Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 9836–9840, October 1992
Genetics

Roles of selection and recombination in the evolution of type I restriction–modification systems in enterobacteria

(molecular evolution/restriction enzymes/frequency-dependent selection/horizontal transfer/*Enterobacteriaceae*)

PAUL M. SHARP*, JULIA E. KELLEHER†, ANNE S. DANIEL†, GILL M. COWAN†, AND NOREEN E. MURRAY†

*Department of Genetics, Trinity College, Dublin 2, Ireland; and †Institute of Cell and Molecular Biology, University of Edinburgh, Edinburgh EH9 3JR, Scotland

Communicated by J. Maynard Smith, June 12, 1992

ABSTRACT **Restriction–modification systems can protect bacteria against viral infection. Sequences of the *hsdM* gene, encoding one of the three subunits of type I restriction–modification systems, have been determined for four strains of enterobacteria. Comparison with the known sequences of *Eco*K and *Eco*R124 indicates that all are homologous, though they fall into three families (exemplified by *Eco*K, *Eco*A, and *Eco*R124), the first two of which are apparently allelic. The extent of amino acid sequence identity between *Eco*K and *Eco*A is so low that the genes encoding them might be better termed pseudoalleles; this almost certainly reflects genetic exchange among highly divergent species. Within the *Eco*K family the ratio of intra- to interspecific divergence is very high. The extent of divergence between the genes from *Escherichia coli* K-12 and *Salmonella typhimurium* LT2 is similar to that for other genes with the same level of codon usage bias. In contrast, intraspecific divergence (between *E. coli* strains B and K-12) is extremely high and may reflect the action of frequency-dependent selection mediated by bacteriophages. There is also evidence of lateral transfer of a short sequence between *E. coli* and *S. typhimurium*.**

Hundreds of different sequence-specific restriction–modification (R-M) systems have been identified in bacteria (1). We are concerned with type I R-M systems (reviewed in refs. 1 and 2). These enzymes comprise three different subunits encoded by adjacent genes; the archetypal system is *Eco*K encoded by the *hsdR*, *hsdM*, and *hsdS* genes, located at 99 min on the *Escherichia coli* K-12 chromosome. The resulting complex is both an endonuclease and a methyltransferase. One of the three polypeptides (S) dictates the sequence specificity of the enzyme, so that merely changing the S subunit generates an enzyme that recognizes a different sequence of nucleotides. As a consequence, type I R-M systems have greater potential for evolutionary diversification than do the type II systems, in which the modification and restriction enzymes must share a common recognition sequence but are encoded by separate genes. Type I systems are the only R-M systems observed to undergo major changes in specificity, and families of related enzymes conferring different specificities could evolve by changes in the S gene.

Relatives of *Eco*K have been identified in different strains of *E. coli*, and in different serotypes of the closely related enteric *Salmonella typhimurium*. Evidence for relatedness relied initially on complementation tests that indicated the exchange of subunits between enzymes conferring different specificities, and subsequently on molecular comparisons involving nucleic acid hybridization and antibody crossreactivity (see ref. 2). On the basis of the same criteria a second family has been identified, including two members from *E. coli* (*Eco*A and *Eco*E) (2) and one from *Citrobacter freundii* (3). The chromosomal location of the genes for *EcoK*, *EcoA*, and relatives appears to be the same (3, 4). The genes for members of a third family of type I R-M systems (*Eco*R124 and *Eco*DXXI) are plasmid borne (see ref. 2).

The complete coding sequence for *Eco*K is known (5, 6), as is that for *Eco*R124 (7). The specificity genes for many type I R-M systems have been sequenced and compared: the polypeptides have two recognition domains, each defining one component of a bipartite target sequence (4, 6, 8). Recombination between different S genes can reassort recognition domains and generate new specificities. This has been demonstrated for members of both the *Eco*K (9) and *Eco*R124 families (10). [A second kind of specificity change in *Eco*R124 seems to have arisen by unequal crossing-over within a short duplicated sequence in the S gene (7).]

In this paper we report the nucleotide sequences for the *hsdM* genes of *Eco*A and three additional members of the K family. By comparisons of the nucleotide and predicted amino acid sequences, we establish that all three families of genes are homologous and we investigate the evolutionary processes involved in their divergence.‡

MATERIALS AND METHODS

Bacterial Strains, Media, and Microbial Techniques. The general host for phages (λ and M13) and plasmids was NM522 [(*lac–pro*)Δ *hsdMS*Δ/F′ *lacZ*ΔM15 *lacI*^{q}] (6). The *hsdM* genes for the *Eco*B (from *E. coli* B), *Sty*SB (*S. typhimurium* LT2), and *Sty*SP (*S. typhimurium* serotype *potsdam*) systems were cloned in the λ vector NM1149 (11) by A. J. B. Campbell (Edinburgh). The other λ*hsd* phages used and the derivative plasmid pFFP32 (containing the *hsdM* gene of *Eco*A) have been described (12). Media, general methods, and tests for restriction and modification were as before (12).

DNA Manipulations. Preparation, manipulation, and recovery of DNA were as described (13). Template DNA was sequenced by the dideoxy chain-termination method (14) using deoxyadenosine 5′-[α-[^{35}S]thio]triphosphate, and the reactions were analyzed by electrophoresis in buffer gradient gels (15). Most recently T7 polymerase has been used, following the supplier's recommendations (United States Biochemical).

Restriction fragments of *hsdM* genes were subcloned in M13 vectors, and oligonucleotide primers at intervals of about 250 bases were used to prime sequencing reactions. Sequences were already available for the 3′ ends of the *hsdM* genes of *Eco*B, *Sty*SB, and *Sty*SP; in each case the sequence determined fully overlapped that previously published (6, 9, 16). All sequences were determined for both strands.

Sequence Analysis. Protein sequences were aligned by the CLUSTAL multiple alignment program (17). Numbers of amino acid replacements per site between aligned sequences were

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: R-M, restriction–modification.

‡The sequences reported in this paper have been deposited in the GenBank/EMBL data base (accession nos. L02505–L02508).

estimated by the empirical method of Kimura (equation 4.4 in ref. 18). DNA sequence divergence was analyzed by the approach of Li *et al.* (19). In this method, rates of transitions and transversions are estimated separately for 0-fold, 2-fold, and 4-fold degenerate sites and then combined with appropriate weighting to provide estimates of the numbers of nucleotide substitutions per synonymous (K_S), and per nonsynonymous (K_A) site.

RESULTS AND DISCUSSION

Homology of M Proteins. The nucleotide sequences of the *hsdM* genes from *E. coli* strains B (which encodes the *Eco*B system) and 15T$^-$ (*Eco*A), and *S. typhimurium* LT2 (*Sty*SB), and *S. typhimurium* serotype *potsdam* (*Sty*SP) have been determined (the nucleotide sequences are not shown, but have been deposited in the GenBank/EMBL data base). These sequences can be compared with the *hsdM* sequences from *E. coli* K-12 (*Eco*K; ref. 5) and the plasmid-encoded *Eco*R124 (7). Using several criteria (described above), these different systems have previously been grouped into three families, exemplified by *Eco*K (*Eco*B, *Eco*K, *Sty*SB, and *Sty*SP), *Eco*A, and *Eco*R124.

There has been some doubt about whether the two known M gene sequences (*Eco*K and *Eco*R124) are strictly homologous (i.e., sharing a common ancestor), since their protein sequences are reported (7) to be similar only within a region of about 12 amino acids common to all known *E. coli* adenine methyltransferases. We have aligned representative M protein sequences of the K, A, and R124 families: they are highly divergent, but there are several regions of sequence similarity among all three throughout the central region of the polypeptide (Fig. 1). The *Eco*K and *Eco*A M proteins share 32% amino acid sequence identity, a value well above the "twilight zone" (20) in which it is difficult to differentiate convergence and divergence. The sequence identity between *Eco*R124 and *Eco*K is less (26%), but the areas of similarity are specifically in regions where *Eco*K and *Eco*A are conserved (Fig. 1). Therefore, it seems most likely that the three families have diverged from a distant common ancestor.

FIG. 1. Alignment of the M proteins of the *Eco*A, *Eco*K, and *Eco*R124 R-M systems. Amino acid residues identical between *Eco*A and *Eco*K or between *Eco*K and *Eco*R124 are denoted by stars above or below the *Eco*K sequence, respectively.

The extent of amino acid identity among all the aligned M protein sequences confirms that they unambiguously fall into the expected three families (Table 1). Among members of the K family there is <6% amino acid sequence difference, and it is possible to make comparisons at the DNA sequence level (discussed below); members of different families are far too divergent for such an analysis.

***Eco*K Family: Interspecific Divergence.** The nature of the divergence between *E. coli* K-12 and *S. typhimurium* LT2 is better characterized than for any other pair of bacterial species (21). Amino acid sequence identity between these two species varies from about 67% up to 100%, with the average close to 90% (21). Here, the M polypeptides of *E. coli* K-12 and *S. typhimurium* LT2 are identical at 94.3% of residues, and the average interspecific identity within the K family is 94.5% (Table 1); thus, these are quite highly conserved proteins.

The extent of divergence between *E. coli* and *S. typhimurium* at synonymous (or silent) sites also varies considerably among genes. Two distinct factors that influence this variability have been identified. The first, and more important, is the strength of codon usage bias in a gene: genes that are more highly expressed have stronger codon usage bias (22) and have accumulated fewer synonymous differences during the divergence of these two species (21). A secondary factor is map location (21), but that is unimportant in the current context. The extents of interspecific nucleotide divergence among the K-family *hsdM* genes, measured as the estimated numbers of nucleotide substitutions per synonymous (K_S) and per nonsynonymous (K_A) site (19) are given in Table 2. The K_S value for *hsdM* between *E. coli* K-12 and *S. typhimurium* LT2 is 0.50, and the average K_S value in interspecific comparisons is 0.53. These values are much less than the average for 67 genes (0.94) previously reported for *E. coli* and *S. typhimurium* (21). The *hsdM* gene is not known to be particularly highly expressed, but the codon usage bias (measured by the codon adaptation index, CAI; ref. 23) of the *hsdM* genes (average CAI = 0.49) is higher than average (among 1038 *E. coli* genes the mean CAI is 0.377, with a standard deviation of 0.135; A. T. Lloyd and P.M.S., unpublished data), suggesting that *hsdM* codon usage is under some selective constraint. When the K_S values for different genes are plotted as a function of CAI (as in figure 1 of ref. 21), the K_S for *hsdM* is close to the value expected for a gene with this CAI value. That is, the extent of substitution at silent sites is quite compatible with the divergence of *E. coli* K-12 and *S. typhimurium* LT2 *hsdM* genes from a sequence on the chromosome of the common ancestor of these two species.

Table 1. Divergence among M proteins of type I R-M systems

	*Eco*B	*Eco*K	*Sty*SB	*Sty*SP	*Eco*A	*Eco*R124
*Eco*B	—	96.6	94.5	94.9	32.3	26.2
*Eco*K	0.04	—	94.3	94.3	31.9	26.4
*Sty*SB	0.06	0.06	—	95.3	32.9	25.5
*Sty*SP	0.05	0.06	0.05	—	32.5	26.0
*Eco*A	1.46	1.49	1.49	1.45	—	22.9
*Eco*R124	1.89	1.89	1.95	1.86	2.23	—

Values indicate percent amino acid sequence identity (above the diagonal) and estimated number of amino acid replacements per site (18) after correction for multiple hits (below the diagonal).

Table 2. Nucleotide sequence divergence of *hsdM* within the K family

	(CAI*)	*Eco*B	*Eco*K	*Sty*SB	*Sty*SP
*Eco*B	(0.49)	—	0.017 [0.013]	0.030 [0.026]	0.026 [0.024]
*Eco*K	(0.47)	0.24 [0.21]	—	0.033 [0.032]	0.030 [0.027]
*Sty*SB	(0.51)	0.45 [0.44]	0.50 [0.49]	—	0.026 [0.025]
*Sty*SP	(0.50)	0.53 [0.56]	0.62 [0.59]	0.29 [0.28]	—

Estimated numbers of nucleotide substitutions per nonsynonymous site (K_A; above the diagonal) and per synonymous site (K_S; below the diagonal) are given; see text and ref. 21 for details of calculation. Values in brackets exclude codons 78–96, which may have been involved in a recombination event (see text).
*Codon adaptation index (see text and ref. 23).

***Eco*K Family: Intraspecific Divergence.** The extent of intraspecific divergence at *hsdM* is high. Among 29 genes that can be compared between *E. coli* strains K-12 and B (or B/r), *hsdM* is the most divergent in terms of both synonymous and nonsynonymous substitutions (Fig. 2). Compared to the average of the other 28 sequences, the M protein is >7 times as divergent, while the *hsdM* DNA sequence is >4 times as divergent. This is particularly surprising because in the interspecific comparisons *hsdM* is more conserved than average (see above). An unusually high divergence could result from lateral transfer into *E. coli* K-12 or B from a more divergent strain. In the next section, we suggest that there has indeed been such a transfer into an ancestor of the *Eco*B gene. However, the putative transfer involved only a short sequence (50–100 base pairs) and *Eco*B–*Eco*K divergence values calculated by excluding this segment are still exceptionally high (Table 2). For a number of genes it is possible to examine the extent of intraspecific divergence (between *E. coli* B and K-12) relative to the level of interspecific divergence (between *E. coli* and *S. typhimurium* LT2). The values for the ratio of intra- to interspecific divergence (Table 3) further emphasize that the intraspecific divergence of *hsdM* is exceptionally high. The relationships among a large number of *E. coli* strains have been investigated by multilocus enzyme electrophoresis; strains K-12 and B are relatively closely related within one subgroup (24). Therefore, other *E. coli* strains may harbor K-family *hsdM* sequences that are even more divergent.

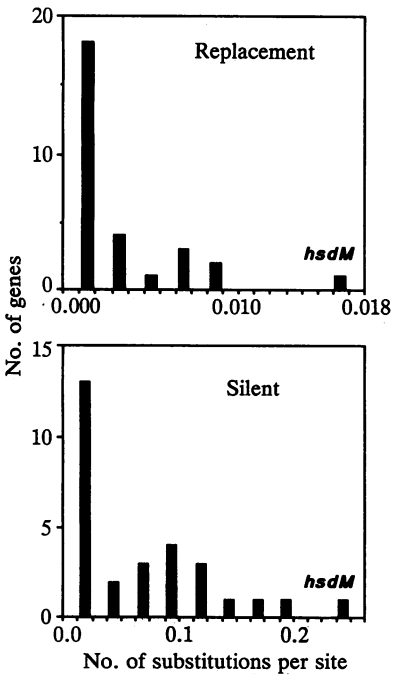

FIG. 2. Histograms of divergence levels at nonsynonymous (*Upper*) and synonymous (*Lower*) sites for 29 genes compared between *E. coli* strains B and K-12. Numbers of nonsynonymous (K_A) and synonymous (K_S) nucleotide substitutions per site were estimated by the method of Li *et al.* (19). The genes compared are *araA*, *araC*, *araD*, *alkB*, *ada*, *tyrA*, *recA*, *cysH*, *cysI*, *argR*, *gltS*, *ilvL*, *polA*, *phnC*, *phnD*, *phnE*, *phnF*, *phnG*, *phnH*, *phnI*, *phnJ*, *phnK*, *phnL*, *phnM*, *phnN*, *phnO*, *phnP*, *gnd*, and *hsdM* (details available on request).

Extremely high intraspecific allelic diversity has been reported among flagellin genes from different clones of *Salmonella* (25), at the major histocompatibility complex in mammals (26, 27), and at the self-incompatibility locus in tobacco and related members of the Solanaceae family (28). In each case it has been convincingly argued that this diversity reflects the action of natural selection, and it is tempting to speculate that the high level of divergence between the *hsdM* alleles of *E. coli* B and K-12 also results from selection.

While R-M systems may have more than one role, the current consensus view is that they protect bacteria against infection by phages (1). The selection pressure exerted by R-M systems on phage genomes is clear (29); in turn, phages will exert a selection pressure on their hosts. Levin (30) has pointed out that phage-mediated selection of bacterial R-M systems is likely to be frequency-dependent, favoring the retention of rare genotypes and thus promoting diversity. Phage-mediated selection is expected to act directly on the *hsdS* gene, which specifies the recognition sequence of the R-M enzyme complex. The high divergence at *hsdM* may result from its tight linkage with *hsdS*, but it is also possible

Table 3. Intra- vs. interspecific divergence

Gene	L*	*E. coli* B vs. K-12	*E. coli* vs. *S. typhimurium*	Ratio†
araA	61	0.000	0.038	0.00
		0.12	0.93	0.13
araC	293	0.000	0.038	0.00
		0.04	1.24	0.03
araD	232	0.006	0.024	0.25
		0.10	0.98	0.10
cysH	245	0.004	0.032	0.12
		0.16	1.10	0.14
cysI	233	0.006	0.069	0.08
		0.02	0.66	0.03
gnd	126	0.000	0.007	0.00
		0.20	0.74	0.27
hsdM‡	510	0.013	0.032	0.41
		0.21	0.49	0.43

The extents of divergence due to synonymous and nonsynonymous nucleotide substitutions are considered separately in each comparison. Numbers of nonsynonymous substitutions per site (K_A; upper value of each pair) and of synonymous (silent) substitutions per site (K_S; lower value) are presented; see text and ref. 19 for details of calculation.
*Length of sequences compared (in codons).
†Ratio of intraspecific/interspecific divergence.
‡Excluding codons 78–96.

that the M subunit contributes to specificity, in which case *hsdM* would also be under direct selection. Alternatively, the antirestriction functions of bacteriophages, in some cases enhancing modification and in others blocking both modification and restriction (29), could exert a direct selection pressure on *hsdM*. That the high intraspecific diversity of *hsdM* involves nonsynonymous substitutions to at least the same extent as synonymous changes (Table 3) may indicate that the M protein *is* under some direct selection (in contrast to *gnd*; see below).

The *hsdM* genes of *S. typhimurium* strains LT2 and *potsdam* are slightly more divergent than those of *E. coli* B and K-12 (Table 2), but this level of divergence cannot yet be compared with that at other loci from the same strains. Nelson *et al.* (31) have examined the extent of diversity among a number of *Salmonella* strains as revealed by sequences of the *gapA* gene and by multilocus enzyme electrophoresis; their results suggest that intraspecific diversity within *Salmonella* is generally much higher than within *E. coli*. Thus, we cannot yet determine whether the extent of divergence at *hsdM* between the two *Salmonella* strains examined here is exceptional.

Interestingly, in the comparison between *E. coli* B and K-12, the second most divergent locus (at silent sites; Fig. 2 *Lower*) is *gnd*. This gene has also been found to be highly divergent (in comparison with *trpB* and *phoA*) in surveys of a range of natural strains from the ECOR collection (32, 33). Among these more divergent *E. coli* strains the great majority of nucleotide substitutions at *gnd* are synonymous and the extent of nonsynonymous divergence is not exceptional, suggesting that the diversity is not due to direct selection at *gnd*. However, *gnd* is located immediately adjacent to the *rfb* locus, encoding the O antigen, and frequency-dependent selection favoring rare O antigens has been invoked to explain diversity at the tightly linked *gnd* locus (32, 33).

***Eco*K Family: Recombination Between *E. coli* and *S. typhimurium*.** Phylogenetic analysis of the four *hsdM* genes of the *Eco*K family clearly clusters *Eco*B with *Eco*K, and *Sty*SB with *Sty*SP. For example, if the maximum parsimony approach (34) is used, there are 86 variable nucleotide sites that are phylogenetically "informative" (i.e., where two members share one nucleotide, while the other two members share another), and 84% of these sites support the grouping of *Eco*B with *Eco*K (and *Sty*SB with *Sty*SP). The remaining small number of "informative" sites would be expected to have resulted from coincidental parallel mutations occurring in two separate lineages. The six sites at which *Eco*B and *Sty*SB share one nucleotide (and *Eco*K and *Sty*SP share another) are scattered through the gene (Fig. 3), as expected. However, six of the eight sites at which *Eco*B and *Sty*SP are identical (and *Eco*K and *Sty*SB share a different base) are clustered within a short region between codons 78 and 94, and there are no contradictory informative sites within this region (Fig. 3). Following Stephens' method (35), this clustering is highly significant ($P < 10^{-5}$) and suggests that there has been a lateral transfer subsequent to the intraspecific divergence events within the *Eco*K family.

Within this region (codons 78–95), the *Eco*B and *Sty*SP sequences differ at only 4 nucleotides, whereas there are 9–17 differences in any other pairwise comparison. Thus, the putative transfer event appears to have involved the ancestors of these two sequences. Furthermore, within this region the *Eco*B sequence appears to be unusually divergent from *Eco*K, suggesting that the transfer was from the *Sty*SP lineage to the *Eco*B lineage.

While several studies have indicated that recombination has occurred among different *E. coli* strains (32, 33, 36, 37), and recombination has been invoked in the generation of flagellin gene diversity among *Salmonella* serovars (25), there has been little evidence of recombination between *E. coli* and *S. typhimurium*. This may be partly because few studies have examined more than one allele from *Salmonella*. An analysis of the *gapA* gene which included multiple alleles from both species produced no evidence of recombination at either the intra- or interspecific levels (31). Analyses of the *gnd* locus have revealed alleles from *E. coli* with unusually high similarity to *S. typhimurium* LT2 (32), but in each case only one *Salmonella* allele was examined and the evidence for interspecific transfer was not as strong as that provided here. The patterns of divergence for different loci discussed above (see Eco*K Family: Interspecific Divergence*) provide indirect evidence against extensive interspecific lateral transfer (21, 38), and a general concordance between traditional taxonomic classifications of enteric bacteria and phylogenies based on various informational macromolecules has also been taken as evidence that lateral transfer among these species has been rare (39).

Divergence of the K and A Families. The *Eco*K and *Eco*A R-M systems are encoded by genes which are allelic in the sense that they are alternative sequences located at the same chromosomal position in their respective strains of origin. However, the divergence between the K and A families is extremely high by comparison with the divergence within the K family. For example, the estimated number of amino acid replacements per site between the M proteins of *Eco*K and *Eco*A is ≈25 times that between *Eco*K and *Sty*SB (Table 1). This ratio must be taken with caution because the K and A families are so divergent that the estimate of divergence has a large error associated with it. However, it is clear that if the *hsdM* genes have been evolving in a roughly clock-like manner, and if the divergence between *Eco*K and *Sty*SB can be used to calibrate that clock, then the level of divergence between (for example) *Eco*K and *Eco*A might be expected to be found only between different phyla of bacteria. It seems unlikely that such a high level of divergence could be maintained intraspecifically over such a long divergence time. [Some *Nicotiana alata* self-incompatibility alleles exhibit only 43% amino acid identity (28), but that protein seems to evolve at a much higher rate than the M protein.] Rather, the high divergence of the *Eco*K and *Eco*A families is more likely to reflect lateral transfer involving the R-M loci. The recognition domains of the S polypeptides of *Eco*K are encoded by sequences with a base composition unlike that of the *E. coli* genome, but similar to that of (for example) species of the

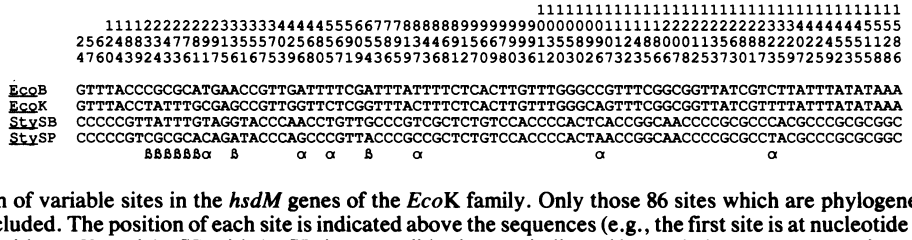

FIG. 3. Distribution of variable sites in the *hsdM* genes of the *Eco*K family. Only those 86 sites which are phylogenetically informative for these four genes are included. The position of each site is indicated above the sequences (e.g., the first site is at nucleotide 24). Most sites support the clustering of *Eco*B with *Eco*K, and *Sty*SB with *Sty*SP; incompatible sites are indicated beneath the sequences: α sites group *Eco*B with *Sty*SB (and *Eco*K with *Sty*SP); β sites group *Eco*B with *Sty*SP (and *Eco*K with *Sty*SB). Six β sites are clustered between codons 78 and 94 and may have resulted from a lateral transfer from the *Sty*SP lineage to the *Eco*B lineage (see text).

genus *Proteus* (40). Since the *Eco*R124 system is encoded by plasmid-borne genes, plasmids may serve as intermediaries in such transfer. It is also possible that plasmid-borne sequences evolve at faster rates, due to a higher mutation rate or to less stringent selective constraints.

Highly divergent, yet allelic, sequences have been reported elsewhere. For example, in *Bacillus subtilis*, genes from strains 168 and W23 typically differ at about 5% of bases (41), yet the genes concerned with teichoic acid biosynthesis are so different that homology between them was not detected by Southern hybridization (42). These sequences have been called "pseudoalleles" (42), and the same terminology may be appropriate here.

Evolution of *hsdM* and of the *E. coli* Chromosome. The *hsdM* genes of type I R-M systems may have been subject to a variety of evolutionary processes. The extent of divergence at *hsdM* between *E. coli* K-12 and *S. typhimurium* LT2 can be most parsimoniously attributed to the simple accumulation of largely neutral changes: the degree of divergence is low because of above-average levels of constraint on both silent sites and those causing amino acid replacements. In contrast, intraspecific divergence (between *E. coli* strains K-12 and B) is exceptionally high, which may reflect phage-mediated frequency-dependent selection. In addition, there is evidence that recombination has played a role in the history of these sequences.

The results of multilocus enzyme electrophoresis have seemed to indicate that different *E. coli* strains have evolved in a largely clonal fashion (43), but DNA sequence analysis of natural isolates indicates that transfer of short sequences (perhaps around 1000 base pairs) has occurred among strains of *E. coli* (33, 36, 37). From attempts to integrate these two different views, a new picture of the population genetics and evolution of *E. coli* is emerging. Periodic selection of occasional advantageous mutations, as well as random genetic drift, may homogenize worldwide populations of this bacterium, yielding a largely clonal population structure; this clonality is disrupted by recombination events (37). To explain the extraordinary diversity at the *hsd* locus, one may consider that the chromosome as a whole is being homogenized due to random drift, or due to selection at loci other than *hsd*, in which case highly divergent *hsd* alleles appear to have been recombined into this background and provided a selective advantage due to phage-mediated selection. Alternatively, focusing on the *hsd* locus, high diversity at the silent sites within the *hsdM* alleles indicates that the clonal frames including these alleles must have been circulating in the *E. coli* population for a very long period of time, but the lower diversity at silent sites in other genes indicates that their alleles have a more recent common ancestry, presumably due to recombination.

We thank Annette Campbell for providing clones. These studies were supported by grants from The Medical Research Council (N.E.M.) and EOLAS (P.M.S.) and by Science and Engineering Research Council Studentships to J.E.K. and G.M.C. This is a publication from the Irish National Centre for Bioinformatics and the Edinburgh Centre for Molecular Recognition.

1. Wilson, G. G. & Murray, N. E. (1991) *Annu. Rev. Genet.* **25**, 585–627.
2. Bickle, T. A. (1987) in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, eds. Neidhardt, F. C., Ingraham, J. L., Low, K. B., Magasanik, B., Schaechter, M. & Umbarger, H. E. (Am. Soc. Microbiol., Washington), pp. 692–696.
3. Daniel, A. S., Fuller-Pace, F. V., Legge, D. M. & Murray, N. E. (1988) *J. Bacteriol.* **170**, 1775–1782.
4. Kannan, P., Cowan, G. M., Daniel, A. S., Gann, A. A. F. & Murray, N. E. (1989) *J. Mol. Biol.* **209**, 335–344.
5. Loenen, W. A. M., Daniel, A. S., Braymer, H. D. & Murray, N. E. (1987) *J. Mol. Biol.* **198**, 159–170.
6. Gough, J. A. & Murray, N. E. (1983) *J. Mol. Biol.* **166**, 1–19.
7. Price, C., Lingner, J., Bickle, T. A., Firman, K. & Glover, S. W. (1989) *J. Mol. Biol.* **205**, 115–125.
8. Cowan, G. M., Gann, A. A. F. & Murray, N. E. (1989) *Cell* **56**, 103–109.
9. Gann, A. A. F., Campbell, A. J. B., Collins, J. F., Coulson, A. F. W. & Murray, N. E. (1987) *Mol. Microbiol.* **1**, 13–22.
10. Gubler, M., Braguglia, D., Meyer, T., Piekarowicz, A. & Bickle, T. A. (1992) *EMBO J.* **11**, 233–240.
11. Murray, N. E. (1983) in *Lambda II*, eds. Hendrix, R. W., Roberts, J. W., Stahl, F. W. & Weisberg, R. A. (Cold Spring Harbor Lab., Cold Spring Harbor, NY), pp. 395–432.
12. Fuller-Pace, F. V., Cowan, G. M. & Murray, N. E. (1985) *J. Mol. Biol.* **186**, 65–75.
13. Midgley, C. A. & Murray, N. E. (1985) *EMBO J.* **4**, 2695–2703.
14. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
15. Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3963–3965.
16. Fuller-Pace, F. V. & Murray, N. E. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 9368–9372.
17. Higgins, D. G. & Sharp, P. M. (1988) *Gene* **73**, 237–244.
18. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, U.K.).
19. Li, W.-H., Wu, C.-I. & Luo, C.-C. (1985) *Mol. Biol. Evol.* **2**, 150–174.
20. Doolittle, R. F., Feng, D. F., Johnson, M. S. & McClure, M. A. (1986) *Cold Spring Harbor Symp. Quant. Biol.* **51**, 447–455.
21. Sharp, P. M. (1991) *J. Mol. Evol.* **33**, 23–33.
22. Ikemura, T. (1985) *Mol. Biol. Evol.* **2**, 13–34.
23. Sharp, P. M. & Li, W.-H. (1987) *Nucleic Acids Res.* **15**, 1281–1295.
24. Herzer, P. J., Inouye, S., Inouye, M. & Whittam, T. S. (1990) *J. Bacteriol.* **172**, 6175–6181.
25. Selander, R. K. & Smith, N. H. (1990) *Rev. Med. Microbiol.* **1**, 219–228.
26. Figueroa, F., Günther, E. & Klein, J. (1988) *Nature (London)* **335**, 265–267.
27. Lawlor, D. A., Ward, F. E., Ennis, P. D., Jackson, A. P. & Parham, P. (1988) *Nature (London)* **335**, 268–271.
28. Ioerger, T. R., Clark, A. G. & Kao, T.-H. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 9732–9735.
29. Krüger, D. H. & Bickle, T. A. (1983) *Microbiol. Rev.* **47**, 345–360.
30. Levin, B. R. (1988) *Philos. Trans. R. Soc. London Ser. B* **319**, 459–472.
31. Nelson, K., Whittam, T. S. & Selander, R. K. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 6667–6671.
32. Bisercic, M., Feutrier, J. Y. & Reeves, P. R. (1991) *J. Bacteriol.* **173**, 3894–3900.
33. Dykhuizen, D. E. & Green, L. (1991) *J. Bacteriol.* **173**, 7257–7268.
34. Fitch, W. M. (1971) *Syst. Zool.* **20**, 406–416.
35. Stephens, J. C. (1985) *Mol. Biol. Evol.* **2**, 539–556.
36. DuBose, R. F., Dykhuizen, D. E. & Hartl, D. L. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 7036–7040.
37. Milkman, R. & Bridges, M. M. (1990) *Genetics* **126**, 505–517.
38. Maynard Smith, J., Dowson, C. G. & Spratt, B. G. (1991) *Nature (London)* **349**, 29–31.
39. Ochman, H. & Wilson, A. C. (1987) in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, eds. Neidhardt, F. C., Ingraham, J. L., Low, K. B., Magasanik, B., Schaechter, M. & Umbarger, H. E. (Am. Soc. Microbiol., Washington), pp. 1649–1654.
40. Dila, D., Sutherland, E., Moran, L., Slatko, B. & Raleigh, E. A. (1990) *J. Bacteriol.* **172**, 4888–4900.
41. Sharp, P. M., Nolan, N. C., Ni Cholmain, N. & Devine, K. M. (1992) *J. Gen. Microbiol.* **138**, 39–45.
42. Young, M., Mauël, C., Margot, P. & Karamata, D. (1989) *Mol. Microbiol.* **3**, 1805–1812.
43. Selander, R. K., Caugant, D. A. & Whittam, T. S. (1987) in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, eds. Neidhardt, F. C., Ingraham, J. L., Low, K. B., Magasanik, B., Schaechter, M. & Umbarger, H. E. (Am. Soc. Microbiol., Washington), pp. 1625–1648.