Article No. mb981886 *J. Mol. Biol.* (1998) **280**, 345–353

JMB

I-*Ppo*I and I-*Cre*I Homing Site Sequence Degeneracy Determined by Random Mutagenesis and Sequential *in Vitro* Enrichment

Gretchen M. Argast[1], Kathryn M. Stephens[1], Mary J. Emond[2] and Raymond J. Monnat Jr[1]*

[1]*Departments of Pathology and* [2]*Biostatistics, University of Washington, Seattle, WA 98195 USA*

**Corresponding author*

Plasmid libraries containing partially randomized cleavage sites for the eukaryotic homing endonucleases I-*Ppo*I and I-*Cre*I were constructed, and sites that could be cleaved by I-*Ppo*I or I-*Cre*I were selectively recovered by successive cycles of cleavage and gel separation followed by religation and growth in *Escherichia coli*. Twenty-one different I-*Ppo*I-sensitive homing sites, including the native homing site, were isolated. These sites were identical at four nucleotide positions within the 15 bp homing site, had a restricted pattern of base substitutions at the remaining 11 positions and displayed a preference for purines flanking the top strand of the homing site sequence. Twenty-one different I-*Cre*I-sensitive homing sites, including the native site, were isolated. Ten nucleotide positions were identical in homing site variants that were I-*Cre*I-sensitive and required the addition of SDS for efficient cleavage product release. Four of these ten positions were identical in homing sites that did not require SDS for product release. There was a preference for pyrimidines flanking the top strand of the homing site sequence. Three of the 24 I-*Cre*I homing site nucleotide positions apparently lacked informational content, i.e. were permissive of cleavage when occupied by any nucleotide. These results suggest that I-*Ppo*I and I-*Cre*I make a large number of DNA-protein contacts across their homing site sequences, and that different subsets of these contacts may be sufficient to maintain a high degree of sequence-specific homing site recognition and cleavage. The sequential enrichment protocol we used should be useful for defining the sequence degeneracy and informational content of other homing endonuclease target sites.

Keywords: intron homing; homing endonuclease; group I intron; *in vitro* enrichment; sequence degeneracy

Introduction

Homing endonucleases are a diverse family of proteins found in unicellular eukaryotes, *Archaea* and eubacteria. A majority are encoded by open reading frames contained within group I self-splicing introns. Several have been identified within inteins (self-splicing protein introns) or as free-standing open reading frames (ORFs; Belfort & Roberts, 1997). Homing endonucleases are able to promote the lateral transfer, or ''homing'', of the intron containing the endonuclease ORF to specific 15 to 40 bp DNA target or homing sites by generating DNA double-strand breaks in copies of the homing site that lack the mobile intron. Double strand break repair of the intron-containing allele transfers an intron copy to the cleaved allele and disrupts the homing site, thus rendering it resistant to additional cycles of cleavage (reviewed by Belfort & Perlman, 1995).

The two homing endonucleases we have studied, the *Physarum polycephalum* I-*Ppo*I and *Chlamydomonas reinhardtii* I-*Cre*I endonucleases, are members of two major classes of homing endonucleases (reviewed by Belfort & Roberts, 1997). I-*Ppo*I is a member of the His-Cys box family of homing endonucleases that are found in myxomy-

Present address: K. M. Stephens, GeneTrace Systems, Inc., Alameda CA, USA.

0022–2836/98/280345–09 $30.00/0

cetes and the amebae. The His-Cys box protein motif consists of two conserved histidine and three conserved cysteine residues within a 30 residue region of protein (Muscarella & Vogt, 1989; Johansen *et al.*, 1993). These conserved His and Cys residues appear to contribute to a Zn^{2+}-binding motif and to the endonuclease active site (Flick *et al.*, 1998). I-*Cre*I is a member of the large family of homing endonucleases that contain one or two copies of the LAGLIDADG protein motif. Site-directed mutagenesis studies (Hodges *et al.*, 1992; Gimble & Stephens, 1995; Henke *et al.*, 1995; Seligman *et al.*, 1997) and the recent X-ray crystal structures of I-*Cre*I and of PI-*Sce*I (Heath *et al.*, 1997; Duan *et al.*, 1997; reviewed by Aggarwal & Wah, 1998) indicate that the LAGLIDADG motif forms a dimer or intramolecular folding interface and contributes an N-terminal aspartic acid residue to the endonuclease active site(s). Both the His-Cys and LAGLIDADG homing endonucleases cleave DNA to generate four base, 3′ extended cohesive ends.

We have devised a general method to determine the DNA sequence requirements for homing site cleavage, and have used this method to determine the homing site sequence degeneracy that can be tolerated by I-*Ppo*I and I-*Cre*I. The 15 bp I-*Ppo*I and 24 bp I-*Cre*I homing sites were originally delineated by cleavage analyses of cloned native homing sites and of small numbers of naturally occurring variant or mutant homing sites (Muscarella *et al.*, 1990; Thompson *et al.*, 1992; Dürrenberger & Rochaix, 1993; Ellison & Vogt, 1993; Muscarella & Vogt, 1993; Seligman *et al.*, 1997; Wittmayer *et al.*, 1998; K. Knoche, personal communication). Subsequent analyses have indicated that other homing endonucleases can tolerate limited homing site sequence degeneracy (Colleaux

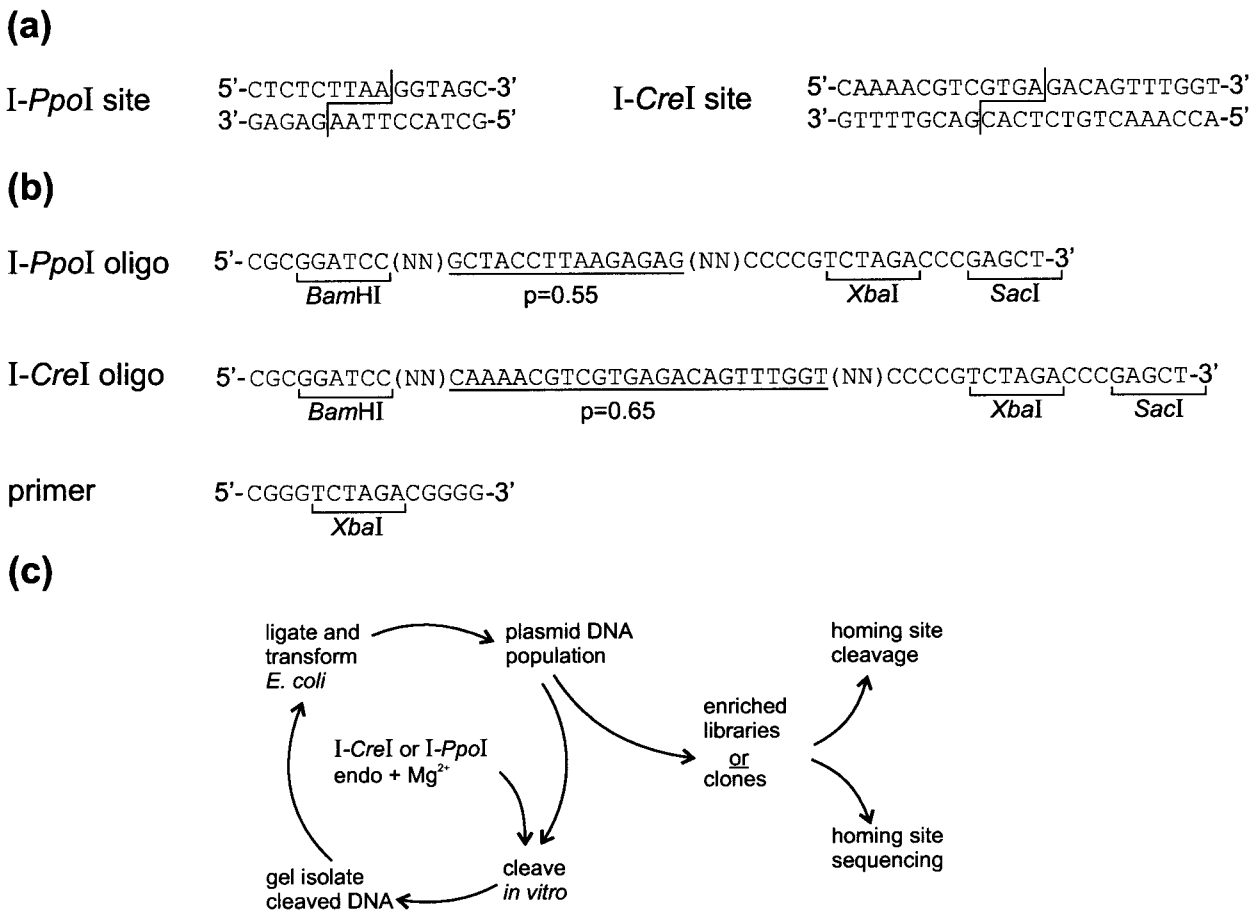

Figure 1. I-*Ppo*I and I-*Cre*I homing sites, oligonucleotides and enrichment cycle. (a) Native homing sites for I-*Ppo*I (Muscarella *et al.*, 1990; Ellison & Vogt, 1993) and I-*Cre*I (Thompson *et al.*, 1992; Dürrenberger & Rochaix, 1993). The site of asymmetric cleavage is indicated by the staggered line. (b) Oligonucleotides used to construct partially randomized homing site libraries in plasmid DNA. The I-*Ppo*I oligonucleotide contains the bottom strand, and the I-*Cre*I oligonucleotide contains the top strand, of the homing sites shown in (a). Underlined sequences were partially randomized during synthesis to contain the native nucleotide with probability p and equal probabilities of the other three nucleotides. Two fully random nucleotides flank each homing site to allow the identification of sibs and flank preferences following enrichment. Each oligonucleotide contains a binding site for the primer oligonucleotide shown at the bottom. The *Sac*I-compatible overhang (right) is generated upon primer annealing. (c) The strategy for sequential *in vitro* enrichment and characterization of cleavage-sensitive homing site variants.

et al., 1988; Wernette *et al.*, 1992; Bryk *et al.*, 1993; Lykke-Andersen *et al.*, 1994; Gimble & Wang, 1996; Aagaard *et al.*, 1997). None of these analyses, however, studied large numbers of potential homing site sequences. We thus extended these analyses by developing a method for isolating and characterizing endonuclease-sensitive homing site variants contained in large populations or ''libraries'' of potential homing sites.

Our results provide a detailed picture of homing site length and of site and flanking DNA sequence requirements for I-*Ppo*I and I-*Cre*I cleavage, and identify potential DNA-protein interactions that may be required for cleavage by these and related members of the His-Cys and LAGLIDADG homing endonuclease families. A similar approach was also recently used to examine the flanking sequence preferences of mutant and wild-type forms of the *Eco*RI restriction endonuclease (Windolph *et al.*, 1997).

Results

Construction and characterization of homing site plasmid libraries

Plasmid libraries containing partially randomized I-*Ppo*I or I-*Cre*I homing site sequences were constructed by modifying an existing protocol (Black *et al.*, 1993). Briefly, two oligonucleotides were synthesized that contained partially randomized I-*Ppo*I or I-*Cre*I homing sites, flanked by two random nucleotides and restriction endonuclease cleavage and common primer binding sites (Figure 1(a) and (b)). The probability of the native nucleotide at each homing site position was 55% for I-*Ppo*I and 65% for I-*Cre*I. A common primer oligonucleotide (Figure 1(b)) was annealed to each homing site oligonucleotide and extended with Klenow DNA polymerase to convert the homing site into double-stranded form. After cleavage with *Bam*HI, the homing site oligonucleotides were ligated into *Bam*HI- and *Sac*I-cleaved pBluescriptII SK (+) plasmid DNA prior to transformation into *Escherichia coli*. The I-*Ppo*I homing site library consisted of 6.4×10^4 independent transformants, while the I-*Cre*I homing site library consisted of 2.4×10^4 independent transformants.

Cleavage of plasmid libraries with *Xba*I, a restriction endonuclease that cleaves uniquely once in the homing site insert (Figure 1(b)), revealed that greater than 90% of the plasmids in each library contained homing site inserts (Figures 2(a) and 3(a)). DNA sequence analysis of 21 independent clones from each homing site library confirmed partial randomization at each nucleotide position within the homing sites, and complete randomization of the two flanking nucleotide positions (data not shown). The mean probability of the native nucleotide at each homing site position, calculated from observed frequencies, was 0.49 for the I-*Ppo*I library (95% confidence interval 0.42 to 0.55), and 0.59 for the I-*Cre*I library (95% confidence interval 0.53 to 0.64). The mean number of altered positions per homing site was 7.7 for the I-*Ppo*I library and 9.8 for the I-*Cre*I library. The potential complexity of the homing site libraries, estimated from the product of independent plasmid transformants in each library and the proportion of each library that contained homing site inserts ($\geqslant$90% for both libraries), was 5.8×10^4 for the I-*Ppo*I library and 2.1×10^4 for the I-*Cre*I library. The estimated frequency of native homing sites in each starting library, calculated from the product of the observed mean frequencies of the native nucleotide at 15 (I-*Ppo*I) or 24 (I-*Cre*I) homing site positions, was 1.5×10^{-5} for I-*Ppo*I and 1.5×10^{-6} for I-*Cre*I. An estimated $\geqslant$99.7% of insert-containing plasmid molecules in both libraries contained unique homing site sequences.

Sequential enrichment for I-*Ppo*I homing site variants

Sequential enrichment (Figure 1(c)) was used to isolate and amplify I-*Ppo*I-sensitive homing site plasmids from the I-*Ppo*I library. Enrichment was achieved by using agarose gel electrophoresis to separate, and thus allow the selective recovery of, homing site plasmids that could be cleaved by I-*Ppo*I. Cleaved plasmid molecules were recovered by in-gel ligation, followed by agarase digestion and electroporation into *E. coli*. Cleaved plasmid molecules were first clearly observed after three rounds of enrichment, when greater than 90% of the plasmid DNA population was I-*Ppo*I-sensitive. This fraction did not change after two additional rounds of enrichment (Figure 2(a)).

In order to determine the sequence complexity of the enriched plasmid population, we sequenced the homing site regions of 77 cleavage-sensitive plasmids. Twenty-one different homing site sequences were observed (Figure 2(b)). The remaining 56 clones were classified as sibs of these on the basis of identical flanking nucleotides. The number of different homing site variants provides a minimum estimate of the proportion of the I-*Ppo*I homing site library that was cleavage-sensitive prior to enrichment: $21/5.8 \times 10^4$ or $\sim$0.04%. The recovery of the native I-*Ppo*I homing site twice, in different sequence contexts, suggests that the complexity of the starting library was close to what we had originally calculated (observed frequency of $2/5.8 \times 10^4$ or 3.4×10^{-5}, *versus* a predicted frequency of 1.5×10^{-5}). All of the variant I-*Ppo*I homing sites could be cleaved in the homing site insert by I-*Ppo*I with only modest differences in cleavage efficiency as compared with the native homing site (data not shown).

In addition to the native I-*Ppo*I homing site, we isolated variant I-*Ppo*I homing sites that contained up to three base substitutions (Figure 2(b)). The mean substitution frequency was 1.5 positions/homing site. All of the homing site variants were identical at four A-T base-pairs, located at (positions -1 or $+2$) or symmetrically around

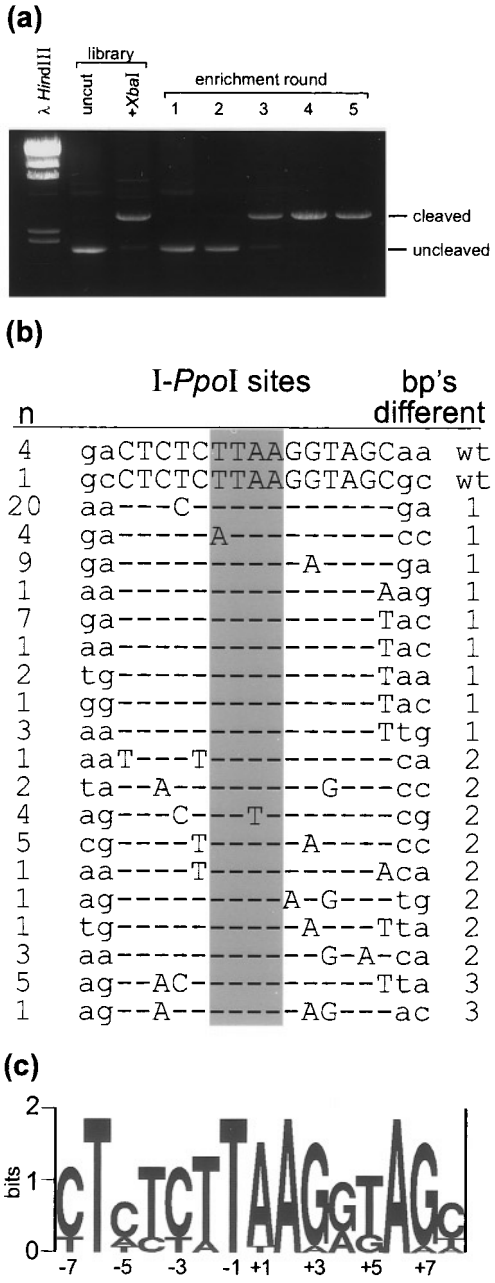

Figure 2. Sequential enrichment of cleavage-sensitive I-*Ppo*I homing site variants. (a) Cleavage sensitivity of the homing site library and successive rounds of enrichment. A majority of the starting library (>90%) contains homing site inserts, as revealed by *Xba*I sensitivity. Enrichment for cleavage-sensitive plasmids reached a plateau after three rounds of enrichment (compare enrichment rounds 1 and 2 with 3 to 5). Each lane contains 250 ng of homing site plasmid DNA. A λ*Hin*dIII size standard is shown to the left: the lower two fragments are 2.2 and 2 kb. (b) Sequence alignment of 77 I-*Ppo*I-sensitive homing sites isolated after four or five rounds of enrichment. Sequences are aligned on the central, cleaved 4 bp region of the I-*Ppo*I homing site (gray shaded). The number of plasmids isolated with a given sequence (*n*) is shown to the left, and the number of base-pair differences *versus* the native homing site is shown to the right (bp's different), of each homing site sequence. (c) Summary of sequences shown in (b) as a sequence logo (Schneider & Stephens, 1990). The informational content of each homing site position is indicated by bit height: identical positions in all sites are the tallest, while positions where any nucleotide can be substituted have no ''bit height'' as they lack informational content. The type and relative proportion of substitutions at other homing site positions are indicated by single-letter nucleotide abbreviations and height, respectively. Homing site positions are numbered to the right (the + positions) or left (−positions) of the center of the cleaved 4 bp region of the I-*Ppo*I homing site.

(positions −6 and +6) the center of the I-*Ppo*I homing site (Figure 2(c)). In the remaining 11 positions, seven of the 12 potential base-pair substitutions were observed with a slight preference for transitions. There was a strong preference for purines at the two nucleotide positions flanking the top strand to the left (88% A + G) though not to the right (55% A + G) half of the homing site sequence.

Sequential enrichment for I-*Cre*I homing site variants

I-*Cre*I endonuclease forms a tight, though noncovalent, complex with cleaved homing site DNA that can be disrupted by SDS (Wang *et al.*, 1997; Seligman *et al.*, 1997). We thus performed parallel enrichments to selectively isolate I-*Cre*I homing site plasmids that did (SDS-dependent sites) or did not (SDS-independent sites) require SDS for cleavage product release. Cleaved plasmid molecules were first observed in both protocols after two rounds of enrichment. After three rounds of enrichment >90% of the SDS-dependent plasmid population and ~75% of SDS-independent plasmid population could be cleaved by I-*Cre*I. These fractions did not change after additional rounds of enrichment (Figure 3(a)).

In order to determine the sequence complexity of our enriched I-*Cre*I-sensitive homing site populations we sequenced the homing sites of 131 SDS-dependent (Figure 3(b), top) and 44 SDS-independent (Figure 3(b), bottom) plasmids. Twelve SDS-dependent I-*Cre*I homing site variants including the native I-*Cre*I homing site and nine SDS-independent homing site variants were identified (Figure 3(b)). The remaining 154 plasmids were classified as sibs of these on the basis of identical flanking nucleotides. The frequency of cleavage-sensitive homing sites in the starting library was estimated from the number of observed site variants divided by the complexity of the starting library: $12/2.1 \times 10^4$ or ~0.06% for SDS-dependent sites, and $9/2.1 \times 10^4$ or ~0.04% for SDS-independent sites. Recovery of the native I-*Cre*I homing site suggests that the starting library may have been somewhat less complex than initially calculated (observed frequency $1/2.1 \times 10^4$ or 4.8×10^{-5}, threefold higher than the predicted frequency of 1.5×10^{-6}). All of the variant homing sites could be cleaved in the homing site insert

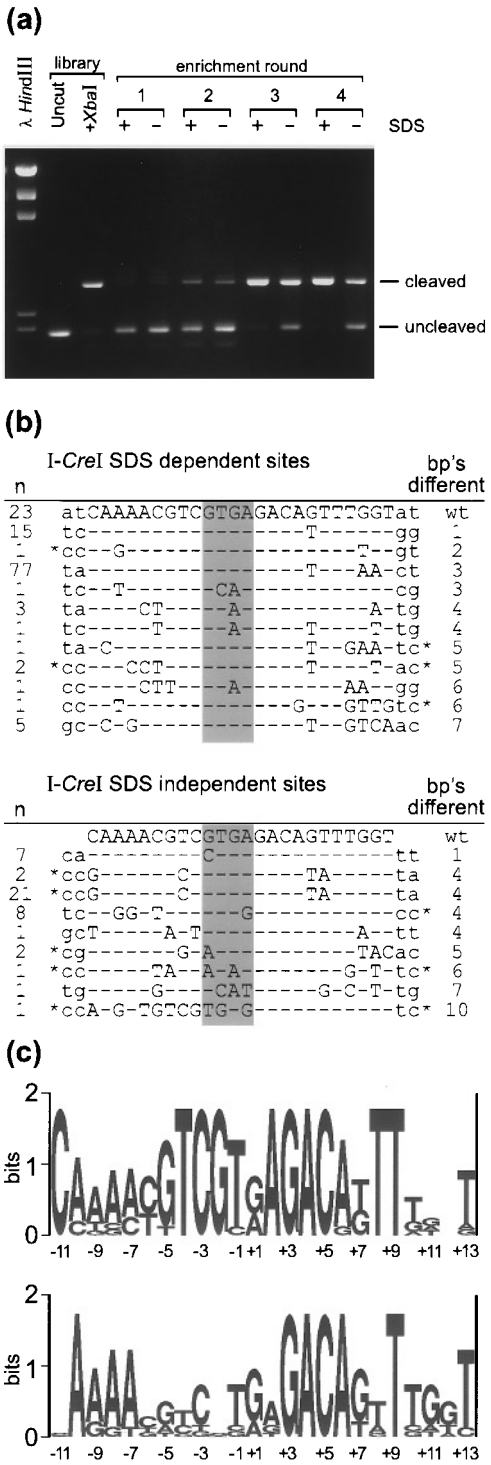

Figure 3. Sequential enrichment of cleavage-sensitive I-*Cre*I homing site variants. (a) Cleavage sensitivity of the homing site library prior to and following successive rounds of enrichment. A majority of the starting library (>90%) contains homing site inserts, as revealed by *Xba*I sensitivity. Enrichment for both SDS-dependent (+SDS) and SDS-independent (−SDS), cleavage-sensitive plasmids reached a plateau after three rounds of enrichment (compare enrichment rounds 1 and 2 with 3 and 4). Each lane contains 250 ng of homing site plasmid DNA. A λ*Hin*dIII size standard is shown to the left: the lower two fragments are 2.2 and 2 kb. (b) Sequence alignments of 131 SDS-dependent (top) and 44 SDS-independent (bottom) homing site plasmids isolated after at least three rounds of enrichment, displayed as described for Figure 2. Asterisks (*) indicate homing sites with deletions in flanking nucleotides of the homing site insert on the indicated side (not shown). An alternative, best fit alignment of the last SDS-independent homing site variant has a G insertion in the center of the homing site and three base substitutions in the left half of the homing site (see the text). (c) SDS-dependent (top) and SDS-independent (bottom) I-*Cre*I homing sites displayed as sequence logos.

with only modest differences in cleavage efficiency as compared with the native homing site. All of the sites isolated as SDS-dependent required the addition of SDS for cleavage product release (data not shown).

The SDS-dependent I-*Cre*I homing site sequences were identical at ten positions and contained up to seven base substitutions at the remaining 14 homing site positions (Figure 3(b)). The mean substitution frequency was 3.8 positions/homing site. Ten different base-pair substitutions were observed, with a slight preference for transversions. There was a preference for pyrimidines at the two nucleotide positions flanking the top strand of the left (79% C + T) though not the right (54% C + T) half of the homing site sequence (Figure 3(b), top). Five homing site variants were identified that contained deletions of one or two base-pairs adjacent to the homing site sequence (Figure 3(b)). The SDS-independent I-*Cre*I homing site sequences were identical at six nucleotide positions. Four of these six positions (+3 to +5 and +9) were identical as well in SDS-dependent I-*Cre*I homing sites (compare the sequence logos in Figure 3(c)). Sites containing as many as ten substitutions at the remaining 18 positions were observed. The mean number of substitutions per SDS-independent homing site was 5.3. Equal numbers of transitions and transversions were observed. Only A-to-C transversions were not seen. There was a modest preference for pyrimidines flanking the homing site. Six homing site variants contained one or two base-pair deletions adjacent to the homing site, and one contained a single base insertion in the center of the homing site.

Discussion

We have used a combination of random mutagenesis and sequential enrichment to selectively recover and amplify rare, cleavage-sensitive I-*Ppo*I and I-*Cre*I homing site variants from large populations of potential homing sites. The length of these homing sites (15 and 24 bp, respectively) precluded the systematic analysis of sequence degeneracy by conventional approaches such as site-directed mutagenesis. We thus developed an alternative approach by combining methods to generate partially randomized plasmid DNA

libraries (Black *et al.*, 1993) and to isolate short consensus sites for DNA-binding proteins (reviewed by Szostak, 1992).

The ability of sequential enrichment to recover rare, cleavage-sensitive molecules is indicated by the estimated frequencies of cleavage-sensitive sites in the starting libraries (~0.04% for the I-*Ppo*I library, and ~0.1% for the I-*Cre*I library) and by differences in the mean number of nucleotide substitutions in cleavage-sensitive homing sites as compared with the starting libraries. This reduction was from 7.7 to 1.5 substitutions/site for I-*Ppo*I-sensitive homing sites, and from 9.8 to 4 (SDS-dependent) or 5.3 (SDS-independent) substitutions/site for I-*Cre*I-sensitive homing sites. Early rounds of enrichment, where there is likely to have been a high enzyme-to-substrate ratio, did not lead to the preferential recovery of degenerate or ''star'' homing site sequences (Roberts & Halford, 1993; Pingoud & Jeltsch, 1997; additional data not shown). Thus, the protocol described here should be useful for determining the sequence degeneracy tolerated by other homing endonucleases.

Previous analysis indicated the need for a 15 bp long I-*Ppo*I homing site for efficient cleavage, although low levels of cleavage activity could be detected in sites that lacked one or two base-pairs at either end (Ellison & Vogt, 1993). Our results are in close agreement with these data, as we found no I-*Ppo*I homing site variant that differed in length by greater than one base-pair (Figure 2(b)). We also knew that insertions or deletions in the native I-*Ppo*I homing site completely abolished cleavage (Muscarella & Vogt, 1993; Wittmayer *et al.*, 1998; K. Knoche, personal communication). This indicates that the correct positioning or phasing of DNA-protein contacts across the I-*Ppo*I homing site is important for cleavage. I-*Ppo*I protects ~25 bp of DNA centered on the homing site (Ellison & Vogt, 1993). These footprinting and protection data, in conjunction with a preference for purines adjacent to the homing site, suggest that I-*Ppo*I also contacts DNA adjacent to the 15 bp homing site. Additional DNA-protein contacts adjacent to the homing site may facilitate homing site recognition or cleavage (Ellison & Vogt, 1993; Pingoud & Jeltsch, 1997).

Several lines of evidence indicate that I-*Ppo*I tolerates little homing site sequence degeneracy. There is no I-*Ppo*I cleavage site in the $\phi\chi174$, pBR322, λ, adenovirus 2 or T7 chromosomes, or in *E. coli* rDNA, and only a single I-*Ppo*I homing site in *Saccharomyces cerevisiae* in the chromosome XII rDNA locus (Lowery *et al.*, 1992). Analyses of the cleavage sensitivity of small numbers of base substitution mutants within the I-*Ppo*I homing site revealed that a majority of the base substitutions that we identified had little or no effect on cleavage sensitivity. In contrast, substitutions that we failed to identify among cleavage-sensitive site variants often inhibited cleavage by 10 to 100-fold (Lowery *et al.*, 1992; Muscarella & Vogt, 1993; Wittmayer *et al.*, 1998; K. Knoche, personal communication).

We found I-*Cre*I-sensitive homing site variants that differed in length from the native homing site due to as many as four base-pair substitutions at the ends of the homing site (Figure 3(b)). Although hydroxyl radical footprinting and interference assays previously indicated that I-*Cre*I protects a slightly asymmetrical region of ~14 bp from homing site positions −7 to +7 (Wang *et al.*, 1997), three lines of evidence suggest that the effective homing site length is at least 20 bp. First, at least 18 bp of the homing site sequence is required for efficient I-*Cre*I binding and cleavage (Thompson *et al.*, 1992; unpublished results). Second, we identified residues outside the central 14 bp region that were identical in SDS-dependent and/or SDS-independent homing site variants (Figure 3 (c)). Third, we have isolated dominant I-*Cre*I endonuclease mutants that inhibit homing site cleavage by native I-*Cre*I. Two of these mutants, K34N and F35C, have amino acid substitutions in a loop between strands β1 and β2 that may make base-specific, major groove contacts at the ends of the I-*Cre*I homing site sequence (Seligman *et al.*, 1997; Heath *et al.*, 1997).

I-*Cre*I may tolerate a higher degree of homing site sequence degeneracy than does I-*Ppo*I. This is suggested by the broader range of substitutions and larger fraction of variable sites (16.7% for I-*Cre*I, as opposed to 10% for I-*Ppo*I) we observed among cleavage-sensitive I-*Cre*I homing site variants. Sequence-tolerant cleavage was especially notable among SDS-independent I-*Cre*I homing site variants, in which each of the central 4 bp positions of the homing site could be substituted without abolishing cleavage, and two nucleotide positions were identified at which any nucleotide was permissive of cleavage (Figure 3(b), bottom). The loss of identical nucleotide positions in the left half of many SDS-independent I-*Cre*I homing sites (Figure 3(c)) is particularly intriguing: it suggests that the left half of the I-*Cre*I homing site plays a role in forming or stabilizing the I-*Cre*I-product complex. Comparable results suggesting a different affinity for homing half sites have been observed for other homing endonucleases, most notably those from *S. cerevisiae* (e.g. I-*Sce*I, I-*Sce*-III and PI-*Sce*I; see Perrin *et al.*, 1993; Schapira *et al.*, 1993; Gimble & Wang, 1996; Wende *et al.*, 1996).

I-*Cre*I, in contrast to I-*Ppo*I, appears to be able to tolerate at least some single base insertions and deletions within its homing site sequence. We found one SDS-independent site with a single base-pair insertion in the center of the homing site in conjunction with three other transitions (Figure 3(b), bottom, last SDS-independent sequence). In previous work we identified several I-*Cre*I homing sites with single base-pair deletions that can be cleaved at reduced efficiency (Seligman *et al.*, 1997). These results indicate that I-*Cre*I can bind and cleave at least a portion of homing sites in which the spacing or phasing of potentially important DNA-protein contacts has been altered.

Our results, in conjunction with previous data, indicate that I-*Ppo*I and I-*Cre*I make a large number of DNA-protein contacts across their homing sites, and that subsets of these contacts may be sufficient to maintain a high degree of sequence-specific homing site binding and cleavage. The tolerance of both homing endonucleases for a limited degree of homing site sequence degeneracy may be evolutionarily advantageous, as it would allow the lateral transfer of the *P. polycephalum* LSU3 and *C. reinhardtii* LSU mobile introns to organisms containing related, though non-identical, homing site sequences. I-*Ppo*I and I-*Cre*I may also be sufficiently tolerant of homing site sequence degeneracy to allow them to be engineered to recognize or cleave distantly related or novel homing site sequences with high specificity. Highly sequence-specific proteins of this type would be useful reagents for locating, cleaving or modifying the expression of specific genes in mammalian and other eukaryotic cells.

Materials and Methods

Library construction

Plasmid libraries containing partially randomized I-*Ppo*I or I-*Cre*I homing site sequences were constructed by modifying an existing protocol (Black *et al.*, 1993). Homing site inserts were generated from 47mer I-*Ppo*I and 56mer I-*Cre*I oligonucleotides (Keystone Labs) that contained partially randomized homing site sequences flanked by two random nucleotides, restriction cleavage sites for *Bam*HI (5′) and *Xba*I (3′) and a common primer binding site (Figure 2(b)). The specified frequency of native nucleotide at each homing site position was 0.55 for I-*Ppo*I and 0.65 for I-*Cre*I. The homing site oligonucleotides were converted into double-stranded form by primed *in vitro* DNA synthesis with Klenow DNA polymerase (Ausubel *et al.*, 1987; Black *et al.*, 1993). Synthesis reactions were incubated at 37°C for 30 minutes, followed by heating to 65°C for 10 minutes to inactivate polymerase. Completed extension reactions were digested with 20 units of *Bam*HI for two hours, followed by the addition of EDTA, phenol-extraction and desalting on a Centricon 10 spin concentrator (Amicon).

Plasmid libraries were constructed by ligating homing site inserts into pBluescriptII SK (+) plasmid DNA that had been digested with *Bam*HI and *Sac*I. Ligation reactions consisted of 10 ng (5 fmol) of cleaved plasmid DNA and an equimolar ratio of homing site insert in 20 μl of ligase buffer containing 400 units of T4 DNA ligase (New England BioLabs). Eight identical ligations were performed overnight at 16°C for each library. Ligase was inactivated by heating reactions to 65°C for ten minutes, and ligation products were desalted and concentrated on a Microcon-100 spin concentrator (Amicon) prior to electroporation into *E. coli* DH5α. Each ligation reaction was electroporated into 100 μl of competent cells using a BioRad electroporator set at 2 kV, 400Ω and 25μF with a 0.2 cm gap. LB broth (900 μl) was added immediately after electroporation, and aliquots representing 10% (or 100 μl) from each of three cultures were plated on LB agar plates supplemented with 100 μg/ml carbenicillin to determine transformation efficiency. The remaining transformation mixtures (3 × 900 μl and 5 × 1 ml) were pooled and grown in 100 ml of LB broth supplemented with 100 μg/ml carbenicillin overnight at 37°C. Plasmid DNA isolated from this culture using a Maxiprep kit (Qiagen) was used as the starting library for enrichment. Standard recombinant DNA methods were used throughout (Ausubel *et al.*, 1987).

Sequential enrichment

Sequential enrichment for I-*Ppo*I substrate plasmids was performed by digesting 2 μg of I-*Ppo*I plasmid library DNA with 210 units (2 μl) of I-*Ppo*I (Promega) at 37°C for one hour in digest buffer supplied by the manufacturer. Cleavage products were size-fractionated on 1% (w/v) low-melting agarose gels (FMC) run in TAE buffer at 1.4 V/cm overnight at room temperature. Size markers, consisting of uncut and linear plasmid DNAs, were used to locate and excise gel regions containing cleaved linear molecules. Excised gel slices were melted at 65°C for ten minutes, and then sterile water and 10 × T4 DNA ligase buffer were added to bring the final volume to 350 μl. T4 DNA ligase (800 units) was added and the reaction incubated at room temperature overnight. DNA ligase was inactivated by heating the samples to 65°C for ten minutes. Gelase (0.2 unit; Epicenter Technologies) and buffer were added to the ligation reactions, followed by incubation at 45°C for three hours to digest the agarose gel matrix. DNA was recovered, purified and concentrated on Microcon 100 spin columns, then electroporated into 100 μl of electrocompetent DH5α cells as described above. Aliquots (50 μl) of the transformation mix were plated, and the remainder was grown overnight at 37°C in 50 ml of LB broth supplemented with 100 μg/ml carbenicillin. Plasmid DNA was prepared from the resulting culture as described above and used for the next round of enrichment.

Two different protocols were used to selectively recover I-*Cre*I-sensitive homing sites. SDS-dependent homing site variants were isolated by separating cleavage products on 1% low melting temperature agarose in the absence of SDS, followed by the isolation of the band corresponding to uncut or supercoiled plasmid DNA. This band was excised from the gel and melted at 65°C for ten minutes, and then digested with Proteinase-K (100 μg/ml) in 200 μl of 10 mM Tris (pH 7.8), 5 mM EDTA containing 0.5% (w/w) SDS at 50°C for two hours. Digest products were phenol-extracted with an equal volume of buffer-equilibrated phenol, ethanol-precipitated and separated a second time by electrophoresis through a 1% low melting temperature agarose gel to isolate cleaved linear plasmid DNA.

Homing sites that did not require SDS for product release (''SDS-independent'' sites) were isolated as described above for I-*Ppo*I: 2 μg of the I-*Cre*I homing site plasmid library was digested with 76 ng (4 pmol) of D56G variant I-*Cre*I protein (Heath *et al.*, 1997) in 25 μl of 20 mM Tris (pH 8.0), 10 mM $MgCl_2$ at 37°C for one hour, followed by agarose gel electrophoresis after the addition of loading dye containing SDS. The remainder of the protocol was identical with that described above for I-*Ppo*I.

Cleavage sensitivity of homing site libraries and individual clones

The cleavage sensitivity of homing site libraries and individual clones was determined by endonuclease

digestion followed by agarose gel electrophoresis. I-*Ppo*I digestions used 0.25 μg of plasmid DNA and 10 units of I-*Ppo*I (Promega) at 37°C for one hour prior to electrophoresis on 1% agarose, TAE gels at 4.5 V/cm (Ausubel *et al.*, 1987). I-*Cre*I digestions used 0.5 μg of plasmid DNA and 0.38 ng (1 pmol) of I-*Cre*I in 20 μl of 20 mM Tris (pH 8.0), 100 mM KCl, 10 mM $MgCl_2$ containing 100 μg/ml BSA at 42°C for one hour. Half of each reaction (10 μl) was then mixed with loading dye that did or did not contain SDS to a final concentration of 1% (w/v) prior to electrophoresis on 1% agarose gels as described above. Cleavage within the homing site insert was confirmed by digesting plasmids with *Asp700* after homing endonuclease digestion, followed by agarose gel electrophoresis.

DNA sequence analysis of homing site variants

Plasmid DNA from individual clones was purified using Qiagen Spin Prep columns, and then sequenced using an ABI Dye Terminator chain terminator sequencing kit. Reactions were purified on Centrisep columns (Princeton Separations) prior to analysis on ABI sequencers in the University of WA Department of Biochemistry or the UW NIEHS Center for Ecogenetics Biomarkers facility (Seattle, WA).

Statistical analysis and sequence display

The observed substitution frequency at each homing site position was used to calculate the mean native nucleotide frequency over n homing site positions where $n = 15$ for the I-*Ppo*I site and $n = 24$ for the I-*Cre*I site. Distribution of the number of changes per site was determined using a binomial distribution, and the expected number of different sequences in each starting library was determined from multinomial distributions. Assumptions in each case were that the native nucleotide frequency at every site was equal to the observed native frequency, that substitutions at different homing site positions were independent, and that the other three nucleotides had equal probabilities of substitution at all homing site positions. The frequency of native homing sites in each starting library was calculated from the product of the observed mean frequencies of the native nucleotide at 15 (I-*Ppo*I) or 24 (I-*Cre*I) homing site positions. Summary or consensus homing site sequences were assembled and displayed by using sequence logo software (Schneider & Stephens, 1990).

Acknowledgments

We thank David Herrin, Baek Kim, Larry Loeb, Premal Patel, Len Seligman and Barry Stoddard for useful discussion, and Karen Flick, David Herrin, Kim Knoche (Promega Corporation) and Barry Stoddard for materials and/or communicating results prior to publication. David Herrin, Premal Patel, Lenny Seligman, Denise Spring and two anonymous reviewers provided useful comments. Figures were generated by Alden Hackmann, and sequence logos with the help of Shmuel Pietrokovski. This work was supported by NIH grant CA48022 (R.J.M. Jr) and NIEHS training grant T32 ES07032 (K.M.S.).

References

Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. & Struhl, K. (1987). *Current Protocols in Molecular Biology*, John Wiley & Sons, New York.

Aagaard, C., Awayez, M. J. & Garrett, R. A. (1997). Profile of the DNA recognition site of the archeal homing endonuclease I-*Dmo*I. *Nucl. Acids Res.* **25**, 1523–1530.

Aggarwal, A. K. & Wah, D. A. (1998). Novel site-specific DNA endonucleases. *Curr. Opin. Struct. Biol.* **8**, 19–25.

Belfort, M. & Perlman, P. S. (1995). Mechanisms of intron mobility. *J. Biol. Chem.* **270**, 30237–30240.

Belfort, M. & Roberts, R. (1997). Homing endonucleases–keeping the house in order. *Nucl. Acids Res.* **25**, 3379–3388.

Black, M., Munir, K. & Loeb, L. (1993). Random sequence selection for generation of active recombinant molecules. *Methods Mol. Genet.* **2**, 26–43.

Bryk, M., Quirk, S. M., Mueller, J. E., Loizos, N., Lawrence, C. & Belfort, M. (1993). The *td* intron endonuclease makes extensive sequence tolerant contacts across the minor groove of its DNA target. *EMBO J.* **12**, 2141–2149.

Colleaux, L., D'Auriol, L., Galibert, F. & Dujon, B. (1988). Recognition and cleavage site of the intron-encoded *omega* transposase. *Proc. Natl Acad. Sci. USA,* **85**, 6022–6026.

Duan, X., Gimble, F. S. & Quiocho, F. A. (1997). Crystal structure of PI-SceI, a homing endonuclease with protein splicing activity. *Cell,* **89**, 555–564.

Dürrenberger, F. & Rochaix, J.-D. (1993). Characterization of the cleavage site and the recognition sequence of the I-*Cre* I DNA endonuclease encoded by the chloroplast ribosomal intron of *Chlamydomonas reinhardtii*. *Mol. Gen. Genet.* **236**, 409–414.

Ellison, E. L. & Vogt, V. M. (1993). Interaction of the intron-encoded mobility endonuclease *I-Ppo*I with its target site. *Mol. Cell. Biol.* **13**, 7531–7539.

Flick, K. E., Jurica, M. S., Monnat, R. J., Jr & Stoddard, B. L. (1998). DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-*Ppo*I. *Nature,* in the press.

Gimble, F. S. & Stephens, B. W. (1995). Substitutions in conserved dodecapeptide motifs that uncouple the DNA binding and DNA cleavage activities of PI-*Sce*I endonuclease. *J. Biol. Chem.* **270**, 5849–5856.

Gimble, F. S. & Wang, J. (1996). Substrate recognition and induced DNA distortion by the PI-*Sce*I endonuclease, an enzyme generated by protein splicing. *J. Mol. Biol.* **263**, 163–180.

Heath, P. J., Stephens, K. M., Monnat, R. J., Jr & Stoddard, B. L. (1997). The structure of I-*Cre*I, a group I intron-encoded homing endonuclease. *Nature Struct. Biol.* **4**, 468–476.

Henke, R. M., Butow, R. A. & Perlman, P. S. (1995). Maturase and endonuclease functions depend on separate conserved domains of the bifunctional protein encoded by the group I intron aI4a of yeast mitochondrial DNA. *EMBO J.* **14**, 5094–5099.

Hodges, R., Perler, F., Noren, C. & Jack, W. (1992). Protein splicing removes intervening sequences in an archaea DNA polymerase. *Nucl. Acids Res.* **20**, 6153–6157.

Johansen, S., Embley, T. M. & Willassen, N. P. (1993). A family of nuclear homing endonucleases. *Nucl. Acids Res.* **21**, 4405–4405.

Lowery, R., Hung, L., Knoche, K. & Bandziulis, R. (1992). Properties of I-*Ppo*I: a rare-cutting intron-encoded endonuclease. *Promega Notes,* **38**, 8–12.

Lykke-Andersen, J., Thi-Ngoc, H. & Garrett, R. (1994). DNA substrate specificity and cleavage kinetics on an archeal homing-type endonuclease from Pyrobaculum organotrophum. *Nucl. Acids Res.* **22**, 4583–4590.

Muscarella, D. E. & Vogt, V. M. (1989). A mobile group I intron in the nuclear rDNA of *Physarum polycephalum*. *Cell,* **56**, 443–454.

Muscarella, D. E. & Vogt, V. M. (1993). A mobile group I intron from *Physarum polycephalum* can insert itself and induce point mutations in the nuclear ribosomal DNA of *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **13**, 1023–1033.

Muscarella, D. E., Ellison, E. L., Ruoff, B. M. & Vogt, V. M. (1990). Characterization of I-*Ppo,* an intron-encoded endonuclease that mediates homing of a group I intron in the ribosomal DNA of *Physarum polycephalum*. *Mol. Cell. Biol.* **10**, 3386–3396.

Perrin, A., Buckle, M. & Dujon, B. (1993). Asymmetrical recognition and activity of the I-*Sce*I endonuclease on its site and on intron-exon junctions. *EMBO J.* **12**, 2939–2947.

Pingoud, A. & Jeltsch, A. (1997). Recognition and cleavage of DNA by type-II restriction endonucleases. *Eur. J. Biochem.* **246**, 1–22.

Roberts, R. & Halford, S. (1993). Type II restriction endonucleases. In *Nucleases* (Linn, S., Lloyd, R. & Roberts, R., eds), pp. 35–88, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Schapira, M., Desdouets, C., Jacq, C. & Perea, J. (1993). I-*Sce*III an intron-encoded DNA endonuclease from yeast mitochondria. Asymmetrical DNA binding properties and cleavage reaction. *Nucl. Acids Res.* **21**, 3683–3689.

Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucl. Acids Res.* **18**, 6097–6100.

Seligman, L. M., Stephens, K. M., Savage, J. H. & Monnat, R. J., Jr (1997). Genetic analysis of the *Chlamydomonas reinhardtii* I-*Cre*I mobile intron homing system in *Escherichia coli*. *Genetics,* **147**, 1653–1664.

Szostak, J. W. (1992). *In vitro* genetics. *Trends Biochem. Sci.* **17**, 89–93.

Thompson, A. J., Yuan, X., Kudlicki, W. & Herrin, D. L. (1992). Cleavage and recognition pattern of a double-strand-specific endonuclease (I-*Cre*I) encoded by the chloroplast 23S rRNA intron of *Chlamydomonas reinhardtii*. *Gene,* **119**, 247–251.

Wang, J., Kim, H.-H., Yuan, X. & Herrin, D. L. (1997). Purification, biochemical characterization and protein-DNA interactions of the I-*Cre*I endonuclease produced in *Escherichia coli*. *Nucl. Acids Res.* **25**, 3767–3776.

Wende, W., Grindl, W., Christ, F., Pingoud, A. & Pingoud, V. (1996). Binding, bending and cleavage of DNA substrates by the homing endonuclease PI-*Sce*I. *Nucl. Acids Res.* **24**, 4123–4132.

Wernette, C., Saldanha, R., Smith, D., Ming, D., Perlman, P. & Butow, R. (1992). Complex recognition site for the group I intron-encoded endonuclease I-SceII. *Mol. Cell. Biol.* **12**, 716–723.

Windolph, S., Fritz, A., Oelgeschläger, T., Wolfes, H. & Alves, J. (1997). Sequence context influencing cleavage activity of the K130E mutant of the restriction endonuclease *Eco*RI identified by a site selection assay. *Biochemistry,* **36**, 9478–9485.

Wittmayer, P. K., McKenzie, J. L. & Raines, R. T. (1998). Degenerate DNA recognition by I-*Ppo*I endonuclease. *Gene,* **206**, 11–21.

Edited by G. Smith

(Received 15 December 1997; received in revised form 13 April 1998; accepted 20 April 1998)