3. Plasmid sequencing and annotation

3.1 Introduction

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

3.2 Results

3.2.1 Overview of methods

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 The sequence reads were assembled using Phrap (Green, P., unpublished) into 4 large contiguous regions of 30788 bp, 23049 bp, 10671 bp and 10372 bp.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

3.2.2 Identification of replication origins

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

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

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

3.2.3 General characteristics

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

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

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

3.2.3.1 Characteristics of pVir

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

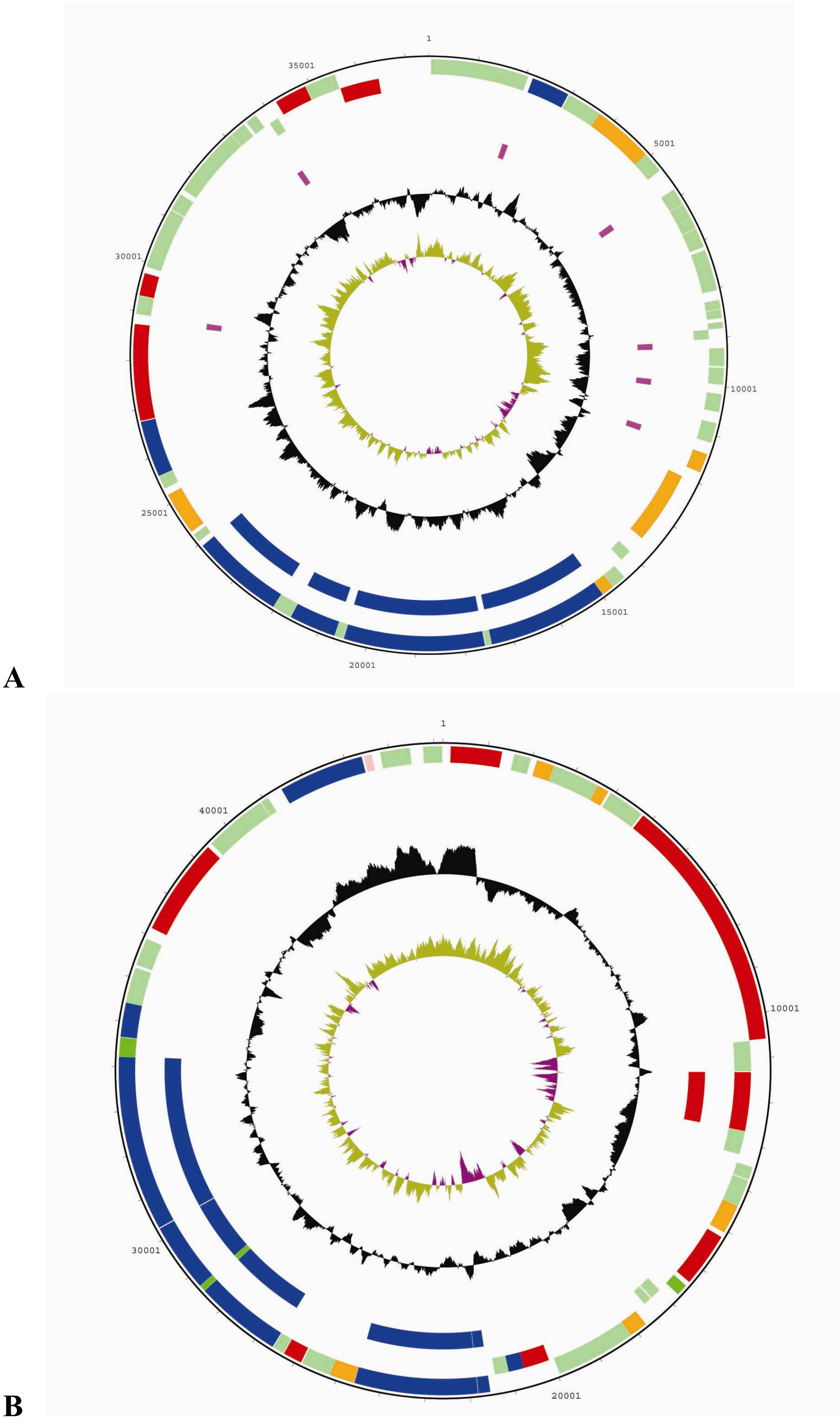


Fig 3.1: A circular representation of the plasmid sequences. A: pVir, B: pTet.

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

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

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

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


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


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

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

3.2.3.2 Characteristics of pTet

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

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

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

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

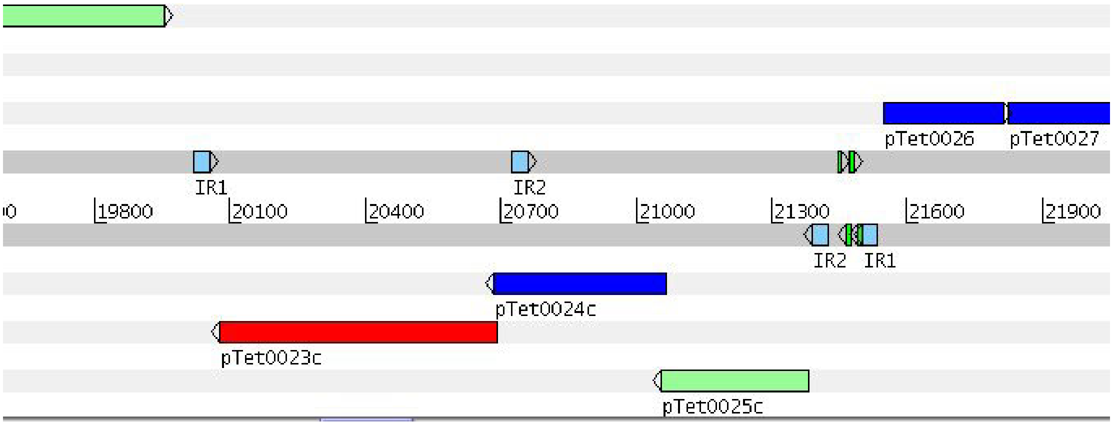


Fig 3.3: Putative invertible region in the plasmid pTet. The region is viewed using Artemis [102], forward and reverse DNA lines are represented by the central dark grey bars. The three forward and three reverse reading frames translated from the DNA sequence are represented by the light grey bars. Open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

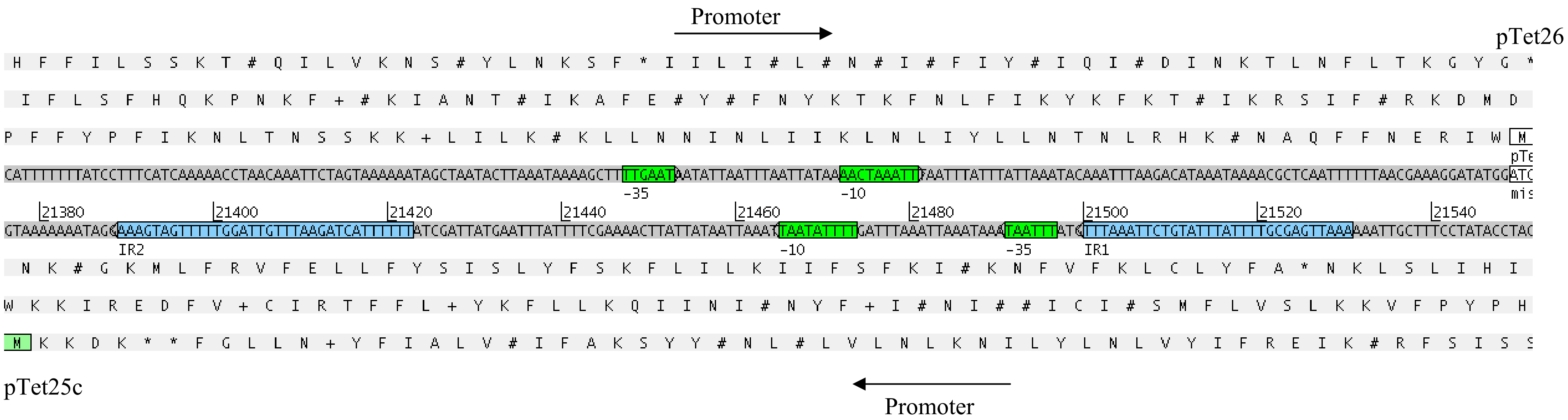


Fig 3.4: Region between pTet25c and pTet26 predicted to contain a promoter. The region is viewed using Artemis [102], forward and reverse DNA lines are represented by the central dark grey bars. The three forward and three reverse reading frames translated from the DNA sequence are represented by the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

3.3 Discussion

3.3.1 Comparison to published sequences

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

3.3.1.1 pVir

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

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

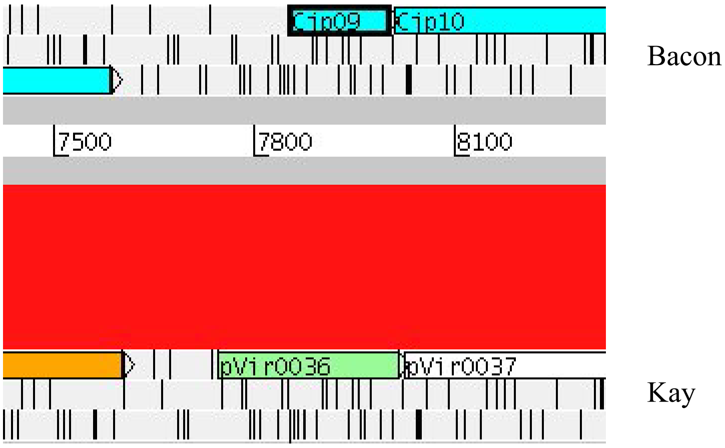


Fig 3.5: A WUBLASTN comparison of the published pVir sequence with the pVir sequence from this study in the region of pVir36. The comparison is viewed using the Artemis Comparison Tool (ACT) where blocks of red indicate sequence homology and the intensity of colour is proportional to the percent identity.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

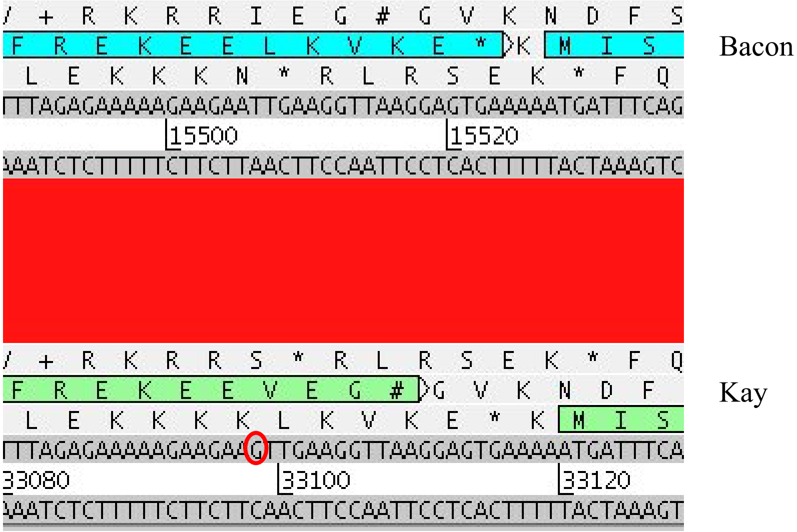
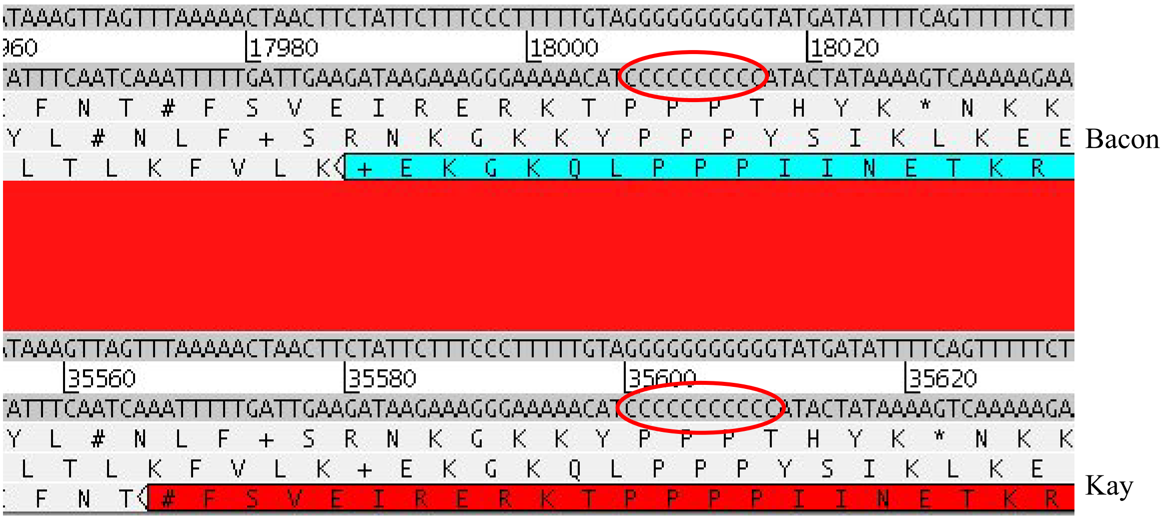

/ + R K R R I E G # G V K N D F S
F R E K E E L K V K E * K M I S
L E K K K N * R L R S E K * F Q
TTTAGAGAAAAAGAAGAATTGAAGGTTAAGGAGTGAAAAATGATTTCAG
15500
15520
AAATCTCTTTTTCTTCTTAACTTCCAATTCCTCACTTTTTACTAAAGTC
Bacon
/ + R K R R S * R L R S E K * F Q
F R E K E E V E G # G V K N D F
L E K K K K L K V K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

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

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

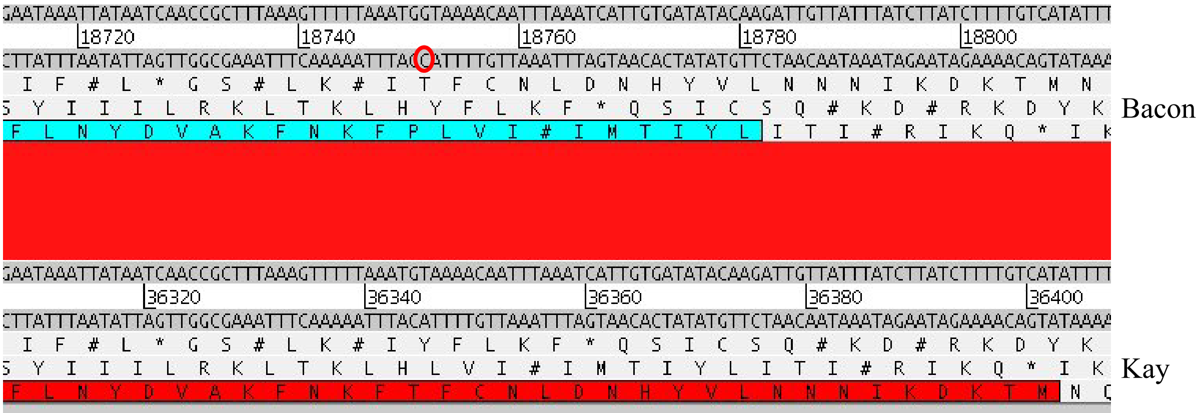
There are 3 bp differences in the putative *repA* gene. There is a homopolymeric tract of guanine residues (G) which appears to vary between G(10-11) before base 35611 (**Fig 3.7**). In the shotgun assembly three reads contained G(10) and seven reads contained G(11) suggesting that this homopolymeric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



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

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

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



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

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

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

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

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

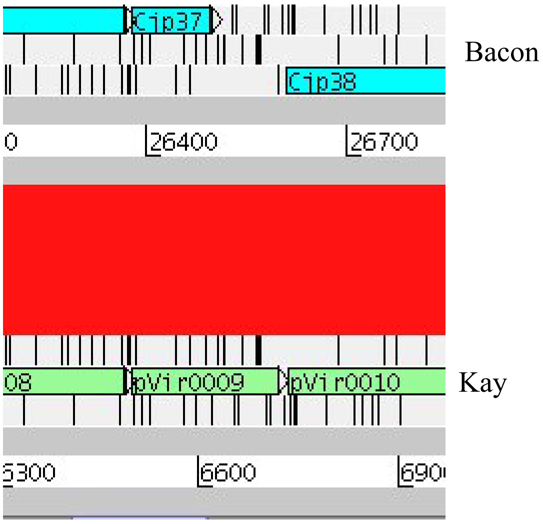


Fig 3.9: A WUBLASTN comparison of the published pVir sequence with the pVir sequence from this study in the region of pVir9. The comparison is viewed using ACT where blocks of red indicate sequence homology and the intensity of colour is proportional to the percent identity. However, single base pair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

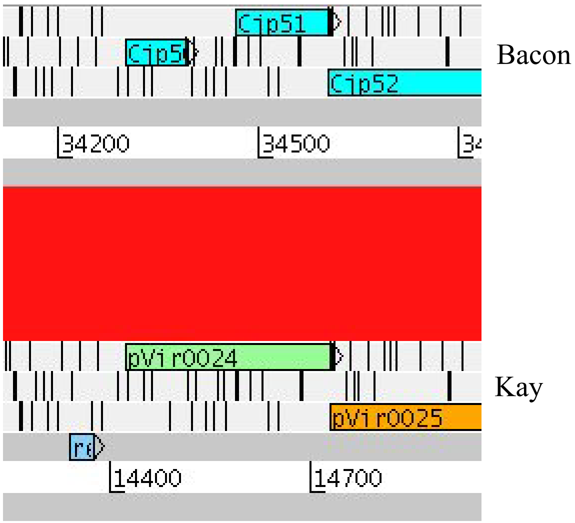


Fig 3.10: A WUBLASTN comparison of the published pVir sequence with the pVir sequence from this study in the region of pVir24. The comparison is viewed using ACT where blocks of red indicate sequence homology and the intensity of colour is proportional to the percent identity. However, single base pair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

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

3.3.1.2 pTet

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

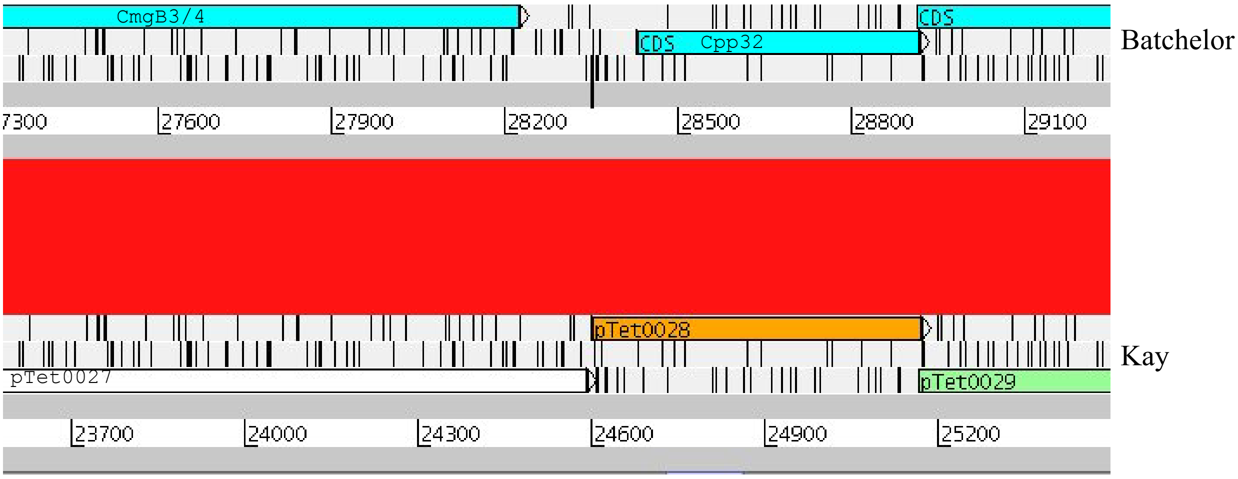


Fig 3.11: A WUBLASTN comparison of the published pTet sequence with the pTet sequence from this study in the region of pTet28. The comparison is viewed using the ACT where blocks of red indicate sequence homology and the intensity of colour is proportional to the percent identity. However, single base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

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

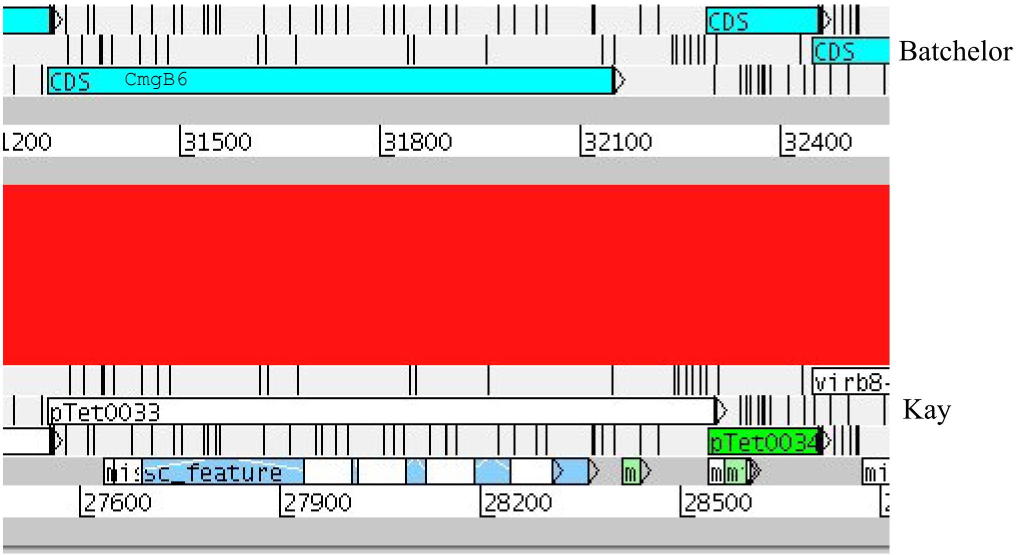


Fig 3.12: A WUBLASTN comparison of the published pTet sequence with the pTet sequence from this study in the region of pTet33. The comparison is viewed using ACT where blocks of red indicate sequence homology and the intensity of colour is proportional to the percent identity. However, single base pair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

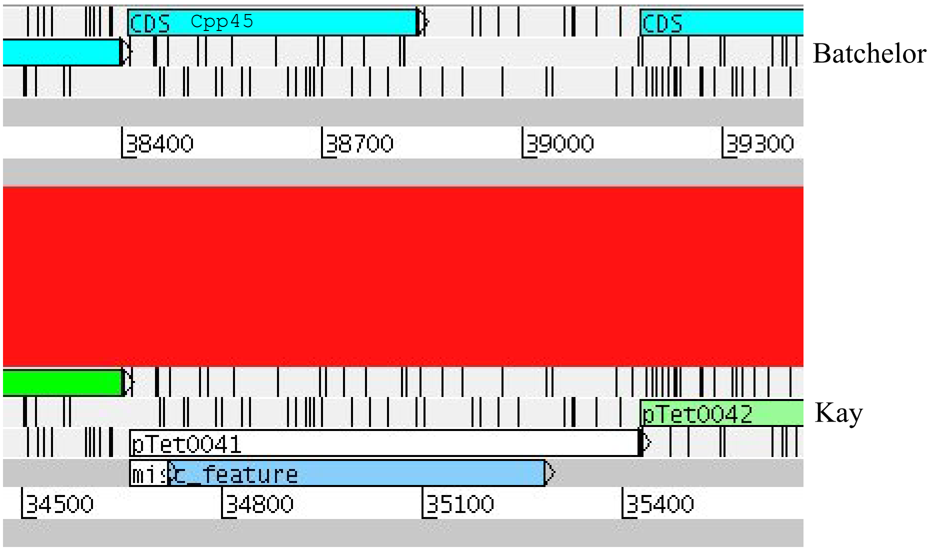


Fig 3.13: A WUBLASTN comparison of the published pTet sequence with the pTet sequence from this study in the region of pTet41. The comparison is viewed using ACT where blocks of red indicate sequence homology and the intensity of colour is proportional to the percent identity. However, single base pair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

3.3.2 Characteristics of pVir and pTet

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

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

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

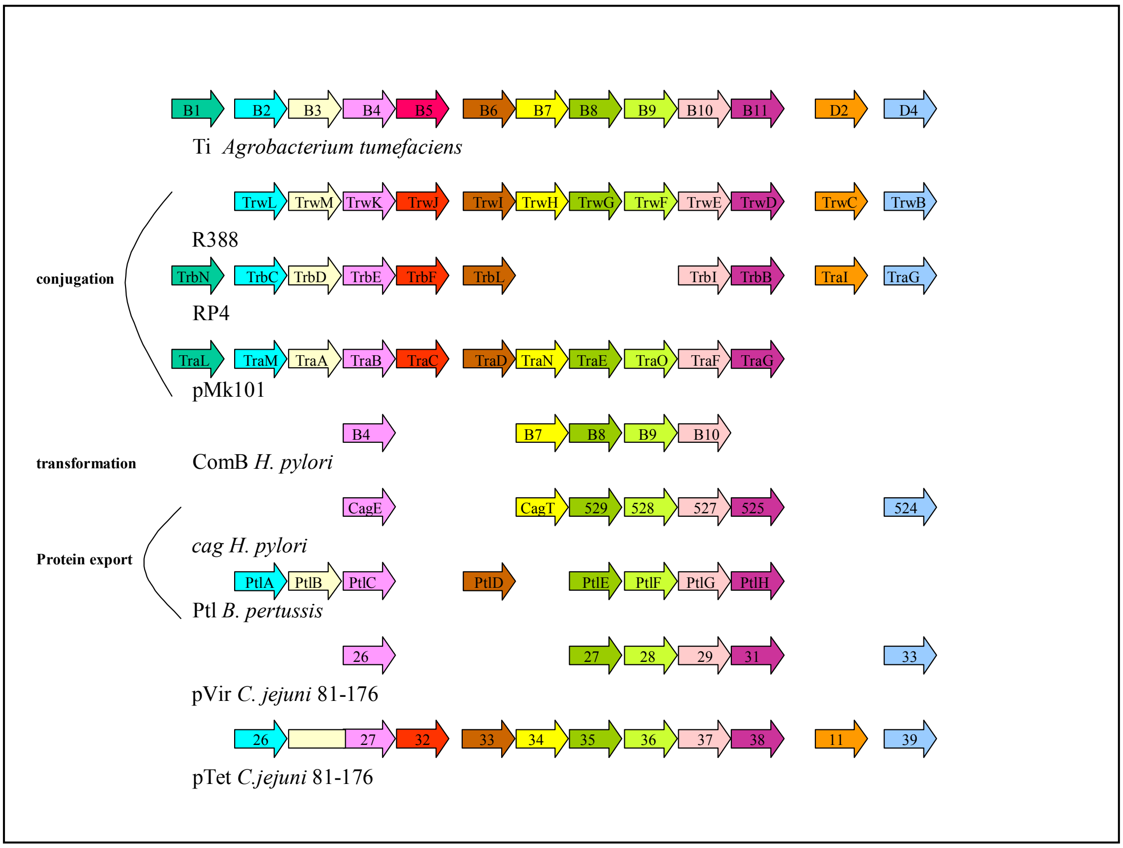


Fig 3.14: Schematic comparison of proteins involved in type IV bacterial secretion systems. Proteins that are homologous are shown with arrows of the same colour. The order of proteins expressed here is not necessarily the order in which the genes appear on respective stretches of DNA with the exception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

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

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