

Families of restriction enzymes: an analysis prompted by molecular and genetic data for type ID restriction and modification systems

Annette J. B. Titheradge, Jonathan King¹, Junichi Ryu¹ and Noreen E. Murray*

Institute of Cell and Molecular Biology, University of Edinburgh, King's Buildings, Edinburgh EH9 3JR, UK and

¹Department of Microbiology and Molecular Genetics, Loma Linda University, Loma Linda, CA 92350, USA

Received June 14, 2001; Revised and Accepted August 17, 2001

DDBJ/EMBL/GenBank accession no. AJ132566

ABSTRACT

Current genetic and molecular evidence places all the known type I restriction and modification systems of *Escherichia coli* and *Salmonella enterica* into one of four discrete families: type IA, IB, IC or ID. *StySBLI* is the founder member of the ID family. Similarities of coding sequences have identified restriction systems in *E.coli* and *Klebsiella pneumoniae* as probable members of the type ID family. We present complementation tests that confirm the allocation of *EcoR9I* and *KpnAI* to the ID family. An alignment of the amino acid sequences of the HsdS subunits of *StySBLI* and *EcoR9I* identify two variable regions, each predicted to be a target recognition domain (TRD). Consistent with two TRDs, *StySBLI* was shown to recognise a bipartite target sequence, but one in which the adenine residues that are the substrates for methylation are separated by only 6 bp. Implications of family relationships are discussed and evidence is presented that extends the family affiliations identified in enteric bacteria to a wide range of other genera.

INTRODUCTION

In *Escherichia coli* K-12, a 15 kb segment of the bacterial chromosome, referred to as the immigration control region (ICR), specifies three restriction endonucleases: one classical type I restriction and modification (R–M) system and two methylation-dependent restriction enzymes (1). Three genes (*hsdR*, *M* and *S*) encode the classical R–M system (*EcoKI*) first identified by Bertani and Weigle (2). The ICR, however, is hypervariable in *E.coli* and its close relatives (3–5). In different strains of *E.coli*, alternative *hsd* genes specify type I R–M systems with different specificities. *Escherichia coli* K-12 and *E.coli* B, for example, have R–M systems specified by allelic genes and complementation tests showed that the subunits

of one system, *EcoKI*, can associate with those of the other, *EcoBI*, to make functional chimaeric enzymes (6,7). *EcoKI* and *EcoBI* differ significantly in only one of their three subunits, HsdS, the subunit that confers sequence specificity to the heterooligomeric complex. *EcoKI* and *EcoBI* each comprise one specificity subunit and two of each of the other subunits, HsdM and HsdR. The alternative activities of the R–M complex are dictated by the methylation state of the target sequence. Unmethylated targets evoke endonuclease activity and hemimethylated targets are the substrates for methylation (8–11).

EcoKI and *EcoBI* are the founder members of a family of restriction and modification strains referred to as type IA. This family also includes members from strains of *Salmonella enterica* (12) and a variety of natural isolates of *E.coli* (for reviews see 11,13). The strictest requirement for membership of a family depends on relatedness as demonstrated by complementation tests in which subunits from different enzymes associate to make a functional enzyme. These tests require partial diploids made in bacterial strains sensitive to tester phages and, therefore, have seldom been extended to different genera. More generally applicable tests rely on molecular evidence derived from hybridisation screens of bacterial DNA using *hsd* sequences as probes, and serological screens of cell extracts with antibodies raised against a representative of a known family of enzymes (5,14).

Escherichia coli 15T⁻ has chromosomal *hsd* genes that behave as alleles of those in *E.coli* K-12 (15) but they share very little sequence similarity as evidenced initially from hybridisation screens (14). The *hsd* genes of *E.coli* 15T⁻ specify *EcoAI*, the first member of a second family of type I R–M systems (IB) in which different HsdS subunits confer specificities for different nucleotide sequences (16,17). A third family (type IC) was identified by plasmid representatives (8). More recently an R–M system originally identified in the *blegdam* serovar of *S.enterica* (12) was identified as the first member of a new family (type ID) of chromosomally encoded type I R–M systems (18). This system, *StySBLI*, is encoded by genes that may be alleles of those for the type IA or IB systems (12). Physical evidence obtained using flanking DNA

*To whom correspondence should be addressed. Tel: +44 131 650 5374; Fax: +44 131 650 8650; Email: noreen.murray@ed.ac.uk

Present address:

Annette J. B. Titheradge, John Hughes Bennett Laboratories, Department of Oncology, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK

sequences as probes identifies a very similar location on the chromosome but no hybridisation was detected between the *hsd* genes of either *E.coli* K-12 or *E.coli* 15T⁻ and those of *S.enterica* serovar *blegdam* (18).

Currently, molecular tests place all the known type I R–M systems of *E.coli* and *S.enterica* into one of four discrete groups or families; those in type IC may be plasmid-borne and those of type IA, IB and ID are specified by genes linked to *serB* in the bacterial chromosome. The subdivision of type I R–M systems is an empirical one. High levels of identity at the level of nucleotide sequences are indicative of relatively recent divergence and conservation at the level of protein subunits. Any comparison between representatives of two families of systems identifies little sequence similarity even at the level of amino acid sequence; commonly only 20–30% amino acid identity. The known exceptions are two target recognition domains (TRDs) that recognise the same nucleotide sequences (19). To date classifications dependent on molecular probes are consistent with the limited information from complementation tests. In this paper we identify the recognition sequence of *StySBLI* and provide evidence based on complementation for the extension of the ID family documented in *Salmonella* to include members from *E.coli* and *Klebsiella pneumoniae*. We then review the status of the family concept as assessed by comparisons of amino acid sequences, an approach more relevant to the present era.

MATERIALS AND METHODS

Bacteria, phages and plasmids

The *hsd* genes specifying *EcoR9I* were cloned from the DNA of ECOR9 (ATCC no. 35328) (20). LB4037 (12), an *E.coli* K-12 derivative, in which the *mcr hsd mrr* region has been replaced with the *hsd* region of *S.enterica* serovar *blegdam* (hereafter abbreviated to *S.blegdam*), was used as a λ -sensitive strain proficient in the *blegdam* (*StySBLI*) R–M system. Mutant derivatives of LB4037 were made for use in complementation tests; in NM856 (*hsdS*), the *hsdS* gene was inactivated by the insertion of *supF* (18), in NM867 (*hsdM*), a mutation changed the sequence of the methyltransferase motif IV from NPPF to NPPC and in NM857 (*hsdR*), a mutation changed the K in the ATP-binding motif to T.

For experiments using M13, F' derivatives of LB4037, the *mutD5* strain RP526 (21) and the *dam*⁻ strain CB51 (provided by Dr A. C. Boyd, Medical Genetics, Western General Hospital, University of Edinburgh, UK) were made by transferring the F':Tn5 donor from EH55 (22). XL1-Blue MRF' (Stratagene) was used as an *r^m-supE* host for M13 phages. DH5 α (23) and XL1-Blue (Stratagene) were hosts for the recovery and amplification of plasmid DNA. ED8654 (24) was the standard λ -sensitive, restriction-deficient strain for the recovery and amplification of λ phages. Either NM679 (25), a Δ (*mcr hsd mrr*) Δ *mcrA* derivative of W3110 or JR300 (26), an *E.coli* C strain that is naturally deficient in *hsd* genes, were used as λ -sensitive, *r^m*⁻ hosts.

Libraries of ECOR9 DNA were made in NM1249, a *cI857* derivative of EMBL3 (27). A 4.5 kb *EcoRI*–*SalI* fragment from a λ *hsd* phage (isolate #8) subcloned in pUC19 generated pECOR9, a plasmid that conferred *EcoR9I* modification proficiency, but not restriction proficiency, to *E.coli* K-12 strains.

The plasmids containing the *hsd* genes of *K.pneumoniae* were pJR41 (*hsdM*⁺*S*⁺), pJR43 (*hsdM*⁺), pNL3 one of four (*hsdR*⁺) clones pNL1-4 (28) and pJR51 (*hsdR*⁺*M*⁺*S*⁺). pJR51 contains two *PstI* fragments in pUC19: one (6.7 kb) from pJR31 includes *hsdM* and *hsdS* and a part of *hsdR*, and the other, a 1.4 kb fragment from pNL3, provides the remainder of *hsdR*. The plasmids containing the *hsd* genes from *S.blegdam* were described by Titheradge *et al.* (18). pAC18 includes the *hsdM* and *hsdS* genes within an *EcoRI* fragment and pAT29, the *hsdR* gene within a *BglII* fragment in pUC9. These plasmids were the substrates for site-directed mutagenesis to generate mutations in *hsdM* and *hsdR*, respectively. In each case an *EcoRI* fragment was then subcloned in NM461 (λ b538cI857srI4^onin srI5^o) an integration-defective, temperature-inducible λ vector (29), and transferred to the bacterial chromosome to generate the *hsdM* and *hsdR* derivatives of LB4037.

λ vir, used to test restriction and modification, was either unmodified (λ vir.0) by propagation in NM679 or JR300, or appropriately modified by propagation in LB4037 (λ vir.*StySBLI*), *E.coli* C/pJR41 (λ vir.*KpnAI*) or NM679 lysogenic for λ *hsd* clone #8 (λ vir.*EcoR9I*). λ vir.0 after propagation on a test strain was checked for modification. All tests dependent on plasmids used freshly transformed strains.

Media and microbial techniques

Media and general methods (28,30) and the use of λ *hsd* phages to transfer mutations to the chromosome have been described (31).

Enzymes, reagents and reactions

Restriction enzymes, T4 DNA ligase, mung bean nuclease and Klenow polymerase were purchased from Boehringer Mannheim or New England Biolabs. Red Hot *Taq* polymerase from Advanced Biotechnologies Ltd was used to amplify DNA to make probes. Site-directed mutagenesis was done using the QuikchangeTM mutagenesis kit of Stratagene. λ packaging extracts were from Promega.

DNA preparations and manipulations used standard methods (32). The reagents and methods for the detection of DNA sequences by hybridisation have been described (5). The ABI PRISM dRhodamine Terminator Cycle Sequencing Ready Reaction Kit from PE Biosystems was used to prepare samples for an automated sequencer (AB1 PRISM 377). Plasmid DNA for sequence determination was prepared using the Biotech Flexiprep Kit (Pharmacia). Oligonucleotides were obtained from Oswel or MWG-Biotech UK Ltd. The sequence of the *hsd* genes of ECOR9 was always determined on both strands. The sequences of DNA fragments were compiled using Gene-Jockey.

Sequence comparisons

Alignments were made initially using the TBLASTN program (33) available on the National Center for Biotechnology Information web site (<http://www.ncbi.nlm.nih.gov/BLAST>). The predicted amino acid sequences of the subunits of *EcoKI*, *EcoAI*, *EcoR124I/II* and *StySBLI* were used to screen databases. The nucleotide sequences specifying these subunits are in the EMBL/GenBank/DBJ databases under the following accession numbers: *EcoKI* (U14003); *EcoAI* HsdR (L18758), HsdM (L02505), HsdS (J03150); *EcoR124I/II* (X13145) and

```

mp8  GAATTC          CCGGGGATCC      GTCGACCTGCAGC   CAAGCTTGGC
mp10 GAATTCGAGCTCG  CCCGGGATCCTCTAGAGTCGACCTGCAGC  CCAAGCTTGGC
mp18 1 3 5 7 9 11 13 15
      GAATTCGAGCTCGGTACCCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCCAAGCTTGGC
      EcoRI      KpnI      BamHI          SphI      HindIII
              SacI      XmaI

```

Figure 1. Polylinker sequences of M13 vectors. The sequences of the polylinkers of mp8, mp10 and mp18 are aligned. The targets for type II endonucleases are identified in the sequence of mp18 and the numbers (1–15) correspond to bases 1–15 in Figure 2.

StySBLI (X99719). The sequence for the HsdR of *EcoR124 I/II* (X13145) was corrected by the addition of a C at position 3064; this changes the C-terminal sequence from FRKSSRLLRSLKA to FQKIVSFIEKFKGVGGKI. In addition to the databases of published sequences, preliminary sequence data were obtained from the NCBI unfinished Microbial Genomes BLAST website at http://www.ncbi.nlm.nih.gov/Microb_blast/unfinishedgenome.html; http://www.sanger.ac.uk/Projects/S_typhi; http://www.sanger.ac.uk/Projects/S_equi.

The multiple sequence alignments for the comparisons in Table 5 used the CLUSTAL W(1.5) program.

RESULTS

The recognition sequence for *StySBLI*

Gann *et al.* (34) introduced an *in vivo* strategy in which M13 clones were used as substrates to define the target sequence for a type I R–M system; only those phages that are susceptible to restriction contain a target sequence. When experiments were initiated to determine the target sequence recognised by *StySBLI*, it was found that two commonly used M13 vectors, mp18 and mp19, that differ in the orientation of their polylinker sequence, had an efficiency of plating (e.o.p.) of ~0.1 on an *E. coli* strain in which the *hsd* genes for *EcoKI* had been replaced by those specifying *StySBLI* (LB4037F'). This implied a single target for *StySBLI* within these M13 vectors. Related vectors, mp8 and mp10, differing from mp18 in their shorter polylinker sequences (Fig. 1), were not restricted (e.o.p. of 1). These results indicate that the polylinker sequence of M13mp18 includes a recognition sequence for *StySBLI*, part of which is within either the *KpnI* or the *SphI* target; targets that are absent in mp8 and mp10 (Fig. 1).

Two methods were used to refine the boundaries of the predicted recognition sequence. First, mutant derivatives of mp18 that have an e.o.p. of 1 on LB4037F' were selected after amplification on a *mutD* strain. Two mutants with base substitutions were identified, and each base change was within the *KpnI* target (residues 13 and 14 in Fig. 1). Secondly, mutations were made within the polylinker sequence after cutting the vector with a type II restriction endonuclease. The projecting 5' single-stranded regions were either removed or used as templates to make double-stranded ends, and circular genomes were restored by DNA ligase. Changes within the *EcoRI*, *BamHI* and *XmaI* target sequences failed to remove the *StySBLI* recognition sequence, while that within the *SacI* sequence generated a restriction-resistant derivative. These experiments identify the relevance of bases within the *SacI* and *KpnI* target sequences.

Known type I recognition sequences are bipartite, consisting of a 3 or 4 bp component, separated by a non-specific spacer of 6–8 bp from a second component of 4 or 5 bp. Each component includes an adenine residue, the substrate for methylation. The adenine residues are on opposite strands and generally ~10 bp apart (8,9). Residue 4 within the *SacI* target and residue 11 within the *KpnI* target are the only candidates for methylation, within the defined region (Fig. 1).

Eleven nucleotides that included the *SacI* and *KpnI* target sequence (nucleotides 1–5 and 10–15) were subjected to site-directed mutagenesis. Each nucleotide was replaced by the three alternatives to determine whether degeneracies at any position provided a sequence that could be recognised by *StySBLI*. The results, summarised in Figure 2, define a target sequence, CGA(N₆)TACC comprising 7 bp with no degeneracies and a spacer of 6 bp. They also show that methylation of the first adenine residue (position 4 in Fig. 2) by the *Dam* methylase prevents attack by *StySBLI*.

Cloning the *hsd* genes of *ECOR9*

DNA from *ECOR9*, a member of the *ECOR* collection of *E. coli* strains (20), has been shown to hybridise with a probe specific to the ICR of *S. blegdam* (5). *ECOR9* is presumed to include *hsd* genes, but the strain is insensitive to the phages commonly available in the laboratory; therefore, no biological tests could be made. Evidence for a functional restriction system required that the putative *hsd* genes were cloned in

	1	2	3	4	5	6-9	10	11	12	13	14	15
	t	C	G	A	g	ctcg	g	T	A	C	C	c
A	√	x	x	n.a.	√	n.d.	√	x	n.a.	x	x	√
C	√	n.a.	x	x	√	n.d.	√	x	x	n.a.	n.a.	n.a.
G	√	x	n.a.	x	n.a.	n.d.	n.a.	x	x	x	x	√
T	n.a.	x	x	x	√*	n.d.	√	n.a.	x	x	x	√

Figure 2. Base substitutions made to identify the nucleotide sequence recognised by *StySBLI*. The upper case letters in bold define the nucleotide sequence recognised by *StySBLI* and the nucleotides are numbered as in Figure 1. Substitutions were made for the bases identified by numbers 1–5 and 10–15. √ identifies a base change that is without effect on the recognition of the nucleotide sequence by *StySBLI*. x identifies a base change that destroys the target sequence. n.a., not applicable; n.d., not done. *When G is replaced with T at this position, the A residues within the sequence GATC become the substrate for the *Dam* methylase; methylation of the A residues blocks restriction by *StySBLI*.

E. coli K-12. A library of DNA fragments, generated as the products of partial digestion by *Sau*3AI, was recovered in the λ vector NM1249, a *ci*857 derivative of EMBL3 (27). Plaques identified by the *Sty*SBLI-specific probe were purified and tested with probes derived from the sequences that flank the *hsd* genes in serovars of *S. enterica*. These probes also hybridise to DNA in *E. coli* K-12. One *hsd* phage (clone 8) included DNA that hybridised to the probe made from sequence extending ~2 kb downstream of *hsdS* in *S. enterica* serovar *typhimurium* LT2 (pAB3 in ref. 18), while another *hsd* phage (clone 9) hybridised to the probe made from a DNA fragment located ~5 kb upstream of *hsdM* in *S. blegdam* (18). The type IA *hsd* genes of serovar *typhimurium* and the type ID *hsd* genes of serovar *blegdam* were known to have the same chromosomal location despite their different gene order (18). The physical evidence from the λ clones supports the same location for the type ID *hsd* genes of *ECOR9*.

A 4.5 kb *Eco*RI-*Sal*I fragment shown to hybridise to the *hsdS* gene of *Sty*SBLI was transferred from λ clone 8 to pUC19. The nucleotide sequence of part of the chromosomal DNA fragment within this plasmid (pECOR9) was determined (EMBL nucleotide sequence database, accession no. AJ132566). The sequence from *E. coli* *ECOR9* was readily aligned with the *hsd* gene sequences of *S. blegdam*. It is consistent with the presence of *hsdM* and *hsdS*, followed by the beginning of *hsdR*; the same gene order as the first member of the ID family. Plasmid pECOR9 is predicted to specify the modification component of the R-M system *Eco*R9I. The comparison of the *hsdS* genes of *Sty*SBLI and *Eco*R9I identifies two variable regions as anticipated for two TRDs that specify a bipartite target sequence, flanking a central conserved region (Fig. 3). There is a conserved sequence at the C-terminus, but no well-conserved sequence was identified at the N-terminus.

Complementation between subunits of different enzymes

The close relatedness of two type I R-M systems was first demonstrated for *Eco*KI and *Eco*BI; in these experiments functional R-M systems resulted when a subunit of one enzyme replaced the subunit of the other, with the HsdS subunit determining the sequence specificity of a chimaeric complex (6). These tests for restriction and modification were done *in vivo* using a partial diploid in which a second set of *hsd* genes was provided on an *F'* plasmid. Such tests are now more generally applicable to genes cloned in a plasmid vector.

The *hsd* genes for *Eco*R9I were previously identified by hybridisation screens, but those in *K. pneumoniae*, specifying *Kpn*AI have already been cloned. Sequence comparisons identified the latter as a likely member of the ID family (28); the predicted HsdM sequences of *Kpn*AI and *Sty*SBLI share 97% identity and those for HsdR 94% identity. The *hsd* genes of *S. blegdam* had been transferred to *E. coli* (12) and a mutation in *hsdS* made by the insertion of *supF* (18). Site-directed mutagenesis of *hsdM* (in pAC18) and *hsdR* (in pAT29) was used to generate substitutions in conserved motifs within the active sites for the methyltransferase and ATPase activities, respectively. These mutations were transferred via λ *hsd* phages to the chromosome of LB4037, the *E. coli* strain specifying *Sty*SBLI, to make *hsdR*, *hsdM* and *hsdS* mutants of a λ -sensitive strain (see Materials and Methods). Partial diploids were made by transforming the mutant strains with a plasmid carrying the

	1								50
<i>Sty</i> SBLI								MAFEKT	IPLNEFITLQ
<i>Eco</i> R9I	MGNSGFKLPL	GWNCCKLVDC	TKEGNISYGI	VQPGQHEDG	IGIIRVNNIQ				
	51								100
<i>Sty</i> SBLI	RGFDLQDKR	VMGDIPVVAS	TGVVGYHNEE	KVLAPGVVIG	RSGSIGGGQY				
<i>Eco</i> R9I	NG.....N	IYDDVLKVS	HEIESKFAKT	RLEGGEVLLT	LVGSTGISAI				
	101								150
<i>Sty</i> SBLI	ITTNF..WPL	.NTTLWVKDF	KGHHPRFVYV	LLRSIDFSQF	NVGSV.....				
<i>Eco</i> R9I	TTKALQGNV	ARAVAVIKPC	DEISAEWIHI	CLQS.PFTKY	FLDSRANTTV				
	151								200
<i>Sty</i> SBLI	.PTLNRNHL	GILVADTSYS	YEKEASDIIG	ILDDKIKLNK	ELNHTLEQIS				
<i>Eco</i> R9I	QKTLNLKDVK	EIPLPIPPHE	ERVSLEKIYF	NFENRINLNI	KINKILEEMS				
	201								250
<i>Sty</i> SBLI	QTLFKSWFVD	FDPVIDNALD	AGNPIPEALQ	SRAELRQKIR	NSADFKPLPA				
<i>Eco</i> R9I	QNLFKSWFVD	FDPVVDNALD	AGNPIPEALQ	SRAELRQKVR	NSADFKPLPA				
	251								300
<i>Sty</i> SBLI	DIRALFPAEF	EE TELGWMPK	GW ITTSFNLD	IELIGG.GTP	KTSVEEFWNG				
<i>Eco</i> R9I	EIRSLFPSEF	EE TELGWMPK	GW QIKSLDHI	ANFQNLALQ	KFRPKNMEED				
	301								350
<i>Sty</i> SBLI	DIPWFSVVDA	PSESDVYVLT	TEKKITIEGL	NNSSAKLLRK	GTIISARGT				
<i>Eco</i> R9I	YLPVLKIADL	RAGQ...IT	NDERARTD..	ISDSCKVY.D	GDMIFSWSGT				
	351								400
<i>Sty</i> SBLI	VGKCAMVAVP	MAMNQSCYGV	IGKNNISDEY	IYFQLKNAVQ	TLQQMGHGSV				
<i>Eco</i> R9I	LMIDIWTGGN	AALNQHLKYV	TSKK..YPO	YFPMW..TIQ	HLSRFQHIAT				
	401								450
<i>Sty</i> SBLI	FNTITRDTEK	..NIKVPFC.	..NEELTNSY	SLLVKINYFSK	ILNNYQNI				
<i>Eco</i> R9I	AKAVTMGHK	KGDLNSPFL	IPTSSLITKY	DNIVGGYLAK	IKNQRLNNO				
	451							482	
<i>Sty</i> SBLI	LTNLRDTELLP	KLISGELSLE	DLPNLAQTE	PA					
<i>Eco</i> R9I	MTALRDTELLP	KLISGELSLE	DIPDLNTE	AA					

Figure 3. An alignment of the amino acid sequences of the S subunits of *Sty*SBLI and *Eco*R9I. The alignment was made using PILEUP [Wisconsin Package Version 10, Genetics Computer Group (GCG), Madison, WI, USA]. Conserved amino acids are indicated by bold type. The alignments identify two variable regions (TRDs) flanking a central conserved region and a conserved C-terminus.

hsdM and *hsdS* genes of *ECOR9* or plasmids carrying one or more of the *hsd* genes that specify *Kpn*AI. The HsdR subunit of *Sty*SBLI could substitute for the HsdR subunit of *Eco*R9I, and the HsdM subunit of *Eco*R9I for that of *Sty*SBLI as seen by the presence of two specificities when the *hsdM* strain (NM867) was transformed with pECOR9 (Table 1). Similarly, the subunits of *Kpn*AI could substitute for those of *Sty*SBLI (Table 2). Wherever a strain included functional *hsdR*, *hsdM* and *hsdS* genes, a R-M-proficient strain was obtained. Additional complementation tests using two plasmids confirmed the functional association of HsdR and HsdM of *Sty*SBLI with those of *Kpn*AI (data not shown).

Our results demonstrate that the three R-M systems from different genera maintain sufficient similarity to meet the most demanding requirement for membership of the same family of type I R-M systems.

DISCUSSION

Sequence specificity

All type I R-M systems are likely to have a common origin (35), those allocated to a family being very closely related but illustrating significant diversification only within the specificity gene. The present organisation of an *hsdS* gene that specifies two TRDs is believed to have arisen by either duplication

Table 1. Complementation between subunits of *StySBLI* and *EcoR9I*

Strain	Functional Hsd subunits					e.o.p. of λ vir which is unmodified (v.0) or modified against <i>StySBLI</i> (v.SBLI) or <i>EcoR9I</i> (v.R9I)			Relevant phenotype
	<i>StySBLI</i> (on the chromosome)			<i>EcoR9I</i> (on the plasmid)		v.0	v.SBLI	v.R9I	
	R	M	S	M	S				
NM679						1	1	1	r^+m^-
LB4037	+	+	+			$(1.8 \pm 0.7) \times 10^{-5}$	0.8 ± 0.02	$(4.3 \pm 5.8) \times 10^{-5}$	$r^+m^+_{StySBLI}$
NM856	+	+				0.7 ± 0.06	0.7 ± 0.1	0.8 ± 0.3	r^+m^-
NM867	+		+			0.9 ± 0.2	0.9 ± 0.1	0.9 ± 0.3	r^+m^-
NM679(pECOR9)				+	+	1.1 ± 0.2	1.0 ± 0.3	1.0 ± 0.4	$r^+m^+_{EcoR9I}$
NM856(pECOR9)	+	+		+	+	$(4.0 \pm 3.7) \times 10^{-4}$	$(1.2 \pm 0.7) \times 10^{-4}$	0.5 ± 0.1	$r^+m^+_{EcoR9I}$
NM867(pECOR9)	+		+	+	+	$(3.1 \pm 1.7) \times 10^{-4}$	$(9.9 \pm 3.5) \times 10^{-4}$	$(6.4 \pm 2.9) \times 10^{-4}$	$r^+m^+_{StySBLI}$ $r^+m^+_{EcoR9I}$

Table 2. Complementation between subunits of *StySBLI* and *KpnAI*

Strain	Functional Hsd Subunits						e.o.p. of λ vir which is unmodified (v.0) or modified against <i>StySBLI</i> (v.SBLI) or <i>KpnAI</i> (v.KpnAI)			Relevant phenotype
	<i>StySBLI</i> (on the chromosome)			<i>KpnAI</i> (on the plasmid)			v.0	v.SBLI	v.KpnAI	
	R	M	S	R	M	S				
<i>E. coli</i> C							1	1	1	r^+m^-
LB4037	+	+	+				$(1.2 \pm 0.7) \times 10^{-4}$	0.8 ± 0.6	$(3.5 \pm 0.9) \times 10^{-4}$	$r^+m^+_{StySBLI}$
DH5 α (pJR51)				+	+	+	$(3.6 \pm 1.6) \times 10^{-8}$	$(5.7 \pm 4.2) \times 10^{-8}$	0.8 ± 0.4	$r^+m^+_{KpnAI}$
NM857		+	+				1.2 ± 0.1	1.4 ± 0.5	1.4 ± 0.5	$r^+m^+_{StySBLI}$
NM867	+		+				1.5 ± 0.1	1.4 ± 0.4	1.6 ± 0.6	r^+m^-
NM857(pNL3)		+	+	+			$(1.0 \pm 0.6) \times 10^{-5}$	1.5 ± 0.5	$(9.3 \pm 5.8) \times 10^{-5}$	$r^+m^+_{StySBLI}$
NM867(pJR43)	+		+		+		$(2.8 \pm 2.2) \times 10^{-4}$	1.2 ± 0.2	$(6.5 \pm 4.6) \times 10^{-4}$	$r^+m^+_{StySBLI}$
NM867(pJR41)	+		+		+	+	$(5.8 \pm 3.4) \times 10^{-5}$	$(6.3 \pm 1.3) \times 10^{-3}$	$(6.6 \pm 4.3) \times 10^{-4}$	$r^+m^+_{StySBLI}$ $r^+m^+_{KpnAI}$

within the gene or gene duplication followed by gene fusion. The determinants of the N- and C-terminal TRDs would, therefore, be derived from a common ancestor (36). Gene duplication may have occurred more than once. The TRDs have evolved to recognise a variety of tri-, tetra- and even pentanucleotide sequences, but all known target sequences for type I R–M systems comprise two components separated by a non-specific sequence of 6–8 bp.

Currently, all the sequences recognised by type I R–M systems include two adenosyl residues, one in each strand of DNA, which are the targets for methylation. Early experiments indicated that the adenosyl residues targeted by type I modification enzymes were ~10 bp apart, separated by 8 or 9 bp, consistent with the two TRDs of HsdS making interactions within two successive major grooves of the DNA helix (37). However the two adenine residues that are methylated by *EcoR124I* (type IC) are separated by 7 bp (38) and our experiments determine the sequence recognised by *StySBLI* as CGA(N)₆TACC within which the two available adenosyl residues that identify each strand of the recognition sequence are separated by only 6 bp. Methylation of the adenosyl residue in the CGA sequence was shown to protect the target sequence from restriction by *StySBLI*.

The emergence of families of enzymes, which differ in the distance between the bases that are the targets for methylation, has enhanced the potential for the diversification of target sequences. The present data, summarised in Table 3, are consistent with the methylation of adenosyl residues separated by 9 bp in the IB family, 8 bp in the IA family, 7 or 8 bp in the IC family and 6 bp in the first member of the ID family. The variability in the IC family is dictated by whether a tetrapeptide sequence (TAEL) within the central conserved region is present in duplicate or in triplicate, the additional four amino acids increasing the separation of the target adenines by 1 bp (39). The importance of the correct spacing between the adenine residues is emphasised by the target sequences for *EcoR124IΔ (40) and *EcoDXXI*Δ (41). In these systems, where the 3' half of *hsdS* is lost and the two truncated HsdS subunits associate symmetrically, an additional base pair within the spacer sequence maintains the distance between the target adenines (Table 3).*

The genetic concept of families

The separation of type I R–M systems into families originally relied on genetic tests for complementation, first demonstrated for *EcoKI* and *EcoBI* (6,7). Complementation requires sufficient sequence conservation to permit subunits from one

Table 3. Family-specific distance between target adenines

Family	Enzyme	Recognition sequence	Reference
IB	<i>EcoAI</i>	G A GNNNNNNNGTCA	37
	<i>EcoEI</i>	G A GNNNNNNNATGC	19
	<i>CfrAI</i>	GC A NNNNNNNNGTGG	36
IA	<i>EcoBI</i>	TC A NNNNNNNNTGCT	37
	<i>EcoKI</i>	A A CNNNNNNNGTGC	37
	<i>EcoDI</i>	TT A NNNNNNNNGTCY	37
	<i>StyLTHI</i>	G A GNNNNNNNR T AYG	37
	<i>StySPI</i>	A A CNNNNNNNG T RC	37
IC	<i>EcoR124I</i>	GA A NNNNNNNR T CG	38,55
	<i>EcoR124IA</i>	GA A NNNNNNNN T TC	40
	<i>EcoR124II^a</i>	GA A NNNNNNNNR T CG	37
	<i>EcoDXXI^a</i>	TC A NNNNNNNNR T TC	56
	<i>EcoprrI^a</i>	CC A NNNNNNNNR T GC	43
	<i>EcoDXXIA^a</i>	TC A NNNNNNNN T GA	41
ID	<i>StySBLI</i>	CG A NNNNNN T ACC	This work

Where N = any nucleotide, R is either purine and Y is either pyrimidine and the bold type identifies either the adenine that is the target for methylation or the thymine complementary to the target adenine. For *EcoEI*, *CfrAI* and *StySBLI* the relevant adenine residues are not defined by experiments, but are the sole candidates within the target sequences.

^aThese type IC members have four more amino acids within the central conserved region, the region that links the TRDs, than *EcoR124I*.

complex (e.g. *EcoAI*) to substitute for those in another (e.g. *EcoEI*) (16). The principal differences between two members of one family reside in the TRDs, the regions of HsdS that confer sequence specificity to the enzyme (31). Experimental tests for family relationships have come to rely on the more generally applicable approaches of DNA hybridisation screens for similar gene sequences, or serological tests for similar HsdM and HsdR subunits, rather than genetic tests (5,14). The *hsd* genes for *EcoR9I* were identified by hybridisation with a probe derived from the *hsd* genes of *StySBLI*, the archetypal member of the ID family (5). The complementation tests reported in this paper confirm the allocation of *EcoR9I* to the type ID family.

The amino acid sequences predicted for the subunits of enzymes within the established families have >80% identity if the TRDs are ignored, whereas those between families generally indicate only 20–30% identity (18,35,42). The levels of identity in these sequence comparisons explain the ease with which type I R–M systems have been allocated to a family. Such sequence comparisons placed *KpnAI* in the ID family (28). Our genetic tests for complementation provide the classical evidence that *KpnAI*, like *EcoR9I*, is a member of the ID family. The genes for *EcoR9I* and *StySBLI* have a similar location in the chromosome of their respective bacterial species; the location of those for *KpnAI* remains to be determined. However, allelism is not a necessary requirement for membership of a family; one member of the type IC family is specified by chromosomal genes, apparently as part of a defective

prophage (43,44), others are specified by plasmid genes (37,39).

During the past two decades the concept of families of type I restriction enzymes, in which alternative specificity subunits confer different sequence specificities, has provided a pragmatic basis for the description and understanding of these R–M systems. It seems likely that this concept can be extended from enteric bacteria to other genera. In *Helicobacter*, for example, allelic genes specify putative type I R–M systems for which the predicted HsdS subunits seem likely to confer different specificities (11). Similar conclusions can be drawn for *Lactococcus lactis* (45). However, in this case the genetic experiments add a novel complication to the family status. The data from this genus suggest a discrepancy between the family defined by complementation tests and that dependent upon a high level of amino acid identity (>80%) between polypeptides. Schouler and colleagues (45) found for *Lactococcus* that the HsdR and HsdM subunits specified by two plasmids have the high level of identity (~90%) expected for membership of a family, while these sequences have only 40% identity with those specified by the bacterial chromosome. Despite this relatively low level of identity, the plasmid-encoded HsdS subunits were found to interact with the chromosomally encoded HsdM subunits. Type I R–M systems, which by sequence analysis might be separated into two families, were within one family as assessed by complementation tests. A more detailed analysis of the HsdS subunits identified conserved sequences common to all the HsdS subunits, whether of plasmid or chromosomal origin. In addition, all the HsdM subunits share a conserved C-terminus. It was suggested (45) that sequences conserved in all HsdM subunits, and those conserved in all HsdS subunits, identify the sites of interaction between HsdM and HsdS.

On the basis of the *Lactococcus* enzymes, comparisons of the conserved sequences within HsdS could indicate which HsdS subunits might substitute for others, and hence point to the family relationship, but the general similarities detected between the conserved regions of HsdS subunits in different type I families (36,46) suggest that the experimental test for complementation may remain the only test for exchange of subunits. The high level of identity within the C-termini of the chromosomal and plasmid-encoded HsdM subunits could indicate that convergent evolution has enhanced the potential for the generation of enzymes with different specificities. In this way a reservoir of independent, plasmid-borne *hsdS* genes provides an effective ‘combinational’ system for varying the target specificity of the catalytic subunits provided by the host.

Sequence comparisons

We wished to determine whether the striking subdivision of type I R–M systems found within the enteric bacteria can be extended to other genera, or other phyla, of bacteria. Genetic tests are of limited value in the analysis of putative R–M systems provided by the databases of genomic sequences from an extensive array of bacterial species. Polypeptide sequences provide a general approach and current programs compensate for possible errors in the DNA sequence. We therefore decided to search the databases of completed and incomplete genomic sequences, using the TBLASTN program (33), for amino acid sequences similar to those predicted for the archetypal subunits of the type IA, IB, IC and ID families. All HsdM subunits share the motifs of methyltransferases and all HsdR subunits have

the DEAD-box motifs; the motifs correlate with the catalytic activities of the subunits. In the enteric bacteria the current interfamily levels of identity for HsdM extend from 25 to 33% and those for HsdR from 17 to 26% (18,35,42).

In our first screen we used the amino acid sequences of the HsdM polypeptides of the archetypal representatives of each of the four established families of type I enzymes to identify putative HsdM subunits with identities >45% (Table 4). The figure of >45% was chosen as one that was appreciably higher than any found previously for an interfamily comparison, but it sets a high level of identity for comparisons between widely separated genera, where low levels of identity are found when enzymes involved in basic metabolic pathways are compared. For example, the level of identity for triosephosphate isomerase for *E.coli* and *Xylella fastidiosa* is 48%; that between *E.coli* and *Bacillus stearothermophilus* is only 41%.

The polypeptides identified by our screens are listed in Table 4, each section identifying one of the four screens. The table is simplified by the omission of three sequences identified by the screen with the HsdM polypeptide of *EcoR124I* (Table 4C) and some sequences identified by the screen with *StySBLI* (Table 4D). The putative HsdM sequences omitted from Table 4C are a coding sequence in *Ureoplasma urealyticum* which is interrupted by a rearrangement, a coding sequence in *K.pneumoniae* which is within transposon Tn5708 and the sequence in serotype B of *Neisseria meningitidis* which is nearly identical to that identified in serogroup A. The sequences omitted from Table 4D are identified later.

The sequences of the HsdM polypeptides listed in Table 4 were compared with each other using the TBLASTN program. Comparisons within a group usually showed >45% identity, the lowest level was 43% for that between *Streptococcus pneumoniae* and *Mycobacterium avium*; those between sequences allocated to different groups had <35% identity. On this basis, where comparisons relied on alignments of the major portion (>95%) of the subunit, the HsdM polypeptides identified in Table 4 fall into four discrete groups.

The databases were then screened with each of the four archetypal polypeptide sequences for HsdR and HsdS. In those genomic sequences where the HsdM polypeptides were identified by the IA, IB or IC comparisons, HsdR and HsdS polypeptides were detected (Table 4A–C). The coding sequences for the HsdR, HsdM and HsdS polypeptides identified in Table 4 were closely linked, and in the order consistent with the suggested family affiliation; the gene order for the known members of the IA and IB families is *hsdR*, *hsdM*, *hsdS*, while that for the IC and ID families is *hsdM*, *hsdS*, *hsdR*. The levels of identity found for HsdR were 37% or greater, higher therefore than those (17–26%) reported for interfamily comparisons in *E.coli*. The relatively long (~1000 amino acids) HsdR subunits of enteric bacteria show lower levels of identity than HsdM subunits, probably because much of the HsdR polypeptide sequence is not within a predicted catalytic domain (42). Close relationships between HsdS subunits are obscured by the very variable sequences of the TRDs, but for each of the known families conservation is readily detected within the region of HsdS that separates the two TRDs. This region, the so-called central conserved region, is presumed to be involved in the association of HsdS with the catalytic subunits. With the exception of *M.avium*, for which the sequence data were insufficient, the central conserved region of the HsdS subunit was

readily aligned with that of the archetypal representative of the group designated on the basis of identities within HsdM (Table 4).

The screen using the type ID sequences identified a number of HsdM subunits with >45% identity, but some HsdM sequences (*B.stearothermophilus*, *Mycobacterium tuberculosis* and *Mycobacterium bovis*) were not associated with HsdR subunits and others (e.g. *Acidithiobacillus ferrooxidans*, *X.fastidiosa*, *Campylobacter jejuni*, *Pseudomonas syringae* p.v. tomato and a second putative type I R–M system of *Chlorobium tepidum*) were associated with HsdR subunits with <35% identity. These putative type I R–M systems were, therefore, omitted from Table 4. The R–M system of *Pasteurella haemolytica*, referred to as type 1d by Highlander and Garza (47) did not qualify for entry in Table 4D using our stringent criteria; its HsdM subunit has only 42% identity. The sequence comparisons (Table 4D), as predicted, identify *KpnAI* with scores >90% identity, but they indicate even higher scores (96–100% identity) for three polypeptides in *S.enterica* serovar *enteritidis*. These Hsd sequences in *S.enteritidis* correlate well with the biological information that *S.blegdam* and *S.enteritidis* have an R–M system with the same sequence specificity (12).

Three of the screens identified putative type I R–M systems across a wide range of bacterial species. Each of the four screens identify HsdM polypeptides with >45% identity in different genera of the *Gamma* subdivision of the *Proteobacteria* (e.g. *Shewanella*, *Haemophilus* and *Pseudomonas*) and in the *Firmicutes* (*Bacillus*, *Streptococcus* and *Mycobacterium*). In addition, close relatives of *EcoR124I* (type IC) were identified in the *Beta* (*Neisseria*) and *Epsilon* (*Helicobacter*) subdivisions of the *Proteobacteria* and in a member of the *Green Sulphur Bacteria* (*Chlorobium*). The HsdM and HsdR polypeptides listed in Table 4 were aligned by the CLUSTAL W program and the identities determined after the exclusion of sites that contain a gap in any sequence (48). These comparisons (Table 5) identify groups IA, IB, IC and ID on the basis of identities in HsdM and HsdR. Members of the IA group are more closely related to those of IB than they are to those of IC and ID; similarly members of IC are more closely related to those of ID than they are to those of IA and IB. The extension of sequence comparisons to more distantly related bacteria currently permits the unambiguous association of putative type I R–M systems from a wide range of eubacterial species with one of the known families. Of course, many HsdM sequences in the database showed <45% identity with each of the four test sequences. The simplification of our analyses by selection of only M subunits with >45% identity has excluded some sequences that can be associated with one of the four families (P.M.Sharp, personal communication). The relationships analysed in Table 5 are not obviously complicated by genetic recombination. On the basis of sequence comparisons in Tables 4 and 5, we suggest that the family affiliations extend across the eubacterial kingdom. These affiliations are of evolutionary significance, irrespective of whether the groupings meet the classical genetic criterion for family status.

An examination of the aligned central conserved regions of HsdS polypeptides with that of *EcoR124II* indicated an additional feature that is in accord with the subdivisions based on identities within HsdM and HsdR. The archetypal members of the IC family include a tetrapeptide sequence (TAEL) present

Table 4. Comparisons among sequences identified by TBLASTN

Bacterial strain	HsdR ^a	HsdM ^a	S ^b	Reference ^c
(A) Per cent identity with <i>EcoKI</i> polypeptide sequences (type IA)				
<i>S.typhimurium</i> LT2 ^d	74(81) ^e	92(95)	71	<i>Sty</i> LTIII, WUGS 99287 contig 1424
<i>S.typhi</i> CT18	91(95)	93(94)	80	Sanger ORFS STY4884,3 & 1
<i>S.paratyphi</i> A	65(73) ^e	90(91) ^e	80	WUGSC 32027
<i>S.putrefaciens</i>	39(57)	54(69)	71	TIGR24 6431
<i>B.stearothermophilus</i>	37(55)	49(63)	34	UOKN03 1422 contig 715
(B) Per cent identity with <i>EcoAI</i> polypeptide sequences (type IB)				
<i>E.coli</i> O157:H7EDL933	99(99)	98(98)	88	M. <i>Eco</i> O157 ORF 5947P
<i>E.coli</i> A58	77(87)	90(94)	85	<i>EcoEI</i>
<i>P.putida</i> KT2440	61(77)	67(78)	52	TIGR10787
<i>A.ferrooxidans</i>	56(71)	63(78)	62	TIGR6149
<i>S.pneumoniae</i>	48(65)	49(66)	35	TIGR3836
<i>M.avium</i>	39(55)	48(66)	i.d.	TIGR332
(C) Per cent identity with <i>EcoR124II</i> polypeptide sequences (type IC)				
<i>C.tepidum</i>	73(84)	84(90)	82	TIGR3499. J.Eisen (pers. comm.)
<i>N.gonorrhoea</i>	74(85)	75(87)	49	AEOO4969
<i>N.meningitidis</i> serotype A	72(84) ^e	75(86)	31	M. <i>NmeA</i> ORF1038P
<i>Haemophilus influenzae</i> Rd	76(86) ^e	66(72) ^e	53	M. <i>Hind</i> ORF 215P
<i>S.equi</i>	67(82)	66(79)	61	Sanger 1336 contig 445
<i>X.fastidiosa</i>	42(60)	55(70)	54	M. <i>Xfa</i> ORF2728P ^f
<i>H.pylori</i> J99	42(62)	52(69)	46	M. <i>Hpy99</i> ORF786P
<i>H.pylori</i> 26695	43(62)	54(70)	50	M. <i>HpyA</i> ORF 850P
(D) Per cent identity with <i>StySBLI</i> polypeptide sequences (type ID)				
<i>S.enteritidis</i>	96(96) ^e	100(100)	100 ^e	UIUC 592 contigs 1881 & 2214
<i>K.pneumoniae</i>	94(96)	97(98)	95	<i>KpnAI</i>

Data selected where >45% identity for HsdM. The alignments for HsdR and HsdM include >95% of the length of the respective sequence, with the exception of *S.paratyphi* and *S.enteritidis*. The data are omitted for well-established family members (e.g. *EcoBI* and *EcoDXXI*) if the sequences are not available for all three genes.

^aPer cent similarity indicated in brackets.

^bThe central conserved sequence of HsdS, as defined by Sturrock and Dryden (50), was compared to avoid the contribution made by the TRDs; the conserved regions are relatively short (varying from 56 amino acids for *EcoKI* to 155 for *EcoAI*). i.d., insufficient data.

^cSee REBASE (57) for systems identified as enzyme or protein (P) sequence.

^dSerovar of *S.enterica*.

^eSequence alignment impaired by putative frame shifts. The numbers given are for the longest alignment.

^fThe M and S coding sequences are separated by a short ORF (ORF2727), it is not known whether this ORF is an artefact of cloning, or indicates a natural insertion within the coding sequence.

in duplicate in *EcoR124I* and in triplicate in *EcoR124II* (39). An identical, or related, repeat sequence was found in enzymes affiliated with the IC family. A duplicated TAEL sequence is present in *C.tepidum* and *Streptococcus equi*. Similar repeat sequences are present in triplicate in *Neisseria gonorrhoeae* (EATL), in *Haemophilus* (TSEL), in *Xylella* (EAEL) and in *Helicobacter pylori* 26695 (NTEL).

The specificity subunits

The HsdS subunits identified by searches using the predicted amino acid sequence of the specificity subunit of *StySBLI* provide additional interest (Table 6). Either an N- or C-terminal segment of the polypeptide, each presumed to be a

TRD sequence, often enhances the level of identity. This information is lost when the entire HsdS polypeptides are compared. While the predicted carboxy TRD of the HsdS subunit from *C.tepidum* has 50% identity with *StySBLI*, the predicted amino TRD has merely 10% identity. This is in contrast to HsdS subunits from *K.pneumoniae*, *P.syringae*, *A.ferrooxidans* and *Actinobacillus actinomycetemcomitans* in which the predicted amino TRDs each have ~45% identity and the carboxy TRDs ~20%. All these examples suggest sequence conservation between the test sequence and one TRD. In an early comparison of the predicted amino acid sequence of *EcoAI* (IB) with those of five members of the IA family, marked similarity in the HsdS subunits was detected only with

Table 5. Sequence comparisons within and between groups

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	<i>EcoKI</i> (IA)	–	95	60	55	31	30	28	28	29	22	22	21	22	22	19	19	18	18
2	<i>StyLTIII</i>	92	–	60	56	31	30	28	27	29	22	22	22	23	22	19	19	18	18
3	<i>S.putrefaciens</i>	47	48	–	55	34	33	29	29	30	20	21	20	20	21	19	17	20	20
4	<i>B.stearothermophilus</i>	44	44	51	–	32	31	29	27	29	21	21	20	22	21	21	20	19	19
5	<i>EcoAI</i> (IB)	25	24	23	25	–	70	67	52	51	22	24	23	23	24	20	23	19	19
6	<i>P.putida</i>	25	25	23	24	64	–	70	53	50	21	23	22	22	24	22	22	18	18
7	<i>A.ferrooxidans</i>	25	24	25	25	58	60	–	52	52	20	23	21	22	22	20	21	17	17
8	<i>S.pneumoniae</i>	23	23	22	22	52	53	51	–	41	20	21	21	20	21	19	22	18	18
9	<i>M.avium</i>	23	23	23	23	39	39	39	37	–	22	23	21	21	21	18	19	19	18
10	<i>EcoRI24I</i> (IC)	15	15	15	14	15	15	14	14	14	–	86	82	78	69	56	58	29	29
11	<i>C.tepidum</i>	15	15	14	14	15	16	15	14	15	81	–	81	78	70	55	57	30	29
12	<i>Hind</i> ORF215P	15	16	15	15	15	16	14	15	15	81	79	–	79	73	56	58	30	29
13	<i>N.gonorrhoea</i>	16	16	14	16	15	16	15	15	15	80	82	80	–	70	54	56	29	28
14	<i>S.equi</i>	15	15	14	14	14	15	14	14	15	74	73	75	75	–	59	59	28	28
15	<i>Hpy99</i> ORF786P	15	16	16	15	16	15	16	16	16	49	49	48	49	49	–	62	32	31
16	<i>Xfa</i> ORF2728P	15	15	15	16	15	15	16	16	14	48	50	48	49	48	58	–	31	31
17	<i>StySBLI</i> (ID)	13	14	12	13	15	15	14	13	14	22	22	22	22	20	20	20	–	98
18	<i>KpnAI</i>	14	14	12	13	14	15	14	13	14	21	21	21	21	20	19	21	95	–

The R–M systems, or bacterial species, identified by numbers 1–18 are listed in the left-hand column; for 12, 15 or 16 the putative R–M system is identified by the ORF for HsdM (57). The values are the per cent identity of aligned sequences after the exclusion, from any pairwise comparison, of any site that contains a gap in any sequence; above the diagonal for HsdM and below the diagonal for HsdR. The sequences were aligned by the CLUSTAL W(1.5) multiple sequence alignment program (48), with a minor adjustment for HsdR. Values for comparisons within groups are given in bold. For each group only one representative of any genus was included. The maximum difference for comparisons between two representatives of the same group from one genus is that for *EcoAI* and *EcoEI* (10% for HsdM and 20% for HsdR), the smallest that for *EcoAI* and the polypeptides of *E.coli* O157 (1 and 0.25%, respectively).

Table 6. Sequence comparisons based on HsdS of *StySBLI*

Bacterial strain	Per cent identity	Length of alignment	Per cent identity in ^a			Reference ^b
			N-TRD	Centre	C-TRD	
<i>S.enteritidis</i>	100	434 ^c	100	100	100	UIUC 592 contig 2214
<i>E.coli</i> ECOR9	43	300	^d	89	21	<i>EcoRI</i> (this work)
<i>K.pneumoniae</i>	44	437	41	95	21	<i>KpnAI</i>
<i>C.tepidum</i> ^e	44	327	10	34	50	TIGR 3499
<i>Pasteurella multocida</i> PM70	36	344	12	48	34	CBU MN747 AE006190
<i>Ps.syringae</i> pv tomato	34	433	44	43	21	TIGR 323
<i>A.ferrooxidans</i> ^e	32	427	46	38	15	TIGR 6154
<i>A.actinomycetemcomitans</i>	32	421	43	24	24	OUACGT714

HsdS subunits with >30% identity in HsdS (434 amino acids) and >40% in some region of HsdS.

^aThe per cent identity in each of the three regions was calculated on the basis of the alignments given by TBLASTN for the entire HsdS sequence of *StySBLI*; the regions are defined according to Sturrock and Dryden (50).

^bFor enzymes and proteins see REBASE (57).

^cSequence alignment impaired by putative frame shift.

^dThere is insufficient identity in the N-TRD for the TBLASTN program to make an alignment. For an alignment by PILEUP see Figure 3.

^eThe subunit is not that identified in Table 4.

StyLTIII (previously called *StySB*) (19). Residues within the N-terminal TRDs of *EcoAI* and *StyLTIII* showed 44% identity. The similarity of the sequences of the TRDs of these enzymes

correlates with their recognition of the trinucleotide GAG. This and other evidence (49) indicate that TRDs from different families are of similar sequence if they confer the same

sequence specificity. This similarity between TRDs in distantly related bacteria appears to imply a closer evolutionary relationship than those between dissimilar TRDs in the same family. It could, however, reflect some structural constraint on TRDs that recognise the same nucleotide sequence. A recent analysis suggested that all type I TRDs include a region with a similar conserved tertiary structure at the interface with DNA (50). Whatever the underlying explanation, the sequence similarities imply a conservation of TRDs in type I R–M systems, irrespective of family relationships or bacterial phylogeny.

Comparative analyses of the sequences of type II restriction endonucleases and modification methylases rarely document evidence for the relatedness of amino acid sequences outside of the catalytic domains (51). For many type II enzymes, structures of cocrystals are available in which the enzymes are bound to their target sequence. On the basis of these, diversity in the structure of the polypeptide sequences interfacing with DNA has been emphasised (52). In general, relatively little support for the evolutionary connections between different type II systems has been presented. This contrasts with the information from type I R–M systems where closely related groups, previously obvious for the enteric bacteria, are found in widely different bacterial phyla. This is not unexpected if the type I R–M systems have a common origin. Extremely tight post-translational control of restriction activity in the absence of adequate modification should facilitate the evolution of new specificities (53,54).

ACKNOWLEDGEMENTS

We thank our colleagues for their interest and support, particularly Costel Atanasiu and Erica de Leau. We are indebted to Jonathan Eisen, Rich Roberts, Paul Sharp and Geoff Wilson for their comments on the manuscript, Paul Sharp for generating Table 5, and to Natalie Honeyman and Alix Fraser for their tireless efforts in the preparation of this manuscript. We wish to acknowledge the many contributors to REBASE and the databases of bacterial genomes including the following supplied by personal communication: The Pathogen Sequencing Unit at The Sanger Centre for the sequences of *Salmonella typhi* and *S.equi*; The Institute for Genomic Research for the genomes *C.tepidum*, *M.avium*, *Pseudomonas putida*, *P.syringae*, *Shewanella putrefaciens*, *S.pneumoniae*, *A.ferrooxidans*; the sequencing centres at the University of Washington for WUGS 99287 (*Salmonella typhimurium*), 3207 (*Salmonella paratyphi*); The University of Illinois for UIUC592 (*S.enteritidis*); The University of Oklahoma for OUKN03 1442 (*B.stearothermophilus*) and Ngon 485 (*N.gonorrhoea*). The Gonococcal Genome Sequencing Project was supported by USPHS/NIH grant #AI38399. The work in Edinburgh was supported by the Medical Research Council, that in Loma Linda by grant number NMTBO315-8853-01.

REFERENCES

- Raleigh, E.A. (1992) Organization and function of the *mcrBC* genes of *Escherichia coli* K-12. *Mol. Microbiol.*, **6**, 1079–1086.
- Bertani, G. and Weigle, J.J. (1953) Host-controlled variation in bacterial viruses. *J. Bacteriol.*, **65**, 113–121.
- Sain, B. and Murray, N.E. (1980) The *hsd* (host specificity) genes of *E.coli* K-12. *Mol. Gen. Genet.*, **180**, 35–46.
- Daniel, A.S., Fuller-Pace, F.V., Legge, D.M. and Murray, N.E. (1988) Distribution and diversity of *hsd* genes in *Escherichia coli* and other enteric bacteria. *J. Bacteriol.*, **170**, 1775–1782.
- Barcus, V.A., Titheradge, A.J.B. and Murray, N.E. (1995) The diversity of alleles at the *hsd* locus in natural populations of *Escherichia coli*. *Genetics*, **140**, 1187–1197.
- Boyer, H.W. and Roulland-Dussoix, D. (1969) A complementation analysis of the restriction and modification of DNA in *Escherichia coli*. *J. Mol. Biol.*, **41**, 459–472.
- Glover, S.W. and Colson, C. (1969) Genetics of host-controlled restriction and modification in *Escherichia coli*. *Genet. Res.*, **13**, 227–240.
- Redaschi, N. and Bickle, T.A. (1996) DNA restriction and modification systems. In Neidhart, F.C., Curtiss, R., Ingraham, J.L., Lin, E.C.C., Low, K.B., Magasanik, B., Reznikoff, W.S., Riley, M., Schaechter, M. and Umberger, H.E. (eds), *Escherichia coli and Salmonella*, 2nd Edn. American Society for Microbiology, Washington, DC, pp. 773–781.
- Dryden, D.T.F. (1999) Bacterial methyltransferases. In Cheng, X. and Blumenthal, R.M. (eds), *S-Adenosylmethionine-Dependent Methyltransferases*. World Scientific Publishing, Singapore, pp. 283–340.
- Rao, D.N., Saha, S. and Krishnamurthy, V. (2000) The ATP-dependent restriction enzymes. *Prog. Nucleic Acid Res. Mol. Biol.*, **64**, 1–63.
- Murray, N.E. (2000) Type I restriction systems: sophisticated molecular machines. *Microbiol. Mol. Biol. Revs.*, **64**, 412–434.
- Bullas, L.R., Colson, C. and Neufeld, B. (1980) Deoxynucleic acid restriction and modification systems in *Salmonella*: chromosomally-located systems of different serotypes. *J. Bacteriol.*, **141**, 275–292.
- Barcus, V.A. and Murray, N.E. (1995) Barriers to recombination. In Baumberg, S., Young, J.P.W., Saunders, S.R. and Wellington, E.M.H. (ed.) *Population Genetics of Bacteria*. Society for General Microbiology, Cambridge University Press, pp. 31–58.
- Murray, N.E., Gough, J.A., Suri, B. and Bickle, T.A. (1982) Structural homologies among type I restriction–modification systems. *EMBO J.*, **1**, 535–539.
- Arber, W. and Wauters-Willems, D. (1970) Host specificity of DNA produced by *Escherichia coli*. XII. The two restriction and modification systems of strain 15T⁻. *Mol. Gen. Genet.*, **108**, 203–217.
- Fuller-Pace, F.V., Cowan, G.M. and Murray, N.E. (1985) EcoA and EcoE: alternatives to the EcoK family of type I restriction and modification systems of *Escherichia coli*. *J. Mol. Biol.*, **186**, 65–75.
- Suri, B. and Bickle, T.A. (1985) EcoA: the first member of a new family of type I restriction modification systems. Gene organization and enzymatic activities. *J. Mol. Biol.*, **186**, 77–85.
- Titheradge, A.J.B., Ternent, D. and Murray, N.E. (1996) A third family of allelic *hsd* genes in *Salmonella enterica*: sequence comparisons with related proteins identify conserved regions implicated in restriction of DNA. *Mol. Microbiol.*, **22**, 437–447.
- Cowan, G.M., Gann, A.A.F. and Murray, N.E. (1989) Conservation of complex DNA recognition domains between families of restriction enzymes. *Cell*, **56**, 103–109.
- Ochman, H. and Selander, R.K. (1984) Standard reference strains of *E.coli* from natural populations. *J. Bacteriol.*, **157**, 690–693.
- Fowler, G., Degnen, E. and Cox, E.C. (1974) Mutational specificity of a conditional *Escherichia coli* mutator *mutD5*. *Mol. Gen. Genet.*, **133**, 179–191.
- Hansen, E.B., Atlung, T., Atlung, F.G., Skovgaard, O. and Van Mayenberg, K. (1984) Fine structure genetic map and complementation analysis of mutations in the *dnaA* gene of *Escherichia coli*. *Mol. Gen. Genet.*, **196**, 387–396.
- Grant, S.G.N., Jesse, J., Bloom, F.R. and Hanahan, D. (1990) Differential plasmid rescue from transgenic mouse DNAs into *Escherichia coli* methylation restriction mutants. *Proc. Natl Acad. Sci. USA*, **87**, 4645–4649.
- Bork, K., Beggs, J.D., Brammar, W.J., Hopkins, A.S. and Murray, N.E. (1976) The construction *in vitro* of transducing derivatives of Phage Lambda. *Mol. Gen. Genet.*, **146**, 199–207.
- King, G. and Murray, N.E. (1995) Restriction alleviation and modification enhancement by the Rac prophage. *Mol. Microbiol.*, **16**, 769–777.
- Prakash-Cheng, A., Chung, S.S. and Ryu, J. (1993) The expression and regulation of *hsdK* genes after conjugative transfer. *Mol. Gen. Genet.*, **241**, 491–496.
- Frischauf, A.-M., Lehrach, H., Poustka, A. and Murray, N.E. (1983) Lambda replacement vectors carrying polylinker sequences. *J. Mol. Biol.*, **170**, 827–842.

28. Lee, N.S., Rutebuka, O., Arakawa, T., Bickle, T.A. and Ryu, J. (1997) *KpnAI*, a new type I restriction-modification system in *Klebsiella pneumoniae*. *J. Mol. Biol.*, **271**, 342–348.
29. Murray, N.E. and Murray, K. (1974) Manipulation of restriction targets in phage λ to form receptor chromosomes for DNA fragments. *Nature*, **251**, 476–481.
30. Murray, N.E., Brammar, W.J. and Murray, K. (1977) Lambdoid phages that simplify the recovery of *in vitro* recombinants. *Mol. Gen. Genet.*, **150**, 53–61.
31. Gough, J.A. and Murray, N.E. (1983) Sequence diversity among related genes for recognition of specific targets in DNA molecules. *J. Mol. Biol.*, **166**, 1–9.
32. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd Edn. Cold Spring Harbor Laboratory Press, Plainview, NY.
33. Altschul, S.F., Madden, T.L., Schäffer, J.Z., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3384–3402.
34. Gann, A.A.F., Campbell, A.J.B., Collins, J.F., Coulson, A.F.W. and Murray, N.E. (1987) Reassortment of DNA recognition domains and the evolution of new specificities. *Mol. Microbiol.*, **1**, 13–22.
35. Sharp, P.M., Kelleher, J.E., Daniel, A.S., Cowan, G.M. and Murray, N.E. (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.
36. Kannan, P., Cowan, G.M., Daniel, A.S., Gann, A.A.F. and Murray, N.E. (1989) Conservation of organization in the specificity polypeptides of two families of type I restriction enzymes. *J. Mol. Biol.*, **209**, 335–344.
37. Bickle, T.A. (1987) Restriction and modification systems. In Neidhardt, F.C. (ed.), *Escherichia coli and Salmonella*, Edn 1. ASM Press, Washington DC, pp. 692–696.
38. Taylor, I., Watts, D. and Kneale, G. (1993) Substrate recognition and selectivity in the type IC DNA modification methylase *M.EcoR124I*. *Nucleic Acids Res.*, **21**, 4929–4935.
39. Price, C., Lingner, J., Bickle, T.A., Firman, K. and Glover, S.W. (1989) Basis for changes in DNA recognition by the *EcoR124* and *EcoR124/3* type I DNA restriction and modification enzymes. *J. Mol. Biol.*, **205**, 115–125.
40. Abadjieva, A., Webb, M., Patel, J., Zinkevich, V. and Firman, K. (1993) Deletions within the DNA recognition subunit of *M.EcoR124I* that identify a region involved in protein-protein interactions between HsdS and HsdM. *J. Mol. Biol.*, **241**, 35–43.
41. Meister, J., MacWilliams, M., Hubner, P., Jutle, H., Skrzypek, E., Piekarczyk, A. and Bickle, T.A. (1993) Macroevolution by transposition: drastic modification of DNA recognition by a type I restriction enzyme following *Tn5* transposition. *EMBO J.*, **12**, 4585–4591.
42. Murray, N.E., Daniel, A.S., Cowan, G.M. and Sharp, P.M. (1993) Conservation of motifs within the unusually variable polypeptide sequences of type I restriction and modification enzymes. *Mol. Microbiol.*, **9**, 133–143.
43. Tyndall, C., Meister, J. and Bickle, T.A. (1994) The *Escherichia coli prr* region encodes a functional type IC DNA restriction system closely integrated with an anticodon nuclease gene. *J. Mol. Biol.*, **237**, 266–274.
44. Tyndall, C., Lehnerr, H., Sandmeier, U., Kulik, E. and Bickle, T.A. (1997) The type IC *hsd* loci of the enterobacterial *arr* flanked by DNA with high homology to the phage P1 genome: implications for the evolution and spread of DNA restriction systems. *Mol. Microbiol.*, **23**, 729–736.
45. Schouler, C.M., Gautier, M., Ehrlich, S.D. and Chopin, M.-C. (1998) Combinational variation of restriction-modification specificities in *Lactococcus lactis*. *Mol. Microbiol.*, **28**, 169–178.
46. Kneale, G.G. (1994) A symmetrical model for the domain structure of type I DNA methyltransferases. *J. Mol. Biol.*, **243**, 1–5.
47. Highlander, S.K. and Garza, D. (1996) The restriction-modification system of *Pasteurella haemolytica* is a member of a new family of enzymes. *Gene*, **178**, 89–96.
48. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4672–4680.
49. Thorpe, P.H., Ternent, D. and Murray, N.E. (1997) The specificity of *StySKI*, a type I restriction enzyme, implies a structure with rotational symmetry. *Nucleic Acids Res.*, **25**, 1694–1700.
50. Sturrock, S.S. and Dryden, D.T.F. (1997) A prediction of the amino acids and structures involved in DNA recognition by type I DNA restriction and modification enzymes. *Nucleic Acids Res.*, **25**, 3408–3414.
51. Wilson, G.G. and Murray, N.E. (1991) Restriction and modification systems. *Annu. Rev. Genet.*, **25**, 585–627.
52. Lukacs, C.M. and Aggarwal, A.K. (2001) *BglII* and *MunI*: what a difference a base makes. *Curr. Opin. Struct. Biol.*, **11**, 14–18.
53. Makovets, S., Doronina, V.A. and Murray, N.E. (1999) Regulation of endonuclease activity by proteolysis prevents breakage of unmodified bacterial chromosomes by type I restriction enzymes. *Proc. Natl Acad. Sci. USA*, **96**, 9757–9762.
54. O'Neill, M., Powell, L.M. and Murray, N.E. (2001) Target recognition by *EcoKI*; the recognition domain is robust and restriction-deficiency commonly results from the proteolytic control of enzyme activity. *J. Mol. Biol.*, **307**, 951–963.
55. Price, C., Shepherd, J.C.W. and Bickle, T.A. (1987) DNA recognition by a new family of type I restriction enzymes: a unique relationship between two different DNA specificities. *EMBO J.*, **6**, 1493–1497.
56. Gubler, M., Graguglia, D., Meyer, J., Piekarczyk, A. and Bickle, T.A. (1992) Recombination of constant and variable modules alters DNA sequence recognition by type IC restriction-modification enzymes. *EMBO J.*, **11**, 233–240.
57. Roberts, R.J. and Macelis, D. (2001) REBASE: restriction enzymes and methylases. *Nucleic Acids Res.*, **29**, 268–269.

