4. Analysis of pUC libraries

4.1 Introduction

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

4.2 Results

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

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

length in some instances. In strain 81-176 this accounted for 5 regions, in strain M1 8 regions and in strain 52472 7 regions which were discounted from further analysis unless they occurred next to re-arrangement events compared to the NCTC 11168 chromosome or represented more complete versions of pseudogenes in NCTC 11168.

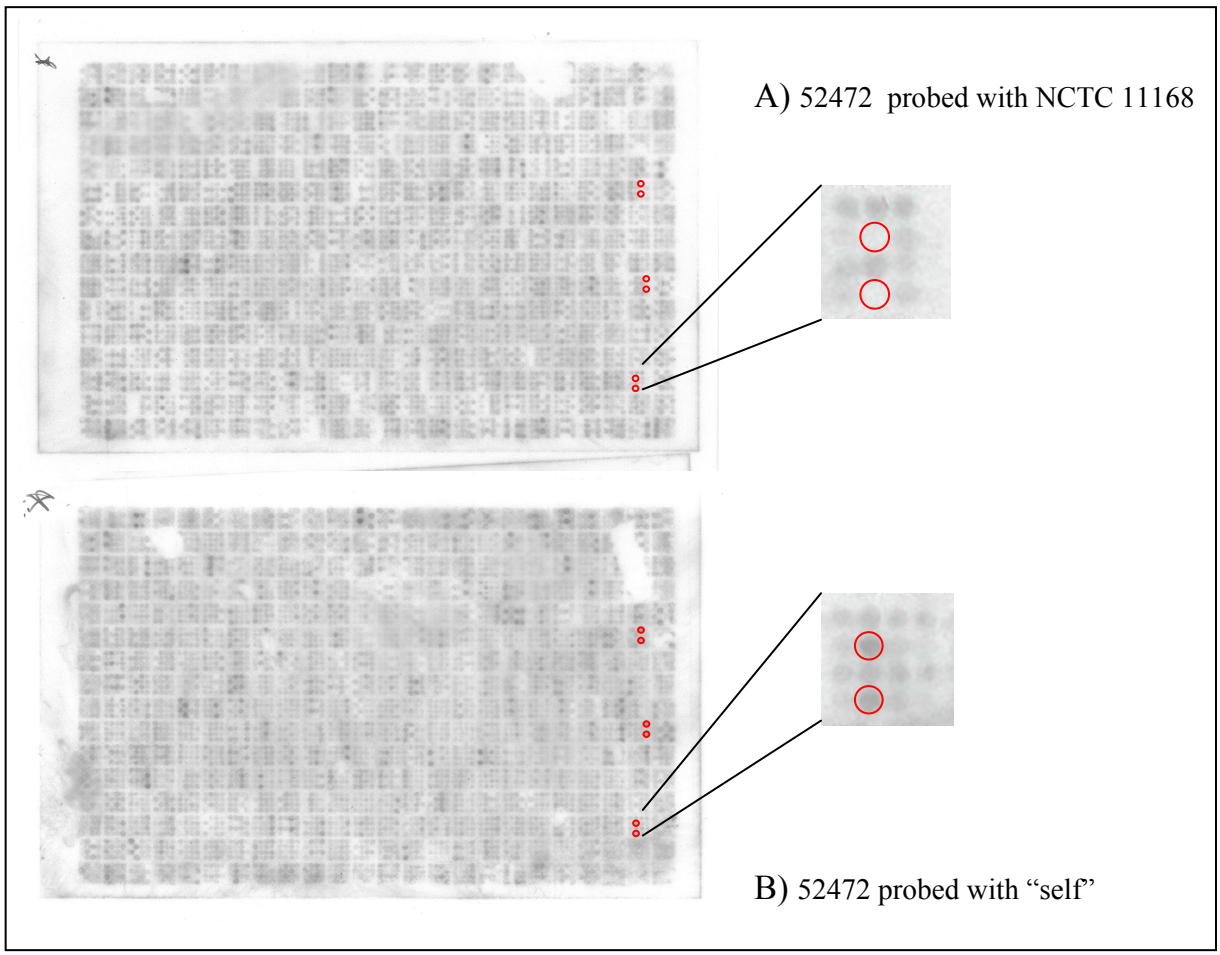


Fig 4.1: Differential genomic DNA hybridization array. A library of DNA fragments from each strain was spotted onto nylon membranes. A) shows an example of a membrane from strain 52472 probed with genomic DNA from NCTC 11168 and B) shows a membrane from strain 52472 probed with its own genomic DNA. The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

4.2.2 False positive and false negative testing

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

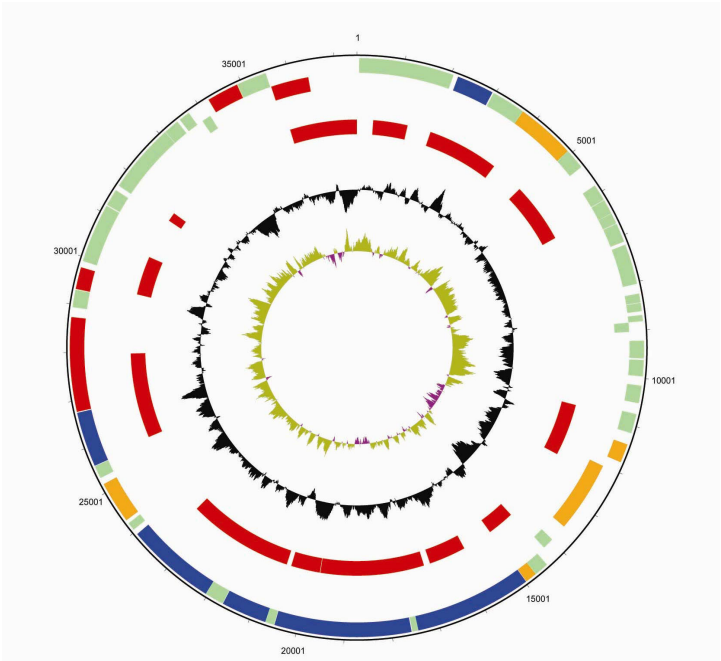
The sequence reads from pUC clones identified as containing end-sequences with less than 85% nucleotide id to NCTC 11168, from the false negative testing screen, were compared to the differential hybridization results. In strain 81-176 only 3 clones out of the 20 identified as novel (15%) had not been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

screen. Of these clones 5 contained DNA already sequenced leaving 19% of novel DNA not previously identified by sequencing other library clones.

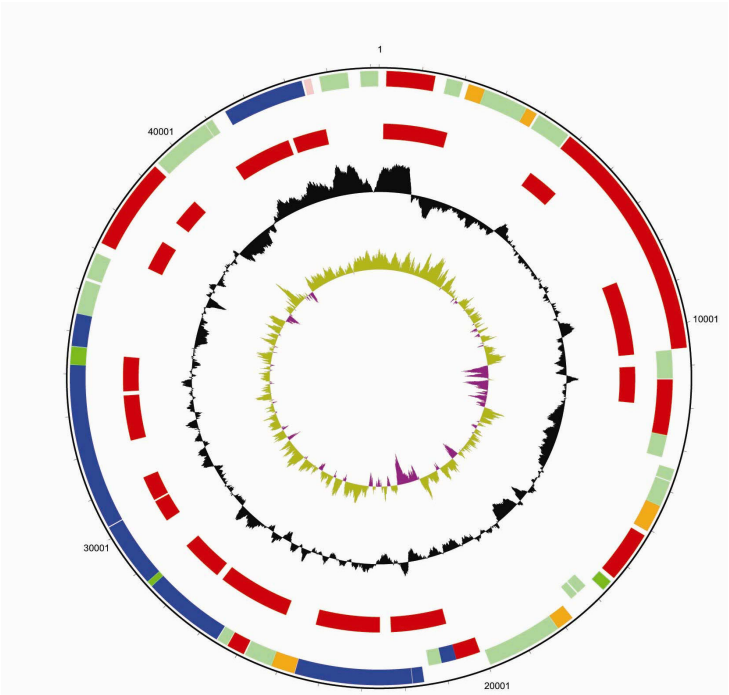
Only 11% of randomly selected library clones from strain 81-176 contained novel sequence whereas 14% of randomly selected library clones from strain M1 contained novel sequence. Using differential hybridization data to select clones for sequencing, 83% of clones from strain 81-176 and 66% of clones from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

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

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



A



B

Fig 4.2: A circular representation of the plasmid sequences. A: pVir, B: pTet.
The circles represent the following features, numbering from the outside in: 1, 2, all CDSs (transcribed clockwise and anticlockwise respectively); 3, the position of 81-176 pUC clone sequence reads identified from the differential hybridization screen with more than 85% similarity to the plasmid sequence; 4, G+C content; 5, GC deviation ((G-C)/ (G+C)) viewed using a window size of 250 bp, with a step size of 10 bp. The 12 o'clock position of each circle represents the predicted origin of replication and CDS colours represent the following putative functions: red, information

transfer (transcription/ translation + DNA/ RNA modification); light green, unknown; dark green, surface; orange, conserved hypothetical; blue, pathogenicity/ adaptation; pink, bacteriophage/ IS elements.

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

4.2.4 General Features of novel pUC assemblies

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

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

Discounting CDS matches with more than 95% amino acid id to NCTC 11168 across entire length, 93 novel genes were discovered in strain 81-176, 137 in strain M1, 97 in strain 40671 and 268 in strain 52472. The CDSs with more than 95% amino acid id were found to be either towards the ends of contiguous regions containing novel sequence or next to

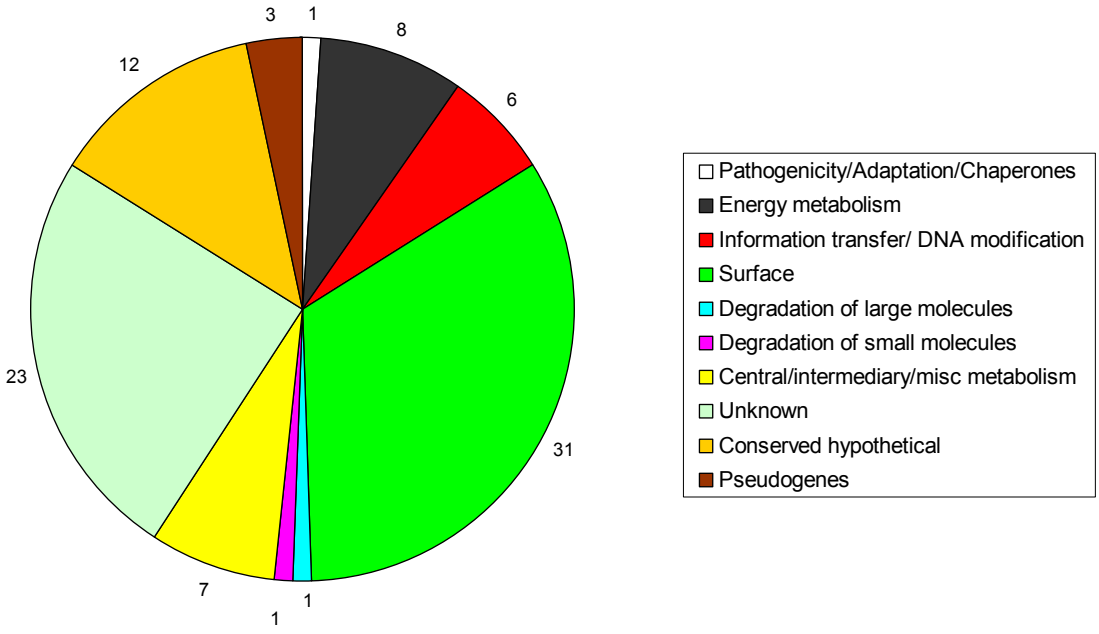
insertion or deletion (indel)/ rearrangement events, which may provide useful positional information.

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

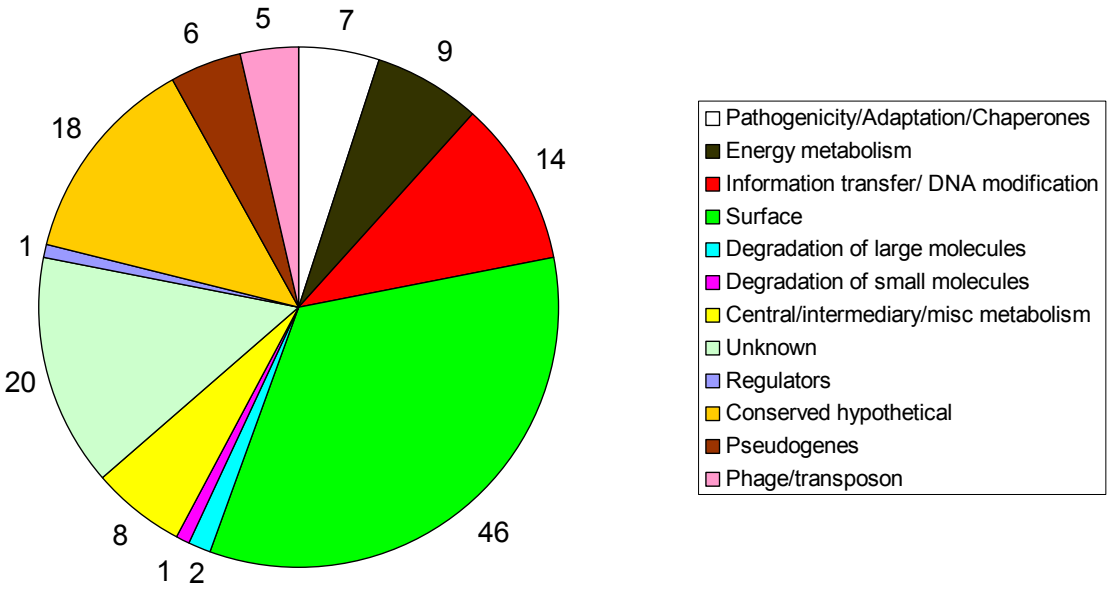
Excluding phage associated CDSs in strain 52472 the largest class of predicted CDSs in all strains tested are hypothetical (**Fig 4.3**). There are 35 hypothetical CDSs in strain 81-176 (38% of novel), 38 in strain M1 (28% of novel), 46 in strain 40671 (47% of novel) and 63 in strain 52472 (24% of novel).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

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

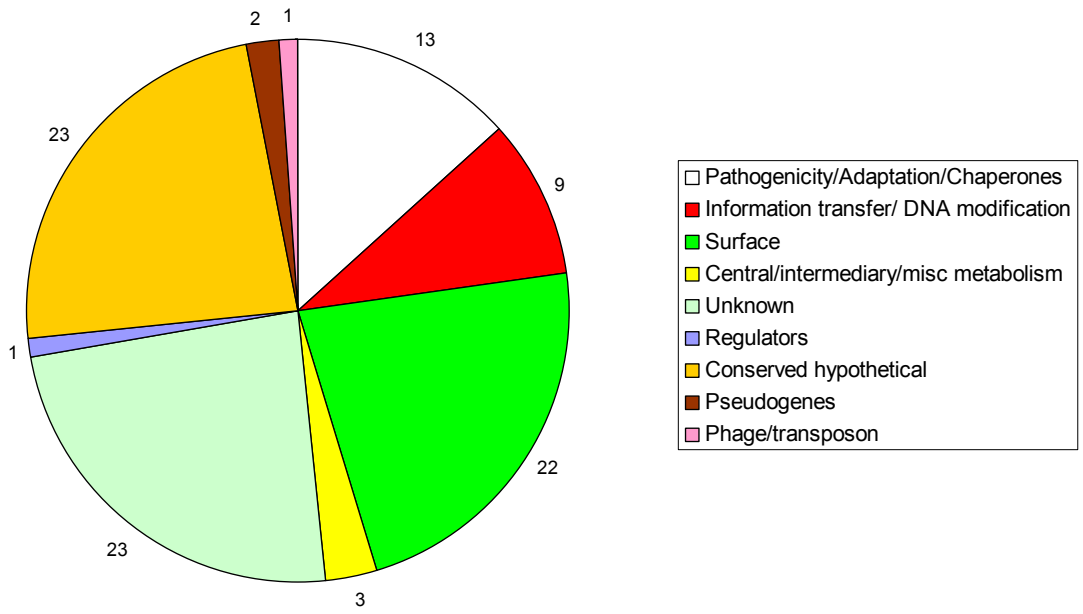
A. Functional categories of predicted CDSs in 81-176 pUC assemblies



B. Functional categories of predicted CDSs in M1 pUC assemblies



C. Functional categories of predicted CDSs in 40671 pUC assemblies



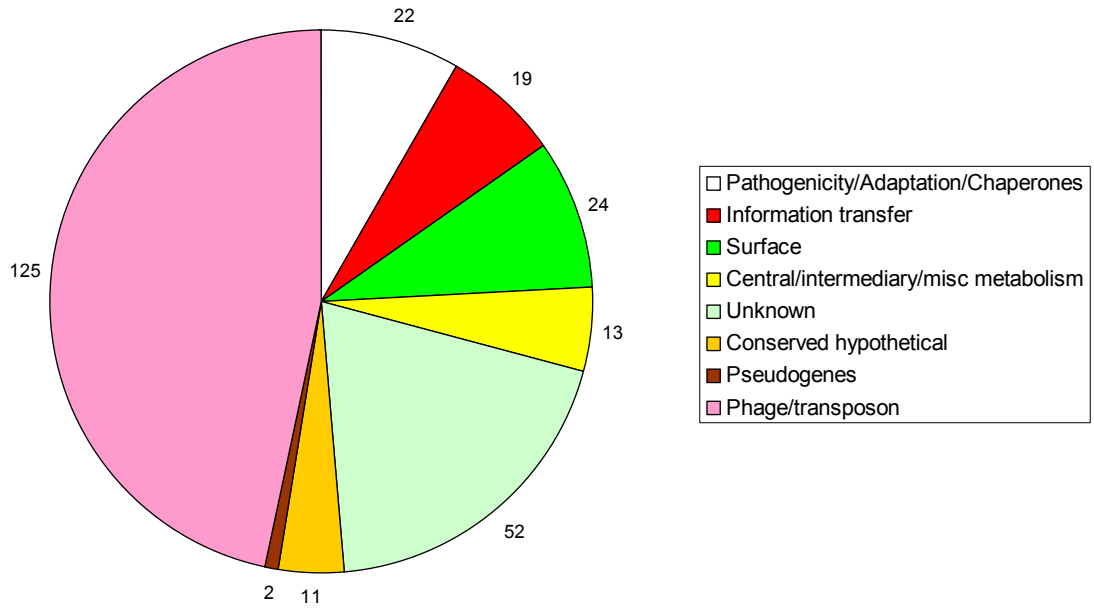


Fig 4.3: Pie-chart diagrams depicting functional categories for novel predicted CDSs. Functional categories are described in the keys beside each chart. Certain categories make up a large proportion of the novel CDSs of all strains: DNA modification; Surface; unknown; conserved hypothetical.

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

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

4.2.5 Regions showing limited identity to genes in NCTC 11168

4.2.5.1 Surface associated

4.2.5.1.1 N-linked glycosylation locus cj1119c-cj1130c

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

4.2.5.1.2 Lipo-oligosaccharide biosynthesis locus cj1131-1152

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

In strain M1 all the predicted CDSs from the LOS region show highest identity to CDSs from strain 81116, also known as strain NCTC 11828 [144], which does not fall into the A, B or C class system (**Fig 4.4**). The predicted CDS MP0121 (MP2b12p) shows high identity to WlaNA and the partial CDSs MP0120 and MP0029 (MP1f05p) both show high identity to WlaNB. CDS MP0028 (MP1f05p) shows high identity to a transferase, RlmA,

and CDS MP0010 (MP5b05p) shows 59% id to a DTPT dehydratase from *H. hepaticus*. Predicted CDS MP0051 (MP3d07q) shows high identity to a hypothetical CDS, MP0052 to an aminotransferase and MP0053 to a membrane protein. Predicted CDSs MP0015 and MP0016 (MP3b05q) show high identity to two glycosyltransferases from strain 81116 with MP0016 showing 63% id to CgtA (Cj1138) from strain NCTC 11168. The predicted CDSs MP0034 (MP2f12q) and MP0019 (MP3d02q) show high identity to an O-acetylation protein and MP0020 (MP3d02q) to a hypothetical protein which are inserted upstream of *gmhA* (MP0021). This arrangement is present in class B1 LOS clusters [59].

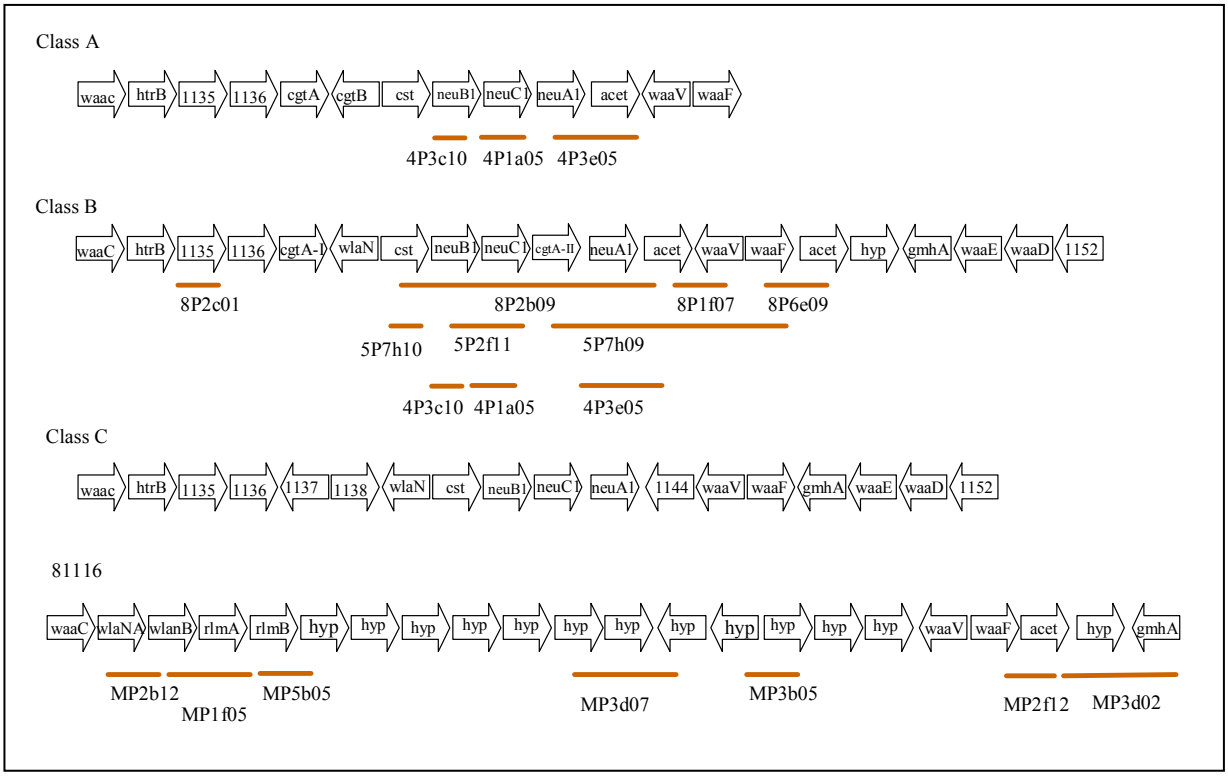


Fig 4.4: **Schematic representation of the genes from different LOS classes** of *C. jejuni* according to Gilbert *et al*. 2002 [143]. Arrow sizes are not representative of gene or protein size. Arrows are labelled with predicted protein products; numbers represent CDS locus id numbers from *C. jejuni*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

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

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

4.2.5.1.3 Flagellar associated genes

In strains 81-176 and 40671 the homologues of cj0043 encoding the flagellar hook protein FlgE appear to be variable. The predicted CDS 8P0054 (8P4a03p) shows high identity to FlgE previously sequenced from strain 81-176 and predicted CDS 4P0012 (4P1c07p) shows

homology to FlgE from strain lio7 and but these two predicted CDSs, 8P0054 and 4P0012, do not share high identity with each other [138].

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

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

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

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

4.2.5.1.4 Capsule locus cj1413-1448

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

which has been associated with GBS [146], and 8P0029 (8P2d02p) which shows 41% id to Cj1431 from strain NCTC 11168 [151].

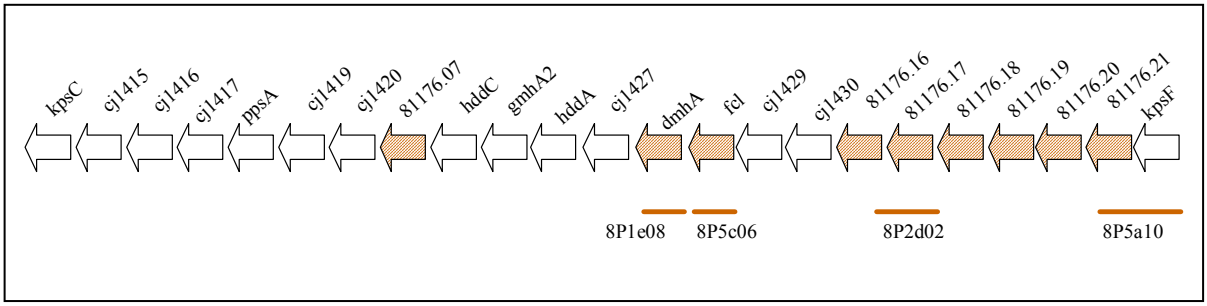


Fig 4.5: Capsule locus of *C. jejuni* strain 81-176. CDSs from the capsule locus of strain 81-176 determined by Karlyshev *et al.* 2005 [151] (accession number BX545858) are depicted by arrows. The size of the arrows is not representative of gene or protein size. Contiguous regions from this study are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

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

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

4.2.5.1.5 Miscellaneous

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

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

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

4.2.5.2 Metabolism

4.2.5.2.1 Molybdate transport region cj0294-cj0310

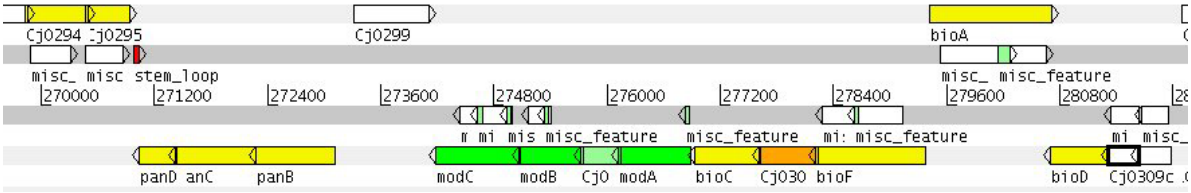


Fig 4.6: Molybdate transport region of *C. jejuni* strain NCTC 11168. The region from cj0294-cj0310 is viewed using Artemis. The forward and reverse DNA lines are represented by the central dark grey lines. The light grey lines represent all three forward and reverse translated reading frames respectively.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

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

4.2.5.2.2 Region cj0807-cj0813

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

4.2.5.2.3 Miscellaneous

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

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

In strain 81-176 a predicted CDS, 8P0092, with homology to *purU* (cj0789) from strain NCTC 11168 appears to be shorter than in NCTC 11168. The predicted CDS 8P0092 is 158 aa long compared to Cj0789 which is 274 aa in NCTC 11168. It is not possible to say whether this gene would still be functional or whether a duplication event may have occurred

and a full length copy of the gene is present elsewhere on the chromosome. The predicted CDS 8P0092 is located on contiguous region 8P7g11p and a rearrangement compared to NCTC 11168 seems to have occurred with a predicted CDS with high identity to the iron uptake transporter, cj0173c, occurring upstream of the *purU* homologue.

4.2.5.3 Hypothetical genes

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

4.2.5.4 Pseudogenes

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

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

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

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

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

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

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

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

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

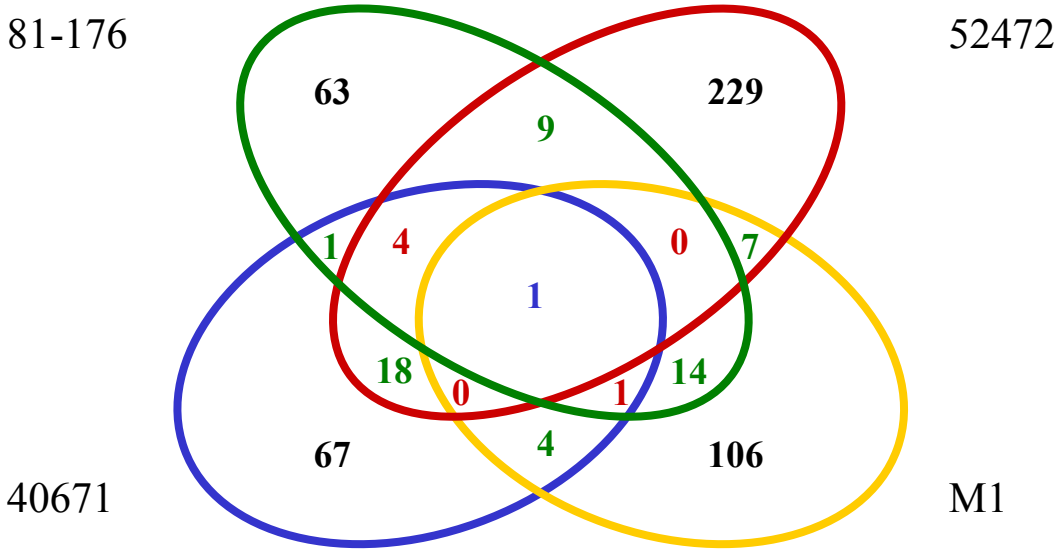


Fig 4.7: Venn diagram of predicted novel CDSs shared between strains in this study. The green ellipse represents novel CDSs identified in strain 81-176; the red ellipse represents novel CDSs identified in strain 52472; the yellow ellipse represents novel CDSs identified in strain M1 and the blue ellipse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

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

4.2.6.1 CDSs present in all test strains

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

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

4.2.6.2 CDSs present in three test strains

4.2.6.2.1 Shared between 81-176, M1 and 40671

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

4.2.6.2.2 Shared between 81-176, 40671, 52472

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

4.2.6.3 CDSs present in two test strains

4.2.6.3.1 Shared between 81-176 and M1

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

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

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

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

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

4.2.6.3.2 Shared between 81-176 and 40671

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

4.2.6.3.3 Shared between 81-176 and 52472

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

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

4.2.6.3.4 Shared between M1 and 40671

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

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

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

4.2.6.3.5 Shared between M1 and 52472

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

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

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

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

4.2.6.3.6 Shared between 40671 and 52472

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

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

Strain 40671 contains 18387 bp of sequence that matches to pTet representing 41% of the plasmid but contains most of the type IV secretion system genes with the exception of the *virB5* and *virD2* homologues. Strain 52472 contains 33034bp of sequence that matches to pTet representing 73% of the plasmid and contains homologues of all the type IV secretion system genes. It may also be possible that part of the plasmid is inserted on the

chromosome in a similar way to the plasmid derived island of RM1221. Further work would be needed to explore this possibility.

4.2.6.4 Predicted novel CDSs present in RM1221

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

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

81-176	M1	40671	52472
21	30	4	108

4.2.6.4.1 Shared between 81-176 and RM1221

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

4.2.6.4.2 Shared between M1 and RM1221

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

Also conserved between the two strains is RM1221 CJE0312 and MP0030 (MP2h03q) which shows 55% id to Cj0262, a methyl-accepting chemotaxis signal transduction protein. There is also a conserved CDS associated with restriction-modification

systems, RM1221 pseudogene CJE1720 and MP0071 (MP2g03q) which shows high identity to HsdR from strain 81116 and appears to be inserted before a homologue of the dehydrogenase cj1548 [154].

4.2.6.4.3 Shared between 40671 and RM1221

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

4.2.6.4.4 Shared between 52472 and RM1221

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

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

There are some plasmid associated genes shared between the two strains with 5P0108 (5P6a01q) putatively encoding a TraC-like protein and showing identity to RM1221

pseudogene CJE1121, and 5P0121 (5P1d01q) showing identity to a hypothetical protein and RM1221. Both 5P0108 and 5P0121 show homology to pTet. RM1221 is known to have a large insert of novel DNA predicted to be of plasmid origin [9].

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

4.2.7 Predicted CDSs unique to each test strain

4.2.7.1 Strain 81-176

4.2.7.1.1 Restriction modification

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

4.2.7.1.2 Hypothetical

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

4.2.7.1.3 Surface (transport)

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

4.2.7.1.4 Miscellaneous

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

4.2.7.2 Strain M1

4.2.7.2.1 Restriction modification

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

4.2.7.2.2 Hypothetical

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

4.2.7.2.3 Surface

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

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

4.2.7.2.4 Miscellaneous

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

There are also various protease matches with MP0001 (MP2d02q) showing 37% id to the serine protease SigA from *Shigella flexneri*, MP0148 (MP2f07q) showing 46% id to a

haemoglobin protease from *Escherichia coli* (this contiguous region shows 92% nucleotide id to cj0223) and MP0054 (MP3e11p) showing 24% id to a haemolysin from *Xanthomonas axonopodis*.

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

4.2.7.3 Strain 40671

4.2.7.3.1 Restriction modification

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

4.2.7.3.2 Hypothetical

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

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

4.2.7.3.3 Surface

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

4.2.7.3.4 Miscellaneous

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

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

4.2.7.4 Strain 52472

4.2.7.4.1 Restriction modification

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

4.2.7.4.2 Hypothetical CDSs

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

Other hypothetical proteins without any information as to where they may located on the chromosome are 5P0059 (5P8b01p), 5P0071 (5P5h08p) which may be phage associated as it shows 41% id across part to a hypothetical protein from *Salmonella enterica* Typhi,

located adjacent to phage genes, and 5P0141 (5P3c11q), 5P0142 which shows 40% id to a hypothetical protein from *Helicobacter hepaticus* and 5P0143.

4.2.7.4.3 Surface

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

4.2.7.4.4 Miscellaneous

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

4.3 Discussion

4.3.1 Surface structures

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

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

The LOS regions have been highly studied and it appears that strains 81-176 and 52472 belong to class B [143] (**Fig 4.4**). Strain M1 appears to have high identity to the LOS

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

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

the capsule from other bacteria it is not possible to be confident that they are involved in capsular polysaccharide production in these strains.

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

4.3.2 Transport

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

Strains 81-176 and M1 contain homologues of DcuC: a dicarboxylate transporter. C4-dicarboxylates like succinate, fumarate and malate can be metabolized by bacteria under both aerobic and anaerobic conditions [161]. In NCTC 11168 there are *dcuA* and *dcuB* homologues but no functional *dcuC* homologue although there are two pseudogenes, cj1528

and cj1389, with homology to *dcuC*. In strain 81-176 it appears likely that the *dcuC* homologue is a more complete version of pseudogene cj1389 whereas in strain M1 the *dcuC* homologue shows some similarity to the pseudogene cj1528. DcuAB are used for electroneutral fumarate:succinate antiport which is required in anaerobic fumarate respiration. DcuC can replace DcuA or DcuB in catalyzing fumarate-succinate exchange and fumarate uptake but usually DcuC carriers function in succinate efflux during fermentation [161].

4.3.3 Restriction modification

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

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

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

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

4.3.4 Metabolism

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

A WUBLASTN comparison of the molybdate transport region, including bioF identified in 81-176 and 52472, to strain RM1221 showed 99% similarity. BioF is involved

in the biosynthesis of biotin which is an essential prosthetic group for carboxylase enzymes which each catalyse an essential metabolic reaction [164]. It is possible that this region in NCTC 11168 is under diversifying selection or that it has been horizontally transferred from another source, although the reasons why this region may vary are unclear. The natural competence of *C. jejuni* and high recombination rate are thought to be involved in generating diversity at the cell surface [51] but may also be involved in generating diversity elsewhere in the genome.

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

4.3.5 Respiration

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

the different ecological niches to which *C. jejuni* is exposed: for example, the avian and mammalian gut which is essentially anaerobic [4].

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

4.3.6 Chemotaxis

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

4.3.7 Pseudogenes

Although this study was not designed to investigate pseudogenes, 8 pseudogenes from NCTC 11168 have been identified through the differential hybridization screen and appear to vary. It should be borne in mind that the depth of coverage in this pUC screen will not give definitive results with regard to pseudogenes in all cases. In addition there are a number of novel genes that are probably pseudogenes. This is perhaps not unexpected as genes that are

not shared between strains are likely to be accessory and perhaps only required under a subset of ecological conditions which the bacterium may encounter and therefore more likely to pick up deleterious mutations.

4.3.8 Characteristics of each strain

4.3.8.1 Strain 81-176

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

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

missed are from the LOS region or capsule region and have a low G+C content [170]. Of the regions identified in this study 59% were missed by Poly *et al.*

4.3.8.2 Strain M1

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

4.3.8.3 Strain 40671

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

4.3.8.4 Strain 52472

Strain 52472 has a large number of bacteriophage associated CDSs. Bacteriophage are known to be highly mosaic in structure with a high rate of horizontal exchange between bacteriophage occupying similar ecological niches [171]. Bacteriophage may pick up

virulence determinants when they excise although none have been found in this study from strain 52472 or in RM1221 [9].

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

4.3.9 Summary

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