Genomic analysis of the PAI ICl3 locus in pathogenic LEE-negative Shiga toxin-producing Escherichia coli and Citrobacter rodentium

Jean Pierre Girardeau, Yolande Bertin and Christine Martin

INRA, UR454 Unité de Microbiologie, F-63122 Saint-Genès Champanelle, France

Shiga toxin-producing Escherichia coli (STEC) causes a spectrum of human illnesses such as haemorrhagic colitis and haemolytic–uraemic syndrome. Although the locus of enterocyte effacement (LEE) seems to confer enhanced virulence, LEE-negative STEC strains are also associated with severe human disease, suggesting that other unknown factors enhance the virulence potential of STEC strains. A novel hybrid pathogenicity island, termed PAI ICl3, has been previously characterized in the LEE-negative O113:H21 STEC strain CL3. Screening for the presence of PAI ICl3 elements in 469 strains of E. coli, including attaching and effacing (A/E) pathogens [enteropathogenic E. coli (EPEC) and enterohemorrhagic E. coli (EHEC)], non-A/E pathogens [LEE-negative STEC, extra-intestinal pathogenic E. coli (ExPEC), enterotoxigenic E. coli (ETEC) and enteropathogenic E. coli (EAEC)] and commensal E. coli isolates, showed that PAI ICl3 is unique to LEE-negative STEC strains linked to disease, providing a new marker for these strains. We also showed that a PAI ICl3-equivalent gene cluster is present in the genome of Citrobacter rodentium, on a 53 kb genomic island inserted into the pheV IRNA locus. While the C. rodentium PAI ICl3 shows high similarities at the nucleotide level and in organization with the E. coli PAI ICl3, the genetic context of the integration differs completely. In addition, BLAST searches revealed that other E. coli pathotypes (O157 : H7 EHEC, ExPEC, EPEC and EAEC) possess incomplete PAI ICl3 elements that contain only the genes located at the extremities of the island. Six of the 16 sequenced E. coli genomes showed deleted PAI ICl3 gene clusters which are carried on mobile genetic elements inserted into pheV, selC or serW IRNA loci, which is compatible with the idea that the PAI ICl3 gene cluster entered E. coli and C. rodentium at multiple times through independent events. The phylogenetic distribution of the PAI ICl3 variants suggests that a B1 genetic background is necessary for the maintenance of the full complement of PAI ICl3 genes in E. coli.

INTRODUCTION

The composition of bacterial genomes can change rapidly and dramatically through a variety of processes, including horizontal gene transfer. Analysis of bacterial genomic sequences has indicated that many of the horizontal gene transfer events observed in bacteria involve clusters of genes referred to as genomic islands (GEIs) (Hacker & Kaper, 2000). Typically, GEIs are blocks of DNA with signatures of mobile genetic elements, i.e. they possess genes encoding factors that are involved in genetic mobility, such as integrases, transposases, phage genes and origins of replication, are frequently flanked on one side by a TRNA gene and by direct repeat sequences, and are often associated with a genetic instability (Hacker & Kaper, 2000; Hacker & Carniel, 2001). GEIs have a modular organization and encode various functions related to virulence, symbiosis, metabolism, resistance to antibiotics and degradation of xenobiotic compounds (Hacker & Carniel, 2001). Pathogenicity islands (PAIs) represent a subclass of GEIs that contain similar features to GEIs but encode various virulence factors and are normally absent from non-pathogenic strains of the same or closely related species (Hacker & Kaper, 2000; Gal-Mor & Finlay, 2006; Schmidt & Hensel, 2004).
Enteropathogenic *Escherichia coli* (EPEC), enterohaemorrhagic *E. coli* (EHEC) and the mouse enteropathogen *Citrobacter rodentium* belong to the family of attaching and effacing (A/E) bacterial pathogens. A/E lesions are mediated by components of a type III secretion apparatus encoded by a PAI called the locus of enteroctye effacement (LEE) (Deng et al., 2004; Frankel et al., 1998; Garmondia et al., 2005). EHEC and EPEC are poorly pathogenic in mice but infect humans and domestic animals. In contrast, *C. rodentium* is a natural mouse pathogen that is related to *E. coli*, hence providing an *in vivo* model for A/E pathogens (Mundy et al., 2005). While *C. rodentium* infection of mice results in colonic hyperplasia, EHEC O157 : H7 strains and other Shiga toxin-producing *E. coli* (STEC) cause a spectrum of human illnesses such as watery diarrhoea, haemorrhagic colitis (HC), haemolytic–uraemic syndrome (HUS) and thrombotic thrombocytopenic purpura. HUS in EHEC infection is the leading cause of acute renal failure in children, and is mainly caused by the production of Shiga toxins (Beutin, 2006; Clarke et al., 2003; Paton & Paton, 1998).

Although the LEE seems to confer enhanced virulence, LEE-negative STEC strains are also associated with severe human disease (Girardeau et al., 2005; Johnson et al., 1996; WHO, 1999). These observations suggest that other unknown factors, possibly GEIs or PAIs, enhance the virulence potential of STEC strains (Boerin et al., 1999; Kaper et al., 1999; Karmali et al., 2003). Compared with the *E. coli* K-12 genome, O157 : H7 strains EDL933 and Sakai contain several additional GEIs (O islands) including the 87 kb O island 48 (OI-48) and the 23 kb OI-122 (Perna et al., 2001). To investigate the structural diversity of GEI OI-122 between the O157 : H7 *E. coli* EDL933 and the LEE-negative O113 : H21 *E. coli* CL3, a region of 27 kb was sequenced by Shen et al. (2004). These authors describe, in *E. coli* CL3 and in other STEC strains, a novel hybrid genomic region composed of three physically distinct portions: a 13 kb segment (left part), which carries a *Yersinia pestis*-like haemolysin/adhesin gene cluster predicted to encode members of the ShlA/HecA/FhaA exoprotein family (secreted by the two-partner secretion pathway); a central 4.4 kb segment bracketed by two 190 bp direct repeat sequences, which carries five ORFs (including transposase genes); and a composite 10 kb segment (right part), which carries three EDL933 OI-122 genes (including the virulence gene Z4321; pagC-like), a 2.2 kb fragment of the enteroaggregative *E. coli* (EAEC) O42 and a 5.3 kb segment of the EDL933 OI-48. Its left and right termini (Z1635 and Z1644, respectively) contain ORFs that show homology to EDL933 OI-48 genes. Given the presence of putative virulence genes (a haemolysin/adhesin gene cluster and pagC) and mobile genetic elements, this GEI was considered to be a PAI. Because this was the first PAI found in the CL3 genome, it was designated PAI ICL3 (Shen et al., 2004).

A large variety of STEC serotypes have been implicated in disease. However, certain STEC serotypes recovered from animals and food have never been associated with serious human disease. For a better understanding of the apparent differences in virulence between groups of STEC serotypes, STEC strains have been classified into five seropathotypes (A–E) by Karmali et al. (2003), according to their incidence and association with HUS and outbreaks. Recent studies have demonstrated that determination of the seropathotype distribution of virulence factors allows identification of DNA targets for selective detection of strains that present a risk to public health (Gilmour et al., 2006; Karmali et al., 2003). As we have shown, there is a link between seropathotype, prevalence of various virulence factors, phylogeny and Shiga toxin gene expression (de Sablet et al., 2008; Girardeau et al., 2005).

The initial objective of the present study was to determine whether the PAI ICL3 locus is a suitable DNA target for the selective detection of LEE-negative STEC strains belonging to seropathotype C that represent a significant risk of disease in humans. With this aim, the presence of PAI ICL3 elements among LEE-negative STEC strains and other *E. coli* pathotypes as well as commensal *E. coli* isolates was investigated in a collection of 469 *E. coli* strains. We show the widespread dissemination of the PAI ICL3 element among LEE-negative STEC strains, and its absence from other *E. coli* pathotypes. In addition, we discover that the PAI ICL3 is also present in *C. rodentium*. We further report that the *C. rodentium* PAI ICL3 is borne on a 53 kb GEI termed in this study GPlpheV-CRSCC168.

**METHODS**

**Bacterial strains.** A total of 469 *E. coli* strains, including clinical isolates, reference strains and intestinal commensal strains, were used in this study. Clinical *E. coli* strains (449 isolates) were provided from five different sets. The first set consisted of 233 STEC isolates collected in France from human disease, bovine faeces, food, dairy herd manure, wastewater treatment plants and pig farm manure samples (Bonnet et al., 1998; Pradel et al., 2000; Vernozy-Roand et al., 2002). The second set, which comprised 54 STEC strains of diverse geographical origins isolated from human disease, was provided by the STEC Center, National Food Safety and Toxicology Center (NFSTC), Michigan State University, USA. Of these 287 STEC isolates, 44 are LEE-positive and 243 are LEE-negative. The third set comprised 77 extra-intestinal pathogenic *E. coli* (ExPEC) isolated from humans with bacteraemia or urinary tract infections and from calves with intestinal or extraintestinal colibacillosis (Girardeau et al., 2003). The fourth set comprised 35 ExPEC strains isolated in Scotland from lambs with nephropathy (Bertin et al., 1998). The last set consisted of 50 enterotoxigenic *E. coli* (ETEC) isolates collected from calves with intestinal colibacillosis (Contrepois et al., 1993). In addition, seven *E. coli* intestinal commensal strains were isolated from healthy human and animal subjects. Thirteen reference strains of different pathotypes, including the neonatal meningitis *E. coli* (NMEC) strain RS218, adherent- invasive *E. coli* (AIEC) strain LF82, EAEC strain O42, ETec strain H10407, human EPEC strains E2348/69, E56/54 and B171-8, rabbit EPEC strains RDEC-1 and E22, ExPEC strains 536 and CFT073, and HEC O157 : H7 strains EDL933 and Sakai, were kindly provided by José Harel (GREMIP, Faculte de Médecine Vétérinaire, Université de Montréal, Canada), Arlette Darfeuille Michaud (Pathogénie Bactérienne Intestinale, Université d’Auvergne, Clermont-Ferrand, France), Chantal Le Bougnec
Table 1. PCR primers used in this study

<table>
<thead>
<tr>
<th>Probe</th>
<th>Target DNA</th>
<th>Primer sequences</th>
<th>Size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ms-1</td>
<td>Z1640::S1 junction (left junction with OI-48)</td>
<td>GCTGATGCGTTACCACACTG, GTAATCCTCAAACCGCACCAG</td>
<td>684</td>
</tr>
<tr>
<td>ms-2</td>
<td>S4 (shiA/secA/phaA-like gene)</td>
<td>CTCGAAATGACAGGTGAAAACAG, TGTATGTCTCGGCTTTCCAG</td>
<td>498</td>
</tr>
<tr>
<td>ms-3</td>
<td>S10/S11 (pagC) junction</td>
<td>GCCTGATGACACCTGATCACG, GACAACACTGACGCAATAC</td>
<td>880</td>
</tr>
<tr>
<td>ms-4</td>
<td>Z4318::S12 junction (left junction with OI-122)</td>
<td>CTTGAATGACGGTGAACAG, GTCGGGCGTTACCTTATAC</td>
<td>324</td>
</tr>
<tr>
<td>ms-5</td>
<td>S14::Z1642 junction (right junction with OI-48)</td>
<td>CTCGGGAGCTACAGGAATATC, GTTGCCGTATTCTCAGATAC</td>
<td>640</td>
</tr>
</tbody>
</table>

Fig. 1. Genetic organization of the GEI and flanking sequences in the E. coli CL3 PAI ICL3 in comparison to the C. rodentium PAI ICL3. ORFs with homology to genes of the EDL933 GEI OI-122 and OI-48 are indicated by black and dark-grey shading, respectively. Y. pestis shiA/secA/phaA-like genes are indicated by light-grey shading. Vertically striped ORFs indicate homologues to the EAEC O42 sequence. The solid black line indicates the central variable region. ORFs within a region without homology at the nucleotide level are white. ORFs of mobile genetic elements (IS element and phage sequence) are indicated by hatched arrows. Black boxes indicate the position of the direct repeats (H8DR). Note the premature stop codon (*) at position 623 of the shiA/secA/phaA-like gene (creating the pseudogene S3 in the prototypic E. coli PAI ICL3). The identities or characteristics of ORFs are detailed in Supplementary Table S1 (bold type). E. coli PAI ICL3 ORFs are based on published sequences (Shen et al., 2004).

Statistical analyses. Statistical analyses were performed with SAS for Unix Windows (version 8.01, SAS Institute). Comparison of the prevalence of a particular characteristic in different populations was evaluated with the chi-squared test and odds ratios (ORs), and 95% confidence intervals (CI) were determined.

RESULTS

Distribution of the PAI ICL3 elements among A/E pathogens and other E. coli pathotypes

The presence of the PAI ICL3 elements among A/E pathogens (including 44 EHEC, 5 EPEC and C. rodentium) and non-A/E pathogens (including 243 LEE-negative STEC strains, 115 ExPEC and 51 ETEC strains, and representative strains of NMEC, AIEC and EAEC) was investigated by dot-blot hybridization. Genomic DNA was hybridized with five probes specific for the left (ms-1 and ms-2), internal (ms-3 and ms-4) and right (ms-5) regions of the PAI ICL3 (Fig. 1). The five PAI ICL3 marker sequences were found in C. rodentium and among 102 of the 243 (42%) LEE-negative STEC strains belonging to 29 serotypes. PCR linkage analysis (not shown) carried out on 18 STEC strains of different serotypes demonstrated that the PAI ICL3 element carried by these strains has a structure similar to that of the prototypic PAI ICL3. Interestingly, the PAI ICL3 appears to be unique to C. rodentium and LEE-negative STEC strains, since it was not detected in intestinal commensal isolates nor in any of the strains belonging to the other E. coli pathotypes.

Seropathotype distribution of the PAI ICL3

Consistent with an earlier study (Shen et al., 2004), PAI ICL3-related elements were not found among strains in seropathotypes A and B that included the LEE-positive strains known to be associated with outbreaks (serotypes O157:H7, O111:H2, O111:H11*, O103:H2 and O26:H11). However, the prevalence of strains carrying PAI ICL3 was significantly higher among isolates in seropathotype C linked to disease [58 of 87 strains (67%)], than among isolates in both seropathotypes D and E not linked to disease [44 of 156 strains (28%)]. This finding revealed PAI ICL3 as a significant predictor of ‘virulent’ status among the LEE-negative strains [P<0.0001; OR=5.5 (95% CI=2.7–8.4)].

Our previous findings, with the same STEC strains, showed that the distribution pattern of virulence factors differed considerably between clonal groups (Girardeau et al., 2005). Accordingly, the present study revealed a close association between PAI ICL3 and certain clonal groups. Indeed, PAI ICL3 was detected in most (90%) strains belonging to serotypes ON: H21, O91:H21, O113:H21 and O174:H21, which are the archetypal strains of the virulent clonal group STEC-1 associated with HUS. Therefore, when only STEC-1 strains in seropathotype C were compared with isolates in seropathotypes D and E, the presence of the PAI ICL3 appeared as an even stronger predictor of virulent status [P<0.0001; OR=25.7 (95% CI=8.6–76.6)].

Identification of a similar PAI ICL3 element in the C. rodentium genome

Sequence data available in the colibase database indicate that the PAI ICL3 gene cluster is virtually identical at the nucleotide level (97% identical within 17.5 kb) to a genomic fragment identified in the genome of C. rodentium ICC168 (positions 5236633 to 5256813 in the genome sequence) (Fig. 1, bold type in Supplementary Table S1). The C. rodentium PAI ICL3 contains 15 ORFs; ORFs ROD49881 and ROD50041 define its left and right boundaries, respectively (corresponding in the E. coli CL3 PAI ICL3 to ORFs Z1635 and S14; accession nos AAQ19121.1 and AAQ19137.1, respectively). In contrast to the conserved ‘core’ of the PAI, the central segment bracketed by the two 190 bp direct repeat sequences differed completely in size and sequence between C. rodentium and E. coli (Fig. 1).

Sequence comparison of PAI ICL3 elements in E. coli and C. rodentium

Whereas the prototypic PAI ICL3 of E. coli CL3 is 21925 bp in length, the PAI ICL3 of C. rodentium spans 20181 bp (Fig. 1). The difference can be accounted for by a larger central segment in E. coli CL3 (ORFs S5–S9) than in the corresponding region of C. rodentium (ORFs CR33–CR28). Comparisons of the nucleotide sequences showed that this variable central segment contains a unique complement of genes. They encode short peptides that show homology with the C terminus of haemagglutinin-like proteins and with several proteins of unknown function (Supplementary Table S1). The fact that this segment is bracketed by direct repeat sequences (Shen et al., 2004) and has a low G+C content (42.8 and 40.2% in E. coli and C. rodentium, respectively) suggests that it has been horizontally transferred.

A second structural difference in the C. rodentium PAI ICL3 is the size of the gene that defines its right-hand boundary (CR21/22 and S14 in C. rodentium and E. coli, respectively). These two ORFs show high similarity to the gene that encodes the putative ImpA-like ATPase from the EAEC strain O42 (1368 bp in length) (locus Ec042-4550). However, while the C. rodentium sequence lying within
CR21 and CR22 ORFs includes 1365 bp from *E. coli* O42 ImpA, the *E. coli* S14 is much shorter (801 bp) (Fig. 1).

*C. rodentium* PAI ICL3 is carried by a 53 kb region with characteristics of a GEI

Using the sequence data available in the *coli* BASE server for *C. rodentium* strain ICC168, the genomic localization of the PAI ICL3 was investigated and the flanking regions were explored for signatures of mobile genetic elements. We found that the *C. rodentium* PAI ICL3 is carried by a 53 kb GEI inserted into the pheV-trNA locus. Moritz & Welch (2006) have proposed that islands be given unique names that comprise their chromosomal location relative to the sequence of *E. coli* strain MG1655 and the host strain number. In addition, when there is no evidence that a PAI is necessary for complete virulence of a pathogen, they propose that the phenotypically neutral abbreviation GI, for 'genomic island', be used. Therefore, the novel genomic island described here was named GI<sub>pheV</sub>-CRICC168 [for *pheV*-associated genomic island of the *C. rodentium* (CR) strain ICC168]. Its genetic organization is described in Supplementary Table S1 and Fig. 2.

About 13 kb downstream of the PAI ICL3 element, we found a prophage P4 integrase gene flanked by the *pheV*-trNA locus next to the yqgA gene. An attB-like site (phage P4 integration site) was identified 134 bp downstream from the integrase gene. The borders of the island can be defined by 23 bp direct repeats that are duplications of the 3' end of the *pheV*-trNA (TTCGATTCCGATCCGGGCACCA), containing part of the *attB*-like site, giving a total size of 53 197 bp for GIpheV-CR<sub>ICC168</sub>. This novel GEI (positions 5218424–5271439 bp in the genome sequence of *C. rodentium*) was predicted to contain 82 ORFs numbered from CR01 to CR82 (only ORFs longer than 65 amino acid residues were extracted). The CR01 gene (ROD50201) encodes a putative 441 aa peptide that has 97.6 % identity with the *Shigella flexneri* 2a strain 301 prophage P4 integrase (GenBank no. AAP18269). This integrase-encoding gene is intact and its start codon is adjacent to a ribosome-binding site.

**Insertion sequence (IS) and prophage elements**

Ten ORFs of GIpheV-CR<sub>ICC168</sub> were homologous to whole or partial IS elements or transposons of the IS3, IS4 or IS66 type (Supplementary Table S1, Fig. 2). Data from the IS Finder database revealed that a 2704 bp stretch of sequence lying within the CR10–CR12 gene cluster shares high similarity (96.5 %) with the IS679 element. The inverted repeat sequence GTAAGGNTTCANNNAACCGTNTT was found at both ends of this sequence, suggesting that the IS element is intact and potentially functional in GIpheV-CR<sub>ICC168</sub>. Additionally, it was flanked by an 8 bp

---

![Fig. 2](image-url)  
**Fig. 2.** Genetic organization of the GIpheV-CR<sub>ICC168</sub> GEI of *C. rodentium*. The GEI appears to consist of four distinct modules bound by mobile genetic elements. Module I exhibits marked similarities to the GIselC<sub>CFT073</sub> GEI (from locus c<sub>4509</sub> to locus c<sub>4517</sub>) of the extraintestinal pathogenic *E. coli* CFT073. Module II spans the PAI ICL3 element. Modules III and IV exhibit marked similarities with the AGI-3 GEI (modules 4 and 5, respectively) of the extra-intestinal APEC BEN2908. Closely related sequences between GEIs are indicated by grey shading. ORFs are coded as described in the key and labelled consecutively from 1 to 82, and correspond to CR01 to CR82. DR, direct repeat.
direct repeat (CCCTGATG), which could be the target sequence of the insertion. The identification of an IS679-like element in virulence plasmid pB171 of EPEC B171 (Tobe et al., 1999), PAI AGI-3 of avian pathogenic E. coli (APEC) (Chouikha et al., 2006), LEE of C. rodentium (Deng et al., 2001) and PAI of S. flexneri (Jin et al., 2002) raises the possibility that IS679 is involved in the transfer of virulence determinants between different bacterial strains and species. Other IS elements appeared to be defective in their transposition capacity, since these sequences were disrupted either by frameshift or by integration of another IS element. Additionally, five non-functional ORFs were homologous to prophage CP4-57 and CP-933L elements. The distribution of many of these mobile elements on GlpheV-CRICC168 suggested that extensive rearrangements have occurred during the evolution of this GEI.

**Genetic features of the GEI GlpheV-CRICC168**

The G+C content variations observed throughout the complete sequence of GlpheV-CRICC168 reflect the mosaic structure of the GEI. These detected variations indicated that GlpheV-CRICC168 comprises four distinct modules bound by insertion elements or bacteriophage sequences (Supplementary Table S1, Fig. 2).

Module I spans a 12399 bp segment (including CR07 to CR18) that is nearly identical (98.2 %) to a region (c4509–c4517) of the selC-associated GEI of the uropathogenic *E. coli* strain CFT073 (termed in this study GlselC_{CFT073}) (Welch et al., 2002). The complete size of this segment in GlpheV-CRICC168 is 2653 bp larger than that of the corresponding region in GlselC_{CFT073}. This difference is due to the presence of the IS679 element (CR10–CR12). The CR07 gene encodes a putative protein that shares similarity with *Bacillus licheniformis* tRNA nucleotidyl transferase. The putative CR14 gene product is identical to the *E. coli* CFT073 TatD-like DNase. Proteins that might be encoded by CR15 and CR17 genes are highly similar to the PP-loop superfamily ATPase proteins. In addition, this segment comprises seven genes of mobile genetic elements, such as a bacteriophage gene (CR18) and IS elements (CR04–CR06 and CR10–CR12), as well as ORFs of unknown function.

Module II spans the PAI I_{CL3} elements previously described in *E. coli* CL3. This 20.2 kb segment that contains 15 ORFs (from CR21 to CR43) represents the ‘core’ of the island. The G+C content (52.5 %) of module II differs markedly from that of module III located at its left flank, which averages only 45 %.

Modules III and IV were found to be present with the same organization in the selC-associated GEI AGI-3, involved in carbohydrate assimilation and virulence of the APEC strain BEN2908 (Chouikha et al., 2006). Module III, which contains 12 ORFs (CR45–CR65), spans a 9.4 kb segment which is nearly identical (94 % identity) to module 4 (ORFs acc-59 to acc-64) of the GEI AGI-3 and to a segment (ORF29–ORF34) of the 111 kb PAI IRW1374 of the EHEC strain RW1374 (Jores et al., 2005). The CR64 gene encodes a putative ERA-like GTP-binding protein involved in 16S rRNA maturation, regulation of the cell cycle, and protein synthesis in *E. coli* (Meier et al., 2000). The CR45 gene flanks the left junction of the PAI I_{CL3}; its predicted product is highly related (97 %) to a transposase of the IS3 family. Although it was truncated at its 5’ end by insertion of two IS elements (CR46 and CR47), this IS-related sequence may be a remnant of the PAI I_{CL3} association with module 4 of the GEI AGI-3.

Module IV has a significantly higher G+C content, which averages 56.6 %. This module spans an 8.5 kb segment (from CR67 to CR82) highly homologous to module 5 of the GEI AGI-3. This gene cluster is also present in two other phev-associated PAIs [ORF35–ORF47 of the 111 kb PAI IRW1374 of EHEC strain RW1374 (Jores et al., 2005) and the she PAI of *S. flexneri* (Al-Hasani et al., 2001)], and in two different GEIs of *E. coli* strain CFT073, inserted at the serX and phev loci (Chouikha et al., 2006). The order and orientation of the corresponding ORFs in GlpheV-CRICC168 are identical to those of their homologues in *E. coli* and *S. flexneri*. The CR67 gene product has high identity (97 %) to the *E. coli* autotransporter adhesin adhesin-involved-in-diffuse-adherence (AIDA-1), which belongs to the autotransporter protein of the type V secretion system family (Maurer et al., 1999). In the GEI AGI-3 of the APEC strain BEN2908, the acc-67 gene that encodes AIDA-1 is truncated at its 5’ end by the insertion of an IS911 transposase gene (Chouikha et al., 2006). In contrast, GlpheV-CRICC168 possesses an intact and potentially functional AIDA-1 gene (CR67) similar to the native AIDA-1 precursor of *E. coli* MG1655.

Other than the AIDA-1-like gene, Module IV contains only phage- and plasmid-related sequences. CR68 is highly related to a gene adjacent to the origin of transfer in plasmids F and R100, while CR69 has clear identity to antirestriction proteins of conjugative plasmids. At the right end of the GlpheV-CRICC168, four ORFs are present (CR77–CR80) with similarities to ORFs L007–L0012 of the putative prophage CP-933L in the EHEC LEE. As reported for the she PAI of *S. flexneri* (Al-Hasani et al., 2001), these phage- and plasmid-related sequences may be remnants of the PAI association with self-transmissible elements.

**PAI I_{CL3} is carried in *C. rodentium* and *E. coli* by different GEIs.**

The O island 48 (OI-48_{EDL933}), also termed the tellurite resistance- and adherence-conferring island (TAI), is inserted 2 bp from the serW-tRNA gene in *E. coli* EDL933 (Perna et al., 2001). As reported elsewhere (Shen et al., 2004), the prototypical *E. coli* PAI I_{CL3} contains EDL933 OI-48 genes Z1635, Z1636 and Z1637 at the left terminus, and Z1642, Z1643 and Z1644 at the right terminus (Fig. 1). From this finding, the *E. coli* PAI I_{CL3} is believed to reside within an *E. coli* CL3-homologous OI-48
GEI (termed in this study OI-48CL3). Although the C. rodentium PAI ICL3 closely resembles the E. coli PAI ICL3, it differs in its location. We show here that the C. rodentium PAI ICL3 is carried by the GEI GpheV-911 (EcO42-3187–EcO42-3188) inserted into the phv-tRNA gene (position 5272 Mb in the C. rodentium genome sequence). IS elements homologous to the putative transposase IS3A and sequence similar to the putative prophage CP-933L flank the left and right ends of the C. rodentium PAI ICL3, respectively (Fig. 1). We hypothesize that these elements have played a role in the integration of the PAI ICL3 element into the GEI of C. rodentium. Taken together, our observations indicate that the C. rodentium PAI ICL3 was inserted differently into the bacterial genome and thus represents an evolutionary lineage different from that of the E. coli PAI ICL3.

The genomes of other E. coli pathotypes contain deleted versions of the PAI ICL3

BLAST searches on the colibase server revealed that other genome-sequenced E. coli strains possess deleted PAI ICL3 sequences in which all that remain are the genes located at the extremities of the island. Six of the 16 E. coli genome sequences showed evidence of PAI ICL3 being inserted in at least three different chromosomal sites, which is compatible with the idea that PAI ICL3 ancestors entered E. coli and C. rodentium genomes at multiple times through independent events. Three patterns of deletion, which removed almost all of the PAI ICL3 sequences, were noted (Fig. 3).

In the O157:H7 EHEC strain EDL933, the PAI ICL3 ancestor was inserted within the serW-associated OI-48_EDEL933 (OI-48serWEDL933) GEI. The unique character of the PAI ICL3 appears to be due to genomic events related to the EDL933 OI-48 gene Z1640. Of particular interest is the observation that the middle portion of the PAI ICL3 consisted of the Z1640 gene separated into three fragments (found in S1, S3 and S14 genes of the PAI ICL3 sequence) by two genomic segments (GSI and GSII) that include the haemolysin/adhesin gene cluster and the EDL933 OI-122 segment, respectively (Shen et al., 2004). We have identified the rearrangements observed in EDL933 as a deletion that fuses the three segments found in S1, S3 and S14, to generate the chimaeric gene Z1640 (Fig. 3a). The presence of the putative functional insertion sequence IS629 (ORFs Z1638 and Z1639) contiguous to Z1640 may indicate that the chimaeric Z1640 gene arose following an IS-mediated deletion. This deletion event has removed a large part (19 kb) of the PAI ICL3 gene cluster carried by the OI-48EDL933 ancestor. An identical rearrangement was found in the EHEC O157:H7 Sakai strain (EcS1382 gene) and in the EPEC O111:H9 E110019 strain (EcolE1_01002508 gene), demonstrating that a similar pattern of deletion has occurred in these strains.

In the ExPEC strain CFT073, the PAI ICL3 ancestor was inserted within a selC-associated GEI (GselC_CFT073) which contains a S. flexneri shiA homologue and the ethanola-mine utilization gene cluster (Welch et al., 2002). As described above (Fig. 2), the genetic organization and DNA content of module I of GpheV-911 are similar to those of the corresponding region (c4509–c4517) of GselC_CFT073. The homology ends immediately upstream of the c4517 locus. Preceding c4517, the c4518 locus is identical to CR47 located at the 5’ end of the PAI ICL3 element of GpheV-911 (Fig. 3b). As the region located between c4517 and c4518 appears considerably shorter than its equivalent in the full PAI ICL3 gene cluster, we concluded that a deletion including a large part (17 kb) of the PAI ICL3 gene cluster had occurred. A 56 bp sequence corresponding to the IS911 5’ end was found adjacent to c4514 (Fig. 3). This IS segment, which contains the canonical C-terminal inverted repeat, might be the remnant of an IS-assisted deletion in a GselC_CFT073 ancestor. A homologous pattern of deletion was also identified within the APEC O1 strain.

In the EAEC strain O42, the PAI ICL3 ancestor was inserted at position 3 405 569 bp within a phv-associated GEI (GpheVo42) which carries the gene cluster encoding the antibiotic peptide microcin H47 (ORFs, EcO42-3195/mchF-EcO42-3199/mchC). Following the microcin gene cluster, a 2 kb segment (region EcO42-3191–EcO42-3188) is identical to the region CR37–CR43 located at the 5’ end of the C. rodentium PAI ICL3 gene cluster (Fig. 3c). ORF EcO42-3188 is followed by a 3.5 kb segment (including the locus EcO42-3187) identical to region CR21–CR24 located at the 3’ end of the C. rodentium PAI ICL3 gene cluster. The region located between EcO42-3188 and EcO42-3187 appears considerably shorter than its equivalent in the full PAI ICL3 gene cluster. We concluded that a deletion removing a large part (17 of 22 kb) of the PAI ICL3 gene cluster had occurred between the two loci CR37 and CR24. The intergenic sequence between EcO42-3187 and EcO42-3188 contains a 305 bp sequence nearly identical to the 3’ end of the Shigella dysenteriae insertion sequence IS911. The presence of IS911 sequences may indicate that the EcO42-3188–EcO42-3187 region arose following an IS911-mediated deletion event that has removed a large part of the PAI ICL3 element in the microcin region of EAEC strain O42.

Different variants of the PAI ICL3 gene cluster are present in C. rodentium and E. coli

On the basis of the host GEI, the integration site and the pattern of deletion, five variant types of PAI ICL3 were identified. Their characteristics are summarized in Table 2. The first variant (CL3_PA1 IC3) contains the full complement of PAI ICL3 genes previously identified in E. coli CL3 (Shen et al., 2004) and resides within an E. coli CL3-homologous OI-48 GEI (OI-48CL3). The second variant (CR-PAI ICL3) contains the C. rodentium PAI ICL3 gene cluster carried by the GEI GpheV-911 inserted at the phv-tRNA locus, described above. In a third variant (EDL933-del/PAI ICL3), the O island OI-48-EDL933 (inserted at the serW-tRNA locus) carries the
deleted derivative of the PAI I\textsubscript{CL3} gene cluster found in different evolutionary lineages of A/E pathogens (EHEC strains EDL933 and Sakai, and atypical EPEC strain E110019). In the fourth variant (CFT073-del/PAI I\textsubscript{CL3}), the GEI GI\textsubscript{selC}\textsubscript{CFT073} (inserted at the \textit{selC}\textsubscript{-tRNA} locus) carries the deleted derivative of the PAI I\textsubscript{CL3} gene cluster found in ExPEC strain CFT073. In the last variant (O42-del/PAI I\textsubscript{CL3}), the GEI GI\textsubscript{pheV-O42} of EAEC strain O42.

**DISCUSSION**

The PAI I\textsubscript{CL3} element is a common component of the genome of LEE-negative STEC strains linked to disease, and thus provides a new marker for these strains.

By screening a collection of 469 \textit{E. coli} strains, we demonstrated that the complete PAI I\textsubscript{CL3} is a widespread genomic element among LEE-negative STEC strains, and is missing in all other pathotypes of \textit{E. coli}, as well as in

---

**Fig. 3.** Patterns of deletion of the PAI I\textsubscript{CL3} gene cluster among different \textit{E. coli} pathotypes. (a) A deletion of 19 kb that fuses three segments of the S1, S3 and S14 genes of the prototypic PAI I\textsubscript{CL3} locus, to generate the chimaeric fusion gene Z1640 of the O island OI-48 from the O157:H7 EHEC strain EDL933. (b) A deletion of 17.3 kb that removes a large region (from S1 to S14) from the PAI I\textsubscript{CL3} locus identified in the GEI GI\textsubscript{selC} of ExPEC strain CFT073. (c) A deletion of 17 kb that removes a large region (from S1 to S11) from the PAI I\textsubscript{CL3} locus identified in the GEI GI\textsubscript{pheV-O42} of EAEC strain O42. IS sequences flanking the GEIs are in dark blue.
commensals. In a recent study we found a clear association of certain virulence factors with seropathotypes linked to disease (Girardeau et al., 2005). Consistent with these findings, this study showed a significantly higher prevalence of the PAI ICL3 in the virulent seropathotype C than in the non-virulent seropathotypes D and E. In addition, the presence of the PAI ICL3 element in most strains belonging to the virulent clonal group STEC-1 indicates that PAI ICL3 is a common component of the genome of certain E. coli lineages. Therefore, when LEE-negative STEC strains in the virulent seropathotype were compared with isolates in non-virulent seropathotypes, PAI ICL3 appeared to be a significant predictor of the virulent status.

### Table 2. Variant types of PAI ICL3

<table>
<thead>
<tr>
<th>PAI ICL3 variant type</th>
<th>PAI ICL3 version</th>
<th>Deletion size (kb)</th>
<th>Species</th>
<th>E. coli pathotype</th>
<th>LEE</th>
<th>Strain</th>
<th>Host GEI</th>
<th>Insertion site</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL3-PAI ICL3</td>
<td>Complete</td>
<td></td>
<td>E. coli</td>
<td>STEC</td>
<td>Absent</td>
<td>CL3</td>
<td>OI-48</td>
<td>Unknown</td>
</tr>
<tr>
<td>CR-PAI ICL3</td>
<td>Complete</td>
<td></td>
<td>C. rodentium</td>
<td>Prese..</td>
<td>Present</td>
<td>ICC168</td>
<td>GI -CRICC168</td>
<td>pheV</td>
</tr>
<tr>
<td>EDL933-del/PAI ICL3</td>
<td>Deleted</td>
<td>19.0</td>
<td>E. coli</td>
<td>EHEC</td>
<td>Present</td>
<td>EDL933</td>
<td>GI -CRICC168</td>
<td>serW</td>
</tr>
<tr>
<td></td>
<td></td>
<td>19.0</td>
<td>E. coli</td>
<td>EHEC</td>
<td>Present</td>
<td>Sakai</td>
<td>GI -CRICC168</td>
<td>serW</td>
</tr>
<tr>
<td></td>
<td></td>
<td>19.0</td>
<td>E. coli</td>
<td>EHEC</td>
<td>Present</td>
<td>E110019</td>
<td>GI -CRICC168</td>
<td>serW</td>
</tr>
<tr>
<td>CFT073-del/PAI ICL3</td>
<td>Deleted</td>
<td>17.5</td>
<td>E. coli</td>
<td>ExPEC</td>
<td>Absent</td>
<td>CFT073</td>
<td>GI -GselCFT073</td>
<td>selC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>17.5</td>
<td>E. coli</td>
<td>ExPEC</td>
<td>Absent</td>
<td>APEC O1</td>
<td>GI -GselCFT073</td>
<td>selC</td>
</tr>
<tr>
<td>O42-del/PAI ICL3</td>
<td>Deleted</td>
<td>17.0</td>
<td>E. coli</td>
<td>EAEC</td>
<td>Absent</td>
<td>O42</td>
<td>GI -GpheO42</td>
<td>pheV</td>
</tr>
</tbody>
</table>

In *C. rodentium*, a 53 kb GEI carries a PAI ICL3-equivalent gene cluster

A PAI ICL3-equivalent gene cluster was also detected in the genome of *C. rodentium*. The fact that in both *E. coli* and *C. rodentium* the PAI ICL3 element is extremely conserved at the nucleotide level and in genetic organization strongly supports the idea that the PAI ICL3 from *E. coli* CL3 and *C. rodentium* ICC168 shares a common origin. Deviation in G+C content of a gene compared with the whole genome is often a valuable marker for identifying genes recently acquired by horizontal transfer. The G+C content of the PAI ICL3 haemolysin/adhesin gene cluster is 56.6 %, compared with ~50.1 % for the *E. coli* genome, suggesting that this DNA segment was recently acquired by *E. coli*.

We further report that *C. rodentium* PAI ICL3 is borne on a 53 kb GEI termed GlpheV-CR ICC168. The presence of a putative functional phage P4-like integrase, the insertion into a tRNA locus (*pheV*), and the presence of 23 bp direct repeat sequences carrying an *attB*-like site, suggest that the GEI GlpheV-CR ICC168 was acquired from a bacteriophage via horizontal transfer. This novel *C. rodentium* GEI exhibits marked similarities to certain chromosomal regions of ExPEC CFT073 (module I), STEC CL3 (module II) and APEC BEN2908 or EHEC RW1374 (modules III and IV) strains, particularly with regard to certain GEIs of these strains. All of these regions were surrounded by intact or fragmented IS elements, suggesting that GlpheV-CR ICC168 acquired these regions through the horizontal transfer of mobile elements. In module I of GlpheV-CR ICC168, located upstream of the PAI ICL3 element, is the intact 2.7 kb IS679. IS679, also found next to the *C. rodentium* LEE, is thought to have played a role in acquisition of the LEE by *C. rodentium* (Deng et al., 2001).

We could identify 82 ORFs within GlpheV-CR ICC168. In addition to the putative virulence factors, including the *Yersinia* haemagglutinin/adhesin gene cluster and pagC carried by the prototypic *E. coli* PAI ICL3 (Shen et al., 2004), the *C. rodentium* GEI contains a gene encoding the diffusely adhering *E. coli* adhesin AIDA-1. Given the presence of these putative virulence genes, this GEI could be considered a PAI. However, because there is no evidence that the GEI GlpheV-CR ICC168 is necessary for complete virulence of *C. rodentium*, according to the Moritz nomenclature (Moritz & Welch, 2006), the phenotypically neutral abbreviation GI, for ‘genomic island’, was used.

*C. rodentium* and *E. coli* possess different evolutionary lineages of the PAI ICL3 elements

The PAI ICL3 variant types of both *E. coli* CL3 and *C. rodentium* possess the full complement of PAI ICL3 genes and are highly conserved. However, they have been shown to be inserted into different host GEIs. While the *C. rodentium* variant type (CR-PAI ICL3) is borne on the GEI GlpheV-CR ICC168, the *E. coli* CL3 variant type (CL3-PAI ICL3) is believed to reside within a homologous OI-48 GEI (OI-48 CL3). It is believed that the LEE has been acquired by A/E pathogens at multiple times, possibly via the horizontal transfer of a putative plasmid (Deng et al., 2001). As reported for the LEE, it appears that the PAI ICL3 element was acquired by *E. coli* and *C. rodentium* after horizontal transfer through independent events. IS elements, similar in sequence to putative prophage CP-933L and other putative DNA-binding proteins that flank the ends of the PAI ICL3 from *C. rodentium* and *E. coli* CL3, may have played a role in the integration of PAI ICL3 into the host GEI.
The PAI ICL3 element is present in other *E. coli* pathotypes, but it has been subjected to extensive deletions

Other than STEC, we discovered that other *E. coli* pathotypes (including A/E and non-A/E pathogens) possess deleted PAI ICL3 sequences in which all that remain are the genes located at the extremities of the genomic segment. In the deleted derivatives of PAI ICL3, the junction boundaries are often marked by truncated genes fused to generate a chimaeric gene (i.e. EDL933 Z1640), revealing that a deletion has occurred. The PAI ICL3 variant types EDL933-del/PAI ICL3, CFT073-del/PAI ICL3, and O42-del/PAI ICL3 (from EHEC/EPEC, ExPEC and EAEC O42, respectively), have undergone extensive deletions that removed almost all of the PAI ICL3. ISSs of several classes (IS3, IS911 and IS629) are found at the sites of deletion, suggesting that homologous recombination between such elements following deletion accounts for the multigene deletions. Such deletion events explain the absence of PAI ICL3 elements among LEE-positive isolates. Hence, the difference between LEE-positive and LEE-negative STEC strains is a deletion in LEE-positive rather than an insertion into LEE-negative STEC strains.

It remains unclear whether IS elements found at the junction sites are responsible for (or remnants of) the observed deletions. However, IS-mediated deletions and genetic instability have been frequently observed in well-known PAIs and GEIs, e.g. *Y. pestis* HPI (Schubert et al., 1998), ETT2 (Ren et al., 2004) and Salmonella Spi (Amavisit et al., 2003). The ETT2 cluster (encoding a second cryptic type III secretion system) simultaneously provides a model of gene flux and of genetic loss, and shows a whole spectrum of reductive evolution, from an apparently intact 27.5 kb cluster in *E. coli* O42 to only two residual gene fragments in *S. flexneri* (Ren et al., 2004). As reported here, the PAI ICL3 cluster provides a new example of gain and loss of genetic elements from one lineage to another.

PAI ICL3 ancestors entered the *E. coli* genomes at multiple times, through independent events

Many tRNA genes are frequently used as integration sites for GEIs (Gal-Mor & Finlay, 2006; Hacker & Carniel, 2001). Among six sequenced *E. coli* genomes, as well as in the genome of *C. rodentium*, the PAI ICL3 gene cluster was found to be carried by five different GEIs associated with three different tRNA loci (*pheV, selC* and *serW*). The observed relationships between the *E. coli* pathotype and the integration site in a tRNA locus (EHEC/EPEC/*serW*, ExPEC/*selC*, EAEC/*pheV*) are compatible with the idea that different host GEIs carrying the PAI ICL3 genes entered the *E. coli* genomes at multiple times, through independent events, and then integrated in different tRNA loci. Various mechanisms could account for the generation of the deleted PAI ICL3 variant types. One scenario is that a host GEI carrying the full complement of PAI ICL3 genes entered the genome first, and then has been subject to IS-mediated deletion. In another scenario, PAI ICL3 assembled on another element (i.e. a resident plasmid or phage) has been subject first to deletion and then integrated into a host genomic island in an insertional hotspot.

PAI ICL3 dissemination throughout *E. coli* strains of different pathotypes (EHEC, EPEC, ExPEC and EAEC) and its wide distribution among distantly related LEE-negative STEC strains imply an efficient mechanism of transfer. Temperate phages or transmissible plasmids are candidates for PAI ICL3 vehicles. The putative functional phage-like integrase, the *att* sites, the direct repeat sequences, and the targeting of tRNA loci as integration sites shared by the different PAI ICL3-host GEIs, suggest that these GEIs may have been acquired from a bacteriophage via horizontal transfer.

A B1 genetic background appears necessary for the maintenance of a full complement of PAI ICL3 genes

We noted an important distinction between *C. rodentium* and LEE-negative STEC which contain a complete PAI ICL3 gene cluster and other *E. coli* pathotypes which contain deleted PAI ICL3 gene clusters. A striking phylogenetic distribution of the PAI ICL3 genotypes could explain this distinction. While six major phylogenetic groups of *E. coli* (A, B1, C, E, D and B2) form the core of the *E. coli* species (Escobar-Paramo et al., 2004), we showed in a previous study that LEE-negative STEC strains in seropathotype C belong exclusively to the ECOR group B1 (Girardeau et al., 2005). Therefore, the observed link between complete PAI ICL3 genotype and seropathotype C demonstrates a striking phylogenetic link between intact PAI ICL3 and ECOR group B1. Consistent with previous studies that suggest that the arrival and stability of a PAI in a genome require a particular genetic background (Escobar-Paramo et al., 2004; Reid et al., 2000), a B1 genetic background seems to be necessary for the maintenance of a full complement of PAI ICL3 genes. The finding that *E. coli* strains belonging to phylogenetic groups E (EHEC EDL933 and Sakai), B2 (ExPEC CFT073 and APEC O1) and D (EAEC O42) carry exclusively the deleted versions of PAI ICL3 supports this concept.

Conclusion

Taken together, our results indicate that (i) the PAI ICL3 gene cluster is a common component of the genome of LEE-negative STEC strains linked to disease, and could provide a new marker for these strains; (ii) a PAI ICL3-equivalent gene cluster is present in the genome of *C. rodentium*; (iii) the *C. rodentium* PAI ICL3 is borne on a genomic region with characteristics typical of a horizontally transferable GEI; (iv) other *E. coli* pathotypes (including A/E and non-A/E pathogens) possess deleted subtypes of PAI ICL3; (v) the PAI ICL3 gene cluster entered *E. coli* genomes at multiple times, through independent
events; and (vi) a B1 genetic background is necessary for the maintenance of a full complement of PAI ICL3 genes in E. coli.

REFERENCES


J. P. Girardeau, Y. Bertin and C. Martin


Edited by: R. J. Lamont