J Mol Evol (1998) 47:258–267

JOURNAL OF MOLECULAR EVOLUTION

Diversifying Selection Governs Sequence Polymorphism in the Major Adhesin Proteins FimA, PapA, and SfaA of *Escherichia coli*

E. Fidelma Boyd, Daniel L. Hartl

Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA

Received: 7 August 1997 / Accepted: 8 March 1998

Abstract. Fimbriae or pili are essential adherence factors usually found in pathogenic bacteria to aid colonization of host cells. Three major structural pilin genes, *fimA, sfaA,* and *papA,* from *Escherichia coli* natural isolates were examined and nucleotide sequence data revealed elevated levels of both synonymous and nonsynonymous site variation at these loci. Examination of synonymous site variation shows a fivefold increase in *fimA* sites, relative to the housekeeping gene *mdh;* and similarly the *sfaA* and *papA* genes have increased synonymous sites variation relative to *fimA.* Nonsynonymous site variation is also elevated at all three loci but, in particular, at the *papA* locus ($k_N = 0.44$). The k_N/k_S ratio for the three genes are among the highest yet reported for *E. coli* genes. Regional variation in nucleotide polymorphism within each of the genes reveal hypervariable segments where nonsynonymous substitutions exceed synonymous substitutions. We propose that at the *fimA, papA,* and *sfaA* genes, diversifying selection has brought about the increase levels of polymorphism.

Key words: Diversifying selection — *Escherichia coli* — Major structural pilin proteins

Introduction

Evidence at the molecular level for the role of diversifying selection in protein evolution is rare and, when found, is usually associated with loci implicated in host defense mechanism such as the major histocompatibility complex, the *Vh* loci, serine proteases of mammals (Hughes and Nei 1988; Tanaka and Nei 1989), and restriction–modification genes and colicins in bacteria (Sharp et al. 1992; Tan and Riley 1996). The best-studied case of diversifying selection in bacteria is work on the colicin proteins of *E. coli;* colicins are toxins produced by, and active against, *E. coli* and related bacteria. Diversifying selection has been proposed as a molecular mechanism for the origin and diversification of one group of colicins. According to this model, novel colicin gene clusters arise from mutations which generate expanded immunity functions, then positive selection favors these new immunities and rapidly drives strains carrying the new colicin gene clusters to fixation in the local population (Tan and Riley 1996).

An alternative mechanism for generating diversity within genes is recombination of horizontally transferred DNA. However, interlocus comparisons of the levels of recombination among bacterial genes have revealed very different patterns depending on the locus under study. For example, recombination among housekeeping or virulence genes of enteric bacteria is infrequent (Hall and Sharp 1992; Selander et al. 1994; Boyd et al. 1994, 1996, 1997; Li et al. 1995). In contrast, genes of the *rfb* region that mediate biosynthesis of a highly antigenic polysaccharide show evidence of frequent recombination from distantly related species of bacteria (Reeves, 1992; Lui and Reeves 1994), and similarly genes of the type I restriction–modification systems have undergone frequent recombination as a mechanism of allelic variation (Sharp et al. 1992; Barkus et al. 1995). Comparative nucleotide sequence analysis of enteric bacteria has revealed many

Correspondence to: D.L. Hartl; *e-mail:* dhartl@oeb.harvard.edu

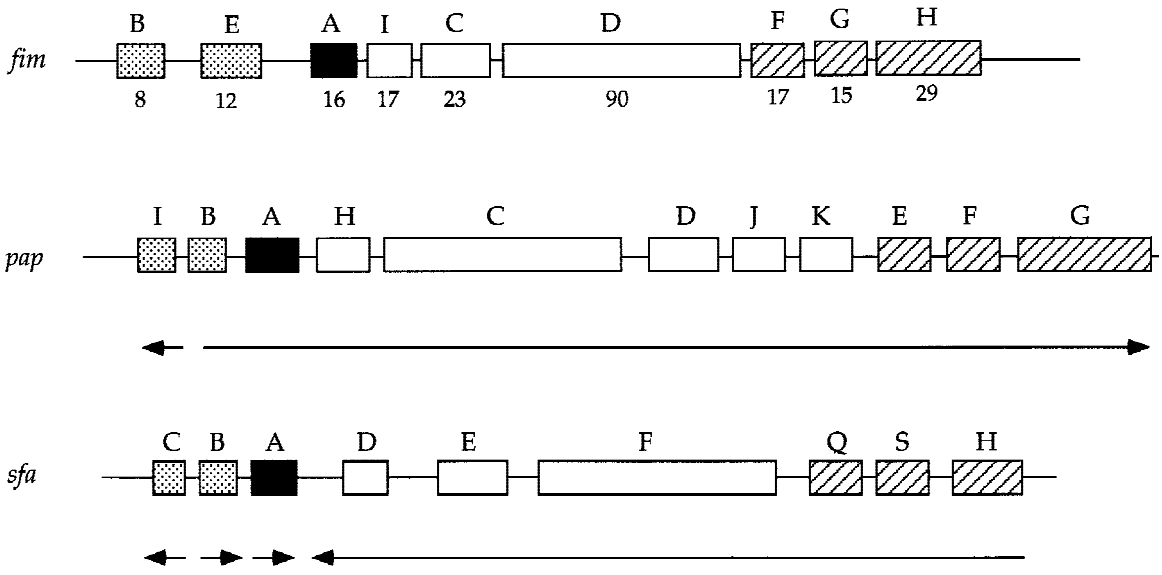

Fig. 1. Organization and structure of *fim*, *pap*, and *sfa* fimbriae gene clusters of *E. coli*. *Filled boxes* indicate major adhesin genes, *hatched boxes* indicate regulatory genes, *shaded boxes* indicate minor adhesin genes, and *open boxes* indicate processing and assembly genes. *Uppercase letters* indicate gene designation.

cases of mosaic segments of DNA indicative of recombination (DuBose et al. 1988; Milkman and Bridges 1990, 1993; Achtman and Hakenbeck 1992).

Pilin adhesins are essential colonization factors usually found in pathogenic enteric bacteria. Fimbrial adhesins facilitate the binding of bacteria to eukaryotic cells, which is the first step in the pathogenic process, and they may play an important role in avoiding the host immune system. Uropathogenic strains of *E. coli* associated with acute pyleonephritis often express P and S pili, which are absent from *E. coli* K-12, whereas type 1 fimbria is associated with upper urinary tract infections and are present in most wild-type *E. coli* (Hacker et al. 1997). The type 1, P, and S fimbriae show similar overall genetic organization and function that consists of four functional clusters (Fig. 1) that encode regulatory factors, the major subunit protein (encoded in *fimA*, *papA*, and *sfaA*) of interest here, transport and assembly proteins, and minor subunit proteins (Hultgren et al. 1996). The major pilin proteins of type-1, P, and S pilin gene clusters are the main structural units and are arranged as repeating units to form the pilin; the precise arrangement of the major proteins differs depending on the pilin type (Gaastra and Svennerholm 1996). The minor pilin proteins of type-1, P, and S pilin gene clusters correspond to the adhesive proteins of each of these attachment factors, in contrast to other adherence factors where the major subunit is the adhesive entity (Hacker 1992).

We examined natural isolates of the *E. coli* reference collection (ECOR) (Ochman and Selander 1984), whose phylogenetic relationships are known (Herzer et al. 1992), to determine allelic variation among the major adhesin protein of three fimbriae gene clusters. We found elevated levels of both synonymous and nonsynonymous site polymorphism across the major pilin genes *fimA*, *papA*, and *sfaA*. Within each gene there was regional variation in polymorphism with distinct hypervariable segments. An unusually high level of polymorphism is characteristic of a variety of bacterial genes encoding cell surface components that directly interact with the extracellular environment. We propose that the hypervariability of the adhesin pili structural genes examined here most likely reflect the action of diversifying selection in adaptation to the variation encountered by the bacteria in their host.

Materials and Methods

Bacterial Strains. From the *E. coli* reference (ECOR) collection of natural isolates (Ochman and Selander 1984), we examined nine strains that encode the *sfa* and *pap* gene clusters. Both the *sfa* and the *pap* gene clusters are confined mainly to two subgroups, B2 and D, of the ECOR collection (Boyd and Hartl 1998). Eight of the nine ECOR strains have previously been examined in the study of the housekeeping gene malate dehydrogenase (*mdh*) (Boyd et al. 1994).

Molecular Techniques. Genomic DNA was extracted with the G-Nome DNA isolation kit from Bio 101 (Vista, CA). Genomic DNA was used as template for polymerase chain reaction (PCR) amplification of *papA*, *papH*, and *sfaA*. Primers for PCR amplification were designed from previously reported *pap* and *sfa* sequences (Marklund et al. 1992; Hacker et al. 1992). PCR products were separated by electrophoresis in agarose gel and extracted from the agarose using the Geneclean kit (Bio 101). The purified PCR products were cloned using the TA cloning kit and the cloned fragments of the three genes *papA*, *papH*, and *sfaA* were sequenced with an Applied Biosystems Model 373A automated DNA sequencing system with a DyeDeoxy terminator cycle sequencing kit. For all strains both strands were sequenced. The nucleotide sequences of the *papA*, *papH*, and *sfaA* genes described in this paper have been deposited in the GenBank database under the accession numbers of AF051360 to AF051364 and AF051810 to AF051815.

Table 1. Nucleotide sequence polymorphism and diversity values for 3 major pilin genes and 11 genes from natural isolates of *E. coli*

Gene	No. of strains	Length (bp)	Poly sites		Mean pairwise value for		k_N/k_S ratio	Nucleotide diversity		Reference
			nt	aa	k_S	k_N		π	θ	
Major pilin										
fim A	7	540–552	80	44	0.157 ± 0.025	0.034 ± 0.007	0.22	0.06 ± 0.03	0.15	This study
papA	6	507–546	306	277	0.836 ± 0.113	0.443 ± 0.032	0.53	0.33 ± 0.16	0.56	This study
sfaA	5	531–537	132	97	0.280 ± 0.043	0.082 ± 0.010	0.29	0.10 ± 0.07	0.25	This study
E. coli										
papH	6	462	22	9	0.053 ± 0.015	0.010 ± 0.004	0.21	0.02 ± 0.01	0.05	Boyd and Hartl (1998)
*papH**	5	462	10	3	0.034 ± 0.013	0.003 ± 0.002	0.09	0.01 ± 0.01	0.02	Boyd and Hartl (1998)
phoA	8	1434	58	8	0.06 ± 0.008	0.003 ± 0.001	0.05	0.02 ± 0.010	0.04	Dubose et al. (1988)
putP	12	1467	108	8	0.09 ± 0.015	0.002 ± 0.001	0.02	0.02 ± 0.013	0.005	Nelson et al. (1992)
trpA	25	807	72	13	0.10 ± 0.014	0.003 ± 0.001	0.03	0.03 ± 0.010	0.09	Milkman and Stoltzfu (1988)
sppA	12	972	36	7	0.05 ± 0.008	0.002 ± 0.001	0.05	0.01 ± 0.006	0.04	Guttman and Dykhuizen (1994)
celC	11	351	15	4	0.05 ± 0.03	0.002 ± 0.004	0.04	0.01 ± 0.009	0.04	Hall and Sharp (1992)
gnd	34	1335	441	114	0.27 ± 0.12	0.009 ± 0.01	0.03	0.07 ± 0.040	0.32	Bisercic et al. (1991); Nelson and Selander (1994)
mdh	46	864	36	6	0.03 ± 0.017	0.001 ± 0.001	0.03	0.01 ± 0.005	0.04	Boyd et al. (1994); Pupo et al. (1997)
aceK	16	1722	166	20	0.13 ± 0.01	0.004 ± 0.001	0.03	0.03 ± 0.020	0.10	Nelson et al. (1997)
icd	17	1212	73	7	0.08 ± 0.036	0.002 ± 0.002	0.02	0.02 ± 0.009	0.06	Wang et al. (1997)
kpsD	8	729	46	12	0.06 ± 0.009	0.005 ± 0.002	0.09	0.02 ± 0.013	0.06	Boyd and Hartl (1998)

Nucleotide Sequence Analysis. In addition to *papA* and *sfaA* pilin genes, we analyzed a third major adhesin gene, *fimA*, of the type-1 fimbria from isolates of *E. coli*. The *fimA* sequences were recovered from the nucleotide databases and represent *E. coli* isolated from a range of sources. The U00096 sequence was isolated from *E. coli* K-12; U20815 from an *E. coli* O157:H7 strain; M27603 and Y10902 (Orndorff and Falkow 1985; van Die et al. 1984) from uropathogenic *E. coli* isolates; D13186 and Z37500 (Sekizaki et al. 1993; Marc and Dho-Moulin 1996) from avian *E. coli* isolates; and X00981 from a laboratory K12 sequence (Klemm 1984). Nucleotide sequences were analyzed using the following computer programs: Molecular Evolutionary Genetics Analysis (MEGA) (Kumar et al. 1993), Molecular Evolution Analysis (MEA) (Etsuko Moriyama, Yale University), and programs written by Thomas S. Whittam, The Pennsylvania State University.

Results

Nucleotide Polymorphism Among the fimA, papA, *and* sfaA *Genes*

There were both size and sequence polymorphism among the *papA* genes sequenced from six closely related *E. coli* isolates from ECOR groups B2 and D (Table 1). The three *papA* alleles found varied in length within and between ECOR subgroups: 546 bp in ECOR 49 and ECOR 50 (ECOR subgroup D); 525 bp in ECOR 48 (subgroup D), ECOR 52 (subgroup B2), and ECOR 53 (subgroup B2); and 507 bp in ECOR 46 (subgroup D). Among the six *papA* sequences studied, there were 306 polymorphic sites, resulting in 277 amino acid substitutions (Table 1, Fig. 2). For comparative purposes, a region 63 bp downstream of the *papA* gene, *papH* was also sequenced. Among the six *E. coli* isolates examined at *papH*, there were 22 polymorphic sites with nine amino acid substitutions, and six of the amino acids substitutions were in ECOR 49 (Fig. 2). Variation among the *sfaA* genes of the five *E. coli* isolates examined was confined mostly to ECOR 52; of the 132 polymorphic sites in the 537-bp region sequenced, which results in 97 amino acid substitutions (Fig. 2), 128 polymorphic sites were unique to ECOR 52 (Table 1). ECOR 53, ECOR 58, ECOR 64, and ECOR 65 were virtually identical in *sfaA* sequence and each had a 6-bp deletion compared with ECOR 52.

Among the seven *fimA* sequences examined from *E. coli* isolates, there were also size and sequence polymorphism. In the 540–552 bp of *fimA* analyzed, there were 80 polymorphic nucleotides, which resulted in 44 amino acid replacements (Table 1 and Fig. 2).

Two measures of nucleotide diversity were calculated: π, which is the observed average proportion of nucleotide differences between sequences; and θ, which is the average number of polymorphic nucleotides per nucleotide site. All three adhesin loci have a high nucleotide diversity compared with the housekeeping gene *mdh* and other loci previously examined from *E. coli* (Table 1).

Synonymous and Nonsynonymous Substitution Among fimA, sfaA, *and* papA

For *fimA*, *papA*, and *sfaA*, we estimated the number of synonymous substitutions per synonymous site (k_S) and

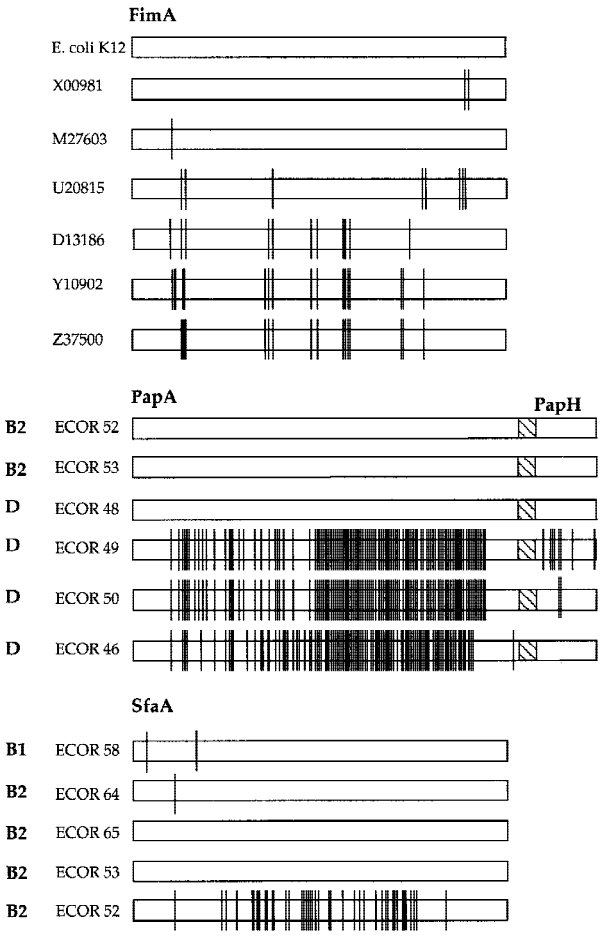

Fig. 2. Comparison of polymorphic amino acid sites among three pilin proteins from closely related natural isolates of *E. coli. Hatched boxes* indicate intergenic region. *Vertical lines* represent amino acid polymorphic sites along the three pilin proteins: FimA, PapA, and SfaA. The strain number on the *left* indicates the ECOR isolates examined and *single uppercase letters* indicate ECOR subgroups.

nonsynonymous substitutions per nonsynonymous site (k_N) for all pairwise comparisons using the Jukes–Cantor correction of proportion of differences. For comparative purposes, we estimated the k_S and k_N of *papH* in the same set of strains examined for *papA* and *sfaA.* Among the three major adhesin proteins studied, k_S was significantly different from other chromosomal genes examined in this species (Table 1). Indeed at the *papA* gene the k_S value of 0.84 ± 0.11 is near saturation, with an appreciable increase in k_N to 0.44 ± 0.03. The k_N/k_S ratio, which is a measure of the selective constraints on a gene, at the *papA* locus is the highest seen for any *E. coli* gene. Also, in both *fimA* and *sfaA*, the levels of synonymous and nonsynonymous site variation are elevated, and similarly, the k_N/k_S ratio was extremely high, 0.22 and 0.29, respectively. Analysis of the *papH* gene, which is 63 bp downstream of *papA*, indicated that this region, too, had high levels of polymorphism; however, from Fig. 2 it can be seen that most of the amino acid variation at the *papH* locus is found in ECOR 49. The polymorphic sites in the *papH* gene of ECOR 49 are confined to approximately the first 100 bp of the gene, and this is indicative of intragenic recombination from an unknown source. When this 100-bp region is removed from the analysis, the remaining region of the *papH* is evolving similarly to *mdh* and other genes from *E. coli* (Table 1).

For comparative purposes levels of nucleotide variation were calculated for 10 genes from natural isolates of *E. coli* (Table 1). Only one locus, *gnd*, shows a level of synonymous site variation similar to that of the three pilin genes; however, at nonsynonymous sites *gnd* shows limited variation similar to all other *E. coli* genes examined here.

Distribution of Polymorphic Nucleotide and Amino Acid Sites

To test for nonrandom clustering of polymorphic nucleotide sites, a pattern that may be indicative of intragenic recombination, we used the Stephens (1989) test. For the 306 polymorphic sites among the six *papA* sequences, 4 statistically significant clusters of polymorphic sites were identified which separated ECOR 46, ECOR 49, and ECOR 50 from all other ECOR strains. Analysis of *papH* by the Stephens test identified one statistically significant partition which involved the 5′ region of the gene in ECOR 49, which we have previously suggested to have been acquired by intragenic recombination. The Stephens test identified 128 unique sites of the *sfaA* sequence from ECOR 52 relative to the others ECOR strains. Among the *fimA* sequences, there were five statistically significant nonrandom runs of polymorphic sites that partition strains D13186, Y10902, and Z37500 from all other *E. coli* sequences.

The distribution of polymorphic amino acids along the protein in both PapA and SfaA was similar, with the majority of amino acid substitutions occurring in the central region of the protein, shown in Fig. 2. The pattern of amino acid polymorphism in the FimA protein was different in that the replacement sites were not concentrated in the central region of the protein. However, among the three adhesin proteins, both the carboxyl and the amino terminals were conserved with few amino acid replacement sites.

For each of the three major pilin genes, *fimA*, *papA*, and *sfaA*, there was marked regional variation in the frequency of both synonymous and nonsynonymous substitutions along the genes (Fig. 3). A striking feature of the regional variation along the *papA* gene is the conspicuous increase in both synonymous and nonsynonymous polymorphism. In the *papA* gene the high peak in the mean proportion of nonsynonymous site differences between pairs of strains in the middle to 3′ region of the gene with replacement sites exceeding silent sites over a 200-bp segment is highly unusual and represents rapid divergence in protein sequence. Further analysis of regional variation in the percentage GC content of *papA*

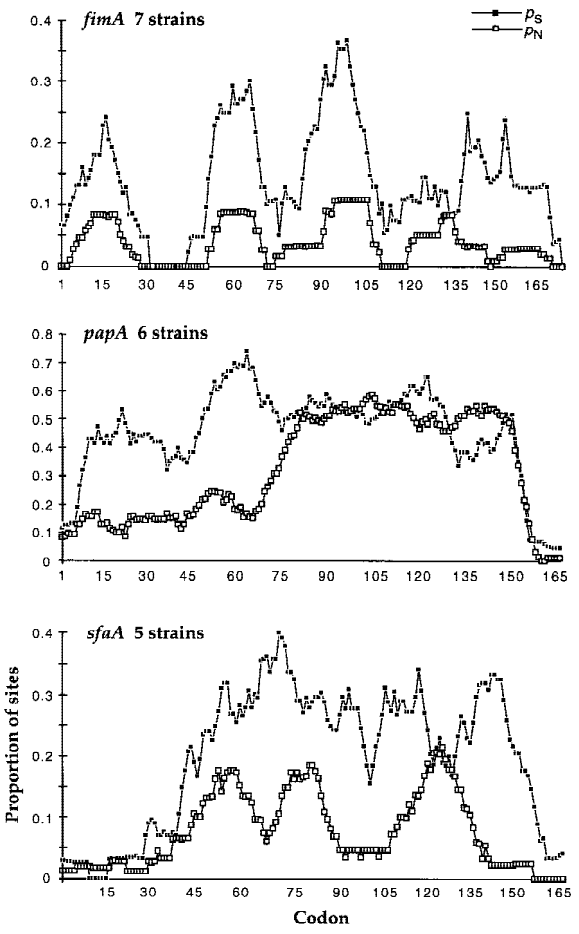

Fig. 3. Distribution of variable sites in the *fimA*, *papA*, and *sfaA* genes among isolates of *E. coli.* Regional variation in the mean proportion of synonymous differences between pairs of strains (p_S) and the mean proportion of nonsynonymous differences between pairs of strains (p_N) based on a sliding window of 45 nucleotides.

and *papH* based on a sliding window of 90 nucleotides indicates that the GC content of the middle to 3′ region of *papA* is higher than the 5′ end of the gene (Fig. 4). Examination of the mean proportion of synonymous and nonsynonymous substitution sites across the *papA* and *papH* genes shows how the levels within *papH* are similar to those of most *E. coli* genes, in contrast to the *papA* gene (Fig. 4).

Codon Usage and GC Content

Within the *FimA*, *PapA*, and *SfaA* protein sequences, all retained two cysteine residues 40 amino acids apart, which probably reflects the constraints imposed by an S–S bridge. We employed three measures to calculate codon bias: the codon adaptation index (AC) (Sharp and Li 1987), codon bias (Shields et al. 1988), and the effective number of codons (ENC) (Wright 1990) for all three adhesin subunit proteins (Table 2). The percentage GC content at each third position and at all positions was also calculated (Table 2).

At the *fimA* locus the GC content for the isolates studied was 52%, which is similar to the overall GC content for *E. coli.* Three measures of codon bias indicate that *fimA* is a moderately expressed gene with no unusual codon usage pattern.

For the ECOR strains examined at the *papA* locus, the CAI was 0.36 for ECOR 48, ECOR 52, and ECOR 53 and 0.33 for ECOR 49 and ECOR 50. However, for ECOR 46 the CAI was only 0.28. The GC content of the *papA* sequences also differed among the ECOR strains, ranging from 41.6% for ECOR 46 to 45.2% for ECOR 49 and ECOR 50 (Table 2). In ECOR 46 there is a statistically significant decrease in the frequency of utilization of the amino acids valine, threonine, glycine, and glutamine. Also, ECOR 46, ECOR 49, and ECOR 50 used codon GCA as the most frequent alanine codon, compared with GCT for ECOR 48, ECOR 52, and ECOR 53. The different patterns of codon usage and GC content between strains may indicate possible intergenic recombination of this gene.

The *sfaA* gene among the five ECOR strains examined had a GC content of approximately 45%, and the CAI ranged from 0.29 for ECOR 52 to 0.33 for all other strains. ECOR 52 also used a different set of preferred codons for arginine (CGA, CGC), glycine (GGG), and proline (CCG) relative to the other ECOR strains. Moreover, ECOR 52 had an increased ENC of 54.2. It is entirely feasible that the *sfaA* gene in ECOR 52 was acquired by horizontal transfer and recombination.

Test for Neutrality at fimA, papA, *and* sfaA

To test for departure from neutrality of molecular polymorphisms among *fimA*, *papA*, and *sfaA*, the Tajima's (1989) *D* statistic was used, which determines if π and θ are significantly different. Under the null hypothesis of selective neutrality, π and θ are equal. From Table 3 two loci, *fimA* and *papA*, show a significant difference between π and θ; in both cases π is smaller, producing a positive *D* value.

Evolutionary Trees for the fimA, papA, *and* sfaA *sequences*

Relationships among the *E. coli* isolates are indicated by neighbor-joining trees (Saitou and Nei 1987) based on synonymous substitution rates using the Jukes–Cantor correction. It is apparent from Fig. 5 that all three major adhesin genes are evolving very differently than *mdh.* Among the *fimA* sequences, two were recovered from avian pathogenic *E. coli* isolates (K12 and U20815) and three were from human uropathogenic *E. coli* isolates (M27603, Y10902, and D13186). The *fimA* gene from the human and avian *E. coli* isolates do not cluster together in the neighbor-joining tree in Fig. 5. Both the *papA* and the *sfaA* sequences are much more diverse than

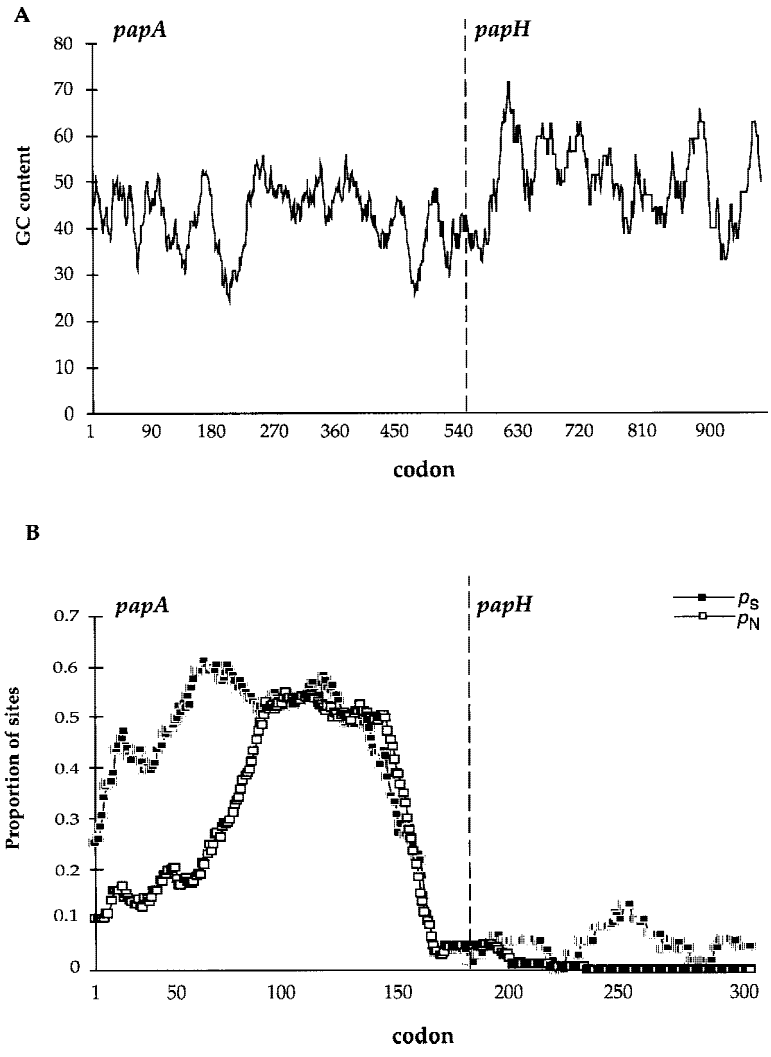

Fig. 4. Comparison of GC content and variable sites in the *papA* and *papH* genes from *E. coli.* **A** Regional variation in the mean proportion of GC content between pairs of strains based on a sliding window of 90 nucleotides. **B** Regional variation in the mean proportion of synonymous differences between pairs of strains (p_s) and the mean proportion of nonsynonymous differences between pairs of strains (p_N based on a sliding window of 90 nucleotides in *papA* and *papH* genes.

the *fimA* gene. It is of interest to note that the *fimA* gene is found in most *E. coli* isolates, whereas both *papA* and *sfaA* are usually associated only with uropathogenic *E. coli* strains.

Discussion

Bacterial genes whose products are subject to diversifying selection in adaptation to host defense systems or other variable aspects of the environment show a pattern of nucleotide and amino acid polymorphism very different from those of housekeeping genes. In the major structural pilin genes *fimA*, *papA*, and *sfaA*, we found levels of both synonymous and nonsynonymous nucleotide substitution higher than those observed for chromosomal genes for which comparable data are available (Table 1). The pattern of extremely high intraspecific polymorphism revealed is similar to that found at loci that are subjected to selective pressure such as flagellin genes in *Salmonella* (Selander and Smith 1990). For comparative purposes we analyzed data from 10 *E. coli* genes for which sequence data from ECOR strains were available (Table 1). Of the genes examined only *gnd* has a level of synonymous sites polymorphism that is comparable to the major pilin genes, however, nonsynonymous site variation is similar to that of other *E. coli* genes (Table 1). Further, it has been shown that the *gnd* locus, which is located near the *rfb* O antigen locus has been subject to frequent recombination events from distantly related types of bacteria (Bisercic et al. 1991; Nelson and Selander, 1994), and this accounts for the diversity seen at this locus.

The function of fimbriae adhesin in bacteria may be threefold: first, pilin allows attachment of the bacterium to the host cell, the first step in the pathogenic process; second, different pilin variants may be better adapted to different environments encountered by the bacterium in their host; and finally, new pilin variants may confer a selective advantage to an isolate due to the increase ability in avoiding the host defense system.

The three adhesin proteins analyzed in this study play an important role in attachment of the bacterium to the host cell and there may be a selective advantage to di-

Table 2. Codon bias and GC content within three adhesin pilin genes of *E. coli*

Gene	Strain	GC content		Codon bias[a]	CAI[b]	ENC[c]
		3rd position	Total			
fimA	K-12	43.7	51.7	0.58	0.40	33.4
	M27603	44.3	51.7	0.58	0.40	33.4
	X00981	44.0	51.8	0.57	0.40	33.5
	U20815	45.3	51.6	0.56	0.43	36.0
	D13186	45.4	52.3	0.50	0.44	37.0
	Y10902	43.0	51.0	0.47	0.42	41.3
	Z37500	42.7	50.5	0.41	0.38	41.0
papA	ECOR 48	31.4	44.4	0.34	0.36	46.4
	ECOR 52	31.4	44.4	0.34	0.36	46.4
	ECOR 53	31.4	44.4	0.34	0.36	46.4
	ECOR 49	37.9	45.2	0.30	0.33	46.6
	ECOR 50	37.9	45.2	0.30	0.33	46.3
	ECOR 46	26.6	41.6	0.25	0.28	47.2
sfaA	ECOR 53	33.3	45.2	0.25	0.33	46.2
	ECOR 64	32.8	45.2	0.26	0.33	45.8
	ECOR 65	32.8	45.2	0.25	0.33	46.0
	ECOR 58	32.8	44.8	0.26	0.33	45.8
	ECOR 52	34.1	44.9	0.18	0.29	54.2

[a] Scaled χ^2/L (L, total no. of codons minus W, M, and stop codons).
[b] Codon adaptation index (Sharp et al. 1987).
[c] Effective number of codons used.

Table 3. Tajima's (1989) *D* test for significant differences between π and θ

Gene	Total	Silent
fimA	0.3	3.2[b]
papA	2.3[a]	27.6[b]
sfaA	−1.2	1.3
mdh	0.3	0.3
papH	−0.3	1.4

[a] $P < 0.01$.
[b] $P < 0.001$.

versity. The type-1 fimbriae consist of identical subunits, FimA, held together in a fixed pattern by subunit–subunit interactions. It is therefore expected that a large fraction of a each subunit will be hidden from the surrounding environment and not subject to the host immune system but specific regions will be exposed to the host defense. The five predicted antigenic determinants of FimA (Klemm 1984) are precisely those segments of the peptide sequence in which we found clusters of amino acid polymorphisms (Figs. 2 and 4). Furthermore, it has been proposed that the C-terminus of FimA is the region that is implicated in the subunit-subunit interaction contributing to the structural integrity of the fimbriae or the fixing of the pili to the cell (Båga et al. 1985), and it is this region among the isolates that is highly conserved with little or no amino acid polymorphisms. The major subunits of P fimbria, PapA, are arranged in a helix to form a rigid rod-shaped structure with a terminal fibrillar structure (Gaastra and Svennerholm 1996). Although the role of attachment is an essential one, the pili may play another important function, that of avoiding the host immune system by presenting alternative protein sequences to the host defense system and this may explain the extraordinary high levels of intraspecific polymorphism we observe at these loci and the regional variation within each of the genes (Fig. 4).

Regional variation in synonymous and nonsynonymous substitution levels among the major pilin genes was examined based on a sliding window of 45 nucleotides (Fig. 3). There is extreme fluctuation in variation within each gene. Among the seven *fimA* sequences examined elevated levels of both synonymous and nonsynonymous substitutions occur in five regions—those predicted to be antigenic. The absence of both synonymous and nonsynonymous variation at the 5′ region of the *fimA* gene probably reflects a high functional constraint on this part of the encoded protein (Fig. 4). In the *papA* gene in the 3′ region of the gene nonsynonymous substitution outnumber synonymous variation in an approximately 200-bp region, this hypervariable region we propose is the region presented on the surface of the pilin (Fig. 3). The *sfaA* gene has a 5′ region that is highly conserved, which may again reflect functional constraints. There is subsequently an increase in both synonymous and nonsynonymous substitution levels with a region where nonsynonymous exceed synonymous variation. This regional variation in substitution levels plays a role in antigenic diversity and host immune system avoidance which strongly supports positive Darwinian selection acting at these sites.

An alternative explanation that could account for the high levels of substitution is the role of horizontal transfer and recombination of divergent sequences at these loci. However, one would have to evoke three independent horizontal transfer events occurring within *sfaA*, *papA*, and *fimA*, at 26, 55 and 98 minutes on the *E. coli* chromosome, with each recombination event occurring precisely within the major structural pilin gene. Gene features indicative of recombination are an unusual GC content and codon usage patterns. The codon adaptation index (CAI), one measure of codon bias, is also associated with levels of gene expression; highly expressed genes use a limited number of codons, whereas lowly expressed genes use codons more randomly. Among the seven *fimA* sequences examined all had a GC content of approximately 52% and a CAI that ranged from 0.38 to 0.44, suggesting moderate expression of this gene in *E. coli*. However, for the *sfaA* locus the percentage GC content was 44 and the gene had a CAI of 0.33; moreover, in ECOR 52, the CAI was only 0.29 and the effective number of codons (ENC) was high (54.2), reflecting an essentially unbiased codon usage. Both *papA* and *sfaA* have a GC content significantly different for the overall 52% of *E. coli*, and among the ECOR strains analyzed at the *papA* locus there was an even greater

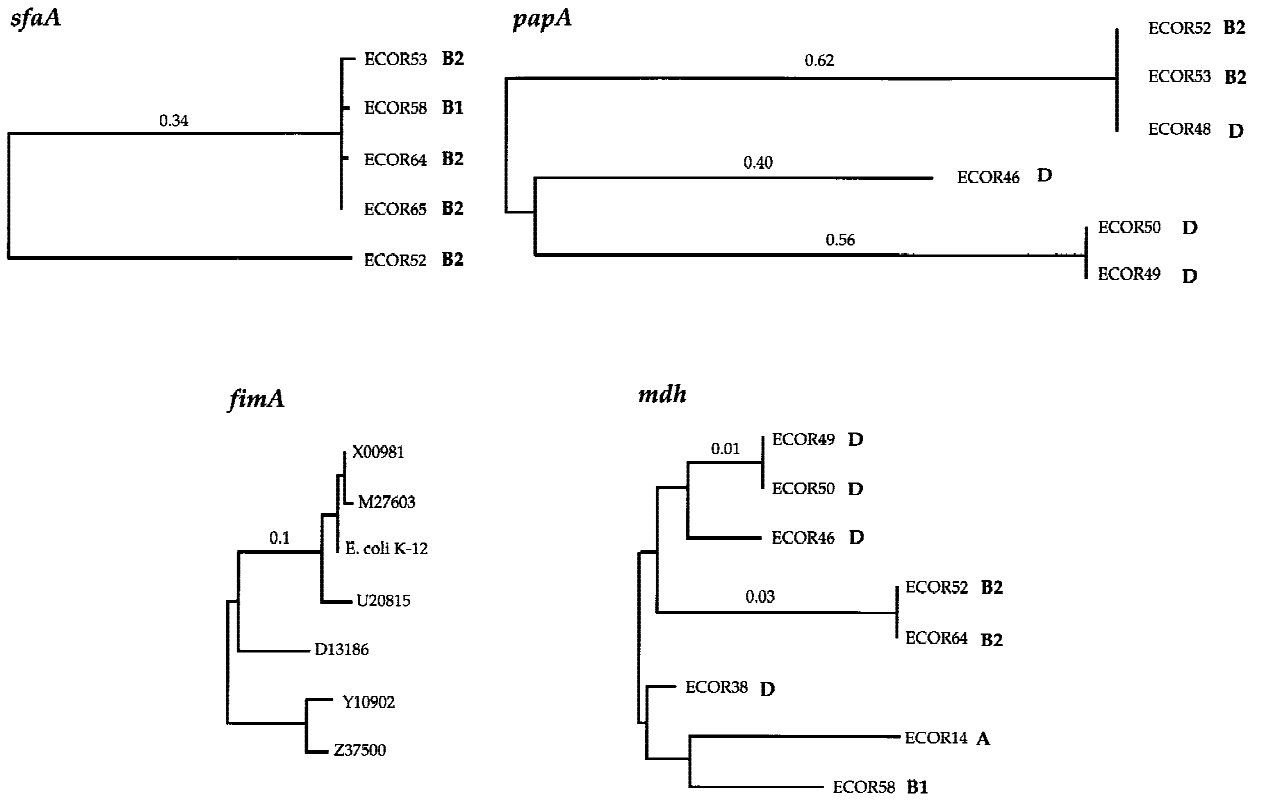

Fig. 5. Individual phylogenetic gene trees for *fimA*, *papA*, *sfaA*, and *mdh* representing the evolutionary relationships of the sequences among the *E. coli* isolates based on synonymous site variation. The strain number on the *right* indicates the ECOR isolate and the subgroup designation of the isolate.

difference in GC content and CAI in ECOR 46 (Table 2). In ECOR 46 the percentage GC content at third position sites was 26.6, which differs considerable from the mean of 33.0, and similarly for all sites, the GC content was 41.6 for ECOR 46 compared to a mean of 44.2. Therefore, one cannot entirely discount a possible role for horizontal DNA transfer and recombination in accounting for some of the diversity seen at the *papA* and *sfaA* loci. Other studies have shown the presence of sequences homologous to *papA* in a diverse range of species such as *Proteus mirabilis*, *Serratia marcescens*, and *Haemophilus influenzae*. Molecular phylogenetic analysis indicates that there was horizontal transfer of the major structural pilin gene among these species (Bijlsma et al. 1995). Further, Marklund and colleagues (1992) have proposed lateral transfer as a mechanism in *pap* gene cluster evolution in *E. coli.*

We examined the *papH* locus 63 bp downstream of *papA* to determine nucleotide diversity and patterns of polymorphism, in order to clarify further the role of horizontal transfer and recombination in generating pilin variants (Fig. 4). From Table 1, at the *papH* locus, both synonymous and nonsynonymous site variations are slightly elevated compared with a typical housekeeping gene such as *mdh*, which was analyzed from the same set of strains as at the *papA* and *sfaA* loci. Comparative sequence analysis revealed that most of the variation at the *papH* locus is found in one strain, ECOR 49. When this isolate was removed from the analysis, *papH* shows a level of variation very similar to that of *mdh.* Of the 19 polymorphic sites in *papH*, 13 where found in ECOR 49 in the first 100 bp of the gene; removal of this region from the analysis gave levels of nucleotide sequence variation similar to *mdh.* Further comparison of the level of synonymous and nonsynonymous substitution within *papA* and *papH* revealed the very stark differences in variation (Fig. 4), contrasting principally in the near-sequence identity observed in the flanking region of *papA* and the entire *papH* gene. This pattern is also observed in the *fimA* and *sfaA* genes (Fig. 3). It is unlikely that there has been enough time for multiple recombination events to have occurred to homogenize the 5′ and and 3′ regions of these genes.

A third hypothesis to explain the unusually high levels of nonsynonymous substitution found in the three adhesin genes is that it may reflect low levels of functional constraint on the FimA, PapA, and SfaA proteins. The neutral mutation hypothesis (Kimura 1983) predicts that a protein experiencing stringent functional constraints accumulates less nonsynonymous substitutions than a protein experiencing less constraints. However, relaxed constraints at the protein level should not affect the rate of substitution at the synonymous sites and therefore could not account for the high levels of synonymous

substitutions found in all three adhesins (Table 1). Furthermore, relaxed constraint at the protein level also does not account for the low levels of variation found at both the 5′ and the 3′ ends of *fimA, papA,* and *sfaA.*

Of the three adhesin loci examined, only *fimA* is found in the majority of *E. coli* isolates, which is indicative of its presence in the most recent common ancestor of the species. The P and S pili are confined to specific lineages of *E. coli:* the *sfa* operon was found only within ECOR subgroup B2 strains, and, similarly, the *pap* gene cluster was identified in strains of subgroup B2 and D, with sporadic occurrence in subgroup A isolates (Boyd and Hartl 1998). Thus, in contrast to *fimA,* it appears that the P and S gene clusters were acquired after speciation and hence have a limited distribution in *E. coli.* As a mechanism for colonizing new hosts and perhaps helping to circumvent the host immune system, the acquisition of these loci in specific lineages of *E. coli* may be a contributing factor in the emergence of a pathogen from a typically commensal organism.

The *fim, pap,* and *sfa* loci impart upon the organism an ability to invade a host, and a strain which possesses the ability to also avoid the host immune system must be at a selective advantage in certain cases. A scenario of frequency-dependent selection, as proposed for colicin immunity proteins and restriction-modification genes, is envisioned here where variants due to mutation in the major pilin genes are chosen for their selective advantage and subsequently having reached high frequency in the *E. coli* population are maintained by frequency-dependent selection in a particular niche where they provide a selective advantage due to their ability of host immune avoidance. It is difficult to envision a recombinational mechanism that can account for the regional patterns of diversity found within the major pilin genes. Certainly recombination could account for some of the variation found, but the accelerated divergence in different regions of the genes suggests some form of positive selection playing a role in antigenic diversity.

Acknowledgment. This work was support by an NIH grant to D.L.H.

References

Achtman M, Hakenbeck R (1992) Recent developments regarding the evolution of pathogenic bacteria. In: Hormache CE, Penn CW, Smyth CJ (eds) Molecular biology of bacterial infection: current status and future perspectives. Cambridge University Press, New York

Båga MS, Normark J, Hardy P, O'Hanley D, Lark O, Olsson G, Schoolonik, Falkow S (1984) Nucleotide sequence of the *papA* gene encoding the *pap* pilin subunit of human uropathogenic *Escherichia coli.* J Bacteriol 157:330–333

Barkus VA, Titheradge AJ, Murray NE (1995) The diversity of alleles at the *hsd* locus in natural populations of *Escherichia coli.* Genetics 140:1187–1197

Bijlsma IG, van Dijk WL, Kusters JG, Gaastra W (1995) Nucleotide sequences of two fimbrial major subunit genes, *pmpA* and *ucaA,* from canine-uropathogenic *Proteus mirabilis* strains. Microbiology 141:1349–1357

Bisercic M, Feutrier JY, Reeves PR (1991) Nucleotide sequence of the *gnd* genes from nine natural isolates of *Escherichia coli:* evidence of intragenic recombination as a contributing factor in the evolution of the polymorphic *gnd* locus. J Bacteriol 173:3894–3900

Boyd EF, Hartl DL (1998) Chromosomal regions specific to pathogenic isolates of *Escherichia coli* have a phylogenetically clustered distribution. J Bacteriol 180:1159–1165

Boyd EF, Wang F-S, Whittam TS, Nelson K, Selander RK (1994) Molecular genetic basis of allelic polymorphism in malate dehydrogenase (*mdh*) in natural isolates of *Escherichia coli* and *Salmonella* enterica. Proc Natl Acad Sci USA 91:1280–1284

Boyd EF, F-S Wang, TS Whittam, Selander RK (1996) Molecular genetics relationships of the salmonellae. Appl Envir Microbiol 62:804–808

Boyd EF, J Li J, Ochman H, Selander RK (1997) Comparative genetics of the *inv-spa* invasion gene complex of *Salmonella enterica.* J Bacteriol 179:1985–1991

Dubose F, Dykhuizen DE, Hartl DL (1988) Genetic exchange among natural isolates of bacteria: recombination within the *phoA* gene of *Escherichia coli.* Proc Natl Acad Sci USA 85:7036–7040

Gaastra W, Svennerholm A (1996) Colonization factors of human enterotoxigenic *Escherichia coli* (ETEC). Trends Microbiol 4:444–452

Guttman DS, Dykhuizen D (1994). Detecting selective sweeps in naturally occurring *Escherichia coli.* Genetics 138:993–1003

Hacker J (1992) Role of fimbrial in the pathogenesis of *Escherichia coli* infections. Can J Microbiol 38:720–727

Hacker J, Oehler-Blum G, Mühlodorfer I, Tschäpe H (1997) Pathogenicity islands of virulent bacteria: Structure, function, and impact on microbial evolution. Mol Microbiol 23:1089–1097

Hall BG, Sharp PM (1992) Molecular population genetics of *Escherichia coli:* DNA sequence diversity at the *celC, crr,* and *gutB* loci of natural isolates. Mol Biol Evol 9:654–665

Herzer PJ, Inouye S, Inouye M, Whittam TS (1990) Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli.* J Bacteriol 172:6175–6181

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167–170

Hultgren SJ, Jones CH, Normark S (1996) Bacterial adhesins and their assembly. In: Neidhardt FC, J LI, Lin ECC, Low KB, Magasanik B, Reznikoff WS, Riley M, Schaechter M, Umbarger HE (eds) *Escherichia coli* and *Salmonella:* cellular and molecular biology. American Society for Microbiology Press, Washington, DC

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Klaasen P, der Graaf FL (1990) Characterization of FapR, a positive regulator of expression of the 987P operon in enterotoxigenic *Escherichia coli.* Mol. Microbiol 4:1779–1779

Klemm P (1984) The *fimA* gene encoding the type-1 fimbrial subunit of *Escherichia coli.* Eur J Biochem 143:395–399

Kumar S, Tamura K, Nei M (1993) MEGA: molecular evolutionary genetics analysis, Version 1.0. Pennsylvania State University, Philadelphia

Li J, Ochman H, Groisman EA, Boyd EF, Solomon F, Nelson K, Selander RK (1995) Relationship between evolutionary rate and cellular location among the Inv/Spa invasion proteins of *Salmonella enterica.* Proc Natl Acad Sci USA 92:7252–7256

Lui D, Reeves PR (1994) Presence of different O antigen forms in three isolates of one clone of *Escherichia coli.* Genetics 138:7–10

Marc D, Dho-Moulin M (1996) Analysis of the *fim* cluster of an avian 02 strain of *Escherichia coli:* serogroup-specific sites within *fimA* and nucleotide sequence of *fimI.* J Med Microbiol 44:444–452

Marklund BI, Tennent JM, Garcia E, Hamers A, Baga M, Lindberg F, Gaastra W, Normark S (1990) Horizontal gene transfer of the *Esch-*

erichia coli pap and *prs* pili operons as a mechanism for the development of tissue-specific adhesive properties. Mol Microbiol 6:2225–2242

Milkman R, Bridges MM (1990) Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. Genetics 126:505–517

Milkman R, Bridges MM (1993) Molecular evolution of the *Escherichia coli* chromosome. IV. Sequence comparisons. Genetics 133:455–468

Milkman R, Stoltzfus A (1988) Molecular evolution of the *Escherichia coli* chromosome. II. Clonal segments. Genetics 120:359–366

Nelson K, Selander RK (1994) Intergenic transfer and recombination of the 6-phosphogluconate dehydrogenase gene (*gnd*) in enteric bacteria. Proc Natl Acad Sci USA 91:10227–10231

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

Nelson K, Wang FS, Boyd EF, Selander RK (1997) Size and sequence polymorphism in the isocitrate dehydrogenase kinase/phosphatase gene (*aceK*) and flanking regions in *Salmonella enterica* and *Escherichia coli.* Genetics 147:1509–1520

Ochman H, and Selander RK (1984) Standard reference strains of *Escherichia coli* from natural populations. J Bacteriol 157:690–693

Orndorff P, Falkow S (1985) Nucleotide sequence of *pilA*, the gene encoding the structural component of type 1 pili in *Escherichia coli.* J Bacteriol 162:454–457

Pupo GM, Karaolis DK, Lan R, Reeves PR (1997) Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and *mdh* sequence studies. Infect Immun 65:2685–2692

Reeves PR (1992) Variation in O-antigen, niche-specific selection and bacterial populations. FEMS Microbiol Lett 100:509–516

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sekizaki T, Ito H, Asawa T, Nonomura I (1993) DNA sequence of type-1 fimbrin, Fpull gene from a chicken pathogenic *Escherichia coli* serotype 078. J Vet Med Sci 55:395–400

Selander RK, Smith NH (1990) Molecular population genetics of *Salmonella.* Rev Med Microbiol 1:219–228

Selander RK, Li J, Boyd EF, Wang FS, Nelson K (1994) DNA sequence analysis of the genetic structure of populations of *Salmonella enterica* and *Escherichia coli*, p. 17–49. In Priest FG, Ramos-Cormenzana A, Tindall BJ (eds), Bacterial diversity and systematics. Plenum Press, New York

Sharp PM, Li WH (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol 4:222–230

Sharp PM, Kelleher JE, Daniel AS, Cowan GM, Murray NE (1992) Roles of selection and recombination in the evolution of type I restriction-modification systems in enterobacteria. Proc Natl Acad Sci USA 89:9836–9840

Shields DC, Sharp PM, Higgins DG, Wright F (1988) Silent sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5:704–716

Stephens JC (1985) Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol Biol Evol 2:539–556

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Tan Y, Riley MA (1996) Rapid invasion by colicinogenic *Escherichia coli* with novel immunity functions. Microbiology 142:2174–2180

Tanaka T, Nei M (1989) Positive Darwinian selection observed at the variable-region genes of immunoglobin. Mol Biol Evol 6:447–459

van Die I, Geffen B, Hoekstra W, Bergmans HEN (1984) Type 1C fimbriae of a uropathogenic *Escherichia coli* strain: cloning and characterization of the gene involved in the expression of the 1C antigen and nucleotide sequence of the subunit gene. Gene 34:187–196

Wang FS, Whittam TS, Selander RK (1997) Evolutionary genetics of the isocitrate dehydrogenase gene (*icd*) in *Escherichia coli* and *Salmonella enterica.* J Bacteriol 179:6551–6559

Wright F (1990) The ‘‘effective number of codons’’ used in a gene. Gene 87:23–29