The Evolution of Insertion Sequences Within Enteric Bacteria

Jeffrey G. Lawrence,[1] Howard Ochman and Daniel L. Hartl[2]

Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63110

Manuscript received October 10, 1991
Accepted for publication January 8, 1992

ABSTRACT

To identify mechanisms that influence the evolution of bacterial transposons, DNA sequence variation was evaluated among homologs of insertion sequences IS*1*, IS*3* and IS*30* from natural strains of *Escherichia coli* and related enteric bacteria. The nucleotide sequences within each class of IS were highly conserved among *E. coli* strains, over 99.7% similar to a consensus sequence. When compared to the range of nucleotide divergence among chromosomal genes, these data indicate high turnover and rapid movement of the transposons among clonal lineages of *E. coli*. In addition, length polymorphism among IS appears to be far less frequent than in eukaryotic transposons, indicating that nonfunctional elements comprise a smaller fraction of bacterial transposon populations than found in eukaryotes. IS present in other species of enteric bacteria are substantially divergent from *E. coli* elements, indicating that IS are mobilized among bacterial species at a reduced rate. However, homologs of IS*1* and IS*3* from diverse species provide evidence that recombination events and horizontal transfer of IS among species have both played major roles in the evolution of these elements. IS*3* elements from *E. coli* and Shigella show multiple, nested, intragenic recombinations with a distantly related transposon, and IS*1* homologs from diverse taxa reveal a mosaic structure indicative of multiple recombination and horizontal transfer events.

INSERTION sequences (IS) are short segments of DNA, typically 1–2 kilobasepairs (kb) in length, with the ability to translocate within and among replicons (GALAS and CHANDLER 1989). IS mediate numerous molecular and genetic phenomena, including gene activation (GLANDSDORFF, CHARLIER and ZAFARULLA 1980), repression (SAEDLER *et al.* 1974), deletion (CHOW and BROKER 1981), rearrangement (SAEDLER *et al.* 1980), recombination (LIAB 1980), and transfer (CHANDLER, CLERGET and CARO 1980), and are also of epidemiological importance, owing to their ability to form composite transposons and mobilize antibiotic resistance determinants (BERG 1977; KLECKNER *et al.* 1975). Although numerous classes of IS have been characterized from the genomes of enteric bacteria, factors that contribute to the evolution of IS within and among bacterial species have not been clearly defined. Elucidating the evolutionary forces acting upon insertion elements is necessary for understanding the genetics of mobile genetic elements in prokaryotes, the impact of IS-mediated events, and the evolution and epidemiology of composite transposons.

Bacterial reproduction confers a clonal population structure if gene exchange is infrequent. Although recent studies have shown that recombination does occur at chromosomal loci among natural isolates of *E. coli* (DUBOSE, DYKHUIZEN and HARTL 1988; STOLTZFUS, LESLIE and MILKMAN 1988), linkage disequilibrium among enzyme electrophoretic types has indicated that large scale genetic exchange of metabolic genes among natural strains of *E. coli* is not common (SELANDER and LEVIN 1980; CAUGANT, LEVIN and SELANDER 1981; OCHMAN and SELANDER 1984a; WHITTAM, OCHMAN and SELANDER 1984). However, the contribution of genetic exchange to the evolution of transposable elements within *E. coli* has not been well characterized. Not only are IS present in variable numbers and positions within bacterial genomes (SAWYER *et al.* 1987; LAWRENCE *et al.* 1989), but a substantial portion of the *E. coli* transposon pool is plasmid borne (SAWYER *et al.* 1987; HALL *et al.* 1989), which increases opportunies for genetic transfer.

SAWYER *et al.* (1987) examined the distribution and abundance of IS*1*, IS2, IS*3*, IS*4*, IS*5* and IS*30*, and HALL *et al.* (1989) the distribution of IS*103*, in natural isolates of *E. coli*, chosen to represent the range of phenotypic variation detected by protein electrophoresis. They determined that IS could be classified into three groups based on the apparent strength of regulation of transposition: IS*1* and IS*5* represented weakly regulated transposons; IS*2*, IS*4*, IS*30* and IS*103* represented moderately regulated transposons; and IS*3* represented a class of strongly regulated transposons (see also HARTL and SAWYER 1988a,b). Although these analyses suggest different patterns of

[1] Current address: Department of Biology, University of Utah, Salt Lake City, Utah 84112.
[2] To whom correspondence should be addressed.

regulation, they did not establish which factors govern the evolution of insertion sequences. In addition, the degree to which these factors differentially influence the evolution of each class of transposons is not known.

To address these issues, isoforms of three insertion sequences, IS*1*, IS*3* and IS*30*, were studied from isolates of the ECOR reference collection of *E. coli* (OCHMAN and SELANDER 1984b). In addition, IS homologs from related species of enteric bacteria were isolated and analyzed. Nucleotide variation among classes of IS was examined to elucidate the mechanisms influencing the evolution of IS within *E. coli*, to evaluate the extent to which genetic exchange contributes to the evolution of transposon populations with *E. coli*, to determine if similar processes influence the evolution of IS among bacterial species, to assess the relative proportions of functional and non-nonfunctional elements, to identify evolutionarily conserved reading frames and sequence motifs, and to ascertain whether patterns of evolution differ among distinct classes of insertion sequences. In this manner, we may assess the evolutionary influences which differ between bacterial IS and metabolic genes, as well as those which differ between prokaryotic and eukaryotic transposons.

MATERIALS AND METHODS

Strains: Strains of the ECOR collection (OCHMAN and SELANDER 1984b), ATCC 35320-35391, were obtained from laboratory collections. The phenotypic (SELANDER, CAUGANT and WHITTAM 1987) and insertion sequence (SAWYER *et al.* 1987) profiles of these strains have been described. *Escherichia fergusonii* ATCC 35469 and ATCC 35471, *Escherichia hermannii* ATCC 33652, *Escherichia vulneris* ATCC 29943, *Shigella dysenteriae* ATCC 13313, *Shigella flexneri* ATCC 29508, *Shigella sonnei* ATCC 29930 and *Serratia odorifera* ATCC 3307 were obtained from laboratory stocks. Relationships among these taxa inferred from chromosomal gene sequences have been described (LAWRENCE, OCHMAN and HARTL 1991).

Southern blotting: Chromosomal DNA was isolated, digested with restriction endonucleases, size fractionated on agarose gels and transferred to nylon membranes as described previously (SAWYER *et al.* 1987). Probes were prepared as internal portions of IS amplified via the polymerase chain reaction (PCR; SAIKI *et al.* 1985, 1988). Primers utilized for amplification annealed internal to the inverted repeats: IS*1* forward: GATTTAGTGTATGATGG; IS*1* reverse: GATAGTGTTTTATGTTC; IS*3* forward: GGACACGCGGCTAAGTG; IS*3* reverse: TGGACACAGGCCTAAGCG; IS*30* forward: GCAACAGTTATGTGAAA; IS*30* reverse: AATGCAACACCCCTTTC. Amplification products were purified by the method of LAWRENCE, HARTL and OCHMAN (1991a), and labeled to high specific activity by the method of FEINBERG and VOGELSTEIN (1983). Membranes were hybridized under moderate stringency conditions as described (SAWYER *et al.* 1987).

DNA sequencing: Internal segments of each IS were amplified by the PCR utilizing either genomic DNA, for strains containing a single copy of any one IS, or DNA fragments, size fractionated by gel electrophoresis and identified by Southern blotting, as a template. The oligonucleotide primers described above allowed amplification of a 729-bp fragment comprising 95% of IS*1*, a 1220-bp fragment encompassing 97% of IS*3*, and an 1185-bp fragment comprising 97% of IS*30*. Amplification products were purified and sequenced according to the methods of DUBOSE and HARTL (1990) and LAWRENCE, HARTL and OCHMAN (1991a). For copies of IS*3* and IS*30*, the nucleotide sequence of the entire region between the amplification primers was determined. For IS*1* resident in *E. coli*, nucleotide sequences downstream of position 400 were determined. In all cases, the nucleotide sequences of both strands of each IS were determined. For genomes containing elements that could not be amplified by the PCR, chromosomal DNA was partially digested with *Sau*3A, ligated to EMBL3 digested with *Bam*HI, and packaged in Gigapack (Stratagene). Bacteriophage DNA was prepared from appropriate clones by the method of HELMS *et al.* (1985), partially digested with *Sau*3A, ligated into M13 vectors digested with *Bam*HI, and introduced into *E. coli* JM101 by electroporation (DOWER, MILLER and RAGSDALE 1988). Appropriate clones were selected for DNA sequencing.

Computer analysis: Parsimony analysis was implemented by PAUP (D. SWOFFORD), and divergence calculations utilized the GCG program package (DEVEREUX, HAEBERLI and SMITHIES 1984). Phylogeny testing employed the program MATRIX2 (LAWRENCE and HARTL 1992).

TABLE 1

Single nucleotide differences within IS*1* copies from *E. coli*

	Strain[a]											
Position	K12	5	28	48	12	32	33	51	52	53	66	60
393	A	C	–	–	C	C	C	C	C	C	C	C
396	C	T	–	–	–	–	–	–	–	–	–	–
418	C	–	–	–	–	–	–	–	A	–	–	–
486	C	–	–	–	–	–	–	G	–	–	–	–

[a] K12, *E. coli* K12 (OHTSUBO and OHTSUBO 1978); numbers refer to ECOR strains (OCHMAN and SELANDER 1984b). Strains 5, 28, and 48 are isoforms of IS*1*R; the remaining strains are isoforms of IS*1*F. Positions which distinguish IS*1*F from IS*1*R were not included (see text). The dash indicates a nucleotide identical to that in K12.

RESULTS

Nucleotide substitutions: The DNA sequences of insertion elements IS*1*, IS*3* and IS*30* revealed little nucleotide variation in the form of base substitutions among strains of *E. coli* (Tables 1-3). Both types of IS*1* were detected: IS*1*R, originally isolated from plasmid R100 (OHTSUBO and OHTSUBO 1978), and IS*1*F, first isolated from *S. flexneri* (OHTSUBO *et al.* 1984). Although they differ by 10% at the nucleotide level, little variation was detected within each isoform. Similarly, IS*3* and IS*30* were also virtually monomorphic, despite their presence in bacterial strains that are quite distinct based on enzyme electrophoretic profiles (Figure 1). Figure 1 also indicates strains containing multiple copies of either IS*1*F or IS*1*R. The isoforms were distinguished by (1) restriction fragment length polymorphism analysis of PCR-amplified regions of IS*1* cleaved with restriction endonucleases that distinguish

TABLE 2

Single nucleotide differences within IS*3* copies from *E. coli*

	Strain[a]											
Position	K12	1	14	18	23	30	43	46	50	58	69	70
38	A	–	–	–	–	–	–	–	–	T	–	–
80	A	–	–	–	–	–	G	–	–	–	–	–
200	A	–	G	–	G	G	G	G	G	G	G	–
250	G	–	–	–	–	–	A	–	–	–	–	–
483	C	–	–	–	–	–	T	–	–	–	–	–
485	C	–	–	A	–	–	–	–	–	–	–	–
716	G	–	–	–	–	–	–	–	–	–	T	–
1019	G	A	–	–	–	–	–	–	–	–	–	–
1125	C	–	–	–	–	–	–	–	T	–	–	–
1214	A	–	C	–	–	–	C	–	–	–	C	–

[a] K12, *E. coli* K12 (TIMMERMAN and TU 1985); numbers refer to ECOR strains (OCHMAN and SELANDER 1984b). The nucleotide sequence of ECOR 63 appears in Figure 3. The dash indicates a nucleotide identical to that in K12.

TABLE 3

Single nucleotide differences within IS*30* copies from *E. coli*

	Strain[a]												
Position	K12	1	8	14	19	23	24	31	35	36	50	56	71
74	T	A	A	A	A	A	A	A	A	A	A	A	A
75	A	T	T	T	T	T	T	T	T	T	T	T	T
141	A	G	G	G	G	G	G	G	G	G	G	G	G
217	G	–	A	–	A	–	–	–	–	–	–	A	–
220	A	C	–	–	–	–	–	–	–	–	–	–	–
236	C	–	–	–	–	–	–	–	–	–	–	T	–
266	A	–	–	–	–	–	–	–	–	–	T	–	T
276	T	–	–	–	–	–	–	–	–	–	C	–	–
461	G	–	–	–	–	–	T	–	–	–	–	–	–
479	G	–	–	–	–	–	–	–	–	–	–	–	T
486	A	–	–	–	–	–	–	–	C	C	–	–	–
594	C	–	–	–	A	–	–	–	–	–	–	–	–
755	A	G	G	G	G	G	G	G	G	–	G	G	G
758	G	–	–	–	–	–	–	A	–	–	A	–	–
899	C	–	–	–	–	–	–	–	–	–	A	–	–
983	C	–	–	–	–	T	–	–	–	–	–	–	–
1046	C	–	–	–	–	T	–	–	–	–	–	–	–

[a] K12, *E. coli* K12 (DALRYMPLE *et al.* 1984); numbers refer to ECOR strains (OCHMAN and SELANDER 1984b). The dash indicates a nucleotide identical to that in K12.

IS*1*R from IS*1*F (strains 2, 24, 29, 42, 55, 57, 59 and 72), and (2) DNA sequencing of a 280-bp region in particular IS*1* copies that included 29 sites distinguishing the two isoforms (strains 3, 6 and 30). The distributions of IS*1*F and IS*1*R are not confined to any one group of *E. coli*, but are widely distributed among divergent strains. For example, although strains 30, 32 and 33 are quite closely related, strain 30 contains IS*1*R, while strains 32 and 33 contain IS*1*F. We have not identified any strain containing both IS*1*R and IS*1*F. The multiple copies of IS*1* in strains 2, 3, 6, 24, 29, 30, 42 and 72 are isoforms of IS*1*R, while those in strains 55, 57 and 59 are IS*1*F. However, the laboratory strain *E. coli* K12 harbors five IS*1*R and one IS*1*F (UMEDA and OHTSUBO 1991).

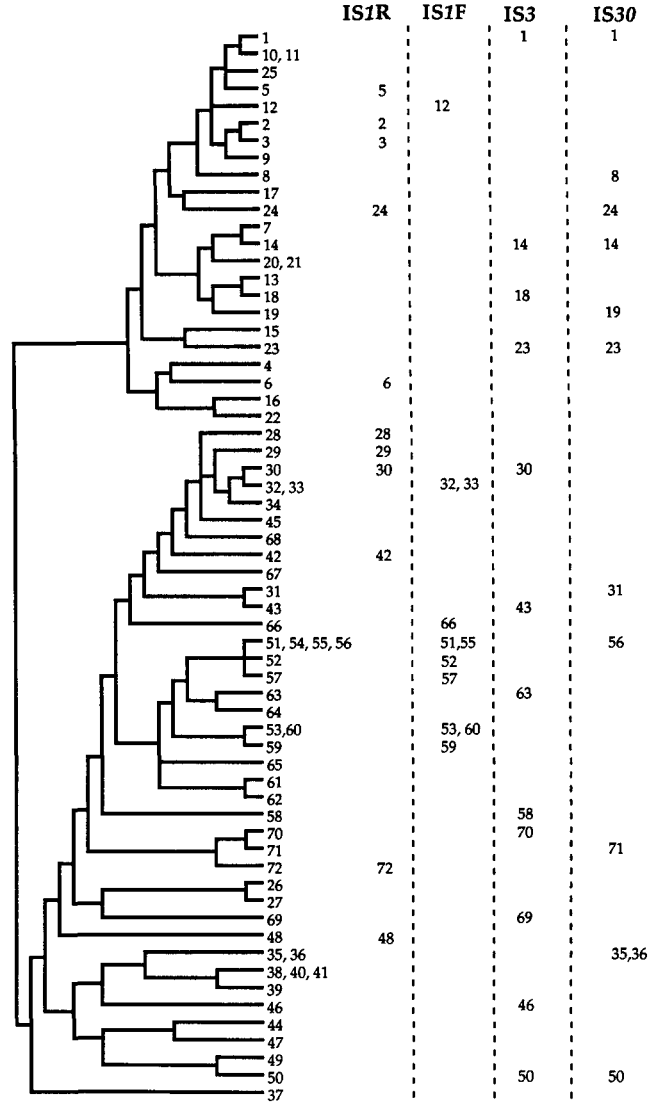


FIGURE 1.—Phylogenetic relationships of *E. coli* strains from which IS were isolated and analyzed [tree after SELANDER, CAUGANT and WHITTAM (1987)]. Strains from which copies of IS*1*R, IS*1*F, IS*3* and IS*30* were isolated are indicated.

Figure 2 shows the nucleotide sequences of IS*1* elements resident in the genomes of related enteric bacteria. IS*1*-related sequences were isolated from three species of Shigella (*S. dysenteriae*, *S. flexneri*, and *S. sonnei*) and their nucleotide sequences agree well with IS*1* sequences previously published from these taxa (OHTSUBO *et al.* 1984), differing by only single substitutions. IS*1* homologs were also isolated from *E. fergusonii*, *E. hermannii* and *E. vulneris*. [Although classified as Escherichia, *E. hermannii* and *E. vulneris* are as divergent from *E. coli* as species of Enterobacter (LAWRENCE, HARTL and OCHMAN 1991b; LAWRENCE, OCHMAN and HARTL 1991).] Relationships among IS*1* homologs isolated from these species are presented in Table 4. The relationships among the elements are not congruent with those inferred from chromosomal loci, notably *gap* and *ompA*, which encode glucose-3-

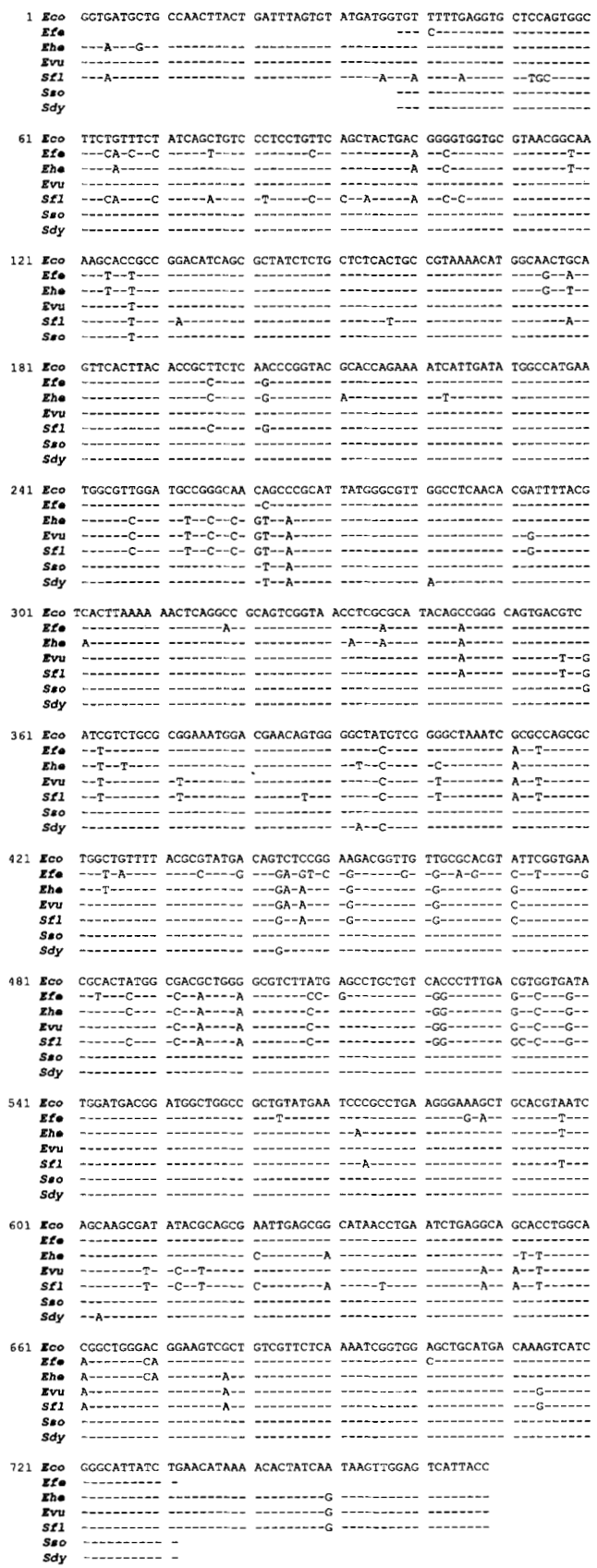


FIGURE 2.—Sequences of IS*1* from enteric bacteria. Eco, *E. coli* (OHTSUBO and OHTSUBO 1978); Sdy, *S. dysenteriae*; Sfl, *S. flexneri*; Sso, *S. sonnei*, Efe, *E. fergusonii*; Ehe, *E. hermannii*, Evu, *E. vulneris*. Numbering begins at the first base of the inverted repeat.

TABLE 4

Relationships among IS*1* homologs from enteric bacteria

Bacteria	Percent nucleotide identity			
	Efe[a]	Ehe	Evu	Sfl
Bases 1–768				
Eco	91.1	92.3	94.4	90.5
Efe		93.1	91.4	90.3
Ehe			94.8	93.4
Evu				95.8
Bases 1–246				
Eco	92.3	94.3	99.6	90.2
Efe		95.7	92.8	91.9
Ehe			94.7	90.7
Evu				90.7
Bases 246–540				
Eco	87.1	88.4	89.1	88.4
Efe		90.1	89.8	89.1
Ehe			95.2	94.6
Evu				98.6
Bases 541–768				
Eco	95.8	95.2	95.6	93.4
Efe		94.8	92.1	90.6
Ehe			94.3	94.7
Evu				97.8

[a] Taxon designations: Eco, *E. coli* K12; Efe, *E. fergusonii*; Ehe, *E. hermannii*; Evu, *E. vulneris*; Sfl, *S. flexneri*.

TABLE 5

Test of recombination within IS*1* and IS*3*

Bootstrap data set[a]	Comparison data set	Distribution		Similarity coefficient	P^b
		μ	σ		
IS*1*s2	IS*1*s1	0.892	0.076	0.600	0.001
IS*1*s1	IS*1*s2	0.910	0.062	0.600	0.000
IS*1*s3	IS*1*s1	0.894	0.073	0.520	0.000
IS*1*s1	IS*1*s3	0.891	0.065	0.520	0.000
IS*1*s3	IS*1*s2	0.910	0.066	0.480	0.000
IS*1*s2	IS*1*s3	0.893	0.065	0.480	0.000
IS*3a*	IS*3c*	0.898	0.051	0.505	0.000
IS*3c*	IS*3a*	0.925	0.042	0.505	0.000

[a] IS*1*s1: nucleotides 1–246; IS*1*s2: nucleotides 247–540; IS*1*s3: nucleotides 541–768 of IS*1* (see Table 4). IS*3a*, IS*3c*: IS*3* ORFs (see Figure 5).

[b] Significance estimated as the number of similarity coefficients generated by resampling methods that are smaller than that observed between the bootstrap and comparison data sets (LAWRENCE and HARTL 1991).

phosphate dehydrogenase and outer membrane protein 3A, respectively (LAWRENCE, OCHMAN and HARTL 1991). Moreover, the evolutionary histories of different segments of IS*1* appear to be substantially different. For example, while the 5′ end of IS*1* from *E. vulneris* is nearly identical to IS*1*R, the remainder of the element is quite different and closely resembles IS*1*F. The three portions of IS*1* defined in Table 4 describe different taxonomic relationships to a high level of significance by the test of LAWRENCE and HARTL (1992) ($P < 0.001$, Table 5).

Homologs of IS*3* were isolated from *S. dysenteriae*,

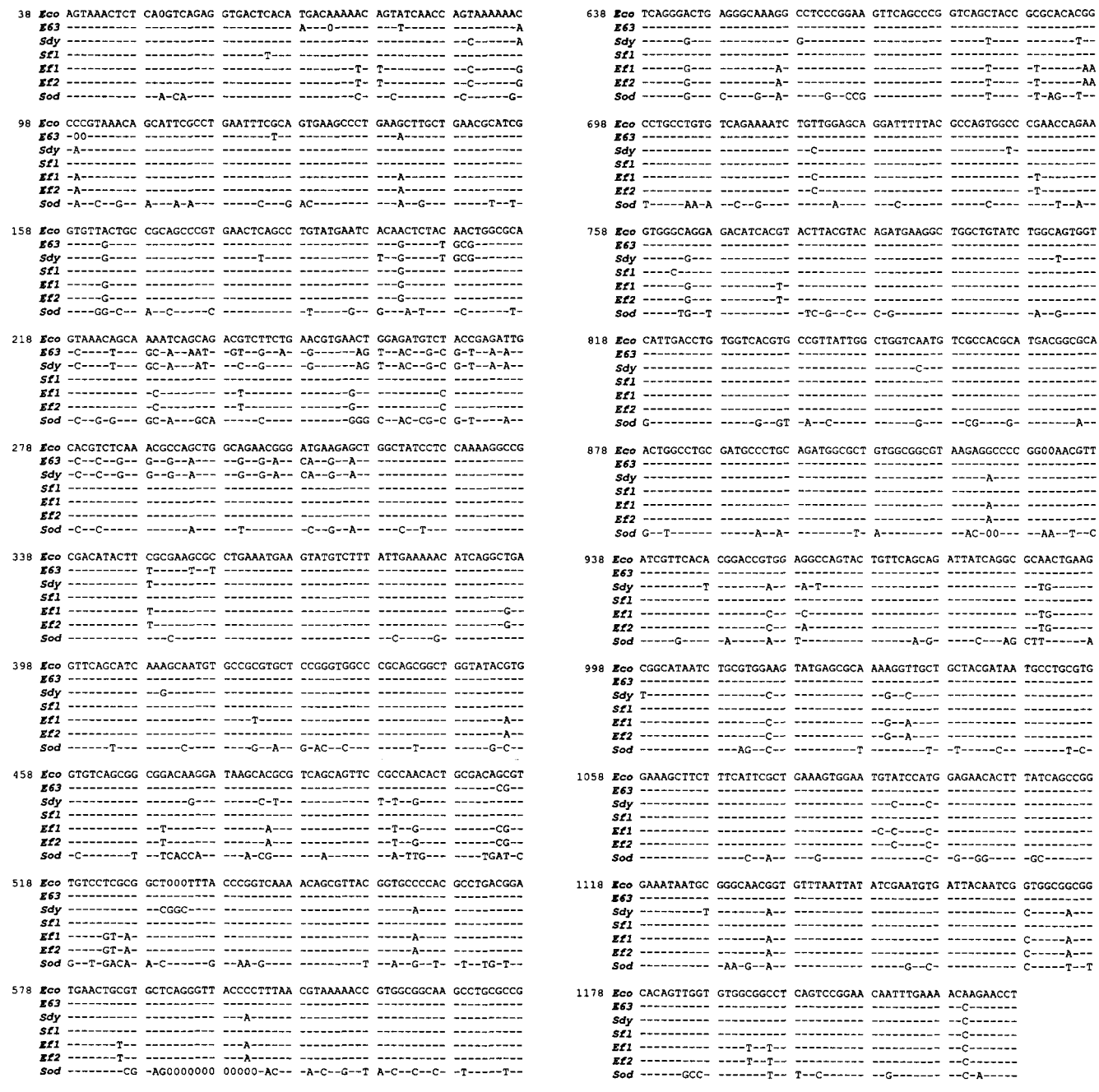
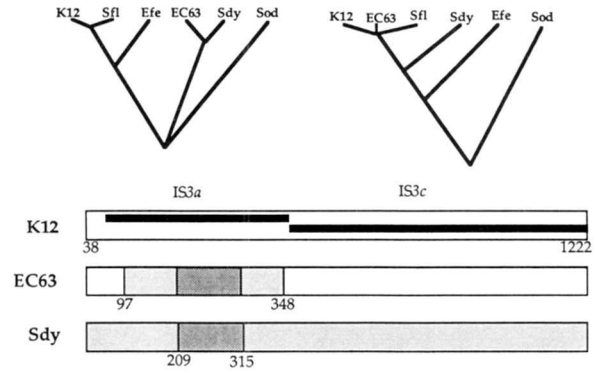


FIGURE 3.—Sequences of IS*3* from enteric bacteria. Eco, *E. coli* K12 (TIMMERMAN and TU 1985); E63, *E. coli* ECOR 63; Sdy, *S. dysenteriae*; Sfl, *S. flexneri*, Ef1, *E. fergusonii* ATCC 35469; Ef2, *E. fergusonii* ATCC 35471; Sod, *S. odorifera*. Numbering begins at the first base of the inverted repeat.

S. flexneri, *E. fergusonii* and *S. odorifera* (Figure 3). While most copies from *E. coli* were monomorphic, the IS*3* from ECOR 63 showed evidence of intragenic recombination. Figure 5 details the regions of IS*3* from *E. coli* K12, ECOR 63, and *S. dysenteriae* involved in the putative recombination events. In the region from 349 to 1222 bp, IS*3* from ECOR 63 is 99.4% identical to its K12 counterpart, and 95% identical to the IS*3* from Shigella. However, both the ECOR 63 and the Shigella elements are only 65% identical to the K12 element in the region from 209 to 315 bp, and 96% identical to each other. The occurrence of a large number of differences in a limited and sharply bounded region strongly suggests the region bounded by 209–315 bp was the result of intragenic recombination between a Shigella-like IS*3* and a distantly related element. Following the initial recombination event, a second intragenic recombination transferred the divergent sequence, as well as additional Shigella-like sequences from 97 to 348 bp, to the IS*3* now resident in ECOR 63. As a result of the recombination, the two putative open reading frames (ORFs), IS*3a* and IS*3c*, support substantially different phylogenetic trees (Figure 5) and the taxonomic relation-

```
   1 Eco TGTAGATTCA ATTGGTCAAC GCAACAGTTA TGTGAAAACA TGGGGTTGCG GAGGTTTTTT
     Ehe ---------- ---------- A-------C- ---------- --TA------ ----A--G--

  61 Eco GAATGAGACG AACTATTACA GCAGAGGAAA AAGCCTCTGT TTTTGAACTA TGGAAGAACG
     Ehe ---------- ---AT----- ---A------ --A-A----- ---------G ----------

 121 Eco GAACAGGCTT CAGTGAAATA ACGAATATCC TGGGTTCAAA ACCCGGAACG ATCTTCACTA
     Ehe ---A------ ------G--- G-T------- -------G-- ---A--G--C --------G-

 181 Eco TGTTAAGGGA TACTGGCGGC ATAAAACCCC ATGAGCGTAA GCGGGCTGTA GCTCACCTGA
     Ehe --------A-- ---------- -------A-A -------C-- ---A-----T ----------

 241 Eco CACTGTCTGA GCGCGAGGAG ATACGAGCTG GTTTGTCAGC CAAAATGAGC ATTCGTGCGA
     Ehe -G-------- ---------A -------C-- --C------- T--------- --------C-

 301 Eco TAGCTACTGC GCTGAATCGC AGTCCTTCGA CGATCTCACG TGAAGTTCAG CGTAATCGGG
     Ehe -C-----AAT ---------T --C-----A- -A-------- ---------- --------T-

 361 Eco GCAGACGCTA TTACAAAGCT GTTGATGCTA ATAACCGAGC CAACAGAATG GCGAAAAGGC
     Ehe -T--G----- ---------C --------G- -------G-- ------G--- --TC----A-

 421 Eco CAAAACCGTG CTTACTGGAT CAAAATTTAC CATTGCGAAA GCTTGTTCTG GAAAAGCTGG
     Ehe ---------- ---------- --G---A-G- -G------G- ---------- --------T-

 481 Eco AGATGAAATG GTCTCCAGAG CAAATATCAG GATGGTTAAG GCGAACAAAA CCACGTCAAA
     Ehe -------G-- ---C------ ---------- ---------- ------G--G --G-------

 541 Eco AAACGCTGCG AATATCACCT GAGACAATTT ATAAAACGCT GTACTTTCGT AGCCGTGAAG
     Ehe ----A--A-- ---------- ---------- ----G----- T--------- --T-------

 601 Eco CGCTACACCA CCTGAATATA CAGCATCTGC GACGGTCGCA TAGCCTTCGC CATGGCAGGC
     Ehe ---------- ------CG-- -----C---- -C--A----- C-----C--T ----------

 661 Eco GTCATACCCG CAAAGGCGAA AGAGGTACGA TTAACATAGT GAACGGAACA CCAATTCACG
     Ehe -C-----T-- ------T--- -----C--T- -C-------- ---------G -----C----

 721 Eco AACGTTCCCG AAATATCGAT AACAGACGCT CTCTAGGGCA TTGGGAGGGC GATTTAGTCT
     Ehe ---------- GC-------- ---------- ----G--A-- C-----A--- ----------

 781 Eco CAGGTACAAA AAACTCTCAT ATAGCCACAC TTGTAGACCG AAAATCACGT TATACGATCA
     Ehe -G-------- ------C--- ---------- ---------- ---------- ----------

 841 Eco TCCTTAGACT CAGGGGCAAA GATTCTGTCT CAGTAAATCA GGCTCTTACC GACAAATTCC
     Ehe ----G----- ---------- --CG--T-A- ----G----- ---------A --T-------

 901 Eco TGAGTTTACC GTCAGAACTC AGAAAATCAC TGACATGGGA CAGAGGAATG GAACTGGCCA
     Ehe ----C----- AC-T------ C-GCG----- ---------- ---------- --G-------

 961 Eco GACATCTAGA ATTTACTGTC AGCACCGGCG TTAAAGTTTA CTTCTGCGAT CCTCAGAGTC
     Ehe -------G-- --------G- -----A--A- ---------- ---------- --------C-

1021 Eco CTTGGCAGCG GGGAACAAAT GAGAACACAA ATGGGCTAAT TCGGCAGTAC TTTCCTAAAA
     Ehe -C-------- ---------C --A-----T- ---------- C--A------ -----C----

1081 Eco AGACATGTCT TGCCCAATAT ACTCAACATG AACTAGATCT GGTTGCTGCT CAGCTAAACA
     Ehe -------C-- ---------- ------T--- -G--G----A ------A--- ----------

1141 Eco ACAGACCGAG AAAGACACTG AAGTTCAAAA CACCGAAAGA GATAATTGAA AGGGGTGTTG
     Ehe ---------- ---------- ---------- ----T----- ---------- ----------

1201 Eco CATTGACAGA TTGAATCTAC A
     Ehe -G-------- ---------- -
```

FIGURE 4.—Sequences of IS*30* from enteric bacteria. Eco, *E. coli* K12 (DALRYMPLE, CASPERS and ARBER 1984); Ehe, *E. hermannii*. Numbering begins at the first base of the inverted repeat.



	Region 209-315 Percent Identity		Region 97-208 Percent Identity		Region 349-1222 Percent Identity	
IS	K12	Sdy	K12	Sdy	K12	Sdy
EC63	63.3	96.3	93.6	96.4	99.4	95.0
Sdy	66.4	100	92.9	100	95.3	100

FIGURE 5.—Representation of intragenic recombination within IS*3*. K12, *E. coli* K12; EC63, *E. coli* ECOR 63; Efe, *E. fergusonii*; Sdy, *S. dysenteriae*; Sfl, *S. flexneri*; Sod, *S. odorifera*. The location of the two putative ORFs, IS*3a* and IS*3c*, are indicated with heavy lines. Dendrograms were constructed using parsimony methods from sequences of either the IS*3a* or IS*3c* reading frames. Consistency indices for the trees are 0.905 and 0.848, respectively.

TABLE 6

Length differences within insertion sequences from *E. coli*

IS	Strain	Position	Event
IS*1*	ECOR32	544-545	Deletion of AT
IS*1*	ECOR33	544-545	Deletion of AT
IS*3*	ECOR63	71	Deletion of A
IS*3*	ECOR63	98-99	Deletion of CC
IS*3*	ECOR18	493/494	Insertion of CAGT
IS*3*	ECOR50	1166-1182	Deletion of 17 bp
IS*30*	ECOR24	1071-1072	Deletion of TT
IS*30*	ECOR24	1070/1071	Insertion of IS*3411*

ships inferred from these two ORFs are significantly different ($P < 0.001$; Table 5).

The single IS*30* isolated from *E. hermannii* was 89.5% identical to the *E. coli* K12 element (Figure 4). Southern blot analysis revealed the presence of sequences related to IS*30* in the genomes of *Escherichia blattae* and *E. fergusonii*; however, the relative hybridization intensity indicated that either the IS*30* sequence was much more divergent than that resident in the *E. hermannii* genome, or only a small fragment of the element remained in that genome (data not shown).

Length variation: In addition to the nucleotide substitutions and recombinations described above, several insertions and deletions were observed within various IS (Table 6). The IS*30* from ECOR 24 carried an insertion of the unrelated transposon IS*3411*. In addition, the IS*3* from ECOR 18 carried a 4-bp duplication which may have arisen as a target duplication from transposon insertion and subsequent excision. In total, five of the 35 elements examined from *E. coli* exhibited variation in length which would disrupt a major ORF.

The IS*30* and IS*1* sequences from other enteric species were the same length as the *E. coli* element, but several IS*3* elements differed in size (Figure 3). The IS*3* copy in *S. dysenteriae* contains a 3-bp insertion conserving the IS*3c* reading frame. The *S. odorifera* element exhibits 12- and 2-bp deletions, and a 2-bp insertion, which together conserve the reading frame of IS*3c*. The 3 bp deleted from IS*3a* in ECOR 63, although conserving the reading frame in this element, probably arose as two separate events (Table 6). Since the putative initiation codon has been mutated to AAG in this isoform, it is unclear whether the second deletion was favored by selection to restore the proper reading frame.

Distribution of IS*3411*: The IS*30* from ECOR 24 exhibits a 1.2-kb insertion which proved to be IS*3411*. Although these sequences were plasmid borne, IS*30*::IS*3411* transposons were not detected in other

TABLE 7

Distribution of IS*3411* within the ECOR strains

Strain[a]	Chromosomal	Plasmid	Strain	Chromosomal	Plasmid
2	6	1	39	13	2
3	6	1	40	2	0
4	1	1	41	0	1
7	0	1	43	0	2
8	3	0	44	0	2
9	6	1	49	0	1
10	6	1	50	8	1
11	11	1	51	1	1
13	0	1	52	1	0
15	1	0	53	0	1
18	1	0	54	0	1
19	5	0	55	1	0
20	2	0	56	1	0
21	1	1	57	1	0
23	0	1	59	0	1
24	7	6	60	1	0
25	5	0	61	1	0
30	1	1	62	6	1
34	4	1	63	1	0
35	0	7	64	8	1
36	9	1	70	2	0
37	0	2	71	2	1
38	0	2	72	2	1

[a] Number refers to ECOR strain (OCHMAN and SELANDER 1984b). Only strains harboring IS*3411* are listed.

ECOR strains. The distribution of IS*3411* within ECOR strains is presented in Table 7. A total of 174 copies of IS*3411*, 127 chromosomal and 47 plasmid, were detected among 46 strains. The copy number of 3.7 per infected strain is typical of other IS (SAWYER *et al.* 1987). Analysis of the distribution of IS*3411* among ECOR strains by the method of SAWYER *et al.* (1987) suggested a pattern of regulation similar to that of IS*1* and IS*5*. Models with no explicit regulation of transposition (that is, a linear increase of transposition rate with copy number), and a moderate decrease in fitness with increasing copy number, fit the data reasonably well ($P \sim 0.25$). Models with constant or deceasing rate of transposition with increasing copy numbers were strongly rejected.

ORF analysis: GALAS and CHANDLER (1989) proposed 8, 7 and 3 potential ORF for IS*1*, IS*3* and IS*30*, respectively (Figure 6). To determine which ORFs were evolutionarily conserved, and therefore the most likely to encode proteins, nucleotide substitutions in each ORF within divergent copies of each IS were classified as nonsynonymous substitutions, which altered the amino acid sequence, or synonymous substitutions. Conserved ORFs were identified as those with an excess of synonymous substitutions relative to nonsynonymous substitutions, an absence of chain terminating substitutions, and preservation of initiation and termination codons. Data for certain comparisons are presented in Tables 8–10.

IS*1* contains two evolutionarily conserved ORFs,

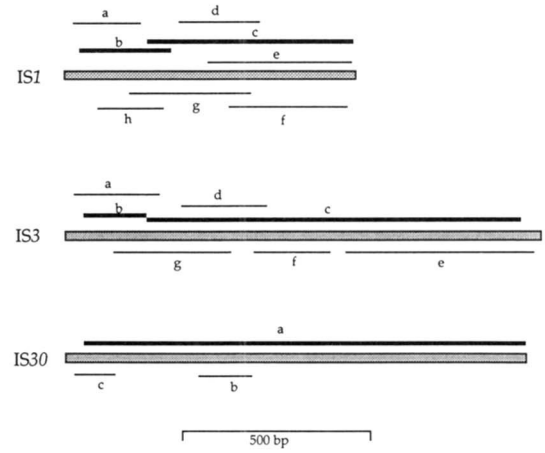


FIGURE 6.—Genetic organization of bacterial insertion sequences (after GALAS and CHANDLER 1989). Reading frames transcribed from the positive strand are placed above each element; those transcribed from the complementary strand are placed below. Heavy lines indicate conserved ORFs.

TABLE 8

Comparison of IS*1* ORFs from *E. coli* and *S. flexneri*

ORF[a]	Begin[b]	End	Stop[c]	Percent Divergence[d] Replacement	Silent	Protein similarity
a	27 (GTG)	239 (TGA)		13.8	2.8	75.4
b	56 (GTG)	331 (TAA)		4.0	32.1	95.2
c	250 (ATG)	753 (TAA)		5.1	33.0	92.0
d	353 (GTG)	511 (TGA)	1	16.6	1.9	71.6
e	376 (ATG)	753 (TAA)		4.2	35.4	93.5
f	719 (ATG)	429 (TAA)	2	7.5	30.7	89.0
g	468 (GTG)	199 (TGA)	2	7.5	11.1	88.4
h	304 (GTG)	89 (GGA)	2	10.4	8.1	83.3
bc	56 (GTG)	753 (TAA)		3.8	36.8	94.6

[a] ORF notation of GALAS and CHANDLER (1989).
[b] Begin, putative initiation codon; End, putative termination codon. Codons in parentheses are present in *S. flexneri*.
[c] Stop = number of nonsense substitutions in the putative ORF.
[d] Divergence values calculated by the method of PERLER *et al.* (1980).

IS*1b* and IS*1c* (Table 8). SEKINE and OHTSUBO (1989) proposed ribosomal frameshifting between these two reading frames as a requisite for the proper production of the IS*1* transposase. This hypothesis is supported by the present data, since both ORFs are conserved among divergent copies. The putative full length protein is described by the IS*1bc* ORF (Table 8), created by a −1 frameshift between these two reading frames. As expected, the 5′ portion of IS*1c*, which would be translated in the frame of IS*1b*, contains an excess of nonsynonymous substitutions in the IS*1c* frame. The remaining potential reading frames in IS*1* are not evolutionarily conserved, bearing either nonsense mutations in the divergent homolog, loss of initiation or termination codons, or an excess of nonsynonymous substitutions which would yield a substantially altered protein product.

TABLE 9

Comparison of IS*3* ORFs from *E. coli* and *S. odorifera*

ORF[a]	Begin[b]	End	Stop[c]	Percent divergence[d] Replacement	Silent	Protein similarity
a	57 (GTG)	362 (TGA)		8.9	109.7	85.4
b	157 (GTG)	357 (TGA)		33.0	6.3	52.9
c	362 (ATG)	1225 (?)		8.5	116.4	88.5
d	415 (GTG)	573 (TGA)	3	44.3	34.4	55.6
e	1240 (?)	968 (TGA)	2	20.3	19.5	74.4
f	691 (ATG)	542 (TGC)	(1)	39.0	26.3	53.2
g	505 (GTC)	233 (TGT)	(1)	18.3	45.0	77.2

[a] ORF notation of GALAS and CHANDLER (1989).
[b] Begin, putative initiation codon; End, putative termination codon. Codons in parentheses are present in *S. odorifera*.
[c] Stop = number of nonsense substitutions in the putative ORF.
[d] Divergence values calculated by the method of PERLER *et al.* (1980).

IS*3* also exhibits two evolutionarily conserved ORFs, IS*3a* and IS*3c* (Table 9). Both ORFS present the excess of synonymous over nonsynonymous substitutions typical of protein coding regions. IS*30* bears a single evolutionarily conserved ORF, IS*30a*, which covers 94% of its length (Table 10). Although IS*30c* appears somewhat conserved, it is encoded on the complementary DNA strand as IS*30a* (Figure 6). It is unlikely that IS*30c* encodes a peptide since it lacks conventional promoters and a Shine-Dalgarno sequence.

DISCUSSION

Variation within *E. coli*: Eleven independent copies of IS*1* representing both major types, 12 copies of IS*3*, and 12 copies of IS*30* were isolated and examined from natural strains of *E. coli* and related enteric bacteria. The data presented in Tables 1–3 indicate that IS present in otherwise divergent isolates of *E. coli* are nearly monomorphic. The copies showed four or fewer substitutions relative to a consensus sequence (0–0.3% different), and these differences were typically restricted to a particular element. In contrast, sequence difference for chromosomal genes in representative strains of *E. coli* ranges from 0 to 3.4% at the *trp* locus (MILKMAN and CRAWFORD 1981), 2–4% at *pho*A (DUBOSE, DYKHUIZEN and HARTL 1988), and 4–16% at *gnd* (DYKHUIZEN and GREEN 1986). The sequence homogeneity of IS copies is also in strong contrast to the diversity in the number and chromosomal position of IS observed among *E. coli* strains (SAWYER *et al.* 1987; LAWRENCE *et al.* 1989).

The high rate of transposition of IS in *E. coli* genomes and their frequent occurrence on plasmids (SAWYER *et al.* 1987) suggests rapid turnover and frequent transfer of IS among strains. This model of population structure is supported by the sequence homogeneity among *E. coli* insertion sequences. Within *E. coli*, each family of IS turns over so rapidly

TABLE 10

Comparison of IS*30* ORFs from *E. coli* and *E. hermannii*

ORF[a]	Begin[b]	End	Percent divergence[c] Replacement	Silent	Protein similarity
a	63 (ATG)	1214 (TGA)	2.5	46.1	96.0
b	557 (GTG)	388 (TCG)	15.4	0.0	72.3
c	202 (ATG)	8 (TGA)	11.4	25.8	83.9

[a] ORF notation of GALAS and CHANDLER (1989).
[b] Begin, putative initiation codon; End, putative termination codon. Codons in parentheses are present in *E. hermannii*.
[c] Divergence values calculated by the method of PERLER *et al.* (1980).

that there is insufficient time to accumulate substantial genetic divergence. High rates of interstrain transfer are also supported by the statistically significant co-occurrence of unrelated IS in *E. coli* genomes, which can be explained quantitatively by simultaneous plasmid-mediated transfer (HARTL and SAWYER 1988a,b). Rapid dissemination and turnover of IS copies in *E. coli* contrasts with the much smaller rate of exchange among conventional chromosomal genes (DUBOSE, DYKHUIZEN and HARTL 1988; STOLTZFUS, LESLIE and MILKMAN 1988; SELANDER and LEVIN 1980; WHITTAM, OCHMAN and SELANDER 1984).

Alternatively, insertion sequences may have been conserved among *E. coli* isolates due to intense selection. Variation at the *gap* locus, encoding the glycolytic enzyme glyceraldehyde-3-phosphate dehydrogenase, is also quite low (NELSON, WHITTAM and SELANDER 1991), as selection on both function and codon usage constrain the evolution of this gene even among distantly related enteric bacteria (LAWRENCE, HARTL and OCHMAN 1991b). However, nucleotide variation among IS homologs from related enteric species do not support this model. Rather, the lack of nucleotide variation among *E. coli* IS supports a model of high turnover within *E. coli* genomes with significant mobilization between isolates.

Variation among enteric bacteria: In contrast to the high rate of transfer of IS among *E. coli* genomes, horizontal transfer of IS among bacterial species is less frequent. For example, the genome of the closely related bacterium *Salmonella typhimurium* contains none of the IS present in *E. coli* K12 (GALAS and CHANDLER 1989), and the only IS isolated from *S. typhimurium* (IS*200*) has not been detected in *E. coli* (LAM and ROTH 1983a,b). The difference in IS pools between *E. coli* and Salmonella implies either that IS are very recent invaders of enteric genomes, that transfer of IS among bacterial species is not common, or that *E. coli* IS cannot proliferate in Salmonella genomes.

The inferred relationships of IS*3* homologs from *E. coli*, *S. dysenteriae*, *E. fergusonii* and *S. odorifera* are consistent with the phylogenetic relationships inferred

from chromosomal genes (LAWRENCE, HARTL and OCHMAN 1991b; LAWRENCE, OCHMAN and HARTL 1991). The IS*3* copies from *E. fergusonii* are virtually identical to each other (Figure 3) and 1.9% and 15.6% divergent from the *E. coli* element at nonsynonymous and synonymous sites, respectively. This level of divergence is consistent with that observed for the *ompA* gene (LAWRENCE, HARTL and OCHMAN 1991b). The IS*3* homolog from *S. odorifera* is substantially different from the *E. coli* sequence. It exhibits 9.0% and 119% divergence (corrected for multiple substitutions) at nonsynonymous and synonymous sites, respectively; these values are also consistent with the *ompA* data (LAWRENCE, HARTL and OCHMAN 1991b). The IS*3* copies from Shigella are more closely related to those from *E. coli* than are those from *E. fergusonii*, which is consistent with the taxonomic placement of these species (EDWARDS and EWING 1962). These data suggest that IS*3* elements have been resident in these species since their separation and have been evolving at approximately the same rate as chromosomal genes.

In constrast, the relationships inferred from IS*1* sequences are not congruent with those inferred from chromosomal genes (Table 4). Sequence divergences among IS*1* copies from *E. hermannii* and *E. vulneris* are substantially smaller than those observed for *gap* and *ompA*, both of which show strong conservation of both synonymous and nonsynonymous sites (LAWRENCE, OCHMAN and HARTL 1991). In addition, while the *ompA* data indicate that *E. fergusonii* is more closely related to *E. coli* (0.8% divergent at nonsynonymous sites) than to either *E. hermannii* or *E. vulneris* (5.0% and 9.8% divergent at nonsynonymous sites, respectively), the IS*1* data indicate just the opposite (3.9% *vs.* 3.0% and 2.1% divergence at nonsynonymous sites for the IS*1bc* ORF. Comparisons with IS*1*R and IS*1*F yield similar results). Although these comparisons are potentially confounded by recombination events among these elements (see below), it is probable that IS*1* was introduced into these genomes by horizontal transfer since their separation. In addition, the divergence of IS*30* in *E. hermannii* is substantially less that that of *ompA* or *gap*, which suggests more recent invasion of this species by IS*30*. Yet the absence of IS which closely resemble *E. coli* elements in other species indicates that the rate of horizontal transfer of IS among bacterial species must be considerably lower than the rate of intraspecific transfer.

Intragenic recombination: Both IS*1* and IS*3* provide evidence that intragenic recombination plays a role in the evolution of insertion sequences. Two nested intragenic recombinations involve IS*3* homologs resident in *E. coli* (ECOR 63) and *S. dysenteriae*, as well as a distantly related element (Figure 5). The taxonomic relationships implicit in sequences within the recombined region (209–315 bp) are significantly different from those implicit in the remainder of the element ($P = 0.000$, Table 5), supporting a transfer of genetic material between the Shigella IS*3* and a distantly related element. Moreover, this region, as well as flanking sequences (97–348 bp), were transferred a second time between the Shigella IS*3* and and element now present in *E. coli* (Figure 5). As another indication of recombination, IS*1* homologs in *E. vulneris* and *E. hermannii* are also chimeras of distantly related elements (Tables 4 and 5). At least three separate regions exhibiting statistically distinct evolutionary histories can be identified among IS*1* sequences present in the genomes of enteric bacteria. Therefore, gene conversion, as well as rapid turnover, may play an important role in maintaining sequence homogeneity among IS within bacterial species. Recombination within bacterial genes has also been reported in *E. coli* (DUBOSE, DYKHUIZEN and HARTL 1988), *Streptococcus pneumoniae* (DOWSON *et al.* 1989), and *Neisseria gonorrhoeae* (SPRATT 1989).

Length polymorphisms: IS copies exhibit variation in length. Six of 35 elements (17%) examined from *E. coli* show length differences (Table 6), all but one being unique, and all the elements appear to be nonfunctional. In one case, IS*3411* was inserted into an unrelated insertion sequence (IS*30* in ECOR 24). The insertion site reveals not only a 3-bp target duplication, but a deletion of 2 bp downstream of the site of insertion. In addition, the 4-bp insertion in the ECOR 18 copy of IS*3* (cag/CAGT/tt) may be a remnant of the target site duplication of an excised transposon. In contrast, in certain natural populations of *Drosophila melanogaster*, the KP element, which contains a deletion of over 1700 bp, accounts for over 50% of the *P* element homologs (BLACK *et al.* 1987). In *Drosophila teissieri*, at least 75% of the *mariner* elements are deleted for more than half of their length (MARUYAMA and HARTL 1991). The few, unique, defective transposons detected among *E. coli* isolates indicate that substantially different factors contribute to the evolution of transposon populations within prokaryotic and eukaryotic genomes.

Ribosomal frameshifting: Analysis of nucleotide sequence variation revealed two evolutionarily conserved ORFs in divergent isolates of IS*1* and IS*3* and a single conserved ORF in IS*30*. It is not surprising that IS*1* maintains two conserved ORFs. SEKINE and OHTSUBO (1989) have demonstrated that ribosomal frameshifting between IS*1b* and IS*1c* is required for the production of the IS*1* transposase (see also LÜTHI *et al.* 1990), and our data support this hypothesis. The single conserved ORF of IS*30* probably encodes that element's transposase. The finding of two conserved ORFs within IS*3* was unexpected. However, a comparison between the IS*1* and IS*3* ORFs reveals striking similarity between the genetic organization of these

TABLE 11

Organization of ORFs and potential frameshifting signals in bacterial insertion sequences

Element	5′ ORF[a]	3′ ORF	Signal[b]	Location[c]
IS*1*	56–331 (b)	250–750 (c)	AAAAAAC	307
IS*3*	57–365 (a)	362–1225 (c)	AAAAGG	326
IS*10*	4–168 (a)	108–1313 (b)	AAAAAT	84
IS*21*	102–1274 (a)	1271–2068 (d)	??	
IS*51*	1261–938 (d)	941–42 (f)	AAAAAAC	1258
IS*150*	48–569 (a)	566–1414 (c)	AAAAAG	561
IS*3411*	55–381 (a)	378–1097 (b)	AAAAAAT	373

[a] Coordinates of 5′ and 3′ open reading frames in putative frameshifting events. Letters indicate ORF notation of Galas and Chandler (1989).

[b] Potential frameshifting signal.

[c] Coordinates of frameshifting signal.

unrelated elements. Both show two conserved open reading frames, the shorter located at the 5′ end of the element and oriented in the −1 reading frame relative to the longer 3′ ORF. Both exhibit the polyadenine frameshifting signal as well as a potential hairpin structure to facilitate ribosome slippage. In addition to IS*1*, these motifs have been implicated as the source of ribosomal frameshifting in the bovine leukemia virus (YOSHINAKA *et al.* 1986), mouse mammary tumor virus (HIZI *et al.* 1987; JACKS *et al.* 1987), and human T cell leukemia virus (SHIMOTOHNO *et al.* 1985) genomes. The 3′ ORF in IS*3*, like that in IS*1*, lacks a conventional promoter region for transcription initiation and a Shine-Dalgarno sequence for translation initiation. These similarities suggest that ribosomal frameshifting may be also responsible for the production of the IS*3* transposase.

The frameshifting sequence motif may also be identified is other bacterial insertion sequences. The 19 IS described by GALAS and CHANDLER (1989) may be separated into three classes: (1) IS bearing one, long, uninterrupted open reading frame, such as IS*4*, IS*5*, IS*6*, IS*30*, IS*50*, IS*186*, IS*701* and IS*903*; (2) IS bearing numerous scattered ORFs, such as IS*66*, IS*136* and IS*600*; and (3) IS bearing two closely situated open reading frames, the smaller 5′ ORF being in the −1 reading frame relative to the 3′ ORF. These elements are listed in Table 11. Although the genetic organization of certain elements resembles that of IS*1*, clearly not every IS is regulated by ribosomal frameshifting. IS*21* exhibits no obvious frameshifting signal between its closely positioned ORFs, while the poly(A) signal in IS*10* would lead to translation termination prior to reaching the second ORF. These cases aside, IS*3*, IS*150* and IS*3411* all exhibit the sets of sequence motifs responsible for frameshifting in IS*1*. Aside from IS*3*, however, it is not clear which ORFs are evolutionarily conserved, and there is no empirical evidence for frameshifting. Nevertheless, the repetition of this motif among bacterial IS suggests that ribosomal frameshifting may be a common mechanism for the regulation of IS transposition.

In summary, DNA sequence variation among bacterial IS within *E. coli* and related species indicates that (1) insertion sequences are rapidly mobilized among strains within a species; (2) horizontal transfer of IS occurs, but is less frequent than intraspecific transfer; (3) few defective transposons are present in the bacterial genome, relative to nonautonomous elements within eukaryotic genomes; (4) recombination occurs between IS at significant rates; and (5) functionally important ORFs can be recognized by their evolutionary conservation.

We thank S. SAWYER for aid with the IS*3411* analysis, M. STURMOSKI for technical assistance, and D. DYKHUIZEN and an anonymous reviewer for helpful comments on the manuscript. This work was supported by grants GM 40322 (D.L.H.) and GM 40995 (H.O.) from the National Institutes of Health. The DNA sequences reported in this paper have been submitted to GenBank and have been assigned accession numbers Z11603 through Z11609.

LITERATURE CITED

BERG, D. E., 1977 Insertion and excision of the transposable kanamycin resistance determinant Tn*5*, pp. 205–212 in *DNA Insertion Elements, Plasmids, and Episomes*, edited by A. BUKHARI, J. A. SHAPIRO and S. ADHYA. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

BLACK, D. M., M. S. JACKSON, M. G. KIDWELL and A. DOVER, 1987 KP elements repress P-induced hybrid dysgenesis in *Drosophila melanogaster* using a novel and general method. Cell **25:** 693–704.

CAUGANT, D. A., B. R. LEVIN and R. K. SELANDER, 1981 Genetic diversity and temporal variation in the *E. coli* population of a human host. Genetics **98:** 467–490.

CHANDLER, M., M. CLERGET and L. CARO, 1980 IS*1*-promoted events associated with drug resistance plasmids. Cold Spring Harbor Symp. Quant. Biol. **45:** 157–165.

CHOW, L. T., and T. R. BROKER, 1981 Adjacent insertion sequences IS*2* and IS*5* in bacteriophage Mu mutants and IS*5* in lambda d*arg* bacteriophage. J. Bacteriol. **133:** 1427–1436.

DALRYMPLE, B., P. CASPERS and W. ARBER, 1984 Nucleotide sequence of the prokaryotic mobile genetic element IS*30*. EMBO J. **3:** 2145–2149.

DEVEREUX, J., P. HAEBERLI and O. SMITHIES, 1984 A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12:** 387–395.

DOWER, W. J., J. F. MILLER and C. RAGSDALE, 1988 High efficiency transformation of *E. coli* by high voltage electroporation. Nucleic Acids Res. **16:** 6127–6145.

DOWSON, C. G., A. HUTCHINSON, J. A. BRANNIGAN, R. C. GEORGE, D. HANSMAN, J. LINARES, A. TOMASZ, J. MAYNARD SMITH and B. G. SPRATT, 1989 Horizontal transfer of penicillin-binding genes in penicillin-resistant clinical isolates of *Streptococcus pneumoniae*. Proc. Natl. Acad. Sci. USA **86:** 8842–8846.

DUBOSE, R. F., D. E. DYKHUIZEN and D. L. HARTL, 1988 Genetic exchange among natural isolates of bacteria: recombination within the *pho*A locus of *Escherichia coli*. Proc. Natl. Acad. Sci. USA **85:** 7036–7040.

DUBOSE, R. F., and D. L. HARTL, 1990 Rapid purification of PCR products for DNA sequencing using Sepharose CL-6B spin columns. Biotechniques **8:** 271–273.

DYKHUIZEN, D. E., and L. GREEN, 1986 DNA sequence variation, DNA phylogeny, and recombination in *E. coli*. Genetics **113** (Suppl.): s71.

EDWARDS, P. R., and W. H. EWING, 1962 *Identification of Enterobacteriaceae*. Burgess, Minneapolis.

FEINBERG, A. P., and B. VOGELSTEIN, 1983 A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. **132:** 6–13.

GALAS, D. J., and M. CHANDLER, 1989 Bacterial insertion sequences, pp. 109–162 in *Mobile DNA*, edited by D. E. BERG and M. M. HOWE. American Society for Microbiology, Washington, D.C.

GLANDSDORFF, N., D. CHARLIER and M. ZAFARULLAH, 1980 Activation of gene expression by IS*2* and IS*3*. Cold Spring Harbor Symp. Quant. Biol. **45:** 153–156.

HALL, B. G., L. L. PARKER, P. W. BETTS, R. F. DUBOSE, S. A. SAWYER and D. L. HARTL, 1989 IS*103*: a new insertion element in *Escherichia coli*. Characterization and distribution in natural populations. Genetics **121:** 423–431.

HARTL, D. L., and S. A. SAWYER, 1988a Multiple correlations among insertion sequences in the genome of natural isolates of *Escherichia coli*, pp. 91–106 in *Transposition*, edited by A. J. KINGSMAN, S. M. KINGSMAN and K. F. CHATER. Cambridge University Press, Cambridge.

HARTL, D. L., and S. A. SAWYER, 1988b Why do unrelated insertion sequences occur together in the genome of *Escherichia coli*? Genetics **118:** 537–541.

HELMS, C., M. Y. GRAHAM, J. E. DUTCHIK and M. V. OLSON, 1985 A new method for purifying lambda DNA from phage lysates. DNA **4:** 39–49.

HIZI, A., L. E. HENDERSON, T. D. COPELAND, R. C. SOWDER, C. V. HIXSON and S. OROSZLAN, 1987 Characterization of mouse mammary tumor virus *gag-pro* gene products and the ribosomal frameshift site by protein sequencing. Proc. Natl. Acad. Sci. USA **84:** 7041–7045.

JACKS, T. K. TOWNSLEY, H. E. VARMUS and J. MAJORS, 1987 Two efficient ribosomal frameshifting events are required for synthesis of mouse mammary tumor virus *gag*-related polyproteins. Proc. Natl. Acad. Sci. USA **84:** 4298–4302.

KLECKNER, N., R. K. CHAN, B. K. TYE and D. BOTSTEIN, 1975 Mutagenesis by insertion of a drug resistance element carrying an inverted repetition. J. Mol. Biol. **97:** 561–575.

LAM, S., and J. R. ROTH, 1983a Genetic mapping of IS*200* copies in *Salmonella typhimurium* LT2. Genetics **105:** 801–811.

LAM, S., and J. R. ROTH, 1983b IS*200*: a *Salmonella*-specific insertion sequence. Cell **34:** 951–960.

LAWRENCE, J. G., and D. L. HARTL, 1992 Inference of horizontal genetic transfer from molecular data: an approach using the bootstrap. Genetics (in press).

LAWRENCE, J. G., D. L. HARTL and H. OCHMAN, 1991a Sequencing products of the polymerase chain reaction. Methods Enzymol. (in press).

LAWRENCE, J. G., D. L. HARTL, and H. OCHMAN, 1991b Molecular considerations in the evolution of bacterial genes. J. Mol. Evol. **33:** 241–250

LAWRENCE, J. G., H. OCHMAN and D. L HARTL, 1991 Molecular and evolutionary relationships among enteric bacteria. J. Gen. Microbiol. **137:** 1911–1921

LAWRENCE, J. G., D. E. DYKHUIZEN, R. F. DUBOSE and D. L. HARTL, 1989 Phylogenetic analysis of *Escherichia coli* based on insertion sequence fingerprinting. Mol. Biol. Evol. **6:** 1–14.

LIAB, M., 1980 IS*5* increases recombination in adjacent regions as shown for the repressor gene of coliphage lambda. Gene **12:** 277–280.

LÜTHI, K.. M. MOSER, J. RYSER and H. WEBER, 1990 Evidence for a role of translational frameshifting in the expression of transposition activity of the bacterial insertion element IS*1*. Gene **88:** 15–20.

MARUYAMA, K., and D. L. HARTL, 1991 Evolution of the transposable element *mariner* in *Drosophila* species. Genetics **128:** 319–329.

MILKMAN, R., and I. P CRAWFORD, 1981 Clustered third base substitutions among wild strains of *Escherichia coli*. Science **221:** 378–379.

NELSON, K., T. S. WHITTAM and R. K. SELANDER, 1991 Nucleotide polymorphism and evolution in the glyceraldehyde-3-phosphate gene (*gapA*) in natural populations of *Salmonella* and *Escherichia coli*. Proc. Natl. Acad. Sci. USA **88:** 6667–6671.

OCHMAN, H., and R. K. SELANDER, 1984a Evidence for clonal population structure in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **81:** 198–201.

OCHMAN, H., and R. K. SELANDER, 1984b Standard reference collection of *Escherichia coli* from natural populations. J. Bacteriol. **157:** 517–524.

OHTSUBO, H., and E. OHTSUBO, 1978 Nucleotide sequence of an insertion element, IS*1*. Proc. Natl. Acad. Sci. USA **75:** 615–619.

OHTSUBO, E., H. OHTSUBO, W. DOROSZKIEWICZ, K. NYMAN, D. ALLEN and D. DAVIDSON, 1984 An evolutionary analysis of iso-IS*1* elements from *Escherichia coli* and *Shigella* strains. J. Gen. Appl. Microbiol. **30:** 359–376.

PERLER, F., A. EFSTRATIADIS, P. LOMEDICO, W. GILBERT, R. KOLODNER and J. DODGSON, 1980 The evolution of genes: the chicken preproinsulin gene. Cell **20:** 555–566.

SAEDLER, H., H. J. REIF, S. HU and N. DAVIDSON, 1974 IS*2*: a genetic element for turn-off and turn-on of gene activity in *E. coli*. Mol. Gen. Genet. **132:** 265–289.

SAEDLER, H., G. CORNELIS, J. CULLUM, B. SCHUMACHER and H. SOMMER, 1980 IS*1* mediated DNA rearrangements. Cold Spring Harbor Symp. Quant. Biol. **45:** 93–98.

SAIKI, R. K., S. SCHARF, F. FALOONA, K. B. MULLIS, G. T., HORN, H. A. ERLICH and N. A. ARNHEIM, 1985 Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science **230:** 1350–1354.

SAIKI, R. K., D. H. GELFAND, S. STOFFEL, S. J. SCHARF, R. HIGUCHI, G. T. HORN, K. B. MULLIS and H. A. ERLICH, 1988 Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science **239:** 487–491.

SAWYER, S., D. DYKHUIZEN, R. DUBOSE, L. GREEN, T. MUTANGADURA-MHLANGA, D. WOLCZYK and D. HARTL, 1987 Distribution and abundance of insertion sequences among natural isolates of *Escherichia coli*. Genetics **115:** 51–63.

SEKINE, Y., and E. OHTSUBO, 1989 Frameshifting is required for production of the transposase encoded by insertion sequence *1*. Proc. Natl. Acad. Sci. USA **86:** 4609–4613.

SELANDER, R. K., and B. R. LEVIN, 1980 Genetic diversity and structure in *Escherichia coli* populations. Science **210:** 545–547.

SELANDER, R. K., D. A. CAUGANT and T. S. WHITTAM, 1987 Genetic structure and variation in natural populations of *Escherichia coli*, pp. 1625–1648 in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, edited by F. C. NEIDHARDT. American Society for Microbiology, Washington, D.C.

SHIMOTOHNO, K., Y. TAKAHASHI, T. GOJOBORI, D. W. GOLDE, I. S. Y. CHEN, M. MIWA and T. SUGIMURA, 1985 Complete nucleotide sequence of an infectious clone of human T-cell leukemia virus type II: an open reading frame for the protease gene. Proc. Natl. Acad. Sci. USA **82:** 3101–3105.

SPRATT, B. G., 1989 Hybrid penicillin-binding proteins in penicillin-resistant strains of *Neisseria gonorrhoeae*. Nature **332:** 173–176.

STOLTZFUS, A., J. F. LESLIE and R. MILKMAN, 1988 Molecular evolution of the *Escherichia coli* chromosome. I. Analysis of structure and natural variation in a previously uncharacterized region between *trp* and *tonB*. Genetics **120:** 345–358.

TIMMERMAN, K., and C. P. D. TU, 1985 Complete sequence of IS*3*. Nuc Acids Res **13:** 2127–2139.

WHITTAM, T. S., H. OCHMAN and R. K. SELANDER, 1984 Geographical components of linkage disequilibrium in natural populations of *Escherichia coli*. Mol. Biol. Evol. **1:** 67–83.

UMEDA, M., and E. OHTSUBO, 1991 Four types of IS*1* with differences in nucleotide sequence reside in the *Escherichia coli* K-12 chromosome. Gene **98:** 1–5.

YOSHINAKA, Y., I. KATOH, T. D. COPELAND, G. W. SMYTHERS and S. OROSZLAN, 1986 Bovine leukemia virus protease: purification, chemical analysis, and in-vitro processing of *gag* precursor polyproteins. J. Virol. **57:** 826–832.

Communicating editor: W.-H. LI