Size and Sequence Polymorphism in the Isocitrate Dehydrogenase Kinase/Phosphatase Gene (*aceK*) and Flanking Regions in *Salmonella enterica* and *Escherichia coli*

Kimberlyn Nelson, Fu-Sheng Wang,[1] E. Fidelma Boyd[2] and Robert K. Selander

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

Manuscript received February 4, 1996
Accepted for publication August 25, 1997

ABSTRACT

The sequence of *aceK*, which codes for the regulatory catalytic enzyme isocitrate dehydrogenase kinase/phosphatase (IDH K/P), and sequences of the 5′ flanking region and part or all of the 3′ flanking region were determined for 32 strains of *Salmonella enterica* and *Escherichia coli*. In *E. coli*, the *aceK* gene was 1734 bp long in 13 strains, but in three strains it was 12 bp shorter and the stop codon was TAA rather than TGA. Strains with the shorter *aceK* lacked an open reading frame (*f728*) downstream between *aceK* and *iclR* that was present, in variable length, in the other strains. Among the 72 ECOR strains, the truncated *aceK* gene was present in all isolates of the B2 group and half of those of the D group. Other variant conditions included the presence of IS*1* elements in two strains and large deletions in two strains. The *aceK-aceA* intergenic region varied in length from 48 to 280 bp in *E. coli*, depending largely on the number of repetitive extragenic palindromic (REP) sequences present. Among the ECOR strains, the number of REP elements showed a high degree of phylogenetic association, and sequencing of the region in the ECOR strains permitted partial reconstruction of its evolutionary history. In *S. enterica*, the normal length of *aceK* was 1752 bp, but three other length variants, ranging from 1746 to 1785 bp, were represented in five of the 16 strains examined. The flanking intergenic regions showed relatively minor variation in length and sequence. The occurrence of several nonrandom patterns of distribution of polymorphic synonymous nucleotide sites indicated that intragenic recombination of horizontally exchanged DNA has contributed to the generation of allelic diversity at the *aceK* locus in both species.

POPULATION genetic studies of sequence variation in several genes among strains of *Salmonella enterica* and *Escherichia coli* recovered from natural populations have demonstrated that within each species there is a large interlocus variance in the effective (realized) rate of recombination of horizontally exchanged DNA (BISERCIC *et al.* 1991; DYKHUIZEN and GREEN 1991; NELSON *et al.* 1991; NELSON and SELANDER 1992, 1994b; MILKMAN and BRIDGES 1993; BOYD *et al.* 1994; LI *et al.* 1994, 1995; THAMPAPAPILLAI *et al.* 1994; review in SELANDER *et al.* 1994). Thus, for example, intragenic and assortative (entire gene) recombination of the *fliC* gene encoding the highly variable antigenic flagellin protein of phase 1 flagella in *S. enterica* occurs frequently enough to constitute a major source of allelic variation and serovar diversity (SMITH *et al.* 1990; LI *et al.* 1994), whereas recombination events involving the genes that encode the transfer protein proline permease (*putP*) (NELSON and SELANDER 1992), the metabolic enzyme malate dehydrogenase (*mdh*) (BOYD *et al.* 1994), and the invasion proteins of the Inv/Spa complex (LI *et al.* 1995) are infrequent. These findings, together with evidence from sequence studies of genes in other species of bacteria, have suggested that, in general, the highest rates of recombination affect loci for which the products are subject to diversifying selection in adaptation to host defense systems or other variable aspects of the environment (NELSON and SELANDER 1992; REEVES 1992; ACHTMAN 1994; MOXON *et al.* 1994; SELANDER *et al.* 1994).

As a basis for testing this and other hypotheses relating to the genetic structure and evolutionary dynamics of bacterial populations, we have studied sequence variation in the isocitrate dehydrogenase kinase/phosphatase gene (*aceK*) and its flanking regions in multiple strains of *S. enterica* and *E. coli*. The *aceK* gene encodes a bifunctional regulatory enzyme (IDH K/P) that catalyzes the phosphorylation and dephosphorylation of isocitrate dehydrogenase (IDH) and thereby controls the flux of isocitrate through the tricarboxylic acid cycle and the glyoxylate bypass. It is part of the acetate (*ace* or *aceBAK*) operon (Figure 1), which also includes the *aceB* and *aceA* genes encoding, respectively, the glyoxylate bypass enzymes malate synthase and isocitrate lyase (LAPORTE 1993). Expression of the *ace* operon is controlled by both culture conditions and the product of

Corresponding author: Robert K. Selander, Institute of Molecular Evolutionary Genetics, 516 Mueller Laboratory, University Park, PA 16802-5301.

[1] *Present address:* Department of Microbiology and Immunology, University of Western Ontario, Ontario, Canada N6A 5C1.

[2] *Present address:* Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138.

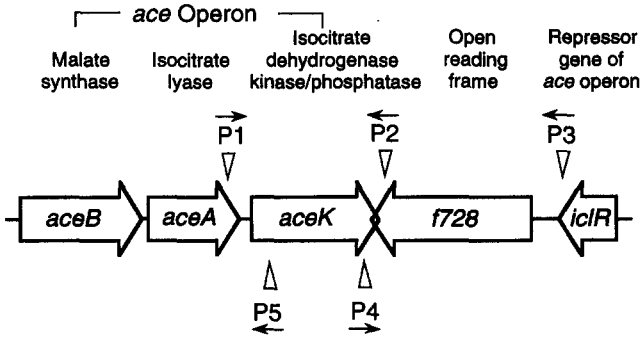

FIGURE 1.—Organization of genes of the *ace* operon (*aceB*, *aceA*, and *aceK*) and the repressor gene (*iclR*) of the operon in *S. enterica* and *E. coli*. In some *E. coli* strains, open reading frame *f728* is present, in whole or in part, in the *aceK-iclR* intergenic region. The positions of five primers (P1–P5) used in PCR amplification of *aceK* and its flanking intergenic regions are shown.

iclR, the repressor gene of the operon, which is located downstream from *aceK*.

Our analysis has demonstrated that the *aceK* gene is polymorphic in length as well as in sequence within and between species and that in *E. coli* the flanking intergenic regions are highly variable in both length and sequence, largely as a result of variation in the number of repetitive extragenic palindromic (REP) elements in the *aceA-aceK* intergenic region and the presence or absence of an open reading frame, *IS1* elements, and other DNA segments in the region between *aceK* and *iclR*. For each species, there is evidence of several recombination events involving the horizontal exchange of gene segments.

MATERIALS AND METHODS

***S. enterica* strains:** Strains of *S. enterica* were the same as those used in previous studies of other genes (NELSON *et al.* 1991; NELSON and SELANDER 1992, 1994b; BOYD *et al.* 1994, 1997; LI *et al.* 1995; WANG *et al.* 1997); this sample of 16 strains, which includes two representatives of each of the eight subspecies, I, II, IIIa, IIIb, IV, V, VI, and VII, constitutes the *Salmonella* Reference Collection C (SARC) (BOYD *et al.* 1996). Ten additional strains of subspecies V were included in the analysis of the *aceA-aceK* intergenic region.

***E. coli* strains:** The sample of *E. coli* included 12 isolates that were used in previous studies (NELSON *et al.* 1991; NELSON and SELANDER 1992, 1994b; BOYD *et al.* 1994; WANG *et al.* 1997); these strains were obtained from the *E. coli* Reference Collection (ECOR) (OCHMAN and SELANDER 1984) and a research collection maintained by T. S. WHITTAM. Three additional ECOR strains and a K-12 strain (CH734) were also studied. The isolates were selected to represent the five major phylogenetic lineages of *E. coli* (A, B1, B2, D, and E) that have been identified by multilocus enzyme electrophoresis (MLEE) (HERZER *et al.* 1990). Strains that are members of ECOR are identified by the designation "EC."

Sixty-one additional strains of ECOR were examined for sequence variation in the *aceA-aceK* and *aceK-iclR* intergenic regions.

Nucleotide sequencing: The *aceK* region in each of 32 strains of *S. enterica* and *E. coli* was directly sequenced from the product of the polymerase chain reaction (PCR) (SAIKI *et al.* 1988; NELSON and SELANDER 1994a). Three primers (P1, P2, and P3), the locations of which are shown in Figure 1, were designed from published sequences for *E. coli* K-12 (KLUMPP *et al.* 1988; RIEUL *et al.* 1988). Primers P1 and P3 were used for PCR amplification from 16 *S. enterica* strains and three *E. coli* strains that lack the segment containing P2, and primer pairs P1 and P2 were used for the other *E. coli* strains. The sequences were determined in both orientations with additional internal primers, and the overlapping sequences were assembled and edited with the SEQMAN and SEQMANED programs (DNASTAR, Madison, WI). The edited sequences were aligned manually with the Eyeball Sequence Editor (ESEE) (CABOT and BECKENBACH 1989). The 32 sequences have been deposited in the GenBank data base (accession nos. U43000–U43315 and U43344–U43359).

Sequences of the *aceA-aceK* intergenic region were obtained for 71 additional strains of *S. enterica* and *E. coli* from segments amplified with primers P1 and P5 (Figure 1). Most of these segments were sequenced in one orientation only.

To study variation in the *aceK-iclR* intergenic region in *E. coli*, a segment was PCR amplified with primers P3 and P4.

RESULTS

Size and sequence variation in *aceK*: The sequences of *aceK* from all 32 strains of *S. enterica* and *E. coli* were directly alignable, without gaps, to nucleotide 1719 (codon 573), but the remaining 3′ part of the gene was variable in length within and between species (Figures 2 and 3). Among strains of *S. enterica*, the normal length was 1752 bp (11 strains), but there were three other length variants: 1746 bp (the two strains of subspecies VI), 1782 bp (one strain of subspecies I), and 1785 bp (the two strains of subspecies IIIa). The three strains with unusually long genes have TGA as the stop codon, whereas all other strains have TAA. The *aceK* gene of *E. coli* is shorter than that of *S. enterica;* in 13 strains, the gene was 1734 bp long, but in the remaining three strains it was 12 bp shorter (1722 bp) and the stop codon was TAA rather than TGA.

Among the 16 strains of *S. enterica*, there were 338 polymorphic nucleotide sites in the 1719-bp aligned segment of *aceK* (Figure 2), and the sequences of pairs of strains differed on average at 97 sites (5.7%). There were 47 polymorphic amino acid positions, with an average pairwise difference of 2.1%. Most of the polymorphism occurred between strains of different subspecies; between pairs of strains of the same subspecies, the average nucleotide difference was only 0.4%.

Among the 16 strains of *E. coli*, 164 polymorphic nucleotide sites were present in the 1719-bp segment of the gene (Figure 3). The sequences of pairs of strains differed on average at 50 sites (2.9%). There were 15 amino acid substitutions, and pairs of strains differed on average at 4.5 (0.8%) of amino acid positions.

Pairwise comparisons of the 1719-bp segment between strains of *S. enterica* and *E. coli* yielded an average difference of 309 nucleotide sites (17.9%) and 43 amino acid positions (7.5%).

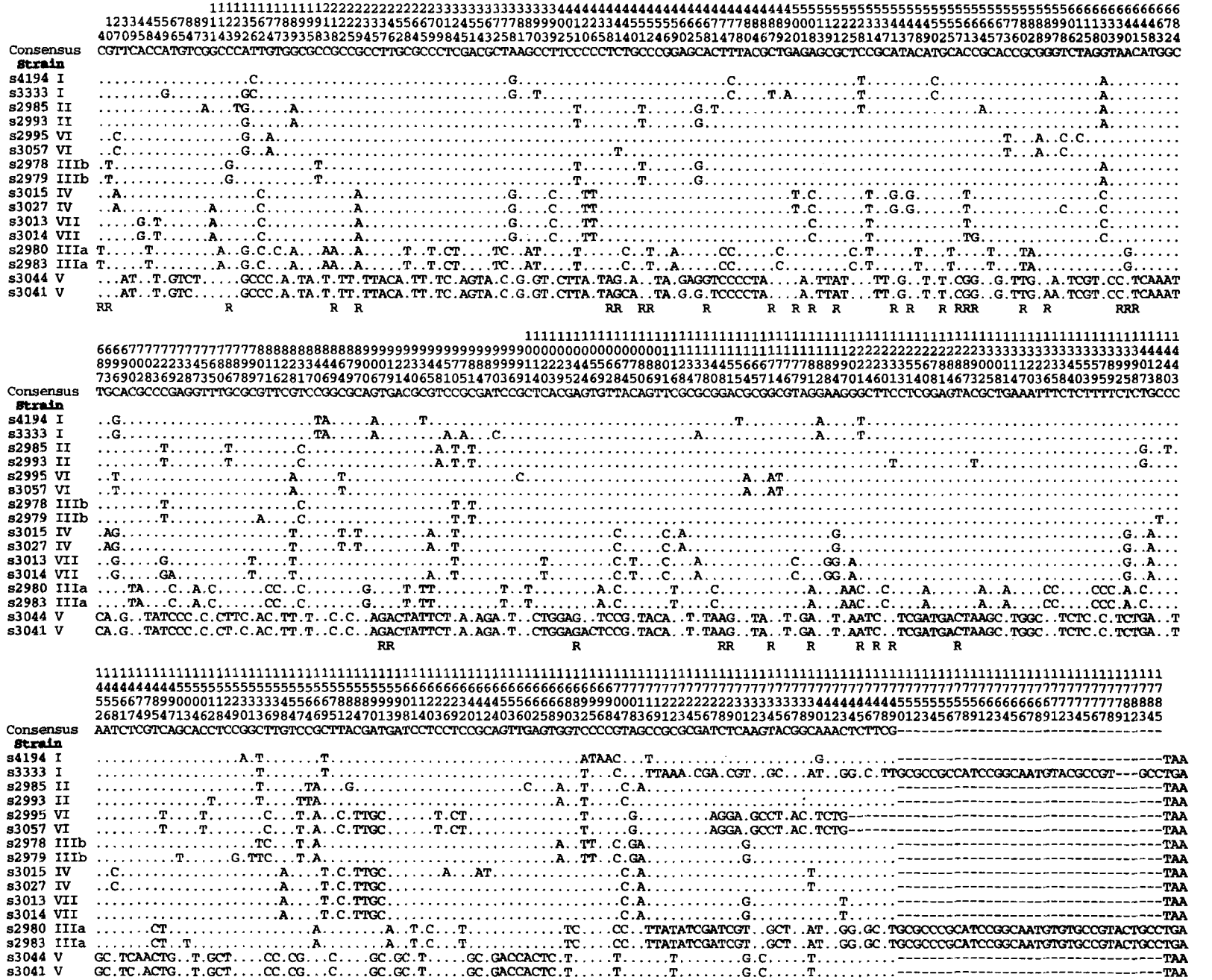

FIGURE 2.—Nucleotide sequence variation in the *aceK* gene among 16 strains of *S. enterica* representing the eight recognized subspecies. Only the polymorphic sites are shown. Dots indicate identity to the consensus sequence. Dashes indicate the absence of nucleotides. R indicates the occurrence of a substitution that results in an amino acid replacement.

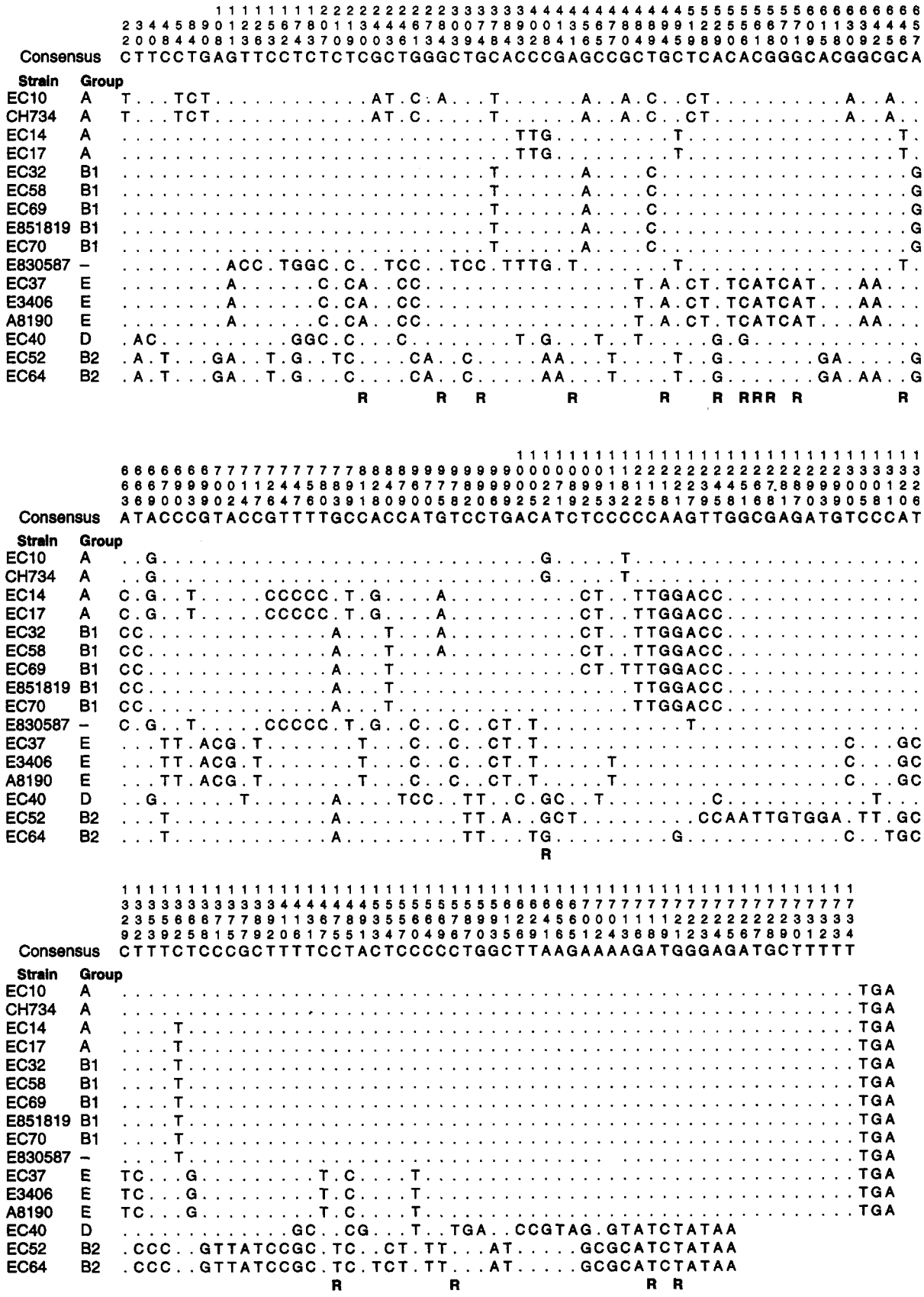

FIGURE 3.—Nucleotide sequence variation in the *aceK* gene among 16 strains of *E. coli* representing the major groups of ECOR. Only the polymorphic sites are shown. Dots indicate identity to the consensus sequence. *R* indicates the occurrence of a substitution that results in an amino acid replacement.

TABLE 1

Phylogenetic partitions of *aceK* sequences identified by the Stephens test

Partition	Value of indicated statistic[a]				
	s	d_0	g_0	$P(d \leq d_0)$	Pg_0
S. enterica (16 strains)					
1. V/others	145	1638	99	0.006	0.013
2. s2979 (IIIb)/others	4	103	93	0.0008	0.22
3. IV, VI, VII/others[b]	5	10	3	<0.0001	0.66
E. coli (16 strains)					
1. EC52/others	11	216	159	<0.0001	<0.0001
2. E830587/others	7	308	141	0.0002	0.24
3. EC14, EC17, E830587/others	10	426	291	<0.0001	0.009
4. EC10, CH734/others	8	623	240	0.004	0.32
5. EC37, E3406, A8190/others	17	1237	518	0.028	0.004
6. EC52, EC64/others	24	1654	735	0.680	<0.0001
7. EC37, E3406, A8190, EC52, EC64/others	4	662	650	0.16	0.0008

[a] s, number of polymorphic sites; d_0 observed distance between the two terminal unique polymorphic sites; g_0, length of the longest segment of consecutive nonpolymorphic sites; $P(d \leq d_0)$, probability that the observed distance between terminal unique polymorphic sites exceeds the expected distance; and Pg_0, probability that at least one of $s - 1$ random, independently observed segments is as long or longer than g_0.

[b] Partition identified after removal of strains of subspecies V (see text).

Distribution of sequence polymorphisms in *aceK*: With the guidance of a statistical test developed by STEPHENS (1985), a number of nonrandom patterns of distribution of polymorphic sites in *aceK* were identified, most of which may reflect intragenic recombination events (Table 1). For *S. enterica*, application of the test to all 16 strains identified only two significant phylogenetic partitions supported by four or more sites (Table 1). Partition 1 separated the strains of subspecies V from those of all other subspecies, and partition 2 identified a 103-bp segment with four unique sites in strain s2979 of subspecies IIIb. Because the polymorphic sites in the subspecies V sequences represent 45% of all polymorphic sites in the *S. enterica* sample, the Stephens test was applied in a separate analysis to 14 strains of *S. enterica* (subspecies V excluded) to search for partitions that might have been obscured. One additional partition was uncovered: a 10-bp segment supported by five sites at the 3′ end of the gene that was shared by subspecies IV, VI, and VII.

In application to the distribution of polymorphic sites among the 16 *E. coli* sequences, the Stephens test distinguished seven significant partitions supported by four or more sites (Table 1). One of these, partition 1, separated EC52 from all other strains and was supported by a total of 11 sites; this partition identifies a 216-bp segment that includes a 57-bp sequence (bp 1248–1305) in which EC52 has 10 unique silent substitutions.

The *aceK* gene of strain E830587 is a complex mosaic of segments apparently derived by recombination from several sources (Figure 3). In the 5′ third of the gene, partition 2 identifies a distinctive 308-bp segment (bp 123–431) containing seven unique sites. Just downstream from this segment (and partially overlapping with it), partition 3 (bp 402–828) is shared by group A strains EC14 and EC17; and further along, 5 sites between bp 960 and 1005 are shared by strains of group E. In the 3′ third of the gene, E830587 is similar in sequence to the strains of groups A and B1.

Partition 4 reflects the fact that EC10 and CH734 differ from EC14 and EC17 in the 5′ two thirds of the gene but are identical in sequence in the 3′ third. This is true even when the 5′ sites in EC14 and EC17 that are shared with E830587 (partition 3), presumably as a result of recombination, are ignored, for EC10 and CH734 share nine unique sites in this segment (Figure 3). Near the center of the gene, between bp 1212 and 1245, there is a distinctive segment of 33 bp containing seven polymorphic sites (three of which are unique) that is present in all five group B1 strains and group A strains EC14 and EC17 but not in group A strains EC10 and CH734 (Figure 3). The Stephens test identified this segment as a significant partition with three unique sites ($P[d \leq d_0] \leq 0.0001$).

Partition 5 separated the strains of group E from those of all other groups. Inasmuch as this partition includes almost all of the gene, the simplest interpretation is that its distinctive character reflects phylogenetic divergence by mutation rather than the acquisition of horizontally transferred DNA.

Partitions 6 and 7 identify segments of moderate length in which there is a paucity of unique polymorphic sites in several strains.

Evolutionary tree for *aceK*: To examine the relationships among the *aceK* alleles of the 32 strains, an evolutionary tree was constructed by the neighbor-joining method (SAITOU and NEI 1987) from a matrix of pairwise genetic distances based on synonymous

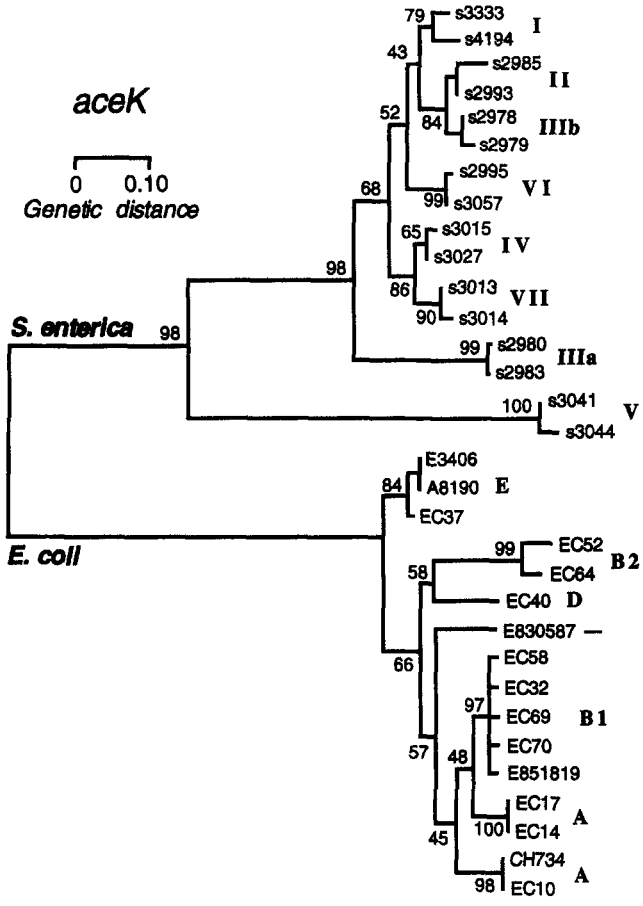

FIGURE 4.—Neighbor-joining tree for the *aceK* gene in 32 strains of *S. enterica* and *E. coli*, based on variation at synonymous nucleotide sites. Subspecies of *S. enterica* are indicated by roman numerals, and ECOR group assignments for the *E. coli* strains are given. Bootstrap values based on 1000 computer-generated trees are indicated at the nodes.

nucleotide sites, with corrections for multiple substitutions by the Jukes-Cantor method (KUMAR *et al.* 1993) (Figure 4).

Within *S. enterica*, the *aceK* sequences of strains of the same subspecies are invariably paired and closely similar. Subspecies V is the most divergent, followed by subspecies IIIa. Subspecies IV and VII cluster together, and there is an association of the sequences of subspecies I, II, IIIb, and VI.

Among the *E. coli* sequences, those of the three group E strains (EC37, E3406, and A8190) are the most divergent. Sequences of the two group B2 strains are much alike, as are those of the five B1 strains. However, the four group A strains are separated into two clusters; EC10 and CH734 differ from EC14 and EC17 at 2.5% of nucleotide sites.

Length and sequence variation in the *aceA-aceK* intergenic region: The region between *aceA* and *aceK* in K-12 *E. coli* laboratory strains is reported to contain three REP sequences (CORTAY *et al.* 1988; KLUMPP *et al.* 1988; MATSUOKA and MCFADDEN 1988). Among the 77 *E. coli* strains from which this region was sequenced (including the 72 strains of the ECOR collection and five additional strains in the core sample), the number of REP sequences ranged from zero to five (although no strain had 2), and a total of 12 length and sequence patterns was observed (Figures 5 and 6).

The most common pattern (labeled 5a in Figure 5) contained five REPs; this pattern was seen in 40 strains distributed in four of the five ECOR groups (Figure 7). Most strains of groups A and B1 had the 5a pattern, which also occurred in strain EC43 of group E, presumably as a result of horizontal transfer. Patterns 5b and 5c are each single-base substitution variants of 5a; and 5d differs from 5a at four nucleotide sites. The occurrence of pattern 5b in two divergent ECOR strains (EC67 of B1 and EC59 of B2) can be attributed to transfer of the intergenic region from a B1 to a B2 strain because EC59 is the only B2 strain that has five REP sequences and the 5b pattern cannot easily be derived by recombination and nucleotide substitution from any of the other sequences observed in the B2 strains.

At least two of the three patterns involving three REP sequences (patterns 3a and 3b) can be derived by intramolecular recombination between the directly repeated REP sequences of pattern 5a. The fact that pattern 3a has been observed only in K-12 strains (including CH734) suggests that it was produced by a recombination event occurring in the laboratory. Pattern 3b was found in four B1 strains, among which the extent of overall phylogenetic divergence is sufficiently large to suggest that the pattern has been generated more than once (Figure 7).

Two additional patterns with three REP sequences were present in strains of ECOR groups D and E. The origin of pattern 3c, which occurs in half the strains of these two groups, is difficult to determine: it can be derived from pattern 5a by one recombination event, three nucleotide substitutions, and a single-base insertion, or from pattern 3b by 5 substitutions and a single-base insertion. In view of the phylogenetic distribution of pattern 3c, it seems likely that it evolved independently from an ancestral pattern 5 sequence. A single-base mutation in 3c generates pattern 3d, which was detected in three strains of group E (Figure 7); and intramolecular recombination between the directly repeated REPs of 3d leaves a single REP, which is pattern 1 found in strain E830587 (unassigned to an ECOR group).

Finally, there are two patterns (4a and 4b) that contain three complete REPs, most of a fourth sequence, and a small segment of a fifth sequence (Figure 5). Pattern 4a was found only in strain EC44 of group D, and pattern 4b occurred in nine strains of B2. These two patterns are alike in having the region from the 3′ end of the fourth REP sequence through the 3′ end of the fifth REP sequence missing, but they differ by 18 nucleotide substitutions, one base insertion, and a 7-bp

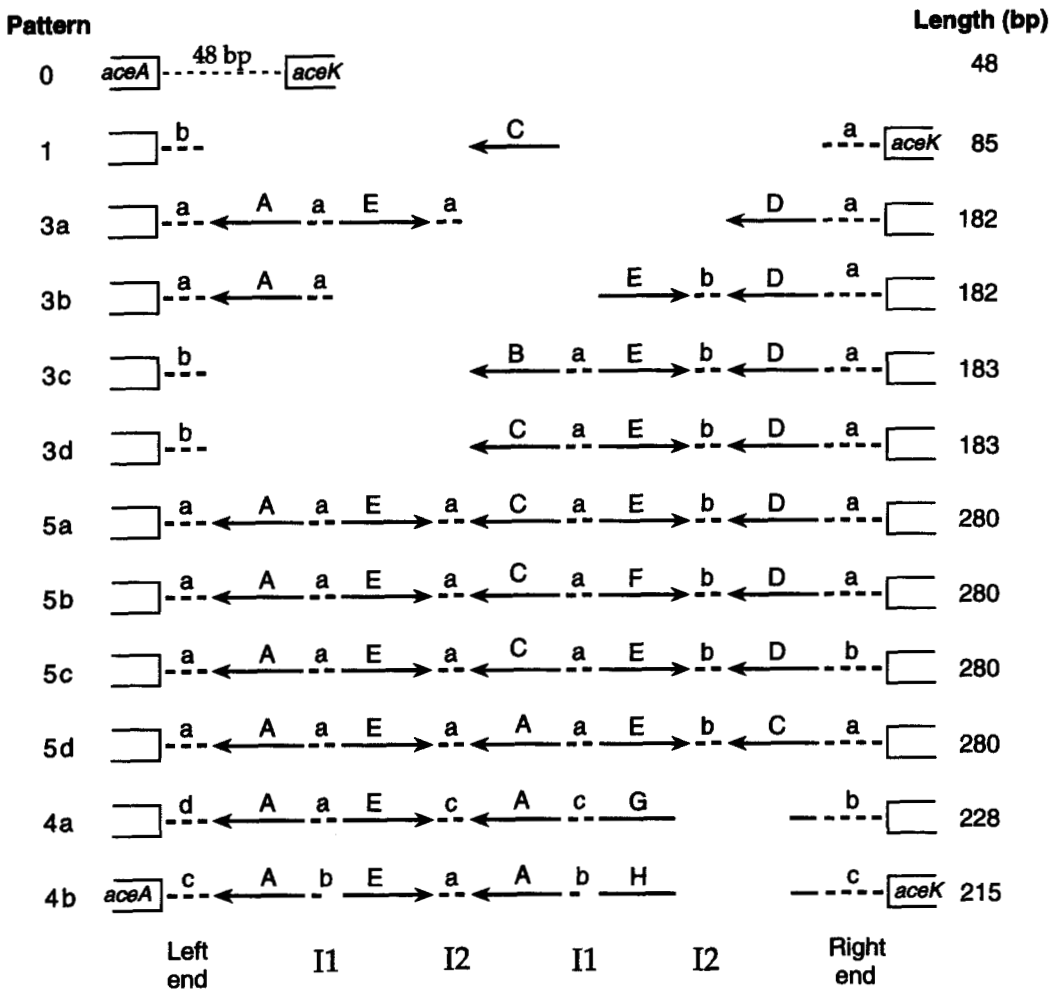

FIGURE 5.—Variation in size, structure, and sequence of the *aceA-aceK* intergenic region among strains of *E. coli*. Arrows represent REP sequences, variants of which are indicated by capital letters. The nucleotide sequences of the REPs and of the left end, right end, and internal elements (I1 and I2), which are indicated by dashes and labeled with lowercase letters, are shown in Figure 6.

deletion, which, significantly, occurs twice in pattern 4b, and almost certainly were produced by independent recombination events (Figures 5 and 6).

In contrast to *E. coli*, all *S. enterica* sequences had a single REP element in the *aceA-aceK* intergenic region. The length of the region varied from 93 bp in strains of subspecies IIIa to 106 bp in strain s3041 (subspecies V), and 49% of sites were polymorphic for nucleotide substitutions. The two strains of subspecies V differed significantly in the sequence of this region. As shown in Figure 8, strain s3044 was almost identical to strains of subspecies I throughout a segment that includes the intergenic region and 195 bp of the upstream *aceA* gene; this clearly reflects a recombination event involving a transfer from subspecies I to subspecies V. To further define this event, we sequenced this segment in 10 additional strains of subspecies V. The subspecies I sequence was present in six of the total of 12 strains, and the subspecies V sequence was observed in the remaining six strains. Mapping of this sequence variation onto the pattern of overall genomic relationships among the subspecies V strains, as indexed by MLEE (Figure 9), suggests that the transfer from subspecies I to V occurred twice, with a subsequent transfer to strain s3048.

Length and sequence variation in the *aceK-iclR* intergenic region: The 72 ECOR strains and the five additional strains of *E. coli* could be partitioned into seven classes on the basis of the size and sequence composition of the PCR product generated with primers 3 and 4 flanking the *aceK-iclR* region (Figure 1). Classes 1, 4, and 6 accounted for 94% of the strains (Figure 10). The class 1 region was present in all 15 strains of the ECOR B2 group and all six strains in one lineage of the D group (Figure 7). This variant was completely sequenced in three strains and was found to have an intergenic region of only 16 bp. The class 6 region, which has been completely sequenced in K-12 (GALINIER *et al.* 1991), is approximately 2400 bp in length and includes a 1836-bp open-reading frame, *f728*, and 601 bp of other sequence adjacent to *iclR*. This type of region was present in 27 strains of groups A, D, and E, as well as in strain E830587. Class 4, which was observed in 24 strains of ECOR groups A and B1, differs in size from class 6 by the absence of approximately 600 bp, and its sequence diverges markedly from that of class 6 about 390 bp from the end of *aceK*.

The remaining four classes were each observed in only one or two strains and presumably represent unique evolutionary events. Class 2, occurring in strains

REP sequence

```
REP-A  GCCGGATGCGGCGTGA--ACGCCTTATCCGGCCTAC
REP-B  ..............AT---T..................
REP-C  ..............A.---T..................
REP-D  ..............A.-----------A..........
REP-E  ...T......A..CTTGCG..T......AT.......
REP-F  ...T......A..CTTGCG..T...T..AT.......
REP-G  ...T......A..CTTGCG..T---------------
REP-H  ...T....T.A..CTTACG..T---------------
```

Left end

```
LE-a   GCAACA-ACAACCGTTGCTGACT
LE-b   .....CT....A...........
LE-c   .....CG.T..A...........
LE-d   ........-............T.
```

Right end

```
RE-a   AATTCTCTGCTCCTGATGAGGGCGCTAA
RE-b   ..............T..............
RE-c   ..............C..C.......AAC..
```

Internal 1

```
I1-a   AATCGGTGCACGAT
I1-b   ....A.C-------
I1-c   ......C.......
```

Internal 2

```
I2-a   AGCCGTTGCCGAAC
I2-b   .A............
I2-c   .AA......T..TT
```

FIGURE 6.—Sequence variation among REPs and their flanking elements—left end, right end, internal 1, and internal 2—in the *aceA-aceK* intergenic region among strains of *E. coli*. See Figure 5 for locations of the elements and their association in various patterns.

EC20 and EC21, was completely sequenced in EC21; it is 993 bp long and has identity to classes 4 and 6 for the first 71 bp beyond *aceK;* this is followed by an *IS*1 element and then 154 bp of sequence upstream of the end of *iclR* (Figure 10). Variation in the size of the intergenic region of EC17 (class 3) is also due in part to an *IS*1 element. In this case, however, the element is inserted 19 bp downstream of the site of insertion in class 2 and in the opposite orientation (Figure 10). Based on the size of the PCR product, this strain must also have a deletion of at least 800 bp relative to the K-12 sequence (class 6) beyond the site of insertion. The remaining two classes (not shown in Figure 10) were observed in strains EC18 (class 5) and EC34 (class 7), and their intergenic regions are approximately 2170 and 3470 bp, respectively. The composition of these two classes has not yet been determined.

There was much less size and sequence variation in the *aceK-iclR* intergenic region in *S. enterica* than in *E. coli* (Figure 10). Isolate s3333 (class 1), which represents serovar *S. e.* Typhi, and the two strains of subspecies IIIa (class 2) have intergenic regions of 65 and 80 bp, respectively. The first 27 bp of these sequences are 89% identical, but in the remaining part of their sequences they show no similarity to each other or to sequences of the other 13 strains of *S. enterica*. Eleven strains of *S. enterica* (excluding the three strains just discussed and the two strains of subspecies VI) have an intergenic region of 402–403 bp, designated as class 3. Within this class, 18% of the sites are polymorphic and strains of the same subspecies are closely similar. Subspecies V strains are the

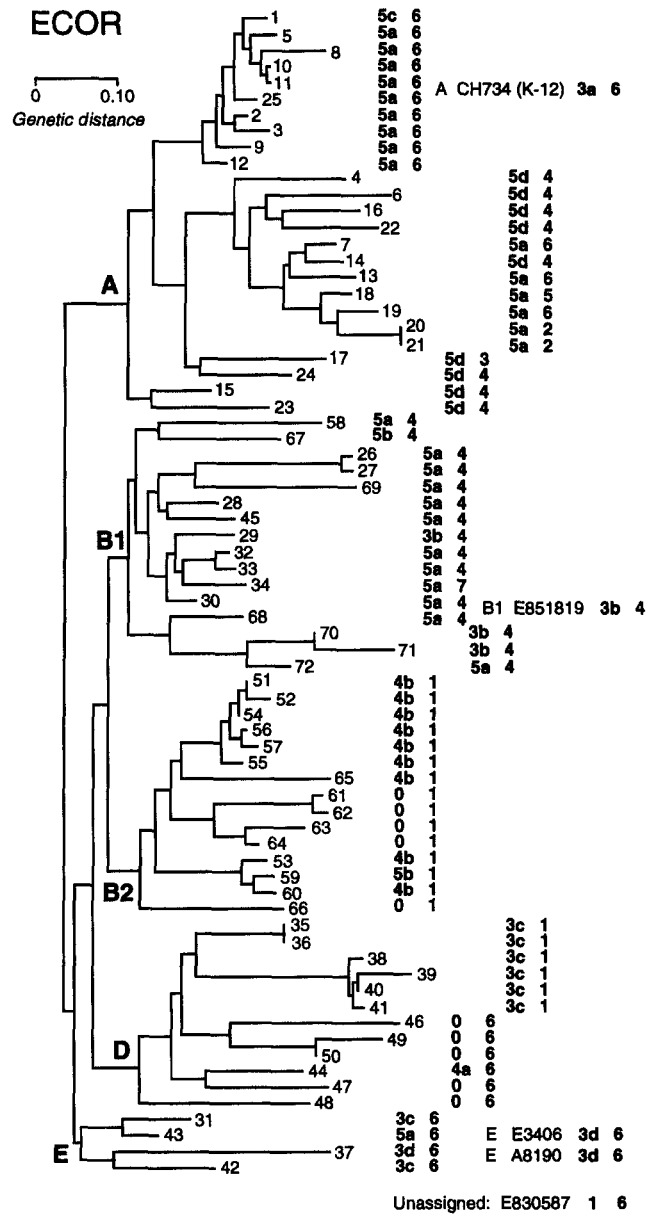

FIGURE 7.—Distribution of patterns of the *aceA-aceK* intergenic region and classes of the *aceK-iclR* intergenic region among the 72 ECOR strains and five additional isolates of *E. coli*. The overall evolutionary genomic relationships of the strains are indicated by their positions in the neighbor-joining tree, which is based on an MLEE analysis of allelic variation at 38 loci (HERZER *et al.* 1990), and the major groups of the ECOR strains are indicated at the major nodes. ECOR strain numbers are shown at the branch ends, and for each strain, the pattern of the *aceA-aceK* intergenic region (see Figure 5) and the class of the *aceK-iclR* intergenic region (see Figure 10) are indicated.

most divergent, and the sequences of subspecies IV and VII differ from one another by 7.6%. The sequences of strains of subspecies VI (class 4) are much longer (approximately 1000 bp) than those of the other salmonellae and have no sequence homology with them. The sequences of the two strains of subspecies VI differ from one another by an insertion/deletion of 51 bp.

	Strain	*aceA*	Intergenic Region	*aceK*
		1111111111111111111111111111111111111111111		
		1111111111111111111112222222222222222222333	1111	1111
		0001122344455667889900012233344667888890001	24446667888889999990000	33556788689 9
		1470628436925479691703981436925695124732 51	12345677367567434678023466789	094955472335
I	s4194	TACGCCGTGTAACCGCTCCTTCGGTTCCACCGATGTAGCGGT	------TGGAGA--CTGCGTGGCCGAAAA	TCCATGTCTGCG
I	s3333	...TT.A.....A.CT.....T..	----------.....T.....-	...G.....
V	s3044	CGTC.TA.....A.C.T.....A...TT..	----------	ATT.GTCTCATA
V	s3041	CGTC.G.CACGGAGTGA..CCTACAGTTGGTCTCACGCGTAC	TCAATGCAA--GGTTACTCCCT-TCGGCG	ATT.GTC.CATA

FIGURE 8.—Sequence variation in the *aceA-aceK* intergenic region and adjacent parts of the *aceA* and *aceK* genes among four strains of *S. enterica* representing subspecies I and V.

DISCUSSION

Analysis of the *aceK* gene and its flanking intergenic regions has revealed variation in size as well as in nucleotide sequence within and between *S. enterica* and *E. coli*. In all, six length variants of *aceK* were identified, ranging in size from 1722 to 1785 bp, and the gene was consistently longer in strains of *S. enterica* than in those of *E. coli*. An especially surprising feature is the use of either TGA or TAA as a stop codon both within and between species. We are unable to construct a detailed evolutionary scheme to account for the observed variation in length, sequence, and stop codon usage, but inasmuch as it involves changes in the 3′ end of the gene, it is likely to be related, in major part at least, to events that have occurred in the *aceK-iclR* intergenic region. Each of the four length variants of *aceK* in *S. enterica* was associated with an intergenic region of distinctive length and nucleotide sequence, and the two length variants of *E. coli* were associated with different types of intergenic regions; the short (1722 bp) variant was consistently paired with a short *aceK-iclR* intergenic region, whereas the long (1734 bp) variant was associated with a variety of intergenic region length and sequence types. The variation in *E. coli* is in part the result of the presence or absence of open reading frame *f728;* and the observation that *f728* has a G + C content of only 37.5% and uses codons that are otherwise rarely employed in *E. coli* (AGA, AGG, CGA, ATA, and CTA) suggests that it was introduced to the *E. coli* genome from an outside source. Because of the AT-richness of this region, it becomes a likely target for *IS*1 insertion, which has occurred at least twice in the lineages of the strains examined.

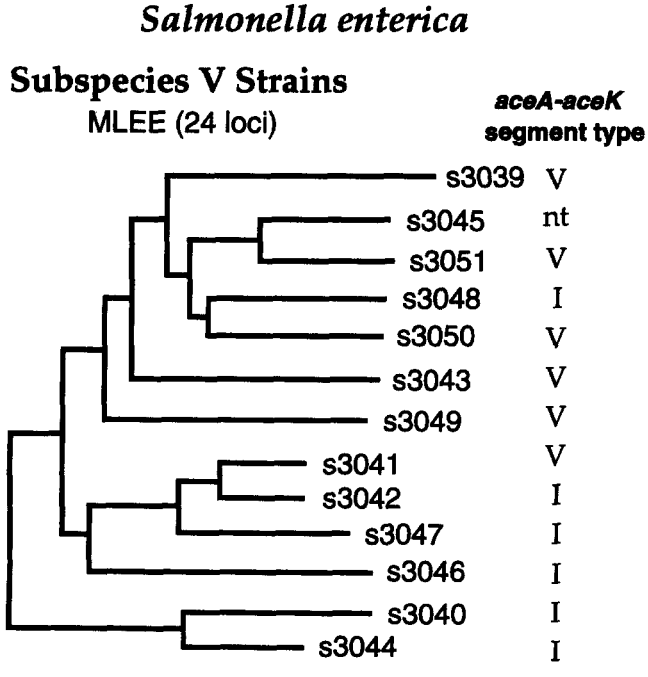

FIGURE 9.—Neighbor-joining tree for 12 strains of *S. enterica* subspecies V, based on an MLEE analysis of allelic variation at 24 loci. For all but one strain, the *aceA-aceK* segment type (see Figure 8) is indicated.

Variation in gene size within a bacterial species or, even, between closely related species is unusual and, indeed, all but unknown in the case of those encoding metabolic enzymes or other essential housekeeping proteins. However, certain other types of gene may vary in size, sometimes markedly. A familiar example is that of the *fliC* gene of *S. enterica* encoding the highly antigenic flagellin protein, which is the polymerizing structural unit of the flagellar filament (WILSON and BEVERIDGE 1993). The central region of this gene, in which the flagellar epitopes are encoded, is hypervariable in both length and sequence (FRANKEL *et al.* 1989; LI *et al.* 1994), and the variation presumably is adaptive in promoting flagellar antigenic diversity within populations (REEVES 1992; BRUNHAM *et al.* 1993; SELANDER *et al.* 1994).

Among strains of *E. coli*, total chromosome length varies from 4.3 to 5.3 megabases, and the deletions, insertions, and duplications underlying this variation are distributed throughout the chromosome (BERGTHORSSON and OCHMAN 1995; H. OCHMAN, personal communication). In contrast, the recorded variation in chromosome length among strains of *S. enterica* is in the range of only a few hundred kilobases and, moreover, apparently involves limited regions of the chromosome (LIU and SANDERSON 1995a,b,c and earlier papers). It must be noted, however, that the available estimates of chromosome size in *S. enterica* pertain to serovars of only a single major lineage (subspecies I), among which the phylogenetic diversity is considerably less than in the species *E. coli* as a whole. But, even so, in the context

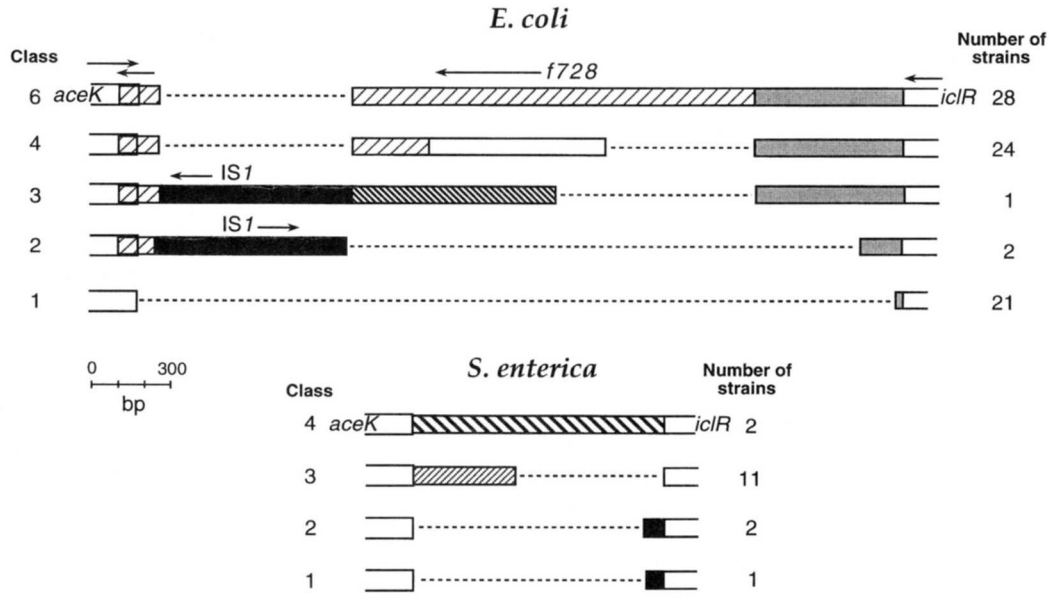

FIGURE 10.—Variation in size, structure, and sequence of the *aceK-iclR* intergenic region among strains of *E. coli* (above) and *S. enterica* (below). See text for explanation.

of this preliminary finding, it is noteworthy that our analysis detected much greater size variation in the intergenic regions flanking *aceK* in *E. coli* than in *S. enterica.*

The clonal model of population structure for *S. enterica* and *E. coli,* initially suggested by MLEE studies (SELANDER and LEVIN 1980; BELTRAN *et al.* 1988; review in SELANDER *et al.* 1994), has been supported by comparative analyses of DNA sequences for the *gapA, putP, mdh,* and *icd* genes in both species (NELSON *et al.* 1991; NELSON and SELANDER 1992; BOYD *et al.* 1994; WANG *et al.* 1997), several other genes in *E. coli* (HALL and SHARP 1992), and *gnd* (NELSON and SELANDER 1994b; THAMPAPAPILLAI *et al.* 1994) and several *inv/spa* genes in *S. enterica* (LI *et al.* 1995). And in the case of *aceK,* the relationships among strains of both species are essentially the same as those indicated by analyses of these other loci. Thus, for both species the realized rate of recombination in housekeeping genes apparently is sufficiently low to permit the long-term, if not permanent, maintenance of differentially adapted, widely distributed chromosomal genotypes in populations.

For *S. enterica,* the most interesting example of recombination is that involving subspecies I and V. In total genetic character, these are by far the two most divergent lineages of *S. enterica* (REEVES *et al.* 1989; SELANDER *et al.* 1994; BOYD *et al.* 1996), yet they have clearly exchanged segments of the *aceK* region and of several other genes. In the *aceA-aceK* intergenic region and the proximal part of the *aceA* gene, six of the 12 subspecies V strains examined had sequences that are nearly identical to those of subspecies I, while the other six strains had a very distinctive sequence, consistent in character with the overall level of genomic divergence of subspecies I and V (Figure 9). There was a strong phylogenetic component to the distribution of these two sequences among the subspecies V isolates that suggests that the exchange has occurred no more than twice (Figure 8). Recombination between subspecies I and V is also evident in the sequences of *putP* (NELSON and SELANDER 1992), *gnd* (THAMPAPAPILLAI *et al.* 1994), and *icd* (WANG *et al.* 1997), with subspecies I the donor and subspecies V the recipient in each case.

In the nucleotide sequence of *aceK,* the relationships

TABLE 2

Sequence variation in six housekeeping genes among 16 strains of *S. enterica*

Gene	No. of basepairs of gene sequenced[a]	No. of polymorphic[a]		Mean pairwise (×100)[b]	
		Nucleotides	Amino acids	d_N	d_S
aceK	1719 (98)	338 (19.7)	47 (8.2)	28.39 ± 1.76	1.05 ± 0.14
icd	1164 (93)	215 (18.5)	11 (2.9)	29.35 ± 2.00	0.25 ± 0.09
gapA	924 (93)	118 (12.8)	14 (4.5)	15.15 ± 1.49	0.61 ± 0.15
putP	1467 (97)	216 (14.7)	21 (4.3)	16.70 ± 1.88	0.60 ± 0.23
mdh	849 (90)	133 (15.7)	11 (3.9)	20.13 ± 1.72	0.48 ± 0.16
gnd	1335 (95)	216 (16.2)	20 (4.5)	21.80 ± 1.40	0.44 ± 0.10

[a] Values in parentheses are percentages.

[b] d_S, the mean number of synonymous (silent) substitutions per silent site between pairs of stains; d_N, the corresponding number of nonsynonymous (replacement) substitutions.

among the four group A strains of *E. coli* differ from those indicated by other genes for which comparable data are available (SELANDER *et al.* 1994), apparently as a result of an episode of intragenic recombination. Two divergent *aceK* sequences were represented (Figure 3), and strains EC14 and EC17 were associated with strains of group B1 in our phylogenetic analysis rather than with the two other group A strains, EC10 and CH734 (Figure 4).

At least in *S. enterica*, the rate of nonsynonymous nucleotide substitution in *aceK* apparently is somewhat higher than those recorded for the several other housekeeping genes for which comparable data are available (Table 2). It is tempting to speculate that both this increased sequence variation and the unusual variation in gene size within and between species are related to the circumstance that *aceK* is not an essential gene, since it functions only to regulate the flux of isocitrate through the tricarboxylic cycle and the glyoxylate bypass.

C. W. HILL provided strain CH734 and T. S. WHITTAM supplied ECOR and other strains of *E. coli* and assisted in the statistical analysis of data. This research was supported by Public Health Service grant AI-22144 from the National Institutes of Health.

LITERATURE CITED

ACHTMAN, M., 1994 Clonal spread of serogroup A meningococci: a paradigm for the analysis of microevolution in bacteria. Mol. Microbiol. **11:** 15–22.

BELTRAN, P., J. M. MUSSER, R. HELMUTH, J. J. FARMER III, W. M. FRERICHS *et al.*, 1988 Toward a population genetic analysis of *Salmonella:* genetic diversity and relationships among strains of serotypes *S. choleraesuis, S. derby, S. dublin, S. enteritidis, S. heidelberg, S. infantis, S. newport,* and *S. typhimurium.* Proc. Natl. Acad. Sci. USA **85:** 7753–7757.

BERGTHORSSON, U., and H. OCHMAN, 1995 Heterogeneity of genome size among natural isolates of *Escherichia coli.* J. Bacteriol. **177:** 5784–5789.

BISERCIC, M., J. Y. FEUTRIER and P. R. REEVES, 1991 Nucleotide sequence of the *gnd* genes from nine natural isolates of *Escherichia coli:* evidence of intragenic recombination as a contributing factor in the evolution of the polymorphic *gnd* locus. J. Bacteriol. **173:** 3894–3900.

BOYD, E. F., K. NELSON, F.-S. WANG, T. S. WHITTAM and R. K. SELANDER, 1994 Molecular genetic basis of allelic polymorphism in malate dehydrogenase (*mdh*) in natural populations of *Escherichia coli* and *Salmonella enterica.* Proc. Natl. Acad. Sci. USA **91:** 1280–1284.

BOYD, E. F., F.-S. WANG, T. S. WHITTAM and R. K. SELANDER, 1996 Molecular genetic relationships of the salmonellae. Appl. Environ. Microbiol. **62:** 804–808.

BOYD, E. F., J. LI, H. OCHMAN and R. K. SELANDER, 1997 Comparative genetics of the *inv-spa* invasion gene complex of *Salmonella enterica.* J. Bacteriol. **179:** 1985–1991.

BRUNHAM, R. C., F. A. PLUMMER and R. S. STEPHENS, 1993 Bacterial antigenic variation, host immune response, and pathogen-host coevolution. Infect. Immunol. **61:** 2273–2276.

CABOT, E. L., and A. T. BECKENBACH, 1989 Simultaneous editing of multiple nucleic acid and protein sequences with ESEE. Comput. Appl. Biosci. **5:** 233–234.

CORTAY, J.-C., F. BLEICHER, C. RIEUL, H. C. REEVES and A. J. COZZONE, 1988 Nucleotide sequence and expression of the *aceK* gene coding for isocitrate dehydrogenase kinase/phosphatase in *Escherichia coli.* J. Bacteriol. **170:** 89–97.

DYKHUIZEN, D. E., and L. GREEN, 1991 Recombination in *Escherichia coli* and the definition of the biological species. J. Bacteriol. **173:** 7257–7268.

FRANKEL, G., S. M. C. NEWTON, G. K. SCHOOLNIK and B. A. D. STOCKER, 1989 Intragenic recombination in a flagellin gene: characterization of the *H1*:j gene of *Salmonella typhi.* EMBO J. **8:** 3149–3152.

GALINIER, L., F. BLEICHER, D. NÈGRE, G. PERRIÈRE, B. DUCLOS *et al.*, 1991 Primary structure of the intergenic region between *aceK* and *iclR* in the *Escherichia coli* chromosome. Gene **97:** 149–150.

HALL, B. G., and P. M. SHARP, 1992 Molecular population genetics of *Escherichia coli:* DNA sequence diversity at the *celC, crr,* and *gutB* loci of natural isolates. Mol. Biol. Evol. **9:** 654–665.

HERZER, P. J., S. INOUYE, M. INOUYE and T. S. WHITTAM, 1990 Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli.* J. Bacteriol. **172:** 6175–6181.

KLUMPP, D. J., D. W. PLANK, L. J. BOWDIN, C. S. STUELAND, T. CHUNG *et al.*, 1988 Nucleotide sequence of *aceK,* the gene encoding isocitrate dehydrogenase kinase/phosphatase. J. Bacteriol. **170:** 2763–2769.

KUMAR, S., K. TAMURA and M. NEI, 1993 MEGA: molecular evolutionary genetics analysis, version 1. The Pennsylvania State University, University Park, PA 16802.

LAPORTE, D. C., 1993 The isocitrate dehydrogenase phosphorylation cycle: regulation and enzymology. J. Cell. Biochem. **51:** 14–18.

LI, J., K. NELSON, A. C. MCWHORTER, T. S. WHITTAM and R. K. SELANDER, 1994 Recombinational basis of serovar diversity in *Salmonella enterica.* Proc. Natl. Acad. Sci. USA **91:** 2552–2556.

LI, J., H. OCHMAN, E. A. GROISMAN, E. F. BOYD, F. SOLOMON *et al.*, 1995 Relationship between evolutionary rate and cellular location among the Inv/Spa invasion proteins of *Salmonella enterica.* Proc. Natl. Acad. Sci. USA **92:** 7252–7256.

LIU, S.-L., and K. E. SANDERSON, 1995a I-*Ceu*I reveals conservation of the genome of independent strains of *Salmonella typhimurium.* J. Bacteriol. **177:** 3355–3357.

LIU, S.-L., and K. E. SANDERSON, 1995b Rearrangement in the genome of the bacterium *Salmonella typhi.* Proc. Natl. Acad. Sci. USA **92:** 1018–1022.

LIU, S.-L., and K. E. SANDERSON, 1995c The chromosome of *Salmonella paratyphi* A is inverted by recombination between *rrnH* and *rrnG.* J. Bacteriol. **177:** 6585–6592.

MATSUOKA, M., and B. A. MCFADDEN, 1988 Isolation, hyperexpression, and sequencing of the *aceA* gene encoding isocitrate lyase in *Escherichia coli.* J. Bacteriol. **170:** 4528–4536.

MILKMAN, R., and M. M. BRIDGES, 1993 Molecular evolution of the *Escherichia coli* chromosome. IV. Sequence comparisons. Genetics **133:** 455–468.

MOXON, E. R., P. B. RAINEY, M. A. NOWAK and R. E. LENSKI, 1994 Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr. Biol. **4:** 24–33.

NELSON, K., and R. K. SELANDER, 1992 Evolutionary genetics of the proline permease gene (*putP*) and the control region of the proline utilization operon in populations of *Salmonella* and *Escherichia coli.* J. Bacteriol. **174:** 6886–6895.

NELSON, K., and R. K. SELANDER, 1994a Analysis of genetic variation by polymerase chain reaction-based nucleotide sequencing. Methods Enzymol. **235:** 174–183.

NELSON, K., and R. K. SELANDER, 1994b Intergeneric transfer and recombination of the 6-phosphogluconate dehydrogenase gene (*gnd*) in enteric bacteria. Proc. Natl. Acad. Sci. USA **91:** 10227–10231.

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

OCHMAN, H., and R. K. SELANDER, 1984 Standard reference strains of *Escherichia coli* from natural populations. J. Bacteriol. **157:** 690–693.

REEVES, M. W., G. M. EVINS, A. A. HEIBA, B. D. PLIKAYTIS and J. J. FARMER III, 1989 Clonal nature of *Salmonella typhi* and its genetic relatedness to other salmonellae as shown by multilocus enzyme electrophoresis, and proposal of *Salmonella bongori* comb. nov. J. Clin. Microbiol. **27:** 311–320.

REEVES, P. R., 1992 Variation in O-antigens, niche-specific selection and bacterial populations. FEMS Microbiol. Lett. **100:** 509–516.

RIEUL, C., F. BLEICHER, B. DUCLOS, J.-C. CORTAY and A. J. COZZONE,

1988 Nucleotide sequence of the *aceA* gene coding for isocitrate lyase in *Escherichia coli*. Nucleic Acids Res. **16:** 5689.

SAIKI, R. K., D. H. GELFAND, S. STOFFEL, S. J. SCHARF, R. HIGUCHI *et al.*, 1988 Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science **239:** 487–491.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

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

SELANDER, R. K., J. LI, E. F. BOYD, F.-S. WANG and K. NELSON, 1994 DNA sequence analysis of the genetic structure of populations of *Salmonella enterica* and *Escherichia coli*, pp. 17–49 in *Bacterial Diversity and Systematics*, edited by F. G. PRIEST, A. RAMOS-CORMENZANA and B. J. TINDALL. Plenum Press, New York.

SMITH, N. H., P. BELTRAN and R. K. SELANDER, 1990 Recombination of *Salmonella* phase 1 flagellin genes generates new serovars. J. Bacteriol. **172:** 2209–2216.

STEPHENS, J. C., 1985 Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol. Biol. Evol. **2:** 539–556.

THAMPAPAPILLAI, G., R. LAN and P. R. REEVES, 1994 Molecular evolution in the *gnd* locus of *Salmonella enterica*. Mol. Biol. Evol. **11:** 813–828.

WANG, F.-S., T. S. WHITTAM and R. K. SELANDER, 1997 Evolutionary genetics of the isocitrate dehydrogenase gene (*icd*) in *Escherichia coli* and *Salmonella enterica*. J. Bacteriol. (in press).

WILSON, D. R., and T. J. BEVERIDGE, 1993 Bacterial flagellar filaments and their component flagellins. Can. J. Microbiol. **39:** 451–472.

Communicating editor: A. G. CLARK