Nonrandom Location of IS1 Elements in the Genomes of Natural Isolates of 
Escherichia coli

E. Fidelma Boyd and Daniel L. Hartl
Department of Organismic and Evolutionary Biology, Harvard University

We have studied the spatial distribution of IS1 elements in the genomes of natural isolates comprising the ECOR reference collection of Escherichia coli. We find evidence for nonrandomness at three levels. Many pairs of IS1 elements are in much closer physical proximity (<10 kb) than can be accounted for by chance. IS1 elements in close proximity were identified by long-range PCR amplification of the genomic sequence between them. Each amplified region was sequenced and its map location determined by database screening or DNA hybridization. Among the ECOR strains with at least two IS1 elements, 54% had one or more pairs of elements separated by <10 kb. We propose that this type of clustering is a result of “local hopping,” in which we assume that a significant proportion of transposition events leads to the insertion of a daughter IS element in the vicinity of the parental element. A second level of nonrandomness is found in strains with a modest number of IS1 elements that were mapped through the use of inverse PCR to amplify flanking genomic sequences: in these strains, the insertion sites tend to be clustered over a smaller region of chromosome than would be expected by chance. A third level of nonrandomness is observed in the composite distribution of IS elements across strains: among 20 mapped IS1 elements, none were found in the region of 48–77 minutes, a significant gap. One region of the E. coli chromosome, at 98 min, had a cluster of IS1 elements in seven ECOR strains of diverse phylogenetic origin. We deduce from sequence analysis that this pattern of distribution is a result of initial insertion in the most recent common ancestor of these strains and therefore not a hot spot of insertion. Analysis using long-range PCR with primers for IS2 and IS3 also yielded pairs of elements in close proximity, suggesting that these elements may also occasionally transpose by local hopping.

Introduction

The genomes of both prokaryotes and eukaryotes are often subject to dynamic changes brought about by transposable elements. Insertion sequence (IS) elements are transposable elements found in bacteria, ranging in size from about 1 to 2 kb (Galas and Chandler 1989). They were originally discovered during investigation of highly polar mutations in the galactose and lactose operons of Escherichia coli K12 (Malamy 1966; Jordan, Saedler, and Starlinger 1968; Galas and Chandler 1989). IS elements were subsequently identified in the genomes, plasmids, and bacteriophages of a wide range of bacterial genera and species (Ohtsubo and Sekine 1996). Natural isolates of E. coli are polymorphic for the presence and copy number of IS sequences (Ajioka and Hartl 1989). In strains of the ECOR reference collection, the average numbers of copies of ISI, IS2, and IS3 (chromosomal and plasmid copies together) are 7.2, 2.9, and 1.9, respectively. The largest variance in copy number was observed for ISI, for which 11 ECOR isolates had no genomic copies and 17 isolates had between 11 and 27 copies (Sawyer et al. 1987). The joint distribution of copy number among unrelated IS elements is consistent with their spread by horizontal transmission, most likely on plasmid vectors (Hartl and Sawyer 1988a, 1988b).

Previous studies of the map positions of IS elements in the laboratory strain E. coli K12 have yielded evidence of a nonrandom distribution of IS elements and an apparent clustered location of particular IS families on the chromosome (Birkenbihl and Vielmetter 1989). It was also observed that there were two regions showing a paucity of IS elements: one around the origin of replication (between 76 and 92 min) and another between 53 and 67 min (Birkenbihl and Vielmetter 1989; Deonier 1996). Escherichia coli K12 has a pedigree of extended laboratory subculture, and some of its derivatives contain chromosomal rearrangements (Deonier 1996). Hence, whether the apparent tendency of members of each IS family to be clustered within the chromosome of the host is a peculiar feature of strain K12 or a general feature of the population dynamics of IS elements in E. coli remains unknown.

The present study is the first determination of the map positions of a sample of IS1 elements in natural isolates of E. coli. To identify very closely linked pairs of IS elements in the ECOR strains (Ochman and Landler 1984), we used long-range PCR (Cheng et al. 1994) to amplify the genomic sequence between them. Most amplification products were <10 kb. The amplified products were sequenced and each assigned a map position by database screening or DNA hybridization with the Kohara lambda clones that cover the entire K12 chromosome (Kohara, Akiyama, and Isono 1987). A total of 27 ECOR strains yielded one or more amplification products with long-range PCR. We propose a model of transposition that can account for this type of IS clustering. We suggest a process of local hopping in which a proportion of the transposition events are assumed to occur locally, leading to the insertion of a daughter IS element in close proximity to the parental element. Long-range PCR analysis with reverse primers for IS2 and IS3 also gave evidence for local hopping of these elements.

The IS1 elements in five other ECOR strains were mapped through study of the flanking genomic region
amplified by inverse PCR. These elements were found to be clustered also: they tend to be present in a smaller region of the chromosome than a random distribution would allow. Still another level of clustering was found in the composite spatial distribution of 20 mapped IS1 elements: no elements were found in the 29% of the chromosome extending from 48 to 77 min.

Materials and Methods

Genomic DNA and PCR

Chromosomal DNA was extracted from all 72 ECOR isolates using G-Nome DNA isolation kits (Bio101, Vista, Calif.). Long-range PCR (Cheng et al. 1994) was carried out as described (Nurminsky and Hartl 1996) using oligonucleotides to the unique internal sequences of IS1, IS2, and IS3 that were directed outward (to amplify the inverted repeats and flanking genomic DNA). The primers were: IS1 IF, 5'-AGAAGCCACTGGAGCACC-3'; IS1 IR, 5'-CTGTGTGGTACGCCGTC-3'; IS2 IF, 5'-ATGTCTGGGATTAGGG-3'; IS2 IR, 5'-CTGTGTGGTACGCCGTC-3'; IS3 IF, 5'-GTCTGCTGATTTTGCTGT-3'; and IS3 IR, 5'-AAGGTTGCTGCTACGATA-3'. For each IS element, a combination of the F and R primers will amplify pairs of elements sufficiently close together and in either direct or inverted orientation. From the ECOR isolates were circularized for amplification by inverse PCR using the primers IS1 IF and IS1 IR. Amplified bands were separated by agarose gel electrophoresis, eluted from the gel, and purified using the Mermaid kit (Bio 101, Vista, Calif.).

Molecular Cloning and Sequencing

Long-range and inverse PCR products were cloned using TA cloning kit (Invitrogen, San Diego, Calif.). Products amplified by long-range PCR were 3' poly-A-tailed prior to cloning. Plasmid DNA was prepared by QiaGen column purification according to the manufacturer's instructions (QiaGen, Chatsworth, Calif.). The inserts were sequenced with reverse primers to determine the genomic sequence adjacent to each IS element. Sequencing of the PCR products was performed with an Applied Biosystems model 373A automated DNA sequencer.

Nucleotide Sequence Analysis

The IS1 flanking sequences obtained from each strain were identified by translating in all six possible reading frames and using the BLAST search to identify similarities in the databases (Altschul et al. 1995).

DNA Hybridization Analysis

The IS1 flanking sequences for which no similarity was obtained in the databases were further analyzed by DNA hybridization studies. IS1 flanking sequences for use as probes were labeled to high specific activity with [α-32P]dATP using Prime-It random primer labeling kit (Stratagene, La Jolla, Calif.). Probe hybridization to the nylon-bound DNA of the 476 Kohara set of ordered lambda clones (Kohara, Akiyama, and Isono 1987) was carried out in 29% formamide, 5 × SSPE, 0.1% sodium lauryl sulfate (SDS), salmon sperm DNA (0.1mg/ml), and Denhardt reagent at 42°C overnight. Three SSPE/SDS washes were done at 50°C.

Phylogenetic Analysis

The MEGA software was used to construct phylogenetic trees from IS1 flanking sequences (Kumar, Tamura, and Nei 1993).

Statistical Analysis

One test for clustering in genomic location is based on the principle that, if a natural isolate of E. coli contains n IS elements that are located at random around a chromosome of 100 min in length, the probability that each IS element is located at a distance of a min or more from its nearest neighbor, is given by

\[
P_L = \prod_{i=1}^{n-1} \left[ 1 - 2i \left( \frac{a}{100} \right) \right].
\]

This test is appropriate when pairs of IS elements are found that are located so close together that the genomic region between them can be amplified by long-range PCR.

A second test for clustering is based on a theorem in Feller (1966, p. 28, theorem 3) that deals with random splitting. This theorem yields the probability distribution of the size of the largest gap when a line segment is split into a number of subsegments of random length. In the present context, the theorem reads as follows. If a bacterial chromosome of 100 min is divided into n subintervals by n randomly placed IS elements, the probability that none of the subinterval exceeds a min in length equals

\[
\sum_{i=0}^{n} (-1)^i \binom{n}{i} \left( 1 - i \left( \frac{a}{100} \right) \right)^{n-1},
\]

where the subscript + means that the factor enclosed in the square brackets equals 0 unless it is positive. We designate the P value derived from this test as P15S because it tests the length of the longest unoccupied segment of chromosome. A significant value of P15S means that the IS elements are more clustered in one region of the chromosome than would be expected by chance.

A third test is due to Watson (1961). This test is for a different type of nonrandomness because, in the context of IS elements, the null hypothesis is that the map positions of IS elements were all chosen from an identical underlying continuous distribution. We designate the P value derived from Watson's test as PW. A significant value of PW means that the distribution of segment lengths between IS elements, taken as a group, is incompatible with their being chosen from the same underlying distribution.
Fig. 1.—Size distribution of long-range PCR fragments among 27 ECOR strains.

Results

Tightly Linked ISI Elements

Of the 72 ECOR isolates examined by long-range PCR using ISI primers, 27 gave one or more amplified bands. Each PCR band corresponds to flanking sequence between two ISI elements. Of the positive strains, 9 strains yielded only one band, 12 yielded two bands, 2 yielded three bands, 2 yielded four bands, and 2 yielded five bands. The total number of amplified bands was 57, ranging in size from approximately 0.5 to 23.5 kb, with a mean and standard deviation of 3.4 and 3.6 kb, respectively (fig. 1). About 75% of the bands were smaller than 5 kb, corresponding to a distance of about 0.1 min on the genetic map. The small distances between certain pairs of IS elements is statistically highly significant: 30% of the strains with two to eight copies of ISI had at least one pair of closely linked copies; with a random distribution of IS elements around the chromosome, the probability that a strain with as many as eight copies would have at least one pair within 0.1 min is only 5%.

There may be other closely linked pairs of ISI elements that were not identified by long-range PCR, either because the DNA fragment between them was too long or because it was too inefficiently amplified. Undoubtedly there is a bias toward the amplification of shorter fragments, so our detection of 57 closely linked pairs of closely linked elements in 27 strains is a minimum estimate. Although the largest amplified band was 23.5 kb, only one band was this large; all the others were smaller than 10 kb (fig. 1).

Our physical mapping method is based on comparison with E. coli K12. Genome size variation among ECOR groups has been shown for a number of strains. These size variations are believed to be mainly due to insertions and deletions without any significant alteration in the gene organization of the chromosome (H. Ochman, personal communication). Of the 27 strains scoring positive by long-range PCR, 14 were chosen for detailed mapping of the ISI elements by cloning and sequencing flanking regions. These 14 strains yielded 32 amplified bands, of which 17 were successfully mapped (table 1). The six unmapped sequences denoted as NP in tables 1 and 2 either are not present in E. coli K12 or are too divergent to hybridize with the phage clones from K12. Of the 17 pairs of ISI elements that could be assigned a location, 6 were present in plasmids. It should be noted that closely linked pairs of ISI elements from the same strain are not necessarily closely linked to each other (for example, in ECOR25, ECOR50, and

Table 1

<table>
<thead>
<tr>
<th>ECOR Strain</th>
<th>No. PCR</th>
<th>Locus Bands</th>
<th>Amino Acid Identity</th>
<th>Percent Identity</th>
<th>Map Positiona (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECOR6 .......</td>
<td>4</td>
<td>f276 rib sopA vacJ mglA</td>
<td>58/65 71/102 NA 57/59 65/71</td>
<td>89 70 NA 96 94</td>
<td>98 48</td>
</tr>
<tr>
<td>ECOR13 ......</td>
<td>4</td>
<td>f276</td>
<td>20/23 18/36 119/131</td>
<td>91</td>
<td>98</td>
</tr>
<tr>
<td>ECOR14 ......</td>
<td>1</td>
<td>IS1 and ompP</td>
<td>27/33 and 17/22</td>
<td>80 and 77</td>
<td>NP</td>
</tr>
<tr>
<td>ECOR17 ......</td>
<td>1</td>
<td>Plasmid</td>
<td>NA</td>
<td>NA</td>
<td>Plasmid</td>
</tr>
<tr>
<td>ECOR18 ......</td>
<td>2</td>
<td>NA</td>
<td>40/46</td>
<td>87</td>
<td>Plasmid</td>
</tr>
<tr>
<td>ECOR25 ......</td>
<td>2</td>
<td>f276 mglb</td>
<td>132/137 18/36</td>
<td>96 50</td>
<td>98 48</td>
</tr>
<tr>
<td>ECOR35 ......</td>
<td>3</td>
<td>f276 sopA</td>
<td>119/131 NA</td>
<td>91 91</td>
<td>98</td>
</tr>
<tr>
<td>ECOR36 ......</td>
<td>2</td>
<td>f276</td>
<td>132/137</td>
<td>96</td>
<td>98</td>
</tr>
<tr>
<td>ECOR37 ......</td>
<td>2</td>
<td>f276</td>
<td>132/137</td>
<td>96</td>
<td>98</td>
</tr>
<tr>
<td>ECOR49 ......</td>
<td>2</td>
<td>f276</td>
<td>118/131</td>
<td>90</td>
<td>98</td>
</tr>
<tr>
<td>ECOR50 ......</td>
<td>3</td>
<td>f276 ndh afa</td>
<td>125/139 50/71</td>
<td>90 70</td>
<td>27 83</td>
</tr>
<tr>
<td>ECOR55 ......</td>
<td>2</td>
<td>f56I narG and cheR</td>
<td>24/32 and 38/43</td>
<td>75 and 85</td>
<td>28/42</td>
</tr>
<tr>
<td>ECOR62 ......</td>
<td>2</td>
<td>vacI sopA</td>
<td>63/67 63/67</td>
<td>78 94</td>
<td>Plasmid</td>
</tr>
<tr>
<td>ECOR71 ......</td>
<td>2</td>
<td>sfa</td>
<td>NA</td>
<td>NA</td>
<td>NP</td>
</tr>
</tbody>
</table>

Note.—NA, not applicable; NP, not present in E. coli K12.

a Map position is based on E. coli K12 chromosome.
ECOR55). We did not find any instances in which three or more IS1 elements were all very tightly clustered.

Loosely Linked IS1 Elements

We also chose five ECOR strains (ECOR39, ECOR40, ECOR61, ECOR64, and ECOR72) that were negative by long-range PCR but contained multiple IS1 copies for analysis by inverse PCR (Ochman, Ayala, and Hartl 1993). The number of IS1 copies among these strains ranges from two to five, and none has plasmid-borne copies (Sawyer et al. 1987). Inverse PCR amplified 5, 5, 5, 2, and 3 bands respectively, from the strains. Each PCR band corresponds to an IS1 element and its adjacent sequence. These flanking sequences were also mapped by database similarity searches or by DNA hybridization to the Kohara clones (table 2). In these strains, all but one of the flanking sequences could be assigned a map position on the *E. coli* K12 chromosome.

The map positions of the IS1 elements in five ECOR strains are shown in figure 2. In each case, the $P_{LS}$ value is the probability that *n* IS elements would all lie within a region as small as or smaller than that observed, given random and independent locations in the genome. The $P_W$ value is the probability that the IS elements are nonrandomly distributed in the sense that their map positions do not conform to independent samples from the same distribution. Among these strains, the $P_{LS}$ value for ECOR39 is significant at the 1% level and the $P_W$ values for ECOR39 and ECOR61 are significant at the 5% level. Although none of the values for ECOR64 and ECOR72 are significant, the number of elements is so small that the tests are almost meaningless.

Because figure 2 includes multiple significance tests, the nominal significance levels should be adjusted, but there is still statistical significance overall. Given five independent tests of the "LS" type, the probability that one or more would be significant at the 1% level or less owing to chance alone is 0.049; given five independent tests of the "W" type, the probability that two or more would be significant at the 5% level or less owing to chance alone is 0.022. Evidently, the nonrandomness detected in the IS1 map positions in ECOR39 and ECOR61 cannot be attributed to chance.

The Composite Genetic Map

Figure 3 shows the locations of 20 IS1 elements mapped in 11 ECOR strains. Given a total of 20 randomly located IS elements, any gap greater than 27 min would be statistically significant at the 5% level. In the composite IS1 map, there is a slightly longer gap of 29 min, extending from 48 to 77 minutes, in which no IS1 elements were found. The length of this gap is significant ($P_{LS} = 0.03$). Four IS1 insertions had flanking sequence that mapped to the region 76–92 min, which was previously believed to lack IS elements based on the findings in *E. coli* K12.

The Watson test for nonrandomness in the composite IS distribution yields $P_W = 0.48$. The discrepancy between the two tests is expected because the null hypotheses differ. On the one hand, the longest unoccupied segment is too long to be accounted for by chance; on
Table 3

<table>
<thead>
<tr>
<th>Map Position</th>
<th>Locus</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>gef</td>
<td>Toxic killing protein</td>
</tr>
<tr>
<td>3.7</td>
<td>flhB</td>
<td>Transport of ferrichrome</td>
</tr>
<tr>
<td>6.5</td>
<td>papA</td>
<td>P-type fimbriae, pilus subunit</td>
</tr>
<tr>
<td>12</td>
<td>Bacteriophage P2</td>
<td>Tail fiber region</td>
</tr>
<tr>
<td>17.5</td>
<td>bioF</td>
<td>Biotin synthetase</td>
</tr>
<tr>
<td>22.2</td>
<td>serT</td>
<td>Serine tRNA</td>
</tr>
<tr>
<td>75.7</td>
<td>ndh</td>
<td>NADH dehydrogenase</td>
</tr>
<tr>
<td>27.5</td>
<td>narG</td>
<td>Nitrate reductase</td>
</tr>
<tr>
<td>35.4</td>
<td>relP</td>
<td>Regulation of RNA synthase</td>
</tr>
<tr>
<td>42.4</td>
<td>cheR</td>
<td>Chemotaxis</td>
</tr>
<tr>
<td>43</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>45</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>48.1</td>
<td>mglA</td>
<td>Methylgalactoside transporter</td>
</tr>
<tr>
<td>48.2</td>
<td>mglB</td>
<td>Usher protein</td>
</tr>
<tr>
<td>77</td>
<td>f400</td>
<td>43.8 kDa anonymous protein</td>
</tr>
<tr>
<td>78</td>
<td>NA</td>
<td>111 kDa anonymous protein</td>
</tr>
<tr>
<td>83</td>
<td>f561</td>
<td>58.9 kDa anonymous protein</td>
</tr>
<tr>
<td>89</td>
<td>f485</td>
<td>Phosphotransferase system</td>
</tr>
<tr>
<td>98</td>
<td>f276</td>
<td>Hexuronate degradation</td>
</tr>
<tr>
<td>NA</td>
<td>afa</td>
<td>Afimbrial adhesin</td>
</tr>
<tr>
<td>NA</td>
<td>sfa</td>
<td>S-type fimbriae, adhesin</td>
</tr>
<tr>
<td>NA</td>
<td>vacJ</td>
<td>28 kDa membrane protein</td>
</tr>
<tr>
<td>NA</td>
<td>ompP</td>
<td>Outer membrane protease</td>
</tr>
<tr>
<td>NA</td>
<td>HK3111</td>
<td>Methyltransferase</td>
</tr>
<tr>
<td>NA</td>
<td>rib</td>
<td>Ribulose-1-5-bisphosphate</td>
</tr>
<tr>
<td>NA</td>
<td>sopA</td>
<td>Plasmid protein A</td>
</tr>
</tbody>
</table>

NOTE.—NA, not applicable.

the other hand, the map positions of the IS elements seem to be independent samples from an identical underlying distribution. The unoccupied-segment test asks whether the entire chromosome is an equal target for IS insertion; whereas the Watson test asks whether, within the occupied segment of chromosome, the map positions are random. The latter is expected to be nonsignificant because the IS positions in figure 3 have been pooled across 11 unrelated strains.

Sequences Not Present in E. coli K12

Altogether, the sequence databases were searched for similarity to IS1 flanking sequences amplified from 19 ECOR isolates (14 subjected to amplification with long-range PCR and 5 subjected to amplification with inverse PCR). Of the total of 52 IS1 flanking sequences analyzed, 38 showed similarity in the databases. To assign map positions to these sequences, we used the data from E. coli K12 in table 3 (Berlyn et al. 1996). One flanking sequence from ECOR6 gave significant database similarity to ribulose-1-5-bisphosphate, a gene isolated from Thiobacillus denitrificans.

For flanking sequences that showed no similarity in the databases, we attempted to identify the map position by DNA hybridization with a set of ordered phage clones (Kohara, Akiyama, and Isono 1987). Among the 14 flanking sequences not locatable by database searching, 5 were assigned positions by this method. The remaining 9 flanking sequences (17% of the total) failed to hybridize, from which we infer that these sequences are not present in the genome of E. coli K12. The G+C content of these sequences was unusual. Most of the flanking sequences had a G+C content atypical of the E. coli K12 chromosome, with values ranging from 31% to 49% (mean 39%).

IS1 Elements at the Same Location in Different Strains

Several ECOR strains had IS1 elements that mapped to the same location on the chromosome. At map position 98 min, IS1 elements were clustered at locus f276 in seven strains (ECOR6, ECOR25, ECOR35, ECOR36, ECOR37, ECOR49, and ECOR50). These strains represent phylogenetically diverse subgroups (A, D, and E). In order to judge whether the IS elements at f276 were present in the common ancestor of these strains or derive from more recent horizontal transmission of a larger segment of chromosome, the flanking sequence was used to construct a neighbor-joining tree based on all sites (Saitou and Nei 1987). This tree is shown in figure 4. Since it is virtually identical to the trees based on the chromosomal gene for malate dehydrogenase (mdh, Boyd et al. 1994), we infer that the insertion near f276 was present in the common ancestor.

In addition to the f276 situation at 98 min, IS1 insertions near seven loci (vacJ, ompP, bioF, f561, cheR, f485, and relP) were each found in two or more E. coli strains. Three of these loci (vacJ, ompP, and f561) had identical IS1 insertion points in different strains, whereas two loci (bioF and relP) had unique IS1 insertion points in different isolates. The flanking sequence from one IS1 insertion, found in ECOR55, was similar along part of its length to cheR, located at 41 min, and along another part of its length to narG, located at 26 min. Similarly, in ECOR72, a flanking sequence was similar along one part of its length to cheR (41 min) and along a different part to rlpA (15 min).

Long-Range PCR with IS2 and IS3

Examination by long-range PCR using IS2 and IS3 primers with the 72 ECOR strains gave 26 and 27 positive strains, respectively. Among the strains positive with IS2 primers, 19 yielded one band and 7 yielded two bands. With the IS3 primers, 14 ECOR strains yielded one band, 12 yielded two bands, and 1 yielded three bands. The bands ranged in size from <1 kb to >6 kb. Seven strains (ECOR7, ECOR10, ECOR16, ECOR17, ECOR18, ECOR23, ECOR58) were analyzed by sequencing the long-range PCR products from the IS3 primers. The flanking sequence of the closely linked IS3 elements in ECOR7, ECOR10, and ECOR23 was
an IS1 element, whereas the flanking sequence of the closely linked IS3 elements in ECOR16 was another IS3 element (the only example we observed in which three IS elements were closely linked). Flanking sequence from ECOR58 showed similarity to pin, a DNA inverterase gene which maps to 26 min. The amplified flanking sequence from ECOR17 and ECOR18 showed no similarity to sequences in the databases.

Discussion

How can one account for the very close proximity of many pairs of IS1 elements? About 75% of the long PCR bands are <5 kb, or less than 0.1 min distant. Among the ECOR strains, 31 have 2–10 copies of IS1 and 17 have 11–27 copies (Sawyer et al. 1987). These are undoubtedly underestimates because some very closely linked copies would be present in the same restriction fragment and would, with DNA hybridization, be scored as a single copy. Nevertheless, if each of 31 strains had as many as 10 copies distributed at random, the probability that each would be at least 0.1 min from the next nearest copy is 0.913, and so the probability of such close linkage by chance is 0.087; we would therefore expect close linkage in only 3 strains among 31. The observed number is 9. The chi-square for goodness of fit is 13.7 and highly significant. Likewise, among strains with 11–27 copies, if each strain had as many as 27, the probability of no close linkage with a random distribution is 0.463, so the probability of at least one close linkage is 0.537. However, the probability that all of 17 strains would show at least one close linkage is $3 \times 10^{-5}$ and thus very unlikely to be due to chance.

We propose that the close linkage results from local hopping. In this model, each transposition event has a probability of resulting in a new insertion in very close proximity to the parental element. This probability need not be large to yield a significant number of close linkages. Although we do not have a direct estimate of the probability of local hopping, we can suggest an approximate figure based on the observed number of close linkages. Among approximately 500 IS1 elements that have been detected in the ECOR strains, we find 57 close linkages; hence, the proportion of elements with a closely linked partner is about 20%. This may be taken as a rough estimate of the probability of local hopping. It should considered a very tentative estimate. It fails to take into account the possible undercounting of total IS1 elements and the possible undercounting of closely linked IS1 elements. It also fails to distinguish between IS sites that are occupied by transposition within replicons versus those occupied by transposition between replicons (plasmid to chromosome or vice versa). Some transposable elements from flies and plants also show “local hopping” (reviewed in Craig 1997).

There is also a nonrandomness in the spatial distribution of more loosely linked elements (fig. 2). We chose to analyze ECOR strains having a modest number of IS1 elements so that all of them could be mapped. We also chose strains lacking plasmids so that the IS1 elements in each strain have a good chance of being descendants of a single horizontal transmission rather than coming from multiple independent events. There is no assurance that this assumption is correct, but it is consistent with the finding that the rate of horizontal transmission between strains seems to be small relative to the rate of transposition within strains (Sawyer et al. 1987).

The tendency for loosely linked IS1 elements to be present in a region of the chromosome smaller than that expected by chance suggests some spatial limitation even when transposition does not result in local hopping. What this limitation may be one can only speculate. Perhaps it has something to do with the nucleoid structure of the bacterial chromosome: maybe, at any one time, only portions of the chromosome are accessible to transposition.

There is a still more general level of nonrandomness apparent in the IS1 data. The composite distribution lacks insertions in the region between 48 and 77 min (fig. 3). It is of interest that the region between 76 and 92 min, in which IS elements are not present in E. coli K12, is not underrepresented in natural isolates. However, the gap between 48 and 77 min includes that between 53 and 67 min, in which no IS elements are found in E. coli K12 (Deonier 1996). What is the biological significance of this region? It contains neither the origin nor the terminus of replication. In gene content and density it seems to be quite unremarkable. Nevertheless, the gap suggests that, compared to other regions, the 53–67 interval may be less physically accessible to insertion by IS elements or perhaps have a lower density of target sites. A different sort of hypothesis is that IS insertions in this interval, perhaps because of the genes it contains, tend to be more detrimental to fitness than those in other intervals. Even a small difference in fitness would be sufficient to account for the gap.

The potential for IS transfer among genomes gives rise to the problem of inferring the source and ancestry of elements held in common. Comparative sequence analysis was used to determine the phylogenetic history of a region near 98 minutes sandwiched between a pair of closely linked IS1 elements in each of seven strains. The presence of these IS1 elements at the position in related lineages could be the consequence of either common ancestry or independent acquisition at a hot spot of insertion. The inferred phylogeny of the flanking sequences from the seven ECOR strains were congruent with the relationships determined from chromosomal genes, including mdh (Boyd et al. 1994), as might be expected if the acquisition of the IS elements occurred prior to the divergence of the strains (fig. 4). It is of interest to note that the region around 98 min has been identified, by analysis of base composition and codon usage bias, as a region that has been horizontally transferred into E. coli (Lawrence and Ochman 1996).

Certain other ECOR strains also share IS1 elements at identical locations, including insertions near the locus f56j in three isolates and near ompp, f485, and vacJ in two isolates each. In these cases, sequence analysis of the flanking region also supports common descent rather than independent insertions. In contrast, in two other
cases, insertions near bioF in ECOR39 and ECOR40 and near relF in ECOR61 and ECOR64, independent insertion of the IS1 elements is indicated by the fact that the exact insertion points differ.

In strain ECOR55, the long-range PCR product gave a region of sequence matching cheR (41 min) and a region matching narG (26 min). From ECOR72, an IS1 flanking sequence amplified by inverse PCR gave a region matching cheR and one matching rlpA (15 min). One possible explanation for such results is nonhomologous recombination at the cheR locus.

In a more general context, IS elements have the propensity to cause chromosomal rearrangements, generate mutations, and potentially contribute to the evolution of bacterial genomes, in particular, genome size evolution. We were unable to detect similarity in the databases or to assign map positions to 9 IS1 flanking sequences, indicating the absence of this DNA from E. coli K12. In a recent study of genome size variation, Bergsthorsson and Ochman (1995) showed that the genomes of natural isolates of E. coli may vary in size by up to one megabase. They identified genome size differences among the subgroup of the ECOR collection, with strains of subgroup A and D having the smallest and largest genomes, respectively. Laboratory strains, like E. coli K12 and its derivatives, were generally smaller than natural isolates, reflecting the fact that they were derived from ECOR subgroup A strains that have the smallest genomes among natural isolates (Bergthorsson and Ochman 1995). It is noteworthy that our analysis revealed a large number of IS1 flanking sequences that are not found in E. coli K12.

Base composition among bacterial species varies widely but is relatively uniform along the chromosome and conserved within and among related lineages. Sequences that result from recent horizontal transfer of DNA from a divergent species are often atypical of the genome as a whole in their G+C content or codon usage patterns (Lawrence and Ochman 1996; Ochman and Lawrence 1996). To investigate whether anonymous IS1 flanking regions are native to E. coli, their G+C contents were calculated to determine if they fell within the range of the E. coli chromosome. The flanking sequences usually had a low G+C content (mean 39%), giving additional support to the inference of horizontal transmission.

Médigue et al. (1991) classified E. coli genes into three main groups based on codon usage. Class III genes which have low codon bias consist of genes acquired by horizontal transfer and include genes corresponding to surface elements of the cell and genes coming from mobile elements. Analysis of the spectrum of chromosomal genes near which IS1 elements are located did not reveal any clear patterns (table 3). However, four of the loci (afa, pap, sfa, and vacA) are known virulence loci of E. coli pathogenic isolates, and two (afa and pap) were shown previously to have IS1 elements associated with them (Marklund et al. 1992; Garcia, Labigne, and Le Bouguenec 1994). Indeed, the afa gene cluster was demonstrated to translocate from a recombinant plasmid to the E. coli chromosome via IS1-specific recombination (Garcia, Labigne, and Le Bouguenec 1994). The afa, pap, and sfa operons are members of a group of adhesin loci, which includes csv, fae, fap, and fim, that are all involved in pilin biosynthesis in pathogenic E. coli isolates. Also, the loci csv, fae, fap, and fim have been previously shown to have IS1 elements associated with them (Klaasen and Graaf 1990; Bloomfield et al. 1991; De Haan et al. 1991; Marklund et al. 1992; Huisman et al. 1994). Whether this association is a general feature of IS1 or reflects the intense study of loci involved in virulence remains to be determined. Furthermore, a recent theory on operon evolution, the selfish operon model, dictates horizontal transfer as the driving force behind the evolution of gene clusters in bacteria, particularly for nonessential functions like those involved in virulence (Lawrence and Roth 1996).

Acknowledgments

We thank S. M. Rich for initial work, D. Nurminsky for suggestions and discussion, and J. Adams for calling Watson's goodness-of-fit test to our attention. This research was supported by grants from the National Institute of Health to D.L.H.

LITERATURE CITED


JULIAN P. ADAMS, reviewing editor
Accepted March 26, 1996