CHAPTER 6

Identification and characterisation of part of the *intD* element from *D. nodosus* strain C305

The important thing is not to stop questioning. Curiosity has its own reason for existing. One cannot help but be in awe when he [or she] contemplates the mysteries of eternity, of life, of the marvellous structure of reality.

Albert Einstein

6.1 Introduction

The benign *D. nodosus* strain C305 does not contain any copies of the *vap* regions. A different genetic element, the *intC* element, is integrated into the *tRNAser* gene downstream from *askA* in strain C305 (Section 1.6.8.1, Figure 6.1).

Southern blot hybridisation experiments suggest that our original laboratory strain of C305 (C3051) contained a DNA sequence between *pnpA* and $intB_N$, and this was lost, or changed position in the genome, to generate the current laboratory strain (C3052), in which *pnpA* and $intB_N$ are juxtaposed (Section 1.6.8.2). This hypothesis was supported by the isolation, from a C305 genomic library, of three different lambda clones which hybridise to an *intB* probe (Figures 6.1 and 6.2). One of these clones (λGB300) contains a 3.3 kb *Hin*dIII fragment, on which there are three copies of the 102 bp repeats found within the *vap* regions of strain A198. This clone also contains part of the *attR* from *vap* region 2 of strain A198, and a copy of $intB_N$ (Figure 6 2).

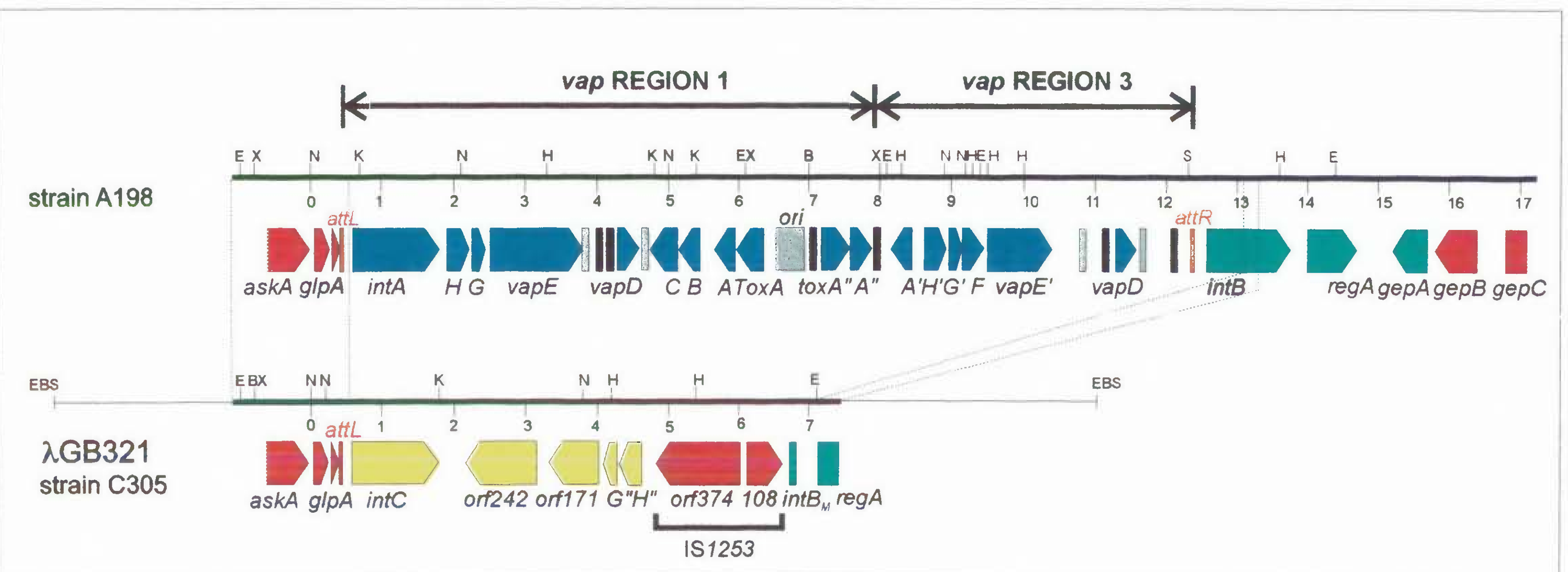

Figure 6.1: Alignment of restriction maps of regions downstream of *askA* in virulent strain A198 and benign strain C305. The C305 sequences were isolated on the lambda clone λGB321 as shown. The numbers indicates the distance in kb. Restriction enzyme sites shown are *Bam*HI (B), *Eco*RI (E), *Hin*dIII (H), *Kpn*I (K), *Nru*I (N), *Sac*I (S) and *Xho*I (X). Open reading frames are indicated by shaded arrows and are classified as follows: belonging to the *intB* element (green), belonging to the *vap* region (blue), belonging to the *intC* element (yellow)whilst genes that are not part of an integrated genetic element are also distinguished (red) and include *tRNA* genes (red triangles). Repeated sequences are as follows: attachment sites, *attL* or *attR* (narrow red boxes), 102 bp repeats or partial copies thereof (filled boxes), 103 bp repeats or partial copies of (narrow grey boxes), and a putative origin of replication is labelled *ori* (grey box). Those regions that have been sequenced are distinguished by bold lines above open reading frames. The extent of the sequences that are similar in strain A198 and strain C305 are indicated by dotted lines.

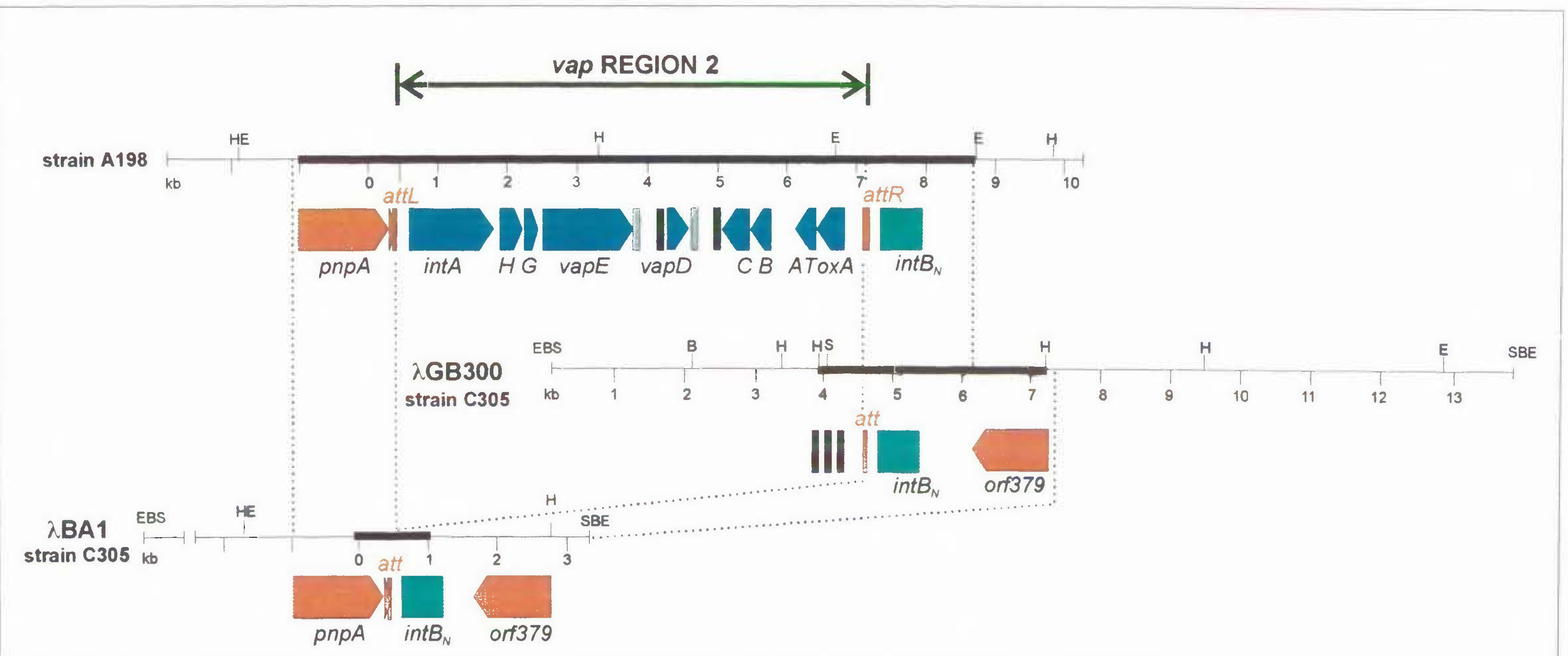

Figure 6.2: Maps showing the alignment of *intB$_N$* genes from *D. nodosus* strains A198 and C305. The two lambda clones λGB300 (middle) and λBA1 (bottom) that both contain copies of *intB$_N$* were isolated from the same C305 genomic DNA library are shown. λGB300 contains a 3.3 kb *Hin*dIII fragment that hybridised to the *intB* probe in the original laboratory strains of C305 (C3051), whilst λBA1 contains the 4.2 kb *Hin*dIII fragment that hybridises to the *intB* probe in current laboratory stocks of strain C305 (C3052). Restriction sites shown include *Bam*HI (B), *Eco*RI (E), *Hin*dIII (H) and *Sac*I (S). Integrated genetic elements are distinguished as follows: *vap* element sequences (blue), *intB* element sequences (green) and sequences that are not part of an integrated element are shown also (red). EBS correspond to the multiple cloning site at the left- and right-hand ends of the insert in λGEM-11. Putative attachment sites, *att*, are indicated (narrow red boxes). Those regions that have been sequenced are distinguished by a thick black line above the appropriate region. Regions of identity between strains A198 and C305 are indicated by lines.

Since *intB* probes do not hybridise to a 3.3 kb *Hin*dIII fragment in genomic DNA from strain C3052, it was propcsed that the lambda clone which contains the 3.3 kb *Hin*dIII fragment contains part of the sequences which were lost or moved to generate strain C3052.

In order to investigate the nature, origin and prevalence of the sequences that separated *pnpA* and *intB*$_N$ in strain C3051, a sequence of 3.7 kb immediately upstream of the 3.3 kb *Hin*dIII fragment (Figure 6.2) from the lambda clone λGB300 (Bloomfield, 1997) was determined. Sequences from this region were subsequently utilised as probes in Southern blot experiments. The nature of these sequences is of interest primarily because three copies of the 102 bp repeats that are found within the *vap* regions of strain A198 are located immediately upstream from the copy of *intB*$_N$ that was present in C3051 (Figure 6.2, λGB300). The presence of these 102 bp repeats led to the hypothesis that some remnants of the *vap* region may have originally separated *pnpA* and *intB*$_N$ in the C3051 genome. However, the results of the Southern blot experiments show that these sequences are not remnants of the *vap* regions but instead are part of a new genetic element, designated the *intD* element.

Within these *intD* element sequences, open reading frames that share similarity, at both the sequence level and in their arrangement, to mobilisation (*mob*) genes from non-conjugative and conjugative plasmids, and transposons, were identified. Putative *oriV* and *oriT* regions were also identified. Analyses confirmed that these *intD* element sequences are no longer present in the C3052 genome, but all orfs were present in two of seventeen strains analysed, whilst four other strains contained some of these *mob* sequences. These results indicate that the *intD* element was acquired horizontally *via* the integration of a mobilisable or conjugative plasmid or transposon into the *D. nodosus* genome. The significance of this *intD* element in the acquisition and dissemination of other genetic elements in *D. nodosus* is discussed.

6.2 Results

6.2.1 Sequence of a 3.7 kb region from left of *intB*$_N$ from C305 lambda clone λGB300

DNA from the lambda clone λGB300 was digested with *Sac*I and *Eco*RI and a 3.7 kb fragment was purified and cloned into pUC18, generating pGW47.2 (Figure 6.3). Subclones of pGW47.2 were subsequently made such that overlapping sequences could be determined from each strand (Figure 6.3). In addition, due to the lack of appropriate restriction sites four oligonucleotides were constructed in order to sequence across restriction enzyme sites, and their approximate positions are also indicated (Figure 6.3).

A sequence of 3687 bp was determined (Figure 6.4). Base composition analysis indicates that the region has a very low G + C content of only 31.8%, which contrasts markedly with the 45% G + C content characteristic of the *D. nodosus* chromosome (Holdeman, Kelley & Moore, 1984), and thus suggests that these sequences are likely to have been acquired horizontally. For reasons that will be discussed in Section 6.3, this DNA region will be referred to from here on as the *intD* element.

The *intD* element sequence was analysed for open reading frames with start codons preceded by Shine-Dalgarno ribosome binding sites (Shine & Dalgarno, 1974), which led to the identification of six potential open reading frames, that have been designated *int*D element-associated proteins (*dap*) *A* to *E* (Table 6.1, Figure 6.5). No putative motifs were identified in any of the *dap*-encoded proteins using the WAG motif program and Prosite patterns database available at ANGIS.

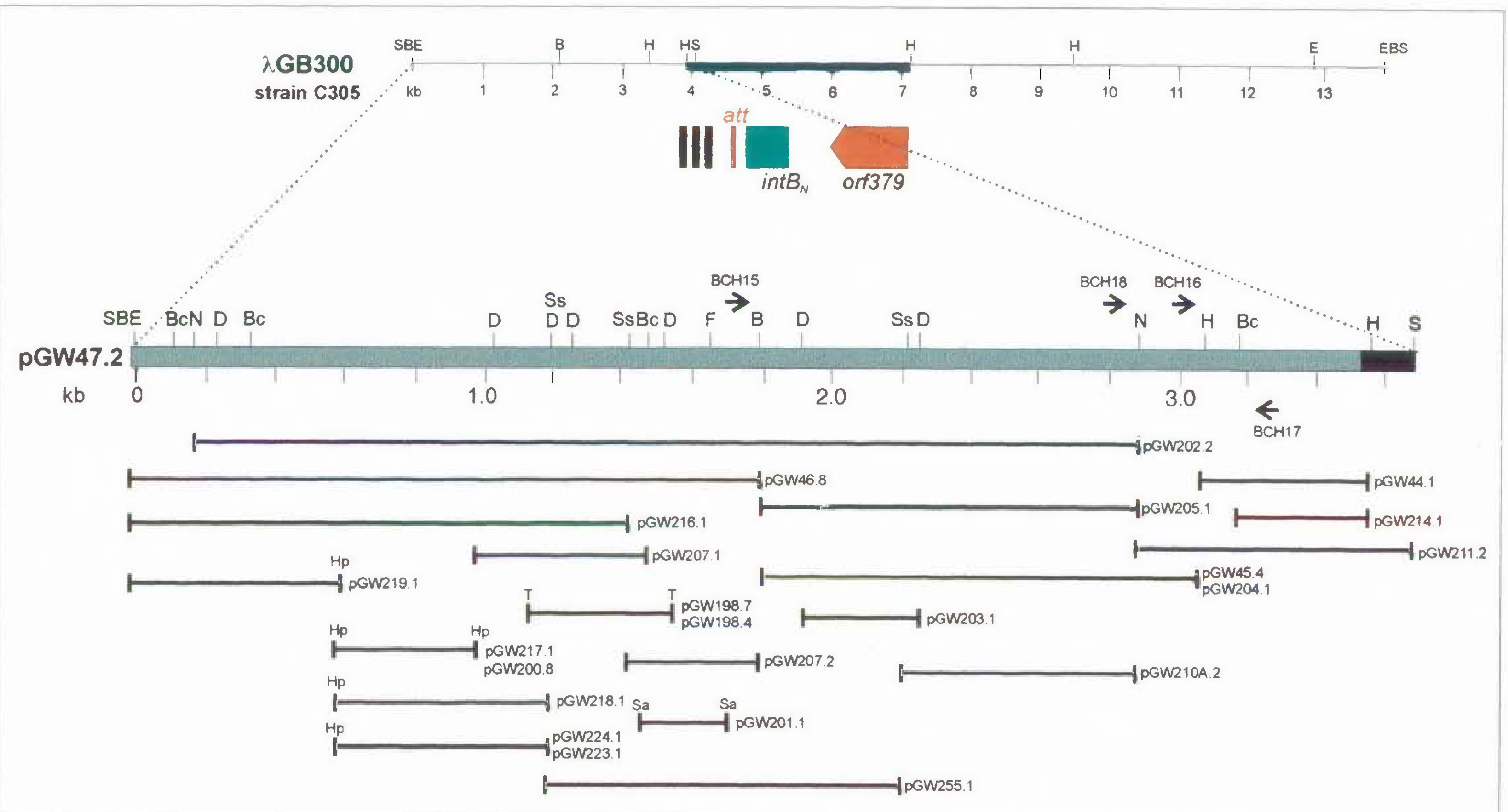

Figure 6.3: Restriction map of subclones from lambda clone λGB300, isolated from a genomic DNA library of *D. nodosus* strain C305. Restriction sites shown include *Bam*HI (B), *Bcl*I (Bc), *Dra*I (D) *Eco*RI (E), *Fsp*I (F), *Hin*dIII (H), *Nsi*I (N), *Sac*I (S) and *Ssp*I (Ss). Note that not all restriction enzyme sites are indicated on the map; restriction enzymes that cut at 4 bp recognition sites are not shown on the map, instead where 4 base recognition sites were used to clone a given fragment it is indicated at the end of the appropriate fragment as follows: *Hpa*II (Hp), *Taq*I (T), *Sau*3AI (Sa). Those regions that have been sequenced in this work are indicated by a grey box, whilst those regions that had been determined previously are distinguished by a black box. Arrows indicate the positions of oligonucleotides used for sequencing, PCR experiments, or both and are labelled BCH15-BCH18. The leftmost S, B, and E sites are from within the multiple cloning site of λ-GEM12.

```
---------+---------+---------+---------+---------+---------+---------+
TTTGGCGGCGACATTTCTCTCATGTGATGATGTGTACAACTCTTCAAATGTGATTTTTTCATTTCTGACC 70
AAACCGCCGCTGTAAAGAGAGTACACTACTACACATGTTGAGAAGTTTACACTAAAAAAGTAAAGACTGG
  K  A  A  V  N  R  E  H  S  S  T  Y  L  E  E  F  T  I  K  E  N  R  V

ATGTCAAGGCATTGATCATTTTGCAGTAGATTCGGCATAGATAATGGTTTTCGTTGCCCGCGATTTCCAT 140
TACAGTTCCGTAACTAGTAAAACGTCATCTAAGCCGTATCTATTACCAAAAGCAACGGGCGCTAAAGGTA
 M  D  L  C  Q  D  N  Q  L  L  N  P  M  S  L  P  K  R  Q  G  R  N  G
                      NsiI
TACCACGGCGTGTTCGCAACTCATGCATACTGAATCGACTGTTTCTCTGCTCAATTTGGGTTGTATCATT 210
ATGGTGCCGCACAAGCGTTGAGTACGTATGACTTAGCTGACAAAGAGACGAGTTAAACCCAACATAGTAA
N  G  R  R  T  R  L  E  H  M  S  F  R  S  N  R  Q  E  I  Q  T  T  D  N

TAAATCTCCATAAATTTCATTAATATGCGGTGCTTCATCAAGGTTGGGAATTTCATCAAGATGACGACAA 280
ATTTAGAGGTATTTAAAGTAATTATACGCCACGAAGTAGTTCCAACCCTTAAAGTAGTTCTACTGCTGTT
  L  D  G  Y  I  E  N  I  H  P  A  E  D  L  N  P  I  E  D  L  H  R  C

TCCCCAATATATGGCACATCATCAATATGCGGAATGTGATCAATATCAACATAGGTAACTGTAATAGTTT 350
AGGGGTTATATACCGTGTAGTAGTTATACGCCTTACACTAGTTATAGTTGTATCCATTGACATTATCAAA
 D  G  I  Y  P  V  D  D  I  H  P  I  H  D  I  D  V  Y  T  V  T  I  T

CATAATTAAAGTTATGTGTGCTTTGTGGCGAATCTTCATAACCGATGTGTTTGAGATGATGATGCCACTG 420
GTATTAATTTCAATACACACGAAACACCGCTTAGAAGTATTGGCTACACAAACTCTACTACTACGGTGAC
E  Y  N  F  N  H  T  S  Q  P  S  D  E  Y  G  I  H  K  L  H  H  H  W  Q

TGCTGCTAACCATGATTTTTTTTCTGGCAAATTCATTGCTACATAGGCATCCGCGTTCATGTTTTGCTGC 490
ACGACGATTGGTACTAAAAAAAAGACCGTTTAAGTAACGATGTATCCGTAGGCGCAAGTACAAAACGACG
  A  A  L  W  S  K  K  E  P  L  N  M  A  V  Y  A  D  A  N  M  N  Q  Q

TTTTTCGACAAATCGCGCTTCGCAGGCAGCGTATTCATACCATTCATCAAGCAAATGTTGATTGTCATGA 560
AAAAAGCTGTTTAGCGCGAAGCGTCCGTCGCATAAGTATGGTAAGTAGTTCGTTTACAACTAACAGTACT
 K  K  S  L  D  R  K  A  P  L  T  N  M  G  N  M  L  C  I  N  I  T  M   ←dapA

ACAAACTCAACGTTCGCGGTGCGGTTTTCTAAATAATGATTGAGAAAAACCGGGATCTTTAAGATTAATT 630
TGTTTGAGTTGCAAGCGCCACGCCAAAAGATTTATTACTAACTCTTTTTGGCCCTAGAAATTCTAATTAA
           R  E  R  H  P  K  R  F  L  S  Q  S  F  G  P  D  K  L  N  I

CCTTTATCTTTACCAACAATTTTTAGATTGATATAAGGTTGCCCAAAACTCTTTGAATCGCGTACTGTGA 700
GGAAATAGAAATGGTTGTTAAAAATCTAACTATATTCCAACGGGTTTTGAGAAACTTAGCGCATGACACT
 G  K  D  K  G  V  I  K  L  N  I  Y  P  Q  G  F  S  K  S  D  R  V  T

CAGTACCAAAGCCTTTAATAACTTGTGCAAGTTCTGTAACTGATGAAATCTGCGGATTATTTTGAATGAT 770
GTCATGGTTTCGGAAATTATTGAACACGTTCAAGACATTGACTACTTTAGACGCCTAATAAAACTTACTA
V  T  G  F  G  K  I  V  Q  A  L  E  T  V  S  S  I  Q  P  N  N  Q  I  I

AGAAAATATGGTTTGCTTGATTTCTTTGTTAGTCATATCTGGTTTTATCTCATACCGTGCGATGACTTCT 840
TCTTTTATACCAAACGAACTAAAGAAACAATCAGTATAGACCAAAATAGAGTATGGCACGCTACTGAAGA
  S  F  I  T  Q  K  I  E  K  N  T  M  D  P  K  I  E  Y  R  A  I  V  E

TCTCGCCCTGTCTTAATTGTGCGCTGATTGTCTTTTGGGCTCTCCAATTCATATTTTTCATTAATATATT 910
AGAGCGGGACAGAATTAACACGCGACTAACAGAAAACCCGAGAGGTTAAGTATAAAAAGTAATTATATAA
 E  R  G  T  K  I  T  R  Q  N  D  K  P  S  E  L  E  Y  K  E  N  I  Y

CTTGAAACGCATTGATATATCGTTCCCTTTTATCTACTAATCCAAGTGGATTATCTCTTTTTCCGGTATA 980
GAACTTTGCGTAACTATATAGCAAGGGAAAATAGATGATTAGGTTCACCTAATAGAGAAAAAGGCCATAT
E  Q  F  A  N  I  Y  R  E  R  K  D  V  L  G  L  P  N  D  R  K  G  T  Y

                                   EcoRV
TAAATTATATTCCGGAATTACAACATGGATATGCGGATATCTTTTAAAGTTTTTTCCCTGCATATCAATT 1050
ATTTAATATAAGGCCTTAATGTTGTACCTATACGCCTATAGAAAATTTCAAAAAAGGGACGTATAGTTAA
  L  N  Y  E  P  I  V  V  H  I  H  P  Y  R  K  F  N  K  G  Q  M  D  I

AGCGATTGTATTTTTGGAAAATGAATTTCACTAAAATAATAAAATTCACCTTCAACACAAGATGAAAAAA 1120
TCGCTAACATAAAAACCTTTTACTTAAAGTGATTTTATTATTTTAAGTGGAAGTTGTGTTCTACTTTTTT
 L  S  Q  I  K  P  F  H  I  E  S  F  Y  Y  F  E  G  E  V  C  S  S  F
```

```
---------+---------+---------+---------+---------+---------+---------+

TAAATTTTCGAAACTCAACATCTATATTTTTCAATACTTCTTCACTAATGTTTTTTTCTTTAAAGGATAA 1190
ATTTAAAAGCTTTGAGTTGTAGATATAAAAAGTTATGAAGAAGTGATTACAAAAAAAGAAATTTCCTATT
I  F  K  R  F  E  V  D  I  N  K  L  V  E  E  S  I  N  K  E  K  F  S  L

ACTGATATGAAAATATTTTTGTGTTTTGCATTCTCTATCAAAACTATTTAAAACCGTATCTAATTCGGTT 1260
TGACTATACTTTTATAAAAACACAAAACGTAAGAGATAGTTTTGATAAATTTTGGCATAGATTAAGCCAA
  S  I  H  F  Y  K  Q  T  K  C  E  R  D  F  S  N  L  V  T  D  L  E  T

AAATTTCCGCATAATTGAACGCGTTGATCTAATTCATCACGCGAATAATATCTACCGCTCTTTTTACCAT 1330
TTTAAAGGCGTATTAACTTGCGCAACTAGATTAAGTAGTGCGCTTATTATAGATGGCGAGAAAAATGGTA
 L  N  G  C  L  Q  V  R  Q  D  L  E  D  R  S  Y  Y  R  G  S  K  K  G

CGCGTAAATAAGTAACAACGCCATGAGATGCGCCTCTAAATCTAATAATCATTAAACATCTTCAAAAAAA 1400
GCGCATTTATTCATTGTTGCGGTACTCTACGCGGAGATTTAGATTATTAGTAATTTGTAGAAGTTTTTTT
D  R  L  Y  T  V  V  G  H  S  A  G  R  F  R  I  I  M  <- dapB
                                                * Y  D  N  F  M  K  L  F

TATGTTGTTGCGCTTCAATCGTATCTAATATTCTCAAATAATGAATACATTGATCATAATGAAGCTGATT 1470
ATACAACAACGCGAAGTTAGCATAGATTATAAGAGTTTATTACTTATGTAACTAGTATTACTTCGACTAA
I  H  Q  Q  A  E  I  T  D  L  I  R  L  Y  H  I  C  Q  D  Y  H  L  Q  N

ATAGAGCGCTAATGAATTAAGCTGCCTTGCAATTTGATTTAAATTATTACTACTTTTTGAAATCAATCGA 1540
TATCTCGCGATTACTTAATTCGACGGAACGTTAAACTAAATTTAATAATGATGAAAAACTTTAGTTAGCT
  Y  L  A  L  S  N  L  Q  R  A  I  Q  N  L  N  N  S  S  K  S  I  L  R
                                      IHF binding site
ATGATTTTTAATCTATCTTCAATTGCTAATTTATCAATAAATTGAGGATTGGATTCTAGTAAAACTTGTC 1610
TACTAAAAATTAGATAGAAGTTAACGATTAAATAGTTATTTAACTCCTAACCTAAGATCATTTTGAACAG
 I  I  K  L  R  D  E  I  A  L  K  D  I  F  Q  P  N  S  E  L  L  V  Q

TTATGACTTCAGATAATGTATTGCCCGTTTGATCAGCAATTGCTGCATATTTTTCCTGCAATTCATCTGA 1680
AATACTGAAGTCTATTACATAACGGGCAAACTAGTCGTTAACGACGTATAAAAAGGACGTTAAGTAGACT
R  I  V  E  S  L  T  N  G  T  Q  D  A  I  A  A  Y  K  E  Q  L  E  D  S
                                                         -> DnaA box
TAATCGCAGAGTAATGCGAGTCTCTAATGCCATCTCATCTTCCTTACGTCAGTAAGTCTCCACATTGATT 1750
ATTAGCGTCTCATTACGCTCAGAGATTACGGTAGAGTAGAAGGAATGCAGTCATTCAGAGGTGTAACTAA
 L  R  L  T  I  R  T  E  L  A  M  <- dapC
                                  BamHI
TTTCCGCAATAGGGTTTCAAAGGGGGATCCCCTTTGGCACGCGACGTCTGAAAACGTCAGTTTGTCAGAC 1820
AAAGGCGTTATCCCAAAGTTTCCCCCTAGGGGAAACCGTGCGCTGCAGACTTTTGCAGTCAAACAGTCTG

AGTCGGCAACGTGCCTCAATGCTGCGGAAATAATAAATTATTAGTAAAAAATAATCAAATAAAATCCAAA 1890
TCAGCCGTTGCACGGAGTTACGACGCCTTTATTATTTAATAATCATTTTTTATTAGTTTATTTTAGGTTT
                <- 5' end
TTTATTAAATTTAAAGTTTATATTTGTGGGAAAAATATGTAATATGTGCAAAATGCACAAATAGTATTTG 1960
AAATAATTTAAATTTCAAATATAAACACCCTTTTTATACATTATACACGTTTTACGTGTTTATCATAAAC
                                                  DnaA box <-
AACACATGTGCAAAATGCACAAATAGTATTTGAACACATGTGCAAAATGCACAAATAGTATTTGAACACA 2030
TTGTGTACACGTTTTACGTGTTTATCATAAACTTGTGTACACGTTTTACGTGTTTATCATAAACTTGTGT
     DnaA box <-                         DnaA box <-

TGTGCAAAATGCACAAATAGTATTTGAACACATGTGCAAAATGCACAAATAGTATTTGAACACATGTGCA 2100
ACACGTTTTACGTGTTTATCATAAACTTGTGTACACGTTTTACGTGTTTATCATAAACTTGTGTACACGT
 DnaA box <-                          DnaA box <-                DnaA box <-
AAATGCACAAATAAGAGAAATGTTATGAAAAAAATAGAAATTGATATAAATTTAGTAAGATATTTTTTGT 2170
TTTACGTGTTTATTCTCTTTACAATACTTTTTTTATCTTTAACTATATTTAAATCATTCTATAAAAAACA
                   dapD ->  M  K  K  I  E  I  D  I  N  L  V  R  Y  F  L  S

CTCTTGGTTATTCAAGAAAAAAAATAGCACTAAAACTAAATATTCATAGAAATACATTAAGAAAATATTT 2240
GAGAACCAATAAGTTCTTTTTTTTATCGTGATTTTGATTTATAAGTATCTTTATGTAATTCTTTTATAAA
  L  G  Y  S  R  K  K  I  A  L  K  L  N  I  H  R  N  T  L  R  K  Y  L

AAAGATGAATAATATTTCCGGAGAAGTAGAATTAGAACATGATACATTAGATAAAAAATATAAGAATAAT 2310
TTTCTACTTATTATAAAGGCCTCTTCATCTTAATCTTGTACTATGTAATCTATTTTTTATATTCTTATTA
 K  M  N  N  I  S  G  E  V  E  L  E  H  D  T  L  D  K  K  Y  K  N  N
```

```
---------+---------+---------+---------+---------+---------+---------+

ATTTCATCACTAGAAAATATGAAAGAATTAAATAATTCTAAACCCAAAGAAAAATTAGATTTATCTATTT 2380
TAAAGTAGTGATCTTTTATACTTTCTTAATTTATTAAGATTTGGGTTTCTTTTTAATCTAAATAGATAAA
I  S  S  L  E  N  M  K  E  L  N  N  S  K  P  K  E  K  L  D  L  S  I  L

TAGATAAAGCTATTGAAAGAATGAAAAATAAAATGGAGTAAATATGAAAATTGCGTTGTTAAATTACAGC 2450
ATCTATTTCGATAACTTTCTTACTTTTTATTTTACCTCATTTATACTTTTAACGCAACAATTTAATGTCG
  D  K  A  I  E  R  M  K  N  K  M  E  *
                                  dapE    -> M  K  I  A  L  L  N  Y  S

GGAAATGTCGGAAAAACAACTTTAGCAAGAGATTTACTTAAATTTCGCTTACAAGAATATGAGTTAATTA 2520
CCTTTACAGCCTTTTTGTTGAAATCGTTCTCTAAATGAATTTAAAGCGAATGTTCTTATACTCAATTAAT
G  N  V  G  K  T  T  L  A  R  D  L  L  K  F  R  L  Q  E  Y  E  L  I  T
                                                              -> DnaA
CTATTGAAAGCAGTAATGCAGATGGTAAAGAAGAGTTGATTATAAGAGGTGAAGATCGTGATAAATTATA 2590
GATAACTTTCGTCATTACGTCTACCATTTCTTCTCAACTAATATTCTCCACTTCTAGCACTATTTAATAT
  I  E  S  S  N  A  D  G  K  E  E  L  I  I  R  G  E  D  R  D  K  L  Y
 Box
TACAGAATTATTATTAGCTGATGATGTAATACTAGACATCGGATCTTCTAATCTTGAATCTTTTTTTAAG 2660
ATGTCTTAATAATAATCGACTACTACATTATGATCTGTAGCCTAGAAGATTAGAACTTAGAAAAAAATTC
 T  E  L  L  L  A  D  D  V  I  L  D  I  G  S  S  N  L  E  S  F  F  K

TCAAGTAAAAAGAATCTGAGATTATCAGCAATATTAATATATTCATCGTACCGACCACGCCGGAGCTTA 2730
AGTTCATTTTTCTTAGACTCTAATAGTCGTTATAATTATATAAGTAGCATGGCTGGTGCGGCCTCGAAT
S  S  K  K  E  S  E  I  I  S  N  I  N  I  F  I  V  P  T  T  P  E  L  K

AGCAACAGCAAGATACAATTAAGACAATAAAAGATTTATTAGAAAATAATGTGAACGCTAATCAAATTCA 2800
TCGTTGTCGTTCTATGTTAATTCTGTTATTTTCTAAATAATCTTTTATTACACTTGCGATTAGTTTAAGT
  Q  Q  Q  D  T  I  K  T  I  K  D  L  L  E  N  N  V  N  A  N  Q  I  H

TATTATTGCAAATCAAGTAGTTTTTGACACATACACAACAGTTACCGATACATTTAGTGACTTGATCGCT 2870
ATAATAACGTTTAGTTCATCAAAAACTGTGTATGTGTTGTCAATGGCTATGTAAATCACTGAACTAGCGA
 I  I  A  N  Q  V  V  F  D  T  Y  T  T  V  T  D  T  F  S  D  L  I  A
        NsiI
GTAACTAATGCATTGAATATCAACTTTAATTTCAAATACGCTATTGAACGCCATGACTTGTATAAAAACG 2940
CATTGATTACGTAACTTATAGTTGAAATTAAAGTTTATGCGATAACTTGCGGTACTGAACATATTTTTGC
V  T  N  A  L  N  I  N  F  N  F  K  Y  A  I  E  R  H  D  L  Y  K  N  G

GGCAAATTTTAGCTGAAATGCTTAGTGAAGAAGATTATCGCGCAAAAATGGAAGAAGCAAAAACTCACGG 3010
CCGTTTAAAATCGACTTTACGAATCACTTCTTCTAATAGCGCGTTTTTACCTTCTTCGTTTTTGAGTGCC
  Q  I  L  A  E  M  L  S  E  E  D  Y  R  A  K  M  E  E  A  K  T  H  G

                                                         HindIII
AGATAAACGTAACGCACGAATCTTAGCAAATAAATATATTCGACAAAAAAAAATTCGAAAGCTTAATGAA 3080
TCTATTTGCATTGCGTGCTTAGAATCGTTTATTTATATAAGCTGTTTTTTTTTAAGCTTTCGAATTACTT
 D  K  R  N  A  R  I  L  A  N  K  Y  I  R  Q  K  K  I  R  K  L  N  E
          -> DnaA box
CATTATCAAATGATTTTCGACAATGTAATCAACAGCAGAGAGAAAAATTATGAATGAACATGAAAGCAGC 3150
GTAATAGTTTACTAAAAGCTGTTACATTAGTTGTCGTCTCTCTTTTTAATACTTACTTGTACTTTCGTCG
H  Y  Q  M  I  F  D  N  V  I  N  S  R  E  K  N  Y  E  *
                                      dapF   -> M  N  E  H  E  S  S

TATGCTACTGCTGATCACAAAATGCTTTTAGCCGATTGTATTTTAGAAATCAATGAATTATCAAAAATAG 3220
ATACGATGACGACTAGTGTTTTACGAAAATCGGCTAACATAAAATCTTTAGTTACTTAATAGTTTTTATC
Y  A  T  A  D  H  K  M  L  L  A  D  C  I  L  E  I  N  E  L  S  K  I  V

TTAAAGAATTACATAACGAATTTACGCTCATTCAAGAAGATATTGATACACTCAATCGATCACTTAATGA 3290
AATTTCTTAATGTATTGCTTAAATGCGAGTAAGTTCTTCTATAACTATGTGAGTTAGCTAGTGAATTACT
  K  E  L  H  N  E  F  T  L  I  Q  E  D  I  D  T  L  N  R  S  L  N  D

TTTTTATCAACAAATTCATCGCGTTAAAATAGAAACATCAGAACATCTTAGTAATGTTGTTATTGAAGCA 3360
AAAAATAGTTGTTTAAGTAGCGCAATTTTATCTTTGTAGTCTTGTAGAATCATTACAACAATAACTTCGT
F  Y  Q  Q  I  H  R  V  K  I  E  T  S  E  H  L  S  N  V  V  I  E  A
```

Figure 6.4: DNA Sequence of the 3'-end of the *intD* element from *D. nodosus* strain C3051.

The aa sequences of six putative proteins encoded by *dapA*, *dapB*, *dapC*, *dapD*, *dapE* and *dapF* are aligned with the nucleotide sequence and the orientation of these open reading frames is indicated by an arrow. The following features are indicated: Potential *DnaA* boxes at positions 1736, 2586, 3094 and at 1943, 1975, 2008, 2039, 2072, 2104 in the complementary strand; a putative IHF binding site at position 1572 (boxed and shaded yellow); stop codons for *orfs* (red asterisks); putative Shine-Dalgarno sequences (green); 6 x 32 bp repeats comprising a putative origin of replication (double-headed red arrows)and the 118 nt region (putative *oriT*) with similarity to the *vap* regions 1/3 from *D. nodosus* strain A198 (red type, orientation indicated by an arrow at the 5'-end of the sequence). Within the 118 nt sequence two inverted repeats of 8 bp and 9 bp respectively are underlined. Those nucleotides that differ from the corresponding sequence in the *vap* element sequence are distinguished by black type; part of the region with similarity to the 102 bp repeats from the *vap* element is shown in pink and their inverse orientation is indicated, whilst grey shading shows the limits of these repeats. It should be note that only repeat 3 and part of repeat 2 are shown since the rest of repeat 2 and repeat 1 are to the right of the *Sac*I site shown. 10 bp intervals are indicated (+) at the top of each page. Restriction enzyme sites shown include *Bam*HI, *Eco*RV, *Hin*dIII, *Nsi*I and *Sac*I.

Table 6.1: Sequence analysis of the *intD* element from *D. nodosus* strain C305

Gene	Co-ordinates 5'-3- (nt)	Size (aa)	% identity to putative protein	Homologues/Description	Accession Number	P(*n*)
dapA	3130->1 (comp)	>186	-	NSH	-	-
dapB	1381-568 (comp)	271	38.9/54aa	Colicin E1 plasmid encoded MbeA (517aa) mobilisation protein	X15873	1.1
			34.7/72aa	*Escherichia coli* plasmid pColD-157 MbdA (532aa) mobilisation protein	Y10412	1.1
			20.8/144aa	*Lactococcus lactis* plasmid pC1528 putative mobilisation protein (410aa)	L08601	3.8
			27.1/118aa	*Erwinia stewartii* plasmid encoded MobA (499 aa) mobilisation protein	L42525	8.0
			20.8/144aa	Lactococcal plasmid pNZ4000 MobA (416aa) mobilisation protein	AF036487	13
dapC	1712-1373 (comp)	113	30.7/75aa	*Bacteroides fragilis* MocB from Tn*4399* mobilisation cassette (143 aa)	L20975	6.6e-03
			33.7/77aa	*Pasteurella haemolytica* unidentified orf upstream of MbeA	M8317	2.2e-01
			33.8/70aa	*E. coli* plasmid pColD-157 MbdC (115 aa) mobilisation protein	Y10412	8.2e-01
			32.5/77aa	Plasmid pLS88 unidentified orf upstream of RSF1010-like *oriV*	L23118	6.5e-01
			37.3/51aa	*Erwinia carotovora* plasmid pEC3 MobC (115 aa) mobilisation protein	D45188	3.7
			30.0/71aa	*Erwinia stewartii* plasmid pSW200 MobC (107aa) mobilisation protein	L42525	7.0
dapD	2125-2419	98	26.6/94aa	*Lactobacillus johnsonii* IS*1253* protein OrfA (177aa)	U09558	2.8e-01
			31.4/79aa	*Borellia burgdoferi* plasmid cp18 encoded OrfE protein (190aa)	U42599	8.7e-01
			38.3/60aa	*Streptococcus pyogenes*, 5'-end of RecF protein (369aa)	U07342	9.3e-01
dapE	2424-3135	237	28.8/222aa	*Salmonella typhimurium* plasmid pKM101-encoded StbB (238 aa)	U43676	3.5e-13
			30.6/85aa	*Methanococcus jannaschii* putative orf MJ1054 (895aa)	U67548	6.2e-01
dapF	3130-3439	103	28.9/76aa	Plasmid pWQ799 encoded MdeD (77aa) mobilisation protein	L39794	1.7
			28.6/63aa	Plasmid Colicin E1 MbeD (77aa) mobilisation protein	X15873	2.4
			25.9/77aa	*Enterococcus faecalis* unidentified orf within *tetM* of Tn916-like TnF01	X92947	2.6
			25.9/77aa	*Gardnerella vaginalis* unidentified orf within *tetM*	U58985	5.6
			25.9/77aa	*Neisseria gonhorrhoea* plasmid p02100 unidentified orf within *tetM*	L12241	5.6

NSH indicates that there is no significant homology to sequences in databases.

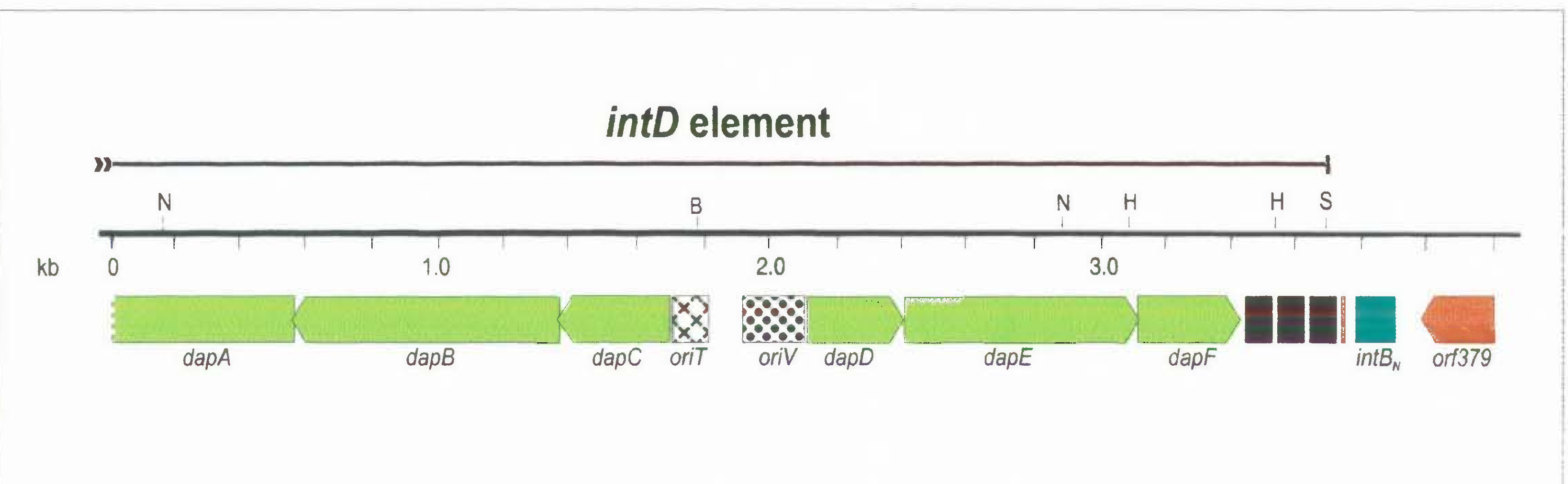

Figure 6.5: Map of the right-hand end of the *intD* element from *D. nodosus* strain C305. The numbers show the distance in kb. Restriction sites shown include *Bam*HI (B), *Hin*dIII (H), *Nsi*I (N) and *Sac*I (S). The major potential genes *dapA-F* are indicated by green arrows. The 118 nt sequence also present in the *vap* element is indicated by a crossed box. Repeated sequences are indicated as follows: 32 bp repeats comprising a putative origin of replication (spotted box), 102 bp repeats (narrow black rectangles), and putative attachment site (narrow red rectangle). The sequences that flank the right hand end of the *intD* element, including the $intB_N$ gene and *orf379*, are also shown.

Of the six putative *dap* genes identified, four share amino acid similarity with mobilisation (*mob*) genes that are, in most cases, clustered together on various genetic elements that include several mobilisable plasmids (Boyd, Archer & Sherratt, 1989; Chang *et al.*, 1992; Dixon, Albritton & Willson, 1994; Dorrington & Rawlings, 1990; Hofinger, Karch & Schmidt, 1998; Lucey, Daly & Fitzgerald, 1993a; Lucey, Daly & Fitzgerald, 1993b; Nomura, Yamashita & Murooka, 1996), a conjugative plasmid (pKM101) (Winnans 1998, GenBank Accession U43676) and a mobilisable transposon (Tn*4399*) (Murphy & Malamy, 1993).

Whilst large conjugative plasmids are transferred autonomously, smaller plasmids are only transferred (mobilised) in the presence of a conjugative element (Dorman, 1994). Such conjugative elements encode proteins that provide the initial pilus-mediated cell contact, induce formation of a conjugative pore, and specify other morphological functions. For the F plasmid, conjugative functions require the products of at least 20 genes and >35 kb of sequence (Ippen-Ihler & Minkley, 1986). In contrast, *mob* genes typically encode proteins that would function in the nicking of the *oriT* region, the subsequent piloting of the 5'-end of the nicked strand into the recipient, and recircularisation and priming of the complementary strand for replication in the recipient (Boyd *et al.*, 1989). These *mob* proteins are encoded by comparatively few genes, and in *E. coli* plasmid ColE1 there are only four genes stretching over 2.1 kb that are essential for mobilisation processes (Boyd *et al.*, 1989), whilst the mobilisation region of a *Bacteroides* insertion element consists of a single *mob* gene and an *oriT* stretching over only 1.65 kb (Li, Shoemaker & Salyers, 1995). *mob* genes are found on both conjugative elements and mobilisable non-self transmissible elements (Dorman, 1994).

6.2.1.1 The *dapA*, *dapB* and *dapC* genes

Putative genes *dapA*, *dapB* and *dapC* are in the same orientation, and are in an operon-like arrangement since the start codon of *dapB* overlaps with the 3'-end of *dapC* by seven amino acids, and similarly the start codon of *dapA* overlaps with the 3'-end of *dapB* by eleven amino acids (Figure 6.4). Although the upstream regions of these putative open reading frames were scanned for potential promoter sequences similar to the *E. coli* σ^{70} consensus sequence (Hawley & McClure, 1983), none were identified.

The *dapA* gene is still open at the end of the region that was sequenced, and consequently only the first 186 aa of the putative DapA protein are available for comparison with known protein sequences. The putative DapA protein shows no similarity to sequences in the GenBank databases.

DapB has amino acid similarity to mobilisation proteins 'A', encoded by *E. coli* ColicinE1 plasmid (Boyd *et al.*, 1989) and colicin-related plasmids (Boyd *et al.*, 1989), pCI528 and pNZ4000 from *Lactococcus* spp.(Lucey *et al.*, 1993a; Lucey *et al.*, 1993b) and *Erwinia stewartii* plasmid, pSW200 (Table 6.1). Although the level of similarity is not high, this similarity is likely to be significant, since other *dap* genes also show similarity to mobilisation proteins (Table 6.1), and in addition, the *dap* mobilisation-like proteins are arranged like related mobilisation regions.

It is notable that the mobilisation proteins, such as *mbeA*, that are most similar to DapB are encoded by significantly larger genes (Table 6.1 and Figure 6.6). However, it has been reported that, in *E. coli* colicin plasmid ColE1 and related-plasmids, the *mbeA* transcript is poorly translated (Boyd *et al.*, 1989), and these authors have proposed that two internal open reading frames (*mbeB* and *mbeD*)are expressed from within the *mdeA* gene in these plasmids to compensate for the low level of *mbeA* translation.

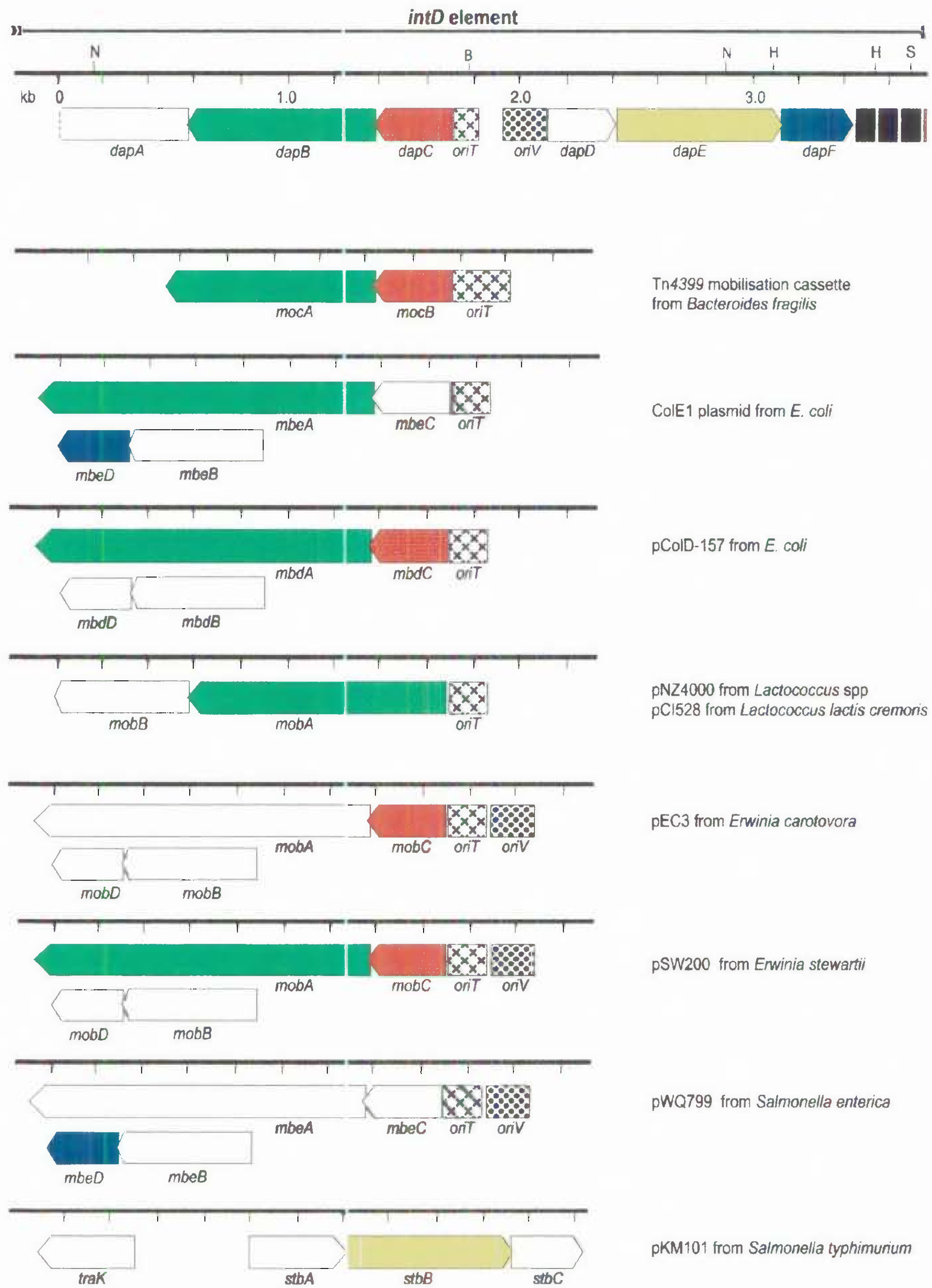

Figure 6.6: Map showing the alignment of sequences at the right-hand end of the *intD* element from *D. nodosus* strain C305 (top) and mobilisation regions from mobilisable elements that have similarity to *intD* element putative proteins. Genes with similarity to *dap* genes are indicated as follows: similarity to *dapB* (green), to *dapC* (red), to *dapE* (yellow), to *dapF* (blue). Where there is no apparent similarity to the mobilisation regions shown in this figure the orfs are unshaded. Confirmed and putative *oriT* regions (crossed boxes) and *oriV* regions (spotted boxes) are also indicated. Restriction sites shown in the *intD* element are as follows: *Eco*RI (E), *Hin*dIII (H), *Nsi*I (N) and *Sac*I (S). The distance in kb is indicated in the map of the *intD* element (top).

The *dapB* gene does not contain smaller open reading frames with similarity to *mbeB* and *mbeD*, however, a gene 1.7 kb upstream from *dapB* (*dapF*, see below) does encode a protein which is similar to MbeD (Figure 6.6). Thus it is possible that collectively the DapB and DapF proteins may perform the functions normally encoded by the larger *mbeA* transcript and similar large homologues. Although the proteins encoded by *dapB* homologues have been shown to be required for mobilisation, their precise functions in the mobilisation process have not been determined.

Like DapB, DapC also has a high degree of aa identity to mobilisation proteins from several plasmids (Table 6.1 and Figure 6.6) but is most similar to a protein encoded by the *mocB* gene from *Bacteroides fragilis* mobilisable transposon, Tn*4399* (Hecht & Malamy, 1989; Hecht, Thompson & Malamy, 1989; Murphy & Malamy, 1993). DapC has 30.7% identity over 75 aa to the MocB (143 aa) protein (Figure 6.7). Although it has been demonstrated that MocB is required for mobilisation (Murphy & Malamy, 1993) the exact role of the MocB protein in mobilisation has not yet been elucidated.

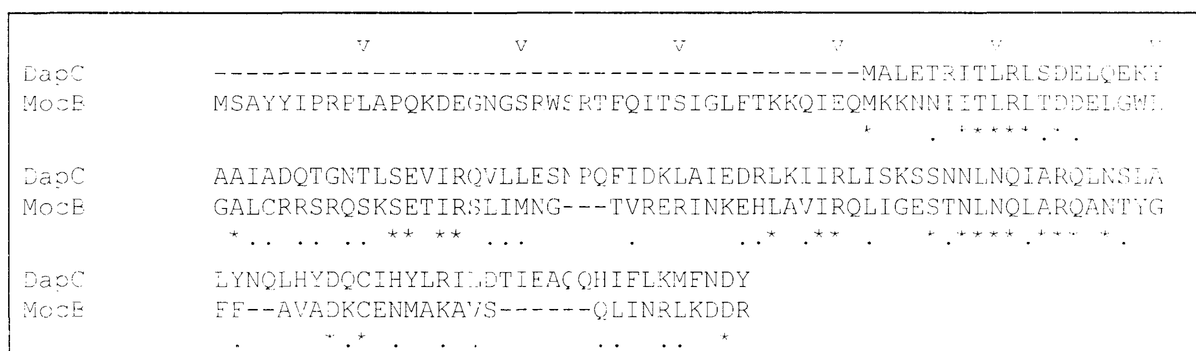

Figure 6.7: ClustalW alignment of putative DapC protein (top) from the *D. nodosus intD* element, and the MocB protein (bottom) from the *B. fragilis* mobilisable transposon Tn*4399*. Amino acids that are identical in both protein sequences are indicated by asterisks; conserved amino acid residues are indicated by a period; 10 aa intervals are shown (v).

It is interesting that the 5'-end of the *mocB* gene (1040-1471) overlaps the origin of transfer region (*oriT* – 991-1189) of Tn*4399* (Murphy & Malamy, 1993), since it is downstream of this region of overlap that the MocB sequence begins to align with the

DapC sequence (Figure 6.7). Furthermore, although MocB is required for mobilisation of Tn*4399*, a 19 aa deletion of the *N*-terminus of MocB still allowed mobilisation at only a slightly decreased frequency (Murphy & Malamy, 1993). Thus, it has been suggested that the truncated *mocB* could encode a 'restart protein' of 102 aa, which compares well with the size of DapC which consists of 113 aa.

The *mocB* gene is located immediately adjacent to an *oriT*, and is in an operon-like arrangement with *mocA* (318 aa), that has similarity to *dapB*. Collectively the *oriT* region, *mocB* and *mocA* genes comprise a 2.8 kb 'mobilisation cassette' (Murphy & Malamy, 1993). Derivatives of non-conjugative plasmids that carry the mobilisation cassette region from Tn*4399* can be mobilised by IncP plasmids RK231 or R751 in *E. coli* (Murphy & Malamy, 1993). In *B. fragilis,* chimeric plasmids containing Tn*4399* are able to be transferred from *B. fragilis* donor cells to either *B. fragilis* or *E. coli* recipient cells in the absence of other plasmids or transfer factors, but not from *E. coli* donor to recipient cells (Murphy & Malamy, 1993). These observations suggest that the requirements for transfer from *B. fragilis* differ from those in *E. coli*.

Although no similarity between DapB and MocA was detected during database searches, *(i)* the high degree of similarity between proteins encoded by *dapC* and *mocB* proteins immediately upstream from *dapB* and *mocA* respectively (Table 6.1); and *(ii)* the similar size and operon-like arrangement of the *D. nodosus dapC/dapB* genes and *B. fragilis mocB/mocA* coding regions; suggest that *mocA*- and *dapB*-encoded proteins may have similarity (Figure 6.6) that was not detected during database searches, since these algorithms are designed to detect regions of identity rather than similarity.

A ClustalW alignment of the DapB and the MocA protein (Figure 6.8) shows that the two proteins are of similar size and share 38% aa similarity and 18.8% identity in 271 aa, and thus do share some similarity. However, MocA contains a 14 aa motif that is

common to nine confirmed DNA relaxases (Murphy & Malamy, 1993) that is not present in the DapB protein (Figure 6.8). This suggests that DapB and MocA may not perform the same function. Confirmed DNA relaxases interact specifically with their respective target regions to create nicked open circular plasmid DNA for conjugative transfer.

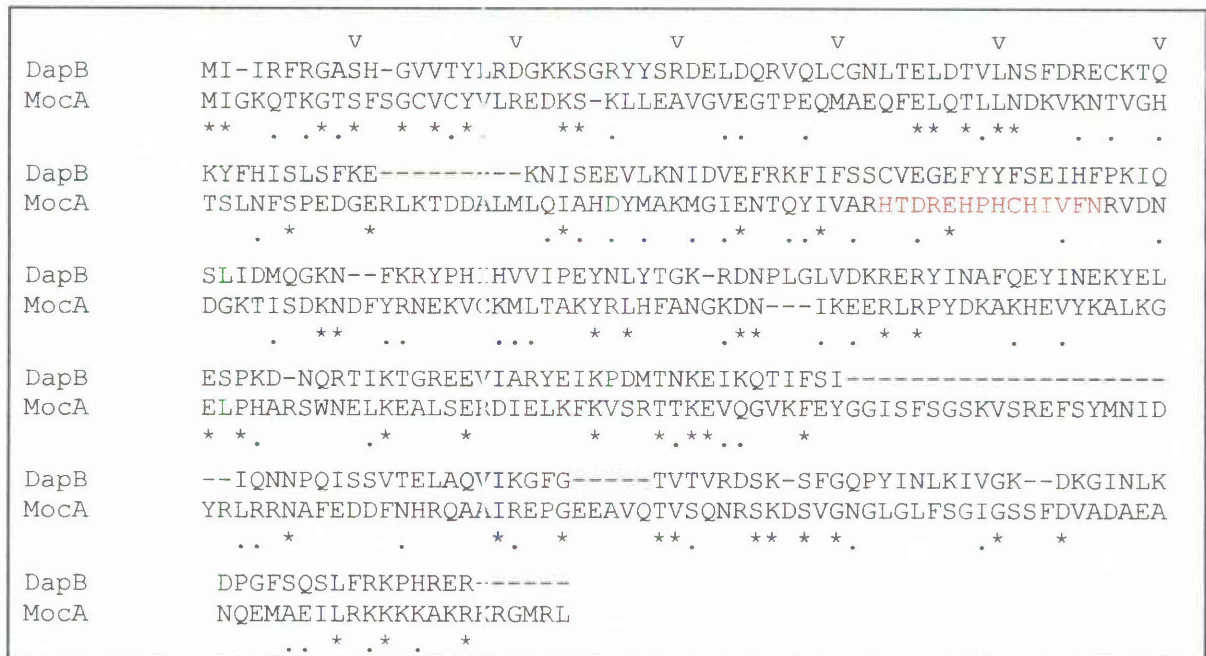

Figure 6.8: ClustalW alignment of putative protein DapB (top) from the *D. nodosus intD* element, and the MocA (bottom) protein from the *B. fragilis* mobilisable transposon Tn*4399*. Amino acids that are identical in both protein sequences are indicated by asterisks; conserved amino acid residues are indicated by fullstops; 10 bp intervals are shown (v); and the 14 aa that correspond to the relaxase motif that is present in the MocA protein, but not present in putative protein DapB, is shown in red type.

To determine whether any of the other mobilisation proteins with similarity to DapB (Table 6.1) contained a relaxase motif (Panseqrau & Lanka, 1991), the protein sequences of these MobA-type proteins were compared with MocA in a multiple sequence alignment and were also subject to a motif search. These analyses showed that none of the other proteins with similarity to DapB contained a motif that matched the consensus sequence for the relaxase motif, though the 2.1 kb region containing these mobilisation genes in related-plasmids is known to be sufficient for mobilisation. No other putative motif was identified during a multiple sequence alignment of these proteins.

It is of interest that in all of the proteins with similarity to DapC (Table 6.1) there is an amino acid sequence of 13 aa that is highly conserved and which therefore is likely to comprise a motif of functional significance for the DapC-like proteins (Figure 6.9). Although most of the DapC-like proteins have been shown to be required for mobilisation in *cis*, the mechanism by which this group of proteins achieve mobilisation has not yet been deduced.

```
                             v         v         v         v         v         v
DapC                LLE-SNPQFIDKLAIEDRLKIIRLISKSSNNLNQIARQLNSLALYNQLHYDQCIHYLRIL

MbdC                CLG-APPSKTSGLPTLAPPLLRQFASV-GNNLNQIARKINSGQWSGHDRVHVVAALMAIG
MobC (c)            CLD-EKPARAGKLPSISPALLRQLAGM-GNNLNQIARQVNAGGGSGHDRVQVVAALMAID
MobC (s)            CLD-EKPARAGKLPSISPALLRQLAGM-GNNLNQIARQVNAGGGSGIDRVQVIASLMAID
MocB                IMNGTVRERIN--KEHL-AVIRQLIGE-STNLNQLARQANTYGFF-AVADKCENMAKAVS
Ph Mob              ALE-QQPKRQP--KVIDPALLFELNRI-GVNLNQIARQCNSQKPS-IDLVSVLATLREIE
Hd Mob              PLE-QQPKRQP--KVIDPALLFELNRI-GVNLNQIARQYNSQKPS-IDLVSVLATLHCTR
                     .                  .         ****.**. *.
consensus sequence                    (S/G) (N/T/V)NLNQ(I/L)AR(K/Q)N(S/A/T)
```

Figure 6.9: ClustalW alignment of putative motif present in proteins that are similar to DapC. The putative motif is shown in blue type, and a 13 aa consensus sequence is shown below the aligned sequences. Sequences shown from top to bottom are as follows: DapC (*D. nodosus intD* element), MbdC (*E. coli* pColD-157), MobC (*E. carotovora* pEC3), MobC (*E. stewartii* pSW200), MocB (*B. fragilis* mobilisable transposon, Tn*4399*), Ph. Mob (*P. haemolytica* unidentified orf upstream of *mbeA*) and Hd. Mob (*H. ducreyi* unidentified orf on pLS88). Amino acids that are identical in all protein sequences are indicated by asterisks; conserved amino acid residues are indicated by a period; 10 bp intervals are shown (v).

dapA, *dapB* and *dapC* are separated from *dapD*, *dapE* and *dapF* by 413 bp of DNA (Figure 6.5). Within this 413 bp sequence, and 9 bp upstream of *dapC*, is a sequence of 118 nt (position 1850-1971 in complement) that has 89.3% identity in 118 nt to a sequence that is found between a putative origin of replication and a 102 bp repeat from within *vap* region 1 (position 6995-7113, GenBank accession L22145) of *D. nodosus* strain A198 (Figure 6.10). Within these 118 nts there is a sequence 5'-GTCTCCACA-3'(position 1736) that corresponds to the consensus sequence for a *DnaA* box, 5'-TTATCCACA-3', (Fujita, Yoshikawa & Ogasawara, 1989). This putative *DnaA* box is not present within the corresponding sequence from *vap* region 1/3 (Figure 6.6).

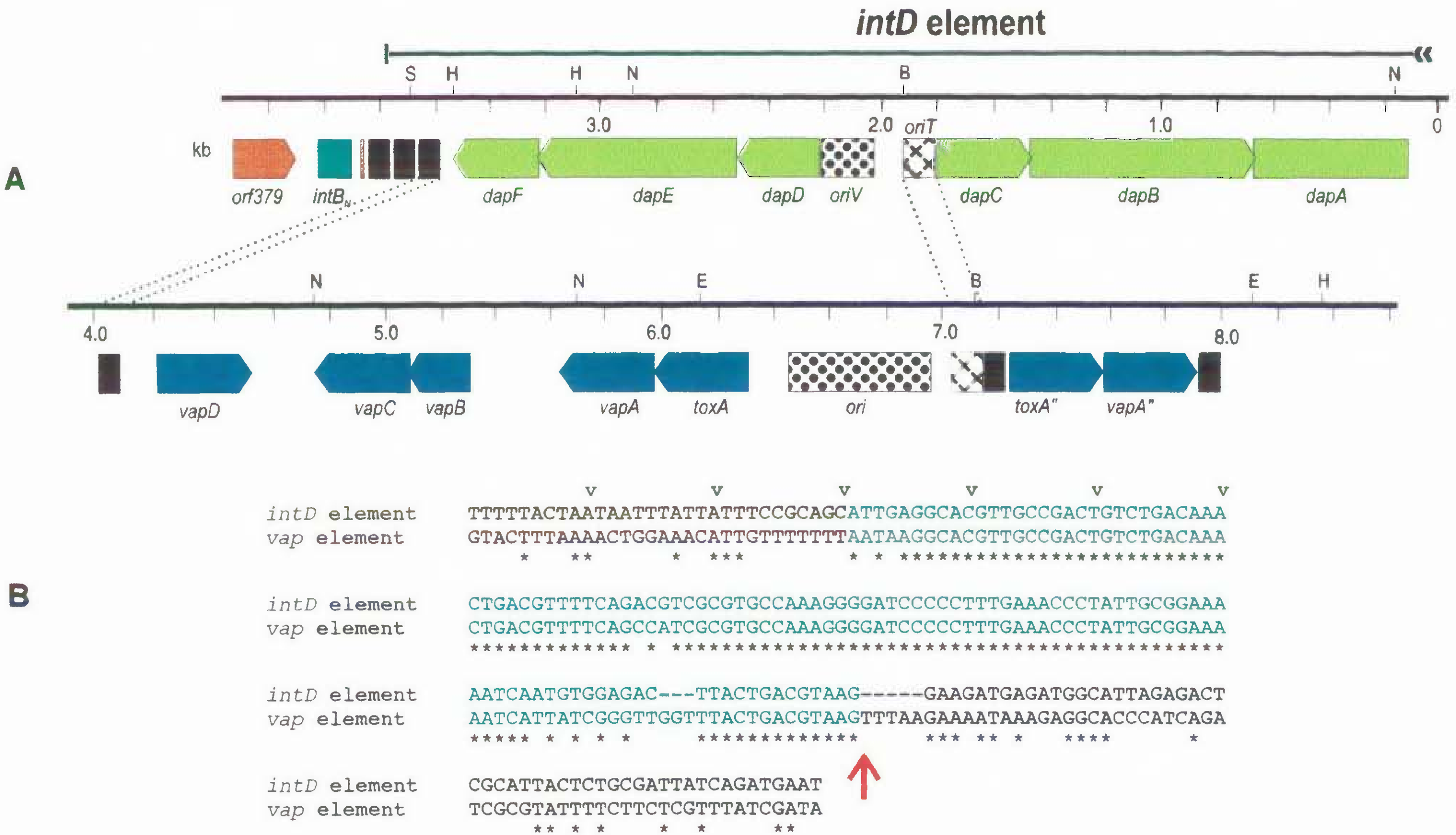

Figure 6.10: Map showing regions of alignment between the right-hand end of the *intD* element from *D. nodosus* strain C305 and *vap* region 1 from *D. nodosus* strain A198 (**A**); and a ClustalW alignment (**B**) of the nucleotide sequence of the putative *oriT* region from the *intD* element (top) with the homologous region from the *vap* element (bottom). In **A** the numbers shown indicate the distance in kb and regions of alignment are indicated by dotted lines. Restriction sites shown include *Bam*HI (B), *Eco*RI (E), *Hin*dIII (H), *Nsi*I (N) and *Sac*I (S). The major potential genes *dapA-F* are represented by green arrows, and genes flanking the *intD* element ($intB_N$ and *orf379*) are also shown. Other features are as follows: putative *oriT* (crossed box), putative *oriV* (spotted box), 102 bp repeats (black boxes), putative *attR* site (red rectangle). In **B** those bases that are identical are indicated by asterisks; 10 bp intervals are shown (v); and the putative *oriT* regions are shown in green type. A red arrow indicates the first nucleotide at the 3'-end of a 102 bp repeat in the *vap* element sequence. Note that the *intD* element is drawn in the opposite orientation to that in previous figures.

6.2.1.2 Identification of a putative *oriV*

At position 1934, there are five perfect direct repeats of the 32 bp sequence 5'-ATGTGCAAAATGCACAAATAGTATTTGAACAC-3', that are followed immediately by a sixth partial repeat in which only the first 19 nts of the 32 bp repeat unit are present (Figure 6.4). The repeated sequences are immediately preceded by 98 bp which is A + T rich (83% A + T), and followed by 98 bp of similarly high A + T content (81% A + T). Within each of these six repeat units there is a 9 bp sequence that corresponds to the consensus sequence for a DnaA box (positions 1943, 1975, 2007, 2039, 2072, 2104 on the complement strand). This region is likely to be an origin of replication (*oriV*).

Initiation of replication at the *E. coli* origin of replication (*oriC*) requires binding of the DnaA protein to the four DnaA binding sites at the *E. coli oriC* (Marians, 1992). Binding of DnaA results in the melting of the A + T rich region found adjacent to the *oriC* and guides the DnaB (helicase) and DnaC (primase) protein complex in solution to their place in the replication bubble (Marians, 1992). In addition, one or more copies of the 9 bp DnaA box sequence are found in the origins of bacteriophage P1 and plasmids ColE1, pBR322, CloDF13, R1, R100 and F, though replication of ColE1, pBR322, R1, F and P1 can occur *in vivo* in the absence of *dnaA* gene function (Fuller, Funnell & Kornberg, 1984). A DnaA box sequence is also required for the transposition of Tn5 (Johnson & Reznikoff, 1983). It is not clear why so many sequences (9 in total) matching the consensus sequences of a *DnaA* box should be present within the *intD* element.

In addition, just upstream (position 1572-1584) of the 413 nt that separates *dapA-C* from *dapD-F* is the sequence 5'-TATCAATAAATTG-3' that corresponds to the consensus sequence for an Integration Host Factor (IHF) binding site, WATCAA(N)$_4$TTR, where N is any base, R is any purine and W is A or T (Birge, 1994). IHF sites are often found at origins of replication, and are involved in DNA bending at or near the origin to facilitate denaturation of the A + T rich region (Marians, 1992).

The binding of IHF to the target DNA introduces a sharp bend of >160° in the DNA which is required in order to bring distant sequences together. The synapse between distant sites increases the probability of reactions that are dependent on interaction between these sites (Dorman, 1994; Friedman, 1988). IHF is also known to be involved in site-specific recombination, in which IHF is required in addition to Int (integrase), Xis (excisionase) and FIS (factor for inversion stimulation) in order to catalyse recombination at normal efficiency (Friedman, 1988). IHF also has a role in transcriptional regulation by bringing transcriptional activators bound to distant enhancer sequences into contact with promoter-bound RNA polymerase in order to form an open transcription complex. IHF is known to be involved in the regulation of virulence genes from *S. flexineri* (*virB* expression), *P. aeruginosa* (*algD* expression), *N. gonorrhoeae* (pilus production) and *E. coli* (type 1 fimbriae synthesis) (Friedman, 1988; Porter & Dorman, 1997).

Often the Factor for inversion stimulation (FIS) also binds at the *oriC* at several sites, and also bends the DNA (Marians, 1992). A sequence that resembles the consensus sequence 5'-GNtcAaaTttTgaNC-3' (Finkel & Johnson, 1992) for an FIS binding site was located within the *intD* sequence (5'-GGGCAAATTTTAGCT-3' at position 2940), although it should be noted that the last base is a T instead of a C. Although Fis is a sequence-specific DNA binding protein, Fis binding sites are difficult to identify (Finkel & Johnson, 1992) because they lack an obvious consensus sequence. The Fis consensus sequence of 15 nt was originally defined as 5'-GNNYRNNTNNYRNNC-3', and, despite its apparent degeneracy, more than 50% of thirty seven sequences that have so far been characterised in DNase footprinting studies vary from the consensus by one or more nucleotides (Finkel & Johnson, 1992). Hence, it has been proposed that there are factors other than the DNA sequence which define a Fis binding site, such as the local shape of the DNA. Some of these include an A + T rich region, and a run of 3-6 A or T residues flanking the core sequence (Finkel & Johnson, 1992). FIS is a histone-like DNA-bending protein which was originally identified as the factor for inversion stimulation of

homologous Hin and Gin site-specific recombinases of *Salmonella* and phage Mu respectively, but has since been shown to function in site-specific recombination of bacteriophage λ, transcriptional activation of *rRNA* and *tRNA* operons, and *oriC*-directed DNA replication (Dorman, 1994).

Due to the presence of multiple iteron-like direct repeats (position 1934-2122), flanked by a *DnaA* box, a region of high A + T content, an IHF binding site and a putative FIS site, this region resembles the *ori* of several plasmids and bacteriophages (Marians, 1992; Nordstrom, 1990; Solar *et al.*, 1998) (Figure 6.11). Such sites are essential components of the origin of replication region of autonomously replicating elements since they are required to organise a functional replisome (Solar *et al.*, 1998). Consequently, it is proposed that the 413 bp region of the *intD* element that separates *dapA-C* from *dapD-F* contains a putative origin of replication.

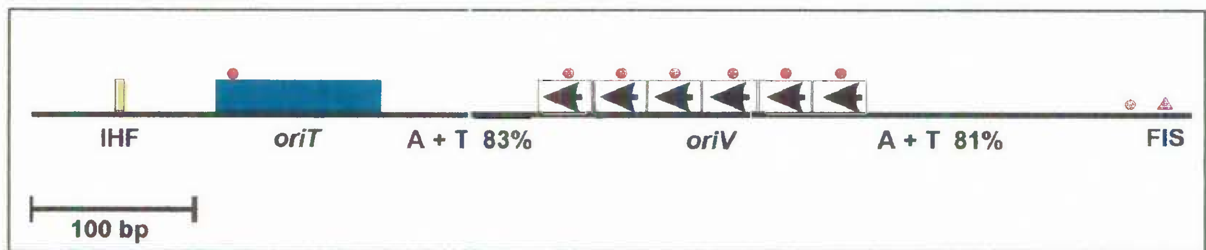

Figure 6.11: The putative origin of replication (*oriV*) region and flanking sequences of the *intD* element. The symbols used are as follows: IHF binding site (yellow rectangle); *DnaA* boxes (red spots); putative *oriT* (blue box); six 32 bp direct repeats are indicated by arrows within six boxes; FIS-binding site (red triangle). A + T rich regions are indicated.

6.2.1.3 Identification of putative *oriT* in the *intD* element

Adjacent to the putative origin of replication regions present in the *intD* element and the *vap* element, there is a 118 nt sequence which is identical in both elements (Figure 6.10). Previous work has already established that the *vap* element was acquired *via* horizontal transfer (Cheetham *et al.*, 1995a; Cheetham & Katz, 1995; Cheetham *et al.*, 1995b; Cheetham, Whittle & Katz, 1998). It is also known that the *intD* element

sequences were lost from C3051 to generate current laboratory strain C3052 (Section 6.1) and thus the *intD* element sequences are apparently mobile (Section 6.2.2.1). They have a G + C content approximately 13% below that of the *D. nodosus* chromosome, encode proteins homologous to proteins encoded by episomes and thus, like the *vap* regions, are also likely to have been acquired horizontally.

Consequently, it is proposed that the 118 nt sequence (Figure 6.10) may be important for the horizontal transfer of some mobile genetic elements in *D. nodosus*, and may comprise an origin of transfer (*oriT*). This hypothesis is supported by the arrangement of genes that show similarity to the *dap* genes (Figure 6.6) since in all cases *dapC*-like genes are located immediately adjacent to an *oriT*. In addition, the *oriT* regions that have so far been described for Gram-negative conjugative and non-conjugative plasmids do consist of short regions of between 100-300 bp and contain inverted repeats (Willetts & Wilkins, 1984). Two such inverted repeats have been identified within the putative *oriT* region of the *intD* element (Figure 6.4).

The putative *oriT* region from the *intD* element was aligned with the putative *oriT* regions of related plasmids (Figure 6.10), but no similarity to these *oriT* regions, nor to other *oriT* sequences available from the GenBank database was identified. However, it has previously been reported that even the *oriT* regions of closely related plasmids, such as those in the *E. coli* IncF group which have genetically similar transfer regions and encode related pili have different *oriT* sequences (Willetts & Wilkins, 1984). Similarly *oriT* regions of non-conjugative plasmids such as ColE1, RSF1010 and pSC101 also differ from each other. Even the *oriT* regions of closely related plasmids ColE1 and ColE2 are different despite being mobilised by the same IncI conjugative plasmids (Warren, Twigg & Sherratt, 1978). Whether the 118 nt sequence from the *intD* element does correspond to an origin of transfer would need to be determined experimentally.

6.2.1.4 The *dapD*, *dapE* and *dapF* genes

Sixteen nucleotides downstream from the putative origin of replication in the *intD* element is the putative gene *dapD*, which is followed by *dapE* and *dapF* respectively. *dapD* encodes a putative protein of 98 aa that has 26.6% identity in 94 aa to OrfA (177 aa) from *Lactobacillus johnsonii* insertion sequence, IS*1223* (Walker & Klaenhammer, 1994). It has been proposed that the 1.5 kb IS*1223* element produces an active transposase *via* translational frameshifting between two tandem overlapping orfs, *orfA* and *orfB*. It seems unlikely that DapD and DapE produce such a transposase given the high similarity between DapE and an unrelated protein from pKM101 (see below). DapD also shows 31.4% identity in 79 aa to part of putative protein OrfE from *Borrelia burgdoferi* plasmid cp18 which has an unknown function (Stevenson *et al.*, 1997) and to the 5'-end of RecF from *Streptococcus pyogenes* which is involved in ssDNA-binding during recombination and DNA repair (DeAngelis, Yang & Weigel, 1995).

dapE begins five nucleotides downstream of *dapD*, and encodes a putative protein of 237 aa that has a high degree of similarity (28.8% identity in 222 aa) to a putative protein, StbB (238 aa), from the *Salmonella typhimurium* IncN conjugative plasmid, pKM101 (Winnans 1998, GenBank Accession U43676). The function of StbB is currently unknown, however the gene is located within a region of pKM101 previously described as having a role in plasmid stability (Iyer, 1989).

dapE is followed by *dapF* that encodes a putative protein of 103 aa, and overlaps with the last five nucleotides of *dapE* and hence the expression of *dapE* and *dapF* is likely to be coupled with the transcription of *dapD*, since *dapD* is only separated from the start of the *dapE/dapF* coding regions by five nucleotides. *dapF* has 28.9% and 28.6% identity to MdeD and MbeD mobilisation proteins from *E. coli* ColicinE1 plasmid and *S. enterica* plasmid pWQ799 respectively (Table 6.1 & Figure 6.10).

6.2.2 Southern blot analysis of the *intD* element in seventeen strains of *D. nodosus*

In order to investigate the prevalence of the *intD* element in *D. nodosus*, seventeen strains of *D. nodosus* were analysed in Southern blot experiments in which three probes from the *intD* element were utilised (Figure 6.12). The strains of *D. nodosus* analysed included virulent strains A198, 1311, B1006, G1220, H1215, D1172, intermediate strain AC3577, and benign strains C3052, 819, 1169, 2483, 1493, 3138, 1469, 1311A, H1204 and AC390.

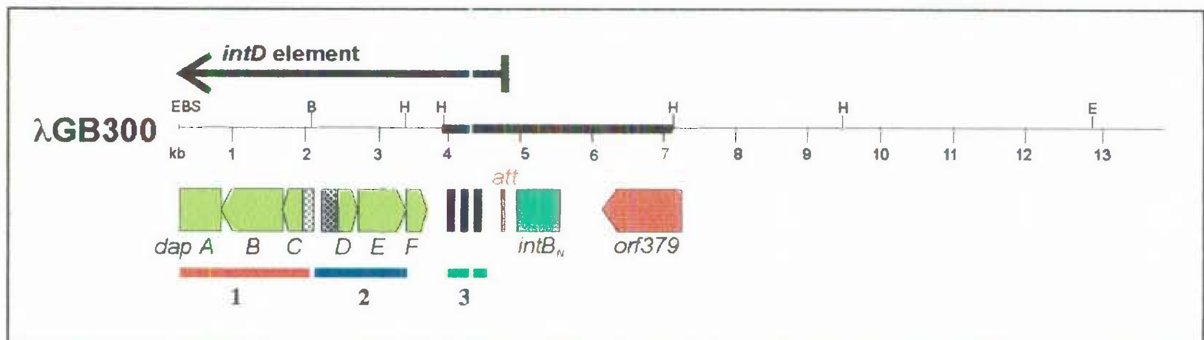

Figure 6.12: Restriction map of lambda clone λGB300 from genomic DNA library of *D. nodosus* strain C305. The DNA sequences hybridising to *intB* were on a 3.3 kb *Hin*dIII fragment which is indicated by a bold line between *Hin*dIII sites. DNA fragments from the *intD* element that were utilised as probes in Southern blot analyses are indicated by bold lines: probe 1 (red), probe 2 (blue) and probe 3 (green). Other features shown are: 102 bp repeats (black rectangles), and putative attachment site, *att* (red rectangle). Restriction fragment sites shown include *Bam*HI (B), *Eco*RI (E) and *Hin*dIII (H). The end of the lambda clone corresponds with the 5'-end of probe 1.

The results of these Southern blot experiments are shown in Figures 6.13 and 6.15, and a complete record of data obtained is shown in Appendix 6. Based on hybridisation to probes 1, 2 (Figure 6.12) and 3 from the *intD* element, strains of *D. nodosus* can be separated into three groups: Those that hybridise to probes 1, 2 and 3 (Group 1), those that hybridise to probe 1 and probe 3 (Group 2), and those that do not hybridise to *intD* element sequences at all (Group3). Of the seventeen strains analysed only strains 819 and 3138 hybridised to all three probes, whilst four strains 1311, AC3577, B1006, D1172 hybridised to probes 1 and 3, but not to probe 2. Most strains of *D. nodosus* analysed, eleven

including strain C3052, did not hybridise to the *intD* element sequences at all (Figure 6.13). These results are tabulated in Appendix 6.

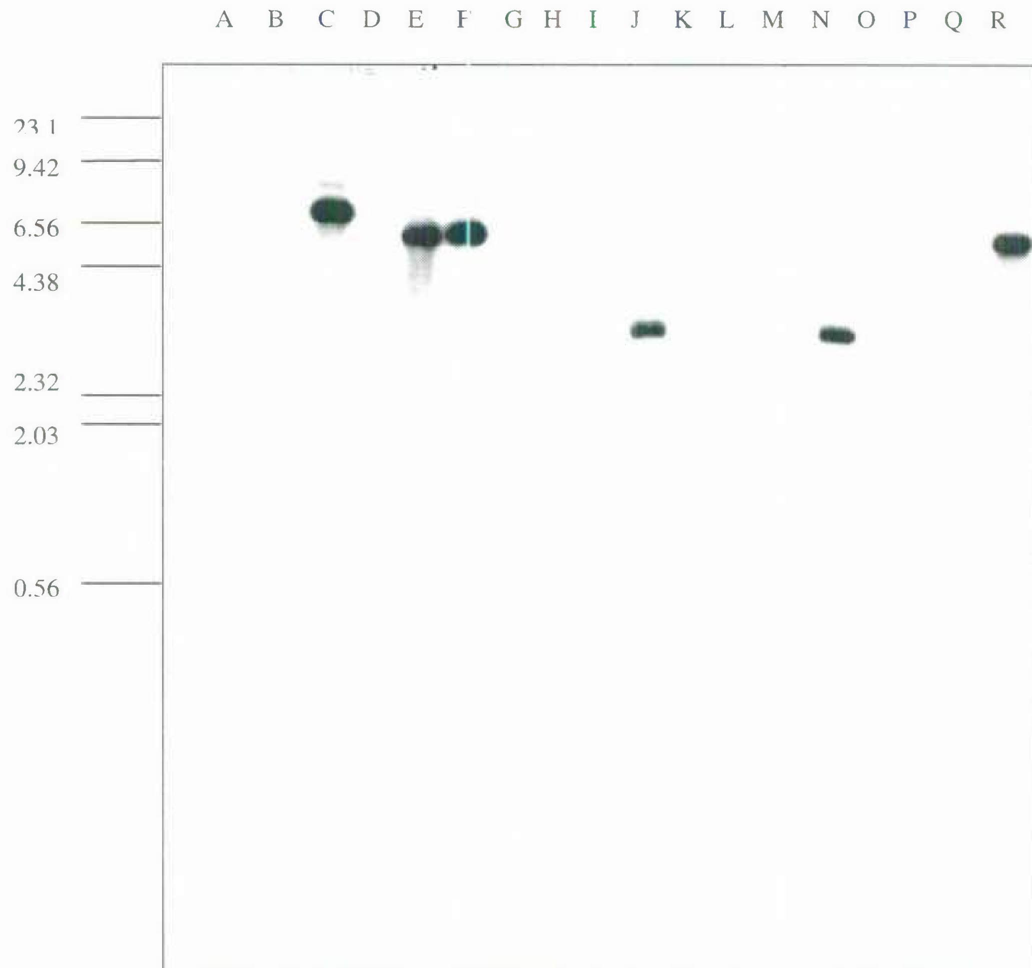

Figure 6.13: Southern blot analysis of probe 1 sequences (Figure 6.12) in *D. nodosus* strains A198 (lane A), C305 (lane B), 1311 (lane C), 1311A (lane D), AC3577 (lane E), B1006 (lane F), G1220 (lane G), H1024 (lane H), H1215 (lane I), 819 (lane J), 1169 (lane K), 2483 (lane L), 1493 (lane M), 3138 (lane N), 1469 (lane O), 1383 (lane P), AC390 (lane Q), D1172 (lane R). Genomic DNA was digested with *Hin*dIII/*Eco*RI. Lambda *Hin*dIII standard sizes in kb are indicated at the left of the panel.

6.2.2.1 *intD* element sequences were lost from strain C3051 to generate C3052

It was hypothesised (Section 6.1) that sequences originally located between the genes *pnpA* and $intB_N$ were lost or moved in strain C3051 to generate strain C3052, and that the sequences that were lost or moved included those that are adjacent to the copy of $intB_N$ isolated on λGB300 (Figure 6.14). The absence of hybridisation to probes 1, 2 or 3 by *D. nodosus* strain C3052 (Group 3) indicates that the sequences were lost rather than relocated in the C3052 genome.

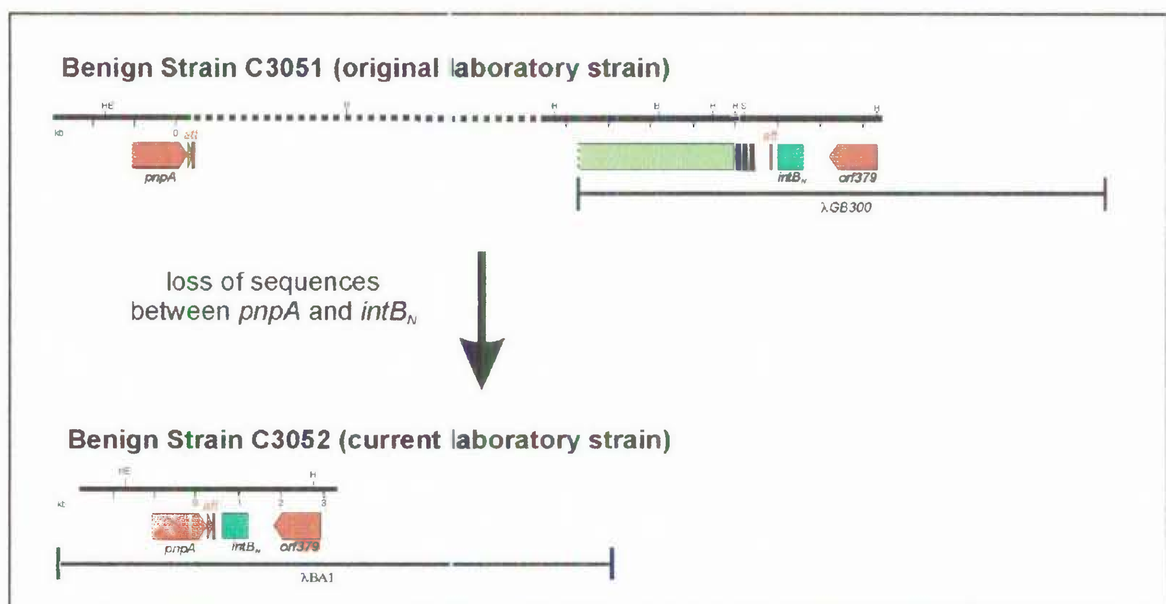

Figure 6.14: Evolution of benign *D. nodosus* strain C305. The figure shows the model for the loss of the *intD* element from adjacent to *pnpA* in the original laboratory strain C3051 to generate the current laboratory strain C3052. Restriction sites indicated include: *Bam*HI (B), *Eco*RI (E) and *Hin*dIII (H). Other features shown include: *tRNA* genes (red triangles), 102 bp repeats (black rectangles) and unknown sequences (dashed lines). The region of the *intD* element sequenced in this work is indicated by a pale green box.

Consequently, the difference in the *intB* hybridisation patterns for strain C3051 and C3052 was indicative of the loss of the *intD* element sequences from the original laboratory strain of *D. nodosus* C305, C3051. Since lambda clones containing the C3051 and C3052 $intB_N$ arrangements were isolated from the same genomic DNA library (Section 1.6.8.2), it suggests that at the time that the library was constructed a small population of C3051 cells had lost the *intD* element generating C3052 cells. We have been unable to acquire a viable stock of C3051 for analysis.

6.2.2.2 Variation in the *intD* element sequences in Group 1 and Group 2 strains

Of the six strains that hybridise to the *intD* element sequences only strains 3138 and 819 hybridise to probe 2 (Figure 6.15) suggesting that strains 1311, AC3577, B1006 and D1172 may contain a deletion in the right-hand end of the *intD* element or conversely 3138, 819 and C3051 may contain an insertion in the *intD* element. In light of sequencing data presented in Section 6.2.1, it seems most likely that there has been a deletion of sequences from strains 1311, AC3577, B1006 and D1172 since in mobilisation regions similar to the *intD* element, *dapF*-like genes are usually associated with *dapB*-like genes. In addition, the observation that sequences hybridising to all three probes were lost together from strain C3051 to generate C3052 also suggests that probe 2 sequences are part of the *intD* element.

Although strains 1311, AC3577, B1006 and D1172 do not hybridise to probe 2, the 102 bp repeats (probe 3) hybridise to the same *Hin*dIII and *Hin*dIII/*Eco*RI fragments as probe 1 in all four strains, suggesting that the 102 bp repeats and the probe 1 sequence are adjacent in these four strains of *D. nodosus* (Figure 6.15). Similarly, the 102 bp repeated sequences (probe 3) are adjacent to probe 1 and 2 sequences in strain 819, 3138 and in strain C3051 (Figure 6.15). These results suggest that the 102 bp repeats found at the *intD* element and $intB_N$ junction are part of the *intD* element, rather than merely remnants from a previous *vap* element integration event. The significance of the presence of the 102 bp repeats in both the *vap* and the *intD* elements will be discussed further in Section 6.3.

The absence of probe 2 sequences in strains 1311, AC3577, B1006, D1172 is of interest, given the characteristics of this subclone during analysis. Probe two (pGW204.1) contains a 1.2 kb *Bam*HI/*Hin*dIII fragment from λGB300 (Figure 6.12). This fragment was cloned into pUC18 and transformed into *E. coli* strain DH5-α, and was observed to be very unstable, even when subcloned afresh. Thus, only very low yields of plasmid DNA were obtained for this recombinant molecule, and consequently it proved difficult to

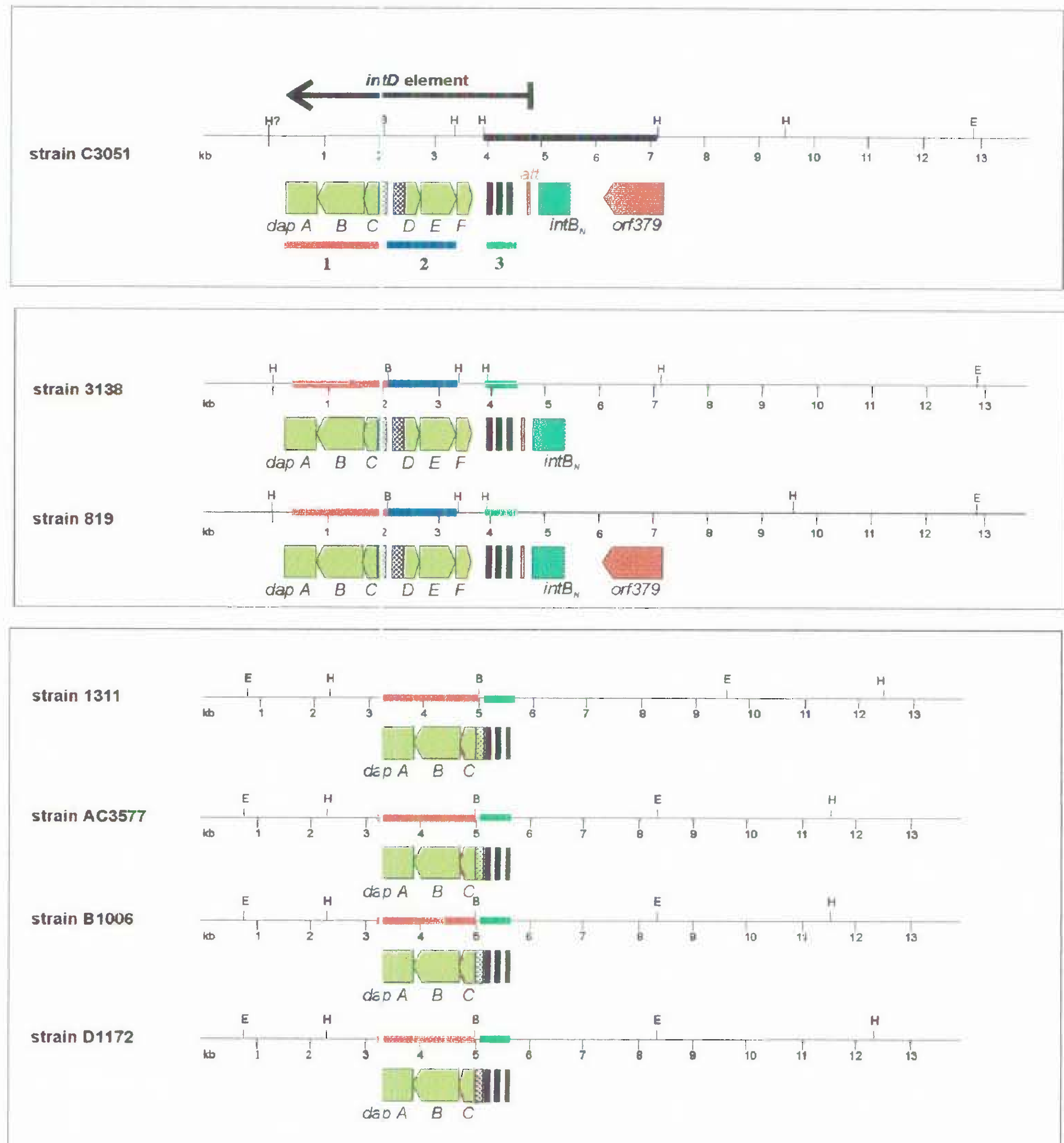

Figure 6.15: Restriction map of *intD* element sequences as determined by Southern blot analyses. A map of the *intD* element sequences from *D. nodosus* strain C3051 is included (top panel). The DNA fragments from strain C3051 utilised as probes in Southern blot analyses in other strains of *D. nodosus* are indicated: probe 1(red), probe 2 (blue) and probe 3 (green). Hybridisation to probe sequences is indicated by a bold line of the appropriate colour. *D. nodosus* strains 3138 and 819 hybridise to all three probes (middle panel), whilst virulent strains 1311, AC3577, B1006 and D1172 hybridise to probes 1 and 3, but do not hybridise to probe 2 (bottom panel). Adjacent sequences are also indicated and include: 102 bp repeats (black boxes), putative attachment sites (*att*, narrow red box) and open reading frames *intB$_N$* and *orf379* as indicated. Restriction sites shown include *Bam*HI (B), *Eco*RI (E) and *Hin*dIII (H).

analyse these sequences. In hindsight, this instability may have been caused by *cis*-acting elements encoded by this sequence, such as a putative origin of replication which may be involved in copy number control, or the numerous DnaA boxes present in this region (Section 6.2.1.2). However, it is puzzling that larger clones that included regions adjacent to the 1.2 kb *Bam*HI/*Hin*dIII fragment were not similarly unstable. The difference in stability but may be attributable to differences in topological constraints imparted by the smaller compared to a larger subclone.

The arrangement of the *intD* element in *D. nodosus* strains 3138 and 819 most resembles the arrangement of these sequences in C3051: *(i)* both 3138 and 819 hybridise to probe 1, 2 and 3 sequences, and like C3051, these sequences are found on a single *Hin*dIII fragment, which has been determined to be of 3.7 kb in strains 3138 and 819; *(ii)* like C3051, in strain 3138 the *intB* probe and a copy of the 102 bp repeats (probe 3) hybridise to a 3.3 kb *Hin*dIII fragment, suggesting therefore that in 3138 the $intB_N$ gene is present adjacent to the *attR* of the *intD* element (Figure 6.15). It is possible that these genetic elements of *D. nodosus* can only integrate at *attL*, and thus the order of the elements present in the *D. nodosus* genome reflects the order of integration. Similarly, in strain 819 a copy of *intB* and the 102 bp repeats hybridise to a 5.7 kb *Hin*dIII fragment, suggesting that the 102 bp repeats and *intB* are adjacent in this strain, as they were for C3051 and strain 3138, though the *Hin*dIII fragment size is larger (Figure 6.15).

However, *orf379* appears to missing from strain 3138, since the *orf379* probe did not hybridise to genomic DNA prepared from this strain. Hence, there is variation downstream of $intB_N$ between strains C3051, 3138 and 819. Comparison of the restriction maps of the *intD* element in strains 1311, AC3577, B1006 and D1172 (Figure 6.15) suggests that regions to the left of probe 1 are identical. However, these sequences are not integrated next to $intB_N$ or *intB* in these four strains, as in Group 1 strains (C3051, 819 and 3138).

The sequences to the left of *dapA* also differ between Group 1 (C3051, 819 and 3138) and Group 2 (1311, AC3577, B1006 and D1172) strains of *D. nodosus*. In Group 2 strains, the next *Eco*RI site is 2 kb left of *dapA*, whilst in Group 1 strains, the next *Eco*RI site to the left of *dapA* is more than 10 kb downstream (Figure 6.12). These differences might reflect the relative integrity of the *intD* element in different strains of *D. nodosus*. Alternatively, in these strains there may be a different integrated genetic element that contains a mobilisation region that is highly related to the mobilisation region of the *intD* element. It is also possible that the divergence may correspond to a restriction site polymorphism. To determine the significance of these divergent regions within the *intD* element would require direct analyses of sequences.

6.3 Discussion

6.3.1 Identification of an integrated genetic element, the *intD* element

In this work, sequences 3.7 kb to the left of $intB_N$ from the lambda clone λGB300 from a genomic library of *D. nodosus* strain C305 were determined (Figure 6.3). Analyses of these sequences have indicated that these sequences were acquired *via* the integration of a genetic element into the C305 genome. Since this genetic element is the fourth mobile genetic element (others include the *vap*, *intB* and *intC* elements) to be identified in *D. nodosus*, it has been designated the *intD* element.

Results from sequence analyses that suggest that the sequences between *pnpA* and $intB_N$ in C3051 comprise part of an integrated genetic element include the following: *(i)* the *intD* element sequences have a G + C content of only 31.8%, which is significantly less than the 45% G + C content characteristic of the *D. nodosus* chromosome (Holdeman *et*

al., 1984); *(ii)* the *intD* element contains a putative origin of replication and a putative *oriT* region, sequences that would be required for autonomous replication and mobilisation respectively; and *(iii)* the putative proteins encoded by the *dap* genes of the *intD* element have sequence similarity to mobilisation proteins (Table 6.1 and Figure 6.6) found on mobilisable plasmids and conjugative plasmids and transposons. The arrangement of the putative mobilisation genes in the *intD* element is very similar to the mobilisation regions in related plasmids and transposons (Figure 6.6).

The putative mobilisation region of the *D. nodosus intD* element has most similarity to a region from within the mobilisable transposon Tn*4399* from *Bacteroides fragilis*. This similarity is not only at the sequence level but also in arrangement and composition of the analogous regions in these two genetic elements. It has been established that derivatives of non-conjugal plasmids that carry Tn*4399* can be mobilised for transfer by broad-host-range IncPα and IncPβ plasmids such as pRK231 or R751 in *E. coli* (Murphy & Malamy, 1993). Thus, in order to determine whether the putative mobilisation region of the *intD* element is a functional mobilisation region, and analogous to the Tn*4399* mobilisation cassette of *B. fragilis*, it is proposed that the putative mobilisation region from the *intD* element of *D. nodosus* could be cloned into a ColE1 derivative (pBR322, pUC18 or p15A derivative, pACYC184). These ColE1 derivatives are nonconjugal plasmids that lack self-transfer sites and functions and are not normally mobilised by IncP plasmids such as RP4 and R751. Hence, an *E. coli* donor strain, which contains an IncP conjugative plasmid and experimental derivatives carrying the *D. nodosus* mobilisation region could be mated with a suitable recipient strain of *E. coli* using the method described previously by Murphy *et al.* (Murphy & Malamy, 1993), and transconjugants subsequently identified.

Southern blot analyses also support the hypothesis that the *intD* element is a mobile genetic element. Southern blot analyses have shown (Section 6.2.2) that the *intD* element

sequences that were present in the original laboratory strain of *D. nodosus* strain C3051, are no longer present in current laboratory strain C3052. This supports the hypothesis that sequences to the left of $intB_N$ in C3051 cells were lost to generate C3052 (Figure 6.14). There is other evidence to suggest that loss of the *intD* element is also occuring in other strains of *D. nodosus* (Section 7.2.4), and suggests that at least under laboratory conditions the *intD* element is unstable in *D. nodosus*. Consequently, the observed loss of the *intD* element from strain C3051 to generate C3052 is not a novel event.

6.3.2 Variation of the *intD* element sequences between *D. nodosus* strains

Southern blot analyses revealed that of the seventeen strains of *D. nodosus* analysed, only six strains, including 819, 3138, AC3577, 1311, B1006 and D1172 contained sequences corresponding to the right hand end of the *intD* element. Furthermore, the latter four strains do not hybridise to the 1.2 kb *Bam*HI/*Hin*dIII fragment (probe 2, Figure 6.12) that contains the sequences corresponding to the *oriV*, *dapD* and *dapE* genes.

Exactly how much of the 1.2 kb *Bam*HI/*Hin*dIII fragment is absent from strain 1311, AC3577, B1006 and D1172 is unclear, because although probe 2 did not hybridise to the genomic DNA from these strains, the *Bam*HI site is still present in these divergent strains (Figure 6.15). Since the *Bam*HI site is still present, some of the sequences to the right of this restriction site might also be present, though based on the lack of hybridisation to the 1.2 kb probe, this would be less than 100 bp.

In order to define the regions between probe 1 and probe 3 (Figure 6.12), the nucleotide sequence of this region in the divergent strains would need to be determined. The absence of the probe 2 sequences is likely to affect the stability of the *intD* element since *dapE*, which may encode stability functions must be non-functional since no

hybridisation to probe 2 was observed. Similarly, one cannot be certain that *dapF* is present strains in 1311, AC3577, B1006 and D1172 since the 500 bp *Hin*dIII fragment containing *dapF* was not used as a probe, and so further Southern blot analyses would be required to determine this. Although, *dapD*, *dapE* and the putative *oriV* are missing from some strains that contain other *intD* element sequences, there is no evidence to suggest that *dapD*, *dapE* and *oriV* sequences are independently mobile.

Non-self-transmissible, mobilisable elements are equipped to exploit the conjugative apparatus of self-transmissible elements by encoding specific *cis*- and *trans*-acting elements that mediate these interactions. The first step in conjugative transfer of DNA involves the introduction of a single-strand nick by *mob* encoded proteins at a site called the *oriT* or *nic* site. When the conjugative element has initiated mating, the nicked strand forms a complex with *mob*-encoded proteins which lead the DNA into the recipient cell (Willetts & Wilkins, 1984).

If the enzyme involved in nicking of the *oriT* region of the *intD* element nicked a linear intermediate rather than a circular intermediate, this could potentially result in loss of sequences to the right of the *oriT* (Figure 6.5), including the *oriV*, *dapD*, *dapE*, *dapF* and the 102 bp repeats. However, Southern blot analysis has shown that although strains 1311, AC3577, B1006 and D1172 do not contain *oriV*, *dapD* and *dapE*, the 102 bp repeats are still present at the right of the *oriT* region in these four strains (Figure 6.15). It is possible that the loss of the *oriV*, *dapD* and *dapE* in strains 1311, AC3577, B1006 and D1172 may have involved more than one step.

In the *vap* element, the region corresponding to the putative *oriT* is immediately followed by the 3'-end of a 102 bp repeat (Figure 6.10). The 102 bp repeat begins the first nucleotide after the alignment between the *oriT* regions of the *vap* element and the *intD* element ceases (Figure 6.10). One possibility is that the 102 bp repeats are recognised and

cleaved by a separate nicking enzyme. Cleavage of a linear or circular molecule at the *oriT* by a *mob*-encoded nickase and of sequences immediately to the left of 102 bp repeats by another protein, perhaps encoded by the *vap* genes, would result in the loss of intervening sequences, as seems to have occurred in the *intD* elements of *D. nodosus* strains 1311, AC3577, B1006 and D1172.

This speculation is supported by the observation that the four strains that have lost these sequences (1311, AC3577, B1006, D1172) contain *vap* regions, whilst the other two strains (819 and 3138) that do not contain *vap* genes, still contain the sequence that separates the putative *oriT* from the 102 bp repeat. The significance of these 102 bp repeats in the *vap* element has not yet been established, though it has been proposed previously that they may have a role in rearrangements of the *vap* genes that they bracket (Cheetham *et al.*, 1995a). Given the variations in the arrangement of the *vap* genes in different strains of *D. nodosus* (Bloomfield *et al.*, 1997), it would not be surprising if there is an enzyme encoded by the *vap* element, that mediates such rearrangements.

6.3.3 The *intD* element is not part of the *intB* or *intC* elements

The *intD* element sequences characterised in this work comprise the right-hand end of an element. Since the right ends of the *intC* element and the *intB* element of *D. nodosus* have not yet been isolated, it was necessary to eliminate the possibility that the region of the *intD* element sequenced was part of either of these genetic elements. In order to definitively show this, chromosome walking from the right end of the *intD* element characterised in this work to the *pnpA* gene at the left end of the *intD* element in a C305 genomic library was carried out by B. Shaw. This chromosome walking was recently completed (Shaw and Cheetham, unpublished). Preliminary analyses suggest that the *intD* element is approximately 14 kb in size and sequencing of the region at the left-end of this element which is adjacent to *pnpA* has led to the identification of another integrase gene,

intD, that is different although related to the integrase genes present in the *vap*, *intB* and *intC* elements of *D. nodosus*. This confirms that the loss of the *intD* element from strain C3051 generated strain C3052, and shows that the *intC* element and the *intD* element are different integrated elements.

Southern blot analyses (Section 6.2.2) suggested that the arrangement of the *intD* element and flanking sequences was similar in C3051 and Group 1 strains (C3051, 819, 3138). In Group 1 (C3051, 819, 3138) strains, the *dap* genes are found on a very large *Eco*RI fragment (>23.1 kb), whilst in the Group 2 strains (1311, AC3577, B1006 and D1172), the *Eco*RI fragment is much shorter (8.3 kb -10.1 kb). Thus, there is divergence to the left of the *dap* genes between Group 1 and Group 2 strains. Since this region definitely contains the *intD* element in Group 1 strains, one possible explanation is that the sequences at the left-hand end in Group 2 strains correspond to a different genetic element, which contains a highly related mobilisation region (the *dap* genes). There is a precedent for different related-elements containing highly homologous mobilisation regions, and the colicin plasmids are one such example (Boyd, *et al.*, 1989). Further analyses of sequences to the left of the regions of divergence between Group 1 (819, 3138) and Group 2 strains (1311, AC3577, B1006, D1172) should be undertaken in order to determine whether these sequences in Group 2 strains correspond to a different integrated element, or to the rearrangement of sequences within the same *intD* element.

Further Southern blot analyses using *intD* as a probe should also be done in order to determine whether the left end of the *intD* element is also present in Group 2 strains (1311, AC3577, B1006, D1172) of *D. nodosus*, and may assist in the identification of the integration sites for the *intD* element in these strains. Inverse PCR could also be utilised to isolate regions flanking the left-end of the *intD* element in these strains of *D. nodosus*.

Preliminary sequence data from the left end of the *intD* element (Shaw &

Cheetham, unpublished) has led to the identification of putative *tra* genes that show similarity to those from plasmid RP4 (also R18, R68, RP1, RK2) and from the *Agrobacterium tumefaciens* Ti plasmid, which are required for the conjugative transfer processes. Also, the putative protein encoded by *dapE* from the *intD* element also has very high similarity (Table 6.1, Figure 6.6) to a protein encoded by *stbB* from pKM101, which may have a role in plasmid stability, though the precise role is currently unknown (Iyer, 1989; Winans 1998, GenBank Accession U43676). Collectively results suggest that the *intD* element may be a conjugative element rather than a non-self-transmissible but mobilisable element.

6.3.4 Similarities between the *intD* element and the *vap* element

The principal reason for investigating the sequences isolated on the lambda clone λGB300 at the left of $intB_N$ was that three copies of 102 bp repeated sequences that are found scattered throughout the *vap* regions of *D. nodosus* strain A198 (Figure 6.1) were immediately adjacent to this copy of $intB_N$. This suggested that a remnant of the *vap* region may have separated *pnpA* and $intB_N$ in the original laboratory strain of C305, called C3051. Analyses of these sequences do not indicate that these sequences comprise part of a *vap* element, but instead demonstrate that these sequences comprise part of a different integrated genetic element of *D. nodosus*.

This new genetic element, the *intD* element, has other features in common with the *vap* regions of *D. nodosus*. The *intD* element integrates into the same position of the chromosome as the *vap* element and integration is likely to be catalysed by the *intD* gene, which encodes an integrase protein which is most similar to the integrases encoded by the *vap*, *intB* and *intC* elements of *D. nodosus* (Shaw & Cheetham, unpublished). Although no genes that are present in other *D. nodosus* elements have so far been identified in the *intD* element, a non-coding sequence of 118 nt (*oriT*) that is present in both the *intD* element

and in the *vap* element was identified 1.7 kb upstream of the 102 bp repeats.

It has previously been established that the *vap* regions of *D. nodosus* have been acquired *via* independent integrations of a genetic element into the *D. nodosus* genome. However, no genes or cis-acting elements that might have been involved in such an acquisition process have been identified within the *vap* regions. Most of the *vap* genes, are instead similar to genes involved in plasmid maintenance functions (Section 1.6.4). The evidence presented here suggests that similarly the *intD* element has also been acquired horizontally.

It may be possible to demonstrate that these mobilisation regions and thus the *oriT* is functional by doing mating experiments in *E. coli*, since mobilisable elements that are related to the *intD* element have been mobilised in *E. coli* by plasmids belonging to the *E. coli* IncP group. Those elements that are related to the *intD* element mobilisation region and have been demonstrated to be mobilised by IncP plasmids include *E. coli* ColE1 and pColD-157 (Guiney & Lanka, 1989), Tn*4399* from *B. fragilis* (Murphy & Malamy, 1993), pWQ799 from *S. enterica* (Keenleyside & Whitfield, 1995), and pEC3 from *E. carotovora* (Nomura *et al.*, 1996).

In addition, since mobilisation proteins are involved in DNA processing during conjugation and, particularly, the nicking of the *oriT* region prior to transfer, then provided an identical *oriT* region was present on a *mob*$^-$ molecule such as the *vap* element, a *mob*$^+$ element like the *intD* element could potentially mobilise a *mob*$^-$ element in *trans*, provided a conjugative element was also present. The presence of an identical putative *oriT* region in the *vap* element and the *intD* element also supports the hypothesis that these *D. nodosus* elements may interact, and suggests that the *intD* element may have been involved in the acquisition and dissemination of the *vap* element in *D. nodosus*. This hypothesis could be investigated in mating experiments in *E. coli*, in which *E. coli* host cells containing *intD*

element sequences could mobilise a plasmid carrying the putative *oriT* in trans.

Conjugation-dependent, *recA*-independent *oriT*-specific recombination has been observed previously in ColE1 and related plasmids (Broome-Smith, 1980; Warren & Clark, 1980), the F plasmid (Gao, Luo & Deonier, 1994), and IncQ plasmid R1162 (Brasch & Meyer, 1986), and reflects the dual endonuclease-ligase activities of the protein that specifically nicks the *oriT* DNA during conjugation. More recently, conjugation-independent *oriT*-specific recombination of IncW plasmid R388 was reported (Llosa, Bolland & de la Cruz, 1994). If the putative *oriT* region of the *intD* element does comprise the nick site for transfer in both elements, then the same *mob* encoded endonuclease-ligase protein from the *intD* element could potentially recognise both sequences and thus catalyse intermolecular recombination between the two elements.

Such a cooperative interaction between different genetic elements in *D. nodosus* may explain why copies of the 102 bp repeats that are found scattered throughout the *vap* regions (Figure 6.1) should also be found within the *intD* element. The copies of the 102 bp repeats that are present in the *intD* element are most similar to the 102 bp repeat found immediately upstream of *vapD* in *vap* region 1 of *D. nodosus* strain A198 (Figure 6.10). Perhaps the presence of the 102 bp repeats in both elements is indicative of a link between the acquisition of the *vap* element and the *intD* element.

It is also interesting that the *intD* element is present in *D. nodosus* strain 1311, which carries the native plasmid pDN1 (Section 3.2.1). pDN1 is highly related to plasmids belonging to *E. coli* incompatibility group Q (Section 3.2.4), which are also mobilised by conjugative plasmids belonging to the *E. coli* IncP group in addition to those from the Incα, IncM and IncX groups (Willetts & Crowther, 1981). Preliminary evidence does suggest that the *intD* element may be a conjugative element, and thus may be responsible for the acquisition of pDN1 in *D. nodosus*. One could investigate the ability of *intD*

element sequences to mobilise antibiotic resistant derivatives of pDN1 in *E. coli*. If this were successful, one could begin *E. coli-D. nodosus* mating experiments, using a strain of *E. coli* carrying a conditional-lethal mutation as a donor. This may allow the introduction of DNA into *D. nodosus*, which is currently not possible (Chapter 3).

6.3.5 What sort of genetic element is the *intD* element?

The *intD* element does contain features common to other genetic elements including bacteriophages, plasmids and conjugative transposons. Such features include: *(i)* the ability to integrate and excise site-specifically from the host genome; *(ii)* presence of an integrase gene related to the phage lambda family of site-specific recombinases; and *(iii)* the observed mobility of these sequences in *D. nodosus*.

However, the presence of genes and *cis*-elements required for conjugative transfer and mobilisation, suggest that the *intD* element is mobilised from one cell to another *via* conjugative transfer rather than by transduction, eliminating the possibility that the *intD* element is an integrated bacteriophage, defective or otherwise. It is less clear whether the *intD* element is a conjugative or mobilisable transposon or an integrated conjugative plasmid.

During transfer, conjugative transposons are thought to be transferred *via* a circular intermediate that does not undergo autonomous replication in hosts so far investigated (Bedzyk *et al.*, 1992; Li *et al.*, 1995; Rice & Carrias, 1994; Salyers *et al.*, 1995; Scott, Kirchman & Carparon, 1988), and thus would not require an *oriV*. However, a putative *oriV* region is present in the *intD* element of Group 1 strains, suggesting that this element does at some stage undergo autonomous replication. However, the absence of the putative *oriV* in Group 2 strains suggests that the *oriV* may not be necessary for the *intD* element to be transferred from one cell to another, such as if conjugation were mediated by a

conservative mechanism, though there may be some other, as yet unidentified, origin within the *intD* element sequence. Alternatively, the absence of such sequences might also suggest that the *mob* genes present in Group 2 strains are part of a different, but related genetic element.

Southern blot evidence suggests that like conjugative transposons, the *intD* element does seem to be present as an integrated element, rather than extrachromosomally as a plasmid. However, the *intD* element might also be an integrated conjugative plasmid, which does undergo autonomous replication prior to transfer, if an alternate *oriV* is present in Group 2 strains.

Conjugative transposons such as Tn*916* and Tn*1545* from Gram positive bacteria, encode integrase genes that are related to phage lambda integrase genes (Clewell, Flannagan & Jaworski, 1995; Poyart-Salmeron *et al.*, 1990). Although they share the aa sequence signature near the active site that is characteristic of the phage lambda integrase family, they have little similarity to the lambda Int protein outside this region (Poyart-Salmeron *et al.*, 1990). In contrast, the integrase genes found in the *intD* element (Shaw & Cheetham, unpublished) and in the *vap*, *intB* and *intC* elements of *D. nodosus* are similar to the bacteriophage integrases over the entire coding region (Bloomfield, 1997; Bloomfield *et al.*, 1997; Cheetham *et al.*, 1995b).

The arrangement of the *intD* element is very similar to that of conjugative transposon Tn*916* (Clewell *et al.*, 1995; Flannagan *et al.*, 1994; Salyers *et al.*, 1995; Senghas *et al.*, 1988; Su & Clewell, 1993) in that: *(i)* an *int* gene is located at the left end of both elements and in the same orientation; *(ii)* both contain a central region encoding *tra* functions; *(iii)* both contain a mobilisation protein gene called *mbeA* or *dapB* in Tn*916* and the *intD* element respectively, which is similar to the ColE1 *mbeA*. Like *dapB*, *mbeA* is located at the far end of the element; and *(iv)* Tn*916* is 18 kb in size, which is comparable

to the size of the *intD* element which is 14 kb.

However, the integration of Tn*916* is relatively random, though preferentially integrating in A + T rich regions, and does not involve the duplication of its target site (Scott *et al.*, 1994; Trieu-Cuot *et al.*, 1993). In contrast, the *intD* element and other *D. nodosus* elements integrate site-specifically in the 3'-end of *tRNA-ser* genes and duplicate the target site at the right-hand end of the integrated element (Bloomfield, 1997; Bloomfield *et al.*, 1997; Cheetham *et al.*, 1995b); integration into a *tRNA* gene with duplication of the target site is characteristic of many bacteriophages and plasmids (Campbell, 1992; Cheetham & Katz, 1995).

However, there are other very large conjugative transposons including Tn*5252* and Tn*5276* that integrate more site-specifically, into a primary target site that is duplicated at each end of the element (Kilic, Vijayakumar & Al-khaldi, 1994; Rauch & de Vos, 1994; Vijayakumar & Ayalew, 1993). In addition, Tc^{r}ERL-like conjugative transposons integrate with some site specificity, having three to seven integration sites (Bedzyk *et al.*, 1992). However, based on the cryptic Tc^{r}ERL-like element XBU4422, no target site duplication is believed to occur during insertion (Bedzyk *et al.*, 1992), although one of the ends of this element is homologous to its target sequence.

The *intD* element might also be an integrated conjugative plasmid, which does undergo autonomous replication prior to transfer, and thus have an *oriV*. However, in at least the four group 2 strains, there does not appear to be an *oriV*.

The *intD* element does seem small (14 kb) given that it needs to encode conjugative functions. For example, the F plasmid of *E. coli* has at least 20 genes that extend over 33 kb of DNA, and are necessary for conjugative processes (Ippen-Ihler & Minkley, 1986). However, the conjugative transposon Tn*916* of *Enterococcus faecalis* is only 18.0 kb (Flannagan *et al.*, 1994) and able to mobilise plasmids in *trans* (Clewell *et al.*, 1995). It

has been proposed that Tn*916* may encode a conjugal transfer system that is simpler than that utilised by plasmids such as F, and may not be mediated by a sex pilus (Salyers *et al.*, 1995), since the genes within the *tra* region of Tn*916* do not have sequence similarity to transfer regions of conjugal plasmids (Flannagan *et al.*, 1994). Such mating systems have been observed previously in some strains of the Gram-positive bacterium *E. faecalis* that harbours pheromone-inducible plasmids such as pAD1 and pCF10. Pheromones induce the mating response by inducing plasmid-encoded genes to produce surface antigens that mediate cell-cell aggregation and formation of a mating pair, rather than sex pilus mediated transfer (Dunny, 1990; Galli, Wirth & Wanner, 1989).

It is therefore significant that preliminary sequence data (Shaw & Cheetham, unpublished) indicates that the *intD* element may contain genes analogous to those from pilus-mediated conjugal transfer systems. It is presently unknown whether the *intD* element contains all of the genetic determinants required to initiate conjugative transfer, and this emphasises the importance of the proposed mating experiments in *E. coli* (above).

In this work, the 3.7 kb region from the region between *pnpA* and $intB_N$ from the original laboratory strain of *D. nodosus* C305 (C3051) were determined. Sequence analyses have indicated that these sequences do not contain remnants of the *vap* region but instead are part of a new genetic element designated the *intD* element that was acquired horizontally *via* the integration of a mobilisable and possibly conjugative plasmid or transposon into the *D. nodosus* genome. Further work is required to demonstrate the functions of these genes, which is likely to be possible in *E. coli*. Such investigations are important because they may allow the development of a conjugation-mediated DNA transfer system in *D. nodosus*, and may tell us more about the acquisition and dissemination of the genetic elements in *D. nodosus* and other pathogenic bacteria in general.