



Conservation of Complex DNA Recognition Domains between Families of Restriction Enzymes

Gill M. Cowan, Alexander A. F. Gann,
and Noreen E. Murray
Department of Molecular Biology
University of Edinburgh
Edinburgh EH9 3JR
Scotland

Summary

One polypeptide, designated S, confers sequence-specificity to the multisubunit type I restriction enzymes. Two families of such enzymes, K and A, include members that recognize diverse, bipartite, target sequences. The S polypeptides of the K family, while having areas of near identity, also contain two extensive regions of variable sequence. We now show that one of these, comprising the N-terminal 150 amino acids, specifies recognition of one component of the bipartite target sequence. We have determined the sequence recognized by EcoE, a member of the A family. This sequence, 5'GAG(N₇)ATGC, has the trinucleotide GAG in common with EcoA and with StySB of the K family. We determined the nucleotide sequences of the S genes of EcoA and EcoE, and compared their predicted amino acid sequences with each other and with those of the five members of the K family. There is no general sequence similarity between families, but the domain of the S polypeptide of StySB, which specifies GAG, shows nearly 50 per cent identity with the amino variable region of the S polypeptides of EcoA and EcoE. A complex domain that recognizes and directs methylation of GAG is therefore common to enzymes of generally dissimilar amino acid sequence.

Introduction

Families of enzymes whose members recognize different, but specific, DNA sequences are presumed to have diverged from a common ancestor. Comparisons of the amino acid sequences of their polypeptides could therefore identify residues that interact with the nucleotides of their target sequences. The chromosomally encoded, multisubunit restriction enzyme of *E. coli* K-12 (EcoK) is a member of such a family of type I restriction and modification (R-M) systems. Relatedness within this family was originally shown by complementation tests that require the interchange of subunits between enzymes (Boyer and Roulland-Dussoix, 1969; Glover and Colson, 1969; Bullas and Colson, 1975), and has been reinforced by molecular evidence indicating cross-hybridization between their genes and cross-reactivity of antibodies raised against their subunits (Murray et al., 1982). Other, apparently allelic, genes encode an alternative family of type I R-M systems, including EcoA and EcoE. Although the overall organization of

these enzymes and the genes encoding them is similar to EcoK, they form a second discrete family on the basis of the three criteria given above (Suri and Bickle, 1985; Fuller-Pace et al., 1985). Complementation tests between mutant members of the K family (Boyer and Roulland-Dussoix, 1969; Glover and Colson, 1969), and more recently those of the A family (Fuller-Pace et al., 1985), indicate that only one of three subunits of the restriction enzyme, the S polypeptide, confers specificity of DNA recognition.

The restriction enzyme, comprising the three polypeptides R, M, and S, can also function as a methylase (modification enzyme), but a complex of M and S alone is sufficient for the latter function. Methylation of either or both of two defined adenine residues of the target sequence prevents restriction (see Bickle, 1987, for a review). A phage with an unmethylated target sequence in its genome is identified by its reduced efficiency of plating on a restricting strain.

The target sequences recognized by all type I R-M enzymes are asymmetric and consist of two defined components, one of 3 bp, and another of 4 or 5 bp, separated by a nonspecific spacer of fixed length (see Bickle, 1987). Consistent with this, the specificity (S) polypeptides of the K family are known to contain two DNA recognition domains, each specifying recognition of one of the two defined components of the target sequence. Reassortment of these two polypeptide domains can result in novel specificities (see Figure 1; Fuller-Pace et al., 1984; Nagaraja et al., 1985a; Gann et al., 1987). The K family includes EcoK, EcoB, and EcoD from *E. coli*, and StySP and StySB from *Salmonella*, and sequence comparisons between the five S polypeptides, each conferring a different specificity, reveal two large regions that vary greatly. We have suggested that these two variable regions may correlate with recognition domains (Gough and Murray, 1983; Gann et al., 1987). In this paper, we demonstrate that one of these, which we define as the amino recognition domain, specifies the trinucleotide component of the target sequence. The S genes of two members of the second family of R-M systems, EcoA and E, whose target sequences include GAG, were sequenced. The only member of the K family with which they show any similarity is StySB, and this is confined to the region in the latter identified as the recognition domain for GAG.

Results

The N-Terminal Domain of S Specifies the Trinucleotide Component of the Target Sequence

The two extensive regions that are very different in S genes of the K family are referred to as variable regions. They are about 450 bp in length and are separated by a well conserved region of approximately 100 bp (Gough and Murray, 1983). Crossing over between the central conserved regions of two S genes, those of StySB and SP, has generated novel specificities StySQ and SJ (Fuller-

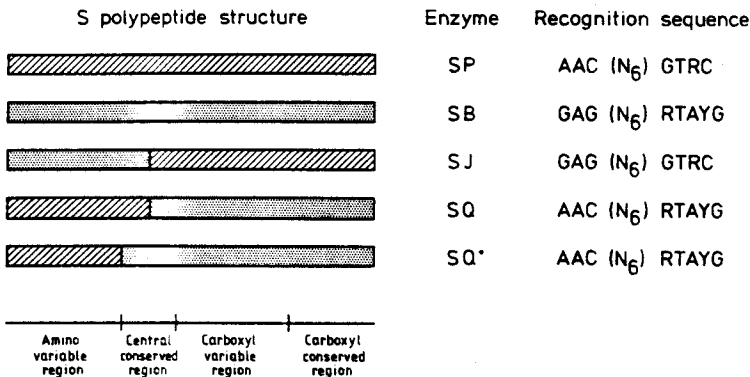


Figure 1. Schematic Diagram of Some Wild-Type and Hybrid S Polypeptides of the K Family, Accompanied by Their Recognition Sequences

StySQ and SJ were produced by homologous recombination (Bullas et al., 1976; Gann et al., 1987). Recognition sequences: StySP and SB (Nagaraja et al., 1985b), StySQ (Nagaraja et al., 1985a), StySJ (Gann et al., 1987), StySQ* (this paper).

StySP and SB are R-M systems from natural isolates of *S. typhimurium* (Bullas and Colson, 1975); StySJ, SQ, and SQ* are encoded by recombinant S genes. Regions originating from StySP are hatched; those from StySB are stippled. When all S polypeptides from the K family are compared, regions of great variation, as well as areas of conservation, are found. The approximate positions of these are indicated at the bottom of the figure. The amino and carboxyl conserved regions are ~30 and 80 amino acids, respectively. The S genes of

Pace et al., 1984; Gann et al., 1987). These recombinant S genes specify target sequences with one component from each of the two parental sequences (see Figure 1; Nagaraja et al., 1985a, 1985b; Gann et al., 1987), clearly demonstrating that there are two structurally independent DNA recognition domains within an S polypeptide, each specifying one defined component of the target sequence. These experiments suggested that recognition is specified by the variable regions, but this has not been proven, since recombination also reassorted minor differences in the central conserved region. Indeed, a model proposed by Argos (1985), based on a computer analysis of the sequences of three S polypeptides, implicated the region including these very residues in DNA recognition.

Between the left end of the central conserved region and the point of exchange that resulted in the formation of the recombinants StySQ and SJ, the parental genes differ in four codons. We changed all four of these in the S gene of StySQ, which has the amino-terminal half of StySP, such that they encoded the equivalent residues of StySB. This derivative, designated StySQ*, therefore encodes a polypeptide whose amino variable region is from StySP, while the remainder of the molecule is identical to that of StySB (Figures 1 and 2; construction described in Experimental Procedures).

The specificity of StySQ* was examined using a bacterial strain in which this newly constructed S gene is expressed from the chromosome in conjunction with the M and R subunits necessary to produce a functional restriction and modification enzyme. This strain was shown to restrict phage, but not if already modified by growth in a host specifying either the StySQ* or the StySQ system. Also, phage modified by StySQ* are not restricted by StySQ (see Experimental Procedures). This demonstrates that the specificities of StySQ and SQ* are identical.

We conclude that the N-terminal domain of 150 amino acids alone is sufficient to confer on the hybrid polypeptide specificity for the trinucleotide component of its target sequence, and therefore infer that the high variability of this domain is, at least in part, a function of the potential for interaction with different target sequences.

The Recognition Sequence of EcoE

Identification of the recognition domains within the S polypeptides of K family enzymes depended on comparisons of both their predicted amino acid and recognition sequences. For all members of this family, the target sequences have been determined (see Bickle, 1987); for the A family, only that of EcoA itself has been reported (Suri et al., 1984). We have now determined the recognition sequence of EcoE, using a biological approach that obviates the need for enzyme purification. This simple strategy (Gann et al., 1987) relies upon the reduced efficiency of plating of those phage that contain an unmodified target sequence when they are assayed on a strain encoding the appropriate restriction system (Arber and Kühnlein, 1967; Franklin and Dove, 1969). Unmodified phage M13 vectors

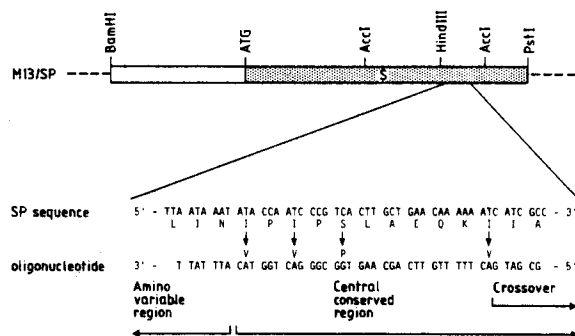


Figure 2. Site Directed Mutagenesis of StySQ*

The top line shows the BamHI-PstI fragment containing part of the S gene from StySP cloned in mp18 that was used as the template for mutagenesis. At the bottom of the figure, the positions of the amino variable and central conserved regions are indicated, as is the region in which crossing over produced the recombinant S genes of StySQ and SJ. The sequence of the 45 base oligonucleotide used for mutagenesis is shown, along with the region of the S gene to which it binds. Arrows identify the four mismatches and these changes alter four codons in StySP such that they encode the equivalent amino acids of StySB. The AccI fragment contains all the changes, and was used to replace the equivalent fragment in the S gene of StySQ (in pAG12) to produce StySQ* (see Experimental Procedures and Figure 1 for details).

Table 1. The Recognition Sequence of EcoE

	Nucleotide Sequence ^a					Source
Positives^b						
1	G	GAG	CACGATC	ATGC	G	pBR322
2	A	GAG	TCGACCG	ATGC	C	pBR322-M13
3	A	GAG	CGAACTG	ATGC	A	hsdK
4	C	GAG	CCAACCTG	ATGC	T	ftsQA
5	T	GAG	GTGAACA	ATGC	T	HBV
6	G	GAG	GCCATTG	ATGC	C	λ
Negatives^c						
7	C	AAG	GAAGAGA	ATGC	G	hsdK
8	C	GGG	CATCCCG	ATGC	C	pBR322
9	T	GAA	TTACCTT	ATGC	G	M13
10	T	GAG	CGAGGGC	GTGC	A	pBR322
11	A	GAG	TTGTTTCG	ACGC	G	hsdSP
12	T	GAG	TACGGTG	ATAC	A	M13
13	T	GAG	CGTCAAA	ATGT	A	M13
Consensus						
	N	GAG	NNNNNNN	ATGC	N	

For clarity, the sequences are written with gaps separating the flanking bases, the trimeric component, the spacer, and the tetrameric component. Positives (1-6) are sequences within DNA fragments that confer sensitivity to restriction. Negatives (7-13) are degenerate versions of the consensus sequence, and do not confer sensitivity to restriction. 1 is in pBR322 (Sutcliffe, 1978); 2, in a sequence created at the junction of a of a *TaqI* fragment of pBR322 with the M13 polylinker of mp18; 3, in the *E. coli* *hsdK* sequence (Gough and Murray, 1983); 4 in the *E. coli* *ftsQA* sequence (Robinson et al., 1984); 5 in Hepatitis B virus (Pugh et al., 1986); 6 in phage λ (Sanger et al., 1982); 7, in the *E. coli* *hsdK* sequence (Loenen et al., 1987); 8 and 10 are in pBR322 (Sutcliffe, 1978); 9, 12, and 13, are in M13 (van Wezenbeck et al., 1980); and 11 is in the *Salmonella* *hsdSP* sequence (Fuller-Pace and Murray, 1986).

plate with an efficiency of one on an E-restricting strain, and therefore are assumed to lack a target sequence for EcoE. Derivatives, including fragments of known nucleotide sequence, were screened, and any recombinant phage that included an EcoE target sequence was detected by its relatively low efficiency of plating (usually $\sim 10^{-1}$) on the E-restricting strain. Libraries of previously sequenced fragments were used for this analysis.

Our computer-aided search for the target sequence for EcoE assumed that it, like those of all other type I R-M systems, would have two components separated by a spacer. Allowing variation of spacer length from 5 to 8 bases identified only one 7 base interrupted sequence, 5'GAG(N₇)ATGC, present in all sensitive phage, but absent from all those that were resistant to restriction (Table 1). All degenerate versions of the candidate sequence were found in phage previously scored as resistant to restriction (Table 1). We conclude that the recognition sequence of EcoE is 5'GAG(N₇)ATGC: EcoE, like EcoA and StySB, recognizes the trinucleotide GAG.

Comparison of S Polypeptides

The S polypeptides of the A family, like those of K, include two large variable regions. In the K family, the amino variable region is the N-terminal 150 amino acids, while the equivalent region in the A family is preceded by approximately 100 amino acids (Cowan and Kannan, unpublished data). We have shown above that the amino variable region of an S polypeptide from the K family specifies recognition of the trinucleotide component of the target sequence. As EcoA and EcoE both recognize GAG, very similar amino variable regions would correlate with these regions specifying recognition of this trinucleotide. Fur-

thermore, StySB from the K family also recognizes GAG. It is of interest to compare recognition domains of identical specificity from enzymes that are only distantly related.

We determined the nucleotide sequences of the S genes of EcoA and E (the nucleotide sequences are not shown here but are available from the GenBank database; accession numbers J03150 and J03162), and compared the predicted amino acid sequences (Figure 3). The N-terminal halves of these, including their amino variable regions, are indeed remarkably similar, while in the carboxyl halves, their variable regions are readily identified.

The amino acid sequences of EcoA and E were compared with those of all five of the K family using the program COMPARE (Devereux et al., 1984). Only with StySB was any obvious similarity detected. Residues 101-247 of either EcoA, or EcoE and the N-terminal 150 amino acids of StySB, show 44% identity (Figure 3). This similarity is within the amino recognition domains of S polypeptides of enzymes that recognize the same trinucleotide sequence. We have therefore identified a domain that is capable of recognizing and directing methylation of the trinucleotide GAG and that is conserved in polypeptides of otherwise dissimilar amino acid sequence.

Discussion

Our site-directed mutagenesis experiment localizes a DNA recognition domain to the N-terminal 150 amino acids of the S polypeptides of the K family of type I R-M enzymes. This domain is one of two that together define the specificity of the enzyme's bipartite, asymmetric target sequence. Similarly, it is one of two extraordinarily variable regions, and while we have here proven that the

	1		50
E.PEP	MAVEKLI	TDH IDIWS	SALQT RSMAGRGSNG KIDLYGIKKL RELILELAVR
A.PEP	MSVEKLI	VDH METWT	SALQT RSTAGRGSNG KIDLYGIKKL RELILELAVR
SB.PEP
	51		100
E.PEP	GKLV	QDPND EPASELLKRI	AAEKTELVKQ GKIKKQKPL RISEDEKPF
A.PEP	GKLV	QDPND EPASELLKRI	AAEKTELVKQ GKIKKQKPL RISEDEKPF
SB.PEPMSGGK
	101		150
E.PEP	LPEG	MEWITL sEIA	TINPKI EVsDDEQEIS FVPMPCISTR FDGAHDOEIK
A.PEP	LPOG	MEWITL tRIA	EINPKI DVsDDEQEIS FIPMPLISTR FDGSHEFEIK
SB.PEP	LPEG	WATSTI nEMCN	LNPKL KL.DDDLDVG FMPMAGVPTT YLgKCNFEtK
	151		200
E.PEP	KWGE	VKKGYT HFAD	GDIALA KITPCFENSK AVIFKGLKGG VGVGTTTELHV
A.PEP	KMKD	VKKGYT HFANG	DIAIA KITPCFENSK AAIFsGLKNG IGVGTTTELHV
SB.PEP	KWSE	VKKGFT QFND	DIIPA KITPCFENSK AVViKEFPNG YGAGsTEyYV
	201		250
E.PEP	ARPI	SsELNL oYILL	NKSP hYLSmGESmM TGSAGQKRV RFFENYPIP
A.PEP	ARPF	SdIINr rYLL	NfKSP nFLksGESmM TGSAGQKRV RFFENYPIP
SB.PEP	LRS	INGLIMP hWLF	ALVktK dFLTNGALNM sGSvGhKRVT KsFLenYGVp
	251		300
E.PEP	FPPN	TEQARI VgtF	sRkLmFL CDQLEQQSLT SLDAAHQQLVE TLLATLTDSQ
A.PEP	FPPLO	EQERI iIRF	TOLmSL CDQLEQQSLT SLDAAHQQLVE TLLgTLTDSQ
SB.PEP	VPP	LAEQxVI aEK	LDTLL.....
	301		350
E.PEP	NAE	EELAE	NWA RISQyFDTLF TTEASIDALK QTILQLAVMG KLVsQDPNDE
A.PEP	NVE	EELAE	NWA RISsHFDTLF TTEASVDALK QTILQLAVMG KLVpQDPNDE
SB.PEPAOVDStK ARLeQIPoIL KRFRQSViVA
	351		400
E.PEP	PA	SELLKRVs	QEKvQLVKEG KIKKQKPLPP VSDDEKPFEL PiGWEHCRIg
A.PEP	PA	SELLKRiA	QEKaQLVKEG KIKKQKPLPP ISDEEKPFEL PEGWEHCRLG
SB.PEP	AVNG	OLTKEL hKKN	KFKLTe LhNISiPSLWK iSEiGOFADV KGGKRLPKG
	401		450
E.PEP	EII	ANMDAGW sPAC	sPSPs NEDiNGVlKT TAVOSLEyRE QENKTLpNSK
A.PEP	SI	VNLNGY. .	APksEwFT sVGLRLLRNA NiARCVTNwK DVVHiPNDMi
SB.PEP	SLIA	ENTG FPYiRAGOLK NGTVLPeGOL YLEeYiOKSi
	451		500
E.PEP	LPR	QYEVHD gDIL	VTRAGP KNRVGVsC.L VEkTRsKLMi sDKiTRFHLi
A.PEP	SDF	ENYLISE nDIV	ISLDRP iINTGLKYAI ISKsDLPCLi LORVAKFNY
SB.PEP	SRY	TvSSGD	L YITiVgACiG DA...GiIPD VYNNANLTEN AAKiCNLNE
	501		550
E.PEP	SDD	ISARyIS LCL	NRGVtAD YLEAsKsGMA ESOMNiSQEN LRSAPiALPP
A.PEP	ANT	VSNsFLT iWLO	sYF... FiNSiDPGRS NGVPHiSTKQ LeMTLFpLLP
SB.PEP	iFR	FLSLWl RssY	LQDiIN SEIKsGAOGK LALARiKSLP LiLpLQEOH
	551		600
E.PEP	TA	IQLVIST iED	FRRVCDQ LKsRLOSaOQ TQLHLADALT DAALN.....
A.PEP	QSE	QDRiISK nDEL	LiOTCNK LKYiIKtAKQ TQLHLADALT DAAIN.....
SB.PEP	Ei	VRRVQLP AYAD	TIEKOV NNALTRVNSL TQSiLAKAFR GELTAQWRAE
	601		635
E.PEP
A.PEP
SB.PEP	NP	ELISGENS	AAALLEKIRA ERAASGGKRT SRKKA

Figure 3. Comparison of S Polypeptides

An optimal alignment of the predicted amino acid sequences of the S polypeptides of EcoA, EcoE, and StySB was generated by use of the programs of the University of Wisconsin Genetics Computer Group (Devereux et al., 1984). Gaps (insertions/deletions) were allowed, to optimize the alignment. Bold characters represent conserved residues: identities and conservative substitutions are indicated, and both are used to calculate similarity. The conservative substitutions allowed were I/L, I/V, L/M, E/D, F/Y, and K/R, as recommended by Collins and Coulson (1987). The amino variable regions of the three polypeptides occur between positions 101 and 247 in the figure. Conservation of the proline, glutamate, and glutamine residues occurs at the junction of the amino variable and central conserved regions in all the S polypeptides analyzed, and these residues are not included in the comparisons of the amino recognition domains. EcoA and EcoE show 80% identity (84% similarity) within this region, EcoA and StyB, and EcoE and StySB show 44% identity (54% similarity). By contrast, the carboxyl variable regions of the polypeptides (positions 400-579) EcoA and E show 14% identity (26% similarity); EcoA and StySB show 5% identity (11% similarity); and EcoE and StySB show 7% identity (11% similarity).

The sequence for StySB has been published (Gann et al., 1987).

N-terminal one specifies the trinucleotide component, it would now appear inevitable that the tetra- or pentanucleotide component is recognized by the carboxyl variable region (see Figure 1; Fuller-Pace and Murray, 1986). The A family of enzymes, whose overall organization and function are in many ways identical to those of K, have been judged unrelated by genetic and molecular criteria (Murray et al., 1982). A comparison of either the nucleotide or predicted amino acid sequences corroborates this sharp distinction in showing no general sequence similarity between the families, even in regions that are conserved within one family. This is seen for the conserved regions of S (see Figure 3), and is true for the other subunits of the complex (unpublished data). Nevertheless, the S polypeptides of the A family, like those of K, contain two variable domains of ~150 amino acids, and our data now implicate the N-terminal one in the specification of GAG. Details of the organization of the S genes and polypeptides of the A family will be reported elsewhere (Cowan and Kannan, unpublished data).

Both the size and general variability of recognition domains within one family of otherwise well conserved polypeptides contrast with the similarity found between those specifying recognition of the same trinucleotide, even when present in otherwise divergent systems. Though the nucleotide sequences recognized are short (3, 4, or 5 bp), the amino acid sequences involved in recognition are extensive, suggesting a more complex interaction than a linear segment of amino acids with a linear sequence of bases. The structure of the type II restriction enzyme, EcoRI (McClarin et al., 1986), shows that adjacent bases are contacted by residues well separated in the amino acid sequence. Also, the importance of the precise presentation of amino acids is emphasized by the finding that arginine can interact with both AT and GC base pairs. DNA recognition can therefore involve extensive regions of polypeptide, particularly when the protein acts on a specific nucleotide sequence, rather than merely binding to it. Type I R-M systems are functionally more complex than type II restriction enzymes. A single enzyme consisting of three subunits, R, M, and S, acts as both a DNA methylase and an endonuclease. Modification of the target sequence specified by S involves methylation of one adenine within each defined component of that target sequence. In turn, the methylation state of the sequence dictates the bound enzyme's subsequent behavior, by acting as an allosteric effector. When the sequence is fully methylated, the enzyme dissociates from the DNA. When hemimethylated, the complex methylates the complementary strand. Only when completely unmethylated is the DNA cut, following an ATP-dependent translocation of the DNA through the bound enzyme (for review, see Bickle, 1987). It may be only in the light of such complexity that the extensive nature of the recognition domains will be understood.

We have no direct evidence to implicate all the residues within a recognition domain in defining specificity, although when two from the same family specify recognition of different trinucleotides, it is hard to detect any similarity between them. In contrast, EcoK and StySP of the K family both recognize 5' AAC and show 90% identity throughout

their amino recognition domains (Fuller-Pace and Murray, 1986), and, as we have shown here, EcoA and E show 80% identity throughout theirs. The much lower identity (44%) seen between those of StySB and either EcoA or E may reveal a more strict definition of the minimum recognition domain. However, the M subunits that these S polypeptides direct to methylate GAG are quite different. Thus, the amino recognition domains of StySB and EcoA, while having identical specificities, may vary slightly to accommodate their different M subunits. It therefore is possible that, within the context of each enzyme, most residues in a recognition domain are important in defining specificity.

The relationships between these different type I R-M systems present an interesting evolutionary puzzle. Within a family the members must have originated from a common ancestor with most of the subsequent divergence being associated with the development of new specificities, the pressure for which clearly exists (Levin, 1986). The considerable similarities between the different families, including the apparent allelism of their genes, would perhaps favor both having originated from the same common ancestor. Why such great divergence has occurred between the two groups of enzymes, while members within each have remained so conserved, is not clear. Nevertheless, evidence already suggests further diversification. A third family has been described in *E. coli* (Price et al., 1987), and others may exist in *Salmonella* (Bullas et al., 1980).

Experimental Procedures

Bacteria, Plasmids, and Phage

The chromosomal genes encoding the type I R-M systems of the K and A families are designated *hsdR*, *M*, and *S*. All three genes are needed to encode the restriction endonuclease and confer the r^+m^+ phenotype, but *hsdM* and *S* are sufficient for modification (r^+m^-). The bacterial strains used in this paper are derived from *E. coli* K-12 and, other than the mutL derivative of BMH71-18 (Kramer et al., 1984) used to recover the products of site directed mutagenesis, they lack part or all of the original *hsd* genes.

The general host for λ phages, plasmids, and M13 was NM522, (*lac-pro*) Δ *hsdMS* Δ *F' lacZ* M15 *lacI^q* (Gough and Murray, 1983). This r^+m^- strain retains a functional *hsdR* gene and, when lysogenic for a λ *hsdMS* phage that includes the *M* and *S* genes from a member of the K family, complementation results in a restricting strain (r^+m^+), whose specificity is determined by the *S* gene within the prophage. Virulent phage infecting such a restricting strain plate with low efficiency if their restriction targets are not methylated by the cognate modification enzyme. To provide an E-restricting strain sensitive to both λ and M13 phages, NM522 was transformed with pGC1, a plasmid encoding the complete EcoE system (Fuller-Pace et al., 1985). NM551 has the *hsd* genes of StySQ substituted for those of *E. coli* K-12 and is $r_{SO}^+m_{SO}^+$ (Fuller-Pace et al., 1984).

The genes encoding all the type I R-M systems discussed were cloned in λ vectors and have been subcloned in plasmids and M13. pGC1 and pFFP19 (Fuller-Pace et al., 1985) were used as a source of the *S* genes of EcoE and EcoA, respectively; fragments of these were cloned in M13 vectors for sequence determination. λ *hsdMS* StySP and StySQ (Fuller-Pace et al., 1984) were the source of their respective *S* genes. The 5.1kb BamHI fragment including the *S* gene of StySQ (Fuller-Pace and Murray, 1986) was cloned in pAG11 to produce pAG12. pAG11 is a derivative of pEMBL8⁺ (Dente et al., 1983), in which the Accl site in the polylinker cloning sequence has been destroyed. This was done by cutting pEMBL8⁺ with Accl, filling in the resulting cohesive ends using Klenow polymerase and all four de-

oxynucleotides (Maniatis et al., 1982), and ligating the blunt ends produced. The *lacZ* reading frame is disrupted, resulting in LacZ⁻ transformants of NM522. Analysis of plasmid DNA, by digestion with restriction enzymes and by nucleotide sequencing, identified an isolate in which the Accl site had been filled in, and no other change had occurred. The removal of this Accl site was essential to allow the replacement of the internal Accl fragment in pAG12 with one from the M13 derivative following site-directed mutagenesis (see Figure 2). The flanking BamHI sites were used to excise the entire restructured StySQ⁺ *S* gene from the plasmid, and this fragment was substituted for its equivalent in λ *hsdMS* StySP imm²¹ nin (Fuller-Pace et al., 1984), thereby changing the specificity of the encoded modification enzyme.

Phage P3 (Bullas et al., 1976) was used to detect the StySQ and SQ⁺ restriction systems, since it itself is restricted very poorly by these systems.

The heteroimmune helper phage λ imm⁴³⁴ was used to integrate λ *hsdMS* phages into the chromosome of the *hsdMS* Δ host and standard λ testers: λ imm²¹cl, λ imm²⁴cl, and λ vir, were required to check for lysogens. The latter was also used to confirm the $r_{E}^+m_{E}^+$ phenotype of NM522 transformed with pGC1.

Enzymes and Chemicals

DNA polymerase (Klenow fragment), T4 DNA ligase, and deoxy- and dideoxynucleotide triphosphates were from Boehringer Mannheim; restriction enzymes were from Boehringer Mannheim and NBL Enzymes; T4 polynucleotide kinase was a generous gift from S. Bruce (University of Edinburgh); M13 universal primer was from New England Biolabs Inc., and other oligonucleotide primers from Oswel DNA Service (Edinburgh). Adenosine 5'-[α -³⁵S]thiotriphosphate (15.2 TBq mmol⁻¹) was from New England Nuclear, and deoxyadenosine [γ -³²P] triphosphate (110 TBq mmol⁻¹) was from Amersham International. IPTG (iso-propyl- β -D-thiogalactopyranoside) and Xgal (5-bromo-4-chloro-3-indolyl- β -D-galactoside) were from United States Biochemical Corporation.

Media and Microbial Methods

Media and general methods (Murray et al., 1977) and tests for estimating restriction and modification have been described (Fuller-Pace et al., 1985).

Preparation, Manipulation, and Recovery of DNA

The methods were those described by Midgley and Murray (1985).

Site-Directed Mutagenesis

The *S* gene of StySQ⁺ was made in two steps. First, a portion of the *S* gene of StySP was cloned as an ~840 bp BamHI-PstI fragment in mp18. This was used as a template for mutagenesis (Zoller and Smith, 1983) directed by a 45 base oligonucleotide in which four base changes were introduced into the first half of the central conserved region of the *S* gene. The template, oligonucleotide sequence, and changes produced are shown in Figure 2. Phage were recovered in the mutL strain (Kramer et al., 1984), and those containing all four mutations were identified by differential hybridization and DNA sequencing. An ~250 bp Accl fragment (see Figure 2) containing all four changes was excised and used to replace the equivalent Accl fragment in the StySQ *S* gene in pAG12 (see above). A derivative containing all four mutations, and thus the *S* gene of StySQ⁺ (identified by cleavage with RsaI, since one of the changes creates a new RsaI site, and DNA sequencing) was designated pAG13. To enable identification of its specificity, the StySQ⁺ *S* gene was transferred into a λ phage, as described above. Phage carrying StySQ⁺ were identified by differential hybridization using as a probe a recombinant M13 containing a fragment specific to the distal variable region of StySB (and thus StySQ⁺; see Figure 1). The presence on the phage of all four mutations was confirmed by subcloning and sequencing the appropriate region, and the specificity conferred by the StySQ⁺ *S* gene determined as described below.

The Specificity of StySQ⁺

λ *hsdSQ⁺* contains the newly constructed *S* gene and a complementary *M* gene (see above). As this phage is att⁺, a λ imm⁴³⁴ att⁺ helper phage was used to construct a lysogen in the *hsdMS* Δ strain, NM522. This strain provides R polypeptides with which the phage-encoded M

and S polypeptides can interact, thereby producing a host that modifies and restricts with the specificity of StySQ. Phage P3 showed a decreased plating efficiency on the dilysogen, due to restriction. However, it plated with an efficiency of ~1 on this strain if it had previously been propagated on, and hence modified by, a strain encoding the StySQ system (NM551). Likewise, P3 modified by growth on the dilysogen was not restricted by NM551.

DNA Sequencing

Sonicated fragments of *hsdE* DNA were ligated to SmaI cut M13mp19, and restriction fragments of *hsdA* DNA were ligated to the mp18 and mp19 M13 vectors. Template DNA was sequenced by the dideoxy chain termination method (Sanger et al., 1977) using deoxyadenosine 5'-[α - 32 S] thiotriphosphate, and reactions were analyzed on buffer gradient gels (Biggin et al., 1983). Sequences of sonicated fragments were generated using universal primer; other fragments were sequenced using oligonucleotide primers. The sequences were compiled by the computer programs of Staden (1982), and analyzed using the programs of Devereux et al. (1984).

Computing Methods

Searches for possible recognition sequences utilized a program that generates, compiles, and sorts ordered lists of subsequences to identify those present in restricted substrates but absent from unrestricted sequences (J. Crook, personal communication).

Acknowledgments

We thank H. D. Braymer, K. Chapman, D. J. Finnegan, K. Kaiser, K. Murray, J. G. Scaife, and our colleagues for constructive criticism of the manuscript, the OSWEL DNA Service and Wellcome Trust for synthetic oligonucleotides, Fiona Govan for typing, and Annie Wilson for figures. This work was supported by the Medical Research Council and by Science and Engineering Research Council studentships to G. M. C. and A. A. F. G.

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

Received August 19, 1988; revised October 12, 1988.

References

Arber, W., and Kühnlein, U. (1967). Mutationeller Verlust B-spezifischer Restriktion des Bakteriophagen ϕ d. *Pathol. Microbiol.* **80**, 946-952.

Argos, P. (1985). Evidence for a repeating domain in the type I restriction enzymes. *EMBO J.* **4**, 1351-1354.

Bickle, T. A. (1987). Restriction and Modification Systems. In *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger, eds. (Washington, D. C.: American Society for Microbiology), pp. 692-696.

Biggin, M. D., Gibson, T. J., and Hong, G. F. (1983). Buffer gradient gels and 35 S label as an aid to rapid DNA sequence determination. *Proc. Natl. Acad. Sci. USA* **80**, 3963-3965.

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.

Bullas, L. R., and Colson, C. (1975). DNA restriction and modification systems in *Salmonella*. III. SP, a *Salmonella potsdam* system allelic to the SB system in *Salmonella typhimurium*. *Mol. Gen. Genet.* **139**, 177-188.

Bullas, L. R., Colson, C., and Van Pel, A. (1976). DNA restriction and modification systems in *Salmonella*. SQ, a new system derived by recombination between the SB system in *Salmonella typhimurium* and the SP system in *S. potsdam*. *J. Gen. Microbiol.* **95**, 166-172.

Bullas, L. R., Colson, C., and Neufeld, B. (1980). Deoxyribonucleic acid restriction and modification systems in *Salmonella*: chromosomally located systems of different serotypes. *J. Bacteriol.* **141**, 275-292.

Collins, J. F., and Coulson, A. F. W. (1987). Molecular sequence com-

parison and alignment. In *Nucleic Acids and Protein Sequence Analysis: A Practical Approach*, M. J. Bishop and C. F. Rawlings, eds. (Oxford: I. R. L. Press), pp. 323-358.

Dente, L., Cesareni, G., and Cortese, R. (1983). pEMBL: a new family of single stranded plasmids. *Nucl. Acids Res.* **11**, 1645-1655.

Devereux, J., Haeblerli, P., and Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucl. Acids Res.* **12**, 387-395.

Franklin, N. C., and Dove, W. F. (1969). Genetic evidence for restriction targets in the DNA of phages lambda and Φ 80. *Genet. Res. Camb.* **14**, 151-157.

Fuller-Pace, F. V., and Murray, N. E. (1986). Two DNA recognition domains of the specificity polypeptides of a family of type I restriction enzymes. *Proc. Natl. Acad. Sci. USA* **83**, 9368-9372.

Fuller-Pace, F. V., Bullas, L. R., Delius, H., and Murray, N. E. (1984). Genetic recombination can generate altered restriction specificity. *Proc. Natl. Acad. Sci. USA* **81**, 6095-6099.

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 *E. coli*. *J. Mol. Biol.* **185**, 65-75.

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.

Glover, S. W., and Colson, C. (1969). Genetics of host-controlled restriction and modification in *Escherichia coli*. *Genet. Res. Camb.* **13**, 227-240.

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**, 1-19.

Kramer, W., Drutsa, V., Jansen, H.-W., Kramer, B., Pflugfelder, M., and Fritz, H.-J. (1984). The gapped duplex DNA approach to oligonucleotide-directed mutation construction. *Nucl. Acids Res.* **12**, 9441-9456.

Levin, B. R. (1986). Restriction-modification immunity and the maintenance of genetic diversity in bacterial populations. In *Evolutionary Processes and Theory*, S. Karlin and E. Nero, eds. (New York: Academic Press), pp. 669-688.

Loenen, W. A. M., Daniel, A. S., Braymer, H. D., and Murray, N. E. (1987). Organization and sequence of the *hsd* genes of *Escherichia coli* K-12. *J. Mol. Biol.* **198**, 159-170.

Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982). *Molecular Cloning: A Laboratory Manual*. (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory).

McClarin, J. A., Frederick, C. A., Wang, B.-C., Greene, P., Boyer, H. W., Grable, J., and Rosenberg, J. M. (1986). Structure of the DNA-EcoRI endonuclease recognition complex at 3A resolution. *Science* **234**, 1526-1541.

Midgley, C. A., and Murray, N. E. (1985). T4 polynucleotide kinase; cloning of the gene (*psfT*) and amplification of its product. *EMBO J.* **4**, 2695-2703.

Murray, N. E., Brammar, W. J., and Murray, K. (1977). Lambdaoid phages that simplify the recovery of in vitro recombinants. *Mol. Gen. Genet.* **150**, 53-61.

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.

Nagaraja, V., Shepherd, J. C. W., and Bickle, T. A. (1985a). A hybrid recognition sequence in a recombinant restriction enzyme and the evolution of DNA sequence specificity. *Nature* **316**, 371-372.

Nagaraja, V., Shepherd, J. C. W., Pripfl, T., and Bickle, T. A. (1985b). Two type I restriction enzymes from *Salmonella* species. Purification, and DNA recognition sequences. *J. Mol. Biol.* **182**, 579-587.

Price, C., Pripfl, T., and Bickle, T. A. (1987). EcoR124 and EcoR124/3: the first members of a new family of type I restriction and modification systems. *Eur. J. Biochem.* **167**, 111-115.

Pugh, J. C., Weber, C., Houston, H., and Murray, K. (1986). Expression of the X gene of hepatitis B virus. *J. Med. Virol.* **20**, 229-246.

Robinson, A. C., Kenan, D. J., Hatfull, G. F., Sullivan, N. F., Spiegel-

- berg, R., and Donachie, W. D. (1984). DNA sequence and transcriptional organisation of essential cell division genes *ftsQ* and *ftsA* of *E. coli*. *J. Bact.* **160**, 546-555.
- Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**, 5463-5467.
- Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F., and Peterson, G. B. (1982). Nucleotide sequence of bacteriophage λ DNA. *J. Mol. Biol.* **162**, 729-773.
- Staden, R. (1982). Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. *Nucl. Acids Res.* **10**, 4731-4751.
- Suri, B., and Bickle, T. A. (1985). EcoA: the first member of a new family of type I restriction and modification systems. *J. Mol. Biol.* **186**, 77-85.
- Suri, B., Shepherd, J. C. W., and Bickle, T. A. (1984). The EcoA restriction and modification system of *Escherichia coli* 15T⁻: enzyme structure and DNA recognition sequence. *EMBO J.* **3**, 575-579.
- Sutcliffe, J. G. (1978). Complete nucleotide sequence of the *Escherichia coli* plasmid pBR322. *Cold Spring Harbor Symp. Quant. Biol.* **43**, 77-79.
- van Wezenbeck, P. M. G. F., Hulsebos, T. J. M., and Schoenmakers, J. G. G. (1980). Nucleotide sequence of the filamentous bacteriophage M13 genome: comparison with phage fd. *Gene* **11**, 129-148.
- Zoller, M. J., and Smith, M. (1983). Oligonucleotide directed mutagenesis of DNA fragments cloned into M13-derived vectors. *Meth. Enzymol.* **100**, 468-500.