

# Molecular Evolution and Mosaic Structure of $\alpha$ , $\beta$ , and $\gamma$ Intimins of Pathogenic *Escherichia coli*

Elizabeth A. McGraw, Jia Li, Robert K. Selander, and Thomas S. Whittam

Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University

Two types of pathogenic *Escherichia coli*, enteropathogenic *E. coli* (EPEC) and enterohemorrhagic *E. coli* (EHEC), cause diarrheal disease by disrupting the intestinal environment through the intimate attachment of the bacteria to the intestinal epithelium. This process is mediated by intimin, an outer membrane protein that is homologous to the invasins of pathogenic *Yersinia*. The intimin (*eae*) gene is part of a pathogenicity island, a 35-kb segment of DNA that has been acquired independently in different groups of pathogens. Nucleotide sequences of *eae* of three EPEC and four EHEC strains representing distinct clonal lineages revealed an exceptionally high level of divergence (15%) in the amino acid sequences of  $\alpha$ ,  $\beta$ , and  $\gamma$  intimin molecules, most of which is concentrated in the C-terminal region. The  $\gamma$  intimin sequences from *E. coli* strains with serotypes O157:H7, O55:H7, and O157:H- are virtually identical, supporting the hypothesis that these bacteria belong to a single clonal lineage. Sequences of  $\beta$  intimin of EPEC strains of serotypes O111:H2 and O128:H2 show substantial differences from  $\alpha$  and  $\gamma$  intimins, indicating that these strains have evolved independently. Strong nonrandom clustering of polymorphic sites indicates that the intimin genes are mosaics, suggesting that protein divergence has been accelerated by recombination and diversifying selection.

## Introduction

Although most *Escherichia coli* that inhabit the intestinal tracts of humans are harmless, there are certain strains that are pathogenic, causing a variety of diseases from cholera-like diarrhea to invasive dysentery (Nataro and Kaper 1998). To become pathogenic, these bacteria often have acquired virulence factors encoded on mobile genetic elements, such as plasmids and bacteriophages, or on distinct DNA segments, called pathogenicity islands, which are integrated into the bacterial chromosome (Hacker et al. 1997).

One factor implicated in the virulence of enteropathogenic *E. coli* (EPEC) is the expression of intimin, an outer membrane protein that mediates the intimate attachment of the bacterial cell to eukaryotic cells. EPEC strains occur throughout the world and are a major source of infant morbidity and mortality in developing countries (Nataro and Kaper 1998). In the course of an EPEC infection, the intimate attachment of bacterial cells damages the intestinal epithelium, disrupts the enteric environment, and results in severe diarrhea (Nataro and Kaper 1998). Volunteer studies with isogenic deletion mutants have demonstrated that intimin expression is required for the full virulence of EPEC (Donnenberg and Kaper 1992).

Intimin is encoded by the *eae* gene that is part of a 35-kb pathogenicity island designated LEE (locus of enterocyte effacement) (McDaniel et al. 1995). LEE occurs in a minority of *E. coli* strains and has integrated into different sites in the chromosome in different clonal lineages of *E. coli* (Wieler et al. 1997). In addition to intimin, LEE encodes about 40 different proteins, in-

cluding a contact-dependent secretion system and a number of secreted products (Elliot et al. 1998).

A second type of pathogenic *E. coli* also has acquired the intimin gene as part of its repertoire of virulence factors. These pathogens are the enterohemorrhagic *E. coli* (EHEC), which cause bloody diarrhea as well as the life-threatening complication called hemolytic uremic syndrome. EHEC strains such as serotype O157:H7 have been responsible for serious outbreaks of food-borne disease in the United States, Great Britain, and Japan (Armstrong, Hollingsworth, and Morris 1996).

The origin of the intimin gene and LEE island is unknown. Intimin shows homology to proteins found in other pathogens, including *Citrobacter rodentium* (Schauer and Falkow 1993; Schauer et al. 1995). It also has some similarity in amino acid sequence to the invasins of pathogenic *Yersinia* species (Isberg, Vorrhis, and Falkow 1987; Young et al. 1990; Simonet et al. 1996). Genes of the LEE island have a low GC content and an unusual codon usage pattern, suggesting that they are foreign DNA that has spread into the *E. coli* population.

Intimin appears to be a highly variable protein among the variety of *E. coli* serotypes (Agin and Wolfe 1997). There are at least five antigenic variants that have been identified (Abu-Bobie et al. 1998), and there is an association of specific intimin variants with human pathogens. There are two groups of EPEC that are recognized by distinct multilocus enzyme genotypes and the conservation of flagellar antigens (Whittam and McGraw 1996). EPEC 1 strains typically produce  $\alpha$  intimin, and EPEC 2 strains produce  $\beta$  intimin (Abu-Bobie et al. 1998). These EPEC strains are distinct from *E. coli* that produce  $\gamma$  intimin, including strains of O157:H7 and O55:H7 (Abu-Bobie et al. 1998). It is not known how these distinct lineages have each evolved a common mechanism for intimate adherence, nor is it clear to what extent virulence is affected by structural and regulatory differences between the genes and proteins of the pathogenicity island.

Key words: mosaic gene structure, recombination, pathogenicity islands, *eae*, intimin, enteropathogenic *E. coli*, *E. coli* O157:H7.

Address for correspondence and reprints: Thomas S. Whittam, IMEG, Department of Biology, 208 Mueller Laboratory, Pennsylvania State University, University Park, Pennsylvania 16802. E-mail: tsw1@psu.edu.

*Mol. Biol. Evol.* 16(1):12–22, 1999

© 1999 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

**Table 1**  
**Intimin Genes (*eae*) Sequenced from Human Pathogenic**  
***Escherichia coli***

Intimin	Strain	Locality (year)	Serotype	ET Group <sup>a</sup>
Int- $\alpha$ . . .	E2348/69 <sup>b</sup>	U.K. (1969)	O127:H6	EPEC 1
Int- $\beta$ . . .	DEC 11a	U.S.A. (1975)	O128:H2	EPEC 2
	DEC 12a	U.K. (1950)	O111:H2	EPEC 2
Int- $\gamma$ . . .	DEC 3a	U.S.A. (1985)	O157:H7	EHEC 1
	DEC 3f	Germany (1989)	O157:H-	EHEC 1
	DEC 5d	Sri Lanka (1965)	O55:H7	EHEC 1
	ECOR 37	U.S.A. (1972)	O--H-	EHEC 1

<sup>a</sup> Based on electrophoretic type (ET) defined by multilocus enzyme electrophoresis (Whittam et al. 1993; Whittam and McGraw 1996). EPEC refers to enteropathogenic *E. coli* that cause infant diarrhea, and EHEC refers to enterohemorrhagic *E. coli* that cause hemorrhagic colitis (Nataro and Kaper 1998).

<sup>b</sup> Originally sequenced from a cloned fragment of this strain by Jerse and Kaper (1991), and sequenced here after PCR amplification.

To begin to address these issues, we sequenced the intimin (*eae*) genes of three EPEC strains and four EHEC strains that, based on multilocus enzyme electrophoresis, are only distantly related to the model organism (strain E2348/69) of EPEC infection, which produces  $\alpha$  intimin. Sequence comparisons support the hypothesis that the O157:H7 serotype very recently evolved from an EPEC-like ancestor and indicate that divergent intimins are encoded by mosaic alleles composed of DNA segments with different evolutionary histories.

## Materials and Methods

### Bacterial Strains

We sequenced the intimin genes of seven strains of *E. coli* (table 1). Six of these strains were originally isolated from patients with diarrheal disease, and one (ECOR37) was originally isolated from a marmoset (Ochman and Selander 1984). Three EPEC strains were from cases of infantile diarrhea and produced either  $\alpha$  or  $\beta$  intimin, depending on whether they belong to the EPEC 1 or EPEC 2 clonal group. We also examined four strains that belong to the EHEC O157:H7 clone complex (Feng et al. 1998), all of which are shown here to have an Int- $\gamma$  gene. The Int- $\gamma$ -producing strains include DEC 3a, a typical O157:H7 strain, DEC 3f, a non-motile O157 strain, and DEC 5d, an O55:H7 strain that represents the immediate ancestor to O157:H7. In addition, we included ECOR37, a natural *E. coli* isolate that has a LEE island (Berghthorsson and Ochman 1998) and is a member of the O157:H7 clone complex based on the results of multilocus enzyme electrophoresis (unpublished data).

### PCR Gene Amplification and DNA Sequencing

Primers for PCR were designed based on GenBank sequences, and internal sequencing primers were constructed based on generated sequence data. The PCR primers, labeled by the base pair locations relative to the start codon and as forward (F) or reverse (R) primers, are as follows: -30F, 5'-CATTCTAACTCA-TTGTGGTGGAGC-3'; 2206F, 5'-TGAGTGGAG-TAGGTACAG-3'; 2470F, 5'-CCACATCTGGTGA-

TAAGC-3'; 2645R, 5' -ATATTTATTTGCAGCCCC CCA-3'; 2660R, 5'-ATATTTATTTGCAGCCCCA-3'; 2983R, 5'-GGTATACAGCGTGGTTGGA-3'; and 3116R, 5'-GTCTACGACGCAATTGATCC-3'.

A single primer pair (-30F, 3116R) amplified the entire coding region of the Int- $\alpha$  gene. For the Int- $\beta$  gene, separate pairs of PCR primers were used to amplify the conserved region (-30F, 2660R) and the variable 3' end (2206F, 3116R) of the gene. For the Int- $\gamma$  gene, separate sets of PCR primers were also used for the central region (-30F, 2645R) and the 3' variable end (2470F, 2983R). Primers were synthesized either by Oligos Inc., the Penn State Biotechnology Center, or by a Beckman 1000 oligonucleotide synthesizer.

For PCR amplification, genomic DNA was prepared by standard protocols (Wilson 1990). Template for cycle sequencing was obtained through amplification for 30 cycles as follows: 94°C for 1 min, 50–55°C (depending on the strain) for 2 min, and 72 °C for 3 min, with an initial denaturing step at 94°C for 5 min. PCR products were purified with Qiaquick spin columns (Qiagen) and suspended in 10% TE buffer to suitable concentration, as determined by agarose gel electrophoresis.

Cycle sequencing was performed with a Prism Ready Reaction DyeDeoxy Terminator Cycle Sequencing kit (Applied Biosystems). Comparative allele sequencing was conducted on the ABI 373A automated sequencer (Bisercic, Feutrier, and Reeves 1991). Raw sequences of both DNA strands were analyzed and concatenated by DNASTAR (Madison). All conflicting and putative polymorphic nucleotide sites were sequenced at least three times to correct sequencing errors. All sequences were deposited in GenBank, Accession numbers AF081182–AF081187.

### Sequence Alignment

The published  $\alpha$  intimin (*eae*) sequence of E2348/69 (Jerse et al. 1990) and the *eae* homolog of *C. rodentium* (Schauer and Falkow 1993) were included in the sequence analysis. Three invasin sequences from *Yersinia pseudotuberculosis* (Isberg, Vorhies, and Falkow 1987), *Yersinia enterocolitica* (Young et al. 1990), and *Yersinia pestis* (Simonet et al. 1996) were also used in the phylogenetic analysis. The invasin gene of *Y. pestis*, which is interrupted by insertion of a 708-bp IS200-like element (Simonet et al. 1996), was aligned after removal of the IS-like sequence.

Multiple alignments were performed on the inferred amino acid sequences using the CLUSTAL method (Higgins and Sharp 1988) with a PAM250 residue weight table. Phylogenetic trees were constructed with MEGA (Kumar, Tamura, and Nei 1993) and PHYLIP 3.57c (Felsenstein 1989). Rates (per site) of synonymous and nonsynonymous substitution were calculated by the method of Nei and Gojobori (1986).

The molecular-clock hypothesis was tested by Tajima's (1993) chi-square method, a test that is appropriate in cases in which the pattern of nucleotide substitution is unknown. To test whether the rate of divergence of two sequences was uniform, we tabulated the number

of sites ( $m_1$ ) in which nucleotides in sequence 1 were different from those in sequence 2 and an outgroup sequence, and we similarly tabulated the number of sites with different nucleotides for sequence 2 ( $m_2$ ) and the outgroup ( $m_3$ ). Significant departures from a uniform substitution rate were detected by comparing the test statistic  $X^2 = (m_1 - m_2)^2 / (m_1 + m_2)$  with a theoretical chi-square distribution with one degree of freedom (Tajima 1993). We also tested the molecular-clock hypothesis by constructing maximum-likelihood trees with and without the clocklike assumption. The trees were produced by the PHYLIP programs DNAML and DNAMLK, and the log-likelihood values were compared with a chi-square distribution with  $(n - 2)$  degrees of freedom, where  $n$  is the number of sequences compared (Felsenstein 1989).

### Analysis of Mosaic Structure

To detect the clustering of polymorphic sites, we used the Stephens (1985) method in which sites are classified by how they partition the sequences into groups. The statistics  $d_0$  (the maximum distance in base pairs between two sites supporting a given partition) and  $g_0$  (the length in base pairs of the longest run of adjacent sites supporting a partition) were used to test the hypothesis that the sites that support a given partition are clustered along the length of the sequence. The test was also applied to all polymorphic sites, without reference to partitions, to determine whether clustering results from variation in the mutation rate along the gene. To identify the ends of segments of a mosaic allele, we developed a computer program called MAXCHI that implements the maximum chi-square method of Maynard Smith (1992). The program compared each sequence with a reference sequence and found the point  $k_{MAX}$  at which the chi-square statistic achieved a maximum. The sequence was then divided into two segments, and a new maximum was found within each segment. This cycle was repeated four times so that 16 maxima were identified. The significance of the  $k_{MAX}$  values for the nested segments were tested by a Monte Carlo procedure (Maynard Smith 1992) in which sites were placed randomly along the sequence 1,000 times and the null distribution of  $k_{MAX}$  was tabulated. Observed  $k_{MAX}$  values that exceeded values in the 5% tail of the null distribution were considered significant.

## Results

### Sequencing Results

Using the specific PCR primers, we amplified and sequenced the intimin gene (*eae*) from DEC 3a and DEC 3f. The sequences were identical to the published Int- $\gamma$  sequence from *E. coli* O157:H7 (Beebakhee et al. 1992; Yu and Kaper 1992) once the errors in the original sequences were identified and corrected (see appendix). There were only two base differences between DEC 5d and the published Int- $\gamma$ , both of which result in single amino acid replacements (R912S and G930R). The *eae* sequence of ECOR 37 was nearly identical to that of DEC 5d, with only one base difference, also responsible

for an amino acid change (T923I). In addition, we sequenced a 3,146-bp PCR fragment from strain E2358/69 (Int- $\alpha$ ) and found that it was identical to the *eae* gene cloned and sequenced from the same strain (Jerse and Kaper 1991).

We sequenced *eae* from two epidemiologically unrelated EPEC 2 strains (DEC 11a and DEC 12a) that were originally collected 15 years apart (table 1). These strains were predicted to have Int- $\beta$ , because they belong to EPEC 2, a group of bacteria in which  $\beta$  intimin has been detected both antigenically and by PCR assays (Abu-Bobie et al. 1998). The *eae* sequences of DEC 11a and DEC 12a were identical to each other but showed many nucleotide differences from the sequences encoding Int- $\alpha$  and Int- $\gamma$ , particularly in the C-terminus of the predicted protein.

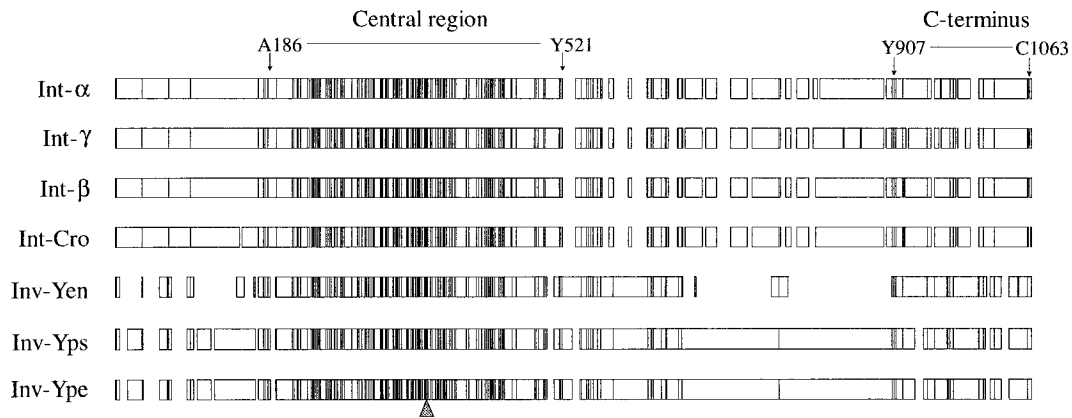
### Comparison of Intimin and Invasin

To determine the evolutionary relationships of  $\alpha$ ,  $\beta$ , and  $\gamma$  intimins, the nucleotide sequences were aligned at the amino acid level with homologous sequences from GenBank. The multiple alignment included three intimins sequenced here from human pathogenic *E. coli* (Int- $\alpha$ , Int- $\beta$ , and Int- $\gamma$ ), an intimin homolog from *C. rodentium* (Int-Cro), and three invasins from the pathogenic *Yersinia* species. The alignment introduced many gaps, particularly in the ends of the sequences, and resulted in a total aligned length of 1,064 amino acid positions (fig. 1, top).

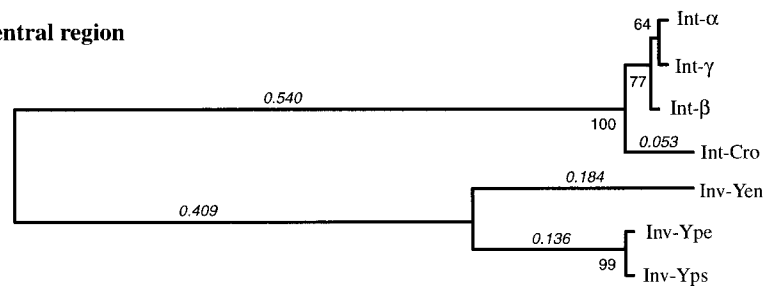
There are two parts of the intimin and invasins molecules that are relatively conserved in amino acid sequence. The first region is a conserved central region located between alanine 186 and tyrosine 524. Within this region, there are four alignment gaps covering seven codon positions, and 34% of the 328 amino acids are conserved among the intimins and invasins. The second region is the C-terminus of the protein and includes 157 amino acid positions between a conserved tyrosine (Y907) and cysteine (C1063). This region has 13 alignment gaps covering 32 codon positions, and only 11% of 125 amino acids are conserved (fig. 1).

The phylogenetic trees for the two regions differ in both their topologies and branch lengths (fig. 1). The trees were constructed by the NJ method with the gamma distance to account for rate variation among amino acid residues (Gu and Zhang 1997). The estimate of the shape parameter ( $\alpha$ ) of the gamma distribution was different for the two regions ( $\alpha = 1.36$  for the central region and  $\alpha = 4.45$  for the C-terminus). These  $\alpha$  values suggest that there is a higher level of among-sites variation in the substitution rate in the central region than there is in the carboxyl end of the protein (Gu and Zhang 1997).

The phylogeny based on variation in the conserved region shows greater divergence among invasins than among intimins. The tree predicts an order of branching for the intimin genes in which Int-Cro splits first, followed by Int- $\beta$ , followed by the divergence of Int- $\alpha$  and Int- $\gamma$ . Neither of the interior nodes is strongly supported by the bootstrap values, although each node is found in a majority of the bootstrap replicates (fig. 1, bottom).



### A. Central region



### B. C-terminus

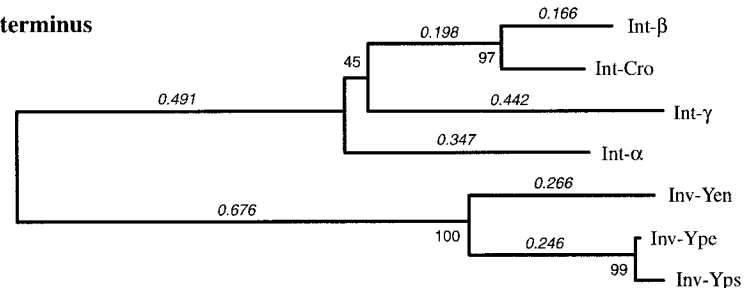


FIG. 1.—*Top*, Multiple-sequence alignment of  $\alpha$ ,  $\beta$ , and  $\gamma$  intimins of pathogenic *E. coli*, an intimin homolog from *C. rodentium* (Cro), and invasins from *Y. enterocolitica* (Yen), *Y. pseudotuberculosis* (Yps), and *Y. pestis* (Ype). Vertical lines mark amino acid positions that are conserved across intimins and invasins. Alignment gaps are shown as open spaces. The solid triangle marks the site of insertion of an IS200-like element in *Y. pestis*. *Bottom*, Neighbor-joining trees based on amino acid positions in the central region (A186–Y521) and in the C-terminus (Y907–C1063). Gamma distances were used with shape parameters of 1.36 and 4.45 for the central region and C-terminus, respectively. Branch lengths and bootstrap values for 10,000 replicate trees are given.

The phylogeny of the invasins based on the C-terminal residues has the same branching as the central region, although the amount of divergence is 1.4–1.8 times greater in terms of the branch lengths (fig. 1, bottom). In contrast, the phylogeny of the intimins based on the C-terminus has a different topology and longer branch tips than the tree based on the central region. The disproportionately long branches leading to Int- $\alpha$  and Int- $\gamma$  reflect a greater level of divergence for the C-terminus of the protein than for the central region. As a consequence of this extensive divergence, the C-terminus of Int- $\beta$  shares a most recent ancestor with Int-Cro,

and the overall level of divergence of the intimins far exceeds that of the invasins (fig. 1, bottom).

There are two possible explanations for the discordance between the intimin phylogenies. One possibility is that the intimin genes are mosaics, composed of segments with different histories; these segments have been brought together by past horizontal transfer and intragenic recombination events. The second possibility is that natural selection has drastically altered the rate of molecular evolution across the gene, either through variation in the level of functional constraint or through the acceleration of amino acid substitutions by diversifying selection.

**Table 2**  
**Rates of Synonymous ( $d_S$ ) and Nonsynonymous ( $d_N$ ) Variation in Two Segments of Intimin and Invasin**

PROTEIN	CENTRAL REGION (335 codons)			C-TERMINUS (157 codons)		
	$d_S \times 100$	$d_N \times 100$	$d_S/d_N$	$d_S \times 100$	$d_N \times 100$	$d_S/d_N$
Intimin (4).....	24.26 $\pm$ 2.72	2.43 $\pm$ 0.41	9.8	116.47 $\pm$ 18.91	39.56 $\pm$ 3.09	2.9
Invasin (3).....	85.71 $\pm$ 11.74	10.29 $\pm$ 1.03	8.3	76.75 $\pm$ 14.35	18.89 $\pm$ 2.25	4.1

NOTE.— $n$  is the number of sequences compared.

Evidence for different histories of DNA segments encoding the central region and C-terminus comes from the comparison of the rates of synonymous and nonsynonymous substitution (table 2). If the rate of amino acid replacement were accelerated by natural selection on residues in the carboxyl end of the protein, then the  $d_N$  of the 3' end of the gene should be significantly greater than that for the central segment, whereas  $d_S$  should be unaffected and approximately equal. This is the pattern that is seen in the three invasins (table 2). In contrast, both the average of  $d_N$  and that of  $d_S$  are significantly inflated at the 3' ends of the intimin genes. It is not clear how diversifying selection, which operates on nonsynonymous variation, would effect a nearly five-fold increase in the divergence at synonymous sites without intragenic recombination.

To determine whether the rate of substitution varies across the gene, we tested the observed number of nucleotide differences for pairs of sequences to an outgroup with Tajima's (1993) chi-square test. There were no significant departures from a uniform rate of nucleotide substitution for four comparisons within each region of the gene (table 3). However, for these tests, we assumed that the outgroups were known based on the phylogenies in figure 1, so that the comparisons and outgroups for the central region were different from those for the C-terminus. If, however, the phylogeny for the central region is used to test the molecular clock in the C-terminus, then significant departures from the uniform rate of evolution are obtained. For example, the molecular-clock hypothesis can be rejected for the C-terminus if Int-Cro is used as an outgroup for Int- $\alpha$  and

Int- $\beta$ , as it is in the central region ( $m_1 = 121$ ,  $m_2 = 29$ ,  $m_3 = 26$ ,  $\chi^2 = 56.24$ ). Our conclusion, therefore, is that there is no basis for rejecting the hypothesis of a uniform rate of amino acid substitution for the divergence of each segment, given that the phylogenies (or histories) for the segments are correct but different from one another.

#### Analysis of Mosaic Gene Structure

To analyze the mosaic gene structures of intimins, we used two statistical tests. First, we examined the distribution of polymorphic nucleotide sites along the gene and used the Stephens (1985) test to detect the significant clustering of sites that can result from past intragenic recombination or gene conversion events. We then applied Maynard Smith's (1992) maximum chi-square method (MAXCHI) to identify the break points between distinct segments within the genes.

For the analysis, the four intimin sequences (Int- $\alpha$ , Int- $\beta$ , Int- $\gamma$ , and Int-Cro) were realigned without the invasins to remove extraneous alignment gaps which could influence the statistical results. The intimin alignment has a length of 944 codon positions (fig. 2). The first step in the Stephens test was to tabulate the number and position of sites that support each partition or grouping of the sequences. The analysis was based on polymorphic sites that involved synonymous differences between the sequences. There were a total of 14 partitions, of which 6 were trivial and supported by less than 4 synonymous sites. Of the remaining 8 partitions, supported by a total of 243 (93%) of the 260 synonymous sites, 5 showed significant nonrandom clustering of sites

**Table 3**  
**Results of the Tajima (1993) Test of the Molecular-Clock Hypothesis for Two Regions of Intimin and Invasin**

Comparison	Sequence 1	Sequence 2	Outgroup	$m_1$	$m_2$	$m_3^a$	$\chi^2$
Central region							
1.....	$\alpha$	$\gamma$	$\beta$	10	8	25 (0)	0.22
2.....	$\alpha$	$\beta$	Cro	16	18	79 (1)	0.12
3.....	$\alpha$	Cro	Yen	25	36	411 (35)	1.98
4.....	Yen	Yps	$\alpha$	73	87	307 (91)	1.23
C-terminus							
1.....	$\beta$	Cro	$\gamma$	23	29	121 (23)	0.69
2.....	$\beta$	$\gamma$	$\alpha$	61	61	65 (41)	0.00
3.....	$\beta$	$\alpha$	Yen	50	46	131 (66)	0.17
4.....	Yen	Yps	$\alpha$	30	43	159 (53)	2.32

NOTE.—The chi-square test statistic was compared with a theoretical  $\chi^2$  with 1 df. None of the values presented in the table were significant at the  $P < 0.05$  level.<sup>a</sup> The numbers in parentheses are the numbers of sites with three different nucleotides among the three sequences.

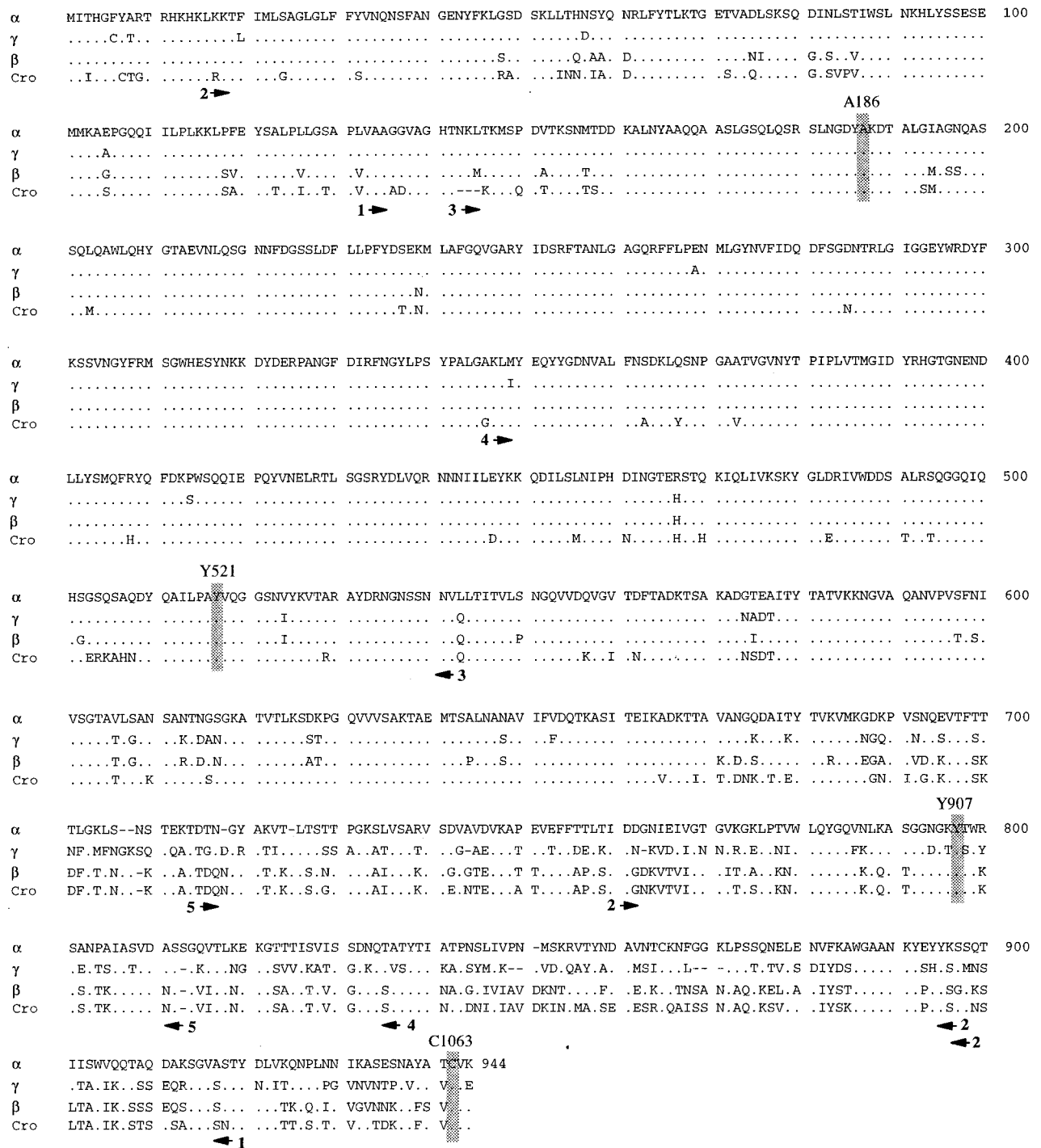


FIG. 2.—Multiple alignment of intimins. Bold numbers and arrows refer to significant partitions detected by the Stephens (1985) test (table 4). Conserved amino acids identified in figure 1 are noted with shading in their corresponding locations in the alignment.

(table 4). Partitions 1 and 2 involve sites supporting a single sequence separated from the remaining three genes. The positions of the sites associated with the significant  $d_0$  and  $g_0$  values are shown on the amino acid alignment in figure 2.

The most convincing evidence for intragenic recombination derived from the Stephens test occurs when two distinct partitions are both supported by significantly clustered sites. This condition is seen in partitions

3 and 4 (table 4). Partition 3 groups Int- $\alpha$  and Int- $\beta$  together and is supported by an improbable run of eight synonymous sites between codons 141 and 553. Significant clustering in support of partition 3 was also found when all sites ( $n = 19$ ) were considered. In contrast to partition 3, partition 4 groups Int- $\gamma$  and Int- $\beta$  together, and 12 synonymous sites between codons 345 and 838 support this partition (table 4). The grouping of Int- $\gamma$  and Int- $\beta$  can also be seen in the amino acid align-

**Table 4**  
**Results of the Stephens Test for Clustering of Polymorphic Sites that Support Specific Partitions**

Partition	No. of sites <sup>a</sup>	$d^0$ (Pr[ $d \leq d_0$ ])	Codon Position	$g_0$ (Pr[ $g \leq g_0$ ])	Codon Position
1. $\alpha/\gamma\beta$ Cro.....	27	2,335 (0.037)	142–920	482 (0.07)	162–536
2. Cro/ $\alpha\gamma\beta$ .....	85	2,622 (0.010)	15–889	383 (0.000)	760–888
3. $\alpha\beta/\gamma$ Cro.....	8	2,038 (0.296)	138–817	1,235 (0.026)	141–553
4. $\gamma\beta/\alpha$ Cro.....	12	1,480 (0.005)	345–838	491 (0.176)	115–169
5. $\alpha/\gamma/\beta$ Cro.....	8	301 (0.000)	714–814	110 (0.365)	740–777

<sup>a</sup> The number of polymorphic sites that support each partition.

ment (see codons 606–648 in fig. 2). To explain this inconsistency in partitioning, one must either invoke parallel mutations at multiple silent sites or hypothesize past lateral transfer and recombination of DNA segments. The latter explanation appears more likely, because multiple synonymous sites support both partitions.

To identify segments involved in lateral transfer and recombination, we compared each *E. coli* intimin gene to Int-Cro and identified the positions of breakpoints by the maximum of the chi-square values calculated for every nucleotide in the sequence (Maynard Smith 1992). For each sequence, there were four significant  $k_{MAX}$  values that divided each gene into five pieces (fig. 3). There is a general concordance in the break points in two locations. All three intimins have a significant  $k_{MAX}$  at or near codon 202, and Int- $\alpha$  and Int- $\gamma$  both have a significant break point at codon 668 (fig. 3).

The complex history of intimin evolution is shown by comparing the divergence at synonymous (fig. 4A) and nonsynonymous (fig. 4B) sites for different segments of the gene. In the 5' end and central region of the gene (codons 1–478), Int- $\alpha$  and Int- $\gamma$  are most similar in sequence, e.g., with  $5.5 \pm 1.8$  synonymous differences per 100 sites for codons 203–478. In this same region, Int- $\beta$  is about twofold more divergent ( $d_s \times 100 = 12.8 \pm 2.8$ ). The segment of codons 44–202 is more highly divergent between Int- $\beta$  and the other intimins, with  $d_s > 30$  per 100 sites, possibly as a consequence of a recombination event involving a 1.5-kb piece of DNA in Int- $\beta$ .

Between codons 478 and 695, there are five short segments in which the relative divergence of the genes varies widely.  $\beta$  and  $\gamma$  intimins are most closely related in two regions, codons 478–510 and 584–668, suggesting the possibility of at least two past recombination

events to create these mosaics. Embedded between these regions is a stretch of 73 codons in which Int- $\beta$  is highly divergent from the other intimins. Finally, all three proteins show extraordinary levels of divergence in the C-terminal 250 amino acid positions. The level of divergence between  $\alpha$ ,  $\beta$ , and  $\gamma$  intimins exceeds 50% of the synonymous sites and more than 40% of the nonsynonymous sites. As observed in the comparison with *invasin* (fig. 1), Int- $\beta$  is more closely related to *C. rodentium* intimin in the 3' end of the gene (differing at 35% of the synonymous sites) than it is to  $\alpha$  and  $\gamma$  intimins of *E. coli*.

## Discussion

There is growing evidence that many different bacterial pathogens evolve by acquiring large blocks of genes with integrated functions called pathogenicity islands (Ochman and Groisman 1996; Hacker et al. 1997). Intimin is one protein encoded on the LEE pathogenicity island, a 35-kb DNA chromosomal region that specifies the mechanism for producing attaching-effacing lesions for pathogenic *E. coli* (McDaniel et al. 1995). The island includes a type III secretion system and a series of secreted proteins with homology to similar proteins in *Shigella*, *Salmonella*, and *Yersinia* (Jarvis and Kaper 1996; Elliot et al. 1998). Although there is no direct evidence that the LEE island has spread between strains in nature, the fact that it has a 38.4% G+C content (Elliot et al. 1998), compared with 50.8% for the *E. coli* chromosome (Blattner et al. 1997), and that it occurs at different genomic locations in divergent clonal groups (Wieler et al. 1997) suggests that this large block of genes has been horizontally transferred in the past. The implication of such transfers is that receptive bacteria can acquire a

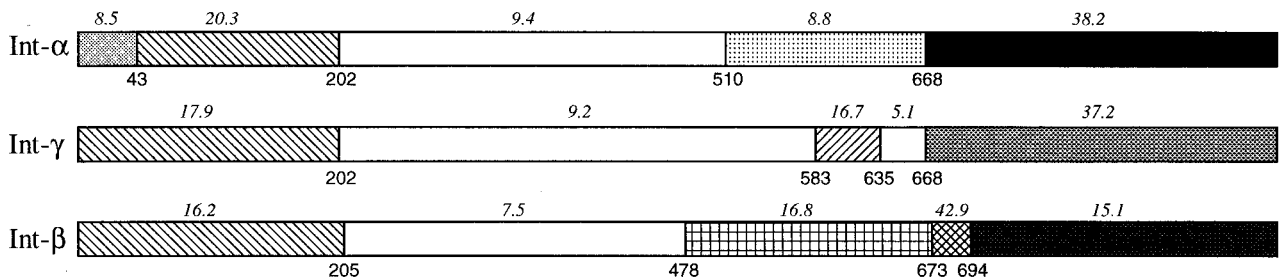


FIG. 3.—The locations of the breakpoints ( $k_{MAX}$ ) and the percentages of sequence divergence between segments detected by the maximum chi-square method. The numbers above the segments are percentages of nucleotide differences in comparison to Int-Cro

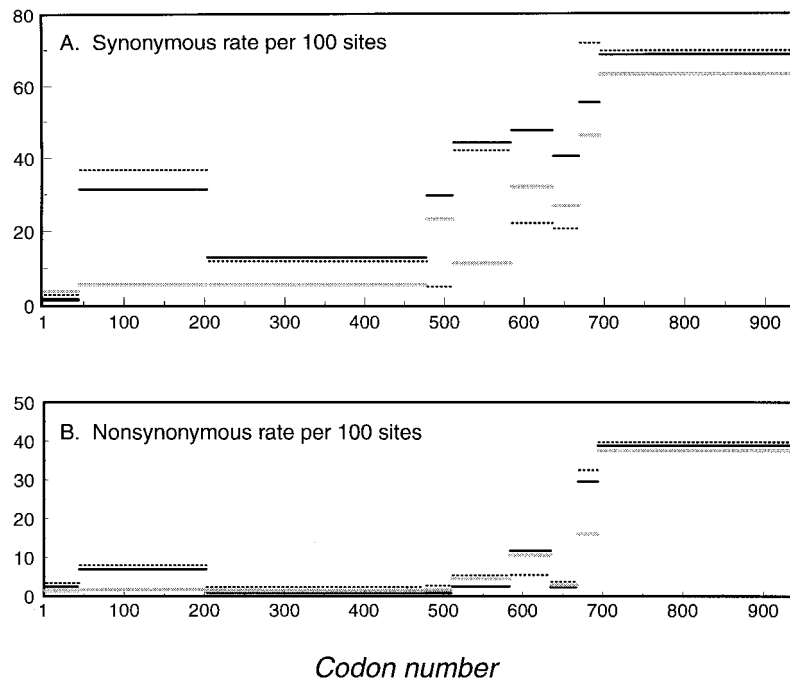


FIG. 4.—Variation in (A) synonymous ( $d_S \times 100$ ) and (B) nonsynonymous ( $d_N \times 100$ ) rates of substitution for mosaic segments identified by maximum chi-square analysis. The lines denote the average value obtained for codons in a segment. Each pair of intimins was compared separately as follows: Int- $\alpha$  versus Int- $\beta$  (solid black), Int- $\alpha$  versus Int- $\gamma$  (thick gray), and Int- $\beta$  versus Int- $\gamma$  (dotted black).

new complex multigene phenotype in the chromosome in a single evolutionary step.

The insertion of LEE into different chromosomal backgrounds sets the stage for acquisition of additional virulence factors and the divergence of new pathogens. In EPEC strains, the spread of a virulence plasmid that encodes a bundle-forming pilus and a locus that upregulates intimin expression (Whittam and McGraw 1996) produced two groups of pathogens that cause infantile diarrhea and have spread worldwide. In EHEC strains, the spread of Shiga toxin genes by bacteriophages into a  $\gamma$  intimin-producing ancestor created a highly virulent pathogen (*E. coli* O157:H7) that causes a new disease (hemorrhagic colitis) (Whittam, McGraw, and Reid 1998). The evidence obtained here suggests that the divergence of intimin genes has been enhanced by recombination, particularly in the 3' end of the gene.

One implication of the finding of mosaic intimins is that other portions of the LEE island have been involved in these past recombination events. This means that the islands carrying the distinct intimin genes are probably larger mosaics. The structures and evolutionary histories of these mosaic segments will be elucidated as more comparative sequencing of pathogenicity islands is completed.

We also found that the Int- $\gamma$  genes of members of the O157:H7 complex, including an O157:H7 strain, the Shiga-like toxin-producing sorbitol-positive O157:H-strain, the nontoxic O55:H7 strain, and ECOR37 are virtually identical in sequence. This result extends the earlier observations of Louie et al. (1994), who sequenced a 0.8-kb fragment of the 3' end of the *eae* gene and found that O55:H7 and O157:H7 differed by only

a single nucleotide. The accumulation of only two point mutations among their *eae* genes supports the model that O157:H7 has very recently evolved from an *eae*-positive O55:H7-like ancestor with the acquisition of toxin genes. One consequence of this model is that observed differences in the pathogenesis of these strains cannot be readily attributed to structural differences in the intimin molecule.

A second finding is that EPEC 2 strains of serotypes O128:H2 (Dec 11a) and O111:H2 (Dec 12a), which have been shown by multilocus enzyme electrophoresis to mark a widespread clonal lineage (Whittam et al. 1993), have the same *eae* gene, encoding Int- $\beta$ , which differs by more than 10% in sequence from Int- $\alpha$  of EPEC strain E2348/69. This result strongly supports the conclusion that classical EPEC strains of a variety of serotypes represent two distantly related clonal lineages that have long been separated (Ørskov et al. 1990; Whittam 1995). In the course of the divergence of these pathogenic clones, many nonsynonymous changes have accumulated at the *eae* locus, by both mutation and recombination, which have contributed to extensive divergence of  $\alpha$  and  $\beta$  intimin, especially in the C-terminal region. Further support for the role of recombination in generating mosaic intimins comes from the characterization of an *eae* gene from a canine *E. coli* strain (An et al. 1997). This intimin is most similar to Int- $\alpha$  at the 5' end of the gene, but resembles Int-Cro at the 3' end. Interestingly, MAXCHI analysis shows that although the break points are in the same places as those in the intimins from human strains, the combination of diverse segments is novel (data not shown). The differences between  $\alpha$  and  $\gamma$  intimin are also concentrated at

the 3' end of the gene, as noted by Yu and Kaper (1992). These findings suggest that diversity among the intimins in the C-terminus has been accelerated by the acquisition of gene segments from sources outside of the *E. coli* population.

Because intimins are outer membrane proteins, they may be subjected to strong selective pressure for amino acid diversity, either through adaptive shifts in tropism or through antigenic shifts to evade the immune system. For invasins and intimins, the C-terminal domain contains the receptor-binding part of the protein (Leong, Fournier, and Isberg 1990; Frankel et al. 1994), suggesting that high C-terminal divergence between invasins and intimin molecules reflects binding to different eukaryotic receptors (Yu and Kaper 1992). Likewise, the divergence among the intimin alleles in the C-terminal portion of the protein may be an evolutionary consequence of shifts in the recognition and binding of the cell receptors. Surprisingly, Kenny et al. (1997) recently discovered that the receptor for Int- $\gamma$  is not a eukaryotic protein, but is, in fact, a bacterial protein that is inserted into the mammalian cell surface. The receptor protein is encoded upstream of *eae* in the LEE island of EPEC E2348/69. Because of the extensive C-terminal divergence, we expect that the corresponding receptor protein of Int- $\gamma$  and Int- $\beta$  is distinct, possibly as a result of recombination within LEE or because these intimins interact with separate eukaryotic surface proteins. The diversity in this part of the molecule could also be favored by immune system selection for novel antigenic variants. In either case, under conditions of diversifying selection, mutations or recombinations that create variation in amino acid sequence can sweep new mosaic alleles to high frequencies.

There are now many examples of bacterial proteins that are highly variable (presumably subjected to diversifying selection) and encoded by genes with mosaic structures, composed of diverse segments with different histories. Often, these proteins are exposed to clear selective pressure for variation. For example, mosaic genes encode the penicillin-binding proteins that confer antibiotic resistance in *Streptococcus pneumoniae* (Martin, Sibold, and Hakenbeck 1992) and *Neisseria meningitidis* (Spratt et al. 1991). Mosaic gene structures have been described for immunogenic virulence factors such as the streptokinase gene in *Streptococcus pyogenes* (Kapur et al. 1995), IgA1 protease (Morelli et al. 1994) and Opa outer membrane proteins (Hobbs et al. 1994) of pathogenic *Neisseria* species, and the anti-phagocytic M protein of group A streptococci (Whatmore et al. 1995). Highly variable cell-surface structures such as phase 1 flagellins (Li et al. 1994) and somatic O antigens of *Salmonella enterica* (Reeves 1992) are encoded by genes with complex mosaic structures. Like intimin genes, these mosaic structures are composed of diverse segments derived through multiple recombination events. Further research is needed into the development of phylogenetic methods for detecting and analyzing mosaic genes and reticulate evolution of DNA segments.

## Acknowledgments

The authors thank Sheheila Plock and Stephanie O'Bryan for technical assistance. This research was supported by Public Health Service Grants AI 24566 (TSW), AI 42391 (TSW), and AI 22144 (RKS) from the National Institutes of Health.

## APPENDIX

Because there were 11 conflicts between the *eae* sequences determined here and those previously reported for the same strains in GenBank, we resequenced parts of the *eae* gene from O157:H7 strains EDL 933 (Yu and Kaper 1992) and CL-8 (Beehahee et al. 1992). The conflicting nucleotide bases at these sites were identical to those of strains DEC 3a, DEC 3f, DEC 5d, and ECOR 37; therefore, the conflicts were attributed to errors in the previously published sequences. The corrected sites (reported in GenBank  $\rightarrow$  corrected) are numbered with respect to the start codons and are as follows. For strain EDL933 (GenBank accession number Z11541): 661 (G $\rightarrow$ A), 933 (G $\rightarrow$ C), 934 (C $\rightarrow$ G), 952 (C $\rightarrow$ A), and 1925 (G $\rightarrow$ C). For strain CL-8 (GenBank accession number X60439): 2319 (G $\rightarrow$ A), 2320 (T $\rightarrow$ G), 2321 (C $\rightarrow$ G), 2322 (G $\rightarrow$ C), 2323 (A $\rightarrow$ G), and 2324 (T $\rightarrow$ A).

## LITERATURE CITED

- ABU-BOBIE, J., G. FRANKEL, C. BAIN, A. G. CONCALVES, L. R. TRABULSI, G. DOUCE, S. KNUITON, and G. DUNCAN. 1998. Detection of intimins  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\delta$ , four intimin derivatives expressed by attaching and effacing microbial pathogens. *J. Clin. Microbiol.* **36**:662–668.
- AGIN, T. S., and M. K. WOLFE. 1997. Identification of a family of intimins common to *Escherichia coli* causing attaching-effacing lesions in rabbits, humans, and swine. *Infect. Immun.* **65**:320–326.
- AN, H., J. M. FAIRBROTHER, J. D. DUBREUIL, and J. HOREL. 1997. Cloning and characterization of the *eae* gene from a dog attaching and effacing *Escherichia coli* strain 4221. *FEMS Microbiol. Lett.* **148**:239–245.
- ARMSTRONG, G. L., J. HOLLINGSWORTH, and J. J. G. MORRIS. 1996. Emerging foodborne pathogens: *Escherichia coli* O157:H7 as a model of entry of a new pathogen into the food supply of the developed world. *Epidemiol. Rev.* **18**:29–51.
- BEEBAKHEE, G., M. LOUIE, J. D. ACAREDO, and J. BRUNTON. 1992. Cloning and nucleotide sequence of the *eae* gene homologue from enterohemorrhagic *Escherichia coli* serotype O157:H7. *FEMS Microbiol. Lett.* **91**:63–68.
- BERGTHORSSON, U., and H. OCHMAN. 1998. Distribution of chromosomal length variation in natural isolates of *Escherichia coli*. *Mol. Biol. Evol.* **15**:6–16.
- BISERCIC, M., J. Y. FEUTRIER, and P. R. REEVES. 1991. Nucleotide sequences of the *gnd* genes from nine natural isolates of *Escherichia coli*: evidence of intragenic recombination as a contributing factor in the evolution of the polymorphic *gnd* locus. *J. Bacteriol.* **173**:3894–3900.
- BLATTNER, F. R., G. PLUNKETT III, C. A. BLOCH et al. (17 co-authors). 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**:1453–1462.
- DONNENBERG, M. S., and J. B. KAPER. 1992. Enteropathogenic *Escherichia coli*. *Infect. Immun.* **60**:3953–3961.
- ELLIOT, S. J., L. A. WAINWRIGHT, T. K. MCDANIEL, K. G. JARVIS, Y. DENG, L. C. LAI, B. P. MCNAMARA, M. S. DON-

- NENBERG, and J. B. KAPER. 1998. The complete sequence of the locus of enterocyte effacement (LEE) from enteropathogenic *Escherichia coli* E2348/69. *Mol. Microbiol.* **28**:1–4.
- FELSENSTEIN, J. 1989. PHYLIP—phylogeny inference package (version 3.2). *Cladistics* **5**:164–166.
- FENG, P., K. A. LAMPEL, H. KARCH, and T. S. WHITTAM. 1998. Genetic and phenotypic changes in the emergence of *Escherichia coli* O157:H7. *J. Infect. Dis.* **177**:1750–1753.
- FRANKEL, G., D. C. A. CANDY, P. EVEREST, and G. DOUGAN. 1994. Characterization of the C-terminal domains of intimin-like proteins of enteropathogenic and enterohemorrhagic *Escherichia coli*, *Citrobacter freundii*, and *Hafnia alvei*. *Infect. Immun.* **62**:1835–1842.
- GU, X., and J. ZHANG. 1997. A simple method for estimating the parameter of substitution rate variation among sites. *Mol. Biol. Evol.* **14**:1106–1113.
- HACKER, J., G. BLUM-OEHLER, I. MUHLDOERFER, and H. TSCHAPE. 1997. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol. Microbiol.* **23**:1089–1097.
- HIGGINS, D. G., and P. M. SHARP. 1988. CLUSTAL: a package for performing multiple sequence alignments on a microcomputer. *Gene* **73**:237–244.
- HOBBS, M. M., A. SEILER, M. ACHTMAN, and J. G. CANNON. 1994. Microevolution within a clonal population of pathogenic bacteria: recombination, gene duplication and horizontal genetic exchange in the *opa* gene family of *Neisseria meningitidis*. *Mol. Microbiol.* **12**:171–180.
- ISBERG, R. R., D. L. VORRHIS, and S. FALKOW. 1987. Identification of invasins: a protein that allows enteric bacteria to penetrate cultured mammalian cells. *Cell* **50**:769–778.
- JARVIS, K. G., and J. B. KAPER. 1996. Secretion of extracellular proteins by enterohemorrhagic *Escherichia coli* via a putative type III secretion system. *Infect. Immun.* **64**:4826–4829.
- JERSE, A. E., and J. B. KAPER. 1991. The *eae* gene of enteropathogenic *Escherichia coli* encodes a 94-kilodalton membrane protein, the expression of which is influenced by the EAF plasmid. *Infect. Immun.* **59**:4302–4309.
- JERSE, A. E., J. YU, B. D. TALL, and J. B. KAPER. 1990. A genetic locus of enteropathogenic *Escherichia coli* necessary for the production of attaching and effacing lesions on tissue culture cells. *Proc. Natl. Acad. Sci. USA* **87**:7839–7843.
- KAPUR, V., S. KANJILAL, M. R. HAMRICK, L.-L. LI, T. S. WHITTAM, S. A. SAWYER, and J. M. MUSSER. 1995. Molecular population genetic analysis of the streptokinase gene of *Streptococcus pyogenes*: mosaic alleles generated by recombination. *Mol. Microbiol.* **16**:509–519.
- KENNY, B., R. DEVINNY, M. STEIN, D. J. REINSCHIED, E. A. FREY, and B. B. FINLAY. 1997. Enteropathogenic *E. coli* (EPEC) transfer its receptor for intimin adherence into mammalian cells. *Cell* **91**:511–520.
- KUMAR, S., K. TAMURA, and M. NEI. 1993. MEGA: molecular evolutionary genetics analysis. version 1.0. Pennsylvania State University, University Park.
- LEONG, J. M., R. S. FOURNIER, and R. R. ISBERG. 1990. Identification of the integrin binding domain of the *Yersinia pseudotuberculosis* invasins protein. *EMBO J.* **9**:1979–1989.
- LI, J., K. NELSON, A. C. MCWHORTER, T. S. WHITTAM, and R. K. SELANDER. 1994. Recombinational basis of serovar diversity in *Salmonella enterica*. *Proc. Natl. Acad. Sci. USA* **91**:2552–2556.
- LOUIE, M., J. D. ALZAVEDO, R. CLARKE, A. BORZCYK, H. LIOR, M. RICHTER, and J. BRUNTON. 1994. Sequence heterogeneity of the *eae* gene and detection of verotoxin-producing *Escherichia coli* using serotype-specific primers. *Epidemiol. Infect.* **112**:449–461.
- MCDANIEL, T. K., K. G. JARVIS, M. S. DONNENBERG, and J. B. KAPER. 1995. A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens. *Proc. Natl. Acad. Sci. USA* **92**:1664–1668.
- MARTIN, C., C. SIBOLD, and R. HAKENBECK. 1992. Relatedness of penicillin-binding protein 1a genes from different clones of penicillin-resistant *Streptococcus pneumoniae* isolated in South Africa and Spain. *EMBO J.* **11**:3831–3836.
- MAYNARD SMITH, J. 1992. Analyzing the mosaic structure of genes. *J. Mol. Evol.* **34**:126–129.
- MORELLI, G., L. D. VALLE, C. J. LAMMEL et al. (11 co-authors). 1994. Immunogenicity and evolutionary variability of epitopes within IgA1 protease from serogroup A *Neisseria meningitidis*. *Mol. Microbiol.* **11**:175–187.
- NATARO, J. P., and J. B. KAPER. 1998. Diarrheagenic *Escherichia coli*. *Clin. Microbiol. Rev.* **11**:142–201.
- NEI, M., and T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**:418–426.
- OCHMAN, H., and E. A. GROISMAN. 1996. Distribution of pathogenicity islands in *Salmonella* spp. *Infect. Immun.* **64**:5410–5412.
- OCHMAN, H., and R. K. SELANDER. 1984. Standard reference strains of *Escherichia coli* from natural populations. *J. Bacteriol.* **157**:690–693.
- ØRSKOV, F., T. S. WHITTAM, A. CRAVIOTO, and I. ØRSKOV. 1990. Clonal relationships among classic enteropathogenic *Escherichia coli* (EPEC) belonging to different O groups. *J. Infect. Dis.* **162**:76–81.
- REEVES, P. R. 1992. Variation in O-antigens, niche-specific selection and bacterial populations. *FEMS Microbiol. Lett.* **100**:509–516.
- SCHAUER, D. B., and S. FALKOW. 1993. Attaching and effacing locus of a *Citrobacter freundii* biotype that causes transmissible murine colonic hyperplasia. *Infect. Immun.* **61**:2486–2492.
- SCHAUER, D. B., B. A. ZABEL, I. F. PEDRAZA, C. M. O'HARA, A. G. STEIGERWALT, and D. J. BRENNER. 1995. Genetic and biochemical characterization of *Citrobacter rodentium* sp. nov. *J. Clin. Microbiol.* **33**:2064–2068.
- SIMONET, M., B. RIOT, N. FORTINEAU, and P. BERCHE. 1996. Invasin production by *Yersinia pestis* is abolished by insertion of an IS200-like element within the *inv* gene. *Infect. Immun.* **64**:375–379.
- SPRATT, B. G., C. G. DOWSON, Q.-Y. ZHANG, L. D. BEWLER, J. A. BRANNIGAN, and A. HUTCHISON. 1991. Mosaic genes, hybrid penicillin-binding proteins, and the origins of penicillin resistance in *Neisseria meningitidis* and *Streptococcus pneumoniae*. Pp. 73–83 in J. CAMPISI, D. D. CUNNINGHAM, M. INOUE, and M. RILEY, eds. *Perspectives on cellular regulation: from bacteria to cancer*. Wiley-Liss, NY.
- STEPHENS, J. C. 1985. Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Mol. Biol. Evol.* **2**:539–556.
- TAJIMA, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135**:599–607.
- WHATMORE, A. M., V. KAPUR, J. M. MUSSER, and M. A. KEHOE. 1995. Molecular population genetic analysis of the *emm* subdivision of group A streptococcal *emm*-like genes: horizontal gene transfer and restricted variation among *emm* genes. *Mol. Microbiol.* **15**:1039–1048.
- WHITTAM, T. S. 1995. Genetic population structure and pathogenicity in enteric bacteria. Pp. 217–245 in S. BAUMBERG, J. P. W. YOUNG, E. M. H. WELLINGTON, and J. R. SAUN-

- DERS, eds. Population genetics of bacteria. Cambridge University Press, Cambridge, England.
- WHITTAM, T. S., and E. A. MCGRAW. 1996. Clonal analysis of EPEC serogroups. *Rev. Microbiol. Sao Paulo.* **27**:7–16.
- WHITTAM, T. S., E. A. MCGRAW, and S. D. REID. 1998. Pathogenic *Escherichia coli* O157:H7: A model for emerging infection diseases. Pp. 163–183 in R. M. KRAUSE, eds. Emerging infections. Academic Press, New York.
- WHITTAM, T. S., M. L. WOLFE, I. K. WACHSMUTH, F. ØRSKOV, I. ØRSKOV, and R. A. WILSON. 1993. Clonal relationships among *Escherichia coli* strains that cause hemorrhagic colitis and infantile diarrhea. *Infect. Immun.* **61**: 1619–1629.
- WIELER, L. H., T. K. MCDANIEL, T. S. WHITTAM, and J. B. KAPER. 1997. Insertion site of the locus of enterocyte effacement in enteropathogenic and enterohemorrhagic *Escherichia coli* differs in relation to the clonal phylogeny of strains. *FEMS Microbiol. Lett.* **156**:49–53.
- WILSON, K. 1990. Preparation of genomic DNA from bacteria. Pp. 2.4.1–2.4.5 in F. M. AUSUBEL, R. BRENT, R. E. KINGSTONET, D. D. MOORE, J. A. SMITH, J. G. SEIDMAN, and K. STRUHL, eds. Current protocols in molecular biology. Wiley-Interscience, New York.
- YOUNG, V. B., V. L. MILLER, S. FALKOW, and G. K. SCHOOLNIK. 1990. Sequence, localization and function of the invasion protein of *Yersinia enterocolitica*. *Mol. Microbiol.* **4**: 119–128.
- YU, J., and J. B. KAPER. 1992. Cloning and characterization of the *eae* gene of enterohaemorrhagic *Escherichia coli* O157:H7. *Mol. Microbiol.* **6**:411–417.
- JULIAN P. ADAMS, reviewing editor
- Accepted September 15, 1998