JOURNAL OF BACTERIOLOGY, Oct. 2000, p. 5381–5390 Vol. 182, No. 19
0021-9193/00/$04.00+0

Gene Conservation and Loss in the *mutS-rpoS* Genomic Region of Pathogenic *Escherichia coli*

CORINNE J. HERBELIN, SAMANTHA C. CHIRILLO, KRISTEN A. MELNICK, AND THOMAS S. WHITTAM*

Institute of Molecular Evolutionary Genetics, Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802

Received 17 May 2000/Accepted 5 July 2000

The extent and nature of DNA polymorphism in the *mutS-rpoS* region of the *Escherichia coli* genome were assessed in 21 strains of enteropathogenic *E. coli* (EPEC) and enterohemorrhagic *E. coli* (EHEC) and in 6 strains originally isolated from natural populations. The intervening region between *mutS* and *rpoS* was amplified by long-range PCR, and the resulting amplicons varied substantially in length (7.8 to 14.2 kb) among pathogenic groups. Restriction maps based on five enzymes and sequence analysis showed that strains of the EPEC 1, EPEC 2, and EHEC 2 groups have a long *mutS-rpoS* region composed of a ~6.0-kb DNA segment found in strain K-12 and a novel DNA segment (~2.9 kb) located at the 3′ end of *rpoS*. The novel segment contains three genes (*yclC*, *pad1*, and *slyA*) that occur in *E. coli* O157:H7 and related strains but are not found in K-12 or members of the ECOR group A. Phylogenetic analysis of the common sequences indicates that the long intergenic region is ancestral and at least two separate deletion events gave rise to the shorter regions characteristic of the *E. coli* O157:H7 and K-12 lineages.

The acquisition of new genes by horizontal transfer has played a major role in the adaptation and ecological specialization of bacterial lineages (17). It has been estimated, for example, that ~18% of the current genome of *Escherichia coli* K-12 represents foreign DNA acquired by horizontal transfers since the divergence of *E. coli* and *Salmonella enterica* (18). Gene acquisitions have also contributed to the variation in virulence among strains and closely related bacterial species (11, 38). In *E. coli* and *S. enterica*, blocks of virulence genes, called pathogenicity islands, have been acquired at different times, thus generating a variety of pathogens with distinct virulence genes and mechanisms of pathogenesis (12, 31, 32). In some cases, loss of genes has been important in adaptive radiation and the evolution of bacterial virulence. For example, Maurelli and coworkers (25) present evidence that the universal deletion of the lysine decarboxylase gene (*cadA*) has enhanced the virulence of *Shigella* species because cadaverine, a product of the reaction catalyzed by lysine decarboxylase, inhibits the activity of *Shigella* enterotoxin.

One active region of genomic evolution is located between 61 and 62 min in the *E. coli* genome (19). This region includes two essential genes (Fig. 1): *mutS*, which encodes one of the four proteins required for DNA mismatch repair (39); and *rpoS*, which encodes a sigma factor (σ^{38}) that regulates many stationary-phase and environmental stress response genes (13). Both *mutS* and *rpoS* are highly conserved in sequence between *E. coli* and *S. enterica*; however, the nearby genomic regions have diverged in a variety of ways. In *S. enterica*, there is a 40-kb pathogenicity island (SPI-1 [Fig. 1]) that is inserted next to *mutS* and is required for epithelial cell invasion (11, 27). SPI-1 has been detected in all *Salmonella* groups but is absent in *E. coli*, suggesting that the island was acquired early in the evolutionary radiation of the salmonellae (31). Among different *E. coli* strains, the length of the genomic sequence between the *mutS* and *rpoS* genes is variable (Fig. 1). The region is 6.9 kb long in the *E. coli* K-12 genome (1) but varies in length among pathogenic strains of *E. coli* and *Shigella* (19, 20; P. E. Carter, L. Butler, I. R. Booth, and F. M. Thomson-Carter, Abstr. 99th Gen. Meet. Am. Soc. Microbiol., p. 32, 1999). Alterations in this region have been correlated with an enhanced mutation rate and are implicated in the emergence of new pathogenic clones (19).

The purpose of this study was to assess DNA polymorphism in the *mutS-rpoS* region and to infer the evolutionary history of divergence of this region among pathogenic strains of *E. coli*. The study focuses on four groups of pathogenic strains (45) representing enteropathogenic *E. coli* (EPEC), a prominent cause of infantile diarrhea in the developing world (29), and enterohemorrhagic *E. coli* (EHEC), a major cause of foodborne illness (29). We used a combination of restriction fragment length polymorphism (RFLP) analysis and nucleotide sequencing to characterize the genetic variation in the *mutS-rpoS* region among the EPEC and EHEC strains and compared the variation to that in nonpathogenic strains isolated from natural populations and strains closely related to laboratory strain K-12.

MATERIALS AND METHODS

Bacterial strains. Of the 27 strains used in this study (Table 1), 21 were implicated in diarrheal diseases. Nineteen were originally isolated from patients, one (EDL-933) was isolated from hamburger implicated in an 1982 outbreak of hemolytic colitis, and DEC 8c was isolated from a calf with scours. The laboratory strain K-12 and five ECOR (*E. coli* reference collection) group A strains originally isolated from natural populations (33) were also included. Twenty of the 21 pathogenic strains represent classical serotypes of EPEC and EHEC (Table 1). The pathogenic strains have been classified previously into four clonal groups (EPEC 1, EPEC 2, EHEC 1, and EHEC 2) based on analysis by multilocus enzyme electrophoresis (MLEE) (43–45). Another pathogenic strain included in this study (921-B4, serotype O111:H9), originally recovered from a disease outbreak in Finland (42), does not fall into one of the four groups (T. S. Whittam, unpublished data). All isolates are epidemiologically unrelated. Bacteria were maintained in Luria-Bertani broth supplemented with 20% glycerol at −70°C.

Preparation of genomic DNA. Genomic DNA was prepared from 1 ml of bacterial culture grown in Luria-Bertani broth (37°C, 16 h, 150 rpm) with a PUREGENE genomic DNA isolation kit (Gentra Systems Inc., Minneapolis, Minn.) and was stored at 4°C.

* Corresponding author. Mailing address: IMEG, Department of Biology, 208 Mueller Laboratory, Pennsylvania State University, University Park, PA 16802-3500. Phone: (814) 863-1970. Fax: (814) 865-9131. E-mail: tswl@psu.edu.

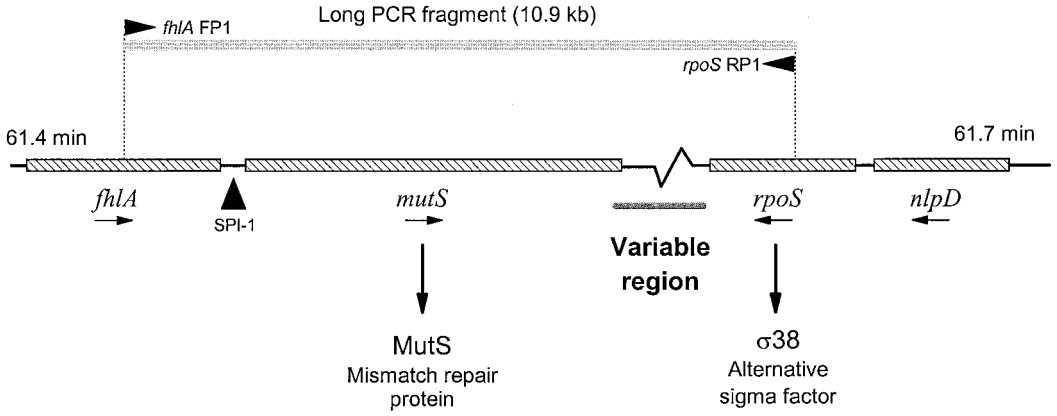

FIG. 1. The *mutS-rpoS* genomic region located at 61 min on *E. coli* K-12 chromosome. Primers designed from *fhlA* (FP1) and *rpoS* (RP1) of the K-12 genomic sequences (1) produce a long PCR amplicon of 10.9 kb. Previous studies have shown that *Salmonella* strains have a 40-kb pathogenicity island (SPI-1) between *fhlA* and *mutS* (27) and that the genomic region between *mutS* and *rpoS* in *E. coli* O157:H7 and *Shigella* strains is variable in length (19, 20).

Restriction enzyme analysis. An amplicon extending from the 3′-end region of *fhlA* to the 5′-end region of *rpoS* including the entire *mutS* gene was produced using primers *fhlA* FP1 and *rpoS* RP1 (Fig. 1). The predicted PCR amplicon in K-12 contains eight open reading frames (ORFs) (including *mutS*) and has a size of 10,950 bp. The primers were designed on the *fhlA* and *rpoS* sequences from the *E. coli* K-12 genome (GenBank accession no. U29579). The primer sequences and positions in the K-12 genome (indicated in parentheses) are as follows: *fhlA* FP1, 5′-CGCGCGGTATTGCTAACACG-3′ (28461 to 28481); and *rpoS* RP1, 5′-GATTCGCCAGACGATTGAAC-3′ (39391 to 39411). The DNA was amplified with the Taq Plus Long PCR system (Stratagene, La Jolla, Calif.). While kept on ice, 50 ng (in 1 μl) of purified genomic DNA template was mixed with 49 μl of PCR mixture containing 20 mM Tris-HCl (pH 9.2), 60 mM KCl, 2 mM $MgCl_2$, 0.5 μM each primer, 250 μM each deoxynucleoside triphosphate (dNTP), and 5 U of Taq Plus Long polymerase mixture. Samples were heated at 94°C for 5 min and then subjected to 30 PCR cycles consisting of 30 s at 94°C, 1 min at 56°C, and 20 min at 72°C. Amplicons were electrophoresed in 0.4% SeaKem Gold agarose gel in 1× Tris-borate-EDTA (TBE) buffer with ethidium bromide at 0.5 μg/ml. The amplicons (2 μl) and the ladder (1 μg) were heated for 10 min at 65°C prior to gel loading, and electrophoresis was performed under ~5 mm of buffer overlay at 1.5 V/cm for 20 h. The fragment sizes were determined using the DNA ProScan molecular weight software (DNA ProScan, Inc., Nashville, Tenn.) with lambda DNA/*Hin*dIII (GIBCO-BRL, Life Technologies, Rockville, Md.) as a standard.

To estimate the length of the *mutS-rpoS* genomic region and to map the length variation, the long PCR amplicons were digested with five restriction enzymes, *Eco*RV and *Nde*I (New England Biolabs, Beverly, Mass.), *Csp*45 (Promega, Madison, Wis.), and *Acc*I and *Nsp*I (GIBCO-BRL). A total of 2 μl of product was digested with 20 U of enzyme for 16 h at 37°C. Restriction fragments were electrophoresed in a 0.75% SeaKem GTG agarose gel (1× TBE, 5 h, 6 V/cm) with 1 μg of 1-kb DNA ladder (GIBCO-BRL) as a size standard. DNA fragments stained with ethidium bromide were detected under UV illumination, and fragment sizes were estimated with DNA ProScan molecular weight software.

RFLP in the *mutS-rpoS* genomic region was further assessed by digestion of 5 μl of the product (containing the long PCR amplicon) with 10 U of four-base cutter restriction endonucleases (*Alu*I and *Sau*3A1; Promega) at 37°C for 2 h. Restriction fragments were electrophoresed in a 4% NuSieve GTG agarose gel (1× TBE; 6 h, 6 V/cm) with 1 μg of the 100-bp DNA ladder (GIBCO-BRL) as a size standard.

Nucleotide sequencing of DNA located in the *mutS-rpoS* intergenic region. Strains DEC 1a and E2348/69 (EPEC 1), DEC 12e (EPEC 2), and DEC 9f (EHEC 2) were used for nucleotide sequencing. The region extending from the 5′ end of ORF *o388* (see Fig. 3) to the 3′ end of *rpoS* was amplified with the Taq Plus Long PCR system using two primers, *o388* FP2 (5′-CCGGAAGCAATCG ACGCACT-3′; positions 34758 to 34778) and *rpoS* RP2 (5′-GTGTTCGCCAG ATTCAGGTT-3′; positions 38936 to 38956), designed from the *E. coli* K-12 genome. For long PCR, 50 ng (in 1 μl) of purified genomic DNA template was mixed on ice with 49 μl of PCR mixture containing 20 mM Tris-HCl (pH 8.75), 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 2 mM $MgCl_2$, 0.1% Triton X-100, 0.1 mg of nuclease-free bovine serum albumin per ml, 0.5 μM each primer, 250 μM each dNTP, and 5 U of Taq Plus Long polymerase mixture. Samples were heated at 94°C for 3 min and then subjected to 30 PCR cycles consisting of 30 s at 94°C, 1 min at 56°C, and 6.5 min at 72°C.

Cycle sequencing PCR was performed with a Prism Ready Reaction DyeTerminator cycle sequencing kit from Applied Biosystems. Sequencing gels were run on an Applied Biosystems 373A automated sequencer. Raw sequences of both DNA strands were analyzed and concatenated by DNASTAR (Madison, Wis.) software. Additional internal sequencing primers were sequentially designed as sequence data were generated. All conflicting and putative polymorphic nucleotides sites were sequenced at least three times on both strands with multiple primers to eliminate sequencing errors.

TABLE 1. Serotypes and sources of 26 *E. coli* strains

ET group[a]	Strain	Serotype[b]	Location	Reference or source
EPEC 1				
1	DEC 1a	O55:H6	Pennsylvania	36
2	DEC 2a	O55:H6	Congo	36
3	E2348/69	O127:H6	United Kingdom	21
4	D55	O127:ND	Thailand	9
5	E851/71	O142:H6	Scotland	21
EPEC 2				
6	DEC 11a	O128:H2	Montana	26
7	DEC 12c	O111:NM	Panama	36
8	DEC 12d	O111:H2	Peru	36
9	DEC 12e	O111:H−	Kenya	36
10	124-55	O111:H−	Florida	28
EHEC 1				
11	EDL-933	O157:H7	Oregon	37
12	86-24	O157:H7	Washington	10
13	93-111	O157:H7	Washington	P. Tarr
14	OK-1	O157:H7	Japan	T. Takeda
15	DEC 5d	O55:H7	Sri Lanka	26
EHEC 2				
16	DEC 8b	O111:H8	Idaho	2
17	DEC 8c	O111:NM	South Dakota	5
18	CL 37	O111:H8	Canada	15
19	DEC 9f	O26:NM	South Dakota	CDC[c]
20	928/91	O111:H−	Germany	H. Karch
Others				
22	ECOR 1	ON:H−	Iowa	33
23	ECOR 2	ON:H32	New York	33
24	ECOR 3	O1:NM	Massachusetts	33
25	ECOR 7	O85:H−	Washington	33
26	ECOR 10	O6:H10	Sweden	33
27	921-B4	O111:H9	Finland	42

[a] Strains are grouped by electrophoretic type (ET) based on MLEE (45). Laboratory strain K-12 was included as "21" in "Others."

[b] NM, nonmotile; H−, serotype negative for flagellar antigens; ND, not determined.

[c] CDC, Centers for Disease Control and Prevention.

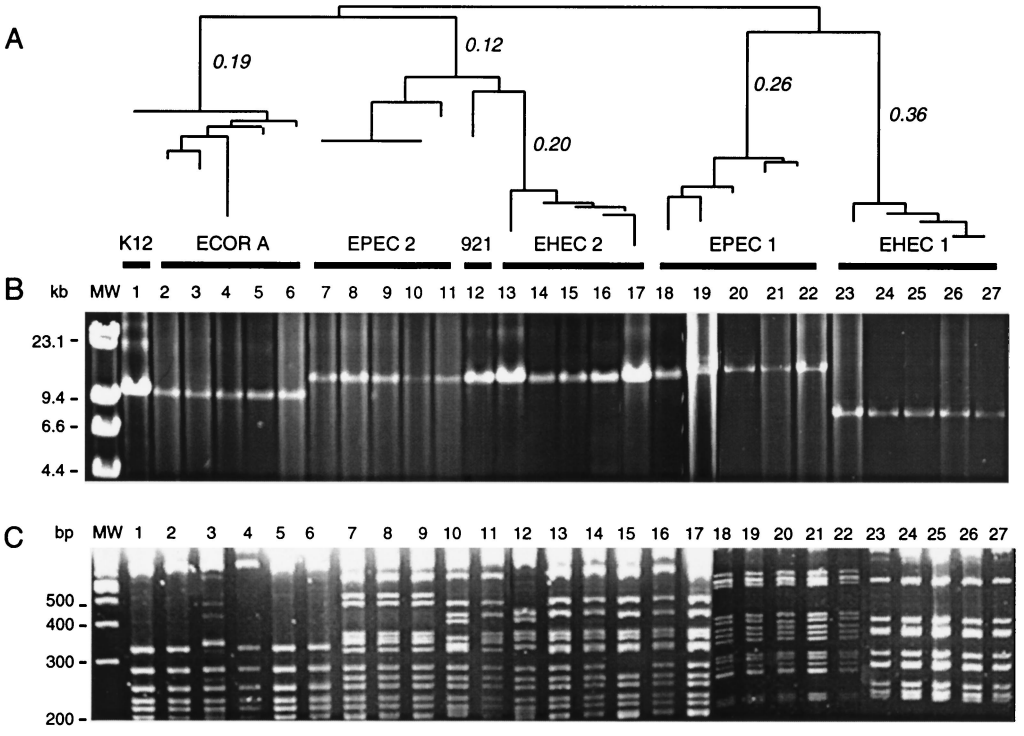

FIG. 2. (A) Neighbor-joining tree of 27 *E. coli* strains based on genetic distances estimated from MLEE data. The clonal groups based on electrophoretic type (ET) and strain names are, from left to right: K-12, ECOR group A (ECOR 1, ECOR 10, ECOR 7, ECOR 2, and ECOR 3), EPEC 2 (DEC 12e, 125-55, DEC 12c, DEC 12d, and DEC 11a), 921-B4, EHEC 2 (DEC 9e, DEC 8c, DEC 8b, 928/91, and CL 37), EPEC 1 (DEC 2a, D55, DEC 1a, E2348/69, and E851/71), and EHEC 1 (DEC 5d, OK-1, 86-24, EDL-933, and 93-111). (B) Resolution of long PCR amplicons corresponding to the genomic region extending from *mutS* to *rpoS*. The molecular weight marker (MW) is the lambda DNA/*Hin*dIII ladder. (C) Restriction fragments (>200 bp in length) of the *mutS-rpoS* long PCR products digested with *Alu*I. The molecular weight marker is a 100-bp DNA ladder.

Phylogenetic analysis. For comparative purposes, sequences of the *mutS-rpoS* genomic region from GenBank were included in the analysis for *E. coli* K-12 (AE000357 and AE000358) and O157:H7 strains (ECAJ6210) and for *Shigella* (AF055472). Rates of synonymous and nonsynonymous substitutions were estimated by the Nei-Gojobori method (30), and gene phylogenies were constructed using the neighbor-joining method (40) in MEGA (16).

Nucleotide sequence accession numbers. The sequences reported here were deposited in GenBank with accession numbers AF242208 to AF242211.

RESULTS

Size variation and DNA polymorphism in the *mutS-rpoS* region. The genomic region between *mutS* and *rpoS* was amplified from strain K-12 as a single ~10.9-kb PCR product as predicted from the genomic sequence (Fig. 1). Application of the long PCR primers to 27 strains of *E. coli* resulted in amplicons that ranged in size between 8.0 and 14.5 kb (Fig. 2B). A comparison of the amplified DNA among strains of distinct phylogenetic groups defined previous by MLEE (Fig. 2A) shows that the long PCR amplicons were consistent in size within the major groups but different in size between groups (Fig. 2B). The average sizes based on electrophoresis were ~11.0 kb for *E. coli* K-12 and ECOR group A strains, ~14.5 kb for each EPEC 1, EPEC 2, and EHEC 2 strain, and ~8.0 kb for O157:H7 and other strains of the EHEC 1 group (Fig. 2B). Strain 921-B4, an O111:H9 pathogen that does not belong to the EPEC or EHEC groups (Fig. 2A), produced a 14.5-kb amplicon (Fig. 2B).

We digested the *mutS-rpoS* amplicons with five six-base cutter restriction enzymes to estimate more accurately the total length and to create maps of the restriction sites (Table 2 and Fig. 3). Restriction digests with four of the enzymes (*Eco*RV, *Nde*I, *Acc*I, and *Csp*45) could not distinguish strains of the EHEC 2 and EPEC 2 groups; however, some genetic differences between strains of these groups were obtained using *Nsp*I (Table 2). The *Nsp*I digest also revealed that the *mutS-rpoS* sequence in strain 921-B4 differs from that of EPEC 2 and EHEC 2 strains (Table 2).

Comparisons of the restriction maps (Fig. 3) show that bacteria from the EPEC 1, EPEC 2, and EHEC 2 groups share a distinct DNA segment located between *o454* and *rpoS*; this ~2.9-kb novel DNA segment is not found in K-12 or strains of the ECOR group A. The maps reveal that the genes occur in the same order in all strains of the EPEC 1, EPEC 2, and EHEC 2 groups (Fig. 3). In addition, there is an extra 400-bp segment at the 5′ end of *mutS* in EPEC 1 strains (Fig. 3). In contrast, strains from the EHEC 1 group have a shorter *mutS-rpoS* region with 3.1 kb less than K-12 and its relatives and ~6.0 kb less than the EPEC 1, EPEC 2, and EHEC 2 groups. Absence of the restriction fragments predicted from the K-12 sequence for *Nde*I, *Eco*RV, *Nsp*I, and *Acc*I between positions 6797 bp (*Nde*I) and 11793 bp (*Nsp*I) indicates that part of the genomic DNA between *mutS* and *rpoS* in EHEC 1 strains is missing or changed. The fact that restriction sites for *Nsp*I (position 5843 in K-12) and *Nde*I (position 5982 in K-12) as well as several upstream sites are intact suggests that the sequence from *mutS* to ORF *o218* has been conserved in both O157:H7 and DEC 5d (O55:H7) strains (Fig. 3). The remaining DNA is highly divergent compared to the K-12 group (Fig. 3). A short segment of DNA between *mutS* and *rpoS* in O157:H7 has been reported previously (4; Carter et al., Abstr. 99th Gen. Meet. Am. Soc. Microbiol., 1999).

TABLE 2. Fragment sizes from six-base cutter restriction enzyme digests of the *mutS-rpoS* genomic region

Strain or group (*n*)	*mutS-rpoS* PCR fragment size (kb)	Fragment sizes (kb)				
		*Eco*RV	*Nde*I	*Csp*45	*Acc*I	*Nsp*I
K-12	10.9	1.4, 2.2, 2.5, 4.8	0.4, 0.8, 1.9, 2.3, 5.6	0.7, 2.1, 3.9, 4.2	0.2, 1.3, 1.5, 1.7, 6.2	0.3, 0.6, 1.0, 1.3, 1.7, 2.7, 3.3
ECOR A (5)	10.9	1.4, 2.2, 2.5, 4.8	0.4, 0.8, 1.9, 2.3, 5.6	0.7, 2.1, 3.9, 4.2	0.2, 1.3, 1.5, 1.7, 6.2	0.3, 0.6, 1.0, 1.3, 1.7, 2.7, 3.3
EPEC 2 (5)	13.8	1.4, 2.2, 2.5, 2.8, 4.8	0.4, 0.8, 1.9, 2.3, 8.4	0.7, 2.9, 4.2, 6.0	0.2, 1.0, 1.3, 1.5, 3.6, 6.2	0.3, 0.6, 1.3, 1.7, 2.7, 7.2
921-B4	13.8	1.4, 2.2, 2.5, 2.8, 4.8	0.4, 0.8, 1.9, 2.3, 8.4	0.7, 2.9, 4.2, 6.0	0.2, 1.0, 1.3, 1.5, 3.6, 6.2	0.5, 0.6, 1.6, 1.7, 2.7, 2.7, 3.9
EHEC 2 (5)	13.8	1.4, 2.2, 2.5, 2.8, 4.8	0.4, 0.8, 1.9, 2.3, 8.4	0.7, 2.9, 4.2, 6.0	0.2, 1.0, 1.3, 1.5, 3.6, 6.2	0.6, 1.6, 1.7, 2.7, 7.2
EPEC 1 (5)	14.2	1.8, 2.4, 2.5, 2.7, 4.8	0.4, 0.8, 2.3, 2.3, 8.4	0.7, 2.1, 2.9, 4.2, 4.5	0.2, 0.5, 1.0, 1.3, 1.4, 1.5, 2.1, 6.2	0.3, 1.0, 1.2, 1.2, 1.7, 2.7, 2.7, 3.3
EHEC 1 (5)	7.8	1.4, 2.2, 4.2	1.9, 2.3, 3.6	2.9, 5.0	0.2, 1.0, 6.6	1.0, 2.8, 4.0

Digestion of the long PCR amplicons with the four-base cutting enzyme *Alu*I resolves the 27 *E. coli* isolates into 11 different RFLP types (Fig. 2C). A phylogenetic analysis (not shown) based on the presence and absence of restriction sites showed that the EPEC and EHEC strains can be separated into three groups consistent with the previously defined phylogeny based on MLEE (Fig. 2A). RFLP data generated by *Sau*3A1 revealed a similar phylogeny (data not shown). The RFLP analysis shows that the O111:H9 strain (921-B4) belongs to a branch between the EPEC 2 and EHEC 2 groups, consistent with the MLEE-based dendrogram (Fig. 2A); however, the RFLP analysis could not resolve the relationship between EPEC 2 and EHEC 2 strains (C. Herbelin, S. D. Reid, and T. S. Whittam, Abstr. 99th Gen. Meet. Am. Soc. Microbiol., p. 237, 1999).

Sequence analysis of genes in *mutS-rpoS* region. We sequenced genes located in the *mutS-rpoS* intervening region with genomic DNA isolated from four strains, DEC 1a and E2348/69 (EPEC 1 group), DEC 9f (EHEC 2), and DEC 12e (EPEC 2). The novel DNA sequence found in these strains contains the same ORFs (*yclC*, *pad1*, and *slyA*) found in the *mutS-rpoS* region of *E. coli* O157:H7 (Carter et al., Abstr. 99th Gen. Meet. Am. Soc. Microbiol., 1999). The 5′ end of the intergenic sequence between *o454* and the *yclC-slyA* segment contains 100 bp with more than 90% identity to the 3′-end sequence of *rpoS* from *S. enterica* serovar Typhi and a sequence with homology to sequence directly downstream from the 3′ end of *rpoS* in serovars Typhimurium and Typhi. This sequence appears to be a remnant of an ancestral inversion (see Discussion) and is oriented in the opposite direction of the *E. coli* *rpoS* gene (Fig. 1).

Although the restriction analysis indicates substantial variation in the size and gene content of the *mutS-rpoS* region among pathogenic groups, sequence analysis reveals that individual genes are highly conserved (Fig. 4). Pairwise comparison of the sequences shows that the percentage of polymorphic nucleotides ranges from 1.3 to 3.7% for genes in the *mutS-rpoS* region (Table 3). This level of nucleotide polymorphism is intermediate between those for the highly conserved *rpoS* sequences and the more variable *mutS* sequences that flank the region (Table 3).

The degree of selective constraint on sequence divergence can also be seen in comparison of the differences at synonymous (d_S) and nonsynonymous (d_N) sites. For all comparisons (Table 3), d_S exceeds d_N ($d_N - d_S < 0$), a pattern indicating the past action of purifying selection against mutations resulting in amino acid replacements. These results indicate that the genes between *mutS* and *rpoS* have levels of polymorphism similar to those for *mutS* and *rpoS* and are conserved at the amino acid level, with divergence attributable to the accumulation of silent substitutions (Fig. 4).

To analyze the history of sequence divergence in *mutS-rpoS* region, sequence data corresponding to the ORFs common to each group of strains were combined. For comparison of K-12 to the EPEC and EHEC strains, we combined the coding sequences for two adjacent genes, *o258* and *o454* (Fig. 3), which covered 2,136 nucleotides and included 48 variable sites. We inferred a neighbor-joining phylogenetic tree for the *o258-o454* genomic region from the divergence at synonymous sites for the 712 codons of the combined genes (Fig. 5A). The phylogeny shows that the EPEC 1 strains (DEC 1a and E2348/69) are most divergent, differing from the sequences of the other three strains at more than 2% of the synonymous sites. The *o258-o454* sequence of K-12 shares its most recent ancestor with the homologous region of DEC 9f (EHEC 2) and DEC 12e (EPEC 2) strains, which are themselves very similar (Fig. 5A). The results suggest that the *o258-o454* region was present in the ancestral genome prior to the divergence of the EPEC and EHEC 2 groups and the K-12 lineage.

To compare the O157:H7 sequence with those of the EPEC and EHEC 2 groups, the coding sequences of three genes (*yclC*, *pad1*, and *slyA*) were combined. The combined sequence is 2,421 bp long (807 codons), with a total of 133 variable nucleotide sites including 17 that predict amino acid differences. A neighbor-joining tree, constructed using divergence at synonymous sites, separates the five strains into three divergent branches differing at ~6% of the synonymous sites based on this region (Fig. 5B). The EPEC 1 strains (DEC 1a and E2348/69) are closely related to each other, as are the EPEC 2 (DEC 12e) and EHEC 2 (DEC 9f) strains (Fig. 5B). Although the O157:H7 sequence joins with the EPEC 1 branch in the phylogeny (Fig. 5B), this node is not supported by a significant

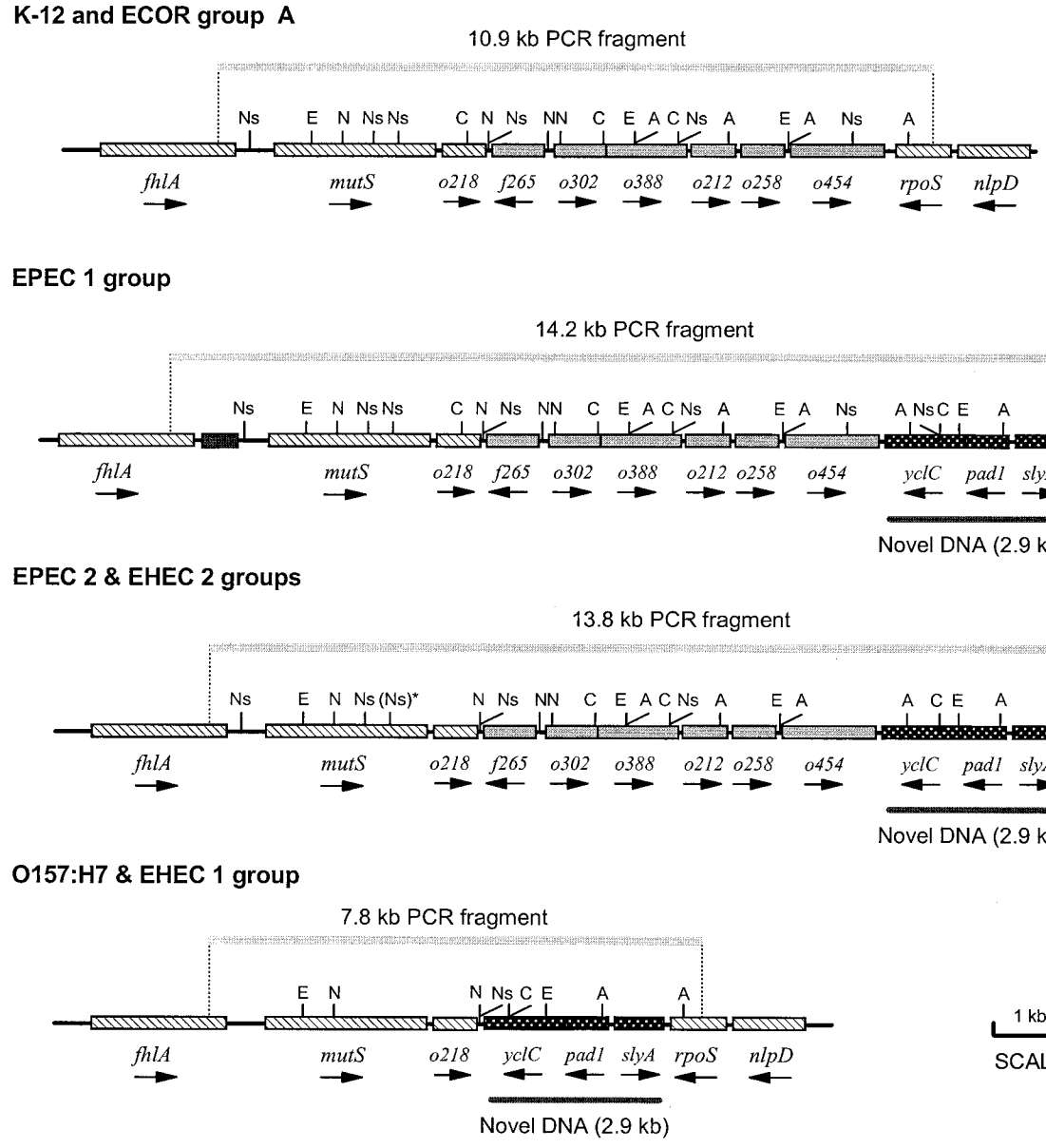

FIG. 3. Restriction maps of the *mutS-rpoS* chromosomal region. Approximate locations of restriction sites for five restriction enzymes: *Eco*RV (E), *Nde*I (N), *Acc*I (A), *Csp*45 (C), and *Nsp*I (Ns). The pattern of restriction sites is conserved among strains of each pathogenic group with the exception of the second *Nsp*I site in *mutS* [(Ns)*], which is present in EPEC 2 strains but absent in EHEC 2 strains. A distinct *Nsp*I map was obtained for 921-B4 (not shown). The novel DNA segment found in EPEC and EHEC strains is located at the 3′ end of *rpoS* and is highlighted with the gray bar.

bootstrap value. In this case, the analysis suggests that the *yclC-slyA* region was also present in the common ancestral strain prior to the radiation of the pathogenic lineages.

DISCUSSION

A key observation of this study is the substantial size variation occurring in the *mutS-rpoS* region between clonal groups of pathogenic *E. coli*. Strains from the EHEC 1 group, including O157:H7 and O55:H7, have a distinctively short intervening region. In a previous study, LeClerc et al. (19) observed a similar length for the *mutS-rpoS* region in O157:H7 and O55:H7 strains, which are closely related. They found larger deletions at the 3′ end of *mutS* in the "mutator" phenotype of an O157:H7 strain, which is characterized by a higher frequency of mutations that confer antibiotic resistance. They also reported the presence of a novel DNA sequence (~2.7 kb) in the *mutS-rpoS* region in nonmutator strains of O157:H7 serotype, as well as in O55:H7 and two related ECOR strains. Here we also found the novel DNA sequence (~2.9 kb) reported by LeClerc et al. (19) in the *mutS-rpoS* region of EPEC groups and EHEC 2 strains but not in strain K-12 or related isolates of the ECOR group A. The extent to which the length and composition of the *mutS-rpoS* region influence local or genomewide mutation rates remains to be elucidated.

Evolutionary model. Several lines of evidence support the idea that the ancestral *E. coli* had a long *mutS-rpoS* region that contained both of the segments that are now found separately in *E. coli* K-12 and O157:H7 strains. First, a long genomic

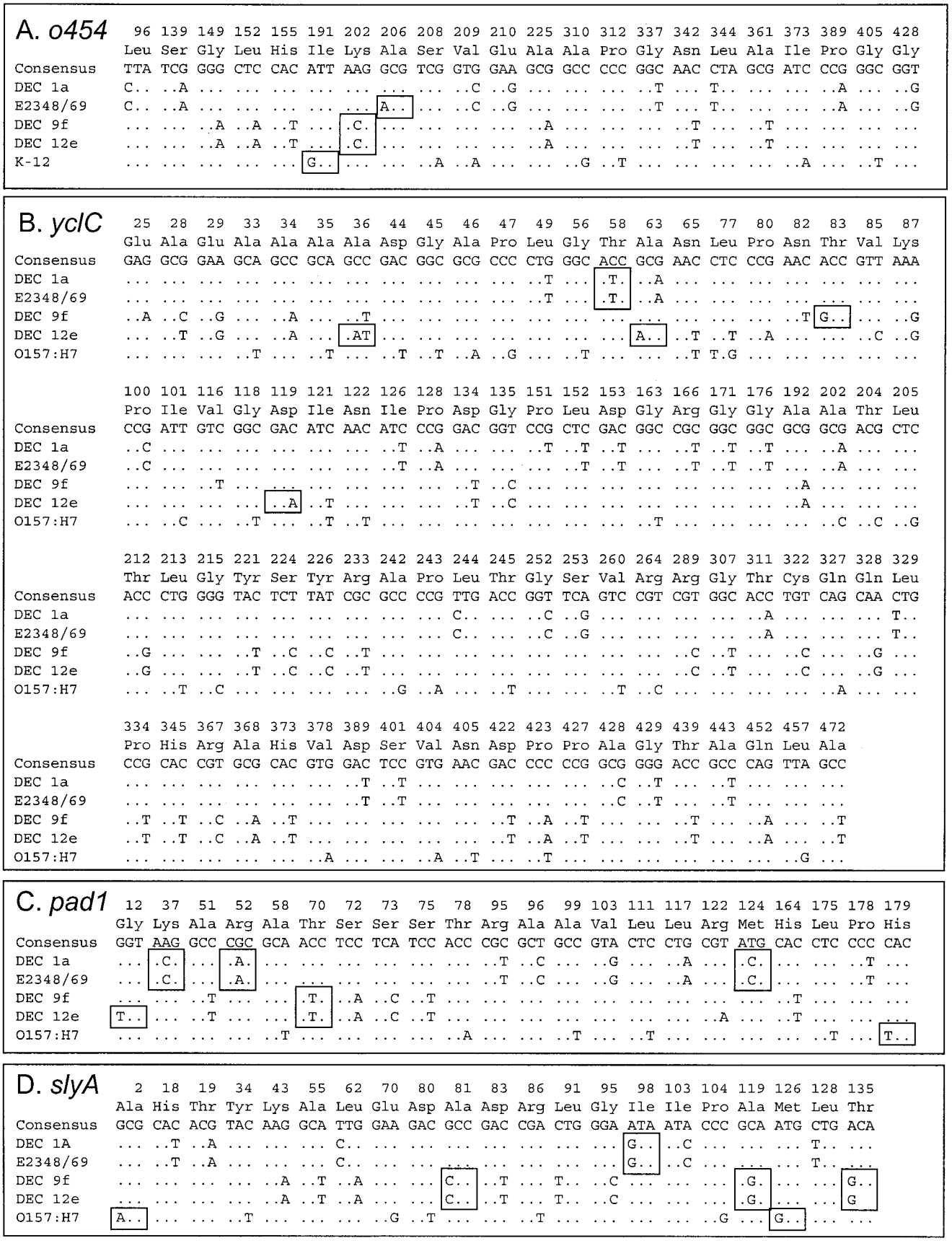

FIG. 4. Polymorphic codons (pc) of four protein-coding genes in the *mutS-rpoS* region. Boxes highlight amino acid replacements. (A) Three out of 22 pc predict replacements (I191V, K202T, and A206T) in *o454* (1,362 bp); (B) 5 out of 86 pc predict replacements (A36D, T58I, A63T, T83A, and D119E) in *yclC* (1,425 bp); (C) 6 out of 22 pc predict replacements (C12C, K37T, R52H, T70I, M124T, and H179Y) in *pad1* (591 bp); (D) 6 out of 21 pc predict replacements (A2T, A81P, I98V, A119G, M126V, and T135A) in *slyA* (405 bp).

TABLE 3. Sequence divergence[a] between genes in the *mutS-rpoS* region

Comparison	% *p*	Mean ± SD		$d_N - d_S$[b]
		d_S	d_N	
DEC 1a vs DEC 9f				
mutS	5.5	25.02 ± 2.49	0.19 ± 0.11	−24.8
o258	2.7	6.65 ± 1.55	1.55 ± 0.52	−5.1
o454	1.2	4.22 ± 1.10	0.10 ± 0.10	−4.1
yclC	3.8	17.10 ± 2.43	0.19 ± 0.13	−16.9
pad1	2.4	7.12 ± 2.28	0.91 ± 0.45	−6.2
slyA	3.7	13.18 ± 4.07	1.28 ± 0.64	−11.9
rpoS	1.3	5.21 ± 1.75	0.00 ± 0.00	−5.2
DEC 1a vs K-12				
mutS	5.9	27.17 ± 2.63	0.26 ± 0.13	−26.9
o258	2.5	5.50 ± 1.76	1.55 ± 0.52	−4.0
o454	1.3	4.52 ± 1.14	0.10 ± 0.14	−4.4
rpoS	1.3	5.21 ± 1.75	0.00 ± 0.00	−5.2
DEC 1a vs O157:H7				
mutS	5.6	25.53 ± 2.52	0.19 ± 0.11	−25.3
yclC	3.9	16.38 ± 2.37	0.19 ± 0.13	−16.2
pad1	2.4	7.10 ± 2.27	0.91 ± 0.45	−6.2
slyA	3.2	11.80 ± 3.81	0.96 ± 0.56	−10.8
rpoS	0.4	5.81 ± 1.86	0.00 ± 0.00	−5.8

[a] Measured by the percentage of polymorphic nucleotide sites (% *p*), the number of synonymous substitutions per 100 synonymous sites (d_S), and the number of nonsynonymous substitutions per 100 nonsynonymous sites (d_N).

[b] A measure of the level of selective constraint and conservation of amino acid sequence.

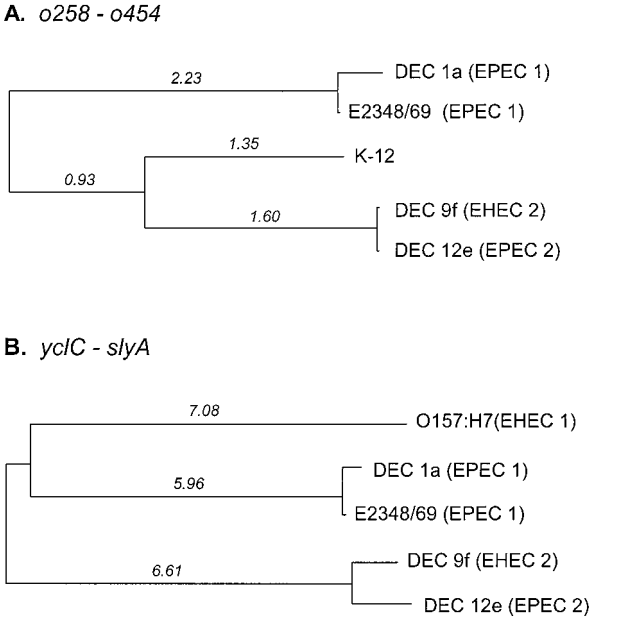

FIG. 5. Phylogenetic trees of the genomic region between *mutS* and *rpoS*. The trees were constructed by the neighbor-joining algorithm, with genetic distance measured by the number of synonymous substitutions per 100 synonymous sites. (A) Gene phylogeny for the combined coding sequences *o258* and *o454* found in K-12, EPEC, and EHEC strains; (B) gene phylogeny for the combined sequences of *yclC*, *pad1*, and *slyA* found in *E. coli* O157:H7, EPEC, and EHEC strains.

region, with the genes in identical order, is found in the two highly divergent EPEC lineages. Individual gene phylogenies (Fig. 5) are compatible with the phylogeny of the pathogenic clones (Fig. 2A). The phylogenies each have four deep branches, consistent with the idea that the sequences were present in the most recent common ancestor of the pathogenic groups. Second, the genes of the regions are conserved, with an average rate of synonymous substitution that greatly exceeds the nonsynonymous rate (Table 3), indicating that the coding sequence is under purifying selection. The pattern of synonymous codon usage is similar to that found for the majority of *E. coli* conserved genes, an observation that is counter to the hypothesis that the region was recently acquired as foreign DNA (20). The sequence analysis showed that most of the divergence at the sequence level is silent and that the function of the proteins has been, for the most part, conserved. Both observations support the idea that the genomic region is old.

The hypothesis that the primitive *mutS-rpoS* genomic region contained all of the genes now found in the various *E. coli* lineages implies that sequence divergence was accompanied by major deletions that eventually gave rise to the shortened intergenic region now seen in strains K-12 and O157:H7. An evolutionary scenario for these changes is outlined in Fig. 6. The cladogram to the left depicts the branching pattern for the clonal frames of the pathogenic groups supported by previous data from MLEE (45). In this phylogeny, the ancestral lineages leading to EPEC 1 strains split first followed by the O157:H7 lineage (EHEC 1), the K-12 and ECOR group A lineages, and finally the split of the EPEC 2 and EHEC 2 groups. Given that EPEC 1 is the most basal group, the primitive ancestor is posited to have a long intergenic region with both *f265-o454* and *yclC-slyA* segments. This ancestral arrangement of genes is conserved as lineages diverge and is found in the contemporary EHEC 2 and EPEC groups. To account for the shortened regions, two major deletions are hypothesized to have occurred: the loss of *yclC-slyA* in the branch leading to the ECOR group A and the loss of *f265-o454* in the branch leading to EHEC 1. The *yclC-slyA* deletion must have happened before the recent radiation of the ECOR group A. The *f265-o454* deletion also must have occurred before the most recent common ancestor of O157:H7 and other members of the EHEC 1 group (8). This deletion must also have preceded the divergence of EHEC 1 from related ECOR strains (ECOR 37 and 42) (26, 35) that have the same proximal borders as O157:H7 based on colony hybridizations (20). Finally, we hypothesize that the EPEC 1 group acquired a small insert upstream of *mutS*; however, this could be an ancient remnant that was lost near the base of the cladogram. The nature of this event can eventually be resolved by sequencing of this region and comparison to more divergent outgroups.

The evolutionary model (Fig. 6) is a parsimonious explanation for the contemporary DNA polymorphism in the *mutS-rpoS* region. Other possible scenarios would require multiple gains and losses of the entire region or pieces of the region in independent lineages. Moreover, the source of the imported DNA would have to be such that the mutations in the sequences would be consistent with the chromosomal background that followed the divergence of the clonal groups (17). These alternative models cannot be ruled out at this point. The simple model (Fig. 6), requiring two independent deletions, can be tested against these alternatives by examining the *mutS-rpoS* region in other groups of *E. coli*.

Testing the model with *Salmonella* genomic sequences. One prediction from the phylogenetic analysis is that the long intergenic region was present in the most recent ancestor of the *E. coli* groups. To test this idea, we did a BLAST comparison of the *mutS-o454* segment of K-12 and the *o454-slyA* segment

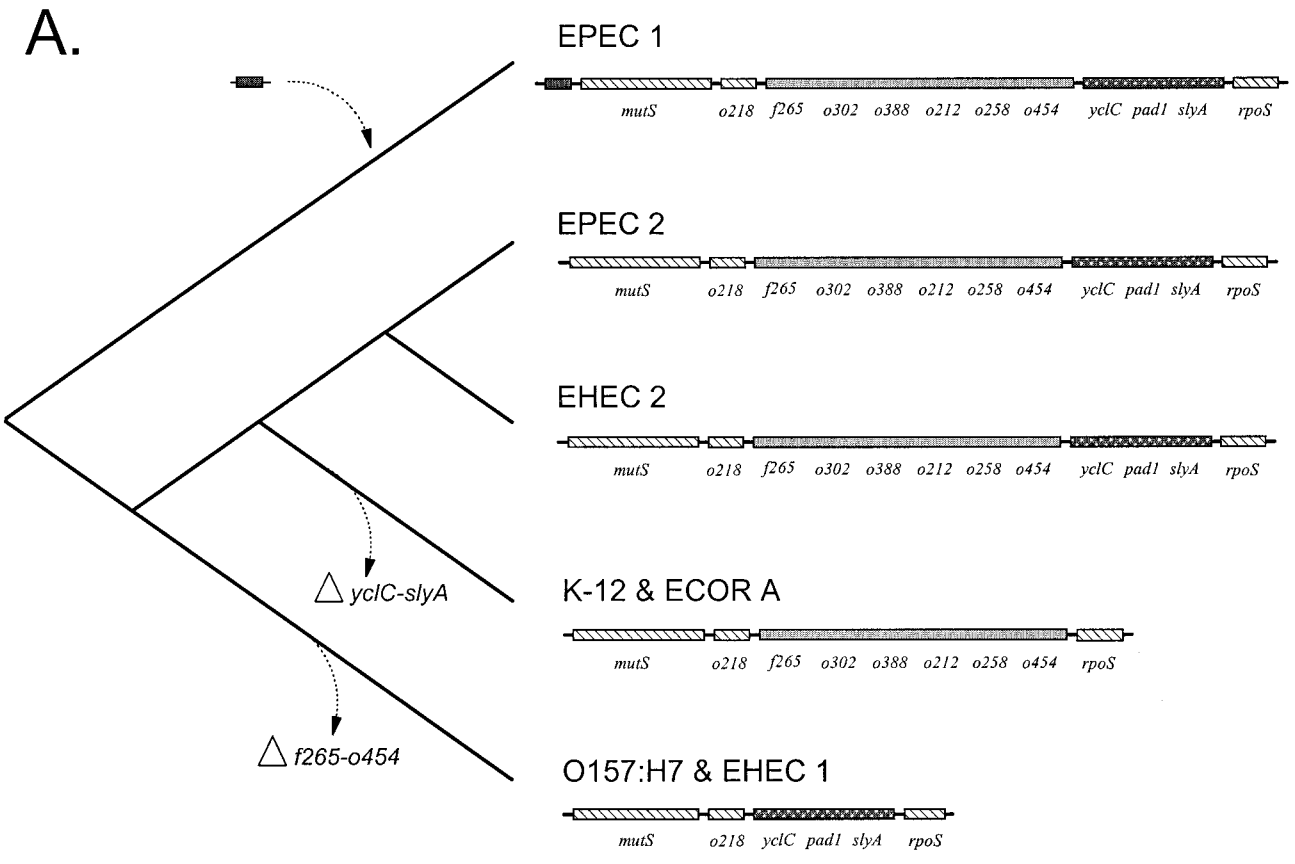

FIG. 6. Evolutionary model of the *mutS-rpoS* genomic region. (A) The left side is a cladogram of the phylogeny of the groups; the right side shows a diagram of the genes in the *mutS-rpoS* region. The EPEC 1, EPEC 2, and EHEC 2 strains have a conserved ancestral sequence in both the *f265-o454* and *yclC-slyA* gene segments. The model predicts two independent deletions of gene segments: the loss of *yclC-slyA* in the branch leading to K-12 and the ECOR group A strains and the loss of the *f265-o454* segment in the branch leading to O55:H7, O157:H7, and other EHEC 1 strains. It is not clear if the additional DNA upstream of *mutS* in the EPEC 1 strains is acquired (as marked here) or is ancestral and has been lost early in divergence. (B) Orientation of the genes for ancestral *E. coli* and *Salmonella*. The location of a short sequence with high homology to the *Salmonella rpoS* gene is marked as the *rpoS* remnant.

from EPEC1 against the unfinished microbial genomes of *S. enterica* serovars Typhimurium, Typhi, Paratyphi A, and Enteritidis. The purpose of this search was to determine if homologous sequences occur in *S. enterica* and to infer the extent to which the arrangement has been preserved in the 100 million years of separation of these bacterial species. The search produced 30 sequences with significant alignments (results not shown). These sequences were assembled into a single contig with DNASTAR. The National Center for Biotechnology Information ORF finder and subsequent BLAST searches were used to identify homologous genes in *E. coli*.

The search of the unfinished *Salmonella* genomes yielded

two important results. First, homologs to all of the genes in the *mutS-rpoS* intervening region of *E. coli* occur in *Salmonella* genomes, and the sequences can be assembled into a contig. This finding supports the evolutionary model that the ancestral *mutS-rpoS* region contained both *f265-o454* and *yclC-slyA* segments and that the short *mutS-rpoS* regions of K-12 and O157:H7 groups are derived states. The level of sequence divergence also supports the hypothesis that these segments are ancestral; for example, the *yclC* genes from *E. coli* O157:H7 and EPEC 1 are 0.84% divergent in amino acid sequence and 2.5% divergent from the *Salmonella yclC* homolog. Second, the orientation of the *f265-o454* and *yclC-slyA* segments, relative to *mutS* and *rpoS* (Fig. 6B), has changed in such a way that there have been at least two inversions since *S. enterica* and *E. coli* shared a common ancestor. The order of the homologous genes in the *f265-o454* and *yclC-slyA* segments is conserved; however, each of these segments lies in the opposite orientation relative to *mutS* and *rpoS* in *S. enterica* (Fig. 6B). This inverted arrangement may also account for the remnant *Salmonella rpoS* sequence located between *o454* and *yclC* in the *E. coli* genome. We suggest that this remnant is a piece of *rpoS* that was carried with the ancient inversion that resulted in the present orientation of *yclC-slyA* in *E. coli*.

Putative function of the *yclC-slyA* genes. The observation that the novel DNA is conserved in the EPEC and EHEC pathogenic groups suggests that the products of *yclC*, *pad1*, and *slyA* function in pathogenesis. Comparison to homologous proteins yields few insights into the role of these genes in pathogenic *E. coli*. For example, in *Saccharomyces cerevisiae*, the *yclC* gene encodes a transmembrane voltage-gated Cl^- protein with 13 hydrophobic domains (14), and the *pad1* gene encodes phenylacrylic acid decarboxylase (PAD), which confers resistance to phenylacrylic acids (6). The predicted 242-amino-acid PAD polypeptide is 48% identical to the product of *dedF* of *E. coli* (6). It is a single-copy gene in the yeast genome and not essential for viability (6). *pad*-related genes have also been described for *Bacillus subtilis*, *Bacillus pumilus*, and *Lactobacillus plantarum* (4).

A function of SlyA in pathogenesis is suggested by results from *Salmonella*, where it has been shown to play a role in bacterial survival in the intracellular environment of host macrophages (3, 22). In *Salmonella* infection, SlyA regulates expression of multiple proteins during stationary phase and upon phagocytosis by macrophages (3). Its expression is required for the destruction of M cells but not for invasion or colonization of the murine small intestine (7). A homologous gene has been found at 37 min on the *E. coli* K-12 genome, and its product confers a hemolytic phenotype by activating expression of *clyA*, which encodes a member of the RTX toxin family (23, 24, 34).

Moreover, SlyA is distantly related to a broad family of bacterial regulatory proteins affecting diverse aspects of bacterial physiology, such as repression of microcin production, intrinsic multiple antibiotic resistance, and repression of growth in *E. coli*. The family includes MrpA, HpcR, MarA, and Prs from *E. coli*, Hpr from *B. subtilis*, and PecS from *Erwinia chrysanthemi* (41). Globally, members of this broad family play a role in the internal economy of the cell and govern functions crucial for survival, inactivation of deleterious exogenous compounds, cytotoxicity to the host, and acquisition of a resistant phenotype. Despite the sequence similarity, the nature of expression of *yclC*, *pad1*, or *slyA*, as well as the function of the proteins in pathogenesis and bacterial survival, remains to be evaluated.

Conclusion. The *mutS-rpoS* region has diverged dramatically among pathogenic groups of *E. coli*, accumulating many point mutations in conserved genes, as well as undergoing changes in gene content. Strains of three pathogenic groups (EPEC 1, EPEC 2, and EHEC 2) contain a full array of genes between *rpoS* and *mutS* which is hypothesized to reflect the primitive state found before *E. coli* separated from *Salmonella*. The evolutionary model proposed here invokes two separate deletion events that resulted in the shorter *mutS-rpoS* genomic region, characteristic of the *E. coli* O157:H7 and K-12 lineages, and may contribute to the ecological specialization of these bacteria.

ACKNOWLEDGMENTS

We thank Andrew Clark and Heidi Waldrip for use of the DNA ProScan program and Sheila Plock for assistance with the Applied Biosystems 373A automated sequencer. Preliminary sequence data were obtained from The Institute for Genomic Research website at http://www.tigr.org.

This research was supported by Public Health Service grant AI 42391.

REFERENCES

1. **Blattner, F. R., G. Plunkett III, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glassner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao.** 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277:**1453–1462.
2. **Bopp, C. A., K. D. Greene, F. P. Downes, E. G. Sowers, J. G. Wells, and I. K. Wachsmuth.** 1987. Unusual verotoxin-producing *Escherichia coli* associated with hemorrhagic colitis. J. Clin. Microbiol. **25:**1486–1489.
3. **Buchmeier, N., S. Bossie, C. Y. Chen, F. C. Fang, D. G. Guiney, and S. J. Libby.** 1997. SlyA, a transcriptional regulator of *Salmonella typhimurium*, is required for resistance to oxidative stress and is expressed in the intracellular environment of macrophages. Infect. Immun. **65:**3725–3730.
4. **Cavin, J. F., V. Dartois, and C. Divies.** 1998. Gene cloning, transcriptional analysis, purification, and characterization of phenolic acid decarboxylase from *Bacillus subtilis*. Appl. Environ. Microbiol. **64:**1466–1471.
5. **Christopher-Hennings, J., J. A. Willgohs, D. H. Francis, U. A. K. Raman, R. A. Moxley, and D. J. Hurley.** 1993. Immunocompromise in gnotobiotic pigs induced by verotoxin-producing *Escherichia coli* (O111:NM). Infect. Immun. **61:**2304–2308.
6. **Clausen, M., C. J. Lamb, R. Megnet, and P. W. Doerner.** 1994. PAD1 encodes phenylacrylic acid decarboxylase which confers resistance to cinnamic acid in *Saccharomyces cerevisiae*. Gene **142:**107–112.
7. **Daniels, J. J., I. B. Autenrieth, A. Ludwig, and W. Goebel.** 1996. The gene *slyA* of *Salmonella typhimurium* is required for destruction of M cells and intracellular survival but not for invasion or colonization of the murine small intestine. Infect. Immun. **64:**5075–5084.
8. **Feng, P., K. A. Lampel, H. Karch, and T. S. Whittam.** 1998. Genetic and phenotypic changes in the emergence of *Escherichia coli* O157:H7. J. Infect. Dis. **177:**1750–1753.
9. **Fletcher, J. N., H. E. Embaye, B. Getty, R. M. Batt, C. A. Hart, and J. R. Saunders.** 1992. Novel invasion determinant of enteropathogenic *Escherichia coli* plasmid pLV501 encodes the ability to invade intestinal epithelial cells and HEp-2 cells. Infect. Immun. **60:**2229–2236.
10. **Griffin, P. M., S. M. Orstoff, R. V. Tauxe, K. D. Greene, J. G. Wells, J. R. Lewis, and P. A. Blake.** 1988. Illnesses associated with *Escherichia coli* O157:H7 infections. Ann. Intern. Med. **109:**705–712.
11. **Groisman, E. A., and H. Ochman.** 1993. Cognate gene clusters govern invasion of host epithelial cells by *Salmonella typhimurium* and *Shigella flexneri*. EMBO J. **12:**3779–3787.
12. **Hacker, J., and J. B. Kaper.** 1999. The concept of pathogenicity islands, p. 1–11. *In* J. Hacker and J. B. Kaper (ed.), Pathogenicity islands and other mobile virulence elements. American Society for Microbiology, Washington, D.C.
13. **Hengge-Aronis, R.** 1996. Regulation of gene expression during entry into stationary phase, p. 1497–1512. *In* F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd ed. American Society for Microbiology, Washington, D.C.
14. **Huang, M. E., J. C. Chuat, and F. Galibert.** 1994. A voltage-gated chloride channel in the yeast *Saccharomyces cerevisiae*. J. Mol. Biol. **242:**595–598.
15. **Karmali, M. A., B. T. Steele, M. Petric, and C. Lim.** 1983. Sporadic cases of haemolytic-uremic syndrome associated with faecal cytotoxin and cytotoxin-producing *Escherichia coli* in stools. Lancet **i:**619–620.
16. **Kumar, S., K. Tamura, and M. Nei.** 1993. MEGA: molecular evolutionary genetics analysis, version 1.0. The Pennsylvania State University, University Park, Pa.
17. **Lawrence, J. G.** 1999. Gene transfer, speciation, and the evolution of bacterial genomes. Curr. Opin. Microbiol. **2:**519–522.

18. **Lawrence, J. G., and H. Ochman.** 1998. Molecular archaeology of the *Escherichia coli* genome. Proc. Natl. Acad. Sci. USA **95:**9413–9417.
19. **LeClerc, J. E., B. Li, W. L. Payne, and T. A. Cebula.** 1996. High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. Science **274:** 1208–1211.
20. **LeClerc, J. E., B. Li, W. L. Payne, and T. A. Cebula.** 1999. Promiscuous origin of a chimeric sequence in the *Escherichia coli* O157:H7 genome. J. Bacteriol. **181:**7614–7617.
21. **Levine, M. M., D. R. Natlin, R. B. Hornick, E. J. Bergquist, D. H. Waterman, C. R. Young, and S. Sotman.** 1978. *Escherichia coli* strains that cause diarrhoea but do not produce heat-labile or heat-stable enterotoxins and are non-invasive. Lancet **i:**1119–1122.
22. **Libby, S. J., W. Goebel, A. Ludwig, N. Buchmeier, F. Bowe, F. C. Fang, D. G. Guiney, J. G. Songer, and F. Heffron.** 1994. A cytolysin encoded by *Salmonella* is required for survival within macrophages. Proc. Natl. Acad. Sci. USA **91:**489–493.
23. **Ludwig, A., S. Bauer, R. Benz, B. Bergmann, and W. Goebel.** 1999. Analysis of the SlyA-controlled expression, subcellular localization and pore-forming activity of a 34 kDa haemolysin (ClyA) from *Escherichia coli* K-12. Mol. Microbiol. **31:**557–567.
24. **Ludwig, A., C. Tengel, S. Bauer, A. Bubert, R. Benz, H. J. Mollenkopf, and W. Goebel.** 1995. SlyA, a regulatory protein from *Salmonella typhimurium*, induces a haemolytic and pore-forming protein in *Escherichia coli*. Mol. Gen. Genet. **249:**474–486.
25. **Maurelli, A. T., R. E. Fernandez, C. A. Bloch, C. K. Rode, and A. Fasano.** 1998. "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. Proc. Natl. Acad. Sci. USA **95:**3943–3948.
26. **McGraw, E. A., J. Li, R. K. Selander, and T. S. Whittam.** 1999. Molecular evolution and mosaic structure of α, β, and γ intimins of pathogenic *Escherichia coli*. Mol. Biol. Evol. **16:**12–22.
27. **Mills, D. M., V. Bajaj, and C. A. Lee.** 1995. A 40 kb chromosomal fragment encoding *Salmonella typhimurium* invasion genes is absent from the corresponding region of the *Escherichia coli* K-12 chromosome. Mol. Microbiol. **15:**749–759.
28. **Moyenuddin, M., I. K. Wachsmuth, S. L. Moseley, C. A. Bopp, and P. A. Blake.** 1989. Serotype, antimicrobial resistance, and adherence properties of *Escherichia coli* strains associated with outbreaks of diarrheal illness in children in the United States. J. Clin. Microbiol. **27:**2234–2239.
29. **Nataro, J. P., and J. B. Kaper.** 1998. Diarrheagenic *Escherichia coli*. Clin. Microbiol. Rev. **11:**142–201.
30. **Nei, M., and T. Gojobori.** 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:**418–426.
31. **Ochman, H., and E. A. Groisman.** 1996. Distribution of pathogenicity islands in *Salmonella* spp. Infect. Immun. **64:**5410–5412.
32. **Ochman, H., and E. A. Groisman.** 1994. The origin and evolution of species differences in *Escherichia coli* and *Salmonella typhimurium*, p. 479–493. *In* B. Schierwater, B. Streit, G. P. Wagner, and R. DeSalle (ed.), Molecular ecology and evolution: approaches and applications. Birkhauser Verlag, Basel, Switzerland.
33. **Ochman, H., and R. K. Selander.** 1984. Standard reference strains of *Escherichia coli* from natural populations. J. Bacteriol. **157:**690–693.
34. **Oscarsson, J., Y. Mizunoe, B. E. Uhlin, and D. J. Haydon.** 1996. Induction of haemolytic activity in *Escherichia coli* by the *slyA* gene product. Mol. Microbiol. **20:**191–199.
35. **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.
36. **Reid, S. D., D. J. Betting, and T. S. Whittam.** 1999. Molecular detection and identification of intimin alleles in pathogenic *Escherichia coli* by multiplex PCR. J. Clin. Microbiol. **37:**2719–2722.
37. **Riley, L. W., R. S. Remis, S. D. Helgerson, H. B. McGee, J. G. Wells, B. R. Davis, R. J. Hebert, E. S. Olcott, L. M. Johnson, N. T. Hargrett, P. A. Blake, and M. L. Cohen.** 1983. Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. N. Engl. J. Med. **308:**681–685.
38. **Rode, C. K., L. J. Melkerson-Watson, A. T. Johnson, and C. A. Bloch.** 1999. Type-specific contributions to chromosome size differences in *Escherichia coli*. Infect. Immun. **67:**230–236.
39. **Rupp, W. D.** 1996. DNA repair mechanisms, p. 2277–2294. *In* F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd ed., vol. 2. American Society for Microbiology, Washington, D.C.
40. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:**406–425.
41. **Thomson, N. R., A. Cox, B. W. Bycroft, G. S. Stewart, P. Williams, and G. P. Salmond.** 1997. The rap and hor proteins of Erwinia, Serratia and Yersinia: a novel subgroup in a growing superfamily of proteins regulating diverse physiological processes in bacterial pathogens. Mol. Microbiol. **26:**531–544.
42. **Viljanen, M. K., T. Peltola, S. Y. T. Junnila, L. Olkkonen, H. Järvinen, M. Kuistila, and P. Huovinen.** 1990. Outbreak of diarrhoea due to *Escherichia coli* O111:B4 in schoolchildren and adults: association of Vi antigen-like reactivity. Lancet **336:**831–834.
43. **Whittam, T. S.** 1998. Evolution of *Escherichia coli* O157:H7 and other Shiga toxin-producing *E. coli* strains, p. 195–209. *In* J. B. Kaper and A. D. O'Brien (ed.), *Escherichia coli* O157:H7 and other Shiga toxin-producing *E. coli* strains. American Society for Microbiology, Washington D.C.
44. **Whittam, T. S., and E. A. McGraw.** 1996. Clonal analysis of EPEC serogroups. Rev. Microbiol. **27**(Suppl. 1)**:**7–16.
45. **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.