4196–4204 *Nucleic Acids Research, 1998, Vol. 26, No. 18*

The complete DNA sequence and analysis of the large virulence plasmid of *Escherichia coli* O157:H7

Valerie Burland*, Ying Shao, Nicole T. Perna, Guy Plunkett, Heidi J. Sofia[1,+] and Frederick R. Blattner

Laboratory of Genetics, University of Wisconsin, 445 Henry Mall, Madison, WI 53706, USA and
[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health,
Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received May 27, 1998; Revised and Accepted August 3, 1998

DDBJ/EMBL/GenBank accession nos AF074613 and AF080442

ABSTRACT

The complete DNA sequence of pO157, the large virulence plasmid of EHEC strain O157:H7 EDL 933, is presented. The 92 kb F-like plasmid is composed of segments of putative virulence genes in a framework of replication and maintenance regions, with seven insertion sequence elements, located mostly at the boundaries of the virulence segments. One hundred open reading frames (ORFs) were identified, of which 19 were previously sequenced potential virulence genes. Forty-two ORFs were sufficiently similar to known proteins for suggested functions to be assigned, and 22 had no convincing similarity with any known proteins. Of the newly identified genes, an unusually large ORF of 3169 amino acids has a putative cytotoxin active site shared with the large clostridial toxin (LCT) family and proteins such as ToxA and B of *Clostridium difficile*. A conserved motif was detected that links the large ORF and the LCT proteins with the OCH1 family of glycosyltransferases. In the complete sequence, the mosaic form can be observed at the levels of base composition, codon usage and gene organization. Insights were obtained from patterns of DNA composition as well as the pathogenic and 'housekeeping' gene segments. Evolutionary trees built from shared plasmid maintenance genes show that even these genes have heterogeneous origins.

INTRODUCTION

Enterohemorrhagic *Escherichia coli* (EHEC) O157:H7 strains are the predominant causative agent of hemorrhagic colitis, a bloody diarrhea which can lead to potentially lethal renal failure in hemolytic-uremic syndrome (HUS) (1,2). EHEC have become increasingly recognized as an important public health concern world-wide since they were first implicated in these diseases in 1983. The Center for Disease Control and Prevention (CDC) currently estimate that O157:H7 causes 25 000 cases of illness with 100 deaths each year in the US alone (3), and new outbreaks are reported every few months. Some virulence determinants contributing to EHEC infections have been characterized, but clearly, a search for all the underlying genes of EHEC pathogenesis will be facilitated by complete DNA sequencing of the O157:H7 genome, now underway in this laboratory. Since the genome sequence of a benign strain, *E.coli* K-12 is known (4), we are in a position to discover the genetic factors that distinguish the pathogenic *E.coli* genome from the benign, keeping in mind that 'virulence' is multigenic, that determinants may act in concert to produce a pathogenic effect, and that benign strains may also encode parts of pathogenic pathways or mechanisms. Most of the Gram-negative pathogens contain plasmids, which frequently carry virulence determinants. In this paper we present the results of the first part of this project, the complete plasmid sequence.

EHEC infections are characterized by binding of the infecting bacteria to intestinal epithelial cells, inducing structural changes by disruption of actin organization. These changes, and production of the resulting lesions, are known as attachment and effacement, and occur in both EHEC and EPEC (enteropathogenic *E.coli*) infections (5,6; reviewed in 7,8). Several virulence factors contributing to pathogenesis have been identified, including Shiga-like toxins (SLTs) (9), and intimin, essential for the tight binding of bacteria to target cells (10). These genes are encoded on 'pathogenicity islands' (PAIs) in the O157 chromosome. In the case of the SLTII, the island is an integrated phage, 933W (11), and intimin is encoded by the LEE element (8,12), also thought to have been acquired by horizontal transfer from an outside source. The large plasmid pO157 is found in most EHEC O157:H7 strains. Its role in pathogenicity is not clearly defined, partly because of conflicting results from studies in various animal models, none of which consistently displays all the characteristics of infection in humans (7). Several reports do correlate pO157 with virulence, for example, with hemolytic activity (13), and adherence to intestinal cells (14,15). Some pO157 putative virulence genes have already been sequenced. A catalase-peroxidase (16), an extracellular serine protease (17), the four genes of the hemolysin locus (18), and 13 genes comprising a type II secretory system (19) have been reported, though the functions of the protease and the putative secretory system have not been examined. A heat stable enterotoxin, EAST1, is encoded by *astA* on the large plasmid of entero-aggregative *E.coli* where

*To whom correspondence should be addressed. Tel: +1 608 262 2534; Fax: +1 608 263 7459; Email: ecoli@genetics.wisc.edu

[+]Present address: Whitman College, Walla Walla, WA, USA it was discovered. In O157:H7 strains, including EDL933, *astA* was found in two copies on the chromosome (20).

Obtaining the complete DNA sequence of the plasmid may help to raise the level of our understanding of the plasmid itself, and its roles in virulence and gene transfer. The sequence offers a catalog of the potential genes and the opportunity to identify new potential virulence determinants, and may point the way to experimental approaches that address the complexities of *E.coli* pathogenesis.

MATERIALS AND METHODS

Sequencing of pO157

The plasmid was was obtained in host C600 from R. A. Welch. It was initially isolated from EHEC O157:H7 strain EDL933 and marked for genetic experiments with the ampicillin-resistance transposon Tn*801* from RP1 (13,15). It was known in this form as pSK3. To avoid confusion with other unrelated plasmids called pSK3, the name pO157 has been used here. The marker transposon (previously unsequenced) was included as part of the entire plasmid project, then spliced out of the finished pO157 sequence file and submitted to the database separately. The insertion point is included in the plasmid annotation file. Both the plasmid and the transposon sequences were deposited in GenBank, accession numbers AF074613 (pO157) and AF080442 (Tn*801*).

Sequences were initially collected from random library clones in a genomic library of O157:H7 EDL933 containing the natural plasmid unmarked by Tn*801*, and also from a library of purified pO157::Tn*801* DNA. Libraries were prepared (21) in M13 Janus (22). DNA templates were purified from library clones (23) and sequences collected using dye-terminator labelled fluorescent cycle sequencing (Prism™ reagents and ABI377 automated sequencer; Applied Biosystem Division of Perkin-Elmer). After assembly into contigs using DNASTAR software, clones were selected for sequencing from the opposite end by the Janus method (22), followed by primer walking to close the last gap. The accuracy of the sequence is estimated at better than one error in 10 000 bases. Any errors persisting after rigorous efforts to detect and resolve frameshifts are most likely to be base substitutions.

Annotation

We identified the open reading frames (ORFs), promoters, operons, regulatory sites, mobile genetic elements and repeated sequences in the plasmid genome, using DNASTAR software and programs written in the laboratory (4). ORFs of at least 50 amino acids were selected from among all possible ORFs by a combination of codon usage analysis, database matches and presence of upstream elements such as promoter-like sequences and ribosome binding sites. Codon usage statistics were used in Geneplot or GeneQuest (DNASTAR) to help predict ORFs, with codon usage matrices developed by Borodovsky (24). Operons were predicted from the arrangement of ORFs and promoters and should be considered as putative. In the absence of any other information, the translation initiation codon giving the longest coding region was chosen, including GUG and UUG. The latter two codons are becoming more frequently validated, especially for genes that apparently have been horizontally transferred. The DNA sequence was searched against GenBank to identify non-coding features. ORF amino acid sequences were searched against SWISS-PROT 34 and GenPept 104 using the BLOSUM62 matrix and the Smith–Waterman algorithm accelerated by the DeCypher II system (TimeLogic Inc.) to assign known functions or suggest functions for new ORFs (4). The criteria for good matches to database entries for assignment of putative function were >30% similarity over >60% of each sequence. ORFs are labelled with numbers L7001–L7100, clockwise as shown on the map in Figure 1 starting with FinO; they are used in the text to indicate ORFs with only putative functions. We have assigned these labels from a series of unique identifiers for ORFs of O157:H7 EDL933 and its extrachromosomal elements.

RESULTS AND DISCUSSION

Organization and physical properties

The genome of pO157-EDL933, shown in Figure 1, consists of 92 077 bp forming a circular duplex. Since data was collected from both the natural and Tn*801*-containing versions of pO157, we were able to precisely identify the insertion point of the transposon, at position 5143 in the sequence. This interrupts ORF L7011, a homolog of hypothetical ORF 2 in the incFII stability locus of *E.coli* plasmid NR1. The sequence of the transposon was also determined.

We found 100 ORFs, most of them oriented in the same direction. Operon arrangements are indicated in Figure 1. The coding regions account for 87% of the plasmid genome, with the previously known genes accounting for 27% of the coding sequence. Codon adaption index (CAI) was calculated for each ORF according to the method of Sharp and Li (25). This index measures the similarity of codon usage to that of a reference set of *E.coli* K-12 genes. The polar coordinate plot of CAI in Figure 1 is designed to highlight regions of the genome with unusual codon usage, which may signify recent acquisition by horizontal transfer. The polar plot shows groups of genes with similar CAI, that correspond roughly with functionally coherent segments, with insertion sequences occurring often at their borders. This analysis suggested a plasmid of mosaic structure, like that reported for other natural isolates of *E.coli* plasmids (26).

Analysis of base composition showed the G+C content of the plasmid to be significantly lower than in *E.coli* K-12, with A (26.74%), T (25.53%), C (21.97%) and G (25.73%), giving G+C as 47.7%. The rare palindromic tetramer, CTAG, found in K-12 strain MG1655 at only 5% of the frequency predicted from base composition (4) occurs at 8% of predicted frequency in pO157.

A restriction map of the plasmid for seven enzymes (27) was compared with the map predicted from the complete sequence. Several *Bam*HI, *Cla*I and *Eco*RI sites we found in the sequence were not observed by Schmidt *et al*. (27), but would have produced very small restriction fragments which could easily be lost on a mapping gel. Three other missing *Cla*I fragments are explained by *dam* methylase-sensitive sites. The only unexplained difference is that one of the two *Xba*I sites in the digest is not found in the sequence.

New putative virulence gene

It was reported several years ago that the intestinal lesions produced in O157:H7 colitis, and the toxin-mediated damage to intestinal cells in *Clostridium difficile* infections, are similar enough to confuse (28). It was suggested by these authors that O157:H7 toxins were responsible. Intriguingly, the most striking newly sequenced gene in pO157 is an unusually large ORF of 3169 amino acids (L7095) which shows strong sequence similarity in a BLAST search, within the first 700 residues, to the

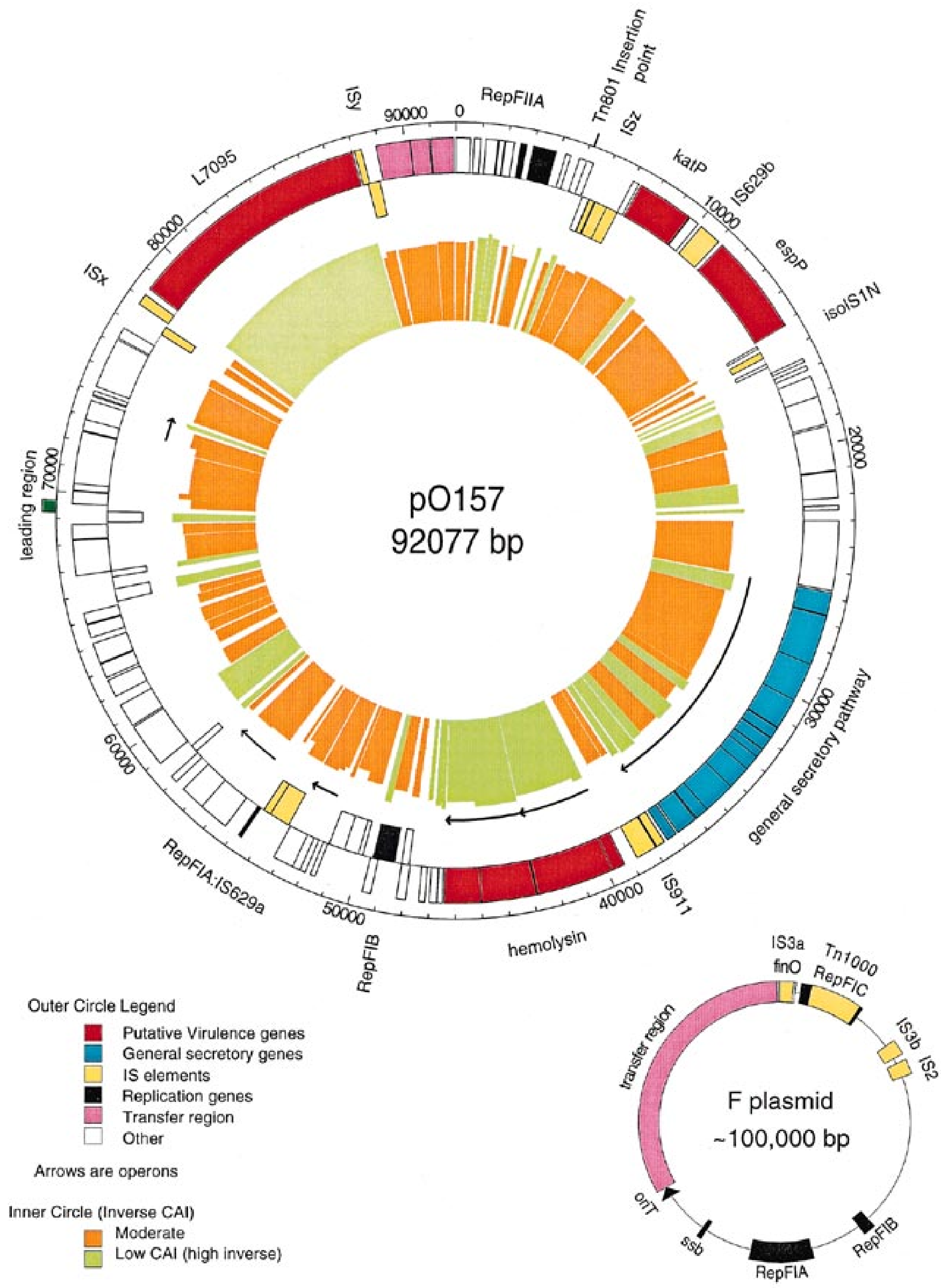
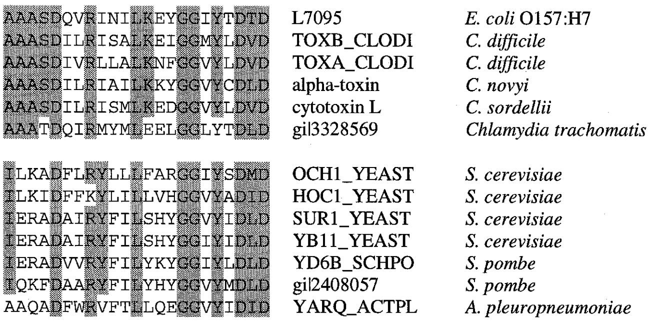


Figure 1. Map of pO157, with small scale map of the F plasmid for comparison. Outer circle of pO157 represents scale in nt. Middle circle shows ORFs color coded for function as indicated by the key. Orientation of transcription is clockwise for the outer ORFs, counter-clockwise for the inner ORFs. Operons are indicated by arrows. Inner circle shows the inverse of the CAI score for each ORF, color coded so that unusually low scores are emphasized (light green). The F plasmid map is approximately scaled (not yet fully sequenced) and color coded for comparison with pO157.

activity-containing N-terminal domain of a toxin family known as the large clostridial toxins (LCT) that includes ToxA (2710 amino acids) and ToxB (2366 amino acids) from *C.difficile*. The four known LCT proteins include *Clostridium novyi* alpha toxin (2178 amino acids) and *Clostridium sordelli* cytotoxin L (2364 amino acids) as well as the two *Clostridium difficile* proteins. These AB cytotoxins share a C-terminal domain that functions in toxin entry into the cell and an N-terminal glucosyltransferase that modifies G proteins such as rho, rac and ras, resulting in the altered regulation of the cytoskeleton, with these two fundamental domains separated by a proposed membrane spanning region. The glucosyltransferase activity of these LCT proteins has been



AAASDQVRINILKEYGGIYTDTD	L7095	*E. coli* O157:H7
AAASDILRISALKEIGGMYLDVD	TOXB_CLODI	*C. difficile*
AAASDIVRLLALKNFGGVYLDVD	TOXA_CLODI	*C. difficile*
AAASDILRIAILKKYGGVYCDLD	alpha-toxin	*C. novyi*
AAASDILRISMLKEDGGVYLDVD	cytotoxin L	*C. sordellii*
AAATDQIRMYMLEELGGLYTDLD	gi|3328569	*Chlamydia trachomatis*
ILKADFLRYLLLFARGGIYSDMD	OCH1_YEAST	*S. cerevisiae*
ILKIDFFKYLILLVHGGVYADID	HOC1_YEAST	*S. cerevisiae*
IERADAIRYFILSHYGGVYIDLD	SUR1_YEAST	*S. cerevisiae*
IERADAIRYFILSHYGGIYIDLD	YB11_YEAST	*S. cerevisiae*
IERADVVRYFILYKYGGIYLDLD	YD6B_SCHPO	*S. pombe*
IQKFDAARYFILYHYGGVYMDLD	gi|2408057	*S. pombe*
AAQADFWRVFTLLQEGGVYIDID	YARQ_ACTPL	*A. pleuropneumoniae*

Figure 2. A conserved motif is shared between L7095 and the LCT proteins and the OCH1 glycosyltransferase family. A Probe search of the nr database derived from GenBank (4.16.98) using the N-terminal domain of L7095 detected a conserved motif that is shared with the four known LCT proteins and the yeast OCH1 family. The sequences shown in the motif are identified by the following accession or GenInfo (gi) numbers and can be located using the SwissProt or GenBank databases: ToxA from *C.difficile* (sp|P16154|TOXA_CLODI); ToxB from *C.difficile* (sp|P18177|TOXB_CLODI); alpha-toxin from *C.novyi* (gi|755724); cytotoxin L from *C.sordellii* (gi|1000695); hypothetical protein from *C.trachomatis* gi|3328569; OCH1 from *Saccharomyces cerevisiae* (sp|P31755|OCH1_YEAST|); HOC1 from *S.cerevisiae* (sp|P47124|HOC1_YEAST); SUR1 from *S.cerevisiae* (sp|P33300|SUR1_YEAST); 44.4 kDa protein from *S.cerevisiae* (sp|P38287|YB11_YEAST); 42.2 kDa protein from *Schizosaccharomyces pombe* (sp|P10323|YD6B_SCHPO); hypothetical protein from *S.pombe* (gi|2408057); 21.6 kDa protein from *Actinobacillus pleuropneumoniae* (sp|P46393|YARQ_ACTPL).

demonstrated biochemically, but they have not been previously described as part of a known glycosyltransferase family based on sequence similarity (29–32). Strong similarity is also seen between the N-terminal domain of L7095 protein over the full extent of a 639 amino acids protein from *Chlamydia trachomatis* (gi|3328569).

A conserved motif from the OCH1 glycosyltransferase family was detected in the N-terminal domains of the L7095 protein and the four LCT toxins using Probe program which performs an iterative profile search (33) (Fig. 2). OCH1 is an alpha-1,6-mannosyltransferase from yeast that functions in mannose outer chain elongation (34). HOC1 (homology to Och1) is expressed in the Golgi apparatus; mutations in this protein show phenotypes typical of defects in cell wall integrity and protein glycosylation (35). The Sur1 protein from yeast is involved in budding, cell polarity and filament formation, and has been shown to be required for mannosylation of inositolphosphorylceramide (36).

Both the OCH1 family and the toxin proteins show an absolutely conserved DXD pattern of aspartates in the motif. Highly conserved aspartate or glutamate residues are expected to be characteristic of glycosyltransferases (37). After submission of this manuscript, work was published showing these two aspartate residues to be essential for the enzymatic activity of the *C.sordelli* cytotoxin L (D286 and D288) (38). The region in the *C.sordelli* cytotoxin L between residues 364 and 516 has recently been shown to play an important role in determining the specificity of the small G protein substrate for the glycosylation reaction (39). These findings are entirely consistent with the putative function of L7095 as a cytotoxin.

Prediction of transmembrane spanning domains using the PHD software on the PredictProtein server (40) showed a single membrane-spanning helix in each of the LCT proteins at ~1100 residues (CdA, 1075..1099; CdB, 1073..1098; CnA, 1081..1105; CsL, 1073..1097) and in the L7095 protein a single predicted transmembrane region is located at 1961..1980. In each case the putative transmembrane domain is contained within a larger region predicted by the Seg program to be non-globular in structure (41,42). In the clostridial proteins the putative transmembrane region is located within a low-complexity region of ~400 residues (CdA, 764..1134; CdB, 722..1131; CnA, 767..1186; CsL, 746..1131). The predicted transmembrane region in the L7095 protein is located within a smaller medium-complexity region of ~40 residues (1942..1981), but an additional low-complexity domain is seen upstream at 1659..1813.

The sequence in the vicinity of L7095 was examined for regulatory elements. A putative promoter was found to overlap the first possible start codon, and the next upstream ORF, a transposase gene in the same orientation, has a good promoter which could conceivably read through L7095. Curiously, the strongest promoter consensus in the region (identical with the sigma 70 consensus sequence) was found in the middle of L7095 oriented with the ORF. No rho-independent transcription terminators were found. Codon usage by L7095 is more like that of genes of phages and mobile elements than *E.coli* genes, while the CAI (Fig. 1) strongly indicates that the ORF did not evolve in *E.coli*.

Previously sequenced virulence genes

The plasmid-encoded hemolysin was the first pO157 sequence to be determined (X86087) (18,27). The term EHEC-hemolysin was used to distinguish it from alpha-hemolysin to which it is related but not identical. EHEC-hemolysin belongs to the RTX (repeats in toxin) family of exoproteins (43,44). Four gene products of the *hlyCABD* operon (L7047–L7050), encode a pore-forming cytolysin and its secretion apparatus. Toxicity results from the insertion of HlyA into the cytoplasmic membrane of the target mammalian cells disrupting permeability control (45). Our sequence of pO157 has several differences from the original determination (X86087) by Schmidt *et al.* (46). However, more recent determinations for *hlyA*, *B* and the 5′ part of *hlyC* (U12572 and Y09824) are in complete agreement with our sequence, including correction of a frameshift error at the start of *hlyC*. There are no other recent determinations for the plasmid-encoded *hlyD*, in which our sequence has four differences from X86087, three resulting in a different amino acid. The product of the *tolC* gene is also required for secretion of this toxin. In O157:H7 EDL933 TolC is encoded on the chromosome (preliminary data, this laboratory).

An EHEC catalase-peroxidase is encoded on pO157 by *katP* (L7017) (16) (X89017), whose product is a bifunctional periplasmic enzyme that protects the bacterium against oxidative stress, one possible hazard of infecting mammalian cells. The occurrence of the *katP* gene in EHEC strains is associated with the EHEC hemolysin operon (16,17). Brunder *et al.* also located an extracellular serine protease gene, *espP* (L7020), on pO157 (X97542). EspP is homologous to EspC (U69128), a secreted protein of EPEC whose role in pathogenicity is unknown. The proteases are Type IV secreted proteins (autotransporters). A domain for effecting translocation across the cell membrane is encoded within the polypeptide. This protein belongs to a family of virulence factors that are secreted in this manner (47). Widely observed in O157 strains, its demonstrated ability to cleave human coagulation factor V probably exacerbates the epithelial degradation and hemorrhage resulting from EHEC infection.

General secretory pathway

A gene cluster closely related to the type II secretion pathway of Gram-negative bacteria has recently been sequenced in pO157

(19); Y09824. The 13 genes of this operon, *etpC* through *etpO* (L7032–L7044), show a high degree of similarity to the genes of the general secretory pathway (GSP) in *Klebsiella pnueumoniae* (M32613), *Aeromonas hydrophyla* (X66504), *Erwinia chrysanthemi* (L02214) and *Erwinia caratova* (X70049), ranging from 40% to >90% identity over at least half of each gene. This pathway provides a mechanism of exoprotein secretion for these and other pathogenic Gram-negative bacteria (48,49). Although the operon appears to contain the necessary genes, its ability to secrete proteins has not yet been demonstrated, nor is its substrate known. Its correlation with disease is also unclear. Curiously, *E.coli* K-12 also has a chromosomally encoded GSP operon, though only two of its genes appear to be expressed in the laboratory strain (50). We have preliminary sequence data showing that this K-12 chromosomal operon is also present in the O157:H7 EDL933 genome. The similarities between the genome-encoded GSP and corresponding genes of the plasmid *etp* operon are in the range 15–60%, i.e. much lower than the matches to operons in other organisms; clearly this operon is more likely to have been acquired from outside than to have derived directly from the K-12 genome. We note that in Y09824, the translation starts chosen for EtpD, I, L and N are AUG codons giving shorter ORFs than those we defined. In these four cases we chose upstream in-frame GUG starts, since the amino acid sequence between the GUG and AUG codons showed a high degree of similarity with the homologous ORFs in *Klebsiella* or *Erwinia*; the similarity scores are, respectively, 65.5% over 57 amino acids, 50% over 6 amino acids, 33.1% over 143 amino acids and 53.9% over 39 amino acids.

Other putative proteins

We found a homolog (L7004; 90% identity, full length alignment) of the *E.coli* regulator Hha, a modulator of hemolysin expression (P23870) (51). Of two ORFs encoded on opposite strands of the same DNA region, L7023 shows 85% similarity to 46 residues of an ORF that was originally identified as PerD (P43476) of EPEC strain E2348/69 (O127:H6). The *per* operon was shown to regulate expression of intimin, an essential factor in the attaching and effacing lesions of EHEC infection (52). The *perD* gene was also reported to show similarity to Tn*7* and IS*630* sequences, though our ORF does not. No role in regulation has been established for this gene, and a recent account of the *per* operon no longer shows the fourth gene (7). No other plasmid ORFs had any match to the other *per* products, and neither did the ORF on the opposite strand show any striking match.

Downstream of the hemolysin operon, two adjacent ORFs showed ~50% similarity with 20 and 42 residues from different parts of a second *E.coli* regulator, PapX, that controls pilus production (P42193). The two ORFs (L7051 and L7052) represent parts of the gene separated by a frameshift. Pili are integral to the recognition and initial attachment of EHECs to the target cells. While there are no obvious pilus components encoded on pO157, two ORFs have some similarity to pilin-related proteins; one (L7081) shows 27% similarity over 67 residues to OrfQ (Q46527), of unidentified function in the *fim* operon of *Bacteroides*, and the second (L7024) shows 49% similarity to 43 amino acids in part of the fimbrial regulator FimB of *E.coli* (P04742).

Two ORFs show similarity to proteins involved in cell membrane or its production. L7029 is 66% identical over 323 amino acids with MsbB, an *E.coli* acyltransferase (P24205). Though not acting on its normal substrate, this protein can take part in LPS production by substituting for HtrB, an acetyltransferase that adds the myristoyl fatty acid to lipid A (53). The myristate component of lipid A, which is the endotoxin portion of LPS, appears to activate and antagonize immune cells (54). L7092 has a region that is 34% similar over 77 amino acids to the functional domain of TolA (P19934), a membrane receptor in *E.coli* for colicins and filamentous phages, though the structural domains do not appear to be present.

Two ORFs are similar to genes of *Vibrio cholerae*. A 324 amino acid segment of L7031 shows 43% identity to 131 amino acids of the C-terminal half of TagA of *V.cholerae* (Q56595). This is a lipoprotein regulated by ToxR (55), thought to be involved in fimbrial biosynthesis because of its similarity to FanD of *E.coli* K99 (P12050) (56), but not essential for virulence. The similarities between L7031, TagA and FanD do not overlap, and neither the signal peptide nor the lipid attachment site of TagA are found in L7031, so no function may be deduced. L7074 is 33% similar over 204 amino acids to hemagglutinin-associated protein in *V.cholerae*, whose function is unknown (Q56638). This pO157 ORF has a homolog in the large virulence plasmid pMT1 of *Yersinia pestis* (Lindler,L.E., Plano.G.V., Burland,V., Mayhew,G.F. and Blattner,F.R., submitted to *Infect. Immunol.*) suggesting a common function. However, both proteins also show similar identities to methylases of several organisms, so a virulence function cannot yet be assigned.

In summary, 19 of the 100 ORFs on the plasmid have confirmed functions, 42 may be assigned putative functions on the basis of their database matches, and 17 have good matches to hypothetical genes in the database. The remaining 22 ORFs had only poor matches to database entries and cannot be assigned functions with any confidence.

Replication regions and plasmid maintenance

Three possible replication origins were found, their structures consistent with pO157 belonging to the incFII group of F plasmids (57,58). In Figure 1, the organization of the the F plasmid is shown for comparison. The origins are compared in more detail with similar replication origins in Figure 3. One putative origin was found with very high similarity to RepFIIA of the incFII plasmid NR1 in both the nt sequence (95% identical) and the translated ORFs (96% to RepA1 and 99% to RepA2). We found a replication origin, oriR, 89% identical to the minimum sequence of the functional origin of pNR1 (X12776) (59).

The second replication origin is 97% similar to the RepFIB replicon region of the enterotoxinogenic (incFI) plasmid p307 (M26308) (60), which differs by only a few bases from that of the F plasmid (61). RepFIB is conserved as a functional replicon in most of the incF plasmids (62). A pO157 ORF (L7056) in this region is 99% identical to an essential replication protein Orf1 of p307 (gi|516615). By comparison with the essential replication region of p307, the 'rep' region of pO157 includes all the essential repeats (BCD and EFG), but lacks 209 bp at the left end (Fig. 3). However, the 141 bp remaining at the left of repeat B could be enough to preserve the functional replication origin.

The third replication origin in pO157 corresponds to the RepFIA region of the F plasmid. It is disrupted in pO157 by insertion of IS*629* (Fig. 3), which replaced the incB repeats and part of *repE* (L7067) in the initiation region. Adjacent genes in this region are homologous to the *ccd* and *sopAB* operons of F (Fig. 3), encoding stable maintenance and partitioning functions,

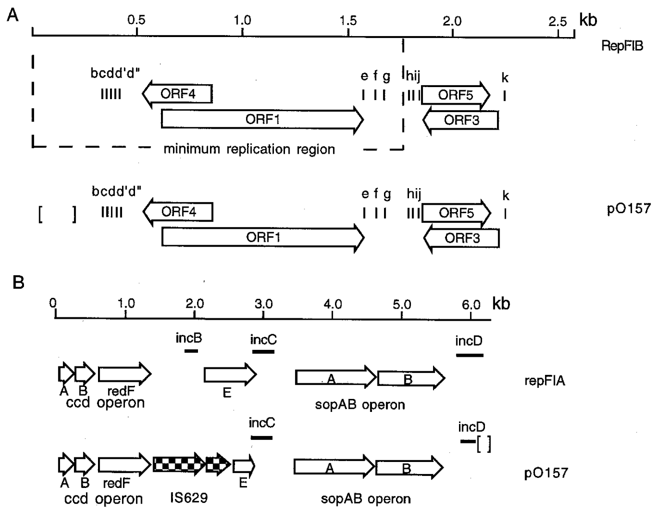


Figure 3. (**A**) Comparison of the RepFIA regions of pO157 and the F plasmid. Deletions are represented by [ ]. In pO157, IS*629* replaces *incB* and the N-terminus of the protein E gene, and part of *incD* is deleted. (**B**) Comparison of the RepFIB regions of pO157 and p307. Short vertical lines represent repeats. 209 bp are deleted from the left end of the minimum replication region, but all the repeats required for function are present.

respectively. Downstream of *sopB*, part of the incD (*sopC*) region is deleted in pO157, but three repeats still remain, with 95–100% identity to incD (M12987) of the F plasmid. Since a single 43 bp repeat of F is sufficient to allow assembly of an active nucleoprotein partition complex (63), it is likely that the *sopAB* operon is also functional. Homologs were found in the RepFIA region, to the *flm* (*stm*) gene cluster of the F plasmid (M97768) as well as the *ccd* operon of F (U51588). These operons both encode ‘toxin/antitoxin’ systems in which the antitoxin encoded by *flmB* is an unstable antisense RNA that prevents expression of the lethal protein FlmA by binding to more stable mRNA (64). The ccd system uses a stable ‘killer’ protein and an unstable regulator protein, sensitive to protease digestion, to achieve the same end (65,66). Such systems occur in many plasmids and genomes (67), to eliminate plasmid-free segregants from the bacterial population. The homology with F through these regions is very high. The one base difference in the *flmA* gene is a third codon position substitution having no effect on the protein. Homologs of other F plasmid genes associated with plasmid maintenance include *psiA* and *B* (P18148, P10031), encoding inhibitors of the SOS cascade, L7084 (*ssbF*, P18310), encoding a single-strand DNA binding protein, and L7091 (*nikB*, P52336), a putative nickase. Two ORFs, L7077 and L7060, showed more distant matches to plasmid R751 genes *klcA* and *kfrA* (P52602 and P71173), also encoding maintenance and partitioning functions. Several other ORFs match hypothetical genes in replication and stability regions.

Transfer region

pO157 was known to be non-conjugative (57). In comparision with F (Fig. 1), a large portion of the transfer region is replaced by two clusters of IS elements and the large ORF L7095. L7099 was found to be a homolog of *traX* (P22709, encoding F-pilin acetylase) and L7001 is a homolog of *finO* (P22707, fertility inhibition protein), and three more ORFs are homologous to *tra* region genes of unknown function (L7096, L7098 and L7100). The leading region is preserved, 98% similar to the corresponding F region. An IS element replaces the N-terminal two-thirds of L7098, a partial homolog of *traI*, which encodes a helicase in F.

Insertion sequence elements

pO157 contains several insertion sequence (IS) elements and related sequences at seven different loci, shown in Figure 1 and summarized in Table 1. Two are intact IS elements, IS*911* adjacent to the hemolysin operon and IS*629*a in RepFIA, with matches to reported IS*911* and IS*629* of 91 and 96% identity, respectively. We found three apparently composite elements or multi-component clusters (details in Table 1), one adjacent to *katP* (ISz), and at each end of L7095 (ISx and ISy). In addition, there are two partial IS elements located between *katP* and *espP* (IS*629*b, 78% identical with the reported IS) and at the other end of *espP* (isoIS*1N*, 94% identical to the reported IS). Both of these encode transposases, the IS*629* protein (L7019) being 96% identical, and that of IsoIS*1N* (L7022), is 84% identical with the corresponding reported proteins, and thus are likely to be functional.

Table 1. IS elements in pO157

IS element	Similarity % (nt)[a]	GenBank accession no.	Location	Coding regions[b]	ORF length (aa)	Inverted repeats
IS*629*a	95.7	X51586	53269..54578	(53311..54201)	296	27 bp
				(54201..54524)	108	
IS*629*b	80.4	X51586	9831..11103	10174..11061	295	one only 27 bp
IS911	90.7	X17613	37521..38772	37557..37898	113	36 bp
				37943..38764	273	
ISx (IS*3*)	51.0	X02311	77032..77764	(77322..77681)	119	none
ISx (IS*21*)	71.2	X14793	77768..78257	77863..78273	136	none
ISy (IS*3*)	91.2	X02311	88137..88348	88143..88415	90	none
ISy (IS*629*)	69.0	X51586	88476..88899	(88342..88845)	167	one only 27 bp
ISz (IS*91*)	59.5	X17114	5661..7215	(5755..5976)	73	one only 8 bp
				(6547..7056)	169	
ISz (IS*600*)	89.1	X05952	5986..6667	(6029..6763)	244	one only 27 bp
isoIS*1N*	77.7	J01737	15462..15926	(15609..15881)	90	none

[a]Similarity to the published nucleotide sequence.
[b]Parentheses denote anticlockwise orientation with respect to the map in Figure 1.

Tn801

This ampicillin resistance transposon is closely related to Tn*3*. On insertion, the 5 bp target site TTATA, different from the Tn*3* target, is duplicated as a direct repeat. The 4949 bp element is 8 bp shorter than Tn*3* (68), and 97.7% similar across the whole element at the nucleotide level. Inverted repeats of 38 bp are at each end, with three genes encoded in between, a transposase (*tnpA*), the ampicillin resistance (β-lactamase, *bla*) and a repressor (*tnpR*), showing the same organization as Tn*3*. The transposase is 14 amino acids shorter at the N-terminus than the Tn*3* protein, the repressor is identical, and β-lactamase is 99% identical. The internal resolution site in the N-terminal portion of *tnpA* is intact and differs from the Tn*3* site by only two out of 23 bp. The site also contains 18 bp identical to a part of the inverted repeats, compared with 19 bp for Tn*3* (69).

Phylogeny

Boyd *et al.* (26) have suggested that the genes of plasmids are more heterogeneous than those of the chromosome, as a result of higher rates of recombination, and frequent plasmid transfer. To understand the relationship between pO157 and other F-like plasmids, and place pO157 in the phylogenic trees constructed by Boyd, we compared their RepA and FinO sequences. The proteins of pO157 were aligned with previously published homologs from F-like plasmids in reference collections of *E.coli* (ECOR) and *Salmonella* (SARA) strains (26,70). Phylogenetic trees, shown in Figure 4, were constructed using a Tamura–Nei distance matrix and the Neighbor Joining tree-building algorithm (71). The pO157 sequences fall within the range of naturally occurring variation for the species. However, the grouping of RepA from pO157 with EC 25A, and FinO of pO157 with SARA 26 is consistent with frequent genetic exchange between *E.coli* and *Salmonella* plasmids. The fact that neither of the pO157 genes cluster with ECOR 37, the closest relative to O157:H7 EDL933 that is included in the tree (72), supports the idea that the genome and plasmid evolved by distinct mechanisms.

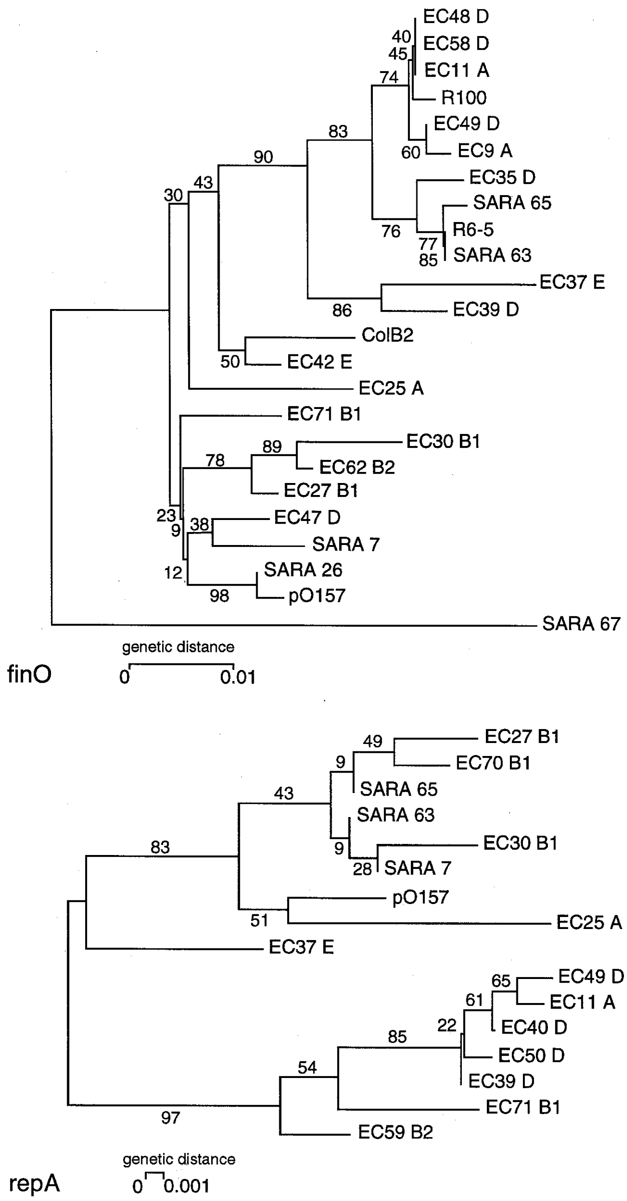


Figure 4. Phylogenetic trees show the relationships among SARA (*Salmonella*) and ECOR *(E.coli)* strains (see text) for RepA and FinO sequences from the F-like plasmids of these strains. Alignments were made by MegAlign (DNASTAR) and the phylogeny was reconstructed by MEGA (75). Genetic distances are maximum likelihood estimates of the number of nucleotide substitutions between two sequences, corrected for multiple substitutions. The numbers shown along the branches indicate the percentage of 2000 bootstrap replicates supporting each cluster. The alignments are available for inspection on request.

Base distribution

The analysis of base distribution revealed two notably AT-rich regions. A 7.7 kb region of 62.6% A+T corresponded to the genes of the hemolysin operon with the adjacent broken reading frames of *papX*. A second region, 9.5 kb of 68.6% A+T, corresponded to ORF L7095 encoding the putative cytotoxin. To observe more closely these compositional shifts in the different segments of pO157, we studied the variation in the third position of the codons of each gene. The third position is less constrained by selection than the first two, and base composition at this position is indicative of rates and patterns of mutation (73). Figure 5 shows the distribution of third position bases plotted for each ORF of the plasmid. A pattern of consistency within segments was revealed, which roughly corresponded to segments of plasmid maintenance genes or virulence determinants, or genes of unknown function. Several of the segments are bounded by IS elements. The pattern of third position base distribution for any particular gene is typical of its segment neighbors, for example, the genes of the Type II secretory pathway (*etp* operon) all use G most frequently in the third position, whereas the *hly* operon, separated from *etp* by IS*911*, use A or T much more frequently than G or C. These variations are consistent with distinct evolutionary histories for each segment.

The complete sequence of this plasmid has revealed a number of potential virulence genes, consistent with a role for the plasmid in the pathogenic mechanism of O157:H7. From this point of view, the most significant discovery in pO157 is the putative new cytotoxin. This molecule potentially represents a new type of toxin since its third domain has no clear resemblance to any of the known motifs that mediate binding to the target cell. The

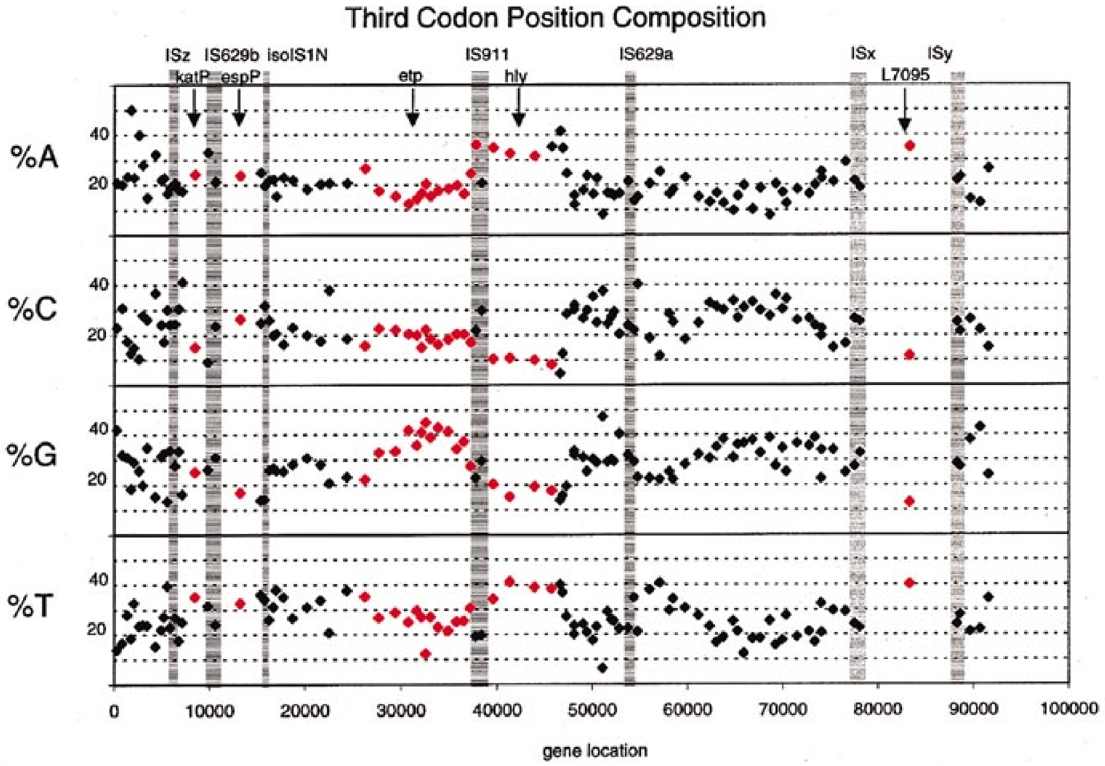


Figure 5. Third codon position composition. For each gene of pO157, the percentage of each base occurring in the third codon positions was calculated, and plotted against the coordinate on the pO157 sequence at the center of each gene. The composition of the *etp* and *hly* operons, and that of L7095 (red symbols) is different from that of adjacent genes. The values for *katP* and *espP* are also shown in red for comparison. The positions of the IS elements are shown by shaded bars.

similarity with the LCTs suggests a common effector mechanism leading to disorganization of the target cell's cytoskeleton. EHEC infection induces morphological changes in intestinal cells, resulting in the characteristic pedestal structure to which the bacteria attach. This is accompanied by changes in actin distribution and leads to new actin structures forming beneath the pedestal. Several genes involved in this process have been identified (reviewed in 8), but the biochemical details are not yet known. Thus, experimental investigation of the L7095 protein may add crucial understanding of EHEC disease.

The emerging picture from partial sequences indicates that Gram-negative plasmid genomes are dynamic, with frequent transfer and recombination events among and between strains and even species. The complete plasmid sequence reported here is consistent with this, and with the notion that extrachromosomal elements such as plasmids and bacteriophages play a significant role in bringing together the virulence factors that make up the disease-potential of the organism. The sequence reveals a backbone of typical F-like replication and maintenance features like those previously reported for plasmids bearing virulence factors and drug resistances (57,58). Interspersed among the plasmid functions are genes and gene groups that have confirmed (e.g. *hlyABCD*) or putative (e.g. L7095) virulence potential, together forming a mosaic of elements of widely varying size assembled in the course of evolution (26), reflecting the situation in the O157:H7 chromosome (74). Although the transfer functions that would have mobilized these elements have been destroyed by insertion of the putative cytotoxin gene, the IS elements may be equally effective in mediating genetic mobility. We are beginning to discern the dynamic exchange of virulence determinants resulting in a reassortment of different pathogenic mechanisms. These observations suggest that the sum of all the pathogenic genes shared among related enteric bacteria form a pool or ‘pathosphere’, and that plasmids and genomes contain genes from this pool assembled by mechanisms that are, in evolutionary terms, relatively fast.

ACKNOWLEDGEMENTS

This is paper 3508 from the Laboratory of Genetics. The work was supported by a research grant from the Ronald McDonald House Charities. N.T.P. is supported by NSF-Sloan Foundation Postdoctoral Fellowship BIR-9626042. We thank D. Rose, G. Mayhew and the technical staff of the Wisconsin Genome Project for help with sequencing, and Dr R. A. Welch for useful discussions and critical reading of the manuscript.

REFERENCES

1 Karmali,M.A., Steele,B.T., Petric,M. and Lim,C. (1983) *Lancet*, **2**, 619–620.
2 Boyce,T.G., Swerdlow,D.L. and Griffin,P.M. (1995) *N. Engl. J. Med*, **333**, 364–368.
3 Wilkinson,S.L. (1997) *Chem. Eng. News*, November 10, 1997, pp. 24–29.

4 Blattner,F.R., Plunkett,G., III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y. (1997) *Science*, **277**, 1453–1474.

5 Levine,M.M. (1987) *J. Infect. Dis*., **155**, 377–389.

6 Moon,H.W., Whipp,S.C., Argenzio,M., Levine,M. and Gianella,R.A. (1983) *Infect. Immun*., **41**, 1340–1351.

7 Nataro,J.P. and Kaper,J.B. (1998) *Clin. Microbiol. Rev*., **11**, 142–201.

8 Kaper,J.B., Elliot,S., Sperandio,V., Perna,N.T., Mayhew,G.F. and Blattner,F.R. (1998) In Kaper,J.B. and O'Brien,A.D. (eds), *Escherichia coli O157:H7 and other Shiga Toxin-Producing E.coli Strains*. American Society for Microbiology, Washington, DC, pp. 163–181.

9 Tesh,V.L. and O'Brien,A.D. (1991) *Mol. Microbiol*., **5**, 1817–1822.

10 McKee,M.L., Melton-Celsa,A.R., Moxley,R.A., Francis,D.H. and O'Brien,A.D. (1995) *Infect. Immunol*., **63**, 3739–3744.

11 O'Brien,A.D., Newland,J.W., Miller,S.F., Holmes,R.K., Smith,H.W. and Formal,S.B. (1984) *Science*, **226**, 694–696.

12 Perna,N.T., Mayhew,G.F., Posfai,G., Elliot,S., Donnenberg,M.S., Kaper,J.B. and Blattner,F.R. (1998) *Infect. Immunol*., **66**, 3810–3818.

13 Schmidt,H., Karch,H. and Beutin,L. (1994) *FEMS Microbiol. Lett*., **117**, 189–196.

14 Fratamico,P.M., Bhaduri,S. and Buchanan,R.L. (1993) *J. Med. Microbiol*., **39**, 371–381.

15 Karch,H., Heesemann,J., Laufs,R., O'Brien,A.D., Tacket,C.O. and Levine,M.M. (1987) *Infect. Immun*., **55**, 455–461.

16 Brunder,W., Schmidt,H. and Karch,H. (1996) *Microbiology*, **142**, 3305–3315.

17 Brunder,W., Schmidt,H. and Karch,H. (1997) *Microbiology*, **24**, 767–778.

18 Schmidt,H., Beutin,L. and Karch,H. (1995) *Infect. Immun*., **63**, 1055–1061.

19 Schmidt,H., Henkel,B. and Karch,H. (1997) *FEMS Microbiol. Lett*., **148**, 265–272.

20 Savarino,S.J., McVeigh,A., Watson,J., Cravioto,A., Molina,J., Echeverria,P., Bhan,M.K., Levine,M.M. and Fasano,A. (1996) *J. Infect. Dis*., **173**, 1019–1022.

21 Mahillon,J., Kirkpatrick,H.A., Kijenski,H.L., Bloch,C.A., Rode,C.K., Mayhew,G.F., Rose,D.J., Plunkett,G., III, Burland,V. and Blattner,F.R. (1998) *Gene*, in press.

22 Burland,V., Daniels,D.L., Plunkett,G., III and Blattner,F.R. (1993) *Nucleic Acids Res*., **21**, 3385–3390.

23 Olson,C.H., Blattner,F.R. and Daniels,D.L. (1991) *Methods*, **3**, 27–32.

24 Borodovsky,M. and McIninch,J. (1993) *Comput. Chem*., **17**, 123–133.

25 Sharp,P.M. and Li,W.H. (1987) *Nucleic Acids Res*., **15**, 1281–1295.

26 Boyd,E.F., Hill,C.W., Rich,S.M. and Hartl,D.L. (1996) *Genetics*, **143**, 1091–1100.

27 Schmidt,H., Kernbach,C. and Karch,H. (1996) *Microbiology*, **142**, 907–914.

28 Griffin,P.M., Olmstead,L.C. and Petras,R.E. (1990) *Gastroenterology*, **99**, 142–149.

29 Just,I., Selzer,J., Wilm,M., von Eichel-Streiber,C., Mann,C. and Aktories,K. (1995) *Nature*, **375**, 500–503.

30 Just,I., Wilm,M., Selzer,J., Rex,G., von Eichel-Streiber,C., Mann,M. and Aktories,K. (1995) *J. Biol. Chem*., **270**, 13932–13936.

31 Hofmann,F., Busch,C., Prepens,U., Just,I. and Aktories,K. (1997) *J. Biol. Chem*., **272**, 11074–11078.

32 Wagenknecht-Weisner,A., Weidmann,M., Braun,V., Leukel,P., Moos,M. and von Eichel-Streiber,C. (1997) *FEMS Microbiol. Lett*., **152**, 109–116.

33 Neuwald,A.F., Liu,J.S., Lipman,D.J. and Lawrence,C.E. (1997) *Nucleic Acids Res*., **25**, 1665–1677.

34 Nakayama,K., Nagasu,T., Shimma,Y., Kuromitsu,J. and Jigami,Y. (1992) *EMBO J*., **11**, 2511–2519.

35 Neiman,A.M., Mhaiskar,V., Manus,V., Galibert,F. and Dean,N. (1997) *Genetics*, **145**, 637–645.

36 Beeler,T.J., Fu,D., Rivera,J., Monaghan,E., Gable,K. and Dunn,T.M. (1997) *Mol. Gen. Genet*., **225**, 570–579.

37 Campbell,J.A., Davies,G.J., Bulone,V. and Henrissat,B. (1997) *Biochem. J*., **326**, 929–939.

38 Busch,C., Hofmann,F., Selzer,J., Munro,S., Jeckel,D. and Aktories,K. (1998) *J. Biol. Chem*., **273**, 19566–19572.

39 Hofmann,F., Busch,C. and Aktories,K. (1998) *Infect. Immun*., **66**, 1076–1081.

40 Rost,B. (1996) *Methods Enzymol*., **266**, 525–539.

41 Wootton,J.C. (1994) *Comput. Chem*., **18**, 269–285.

42 Wootton,J.C. and Federhen,S. (1996) *Methods Enzymol*., **266**, 554–571.

43 Welch,R.A. (1991) *Mol. Microbiol*., **5**, 521–528.

44 Welch,R.A., Forestier,C., Lobo,A., Pellett,S., Thomas,W.,Jr and Rowe,G. (1992) *FEMS Microbiol. Immunol*., **5**, 29–36.

45 Schmidt,H., Maier,E., Karch,H. and Benz,R. (1996) *Eur. J. Biochem*., **241**, 594–601.

46 Schmidt,H. (1995) *B933WSLT Bacteriophage 933W slt-IIB gene*. DDBJ/EMBL/GenBank databases.

47 Loveless,B.J. and Saier,M.H. (1997) *Mol. Memb. Biol*., **14**, 113–123.

48 Pugsley,A. (1993) *Microbiol. Rev*., **57**, 50–108.

49 Pugsley,A.P., Francetic,O., Possot,O.M., Sauvonnet,N. and Hardie,K.R. (1997) *Gene*, **192**, 13–19.

50 Francetic,O. and Pugsley,A. (1996) *J. Bacteriol*., **178**, 3544–3549.

51 Jubete,Y., Zabala,J.C., Juarez,A. and de la Cruz,F. (1995) *J. Bacteriol*., **177**, 242–246.

52 Gomez-Duarte,O.G. and Kaper,J.B. (1995) *Infect. Immun*., **63**, 1767–1776.

53 Clementz,T., Zhou,Z. and Raetz,C.R.H. (1997) *J. Biol. Chem*., **272**, 10353–10360.

54 Somerville,J.E.,Jr, Cassiano,L., Bainbridge,B., Cunningham,M.D. and Darveau,R.P. (1996) *J. Clin. Invest*., **97**, 359–365.

55 Parsot,C. and Mekalanos,J.J. (1991) *J. Bacteriol*., **173**, 2842–2851.

56 Harkey,C.W., Everiss,K.D. and Parsot,K.M. (1995) *Gene*, **153**, 81–84.

57 Hales,B.A., Hart,C.A., Batt,R.M. and Saunders,J.R. (1992) *Plasmid*, **28**, 183–193.

58 Silva,R.M., Saddi,S. and Maas,W.K. (1988) *Infect. Immun*., **56**, 836–842.

59 Womble,D.D., Sampathkumar,D.D, Easton,A.M., Luckow,V.A. and Rownd,R.H. (1985) *J. Mol. Biol*., **181**, 395–410.

60 Saul,D., Spiers,A.J., McAnulty,J., Gibbs,M.G., Bergquist,P.L. and Hill,D.F. (1989) *J. Bacteriol*., **171**, 2697–2707.

61 Bergquist,P.L., Saadi,S.S. and Maas,W.K. (1986) *Plasmid*, **15**, 19–34.

62 Gibbs,M.D., Spiers,A.J. and Bergquist,P.L. (1993) *Plasmid*, **29**, 165–179.

63 Biek,D.P. and Shi,J. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 8027–8031.

64 Loh,S.M., Cram,D.S. and Skurray,R.A. (1988) *Gene*, **66**, 259–268.

65 Salmon,M.A., Melderen,L.V., Bernard,P. and Couturier,M. (1994) *Mol. Gen. Genet*., **244**, 530–538.

66 Jensen,R.B. and Gerdes,K. (1995) *Mol. Microbiol*., **17**, 205–210.

67 Gultyaev,A.P., Franch,T. and Gerdes,K. (1997) *J. Mol. Biol*., **273**, 26–37.

68 Heffron,F., McCarthy,B.J., Otsubo,H. and Otsubo,E. (1979) *Cell*, **18**, 1153–1163.

69 Gill,R., Heffron,F., Dougan,G. and Falkow,S. (1978) *J. Bacteriol*., **136**, 742–756.

70 Boyd,E.F., Li,J., Ochman,H. and Selander,R.K. (1997) *J. Bacteriol*., **179**, 1985–1991.

71 Saitou,N. and Nei,N. (1987) *Mol. Biol. Evol*., **4**, 406–425.

72 Pupo,G.M., Karaolis,D.K.R., Lan,R. and Reeves,P.R. (1997) *Infect. Immun*., **65**, 2685–2692.

73 Sueoka,N. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2653–2657.

74 Blattner,F.R., Burland,V., Plunkett,G., III, Mayhew,G.F., Rose,D., Perna,N. and Gregor,J. (1998) *Microb. Comp. Genomics*, **3**, 45.

75 Kumar,S., Tamura,K. and Nei,M. (1993). *MEGA: Molecular Evolutionary Genetics Analysis.* The Pennsylvania State University, University Park, PA.