048	169	type I RM	*Archaeoglobus fulgidus*	O28563	45.94	3.40E-13	111
			MP0049	226	type I RM	*Wolinella succinogenes*	CAE10680	32.57	1.30E-07	221
1h01.q	8	1192	MP0050	381	cytochrome c	*Shewanella oneidensis*	Q8EJI6	54	5.70E-71	400
3d07.q	8	1703	MP0051	116	hypothetical (los locus)	*C.jejuni* strain 11828	Q9ALT0	95.69	7.40E-35	116
			MP0052c	361	aminotransferase	*C.jejuni* strain 11828	Q9ALS9	98.6	4.40E-139	358
			MP0053c	77	membrane protein	*C.jejuni* strain tgh9011	Q6EB21	84.5	2.10E-20	71
3e11.p	8	1247	MP0054	375	weak match to hemolysin	*Xanthomonas axonopodis*	Q8PHP1	23.89	5.30E-02	318
5h04.p	10	1763	MP0055c	69	iron binding protein	11168	Q7AR79	79.7	6.50E-19	69
			MP0056c	220	hypothetical	*Helicobacter hepaticus*	Q7VK87	34.32	6.70E-16	201
			MP0057c	206	hypothetical	*Helicobacter hepaticus*	Q7VK87	36.22	3.10E-19	196
5d03.p	8	1526	MP0058	432	UGDH glucose dehydrogenase	*Agrobacterium tumefaciens*	Q8U8E3	48.84	4.10E-78	434
			MP0059	34	UDP-glucose 4-epimerase	*Fusobacterium nucleatum*	Q8RGC6	67.64	1.50E-05	34
3c05.q	9	1729	MP0060	183	ribosomal protein	*Vibrio vulnificus*	Q8DF32	32.96	1.90E-06	179
			MP0061c	115	no matches					
			MP0062c	209	putative phage repressor protein	Bacteriophage phi ETA	Q9G039	28.89	5.20E-05	180
1b10.q	10	1565	MP0063	45	cj1337 hypothetical	*C.jejuni* strain 81-176	Q7X517	100	2.40E-14	45
			MP0064c	464	FlaB	*C.jejuni* strain 81116	Q9RF25	100	3.90E-144	462

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
3b09.p	10	1924	MP0065c	81	no matches					
			MP0066c	559	hypothetical	*Fusobacterium nucleatum*	Q8REK3	23.71	1.50E-04	485
5c01.p	11	1646	MP0067c	85	hypothetical	*Wolinella succinogenes*	CAE10494	38.09	2.80E-06	84
			MP0068c	288	DmsC type gene (MraY)	*Wolinella succinogenes*	CAE10493	42.5	8.10E-40	287
			MP0069c	179	oxidoreductase FdhB	*Wolinella succinogenes*	CAE10492	62.77	1.50E-43	180
2g03.q	11	2942	MP0070c	66	RloA	*C.jejuni* strain Rm1551 & rm1850	Q8G8E7	100	1.90E-22	66
			MP0071c	769	HsdR	*C.jejuni* strain 81116	Q8RIX1	100	0	769
			MP0072	71	cj1548c dehydrogenase	11168	Q9PMC1	100	3.40E-30	71
4h06.p	13	1792	MP0073	110	cj0123c	11168	Q9PJ04	90.9	3.70E-36	110
			MP0074c	446	hypothetical	*Plasmodium falciparum*	Q8IHQ0	19.2	0.012	453
3a05.q	12	1401	MP0075	39	periplasmic protein cj0770c	11168	Q9PPE9	100	1.50E-05	22
			MP0076	149	hypothetical reP	*Treponema denticola*	Q9AQF2	39.59	8.00E-14	149
			MP0077	60	hypothetical TnpV	*Clostridium difficile*	O05416	46.42	6.00E-06	56
3e01.p	14	1779	MP0078	146	glucose epimerase	*Pyrococcus furiosus*	Q8U170	34.09	1.70E-07	132
			MP0079	376	glucose dehydrogenase	*Pyrococcus abyssi*	Q9UZI8	38.33	1.50E-42	373
4g01.p	15	1955	MP0080c	85	RlfA	Bacteriophage P1	Q71TB8	44.57	2.10E-07	85
			MP0081c	552	type I RM	*Wolinella succinogenes*	CAE10680	70.27	6.60E-149	555
1g05.q	15	2785	MP0082	238	cj0414 oxidoreductase	11168	Q9PI91	44.03	3.20E-34	243
			MP0083	571	cj0415 oxidoreductase (disrupted)	11168	Q9PI90	57.14	7.70E-131	574
2c11.p	15	3856	MP0084	67	hypothetical	*C. jejuni* strain rm1221	Q8RN32	100	8.20E-22	65
			MP0085	149	hypothetical	*C.jejuni* strain rm1221	Q8RN33	97.84	2.20E-51	139
			MP0086	251	decarboxylase pcac	*Methanosarcina acetivorans*	Q8TTM1	42.57	1.30E-37	249
			MP0087c	496	HsdM	*C.jejuni* strain rm2227	Q8RN18	96.77	2.60E-181	496
			MP0088c	198	HsdS	*C.jejuni* strain rm1163 & rm1508	Q8G8A9	99.48	6.40E-74	194
4e08.q	16	1909	MP0089	164	cytochrome C	*Shewanella oneidensis*	Q8EJI6	49.08	1.20E-26	163
			MP0090	457	hpothetical/ possible cyt C	*Shewanella oneidensis*	Q8EJI5	39.43	1.70E-12	142
3h01.q	16	2537	MP0091c	118	permease protein	*Rhodopseudomonas palustris*	Q6NDI1	43.75	7.00E-15	112
			MP0092c	285	ABC transporter permease	*Rhizobium loti*	Q98JZ2	48.54	1.00E-49	274
			MP0093c	372	ABC transporter	*Agrobacterium tumefaciens*	Q8UIA7	45.43	1.10E-48	372
			MP0094c	41	cj1687	11168	Q9PLZ1	100	1.50E-16	41
3d08.p	18	2768	MP0095	153	Cj1431c hypothetical	11168	Q9PMM6	28.32	2.90E-04	173
			MP0096	264	DdhA (los)	*Yersinia enterocolitica*	Q56860	59.47	4.80E-60	264
			MP0097	452	glucose dehydratase	*Fusobacterium nucleatum*	EAA24619	60.67	6.00E-109	445
			MP0098	50	no matches					

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
1d11.p	18	2713	MP0099c	124	no matches					
			MP0100c	57	hypothetical from transposon	*Enterococcus faecalis*	Q56396	66.66	4.10E-14	57
			MP0101c	582	TetO	*C.jejuni* plasmid pCjA13	Q84FM6	99.48	1.40E-202	577
1g06.p	21	2261	MP0102c	55	cj1584c periplasmic	11168	Q9PM91	83.33	9.80E-16	54
			MP0103	600	DmsA	*Wolinella succinogenes*	CAE10491	61.69	6.00E-155	603
4e01.q	35	4308	MP0104c	881	TraG pseudogene	*Vibrio vulnificus*	BAC97743	21.04	4.10E-11	879
			MP0105	51	cj0937 membrane protein	11168	Q9PNZ6	100	8.40E-20	51
4d08.p	42	4924	MP0106	813	cytochrome C	*Wolinella succinogenes*	CAE11153	54.26	5.90E-172	820
			MP0107c	556	GGT	H. pylori J99	Q9ZK95	67.74	9.40E-136	558
			MP0108	182	cj0031 RM	11168	Q9PJ80	53.8	3.70E-30	184
1c08.p	2	715	MP0109	197	hypothetical (disrupted)	*Helicobacter hepaticus*	Q7VK87	39.28	1.40E-19	196
2g10.p	2	1010	MP0110c	202	Ansa cj0029	11168	Q9PJ82	86.13	8.00E-59	202
			MP0111c	87	RecJ cj0028	11168	Q9PJ83	96.55	1.20E-32	87
1f03.p	2	1035	MP0112c	166	HsdS	c.j strain rm1049, rm1861, 81116	Q8RJ16	100	2.70E-64	166
			MP0113c	179	RloB	c.j strain rm1049, rm1861, 81116	Q8RIW9	100	1.70E-66	179
1b04.q	2	760	MP0114c	170	ABC transporter (disrupted)	*Rhizobium loti*	Q98JZ4	36.25	1.20E-13	160
			MP0115c	60	ABC transporter permease	*Rhizobium loti*	Q98JZ3	56.66	1.10E-10	60
2d06.q	2	593			no predicted CDSs					
2d03.p	2	762	MP0116c	206	hypothetical	*Helicobacter hepaticus*	Q7VIF8	51.33	6.50E-33	187
2e03.p	2	824	MP0117c	70	hypothetical	*Wolinella succinogenes*	Q7MQN4	39.34	2.70E-04	61
			MP0118c	187	formate dehydrogenase	*Vibrio cholerae*	Q9KRX2	28	9.00E-04	
3a03.p	1	646	MP0119c	151	hypothetical	*S. typhimurium* phage ST64B	Q8HAA0	30.87	1.30E-06	149
2b12.p	1	591	MP0120c	118	WlanB glycosyltransferase	*C.jejuni* strain 81116	Q9K5D1	100	1.00E-41	118
			MP0121c	78	WlanA (lipid A sysnthesis cluster)	*C.jejuni* strain 81116	Q9K5D2	100	1.00E-32	78
5b01.p	1	585	MP0122	93	cj1305c hypothetical	11168	Q9PMZ4	57.81	8.00E-12	64
2c05.p	1	425	MP0123c	104	no matches					
1e03.q	1	358	MP0124	119	hypothetical	*Pasteurella multocida*	Q9CKR7	39.02	1.10E-03	82
2a08.q	1	471	MP0125c	99	NADH dehydrogenase	*Strongyloides stercoralis*	CAD90562	36.45	3.30E-03	96
4d09.p	1	814	MP0126c	49	no matches					
			MP0127c	222	hypothetical	*Plasmodium yoelii yoelii*	EAA18980	24.27	0.0093	173
2e10.p	1	805	MP0128	20	transferase cj1050c	11168	Q9PNP1	95	6.60E-06	20
			MP0129	199	membrane protein cj1049c	11168	Q9PNP2	87.94	2.90E-65	199
			MP0130	48	Dape or Cj1048c	11168	Q9PNP3	100	1.70E-17	47
2g02.q	2	664	MP0131c	220	pgi cj1535c pseudogene	11168	G6PI_CAMJE	82.27	4.00E-64	220

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
3b01.p	4	1024	MP0132	89	EspC	*Escherichia coli*	P77070	43.18	4.40E-11	88
			MP0133	225	Argc cj0224	11168 & TGH9011	ARGC_CAMJ	98.22	6.80E-86	225
3e02.q	1	733	MP0134	238	membrane protein cj0692c	11168	Q9PPL5	85.71	4.00E-54	238
4h07.p	2	932	MP0135	154	cj1295 hypothetical	11168	Q9PN04	88.88	6.40E-53	153
			MP0136	157	cj1296 & cj1297	11168	Q9PN03	79.04	1.30E-31	105
							Q9PN02	56.75	2.10E-03	37
5h05.p	4	1775	MP0137	53	hydrophobic protein	11168	Q9PLV0	97.5	4.40E-13	40
			MP0138	127	cj1724c hypothetical	11168	Q9PLV4	100	8.80E-51	127
			MP0139	214	cj1721c outer membrane protein	11168	Q9PLV7	65.42	2.10E-55	214
			MP0140c	106	cj1720 hypothetical	11168	Q9PLV8	100	8.80E-38	107
4c04.p	3	1057	MP0141	202	Cj0967 periplasmic protein	11168	Q9PNW9	96.42	1.90E-31	112
			MP0142	115	hemagglutinin-related protein/ adhesin	*Ralstonia solanacearum*	Q8XQ42	36.28	6.00E-05	113
2g07.q	8	1844	MP0143	470	cj0970, cj0971, cj0972, cj0973	11168	Q9PNW7	95.31	4.90E-34	128
							Q9PNW8	85.85	1.20E-21	99
							Q9PNW6	55.78	2.30E-09	95
							Q9PNW5	93.54	1.60E-03	31
			MP0144	65	Cj0975	11168	Q7AR82	97.29	6.90E-09	37
4f07.p	2	730	MP0145c	31	hypothetical	11168	Q9PNW8	80	7.50E-05	30
			MP0146c	151	ceub uptake permease cj1352	11168	Q9PMU7	98.01	1.10E-50	151
			MP0147c	35	pldA	11168	Q9PMU8	97.14	7.90E-14	35
2f07.q	4	1193	MP0148	89	haemoglobin protease	*Escherichia coli*	Q8FKM0	45.97	6.10E-07	87
			MP0149	162	no matches					
			MP0150	89	no matches					
2b05.p	4	1063	MP0151c	255	dicarboxylate transporter	*Vibrio vulnificus*	BAC95008	35.77	1.00E-20	232
			MP0152	31	hypothetical Cj1523c	11168	Q9PME1	96.77	5.00E-11	31
4d12.p	1	788	MP0153c	97	cj0865 oxidoreductase DsbB	11168	DSBI_CAMJE	95.78	2.40E-39	95
			MP0154c	167	Cj0864 periplasmic protein	11168	Q9PP59	91.76	7.30E-23	85
4e06.p	10	2078	MP0155c	423	Bll0816 hypothetical	*Bradyrhizobium japonicum*	Q89W77	33.48	1.40E-39	427
			MP0156c	266	cj1394 fumarate lyase	11168	Q9PMR1	95.11	1.70E-95	266

APPENDIX 5

Predicted CDSs for sequenced pUC library clones of strain 40671

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
3e04.q	2	606	4P0001	86	TraH	*Comamonas acidovorans* pUO1	BAC82023	41.79	1.60E-08	67
			4P0002	110	no matches					
3d01.p	2	1132	4P0003	115	cj0138	11168	Q9PIY9	90.38	2.00E-30	104
			4P0004	95	no matches					
			4P0005	111	no matches					
1d01.p	2	1060	4P0006	345	mcp-type signal transduction	11168	Q9PMF7	40	5.60E-47	345
3f04.p	4	2259	4P0007c	570	cj1440c sugar transferase	11168	Q9PML7	49.64	2.00E-46	423
			4P0008	139	cj1421c sugar transferase	11168	Q9PMN6	84.17	7.10E-43	139
1b09.p	4	988	4P0009c	342	MagB10	*Actinobacillus actinomycetemcomita*	Q9F252	40.35	2.20E-34	342
3c10.p	4	1147	4P0010c	333	sialic acid biosynthesis	*C. jejuni* strain 43446	Q9L9Q4	99.09	4.20E-123	332
3c08.q	4	1341	4P0011	315	no matches					
1c07.p	4	1052	4P0012c	350	FlgE	*C.jejuni* strain lio7	O86148	99.41	6.90E-122	344
3f07.p	4	1011	4P0013c	336	MagB12	*Actinobacillus actinomycetemcomita*	Q9F254	40.95	8.10E-44	337
2d04.p	4	1442	4P0014c	95	no matches					
1d05.p	4	1179	4P0015	101	hypothetical cj1724c	11168	Q9PLV4	100	1.40E-39	101
			4P0016	213	cj1721c membrane protein	11168	Q9PLV7	64.01	5.00E-54	214
1h08.q	4	1079	4P0017	140	hypothetical	*Helicobacter hepaticus*	Q7VGU0	35.43	4.80E-08	127
			4P0018	97	no matches					
3d10.p	4	1236	4P0019	384	hypothetical cj1341c	11168	Q9PMV8	48.55	3.40E-62	381
1d03.p	4	958	4P0020	143	hypothetical	*Wolinella succinogenes*	Q7MQT2	32.37	3.80E-09	139
			4P0021	174	hypothetical jhp0950	*H. pylori* J99	Q9ZKI5	46.7	3.20E-20	167
3g08.p	4	1498	4P0022	494	cj1431c hypothetical	11168	Q9PMM6	28.14	2.20E-22	430
1e07.p	4	1047	4P0023	87	VirB2	*Escherichia coli*	Q91UX6	35.36	9.90E-07	82
			4P0024	147	TriC	*Yersinia enterocolitica*	CAD58564	39.16	1.20E-12	143
1f06.p	4	1038	4P0025	309	FlaA	*C. jejuni* serotype 0:19	Q99QL6	100	2.70E-89	309
1b06.q	6	2049	4P0026	111	hypothetical	*Pseudomonas syringae*	Q889N9	58.76	6.50E-20	97
			4P0027	241	lipopolysaccharide biosynthesis	*Pseudomonas syringae*	Q889P3	40.49	2.40E-23	242
			4P0028	132	hypothetical	*Actinobacillus suis*	Q84CG6	57.93	9.00E-27	126
			4P0029	142	hypothetical	*Actinobacillus suis*	Q84CG5	40.55	8.60E-16	143
3g02.p	6	1643	4P0030c	521	hypothetical	*Actinobacillus suis*	Q84CG8	26.03	2.40E-11	338

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
2d09.p	6	1199	4P0031c	338	no matches					
			4P0032c	41	cj0121	11168	Q9PJ06	95.12	1.40E-14	41
3c01.q	7	1537	4P0033	243	hypothetical	*Actinobacillus suis*	Q84CG7	53.3	4.70E-42	242
			4P0034	265	c-methyltransferase	*Bordetella bronchiseptica*	Q7WR30	29.16	1.40E-06	216
1b12.q	8	2219	4P0035	633	hypothetical	*Chromobacterium violaceum*	Q7NTJ9	51.42	7.70E-120	634
			4P0036c	53	cj1161c ATPase	11168	Q9PND4	70.21	3.80E-11	47
1g01.p	8	1564	4P0037c	122	no matches					
			4P0038c	89	no matches					
2b07.p	8	1572	4P0039	446	oxidoreductase	*Bacteroides thetaiotaomicron*	Q8A7I2	44.61	7.70E-73	455
			4P0040c	82	cj1069	11168	Q9PNM5	83.54	2.60E-22	79
3f05.q	8	2881	4P0041c	295	virB9-like protein	*C.jejuni* plasmid pCjA13	Q847A7	97.28	3.50E-112	295
			4P0042c	220	virB8-like protein	*C.jejuni* plasmid pCjA13	Q847A8	100	1.10E-77	220
			4P0043c	333	magb06	*Actinobacillus actinomycetemcomita*	Q9F248	26.33	1.60E-15	319
1g12.p	9	2582	4P0044	74	hypothetical	*Bacteroides thetaiotaomicron*	Q8A5B1	47.22	5.80E-06	72
			4P0045	167	hypothetical	*Shewanella oneidensis*	Q8E9K9	26.61	2.60E-05	139
			4P0046	480	type I RM	*Archaeoglobus fulgidus*	O28563	38.63	6.70E-16	176
			4P0047	63	cj1047c	11168	Q9PNP4	88.88	1.00E-18	63
			4P0048	33	cj1046c Moeb	11168	Q9PNP5	93.93	3.10E-14	33
3e05.q	9	1813	4P0049c	276	acetyltransferase	*C.jejuni* strain 43432	Q9F0M5	98.91	3.90E-106	277
			4P0050c	221	NeuA1	*C.jejuni* strain 81-176, 43456, 4344	Q933W2	98.64	3.90E-82	221
1e06.p	10	2212	4P0051c	116	hydrolase	*Pseudomonas syringae*	Q889P1	62.28	8.20E-26	114
			4P0052c	211	hypothetical	*Pseudomonas syringae*	Q889P2	40.67	6.10E-29	209
			4P0053	295	c-methyltransferase	*Leptospira interrogans*	Q8F5S5	25	2.00E-09	276
3g05.p	10	3379	4P0054	655	MagB03	*Actinobacillus actinomycetemcomita*	Q9F245	44.82	9.20E-103	647
			4P0055	188	hypothetical	*C.jejuni* plasmid pCjA13	Q847A4	44.8	1.30E-14	183
			4P0056	221	no matches					
			4P0057	45	SSB	*C.jejuni* plasmid pVir	Q8GJE0	48.88	7.60E-06	45
1d02.q	10	1988	4P0058	402	cj1421c sugar transferase	11168	Q9PMN6	69.38	4.10E-86	343
			4P0059	228	Cst-I (disrupted)	*C.jejuni* strain 0h4384	Q9RGF1	57.85	7.60E-39	242
3a10.q	10	3859	4P0060c	830	hypothetical jhp1285	*H. pylori* J99	Q9ZJM1	30.28	5.60E-43	885
			4P0061c	413	no matches					
3f10.p	11	1692	4P0062c	539	Cj1334 hypothetical	*C.jejuni* strain 81-176	Q7X519	94.83	2.00E-168	465
1a10.p	12	3065	4P0063c	331	DmhA	*Yersinia pseudotuberculosis*	Q8G8E4	78.46	9.30E-97	325
			4P0064c	351	Fcl cj1428c	11168	Q9PMM9	59.07	2.90E-75	347
			4P0065c	181	cj1430c sugar epimerase	11168	Q9PMM7	80.66	4.40E-59	181
			4P0066c	126	cj1421c sugar transferase	11168	Q9PMN6	37.39	7.60E-05	115

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
3b11.p	15	2239	4P0067c	60	phage hypothetical	Bacteriophage P1	Q9XJP4	46.42	1.10E-03	56
			4P0068c	683	type I RM	*Wolinella succinogenes*	Q7M8H9	71.04	8.20E-168	632
1g09.q	25	3249	4P0069	501	FlaB	*Campylobacter coli*	P18245	91.18	9.80E-139	499
			4P0070c	569	cj1337 hypothetical	11168	Q9PMW2	61.13	1.90E-136	566
1h01.p	2	1191	4P0071	382	hypothetical jhp0928	*H. pylori* J99	Q9ZKK7	42.96	8.80E-50	384
1b02.q	2	725	4P0072c	152	hypothetical ydaa	*Wolinella succinogenes*	Q7MQX6	39.43	6.30E-14	142
			4P0073c	89	Vap2	*Riemerella anatipestifer*	O85171	35.59	1.30E-01	59
1d06.p	2	1050	4P0074	246	hypothetical	*H. pylori* 26695	O25892	45	4.00E-30	220
3f11.p	2	1346	4P0075c	436	putative DNA methylase	*Sinorhizobium meliloti* phage PBC5	Q8W6K4	44.87	7.20E-39	312
1d10.p	2	1151	4P0076	330	VirB11-like protein (disrupted)	*C.jejuni* plasmid pCjA13	Q847A5	99.69	4.70E-117	329
2c02.p	2	633	4P0077c	199	TraE (virB8)	*Escherichia coli*	Q60215	36.54	6.90E-14	197
3e03.p	2	744	4P0078c	247	ABC transporter	*Photorhabdus luminescens*	Q7N5W6	52.67	1.90E-55	243
1c06.p	2	1352	4P0079c	304	no matches					
			4P0080c	97	no matches					
			4P0081c	39	hypothetical	*Wolinella succinogenes*	Q7MQT0	50	0.00017	36
2b11.p	2	791	4P0082c	262	hypothetical	*H. pylori* J99	Q9ZKK7	48.47	4.00E-43	262
2c10.p	1	800	4P0083c	201	hypothetical	*Clostridium perfringens*	Q93M99	26.15	1.20E-02	195
			4P0084c	78	no matches					
1a12.q	1	772	4P0085c	205	ATPase 6	*Leishmania tarentolae*	Q33561	22.87	1.10E-02	188
3a12.p	1	819	4P0086	32	no matches					32
			4P0087	67	hypothetical	*H. pylori* J99	Q9ZKH5	66.66	1.40E-12	57
			4P0088	131	no matches					131
3a12.q	1	769	4P0089	252	TrbM-like protein	*Haemophilus aegyptius*	Q8VRC6	37.17	4.50E-11	191
1b05.p	1	696	4P0090c	230	type II RM (cj0032)	11168	Q9PJ79	60.08	2.40E-43	228
1a06.p	1	827	4P0091c	225	CfrA cj0755	11168	Q9PPG3	88	3.80E-78	225
1a05.q	1	695	4P0092	231	sialic acid biosynthesis	*C.jejuni* strain atcc43432	Q9F0M7	99.56	5.20E-82	231
1g03.p	1	810	4P0093c	87	no matches					
			4P0094c	194	no matches					
2a08.p	1	229	4P0095	70	acetyltransferase	*C.jejuni* strain 43446	Q9K379	38.57	1	70
2e08.p	1	728	4P0096c	151	no matches					
			4P0097c	90	no matches					
1g08.p	1	847	4P0098c	281	cj1305c hypothetical protein	11168	Q9PMZ4	75.97	2.10E-80	283
3d03.p	1	770	4P0099c	158	hypothetical cj1337	11168	Q9PMW2	70.77	6.30E-39	154
			4P0100c	45	efflux protein cj1174	11168	Q9PNC1	100	1.40E-15	45
1g10.p	1	151			no predicted CDSs					
2c04.p	1	122			no predicted CDSs					
2b06.p	1	90			no predicted CDSs					

APPENDIX 6

Predicted CDSs for sequenced pUC library clones of strain 52472

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
7h02.q	1	515	5P0001c	88	hypothetical	RM1221	Q5HWU3	98.33	1.10E-19	60
6e05.q	2	726	5P0002	241	di-/tripeptide transporter	RM1221	Q5HVB7_CA	96.9	6.30E-84	226
2f05.p	2	1099	5P0003c	316	type I RM	*Staphylococcus aureus*	Q6GD64_ST	35.5	2.90E-31	307
6d05.q	2	671	5P0004	223	cj0929 Pepa	11168	AMPA_CAMJ	97.76	5.40E-80	223
2g09.p	2	1098	5P0005	139	cj0807 oxidoreductase	11168	Q9PPB3_CA	98.51	1.10E-47	135
			5P0006c	104	cj0808c hydrophobic hypothetical	11168	Q9PPB2_CA	77.22	3.70E-30	101
			5P0007c	83	cj0809c hydrolase	11168	Q9PPB1_CA	90.36	4.30E-31	83
5b12.p	2	1017	5P0008	144	hypothetical	RM1221	Q5HTG8_CA	71.05	1.20E-04	38
			5P0009	78	hypothetical	RM1221	Q5HTH0_CA	92.3	2.90E-24	78
6c07.q	2	625	5P0010c	197	hypothetical	RM1221	Q5HW50_CA	98.96	1.50E-63	126
6f10.p	2	918	5P0011	300	cj0765c hiss	11168	SYH_CAMJE	87.29	1.30E-102	299
3g06.p	2	799	5P0012c	265	base plate assembly	RM1221	Q5HWS9_CA	98.11	3.70E-88	265
3d03.p	2	1137	5P0013	235	type II RM	RM1221	Q5HXC7_CA	73.39	2.20E-63	233
			5P0014	144	hypothetical	*H. pylori*	O26049_HEL	56.55	1.20E-21	145
3a04.p	2	763	5P0015	192	TrbM (cpp45)	*C. coli*	Q69BE2_CA	71.74	1.70E-52	184
			5P0016	60	hypothetical cpp46	*C. jejuni* pTet	Q69B91_CA	98.3	4.50E-17	
6a09.p	2	906	5P0017c	71	hypothetical	RM1221	Q5HTH6_CA	97.02	7.30E-22	67
			5P0018c	74	hypothetical	RM1221	Q5HTH5_CA	93.24	6.20E-27	74
			5P0019c	134	hypothetical	RM1221	Q5HTH4_CA	98.51	1.20E-50	134
4e02.q	2	543	5P0020c	179	cj1218c Riba	11168	Q9PN77_CA	95.5	2.40E-60	178
5f10.q	2	675	5P0021	224	cj0411 ATP/GTP binding protein	11168	Q9PI94_CAM	97.3	7.90E-68	223
4e01.p	2	764	5P0022	104	cj0578c Tatc sec-independent transloca	11168	TATC_CAMJ	97.08	1.50E-39	103
			5P0023	146	cj0577c QueA	11168	QUEA_CAMJ	97.26	1.40E-53	146
5e08.p	2	889	5P0024	243	HsdM	*C.jejuni* strain rm 1170	Q8RN38_CA	100	2.90E-90	242
4d12.p	2	974	5P0025	55	hypothetical cpp2	*C. jejuni* pTet	Q69BD4_CA	97.73	4.90E-17	44
			5P0026	117	hypothetical cpp8	*C. jejuni* pTet	Q69BC8_CA	99.14	1.00E-42	116
			5P0027	132	hypothetical cpp9	*C. jejuni* pTet	Q69BC7_CA	100	9.10E-47	132
3a03.q	3	1058	5P0028c	237	cj0812 Thrc	11168	Q9PPA8_CA	78.48	5.70E-70	237
			5P0029c	118	cj0811 Lpxk tetraacyldisaccharide kinas	11168	LPXK_CAMJ	84.21	4.90E-37	114
6c11.p	3	1062	5P0030c	323	hypothetical	RM1221	Q5HXA9_CA	99.69	1.90E-106	323
4d02.q	3	979	5P0031	200	hypothetical cpp46	*C. jejuni* pTet	Q69B91_CA	99	8.40E-67	200
			5P0032	102	hypothetical cpp47	*C. jejuni* pTet	Q69B90_CA	98	1.20E-34	102

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
8a03.p	4	815	5P0033c	80	hypothetical	RM1221	Q5HTH3_CA	98.63	1.10E-24	73
			5P0034c	106	hypothetical	RM1221	Q5HTH2_CA	99.06	7.20E-35	106
6a05.p	4	1794	5P0035c	124	MloA	*C.jejuni* strain rm 1852	Q8RN19_CA	100	1.40E-38	124
			5P0036c	395	HsdS	*C.jejuni* strain rm 1170	Q8RN40_CA	100	6.00E-152	395
5d07.p	4	1131	5P0037	70	type III RM	*H. pylori*	O25314_HEL	56.52	1.30E-10	69
			5P0038	304	DNA methyltransferase	*H. pylori*	O25315_HEL	61.98	8.60E-48	242
5g12.q	4	788	5P0039	74	cj0762c aspb aspartate aminotransferas	11168	Q9PPF7_CA	91.89	2.90E-24	74
			5P0040	185	hypothetical	*Nitrosomonas europea*	Q82T36_NIT	33.15	9.90E-10	184
5h07.q	4	1432	5P0041	324	hypothetical	RM1221	Q5HWQ1_CA	99.68	1.90E-116	320
			5P0042	89	hypothetical	RM1221	Q5HWQ2_CA	100	3.20E-37	89
			5P0043	57	hypothetical	RM1221	Q5HWQ3_CA	98.11	1.00E-16	53
8h04.p	4	997	5P0044	331	periplasmic protein cj0737	11168	Q7AR90_CA	38.02	2.90E-22	334
8c09.p	4	1085	5P0045	102	virion morphogenesis protein	RM1221	Q5HWU1_CA	96.94	1.30E-36	98
			5P0046	212	dam DNA adenine methylase	RM1221	Q5HWU2_CA	96.49	8.00E-63	171
7a07.p	4	1280	5P0047c	239	cj0813 KdsB	11168	Q9PPA7_CA	82.85	8.20E-75	239
			5P0048c	157	cj0812 Thrc	11168	Q9PPA8_CA	75.48	1.00E-41	155
4g03.q	4	993	5P0049c	158	phage tail protein	RM1221	Q5HWTo_CA	96.81	3.00E-57	157
			5P0050c	170	base plate assembly	RM1221	Q5HWS9_CA	98.82	3.10E-55	170
5c07.q	6	1056	5P0051c	206	cj0293 Sure	11168	SURE_CAMJ	93.78	6.60E-67	193
			5P0052	98	transporter	*Escherichia coli*	Q8FAP1_EC	51.06	1.30E-14	94
6b10.q	5	949	5P0053c	298	transport system permease	*Escherichia coli*	Q8X8T6_EC	52.03	9.00E-59	296
5f02.p	5	1246	5P0054	65	di-/tripeptide transporter	RM1221	Q5HVB7_CA	60.66	2.40E-11	61
			5P0055	282	di-/tripeptide transporter	RM1221	Q5HVB7_CA	99.65	5.40E-115	282
5f11.q	5	2192	5P0056c	215	hypothetical					
			5P0057c	224	signal peptidase I	RM1221	Q5HTF9_CA	33.18	1.40E-17	223
			5P0058	177	dna transition protein a	RM1221	Q5HWP2_CA	87.82	6.20E-43	156
8b01.p	6	875	5P0059	221	hypothetical					
7e11.p	6	1591	5P0060	287	HsdR	*C.jejuni* strain rm 1170	Q8RN42_CA	99.29	1.50E-94	283
			5P0061	238	RloF	*C.jejuni* strain rm 1170	Q8RN41_CA	100	7.10E-89	238
5d04.q	6	1564	5P0062	248	hypothetical	RM1221	Q5HWR6_CA	98.79	1.40E-90	248
			5P0063	152	hypothetical	RM1221	Q5HWR5_CA	98.49	2.80E-45	132
			5P0064	83	hypothetical	RM1221	Q5HWR4_CA	100	3.90E-28	82
8c04.p	6	2259	5P0065	752	type III RM r protein	*H. pylori*	O25314_HEL	52.78	7.30E-58	773
5h03.q	6	1528	5P0066	470	VacA autotransporter domain	*H. pylori*	Q9ZHT4_Vac	23.05	9.20E-06	192
8g05.q	6	1619	5P0067c	94	DNA binding protein	RM1221	Q5HWQ7_CA	97.87	5.30E-31	94
			5P0068	223	DNS extracellular deoxyribonuclease	RM1221	Q5HWQ6_CA	99.55	2.80E-91	223
			5P0069c	91	hypothetical	RM1221	Q5HWQ5_CA	100	2.30E-31	91
			5P0070c	84	hypothetical	RM1221	Q5HWQ4_CA	100	2.80E-33	84

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
5h08.p	6	1272	5P0071c	207	hypothetical	*Salmonella typhi*	Q8Z3Y2_SAI	41.44	8.10E-19	152
7g06.p	6	1220	5P0072	382	hypothetical	RM1221	Q5HWR2_CA	99.19	6.80E-134	369
1e12.q	7	1723	5P0073c	493	type I RM	*Methanosarcina mazei*	Q8PSV0_ME	45.52	1.80E-75	503
6f05.q	7	1270	5P0074	423	cj0629 possible lipoprotein	11168	Q9PHN8_CA	76.21	5.60E-46	269
4e07.q	7	2026	5P0075	77	cj0303c ModA	11168	Q9PIJ6_CAM	81.81	5.60E-22	77
			5P0076	133	cj0302c	11168	Q9PIJ7_CAM	64.61	9.80E-29	130
			5P0077	222	cj0301c ModB	11168	Q9PIJ8_CAM	84.68	3.40E-70	222
			5P0078	240	cj0300c ModC	11168	Q9PIJ9_CAM	78.33	1.80E-59	240
5e04.p	8	1871	5P0079c	73	cj0298c PanB	11168	PANB_CAMJ	98.59	1.10E-22	71
			5P0080c	236	hypothetical	*Helicobacter hepaticus*	Q7VI60_HEL	44.29	3.30E-33	228
			5P0081c	104	hypothetical					
8g09.p	8	2154	5P0082	93	DnaK	11168	DNAK_CAMJ	94.318	1.10E-20	88
			5P0083	579	HsdM (disrupted)	*Vibrio cholerae*	Q9KR74_VIB	47.03	1.80E-68	608
5a07.q	8	2062	5P0084	122	cj1343c putative periplasmic protein	11168	Q9PMV6_CA	98.36	2.20E-42	122
			5P0085	416	cj1342c hypothetical	11168	Q9PMV7_CA	60.24	7.70E-97	415
			5P0086	144	cj1341c hypothetical	11168	Q9PMV8_CA	94.44	2.10E-49	144
4h09.p	8	1368	5P0087	338	DNA methyltransferase	*H. pylori*	O25315_HEL	46	3.10E-39	313
			5P0088	112	serine-threonine protein kinase	*Debaryomyces hansenii*	Q6BHW6_DE	31.13	7.30E-04	106
6c03.q	8	1506	5P0089c	55	Glx2 putative hydrolase	11168	Q9PPB1_CA	78.182	2.90E-15	55
			5P0090	248	cj0810 Nade	11168	NADE_CAMJ	74.07	1.20E-59	243
			5P0091	164	cj0811 Lpxk tetraacyldisaccharide kinas	11168	LPXK_CAMJ	82.31	5.90E-52	164
4g04.p	8	1892	5P0092c	271	Mu-like prophage I protein	RM1221	Q5HWR8_CA	99.26	3.80E-91	271
			5P0093c	144	hypothetical	RM1221	Q5HWR9_CA	100	6.10E-52	144
			5P0094	131	hypothetical	RM1221	Q5HWS1_CA	96.12	7.20E-51	129
6c04.p	8	1547	5P0095	515	Cmgb3/4	*C. jejuni* pTet	Q69BA6_CAI	96.89	9.50E-195	515
2g11.q	8	1062	5P0096c	152	hypothetical	RM1221	Q5HVS2_CA	100	1.80E-58	152
			5P0097c	123	hypothetical	RM1221	Q5HVS4_CA	100	2.10E-49	123
8d12.p	9	2441	5P0098	412	prophage muso1 f protein	RM1221	Q5HWR1_CA	100	1.50E-148	412
			5P0099	124	phage tail protein	RM1221	Q5HWR0_CA	100	9.10E-45	124
			5P0100	140	tail protein D	RM1221	Q5HWQ8_CA	99.28	2.80E-49	140
2c11.q	10	1275	5P0101c	63	hypothetical	RM1221	Q5HWS7_CA	98.41	8.70E-23	63
			5P0102c	210	base plate assembly protein V	RM1221	Q5HS6_CAM	98.57	4.20E-75	210
			5P0103c	86	hypothetical	no matches				
2e10.p	10	1854	5P0104c	149	hypothetical	*C.jejuni* strain rm 1221	Q8RN33_CA	97.84	1.10E-50	139
			5P0105c	391	transporter	*C.jejuni* strain rm 1221	Q5HSN2_CA	98.72	1.20E-139	391
7c10.p	11	1615	5P0106	501	hypothetical phage protein	RM1221	Q5HWR3_CA	100	7.20E-174	442
6a01.q	12	1925	5P0107c	87	cpp23	*C. jejuni* pTet	Q69BB4_CAI	97.7	4.30E-30	87
			5P0108c	409	cpp22 (TraC like)	*C. jejuni* pTet	Q69BB5_CAI	85.92	2.60E-128	412
2f11.q	12	2036	5P0109	298	sialic acid synthase	*C.jejuni* strain oh4384	Q9LAK2_CAI	99.66	2.00E-115	298
			5P0110	374	NeuC1	*C.jejuni* strain atcc43456	Q93D03_CAI	98.66	7.20E-132	374

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
5a06.p	12	3535	5P0111c	238	Cmgb3/4 (virB4)	*C. coli*	Q69BF6_CAI	93.25	7.50E-87	237
			5P0112c	87	Cmgb2 (VirB2)	*C. jejuni* pTet	Q69BA7_CAI	90.8	7.20E-27	87
			5P0113	107	cpp29 hypothetical	*C. jejuni* pTet	Q69BA8_CAI	99.07	2.20E-41	107
			5P0114	125	virulence-associated protein d	*C. jejuni* pTet	Q69BA9_CAI	98.4	1.90E-45	125
			5P0115	204	site-specific recombinase	*C. jejuni* pTet	Q69BB0_CAI	99.02	1.40E-69	204
			5P0116c	286	cpp26 hypothetical	*C. jejuni* pTet	Q69BB1_CAI	94.38	1.70E-97	285
7d08.q	12	2314	5P0117c	269	cj0021c hypothetical	11168	Q9PJ90_CAM	85.82	1.20E-90	268
			5P0118c	298	cj0022c ribosomal pseudouridine syntha	11168	Q9PJ89_CAM	82.37	2.60E-94	295
			5P0119	130	cj0023 purb	11168	Q9PJ88_CAM	93.7	4.60E-41	127
1d01.q	13	2884	5P0120c	844	cpp14 hypothetical	*C. jejuni* pTet	Q69BC2_CA	99.39	0.00E+00	824
			5P0121c	88	cpp13 hypothetical	*C. coli*	Q69BH3_CA	100	1.10E-29	88
6g02.q	14	2648	5P0122c	120	cj0304c BioC	11168	Q9PIJ5_CAM	74.16	3.40E-33	120
			5P0123c	203	cj0305c hypothetical	11168	Q9PIJ4_CAM	68.47	1.10E-51	203
			5P0124c	380	cj0306c BioF	11168	Q9PIJ3_CAM	75.78	4.70E-111	380
			5P0125	156	cj0307 BioA	11168	Q9PIJ2_CAM	96.15	1.40E-60	156
2e12.p	14	3074	5P0126	198	site-specific DNA-methyltransferase	RM1221	Q5HVW9_CA	90.91	8.40E-70	198
			5P0127	117	hypothetical	RM1221	Q5HTH9_CA	100	1.17E+02	
			5P0128c	391	site-specific recombinase	RM1221	Q5HTI1_CAM	100	1.60E-143	391
			5P0129c	144	hypothetical	no matches				
7h09.p	15	3530	5P0130	309	Cgta-II (disrupted)	*C. jejuni* strain atcc 43449	Q934C5_CAI	99.68	5.90E-125	309
			5P0131	245	NeuA1	*C. jejuni* strain atcc 43438	Q93MP7_CA	97.28	6.80E-81	221
			5P0132	277	acetyltransferase (disrupted)	*C. jejuni* strain atcc 43446	Q9L9Q2_CAI	97.83	1.10E-103	277
			5P0133c	270	WaaV	*C. jejuni* lio87	Q6T5A5_CAI	95.17	2.70E-102	269
			5P0134	109	WaaF	*C. jejuni* strain nctc 11828	Q6TDC6_CA	97.96	2.90E-34	98
3e03.p	15	2697	5P0135	315	cj0259 Pyrc	11168	Q9PIN6_CAM	77.84	2.40E-97	316
			5P0136	576	DNA methyltransferase	RM1221	Q5HWK5_CA	97.24	6.30E-209	579
3e06.p	16	2317	5P0137	211	hypothetical	RM1221	Q5HTE9_CA	99.05	6.30E-67	211
			5P0138	127	hypothetical	RM1221	Q5HTF0_CA	97.64	1.30E-40	127
			5P0139	124	hypothetical	RM1221	Q5HVS5_CA	99.19	5.20E-52	124
			5P0140	294	hypothetical	RM1221	Q5HTF2_CA	99.66	4.50E-96	294
3c11.q	17	3353	5P0141c	51	hypothetical	no matches				
			5P0142c	704	hypothetical	*Helicobacter hepaticus*	Q7VI58_HEL	40.29	6.20E-81	752
			5P0143c	103	hypothetical	no matches				

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
1c03.q	19	3151	5P0144c	90	hypothetical	RM1221	Q5HWP9_CA	96.66	1.40E-27	90
			5P0145c	161	gam protein (phage)	*C. coli*	Q9K5D6_CA	98.75	8.90E-53	160
			5P0146c	112	hypothetical	*C. coli*	Q9K5D7_CA	94.64	5.80E-37	112
			5P0147c	143	hypothetical	no matches				
			5P0148c	242	hypothetical	*Helicobacter hepaticus*	Q7VI56_HEL	48.73	5.60E-36	236
8g04.q	21	2356	5P0149c	130	phage terminase	RM1221	Q5HTC7_CA	98.46	8.90E-40	130
			5P0150c	113	HNH endonuclease domain protein	RM1221	Q5HTC6_CA	92.04	3.90E-41	113
			5P0151c	96	hypothetical	RM1221	Q5HTC5_CA	100	2.10E-37	96
			5P0152	174	hypothetical	no matches				
5e02.q	23	2304	5P0153	297	Cmgb5 (virB5)	*C. jejuni* pTet	Q69BA1_CA	98.65	1.90E-98	297
			5P0154	332	Cmgb6 (virB6)	*C. coli*	Q69BF0_CA	85.46	3.60E-98	330
			5P0155	55	Cmbg7 (virB7)	*C. jejuni* pTet	Q69B99_CA	100	8.50E-23	55
			5P0156	89	Cmgb8 (virB8)	*C. jejuni* pTet	Q847A8_CA	100	2.30E-31	89
3g09.q	24	4636	5P0157c	110	hypothetical	RM1221	Q5HWQ0_C	100	5.30E-43	110
			5P0158c	90	hypothetical	RM1221	Q5HWP9_C	100	1.90E-28	90
			5P0159c	161	gam protein	RM1221	Q5HWP7_C	100	3.70E-53	161
			5P0160c	113	hypothetical	*C. coli* plasmid pBT9810	Q9K5D7_CA	95.57	2.30E-38	113
			5P0161c	307	DNA transposition protein B	RM1221	Q5HWP3_C	97.07	1.40E-103	307
			5P0162c	419	DNA transposition protein A	RM1221	Q5HWP2_C	97.85	1.40E-145	419
3b02.p	24	2873	5P0163	730	Cpp49 (VirB8)	*C. coli*	Q69BD8_CA	98.77	0.00E+00	730
			5P0164	141	Cpp50 hypothetical	*C. coli*	Q69BD7_CA	100	5.00E-50	141
5d09.p	28	2775	5P0165c	617	TetO	*C. jejuni* pTet	Q69BD5_CA	99.83	0.00E+00	617
			5P0166c	59	hypothetical Cpp51	*C. coli*	Q69BD6_CA	100	5.90E+01	
			5P0167c	113	hypothetical Cpp50	*C. coli*	Q69BD7_CA	100	2.20E-38	112
2f06.p	39	6771	5P0168	198	Cpp18 hypothetical	*C. coli*	Q69BG8_CA	100	2.40E-54	183
			5P0169	462	(cpp17) nickase MagA2	*C. coli*	Q69BG9_CA	98.92	2.10E-164	462
			5P0170	234	Cpp16 hypothetical	*C. coli*	Q69BH0_CA	100	8.00E-93	234
			5P0171c	242	Cpp15 hypothetical	*C. coli*	Q69BH1_CA	100	1.80E-90	242
			5P0172c	1057	Cpp14 hypothetical	*C. coli*	Q69BH2_CA	100	0.00E+00	1054
3c07.q	41	5026	5P0173	206	Virb9-like protein	*C. jejuni* plasmid pCjA13	Q847A7_CA	100	1.90E-76	206
			5P0174	398	Cmgb10 (VirB10)	*C. jejuni* pTet	Q69B96_CA	100	1.00E-143	398
			5P0175	330	Virb11-like protein	*C. jejuni* pTet	Q69B95_CA	100	4.40E-119	348
			5P0176	603	MagB12 (virD4)	*C. jejuni* pTet	Q69B94_CA	100	0.00E+00	603
			5P0177	145	Cpp44 cag island protein	*C. jejuni* pTet	Q69B93_CA	100	2.70E-54	145

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
7c07.p	43	4998	5P0178c	62	hypothetical	RM1221	Q5HTG6_CA	95.16	4.10E-20	62
			5P0179c	95	hypothetical	RM1221	Q5HTG5_CA	100	5.10E-37	95
			5P0180c	244	dna binding protein Roi	RM1221	Q5HTG4_CA	97.54	6.80E-76	244
			5P0181c	94	hypothetical	RM1221	Q5HTG3_CA	95.75	1.10E-26	94
			5P0182	130	hypothetical	no matches				
			5P0183	105	hypothetical	no matches				
			5P0184	326	hypothetical	*Lactobacillus johnsonii*	Q74HW8_LA	30.2	2.00E-13	308
			5P0185	206	hypothetical	no matches				
			5P0186	71	hypothetical	no matches				
7f04.p	48	6568	5P0187c	740	tail tape measure protein	RM1221	Q5HWU0_CA	98.92	0	738
			5P0188	108	hypothetical	RM1221	Q5HWT9_CA	98.15	1.20E-37	108
			5P0189c	104	hypothetical	RM1221	Q5HWT7_CA	100	4.80E-23	79
			5P0190c	169	major tail tube protein	RM1221	Q5HWt6_CA	100	1.10E-60	169
			5P0191c	397	major tail sheath protein	RM1221	Q5HWT5_CA	98.24	7.30E-147	397
			5P0192c	335	hypothetical protein	RM1221	Q5HWT4_CA	94.93	1.70E-120	335
			5P0193c	128	hypothetical	RM1221	Q5HWT3_CA	90.08	8.70E-47	121
			5P0194c	104	hypothetical	RM1221	Q5HWT2_CA	95.15	5.00E-37	103
7b11.p	55	5186	5P0195	104	hypothetical	RM1221	Q5HWS5_CA	98.08	8.20E-34	104
			5P0196	508	hypothetical	Bacteriophage D3112	Q6TM76_BP	29.48	1.50E-22	502
			5P0197	460	hypothetical	*Shewanella oneidensis*	Q8EDR3_SH	21.27	1.60E-08	470
			5P0198	377	prophage muso1 F protein	RM1221	Q5HWR1_CA	27.67	8.40E-20	365
			5P0199c	167	phage virion morphogenesis protein	RM1221	Q5HWU1_CA	28.74	1.20E-04	167
5g07.q	51	7892	5P0200	81	hypothetical	no matches				
			5P0201c	160	phage virion morphogenesis protein	RM1221	Q5HWU1_CA	28.57	8.50E-05	168
			5P0202c	142	hypothetical	no matches				
			5P0203c	86	hypothetical	no matches				
			5P0204c	128	hypothetical	RM1221	Q5HWQ0_CA	100	4.00E-50	128
			5P0205c	90	hypothetical	RM1221	Q5HWP9_CA	98.89	5.20E-28	90
			5P0206c	161	host-nuclease inhibitor protein gam	RM1221	Q5HWP7_CA	100	3.70E-53	161
			5P0207c	112	hypothetical	*C. coli*	Q9K5D7_CA	94.64	5.80E-37	112
			5P0208c	143	hypothetical	no matches				
			5P0209c	285	transposition protein	*Helicobacter hepaticus*	Q7VI56_HEL	46.02	1.00E-41	289
			5P0210c	705	DNA transposition protein A	RM1221	Q5HWP2_CA	27.14	1.10E-17	689

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
2b12.p	82	9176	5P0211c	113	hypothetical	RM1221	Q5HWS2_CA	97.35	1.10E-36	113
			5P0212c	129	hypothetical	RM1221	Q5HWS1_CA	99.23	3.00E-52	129
			5P0213c	121	hypothetical	no matches				
			5P0214c	117	hypothetical	no matches				
			5P0215c	89	hypothetical	no matches				
			5P0216c	295	major head subunit	Bacteriophage D3112	Q6TM67_BP	35.59	1.50E-14	295
			5P0217c	346	hypothetical	no matches				
			5P0218	154	hypothetical	no matches				
			5P0219	210	baseplate assembly protein V	RM1221	Q5HWS6_CA	99.52	4.60E-76	210
			5P0220	63	hypothetical	RM1221	Q5HWS7_CA	98.41	8.70E-23	63
			5P0221	96	baseplate assembly protein w	*C. coli*	Q9K5E0_CA	97.92	7.10E-34	96
			5P0222	388	baseplate assembly protein J	RM1221	Q5HWS9_CA	99.49	1.50E-129	388
			5P0223	206	phage tail protein	RM1221	Q5HWT0_CA	93.69	7.90E-72	206
			5P0224	343	tail fibre protein H	RM1221	Q5HWT1_CA	75.29	8.90E-80	340
			5P0225	168	hypothetical	RM1221	Q5HWT2_CA	95.83	3.20E-56	168
			5P0226	69	hypothetical	RM1221	Q5HWT3_CA	98.55	9.00E-29	69
4h04.p	102	8165	5P0227c	107	hypothetical	RM1221	Q5HTE8_CA	97.26	2.50E-25	73
			5P0228c	521	hypothetical	RM1221	Q5HTE7_CA	99.62	1.30E-167	521
			5P0229c	210	hypothetical	RM1221	Q5HTE6_CA	99.52	6.20E-67	210
			5P0230c	107	phage head-tail adaptor	RM1221	Q5HTE5_CA	100	6.40E-39	105
			5P0231c	145	hypothetical	RM1221	Q5HTE4_CA	100	1.60E-37	104
			5P0232c	83	hypothetical	RM1221	Q5HTE2_CA	100	7.90E-25	83
			5P0233c	388	major capsid protein, hk97 family	RM1221	Q5HTE1_CA	100	4.00E-136	388
			5P0234c	185	hypothetical	RM1221	Q5HTE0_CA	100	3.50E-64	185
			5P0235c	289	hypothetical	RM1221	Q5HTD9_CA	100	7.20E-119	289
			5P0236c	639	hypothetical	RM1221	Q5HTD8_CA	99.53	5.30E-185	639

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
7e05.p	124	10407	5P0237	67	phage terminase, small subunit	RM1221	Q5HTC7_CA	83.08	4.30E-17	65
			5P0238	541	phage terminase, large subunit	RM1221	Q5HTC8_CA	100	0.00E+00	541
			5P0239	144	toxin-antitoxin protein	RM1221	Q5HTC9_CA	98.61	1.20E-51	144
			5P0240	390	portal protein, hk97 family	RM1221	Q5HTD0_CA	100	1.20E-143	390
			5P0241	188	phage protein, hk97 gp10 family	RM1221	Q5HTD1_CA	100	7.90E-61	180
			5P0242	116	hypothetical	RM1221	Q5HTD2_CA	99.14	7.50E-39	116
			5P0243	326	hypothetical	RM1221	Q5HTD3_CA	99.39	7.20E-114	326
			5P0244	118	hypothetical	RM1221	Q5HTD4_CA	100	2.10E-37	118
			5P0245	71	hypothetical	RM1221	Q5HTD5_CA	100	1.80E-25	71
			5P0246	124	hypothetical	RM1221	Q5HTD7_CA	97.67	7.30E-13	43
			5P0247	1224	hypothetical	RM1221	Q5HTD8_CA	94.2	0	1224
3b03.q	175	15477	5P0248	222	phage repressor protein	RM1221	Q5HWU7_CA	97.61	8.90E-79	209
			5P0249	106	hypothetical protein	RM1221	Q5HWU6_CA	97.17	1.10E-31	106
			5P0250	95	hypothetical	RM1221	Q5HWU3_CA	96.67	5.80E-18	60
			5P0251c	276	dam DNA adenine methylase	RM1221	Q5HWU2_CA	98.52	1.30E-103	271
			5P0252c	322	tail protein d	RM1221	Q5HWQ8_CA	47.1	2.50E-49	327
			5P0253c	124	phage tail protein	RM1221	Q5HWR0_CA	57.26	2.50E-25	124
			5P0254c	654	tail tape measure protein, tp901 family	RM1221	Q5HWU0_CA	26.06	4.70E-22	765
			5P0255c	78	hypothetical	RM1221	Q5HWT7_CA	31.51	1.30E-02	73
			5P0256c	171	major tail tube protein	RM1221	Q5HWt6_CA	41.92	1.90E-20	167
			5P0257c	396	major tail sheath protein	RM1221	Q5HWT5_CA	96.97	2.10E-144	396
			5P0258c	337	hypothetical	RM1221	Q5HWT4_CA	98.52	2.20E-123	337
			5P0259c	123	hypothetical	RM1221	Q5HWT3_CA	98.37	4.80E-52	123
			5P0260c	168	hypothetical	RM1221	Q5HWT2_CA	95.83	1.20E-55	168
			5P0261c	343	tail fiber protein H	RM1221	Q5HWT1_CA	75.59	1.60E-80	340
			5P0262c	206	tail protein	RM1221	Q5HWT0_CA	91.26	3.30E-70	206
			5P0263c	388	baseplate assembly protein J	RM1221	Q5HWS9_CA	98.2	6.90E-129	388
			5P0264c	96	baseplate assembly protein W	*C. coli*	Q9K5E0_CA	97.92	5.10E-35	96
			5P0265c	63	hypothetical	RM1221	Q5HWS7_CA	100	3.70E-23	63
			5P0266c	210	baseplate assembly protein V	RM1221	Q5HWS6_CA	99.05	3.10E-76	210
			5P0267c	104	hypothetical	RM1221	Q5HWS5_CA	100	2.40E-34	104
7h10.p	1	717	5P0268	237	Cst-II, alpha-2,3-sialyltransferase	*C. jejuni* strain 43432	Q9F0M9_CA	95.28	6.20E-91	233
5h05.p	1	781	5P0269	60	cj0168c periplasmic protein	11168	Q9PIW0_CA	90	3.10E-16	60
			5P0270	23	cj0167c integral membrane protein	11168	Y167_CAMJ	95.65	8.50E-10	23
6e09.p	1	510	5P0271	144	cj1624c sdaa L-serine dehydratase	11168	Q9PM51_CA	96.52	7.90E-52	144
7h07.p	1	689	5P0272c	52	hmcd domain protein	RM1221	Q5HXA6_CA	94	1.70E-17	50
			5P0273	126	hypothetical	RM1221	Q5HXA8_CA	96.15	1.60E-14	52
							Q5HXA7_CA	57.38	3.20E-06	61

contig	no. of reads	length/ bp	systematic_id	length in aa	match	organism with match	swall	% id (aa)	e-value	no. of aa in match
7h08.q	1	438	5P0274c	145	tail protein d	RM1221	Q5HWQ8_CA	97.86	4.70E-49	140
6b02.q	1	678	5P0275c	157	hypothetical cpp32	*C. coli*	Q69BF5_CA	67.68	6.10E-36	164
			5P0276c	64	cmgb3/4	*C. coli*	Q69BF6_CA	93.75	2.60E-18	64
5g10.q	1	752	5P0277c	88	Bll0816 protein (propionate catabolism?	*Bradyrhizobium japonicum*	Q89W77	40	2.10E-05	85
			5P0278c	158	cj1394 fumarate lyase	11168	Q9PMR1_CA	94.93	7.30E-52	158
7e10.q	1	272	5P0279	50	type I RM fragment	uncultured Archaeon	Q64AS4_9A	33.8	6.40E-04	71
5d06.q	1	22			no predicted CDSs					
1b11.p	1	107			no predicted CDSs					
6h07.q	1	32			no predicted CDSs					
1d11.p	1	430			no predicted CDSs					
1f11.p	1	206			no predicted CDSs					
5c01.q	1	47			no predicted CDSs					
6b11.p	1	30			no predicted CDSs					
6d11.p	1	326			no predicted CDSs					
4c05.p	1	302			no predicted CDSs					
5f01.p	1	252			no predicted CDSs					