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.

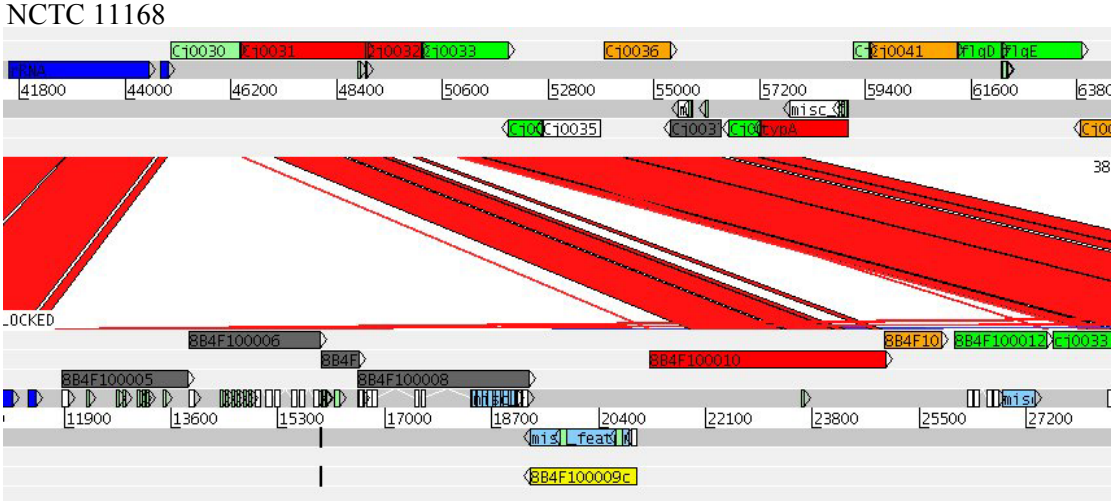
BAC sequences were named in the following way: first the strain designator character, 8 for strain 81-176, M for strain M1, 4 for strain 40671 and 5 for strain 52472; next the library designator character, B for BAC library; then the library plate number

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**).

NCTC 11168



81-176

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**).

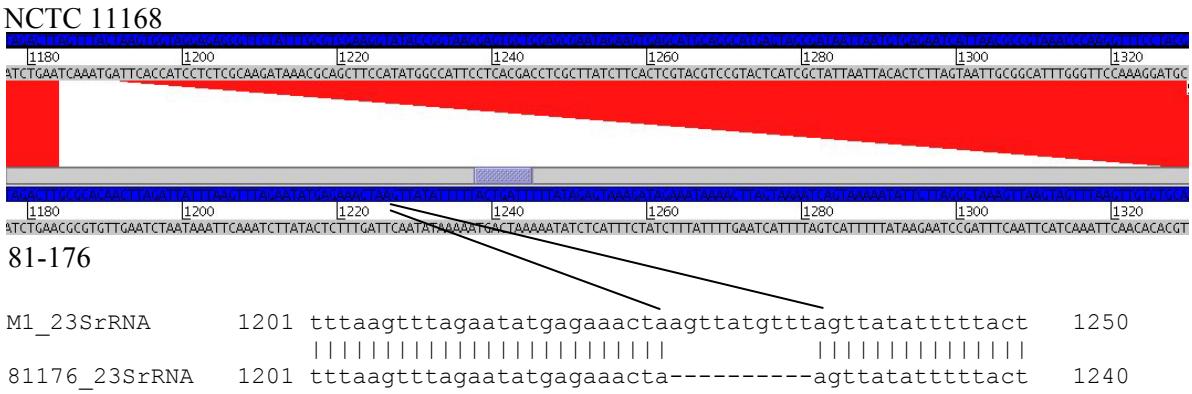


Fig 5.2: Blastn comparison of strain NCTC 11168 and strain 81-176 BAC clone 8B4F10 23S rDNA sequence. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. The sequence of the 23S rDNA is marked by the dark 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.

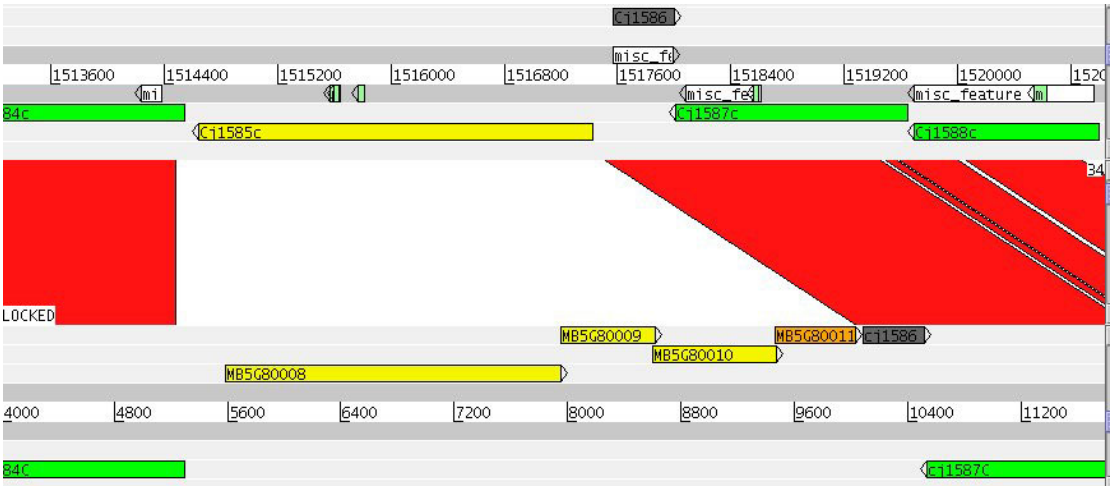
As the sequences in strain 81-176 and M1 are so similar, only the CDSs from strain 81-176 will be discussed further. Downstream of the rDNA there are four predicted cytochrome C associated genes, in the first cytochrome C homologue 8B4F10_5 there are 6 prosite cytochrome c family heme-binding site signatures 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**).

NCTC 11168



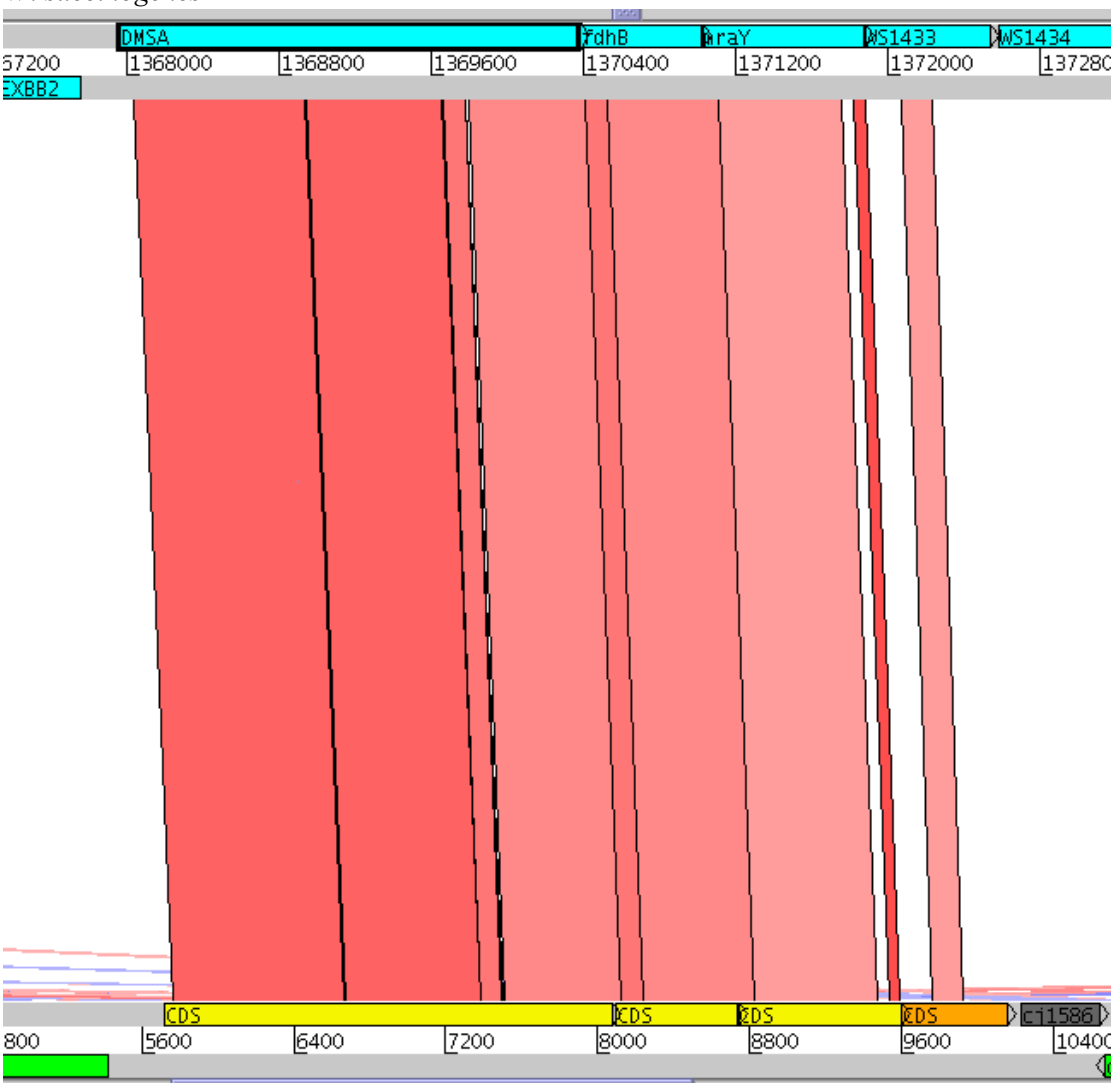
M1

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

W. succinogenes



M1

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**).

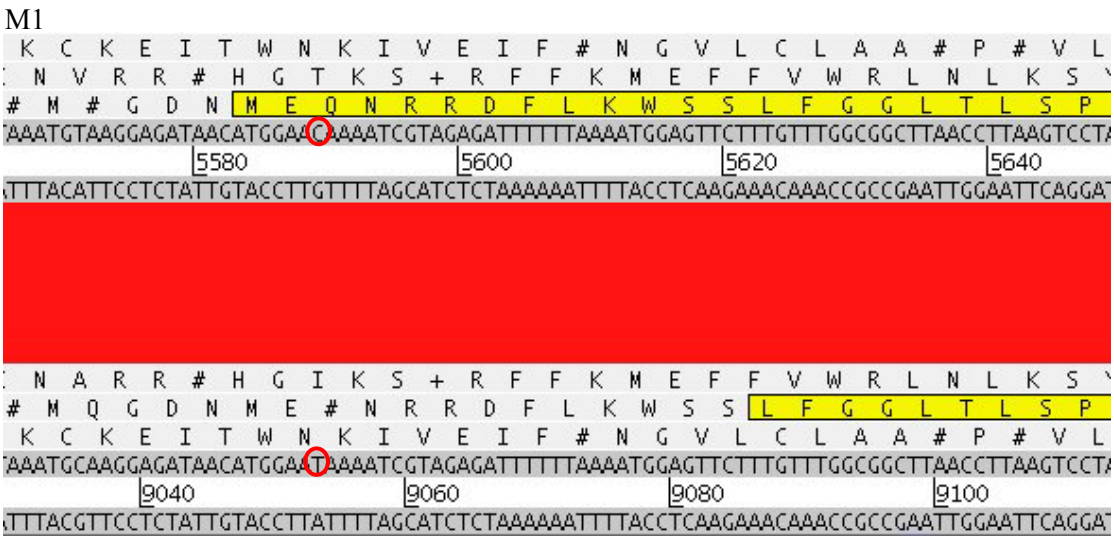


Fig 5.5: Blastn comparison of the sequence from strain M1 and strain 81-176 BAC clones MB5G8 and 8B1E5. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match, however, single base pair differences cannot be accurately 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.

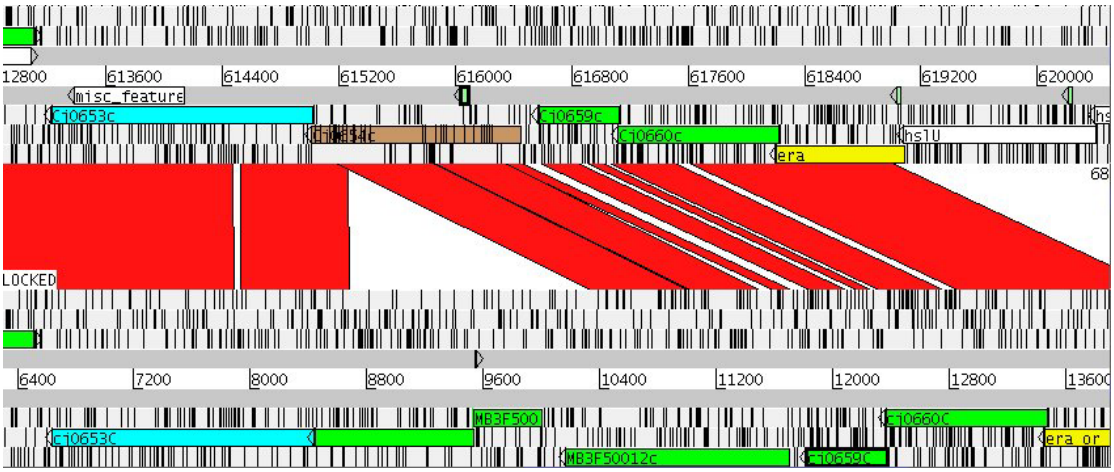
5.2.3 Transport

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).

NCTC 11168



M1

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

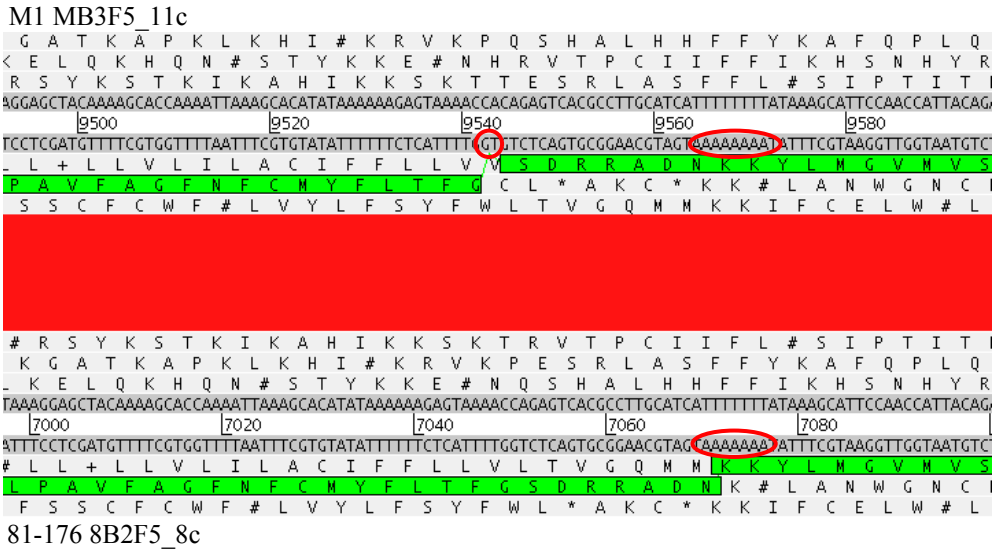


Fig 5.7: Blastn comparison of sequence from strain M1 and strain 81-176 BAC clones MB3F5 and 8B2F5 in the region of MB3F5_11c. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match, however, small regions of 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.

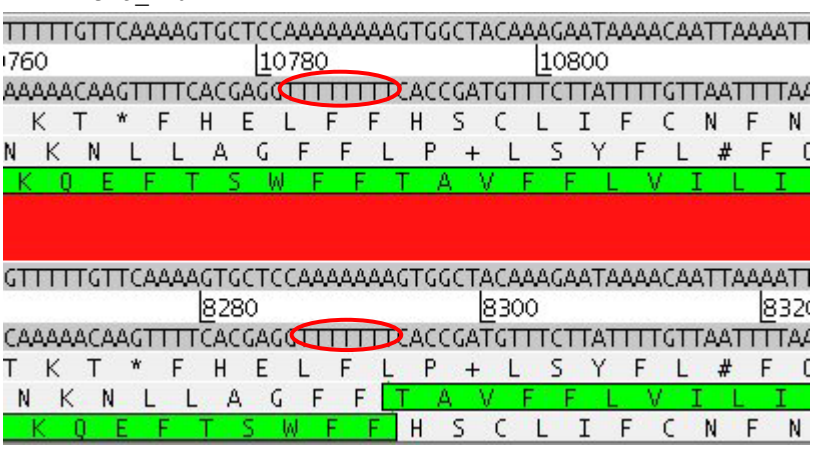


Fig 5.8: Blastn comparison of sequence from strain M1 and strain 81-176 BAC clones MB3F5 and 8B2F5 in the region of MB3F5_12c. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match, however single bp changes 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

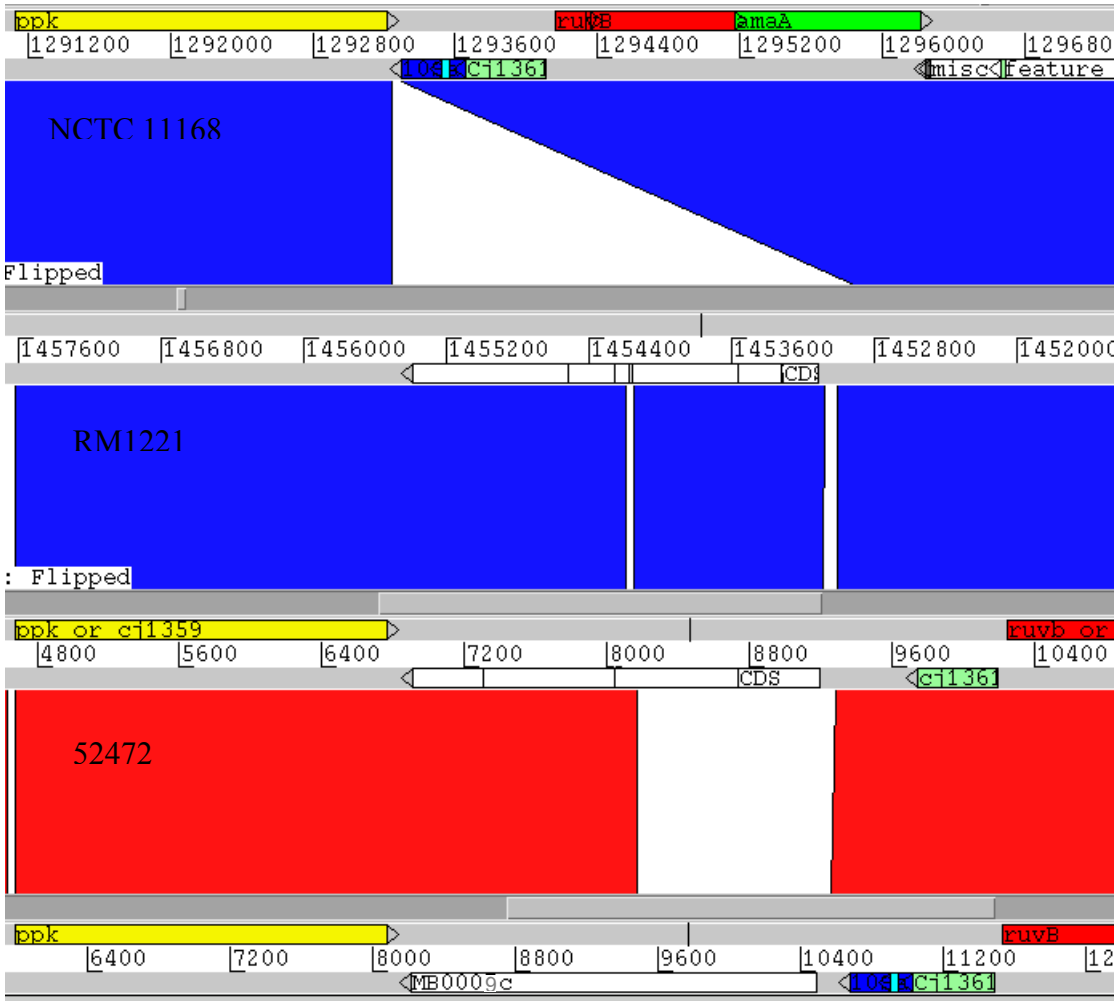
From the pUC assemblies there appeared to be an autotransporter with homology to part of VacA from *Helicobacter pylori* present in strains M1 and 52472, although it was unclear whether this was complete in strain 52472. A probe 5P1h08 was designed to study this region in more depth in strains 52472 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



M1

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.

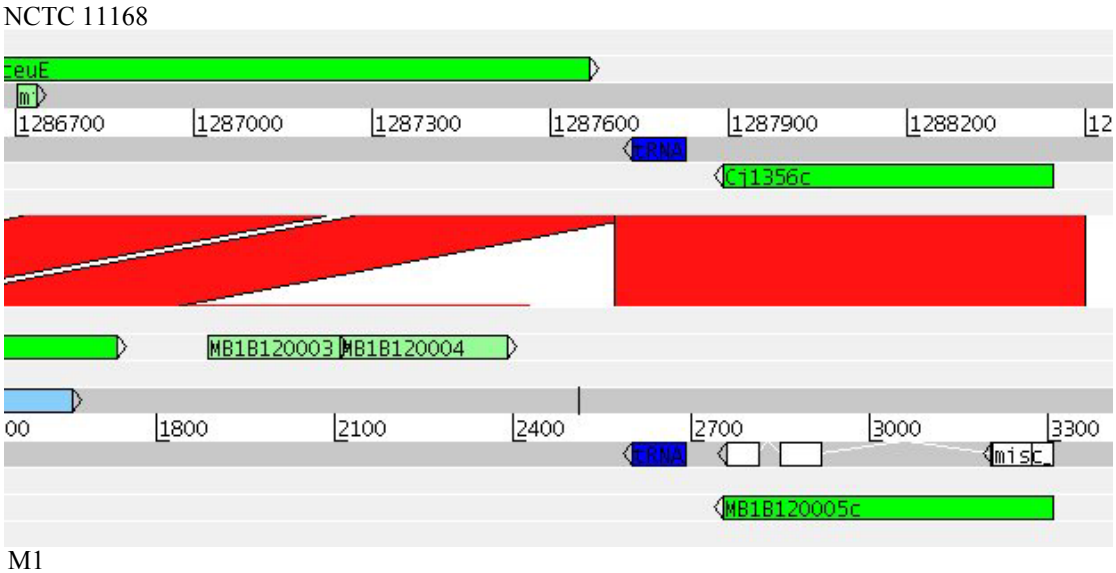


Fig 5.10: Blastn comparison of sequence from strain NCTC 11168 and strain M1 BAC clone MB1B12. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match, however single bp changes cannot be accurately represented. 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.

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

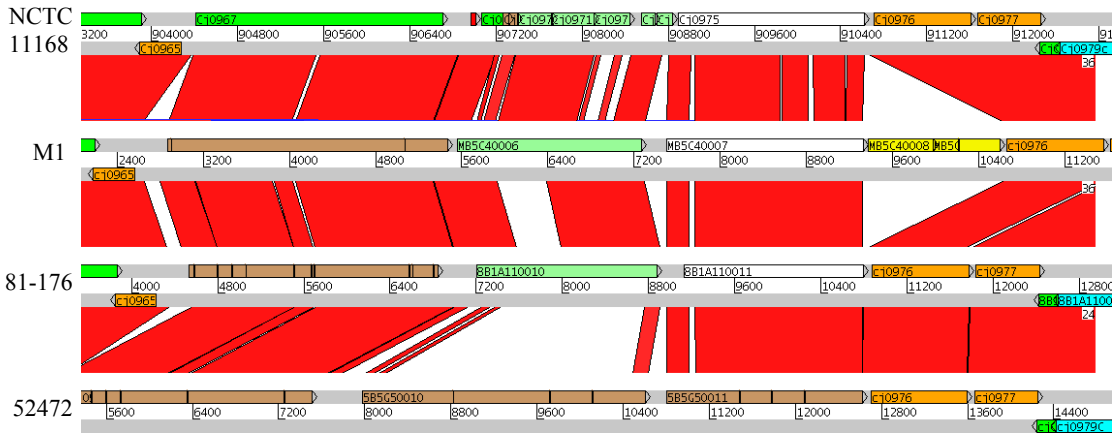


Fig 5.12: Blastn comparison of sequence from strain NCTC 11168, strain M1 BAC clone MB5C4, strain 81-176 BAC clone 8B1A11 and strain 52472 BAC clone 5B5G5. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. 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

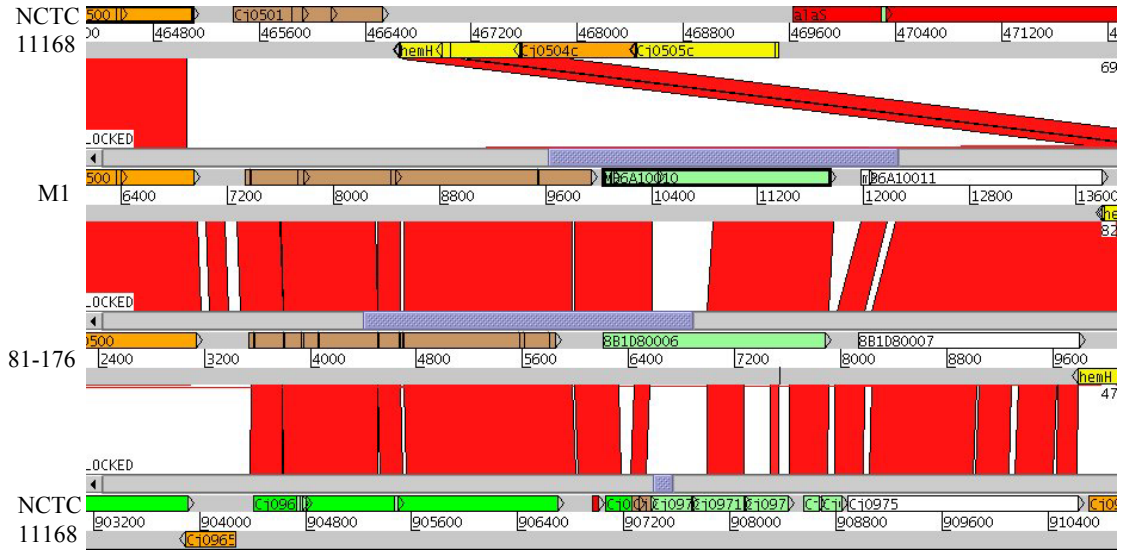


Fig 5.13: Blastn comparison of sequence from strain NCTC 11168, strain M1 BAC clone MB6A1 and strain 81-176 BAC clone 8B1D8. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. CDSs are represented by open 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.

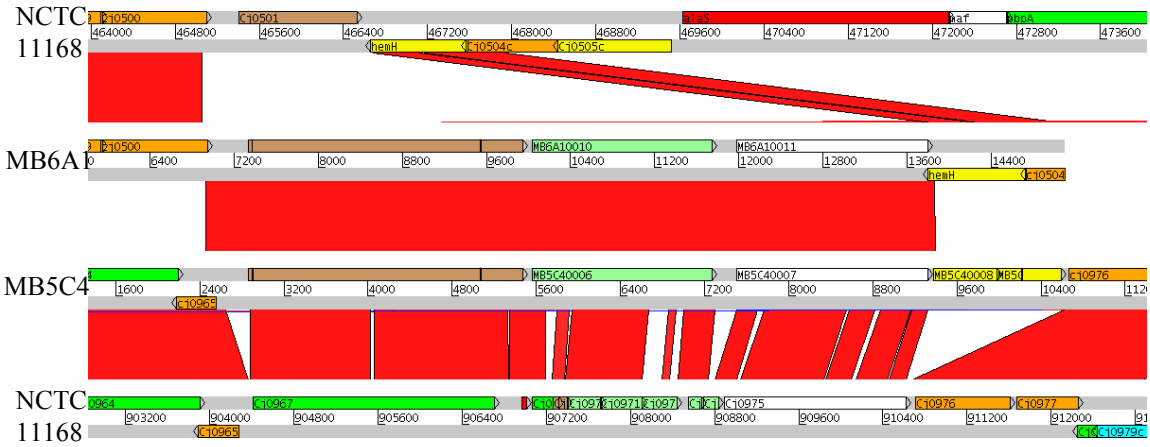


Fig 5.14: Blastn comparison of sequence from strain NCTC 11168 and strain M1 BAC clones MB6A1 and MB5C4. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. CDSs are represented by open boxes and are coloured 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	-			

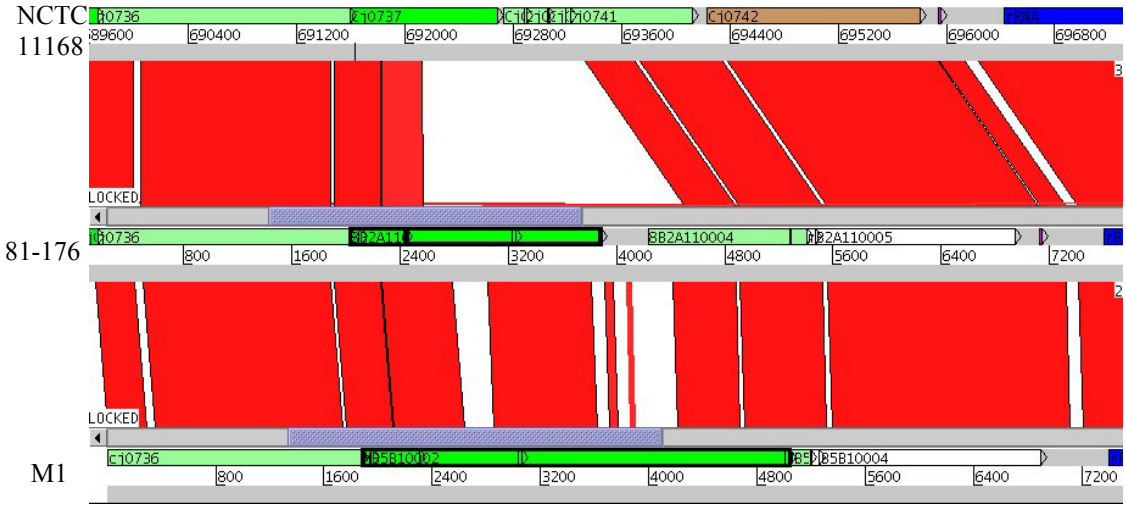


Fig 5.15: Blastn comparison of sequence from strain NCTC 11168, strain 81-176 BAC clone 8B2A11 and strain M1 BAC clone MB5B1. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. CDSs are represented by open 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.

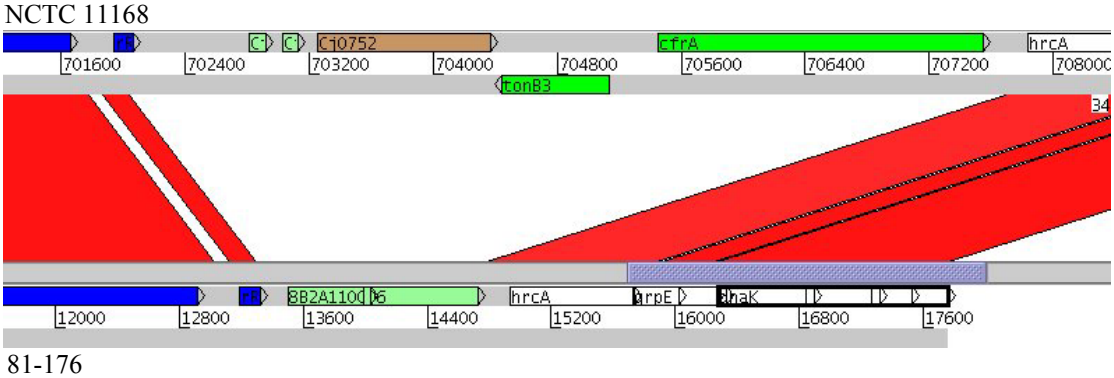


Fig 5.16: Blastn comparison of sequence from strain NCTC 11168 and strain 81-176 BAC clone 8B2A11. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. CDSs are represented by open boxes and are coloured according 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.

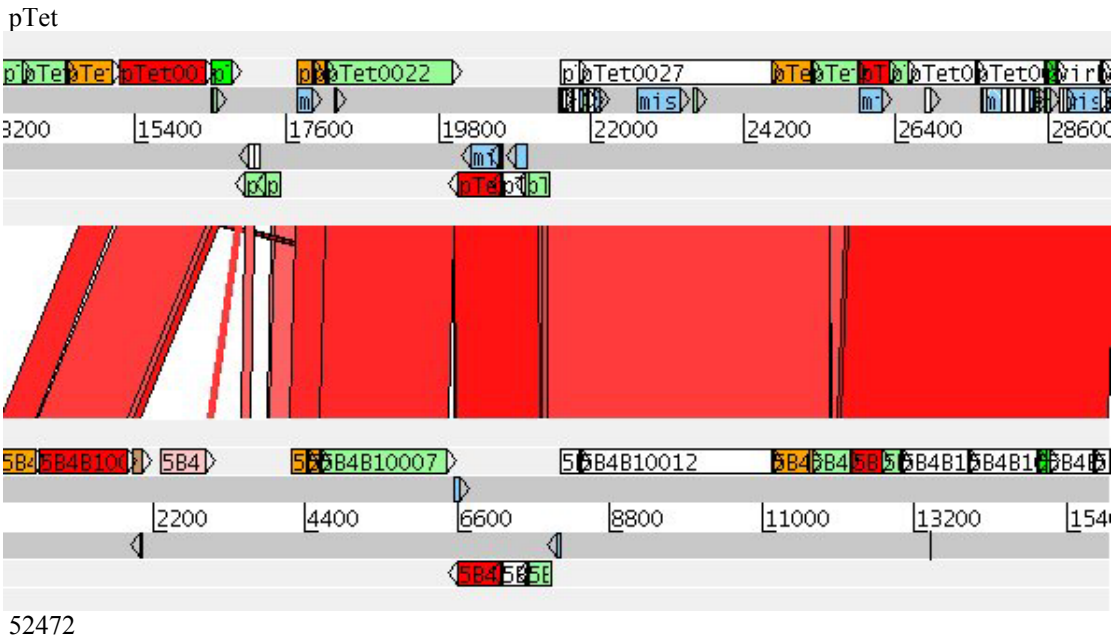


Fig 5.17: tblastx comparison of sequence from pTet and strain 52472 BAC clone 5B4B1. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark grey lines; 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

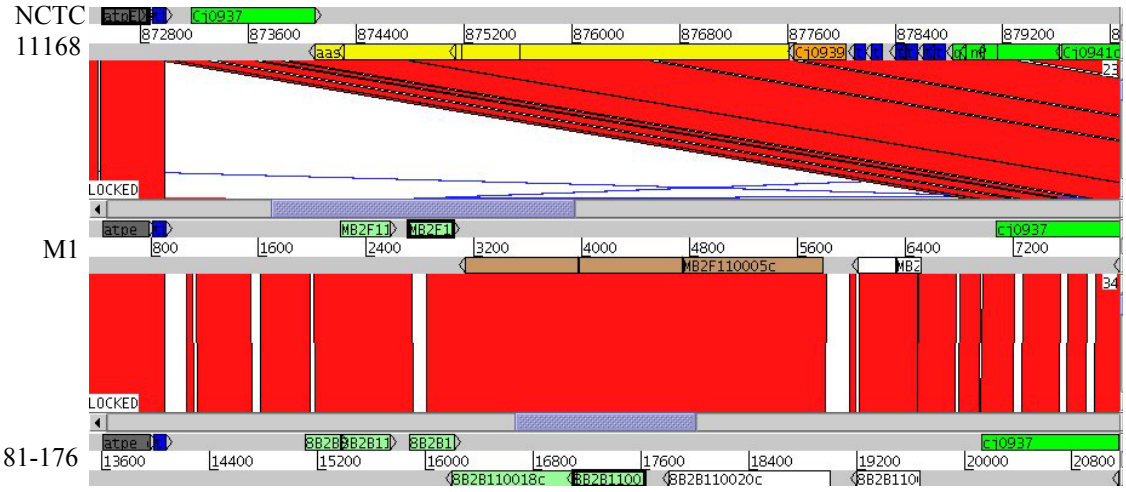


Fig 5.18: Blastn comparison of sequence from strain NCTC 11168, strain M1 BAC clone MB2F11 and strain 81-176 BAC clone 8B2B11. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. CDSs are represented by open 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

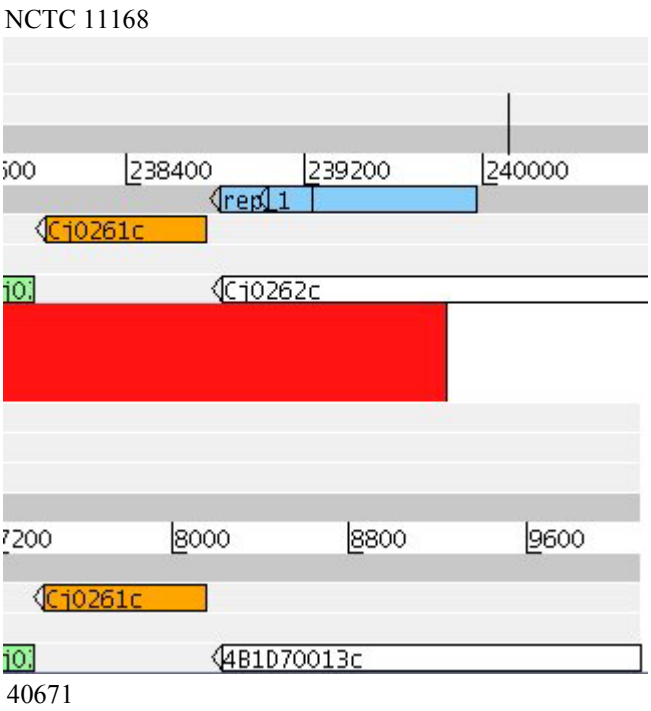


Fig 5.19: Blastn comparison of sequence from strain NCTC 11168 and strain 40671 BAC clone 4B1D7. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark 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.

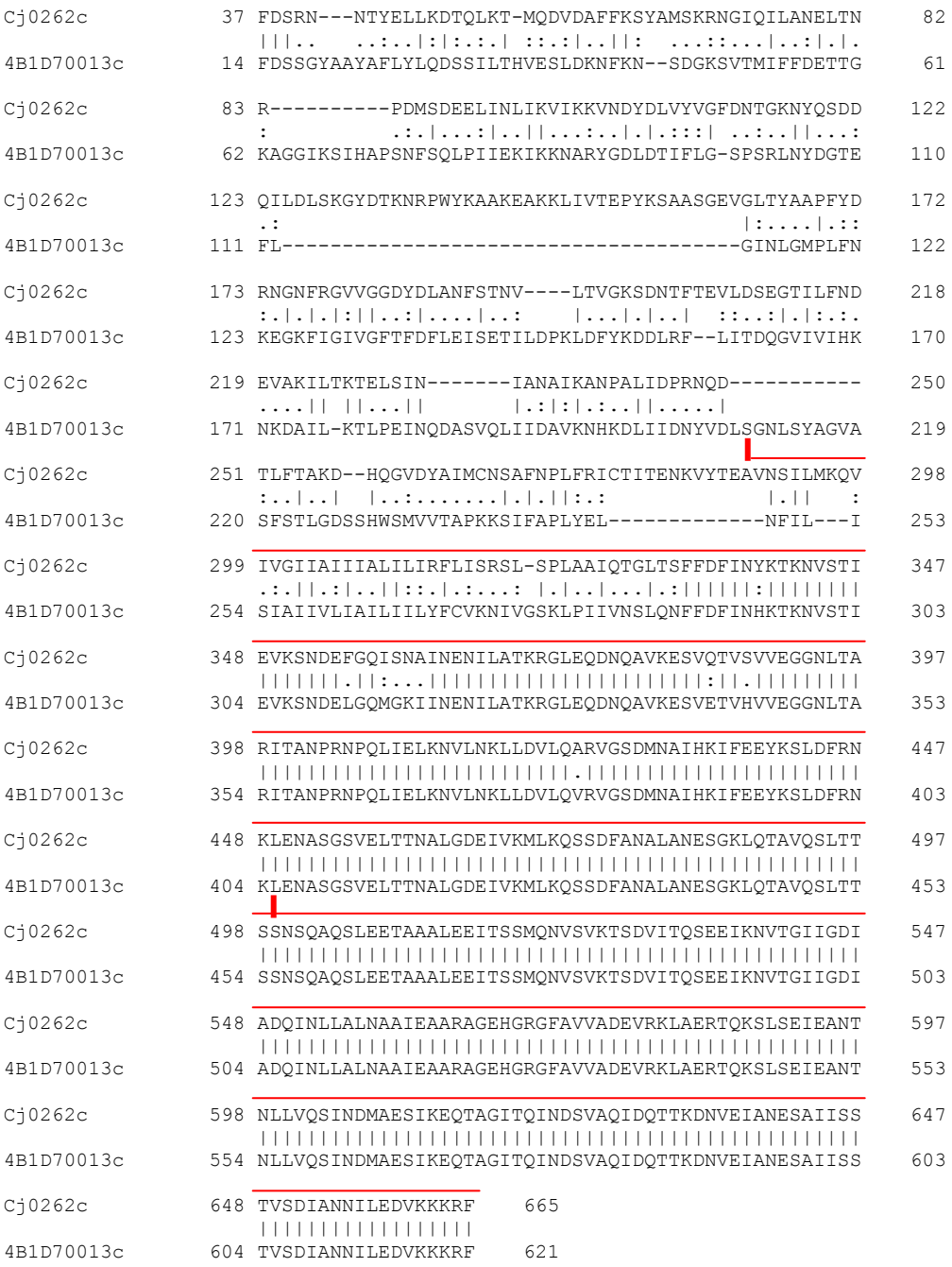


Fig 5.20: Alignment of the predicted MCP-type chemotaxis proteins Cj0262c and 4B1D7_13c. The EMBOSS program 'water' was used to align the sequences using the Smith-Waterman algorithm. The first red mark signifies the beginning of the repeat domain in NCTC 11168 and the second red mark signifies the beginning 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

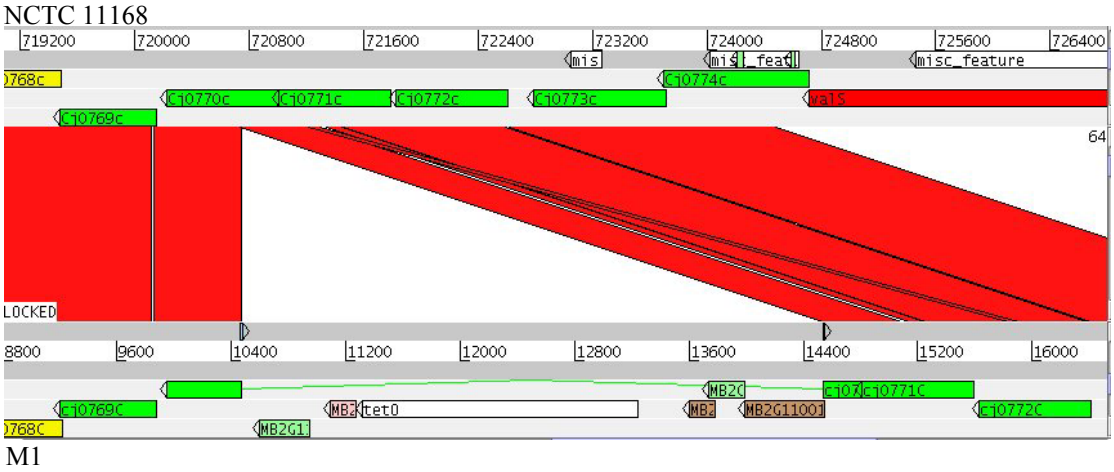


Fig 5.21: Blastn comparison of sequence from strain NCTC 11168 and strain M1 BAC clone MB2G11. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark 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

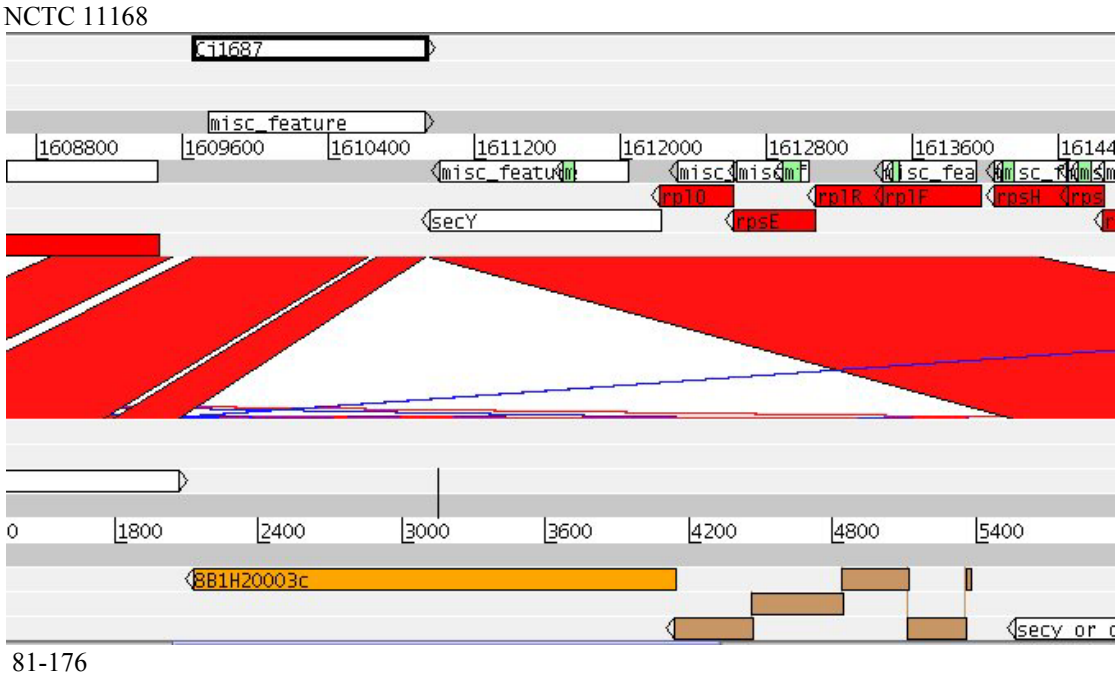


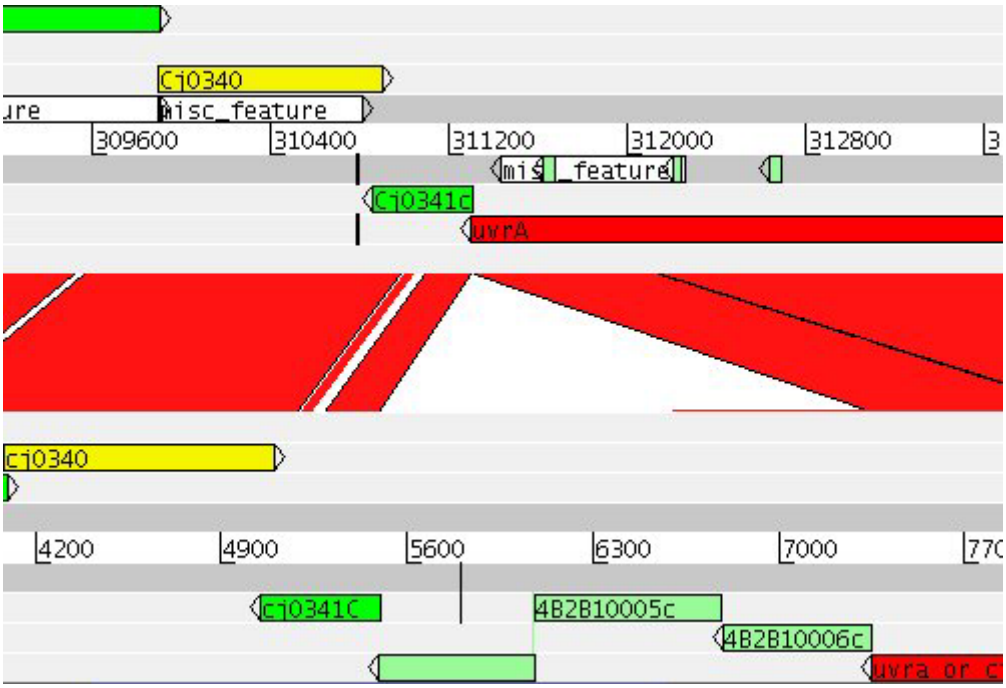
Fig 5.22: Blastn comparison of sequence from strain NCTC 11168 and strain 81-176 BAC clone 8B1H2. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark 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

NCTC 11168



40671

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

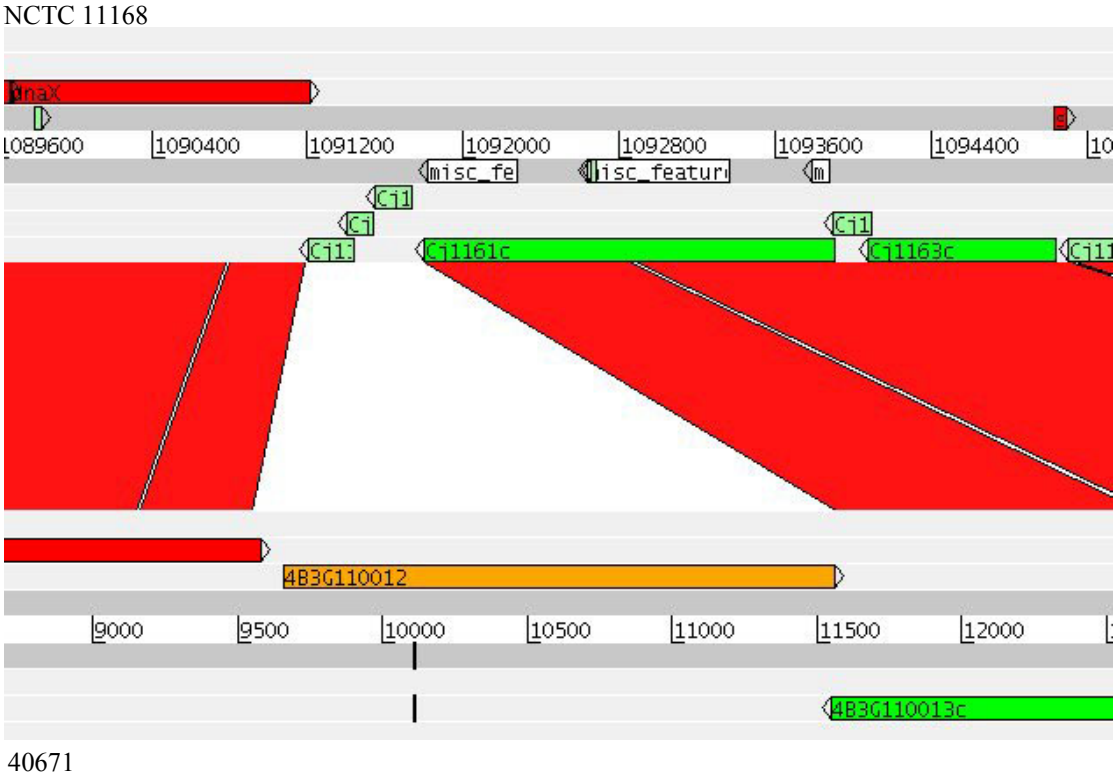


Fig 5.24: Blastn comparison of sequence from strain NCTC 11168 and strain 40671 BAC clone 4B3G11. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark 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

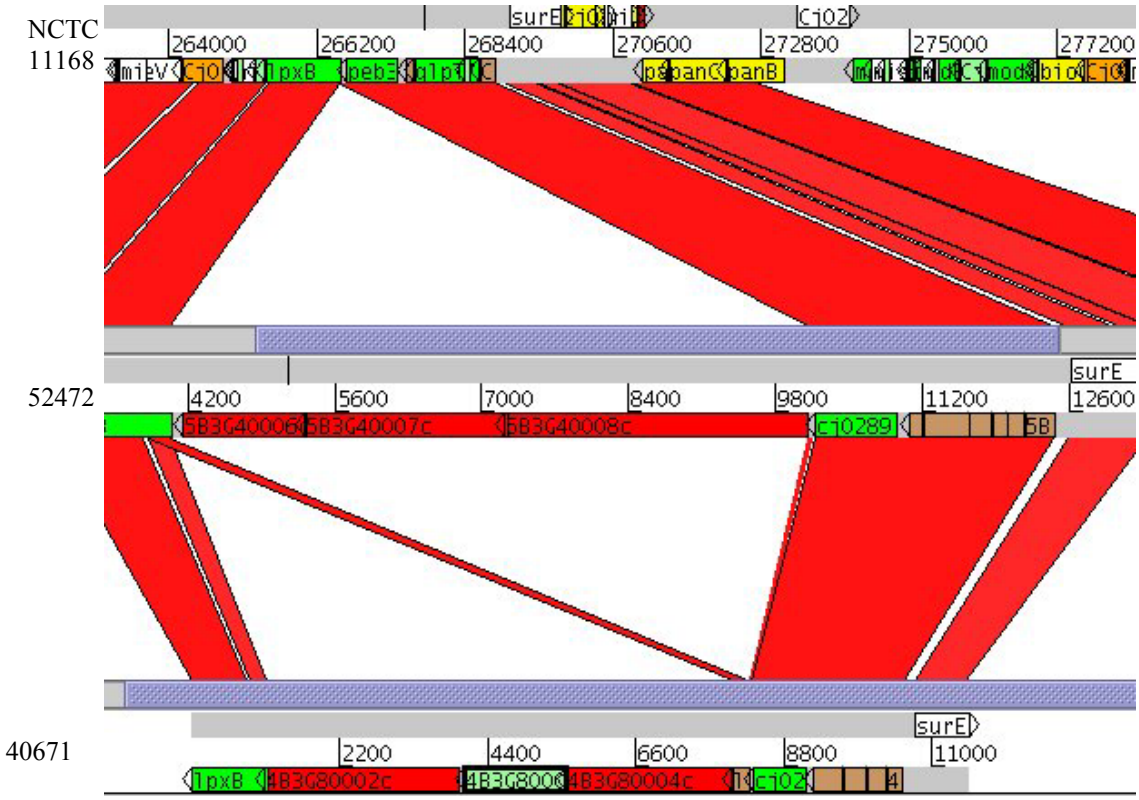


Fig 5.25: Blastn comparison of sequence from strain NCTC 11168, strain 52472 BAC clone 5B3G4 and strain 40671 BAC clone 4B3G8. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. CDSs are represented by open 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

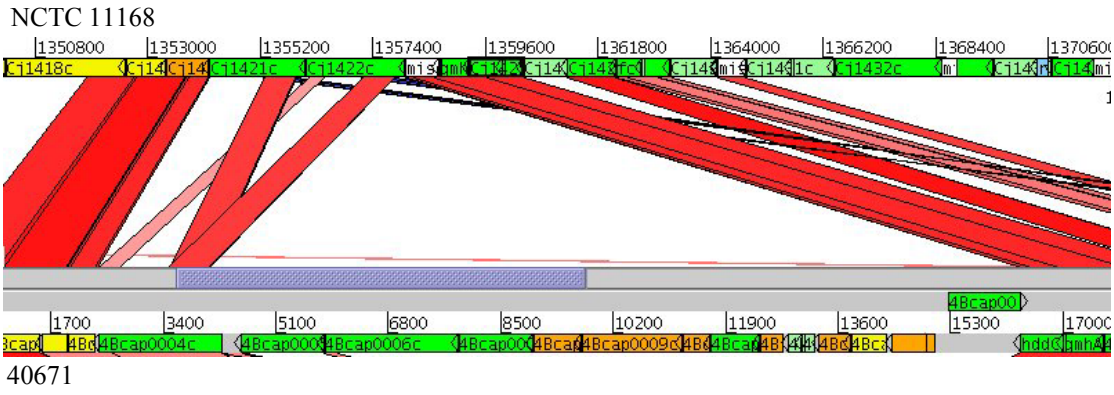


Fig 5.26: tblastx comparison of sequence from strain NCTC 11168 and strain 40671 capsular polysaccharide locus. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. CDSs are represented by open boxes and are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

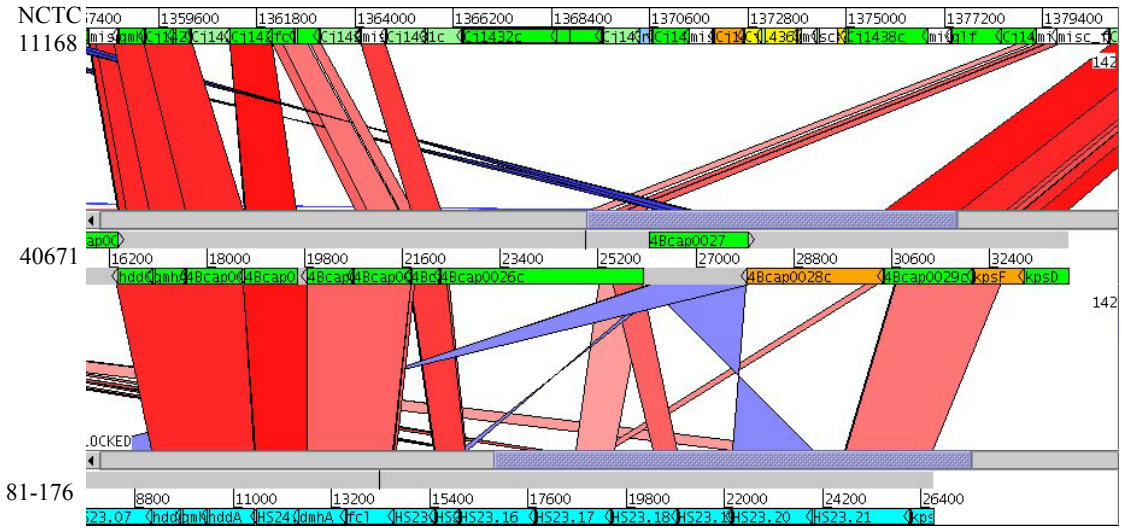


Fig 5.27: tblastx comparison of sequence from strain NCTC 11168, strain 40671 and strain 81-176 capsular polysaccharide locus. The comparison of *hddC* to *kpsF* is viewed using ACT; blocks of red or blue indicate sequence homology with the colour intensity proportional to the percent id of the match. 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

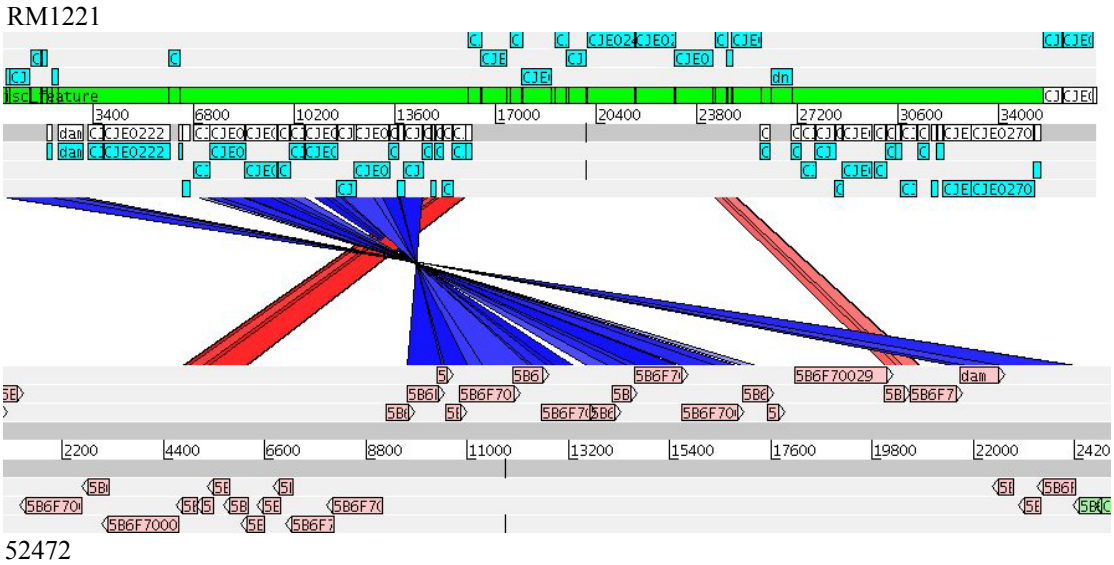


Fig 5.28: tblastx comparison of bacteriophage sequence from strain RM1221 and strain 52472 BAC clone 5B6F7. The comparison is viewed using ACT; blocks of red or blue indicate sequence homology with the colour intensity proportional to the percent id of the match. Blocks of blue indicate that the homologous 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.

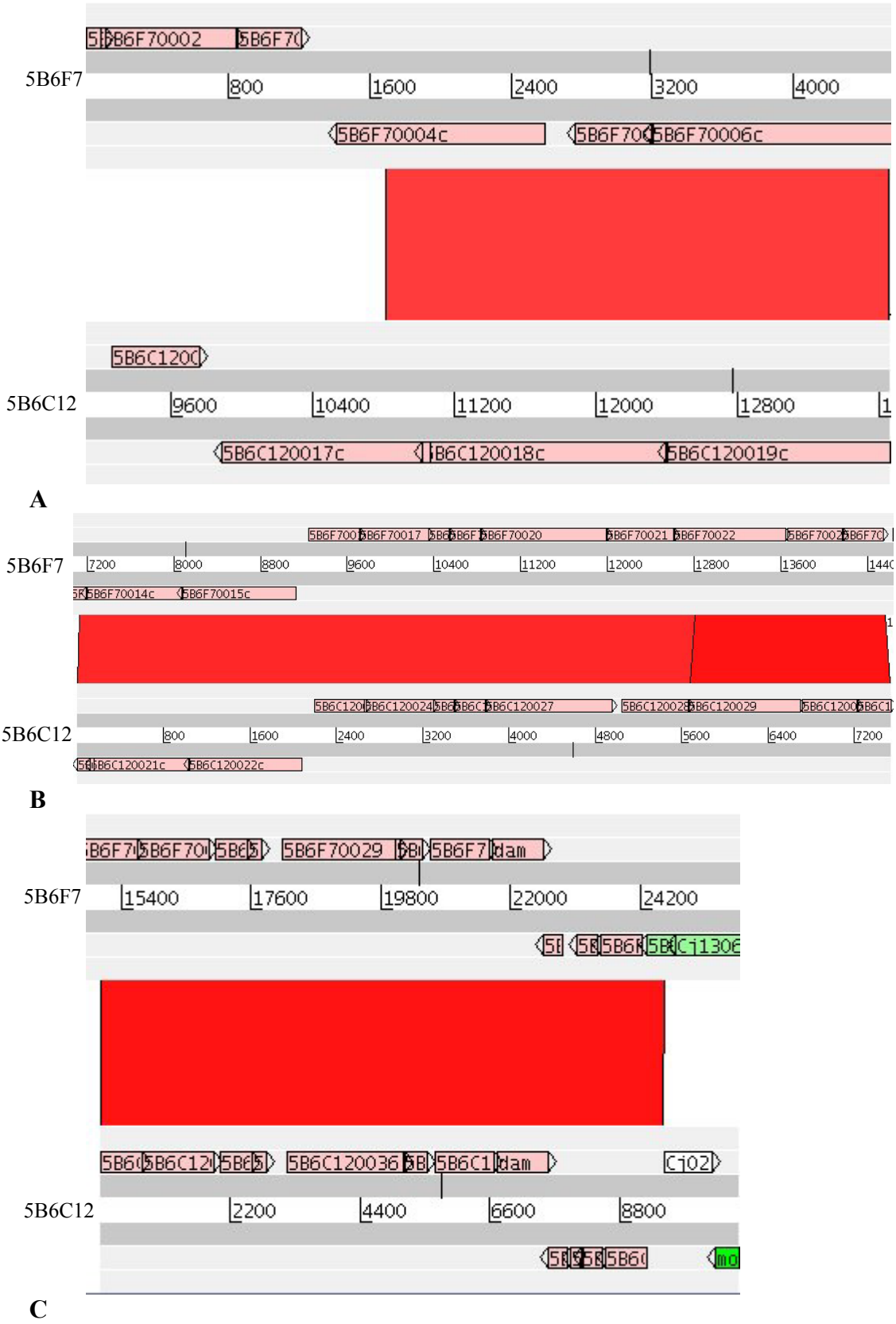


Fig 5.29: WUBLASTN comparison of sequence from strain 52472 BAC clones 5B6F7 and 5B6C12. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are

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

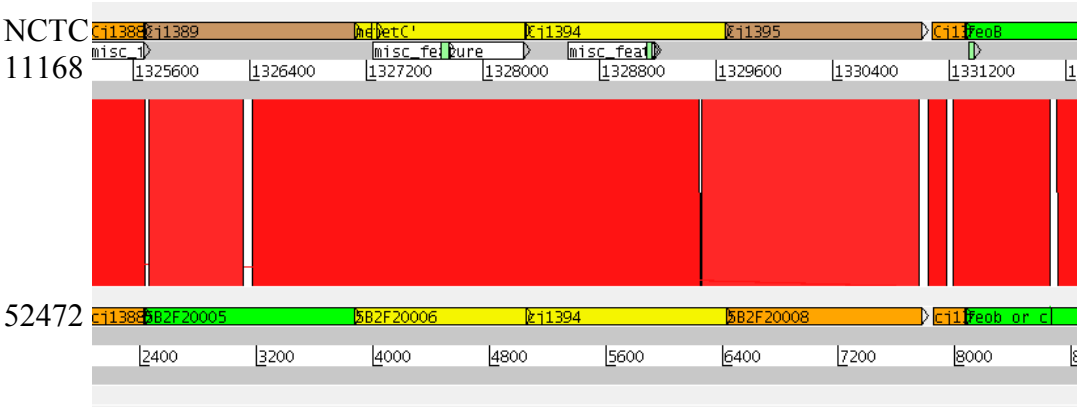


Fig 5.30: Blastn comparison of sequence from strain NCTC 11168 and strain 52472 BAC clone 5B2F2. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark 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

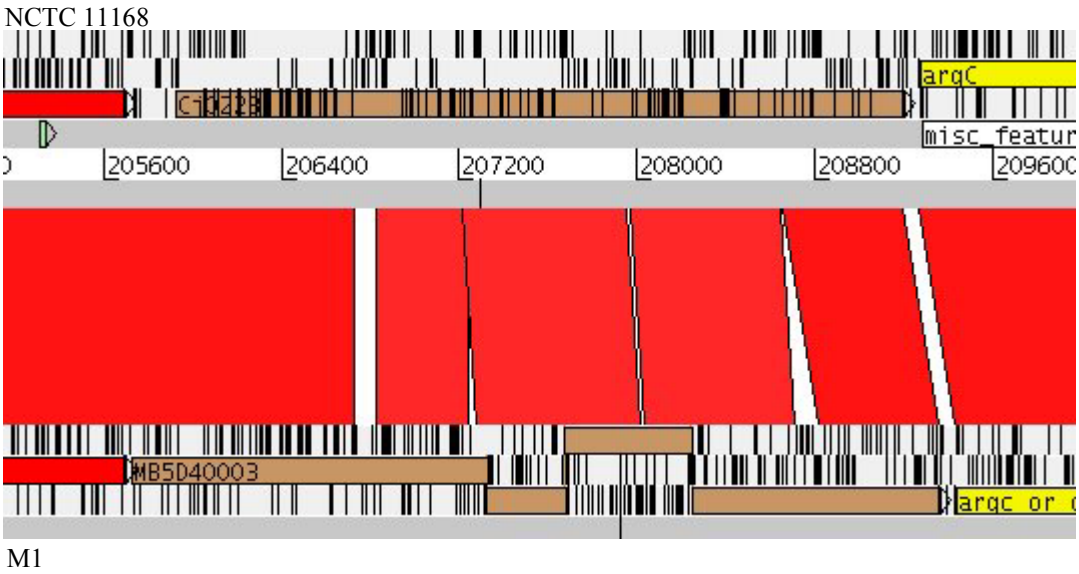


Fig 5.32: Blastn comparison of sequence from strain NCTC 11168 and strain M1 BAC clone MB5D4. The comparison is viewed using ACT; blocks of red indicate sequence homology with the colour intensity proportional to the percent id of the match. Forward and reverse DNA sequences are represented by dark grey 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.

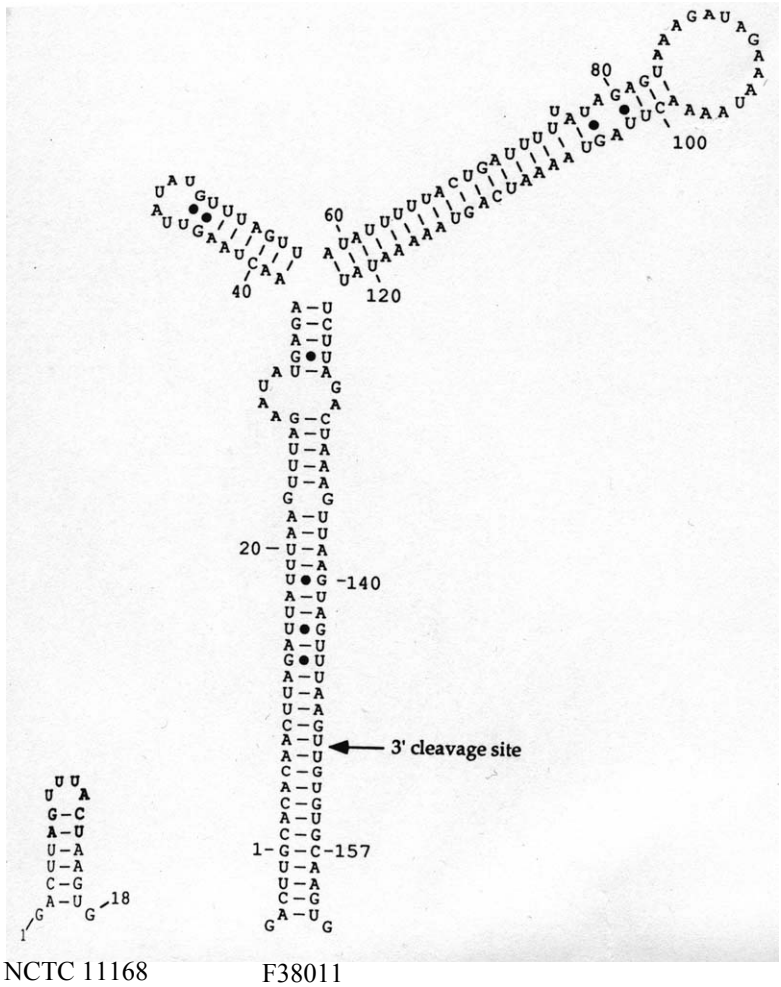


Fig 5.34: Predicted secondary structure of IVS from *C. jejuni* strain F38011 23s rRNA. Figure reproduced from Konkel *et al.* 1994 [172]. The 8 bp in strain NCTC 11168 that are replaced by the IVS are shown in bold on the left hand loop. The right hand loop shows the sequence and predicted secondary 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.

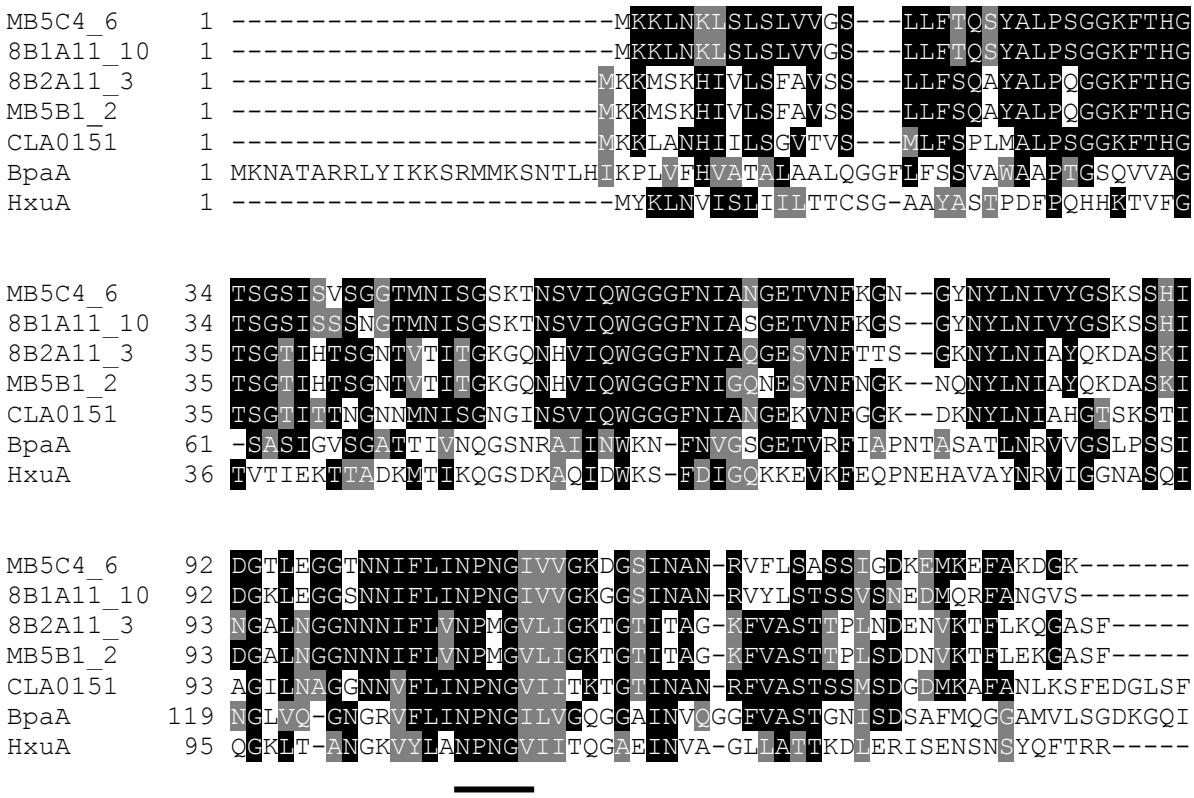
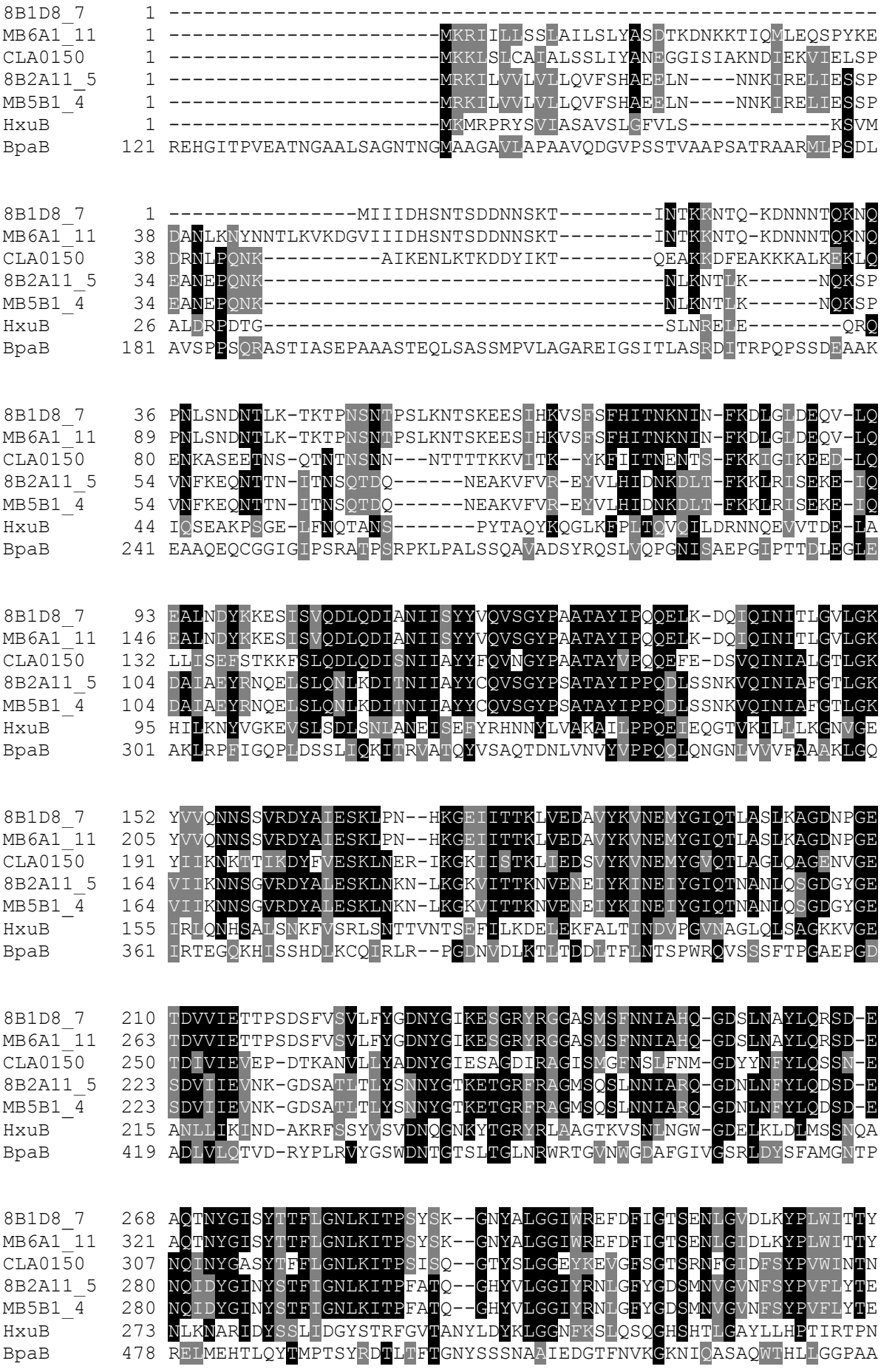


Fig 5.35: Alignment of the N-terminal region of TpsA proteins. Protein sequences were aligned using clustal X. Protein sequences of predicted CDSs from this study are shown on the top 4 lines. CLA0151 is a protein from *Campylobacter lari* RM2100 (accession number AEL54392); BpaA is a protein from *Burkholderia 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
MB6A1_11
CLA0150
8B2A11_5
MB5B1_4
HxuB
BpaB

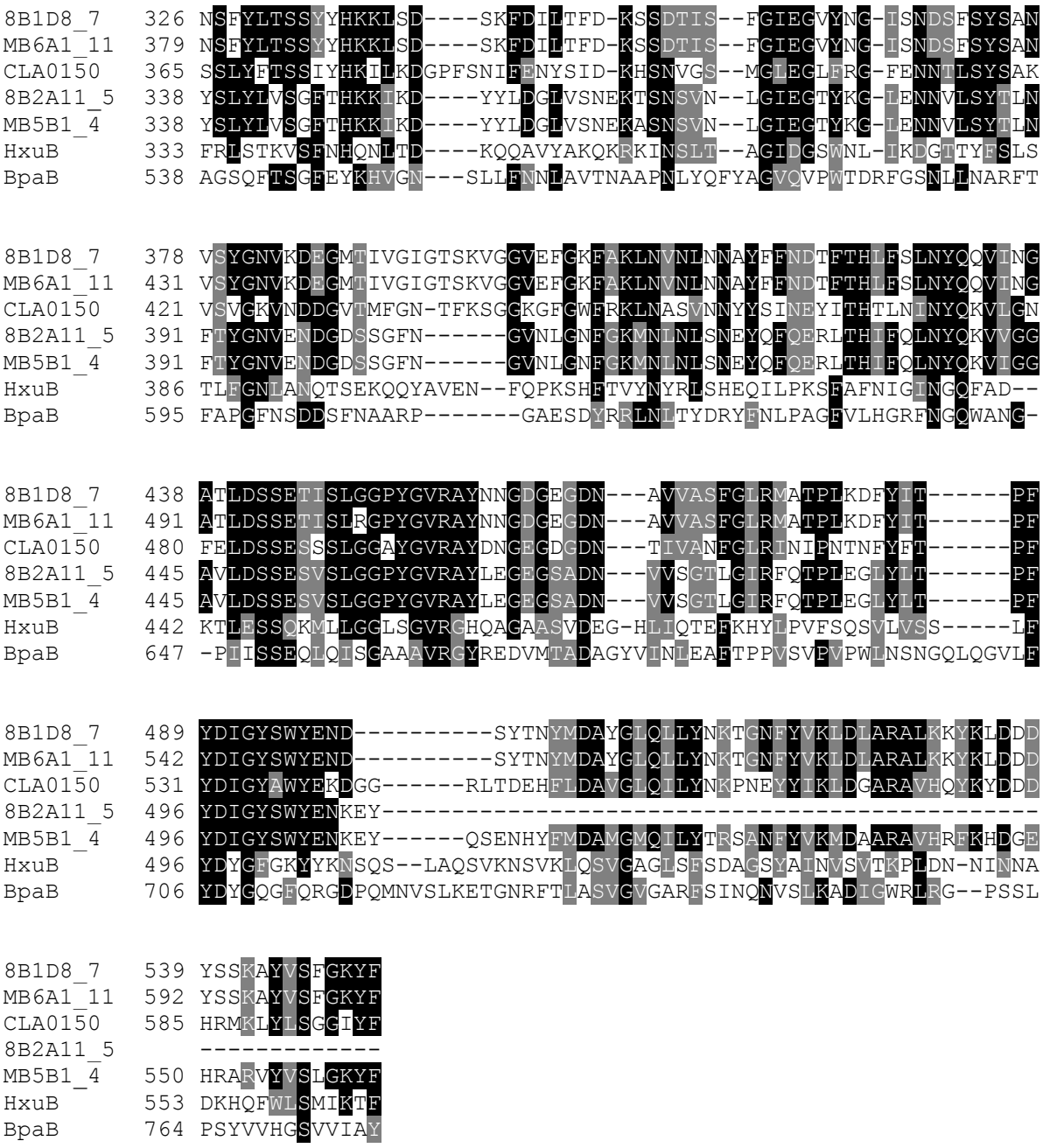


Fig 5.36: Alignment of TpsB secretor proteins. The protein sequences were aligned using clustal X. The CDSs from this study are 8B1D8_7, MB6A1_11, 8B2A11_5 and MB5B1_4. CLA0150 is a protein from *Campylobacter lari* RM2100 (accession number EAL54391); HxuB is a protein from *Haemophilus influenzae* (accession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