Four new type I restriction enzymes identified in *Escherichia coli* clinical isolates

Julie K. A. Kasarjian, Yoshiaki Kodama, Masatake Iida, Katsura Matsuda[1] and Junichi Ryu*

Division of Microbiology and Molecular Genetics, Department of Biochemistry and Microbiology and [1]Department of Earth and Biological Sciences, Loma Linda University, Loma Linda, CA 92350, USA

ABSTRACT

Using a plasmid transformation method and the RM search computer program, four type I restriction enzymes with new recognition sites and two isoschizomers (EcoBI and Eco377I) were identified in a collection of clinical *Escherichia coli* isolates. These new enzymes were designated Eco394I, Eco826I, Eco851I and Eco912I. Their recognition sequences were determined to be GAC(5N)RTAAY, GCA(6N)-CTGA, GTCA(6N)TGAY and CAC(5N)TGGC, respectively. A methylation sensitivity assay, using various synthetic oligonucleotides, was used to identify the adenines that prevent cleavage when methylated (underlined). These results suggest that type I enzymes are abundant in *E.coli* and many other bacteria, as has been inferred from bacterial genome sequencing projects.

INTRODUCTION

Restriction endonucleases recognize specific DNA sequences comprising various combinations of 4–8 nt. A total of 280 distinctive recognition sequences have been found so far (1). Most of the known recognition sequences belong to type II enzymes, although many more type I, type III and type IV enzymes are predicted from current genome data. In the case of type I enzymes, only 35 unique recognition sequences are known. Among those, 27 sequences have been found in nature and the remaining 8 were constructed artificially by phage transduction or *in vitro* gene manipulations (2,3).

The recognition sequences for type II enzymes are relatively easy to identify using crude extracts to digest fixed DNA sequences to produce distinct DNA bands. DNA recognition sequences can then be predicted using a computer program, such as REBpredictor (1). However, no simple method has been identified for finding type I recognition sequences mainly because the enzymes produce DNA fragments with random sequences (4). New methods are essential to predict the recognition sequences of type I enzymes. Previously, we developed a plasmid transformation method (5) and the computer program, RM search for this purpose (6). We have shown the effectiveness of this plasmid transformation method by detecting not only type I and type II enzymes but also type III enzymes (5). Using this new method, we have found two previously unknown recognition sequences from *Klebsiella*: [KpnAI (7) and KpnBI (8)] and from *Salmonella*: StySEAI, StySENI and StySGI (7). Furthermore, this plasmid transformation method was used to screen clinical *Escherichia coli* strains, and four strains, each of which contains a new prototype recognition sequence, were identified (5).

In this paper, we combined the plasmid R-M test with several new strategies to identify four additional new type I restriction enzymes. This supports the notion, predicted from the genome data, that type I enzymes are widespread in bacteria and many with unique recognition sequences can be identified using this plasmid transformation method.

MATERIALS AND METHODS

Bacterial strains and media

All bacterial strains used in this study are clinical isolates of *E.coli*. These clinical strains were obtained from Loma Linda University Medical Center during the period between 2000 and 2004. *E.coli* C strains and DH5α were used as restriction-minus controls. L-agar and L-agar with ampicillin (200 µg/ml) were used for bacterial growth medium and selection medium, respectively.

Plasmid series and plasmid transformation method

Two plasmid series were previously developed. The plasmid pL series contain various DNA from phage lambda (5) and the

pE series contain various fragments from *E.coli* K-12 strain W3110 (7). To obtain a nested series of deletions from plasmid pL34, the plasmid was linearized using EcoRV. The resulting DNA was progressively shortened by about five bases using ExoIII/S1 kit (Fermentas, Ltd). Plasmid DNA was separated and visualized using a 0.8% agarose gel.

A standard cold calcium-heat shock method (5) was used for plasmid transformation and plasmid pMECA (9) was used as the negative control. This plasmid is a pUC derivative and does not contain any restriction sites for any known type I enzymes. The results of the plasmid transformation data [plasmid R-M (restriction-modification) tests] are expressed as units of relative efficiency of transformation (EOT) (7).

Oligonucleotide synthesis and cloning

Oligonucleotides containing each predicted recognition sequence were synthesized at our University core facility in the Center for Molecular Biology and Gene Therapy, Loma Linda University. These oligonucleotides were cloned in pMECA, as described previously (5,7) and used to confirm the recognition sites and positions of the methylated adenine in the recognition sequences. Plasmids containing the synthetic

oligonucleotide with the predicted type I recognition sequences are referred to as 'reference plasmids'. Plasmid subcloning was carried out as described previously (5).

RESULTS AND DISCUSSION

Efficiency of transformation assay using *E.coli* clinical strains

Approximately 500 ampicillin-sensitive clinical isolates of *E.coli* were tested for transformation capability using pMECA. Among those, six strains that showed high frequencies of transformation when compared with the control strain *E.coli* C were analyzed further.

The initial EOT values obtained from pL1 to pL6 plasmids clearly indicated that all six strains have strong restriction activity (EOT < 10^{-1}). Thus, additional EOT analyses were performed using both the pL series and pE series plasmids (Table 1). When plasmids showed EOT values <0.1, they were classified as positive (+), whereas plasmids with EOT values >0.5 were categorized as negative (−) (5,7). These data were analyzed using the RM search program (6). The results

Table 1. Plasmid restriction tests for clinical *E.coli* strain

| Plasmids | Strains | | | | | |
	EC37	EC394	EC826	EC851	EC912	EC1265
pMECA	−	−	−	−	−	−
PL1	+	+	+	+	+	+
PL2	+	+	+	+	+	+
PL3	+	+	−	−	−	+
PL4	+	−	−	−	+	+
PL6	+	−	+	+	−	+
PE2	−	−	−	−	+	−
PE3	+	−	+	−	−	−
PE4	−	+	+	−	−	−
PE5	+	−	+	−	+	+
PE6	+	+	+	+	−	−
PE8	−	+	+	+	+	+
PE9	−	−	−	−	−	−
PE10	−	−	−	−	−	−
PE11	−	−	+	+	+	+
PE12	−	−	−	+	−	−
PE14	+	+	+	−	−	+
PE15	−	−	−	−	−	+
PE16	−	−	−	−	+	−
PE17	+	−	−	−	−	−
PE18	−	+	+	+	−	−
PE19	−	−	+	+	+	−
PE22	+	−	−	−	+	+
PE23	+	−	+	−	−	−
PE24	−	−	−	+	−	−
PE26	+	−	+	−	−	+
PE28	−	−	−	−	−	−
PE29	+	−	+	+	+	−
PE31	−	−	−	−	+	−
PE32	+	+	+	−	+	+
PE33	−	−	+	−	−	+
PE38	−	−	−	−	+	−
PE41	+	−	+	−	+	−
PE44	−	−	−	−	−	−
PE45	+	−	+	−	+	+
Recognition sequence	GGA(8N)ATGC	GAC(5N)RTAAY	GCA(6N)CTGA	GTCA(6N)TGAY	CAC(5N)TGGC	TGA(8N)TGCT
Type (prototype)	I (Eco377I)	I (New)	I (New)	I (New)	I (New)	I (EcoBI)

The presence (+) or absence (−) of the recognition sequence in each plasmid are shown.

predicted that the recognition sequences for the five restriction enzymes, Eco37I, Eco826I, Eco851I, Eco912I and Eco1265I are GGA(8N)ATGC, GCA(6N)CTGA, GTCA(6N)TGAY, CAC(5N)TGGC and TGA(8N)TGCT, respectively (Table 1). The Eco37I enzyme is an isoschizomer of Eco377I, which was discovered earlier in our clinical *E.coli* collection (5), whereas Eco1265I is an isoschizomer of EcoBI. The remaining four enzymes are all prototypes. Their non-palindromic recognition sequences seem to be typical bipartite type I recognition sequences.

Recognition sequence determination for Eco394I

Although the RM search program produced a single plausible recognition sequence from the data in Table 1 for five other enzymes, analysis of the Eco394I recognition sequence was more complicated. Initially, <9 positive plasmid sequences for Eco394I were identified using our standard set of plasmids, compared with >10 positive sequences for the other enzymes (Table 1). Further testing using additional plasmids identified two more positive plasmids (pL5 and pE25) but RM search still produced a larger number of candidate sequences.

Since short positive plasmids are most helpful in identifying recognition sequences, we focused on plasmid pL2 (16 kb BamHI fragment of phage lambda) and made a series of subclones. First, this fragment was cut into two segments by SphI and BamHI. The subclone that included the recognition sequence was named pL9 (Figure 1). The plasmid pL9

was further subcloned into pL31 to pL35 by PstI cleavage. The plasmid R-M tests revealed two Eco397I sites on pL9, which were designated as site A (on pL32) and site B (on pL34). Further subclones were derived from those two positive clones and named pL49, pL50, pL51 (from pL31) and pL54, pL55, pL57 (from pL34). Further R-M tests narrowed the region down to 535 bp between a HincII and MluI site in pL50 (site A) and 150 bases between EcoRV and KpnI site on pL54 (site B) (Figure 1).

When these two positive DNA fragments were added to the RM search program, all the simple candidate sequences were eliminated. However, there were still numerous candidates containing a degenerate base in the 3′ component region. To identify a smaller positive fragment, a series of ExoIII/S1 nested deletion was constructed from pL34 (B site). The plasmid restriction test was performed on ∼30 deletion plasmids and 5 of these clones were sequenced. This revealed a critical 5 bp sequence (CGTCC) between subclones pL34-96 (positive plasmid containing the entire site) and pL34-97 (negative plasmid with an incomplete site) (Figure 1). These five bases contain one end of the recognition sequence. Surprisingly, when these data were put into the RM search, several possible degenerate sequences still remained. To further minimizing possibilities, we constructed an 18mer oligonucleotide sequence, with three bases overlapping the five critical base pair region, which was cloned into pMECA (Figure 1). Plasmid restriction tests showed that this plasmid, pC2, did not contain the complete recognition sequence, therefore,

Figure 1. Plasmid R-M tests for Eco394I on a series of subclones identifying two Eco394I sites (**A** and **B**) on plasmid pL9. Positive plasmids (with Eco394I site) are darkened, whereas negative plasmids (no site) are hatched. Plasmid pL9 is a subclone of pL2 (Table 1). Plasmids pL31 to pL35 are PstI subclones of pL9. Plasmids pL49, pL50 and pL51 are derived from pL32, whereas plasmids pL54, pL55 and pL57 are derived from pL34. Two plasmids pL34-96 and pL34-97 were obtained from an ExoIII/S1 nested deletion experiment, whereas the pC2 is a plasmid containing an 18 bp oligonucleotide cloned into the pMECA EcoRV site. Only relevant DNA sequences are shown. Note that the Eco394I sequence described here at B site is a complementary sequence of the consensus sequence, GAC(5N)RTAAY described in Table 2.

narrowing one end of the recognition sequence to the remaining two bases (CC) (Figure 1). After this result, the recognition sequence of the one remaining candidate was GACNNNNNRTAAY. Table 2 shows the location of this plausible Eco394I sequence on various positive plasmids. This sequence represents a new pattern for a type I sequence, with a central five random sequence and two degenerate bases, R and Y, in the 3′ component. In the process of narrowing down the location of this sequence, one additional new site (pL2C) for Eco394I was discovered just next to the B site described in Figure 1. The predicted sequences were confirmed by constructing a series of oligonucleotide sequences described below.

Type I sequences with more than one degeneracy and previously unidentified 'patterns' (see Table 5), such as Eco394I, are much more difficult to identify. We found that our standard sets of plasmids (pL and pE series) are not sufficient to identify those complicated sequences. An additional new set of multiple overlapping clones, such as those used in shotgun cloning and sequencing experiments, would be ideal candidates to supplement our plasmid series.

Confirmation of the predicted recognition sequence and methylation site

The predicted recognition sequences described above were confirmed by synthesizing a series of oligonucleotides (Table 3, column 3). These oligos were cloned into pMECA at the SspI or EcoRV site in the MCS region. These plasmids were designated pEco394AC, pEco394AT, etc. (Table 3, column 1). Plasmid R-M tests were performed by transforming each newly constructed plasmid into the original host strains that produce the corresponding restriction enzymes. An example of the restriction test results using the pEco394I series plasmids are shown in Table 4. All plasmids showed strong restriction activity (EOT values at 10^{-3} level). Modified (methylated) plasmid DNA was recovered and subjected to plasmid modification tests. All the modified plasmids were methylated and protected from restriction (EOT values close to 1.0). Similar results were also obtained for Eco826I, Eco851I and Eco921I.

For the suspected isoschizomers (Eco37I and Eco1265I), plasmids with each recognition sequence constructed previously (reference plasmids) (5) were used. In summary, all the sequences predicted from the RM search program were confirmed by plasmid R-M tests using the corresponding reference plasmids.

Table 2. Eco394I sequences and surrounding sequences in positive plasmids

Plasmid sites	Eco394I recognition sequences
pL1	GCCTGTTGCA **GAC** GGGCG **ATAAT** GCCGTTGTAA
pL2A	CCCGTATACA **GAC** AACGG **ATAAC** GGTAAGCAGA
pL2B	CGAACTCCGG **GAC** GCTCA **GTAAT** GTGACGATAG
pL2C	AAGTCGAGCT **GAC** GGAGG **ATAAC** GCCAGCAGAC
pL3	ATAGATGAAA **GAC** TTCAG **GTAAT** TGGAATTGAT
pL5	TCAGGCCACT **GAC** TAGCG **ATAAC** TTTCCCCACA
pE4	TCGAAGCCGC **GAC** GACAC **ATAAT** GCGCATCACC
pE6	CACCAAGGTT **GAC** GCCAA **ATAAC** CCGGTCAGAA
pE8A	ACGTTAATAA **GAC** AACCG **ATAAC** GCCTTCGTAA
pE8B	CGAAGTTAGC **GAC** CGCCC **ATAAT** GGCAATGTAA
pE14	AATTGATATG **GAC** GTCGC **GTAAC** GTTAATAACG
pE18	CAAATCGTTT **GAC** ATACT **GTAAC** TGGCGTTTGA
pE25	CGCACAACTG **GAC** TTTCG **ATAAC** GGCTATATGC
pE32A	GGAACAAGGT **GAC** AATCA **ATAAC** AAAATGCGGC
pE32B	CCGGTTGGGC **GAC** GGACA **GTAAC** GACCCGGACA
Consensus	NNNNNNNNNN **GAC** NNNNN **RTAAY** NNNNNNNNNN

Table 3. Design of oligonucleotides containing predicted recognition sequences as well as restriction sites for methylation sensitivity assays

R-M system plasmids constructed	Recognition sequence (degeneracy)	Designed oligos	Reasoning behind the methylation sensitivity assay	Predicted digestion of plasmid with type I sequence by the constructed corresponding type II enzyme	
				Before modification	After modification
Eco394I	5′GAC(5N)RTAAY3′ 3′CTG(5N)CATTR5′				
pEco394AC	(R-Y = A-C)	5′**GAC**GCGTCATAAC3′	MluI does not cut ACGCGT	Cut	Not cut
pEco394AT	(R-Y = A-T)	5′**GAC**TAGTCATAAT3′	SpeI does not cut ACTAGT	Cut	Not cut
pEco394GCr	(R-Y = G-C)	3′CTGA**TGCGC**ATTG5′	MluI does not cut ACGCGT	Cut	Not cut
pEco394GTr	(R-Y = G-T)	3′**CTG**CTGATCATTA5′	SpeI does not cut ACTAGT	Cut	Not cut
Eco826I	5′GCA(6N)CTGA3′ 3′CGT(6N)GACT5′				
pEco826		5′**GCA**GTACTCCTGA3′	ScaI does not cut AGTACT	Cut	Not cut (Fig.3B)
pEco826r		3′**CGT**CGTCATGACT5′	ScaI does not cut AGTACT	Cut	Not cut
Eco851I	5′GTCA(6N)TGAY3′ 3′CAGT(6N)ACTR5′				
pEco851T	(Y = T)	5′**GTCA**GTACTCTGAT3′	ScaI does not cut AGTACT	Cut	Not cut (Fig.3B)
pEco851Tr	(Y = T)	3′**CAGT**GTCATGACTA5′	ScaI does not cut AGTACT	Cut	Not cut
pEco851C	(Y = C)	5′**GTCA**GTACTCTGAC3′	ScaI does not cut AGTACT	Cut	Not cut
pEco851Cr	(Y = C)	3′**CAGT**GTCATGACTG5′	ScaI does not cut AGTACT	Cut	Not cut
Eco912I	5′CAC(5N)TGGC 3′GTG(5N)ACCG				
pEco912		5′**CAC**TAGTCTGGC3′	SpeI does not cut ACTAGT	Cut	Not cut (Fig.3B)
pEco912r		3′**GTG**TCATGACCG5′	ScaI does not cut AGTACT	Cut	Not cut

The type I recognition sequences are underlined in the designed oligo column. The type II recognition sites in the same column are shown in bold. Target adenines in each recognition sequence are bold and underlined. Methylated adenines in the 'Reasoning' column are underlined. In the cloning experiments, either an SspI site or EcoRV (blunt end) site was added to each designed oligo and ligated into the either unique pMECA SspI site or EcoRV site.

Table 4. Plasmid restriction tests of pEco394 series plasmids to confirm the recognition sequences

Plasmid[a]	Amp^R transformants			
	DH5α	EOT	EC394	EOT
pMECA	2000	1.0	2500	1.3
pEco394AC	1800	1.0	2	1×10^{-3}
pEco394AT	2500	1.0	8	3×10^{-3}
pEco394GC	2000	1.0	2	1×10^{-3}
pEco394GT	3000	1.0	20	7×10^{-3}

[a]Plasmids are described in Table 3.

To facilitate the process of identifying methylated adenines in each recognition sequence, suitable restriction sites were also designed into the oligos (Table 3). The rationale of adenine methylation using methylation sensitivity assay is also described (Table 3, column 4). According to our general experimental scheme, when the target adenine is methylated, the methylated plasmid DNA cannot be cut at the cloning site but will be cut only once at the corresponding restriction enzyme site (RE site such as ScaI) on a reference plasmid (Figure 2A). Alternatively, when the target adenine is not methylated, the plasmid will also be cut at the cloning site and produce two bands. Examples of these experiments are shown in Figure 2B. When unmodified pEco826, pEco851 and pEco912 are cut with ScaI (lanes 4, 6 and 8, respectively), they produce two bands, whereas after the modification all the plasmids produce only one band (lanes 5, 7 and 9, respectively). Thus, modification methylation protected the type I recognition sites from restriction. Table 3 summarizes the prediction of the restriction digestion experiment before and after modification (columns 5 and 6). As predicted, all the designed plasmids each with predicted recognition sequence were cut with the corresponding type II enzymes whereas all the modified plasmids were not. Therefore, we concluded that all the adenines predicted from the recognition sequences (column 2) are target adenines for methylation.

Type I recognition sequences and the inferred rules

Table 5 summarizes all prototype type I recognition sequences identified to date. In addition to 27 naturally occurring sequences, eight artificially constructed sequences are also included. Also included are the target adenines for methylation and 'base patterns' (combination of the number of bases in the 5′, random and 3′ components). Further, Table 5 summarizes the total length of the recognition sequences, the distance between two methylated adenines (A-T distance) and known type I family members.

When all 27 naturally occurring sequences were compared, the following characteristics emerged. Some of those characteristics were previously noted.

(i) All the type I recognition sequences are within 12 and 15 bases in length, most (85%) of them consist of either 13 or 14 bases (Figure 3).
(ii) Type I sequences consist of three components: 5′ (3 to 4 bp + 5 to 8 random sequence + 3 to 5 bp) 3′. The sequences in Table 1 can be categorized into several 'patterns', such as 3-8-4 or 3-5-5. A total of seven patterns and their frequency are shown in Figure 4. Most (85%) of

A

Type I recognition sequence
Cloning site (protected by methylation)
0.9kb
RE site
Reference Plasmid (2.9kb)

B
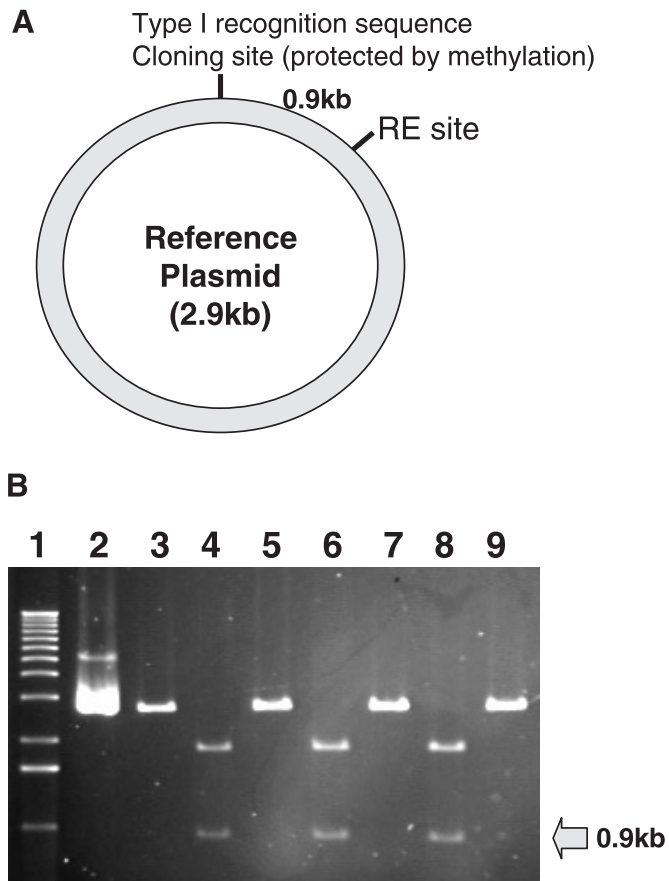
Figure 2. Methylation sensitivity assay to determine the methylated adenine in the target sequence. (**A**) General design of the 'reference' plasmid. An oligonucleotide containing a type I recognition sequence was cloned into the EcoRV site of the MCS region of pMECA. The 'reference' plasmids were modified in the corresponding bacteria. (**B**) Restriction digestions before and after methylation. All plasmids, except for lane 2, were digested with ScaI and subjected to 0.8% agarose gel electrophoresis. Lane 1, 1 kb ladder; lanes 2 and 3, pMECA control digested with EcoRV and ScaI, respectively; lanes 4 and 5, pEco826 unmodified and modified, respectively; lanes 6 and 7, pEco851, unmodified and modified, respectively; lanes 8 and 9, pEco912, unmodified and modified, respectively. Modified plasmids were digested only once and linearized, whereas unmodified plasmids were cut in two places resulting in two bands, one ∼0.9 kb (arrow).

the patterns are either 3-6-4, 3-7-4 or 3-8-4. We designated those patterns as 'core patterns'. The pattern for the Eco394I sequence (3-5-5) elucidated in this study was not included in the core patterns, making its identification more difficult. Another unusual pattern (3-5-4) and the shortest recognition sequence (12 bases) was also found in the present study (Eco921I). Only two sequences (KpnBI and Eco851I) contain 4 bp in the 5′ components. Two other unusual recognition sequences are also noteworthy. One is the NgoAV sequence (3-8-3. GCA-8N-TGC) (19), which contains a palindromic sequence that is extremely rare in type I enzymes. In our present study, a completely new 'prototype' sequence (4-8-4) was found in Eco851I. This sequence (GTCA-8N-TGAY) is close to the palindromic sequence except for the degeneracy, Y. These enzymes may be a bridge between the type I and type IIP restriction enzymes, which show a typical palindromic recognition sequence.

**Table 5.** Type I enzymes prototype recognition sequences

Type I enzyme	Recognition sequence[a]	Pattern (total bp)	A-T distance	Family[b]	Reference
Natural					
CfrAI	GCANNNNNNNNGTGG	3-8-4 (15)	9	B	(10)
Eco377I	GGANNNNNNNNATGC	3-8-4 (15)	9	–	(5)
Eco394I	GACNNNNNRTAAY	3-5-5 (13)	7	–	This work
Eco585I	GCCNNNNNNTGCG	3-6-4 (13)	–	–	(5)
Eco646I	CCANNNNNNNCTTC	3-7-4 (14)	8	–	(5)
Eco777I	GGANNNNNNTATC	3-6-4 (13)	6	–	(5)
Eco851I	GTCANNNNNNTGAY	4-6-4 (14)	6	–	This work
Eco826I	GCANNNNNNCTGA	3-6-4 (13)	7	–	This work
Eco912I	CACNNNNNTGGC	3-5-4 (12)	6	–	This work
EcoAI	GAGNNNNNNNGTCA	3-7-4 (14)	9	B	(11)
EcoBI	TGANNNNNNNNTGCT	3-8-4 (15)	8	A	(12)
EcoDI	TTANNNNNNNGTCY	3-7-4 (14)	8	A	(11;13)
EcoDXXI	TCANNNNNNNRTTC	3-7-4 (14)	8	C	(14)
EcoEI	GAGNNNNNNNATGC	3-7-4 (14)	9	B	(15)
EcoKI	AACNNNNNNGTGC	3-6-4 (13)	8	A	(16)
EcoR124I	GAANNNNNNRTCG	3-6-4 (13)	7	C	(17)
EcoR124II	GAANNNNNNNRTCG	3-7-4 (14)	8	C	(17)
EcoprrI	CCANNNNNNNRTGC	3-7-4 (14)	8	C	(18)
KpnAI	GAANNNNNNTGCC	3-6-4 (13)	6	D	(7)
KpnBI	CAAANNNNNNRTCA	4-6-4 (14)	7	E	(8)
NgoAV	GCANNNNNNNNTGC	3-8-3 (14)	8	C	(19)
StyLTIII	GAGNNNNNNRTAYG	3-6-4 (13)	8	A	(18;20)
StySEAI	ACANNNNNNTYCA	3-6-4 (13)	6	–	(7)
StySBLI	CGANNNNNNTACC	3-6-4 (13)	6	D	(21)
StySGI	TAANNNNNNRTCG	3-6-4 (13)	7	–	(7)
StySKI	CGATNNNNNNNGTTA	3-7-4 (14)	9	–	(7)
StySPI	AACNNNNNNGTRC	3-6-4 (13)	8	A	(18;20)
Artificial constructs					
EcoDR2	TCANNNNNNGTCG	3-6-4 (13)	7	C	(22)
EcoDR3	TCANNNNNNNATCG	3-7-4 (14)	8	C	(22)
EcoRD2	GAANNNNNNRTTC	3-6-4 (13)	7	C	(22)
EcoRD3	GAANNNNNNNRTTC	3-7-4 (14)	8	C	(22)
EcoDXXIΔ	TCANNNNNNNNTGA	3-8-3 (14)	8	C	(23)
Eco124IΔ	GAANNNNNNNTTC	3-7-3 (13)	7	C	(24)
StySJI	GAGNNNNNNNGTRC	3-6-4 (13)	8	A	(25)
StySQI	AACNNNNNNRTAYG	3-6-5 (14)	8	A	(26)

[a]The modified adenine is not experimentally verified for CfrAI, EcoEI and StySBLI, but each has only one candidate adenines in their recognition sequences.
[b]Family yet to be determined.
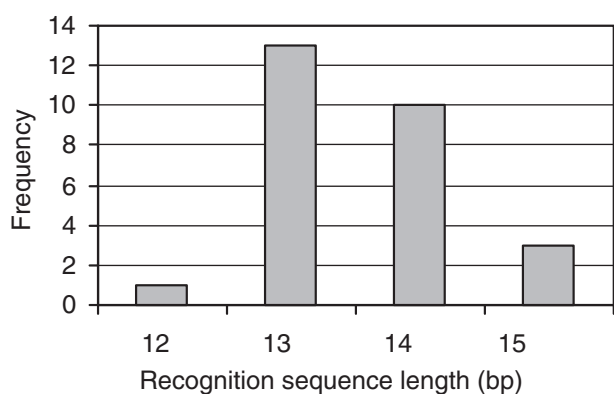
Figure 3. Frequencies of the length of known type I recognition sequences.
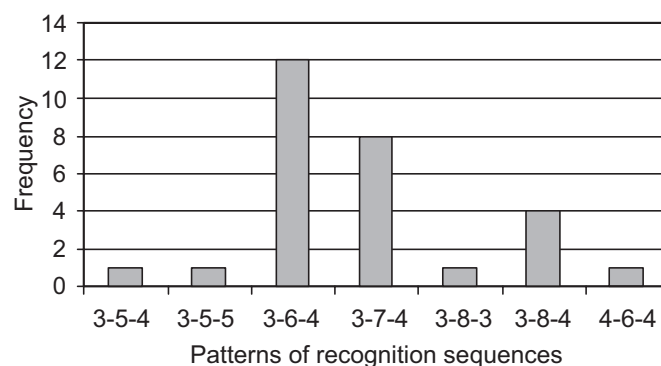
Figure 4. Frequencies of the patterns (number of bases in 5′, random and 3′ component) of type I recognition sequences.

(iii) About a half (48%) of the recognition sequences contain degenerate bases, either R or Y, or both. Only two sequences (Eco394I and StyLTIII) contain two degenerate bases (R and Y) in the sequence. All the degenerate bases are located in the 3′ component of the sequence. There is a strong tendency that R degeneracy is located close to the center random sequence, whereas the Y degeneracy usually is located at the end of the 3′ component. When two degenerate bases exist, the order of the degenerate base is always R then Y (Eco394I and StyLTIII). We designated these as the 'R-Y degeneracy rules'.

(iv) Some 5′ component sequences are shared by different type I enzymes. They are GCA (CfrAI, Eco826I

and NgoAV), GGA (Eco377I and Eco777I), CCA (Eco646I and EcoprrI), GAG (EcoAI, EcoEI and StyLTIII), AAC (EcoKI and StySPI), GAA (Eco124I, Eco124II and KpnAI) and CGA (StySBLI and StySKI). The remaining 10 sequences are unique to each restriction enzyme, making a total of 17 unique 5′ component sequences.

(v) Comparison of the amino acids sequences may reveal several domains that recognize identical DNA sequences, which will be useful to predict recognition sequences of type I enzymes in the future.

(vi) Only a few type I sequences share common 3′ component sequences. Some examples are ATGC (Eco377I and EcoE1) and RTCG (Eco124I, Eco124II and StySGI).

(vii) Artificially constructed hybrid type I sequences retain the original either 5′ or 3′ component sequences.

(viii) Methylated bases are, so far, always adenines. The methylated adenines (thymine in the 3′ component in the Table 4) have a strong tendency to be located next to the central random DNA spacer. When two or three adenines are in the recognition sequence, the one closest to the central random DNA spacer is the target adenine. We refer to this as the 'methyl adenine rule'. Because of this trend, the frequency of adenine at the third base position in the 5′ component region is as high as 62% (17/27). Similarly, the frequency of thymine at the third base position (counting from the end) in the 3′ component region is 59% (16/27)

(ix) The number of bases between methylated adenines seems to be unique to each family (3). The distance for members of the type IA family is 8, type IB is 9, type IC is 7 or 8 and type ID is 6. When additional type I recognition sequences are found, we may find that different families have the same number of base pairs between two methylated adenines. An example is the KpnBI system, which contains 7 bp distance but does not belong to type IC. We named this system as a prototype of a new type IE family (8). However, it is very likely that the same family members contain the same distance between two adenines, thus this number can be used as a discriminatory factor. We named this rule the 'adenine distance' rule.

(x) One type I recognition sequence, Eco585I, does not contain any adenine in the 5′ component (GCC). We examined the possibility that the enzyme may be one of the type II enzymes, which contains methylcytosine as a methylation product and tried to purify the enzymes, but so far without success. Thus, this strain remains to be further characterized.

Using a plasmid transformation method, we have so far revealed a total of eight new prototype type I restriction enzymes and three isoschizomers from our collection of *E.coli* clinical strains. Because we have used only highly transformable strains for this test, additional type I prototypes or isoschizomers may exist in this collection. It is also possible that further studies will identify both type II and type III enzymes in this collection. The clinical strains characterized so far in our studies contain only a single type I restriction enzyme and do not accommodate any other type I, type II or type III enzymes within the cell. We have also used this

transformation method both in *Salmonella* and *Klebsiella* strains, and found four additional prototypes and one isoschizomer (7,8). Thus, in the past several years, a total of 13 prototypes, which correspond to roughly a half of the prototype sequences in Table 4, were found using this method. We believe that this method has the potential to screen for and identify new type I recognition sequences. Since many putative type I sequences have been discovered in various genomes (1), we may be able to apply this method to elucidate activities and recognition sequences of those putative type I enzymes.

REFERENCES

1. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2005) REBASE—restriction enzymes and DNA methyltransferases. *Nucleic Acids Res.*, **33**, D230–D232.
2. 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 of *Salmonella* typhimurium and the SP system of *Salmonella* potsdam. *J. Gen. Microbiol.*, **95**, 166–172.
3. Murray,N.E. (2002) 2001 Fred Griffith review lecture. Immigration control of DNA in bacteria: self versus non-self. *Microbiology*, **148**, 3–20.
4. Bickle,T.A. and Kruger,D.H. (1993) Biology of DNA restriction. *Microbiol. Rev.*, **57**, 434–450.
5. Kasarjian,J.K.A., Iida,M. and Ryu,J. (2003) New restriction enzymes discovered from *Escherichia coli* clinical strains using a plasmid transformation method. *Nucleic Acids Res.*, **31**, e22.
6. Ellrott,K.P., Kasarjian,J.K.A., Jiang,T. and Ryu,J. (2002) Restriction enzyme recognition sequence search program. *BioTechniques*, **33**, 1322–1326.
7. Kasarjian,J.K.A., Hidaka,M., Horiuchi,T., Iida,M. and Ryu,J. (2004) The recognition and modification sites for the bacterial type I restriction systems KpnAI, StySEAI, StySENI and StySGI. *Nucleic Acids Res.*, **32**, e82.
8. Chin,V., Valinluck,V., Magaki,S. and Ryu,J. (2004) KpnBI is the prototype of a new family (IE) of bacterial type I restriction-modification system. *Nucleic Acids Res.*, **32**, e138.
9. Thomson,J.M. and Parrott,W.A. (1998) pMECA: a cloning plasmid with 44 unique restriction sites that allows selection of recombinants based on colony size. *BioTechniques*, **24**, 922–924.
10. Kannan,P., Cowan,G.M., Daniel,A.S., Gann,A.A. 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.
11. Suri,B., Shepherd,J.C. and Bickle,T.A. (1984) The *Eco*A restriction and modification system of *Escherichia coli* 15T⁻: enzyme structure and DNA recognition sequence. *EMBO J.*, **3**, 575–579.
12. Lautenberger,J.A., Kan,N.C., Lackey,D., Linn,S., Edgell,M.H. and Hutchison,C.A.,III (1978) Recognition site of *Escherichia coli* B restriction enzyme on phi XsB1 and simian virus 40 DNAs: an interrupted sequence. *Proc. Natl Acad. Sci. USA*, **75**, 2271–2275.

13. Nagaraja,V., Stieger,M., Nager,C., Hadi,S.M. and Bickle,T.A. (1985) The nucleotide sequence recognized by the *Escherichia coli* D type I restriction and modification enzyme. *Nucleic Acids Res*., **13**, 389–399.

14. Piekarowicz,A. and Goguen,J.D. (1986) The DNA sequence recognized by the *Eco*DXX1 restriction endonuclease. *Eur. J. Biochem*., **154**, 295–298.

15. Cowan,G.M., Gann,A.A. and Murray,N.E. (1989) Conservation of complex DNA recognition domains between families of restriction enzymes. *Cell*, **56**, 103–109.

16. Kan,N.C., Lautenberger,J.A., Edgell,M.H. and Hutchison,C.A.,III (1979) The nucleotide sequence recognized by the *Escherichia coli* K12 restriction and modification enzymes. *J. Mol. Biol*., **130**, 191–209.

17. Bickle,T.A. (1987) Restriction and modification systems. In Neidhardt,F.C *et al*. (ed.), *Escherichia coli and Salmonella, 1st edn*. American Society for Microbiology, Washington, DC, pp. 692–696.

18. 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.

19. Piekarowicz,A., Klyz,A., Kwiatek,A. and Stein,D.C. (2001) Analysis of type I restriction modification systems in the Neisseriaceae: genetic organization and properties of the gene products. *Mol. Microbiol*., **41**, 1199–1210.

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

21. Titheradge,A.J., King,J., Ryu,J. and Murray,N.E. (2001) Families of restriction enzymes: an analysis prompted by molecular and genetic data for type ID restriction and modification systems. *Nucleic Acids Res*., **29**, 4195–4205.

22. Gubler,M., Braguglia,D., Meyer,J., Piekarowicz,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.

23. Meister,J., MacWilliams,M., Hubner,P., Jutte,H., Skrzypek,E., Piekarowicz,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.

24. Abadjieva,A., Patel,J., Webb,M., Zinkevich,V. and Firman,K. (1993) A deletion mutant of the type IC restriction endonuclease EcoR1241 expressing a novel DNA specificity. *Nucleic Acids Res*., **21**, 4435–4443.

25. Gann,A.A., Campbell,A.J., Collins,J.F., Coulson,A.F. and Murray,N.E. (1987) Reassortment of DNA recognition domains and the evolution of new specificities. *Mol. Microbiol*., **1**, 13–22.

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