JOURNAL OF BACTERIOLOGY, Nov. 1997, p. 6551–6559 Vol. 179, No. 21
0021-9193/97/$04.00+0
Copyright © 1997, American Society for Microbiology

# Evolutionary Genetics of the Isocitrate Dehydrogenase Gene (*icd*) in *Escherichia coli* and *Salmonella enterica*

FU-SHENG WANG,* THOMAS S. WHITTAM, AND ROBERT K. SELANDER

*Institute of Molecular Evolutionary Genetics, Mueller Laboratory, Pennsylvania State University, University Park, Pennsylvania 16802*

Received 12 May 1997/Accepted 19 August 1997

**Sequences of the *icd* gene, encoding isocitrate dehydrogenase (IDH), were obtained for 33 strains representing the major phylogenetic lineages of *Escherichia coli* and *Salmonella enterica*. Evolutionary relationships of the strains based on variation in *icd* are generally similar to those previously obtained for several other housekeeping and for invasion genes, but the sequences of *S. enterica* subspecies V strains are unusual in being almost intermediate between those of the other *S. enterica* subspecies and *E. coli*. For *S. enterica*, the ratio of synonymous (silent) to nonsynonymous (replacement) nucleotide substitutions between pairs of strains was larger than comparable values for 12 other housekeeping and invasion genes, reflecting unusually strong purifying selection against amino acid replacement in the IDH enzyme. All amino acids involved in the catalytic activity and conformational changes of IDH are strictly conserved within and between species. In *E. coli*, the level of variation at the 3′ end of the gene is elevated by the presence in some strains of a 165-bp replacement sequence supplied by the integration of either lambdoid phage 21 or defective prophage element e14. The 72 members of the *E. coli* Reference Collection (ECOR) and five additional *E. coli* strains were surveyed for the presence of phage 21 (as prophage) by PCR amplification of a phage 21-specific fragment in and adjacent to the host *icd*, and the sequence of the phage 21 segment extending from the 3′ end of *icd* through the integrase gene (*int*) was determined in nine strains of *E. coli*. Phage 21 was found in 39% of *E. coli* strains, and its distribution among the ECOR strains is nonrandom. In two ECOR strains, the phage 21 *int* gene is interrupted by a 1,313-bp insertion element that has 99.3% nucleotide sequence identity with IS*3411* of *E. coli*. The phylogenetic relationships of phage 21 strains derived from sequences of two different genomic regions were strongly incongruent, providing evidence of frequent recombination.**

NADP$^+$-dependent isocitrate dehydrogenase (IDH), encoded by the *icd* gene, catalyzes the conversion of isocitrate to α-ketoglutarate and $CO_2$, which can be a rate-limiting step in the citric acid cycle, and is also involved in the regulation of carbon flux at the branch point between the citric acid and glyoxylate cycles. In *Escherichia coli*, IDH inactivation is achieved by phosphorylation of serine 113 in the active site (3, 51), a reversible process mediated by isocitrate dehydrogenase phosphatase/kinase, which is encoded by the *aceK* gene (12). With the inactivation of IDH, as much as 30% of the isocitrate pool may be diverted to the glyoxylate bypass (52).

The *icd* gene is at 26 min on the linkage map of *E. coli* (2) and has a coding sequence of 1,248 bp (51). In strains of *E. coli*, a lambdoid phage 21 or a defective prophage element, e14, may be integrated at the same position in the 3′ end of *icd*, to which they introduce an alternative 165-bp terminal segment of the gene and an adjacent 113-bp segment that contains the *icd* transcriptional terminator (7, 8, 10, 17, 21, 27, 45). As a consequence of this replacement, the structure and function of *icd* and IDH are maintained (21).

We report here the results of a comparative analysis of the *icd* sequences of 33 strains of *E. coli* and *Salmonella enterica*, a survey of the distribution of *icd*-integrated phage 21 and e14 among natural isolates of *E. coli*, and a study of sequence variation in part of the phage 21 genome.

* Corresponding author. Present address: Department of Microbiology and Immunology, University of Western Ontario, London, Ontario, Canada N6A 5C1. Phone: (519) 661-3433. E-mail: fswang@julian.uwo.ca.

## MATERIALS AND METHODS

**Bacterial isolates.** We examined 17 strains of *E. coli*, most of which were previously studied for sequence variation in five other housekeeping genes (4, 35, 37–39). From the *E. coli* Reference Collection (ECOR) (40) and the research collection of T. S. Whittam, we selected 15 strains representing the five major evolutionary lineages that have been identified by multilocus enzyme electrophoresis (MLEE) (18, 46), as follows: EC10, EC14, EC15, and EC17 (representing ECOR group A, the lineage to which the laboratory strain K-12 belongs); EC32, EC58, EC69, EC70, and E851819 (ECOR group B1); EC52 and EC64 (ECOR group B2); EC40 (ECOR group D); and EC37, A8190, and E3406 (ECOR group E). Two K-12 strains were provided by C. W. Hill: CH734, which carries an e14 element, and CH1332, a derivative of CH734 that lacks the element (21).

Two strains of each of the eight subspecies of *S. enterica*, I, II, IIIa, IIIb, IV, V, VI, and VII (48), were selected for study. These 16 strains, which constitute the *Salmonella* Reference Collection C (5), were previously examined for sequence variation in housekeeping genes and invasion genes of the *inv-spa* complex (6, 28).

For comparative purposes, the *icd* sequence was also obtained for a strain (CIT42) of *Citrobacter diversus*.

**PCR and sequencing of *icd*.** Primers for PCR of the *icd* gene in strains of *E. coli* were designed from the *icd* sequence of *E. coli* laboratory strain K-12 reported by Thorsness and Koshland (51), and primers for amplification of *icd* in *S. enterica* and *C. diversus* were based on the consensus sequence of the gene for the *E. coli* strains.

Single-stranded DNA was generated by the λ-exonuclease method (20, 36), and the sequence was determined for both strands by the dideoxynucleotide chain termination method, as supplied with Sequenase (United States Biochemical), with the use of sets primers spaced at 200- to 250-bp intervals.

**PCR and sequencing of phage segments.** The presence of phage 21 (as prophage) in strains of ECOR and five additional strains of *E. coli* was detected by PCR amplification of a segment extending from the mid-region of *icd* through most of the integrase gene (*int*) of phage 21 with use of primers representing nucleotides 869 to 886 of the *E. coli* K-12 *icd* gene (51) and a segment that includes the 3′ end of the phage 21 excisionase gene (*xis*) and the 5′ end of the integrase gene (*int*) (Fig. 1). The latter primer corresponds to nucleotides 941 to 964 of the phage 21 sequence deposited in GenBank (accession no. M61865) by S. J. Schneider and A. M. Campbell.

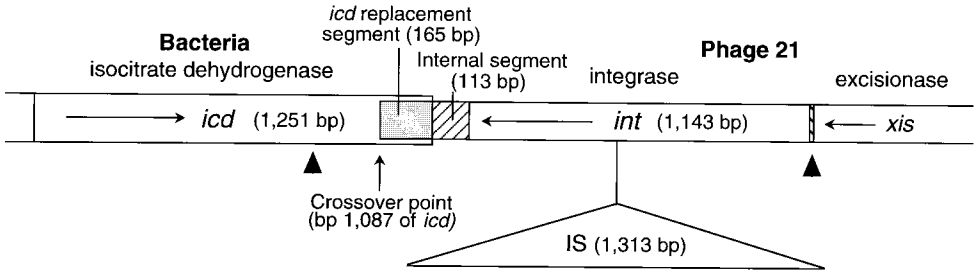


FIG. 1. Spatial relationships of the *E. coli icd* gene and integrated phage 21. There is an 11-bp overlap between the phage *xis* and the *int* genes. The insertion sequence (IS) element was identified in strain EC38 and presumably also occurs in strain EC39. Triangles indicate positions of PCR primers used to survey strains for the presence of phage 21.

A 1,068-bp segment of the *int* gene was sequenced in each of nine ECOR strains by the methods used to sequence *icd*.

Strains of *E. coli* were tested for the presence of element e14 with use of the *icd* primer employed to screen for phage 21 and a primer corresponding to nucleotides 781 to 803 of the e14 sequence located downstream from *icd* in *E. coli* K-12 strain CH734 (45).

**Data analysis.** The nucleotide sequence data were assembled with the SEQ MAN and SEQMANED programs (DNASTAR, Madison, Wis.), aligned manually with the Eyeball Sequence Editor, version 1.09 (9), and analyzed with the MEGA program (26) and the IMEG program, written by T. S. Whittam.

**Nucleotide sequence accession numbers.** The GenBank accession numbers for the sequences reported here are AF017587 through AF017629.

## RESULTS

***icd*. (i) Sequence variation in *icd*.** For each of 17 strains of *E. coli*, the sequence of a 1,212-bp segment (codons 11 to 414) representing 97% of the 1,248-bp coding region of the *icd* gene was determined. (For K-12 strain CH734, the nucleotide sequence was identical to that of the K-12 strain previously sequenced by Thorsness and Koshland [51].) There were 73 polymorphic nucleotide sites (Fig. 2), and the sequences of pairs of strains differed, on average, at 1.8% of nucleotide sites and 0.3% of amino acid positions.

In *E. coli*, the level of variation was higher at the 3′ end of the gene (codons 362 to 414) than elsewhere, with an average pairwise difference of 5.0% of nucleotides and 0.8% of amino acids. This increase is attributable, in part at least, to the presence in many strains of either phage 21 or e14, both of which insert at bp 1087 (in codon 362) and supply an alternative 216-bp segment of the host chromosome, which includes the 165-bp 3′-terminal part of the *icd* gene (including the stop codon 417) and a stretch of 51 bp beyond the *icd* coding sequence (10, 21). In both species, the frequency of synonymous-site polymorphism is relatively low at the 5′ end of *icd*, a common feature of enterobacterial genes (16).

The phage 21 *icd* replacement segments did not contribute any nonsynonymous substitutions, but the replacement segment of the e14 element in strain CH734 carries an amino acid variant, a conservative Glu→Asp substitution at codon 398 (Fig. 3).

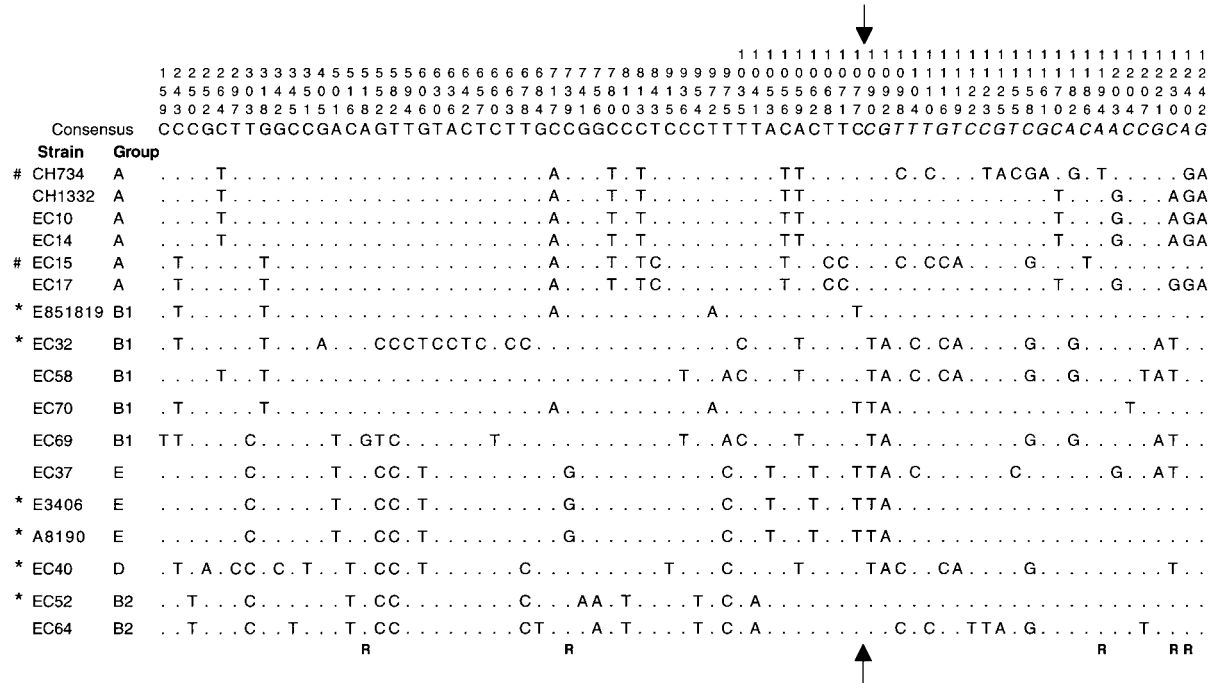


FIG. 2. Nucleotide sequence variation in *icd* among 17 strains of *E. coli*. Only the polymorphic sites are shown. Dots indicate identify with the consensus sequence; arrows indicate the sites of insertion (nucleotide 1087) of phage 21 in strains E851819, EC32, E3406, A8190, EC40, and EC52 and of e14 in strain CH734. Additionally, EC15 may harbor an e14-like element (see text).

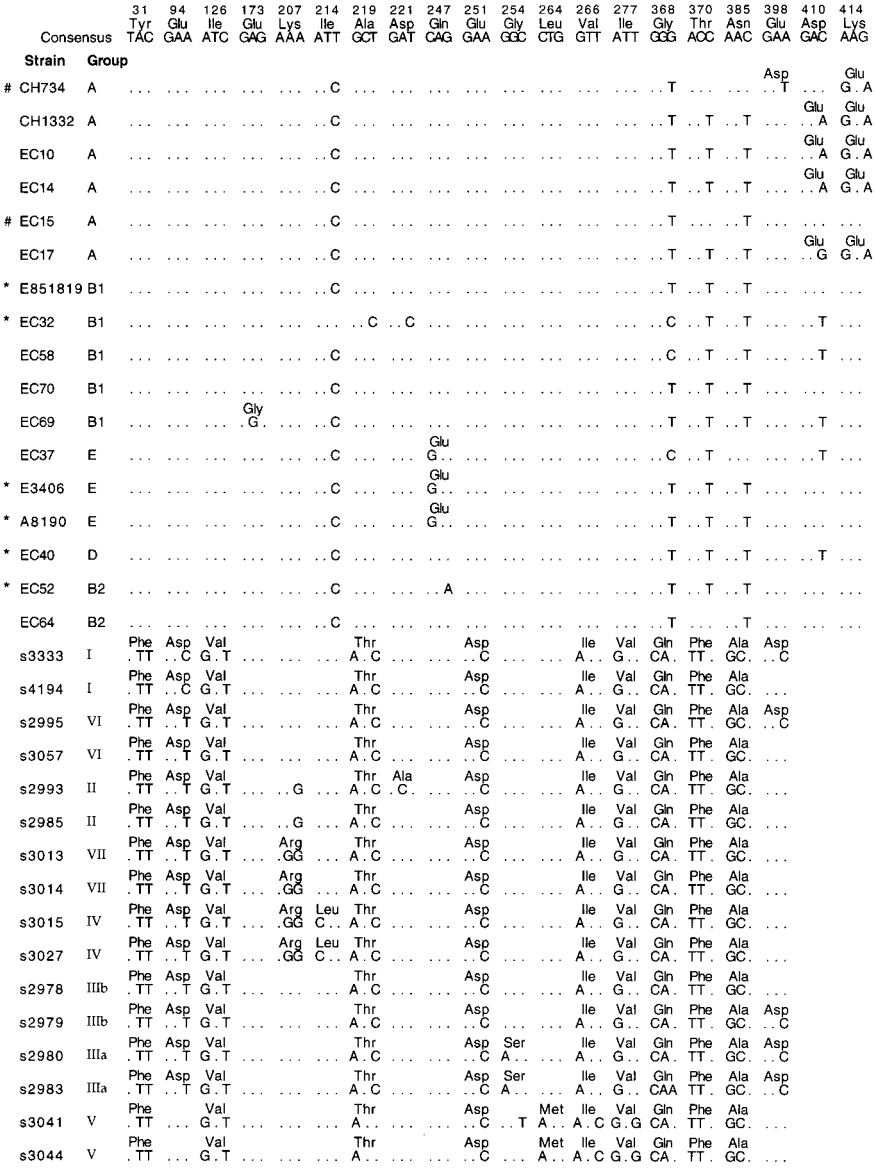


FIG. 3. Amino acid sequence variation in IDH among 17 strains of *E. coli* and 16 strains of *S. enterica*. Only those codons at which amino acid replacement nucleotide substitutions occur are shown. Dots indicate identity with the consensus nucleotide sequence. Strains marked by * carry integrated phage 21; those marked with # carry integrated e14 or an e14-like element.

For 16 strains of *S. enterica*, a 1,164-bp segment (codons 17 to 404), representing 93% of the *icd* gene, was sequenced. There were 215 polymorphic nucleotide sites (Fig. 4), and pairs of strains differed, on average, at 5.6% of nucleotide sites and 0.5% of the amino acid positions.

There were 285 polymorphic nucleotide sites among the 33 sequences of *E. coli* and *S. enterica*, and strains of the two species differed, on average, at 13.3% of nucleotide sites and 2.9% of the amino acid positions. For the 3′ end of the *icd* gene involved in the integration of phage 21 and e14 in *E. coli*, pairs of strains of *E. coli* and *S. enterica* showed an average difference of 16.1% of nucleotides and 7.9% of amino acids.

**(ii) Amino acid sequence variation.** There are nine fixed amino acid sequence differences between *E. coli* and *S. enterica* (Fig. 3), all of which involve conservative substitutions. Additionally, all subspecies of *S. enterica* except V differ from *E. coli* in having Asp rather than Glu at codon 94. Note that the Asp codon is GAC in subspecies I but GAT in the other subspecies of *S. enterica*.

Studies of the molecular structure and function of IDH in *E. coli* have identified 36 amino acid residues that are critically important in catalysis or are involved in catalysis-induced conformational changes in enzyme configuration (14, 15, 22–24). None of these amino acids was polymorphic within or between *E. coli* and *S. enterica*, which is not surprising, inasmuch as many of these amino acids are conserved even among very distantly related bacteria (13, 31, 32).

**(iii) Divergence within and between species.** To compare the relative extent of sequence divergence in *icd* within and between species, the method of Whittam and Nei (53) was applied to a total of 31 sequences of *E. coli* and *S. enterica*. This analysis involved a comparison of the ratio of estimated num-

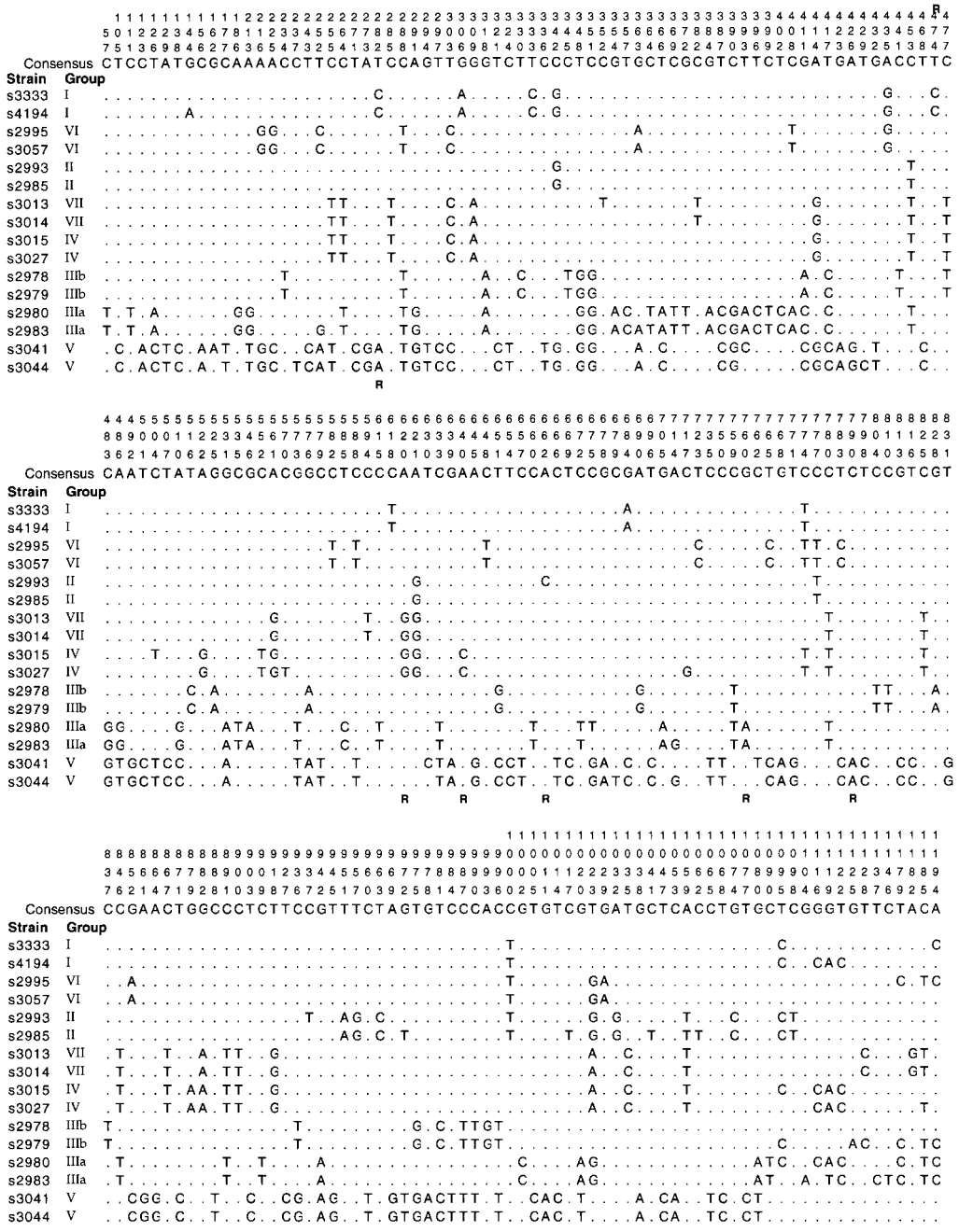


FIG. 4. Nucleotide sequence variation in *icd* among 16 strains of *S. enterica*. Only the polymorphic nucleotide sites are shown. Dots indicate identity with the consensus sequence. R indicates the occurrence of a substitution that results in an amino acid replacement.

bers of nonsynonymous to synonymous substitutions between species to the ratio for pairs of alleles within species. In the absence of natural selection (neutral theory of molecular evolution), the two ratios are expected to be equal.

The number of synonymous substitutions per 100 synonymous sites estimated to have occurred since the time of divergence of *E. coli* and *S. enterica* from a common ancestor ($d_A$) is 63.3 ± 8.2, and the average within-species level of diversity ($\pi$) is 18.3 ± 1.3. Comparable values for nonsynonymous substitutions are $d_A = 1.54 \pm 0.5$ and $\pi = 0.15 \pm 0.0$, respectively. The ratio of nonsynonymous to synonymous substitutions between species is 0.025 ± 0.009, and the ratio of the within-species diversities is 0.008 ± 0.003, which is significantly lower. Inasmuch as the between-species ratio is similar to that reported for the malate dehydrogenase gene (*mdh*) (4), the tentative inference from this analysis is that selection has acted in an unusually strong manner against nonsynonymous substitutions in *icd* within species.

Another, more direct way of analyzing sequence divergence is to construct a 2 × 2 contingency table comparing the actual numbers of nonsynonymous and synonymous sites that are fixed between species to those that are polymorphic within

species (30, 44). In the sample of 31 strains, 14 nonsynonymous and 29 synonymous sites showed fixed differences between *E. coli* and *S. enterica*, and 20 nonsynonymous and 221 synonymous sites were polymorphic within species. The difference in proportions of nonsynonymous and synonymous sites between and within species is highly significant ($\chi^2 = 20.38$, $P = 0.00001$).

The results of these analyses suggest that the divergence of the two species involved an episode of relatively rapid amino acid substitution.

**(iv) Distribution of synonymous polymorphic sites.** A test developed by Stephens (50) was used as a guide to the identification of nonrandom clusters of synonymous polymorphic nucleotide sites, which may be indicative of intragenic recombination. For the 66 silent polymorphic sites in the 17 *E. coli* *icd* sequences, the test identified only one partition for which the distribution of sites is significantly nonrandom. This partition, which separates the sequence of strain EC32 from those of all other strains, includes a cluster of five unique polymorphic sites in a 33-bp segment (bp 630 to 663). If this clustering is, in fact, the result of an intragenic recombination event, the source of the segment is unknown.

Among the 16 *S. enterica* sequences, there were 206 silent polymorphic sites. The Stephens test identified four partitions with significantly nonrandom distributions of sites. The first partition, which separates the two strains of subspecies V from all other strains and was supported by a total of 58 sites distributed over a 972-bp segment beginning near the 5′ end of the gene (bp 105 to 1077), reflects the fact that the 3′ end (bp 1081 to 1164) in subspecies V is identical in sequence to the corresponding segment of the genes of certain strains of subspecies I, II, and VI. This clearly points to the acquisition of this segment by subspecies V through horizontal transfer. The second partition, which grouped the two sequences of subspecies IIIb, was supported by a total of 12 sites in a 759-bp segment (bp 234 to 993); this partition identifies a 170-bp segment at the 3′ end of the gene (bp 994 to 1164) in which there are no unique polymorphic sites. The third partition, which separated subspecies II from all other subspecies, was supported by a total of six sites distributed over a 147-bp segment (bp 951 to 1098). This small segment appears to have been acquired by horizontal transfer, but the source is unknown. The fourth partition, which grouped the four strains of subspecies IIIb and V, identified a cluster of five sites in a 21-bp segment (bp 975 to 996) that apparently has been exchanged between the two subspecies.

In addition to the nonrandom clusters detected by the Stephens test, there is a 48-bp segment (bp 354 to 402) containing 12 unique polymorphic sites in the sequences of both strains of subspecies IIIa. Finally, as noted below, there is reason to believe that a large part of the *icd* sequence of the subspecies V strains was acquired by horizontal transfer from an enterobacterial species other than *S. enterica* or *E. coli*.

**(v) An evolutionary tree for *icd* sequences.** To examine the relationships among strains, an evolutionary tree for *icd* was constructed by the neighbor-joining method (43) from a matrix of pairwise estimated genetic distances (with Jukes-Cantor correction) based on synonymous nucleotide sites (Fig. 5). A strain of *C. diversus* served as an outgroup, and the reliability of the branching order was determined by bootstrap analysis of 1,000 computer-generated trees.

The analysis of *icd* variation placed most of the ECOR strains in the same groups to which they have been assigned by MLEE (18). Thus, ECOR group A strains clustered together, as did those of group B2 and of group E. However, group B1 strain EC32 is not associated with the other group B1 strains,

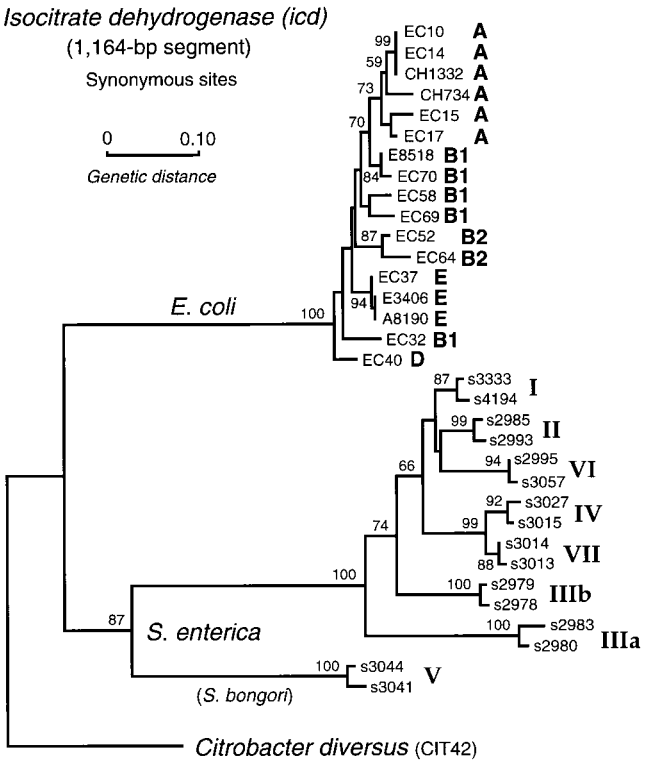


FIG. 5. Neighbor-joining evolutionary tree for strains of *E. coli* and *S. enterica* and a strain of *C. diversus*, based on variation at synonymous nucleotide sites in a 1,164-bp segment of the *icd* gene. The ECOR group assignments of strains of *E. coli* are indicated by letter, and the subspecies of *S. enterica* are designated by roman numerals. Bootstrap values based on 1,000 computer-generated trees are indicated at the nodes.

apparently as a consequence of the recombinant acquisition of a horizontally transferred segment, as noted above.

For *S. enterica*, *icd* sequences of strains of the same subspecies were much more similar to one another than to sequences from other subspecies. Compared with consensus trees for five other housekeeping genes and seven invasion genes (6, 47, 48), the topology of the *icd* tree shows several unusual features. First, the degree of synonymous-site divergence of subspecies V from the other seven subspecies is unusually large, despite the fact that, as previously noted, the 3′ end of the gene in subspecies V is identical in sequence to the genes of certain strains of subspecies I, II, and VI, clearly as a consequence of intragenic recombination. In contrast, the degree of divergence of subspecies V at nonsynonymous sites is not unusual; it shares with the other subspecies most (nine) of the fixed amino acids by which they differ from *E. coli* and has only two unique amino acid substitutions (Fig. 3). Because we are reluctant to invoke an increased synonymous-site mutation rate to account for the marked divergence of the subspecies V sequences, we are forced to postulate the acquisition by horizontal transfer of a major part of the gene from an (unidentified) donor that is a close relative of *S. enterica*.

Second, subspecies IIIb is distant from I, II, and VI, with which it normally clusters. This change reflects the apparent recombinational acquisition of a 794-bp segment of the 5′ end and middle parts of the gene, as described above.

Finally, subspecies IV and VII show relatively little differentiation in *icd* sequence, being in this regard similar to the invasion genes (6, 28, 48).

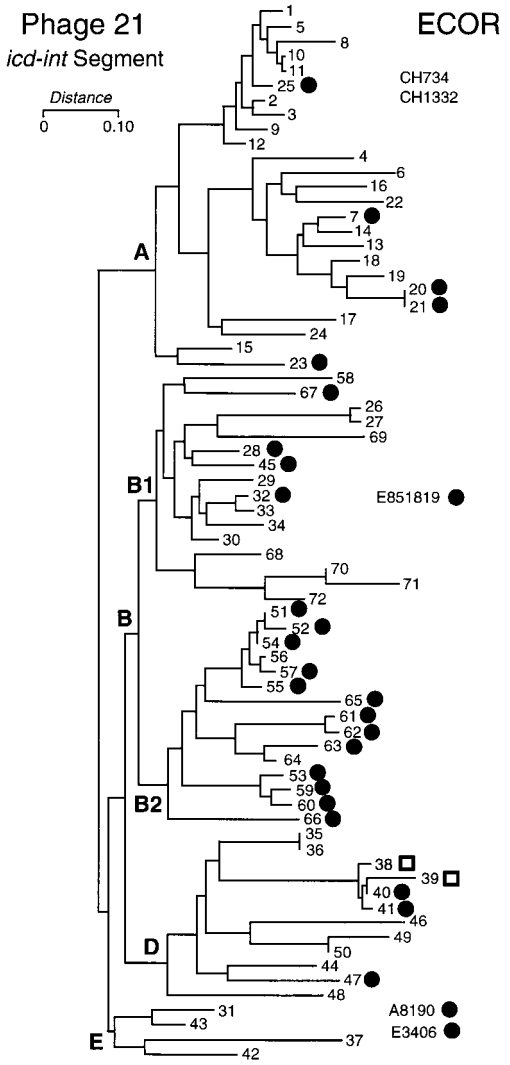


FIG. 6. Distribution of phage 21 (as prophage) among the 72 ECOR strains and 5 additional strains of *E. coli*. The overall genomic evolutionary relationships of the ECOR strains are indicated by their positions in the neighbor-joining tree, which is based on an MLEE analysis of allelic variation at 38 enzyme loci (18). The relationships of the five strains that are not members of the ECOR collection are indicated by their placement in the diagram; for example, strain E851819 is assignable to ECOR group B1. Dots and open squares indicate the presence of normal-sized phage *int* genes and those containing a 1.3-kb insertion sequence, respectively.

**Phage 21 and e14 in *E. coli*. (i) Natural occurrence of phage 21.** A sample of 77 strains of *E. coli*, including the ECOR collection, were screened for the presence of phage 21 (as prophage) by PCR amplification of a segment extending from the mid-region of the bacterial *icd* gene through the phage 21 integrase gene (*int*) (Fig. 1). A total of 30 (39%) of the strains yielded a PCR product (Fig. 6), indicating the presence of at least part of the phage genome. In 28 of the positive strains, the product was 1.6 kb in length, and in the two remaining strains (EC38 and EC39), which are closely related members of ECOR group D, a 3.0-kb segment was amplified.

Although phage 21 was distributed throughout the major lineages of the ECOR collection, it was relatively infrequent among strains of groups A and B1 (Fig. 6). In contrast, it was detected in all but two of the 15 strains of group B2.

**(ii) Sequence variation in phage 21.** Among the *E. coli* strains that yielded phage 21 PCR amplification products, nine strains, representing members of the five major phylogenetic lineages of *E. coli*, were chosen for sequencing of the 3′ *icd-int* region of their integrated phages. They are EC20 (representing ECOR group A); EC67 and E851819 (group B1); EC52 and EC53 (group B2); EC38, EC40, and EC47 (group D); and E3406 (group E).

In all nine strains, the amplified segment included a 165-bp 3′ end of *icd*, a 113-bp internal region (in which the *icd* transcriptional terminator is located), and a 1,068-bp portion of the *int* gene, with a combined length that is consistent with that of the sequence of K-12 strain W3350(21), determined by Schneider (45) and deposited in GenBank (accession no. M61865). However, in strain EC38, a 1,313-bp insertion sequence was located at nucleotide 691 of the *int* gene. This element has 99.3% nucleotide sequence identity with IS*3411* of *E. coli* (25), 99.5% identity with IS*629* of *Shigella sonnei* (29), and 95.6% identity with IS*1203* of *E. coli* (41).

The following analysis of the *icd-int* region of phage 21 is based on the nine sequences that we generated and the sequence from *E. coli* K-12 strain W3350(21) obtained by Schneider (45). In the 165-bp *icd* replacement segment, there are 20 polymorphisms, 19 of which involved synonymous substitutions (Fig. 7). The average proportion of nucleotide differences between pairs of strains was 3.9%, which is three times larger than the comparable value (1.3%) for the remaining 1,047-bp (bacterial) portion of *icd* ($P < 0.05$).

A total of 32 polymorphic sites were observed among the 10 strains in the 1,068-bp segment of phage 21 *int* gene extending from codon 25 through the end of gene (Fig. 7). The sequences of pairs of strains differed on average at 0.9% of nucleotide sites and 0.9% of amino acid positions. The estimated numbers (per 100 sites) of synonymous and nonsynonymous nucleotide substitutions ($d_S$ and $d_N$) (33, 34) were 2.63 ± 0.62 and 0.40 ± 0.13, respectively. The $d_S/d_N$ ratio is 6.6, which is much smaller than the comparable value of 107 for the *icd* gene of *E. coli* and for the average of 24 reported for enterobacterial genes in general (49).

**(iii) Natural occurrence of e14.** Previous studies have shown that the e14 element is present in some strains of *E. coli* K-12 but not in laboratory strains *E. coli* B/5 or *E. coli* C (8, 17, 27). In our study, 77 *E. coli* strains, almost all of which were recovered from natural populations, were screened for the presence of e14 by PCR amplification. Only two strains, CH734 and EC15, yielded a PCR product. The sequence of the segment amplified from CH734 is identical to that previously determined by Schneider (45) for the same strain. For strain EC15, the 3′ *icd* replacement sequence was identical to that we had earlier obtained for same strain, but the sequence beginning at nucleotide 124 downstream from *icd* and extending for a stretch of 500 bp showed only 54% identity with the e14 sequence of CH734. A search of GenBank failed to identify a homolog of the distinctive 500-bp EC15 sequence.

Although strain EC64 was negative when tested by PCR for the presence of e14, the fact that the sequence of the 3′ end of its *icd* gene has certain features in common with the sequence of strain CH734 (Fig. 2) suggests that it also harbors an e14-like element.

## DISCUSSION

For a collection of 33 strains of *E. coli* and *S. enterica*, we have shown that the mean pairwise strain difference in amino acid sequence of IDH is less than that of the average housekeeping gene. In the case of *S. enterica*, the amino acid diver-

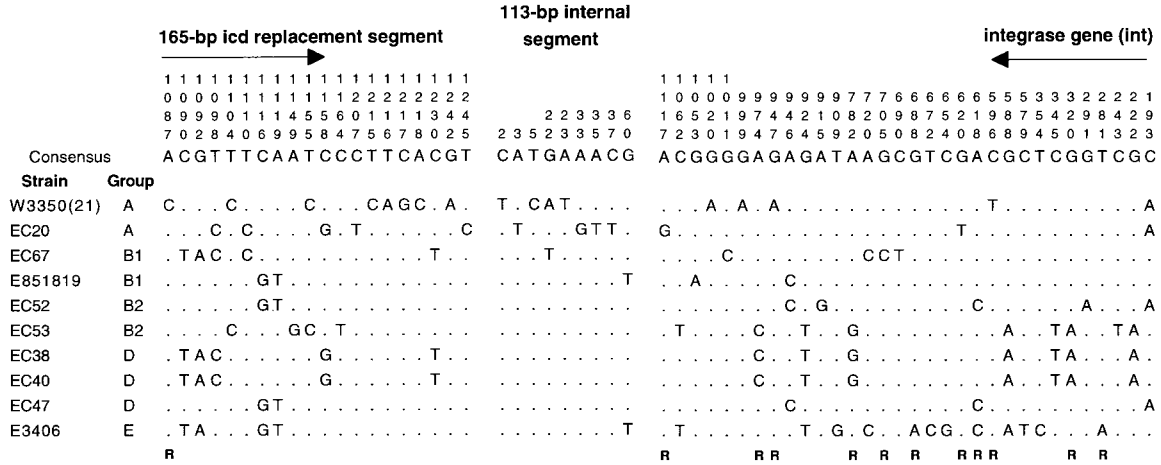


FIG. 7. Nucleotide sequence variation in a portion of the phage 21 genome including the last 165 bp of the *icd* gene (*icd* replacement segment), the 113-bp internal segment, and 1,068 bp of the integrase gene (*int*) among 10 strains of *E. coli*. Only the polymorphic sites are shown. Dots indicate identity with the consensus sequence. R indicates the occurrence of a substitution that results in an amino acid replacement. The sequence of the K-12 strain W3350(21) lysogen was obtained from Schneider (45). The other sequences are from the prophages of strains sequenced in the present study. The *int* gene of strain EC38 contained a 1.3-kb insertion sequence (not shown).

gence in IDH (0.5%) is less than half the mean level (1.2%) observed for housekeeping enzymes, whereas the mean nucleotide difference in *icd* (5.6%) is larger than the average value (4.7%) for housekeeping genes. Consequently, the estimated number of synonymous substitutions per synonymous sites ($d_S$) of *icd* greatly exceeds the number of nonsynonymous substitutions per nonsynonymous site ($d_N$), with a ratio of 117. This ratio is larger than those reported for other housekeeping genes and for the *inv* and *spa* invasion genes in *S. enterica* (Table 1). For *E. coli*, with exclusion of the 165-bp phage 21 integrating region of *icd*, the $d_S/d_N$ ratio is 107, which is about three times greater than the average value for housekeeping genes. Thus, it would seem that the *icd* locus has experienced unusually strong purifying selection against amino acid replacement. There were only 18 polymorphic amino acid codons among the 33 strains of *E. coli* and *S. enterica*, none of which is in the enzyme's active site. All binding sites of IDH with $Mg^{2+}$-isocitrate and $NADP^+$ were completely conserved within and between *E. coli* and *S. enterica*.

**Role of horizontal transfer and recombination.** Although several examples of horizontal transfer and recombination were identified by our analysis, these processes have not played a dominant role in the evolution of the *icd* gene. For the strains of *E. coli* studied, only a single putative case of horizontal transfer, involving a 33-bp segment, was identified. The *S. enterica* tree for *icd* (Fig. 5) is topographically similar to consensus trees based on other housekeeping genes and on invasion genes. However, subspecies IIIb is in an unusual position, being distant from I, II, and VI, with which it is normally associated, but what part of the gene has been recombined is not clear, and a donor has not been identified. It is noteworthy that subspecies IV and VII are relatively weakly differentiated in *icd* sequence. This is also the case with the *inv* and *spa* genes but not with other housekeeping genes or with data from MLEE analysis. Hence, the *icd* gene provides further evidence that the genome of subspecies VII is a megamosaic of segments with very different evolutionary histories (5, 48).

In the *icd* sequence, subspecies V is unusually divergent

TABLE 1. Sequence variation in 13 genes among 16 strains of *S. enterica*

| Gene | No. of bp sequenced | Mean pairwise value ($10^2$) for: | | $d_S/d_N$ ratio | Reference |
|---|---|---|---|---|---|
| | | $d_S$ | $d_N$ | | |
| Housekeeping | | | | | |
| *icd* | 1,164 | 29.35 ± 2.00 | 0.25 ± 0.09 | 117 | This study |
| *putP* | 1,467 | 16.70 ± 1.88 | 0.60 ± 0.23 | 28 | 35 |
| *mdh* | 849 | 20.13 ± 1.72 | 0.48 ± 0.16 | 42 | 4 |
| *gapA* | 924 | 15.15 ± 1.49 | 0.61 ± 0.15 | 25 | 38 |
| *gnd* | 1,335 | 21.80 ± 1.40 | 0.44 ± 0.10 | 49 | 37 |
| *aceK* | 1,719 | 28.39 ± 1.76 | 1.05 ± 0.14 | 27 | 39 |
| Invasion | | | | | |
| *invE* | 1,119 | 21.40 ± 1.33 | 0.50 ± 0.10 | 43 | 6 |
| *invA* | 1,950 | 22.87 ± 1.33 | 0.68 ± 0.12 | 34 | 6 |
| *spaM* | 444 | 21.50 ± 2.89 | 2.84 ± 0.48 | 8 | 6 |
| *spaN* | 1,011 | 25.41 ± 2.07 | 5.99 ± 0.48 | 4 | 6 |
| *spaO* | 909 | 24.30 ± 2.04 | 3.55 ± 0.41 | 7 | 28 |
| *spaP* | 672 | 19.70 ± 2.08 | 0.38 ± 0.15 | 52 | 28 |
| *spaQ* | 258 | 13.78 ± 2.78 | 0.25 ± 0.17 | 55 | 28 |

from all other subspecies, which strongly suggests that it received much or all of its *icd* gene from a source other than *S. enterica*. The situation is similar to that reported for *gapA* (38) and *gnd* (37), in each of which a segment appears to have been imported from *Klebsiella* sp. In any event, the evidence from *icd* analysis strengthens the case for recognizing strains of subspecies V as a distinct species, *S. bongori* (42).

In sum, with respect to synonymous sites, the *icd* gene has evolved at a rate similar to that of the average for other housekeeping genes and the invasion genes.

**Distribution and variation of phage 21.** Among the phage 21 sequences from various strains of *E. coli*, there has been some mutational divergence, but in the *int* gene and the internal region, it has been trivial. Additionally, there has been some modification in the core recognition sequence (11) of the phage and bacterial attachment site (nucleotides 1089 to 1093), such that two core sequences, differing at nucleotide sites 1090 and 1092 (Fig. 7), are present.

The occurrence of prophages of phage 21 of closely similar sequence (as judged by the *int* gene, the internal region, and, to lesser extent, the replacement region) in 30 of 77 strains of *E. coli*, representing all the major ECOR groups, may be interpreted as follows. It is possible that phage 21 is a newly evolved species or has only relatively recently acquired the capacity to infect *E. coli*, so that most of the *E. coli* lysogens represent recent primary infections. However, it is perhaps more likely that phage 21 is a long-standing parasite of *E. coli*, but that a strain that is insensitive to the immunity systems of earlier versions of phage 21 has replaced earlier resident prophages at the *icd* integration site in a high proportion of *E. coli* strains. This scenario is the frequency-dependent model of phage evolution proposed by Campbell et al. (10). Perhaps the new phage variant has in some cases displaced resident prophages downstream. It is also possible that a newly infecting phage could cause the excision of a resident prophage before itself integrating. The repressor system of the new prophage would prevent lytic activity or reintegration of the displaced prophage, which subsequently would be lost in the course of repeated cell division. There well may be some older versions of phage 21 in some strains that were not detected in our PCR survey because the sequence in the primer region (overlap region of the *int* and *xis* genes) has been deleted or has diverged in sequence. However, the important point is that essentially the same version of phage 21 has infected half of the extant lineages of *E. coli*.

**Evidence of phage 21 recombination.** For the 10 strains, separate neighbor-joining evolutionary trees based on variation in the 1,068-bp segment of *int* and in a stretch of sequence that includes the 165-bp *icd* replacement segment and 51 bp of the adjacent internal segment (Fig. 1) are shown in Fig. 8. If recombination among individual phages is absent or infrequent, trees for these two segments should be topologically similar, although rates of substitution may differ.

Comparison of the trees reveals a number of differences. Strains EC851819, EC47, EC52, and E3406, each representing a different ECOR group, cluster together in the tree for the 3′ *icd* segment (Fig. 8B), but the *int* sequence of E3406 is very different from those of all other strains (Fig. 8A). The *icd* sequence of EC53 shares no polymorphic site with those of EC38 or EC40, but the *int* sequence of EC53 is closely similar to those of EC38 or EC40. Furthermore, the *icd* segment of EC67 is like those of EC38 and EC40, but EC67 is strongly divergent from EC38 and EC40 in *int* sequence. These differences provide evidence of frequent recombination of phage 21 genomes. Among lambdoid phages in general, an important role for recombination in generating genomic diversity is well established (1, 10, 19).

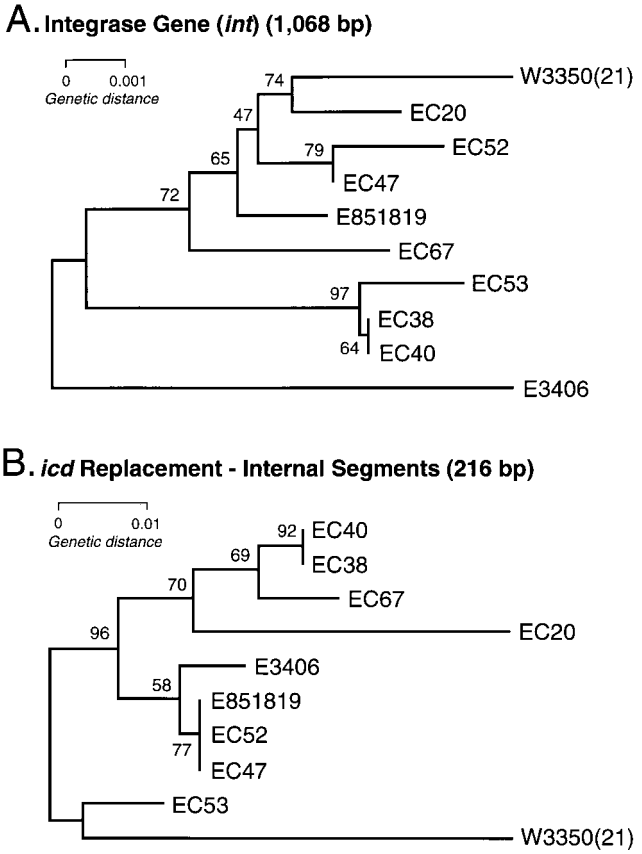


FIG. 8. Neighbor-joining trees based on the nucleotide sequences (all sites) of phage 21 in 10 lysogenic strains of *E. coli*. (A) Integrase gene. The phage 21 *int* sequence from strain W3350 is from Schneider (45); the other *int* sequences are from the prophages in the indicated ECOR and other *E. coli* strains. The *int* gene of the EC38 prophage contains a 1.3-kb insertion sequence. (B) A segment composed of the 165-bp replacement segment of the 3′ end of the *icd* gene and the proximal 51 bp of the 113-bp internal segment.

## ACKNOWLEDGMENT

This research was supported by grant AI22144 from the National Institutes of Health.

## REFERENCES

1. **Baker, J., R. Limberger, S. J. Schneider, and A. Campbell.** 1991. Recombination and modular exchange in the genes of new lambdoid phages. New Biol. **3:**297–308.
2. **Berlyn, M. K. B., K. B. Low, and K. E. Rudd.** 1996. Linkage map of *Escherichia coli* K-12, edition 9, p. 1715–1902. *In* F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology, 2nd ed. ASM Press, Washington, D.C.
3. **Borthwick, A. C., W. H. Holms, and H. G. Nimmo.** 1984. Amino acid sequence round the site of phosphorylation in isocitrate dehydrogenase from *Escherichia coli* ML308. FEBS Lett. **174:**112–115.
4. **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.
5. **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.
6. **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.

7. **Brody, H., and C. W. Hill.** 1988. Attachment site of the genetic element e14. J. Bacteriol. **170:**2040–2044.
8. **Brody, H., A. Greener, and C. W. Hill.** 1985. Excision and reintegration of the *Escherichia coli* K-12 chromosomal element e14. J. Bacteriol. **161:**1112–1117.
9. **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.
10. **Campbell, A., S. J. Schneider, and B. Song.** 1992. Lambdoid phages as elements of bacterial genomes (integrase/phage 21/*Escherichia coli* K-12/*icd* gene). Genetica **86:**259–267.
11. **Campbell, A. M.** 1992. Chromosomal insertion sites for phages and plasmids. J. Bacteriol. **174:**7495–7499.
12. **Cronan, J. E., Jr., and D. LaPorte.** 1996. Tricarboxylic acid cycle and glyoxylate bypass, p. 206–216. *In* F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology, 2nd ed. ASM Press, Washington, D.C.
13. **Dean, A. M., and G. B. Golding.** 1997. Protein engineering reveals ancient adaptive replacements in isocitrate dehydrogenase. Proc. Natl. Acad. Sci. USA **94:**3104–3109.
14. **Dean, A. M., and D. E. Koshland, Jr.** 1990. Electrostatic and steric contributions to regulation at the active site of isocitrate dehydrogenase. Science **249:**1044–1046.
15. **Dean, A. M., M. H. I. Lee, and D. E. Koshland, Jr.** 1989. Phosphorylation inactivates *Escherichia coli* isocitrate dehydrogenase by preventing isocitrate binding. J. Biol. Chem. **264:**20482–20486.
16. **Eyre-Walker, A., and M. Bulmer.** 1993. Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res. **21:**4599–4663.
17. **Greener, A., and C. W. Hill.** 1980. Identification of a novel genetic element in *Escherichia coli* K-12. J. Bacteriol. **144:**312–321.
18. **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.
19. **Highton, P. J., Y. Chang, and R. J. Myers.** 1990. Evidence for the exchange of segments between genomes during the evolution of lambdoid bacteriophages. Mol. Microbiol. **4:**1329–1340.
20. **Higuchi, R. G., and H. Ochman.** 1989. Production of single-stranded DNA templates by exonuclease digestion following the polymerase chain reaction. Nucleic Acids Res. **17:**5865.
21. **Hill, C. W., J. A. Gray, and H. Brody.** 1989. Use of the isocitrate dehydrogenase structural gene for attachment of e14 in *Escherichia coli* K-12. J. Bacteriol. **171:**4083–4084.
22. **Hurley, J. H., P. E. Thorsness, V. Ramalingam, N. H. Helmers, D. E. Koshland, Jr., and R. M. Stroud.** 1989. Structure of a bacterial enzyme regulated by phosphorylation, isocitrate dehydrogenase. Proc. Natl. Acad. Sci. USA **86:**8635–8639.
23. **Hurley, J. H., A. M. Dean, J. L. Sohl, D. E. Koshland, Jr., and R. M. Stroud.** 1990. Regulation of an enzyme by phosphorylation at the active site. Science **249:**1012–1016.
24. **Hurley, J. H., A. M. Dean, D. E. Koshland, Jr., and R. M. Stroud.** 1991. Catalytic mechanism of $NADP^+$-dependent isocitrate dehydrogenase: implications from the structures of magnesium-isocitrate and $NADP^+$ complexes. Biochemistry **30:**8671–8678.
25. **Ishiguro, N., and G. Sato.** 1988. Nucleotide sequence of insertion sequence IS*3411*, which flanks the citrate utilization determinant of transposon Tn*3411*. J. Bacteriol. **170:**1902–1906.
26. **Kumar, S., K. Tamura, and M. Nei.** 1993. MEGA: molecular evolutionary genetics analysis, version 1.0. Pennsylvania State University, University Park, Pa.
27. **Kutsukake, K., T. Nakao, and T. Iino.** 1985. A gene for DNA invertase and an invertible DNA in *Escherichia coli* K-12. Gene **34:**343–350.
28. **Li, J., H. Ochman, E. A. Groisman, E. F. Boyd, F. Solomon, K. Nelson, and R. K. Selander.** 1995. Relationships between evolutionary rate and cellular location among the Inv/Spa invasion proteins of *Salmonella enterica*. Proc. Natl. Acad. Sci. USA **92:**7252–7256.
29. **Matsutani, S., and E. Ohtsubo.** 1990. Complete sequence of IS*629*. Nucleic Acids Res. **18:**1899.
30. **McDonald, J. H., and M. Kreitman.** 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature **351:**652–654.
31. **Miyazaki, K., H. Eguchi, A. Yamagishi, T. Wakagi, and T. Oshima.** 1992. Molecular cloning of the isocitrate dehydrogenase gene of an extreme thermophile, *Thermus thermophilus* HB8. Appl. Environ. Microbiol. **58:**93–98.
32. **Muro-Pastor, M. I., and F. J. Florencio.** 1994. $NADP^+$-isocitrate dehydrogenase from the cyanobacterium *Anabaena* sp. strain PCC 7120: purification and characterization of the enzyme and cloning, sequencing, and disruption of the *icd* gene. J. Bacteriol. **176:**2718–2726.
33. **Nei, M., and T. Gojobori.** 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:**418–426.
34. **Nei, M., and L. Jin.** 1989. Variances of the average numbers of nucleotide substitutions within and between populations. Mol. Biol. Evol. **6:**290–300.
35. **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.
36. **Nelson, K., and R. K. Selander.** 1994a. Analysis of genetic variation by polymerase chain reaction-based nucleotide sequencing. Methods Enzymol. **235:**174–183.
37. **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.
38. **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.
39. **Nelson, K., F.-S. Wang, E. F. Boyd, and R. K. Selander.** Size and sequence polymorphism in the isocitrate dehydrogenase kinase/phosphatase gene (*aceK*) and flanking regions in *Salmonella enterica* and *Escherichia coli*. Genetics, in press.
40. **Ochman, H., and R. K. Selander.** 1984. Standard reference strains of *Escherichia coli* from natural populations. J. Bacteriol. **157:**690–693.
41. **Paton, A. W., and J. C. Paton.** 1994. Characterization of IS*1203*, an insertion sequence in *Escherichia coli* O111:$H^-$. Gene **150:**67–70.
42. **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.
43. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:**406–425.
44. **Sawyer, S. A., and D. L. Hartl.** 1992. Population genetics of polymorphism and divergence. Genetics **132:**1161–1176.
45. **Schneider, S. J.** 1992. Site-specific recombination of lambdoid phage 21 into the *icd* gene of *Escherichia coli*. Ph.D. dissertation. Department of Biological Sciences, Stanford University, Stanford, Calif.
46. **Selander, R. K., D. A. Caugant, and T. S. Whittam.** 1987. Genetic structure and variation in natural populations of *Escherichia coli*, p. 1625–1648. *In* F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology. ASM Press, Washington, D.C.
47. **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*, p. 17–49. *In* F. G. Priest, A. Ramos-Cormenzana, and B. J. Tindall (ed.), Bacterial diversity and systematics. Plenum Press, New York, N.Y.
48. **Selander, R. K., J. Li, and K. Nelson.** 1996. Evolutionary genetics of *Salmonella enterica*, p. 2691–2707. *In* F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd ed. ASM Press, Washington, D.C.
49. **Sharp, P. M.** 1991. Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. J. Mol. Evol. **33:**23–33.
50. **Stephens, J. C.** 1985. Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol. Biol. Evol. **2:**539–556.
51. **Thorsness, P. E., and D. E. Koshland, Jr.** 1987. Inactivation of isocitrate dehydrogenase by phosphorylation is mediated by the negative charge of the phosphate. J. Biol. Chem. **262:**10422–10425.
52. **Walsh, K., and D. E. Koshland, Jr.** 1984. Determination of flux through the branch point of two metabolic cycles: the tricarboxylic acid cycle and the glyoxylate shunt. J. Biol. Chem. **159:**9646–964.
53. **Whittam, T. S., and M. Nei.** 1991. Neutral mutation hypothesis test. Nature **354:**115–116.