Enterobacterial Repetitive Intergenic Consensus (ERIC) Sequences in Escherichia coli: Evolution and Implications for ERIC-PCR

Lindsay A. Wilson¹ and Paul M. Sharp

Institute of Genetics, University of Nottingham, Queen’s Medical Centre, Nottingham, United Kingdom

Enterobacterial repetitive intergenic consensus (ERIC) sequences are 127-bp imperfect palindromes that occur in multiple copies in the genomes of enteric bacteria and vibrios. Here we investigate the distribution of these elements in the complete genome sequences of nine Escherichia coli (including Shigella species) strains. There is a significant tendency for copies to be adjacent to more highly expressed genes. There is considerable variation among strains with respect to the presence of an element in any particular intergenic region, but some copies appear to have been conserved since before the divergence of E. coli and Salmonella enterica. In comparisons of orthologous copies between these species, ERIC sequences are surprisingly conserved, implying that they have acquired some function, perhaps related to mRNA stability. The relationships among copies within E. coli are consistent with a master copy mode of generation. Insertion of new copies seems to occur at, and involve duplication of, the dinucleotide TA. Two classes of inserts of about 70 bp each occur at different specific sites within ERIC sequences; these inserts evolve independently of the ERIC sequences. The small number of ERIC sequences in E. coli genomes indicates that a widely used bacterial fingerprinting method using primers based on ERIC sequences (ERIC-PCR) does not rely on the presence of ERIC sequences.

Introduction

Bacterial genomes are generally considered to be streamlined, and yet numerous families of short (30–150 bp) interspersed repetitive sequences have been described in bacteria (Lupski and Weinstock 1992; Bachellier et al. 1996; Tobes and Ramos 2005). Little is known about the origins, evolution, mode of generation, or possible function of these elements. Most families are restricted to single species or very closely related species, while many other species appear to have no such elements. This suggests that if these repeats have any functions they have been acquired recently, may not apply to all members of the family, and are unlikely to concern fundamental aspects of bacterial growth, survival, and replication. Thus, while some repetitive sequences have been reported to act as binding sites for a variety of proteins, including DNA polymerase and DNA gyrase (Gilson, Perrin, and Hofnung 1990), this may be incidental. Most short bacterial repetitive sequences are imperfect palindromes, with the potential to form secondary structures, which may enhance mRNA stability (Newbury et al. 1987). Alternatively, most repetitive elements may be nonfunctional junk.

Enterobacterial repetitive intergenic consensus (ERIC) sequences, also described as intergenic repetitive units, differ from most other bacterial repeats in being distributed across a wider range of species. ERIC sequences were first described in Escherichia coli, Salmonella typhimurium (now Salmonella enterica serovar Typhimurium), and other members of the Enterobacteriaceae, as well as Vibrio cholerae (Sharples and Lloyd 1990; Hulton, Higgins, and Sharp 1991). The ERIC sequence is an imperfect palindrome of 127 bp (fig. 1). In addition, shorter sequences produced by internal deletions have also been described (Sharp and Leach 1996), as well as longer sequences due to insertions of about 70 bp at specific internal sites (Cromie, Collins, and Leach 1997; Sharp 1997). ERIC sequences have been found only in intergenic regions, apparently only within transcribed regions (Hulton, Higgins, and Sharp 1991). The number of copies of the ERIC sequence varies among species: it was initially estimated by extrapolation that there may be about 30 copies in E. coli K-12 and perhaps 150 in S. enterica Typhimurium LT2 (Hulton, Higgins, and Sharp 1991), while the genome sequence of Photobacterium luminescens has been reported to contain over 700 copies (Duchaud et al. 2003).

These copy number differences imply that orthologous intergenic regions may contain an ERIC sequence in one species but not in another, and this was found for comparisons between E. coli and S. enterica (Hulton, Higgins, and Sharp 1991). However, nothing is known about the nature of the mobility of these elements, such as the rate or means of generation of copies. It is also not clear whether any copies have a function. The extent of sequence similarity between copies in E. coli and V. cholerae implied either conservation or horizontal transfer, but in the one instance where orthologous copies were compared between E. coli and S. enterica, the sequence seemed to have accumulated substitutions at the neutral rate (Hulton, Higgins, and Sharp 1991). These observations could be reconciled if there were one or more “master copies” that are responsible for the generation of new copies, where the master copies are subject to selective constraint, but most new copies are effectively pseudogenes.

To date, the most extensively analyzed family of bacterial short repetitive sequences is that of the 30- to 40-bp REP/PU sequences found in E. coli, S. enterica, and their close relatives (Stern et al. 1984; Gilson et al. 1991). However, little is understood about the evolution of these elements (Bachellier, Clement, and Hofnung 1999). ERIC sequences may offer greater potential as an example for the study of the evolution of bacterial interspersed repetitive sequences because they are longer and thus more informative in comparative analyses and are found in a wider range of species. Here we begin to investigate the evolution of ERIC sequences, by characterizing all the copies of ERIC

¹ Present address: Department of Zoology, University of British Columbia, Vancouver, Canada.

Key words: ERIC sequences, Escherichia coli, repetitive sequences, master copy model, REP-PCR.

E-mail: paul@evol.not.ac.uk.

doi:10.1093/molbev/msj125
Advance Access publication March 13, 2006

© The Author 2006. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.

For permissions, please e-mail: journals.permissions@oxfordjournals.org
sequences in the genome sequences of *E. coli* strains, examining their distribution and conservation, using the closely related species *S. enterica* as an outgroup. ERIC sequences are also of interest because they have been used as the basis of a technique for fingerprinting bacterial genomes (Versalovic, Koeuth, and Lupski 1991). Polymerase chain reaction (PCR) primers were designed to amplify between copies of the ERIC sequence at nearby locations in the bacterial genome. This method was found to produce results in a very wide range of bacterial species (Versalovic, Koeuth, and Lupski 1991), which was interpreted as indicating that ERIC sequences occur throughout the bacterial kingdom (Lupski and Weinstock 1992). This ERIC-PCR approach has subsequently been very widely used to analyze a very broad range of species. This is surprising because we are not aware of any ERIC sequences characterized from species outside the Enterobacteriaceae and Vibrionaceae. Here we consider the implications of the distribution of ERIC sequences in *E. coli* concerning how ERIC-PCR works.

**Materials and Methods**

Fourteen genome sequences were examined (table 1). These include four sequences of *E. coli* strains and five of *Shigella* species. Despite their traditional classification into a separate genus, *Shigella* strains are not monophyletic and lie within the radiation of *E. coli* (Ochman et al. 1983; Pupo, Lan, and Reeves 2000; Escobar-Paramo et al. 2003), and so we consider them here as members of the latter species. In addition, five strains of *S. enterica* were used as outgroups.

The GenBank DNA sequence database was accessed using the ACNUC retrieval system (Gouy et al. 1985). Copies of the ERIC sequence were found using a combination of search programs, including Blast (Altschul et al. 1990), FASTA (Pearson and Lipman 1988), and specifically written software. In particular, genome sequences were searched exhaustively for matches above a certain threshold to ERIC sequences containing a single deletion at any position between sites 7 and 120. To evaluate the significance of the matches found, the same search was conducted against random sequences generated with the same length and dinucleotide content as the intergenic component of the *E. coli* K-12 genome. The search criteria chosen, on the basis of search results with real and random sequences, were a minimum of 55% identity to ERIC sequences containing a deletion of 1–77 bp.

Loci with a copy of the ERIC sequence are referred to by their map location (0–100 min) in the *E. coli* K-12 genome. Genomic rearrangements mean that these map locations are not the same for the other genomes considered, but the K-12 positions are used here for all strains to ease comparison. For loci in other *E. coli* strains that do not exist in K-12, the position was deduced from the nearest flanking genes with orthologues in the K-12 genome.

Sequences were aligned using ClustalV (Higgins, Bleasby, and Fuchs 1992). Comparisons of gene sequences used the method of Li, Wu, and Luo (1985) to estimate the number of, and number of differences at, fourfold degenerate sites at the third position in codons. Codon adaptation index (CAI; Sharp and Li 1987) values were calculated using the CODONS program (Lloyd and Sharp 1992). Phylogenetic relationships among copies of ERIC sequences were estimated by the maximum likelihood method implemented in DNAML in the PHYLIP package (Felsenstein 2004), searching with multiple randomized sequence input orders, and a range of transition/transversion ratios.

**Results**

The distribution and evolution of copies of the ERIC sequence were investigated in the genome sequences of nine strains of *E. coli*. In addition, the genome sequences of five strains of the closely related species, *S. enterica*, were used as an outgroup. To gauge the relationships among these strains, their extent of divergence was measured at fourfold degenerate sites (at third codon positions) in genes. A sample of 21 genes was used; these were those flanking regions occupied by an ERIC sequence in both *E. coli* and *S. enterica* (see below) and having low codon usage bias (here defined as CAI values less than 0.5). The pairs of strains of *E. coli* O157:H7, of *Shigella flexneri* serotype 2a and of *S. enterica* serovar Typhi (table 1), are each extremely closely related: across more than 3,300 fourfold degenerate sites, the pairs of O157 and Typhi strains were identical, while the two *S. flexneri* strains showed just two nucleotide differences. No differences were found between members of each pair with regard to the ERIC sequences investigated here, and so only one representative of each pair (strains EDL933, 301, and Ty2) is considered below. Among the seven remaining *E. coli* genomes, CFT073 was the most divergent, differing at 9% of sites; the other six strains differed by 4%–6%. At the same loci, the difference between *E. coli* and *S. enterica* was 46%, while the *S. enterica* strains differed by 3%–4%.

The approximate relationships among the *E. coli* strains, derived from this sample of 21 genes, are shown in figure 2. Note that *E. coli* strains have a clonal chromosome backbone, peppered with numerous short recombined...
regions (Milkman and McKane Bridges 1993), such that the evolutionary relationships of most genes conform to a consensus phylogeny, but for any particular gene, the positions of some strains may vary.

Copy Number in \textit{E. coli} K-12

Thirty copies of the ERIC sequence were found in the \textit{E. coli} K-12 genome, in 29 different intergenic regions (table 2). Twenty are (near) full length (124–128 bp), while 10 are shorter copies (19–99 bp) produced by deletions; the K-12 genome contains no copies of ERIC sequences containing inserts. Bachellier, Clement, and Hofnung (1999) reported 21 copies of ERIC in the \textit{E. coli} K-12 genome, including the 20 full-length copies plus the partial copy at 52.4 min.

The full-length copies exhibit 60%–88% identity to the original ERIC sequence (as in fig. 1). The copies at 12.9, 17.3, and 66.2 min are the most similar to the consensus, each differing at 15 nucleotides, and differing from each other at only 1–3 nucleotides. The shorter copies have 64%–84% identity to the original ERIC sequence across alignable sites and exhibit a range of internal deletions. Most appear to have undergone a single internal deletion of the kind described previously (Sharp and Leach 1996). In several cases (15.5, 47.5, 52.4, and 82.5 min), the deletion removes a region bounded by short (3–5 bp) repeats, consistent with palindrom-induced deletion (Sharp and Leach 1996). The 19-bp-long copy at 47.5 min in K-12 would not appear as a significant match in normal searches. However, a 55-bp copy was found in other strains, and alignment of this region revealed that K-12 has undergone a deletion of 325 bp by comparison with \textit{S. flexneri}, one end of which lies within the ERIC copy in \textit{S. flexneri}.

This raises the question of the ability to differentiate between highly degenerate copies of the ERIC sequence and random sequences, a problem that is exacerbated for copies that have undergone internal deletions. The best match across 127 bp with no gaps in the K-12 genome, after those in table 2, was 47% and occurred in the middle of a gene (\textit{ydaU}); this was followed by two matches of 46%, and many more in the region of 40%–45%, and so appeared to represent the best of the random matches.

We compared the results of our searches to those in 10 random sequences generated with the same approximate length (500 kb) and dinucleotide composition as the intergenic regions of the \textit{E. coli} K-12 genome. The best match across 127 bp with no gaps in the K-12 genome, after those in table 2, was 47% and occurred in the middle of a gene (\textit{ydaU}); this was followed by two matches of 46%, and many more in the region of 40%–45%, and so appeared to represent the best of the random matches.

We compared the results of our searches to those in 10 random sequences generated with the same approximate length (500 kb) and dinucleotide composition as the intergenic regions of the \textit{E. coli} K-12 genome. The best match across 127 bp with no gaps in the K-12 genome, after those in table 2, was 47% and occurred in the middle of a gene (\textit{ydaU}); this was followed by two matches of 46%, and many more in the region of 40%–45%, and so appeared to represent the best of the random matches.

Locations of ERIC Sequences in \textit{E. coli} K-12

The copies of the ERIC sequence lie within intergenic regions, and all appear to be within transcribed regions. ERIC sequences are imperfect palindromes, and so their
orientation can be described; we refer to that shown in figure 1 as ERIC and its complement as CIRE. Among the 30 copies in K-12, there are 15 ERIC and 15 CIRE sequences on the leading strand with respect to chromosome replication, as expected if orientation is random in this regard. Twenty-five copies lie between genes orientated in the same direction, 15% are convergent, and none between convergently transcribed genes, whereas in the genome as a whole, 70% of neighboring genes are not overlapping the ERIC sequence, is annotated for the gene with the higher value. Nevertheless, the CAI values range from 0.21 to 0.80, indicating a wide range of expression levels, and 15 ERIC copies are at loci where both of the flanking genes have CAI values less than 0.4. At the two loci where the flanking genes have substantially different CAI values (13.8 and 42.0 min), the ERIC sequence lies downstream of the gene with the higher value.

Some copies of the ERIC sequence appear to overlap flanking genes (table 2). In one case, this may be artefactual: for yajO (9.4 min), the GTG start codon annotated for K-12 leads to a 58-bp overlap, but a later (ATG) start codon, not overlapping the ERIC sequence, is annotated for the other genomes. However, at two other loci, the overlap appears to have resulted from insertion of an ERIC sequence within a gene. At 42.0 min, the 19-bp overlap with the start of the nptA gene is found in both E. coli and S. enterica, and the ERIC sequence covers all but 9 bp of the nptA-aspS intergenic region in both species. In Y. pestis, the most closely related outgroup available, the two genes overlap by 1 bp, while in Erwinia carotovora (another member of the Enterobacteriaceae), there is 1 bp between the genes.

### Table 2

**ERIC Sequences in Escherichia coli K-12 Genome**

<table>
<thead>
<tr>
<th>Map</th>
<th>Sitea</th>
<th>Lb</th>
<th>IDd</th>
<th>Location</th>
<th>Distance</th>
<th>CAl</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.8</td>
<td>accF-E-E-ldpA&gt;</td>
<td>127</td>
<td>78.7 (0)</td>
<td>127755–127881</td>
<td>167, 30</td>
<td>0.61 0.65</td>
</tr>
<tr>
<td>3.8</td>
<td>&lt;hemL-C-yaiQ&gt;</td>
<td>58</td>
<td>81.6 (3)</td>
<td>174946–175003</td>
<td>63, 103</td>
<td>0.44 0.27</td>
</tr>
<tr>
<td>4.1</td>
<td>rpsB-E-E-tsF&gt;</td>
<td>127</td>
<td>78.0 (0)</td>
<td>190613–190739</td>
<td>13, 117</td>
<td>0.77 0.77</td>
</tr>
<tr>
<td>5.5</td>
<td>yajO-E-E-ykff&gt;</td>
<td>127</td>
<td>86.6 (0)</td>
<td>253339–253465</td>
<td>177, 1</td>
<td>0.26 0.27</td>
</tr>
<tr>
<td>9.4</td>
<td>&lt;yajC-C-C-&lt;dxx&gt;</td>
<td>127</td>
<td>86.0 (0)</td>
<td>437374–437500</td>
<td>58, 38</td>
<td>0.30 0.39</td>
</tr>
<tr>
<td>12.9</td>
<td>cataC-C-YfeC&gt;</td>
<td>127</td>
<td>88.2 (0)</td>
<td>596203–596329</td>
<td>6, 24</td>
<td>0.26 0.45</td>
</tr>
<tr>
<td>13.8</td>
<td>abpC-C-C</td>
<td>127</td>
<td>74.8 (0)</td>
<td>638731–638857</td>
<td>118, 80</td>
<td>0.80 0.80</td>
</tr>
</tbody>
</table>

a. The genes flanking the ERIC sequence, with transcription direction indicated by the arrows. ERIC sequences are indicated by E or C if on the complementary strand.

b. The length of the ERIC sequence.

c. The percent identity to the original ERIC sequence, with the numbers of insertions/deletions in parentheses.

d. The location of the ERIC sequence in GenBank entry U00096.

Notes:

1. The distance in nucleotides between the flanking genes and the ERIC sequence; negative values indicate overlap.

2. The CAI values for the flanking genes.
Thus, there appears to have been an insertion of an ERIC sequence into the region immediately downstream of the plsX gene in the ancestor of E. coli and S. enterica, creating a C-terminal extension to the plsX-encoded fatty acid/phospholipid synthesis protein by comparison with Y. pestis and E. carotovora.

Distribution of ERIC Sequences Among E. coli Strains

While the seven E. coli genomes are closely related in terms of gene sequences, the other strains all differ substantially from K-12 with respect to their complement of ERIC sequences (table 3). An additional 14 intergenic regions with copies of the ERIC sequence were found among the other strains of E. coli, including two (at 3.2 and 48.1 min) where no insertion of ERIC sequences was found. This suggests that the ERIC sequences contain insertions similar to those reported from other loci in S. enterica Typhimurium LT2 (Cromie, Collins, and Leach 1997). The Shigella strains

Thus, there appears to have been an insertion of an ERIC sequence into the region immediately downstream of the ntpA start codon in the ancestor of E. coli and S. enterica, creating a new intergenic region and altering the NH-terminus of the ntpA-encoded deoxyadenosine triphosphate pyrophosphohydrolase. At 24.7 min, the ERIC sequence appears to have inserted within the 3' end of the plsX gene in the ancestor of E. coli and S. enterica, creating a C-terminal extension to the plsX-encoded fatty acid/phospholipid synthesis protein by comparison with Y. pestis and E. carotovora.

Distribution of ERIC Sequences Among E. coli Strains

While the seven E. coli genomes are closely related in terms of gene sequences, the other strains all differ substantially from K-12 with respect to their complement of ERIC sequences (table 3). An additional 14 intergenic regions with copies of the ERIC sequence were found among the other strains of E. coli, including two (at 3.2 and 48.1 min) where the ERIC sequences contain insertions similar to those reported from other loci in S. enterica Typhimurium LT2 (Cromie, Collins, and Leach 1997). The Shigella strains

Thus, there appears to have been an insertion of an ERIC sequence into the region immediately downstream of the ntpA start codon in the ancestor of E. coli and S. enterica, creating a new intergenic region and altering the NH-terminus of the ntpA-encoded deoxyadenosine triphosphate pyrophosphohydrolase. At 24.7 min, the ERIC sequence appears to have inserted within the 3' end of the plsX gene in the ancestor of E. coli and S. enterica, creating a C-terminal extension to the plsX-encoded fatty acid/phospholipid synthesis protein by comparison with Y. pestis and E. carotovora.

Distribution of ERIC Sequences Among E. coli Strains

While the seven E. coli genomes are closely related in terms of gene sequences, the other strains all differ substantially from K-12 with respect to their complement of ERIC sequences (table 3). An additional 14 intergenic regions with copies of the ERIC sequence were found among the other strains of E. coli, including two (at 3.2 and 48.1 min) where the ERIC sequences contain insertions similar to those reported from other loci in S. enterica Typhimurium LT2 (Cromie, Collins, and Leach 1997). The Shigella strains

Thus, there appears to have been an insertion of an ERIC sequence into the region immediately downstream of the ntpA start codon in the ancestor of E. coli and S. enterica, creating a new intergenic region and altering the NH-terminus of the ntpA-encoded deoxyadenosine triphosphate pyrophosphohydrolase. At 24.7 min, the ERIC sequence appears to have inserted within the 3' end of the plsX gene in the ancestor of E. coli and S. enterica, creating a C-terminal extension to the plsX-encoded fatty acid/phospholipid synthesis protein by comparison with Y. pestis and E. carotovora.

Distribution of ERIC Sequences Among E. coli Strains

While the seven E. coli genomes are closely related in terms of gene sequences, the other strains all differ substantially from K-12 with respect to their complement of ERIC sequences (table 3). An additional 14 intergenic regions with copies of the ERIC sequence were found among the other strains of E. coli, including two (at 3.2 and 48.1 min) where the ERIC sequences contain insertions similar to those reported from other loci in S. enterica Typhimurium LT2 (Cromie, Collins, and Leach 1997). The Shigella strains

Thus, there appears to have been an insertion of an ERIC sequence into the region immediately downstream of the ntpA start codon in the ancestor of E. coli and S. enterica, creating a new intergenic region and altering the NH-terminus of the ntpA-encoded deoxyadenosine triphosphate pyrophosphohydrolase. At 24.7 min, the ERIC sequence appears to have inserted within the 3' end of the plsX gene in the ancestor of E. coli and S. enterica, creating a C-terminal extension to the plsX-encoded fatty acid/phospholipid synthesis protein by comparison with Y. pestis and E. carotovora.
ERIC Sequences in Escherichia coli

A very recent deletion would lead to the presence of an ERIC sequence in six of the seven strains. At five loci, there is a full-length ERIC sequence in six strains but none in the seventh despite the presence of the orthologous intergenic region. In each case, the strain lacking a copy is CFT073; because this is the most divergent of the E. coli strains, this configuration may reflect insertion in the common ancestor of the other six E. coli strains, rather than a recent deletion. This may be the case at 38.1 min, where there is no ERIC sequence present in S. enterica, but an alignment of the E. coli strains shows a gap of 166 bp in CFT073, one end of which coincides with the end of the ERIC sequence in the other strains. This seems more likely to reflect a deletion, rather than an insertion of both an ERIC sequence and additional nucleotides. At two other sites (4.1 and 86.4 min), there is a full-length copy present in all four S. enterica genomes, and in both cases, the gap in CFT073 is approximately the same length as an ERIC sequence. However, while alignment on one side of the ERIC sequence is perfect, on the other side, there is a short region (10–20 bp) of poor alignment. At a fourth site (77.1 min), there is a full-length copy in Typhi, although not in the other S. enterica strains; the alignment has the same characteristics as at 4.1 and 86.4 min. Finally, at 42.0 min, there is an ERIC sequence with short internal deletions present in all four S. enterica strains. Alignment of the E. coli strains shows a gap of 125 bp in CFT073, of the same nature as seen due to ERIC insertion (fig. 3): the ERIC sequences in E. coli and S. enterica occur at precisely the same position and do not show unusually high divergence (see table 4, discussed below), and so it is unlikely that insertion of the ERIC sequence occurred independently in the two species. This suggests that at 42.0 min, there may have been a deletion in CFT073, which was a precise reversal of the insertion.

Relationships Among ERIC Sequence Copies

The phylogenetic relationship among the full-length copies was investigated. Where multiple E. coli strains have orthologous copies, those copies are very similar, differing
from K-12 at fewer than 5% of sites. Thus, only one representative was included in the analysis; arbitrarily, the K-12 copy was chosen, when available. Copies from representative was included in the analysis; arbitrarily, from K-12 at fewer than 5% of sites. Thus, only one

also included; for one locus (90.0 min), the copies from both LT2 and Typhi were included because they differ at 29% of sites.

The phylogeny must be taken with some caution due to the short length of the sequences compared; nevertheless, two interesting patterns emerged (fig. 4). First, the seven ERIC sequences and from 9% to 57% in genes. The variation among genes in levels of divergence can be attributed mainly to differences in selective constraint on codon usage bias (Sharp et al. 1989; Mira and Ochman 2002). In comparisons of an ERIC sequence with its flanking genes, only the copy at 90.0 min was more divergent; there the flanking genes are both very highly expressed ribosomal protein genes with strong codon usage bias (table 2).

The four most conserved ERIC sequences are all full-length copies, but so is the most divergent. The full-length copies at 4.1 and 13.8 min, with less than 10% difference between species, are striking. However, while these two copies are conserved with respect to sequence between E. coli and S. enterica, neither of them is conserved with respect to presence in all seven E. coli strains: the copy at 4.1 min is absent in CFT073, while that at 13.8 has an internal deletion in three strains, and is entirely missing in two others.

Also striking is the difference of only 4% between copies at 90.0 min in E. coli and S. enterica Typhi. As noted above, the copies at this locus in different S. enterica strains show remarkable divergence: the copies in LT2 and Choleraesuis are very similar, and the copies in Typhi and Paratyphi are identical, but LT2 and Typhi differ at 29% of sites. The region of unusual similarity between E. coli and Typhi, rather than the expected similarity between Typhi and LT2, does not extend far beyond the ERIC sequence and does not cover the entire rplA-rplJ intergenic region. If the E. coli-Typhi comparison reflects orthologous divergence, the copy in the LT2/Choleraesuis ancestor must have undergone exchange, either with an orthologous copy in a more distantly related species or with a nonorthologous copy. Alternatively, if the E. coli-LT2 comparison reflects orthologous divergence, there must have been horizontal transfer from E. coli to Typhi or Paratyphi (or their common ancestor) quite recently. This second explanation seems more likely. There is no copy of ERIC elsewhere in the LT2

<table>
<thead>
<tr>
<th>ERIC</th>
<th>Difference</th>
<th>Gene</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>0.06 (125)</td>
<td>rpsB</td>
<td>0.11 (119)</td>
</tr>
<tr>
<td>13.8</td>
<td>0.31 (157)</td>
<td>abpC</td>
<td>0.31 (88)</td>
</tr>
<tr>
<td>13.8</td>
<td>0.23 (66)</td>
<td>abpC</td>
<td>0.31 (88)</td>
</tr>
<tr>
<td>24.7</td>
<td>0.23 (99)</td>
<td>plnX</td>
<td>0.31 (195)</td>
</tr>
<tr>
<td>42.0</td>
<td>0.28 (112)</td>
<td>rnpA</td>
<td>0.42 (73)</td>
</tr>
<tr>
<td>49.2</td>
<td>0.31 (49)</td>
<td>yejK</td>
<td>0.54 (157)</td>
</tr>
<tr>
<td>57.2</td>
<td>0.32 (126)</td>
<td>rpsB</td>
<td>0.51 (229)</td>
</tr>
<tr>
<td>58.2</td>
<td>0.30 (61)</td>
<td>rsc</td>
<td>0.34 (105)</td>
</tr>
<tr>
<td>77.1</td>
<td>0.29 (126)</td>
<td>gntR</td>
<td>0.50 (197)</td>
</tr>
<tr>
<td>81.5</td>
<td>0.25 (68)</td>
<td>gscC</td>
<td>0.35 (37)</td>
</tr>
<tr>
<td>85.4</td>
<td>0.17 (109)</td>
<td>gppA</td>
<td>0.57 (259)</td>
</tr>
<tr>
<td>86.4</td>
<td>0.31 (126)</td>
<td>metR</td>
<td>0.40 (170)</td>
</tr>
<tr>
<td>90.0</td>
<td>0.28 (126)</td>
<td>rplA</td>
<td>0.25 (133)</td>
</tr>
<tr>
<td>90.0</td>
<td>0.04 (126)</td>
<td>rplA</td>
<td>0.25 (133)</td>
</tr>
<tr>
<td>94.6</td>
<td>0.10 (127)</td>
<td>orn</td>
<td>0.29 (85)</td>
</tr>
</tbody>
</table>

**Table 4**

**Comparison of Orthologous ERIC Sequences in Escherichia coli and Salmonella enterica**

<table>
<thead>
<tr>
<th>ERIC</th>
<th>Difference</th>
<th>Gene</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>0.06 (125)</td>
<td>rpsB</td>
<td>0.11 (119)</td>
</tr>
<tr>
<td>13.8</td>
<td>0.31 (157)</td>
<td>abpC</td>
<td>0.31 (88)</td>
</tr>
<tr>
<td>13.8</td>
<td>0.23 (66)</td>
<td>abpC</td>
<td>0.31 (88)</td>
</tr>
<tr>
<td>24.7</td>
<td>0.23 (99)</td>
<td>plnX</td>
<td>0.31 (195)</td>
</tr>
<tr>
<td>42.0</td>
<td>0.28 (112)</td>
<td>rnpA</td>
<td>0.42 (73)</td>
</tr>
<tr>
<td>49.2</td>
<td>0.31 (49)</td>
<td>yejK</td>
<td>0.54 (157)</td>
</tr>
<tr>
<td>57.2</td>
<td>0.32 (126)</td>
<td>rpsB</td>
<td>0.51 (229)</td>
</tr>
<tr>
<td>58.2</td>
<td>0.30 (61)</td>
<td>rsc</td>
<td>0.34 (105)</td>
</tr>
<tr>
<td>77.1</td>
<td>0.29 (126)</td>
<td>gntR</td>
<td>0.50 (197)</td>
</tr>
<tr>
<td>81.5</td>
<td>0.25 (68)</td>
<td>gscC</td>
<td>0.35 (37)</td>
</tr>
<tr>
<td>85.4</td>
<td>0.17 (109)</td>
<td>gppA</td>
<td>0.57 (259)</td>
</tr>
<tr>
<td>86.4</td>
<td>0.31 (126)</td>
<td>metR</td>
<td>0.40 (170)</td>
</tr>
<tr>
<td>90.0</td>
<td>0.28 (126)</td>
<td>rplA</td>
<td>0.25 (133)</td>
</tr>
<tr>
<td>90.0</td>
<td>0.04 (126)</td>
<td>rplA</td>
<td>0.25 (133)</td>
</tr>
<tr>
<td>94.6</td>
<td>0.10 (127)</td>
<td>orn</td>
<td>0.29 (85)</td>
</tr>
</tbody>
</table>

**Note.**—Unless otherwise indicated, all comparisons were between E. coli K-12 and S. enterica Typhimurium LT2. Difference indicates the fraction of different bases. For genes, the difference refers to fourfold degenerate sites. Values in parentheses are numbers of sites (bp) compared.

a E. coli K-12 and S. enterica Typhi Ty2.

b Shigella flexneri and S. enterica Typhimurium LT2.

**Sequence Conservation Between E. coli and S. enterica**

To investigate whether any copies of the ERIC sequence are subject to selective constraint, we compared the extent of sequence divergence between orthologous copies in E. coli and S. enterica to that at fourfold degenerate sites (at third codon positions) in their flanking genes (table 4). Surprisingly, the ERIC sequences showed less difference between species than the flanking genes: the average difference (weighted by length) at ERIC sequences was 20% compared to 42% at fourfold degenerate sites. These values are not corrected for multiple hits, and so they underestimating the disparity between the two figures. The extent of difference varied among loci, from 6% to 31% in ERIC sequences and from 9% to 57% in genes. The variation among genes in levels of divergence can be attributed mainly to differences in selective constraint on codon usage bias (Sharp et al. 1989; Mira and Ochman 2002). In comparisons of an ERIC sequence with its flanking genes, only the copy at 90.0 min was more divergent; there the flanking genes are rplA and rplJ, two very highly expressed ribosomal protein genes with strong codon usage bias (table 2). The four most conserved ERIC sequences are all full-length copies, but so is the most divergent. The full-length copies at 4.1 and 13.8 min, with less than 10% difference between species, are striking. However, while these two copies are conserved with respect to sequence between E. coli and S. enterica, neither of them is conserved with respect to presence in all seven E. coli strains: the copy at 4.1 min is absent in CFT073, while that at 13.8 has an internal deletion in three strains, and is entirely missing in two others.

Also striking is the difference of only 4% between copies at 90.0 min in E. coli and S. enterica Typhi. As noted above, the copies at this locus in different S. enterica strains show remarkable divergence: the copies in LT2 and Choleraesuis are very similar, and the copies in Typhi and Paratyphi are identical, but LT2 and Typhi differ at 29% of sites. The region of unusual similarity between E. coli and Typhi, rather than the expected similarity between Typhi and LT2, does not extend far beyond the ERIC sequence and does not cover the entire rplA-rplJ intergenic region. If the E. coli-Typhi comparison reflects orthologous divergence, the copy in the LT2/Choleraesuis ancestor must have undergone exchange, either with an orthologous copy in a more distantly related species or with a nonorthologous copy. Alternatively, if the E. coli-LT2 comparison reflects orthologous divergence, there must have been horizontal transfer from E. coli to Typhi or Paratyphi (or their common ancestor) quite recently. This second explanation seems more likely. There is no copy of ERIC elsewhere in the LT2...
FIG. 4.—Phylogenetic relationships among *Escherichia coli* copies of the ERIC sequence and their *Salmonella enterica* orthologues; only full-length copies are included. Each copy is denoted by strain abbreviation and map position (see table 3); ins shows those containing inserts. The tree was rooted on the original consensus sequence (ERIC). The scale bar indicates 0.1 substitutions per site.
genome that is very similar to that at 90.0 min, which could have been the source of an exchange event, while more distantly related species (Yersinia and Erwinia) do not have a copy in this intergenic region. This does not rule out the possibility of nonorthologous exchange involving horizontal transfer from another species, but the divergence between LT2 and E. coli is within the range of values seen at other loci, whereas that between Typhi and E. coli is exceptionally low (table 4).

ERIC Sequences in S. enterica and Inserts in ERIC Sequences

The S. enterica genomes contain approximately 100 copies of the ERIC sequence (data not shown); obviously, most occur at sites not occupied by a copy in E. coli. Although here we do not discuss either those sites or general aspects of ERIC evolution in S. enterica, there are two interesting features of S. enterica ERICs at sites orthologous to those occupied in E. coli. First, at 94.6 min, where there is one copy of ERIC in three of the E. coli strains and in S. enterica Typhi, there are two very similar copies (fig. 4) present in the three other S. enterica strains (table 3). This orn-yeS intergenic region contains multiple (2–4) tRNA genes, with the ERIC sequence lying between them. The extra ERIC sequence in the three S. enterica strains is part of a tandem duplication of 232 bp including a tRNA gene. Thus, while this represents a clear case where the origin of one ERIC copy can be traced to another in the same genome, the second copy was not generated by the normal insertion process.

Second, within the gntR-yhhW intergenic region at 77.1 min, where there is a simple ERIC sequence in six of the seven E. coli strains, the copy in S. enterica Typhi contains two separate inserts. One insert is between nucleotides 45 and 46 and, in comparison to the inserts described previously (Cromie, Collins, and Leach 1997), is most similar (91% identity) to that at 45–46 in the ERIC sequence from the S. enterica Typhimurium LT2 nirD-nirC region (previously termed “E2”). The second, at 86–87, is most similar (86% identity) to that at 86–87 in the ERIC sequence from the S. enterica Typhimurium LT2 cysB-topA region (“E3”); the two Typhi inserts differ from each other at 38% of sites.

Only two such inserts were found in the E. coli genomes. One, between nucleotides 45 and 46 within the ERIC sequence in the yadD-panC intergenic region (3.2 min; in Shigella sonnei, Shigella boydii, and E. coli O157), is most similar (91% identity) to E2. The other, at 86–87 within the ERIC sequence in the yetS-yeiT intergenic region (48.1 min; in S. boydii), is similar (81% identity) to E3 but even more similar (87% identity) to the second insert at 77.1 min in Typhi. Thus, in all four cases, the E. coli and Typhi inserts occur at the same location and with the same orientation within the ERIC sequence as the similar sequences previously described. However, the E. coli ERIC sequences flanking these inserts share less identity (64%–78%) to those in which the inserts were first described and much higher levels of similarity (up to 96% identity) to several copies without inserts at different locations in the E. coli genomes (fig. 4).

Discussion

In order to gain insights into the movement, possible role, and evolution of a family of short interspersed repetitive sequences in bacteria, we have examined the distribution and conservation of copies of ERIC sequences in complete genome sequences of seven strains of E. coli. The copy number was found to vary greatly among strains, and there appear to be examples of both recently inserted and recently deleted copies as well as copies that have recently undergone internal deletion. From comparison with the closely related species S. enterica, there are also apparently ancient copies that have been surprisingly well conserved during the approximately 100-Myr divergence (Ochman, Elwyn, and Moran 1999) of these two species.

Our interpretation of the distribution of ERIC sequence copies in E. coli and S. enterica relies on two previous observations concerning the evolutionary history of these species. First, there has apparently been little genetic exchange between E. coli and S. enterica. While there are many laterally transferred genes in gamma proteobacterial genomes, these appear to have originated from more distant species, and single copy genes with orthologues in many species of gamma proteobacteria do not appear to have been involved (Lerat et al. 2005). Also, there is little evidence of surprisingly similar gene sequences shared by E. coli and S. enterica (Mira and Ochman 2002). Second, in contrast, there has been frequent recombination among E. coli strains (Guttman and Dykhuizen 1994), although not so much as to disrupt the evidence of a predominantly clonally evolved chromosome backbone (Milkman and McKane Bridges 1993). Thus, among the sites with informative variation across E. coli strains in the presence or absence of copies, there are multiple discordant distributions (table 3) more likely due to recombination than to coincidental insertion/deletion events.

Although recombination between E. coli and S. enterica appears to have been rare, ERIC sequences seem to provide a surprising example of such an event. The copy at 90.0 min in S. enterica Typhi and Paratyphi is extremely similar to that in the E. coli strains, whereas that in S. enterica Typhimurium and Choleraesuis shows a more typical level of divergence from E. coli (table 4 and fig. 4). The anomalous similarity extends over a short fragment of the intergenic region little longer than the ERIC sequence, and the flanking genes rplA and rplJ encoding essential ribosomal proteins are not involved.

Generation of New Copies

Studies of short interspersed repetitive sequences in eukaryotic genomes have focussed on two possible models for the generation of new copies; it may be that a large number of copies are capable of providing the template for new copies (the “transposon model”), or there may be only one active element (the “master copy model”) (Deininger et al. 1992). The nature of the relationship among different copies of the ERIC sequence (fig. 4) seems consistent with a master copy mode of generation of new copies. That is, there is a group of 10 copies that are very similar to each other and are also the most similar to the original consensus sequence, even though only one of these 10 copies was
included in the compilation used to derive that consensus (Hulton, Higgins, and Sharp 1991). Most of these copies are present in only one or a few of the *E. coli* strains. Thus, these appear to represent recently inserted copies, generated from a common source. In contrast, the copies that clearly arose before the divergence of *E. coli* and *S. enterica* are widely dispersed across the ERIC sequence phylogeny and quite divergent from the group of putative recent copies.

Within the group of 10 copies, the tree is comblike with each branch coming from a single lineage leading to the copies at 12.9 and 66.2 min. This tree shape is as expected under the master copy model, although it has been shown that it can also arise when there are multiple active elements (Brookfield and Johnson 2006). It is not clear whether any one member of this group of closely related copies was the source for the others. From the tree, the element at 66.2 min would be the best candidate because it is present in all seven strains of *E. coli*, whereas the copy at 12.9 min is found only in K-12. However, the copy at 66.2 min is not present in *S. enterica*, where there would have to be a different master copy. Alternatively, the master copy may have been in another species and so not present in the tree. Several of the putative recent copies lie within genomic regions present in only one *E. coli* strain, which appear to be due to horizontal transfer. However, we have not found ERIC sequences within plasmids or obviously associated with phage genomes, which might provide the vehicle for such movement.

Examination of sites of presumed recent ERIC sequence insertion suggested that insertion occurs predominantly at, and involves duplication of, the dinucleotide TA. This is similar to the RUP sequence described in *Streptococcus pneumoniae* (Oggoni and Claverys 1999). RUP sequences are 107 bp long and bounded by 7-bp inverted repeats in which the terminal four bases (TATA) are the same as in the ERIC sequence; the sequence similarity does not extend beyond this. RUP is often found inserted within copies of the insertion sequence IS630-Spn1, but we have not found any association between ERIC sequences and IS elements. Thus, while it seems likely that the generation of new ERIC sequence copies is facilitated by some transacting factor, we do not know what that might be.

Insertions within ERIC sequences were originally reported from *Salmonella*, *Klebsiella*, and *Yersinia* (Cromie, Collins, and Leach 1997; Sharp 1997). Only two such inserts were found in the *E. coli* genomes, although three more were found in *S. enterica* orthologues of *E. coli* ERIC sequences (table 3). Three of these inserts resemble the previously described E2 insert in sequence and occur at the same location as E2 between nucleotides 45 and 46 of the ERIC sequence. The other two resemble the E3 insert in both sequence and location between nucleotides 86 and 87 of the ERIC sequence. The ERIC copy at 77.1 min in *S. enterica* Typhi contains both. The ERIC sequences flanking inserts of either type are not closely related, whereas those at 3.2 and 48.1 min in the *E. coli* strains, containing two different types of insert, are closely related to each other and to other members of the group of putative recent copies without inserts (fig. 4). These observations point to two different classes of inserts, targeted to two different specific sites within the ERIC sequence, but evolving independently of ERIC sequences.

Loss of Old Copies

Gain of ERIC sequences must be counterbalanced by loss, although the apparent wide variation in copy number among members of the Enterobacteriaceae indicates that this balance is not precise. *Buchnera* species seem to have no copies (data not shown), but have substantially reduced genomes of less than 700 kb, in contrast to the 4- to 5-Mbp genomes of other sequenced Enterobacteriaceae. *Escherichia coli* strains have many fewer copies of the ERIC sequence than the other species, including the closest relative among them, *S. enterica*. Thus, loss of copies appears to have outweighed gain of copies during the recent evolution of *E. coli*. Loss could occur by deletion, either precise deletion of the element or imprecise deletion of a larger region, or by decay due to accumulation of mutations making the element unrecognizable. The latter process would be slow: because *E. coli* and *S. enterica* genes differ at nearly 50% of fourfold degenerate sites (in genes with low codon bias), a copy inserted immediately prior to the divergence of these two species would be expected to differ from its progenitor at only about 25% of sites. In contrast, the palindromic nature of the ERIC sequence is expected to make copies susceptible to deletion (Sharp and Leach 1996).

There is evidence of recent deletion of complete copies. One example of a seemingly precise deletion of a complete ERIC sequence was found, but the other putative deletion events were less precise. Multiple copies with internal deletions, typically lacking the central 60–75 bp of the element, are present in the *E. coli* genomes (tables 2 and 3). Comparisons among strains indicate that there are partial copies due to deletion events during the recent divergence of the *E. coli* strains, as well as prior to the common ancestor of the *E. coli* strains, and even before the divergence of *E. coli* and *S. enterica*. Thus, while a partial deletion could be followed by another event removing the remainder of the ERIC sequence, there are cases where the presence of the partial copy has been conserved for a long time.

Based on a sample of gene sequences (fig. 2), the two most closely related strains examined here are *E. coli* O157 and *S. dysenteriae*; these two strains are also the most similar with regard to the presence/absence of ERIC sequence copies (table 2). The most divergent strain is *E. coli* CFT073, in terms of both gene sequences and ERIC distribution. This suggests that ERIC insertion/deletion events occur at a fairly steady rate, which can be quantified relative to the rate of nucleotide substitution. The total branch length of the tree for the seven *E. coli* strains based on fourfold degenerate sites (fig. 2) is about 21%. The distribution of ERIC sequences (table 3) implies at least 35 insertion/deletion events (including partial deletions) during the divergence of the seven strains. Thus, the rate of insertion/deletion has been approximately 1.7 events per 1% nucleotide substitution.

A Function for ERIC Sequences

Although numerous potential roles for interspersed repetitive sequences have been postulated, none are widely supported. A function for ERIC sequences might be evident from their locations within the genome and/or their conservation between species. We have found two features of the location of ERIC sequences. First, there
was an underrepresentation of ERIC sequences lying between convergently transcribed genes. Although this observation was not statistically significant, it is in sharp contrast to the situation reported for other short repeated intergenic sequences in other genomes. Tobes and Ramos (2005) reported diverse 31- to 60-bp species-specific imperfectly palindromic REP-like sequences in eight widely divergent bacterial species; these were found to occur between convergent genes 1.6–3.4 times more often than expected. Second, although ERIC sequences are found near genes with a wide range of expression levels (as assessed by codon usage bias), there was a strongly significant excess of highly expressed genes among their flanking sequences.

Where copies of the ERIC sequence were found at orthologous positions in E. coli and S. enterica, the sequences exhibited substantially less divergence than seen at fourfold degenerate sites in the flanking genes (table 4). The copies at different sites, while conserved between species, are not closely related to each other (fig. 4). This implies some constraint on the divergence of these ERIC sequences subsequent to insertion in the common ancestor of E. coli and S. enterica, as if they have acquired a function at that point. Indeed, it might be expected that copies that have not acquired some role would have been deleted during the 100-Myr divergence of these two species. Surprisingly, however, none of the four most conserved ERIC sequences is present in all the strains of E. coli examined here (table 3), suggesting that any function is not essential.

The presence of an ERIC sequence might enhance the expression of a flanking gene due to the palindrome providing a binding site for some protein or increasing the longevity of its mRNA. For example, it has been found that the presence of REP/PU sequences can stabilize the upstream mRNA (Newbury et al. 1987). It has been suggested that a recently inserted ERIC sequence in a promoter region might increase the expression of the ybtA gene in Yersinia enterocolitica (Anisimov et al. 2005). This proposed function seems consistent with the preponderance of ERIC sequences located near highly expressed genes in E. coli.

Bacterial Fingerprinting Using ERIC Sequence Primers (ERIC-PCR)

A method for distinguishing among bacterial strains using PCR primers derived from within ERIC sequences (Versalovic, Koeuth, and Lupski 1991) has been very widely used. The primers are designed so that amplification occurs between copies of the ERIC sequence; if the positions of copies vary among different strains, the amplification products provide each with a unique fingerprint when run on a gel. The results presented above (table 3) indicate that ERIC sequences do indeed exhibit intraspecific variation in their locations. However, two points indicate that the method is not working as initially envisaged.

First, it is clear that E. coli strains contain insufficient numbers of ERIC sequences, too widely spaced. The products amplified in ERIC-PCR are fragments in the size range of 0.5–5 kb. The 22-bp ERIC primers match sequences within the middle of the element, a region only present in full-length copies. Because there are only 20 full-length copies in the E. coli K-12 genome, the average distance between adjacent copies is about 230 kb. The closest copies are at 12.9 and 13.8 min, but even these are more than 42 kb apart (table 2). The other E. coli genomes examined here have even fewer full-length copies (table 3). ERIC-PCR was initially demonstrated using E. coli K-12 W3110 (Versalovic, Koeuth, and Lupski 1991), a strain thought to have been derived from the same source as E. coli K-12 MG1655 (the genome sequence used here) less than 50 years ago (Itoh et al. 1999), as well as eight other strains of E. coli (including Shigella species). ERIC-PCR has subsequently been used in several studies of E. coli diversity (Lipman et al. 1995; Manges, Dietrich, and Riley 2004; Jeong et al. 2005; Ramchandani et al. 2005). While each of these studies produced a range of fragments in the size range less than 10 kb, it is clear that the fragments amplified in ERIC-PCR cannot be due to primers hybridizing to ERIC sequences. In fact, amplification must occur despite extensive mismatch because there are no sequences in the E. coli K-12 genome with fewer than five differences from the 22-bp primer sequences, except within the defined copies of the ERIC sequence. This may explain why studies of the repeatability of ERIC-PCR for E. coli isolates have yielded poor results (Meacham et al. 2003).

Second, the method apparently works for many different species that do not appear to contain any copies of ERIC sequences within their genomes. Our searches have found evidence of ERIC sequences only within the genomes of Enterobacteriaceae and Vibrio species, representing just two families within the gamma Proteobacteria. However, in the original description of ERIC-PCR, genomic fragments were successfully amplified using ERIC-based primers from a much wider range of species, including other Proteobacteria, as well as members of other highly divergent bacterial phyla, such as Treponema, Deinococcus, Thermus, and even a member of the Archaea (Versalovic, Koeuth, and Lupski 1991). Subsequently, ERIC-PCR has been used for investigating a very wide range of other bacterial species and even eukaryotes, despite the fact that none of these genomes appear to contain copies of the ERIC sequence.

Others have recognized this second point, noting that for most species the ERIC-PCR primers are effectively working as the arbitrary primers do in randomly amplified polymorphic DNA methods (Gillings and Holley 1997; Niemann et al. 1999). For example, Niemann et al. (1999) used ERIC primers to fingerprint strains of Sinorhizobium meliloti, a member of the alpha Proteobacteria. Recognizing that this species was unlikely to contain copies of ERIC sequence, they determined the terminal sequences of some of the fragments that had been amplified: these fragments had similarity only to the primer and not beyond that. More recently, Wei et al. (2004) determined the sequences of genomic fragments amplified using ERIC-PCR primers from unidentified microbial strains within human fecal samples. Although they did not specifically note it, again these sequences showed no similarity to ERIC sequences outside the terminal regions hybridized by the primers.

The question arises whether ERIC-PCR primers are amplifying between copies of the ERIC sequence in any of the analyses of Enterobacterial genomes. While the E. coli genome does not contain enough copies of the element,
species in other genera, including \textit{Salmonella}, \textit{Yersinia}, \textit{Erwinia}, \textit{Photobacterium}, and \textit{Vibrio}, have a far higher copy number, with some copies sufficiently closely located that amplification between them is feasible. However, the similarity of the results obtained for organisms whose genomes do and do not contain ERIC sequences might suggest that the method is working in the same way in all species. Certainly, the conclusion that successful amplification using ERIC-PCR primers indicates that ERIC sequences are widespread among bacteria (Lupski and Weinstock 1992) is invalid.

**Acknowledgments**

We thank John Brookfield for discussion and Liz Bailes for assistance with the figures. L.A.W was supported by a Biotechnology and Biological Sciences Research Council studentship.

**Literature Cited**


Brookfield, J. F. Y., and L. J. Johnson. 2006. The evolution of mobile DNAs—when will transposons create phylogenies that look as if there is a master gene? Genetics (in press).


Jennifer Wernegreen, Associate Editor

Accepted March 6, 2006