INFECTION AND IMMUNITY, Aug. 1998, p. 3810–3817 Vol. 66, No. 8
0019-9567/98/$04.00+0

Molecular Evolution of a Pathogenicity Island from Enterohemorrhagic *Escherichia coli* O157:H7†

NICOLE T. PERNA,[1]* GEORGE F. MAYHEW,[1] GYÖRGY PÓSFAI,[2] SIMON ELLIOTT,[3] MICHAEL S. DONNENBERG,[4] JAMES B. KAPER,[3] AND FREDERICK R. BLATTNER[1]

Laboratory of Genetics, University of Wisconsin-Madison, Madison Wisconsin 53706[1]; Institute of Biochemistry, Biological Research Center, H-6701 Szeged, Hungary[2]; and Center for Vaccine Development[3] and Division of Infectious Disease,[4] University of Maryland School of Medicine, Baltimore, Maryland 21201

Received 29 December 1997/Returned for modification 6 March 1998/Accepted 27 May 1998

We report the complete 43,359-bp sequence of the locus of enterocyte effacement (LEE) from EDL933, an enterohemorrhagic *Escherichia coli* O157:H7 serovar originally isolated from contaminated hamburger implicated in an outbreak of hemorrhagic colitis. The locus was isolated from the EDL933 chromosome with a homologous-recombination-driven targeting vector. Recent completion of the LEE sequence from enteropathogenic *E. coli* (EPEC) E2348/69 afforded the opportunity for a comparative analysis of the entire pathogenicity island. We have identified a total of 54 open reading frames in the EDL933 LEE. Of these, 13 fall within a putative P4 family prophage designated 933L. The prophage is not present in E2348/69 but is found in a closely related EPEC O55:H7 serovar and other O157:H7 isolates. The remaining 41 genes are shared by the two complete LEEs, and we describe the nature and extent of variation among the two strains for each gene. The rate of divergence is heterogeneous along the locus. Most genes show greater than 95% identity between the two strains, but other genes vary more than expected for clonal divergence among *E. coli* strains. Several of these highly divergent genes encode proteins that are known to be involved in interactions with the host cell. This pattern suggests recombinational divergence coupled with natural selection and has implications for our understanding of the interaction of both pathogens with their host, for the emergence of O157:H7, and for the evolutionary history of pathogens in general.

The locus of enterocyte effacement (LEE) is a 35-kb cluster of genes involved in the intimate adherence of pathogens to intestinal epithelial cells, the initiation of host signal transduction pathways, and the formation of attaching and effacing lesions (32, 33). Colony hybridization studies indicate that sequences homologous to the entire element are found in numerous enteropathogenic *Escherichia coli* (EPEC) and enterohemorrhagic *E. coli* (EHEC) strains and in other related bacteria (32). However, sequence data have been available for only a limited number of genes, and often from just one strain.

EPEC is an important cause of nonbloody infantile diarrhea in developing countries (8). EHEC O157:H7 causes both nonbloody diarrhea and a clinically distinct form of diarrheal disease known as hemorrhagic colitis that can lead to hemolytic uremic syndrome (17). These differences in clinical syndromes can be at least partly explained by the presence or absence of specific virulence factor genes. For example, the possession of bacteriophage-encoded Shiga toxins is a crucial distinction in the pathogenesis of disease due to these two pathogens, and acquisition of the phage-encoded toxins played a significant role in the evolution of EHEC from EPEC (48). Apart from the differences in virulence factors, EPEC and EHEC share the important intestinal histopathological phenotype known as attaching and effacing. This phenotype is distinguished by effacement of intestinal epithelial cell microvilli, intimate adherence of the bacteria to the epithelial cells, and marked changes in the host cell cytoskeleton. Both EPEC and EHEC have been shown to produce attaching and effacing lesions in tissue culture (26, 27) and animal models (15, 34, 45, 46). The genes responsible for this phenotype are contained in the LEE pathogenicity island, and sequence variation in one LEE gene has been implicated in different intestinal colonization sites (44, 50). How much of the similarities and differences in the pathogenesis of EPEC and EHEC disease can be attributed to sequence variation in the LEE is unknown.

The majority of information on genes contained in the LEE has been generated with EPEC strain E2348/69, and the complete LEE sequence for this strain has recently been reported (14). The *eae* gene encodes a cell surface protein, intimin, involved with the intimate interaction of the pathogen and host epithelial cells (21). Three genes, *espA* (24), *espB* (12), and *espD* (29), encode secreted proteins involved with host signal transduction pathways. The *esc* (20) and *sep* (39) genes encode components of a type III secretory apparatus. The recently identified *tir* gene encodes a protein that is translocated from the bacterium to the host where it serves as a receptor for intimin (22).

Only three genes of the O157:H7 LEE have been previously sequenced, *eae* (3, 50), *espB* (13), and an open reading frame (ORF) of unknown function immediately upstream of *eae* (*orfU*) (52). We now report the complete sequence of the LEE for EHEC O157:H7 strain EDL933. This information provides a rare opportunity for a comparative analysis of pathogenicity islands from two organisms that have important similarities as well as important differences in the pathogenesis of disease.

* Corresponding author. Mailing address: Laboratory of Genetics, University of Wisconsin-Madison, 445 Henry Mall, Rm. B44, Madison, WI 53706. Phone: (608) 262-2534. Fax: (608) 263-7459. E-mail: nicole@genetics.wisc.edu.

† Laboratory of Genetics paper 3516.

MATERIALS AND METHODS

Bacterial strains. Eight strains of *E. coli* were used in this study. EDL933, an O157:H7 serotype obtained from the American Type Culture Collection (ATCC 43895), was originally isolated from contaminated hamburger implicated in an

TABLE 1. Primers used in this study

Primer name	Sequence (5′ to 3′)	Use(s)	Melting temp (°C)
leephage-f	GATCTTCCGCCAGTTTGCCTCTCC	PCR, sequencing	68
leephage-r	AGGTATACTGGCAATAGCGGACAC	PCR, sequencing	65
Mdh-a	ATGAAAGTCGCAGTCCTCGGC	PCR, sequencing	66
Mdh-c	TTACTTATTAACGAACTCTTCGCCC	PCR, sequencing	61
Mdh274	GTATGGATCGTTCCGACCTGTTTA	Sequencing	63
Mdh-r327	ATAATACCAATGCACGCTTTCGGG	Sequencing	63
Mdh578	TGTCACAGGTTCCTGGCGTTAG	Sequencing	65
Mdh-r591	CGTTTGGTCAGATCAGCCACTTCC	Sequencing	66

outbreak of hemorrhagic colitis (47). EPEC strain E2348/69 (O127:H6) and the five diarrheal *E. coli* (DEC) strains are from the University of Maryland Center for Vaccine Development collection. The DEC strains (48), kindly provided by Tom Whittam (The Pennsylvania State University) are C54-58 (O55:H6), F60-51 (O55:H7), 5625-50 (O55:H7), 3077-88 (O157:H7), and C374-83 (O157:H7). The K-12 strain MG1655, from the University of Wisconsin collection, was described by Guyer et al. (18).

LEE isolation and sequencing. Isolation of the LEE from EDL933 is discussed in detail elsewhere (37). In short, targeting vectors containing approximately 800 bp of known chromosomal sequence flanking the LEE were used to introduce a novel Flp recombinase target site on each side of the EDL933 LEE by homologous recombination. Expression of Flp from a helper plasmid promotes excision of the target as a plasmid. Random shotgun cloning into the Janus M13 vector (7) provided templates for automated sequencing (Applied Biosystems; model 377) with the ABI PRISM DyeTerminator Cycle Sequencing Ready Reaction Kit under standard conditions. Random clones were sequenced to provide an average of eightfold coverage of the entire element with a minimum quality of threefold coverage, including both strands, for all regions. The sequences were assembled with SeqManII software (DNASTAR) and edited manually. We note that the EDL933 sequence is presented in the conventional clockwise orientation of the K-12 genome. This representation is the reverse complement of that published for the EPEC E2384/69 LEE, which was oriented with respect to the transcription of the *eae* gene.

LEE gene identification and comparative analysis. ORFs were located with GeneQuest software (DNASTAR). Each ORF was annotated based on protein sequence searches (DeCypherII hardware/software system; TimeLogic, Inc.) against a combined SwissProt release 34 and trEMBLSP release 1 database. Each ORF in the sequence was assigned a unique identifier (L0001 to L0057), which appears in the GenBank submission. All sequence alignments of the EDL933 and E2348/69 LEEs were done with MegAlign or Align software (DNASTAR), and levels of divergence were assessed with the Molecular Evolutionary Genetic Analysis software package (28). The codon adaptation index was calculated by the method of Sharp and Li (41).

LEE phage PCR and sequencing. The *selC*-LEE junctions of EDL933, E2348/69, and five DEC strains were amplified by PCR. Genomic DNA was isolated in 2% Incert agarose (FMC BioProducts), as described by Kirkpatrick and Blattner (25). Agarose plugs were melted in equal volumes of 10 mM Tris-Cl (pH 8), and 1 μl was used as the template for a 50-μl PCR reaction mixture containing 2.5 U of *TaKaRa LA Taq* polymerase (PanVera), a 200 μM concentration of each dNTP, and a 2.5 μM concentration of each primer. The primers were designated leephage-f (located in *selC*) and leephage-r (in the nearest region of the LEE conserved between EDL933 and E2348/69) (Table 1). These primers were chosen based on the EDL933 LEE sequence generated in this study. Strains with a LEE lacking a prophage generate a fragment of approximately 700 bp, while those with a full, intact sequence generate an approximately 8-kb fragment. An initial melting step of 94°C for 2 min was followed by 30 cycles of 94°C for 1 min and 68°C for 10 min, followed by a 72°C extension for 10 min. The PCR products were gel purified (Microcon/Micropure) according to the manufacturer's instructions and sequenced directly with the amplification primers.

Malate dehydrogenase (*mdh*) gene sequencing. The *mdh* genes from EDL933, E2384/69, and the DEC strains were amplified by PCR in a 50-μl PCR reaction mixture containing 1 μl of the genomic template described above, 2.5 U of *TaKaRa Ex Taq* polymerase (PanVera), a 200 μM concentration of each dNTP, and a 2.5 μM concentration of each primer. The primers were designated mdh-a and mdh-c (Table 1). An initial melting step of 94°C for 1 min was followed by 25 cycles of 94°C for 15 s, 64°C for 15 s, and 72°C for 4 min. PCR products of the correct size (~900 bp) were sequenced by primer walking. Primers were selected from the known MG1655 sequence for *mdh* (Table 1). Templates were sequenced under standard reaction conditions. Generated trace data files were then assembled and edited with SeqManII (DNASTAR). Consensus sequences were used to reconstruct an *mdh* phylogeny with the Kimura two-parameter distance matrix and the neighbor-joining algorithm of the Molecular Evolutionary Genetic Analysis software package (28).

Nucleotide sequence accession numbers. The sequences determined in this study have been assigned GenBank accession numbers AF071027 to AF071034.

RESULTS

Location and length of O157:H7 LEE. The region sequenced from EDL933 included chromosomal sequence flanking the LEE itself, allowing verification of the location by comparison with the K-12 sequence (4). A diagram of the EDL933 LEE and intact genes homologous to K-12 is shown in Fig. 1A. Near identity between the O157:H7 and K-12 chromosomes ends just beyond (16 bp) the 3′ end of the mature *selC* tRNA at bp 3833943 in the K-12 genome (GenBank accession no., U00096). In the K-12 chromosome, *selC* is followed by two hypothetical ORFs, *yicK* and *yicL*. In the EDL933 chromosome, the LEE is substituted for a segment (792 bp) of the K-12 chromosome including the *selC-yicK* intergenic region and 5′ end of *yicK*. The substitution of the LEE is accompanied by at least one other local rearrangement relative to K-12, a 781-bp internal deletion in the *yicK* ORF (bp 3834812 to 3835592 in the K-12 genome). As inferred by a PCR assay, both junctions are conserved between EDL933 and EPEC strain E2348/69 (32).

The EDL933 LEE is 43,359 bp (bp 903 to 44,262 of the sequence with GenBank accession no. AF071034), which is considerably greater than the 35,624-bp element sequenced from the EPEC strain (14). A 7.5-kb putative prophage near the *selC* end of the locus (Fig. 1A) accounts for most of the size difference. Although the 5′ ends of the two LEEs are nearly identical, they diverge abruptly after 118 bp. Excluding the 13 bp at the *selC* junction, this 105-bp region (bp 916 to 1020) is duplicated in the EDL933 LEE, 7,539 bp downstream (bp 8454 to 8558). An alignment of the two copies from EDL933 and the single copy from E2348/69 is shown in Fig. 2. The sequence bounded by these repeats in EDL933 has no homolog in the E2348/69 LEE but contains several ORFs similar to ones found in retronphage phiR73 (42) and other members the CP4 family of cryptic prophage from K-12 previously described (4). The LEE prophage, with the direct repeats delimiting *attR* and *attL*, has been designated 933L.

We have confirmed that the putative prophage is present in the genomic copy of the EDL933 LEE and absent from E2348/69 by PCR (Fig. 3), with primers specific to *selC* and the nearest region of the locus shared by the two strains. The EDL933 PCR product is approximately 7.5 kb larger than the E2348/69 product. We also tested five additional strains, representing each major DEC clone (48) known to have its LEE adjacent to *selC* (49). Of these, two O157:H7 serovars exhibited PCR products identical in size to that of EDL933. Two others, both O55:H6 serovars, lack the LEE prophage and produce a band identical in size to that of E2348/69. One O55:H7 serovar yielded a product much larger than that of E2348/69 but roughly 1 kb smaller than that of EDL933 as a result of an internal deletion of a segment of the prophage.

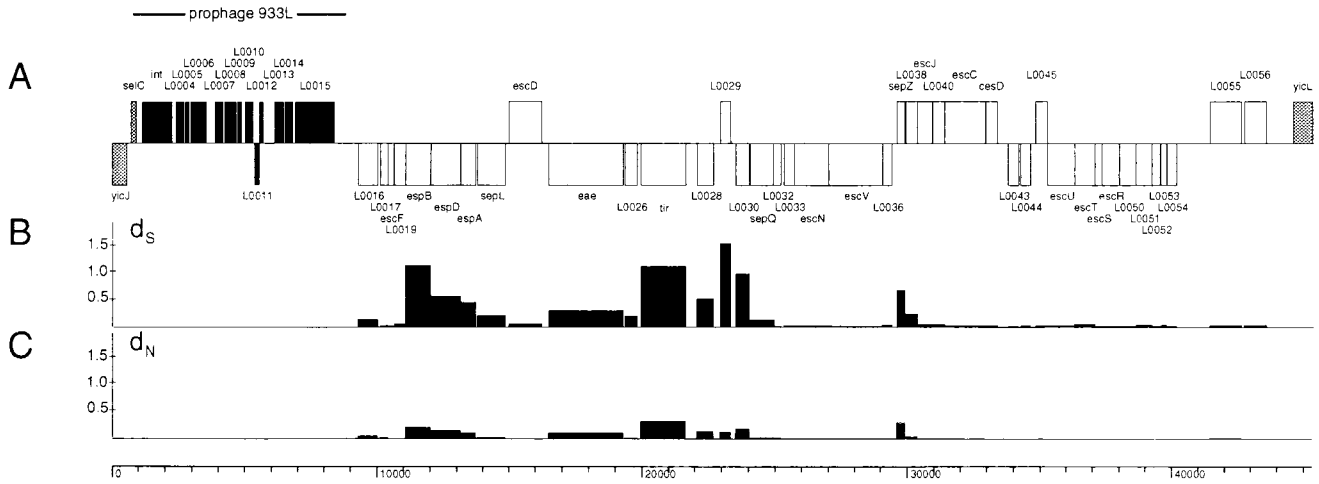

FIG. 1. Diagram of the EDL933 LEE. (A) ORFs are shown above and below the line to indicate the direction of transcription. Genes of the putative prophage are shown in black. Genes common to both the EDL933 and E2348/69 LEEs are shown in white. Genes homologous to those on the K-12 chromosome are hatched. The number of synonymous changes per synonymous site (d_S) and (B) the number of nonsynonymous changes per nonsynonymous site (d_N) (C) are shown for each ORF shared with the EPEC E2348/69 LEE (values are taken from Table 2). A scale (in base pairs) is shown along the bottom.

Single-read sequencing from the leephage-f and leephage-r primers was used to verify the identity of each product.

ORFs of the EDL933 LEE. We have identified a total of 54 ORFs in the extended EDL933 LEE (Fig. 1A). Of these, 13 fall within the putative prophage and 41 correspond to those previously described from the recently completed EPEC E2348/69 LEE sequence (14). The average codon adaptation index for the 54 ORFs, 0.219, is well below the average for the K-12 genome. As observed for EPEC E2348/69, the overall GC content (40.91%) is also below the K-12 average (50.80%). The prophage base composition is 51.72% G+C, while the remainder of the element is only 39.59% G+C.

The prophage ORF nearest to *selC*, L0003, encodes a 393-amino-acid (aa) protein with greater than 40% identity to integrases from a number of known bacteriophages including SF6 (SwissProt accession no., P37317), P4 (P08320), CP4-57 (P32053), and phiR73. L0004 encodes a 116-aa product 48% identical to a hypothetical *Shigella dysenteriae* IS*911* protein (P39213) of similar size. A similar ORF is also found in CP4-6. L0005 encodes a short (60-aa) peptide with no significant matches in the combined SwissProt release 34 and trEMBLSP release 1 database or in the K-12 genome. The L0006 gene product resembles putative transposases from *E. coli* IS*3* (P77673), *Acinetobacter calcoaceticus* (Q43916), and *Erwinia amylovora* (Q57113). The next two ORFs are similar to genes found only in the P4-like family of cryptic prophage from K-12 (4). The first of these two, L0007, encoding a 124-aa product, is shared by CP4-57, CP4-6, and CP4-44, all of which encode comparable-size peptides with >58% of the amino acids identical. L0008 has a match only in CP4-44, which encodes a considerably smaller product (163 versus 60 aa). The next four ORFs (L0009, L0010, L0011, and L0012) are complete unknowns. That is, there are no matches in the existing protein databases exceeding 30% identity alignable across at least 60% of both proteins. The remaining three ORFs within the boundaries of the putative prophage are most similar to hypothetical ORFs annotated in *Agrobacterium* and *Rhizobium* plasmid sequences. The products of all three are similar in length to the database entries (133 aa versus 154 aa for Q52592, 115 aa versus 115 aa for P50359, and 512 aa versus 511 aa for P55504). The first gene in this group, L0013, is also similar to the 5′ end of a hypothetical insertion element IS*2* gene roughly three times its size. The products of the latter two, L0014 and L0015, match proteins that are part of larger families of paralogous hypothetical proteins encoded by a *Rhizobium* plasmid.

The identities of the 41 genes common to both the EHEC EDL933 and EPEC 2348/69 LEEs have been presented in some detail elsewhere (14). Absolute conservation of gene order and number is observed between these elements. Table 2 lists the corresponding ORFs from the EPEC strain and a brief description of the known or putative function. Further comment on individual ORFs will be made in the context of

```
        1                                                                  90
903     ACTCTCCCGACATCAACAAAGCCCCTTATAATCCCTTATATATCAGGTTTTCTTTGTATCCTGAGGTCAAACGACTTTAACCCATATCAA   Repeat 1: EDL933
8454    ------------.....A.....T.....................C.........G.A.A.G.....C.....-C....TG.....   Repeat 2: EDL933
35624   ............G.T.....A.....C.....................C.........G...A..A....T.....-C....TG.....   E2348/69
86      .....A.TG....T-G.T.AT...C.....TT..G..G..T..CG.........G..A..A.....C.A...C.C...G.....   PAI-1

        91                      119
993     GTTGTTTGTGGGGGTGCATT-GGGGGTAT   Repeat 1: EDL933
8530    ....C..T...A..A.A.T..A.......   Repeat 2: EDL933
35535   ....C..T.....A.A.T..A.......   E2348/69
175     .C.T                           PAI-1
```

FIG. 2. Alignment of the two repeats flanking 933L, the *selC* end of the E2348/69 LEE, and the end of PAI-1 from uropathogenic *E. coli* (GenBank accession no., M13943). Residues that match the first copy of the repeat in EDL933 are represented by a dot. Coordinates shown to the left of the alignment correspond to the sequence position in each GenBank entry.

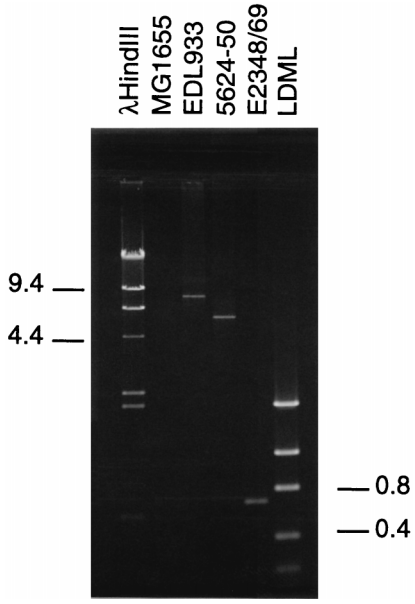

FIG. 3. PCR to detect 933L in genomic DNA from MG1655, EDL933, 5624-50, and E2348/69 with primers leephage-f and leephage-r from Table 1. One-half microliter of the reaction mixture was loaded on a 1% horizontal agarose gel in Tris-acetate-EDTA buffer and electrophoresed at 40 V for 1.5 h to visualize the product. The sizes of bands in Lambda DNA/*Hin*dIII markers (λHindIII) from Promega and a Low DNA MASS ladder (LDML) from GIBCO BRL in kilobase pairs are shown.

exploring the sequence variation among these strains. Additional small ORFs can be found in both the EHEC and EPEC LEEs but are not included in our final determinations.

Sequence comparison with EPEC. Each O157:H7 LEE gene and protein were aligned to their homologs from E2348/69. Protein alignments were used to refine nucleotide sequence alignments by eye. Summary data for each comparison are shown in Table 2. The average level of nucleotide identity is 93.9%. The number of gaps and total gap length in each alignment are shown. We calculated rates of synonymous and nonsynonymous substitution for each gene. These estimates are plotted in Fig. 1B and C to illustrate the distribution of variation across the length of the LEE element. It is clear from both Table 2 and Fig. 1 that there is considerable variation among genes in the level of sequence divergence. In general, variable ORFs are more variable in every way. They have more gaps and higher rates of both synonymous and nonsynonymous substitution.

One group of highly variable genes includes *espB* (25.99% difference), *espD* (19.64% difference), and *espA* (15.37% difference). All three encode secreted proteins implicated in the activation of host epithelial signal transduction, intimate adherence, and formation of attaching and effacing lesions (23, 24, 29). *espB*, *espD*, and *espA* are adjacent in a tightly packed, presumably cotranscribed, cluster that also encodes a potential chaperone, one putative component of the secretion apparatus, and two proteins of unknown function (9). The first gene of this group, one of the unknowns, shows moderate sequence divergence (6.60%) in addition to a variable-number repeat structure described in more detail below. The putative secretion apparatus gene in this region, *escF*, is invariant between EDL933 and E2348/69. A gene of unknown function, L0023, is found on the other side of the variable cluster directly adjacent to *espA*. It varies at 5.68% of the nucleotide sites but has an absolutely conserved length in these two strains. L0023 is followed by a highly conserved secretion system gene, *escD*, contained on the opposite strand.

Another gene that shows marked differences is *eae*, which encodes intimin, perhaps the best characterized of the LEE proteins. This outer membrane protein is required (but not sufficient) for the intimate adherence to epithelial cells characteristic of attaching and effacing enteropathogens and for full virulence of EPEC in human volunteers (10, 11). Cell binding activity is known to reside within the C-terminal 192 aa of the intimin protein (16). Yu and Kaper (50) sequenced *eae* from both EDL933 and E2348/69 and suggested that the C-terminal variability between the corresponding intimins might indicate that the strains bind different eukaryotic receptors. Since then, *eae* sequences from numerous other strains have been reported, and comparative analyses of these have shown divergence patterns similar to that observed in the initial EDL933-E2348/69 comparison and have also shown that there are at least three groupings of *eae* (2, 16). In our study, the two *eae* genes were found to be 87.23% identical, or slightly less variable than the secreted proteins described above.

A second group of contiguous genes is highly variable. The most variable of this group is *tir* (33.52% difference), which was recently discovered to encode a product that is translocated from the bacterium to the host cell where it most likely serves as the intimin receptor (22). For the other three genes, L0028 (17.48% difference), L0029 (21.94% difference), and L0030 (25.30% difference), very little is known about their role in pathogenesis. The L0028 product resembles a hypothetical protein encoded in a *Shigella* virulence-associated cluster (14). Without additional functional characterization, it is impossible to evaluate the adaptive significance of variation at these loci. Among the remaining LEE genes, only one, *sepZ*, exhibits elevated divergence (29.29%) between EDL933 and E2348/69. Interestingly, a *sepZ* mutant of EPEC exhibits certain phenotypic similarities with EHEC, including reduced invasion efficiency and lack of tyrosine phosphorylation of Hp90, yet retains the ability to form attaching and effacing lesions (39).

Alignments of one gene, L0016, revealed a particularly interesting feature. The E2348/69 sequence requires a single 126-bp gap to accommodate the obviously homologous sequences at the 5′ and 3′ ends. The EDL933 sequence in this gap encodes a third copy of a proline-rich repeat noted by Donnenberg et al. (9) in the E2348/69 protein. Figure 4 is an alignment of the three complete copies and one partial copy of the repeat from EDL933 with the two complete copies and one partial copy from E2348/69. The first copies of the repeat in both strains, $R1_{EDL933}$ and $R1_{E2348/69}$, have several bases in common (positions 5, 6, 10, and 12) near the beginning of the alignment that are not conserved among the other complete copies, $R2_{EDL933}$, $R3_{EDL933}$, and $R2_{E2348/69}$. Similarly, the two partial repeats near the ends of the genes, $R4_{EDL933}$ and $R3_{E2348/69}$, are most like each other in the central region of the alignment (positions 24, 55 to 67, 76, and 114). At other sites (positions 96, 99, 102, and 125 to 141) the copies within each strain appear to be more similar to each other than to the copies from the other strain. This suggests either gene conversion to homogenize the repeats within a sequence or separate expansion of repeats in each lineage by either unequal recombination or slipped-strand mispairing. Several other database proteins, mainly eukaryotic, exhibit similar repetitive structures, although there is no functional consensus.

Many of the intergenic regions are quite small, but several additional noncoding regions merit comparison. The large intergenic region between L0054 and L0055 in the EDL933 LEE and that between *rorf2* and *orf1* of the E2348/69 LEE include a member of the ERIC family of repeated elements and are 98.4% identical in the two strains. Previously, similarity between the *selC* end of the EPEC LEE and PAI-1 of uropatho-

TABLE 2. ORFs shared by the EPEC and EHEC LEEs

EDL933 ORF (product length [aa])	E2348/69 ORF (product length [aa])	Gene name (former name)	Protein description	Comparison of EDL933 and E2348/69 coding regions[a]				
				% Identity	No. of variable sites	No. of gaps (length)	$d_S \pm SE$	$d_N \pm SE$
L 0016 (248)	orf30 (206)		Unknown	93.40	41	1 (126)	0.1233 ± 0.0289	0.0505 ± 0.0109
L 0017 (92)	orf29 (92)		Unknown	97.85	6	0	0.0174 ± 0.0175	0.0233 ± 0.0105
L 0018 (73)	orf28 (73)	*escF*	Type III secretion apparatus	100.00	0	0	0.0000 ± 0.0000	0.0000 ± 0.0000
L 0019 (135)	orf27 (135)		Unknown	99.02	4	0	0.0357 ± 0.0208	0.0031 ± 0.0031
L 0020 (312)	orf26 (321)	*espB* (*eaeB*)	Secreted protein	74.01	261	8 (27)[b]	1.0843 ± 0.1432	0.2084 ± 0.0198
L 0021 (374)	orf25 (380)	*espD*	Secreted protein	80.36	222	1 (18)[b]	0.5402 ± 0.0602	0.1503 ± 0.0144
L 0022 (192)	orf24 (192)	*espA*	Secreted protein	84.63	89	0	0.4329 ± 0.0731	0.1110 ± 0.0167
L 0023 (351)	orf23 (351)		Unknown	94.32	60	0	0.1943 ± 0.0327	0.0268 ± 0.0058
L 0024 (406)	rorf12 (406)	*escD*	Type III secretion apparatus	98.28	21	2 (4)[b]	0.0524 ± 0.0141	0.0075 ± 0.0029
L 0025 (934)	orf22 (939)	*eae*	Intimin	87.23	360	6 (33)[b]	0.2850 ± 0.0243	0.1023 ± 0.0073
L 0026 (156)	orf21 (156)	*orfU*	Putative chaperone	95.33	22	0	0.1885 ± 0.0474	0.0138 ± 0.0062
L 0027 (558)	orf20 (550)	*tir*	Translocated intimin receptor	66.48	549	9 (54)	1.0697 ± 0.1023	0.3168 ± 0.0191
L 0028 (203)	orf19 (203)		Similar to *Shigella* IpgB (27%)	82.52	107	0	0.4905 ± 0.0807	0.1369 ± 0.0182
L 0029 (127)	rorf10 (120)		Unknown	78.06	79	1 (21)[b]	1.4865 ± 0.4083	0.1231 ± 0.0221
L 0030 (168)	orf18 (176)		Unknown	74.70	127	4 (42)	0.9421 ± 0.1614	0.1883 ± 0.0247
L 0031 (305)	orf17 (305)	*sepQ*	Type III secretion apparatus	95.32	43	0	0.1145 ± 0.0252	0.0308 ± 0.0067
L 0032 (91)	orf16 (91)		Unknown	98.91	3	0	0.0000 ± 0.0000	0.0142 ± 0.0082
L 0033 (125)	orf15 (125)		Unknown	99.74	1	0	0.0126 ± 0.0126	0.0000 ± 0.0000
L 0034 (446)	orf14 (446)	*escN* (*sepB*)	Type III secretion apparatus	99.40	8	0	0.0188 ± 0.0077	0.0020 ± 0.0014
L 0035 (675)	orf13 (675)	*escV* (*sepA*)	Type III secretion apparatus	99.61	8	0	0.0107 ± 0.0048	0.0019 ± 0.0011
L 0036 (117)	orf12 (117)		Unknown	99.44	2	0	0.0290 ± 0.0206	0.0000 ± 0.0000
L 0037 (99)	rorf9 (98)	*sepZ*	Type III secretion apparatus	70.71	87	1 (3)	0.6527 ± 0.1324	0.2947 ± 0.0435
L 0038 (142)	rorf8 (142)		Unknown	93.01	30	0	0.2223 ± 0.0534	0.0344 ± 0.0104
L 0039 (190)	rorf7 (190)	*escJ* (*sepD*)	Type III secretion apparatus	99.13	5	0	0.0423 ± 0.0191	0.0000 ± 0.0000
L 0040 (151)	rorf6 (151)		Unknown	98.46	7	0	0.0433 ± 0.0218	0.0084 ± 0.0049
L 0041 (512)	rorf5 (512)	*escC* (*sepC*)	Type III secretion apparatus	99.42	9	0	0.0143 ± 0.0064	0.0034 ± 0.0017
L 0042 (151)	rorf4 (151)	*cesD* (sepE)	Chaperone of *espD*	99.56	2	0	0.0200 ± 0.0142	0.0000 ± 0.0000
L 0043 (137)	orf11 (137)		Unknown	99.52	2	0	0.0000 ± 0.0000	0.0063 ± 0.0045
L 0044 (123)	orf10 (123)		Unknown	99.19	3	0	0.0126 ± 0.0126	0.0070 ± 0.0049
L 0045 (152)	rorf3 (152)		Similar to *E. coli* P19 (40%)	98.91	5	0	0.0210 ± 0.0149	0.0084 ± 0.0049
L 0046 (345)	orf9 (345)	*escU* (*sepF*)	Type III secretion apparatus	99.52	5	0	0.0172 ± 0.0086	0.0013 ± 0.0013
L 0047 (258)	orf8 (258)	*escT* (*sepG*)	Type III secretion apparatus	98.84	9	0	0.0459 ± 0.0164	0.0017 ± 0.0017
L 0048 (89)	orf7 (89)	*escS* (sepH)	Type III secretion apparatus	100.00	0	0	0.0000 ± 0.0000	0.0000 ± 0.0000
L 0049 (217)	orf6 (217)	*escR* (*sepI*)	Type III secretion apparatus	99.69	2	0	0.0069 ± 0.0069	0.0020 ± 0.0020
L 0050 (231)	orf5 (231)		Unknown	99.14	6	0	0.0072 ± 0.0072	0.0091 ± 0.0041
L 0051 (199)	orf4 (199)		Unknown	99.33	4	0	0.0233 ± 0.0135	0.0021 ± 0.0021
L 0052 (107)	orf3 (107)		Unknown	100.00	0	0	0.0000 ± 0.0000	0.0000 ± 0.0000
L 0053 (72)	orf2 (72)		Unknown	99.54	1	0	0.0244 ± 0.0245	0.0000 ± 0.0000
L 0054 (129)	orf1 (129)		Similar to *Bordetella* bpH3 (43%, 37 aa)	99.74	1	0	0.0000 ± 0.0000	0.0034 ± 0.0034
L 0055 (398)	rorf2 (398)		Similar to *Shigella* VirA (22%)	98.50	18	0	0.0239 ± 0.0094	0.0126 ± 0.0037
L 0056 (272)	rorf1 (272)		Similar to *E. coli* YijK (36%)	99.02	8	0	0.0262 ± 0.0122	0.0053 ± 0.0029

[a] d_S, rate of synonymous substitution; d_N, rate of nonsynonymous substitution.
[b] One or more gap-containing regions were eliminated from the d_S and d_N analysis; all alignments are available upon request.

genic *E. coli* has been noted (32). This corresponds to the repeated region flanking the prophage in EDL933. The EPEC copy is more like the second repeat in EHEC (97 versus 83% identical). All three LEE repeats are equally similar to PAI-1 (73% identical). An alignment of these sequences is shown in Fig. 2.

Comparisons with other published *espA* and *espB* genes. Our EDL933 *espB* sequence differs from a published one (13) by one synonymous change. Sequence data were available for *espA* and *espB* from a rabbit EPEC strain, RDEC-1 (1). The *espA* sequences of E2348/69 and RDEC-1 are more similar to each other (91.0% similarity) than either is to that of EDL933 (84.6 and 87.6% similarity, respectively). The RDEC-1 *espB* is more similar to both that of EDL933 (76.8% similarity) and E2348/69 (73.4% similarity) than they are to each other (68.6% similarity). In sharp contrast, the *espB* sequences of RDEC-1 and bovine EHEC O26 serovar 413/89-1 (13) are nearly identical (two synonymous differences in 945 bp) (1).

Phylogenetic relationship among strains. To interpet the observed variation between the two complete LEEs and the distribution of the phage, it is helpful to understand the relationship between the strains under consideration. The gene encoding malate dehydrogenase (*mdh*) has been used to describe the relationships among *E. coli* strains (6, 38). We have sequenced PCR-amplified *mdh* from EDL933, E2348/69, and the DEC strains used in this study. None of the *E. coli mdh* sequences differ from another by more than 3% identity. These sequences were used to reconstruct an updated phylogeny (Fig. 5). This tree is consistent with the relationships observed in a multilocus enzyme electrophoresis-based phylogeny of EPEC and EHEC strains (49) but additionally suggests that EDL933 and E2348/69 are both more closely related to strains from the nondiarrheagenic ECOR collection than to each other. As predicted by multilocus enzyme electrophoresis (48, 49), the O55:H7 serovar and all three O157:H7 strains, including EDL933, are very closely related. The distribution of the LEE prophage is indicated on the tree (Fig. 5).

DISCUSSION

We have described the 43.359-kb LEE from EHEC EDL933 and compared it to the entire homologous locus from EPEC E2348/69. The EDL933 LEE has precisely the same chromo-

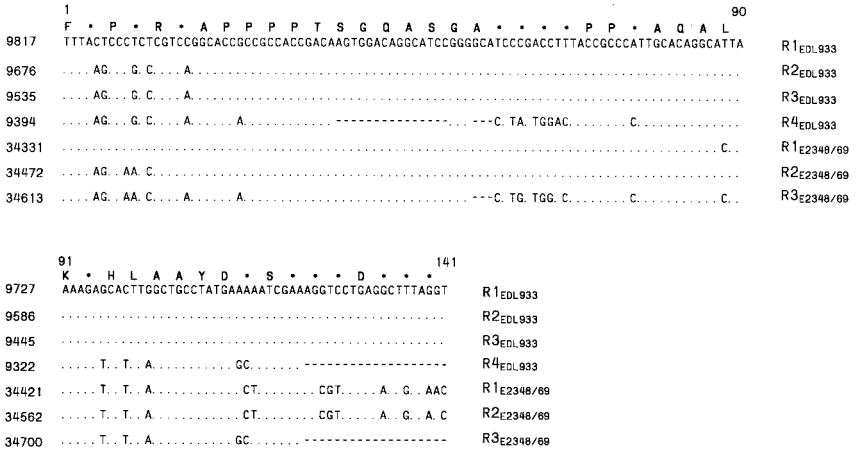

FIG. 4. Alignment of the repetitive motifs of EDL933 gene L0016 and the E2348/69 homolog. Coordinates shown to the left of the alignment correspond to the sequence position in each GenBank entry. Each sequence is labeled with the repeat number and subscripted strain designation ($R1_{EDL933}$, $R2_{EDL933}$, etc.). The repeat is in frame in both strains, and the putative peptide sequence is shown above the alignment with asterisks replacing residues that are not absolutely conserved. Dots indicate nucleotides identical to those in the $R1_{EDL933}$ sequence. Dashes have been added to maximize the similarity among sequences.

somal integration point as the LEE of EPEC E2348/69 and is highly similar for most genes. However, the O157:H7 LEE differs structurally from the 35.5-kb EPEC locus by the inclusion of P4 family cryptic prophage 933L, and the two LEEs exhibit high levels of divergence for some genes.

The base composition and codon usage of the LEE are atypical of *E. coli*, suggesting that the element was horizontally transferred into the species. One very basic question is whether the LEE was acquired once or repeatedly. As predicted by McDaniel et al. (32), EDL933 and E2348/69 LEEs are found in homologous chromosomal positions and have very similar sequences at both junctions of the insertion point relative to the K-12 chromosome. An examination of the *mdh* phylogeny shows that EDL933 and E2348/69 are relatively distantly related *E. coli*, both of which are more closely related to nondiarrheagenic ECOR strains than to each other. In a colony hybridization survey, no nonpathogenic *E. coli* examined exhibited a LEE element (32). If the LEE was acquired only once and is not found among commensal strains, it must have been subsequently lost independently by numerous lineages. Alternatively, the LEEs shared by EDL933, E2348/69, and a variety of other EPEC strains may be the result of lateral exchange. Under this model, LEE is introduced again and again into different clonal lineages by recombination either with a related species or with another strain of *E. coli* that already harbors the element. Wieler et al. (49) have recently shown that there are both EPEC and EHEC strains that contain the LEE but that have an intact *selC*, indicating that there is at least one other chromosomal location for the LEE among *E. coli* strains.

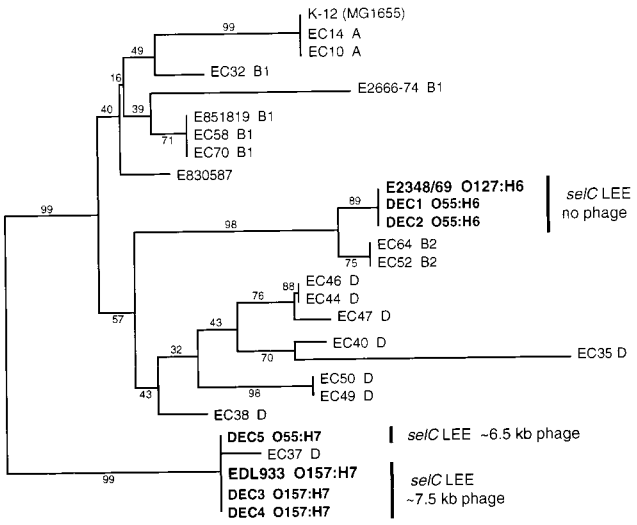

FIG. 5. Evolutionary relationships among strains used in this study and 20 *E. coli* reference strains based on *mdh* sequences. A phylogeny was reconstructed with published data from the ECOR collection (6, 38) (GenBank accession no., U04742 through U04758, U04770, and AF004201) and our EDL933, E2348/69, and DEC strain *mdh* sequences. The tree is based on a MegAlign (DNASTAR) multiple sequence alignment of positions 34 through 878 of the 916-bp *mdh* coding region, excluding positions 652 and 653. The phylogeny was reconstructed with a Kimura two-parameter distance matrix and the neighbor-joining algorithm. The percentages of 2,000 bootstrap replicates supporting each cluster are shown along the branches.

The only previously presented evidence suggesting autonomous mobility of the LEE is the slight similarity of the E2348/69 LEE ends and the IS*600* family of insertion elements, including a small ORF encoding a peptide with similarity to part of the putative transposase (9). These regions are highly conserved between the EDL933 and E2348/69 sequences. If both LEEs were independently mobilized by a precursor of these transposase remnants, we would not expect them to differ from a functional transposon in such similar ways. Although we do not rule out the possibility that the LEE originally entered an *E. coli selC* locus by this mechanism, the conserved remnants suggest that the EDL933 and E2348/69 LEEs have a common ancestor with an already-defective transposase. One possible scenario is transposition of the LEE into one *E. coli selC*, loss of mobility, and subsequent lateral transfer distributing the LEE at the *selC* locus among *E. coli* lineages.

The cryptic P4-like prophage associated with the EDL933 LEE appears to integrate into the very end of the LEE itself rather than the *selC* locus. This is notable since *selC* acts as a recombinational hot spot, serving as the site of integration for the related retronphage phiR73, other phage, and another pathogenicity island, PAI-1, in uropathogenic *E. coli*. PCR and sequencing suggest that this phage integrated into the LEE prior to the divergence of the O55:H7 and O157:H7 strains

shown in the *mdh* tree. Given the ability of prophage to be excised, leaving an intact integration site, it is not possible to tell whether there was ever a phage in the E2348/69 LEE. Regardless, it seems unlikely that the LEE was ever mobilized by this prophage since both putative *att* sites occur on one side of the LEE coding region.

Other than the putative prophage, gene content and organization are highly conserved between the EDL933 and E2348/69 LEEs. There is however, strikingly large and nonrandomly distributed sequence variation within the cluster (Fig. 2). Homologous *E. coli* loci derived from direct clonal descent generally differ at less than 5% of the nucleotide sites. *mdh*, with an average pairwise divergence of 2.1%, is a standard example (36). Many of the LEE genes exhibit comparable levels of sequence divergence, but eight genes vary at more than 15% of the alignable nucleotide sites.

Subdivision of nucleotide variation into synonymous and nonsynonymous changes sheds some additional light on the mechanisms of LEE evolution. Synonymous base substitutions are those changes that do not affect the amino acid sequence of the encoded protein. Nonsynonymous substitutions are reflected in the protein and are subject to natural selection. Several observations can be made about the pattern of divergence for the two complete LEEs. Synonymous substitutions do not occur at a constant rate across the element. High rates of nonsynonymous change are always accompanied by elevated synonymous divergence and in no case does the estimated nonsynonymous rate exceed the synonymous rate, indicating purifying selection on all proteins encoded in the locus.

In the absence of selection constraining the accumulation of synonymous differences, the rate should be constant across sequences that have been evolving independently as intact units. Instead, synonymous substitution rates for the 41 shared LEE proteins are quite heterogeneous. For many of the less variable genes, the estimated synonymous substitution rate is very close to zero indicating very recent divergence of at least some regions of the EDL933 and E2348/69 LEEs. All genes previously described as variable overall have notably higher synonymous substitution rates, as do L0023, L0026, *sepQ*, and L0038, all of which are adjacent to a hypervariable gene. It is difficult to envision a selection-based model of evolution that could lead to the observed disparity in the frequency of synonymous changes in the LEEs if the mutation rate is constant along the entire locus. The patchy distribution of synonymous variation is most easily explained by recombinational events that unite genes or clusters of genes with distinct mutational histories. The number of recombinations and donor sources involved may be discernable by comparisons of more LEE sequences from *E. coli* strains and related species.

Nonsynonymous changes and natural selection of the encoded proteins most likely determines the fate of new alleles. The most striking feature of the divergent genes is that they include all of those encoding proteins known to interact directly with the host: *eae*, *espA*, *espB*, *espD*, and *tir*. Intimin binds to the host cells. Intimin also appears to influence the site of intestinal colonization. Complementation of an O157:H7 *eae* mutant with plasmid-borne EPEC *eae* results in colonization of the distal half of the small intestine and the surface of the large intestine in gnotobiotic pigs (44). This is more typical of EPEC than EHEC, which normally colonizes only the lower bowel. Kenny and Finlay have observed the *espA*-, *espB*-, and *espD*-encoded proteins associated with HeLa cells after EPEC infection (23). The *tir* gene product is translocated into the host cell (22). In EPEC, the *tir* gene product is tyrosine phosphorylated, becoming the protein originally described as Hp90 (40), and is presented on the host cell surface where it serves as the intimin receptor. EHEC O157:H7 fails to induce tyrosine phosphorylation of the receptor protein in HEp-2 and T84 cells (19). Together, these observations suggest that the variability we observe between the EPEC and EHEC LEEs has phenotypic effects that could be subject to natural selection for adaptation to either host specificity or evasion of the host immune system. In contrast to the hypervariability seen in genes encoding proteins known to interact directly with the host, other LEE-encoded proteins that do not interact with the host are highly conserved between both EHEC and EPEC LEEs, notably the *esc* genes encoding the type III secretion system.

Hypervariability of genes directly involved with host interaction is emerging as a paradigm of molecular evolution in pathogens. Two studies (5, 30) have noted a relationship between cellular location of the gene product and rate of divergence of the *inv-spa* invasion gene complex in *Salmonella enterica*. Three adjacent genes in this cluster exhibit elevated nonsynonymous site variation relative to other genes at the locus, among which are components of a type III secretion apparatus similar to that of the LEE. Two of the three, *spaO* and *spaN*, encode secreted proteins, while the product of the third, *spaM*, remains unknown. Boyd et al. (5) ruled out selection for host specificity as a mechanism driving the observed variability because of the lack of polymorphism among strongly host-adapted serovars but suggested that selection for antigenic diversity may be responsible. High levels of polymorphism have been observed at loci involved with host-pathogen interactions in many different bacterial species, including *Borrelia burgdorferi* (43), *Streptococcus pyogenes* (31), *Neisseria* spp. (35), and *Mycoplasma hominis* (51). Future genomic sequencing studies of these and other pathogens will likely yield additional insights into those proteins that are important in host-pathogen interactions.

ACKNOWLEDGMENTS

This work was supported by NSF-Sloan Foundation Molecular Evolution Postdoctoral Fellowship BIR-9626042 (N.T.P.); NIH grants AI41329-01 (F.R.B.), AI32074 (M.S.D.), and AI21637 and AI41325 (J.B.K.); HHMI grant 75195-542102 (G.P.); and Ronald McDonald House Charities.

We thank Guy Plunkett, Val Burland, and Jeremy Glasner for advice and thoughtful consideration of both the data and manuscript. We are grateful for the expert technical assistance provided by Heather Kirkpatrick, Jason Gregor, Guy Peyrot, Pritin Soni, Mike Goeden, and Wayne Davis.

REFERENCES

1. **Abe, A., B. Kenny, M. Stein, and B. B. Finlay.** 1997. Characterization of two virulence proteins secreted by rabbit enteropathogenic *Escherichia coli*, EspA and EspB, whose maximal expression is sensitive to host body temperature. Infect. Immun. **65:**3547–3555.
2. **Agin, T. S., and M. K. Wolf.** 1997. Identification of a family of intimins common to *Escherichia coli* causing attaching-effacing lesions in rabbits, humans, and swine. Infect. Immun. **65:**320–326.
3. **Beebakhee, G., M. Louie, J. D. Azavedo, and J. Brunton.** 1992. Cloning and nucleotide sequence of the *eae* gene homologue from enterohemorrhagic *Escherichia coli* serotype O157:H7. FEMS Microbiol. Lett. **91:**63–68.
4. **Blattner, F. R., G. Plunkett III, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. Goeden, D. Rose, B. Mau, and Y. Shao.** 1997. The complete sequence of *Escherichia coli* K-12. Science **277:**1453–1462.
5. **Boyd, E. F., J. Li, H. Ochman, and R. K. Selander.** 1997. Comparative genetics of the *inv-spa* invasion gene complex of *Salmonella enterica*. J. Bacteriol. **179:**1985–1991.
6. **Boyd, E. F., K. Nelson, F.-S. Wang, T. S. Whittam, and R. K. Selander.** 1994. Molecular genetic basis of allelic polymorphism in malate dehydrogenase (*mdh*) in natural populations of *Escherichia coli* and *Salmonella enterica*. Proc. Natl. Acad. Sci. USA **91:**1280–1284.
7. **Burland, V., D. L. Daniels, G. Plunkett III, and F. R. Blattner.** 1993. Genome sequencing on both strands: the Janus strategy. Nucleic Acids Res. **21:**3385–3390.

8. **Donnenberg, M. S.** 1995. Enteropathogenic *Escherichia coli*, p. 709–726. *In* M. J. Blaser, P. D. Smith, J. I. Ravdin, H. B. Greenberg, and R. L. Guerrant (ed.), Infections of the gastroinstestinal tract. Raven Press, New York, N.Y.
9. **Donnenberg, M. S., L.-C. Lai, and K. A. Taylor.** 1997. The locus of enterocyte effacement pathogenicity island of enteropathogenic *E. coli* encodes secretion functions and remnants of transposons at its extreme right end. Gene **184:**107–114.
10. **Donnenberg, M. S., C. O. Tacket, S. P. James, G. Losonsky, J. P. Nataro, S. S. Wasserman, J. B. Kaper, and M. M. Levine.** 1993. Role of the *eaeA* gene in experimental enteropathogenic *Escherichia coli* infection. J. Clin. Invest. **92:**1412–1417.
11. **Donnenberg, M. S., S. Tzipori, M. L. McKee, A. D. O'Brien, J. Alroy, and J. B. Kaper.** 1993. The role of the *eae* gene of enterohemorrhagic *Escherichia coli* in intimate attachment *in vitro* and in a porcine model. J. Clin. Invest. **92:**1418–1424.
12. **Donnenberg, M. S., J. Yu, and J. B. Kaper.** 1993. A second chromosomal gene necessary for intimate attachment of enteropathogenic *Escherichia coli* to epithelial cells. J. Bacteriol. **175:**4670–4680.
13. **Ebel, F., C. Deibel, A. U. Kresse, C. A. Guzman, and T. Chakraborty.** 1996. Temperature- and medium-dependent secretion of proteins by Shiga toxin-producing *Escherichia coli*. Infect. Immun. **64:**4472–4779.
14. **Elliott, S., L. A. Wainwright, T. McDaniel, B. MacNamara, L.-C. Lai, M. Donnenberg, and J. B. Kaper.** 1998. The complete sequence of the locus of enterocyte effacement (LEE) from enteropathogenic *E. coli* E2348/69. Mol. Microbiol. **28:**1–4.
15. **Francis, D. H., J. E. Collins, and J. R. Duimstra.** 1986. Infection of gnotobiotic pigs with an *Escherichia coli* O157:H7 strain associated with an outbreak of hemorrhagic colitis. Infect. Immun. **51:**953–956.
16. **Frankel, G., D. C. A. Candy, E. Fabiani, J. Adu-Bobie, S. Gil, M. Novakova, A. D. Phillips, and G. Dougan.** 1995. Molecular characterization of a carboxy-terminal eukaryotic-cell-binding domain of intimin from enteropathogenic *Escherichia coli*. Infect. Immun. **63:**4323–4328.
17. **Griffin, P. M., and R. V. Tauxe.** 1991. The epidemiology of infections caused by *Escherichia coli* O157:H7, and other enterohemorrhagic *E. coli*, and the associated hemolytic uremic syndrome. Epidemiol. Rev. **13:**60–98.
18. **Guyer, M., R. E. Reed, T. Steitz, and K. B. Low.** 1981. Identification of a sex-factor-affinity site in *E. coli* as γδ. Cold Spring Harbor Symp. Quant. Biol. **45:**135–140.
19. **Ismaili, A., D. J. Philpott, M. T. Dytoc, and P. M. Sherman.** 1995. Signal transduction responses following adhesion of verotoxin-producing *Escherichia coli*. Infect. Immun. **63:**3316–3326.
20. **Jarvis, K. G., J. A. Girón, A. E. Jerse, T. K. McDaniel, and M. S. Donnenberg.** 1995. Enteropathogenic *Escherichia coli* contains a putative type III secretion system necessary for the export of proteins involved in attaching and effacing lesion formation. Proc. Natl. Acad. Sci. USA **92:**7996–8000.
21. **Jerse, A. E., J. Yu, B. D. Tall, and J. B. Kaper.** 1990. A genetic locus of enteropathogenic *Escherichia coli* necessary for the production of attaching and effacing lesions on tissue culture cells. Proc. Natl. Acad. Sci. USA **87:**7839–7843.
22. **Kenny, B., R. DeVinney, M. Stein, D. J. Reinscheid, E. A. Frey, and B. B. Finlay.** 1997. Enteropathogenic *E. coli* (EPEC) transfers its receptor for intimate adherence into mammalian cells. Cell **91:**511–520.
23. **Kenny, B., and B. B. Finlay.** 1995. Protein secretion by enteropathogenic *Escherichia coli* is essential for transducing signals to epithelial cells. Proc. Natl. Acad. Sci. USA **92:**7991–7995.
24. **Kenny, B., L. C. Lai, B. B. Finlay, and M. S. Donnenberg.** 1996. EspA, a protein secreted by enteropathogenic *Escherichia coli*, is required to induce signals in epthelial cells. Mol. Microbiol. **20:**313–323.
25. **Kirkpatrick, H. A., and F. R. Blattner.** 1997. Isolation of intact, high molecular weight DNA fragments for the *E. coli* genome project. Epicentre Forum **4:**11–13.
26. **Knutton, S., T. Baldwin, P. H. Williams, and A. S. McNeish.** 1989. Actin accumulation at sites of bacterial adhesion to tissue culture cells: basis of a new diagnostic test for enteropathogenic and enterohemorrhagic *Escherichia coli*. Infect. Immun. **57:**1290–1298.
27. **Knutton, S., D. R. Lloyd, and A. S. McNeish.** 1987. Adhesion of enteropathogenic *Escherichia coli* to human intestinal enterocytes and cultured human intestinal mucosa. Infect. Immun. **55:**69–77.
28. **Kumar, S., K. Tamura, and M. Nei.** 1993. MEGA: molecular evolutionary genetics analysis, 1.01 ed. The Pennsylvania State University, University Park, Pa.
29. **Lai, L.-C., L. A. Wainwright, K. D. Stone, and M. S. Donnenberg.** 1997. A third secreted protein that is encoded by the enteropathogenic *Escherichia coli* pathogenicity island is required for transduction of signals and for attaching and effacing activities in host cells. Infect. Immun. **65:**2211–2217.
30. **Li, J., H. Ochman, E. A. Groisman, E. F. Boyd, F. Solomon, K. Nelson, and R. K. Selander.** 1995. Relationship between evolutionary rate and cellular location among the Inv/Spa invasion proteins of *Salmonella enterica*. Proc. Natl. Acad. Sci. USA **92:**7252–7256.
31. **Marciel, A. M., V. Kapur, and J. M. Musser.** 1997. Molecular population genetic analysis of a *Streptococcus pyogenes* bacteriophage-encoded hyaluronidase gene: recombination contributes to allelic variation. Microb. Pathog. **22:**209–217.
32. **McDaniel, T. K., K. G. Jarvis, M. S. Donnenberg, and J. B. Kaper.** 1995. A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens. Proc. Natl. Acad. Sci. USA **92:**1664–1668.
33. **McDaniel, T. K., and J. B. Kaper.** 1997. A cloned pathogenicity island from enteropathogenic *Escherichia coli* confers the attaching and effacing phenotype on *E. coli* K-12. Mol. Microbiol. **23:**399–407.
34. **Moon, H. W., S. C. Whipp, M. Argenzio, M. Levine, and R. A. Gianella.** 1983. Attaching and effacing activities of rabbit and human enteropathogenic *Escherichia coli* in pig and rabbit intestines. Infect. Immun. **41:**1340–1351.
35. **Nassif, X., J. Lowy, P. Stenberg, P. O'Gaora, A. Ganji, and M. So.** 1993. Antigenic variation of pilin regulates adhesion of *Neisseria meningitidis* to human epilethial cells. Mol. Microbiol. **8:**719–725.
36. **Nelson, K., and R. K. Selander.** 1994. Intergeneric transfer and recombination of the 6-phosphogluconate dehydrogenase gene (*gnd*) in enteric bacteria. Proc. Natl. Acad. Sci. USA **91:**10227–10231.
37. **Pósfai, G., M. D. Koob, H. A. Kirkpatrick, and F. R. Blattner.** 1997. Versatile insertion plasmids for targeted genome manipulations in bacteria: isolation, deletion, and rescue of the pathogenicity island LEE of the *Escherichia coli* O157:H7 genome. J. Bacteriol. **179:**4426–4428.
38. **Pupo, G. M., D. K. R. Karaolis, R. Lan, and P. R. Reeves.** 1997. Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and *mdh* sequence studies. Infect. Immun. **65:**2685–2692.
39. **Rabinowitz, R. P., L.-C. Lai, K. Jarvis, T. K. McDaniel, J. B. Kaper, K. D. Stone, and M. S. Donnenberg.** 1996. Attaching and effacing of host cells by enteropathogenic *Escherichia coli*: the absence of detectable tyrosine kinase mediated signal transduction. Microb. Pathog. **21:**157–171.
40. **Rosenshine, I., M. S. Donnenberg, J. B. Kaper, and B. B. Finlay.** 1992. Signal transduction between enteropathogenic *Escherichia coli* (EPEC) and epilethial cells: EPEC induces tyrosine phosphorylation of host cell proteins to initiate cytoskeletal rearrangement and bacterial uptake. EMBO J. **11:**3551–3560.
41. **Sharp, P. M., and W. H. Li.** 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15:**1281–1295.
42. **Sun, J., M. Inouye, and S. Inouye.** 1991. Association of a retroelement with a P4-like cryptic prophage (retronphage ϕR73) integrated into the selenocystyl tRNA gene of *Escherichia coli*. J. Bacteriol. **173:**4171–4181.
43. **Theisen, M., M. Borre, M. J. Mathiesen, B. Mikkelsen, A.-M. Lebech, and K. Hansen.** 1995. Evolution of the *Borrelia burgdorferi* outer surface protein OspC. J. Bacteriol. **177:**3036–3044.
44. **Tzipori, S., F. Gunzer, M. S. Donnenberg, L. DeMontigny, J. B. Kaper, and A. Donohue-Rolfe.** 1995. The role of the *eaeA* gene in diarrhea and neurological complications in a gnotobiotic piglet model of enterohemorrhagic *Escherichia coli* infection. Infect. Immun. **63:**3621–3627.
45. **Tzipori, S., R. M. Robins-Browne, G. Gonis, J. Hayes, M. Whithers, and E. McCartney.** 1985. Enteropathogenic *Escherichia coli* enteritis: evaluation of the gnotobiotic piglet as a model of human infection. Gut **26:**570–578.
46. **Tzipori, S., I. K. Wachsmuth, C. Chapman, R. Birden, J. Brittingham, C. Jackson, and J. Hogg.** 1986. The pathogenesis of hemorrhagic colitis caused by *Escherichia coli* O175:H7 in gnotobiotic piglets. J. Infect. Dis. **154:**712–716.
47. **Wells, J. G., B. R. Davis, I. K. Wachsmuth, L. W. Riley, R. S. Remis, R. Sokolow, and G. K. Morris.** 1983. Laboratory investigation of hemorrhagic colitis outbreaks associated with a rare *Escherichia coli* serotype. J. Clin. Microbiol. **18:**512–520.
48. **Whittam, T. S., M. L. Wolfe, I. K. Wachsmuth, F. Ørskov, I. Ørskov, and R. A. Wilson.** 1993. Clonal relationships among *Escherichia coli* strains that cause hemorrhagic colitis and infantile diarrhea. Infect. Immun. **61:**1619–1629.
49. **Wieler, L. H., T. K. McDaniel, T. S. Whittam, and J. B. Kaper.** 1997. Insertion site of the locus of enterocyte effacement in enteropathogenic and enterohemorrhagic *Escherichia coli* differs in relation to the clonal phylogeny of the strains. FEMS Microbiol. Lett. **156:**49–53.
50. **Yu, J., and J. B. Kaper.** 1992. Cloning and characterization of the *eae* gene of enterohaemorrhagic *Escherichia coli* O157:H7. Mol. Microbiol. **6:**411–417.
51. **Zhang, Q., and K. S. Wise.** 1996. Molecular basis of size and antigenic variation of a *Mycoplasma hominis* adhesin encoded by divergent *vaa* genes. Infect. Immun. **64:**2737–2744.
52. **Zhao, S., S. E. Mitchell, J. Meng, M. P. Doyle, and S. Kresovich.** 1995. Cloning and nucleotide sequence of a gene upstream of the *eaeA* gene of enterohemorrhagic *Escherichia coli* O157:H7. FEMS Microbiol. Lett. **133:**35–39.

Editor: V. A. Fischetti