JB Accepts, published online ahead of print on 27 August 2010
J. Bacteriol. doi:10.1128/JB.00710-10

A commensal gone bad: complete genome sequence of the prototypical enterotoxigenic *Escherichia coli* strain H10407

Lisa C. Crossman[1†§], Roy R. Chaudhuri[2§], Scott A. Beatson[3], Timothy J. Wells[4], Mickael Desvaux[4ƒ], Adam F. Cunningham[4], Nicola K. Petty[1], Vivienne Mahon[5], Carl Brinkley[6], Jon L. Hobman[7], Stephen J. Savarino[6], Susan M. Turner[4], Mark J. Pallen[8], Charles W. Penn[8], Julian Parkhill[1], A. Keith Turner[1], Timothy J. Johnson[9], Nicholas R. Thomson[1], Stephen G.J. Smith[5], Ian R. Henderson[4*]

[1]The Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge, United Kingdom; [2]Department of Veterinary Medicine, University of Cambridge, Cambridge, United Kingdom; [3]School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, Australia; [4]School of Immunity and Infection and [8]School of Biosciences, University of Birmingham, Birmingham, United Kingdom; [5]Department of Clinical Microbiology, School of Medicine, Trinity College Dublin, Dublin, Ireland; [6]Department of Enteric Infections, Division of Communicable Diseases and Immunology, Walter Reed Army Institute of Research, Silver Spring, Maryland, USA; [7]School of Biosciences, The University of Nottingham, Sutton Bonington, United Kingdom; [9]Department of Veterinary and Biomedical Sciences, University of Minnesota, Saint Paul, Minnesota.

*Corresponding author. Mailing address: School of Immunity and Infection, University of Birmingham, Birmingham, B15 2TT, United Kingdom. Phone: +44 121 4144368. Email: i.r.henderson@bham.ac.uk

Present address: [†] The Genome Analysis Centre, Norwich, United Kingdom and [ƒ]INRA, UR454 Microbiology, F-63122 Saint-Genès Champanelle, France

[§]LCC and RRC contributed equally to this investigation.

Running Title: **ETEC genome sequence**

ABSTRACT

In most cases *Escherichia coli* exists as a harmless commensal organism but may on occasion cause intestinal and/or extraintestinal disease. Enterotoxigenic *E. coli* are the predominant cause of *E. coli*-mediated diarrhea in the developing world and are responsible for a significant portion of paediatric deaths. In this study we determined the complete genomic sequence of *E. coli* H10407, a prototypical strain of enterotoxigenic *E. coli*, which reproducibly elicits diarrhea in human volunteer studies. We performed genomic and phylogenetic comparisons with other *E. coli* strains revealing that the chromosome is closely related to the non-pathogenic commensal strain *E. coli* HS and to the laboratory strains *E. coli* K-12 and C. Furthermore, these analyses demonstrated that there were no chromosomally-encoded factors unique to any sequenced ETEC strains. Comparison of the *E. coli* H10407 plasmids with those from several ETEC strains revealed the plasmids had a mosaic structure but that several loci were conserved amongst ETEC strains. This study provides a genetic context for the vast amount of experimental and epidemiological data published thus far.

INTRODUCTION

Current dogma suggests the Gram-negative motile bacterium *Escherichia coli* colonises the infant gut within hours of birth and establishes itself as the predominant facultative anaerobe of the colon for the remainder of life (3, 59). While the majority of *E. coli* strains maintain this harmless existence some strains have adopted a pathogenic lifestyle. Contemporary tenets suggest that pathogenic strains of *E. coli* have acquired genetic elements, which encode virulence factors and enable the organism to cause disease (12). The large repertoire of virulence factors enables *E. coli* to cause a variety of clinical manifestations including intestinal infections mediating diarrhea and extraintestinal infections, such as urinary tract infections, septicaemia and meningitis. Based on clinical manifestation of disease, the repertoire of virulence factors, epidemiology and phylogenetic profiles, strains causing intestinal infections can be divided into six separate pathotypes viz. enteroaggregative *E. coli* (EAEC), enteroinvasive (EIEC), enteropathogenic *E. coli* (EPEC), enterohaemorrhagic *E. coli* (EHEC), diffuse adhering *E. coli* (DAEC) and enterotoxigenic *E. coli* (ETEC) (33, 35, 39).

ETEC is responsible for the majority of *E. coli*-mediated cases of human diarrhea worldwide. It is particularly prevalent amongst children in developing countries where sanitation and clean supplies of drinking water are inadequate, and in travellers to such regions. It is estimated that there are 200 million incidences of ETEC infection annually resulting in hundreds of thousands of deaths in children under the age of 5 (55, 64). The essential determinants of ETEC virulence are traditionally considered to be colonization of the host small intestinal epithelium via plasmid-encoded colonization factors (CFs), and subsequent release of plasmid-encoded heat-stable (ST) and/or heat-labile (LT) enterotoxins that induce a net secretory state leading to profuse watery diarrhea (20, 62). More recently, additional plasmid-encoded factors have been implicated in the pathogenesis of ETEC, namely the EatA serine protease autotransporter (SPATE) and the EtpA protein that acts as an intermediate in the adhesion between bacterial flagella and host cells (23, 32, 42, 46). Furthermore, a number of chromosomal factors are thought to be involved in virulence e.g. the invasin Tia, the TibA adhesin/invasin and LeoA, a GTPase of unknown function (14, 21, 22). *E. coli* H10407 is considered a prototypical ETEC strain; it expresses colonization factor antigen 1 (CFA/I) and the heat-stable and heat labile toxins. Loss of a 94.8-kb plasmid encoding CFA/I and a gene for ST enterotoxin from *E. coli* strain H10407 leads to reduced ability to cause diarrhea (17).

Here we report the complete genome sequence and virulence factor repertoire of the prototypical ETEC strain H10407, the nucleotide sequence and gene repertoire of the plasmids from ETEC strain E1392/75 and we describe a novel conserved secretion system associated with the sequenced ETEC strains.

MATERIALS AND METHODS

Bacterial strains and sequencing.

The ETEC O78:H11:K80 strain H10407 was isolated from an adult with cholera-like symptoms in the course of an epidemiologic study in Dacca, Bangladesh prior to 1973 (19) and was shown to cause diarrhea in adult volunteers (6, 17). The *E. coli* H10407 isolate that was sequenced was from the Walter Reed Army Institute of Research (WRAIR) cGMP stock manufactured in February 1998 as Lot 0519. The whole genome was sequenced to a depth of 8 x coverage from pUC19 (insert size 2.8-5 kb) and pMAQ1b (insert size 5.5-10 kb) small insert libraries. Sanger Sequencing was carried out using Amersham Big-Dye (Amersham, UK) terminator chemistry on ABI3700 sequencing machines. End sequences from larger insert plasmid (pBACe3.6, 20-30 kb insert size) libraries were used as a scaffold. Sequence reads were assembled into contigs with Phrap (Green P, unpublished) and finished using GAP4 as described previously (33). The plasmids from ETEC O6:H16:K15 strain E1392/75, which was isolated from a patient in Hong Kong with diarrhea, expresses the CFA/II (CS1 and CS3) colonization factors and produces the ST and LT toxins, were also sequenced using a similar approach (7, 50, 60). Plasmid DNA for ETEC E1392/75 was provided by Acambis UK.

Gene prediction, annotation, and comparative analysis

Annotation was carried out using the genome viewer Artemis (47). Coding sequences were predicted using the gene prediction programs Orpheus (26), Glimmer2 (11) and Glimmer3 (10), then manually curated. Protein domains were marked up using Pfam (48) and transmembrane domains and signal sequences were predicted using TMHMM and SignalP, respectively (15, 37). Annotation was transferred from previously annotated *E.coli* genomes to orthologous genes and manually curated. A homologue was considered to be present if a hit was found with >60% identity over at least 80% of the length of the query protein. Regions of difference and plasmids were annotated and curated manually. The annotated genome sequence of ETEC H10407 and the plasmids from ETEC H10407 and E1392/75 have been deposited in the EMBL databases (accession number: FN649414 for ETEC H10407 complete chromosome; see Tables 1 and 2 for general features of the nucleotides sequences and accession numbers for the plasmids).

RESULTS AND DISCUSSION

Structure and general features of ETEC H10407 chromosome.

The ETEC H10407 genome consists of a circular chromosome of 5,153,435 bp and four plasmids designated pETEC948, pETEC666, pETEC58 and pETEC52, respectively. The general features of the ETEC H10407 chromosome are presented in Table 1 and the plasmids in Table 2. We identified 4746 protein-coding genes (CDSs) in the chromosome, 33 (0.67%) of which do not have any match in the database, 579 (11.67%) encode conserved hypothetical proteins, with no known function and 503 (10.14%) are genes associated with mobile elements such as integrases, transposases, or phage related. We have identified 25 regions of difference (ROD) that occur in the ETEC H10407 genome and are differentially distributed among the other sequenced *E. coli* chromosomes (Figure 1; Table S1). The combined size of these RODs is 755,359 bp (14.7% of the chromosome) and includes nine prophages, designated ETP29, 33, 86, 128, 216, 284, 295, 468 and 507, where the numeric designations denote their approximate positions (x 10,000 bp) on the chromosome. None appeared to carry cargo genes related to virulence.

Comparative genomics of the ETEC H10407 chromosome.

Previously, a phylogeny was constructed based on the concatenated sequences of 2,173 genes that are conserved in all *E. coli* strains and in *Escherichia albertii* and *Escherichia fergusonii*, which were included as outgroup sequences (4). The established *E. coli* sub-groups (A, B1, B2, D and E) are all monophyletic with the exception of group D, which is divided at the root. In agreement with previous optical mapping experiments (5), *E. coli* H10407 is located in the A subgroup with the non-pathogenic laboratory strains *E. coli* K-12 and C and the non-pathogenic commensal isolate *E. coli* HS. The majority of commensal strains of bacteria belong to the A subgroup (59).

Comparison of *E. coli* H10407 with the closely related non-pathogenic *E. coli* K-12, C and HS strains reveals these chromosomes are largely colinear (Fig S1) and that *E. coli* H10407 chromosome contains 599 CDSs not present in the non-pathogenic strains (Fig. 2 and Table S2). The majority (528) of these are clustered in the 25 RODs and are predicted to encode prophage genes and other mobility factors. Several genes encode previously described loci specifically associated with ETEC virulence viz. *leoA* (ROD 20), *tia* (ROD 20) and *tib* (ROD 13) (13, 14, 22). Other genes encode loci previously noted in ETEC H10407 including the degenerate ETT2 locus (ROD 18) (45), Antigen 43 (ROD 23) (63), a Type 2 protein secretion locus found in many strains of *E. coli* (ROD 19) (4) and the *ecpP* fimbrial gene cluster also found in many *E. coli* strains (ROD 1) (4). Other RODs encode the Sil/Pco efflux system conferring silver/copper resistance (ROD 2), yersiniabactin (ROD 11), and the O78 serotype O-antigen biosynthetic locus (ROD 14). The *sil* operon is closely related to *sil* from IncH2 plasmid pMG101 (30, 38, 53) and is adjacent to a partially interrupted copper resistance operon similar to *pco* from plasmid pRJ1004 (2). The *sil*/*pco* locus is flanked by IS element and phage related sequences suggesting horizontal transfer of these genes. The yersiniabactin iron acquisition locus is widely distributed in *E. coli* and other members of the *Enterobacteriaceae* (49). The remaining *E. coli* H10407-specific CDSs, which are not present on a ROD, and do not encode prophage or mobility factors, encode the H11 flagellin subunits (CDS 2029-2033), an additional copy of Antigen 43 (CDS 2119), and several pseudogenes (CDS 427, 1476, 1573). This data largely agrees with previously published subtractive hybridisation studies (5).

If a particular protein plays an important role in ETEC-mediated disease then one would expect the gene encoding it to have a wide distribution amongst ETEC strains. To determine if there were any chromosomally-encoded genes specific for ETEC strains, comparisons were made with *E. coli* strains E24377A and B7A, the only other ETEC strains for which genome sequence data is available (44). Unlike *E. coli* H10407 both the *E. coli* strains E24377A and B7A belong to the B1 subgroup of the *E. coli* phylogeny, a subgroup from which many commensals are derived but also a number of pathogens (4, 59). Comparison of *E. coli* H10407 with the sequenced ETEC strain E24377A revealed the chromosomes are largely colinear (Fig S2). The genome of ETEC B7A is not finished but experience with other *E. coli* genomes and comparison of the 198 finished ETEC B7A contigs suggests that the chromosome is also largely colinear with the other sequenced ETEC genomes (Fig. S2). Analyses of the gene content of all three strains revealed 3741 genes conserved in all three strains, of which only 188 are not present in the commensal *E. coli* HS (Fig. 2B and Table S3). The 188 genes identified through this comparison included loci encoding xanthine dehydrogenase (CDS 0339-0343), the Mat fimbriae (CDS 0348-0352), conserved proteins of unknown function (CDS 0673-0678), a flavoprotein electron transfer system (CDS 1730-1734), the colanic exopolysaccharide biosynthetic machinery (CDS 2171-2202), the Fec iron citrate uptake system (CDS 3161-3166), a cellulose synthase system (CDS 3776-3779) and a putative sugar utilisation system (CDS 4145-4154) all of which were present in the non-pathogen *E. coli* K-12 and are widely distributed amongst other *E. coli* (data not shown). The remainder of the 188 genes encode prophage or other mobility factors which are predicted to have no role in virulence. Of the 599 *E. coli* H10407-restricted genes identified through comparisons with the non-pathogenic *E. coli* strains above (Fig. 2A), 47 were conserved amongst the three pathogenic ETEC isolates. However, these genes were all related to mobile elements and no putative virulence factors were identified. Notably, no significant homologues of *leoA*, *tibC*, *tibA* or *tia* were detected in either *E. coli* E24377A or B7A strongly suggesting these genes are not essential for ETEC-mediated disease. In conclusion, these data agree with previous observations that the chromosome of *E. coli* H10407 is most closely related to non-pathogenic *E. coli* and the factors mediating diarrhea are not chromosomally encoded thereby indicating the essential virulence factors are encoded on the plasmids (61).

Potential virulence genes encoded on the ETEC plasmids

Since chromosomal comparisons revealed that no chromosomal CDS was unique to all three ETEC strains we next examined the CDSs present on the four plasmids of ETEC H10407. The general characteristics of the plasmids are shown in Table 2.

The two larger plasmids (pETEC948 and pETEC666) are reminiscent of conjugative plasmids that are often associated with the carriage of virulence factors whereas the two smaller plasmids (pETEC58 and pETEC52) are homologous to mobilisable plasmids frequently encountered in a variety of bacterial species (24, 34). The latter plasmids have been shown to be mobilizable in the presence of IncF and other plasmid transfer systems (51). The majority of the CDSs on all four plasmids encode plasmid maintenance and transfer functions, pseudogenes, genes of unknown function not predicted to be involved in virulence and transmissible elements (Table 2). An exhaustive list of the genetic content is unwarranted here, as the complete annotation of the plasmids is provided via the EMBL databases. Nevertheless, there are several noteworthy CDSs, described below, which can be termed “cargo” genes and have a known or putative role in pathogenesis. Thus, analyses revealed *E. coli* H10407 pETEC948 possesses cargo genes encoding the previously described EatA SPATE (*eatA*), heat-stable enterotoxin STa2 (*sta2*), CFA/I fimbriae and associated regulator (*cfaABCD*), Etp two partner secretion system and associated glycosyltransferase (*etpABC*) (Fig. 3) (18, 23, 42, 66). Analyses of the *E. coli* H10407 pETEC666 plasmid revealed it contains the cargo genes encoding the previously described heat stable enterotoxin STa1 (*sta1*) and the two subunits of LT enterotoxin (*eltA* and *eltB*) (Fig 3) (8, 65). In addition, the plasmids contain several loci not previously associated with ETEC strains. ETEC H10407 pETEC948 possesses genes encoding a Type I secretion locus similar to the dispersin secretion locus (*aatA-P*) described for *E. coli* 042 (Fig. 4)(52). Associated with this locus is a gene encoding CexE, a previously described secreted protein of ETEC (43), which bears homology to the *E. coli* 042 dispersin protein (Fig. 4). Furthermore, pETEC666 encodes a two-component sensor-kinase, herein designated *etcA* and *etcB* (***E. coli***

two-**c**omponent), and a three gene locus (herein designated *eor* for ***E. coli*** **o**xido**r**eductase) encoding a protein with homology to cytochrome b-type subunit oxidoreductase protein (*eorA*), a protein with homology to an oxidoreductase molybdopterin binding domain protein (*eorB*) and a periplasmic protein of unknown function (*eorC*). In addition, ETEC H10407 pETEC58 encodes a putative deoxycytidylate deaminase (pETEC58_0005).

As above, if a particular protein plays an important role in ETEC-mediated disease then one would expect it to have a wide distribution amongst ETEC strains. To determine whether the genes encoding the putative and known virulence factors of the ETEC H10407 plasmids, which we identified above, were conserved amongst ETEC strains we next examined their prevalence amongst the available sequenced strains. To aid this process we determined the sequence of the plasmids from ETEC strain E1392/75. *E. coli* E1392/75 possesses five plasmids three large conjugative plasmids designated pETEC1018, pETEC746 and pETEC557 and two mobilizable plasmids termed pETEC75 and pETEC62 (see Table 2 for general characteristics). Included in the prevalence investigations were the ETEC strains E24377A and B7A and the plasmid pCoo from ETEC strain C921b-1, all of which were sequenced in other projects (28, 44). As the ETEC B7A genome is incomplete and no plasmids were resolved, and pCoo is the only plasmid sequenced from ETEC C921b-1, we can only confirm the presence of genes amongst the available DNA sequences and not the absence of particular genes from these strains. The distribution and location of the cargo genes encoding known or putative virulence factors amongst the sequenced ETEC plasmids is depicted in Fig. 3 and also listed in Table S4. Comparative analyses revealed that, like ETEC H10407, the ETEC strains E1392/75, B7A and E24377A possess the ST and LT enterotoxins (none were identified for *E. coli* C921b-1, but previous analyses showed this strain to harbour LT and ST) (54). The EtpABC two-partner secretion system was identified in ETEC E1392/75 and E24377A; homologues may exist in ETEC strains B7A and C921b-1 but their existence or non existence in these strains could not be resolved due to the lack of complete sequence data however other studies have not demonstrated a universal association of the *etpABC* locus with ETEC strains (23). Unlike ETEC strains H10407, E24377A and C921b-1, the autotransporter-encoding *eatA* gene was not present on the ETEC E1392/75 plasmids. A homologue annotated as EatA is found in *E. coli* B7A, however further analyses of this protein reveal that it is more closely related to SepA, a homologous SPATE protein from *Shigella flexneri* (1). No equivalents of the ETEC H10407 *etcAB*, *eorA-C* or of the gene encoding the putative deoxycytidylate deaminase, were detected in any of the other ETEC strains.

Like *E. coli* H10407, the ETEC strains E24377A, E1392/75 and C921b-1 encode dispersin-like proteins previously designated CexE (43). Further analyses reveal that CexE is present in ETEC strains 27D and G427 (two CFA/I$^+$ strains) (43) and ETEC O167:H5, a CS6 and CS5 encoding strain (9). For EAEC, dispersin is secreted via the Aat Type I secretion system, associates non-covalently with the extracellular face of the outer membrane preventing collapse of the AAF/II fimbriae onto the bacterial cell surface by alteration of the surface charge and is required for colonisation (31, 40, 52). Analyses of the nucleotide sequences from ETEC strains B7A, E24377A and E1392/75 reveal the presence of loci encoding Type I secretion systems bearing striking homology to the Aat dispersin secretion system (Fig. 4). The co-occurrence of *cexE* genes with *aat* loci suggests that the CexE proteins are substrates for the Aat-like secretion systems of ETEC. Since, plasmid-borne fimbrial loci are inextricably linked to ETEC-mediated disease (18), CexE may play a similar role to dispersin by maintaining the CFs in a manner such that they can interact with epithelial receptors. However, further studies are required to investigate the function and distribution of CexE and to identify other relatives of this protein hitherto not recognised.

As mentioned above, adherence via plasmid-encoded fimbrial systems is a crucial step in ETEC pathogenesis (62). *E. coli* H10407 pETEC948 possesses the CFA/I chaperone-usher system (Fig. 3). ETEC E24377A possesses two-chaperone-usher fimbrial systems located on pETEC_80 and pETEC_73 encoding the CS3 and CS1 fimbriae, respectively (44). Similarly, *E. coli* E1392/75 possesses the CS3- and CS1-encoding loci on plasmids pETEC1018 and pETEC746 respectively, whereas pCoo possesses the CS1 cluster, all of which have been described previously (28, 57, 58). In addition, *E. coli* E1392/75 pETEC557 also encodes the CFA/III type IV fimbrium (29). To determine whether fimbrial systems other than those mentioned above might play a crucial role in ETEC pathogenesis we investigated conservation of putative fimbrial loci amongst the available *E. coli* sequences. ETEC H10407 contains 12 additional loci predicted to encode fimbriae, all of which were chromosomally located (Table S5). Four of these loci (*mat, sfm, ycb* and *yde*) contain pseudogenes and were considered non-functional. We sought to establish if *E. coli* H10407 harboured ETEC-specific fimbrial loci that might not be expressed by commensal *E. coli, E. coli* K-12 or enteroaggregative *E. coli.* The vast majority of fimbrial operons identified are also located in commensal and laboratory strains with notable exceptions. The *yqi* and *stf-mrf* fimbrial loci are present in *E. coli* H10407 but contain pseudogenes in commensal or laboratory *E. coli.* However, an apparently functional *yqi* operon is also present in enteroaggregative *E. coli* strain 042 and thus a functional *yqi* locus does not appear to be ETEC-specific. Indeed, the *yqi* operon does not appear to be present in ETEC B7A (4). With regard to the *stf-mrf* operon, the *mrfC* gene is a pseudogene in *E. coli* K-12 but not in ETEC H10407. This six gene cluster (*smfA-mrfCD-stfEFG*) is present in ETEC E24377A and EAEC 042, though with some divergence in the *stf* genes.

Finally, the ETEC E1392/75 pETEC62 plasmid possesses CDSs encoding a type II dihydropteroate synthase gene conferring sulfonamide resistance, and CDSs encoding streptomycin phosphotransferase genes conferring streptomycin resistance. This plasmid possesses 99% nucleotide identity with the ETEC E24377A pETEC_6 plasmid and shares high levels of identity with plasmids from a variety of *E. coli* sp including the *Shigella sonnei* pKKTET7 and the EPEC pE2348-2 plasmids However, this plasmid has no homologue in ETEC H10407 and no detectable homology amongst the ETEC B7A sequences suggesting it may not be widespread amongst ETEC strains and thus is not essential for ETEC mediated diarrhea.

In conclusion, the putative and known virulence genes identified on the plasmids of *E. coli* H10407 have a differential distribution amongst the sequenced ETEC strains. In all cases the ETEC strains possess genes encoding the the ST and/or LT toxins (*sta* and/or *eltAB, respectively*), a chaperone-usher fimbrial biogenesis locus (e.g. the *cfa* locus) and components of an *aat-cexE* dispersin-like Type I secretion system. Thus, despite the variation in individual plasmid gene content, comparison of the entire plasmid complement of the sequenced ETEC strains suggests that there is a conserved core of genes contained on the plasmids that are predicted to be involved in virulence and may be essential for the establishment of ETEC-mediated disease.

ETEC plasmids demonstrate a mosaic structure

To determine whether the virulence factors identified above were encoded on a specific plasmid, or repertoire of plasmids, we examined the nucleotide sequence identity shared by the ETEC plasmids. The nucleotide sequence of the conjugative plasmids from each of the ETEC strains H10407, E1392/75 and E24377A were concatenated and compared by BLASTn. The level of nucleotide sequence identity between pCoo and the other ETEC plasmids was determined in a similar manner. These comparisons revealed that while the plasmids all belong to a narrow subset of incompatibility groups (see below), extensive rearrangements and recombination events have occurred, resulting in individual plasmids that vary in their repertoires of virulence genes (Fig. 3 and Table S4). Such recombination is exemplified by examining the distribution of the *eatA* gene. Thus, the *eatA* gene is not present in ETEC strain 1392/75, in ETEC strain E24377A the *eatA* gene is located on pETEC_74 and the *eltAB*, *aatPABC* and *etpABC* loci are located on pETEC_80. In contrast, in ETEC strain H10407 the *eatA* gene is collocated with *etpABC* and *aatPABC* on pETEC948 whereas the *eltAB* locus is located on pETEC666. The eatA gene is present on ETEC C921b-1 pCoo along with *cooABCD* however in ETEC strain E24377A *cooABCD* is located on a separate plasmid (pETEC_73) (Fig. 3 and Table S4)*.* Other virulence-associated genes also display such differential distribution (Table S4) suggesting that the extrachromosomal components of the ETEC genome are in a state of flux (34, 44). Notably, the plasmids contain an extensive repertoire of IS elements and transposons (Table 2)(34); it is likely that the mobility of these genetic elements, or the recombination between these elements, gives rise to the observed mosaic structure of the ETEC plasmids.

Similar comparisons of the small mobilizable plasmids of the ETEC strains did not demonstrate recombination between the mobilizsable plasmids. Furthermore, there did not appear to be any significant exchange of genetic material between the conjugative plasmids and the small mobilizable plasmids (data not shown).

Plasmid stability and maintenance functions of the ETEC plasmids

To determine whether the virulence factors described above were encoded on self transmissible plasmids we examined the CDSs encoding the plasmid maintenance and transfer functions of each ETEC plasmid. A complete description of *E. coli* H10407 pETEC666 has been published previously (41) and the complete repertoire of genes for each ETEC plasmid are given in the EMBL databases (see Table 2 for accession numbers) thus only the most salient features are described here. Plasmid nomenclature utilises a system based on incompatibility groupings; plasmids of the same incompatibility group should not co-exist within the same bacterial cell because of the similarity in their replication systems. (34). However, sequence analyses of the CDSs encoding the plasmid replication functions of the repertoire of ETEC plasmids revealed that the large conjugative-like plasmids of *E. coli* strains H10407, E1392/75 and E24377A belong to a narrow subset of incompatibility groups and possess multiple plasmids with the same replication mechanism (Fig. 3 and Table 2). Thus, *E. coli* H10407 plasmids pETEC948 and pETEC666 belong to the RepFIIA (IncFIIA) subset of incompatibility groupings and have RepA1 proteins which share 94% identity (95% similarity), whereas *E. coli* E1392/75 plasmids pETEC746 and pETEC557 harbour RepI1 (IncI1) replication functions (*E. coli* E1392/75 pETEC557 is an apparent cointegrate of a RepF1B and RepI1 pasmids; such cointegration has previously been noted for *E. coli* C921b-1, where pCoo represents a co-integrate between a RepFIIA and a RepI1 plasmid (28)) with the corresponding RepZ proteins sharing 94% identity (95% similarity). Similarly, the previously described ETEC strain E24377A (44) possesses three plasmids with RepFIIA functions. The basis for these anti-dogmatic observations is not understood and requires further in depth investigation.

Analyses of the nucleotide sequences of the repertoire of large conjugative-like plasmids revealed that they possessed a number of plasmids stability systems including post-segregation killing systems and active partitioning systems. The distribution of these systems amongst the plasmids sequenced in this study is given in Table 2. These stability systems have been described previously (see reviews references (25, 56).

Previous studies have noted that the large plasmids encoding the toxins of ETEC are in some cases self transmissible and in other cases not transmissible (27). To investigate whether the plasmids sequenced in this study possessed transmissibility functions we examined the transfer regions of the conjugative-like plasmids. As noted previously, *E. coli* H10407 pETEC666 has a transfer region which is interrupted by several IScE8 elements severely diminishing the ability of this system to function efficiently (41). In contrast, *E. coli* H10407 pETEC948 only possesses remnants of the conjugation apparatus and is presumably not self transmissible. In addition, the *E. coli* E1392/75 pETEC1018 plasmid also contains an incomplete conjugation apparatus which is presumed to be ineffective at promoting conjugation, however *E. coli* E1392/75 pETEC746 possess an intact conjugation system that is 100% identical to the region encoding the functional R64-like conjugative pilus of pCoo of *E. coli* C921b-1 and thus it is presumed to be functional. *E. coli* E1392/75 pETEC557 lacks CDSs encoding the R64 conjugative pilus and possesses remnants of an F-like conjugation system.

ETEC strains H10407, E1392/75, and E24377A all contain similar small mobilizable plasmids (pETEC52, pETEC75, and pETEC_5, respectively) with *mob* and *rep* regions displaying 100% identity. The *E. coli* E1392/75 pETEC75 plasmid contains an IS100 element not present in the other two plasmids. The distribution of these plasmid types among the sequenced ETEC suggests that they might be common components of ETEC genomes. This plasmid type has been found in a number of other *E. coli* strains and has been shown to increase the fitness of certain *E. coli* host strains (16). Therefore, multiple selective advantages might be conferred on the ETEC strains possessing these small plasmids. The *rep* and *mob* regions (3058 bp) of ETEC H10407 pETEC58 plasmid, which encodes the putative deoxycytidylate deaminase, demonstrates 81% identity with plasmid pHW66 from Rahnella sp. WMR66; the putative deoxycytidylate deaminase is lacking from pHW66. In contrast to the other ETEC plasmids, there are no plasmids homologous to ETEC H10407 pETEC58 amongst the other genome sequenced ETEC isolates.

The *E. coli* E1392/75 pETEC746 plasmid contains a pilin shufflon.

As mentioned above, ETEC E1392/75 pETEC746 contains regions homologous to the *S. enterica* Typhimurium RepI1 plasmid R64 that are also present in *E. coli* C921b-1 pCoo and have been shown to be functional in that system (28). During the finishing of the ETEC genome, dideoxy sequencing of the region from 56,253 bp to 59,961 bp of pETEC746 from *E. coli* E1392/75 identified a nucleotide region undergoing dynamic alteration. The region of DNA consisted of a shufflon similar to that of R64 (36). PilV is a component of a conjugative pilus that expresses different tips involved with attachment to cells. The tips are regulated *via* a DNA shufflon mechanism involving recombination at particular repeating sites. Recombination is mediated by the *rci* recombinase linked to this region. Alternative tip adhesins are involved in attachment to different strains and species and has been elucidated experimentally in *S.* Typhimurium (36). Evidence that the shufflon is functional in the *E.coli* E1392/75 plasmid pETEC746 is provided in the sequences from a small insert library. Within the sequences are examples of *pilV* with alternative C-terminal tips, implying that the plasmids sequenced represented a population in genetic flux. There is direct evidence for sequences of *pilV* with tips *V1*, *V3* and *V4* (Fig. 5). There are also regions of DNA sequence equivalent to tips *shuC1, shuC'* and *shuC2* from *S.* Typhimurium. However, these were only present in a small subpopulation of pETEC746 plasmids and have been discounted from the complete finished sequence.

Conclusion

This study provides a genomic context for the vast amount of experimental and epidemiological data published thus far and provides a template for future diagnostic and intervention strategies. Evidence presented here suggests the prototypical ETEC isolate *E. coli* H10407 was a commensal isolate that acquired a number of plasmids containing a limited repertoire of virulence genes and thereby gained the ability to cause disease. Furthermore, comparisons of the genetic content of *E. coli* H10407 with other ETEC strains reveals only a limited number of conserved genes suggesting that to become pathogenic *E. coli* need only acquire (i) toxins (ST, LT or both) to elicit net secretion from enterocytes, (ii) a fimbrial system that mediates attachment to the intestinal epithelium e.g. CFA/I, and (iii) a novel Type I secretion system the substrate of which (CexE) maintains the fimbriae in the correct physical organisation. This data suggests ETEC vaccine strategies should focus on these plasmid-encoded virulence factors. However, given the relative plasticity of the *E.*

coli genome molecular epidemiological studies are essential to determine whether these factors are widely distributed amongst ETEC strains from geographically diverse locations.

ACKNOWLEDGEMENTS

This work was supported by project grant BB/C510075/1 from the BBSRC to IRH, MJP, CWP, JP and NRT.

REFERENCES

1. **Benjelloun-Touimi, Z., P. J. Sansonetti, and C. Parsot. 1995. SepA, the major extracellular protein of Shigella flexneri: autonomous secretion and involvement in tissue invasion. Mol Microbiol 17:123-35.**
2. **Brown, N. L., S. R. Barrett, J. Camakaris, B. T. Lee, and D. A. Rouch. 1995. Molecular genetics and transport analysis of the copper-resistance determinant (pco) from Escherichia coli plasmid pRJ1004. Mol Microbiol 17:1153-66.**
3. **Chang, D. E., D. J. Smalley, D. L. Tucker, M. P. Leatham, W. E. Norris, S. J. Stevenson, A. B. Anderson, J. E. Grissom, D. C. Laux, P. S. Cohen, and T. Conway. 2004. Carbon nutrition of Escherichia coli in the mouse intestine. Proc Natl Acad Sci U S A 101:7427-32.**
4. **Chaudhuri, R. R., M. Sebaihia, J. L. Hobman, M. A. Webber, D. L. Leyton, M. D. Goldberg, A. F. Cunningham, A. Scott-Tucker, P. R. Ferguson, C. M. Thomas, G. Frankel, C. M. Tang, E. G. Dudley, I. S. Roberts, D. A. Rasko, M. J. Pallen, J. Parkhill, J. P. Nataro, N. R. Thomson, and I. R. Henderson. 2010. Complete genome sequence and comparative metabolic profiling of the prototypical enteroaggregative Escherichia coli strain 042. PLoS One 5:e8801.**
5. **Chen, Q., S. J. Savarino, and M. M. Venkatesan. 2006. Subtractive hybridization and optical mapping of the enterotoxigenic Escherichia coli H10407 chromosome: isolation of unique sequences and demonstration of significant similarity to the chromosome of E. coli K-12. Microbiology 152:1041-54.**
6. **Coster, T. S., M. K. Wolf, E. R. Hall, F. J. Cassels, D. N. Taylor, C. T. Liu, F. C. Trespalacios, A. DeLorimier, D. R. Angleberger, and C. E. McQueen. 2007. Immune response, ciprofloxacin activity, and gender differences after human experimental challenge by two strains of enterotoxigenic Escherichia coli. Infect Immun 75:252-9.**
7. **Cravioto, A. 1980. Ph.D. Thesis. University of London, London.**
8. **Dallas, W. S. 1990. The heat-stable toxin I gene from Escherichia coli 18D. J Bacteriol 172:5490-3.**
9. **de Haan, L. A., G. A. Willshaw, B. A. van der Zeijst, and W. Gaastra. 1991. The nucleotide sequence of a regulatory gene present on a plasmid in an enterotoxigenic Escherichia coli strain of serotype O167:H5. FEMS Microbiol Lett 67:341-6.**
10. **Delcher, A. L., K. A. Bratke, E. C. Powers, and S. L. Salzberg. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23:673-9.**
11. **Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg. 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27:4636-41.**
12. **Duriez, P., O. Clermont, S. Bonacorsi, E. Bingen, A. Chaventre, J. Elion, B. Picard, and E. Denamur. 2001. Commensal Escherichia coli isolates are phylogenetically distributed among geographically distinct human populations. Microbiology 147:1671-6.**
13. **Elsinghorst, E. A., and D. J. Kopecko. 1992. Molecular cloning of epithelial cell invasion determinants from enterotoxigenic Escherichia coli. Infect Immun 60:2409-17.**
14. **Elsinghorst, E. A., and J. A. Weitz. 1994. Epithelial cell invasion and adherence directed by the enterotoxigenic Escherichia coli tib locus is associated with a 104-kilodalton outer membrane protein. Infect Immun 62:3463-71.**
15. **Emanuelsson, O., S. Brunak, G. von Heijne, and H. Nielsen. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2:953-71.**
16. **Enne, V. I., P. M. Bennett, D. M. Livermore, and L. M. Hall. 2004. Enhancement of host fitness by the sul2-coding plasmid p9123 in the absence of selective pressure. J Antimicrob Chemother 53:958-63.**
17. **Evans, D. G., T. K. Satterwhite, D. J. Evans, Jr., and H. L. DuPont. 1978. Differences in serological responses and excretion patterns of volunteers challenged with**

enterotoxigenic Escherichia coli with and without the colonization factor antigen. Infect Immun 19:883-8.

18. Evans, D. G., R. P. Silver, D. J. Evans, Jr., D. G. Chase, and S. L. Gorbach. 1975. Plasmid-controlled colonization factor associated with virulence in Esherichia coli enterotoxigenic for humans. Infect Immun 12:656-67.

19. Evans, D. J., Jr., and D. G. Evans. 1973. Three characteristics associated with enterotoxigenic Escherichia coli isolated from man. Infect Immun 8:322-8.

20. Fleckenstein, J. M., P. R. Hardwidge, G. P. Munson, D. A. Rasko, H. Sommerfelt, and H. Steinsland. Molecular mechanisms of enterotoxigenic Escherichia coli infection. Microbes Infect 12:89-98.

21. Fleckenstein, J. M., D. J. Kopecko, R. L. Warren, and E. A. Elsinghorst. 1996. Molecular characterization of the tia invasion locus from enterotoxigenic Escherichia coli. Infect Immun 64:2256-65.

22. Fleckenstein, J. M., L. E. Lindler, E. A. Elsinghorst, and J. B. Dale. 2000. Identification of a gene within a pathogenicity island of enterotoxigenic Escherichia coli H10407 required for maximal secretion of the heat-labile enterotoxin. Infect Immun 68:2766-74.

23. Fleckenstein, J. M., K. Roy, J. F. Fischer, and M. Burkitt. 2006. Identification of a two-partner secretion locus of enterotoxigenic Escherichia coli. Infect Immun 74:2245-58.

24. Francia, M. V., A. Varsaki, M. P. Garcillan-Barcia, A. Latorre, C. Drainas, and F. de la Cruz. 2004. A classification scheme for mobilization regions of bacterial plasmids. FEMS Microbiol Rev 28:79-100.

25. Friehs, K. 2004. Plasmid copy number and plasmid stability. Adv Biochem Eng Biotechnol 86:47-82.

26. Frishman, D., A. Mironov, H. W. Mewes, and M. Gelfand. 1998. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. Nucleic Acids Res 26:2941-7.

27. Froehlich, B., E. Holtzapple, T. D. Read, and J. R. Scott. 2004. Horizontal transfer of CS1 pilin genes of enterotoxigenic Escherichia coli. J Bacteriol 186:3230-7.

28. Froehlich, B., J. Parkhill, M. Sanders, M. A. Quail, and J. R. Scott. 2005. The pCoo plasmid of enterotoxigenic Escherichia coli is a mosaic cointegrate. J Bacteriol 187:6509-16.

29. Gomez-Duarte, O. G., S. Chattopadhyay, S. J. Weissman, J. A. Giron, J. B. Kaper, and E. V. Sokurenko. 2007. Genetic diversity of the gene cluster encoding longus, a type IV pilus of enterotoxigenic Escherichia coli. J Bacteriol 189:9145-9.

30. Gupta, A., K. Matsui, J. F. Lo, and S. Silver. 1999. Molecular basis for resistance to silver cations in Salmonella. Nat Med 5:183-8.

31. Harrington, S. M., J. Sheikh, I. R. Henderson, F. Ruiz-Perez, P. S. Cohen, and J. P. Nataro. 2009. The Pic protease of enteroaggregative Escherichia coli promotes intestinal colonization and growth in the presence of mucin. Infect Immun 77:2465-73.

32. Henderson, I. R., F. Navarro-Garcia, and J. P. Nataro. 1998. The great escape: structure and function of the autotransporter proteins. Trends Microbiol 6:370-8.

33. Iguchi, A., N. R. Thomson, Y. Ogura, D. Saunders, T. Ooka, I. R. Henderson, D. Harris, M. Asadulghani, K. Kurokawa, P. Dean, B. Kenny, M. A. Quail, S. Thurston, G. Dougan, T. Hayashi, J. Parkhill, and G. Frankel. 2009. Complete genome sequence and comparative genome analysis of enteropathogenic Escherichia coli O127:H6 strain E2348/69. J Bacteriol 191:347-54.

34. Johnson, T. J., and L. K. Nolan. 2009. Pathogenomics of the virulence plasmids of Escherichia coli. Microbiol Mol Biol Rev 73:750-74.

35. Kaper, J. B., J. P. Nataro, and H. L. Mobley. 2004. Pathogenic Escherichia coli. Nat Rev Microbiol 2:123-40.

36. Komano, T., S. R. Kim, and T. Yoshida. 1995. Mating variation by DNA inversions of shufflon in plasmid R64. Adv Biophys 31:181-93.

37. Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567-80.

38. McHugh, G. L., R. C. Moellering, C. C. Hopkins, and M. N. Swartz. 1975. Salmonella typhimurium resistant to silver nitrate, chloramphenicol, and ampicillin. Lancet 1:235-40.

39. Nataro, J. P., and J. B. Kaper. 1998. Diarrheagenic Escherichia coli. Clin Microbiol Rev 11:142-201.

40. Nishi, J., J. Sheikh, K. Mizuguchi, B. Luisi, V. Burland, A. Boutin, D. J. Rose, F. R. Blattner, and J. P. Nataro. 2003. The export of coat protein from enteroaggregative Escherichia coli by a specific ATP-binding cassette transporter system. J Biol Chem 278:45680-9.

41. Ochi, S., T. Shimizu, K. Ohtani, Y. Ichinose, H. Arimitsu, K. Tsukamoto, M. Kato, and T. Tsuji. 2009. Nucleotide sequence analysis of the enterotoxigenic Escherichia coli Ent plasmid. DNA Res 16:299-309.

42. Patel, S. K., J. Dotson, K. P. Allen, and J. M. Fleckenstein. 2004. Identification and molecular characterization of EatA, an autotransporter protein of enterotoxigenic Escherichia coli. Infect Immun 72:1786-94.

43. Pilonieta, M. C., M. D. Bodero, and G. P. Munson. 2007. CfaD-dependent expression of a novel extracytoplasmic protein from enterotoxigenic Escherichia coli. J Bacteriol 189:5060-7.

44. Rasko, D. A., M. J. Rosovitz, G. S. Myers, E. F. Mongodin, W. F. Fricke, P. Gajer, J. Crabtree, M. Sebaihia, N. R. Thomson, R. Chaudhuri, I. R. Henderson, V. Sperandio, and J. Ravel. 2008. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol 190:6881-93.

45. Ren, C. P., R. R. Chaudhuri, A. Fivian, C. M. Bailey, M. Antonio, W. M. Barnes, and M. J. Pallen. 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 186:3547-60.

46. Roy, K., G. M. Hilliard, D. J. Hamilton, J. Luo, M. M. Ostmann, and J. M. Fleckenstein. 2009. Enterotoxigenic Escherichia coli EtpA mediates adhesion between flagella and host cells. Nature 457:594-8.

47. Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream, and B. Barrell. 2000. Artemis: sequence visualization and annotation. Bioinformatics 16:944-5.

48. Sammut, S. J., R. D. Finn, and A. Bateman. 2008. Pfam 10 years on: 10,000 families and still growing. Brief Bioinform 9:210-9.

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

50. Scotland, S. M., N. P. Day, and B. Rowe. 1983. Acquisition and maintenance of enterotoxin plasmids in wild-type strains of Escherichia coli. J Gen Microbiol 129:3111-20.

51. Selvaratnam, S., and M. A. Gealt. 1993. Transcription of ColE1Ap mbeC induced by conjugative plasmids from twelve different incompatibility groups. J Bacteriol 175:6982-7.

52. Sheikh, J., J. R. Czeczulin, S. Harrington, S. Hicks, I. R. Henderson, C. Le Bouguenec, P. Gounon, A. Phillips, and J. P. Nataro. 2002. A novel dispersin protein in enteroaggregative Escherichia coli. J Clin Invest 110:1329-37.

53. Silver, S., A. Gupta, K. Matsui, and J. F. Lo. 1999. Resistance to ag(i) cations in bacteria: environments, genes and proteins. Met Based Drugs 6:315-20.

54. Smyth, C. J. 1982. Two mannose-resistant haemagglutinins on enterotoxigenic Escherichia coli of serotype O6:K15:H16 or H-isolated from travellers' and infantile diarrhoea. J Gen Microbiol 128:2081-96.

55. Steffen, R., F. Castelli, H. Dieter Nothdurft, L. Rombo, and N. Jane Zuckerman. 2005. Vaccination against enterotoxigenic Escherichia coli, a cause of travelers' diarrhea. J Travel Med 12:102-7.

56. Summers, D. K., and D. J. Sherratt. 1985. Bacterial plasmid stability. Bioessays 2:209-211.

57. Svennerholm, A. M., and C. Ahren. 1982. Serological subtypes of Escherichia coli colonization factor antigen II. Eur J Clin Microbiol 1:107-11.

58. Tacket, C. O., R. H. Reid, E. C. Boedeker, G. Losonsky, J. P. Nataro, H. Bhagat, and R. Edelman. 1994. Enteral immunization and challenge of volunteers given enterotoxigenic E. coli CFA/II encapsulated in biodegradable microspheres. Vaccine 12:1270-4.

59. Tenaillon, O., D. Skurnik, B. Picard, and E. Denamur. The population genetics of commensal Escherichia coli. Nat Rev Microbiol 8:207-17.

60. Turner, A. K., T. D. Terry, D. A. Sack, P. Londono-Arcila, and M. J. Darsley. 2001. Construction and characterization of genetically defined aro omp mutants of enterotoxigenic Escherichia coli and preliminary studies of safety and immunogenicity in humans. Infect Immun 69:4969-79.

61. Turner, S. M., R. R. Chaudhuri, Z. D. Jiang, H. DuPont, C. Gyles, C. W. Penn, M. J. Pallen, and I. R. Henderson. 2006. Phylogenetic comparisons reveal multiple acquisitions of the toxin genes by enterotoxigenic Escherichia coli strains of different evolutionary lineages. J Clin Microbiol 44:4528-36.

62. Turner, S. M., A. Scott-Tucker, L. M. Cooper, and I. R. Henderson. 2006. Weapons of mass destruction: virulence factors of the global killer enterotoxigenic Escherichia coli. FEMS Microbiol Lett 263:10-20.

63. van der Woude, M. W., and I. R. Henderson. 2008. Regulation and Function of Ag43 (Flu). Annu Rev Microbiol 62:153-169.

64. Wenneras, C., and V. Erling. 2004. Prevalence of enterotoxigenic Escherichia coli-associated diarrhoea and carrier state in the developing world. J Health Popul Nutr 22:370-82.

65. Yamamoto, T., T. Tamura, and T. Yokota. 1984. Primary structure of heat-labile enterotoxin produced by Escherichia coli pathogenic for humans. J Biol Chem 259:5037-44.

66. Yamamoto, T., and T. Yokota. 1980. Cloning of deoxyribonucleic acid regions encoding a heat-labile and heat-stable enterotoxin originating from an enterotoxigenic Escherichia coli strain of human origin. J Bacteriol 143:652-60.

Table 1. General characteristics of three sequenced *E. coli* chromosomes

Strain	H10407	K-12	HS
Etiology	Pathogen	Lab strain	Commensal
Length (bp)	5153435	4643538	4686137
GC content	50.8%	50.8%	50.8%
Total CDS	4746	4384	4200
tRNA	87	86	86
rRNA	7	7	7

Table 2. General characteristics of the plasmids from ETEC strains H10407 and E1392/75

Strain	*E. coli* H10407				*E. coli* E1392/75				
Plasmid	pETEC948	pETEC666	pETEC58	pETEC52	pETEC1018	pETEC746	pETEC557	pETEC75	pETEC62
Accession No.	FN649418	FN649417	FN649416	FN649415	FN822745	FN822748	FN822746	FN822749	FN822747
Size (bp)	94797	66681	5800	5175	101857	74575	55709	7497	6222
CDS	115	88	7	6	165	117	73	9	13
rep	RepFIIA	RepFIIA	ColE2	ColE1	RepFIIA	RepI1	RepFIB/RepI1	ColE1	ND[a]
Stability genes	StbAB, PsiAB, SopAB, YacAB, RelE	StbAB PsiAB Mok/Hok			StbAB PsiAB CcdAB	StbAB NikAB	SopAB PsiAB		
Insertion elements	IS1, IS2, IS3, IS66, IS91, IS100, IS629, IS911, IS1414, ISEc10, ISEc12, ISSfl4, Tn3	IS1, IS21, IS66, IS600, IS1294, ISEc8			IS1, IS2, IS3, IS21, IS30, IS66, IS91, IS100, IS629, IS630, IS639, IS911, IS1414, ISShdy1	IS2, IS100, IS186, IS1328	IS1, IS30, IS66, IS100, IS911, ISShdy1	IS100	ISCR2

*ND: not determined. pETEC62 has a gene conserved amongst many small plasmids which is annotated as a "probable replication initiation protein" but no experimental evidence exists for this function.

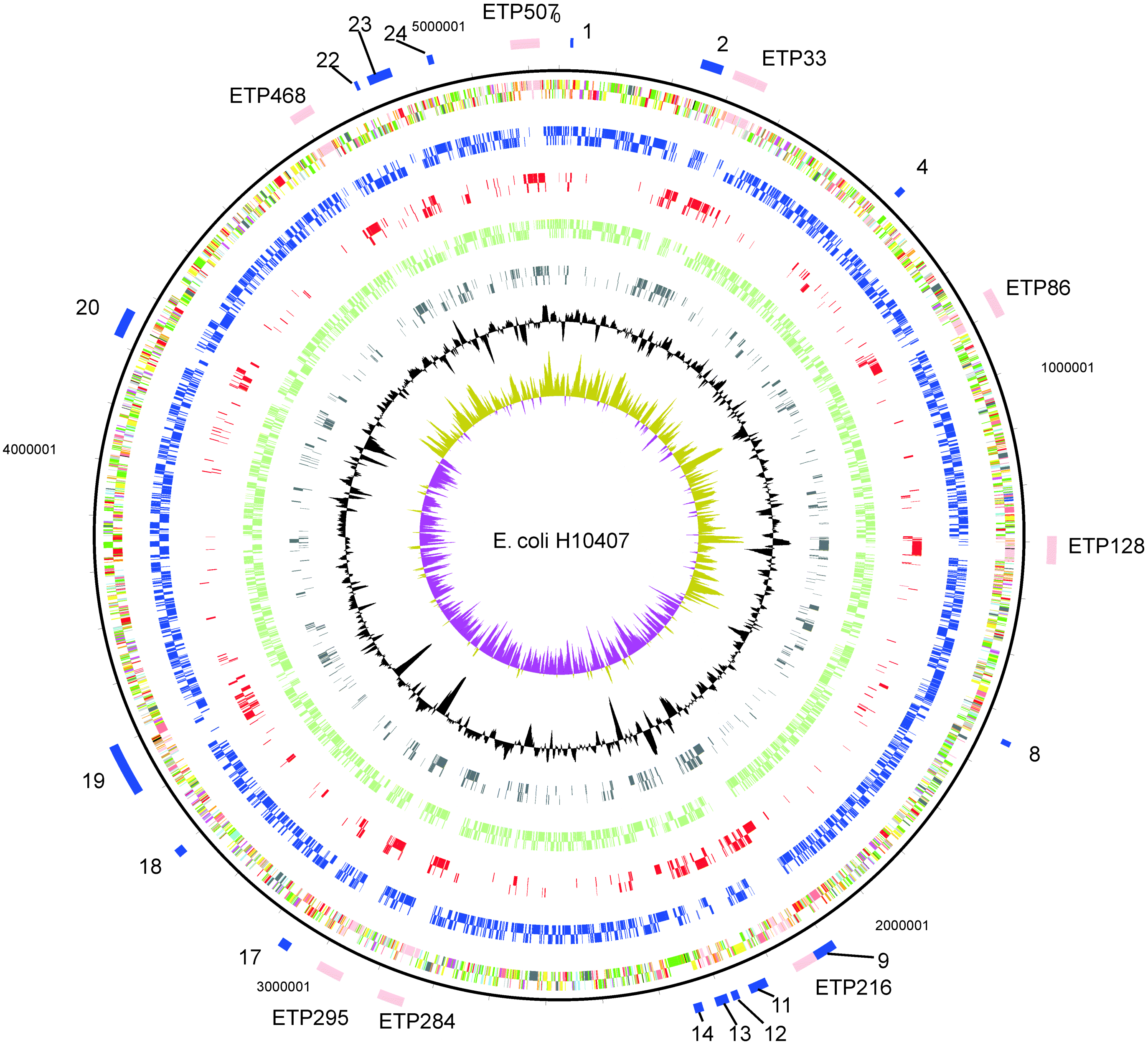
Figure 1. Circular representation of the *E. coli* H10407 chromosome. From the outside in the outer circle 1 marks the position of regions of difference (mentioned in the text) including prophage (light pink) as well as regions differentially present in other *E. coli* strains: blue (See table S1). Circle 2 shows the size in bps. Circles 3 and 4 show the position of CDSs transcribed in a clockwise and anticlockwise direction, respectively (for colour codes see below). Genes in circles 3 and 4 are colour coded according to the function of their gene products: dark green=membrane or surface structures, yellow=central or intermediary metabolism, cyan=degradation of macromolecules, red=information transfer/cell division, cerise =degradation of small molecules, pale blue =regulators, Salmon pink=pathogenicity or adaptation, black=energy metabolism, orange=conserved hypothetical, pale green=unknown, brown=pseudogenes. Circles 5 & 6 and 9 &10 show the position of *E. coli* H10407 genes which have orthologues (by reciprocal FASTA analysis) in *E. coli*: K-12 MG1655 (blue) or *E. coli* 042 (green), respectively. Circles 7 & 8 and 11 & 12 show the position of genes unique to *E. coli* H10407 compared to *E. coli* K-12 MG1655 (red) or *E. coli* 042 (grey), respectively. Circle 13 shows a plot of G+C content (in a 10 Kb window). Circle 14 shows a plot of GC skew ([G-C]/[G+C]; in a 10 Kb window).

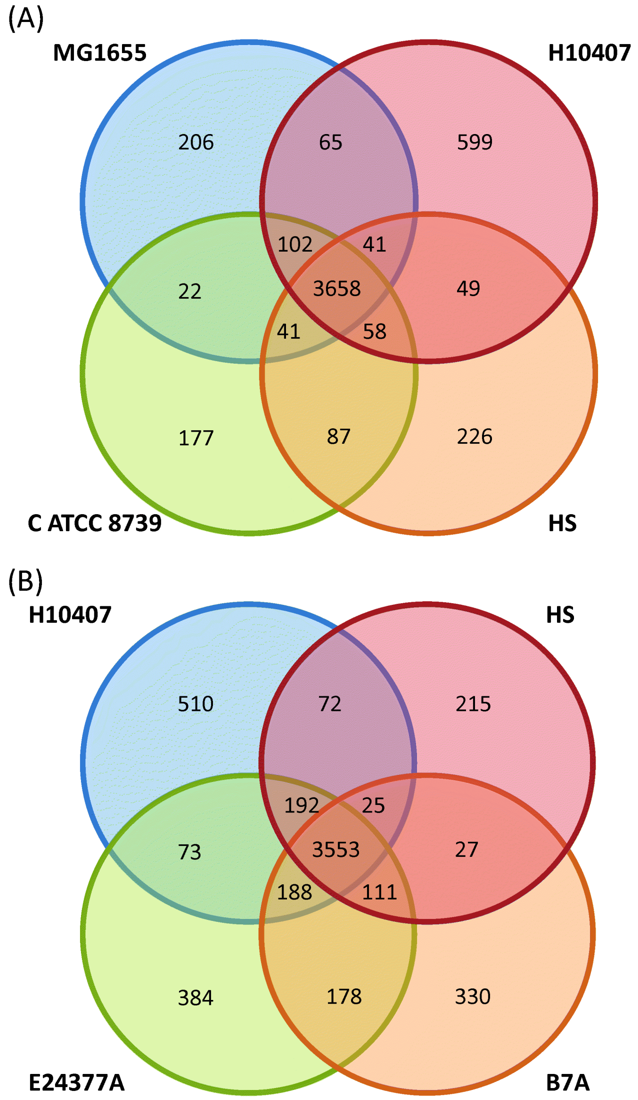
Figure 2. Comparison of the genetic content of *E. coli* H10407 chromosome with the chromosomes of other sequenced strains of *E. coli*. (A) Comparison of *E. coli* H10407 with the three non-pathogenic *E. coli* strains HS, C and K-12 reveals the four strains share a large proportion of common genes. Only 599 *E. coli* H10407 specific genes were identified. The *E. coli* H10407 specific CDS are not thought to be associated with virulence (see text for details). (B) Comparison of *E. coli* H10407 with the genome sequenced ETEC strains E24377A and B7A. The four strains possess 3553 genes in common however the ETEC strains share only 188 genes not present in the commensal strain *E. coli* HS. However, these latter genes are not unique to ETEC and are widely distributed amongst *E. coli* and are largely present among non-pathogenic strains of *E. coli* such as *E. coli* K-12

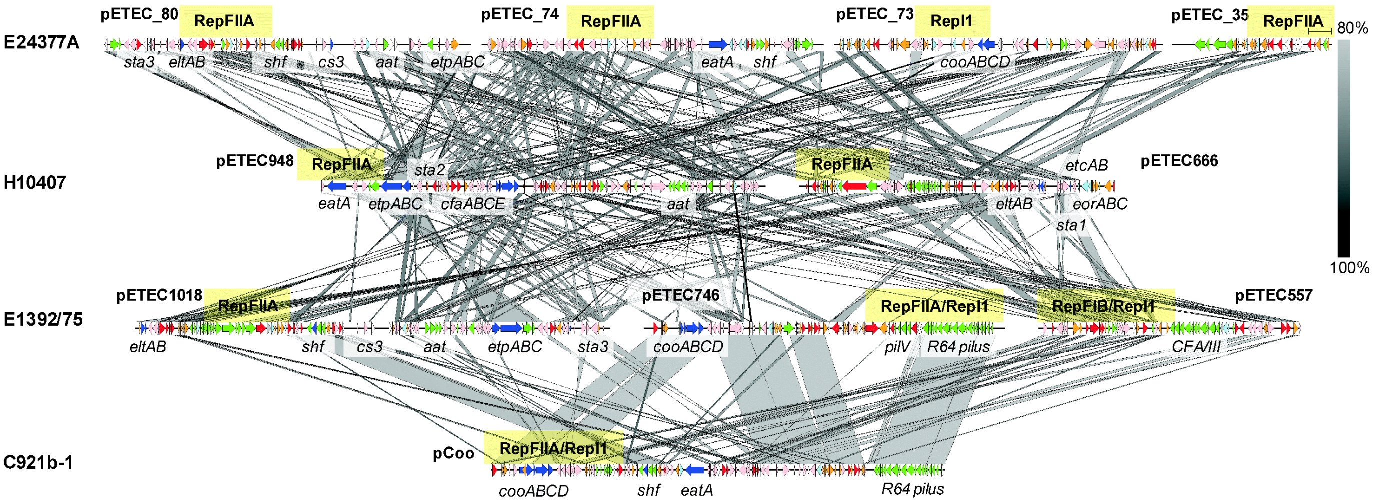
Figure 3. Nucleotide sequence comparison of large conjugative-like plasmids from ETEC strains. Plasmid sequences from each strain were concatenated and compared using BLASTn. BLAST matches longer than 250 bp are shown as grey blocks in a comparison between plasmids from E24377A (pETEC_80, pETEC_74, pETEC_73 and pETEC_35), H10407 (pETEC948 and pETEC666), E1392/75 (pETEC1018, pETEC746 and pETEC557) and C921b-1 (pCoo). Shading of the grey blocks is proportional to the BLAST match (minimum = 80% nucleotide identity, maximum = 100% nucleotide identity). Each plasmid is denoted as a linear black line, the identity of each plasmid is noted above the line and the source ETEC strain from which the plasmids are derived is given on the left side of the figure. Coding sequences are depicted by arrows and are coloured according to known or predicted function: blue, virulence-related; red, plasmid-related protein; green, outer membrane-related (includes conjugal transfer loci); pink, transposase/insertion element-related; light blue, regulatory protein; orange, conserved hypothetical protein; uncoloured, hypothetical protein. The position of genes encoding known or predicted virulence-related proteins is denoted by white boxes harbouring the gene names. In addition, the locus encoding the R64 conjugative pilus and the variant PilV tips is also depicted. The putative origin of replication associated with each of the plasmids is highlighted within yellow shaded boxes. The chimeric nature of the plasmids is clearly visible with recombination between plasmids a frequent occurrence. The unlabeled figure was prepared using a custom script (Sullivan MJ and SA Beatson, unpublished).

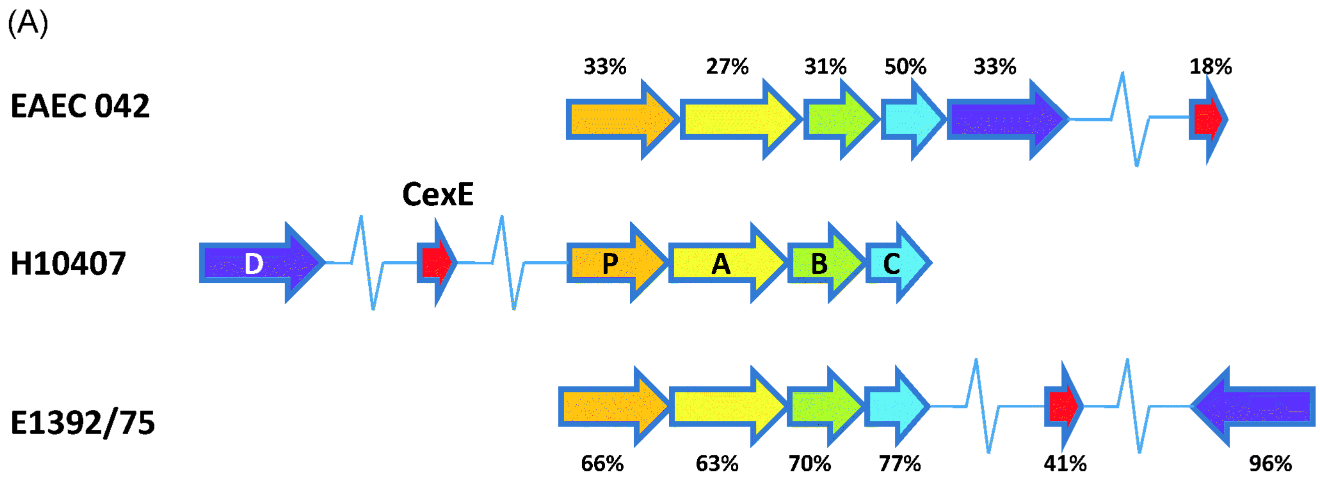
Figure 4. Comparison of the EAEC *aat-aap* locus with the *aat-cexE* loci of ETEC strains. (A) The genetic organisation of the *aat* and *cexE* loci is depicted. The level of amino acid identity for each component of the *aat-cexE* system is shown; figures represent comparison with the *E. coli* H10407 orthologues. Orthologues are coloured coded for ease of identification. Genes which are not juxtaposed are depicted with a blue line separating them. (B) Amino acid sequence alignment of ETEC CexE with the EAEC 042 dispersin. All three proteins possess a signal sequence which is cleaved after the amino acid at position 21 in the alignment. There is limited conservation in the sequences however two cysteine residues which are disulphide bonded in dispersin are conserved. Based on the structure of dispersin, the remainder of the conserved residues appear to represent hydrophobic core residues required for structural integrity of the molecule.

Figure 5. Arrangement of *pilV* shufflon region of *E. coli* E1392/75 pETEC746. Annotation of *pilV* region shown using the Artemis sequence viewer (1). Sequence blocks encoding C-terminal fragments of PilV are found in both orientations between *pilV* and the *rci* recombinase. Identical 13 bp repeats (gtgccaatccggt) are shown as miscellaneous features and mark the predicted sites of recombination between the C-terminal fragments and the *pilV* gene.


ETP507
0
1
23
24
5000001
22
2
ETP33
ETP468
4
ETP86
20
1000001
4000001
E. coli H10407
ETP128
19
8
18
2000001
17
9
ETP216
3000001
11
ETP295
ETP284
14
13
12

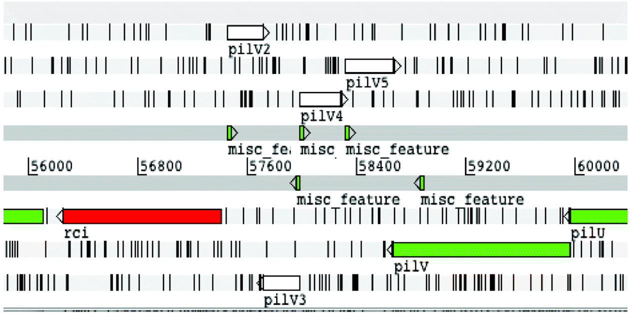

(A)
MG1655
H10407
206
65
599
102
41
22
3658
49
41
58
177
87
226
C ATCC 8739
HS
(B)
H10407
HS
510
72
215
192
25
73
3553
27
188
111
384
178
330
E24377A
B7A


E24377A
pETEC_80
RepFIIA
sta3
eltAB
shf
cs3
aat
etpABC
pETEC_74
RepFIIA
eatA
shf
pETEC_73
RepI1
cooABCD
pETEC_35
RepFIIA
80%
100%
H10407
pETEC948
RepFIIA
sta2
eatA
etpABC
cfaABCE
aat
RepFIIA
etcAB
pETEC666
eltAB
eorABC
sta1
E1392/75
pETEC1018
RepFIIA
eltAB
shf
cs3
aat
etpABC
sta3
pETEC746
cooABCD
RepFIIA/RepI1
pilV
R64 pilus
RepFIB/RepI1
pETEC557
CFA/III
C921b-1
pCoo
RepFIIA/RepI1
cooABCD
shf
eatA
R64 pilus



(B)

```
                     10        20        30        40        50        60
            ....|....|....|....|....|....|....|....|....|....|....|....|
H10407      MKKYILGVI--LAMGSLSAIAGGGNSERPPSVAAGECVTFNSKLGEIGGYSWKYSNDACN
E1392/75    MKKIILALP--FLTSCFSAFAGGSGPEWQPQISPGQCIQY-TEIGETGGYKWHN-IDACN
042         MKKIKFVIFSGILGISLNAFAGGSGWN-ADNVDPSQCIKQ-----SGVQYTYNSGVSVCM
            ***  : :   :   .:.*:***.. :   .: ..:*:       .   *.::   ..*
                     70        80        90       100       110       120
            ....|....|....|....|....|....|....|....|....|....|....|..
H10407      ETVAKGYAIGVAMHRTVNYEGGYSIQSSGIVKPGSDFIMKGGKTYKGHKKVSAGGDTPYWYK
E1392/75    EVVHRGYASGAFVSGKVVYEGGETIEYTGIVKPDAPYTIQAPSTHNGKKKVGHGGAYTYWAR
042         QGLNEGKVRGVSVSGVFYYNDGTTSNFKGVVTPSTPVNTNQDINKTNKVGVQKYRALTEWVK
            : : .* . *. :   . *:.* : : .*:*.*.:    :   . ..:  *      . * :
```


pilV2
pilV5
pilV4
misc_fe
misc_
misc_feature
56000
56800
57600
58400
59200
60000
misc_feature
misc_feature
rci
pilU
pilV
pilV3