Microbiology (2008), **154**, 571–583 DOI 10.1099/mic.0.2007/013086-0

A subset of mucosa-associated *Escherichia coli* isolates from patients with colon cancer, but not Crohn's disease, share pathogenicity islands with urinary pathogenic *E. coli*

Christina Bronowski,[1] Shirley L. Smith,[1,2] Kyoko Yokota,[1] John E. Corkill,[1] Helen M. Martin,[2] Barry J. Campbell,[2] Jonathan M. Rhodes,[2] C. Anthony Hart[1] and Craig Winstanley[1]

Correspondence
Craig Winstanley
C.Winstanley@liv.ac.uk

[1]Division of Medical Microbiology, School of Infection and Host Defence, University of Liverpool, Daulby Street, Liverpool L69 3GA, UK

[2]Division of Gastroenterology, School of Clinical Science, University of Liverpool, Crown Street, Liverpool L69 3BX, UK

Received 10 September 2007
Revised 17 October 2007
Accepted 2 November 2007

Adherent and invasive mucosa-associated *Escherichia coli* have been implicated in the pathogenesis of colon cancer and inflammatory bowel diseases. It has been reported that such isolates share features of extraintestinal *E. coli* (ExPEC) and particularly uropathogenic *E. coli* (UPEC). We used suppression subtractive hybridization (SSH) to subtract the genome of *E. coli* K-12 from that of a colon cancer mucosal *E. coli* isolate. Of the subtracted sequences, 53 % were present in the genomes of one or more of three sequenced UPEC strains but absent from the genome of an enterohaemorrhagic *E. coli* (EHEC) strain. Of the subtracted sequences, 80 % matched at least one UPEC genome, whereas only 4 % were absent from the UPEC genomes but present in the genome of the EHEC strain. A further genomic subtraction against the UPEC strain 536 enriched for sequences matching mobile genetic elements, other ExPEC strains, and other UPEC strains or commensals, rather than strains associated with gastrointestinal disease. We analysed the distribution of selected subtracted sequences and UPEC-associated pathogenicity islands (PAIs) amongst a panel of mucosa-associated *E. coli* isolated from colonoscopic biopsies of patients with colon cancer, patients with Crohn's disease and controls. This enabled us to identify a group of isolates from colon cancer (30–40 %) carrying multiple genes previously categorized as UPEC-specific and implicated in virulence.

INTRODUCTION

There is increasing recognition that mucosa-associated bacteria may play an important role in the pathogenesis of colon cancer (Mager, 2006) and inflammatory bowel diseases (IBDs), including Crohn's disease. Several independent groups have identified mucosa-associated *Escherichia coli* in Crohn's disease (Martin *et al.*, 2004; Darfeuille-Michaud *et al.*, 1998; Swidsinski *et al.*, 2002; Ryan *et al.*, 2004; Mylonaki *et al.*, 2005; Kotlowski *et al.*, 2007). Immunohistochemical studies have shown *E. coli* antigens within tissue macrophages in Crohn's disease tissue (Liu *et al.*, 1995), and *E. coli* DNA can be isolated from the majority of Crohn's disease granulomas in tissue sections (Ryan *et al.*, 2004). Similarly, in a prospective study of patients screened for colorectal cancer, intracellular bacteria, notably *E. coli*, were isolated from both tumour and distant histologically normal mucosa (Swidsinski *et al.*, 1998).

In a previous study, we confirmed that Crohn's disease and colon cancer isolates of *E. coli*, particularly those lying beneath the colonic mucous layer, were more likely to adhere to and invade epithelial cell lines, and we showed that this property correlated well with their ability to agglutinate human red blood cells, regardless of blood group (Martin *et al.*, 2004). In both Crohn's disease and colon cancer, we (Martin *et al.*, 2004) and others (Swidsinski *et al.*, 1998) have found mucosa-associated *E. coli* that resisted gentamicin treatment of the mucosal

Abbreviations: APEC, avian pathogenic *E. coli*; CNF1, cytotoxic necrotizing factor 1; EHEC, enterohaemorrhagic *E. coli*; ExPEC, extraintestinal *E. coli*; IBD, inflammatory bowel disease; PAI, pathogenicity island; SSH, suppression subtractive hybridization; UPEC, uropathogenic *E. coli*.

Supplementary tables showing the *E. coli* isolates used in this study and distribution of sequences according to PCR assays, the oligonucleotide primers used for PCR amplification, and the percentage distribution of PAIs amongst *E. coli* collections are available with the online version of this paper.

samples and were therefore presumed to be intracellular. It has been demonstrated that Crohn's disease-associated *E. coli* adhere to and invade epithelial cell lines *in vitro* (Darfeuille-Michaud *et al.*, 2004; Martin *et al.*, 2004) and replicate inside macrophage phagolysosomes (Bringer *et al.*, 2006), with resulting giant cell formation (Meconi *et al.*, 2007). It has been reported (Kotlowski *et al.*, 2007) that these adherent and invasive *E. coli* (AIEC) belong to phylogenetic groups B2 and D (Clermont *et al.*, 2000), typical of extraintestinal pathogenic *E. coli* (ExPEC), and that they exhibit a diffusely adherent pattern of adherence with Hep-2 cells, thus having many of the features of uropathogenic *E. coli* (UPEC). Like others, we have also shown that they lack the pathogenicity genes associated with *E. coli* that are typically pathogenic in the gut (Martin *et al.*, 2004). Recently, it has been shown that AIEC of phylogenetic groups B2 and D can cause granulomatous colitis in boxer dogs (Simpson *et al.*, 2006).

Thus far, genetic characterization of *E. coli* strains that occupy the submucosal niches in colon cancer and Crohn's disease has been very limited. Some progress has been made with the identification of genes responsible for adherence to and invasion of epithelial cells, but these studies have only been conducted in one ileal isolate from Crohn's disease, LF82 (Barnich *et al.*, 2004; Rolhion *et al.*, 2005, 2007).

In this study we describe the use of suppression subtractive hybridization (SSH) to characterize the genomic content of a mucosal *E. coli* colon cancer isolate and further analyse the distribution of subtracted sequences and UPEC-associated pathogenicity islands (PAIs) amongst a panel of mucosa-associated *E. coli* isolated from colonoscopic biopsies of patients with Crohn's disease and colon cancer, and of controls (patients with irritable bowel syndrome and sporadic polyps).

METHODS

Bacterial strains used in this study. Most of the mucosal isolates of *E. coli* used in this study (Supplementary Table S1) were described in the study of Martin *et al.* (2004). The bacteria were obtained after the overlying mucous layer had been removed by DTT treatment from biopsy samples taken from the sigmoid colon (Martin *et al.*, 2004). The ileal isolate *E. coli* LF82 was a gift from Dr Darfeuille-Michaud, Laboratoire de Bacteriologie, Université d'Auvergne, Clermont-Ferrand, France. The K-12 derivative *E. coli* JM109 was used to obtain driver DNA for subtractive hybridization. The UPEC strains *E. coli* J96 and 536 (gift of Eva Moreno Pujol, Hospital Universitari Vall d'Hebron, Barcelona) were used as controls in PCR assays and, in the case of strain 536, for further subtraction work. All bacteria were maintained by growth on nutrient agar at 37 °C.

Construction and screening of subtraction libraries. Genomic DNA was isolated from *E. coli* strains using the Wizard Genomic DNA Purification kit (Promega). SSH was carried out using the Clontech PCR-Select Bacterial Genome Subtraction kit as recommended by the supplier. In both hybridizations, DNA from *E. coli* HM229 was used as tester. DNA from *E. coli* strains JM109 and 536 was used as the driver in the first and second hybridizations, respectively. All DNAs were digested with *Rsa*I. PCR amplicons obtained following SSH were cloned into pGEM-T (Invitrogen). The subtraction libraries of *Rsa*I fragments thus constructed were screened by sequencing of plasmid DNA extracted from individual clones using M13 forward and reverse vector primers (Cogenics Lark). In order to identify genuinely subtracted sequences, BLASTN searches targeting the genomes of *E. coli* K-12 and 536 were conducted. Sequences sharing >90 % identity with the driver genome were omitted from further study. Sequences sharing <90 % identity with the genome of the relevant driver strain were further analysed using BLASTN and BLASTX searches of the general database. Similar BLASTN searches were used to determine the presence or absence of SSH sequences from the genomes of the UPEC strains CFT073 (Welch *et al.*, 2002), 536 (Brzuszkiewicz *et al.*, 2006) and UTI89 (Chen *et al.*, 2006), and the enterohaemorrhagic *E. coli* (EHEC) strain EDL933 (Perna *et al.*, 2001). All searches were done using the site http://www.ncbi.nlm.nih.gov.

PCR amplification screening of strains. Oligonucleotide primers (Sigma-Genosys) for PCR amplifications are listed in Supplementary Table S2 along with the annealing temperatures used. DNA for PCR amplification was prepared by boiling a suspension of a few colonies in 5 % 200 µl Chelex-100 (Bio-Rad) for 5 min. After centrifugation, the top 150 µl was removed and stored at −20 °C. For PCR amplification, typically, 1 µl DNA was used directly in 25 µl volumes containing 1.25 U *Taq* DNA polymerase (Promega), 1 × TaqMaster (Helena BioSciences), 300 nM each primer, 1 × Taq buffer, 2.5 mM $MgCl_2$ and 100 µM nucleotides (dATP, dCTP, dGTP, dTTP). Amplifications were carried out in an Eppendorf MasterCycler thermal cycler for 30 cycles consisting of 95 °C (1 min), annealing temperature (1 min) and 72 °C (2 min), with an additional extension time at 72 °C (10 min) following completion of the 30 cycles.

PCR assays for UPEC-specific PAIs have been described elsewhere (Sabaté *et al.*, 2006; Johnson & Stell, 2000). In this study we used an amended version of the multiplex PCR assays described by Sabaté *et al.* (2006), whereby multiplex PCR B was split into two separate PCR assays consisting of assays for (1) PAI II_{J96} and PAI I_{536} and (2) PAI II_{536}, PAI I_{CFT073} and PAI I_{J96}. Multiplex PCR A was used to assay for PAI III_{536}, PAI IV_{536} and PAI II_{CFT073}. Amplicon sizes and annealing temperatures are indicated in Supplementary Table S2. For the PAI II_{J96} and PAI I_{536} PCR assays, the extension time was increased to 3 min.

Oligonucleotide primers used in PCR assays for PAI VI_{536} were designed from SSH sequences obtained in this study. PCR assays for 229-4 and ECP1966 were multiplexed together. Phylogenetic groups were determined using assays published elsewhere (Clermont *et al.*, 2000).

The SSH sequence 229-7 was extended by inverse PCR amplification. Genomic DNA from *E. coli* HM229 was digested with *Sal*I, ligated using T4 DNA ligase and subjected to PCR amplification using the primers 229-7up (5′-TGCGCATGATTACCAGAC-3′) and 229-7dn (5′-CGGTTCCTGATGTGTGCTA-3′). The resulting amplicon was sequenced using the same oligonucleotides as primers (Cogenics Lark). Further amplification and sequencing using the primers 5′-GGGCGATTTTTAGAAAGG-3′ and 5′-TGCGCGTCATCAGCTTTC-3′ was used to fill in the remaining gap and enable a complete ORF to be obtained.

RESULTS

Identification of sequences present in the genome of colon cancer isolate *E. coli* HM229 but absent from the genome of *E. coli* JM109

The colon cancer isolate HM229 was chosen for closer analysis as one of only three mucosal isolates from our

previous study that carry the *cnf1* gene. In addition, it has haemagglutinating activity against neuraminidase-treated group O red blood cells, adheres to and invades the I407 intestinal cell line and adheres to the HT29 intestinal cell line (Martin *et al.*, 2004).

The sequences from a total of 116 SSH clones were obtained. Of the sequences, 85 (77 %) were genuinely subtracted from the genome of *E. coli* K-12, but 10 were repeated, leaving an output of 75 different SSH sequences. Of these, only one gave no significant match when used in a BLASTX search of the database. A summary of all SSH sequences obtained, organized according to their putative function by BLASTX match, is shown in Table 1.

Of the 75 SSH sequences obtained, 20 (27 %) gave a best BLASTX match with hypothetical proteins or proteins of unknown function, 15 (20 %) matched putative membrane proteins, six (8 %) matched proteins implicated in adherence or haemagglutination, six (8 %) matched proteins with putative regulatory functions, 10 (7.5 %) were associated with mobile genetic elements and three (4 %) were O antigen-related (Table 1). Two of the SSH sequences (229-G3 and 229-C6) matched the known virulence-related proteins HlyD (involved in the transport of haemolysin A) and cytotoxic necrotizing factor 1 (CNF1), respectively. In the vast majority of cases, the BLASTX matches were with sequences that shared high levels of identity with sequences from the genomes of *E. coli*, *Shigella*, *Yersinia* or *Salmonella*. Notable exceptions to this were SSH sequence 229-H3, sharing only 40 % identity with an O antigen-related protein of *Vibrio cholerae*, and SSH sequence 229-7, which matched two adjacent genes from *E. coli* O157 : H7, but shared only 63 % BLASTX identity with one of these proteins, a putative adhesin/invasion-related protein (Table 1).

Using inverse PCR amplification, we extended this sequence to obtain the whole of the putative gene (submitted to GenBank; accession no. EU046568). The predicted 441 aa protein shared 64 % identity with an IHP1-like protein from a Shiga toxin-producing *E. coli* (AAO83376).

Four of the SSH sequences matched hypothetical proteins encoded by an unpublished putative genomic island (AGI-1) of the avian pathogenic *E. coli* (APEC) strain BEN2908.

Using BLASTN searches, the SSH sequences were screened for their presence in three genome-sequenced UPEC strains (CFT073, 536 and UTI89) and a genome-sequenced EHEC strain (EDL933) (Table 1). Of the SSH sequences, 53 % were present in the genomes of one or more UPEC strains but absent from the EHEC strain. A further 27 % were present in the genomes of one or more UPEC strains and the EHEC strain. Sixteen per cent were absent from all four genomes. Only 4 % of SSH sequences were present in the genome of the EHEC strain but absent from the genomes of all UPEC strains. Amongst the SSH sequences were seven with BLASTX matches to putative proteins encoded by one of the 131 UPEC-specific genes identified by Lloyd *et al.* (2007).

The SSH sequences included sequences that matched parts of the UPEC strain 536 PAIs PAI I_{536}, PAI III_{536}, PAI IV_{536}, PAI V_{536} and PAI VI_{536} (Dobrindt *et al.*, 2002; Brzuszkiewicz *et al.*, 2006; Hochhut *et al.*, 2006) and the UPEC strain CFT073 PAIs PAI I_{CFT073} (Guyer *et al.*, 1998) and PAI II_{CFT073} (Rasko *et al.*, 2001).

Identification of sequences present in the genome of *E. coli* HM229 but absent from the genome of UPEC strain 536

Our findings from the initial subtraction suggested that the accessory genome of *E. coli* HM229 shares more in common with UPEC than with intestinal pathogenic *E. coli*. In an attempt to enrich for genes within the accessory genome that might be specific to bacteria occupying mucosal regions of colon cancer or Crohn's disease, we carried out a second subtraction using the genome-sequenced UPEC strain 536 as the driver.

A total of 100 sequences were obtained, of which 62 were genuinely subtracted from the genome of strain *E. coli* 536. Four sequences were repeated, leaving a total of 58 different subtracted sequences. Of these 58 SSH sequences, 27 matched neither the UPEC strains UTI89 or CFT073 nor the EHEC strain EDL933 (Table 2). Of these 27, eight had a best BLASTX match against *E. coli* APEC O1, two had a best BLAST match against *E. coli* K-12, and five mainly O antigen/capsule-related SSH sequences had a best BLASTX match against other *E. coli* strains. A further eight SSH sequences matched outside the genus *Escherichia*. These included a putative SecA-related protein from *V. cholerae*.

Distribution of UPEC PAIs amongst isolates from colon cancer and Crohn's disease

To further assess how closely related our mucosal *E. coli* isolates were to UPEC *E. coli* we used published PCR assays to screen a panel of isolates for the major reported UPEC PAIs PAI I_{536}, PAI II_{536}, PAI III_{536}, PAI IV_{536}, PAI I_{CFT073}, PAI II_{CFT073}, PAI I_{J96} and PAI II_{J96}. Furthermore, based on SSH sequences, we designed PCR assays for four genes identified as present within PAI VI_{536}. These were designed to target either edge of the PAI [ECP1966 and ECP2038 (229-4)] and the internal ORFs ECP1978 (229-E6) and ECP2007 (229-E4). The results of the PCR assays are presented in Supplementary Table S1 and summarized in Table 3. Notably, there was a group of isolates from colon cancer (30–40 %) that tested PCR-positive for the UPEC PAIs PAI II_{J96}, PAI I_{536}, PAI II_{536} and PAI II_{CFT073}. In contrast, one of six control isolates and no Crohn's disease isolates were PCR-positive for these PAIs, with the exception of one Crohn's disease and two control isolates testing PCR-positive for PAI II_{CFT073}. The PCR assays for four different sequences from PAI VI_{536} indicated considerable variability in the carriage or nucleotide sequence of

Table 1. Summary of SSH using the genome of *E. coli* HM229 as tester and the genome of *E. coli* JM109 as driver

SSH sequence	Length (bp)	Best BLASTX match/comments [accession no. of best match]	ID (%)	Length (aa)*	*E*-value	CFT073 (UPEC)	UTI89 (UPEC)	536 (UPEC)	EDL 933 (EHEC)
		Bacteriophage-related							
229-10	356	Hypothetical protein PhiV10p15 (phage phiV10) [AAZ95908]	100	118	2e-62	None	None	None	None
229-21	385	Hypothetical protein PhiV10p18 (phage phiV10) [AAZ95911]	99	128	3e-70	None	None	None	None
229-D9	>225	Hypothetical protein PhiV10p19 (phage phiV10) [AAZ95912]	93	43	1e-15	None	None	None	None
229-H7	674	Hypothetical protein PhiV10p24 (phage phiV10) [AAZ95916];	97	85	1e-41	None	None	None	None
		Eliminase (phage tail fibre protein) (*E. coli* K5) [CAA65353]	100	45	4e-18	None	None	None	None
229-6	481	Bacteriophage P4 integrase (several *E. coli*) [CAC39282]	99	159	3e-85	c4491	*C2634*	None	Z1652
229-26	594	Bacteriophage CP4/P4 integrase (several *E. coli*) [AAL51003]	97	197	8e-108	c3556	*C4878*	ECP2962 (PAI V)	Z4313
229-E5	487	Putative phage integrase (*Shigella flexneri*) [AAP16001]	100	65	5e-33	*c5371*	*C5085*	None	None
		Plasmid/transposon/insertion sequence (IS)-related							
229-G6	533	RepA4, plasmid-associated (several *E. coli*) [CAA23643]	98	78	2e-31	None	None	None	*L7009*
229-H9	281	Putative transposase (*E. coli* O103 : H2) [CAI43808]	96	92	8e-46	*c1219*	None	*ECP0279*	None
229-C9	425	IS*629* ORF2 (*S. flexneri* 2a strain 301 and others) [AAL72366]	98	63	2e-42	c5178 (PAI II)	None	None	Z3295
		Adhesion/haemagglutination-related							
229-7	734	Hypothetical protein (*E. coli* O157 : H7 and others) [BAB33972];	100	37	2e-13	None	C0522	None	Z0640
		Adhesin/invasin-like protein (*E. coli* O157 : H7 and others) [BAB33971]	63	61	6e-12	None	*C0521*	None	*Z0639*
229-14	582	Type 1 fimbriae regulatory protein FimB (*E. coli* CFT0730 and others) [AAN78911]	100	191	4e-108	c0430	C0338	ECP0382	Z0395
229-H2	300	PapD P-pilus chaperone (*E. coli* CFT073 and several others) [AAN83607]	100	99	2e-51	c5185 (PAI II)	C4892	ECP4538‡	None
229-29	285	Large exoprotein involved in haem utilization or adhesion (*E. coli* UTI89, F11 and CFT073) [ABE10349]	97	95	2e-44	c0345	C4946	ECP4580§	None
229-B5	746	Large exoprotein involved in haem utilization or adhesion (*E. coli* UTI89 and several others) [ABE10349]	100	248	3e-120	c0345	C4946	ECP4580§	None
229-B10	649	Putative adhesin/autotransporter, EaeH (*E. coli* CFT073 and several others) [AAN78896]	99	216	3e-81	c0415	C0321	ECP0367	Z0375
		Regulation							
229-3	306	DNA-binding protein H-NS (*E. coli* CFT073 and F11) [AAN80870]	100	102	4e-51	c2411	*C1434*	ECP1927 (PAI IV)	*Z2013*
229-A3	370	LemA family protein (*E. coli* CFT073 and UTI89) [AAN81785]	100	52	1e-23	c3337	C3142	ECP2756§	None
229-H12	301	Putative signal transduction histidine kinase (*E. coli* UTI89, CFT073, F11 and others) [ABE10335]	99	100	9e-50	c3564 (PAI I)	C4932	ECP4563§	None
229-23	745	Regulatory protein (*E. coli* CFT073 and others) [AAN83624]	98	78	1e-38	c5202	None	*ECP2977*	None
		Phosphoglycerate transporter protein PgtP (*E. coli* CFT073 and others) [AAN83623]	100	67	1e-29	c5201	*C2520*	*ECP2978*	*Z3498*
229-B1	616	Transcriptional regulator YfjR (*E. coli* CFT073 and others) [AAN83000]	99	162	2e-89	c5160 (PAI II)	C4986	ECP4627§	None
229-D6	563	Putative transcriptional repressor (*E. coli* CFT073 and UTI89) [AAN80226]	100	107	2e-53	c1760	C1562	*ECP1343*	Z2511
		Membrane proteins							
229-25	673	Outer-membrane protein, RatA homologue (*E. coli* CFT073) [AAN81479]	100	223	4e-117	c3029	C2828	ECP2512‡	None
229-F9	631	Outer-membrane protein, RatA homologue (*E. coli* CFT073) [AAN81479]	100	210	3e-111	c3029	C2828	ECP2512‡	None
229-28	547	Putative membrane protein (*E. coli* UTI89 and F11) [ABE10586]	100	140	4e-66	None	P010	None	None
229-A6	485	Unknown (uncultured bacterium) [AAO59951];	96	161	2e-89	None	None	None	None

Table 1. cont.

SSH sequence	Length (bp)	Best BLASTX match/comments [accession no. of best match]	ID (%)	Length (aa)*	*E*-value	CFT073 (UPEC)	UTI89 (UPEC)	536 (UPEC)	EDL 933 (EHEC)
		Outer-membrane protein (*Yersinia mollaretii*) [ZP_00827519]	82	161	1e-77	None	None	None	None
229-B8	559	Putative glycoporin (*E. coli* CFT073 and several others) [AAN83277]	99	131	7e-65	c4849	C4483	ECP4110	Z5444
229-D4	336	Putative transmembrane transport protein (*E. coli* CFT073 and UTI89) [AAN83448]	100	112	1e-50	c5022	C4621	None	None
229-D7	711	Hypothetical, putative membrane protein YiaM (*E. coli* CFT073 and UTI89) [AAN82835];	97	96	3e-39	c4399	C4120	ECP3682	None
		Hypothetical protein (*E. coli* CFT073, UTI89 and F11) [AAN82834]	98	77	2e-37	c4398	C4119	ECP3681	None
229-E2	312	Putative outer-membrane colicin Js sensitive receptor protein (*E. coli* UTI89 and others, *Shigella*) [ABE10605]	100	103	7e-55	None	P029	None	None
229-E12	387	Putative outer-membrane channel protein (*E. coli* CFT073 and UTI89) [AAN80231]	100	129	1e-63	c1765	C1565	ECP1346	None
229-G8	725	Outer-membrane protein, SinI-like (*E. coli* UTI89, CFT073 and F11) [ABE08289]	99	163	2e-88	c3030	C2829	ECP2513	None
229-H4	481	Hypothetical ABC transporter permease protein YddQ (*E. coli* CFT073, UTI89 and F11) [AAN83505]	100	102	2e-41	c5079†	C4672	ECP4316‡	None
229-H11	474	Putative ribose ABC transporter (*E. coli* CFT073, UTI89 and F11) [AAN82457]	99	139	2e-69	c4017†	C3694	ECP3346‡	None
229-A8	462	Putative iron compound receptor (*E. coli* CFT073, UTI89, F11 and O157:H7) [AAN82219]	99	154	8e-86	c3775	C3463	ECP3121	Z4386
229-B11	>670	Secondary glycine betaine transporter BetU (several *E. coli* and *Shigella boydii*) [AAQ10261]	100	223	1e-106	None	None	None	None
229-C2	>459	Hypothetical type II secretion protein GspE (*E. coli* UTI89 and several others) [ABE08831]	100	72	2e-33	None	C3384	ECP3045§	None
		O antigen-related							
229-9	567	FnlC, O4 O antigen gene cluster (several *E. coli*) [AAT85657]	99	154	1e-81	None	None	None	None
229-H3	368	Putative *O*-acetyltransferase WavO (*V. cholerae*) [AAL77347]	40	121	1e-15	None	None	None	None
229-27	771	Glucose-1-phosphate thymidylyltransferase RmlA (*E. coli* VW187 and *Salmonella enterica*) [AAC63614] O7-related	88	139	2e-66	None	C2310	None	None
		Virulence-related							
229-C6	211	CNF1 (*E. coli* UTI189 and others) [ABE10324]	100	70	2e-33	None	C4921	None	None
229-G3	453	Haemolysin D, HlyD (*E. coli* CFT073 and several others) [AAN82022]	100	150	6e-76	c3574 (PAI I)	C4924	ECP3829 (PAI I)	L7050
		Other proteins/enzymes							
229-A7	>783	Hydrolases of the alpha/beta superfamily (*Yersinia pestis*) [AAS62069]	85	101	2e-61	*c1938*	None	None	None
229-A11	465	Arylsulfate sulfotransferase (*E. coli* UTI89, CFT073 and F11) [ABE08920]	99	154	2e-87	c3785†	C3473	ECP3130	None
229-B2	480	DEAD box helicase-related protein (*Salmonella* Typhi) [CAD06969]	77	159	1e-61	None	None	None	*Z5898*
229-C12	380	Partitioning protein A (*Y. pestis*) [CAG27535]	96	126	9e-63	None	None	None	None
229-D1	381	Putative dihydrodipicolinate synthase (*E. coli* CFT073, UTI89 and F11) [AAN79234]	100	109	7e-55	c0761†	C0679	ECP0695‡	None
229-E4	495	Putative phosphotriesterase-related protein (*E. coli* CFT073 and several others) [AAN80951]	100	134	7e-73	c2495	C2246	ECP2007 (PAI VI)	None
229-E6	609	Putative polyketide synthase (*E. coli* CFT073, UTI89 and F11) [AAN80927]	100	203	8e-106	c2468	C2221	ECP1978 (PAI VI)	None
229-E9	303	Putative effector of murein hydrolase (*E. coli* CFT073, UTI89 and F11) [AAN83487]	98	76	5e-25	c5061†	C4656	ECP4298‡	None
229-F7	485	Hypothetical oxidoreductase YdfI (*E. coli* UTI89, CFT073 and F11) [ABE08884]	100	124	7e-54	c3751	C3437	ECP3100	None
229-G7	315	Superfamily II helicase (*Shigella dysenteriae* 1012) [ZP_00921822]	100	104	1e-43	None	None	None	*Z1843*

Table 1. cont.

SSH sequence	Length (bp)	Best BLASTX match/comments [accession no. of best match]	ID (%)	Length (aa)*	*E*-value	CFT073 (UPEC)	UTI89 (UPEC)	536 (UPEC)	EDL 933 (EHEC)
229-G11	340	Carbamate kinase YahI (*E. coli* CFT073, UTI89 and F11) [AAN83197]	100	113	7e-60	c4764†	C4405	ECP4034‡	*Z4213*
229-C10	>456	Putative amidase (*E. coli* CFT073 and UTI89) [AAN80916]	99	152	2e-83	c2457	C2212	ECP1969 (PAI VI)	None
		Hypothetical/unknown proteins							
229-1	576	Conserved hypothetical protein (*E. coli* CFT073 and others) [AAN83507] (partial K-12 match over last 49 bp)	100	52	1e-19	c5081†	C4674	ECP4318‡	None
229-4	331	Conserved hypothetical protein (*E. coli* CFT073 and others) [AAN78774]	100	85	1e-40	c0286	C2268	ECP2038 (PAI VI)	Z1652
229-11	458	Hypothetical protein Aec15, AGI-1 (*E. coli* BEN2908 and others) [AAQ96709]	100	152	1e-83	*c1888*	C0253	ECP0240§	*Z0267*
229-12	384	Hypothetical protein Aec7, AGI-1 (*E. coli* BEN2908 and UTI89) [AAQ96701];	100	72	2e-32	None	C0258	ECP0248	None
		Hypothetical protein Aec8, AGI-1 (*E. coli* BEN2908 and UTI89) [AAQ96702]	96	32	9e-10	None	C0257	ECP0247§	None
229-A10	259	Hypothetical protein (*E. coli* CFT073 and UTI89) [AAN80968]	100	41	2e-17	c2513	C2260	ECP2028 (PAI VI)	None
229-A12	755	Hypothetical protein SDY_4475 (*S. dysenteriae* Sd197) [ABB64351]	93	30	2e-7	None	None	one	None
229-B6	551	Hypothetical protein (*E. coli* CFT073 and several others) [AAN80956]	98	50	9e-23	c2500	C2250	*ECP2016* (PAI VI)	None
229-B9	477	Hypothetical protein (*Sal.* Typhi) [AAO70507]	96	32	3e-10	None	None	None	None
229-C3	>731	Hypothetical protein (*E. coli* UTI89 and several others) [ABE08848]	99	165	3e-90	c3716	C3401	ECP3064	None
229-C4	736	Hypothetical protein Aec24 (*E. coli* UTI89 and others) [ABE05748]	97	174	4e-92	None	C0245	ECP0232	*Z0257*
229-D8	733	Hypothetical protein (*E. coli* F11) [ZP_00726067]	99	156	1e-83	c4517	C4942	ECP4576§	None
229-E3	479	Hypothetical protein (*E. coli* CFT073, UTI89 and O157 : H7) [AAN83171]	100	138	6e-81	c4738	C4380	ECP4011	Z5334
229-E11	446	Uncharacterized conserved protein (*E. coli* F11) [ZP_00725602]	87	131	1e-61	None	None	*ECP2816*‡	None
229-F4	239	Hypothetical protein (Aec30) (*E. coli* UTI89 and F11) [ABE05742]	100	79	7e-36	None	C0239	ECP0226	*Z0250*
229-F8	478	Hypothetical protein YkfF (*E. coli* CFT073 and UTI89) [AAN83003]	98	77	4e-39	c4569	C4989	ECP3853§	*L7085*
229-F12	275	Hypothetical protein (*E. coli* UTI89 and F11) [ABE06867]	98	71	8e-35	*c1649*	C1385	ECP1241	Z1963
229-G10	656	Hypothetical protein (*E. coli* CFT073) [AAN82453]	100	116	5e-62	c4013	C3690	ECP3342	None
229-H1	294	Hypothetical protein YiaN (*E. coli* CFT073 and UTI89) [AAN82836]	100	62	3e-27	c4400	C4121	ECP3683	None
229-H6	377	Hypothetical protein (*E. coli* CFT073 and UTI89) [AAN82015]	100	125	2e-71	c3567 (PAI I)	C4929	ECP4560‡	None
229-B4	>222	Hypothetical protein (*E. coli* 536 only) [ABG70017]	98	68	9e-31	None	None	ECP2018 (PAI VI)	None
		No significant hits							
229-G1	545	No significant hits				None	None	None	None

*Length refers to the length in amino acid residues of the BLASTX match; †UPEC-specific as defined by Lloyd *et al.* (2007); UPEC-specific (‡) or 536-specific (§) as defined by Brzuszkiewicz *et al.* (2006). Thus, UPEC-specific indicates present in CFT073 and 536 but absent from two EHEC genomes and K-12; 536-specific indicates absent from all other genome-sequenced *E. coli* strains at the time of publication. ORFs located within strain CFT073 or strain 536 PAIs are indicated by showing the island in parentheses. Italic type indicates a considerably weaker match than the best BLASTX match; matches <50 % identity (ID) to the genomes of CFT073, UTI89 or EDL933 were discounted.

Table 2. Summary of SSH using the genome of *E. coli* HM229 as tester and the genome of *E. coli* 536 as driver

SSH sequence	Length (bp)	Best BLASTX match/comments [accession no. of best match]	ID (%)	Length (aa)*	*E*-value	CFT073 (UPEC)	UTI89 (UPEC)	EDL 933 (EHEC)
		Bacteriophage-related						
229/536-F6	265	Hypothetical protein APECO1_4043 (*E. coli* APEC O1) [ABJ01909]; also matches phage PhiV10p09	98	88	5e-31	None	None	None
229/536-F3	389	Conserved hypothetical protein (*E. coli* APEC O1) [ABJ01903]; also matches phage PhiV10p15	98	129	5e-67	None	None	None
229/536-A6	>418	Hypothetical protein APECO1_4053 (*E. coli* APEC O1) [ABJ01899]; also matches phage PhiV10p19	96	82	9e-37	None	None	None
229/536-G10	278	Hypothetical protein (*E. coli* APEC O1) [ABJ01893]; also matches phage PhiV10p27	98	67	9e-24	None	None	None
229/536-A10	253	Integrase-like protein (*E. coli* APEC O1) [ABJ01932]; also matches phage PhiV10p31	100	83	1e-38	None	None	None
229/536-A2	591	ShiA-like protein (*E. coli* APEC O1) [ABJ03682];	98	73	2e-35	c3557	None	None
		Putative P4-type integrase (*E. coli* APEC O1) [ABJ03683]	100	47	6e-21	c3556	C4878	Z4313
229/536-B5	724	KpLE2 phage-like element; predicted dehydratase (*E. coli* K-12) [AAC77253]	100	146	2e-72	None	None	None
229/536-C7	560	KpLE2 phage-like element, iron-dicitrate transporter subunit (*E. coli* K-12 and others) [AAC77243]	99	186	4e-93	*c0675*	*C0590*	*Z0729*
229/536-F5	563	KpLE2 phage-like element; predicted DNA-binding transcriptional regulator (*E. coli* K-12) [AAC77255];	100	140	9e-77	None	None	None
		Dihydrodipicolinate synthase homologue *yjhH* (*E. coli* K-12) [AAA97194]	100	63	1e-29	None	None	None
229/536-G6	>470	Putative phage small subunit terminase (*Bordetella avium* 197N) [CAJ48920]	63	36	7e-05	None	None	None
229/536-H1	394	Hypothetical bacteriophage protein (*S. boydii* Sb227) [ABB68129]	98	67	4e-30	None	None	None
		Plasmid/transposon/IS-related						
229/536-A3	>722	Putative ATP-binding F pilus assembly protein, TraH, plasmid-related (*E. coli* UTI89 and others) [ABE10708]	100	238	7e-125	None	P132	None
229/536-B2	477	Conjugal transfer protein TrbI (plasmid R100, several *E. coli*) [BAA78865]	100	105	8e-54	None	P118	None
229/536-C6	352	Conjugal transfer protein TraN (plasmid R100; *E. coli*) [BAA78870]	100	117	4e-57	None	P123	None
229/536-H9	207	Mating contact stablization protein TraN (*E. coli* UTI89) [ABE10699]	100	69	2e-35	None	P123	None
229/536-G2	600	Conjugal transfer protein E, plasmid R100 (several *E. coli*) [BAA78853]	100	180	2e-91	None	P108	None
229/536-A11	422	SrnB, plasmid-related, N terminus (*E. coli* APEC O1) [ABD51633]	88	18	0.21	None	None	None
229/536-3	213	AadA, streptomycin/spectinomycin adenyltransferase, plasmid-related (*E. coli* APEC O1) [ABF67771]	100	70	1e-31	None	None	None
229/536-F1	440	Retron-type reverse transcriptase (*E. coli* UTI89 and others) [ABE10613]	99	146	3e-77	None	P037	None
229/536-B3	>611	IS1 ORF2 (*S. flexneri* 2a strain 301) [AAN44567]	98	80	6e-38	None	C4369	Z1192
229/536-C8	346	Putative transposase (*E. coli* CFT073) [AE016767]	100	114	9e-62	c3803	None	*Z5815*
229/536-D1	>650	Putative transposase (Aec65 from APEC AGI-1) (several *E. coli*) [AAW51748]	77	48	4e-15	None	None	Z5089
229/536-E8	394	TnpA/IS*26* transposase [ABF48381];	100	67	6e-31	None	P006	None
		Esterase (*Photobacterium damselae* subsp. *piscicida*) [BAF38031]; also matches esterases in *Yersinia* and *Klebsiella*	100	46	9e-21	*c0464*	*C0375*	*Z0455*
229/536-H10	396	Transposase (*Y. enterocolitica* subsp. *enterocolitica* 8081) [CAL10090]	90	55	5e-34	*c5175*	None	None
		Adhesion-related						
229/536-D7	347	Putative Yqi fimbrial adhesin (*E. coli* UTI89, APEC O1 and CFT073) [ABE08929]	100	115	9e-67	c3794	C3482	None
229/536-H6	522	Putative Fml type I fimbrial adhesin FmlD precursor (*E. coli* UTI89) [ABE07194]; also matches several other *E. coli*, including APEC O1 and CFT073	100	173	2e-85	c1931	C1716	Z2206
		Membrane/transport proteins						
229/536-9	197	Putative transport system protein, YeeE/YedE (*E. coli* B and *S. flexneri*) [EAY46529]	100	65	5e20	None	None	Z3175

Table 2. cont.

SSH sequence	Length (bp)	Best BLASTX match/comments [accession no. of best match]	ID (%)	Length (aa)*	*E*-value	CFT073 (UPEC)	UTI89 (UPEC)	EDL 933 (EHEC)
229/536-B9	>349	Putative outer-membrane protein YieC precursor (*E. coli* UTI89) [ABE07231]; also matches *E. coli* CFT073 and APEC O1	98	115	2e-55	c1956	C1753	None
229/536-C1	277	EntS/YbdA MFS transporter (*E. coli* UTI89) [ABE07105]	95	73	4e-32	None	C1627	None
229/536-D5	504	PTS system, galactitol-specific IIC component (*E. coli* UTI89) [ABE09431]	99	167	9e-87	c4279	C4002	Z4877
229/536-E11	189	Predicted transporter component (several *E. coli*) [ZP_00713448]	100	63	3e-30	None	None	Z3175
229/536-H7	347	SecA-related protein (*V. cholerae* MAK 757) [EAY39281]	45	111	4e-24	None	None	None
		O antigen/capsule-related						
229/536-B4	449	WbuG, O4 antigen glycosyl transferase family protein (*E. coli*) [AAT85652]	100	136	3e-74	None	None	None
229/536-D4	453	FnlC O antigen biosynthesis protein (several *E. coli*) [AAT85657]	100	150	5e-80	None	None	None
229/536-D2	308	KpsD, group III capsular polysaccharide transport protein (*E. coli* CP9) [AAC38077]	98	77	7e-26	None	None	None
229/536-H3	>652	KpsD, group III capsular polysaccharide transport protein (*E. coli* CP9) [AAC38077]	100	217	6e-111	None	None	None
229/536-H5	>720	dTDP-D-glucose-4,6-dehydratase, O antigen or capsule-related (*Aeromonas hydrophila* and many other bacteria) [AAM74474]	91	181	8e-81	*c4708*	*C2312*	*Z5299*
		Virulence-related						
229/536-7	383	Hypothetical protein UTI89_C4920 (*E. coli* UTI89) [ABE10323]	98	68	5e-26	None	C4920	None
		CNF1 (*E. coli*) [CAK02719]	100	36	7e-11	None	C4921	None
		Other proteins/enzymes						
229/536-8	302	Phosphoglycerate dehydrogenase (*E. coli* UTI89 and APEC O1) [ABE09403]	100	100	4e-50	None	C3974	None
229/536-B10	>534	Putative propionate CoA-transferase (*E. coli* CFT073) [AE016770]; also matches *E. coli* UTI89 and APEC O1	99	166	1e-93	c5024	C4623	None
229/536-B11	>432	Putative oxidoreductase (*E. coli* UTI89) [ABE10028]; also matches *E. coli* CFT073 and APEC O1	100	143	4e-78	c5021	C4620	None
229/536-E5	477	Putative oxidoreductase (*E. coli* UTI89 and others) [ABE10028] (different from B11)	100	159	6e-89	c5021	C4620	None
229/536-C5	388	Probable DMSO reductase chain *ynfF* precursor (*E. coli* CFT073 and others) [AE016761]	100	111	2e-59	c1978	C1775	Z2576
229/536-D8	647	*N*-Acetylneuraminate synthase (*Rhodopseudomonas palustris* BisA53) [ABJ08161]	31	203	2e-19	None	None	None
229/536-D10	288	Long-chain fatty acid transport protein precursor (*E. coli* CFT073) [AE016764]	100	95	2e-49	c2889	*C2629*	*Z3608*
229/536-G7	>444	5-Methylthioribose kinase (*Sinorhizobium medicae* WSM419) [ABR64583]	46	124	2e-26	None	None	None
229/536-G8	385	DNA methylase M, restriction–modification-related (*E. coli* UTI89) [ABE10454]; also matches *E. coli* APEC O1 and K12	100	85	3e-44	None	C5051	None
		Hypothetical/unknown proteins						
229/536-10	385	Hypothetical protein APECO1_4055 (*E. coli* APEC O1) [ABJ01897]	98	78	3e-40	None	None	None
229/536-B1	615	Hypothetical protein YberA_01003280 (*Yersinia bercovieri* ATCC 43970) [ZP_00820565]	65	205	7e-76	None	None	None
229/536-D11	>323	Hypothetical protein Bamb_5065 (*Burkholderia cepacia* AMMD) [ABI90614]	36	87	9e-11	None	None	None
229/536-E12	>347	Hypothetical protein (*E. coli* W3110) [BAE78334]	99	115	1e-59	None	None	None
229/536-G5	485	Hypothetical protein (uncultured bacterium, and non-*E. coli* bacteria including *Yersinia*) [AAO59951]	96	161	3e-89	None	None	None
229/536-H8	251	Hypothetical protein, N terminus (*E. coli* CFT073) [AE016766]	93	16	1.8	c3653	None	None
229/536-H11	>330	Hypothetical protein (*E. coli* UTI89) [ABE10538]; also matches *E. coli* CFT073	100	110	5e-59	c3148	C5138	None
		No significant hits (NSH)						
229/536-6	159	NSH				None	None	None
229/536-A5	226	NSH				None	None	None
229/536-G4	>552	NSH				None	None	None
229/536-H2	418	NSH				None	None	None

Table 2. cont.

*Length refers to the length of the BLASTX match. Italic type indicates a considerably weaker match than the best BLASTX match; matches <50 % identity (ID) to the genomes of CFT073, UTI89 or EDL933 were discounted.

this island. Whilst some isolates were PCR-positive for all four of the sequences, others were PCR-positive for none, one, two or three of the sequences (Supplementary Table S1). The proportion of isolates that were PCR-positive for all four sequences was greater amongst the colon cancer isolates than either the Crohn's disease isolates or the controls (Table 3).

Distribution of other selected SSH sequences amongst isolates from colon cancer and Crohn's disease

Using PCR assays we further tested the distribution of three SSH sequences with possible roles in pathogenicity: 229-7, 229-29 and 229-11. The putative adhesion/invasion-related sequence 229-7 was only identified by PCR assay in the group of colon cancer isolates also associated with the carriage of the UPEC PAIs PAI II_{J96}, PAI I_{536}, PAI II_{536} and PAI II_{CFT073}. Apart from one control isolate and one Crohn's disease isolate, the same subset of colon cancer isolates were also the only members of the panel that tested PCR-positive for the putative adhesion/haemagglutination-related SSH sequence 229-29. The PCR analysis suggested that the APEC-associated island AGI-1 has a wider distribution (Table 3).

It has been reported previously that the P-fimbriae-associated *papC* is present in many of the mucosal isolates (Martin *et al.*, 2004). Since the PCR assays based on the study of Sabaté *et al.* (2006) were only designed to identify one *papG* allele type, we used further PCR assays to identify other *papG* allele types. The most common was *papG* allele II (Table 3).

DISCUSSION

The contribution of PAIs to the evolution of bacterial pathogens has been widely acknowledged (Dobrindt *et al.*, 2004; Schmidt & Hensel, 2004; Gal-Mor & Finlay, 2006). *E. coli* is one of the best-studied examples of a bacterial pathogen that exhibits diversity in virulence due to the carriage of specific combinations of virulence genes and PAIs (Kaper *et al.*, 2004). In particular, several groups have sought to identify and characterize UPEC-specific PAIs (Lloyd *et al.*, 2007; Welch *et al.*, 2002; Chen *et al.*, 2006; Oelschlaeger *et al.*, 2002a; Brzuszkiewicz *et al.*, 2006). In the

Table 3. Percentage of isolates from various sources carrying sequences

Abbreviations: CD, Crohn's disease; UC, ulcerative colitis.

Sequence	Source				
	Cancer (*n*=10)	CD (*n*=8)	UC (*n*=1)	Control (*n*=6)	Total (*n*=25)
229-29 (adhesion/haem utilization)	40	13	0	17	24
229-7 (adhesion/invasion)	30	0	0	0	12
229-11 (AGI-1)	60	38	0	50	48
ECP1966 (PAI VI_{536})	40	13	0	33	28
229-E6 (PAI VI_{536})	40	13	0	33	28
229-E4 (PAI VI_{536})	80	63	100	67	72
229-4 (PAI VI_{536})	70	38	100	50	56
hlyD/cnF1 (PAI II_{J96})	40	0	0	17	20
I9/I10 (PAI I_{536})	30	0	0	17	16
Orf1up/down (PAI II_{536})	30	0	0	17	16
RPAi/RPAf (PAI I_{CFT073})	90	67	0	50	72
sfaI.1/2 (PAI III_{536})	0	0	0	0	0
IRP2FP/RP (PAI IV_{536})	90	67	100	100	88
Cft073Ent1/2 (PAI II_{CFT073})	50	13	0	17	32
papGIf/r (PAI I_{J96})	0	0	0	0	0
papG allele II	50	50	0	33	44
papG allele III	10	0	0	17	4
papC	70	63	0	50	60

study of Brzuszkiewicz *et al.* (2006), the 536 islands are described as either 536-specific (for PAI I_{536} and PAI V_{536}) or as part of the 'common ExPEC gene pool' (for PAI II_{536}, PAI III_{536}, PAI IV_{536} and PAI VI_{536}; although some genes from PAI II_{536} and PAI III_{536} fall into the '536-specific' category).

A simple phylogenetic scheme based on the presence or absence of a few selected genes has been used to demonstrate genetic differences between intestinal *E. coli* and ExPEC (Clermont *et al.*, 2000). Several studies have noted the predominance of phylogenetic groups B2 and D amongst UPEC isolates (Johnson *et al.*, 2005; Houdouin *et al.*, 2006; Sabaté *et al.*, 2006; Bidet *et al.*, 2007) or meningitis isolates (Bidet *et al.*, 2007; Ewers *et al.*, 2007). However, few studies have sought to characterize *E. coli* isolates associated with IBD or colon cancer. Kotlowski *et al.* (2007) identified groups B2 and D as being of high prevalence in Crohn's disease, and they also screened Crohn's and ulcerative colitis isolates for the presence of multiple virulence genes, including adhesins. Interestingly, none carried *cnf1*/*hlyA*, an observation supported by our study with respect to Crohn's disease isolates.

Nevertheless, we found a group of colon cancer isolates that carried not only the *cnf1*/*hlyA* associated with PAI II_{J96} but also several other UPEC-associated islands. Interestingly, the I9/I10 putative fimbrial chaperone of PAI I_{536} and the ORF1 (ECP4571) region of PAI II_{536} match only the genomes of UPEC strains 536 and UTI89 when used to search the database. In addition, the CFT073 ferric enterobactin-transport protein/siderophore of PAI II_{CFT073} matches sequences from only four UPEC strains in the database. Thus, our distribution study indicates the presence amongst the colon cancer isolates of a group of *E. coli* isolates that carry multiple genes previously categorized as UPEC-specific. However, it has been reported that the accessory genome of a commensal *E. coli* strain which is a clinically safe, efficient colonizer of the gut also contains *cnf1*, *hlyA* and PAI II_{536} (Hejnova *et al.*, 2005). Those authors suggested that these 'virulence-associated' genes might actually be contributing to fitness or colonization. It is worth noting that although related in the content of their accessory genome, isolates HM213, HM229 and HM334 are different according to typing by PFGE (Martin *et al.*, 2004) and flagellin gene sequence (data not shown).

A closer analysis of the genome of a representative of this group, strain HM229, using SSH against *E. coli* K-12, indicated that the accessory genome of this strain was much more closely related to the genomes of UPECs than to those of intestinal *E. coli*. However, some subtracted sequences were not UPEC-associated. Four of the SSH sequences matched the *E. coli* O157 : H7 bacteriophage PhiV10, which has been completely sequenced (GenBank accession no. DQ126339), but for which there is no publication. Three further SSH sequences matched integrases, including one (SSH229-26) that matched a P4-like PAI-associated integrase similar to those implicated in playing an important role in the plasticity of *E. coli* genomes (Hochhut *et al.*, 2006; Manson & Gilmore, 2006).

Further subtraction against the UPEC strain 536 enriched for sequences matching mobile genetic elements, other ExPEC strains such as APEC O1, and other UPEC strains or commensals, rather than strains associated with gastrointestinal disease. It has been reported elsewhere that the genome sequence of the avian pathogenic *E. coli* strain O1 : K1 : H7 shares strong similarities with human ExPEC genomes (Johnson *et al.*, 2007). Two SSH sequences matched KpsD, a group 2 capsular polysaccharide production protein from a bacteraemia isolate (CP9). Group 2 capsule operons consist of highly conserved regions (e.g. *kpsDMTE*) that encode transport and assembly functions, combined with highly variable type-specific regions that encode the synthesis and/or polymerization of the specific component sugars of the particular polysaccharide (Whitfield & Roberts, 1999).

Sabaté *et al.* (2006) compared 100 UPEC and 50 commensal *E. coli* isolates, according to single PCR assays for each of eight UPEC-associated PAIs. Overall, 93 % of UPEC isolates carried PAIs, compared to 40 % of commensal isolates. However, it is interesting to compare the percentage carriage of individual islands between the isolate sets used in our study and that of Sabaté *et al.* (2006). The percentage carriage of many PAIs was very similar between isolates from UPEC (Sabaté *et al.*, 2006) and colon cancer (Supplementary Table S3). It is also interesting to note that isolates carrying the same number of PAIs generally carry the same combinations of PAIs. This tendency towards specific combinations of PAIs has been reported before for UPEC and commensal isolates (Sabaté *et al.*, 2006), and is indicative of a stepwise sequential programme of evolution rather than a random acquisition of PAIs.

Based on four PCR assays alone, we found some evidence for variations in PAI VI_{536} amongst the isolates in our panel. Instability in UPEC PAIs has been demonstrated before (Middendorf *et al.*, 2004). Again, isolates sharing the same number of PCR-positives for the four assays had the same combination of PCR-positives. Thus, even within a single PAI, there is evidence for sequential evolution.

PAI IV_{536} is the most widely distributed of the islands amongst commensal and UPEC isolates (Sabaté *et al.*, 2006). This island is related to the so-called high-pathogenicity island (HPI), which is widespread amongst the *Enterobacteriaceae* (Schubert *et al.*, 2004).

It has been demonstrated that the ability of UPEC to cause symptomatic infection is enhanced by adhesins, including type 1 and P fimbriae (Klemm & Schembri, 2000; Oelschlaeger *et al.*, 2002b). The SSH analysis of strain HM229 identified sequences from both type 1 and P fimbriae. P fimbriae, encoded by the *pap* operon, enhance the establishment of bacteriuria, activate the innate immune response (Wullt *et al.*, 2000, 2002; Bergsten *et al.*,

2004), and are widely distributed in UPEC and APEC (Johnson & Stell, 2000; Rodriguez-Siek *et al.*, 2005). Binding is mediated by the tip adhesin PapG, which has at least three allelic forms, of which allele II is the most common (Johnson *et al.*, 1998; Johnson & Stell, 2000). *papC* was widely distributed amongst the mucosal isolates from colon cancer (7/10) and Crohn's disease (4/8), with the relative distributions of the *papG* alleles resembling those reported elsewhere for UPEC (Johnson & Stell, 2000; Houdouin *et al.*, 2006). However, for three *papC* PCR-positive isolates we did not get a PCR-positive for any of the *papG* alleles [including the rare $papGI^a$ allele variant (Johnson & Stell, 2000) (data not shown)], suggesting further allelic variation. An association between *papG* allele III and *cnf1* has been reported elsewhere for UPEC (Johnson & Stell, 2000). However, of the three *cnf1*-positive isolates in this study, only one carried *papG* allele III.

The SSH analysis also identified other putative gene/protein sequences with predicted roles in adherence, invasion or haemagglutination. The SSH sequence 229-29 shared similarity with large proteins implicated in haem utilization or adhesion in UPEC. The SSH sequence 229-7 carries a YadA-like C-terminal region. YadA is the major adhesin of *Yersinia enterocolitica*, and is essential for establishing infection by mediating adhesion to host cells and conferring resistance against bactericidal activity. We made a knockout mutant for the full ORF associated with SSH sequence 229-7. However, we could identify no change in either adherence or invasion with I407 or Caco-2 cell lines, nor could we detect any difference in haemagglutination (data not shown). This may be because these are multifactorial phenotypes. For example, UPEC carry multiple adhesion-related genes, but carriage of individual adhesion-related genes, such as *papC*, can be well below 100 % (Johnson & Stell, 2000; Rodriguez-Siek *et al.*, 2005; Houdouin *et al.*, 2006).

PCR screening of Crohn's disease and cancer *E. coli* isolates for the presence of sequences 229-7 and 229-29 suggested a distribution very similar to that of *cnf1* (Supplementary Table S1). In particular, the colon cancer isolates HM213, HM229 and HM334 carry *cnf1*, 229-7 and 229-29. The exotoxin CNF1 is an acknowledged virulence factor of UPEC but its prevalence has been reported as 16–27.5 % of UPEC isolates (Johnson & Stell, 2000; Rodriguez-Siek *et al.*, 2005). Hence, the carriage rates for our colon cancer isolates for this UPEC-associated virulence factor are very similar to the reported figures for UPEC isolates. In contrast, either none or only one of the Crohn's disease isolates carried many of the pathogenicity-related genes.

The possible role of mucosally adherent *E. coli* in colon cancer pathogenesis is currently speculative, but there is growing interest in the possible role of inflammation (Rhodes & Campbell, 2002), and perhaps particularly in NFκB activation (Greten *et al.*, 2004), in colon cancer. Bacterial adhesion, perhaps to dysplastic mucosa lacking an overlying mucus coat, has the potential to induce epithelial cell changes that could promote cancer development (Hope *et al.*, 2005).

It is clear from our study and others that neither the colon cancer nor the Crohn's disease mucosal *E. coli* populations are uniform. However, it is possible that a subset of isolates plays an important role in the pathogenicity of these diseases. We identified a subset of colon cancer isolates, not represented amongst the Crohn's disease isolates, that carry multiple UPEC-associated virulence genes. The clinical relevance of this group remains unknown, but the presence of such strains merits further investigation.

ACKNOWLEDGEMENTS

We would like to acknowledge support from the North West Cancer Research Fund.

We would like to dedicate this paper to the memory of Professor Tony Hart, an inspirational colleague who is sadly missed by us and by many others in the worldwide microbiological community.

REFERENCES

Barnich, N., Bringer, M. A., Claret, L. & Darfeuille-Michaud, A. (2004). Involvement of lipoprotein NlpI in the virulence of adherent invasive *Escherichia coli* strain LF82 isolated from a patient with Crohn's disease. *Infect Immun* **72**, 2484–2493.

Bergsten, G., Samuelsson, M., Wullt, B., Leijonhufvud, I., Fischer, H. & Svanborg, C. (2004). PapG-dependent adherence breaks mucosal inertia and triggers the innate host response. *J Infect Dis* **189**, 1734–1742.

Bidet, P., Metais, A., Mahjoub-Messai, F., Durand, L., Dehem, M., Aujard, Y., Bingen, E., Nassif, X. & Bonacorsi, S. (2007). Detection and identification by PCR of a highly virulent phylogenetic subgroup among extraintestinal pathogenic *Escherichia coli* B2 strains. *Appl Environ Microbiol* **73**, 2373–2377.

Bringer, M. A., Glasser, A. L., Tung, C. H., Meresse, S. & Darfeuille-Michaud, A. (2006). The Crohn's disease-associated adherent-invasive *Escherichia coli* strain LF82 replicates in mature phagolysosomes within J774 macrophages. *Cell Microbiol* **8**, 471–484.

Brzuszkiewicz, E., Bruggemann, H., Liesegang, H., Emmerth, M., Olschlager, T., Nagy, G., Albermann, K., Wagner, C., Buchrieser, C. & other authors (2006). How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic *Escherichia coli* strains. *Proc Natl Acad Sci U S A* **103**, 12879–12884.

Chen, S. L., Hung, C. S., Xu, J., Reigstad, C. S., Magrini, V., Sabo, A., Blasiar, D., Bieri, T., Meyer, R. R. & other authors (2006). Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. *Proc Natl Acad Sci U S A* **103**, 5977–5982.

Clermont, O., Bonacorsi, S. & Bingen, E. (2000). Rapid and simple determination of the *Escherichia coli* phylogenetic group. *Appl Environ Microbiol* **66**, 4555–4558.

Darfeuille-Michaud, A., Neut, C., Barnich, N., Lederman, E., Di Martino, P., Desreumaux, P., Gambiez, L., Joly, B., Cortot, A. & Colombel, J. F. (1998). Presence of adherent *Escherichia coli* strains in ileal mucosa of patients with Crohn's disease. *Gastroenterology* **115**, 1405–1413.

Darfeuille-Michaud, A., Boudeau, J., Bulois, P., Neut, C., Glasser, A. L., Barnich, N., Bringer, M. A., Swidsinski, A., Beaugerie, L. &

Colombel, J. F. (2004). High prevalence of adherent-invasive *Escherichia coli* associated with ileal mucosa in Crohn's disease. *Gastroenterology* **127**, 412–421.

Dobrindt, U., Blum-Oehler, G., Nagy, G., Schneider, G., Johann, A., Gottschalk, G. & Hacker, J. (2002). Genetic structure and distribution of four pathogenicity islands (PAI I_{536} to PAI IV_{536}) of uropathogenic *Escherichia coli* strain 536. *Infect Immun* **70**, 6365–6372.

Dobrindt, U., Hochhut, B., Hentschel, U. & Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* **2**, 414–424.

Ewers, C., Li, G., Wilking, H., Kiebetaling, S., Alt, K., Antao, E. M., Laturnus, C., Diehl, I., Glodde, S. & other authors (2007). Avian pathogenic, uropathogenic, and newborn meningitis-causing *Escherichia coli*: how closely related are they? *Int J Med Microbiol* **297**, 163–176.

Gal-Mor, O. & Finlay, B. B. (2006). Pathogenicity islands: a molecular toolbox for bacterial virulence. *Cell Microbiol* **8**, 1707–1719.

Greten, F. R., Eckmann, L., Greten, T. F., Park, J. M., Li, Z. W., Egan, L. J., Kagnoff, M. F. & Karin, M. (2004). IKKβ links inflammation and tumorigenesis in a mouse model of colitis-associated cancer. *Cell* **118**, 285–296.

Guyer, D. M., Kao, J. S. & Mobley, H. L. (1998). Genomic analysis of a pathogenicity island in uropathogenic *Escherichia coli* CFT073: distribution of homologous sequences among isolates from patients with pyelonephritis, cystitis, and catheter-associated bacteriuria and from fecal samples. *Infect Immun* **66**, 4411–4417.

Hejnova, J., Dobrindt, U., Nemcova, R., Rusniok, C., Bomba, A., Frangeul, L., Hacker, J., Glaser, P., Sebo, P. & Buchrieser, C. (2005). Characterization of the flexible genome complement of the commensal *Escherichia coli* strain A0 34/86 (O83 : K24 : H31). *Microbiology* **151**, 385–398.

Hochhut, B., Wilde, C., Balling, G., Middendorf, B., Dobrindt, U., Brzuszkiewicz, E., Gottschalk, G., Carniel, E. & Hacker, J. (2006). Role of pathogenicity island-associated integrases in the genome plasticity of uropathogenic *Escherichia coli* strain 536. *Mol Microbiol* **61**, 584–595.

Hope, M. E., Hold, G. L., Kain, R. & El Omar, E. M. (2005). Sporadic colorectal cancer – role of the commensal microbiota. *FEMS Microbiol Lett* **244**, 1–7.

Houdouin, V., Bonacorsi, S., Bidet, P., Bingen-Bidois, M., Barraud, D. & Bingen, E. (2006). Phylogenetic background and carriage of pathogenicity island-like domains in relation to antibiotic resistance profiles among *Escherichia coli* urosepsis isolates. *J Antimicrob Chemother* **58**, 748–751.

Johnson, J. R. & Stell, A. L. (2000). Extended virulence genotypes of *Escherichia coli* strains from patients with urosepsis in relation to phylogeny and host compromise. *J Infect Dis* **181**, 261–272.

Johnson, J. R., Brown, J. J. & Maslow, J. N. (1998). Clonal distribution of the three alleles of the Gal(α1–4)Gal-specific adhesin gene *papG* among *Escherichia coli* strains from patients with bacteremia. *J Infect Dis* **177**, 651–661.

Johnson, J. R., Scheutz, F., Ulleryd, P., Kuskowski, M. A., O'Bryan, T. T. & Sandberg, T. (2005). Phylogenetic and pathotypic comparison of concurrent urine and rectal *Escherichia coli* isolates from men with febrile urinary tract infection. *J Clin Microbiol* **43**, 3895–3900.

Johnson, T. J., Kariyawasam, S., Wannemuehler, Y., Mangiamele, P., Johnson, S. J., Doetkott, C., Skyberg, J. A., Lynne, A. M., Johnson, J. R. & Nolan, L. K. (2007). The genome sequence of avian pathogenic *Escherichia coli* strain O1 : K1 : H7 shares strong similarities with human extraintestinal pathogenic *E. coli* genomes. *J Bacteriol* **189**, 3228–3236.

Kaper, J. B., Nataro, J. P. & Mobley, H. L. (2004). Pathogenic *Escherichia coli*. *Nat Rev Microbiol* **2**, 123–140.

Klemm, P. & Schembri, M. A. (2000). Bacterial adhesins: function and structure. *Int J Med Microbiol* **290**, 27–35.

Kotlowski, R., Bernstein, C. N., Sepehri, S. & Krause, D. O. (2007). High prevalence of *Escherichia coli* belonging to the B2 + D phylogenetic group in inflammatory bowel disease. *Gut* **56**, 669–675.

Liu, Y., van Kruiningen, H. J., West, A. B., Cartun, R. W., Cortot, A. & Colombel, J. F. (1995). Immunocytochemical evidence of *Listeria, Escherichia coli*, and *Streptococcus* antigens in Crohn's disease. *Gastroenterology* **108**, 1396–1404.

Lloyd, A. L., Rasko, D. A. & Mobley, H. L. (2007). Defining genomic islands and uropathogen-specific genes in uropathogenic *Escherichia coli*. *J Bacteriol* **189**, 3532–3546.

Mager, D. L. (2006). Bacteria and cancer: cause, coincidence or cure? A review. *J Transl Med* **4**, 14.

Manson, J. M. & Gilmore, M. S. (2006). Pathogenicity island integrase cross-talk: a potential new tool for virulence modulation. *Mol Microbiol* **61**, 555–559.

Martin, H. M., Campbell, B. J., Hart, C. A., Mpofu, C., Nayar, M., Singh, R., Englyst, H., Williams, H. F. & Rhodes, J. M. (2004). Enhanced *Escherichia coli* adherence and invasion in Crohn's disease and colon cancer. *Gastroenterology* **127**, 80–93.

Meconi, S., Vercellone, A., Levillain, F., Payre, B., Al Saati, T., Capilla, F., Desreumaux, P., Darfeuille-Michaud, A. & Altare, F. (2007). Adherent-invasive *Escherichia coli* isolated from Crohn's disease patients induce granulomas *in vitro*. *Cell Microbiol* **9**, 1252–1261.

Middendorf, B., Hochhut, B., Leipold, K., Dobrindt, U., Blum-Oehler, G. & Hacker, J. (2004). Instability of pathogenicity islands in uropathogenic *Escherichia coli* 536. *J Bacteriol* **186**, 3086–3096.

Mylonaki, M., Rayment, N. B., Rampton, D. S., Hudspith, B. N. & Brostoff, J. (2005). Molecular characterization of rectal mucosa-associated bacterial flora in inflammatory bowel disease. *Inflamm Bowel Dis* **11**, 481–487.

Oelschlaeger, T. A., Dobrindt, U. & Hacker, J. (2002a). Pathogenicity islands of uropathogenic *E. coli* and the evolution of virulence. *Int J Antimicrob Agents* **19**, 517–521.

Oelschlaeger, T. A., Dobrindt, U. & Hacker, J. (2002b). Virulence factors of uropathogens. *Curr Opin Urol* **12**, 33–38.

Perna, N. T., Plunkett, G., III, 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.

Rasko, D. A., Phillips, J. A., Li, X. & Mobley, H. L. (2001). Identification of DNA sequences from a second pathogenicity island of uropathogenic *Escherichia coli* CFT073: probes specific for uropathogenic populations. *J Infect Dis* **184**, 1041–1049.

Rhodes, J. M. & Campbell, B. J. (2002). Inflammation and colorectal cancer: IBD-associated and sporadic cancer compared. *Trends Mol Med* **8**, 10–16.

Rodriguez-Siek, K. E., Giddings, C. W., Doetkott, C., Johnson, T. J., Fakhr, M. K. & Nolan, L. K. (2005). Comparison of *Escherichia coli* isolates implicated in human urinary tract infection and avian colibacillosis. *Microbiology* **151**, 2097–2110.

Rolhion, N., Barnich, N., Claret, L. & Darfeuille-Michaud, A. (2005). Strong decrease in invasive ability and outer membrane vesicle release in Crohn's disease-associated adherent-invasive *Escherichia coli* strain LF82 with the *yfgL* gene deleted. *J Bacteriol* **187**, 2286–2296.

Rolhion, N., Carvalho, F. A. & Darfeuille-Michaud, A. (2007). OmpC and the σ^E regulatory pathway are involved in adhesion and invasion of the Crohn's disease-associated *Escherichia coli* strain LF82. *Mol Microbiol* **63**, 1684–1700.

Ryan, P., Kelly, R. G., Lee, G., Collins, J. K., O'Sullivan, G. C., O'Connell, J. & Shanahan, F. (2004). Bacterial DNA within granulomas of patients with Crohn's disease – detection by laser capture microdissection and PCR. *Am J Gastroenterol* **99**, 1539–1543.

Sabaté, M., Moreno, E., Perez, T., Andreu, A. & Prats, G. (2006). Pathogenicity island markers in commensal and uropathogenic *Escherichia coli* isolates. *Clin Microbiol Infect* **12**, 880–886.

Schmidt, H. & Hensel, M. (2004). Pathogenicity islands in bacterial pathogenesis. *Clin Microbiol Rev* **17**, 14–56.

Schubert, S., Rakin, A. & Heesemann, J. (2004). The *Yersinia* high-pathogenicity island (HPI): evolutionary and functional aspects. *Int J Med Microbiol* **294**, 83–94.

Simpson, K. W., Dogan, B., Rishniw, M., Goldstein, R. E., Klaessig, S., McDonough, P. L., German, A. J., Yates, R. M., Russell, D. G. & other authors (2006). Adherent and invasive *Escherichia coli* is associated with granulomatous colitis in boxer dogs. *Infect Immun* **74**, 4778–4792.

Swidsinski, A., Khilkin, M., Kerjaschki, D., Schreiber, S., Ortner, M., Weber, J. & Lochs, H. (1998). Association between intraepithelial *Escherichia coli* and colorectal cancer. *Gastroenterology* **115**, 281–286.

Swidsinski, A., Ladhoff, A., Pernthaler, A., Swidsinski, S., Loening-Baucke, V., Ortner, M., Weber, J., Hoffmann, U., Schreiber, S. & other authors (2002). Mucosal flora in inflammatory bowel disease. *Gastroenterology* **122**, 44–54.

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.

Whitfield, C. & Roberts, I. S. (1999). Structure, assembly and regulation of expression of capsules in *Escherichia coli*. *Mol Microbiol* **31**, 1307–1319.

Wullt, B., Bergsten, G., Connell, H., Rollano, P., Gebretsadik, N., Hull, R. & Svanborg, C. (2000). P fimbriae enhance the early establishment of *Escherichia coli* in the human urinary tract. *Mol Microbiol* **38**, 456–464.

Wullt, B., Bergsten, G., Samuelsson, M. & Svanborg, C. (2002). The role of P fimbriae for *Escherichia coli* establishment and mucosal inflammation in the human urinary tract. *Int J Antimicrob Agents* **19**, 522–538.

Edited by: D. L. Gally