Genomic analysis of the PAI I_{CL3} 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

Correspondence

Christine Martin

cmartin@clermont.inra.fr

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 I_{CL3}, has been previously characterized in the LEE-negative O113:H21 STEC strain CL3. Screening for the presence of PAI I_{CL3} elements in 469 strains of *E. coli*, including attaching and effacing (A/E) pathogens [enteropathogenic *E. coli* (EPEC) and enterohaemorrhagic *E. coli* (EHEC)], non-A/E pathogens [LEE-negative STEC, extra-intestinal pathogenic *E. coli* (ExPEC), enterotoxigenic *E. coli* (ETEC) and enteroaggregative *E. coli* (EAEC)] and commensal *E. coli* isolates, showed that PAI I_{CL3} is unique to LEE-negative STEC strains linked to disease, providing a new marker for these strains. We also showed that a PAI I_{CL3}-equivalent gene cluster is present in the genome of *Citrobacter rodentium*, on a 53 kb genomic island inserted into the *pheV* tRNA locus. While the *C. rodentium* PAI I_{CL3} shows high similarities at the nucleotide level and in organization with the *E. coli* PAI I_{CL3}, 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 I_{CL3} elements that contain only the genes located at the extremities of the island. Six of the 16 sequenced *E. coli* genomes showed deleted PAI I_{CL3} gene clusters which are carried on mobile genetic elements inserted into *pheV*, *sefC* or *serW* tRNA loci, which is compatible with the idea that the PAI I_{CL3} gene cluster entered *E. coli* and *C. rodentium* at multiple times through independent events. The phylogenetic distribution of the PAI I_{CL3} variants suggests that a B1 genetic background is necessary for the maintenance of the full complement of PAI I_{CL3} genes in *E. coli*.

Received 17 December 2008

Revised 15 January 2009

Accepted 15 January 2009

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).

Abbreviations: A/E, attaching and effacing; AIEC, adherent-invasive *E. coli*; APEC, avian pathogenic *E. coli*; CI, confidence intervals; EAEC, enteroaggregative *E. coli*; EHEC, enterohaemorrhagic *E. coli*; EPEC, enteropathogenic *E. coli*; ETEC, enterotoxigenic *E. coli*; ExPEC, extra-intestinal pathogenic *E. coli*; GEI, genomic island; HUS, haemolytic–uraemic syndrome; IS, insertion sequence; LEE, locus of enterocyte effacement; NMEC, neonatal meningitis *E. coli*; OR, odds ratio; PAI, pathogenicity island; STEC, Shiga toxin-producing *Escherichia coli*.

A supplementary table listing genetic features of the identified *GlpheV*-CR_{ICC168} ORFs is available with the online version of this paper.

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 enterocyte effacement (LEE) (Deng *et al.*, 2004; Frankel *et al.*, 1998; Garmendia *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 (Boerlin *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 I_{CL3} (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 I_{CL3} 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 I_{CL3} 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 I_{CL3} element among LEE-negative STEC strains, and its absence from other *E. coli* pathotypes. In addition, we discover that the PAI I_{CL3} is also present in *C. rodentium*. We further report that the *C. rodentium* PAI I_{CL3} is borne on a 53 kb GEI termed in this study GlpheV-CR_{ICC168}.

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; Vernozzy-Rozand *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 EHEC O157:H7 strains EDL933 and Sakai, were kindly provided by Josée Harel (GREMIP, Faculté de Médecine Vétérinaire, Université de Montréal, Canada), Arlette Darfeuille Michaud (Pathogénie Bactérienne Intestinale, Université d'Auvergne, Clermont-Ferrand, France), Chantal Le Bouguénec

Table 1. PCR primers used in this study

Probe	Target DNA	Primer sequences	Size (bp)
<i>ms-1</i>	Z1640::S1 junction (left junction with OI-48)	GCTGATGCGTTACCACTG, GTAATCCTCAACCGCACCAG	684
<i>ms-2</i>	S4 (<i>shlA/secA/fhaA</i> -like gene)	CCTGAAGTGACCGTGAAACAG, TGTATTGTCTGCGCTTCCAG	498
<i>ms-3</i>	S10/S11 (<i>pagC</i>) junction	GCGTGTAGCAGCTCATGCAG, GACAACACTGACCGGATAATC	880
<i>ms-4</i>	Z4318::S12 junction (left junction with OI-122)	CCTGAAGTGACCGTGAAACAG, GTCGGCCGTCACCTTAATAC	324
<i>ms-5</i>	S14::Z1642 junction (right junction with OI-48)	CTCCGGACGTACAGGAATATC, GTTGGCGTTATCTGACATAC	640

(unité de Pathogénie Bactérienne des Muqueuses, Institut Pasteur, Paris), and the NFSTC. *C. rodentium* strain DBS100 was a kind gift from K. T. Wilson (Vanderbilt University, USA) with authorization of D. Schauer (MIT, USA).

Investigation of *E. coli* strains for the presence of PAI I_{CL3} elements by Southern hybridization. The 469 *E. coli* strains and *C. rodentium* strain DBS100 were surveyed for the presence of five marker sequences specific to the prototypic PAI I_{CL3} . Marker sequences included a first region spanning the 5' end junction with the EDL933 GEI OI-48 (*ms-1*) and the 5' end-haemolysin/adhesin gene (*ms-2*), a second region spanning elements of the EDL933 GEI OI-122 (*ms-3* and *ms-4*), and a third region spanning the 3' end junction with the EDL933 GEI OI-48 (*ms-5*) (Table 1, Fig. 1). Strains were screened by dot-blot hybridization assay using DIG-labelled probes generated with a PCR DIG Probe Synthesis kit (Roche Diagnostics). Southern hybridization was carried out with DNA probes generated from PCR products using *E. coli* CL3 as the template. Primers shown in Table 1 were designated on the basis of the sequence of the prototypic PAI I_{CL3} (GenBank accession no. AY275838). Hybridizations were carried out overnight at 42 °C. ECL labelling and signal detection systems (Amersham/Pharmacia

Biotech) were used for detection. *E. coli* CL3 and MG1655 were used as positive and negative controls, respectively.

Sequence analysis. Comparative analyses of the regions surrounding and containing the PAI I_{CL3} -related loci were carried out and visualized by use of the *coliBASE* server (<http://colibase.bham.ac.uk>) (Chaudhuri *et al.*, 2004). This database contains all of the complete *Buchnera*, *E. coli*, *Photobacterium*, *Serratia*, *Shigella*, *Salmonella* and *Yersinia* genome sequences, together with preliminary data from a number of sequencing projects currently in progress at the Wellcome Trust Sanger Institute. The nucleotide sequences of *C. rodentium* strain ICC168 were found in the *coliBASE* database. The *C. rodentium* genome sequencing project is being carried out by the Wellcome Trust Sanger Institute in collaboration with Gad Frankel and Gordon Dougan of the Centre for Molecular Microbiology and Infection, Imperial College, UK. Sequence data were produced by the Microorganisms Sequencing Group at the Sanger Institute and can be obtained from http://www.sanger.ac.uk/Projects/C_rodentium/.

Each coding sequence annotated in the published PAI I_{CL3} of the archetypal *E. coli* strain CL3 (Shen *et al.*, 2004) was analysed by BLAST searches on the *coliBASE* server, supplemented as needed by

Fig. 1. Genetic organization of the GEI and flanking sequences in the *E. coli* CL3 PAI I_{CL3} in comparison to the *C. rodentium* PAI I_{CL3} . 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* *shlA/secA/fhaA*-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 *shlA/secA/fhaA*-like gene (creating the pseudogene S3 in the prototypic *E. coli* PAI I_{CL3}). The identities or characteristics of ORFs are detailed in Supplementary Table S1 (bold type). *E. coli* PAI I_{CL3} ORFs are based on published sequences (Shen *et al.*, 2004).

visualization of the genomic context of homologues using an Artemis applet (Rutherford *et al.*, 2000) within *coli*BASE, by multiple alignments on the EBI CLUSTAL W server (<http://www.ebi.ac.uk/clustalw>), and by searches of the National Center for Biotechnology Information conserved domain database (<http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>). BLAST, genomic BLAST (with microbial genome) and FASTA searches with the non-redundant protein and nucleotide databases were accessed at <http://www.ncbi.nlm.nih.gov> and <http://www.ebi.ac.uk/fasta3/>. The IS Finder database was accessed at <http://www-is.biotoul.fr>.

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 I_{CL3} elements among A/E pathogens and other *E. coli* pathotypes

The presence of the PAI I_{CL3} 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 I_{CL3} (Fig. 1). The five PAI I_{CL3} 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 I_{CL3} element carried by these strains has a structure similar to that of the prototypic PAI I_{CL3}. Interestingly, the PAI I_{CL3} 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 I_{CL3}

Consistent with an earlier study (Shen *et al.*, 2004), PAI I_{CL3}-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:H⁻, O103:H2 and O26:H11). However, the prevalence of strains carrying PAI I_{CL3} 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 I_{CL3} 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 I_{CL3} and certain clonal groups. Indeed, PAI I_{CL3} 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 I_{CL3} 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 I_{CL3} element in the *C. rodentium* genome

Sequence data available in the *coli*BASE database indicate that the PAI I_{CL3} 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 5 236 633 to 5 256 813 in the genome sequence) (Fig. 1, bold type in Supplementary Table S1). The *C. rodentium* PAI I_{CL3} contains 15 ORFs; ORFs ROD49881 and ROD50041 define its left and right boundaries, respectively (corresponding in the *E. coli* CL3 PAI I_{CL3} 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 I_{CL3} elements in *E. coli* and *C. rodentium*

Whereas the prototypic PAI I_{CL3} of *E. coli* CL3 is 21 925 bp in length, the PAI I_{CL3} of *C. rodentium* spans 20 181 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 I_{CL3} 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 EcO42-4550). 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 I_{CL3} 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 I_{CL3} was investigated and the flanking regions were explored for signatures of mobile genetic elements. We found that the *C. rodentium* PAI I_{CL3} 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 *GlpeV*-CR_{ICC168} [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 I_{CL3} 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 (TTCGATTCCGAGTCCGGGCACCA), containing part of the *attB*-like site, giving a total size of 53 197 bp for *GlpeV*-CR_{ICC168}. This novel GEI (positions 5 218 242–5 271 439 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 *GlpeV*-CR_{ICC168} 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 GTAAGCGNNTCANNNAACCGTNTT was found at both ends of this sequence, suggesting that the IS element is intact and potentially functional in *GlpeV*-CR_{ICC168}. Additionally, it was flanked by an 8 bp

Fig. 2. Genetic organization of the *GlpeV*-CR_{ICC168} 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 *GlseI*/C_{CFT073} GEI (from locus c₄₅₀₉ to locus c₄₅₁₇) of the extraintestinal pathogenic *E. coli* CFT073. Module II spans the PAI I_{CL3} 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*-CR_{ICCI68} suggested that extensive rearrangements have occurred during the evolution of this GEI.

Genetic features of the GEI *GlpheV*-CR_{ICCI68}

The G+C content variations observed throughout the complete sequence of *GlpheV*-CR_{ICCI68} reflect the mosaic structure of the GEI. These detected variations indicated that *GlpheV*-CR_{ICCI68} 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 *GIselC*_{CFT073}) (Welch *et al.*, 2002). The complete size of this segment in *GlpheV*-CR_{ICCI68} is 2635 bp larger than that of the corresponding region in *GIselC*_{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 *aec*-59 to *aec*-64) of the GEI AGI-3 and to a segment

(ORF29–ORF34) of the 111 kb PAI I_{RW1374} 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 I_{RW1374} 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*-CR_{ICCI68} 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-I), 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 *aec*-67 gene that encodes AIDA-I is truncated at its 5' end by the insertion of an IS911 transposase gene (Chouikha *et al.*, 2006). In contrast, *GlpheV*-CR_{ICCI68} possesses an intact and potentially functional AIDA-I gene (CR67) similar to the native AIDA-I precursor of *E. coli* MG1655.

Other than the AIDA-I-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*-CR_{ICCI68}, 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 prototypic *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-48_{CL3}). Although the *C. rodentium* PAI I_{CL3} closely resembles the *E. coli* PAI I_{CL3}, it differs in its location. We show here that the *C. rodentium* PAI I_{CL3} is carried by the GEI *GIpheV*-CR_{ICC168} inserted into the *pheV*-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 I_{CL3}, respectively (Fig. 1). We hypothesize that these elements have played a role in the integration of the PAI I_{CL3} element into the GEI of *C. rodentium*. Taken together, our observations indicate that the *C. rodentium* PAI I_{CL3} was inserted differently into the bacterial genome and thus represents an evolutionary lineage different from that of the *E. coli* PAI I_{CL3}.

The genomes of other *E. coli* pathotypes contain deleted versions of the PAI I_{CL3}

BLAST searches on the *coli*BASE server revealed that other genome-sequenced *E. coli* strains possess deleted PAI I_{CL3} 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 I_{CL3} being inserted in at least three different chromosomal sites, which is compatible with the idea that PAI I_{CL3} 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 I_{CL3} sequences, were noted (Fig. 3).

In the O157:H7 EHEC strain EDL933, the PAI I_{CL3} ancestor was inserted within the *serW*-associated OI-48_{EDL933} (OI-48_{serW}_{EDL933}) GEI. The unique character of the PAI I_{CL3} 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 I_{CL3} consisted of the Z1640 gene separated into three fragments (found in S1, S3 and S14 genes of the PAI I_{CL3} 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 I_{CL3} gene cluster carried by the OI-48_{EDL933} 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 I_{CL3} ancestor was inserted within a *selC*-associated GEI (*GIselC*_{CFT073}) which contains a *S. flexneri* *shiA* homologue and the ethanolamine utilization gene cluster (Welch *et al.*, 2002). As

described above (Fig. 2), the genetic organization and DNA content of module I of *GIpheV*-CR_{ICC168} are similar to those of the corresponding region (c4509–c4517) of *GIselC*_{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 I_{CL3} element of *GIpheV*-CR_{ICC168} (Fig. 3b). As the region located between c4517 and c4518 appears considerably shorter than its equivalent in the full PAI I_{CL3} gene cluster, we concluded that a deletion including a large part (17.5 of 22 kb) of the PAI I_{CL3} 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 *GIselC*_{CFT073} ancestor. A homologous pattern of deletion was also identified within the APEC O1 strain.

In the EAEC strain O42, the PAI I_{CL3} ancestor was inserted at position 3 405 569 bp within a *pheV*-associated GEI (*GIpheV*_{O42}) 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 I_{CL3} 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 I_{CL3} gene cluster. The region located between EcO42-3188 and EcO42-3187 appears considerably shorter than its equivalent in the full PAI I_{CL3} gene cluster. We concluded that a deletion removing a large part (17 of 22 kb) of the PAI I_{CL3} 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 I_{CL3} element in the microcin region of EAEC strain O42.

Different variants of the PAI I_{CL3} 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 I_{CL3} were identified. Their characteristics are summarized in Table 2. The first variant (CL3-PAI I_{CL3}) contains the full complement of PAI I_{CL3} genes previously identified in *E. coli* CL3 (Shen *et al.*, 2004) and resides within an *E. coli* CL3-homologous OI-48 GEI (OI-48_{CL3}). The second variant (CR-PAI I_{CL3}) contains the *C. rodentium* PAI I_{CL3} gene cluster carried by the GEI *GIpheV*-CR_{ICC168} inserted at the *pheV*-tRNA locus, described above. In a third variant (EDL933-del/PAI I_{CL3}), the O island OI-48_{EDL933} (inserted at the *serW*-tRNA locus) carries the

Fig. 3. Patterns of deletion of the PAI I_{CL3} gene cluster among different *E. coli* pathotypes. (a) A deletion of 19 kb that fuses three segments of the S1, S3 and S4 genes of the prototypic PAI I_{CL3} locus, to generate the chimeric 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_{CL3} locus identified in the GEI Glse/C_{CFT073} of ExPEC strain CFT073. (c) A deletion of 17 kb that removes a large region (from S1 to S11) from the PAI I_{CL3} locus identified in the GEI GlpheV_{O42} of EAEC strain O42. IS sequences flanking the GEIs are in dark blue.

deleted derivative of the PAI I_{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_{CL3}), the GEI Glse/C_{CFT073} (inserted at the *selC*-tRNA locus) carries the deleted derivative of the PAI I_{CL3} gene cluster found in ExPEC strains CFT073 and APEC 01. In the last variant (O42-del/PAI I_{CL3}), the GEI GlpheV_{O42} (inserted at the *pheV*-tRNA locus) carries the deleted derivative of the PAI I_{CL3} gene cluster found in the EAEC strain O42.

DISCUSSION

The PAI I_{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 *E. coli* strains, we demonstrated that the complete PAI I_{CL3} is a widespread genomic element among LEE-negative STEC strains, and is missing in all other pathotypes of *E. coli*, as well as in

Table 2. Variant types of PAI I_{CL3}

PAI I _{CL3} variant type	PAI I _{CL3} version	Deletion size (kb)	Species	<i>E. coli</i> pathotype	LEE	Strain	Host GEI	Insertion site
CL3-PAI I _{CL3}	Complete		<i>E. coli</i>	STEC	Absent	CL3	OI-48 _{CL3}	Unknown
CR-PAI I _{CL3}	Complete		<i>C. rodentium</i>		Present	ICC168	GIp _{heV} -CR _{ICC168}	<i>pheV</i>
EDL933-del/PAI I _{CL3}	Deleted	19.0	<i>E. coli</i>	EHEC	Present	EDL933	OI-48 _{EDL933}	<i>serW</i>
			<i>E. coli</i>	EHEC	Present	Sakai	OI-48 _{EDL933}	<i>serW</i>
			<i>E. coli</i>	EHEC	Present	E110019	OI-48 _{EDL933}	<i>serW</i>
			<i>E. coli</i>	ExPEC	Absent	CFT073	GIse _l C _{CFT073}	<i>selC</i>
CFT073-del/PAI I _{CL3}	Deleted	17.5	<i>E. coli</i>	ExPEC	Absent	APEC O1	GIse _l C _{CFT073}	<i>selC</i>
O42-del/PAI I _{CL3}	Deleted	17.0	<i>E. coli</i>	EAEC	Absent	O42	GIp _{heV} _{O42}	<i>pheV</i>

commensal isolates. 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 I_{CL3} in the virulent seropathotype C than in the non-virulent seropathotypes D and E. In addition, the presence of the PAI I_{CL3} element in most strains belonging to the virulent clonal group STEC-1 indicates that PAI I_{CL3} 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 I_{CL3} appeared to be a significant predictor of the virulent status.

In *C. rodentium*, a 53 kb GEI carries a PAI I_{CL3}-equivalent gene cluster

A PAI I_{CL3}-equivalent gene cluster was also detected in the genome of *C. rodentium*. The fact that in both *E. coli* and *C. rodentium* the PAI I_{CL3} element is extremely conserved at the nucleotide level and in genetic organization strongly supports the idea that the PAI I_{CL3} 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 I_{CL3} 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 I_{CL3} is borne on a 53 kb GEI termed GIp_{heV}-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 GIp_{heV}-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 GIp_{heV}-CR_{ICC168} acquired these regions through the horizontal transfer of mobile elements. In module I of GIp_{heV}-CR_{ICC168}, located upstream of the PAI I_{CL3} 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 GIp_{heV}-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 I_{CL3} (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 GIp_{heV}-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 I_{CL3} elements

The PAI I_{CL3} variant types of both *E. coli* CL3 and *C. rodentium* possess the full complement of PAI I_{CL3} 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 I_{CL3}) is borne on the GEI GIp_{heV}-CR_{ICC168}, the *E. coli* CL3 variant type (CL3-PAI I_{CL3}) 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 I_{CL3} 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 I_{CL3} from *C. rodentium* and *E. coli* CL3, may have played a role in the integration of PAI I_{CL3} into the host GEI.

The PAI I_{CL3} 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 I_{CL3} sequences in which all that remain are the genes located at the extremities of the genomic segment. In the deleted derivatives of PAI I_{CL3}, 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 I_{CL3} variant types EDL933-del/PAI I_{CL3}, CFT073-del/PAI I_{CL3}, and O42-del/PAI I_{CL3} (from EHEC/EPEC, ExPEC and EAEC O42, respectively), have undergone extensive deletions that removed almost all of the PAI I_{CL3}. ISs 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 I_{CL3} 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 I_{CL3} cluster provides a new example of gain and loss of genetic elements from one lineage to another.

PAI I_{CL3} 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 I_{CL3} 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 I_{CL3} 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 I_{CL3} variant types. One scenario is that a host GEI carrying the full complement of PAI I_{CL3} genes entered

the genome first, and then has been subject to IS-mediated deletion. In another scenario, PAI I_{CL3} 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 I_{CL3} 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 I_{CL3} 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 I_{CL3}-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 I_{CL3} genes

We noted an important distinction between *C. rodentium* and LEE-negative STEC which contain a complete PAI I_{CL3} gene cluster and other *E. coli* pathotypes which contain deleted PAI I_{CL3} gene clusters. A striking phylogenetic distribution of the PAI I_{CL3} 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 I_{CL3} genotype and seropathotype C demonstrates a striking phylogenetic link between intact PAI I_{CL3} 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 I_{CL3} 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 I_{CL3} supports this concept.

Conclusion

Taken together, our results indicate that (i) the PAI I_{CL3} 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 I_{CL3}-equivalent gene cluster is present in the genome of *C. rodentium*; (iii) the *C. rodentium* PAI I_{CL3} 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 I_{CL3}; (v) the PAI I_{CL3} 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 I_{CL3} genes in *E. coli*.

REFERENCES

- Al-Hasani, K., Rajakumar, K., Bulach, D., Robins-Browne, R., Adler, B. & Sakellaris, H. (2001). Genetic organization of the *she* pathogenicity island in *Shigella flexneri* 2a. *Microb Pathog* **30**, 1–8.
- Amavisit, P., Lightfoot, D., Browning, G. F. & Markham, P. F. (2003). Variation between pathogenic serovars within *Salmonella* pathogenicity islands. *J Bacteriol* **185**, 3624–3635.
- Bertin, Y., Martin, C., Girardeau, J. P., Pohl, P. & Contrepolis, M. (1998). Association of genes encoding P fimbriae, CS31A antigen and EAST 1 toxin among CNF1-producing *Escherichia coli* strains from cattle with septicemia and diarrhea. *FEMS Microbiol Lett* **162**, 235–239.
- Beutin, L. (2006). Emerging enterohaemorrhagic *Escherichia coli*, causes and effects of the rise of a human pathogen. *J Vet Med B Infect Dis Vet Public Health* **53**, 299–305.
- Boerlin, P., McEwen, S. A., Boerlin-Petzold, F., Wilson, J. B., Johnson, R. P. & Gyles, C. L. (1999). Associations between virulence factors of Shiga toxin-producing *Escherichia coli* and disease in humans. *J Clin Microbiol* **37**, 497–503.
- Bonnet, R., Souweine, B., Gauthier, G., Rich, C., Livrelli, V., Sirot, J., Joly, B. & Forestier, C. (1998). Non-O157:H7 Stx2-producing *Escherichia coli* strains associated with sporadic cases of hemolytic-uremic syndrome in adults. *J Clin Microbiol* **36**, 1777–1780.
- Chaudhuri, R. R., Khan, A. M. & Pallen, M. J. (2004). *coliBASE*: an online database for *Escherichia coli*, *Shigella* and *Salmonella* comparative genomics. *Nucleic Acids Res* **32**, D296–D299.
- Chouikha, I., Germon, P., Bree, A., Gilot, P., Moulin-Schouleur, M. & Schouler, C. (2006). A *selC*-associated genomic island of the extraintestinal avian pathogenic *Escherichia coli* strain BEN2908 is involved in carbohydrate uptake and virulence. *J Bacteriol* **188**, 977–987.
- Clarke, S. C., Haigh, R. D., Freestone, P. P. & Williams, P. H. (2003). Virulence of enteropathogenic *Escherichia coli*, a global pathogen. *Clin Microbiol Rev* **16**, 365–378.
- Contrepolis, M., Bertin, Y., Girardeau, J. P., Picard, B. & Goullet, P. (1993). Clonal relationships among bovine pathogenic *Escherichia coli* producing surface antigen CS31A. *FEMS Microbiol Lett* **106**, 217–222.
- Deng, W., Li, Y., Vallance, B. A. & Finlay, B. B. (2001). Locus of enterocyte effacement from *Citrobacter rodentium*: sequence analysis and evidence for horizontal transfer among attaching and effacing pathogens. *Infect Immun* **69**, 6323–6335.
- Deng, W., Puente, J. L., Gruenheid, S., Li, Y., Vallance, B. A., Vazquez, A., Barba, J., Ibarra, J. A., O'Donnell, P. & other authors (2004). Dissecting virulence: systematic and functional analyses of a pathogenicity island. *Proc Natl Acad Sci U S A* **101**, 3597–3602.
- de Sablet, T., Bertin, Y., Vareille, M., Girardeau, J. P., Garrivier, A., Gobert, A. P. & Martin, C. (2008). Differential expression of *stx2* variants in Shiga toxin-producing *Escherichia coli* belonging to seropathotypes A and C. *Microbiology* **154**, 176–186.
- Escobar-Paramo, P., Clermont, O., Blanc-Potard, A. B., Bui, H., Le Bouguenec, C. & Denamur, E. (2004). A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*. *Mol Biol Evol* **21**, 1085–1094.
- Frankel, G., Phillips, A. D., Rosenshine, I., Dougan, G., Kaper, J. B. & Knutton, S. (1998). Enteropathogenic and enterohaemorrhagic *Escherichia coli*: more subversive elements. *Mol Microbiol* **30**, 911–921.
- Gal-Mor, O. & Finlay, B. B. (2006). Pathogenicity islands: a molecular toolbox for bacterial virulence. *Cell Microbiol* **8**, 1707–1719.
- Garmendia, J., Frankel, G. & Crepin, V. F. (2005). Enteropathogenic and enterohaemorrhagic *Escherichia coli* infections: translocation, translocation, translocation. *Infect Immun* **73**, 2573–2585.
- Gilmour, M. W., Tracz, D. M., Andrysiak, A. K., Clark, C. G., Tyson, S., Severini, A. & Ng, L. K. (2006). Use of the *espZ* gene encoded in the locus of enterocyte effacement for molecular typing of Shiga toxin-producing *Escherichia coli*. *J Clin Microbiol* **44**, 449–458.
- Girardeau, J. P., Lalioui, L., Said, A. M., De Champs, C. & Le Bouguenec, C. (2003). Extended virulence genotype of pathogenic *Escherichia coli* isolates carrying the *afa-8* operon: evidence of similarities between isolates from humans and animals with extraintestinal infections. *J Clin Microbiol* **41**, 218–226.
- Girardeau, J. P., Dalmasso, A., Bertin, Y., Ducrot, C., Bord, S., Livrelli, V., Vernozy-Rozand, C. & Martin, C. (2005). Association of virulence genotype with phylogenetic background in comparison to different seropathotypes of Shiga toxin-producing *Escherichia coli* isolates. *J Clin Microbiol* **43**, 6098–6107.
- Hacker, J. & Carniel, E. (2001). Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep* **2**, 376–381.
- Hacker, J. & Kaper, J. B. (2000). Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol* **54**, 641–679.
- Jin, Q., Yuan, Z., Xu, J., Wang, Y., Shen, Y., Lu, W., Wang, J., Liu, H., Yang, J. & other authors (2002). Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res* **30**, 4432–4441.
- Johnson, R., Clarkes, R. C., Wilson, J. B., Read, S. C., Rhan, K., Renwick, S. A., Sandhu, K. A., Alves, D., Karmali, M. A. & other authors (1996). Growing concerns and recent outbreaks involving non-O157:H7 serotypes of verotoxigenic *Escherichia coli*. *J Food Prot* **59**, 1112–1122.
- Jores, J., Wagner, S., Rumer, L., Eichberg, J., Laturnus, C., Kirsch, P., Schierack, P., Tschape, H. & Wieler, L. H. (2005). Description of a 111-kb pathogenicity island (PAI) encoding various virulence features in the enterohaemorrhagic *E. coli* (EHEC) strain RW1374 (O103:H2) and detection of a similar PAI in other EHEC strains of serotype O103:H2. *Int J Med Microbiol* **294**, 417–425.
- Kaper, J. B., Mellies, J. L. & Nataro, J. (1999). Pathogenicity islands and other mobile genetic elements of diarrheagenic *Escherichia coli*. In *Pathogenicity Islands and Other Mobile Virulence Elements*, pp. 33–58. Edited by J. B. Kaper & J. Hacker. Washington, DC: American Society for Microbiology.
- Karmali, M. A., Mascarenhas, M., Shen, S., Ziebell, K., Johnson, S., Reid-Smith, R., Isaac-Renton, J., Clark, C., Rahn, K. & other authors (2003). Association of genomic O island 122 of *Escherichia coli* EDL 933 with verocytotoxin-producing *Escherichia coli* seropathotypes that are linked to epidemic and/or serious disease. *J Clin Microbiol* **41**, 4930–4940.
- Maurer, J., Jose, J. & Meyer, T. F. (1999). Characterization of the essential transport function of the AIDA-I autotransporter and evidence supporting structural predictions. *J Bacteriol* **181**, 7014–7020.
- Meier, T. I., Peery, R. B., McAllister, K. A. & Zhao, G. (2000). Era GTPase of *Escherichia coli*: binding to 16S rRNA and modulation of GTPase activity by RNA and carbohydrates. *Microbiology* **146**, 1071–1103.
- Moritz, R. L. & Welch, R. A. (2006). The *Escherichia coli* *argW-dsdCXA* genetic island is highly variable, and *E. coli* K1 strains commonly possess two copies of *dsdCXA*. *J Clin Microbiol* **44**, 4038–4048.

- Mundy, R., MacDonald, T. T., Dougan, G., Frankel, G. & Wiles, S. (2005). *Citrobacter rodentium* of mice and man. *Cell Microbiol* 7, 1697–1706.
- Paton, J. C. & Paton, A. W. (1998). Pathogenesis and diagnosis of Shiga toxin-producing *Escherichia coli* infections. *Clin Microbiol Rev* 11, 450–479.
- Perna, N. T., Plunkett, G. R., Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J. & other authors (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* 409, 529–533.
- Pradel, N., Livrelli, V., De Champs, C., Palcoux, J. B., Reynaud, A., Scheutz, F., Sirot, J., Joly, B. & Forestier, C. (2000). Prevalence and characterization of Shiga toxin-producing *Escherichia coli* isolated from cattle, food, and children during a one-year prospective study in France. *J Clin Microbiol* 38, 1023–1031.
- Reid, S. D., Herbelin, C. J., Bumbaugh, A. C., Selander, R. K. & Whittam, T. S. (2000). Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* 406, 64–67.
- Ren, C. P., Chaudhuri, R. R., Fivian, A., Bailey, C. M., Antonio, M., Barnes, W. M. & Pallen, M. J. (2004). The ETT2 gene cluster, encoding a second type III secretion system from *Escherichia coli*, is present in the majority of strains but has undergone widespread mutational attrition. *J Bacteriol* 186, 3547–3560.
- Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. & Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics* 16, 944–945.
- Schmidt, H. & Hensel, M. (2004). Pathogenicity islands in bacterial pathogenesis. *Clin Microbiol Rev* 17, 14–56.
- Schubert, S., Rakin, A., Karch, H., Carniel, E. & Heesemann, J. (1998). Prevalence of the “high-pathogenicity island” of *Yersinia* species among *Escherichia coli* strains that are pathogenic to humans. *Infect Immun* 66, 480–485.
- Shen, S., Mascarenhas, M., Rahn, K., Kaper, J. B. & Karmali, M. A. (2004). Evidence for a hybrid genomic island in verocytotoxin-producing *Escherichia coli* CL3 (serotype O113:H21) containing segments of EDL933 (serotype O157:H7) O islands 122 and 48. *Infect Immun* 72, 1496–1503.
- Tobe, T., Hayashi, T., Han, C. G., Schoolnik, G. K., Ohtsubo, E. & Sasakawa, C. (1999). Complete DNA sequence and structural analysis of the enteropathogenic *Escherichia coli* adherence factor plasmid. *Infect Immun* 67, 5455–5462.
- Vernozy-Rozand, C., Montet, M. P., Lequerrec, F., Serillon, E., Tilly, B., Bavai, C., Ray-Gueniot, S., Bouvet, J., Mazuy-Cruchaudet, C. & other authors (2002). Prevalence of verotoxin-producing *Escherichia coli* (VTEC) in slurry, farmyard manure and sewage sludge in France. *J Appl Microbiol* 93, 473–478.
- Welch, R. A., Burland, V., Plunkett, G., III, Redford, P., Roesch, P., Rasko, D., Buckles, E. L., Liou, S. R., Boutin, A. & other authors (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci U S A* 99, 17020–17024.
- WHO (1999). Zoonotic non-O157 Shiga toxin-producing *Escherichia coli* (STEC). In *Report of a WHO Scientific Working Group Meeting*. Berlin, Germany: World Health Organization.

Edited by: R. J. Lamont