Homing endonuclease I-CreI derivatives with novel DNA target specificities

Laura E. Rosen[1,2], Holly A. Morrison[1], Selma Masri[1], Michael J. Brown[1], Brendan Springstubb[1], Django Sussman[2], Barry L. Stoddard[2] and Lenny M. Seligman[1,*]

[1]Department of Biology and Program in Molecular Biology, Pomona College, 175 West 6th Street, Claremont, CA 91711, USA and [2]Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North A3-025 Seattle, WA 98109, USA

ABSTRACT

Homing endonucleases are highly specific enzymes, capable of recognizing and cleaving unique DNA sequences in complex genomes. Since such DNA cleavage events can result in targeted allele-inactivation and/or allele-replacement *in vivo,* the ability to engineer homing endonucleases matched to specific DNA sequences of interest would enable powerful and precise genome manipulations. We have taken a step-wise genetic approach in analyzing individual homing endonuclease I-CreI protein/DNA contacts, and describe here novel interactions at four distinct target site positions. Crystal structures of two mutant endonucleases reveal the molecular interactions responsible for their altered DNA target specificities. We also combine novel contacts to create an endonuclease with the predicted target specificity. These studies provide important insights into engineering homing endonucleases with novel target specificities, as well as into the evolution of DNA recognition by this fascinating family of proteins.

INTRODUCTION

Homing endonucleases are enzymes that recognize and cleave DNA target sequences of sufficient length (14–40 bp) to be unique in complex genomes (1,2). In nature they are often encoded by genes imbedded in introns or inteins, and function to promote lateral transfer of these intervening sequences. The lateral transmission mechanism involves homing endonuclease mediated cleavage of a target allele, followed by homology-based DNA double-strand break repair (2–4).

Homing endonucleases have been introduced into bacterial, plant, insect and mammalian cells, where they have been shown to recognize and cleave their specific targets (5–10). Such homing endonuclease induced DNA double-strand breaks may be lethal, mutagenic or repaired by homologous recombination. The ability to engineer homing endonucleases capable of cleaving specific DNA sequences of interest could enable new allele-specific replacement and inactivation strategies in a wide variety of organisms.

I-CreI has become the most extensively engineered homing endonuclease. Structural studies have defined the key protein/DNA interactions between the homodimeric enzyme and its largely palindromic DNA target site (Figure 1) (11,12). By systematically altering single protein/DNA contacts, we have previously described novel endonucleases that target mutant sites altered at three different positions: ±6, 10 and 11 (Figure 1) (13,14). Here, we employ saturation mutagenesis to examine contacts at three other positions: ±4, 5 and 7 (Figure 1). Two of these positions (±4 and 5) were previously examined in a high-throughput assay in which I-CreI derivatives simultaneously altered at three residues were selected from a non-saturating combinatorial library (15). We describe here a substitution at residue 44, not identified among the hundreds of I-CreI derivatives isolated in the high-throughput assay (15), which imparts specific recognition of a mutant site altered at position ±4. We also describe a specificity broadening substitution at residue 28, and show that the interaction between I-CreI residue 68 and target position ±5 is completely recalcitrant to individual specificity-shifting mutations. We demonstrate enhanced recognition of ±6 mutant sites (14) by altering a residue that normally interacts with DNA via a water-mediated contact, and employ a non-saturation mutagenesis screen to identify I-CreI derivatives with multiple substitutions that target sites altered at position ±9. In all, we report on the screening of over 400 endonuclease/target site combinations (Table 1). Finally, we show that it is possible to mix and
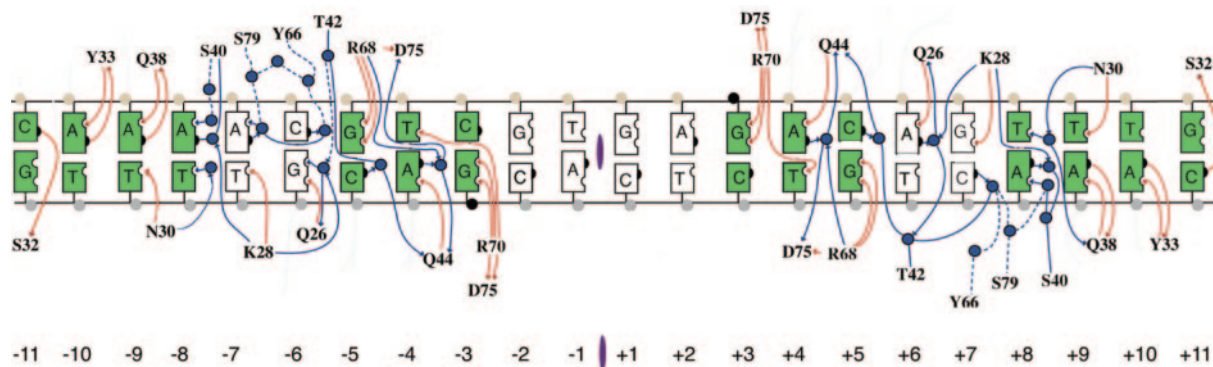
Figure 1. I-CreI/DNA contacts. Direct protein/DNA hydrogen bonds are red; water molecules and water-mediated hydrogen bonds are blue. Palindromic bases are green; non-palindromic are white. Scissile phosphate groups are indicated with black dots.

Table 1. CreI mutants analyzed *in vivo*

Amino acid substitution(s)	Sites assayed
Q44X[a]	±4 sites[b]
R68X[a]	±5 sites[b]
K28X[a]	±7 sites[c]
Q26C/Y66R/T42 to R, S, Q, C, K, N, E, H, F; Q26C/T42E	G:C ±6
T42E/Y66R	WT and C:G and T:A ±6
Q26A/T42E/Y66R, Q26A/T42E	A:T ±6
N30 to A, Q, C, G; Q38 to A, E, G, R; N30A/Q38 to A, C, I, K, L, M, N, P, R, S, T; N30R/Q38A, N30S/Q38A, N30G/Q38A, N30G/Q38Y, N30S/Q38R, N30T/Q38R, N30C/Q38R, N30L/Q38R, N30E/Q38K, N30S/Q38K, N30R/Q38T, N30R/S32G/Q38A, N30R/S32G/Q38Y, N30G/S32Q/Q38A, N30G/S32Q/Q38K, N30R/S32Q/Q38A, N30P/S32L/Q38A	±9 sites[b]

[a]Indicates all 19 mutants with single amino acid substitutions at that I-CreI position.
[b]The wild-type I-CreI site (WT), as well as the three symmetrically altered mutant target sites at that position.
[c]The wild-type I-CreI site, as well as the four symmetrically altered mutant target sites at ±7.

match novel protein/DNA contacts rationally to design an endonuclease that specifically targets a site altered at multiple positions.

MATERIALS AND METHODS

Bacterial strains and media

In vivo assays were performed in the *Escherichia coli* K-12 strain MC1000 [*araD139* Δ(*ara,leu*)*7697* Δ(*lac*)*X74 galE galK thi rpsL*] harboring F′—*lac* episomes with I-CreI target sites situated in place of lacO sequences as described previously (13). Standard growth media were used (16). Where required, media were supplemented with ampicillin (200 µg/ml), chloramphenicol (30 µg/ml), tetracycline (10 µg/ml), X-Gal (40 µg/ml) and arabinose (0.2%).

Plasmids

I-CreI derivatives were constructed in plasmid pA-E using cassette mutagenesis as described previously (13). Mutations

were subcloned to a C-Terminal His-tagged version of the arabinose inducible plasmid pB-E to purify protein for *in vitro* cleavage assays (14). All mutants were sequence verified. I-CreI homing site containing plasmids pKS155 (5) and pBR-O-Xho (13) have been described.

Competitive cleavage assays

His-tagged versions of endonucleases were purified using nickel-affinity chromatography, following a 2 h arabinose induction of mid-exponential cells (14). The abilities of mutant I-CreI homing sites (present on pBR-O-Xho) to serve as substrates for purified I-CreI were determined as described previously (13,14). Assays were performed on 100 ng of each linearized plasmid (pBR-O-Xho for mutant target sites and pKS155 for wild-type) in 10 µl of 20 mM Tris (pH 9.0), 10 mM MgCl$_2$, 1 mM DTT and 50 µg/ml BSA. Minimal enzyme amounts sufficient to achieve complete digestion of each substrate were determined empirically (see Figure 3, 1X samples), and used to initiate a series of 2-fold dilutions (5-fold dilutions were required for the wild-type I-CreI versus A:T ± 4 G:C ± 10 assay). Reactions took place for 60 min at 37°C and were terminated by placing digestions on ice, followed by addition of loading buffer containing SDS (to 0.5% w/v), and electrophoresis through 1.2% agarose gels in 1X TBE buffer. The density of agarose bands was determined using a Kodak Image Station 440CF and Kodak 1D software; the fraction of DNA cleaved was calculated by dividing the density of product bands by the density of substrate plus product bands. Absolute activities were determined for each endonuclease as the amount of protein required for 50% cleavage of a 100 ng sample of cognate target site containing plasmid DNA: w.t I-CreI 0.3 pg, Q44V 0.2 pg, Y33R/Q44V 0.2 pg, N30G/S32Q/Q38K 0.3 pg, N30R/S32G/Q38Y 0.3 pg, K28R 0.3 pg, N30S/Q38R 0.5 pg, Q26C/T42E/Y66R 0.9 pg, Y33R 1.4 pg.

Crystallography

The DNA for co-crystallization was purchased from Oligos etc. (Wilsonville, OR). For the enzyme mutant Q44V, the DNA duplex consisted of two strands of sequences 5′-GCAA-AACG**A**CGTGAG**T**CAGTTTCG-3′ and its complement 3′-CGTTT**T**TGC**T**GCACTC**A**GTCAAAGC-5′ (the two altered bases in each half site are bold and underlined). For the

Table 2. Data processing and refinement statistics

Protein	K28R	Q44V
Space group	P2(1)	P2(1)
Cell parameters (Å)	a = 43.1	a = 43.1
	b = 68.2	b = 68.0
	c = 87.5	c = 87.5
	γ = 91.4	γ = 91.5
Resolution	2.3 Å	2.3 Å
Redundancy	3.3	2.2
Completeness (%)[a]	96.0 (93.9)	85.4 (80.5)
Average I/σ(I)[a]	32.7 (14.7)	25.8 (10.3)
R_{sym} (%)[a]	3.9 (9.7)	3.9 (9.1)
R_{work} (%)	25.0	23.5
R_{free} (%)	30.2	27.9
Ramachandran plot (% of modeled residues)		
Most favored	87.4	87.4
Additionally allowed	12.6	12.6
Disallowed	0.0	0.0
Average B(Å^2) (protein, DNA)	22.3	25.2

[a]Outer resolution bin 2.34–2.30.

enzyme mutant K28R, the DNA duplex consisted of two strands of sequences 5′-GCAAA**T**CGTCGTGAGACA**A**TT-TCG-3′ and its complement 3′-CGTTT**A**GCAGCACTCTG-T**T**AAAGC-5′. Crystals were grown for Q44V and K28R I-CreI bound to their respective cognate DNA constructs using a 1.5:1 molar ratio solution of DNA:protein, by hanging drop vapor diffusion against a reservoir containing 20 mM NaCl, 10 mM calcium chloride, 100 mM MES pH ranging from 6.2 to 6.8 and PEG 400 ranging from 21 to 33.5% v/v. The final concentration of I-CreI in the protein/DNA complex solution was 2.5 mg/ml. For both experiments, the crystal belongs to space group P2$_1$, with unit cell dimensions approximately a = 43 Å, b = 68 Å and c = 88 Å (Table 2).

Crystals were removed directly from the drops in which they grew, suspended in a fiber loop, frozen in liquid nitrogen and maintained at 100 K during data collection. Data from crystals of the Q44V/DNA and K28R/DNA complexes were collected in house using an RAXIS-IV imagine plate area detector (Rigaku, USA) and a rotating anode X-ray generator with a copper source providing X-rays at 1.54 Å wavelengths. Data were reduced using the DENZO/SCALEPACK crystallographic data reduction package (17) (Table 2).

The structures were solved via molecular replacement using EPMR (18) with the wild-type I-CreI/DNA complex structure as an initial search model. All structures were modeled in XtalView (19) and refined using CNS (20) with 5% of the dataset aside for cross-validation (21). The final refinement statistics (Table 2) for Q44V were d_{min} = 2.3 Å, R_{work}/R_{free} = 23.5/27.9 and for K28R were d_{min} = 2.3 Å, R_{work}/R_{free} = 25.0/30.2. Geometric analysis of the structures using PROCHECK (22) indicates no residues in any structure with generously allowed or unfavorable backbone dihedral angles.

RESULTS

Identification of altered specificity I-CreI derivatives

We have identified altered specificity I-CreI derivatives using an *E.coli*-based genetic screen (13,14). Briefly, plasmid-encoded I-CreI alleles are introduced into cells and assayed for the ability to cleave various target sites. These target sites are located on F′ *lac* episomes, in place of *lac*O sequences. Efficient target site cleavage results in cells being converted from lac$^+$ to lac$^-$, as evidenced by white or sectored colonies on media containing the β-galactosidase indicator X-gal (Figure 2).

I-CreI recognizes and cleaves a largely palindromic target sequence, with 7 of 11 bases identical between half-sites. The majority of contacts between the homodimeric enzyme and its DNA substrate occur at these palindromic positions (Figure 1). To ensure that each monomer in an I-CreI homodimer is presented with the same potential contacts, we focused on symmetrical target sequence mutations, where each member of a symmetrical pair is altered in the same fashion. Thus, for each palindromic position in the I-CreI target site, three symmetrical mutant sites were examined; for each non-palindromic position, four symmetrical mutant sites were examined. Site mutants that displayed resistance to cleavage by wild-type I-CreI were used as substrates to screen for novel endonucleases. Endonuclease derivatives were chosen in which mutations were made at amino acid position(s) predicted to interact at or near the altered target site bases.

Residue 44 and position ±4: a specificity-shift via a single side-chain substitution

In wild-type I-CreI, the glutamine residue at position 44 has been shown to form hydrogen bonds directly with the adenines present at target site positions ±4 (11,12). We had previously shown that a Q44A mutant retained some activity towards the wild-type target site, as evidenced by sectored colonies in our *in vivo* assay (13). To better understand the interactions at this position, we created the remaining 18 I-CreI mutants encoding unique amino acids at codon 44 and screened each for activity. Like Q44A, a number of these other mutants (Q44C, Q44I, Q44L, Q44M, Q44S, Q44T, Q44V and Q44W) displayed some activity towards the wild-type target site (Table 3). Each resulted in sectored colonies *in vivo*, in contrast to wild-type I-CreI, which yields white colonies. Thus, the interaction between the glutamine present at residue 44 in wild-type I-CreI and the wild-type T:A ±4 base pair appears to be optimal.

The three symmetrical target site mutants altering target site positions ±4 are all I-CreI resistant *in vivo*. When the 19 I-CreI mutants with substitutions at codon 44 were screened against these target sites, five endonucleases (Q44A, Q44C, Q44T, Q44W and Q44V) with increased affinities for the A:T ±4 site were revealed (Table 3). Each produced sectored colonies in our *in vivo* system. Of these, the Q44V mutant was deemed optimal, as it resulted in the majority of colonies in our assay being completely white, with a minority displaying small blue sectors (Figure 2). No endonuclease displayed increased activity towards either the C:G ±4 or the G:C ±4 sites.

A 6-His tagged version of the Q44V mutant was purified and examined in cleavage competition assays as described previously (13,14). The Q44V mutant displayed a 2.7-fold preference for the A:T ±4 site over the wild-type site *in vitro*; wild-type I-CreI displayed a 3.4-fold preference for the wild-type site over the A:T ±4 site (Table 4 and Figure 3). Thus,
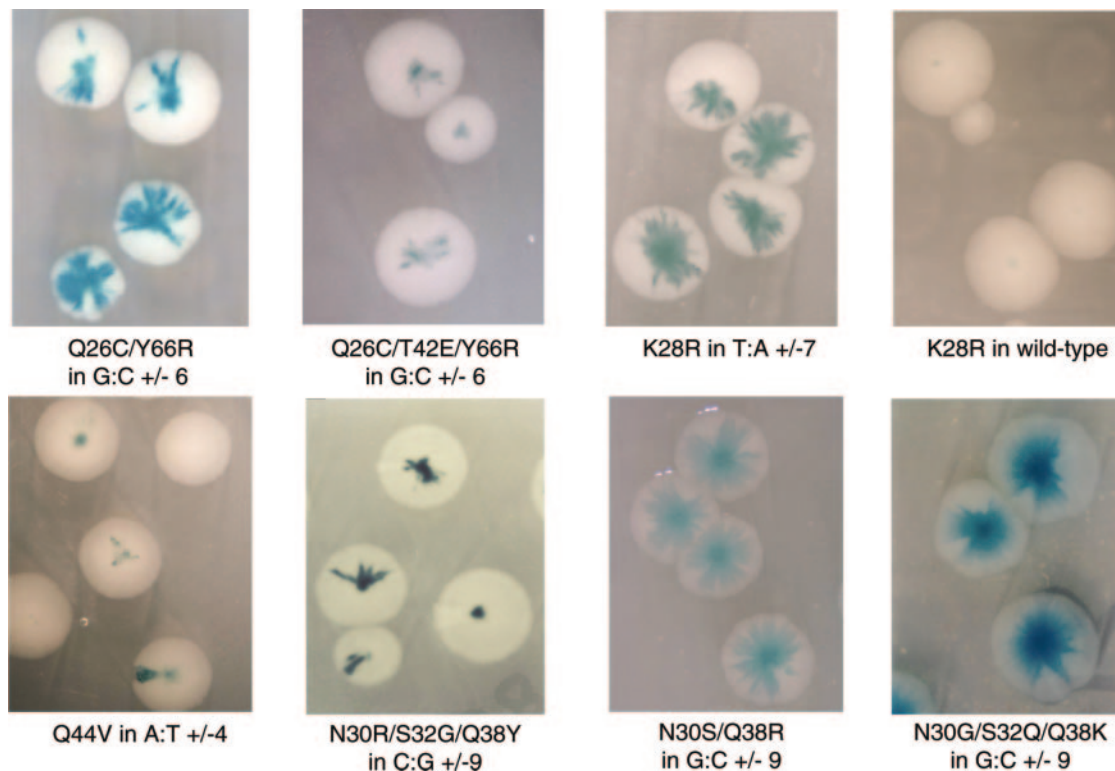
Figure 2. *In vivo* endonuclease assays. Colony phenotypes resulting from transforming indicated I-CreI derivatives into cells harboring the indicated target site mutants (13). Increased proportion of white cells within colonies indicates increased site affinity. The Q26C/Y66R mutant has been described previously (14).

Table 3. Active endonuclease derivatives identified

Target site	Active I-CreI derivatives[a]
A:T ±4	Q44A, Q44C, Q44T, Q44W, Q44V
A:T, C:G and T:A ±7	K28R
C:G ±9	Q38G, N30R/Q38A, N30G/Q38A, N30G/Q38Y, N30R/Q38T, N30R/S32G/Q38A, N30R/S32G/Q38Y
G:C ±9	N30A/Q38R, N30S/Q38R, N30T/Q38R, N30E/Q38K, N30G/S32Q/Q38A, N30G/S32Q/Q38K
Wild-type	Q44A, Q44C, Q44I, Q44L, Q44M, Q44S, Q44T, Q44V, Q44W, R68K, K28R, N30G, N30A, Q38G, Q38A, N30G/Q38A, N30R/S32G/Q38A

[a]All novel interactions identified displayed some level of *in vivo* sectoring; examples are shown in Figure 2.

the Q44 to V substitution resulted in a 9.4-fold specificity shift, with little specificity broadening (Table 4).

The structure of Q44V bound to its cognate target sites clearly illustrates the structural basis behind its altered specificity profile. In the wild-type endonuclease complex, residue Q44 is primarily responsible for base-specific readout of base pairs ±4 in the DNA half-sites, by making two hydrogen-bond contacts to an adenine base: one direct hydrogen bond contact to its N7 nitrogen (Figure 4A), and a water-mediated hydrogen-bond to its extracyclic amine. These contacts make recognition of base pairs ±4 by the wild-type endonuclease among the most specific of all positions in the protein/DNA complex (12). In contrast, the crystal structure of Q44V bound to its novel target demonstrates that the valine side-chain participates in a complementary hydrophobic packing arrangement with the thymine extracyclic methyl

group (Figure 4B). One γ-carbon of the valine side-chain forms a van der Waals contact with the thymine methyl group (distance = 3.7 Å) and with the δ-carbon of neighboring isoleucine 77 (3.6 Å). In addition, the second γ-carbon of V44 is packed against the δ-carbon of isoleucine 24 (3.7 Å) and the γ-carbon of threonine 42 (3.4 Å), completing a hydrophobic pocket surrounding the thymine base that comprises four separate aliphatic side-chains. Modeling of the Q44A mutant (shown above to recognize A:T ±4, albeit less well than Q44V) indicates that this mutation would allow unhindered positioning of the thymine methyl group in a hydrophobic environment, but that the packing scheme described above would be less stable, due to the absence of the two γ-methyl groups of the valine side-chain.

Residue 68 and position ±5: recalcitrance to specificity changes via single side-chain substitutions

In the wild-type I-CreI/DNA complex, the R68 in each monomer forms two hydrogen bonds with the guanine present at ±5 in each half-site (Figure 1) (11,12). We had previously shown that an R68A mutant was inactive against a wild-type target site (13). The remaining 18 I-CreI mutants with unique substitutions at codon 68 were generated and screened for activity. When examined against the wild-type target site, only one mutant (R68K) displayed any activity, producing sectored colonies in our assay. Thus, the wild-type arginine at position 68 appears to be optimal for recognition of the wild-type target site. The three symmetrical target site mutants altering target site positions ±5 are all I-CreI resistant *in vivo* (13). When the 19 I-CreI mutants with substitutions

Table 4. Cognate versus non-cognate cleavage efficiencies

Target site	I-CreI mutant	Wild-type cognate: non[a]	Mutant cognate: non[a]	Specificity shift[b]	Specificity broadening[c]
A:T ±4	Q44V	3.4	2.7	9.4	0.78
G:C ±6	Q26C/T42E/Y66R	8.0[d]	2.9	23.5	0.37
T:A ±7	K28R	5.3	0.8	4.2	0.15
C:G ±9	N30R/S32G/Q38Y	2.6	3.2	8.3	1.26
G:C ±9	N30S/Q38R	2.6	2.4	6.3	0.90
G:C ±9	N30G/S32Q/Q38K	2.6	2.1	5.6	0.81
G:C ±10	Y33R	4.0[d]	1.5	6.1	0.38
A:T ±4 G:C ±10	Y33R/Q44V	>450	3.2	>1440	<0.007

[a]Indicates the ratio of enzyme concentrations yielding 50% cleavage of cognate to non-cognate target sites for the enzyme/site combination indicated. Standard deviations were all <20%, with the exception of the wild-type I-CreI versus A:T ±4 G:C ±10 assay which yielded values of 456, 1622 and 2488; the larger variation in this series is presumably due to the large enzyme dilutions required to achieve 50% cleavage of both wild-type and mutant sites in these assays.
[b]Product of wild-type and mutant cognate to non-cognate cleavage efficiencies, as described previously (14).
[c]Ratio of mutant cognate to non-cognate/wild-type cognate to non-cognate cleavage efficiencies, as described previously (14). Values greater than 1 indicate higher specificity, while values of less than 1 indicate broader specificity.
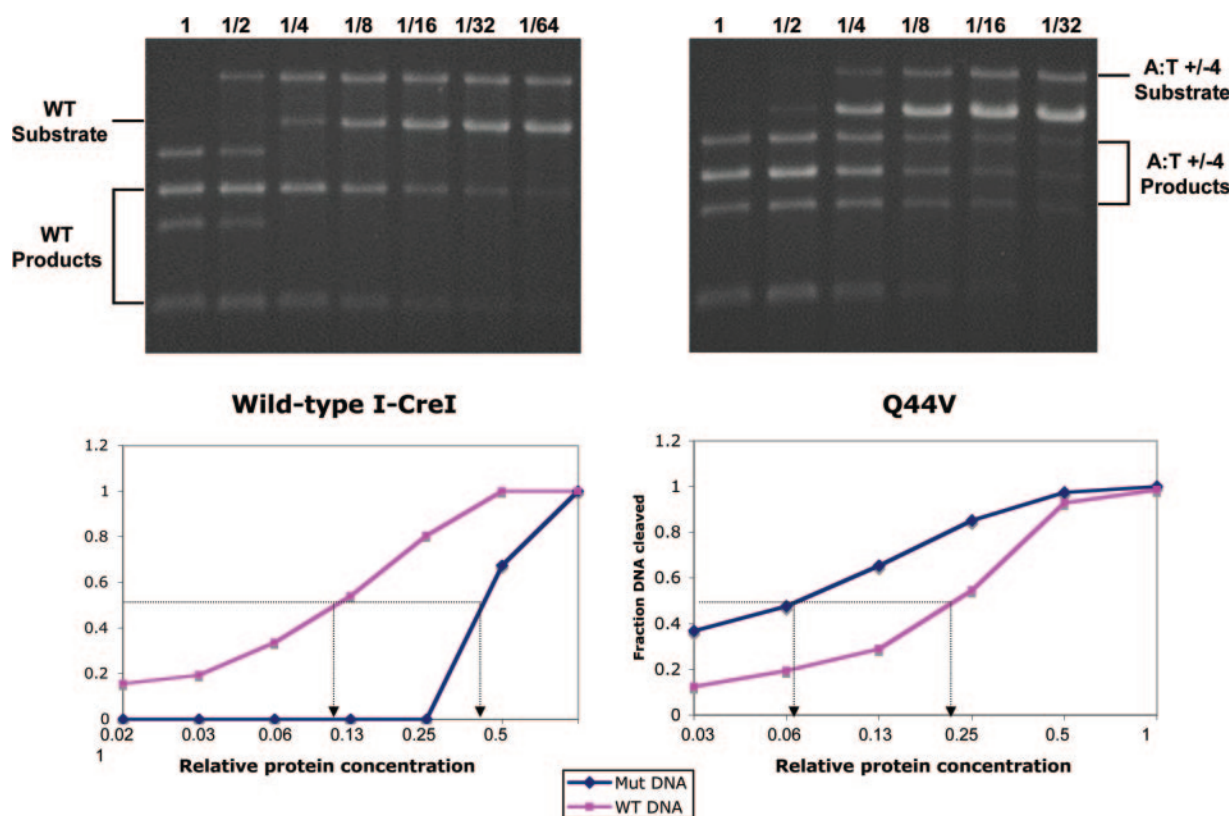[d]From (14).

Figure 3. Relative cleavage of cognate and non-cognate targets. Wild-type (left) and Q44V (right) endonucleases were exposed to linearized plasmids containing wild-type and A:T ±4 homing sites. The left lanes in each gel (upper) indicate the enzyme concentration required for complete digestion of each substrate. Serial 2-fold dilutions were performed, and the relative enzyme concentrations required for 50% cleavage of each substrate determined (lower).

at codon 68 were screened against these target sites, none displayed activity towards any of the mutant sites.

Position ±6: enhanced site recognition via altering a water-mediated contact

Four of eleven I-CreI target site positions are non-palindromic between half-sites (±1, 2, 6 and 7). Bases at two of these sites (±6 and ±7) are involved in important interactions with I-CreI residues (Figure 1) (12). Interactions between residues at I-CreI position 26 and bases at target site positions ±6 have been described previously (14). Briefly, when the four mutant target sites with symmetrical bases at ±6 were examined, the two that shared a base pair with the native site were efficiently cleaved by wild-type I-CreI, yielding completely white colonies *in vivo*. The two completely novel symmetrical target sites at ±6 were each resistant to cleavage by I-CreI. By screening I-CreI
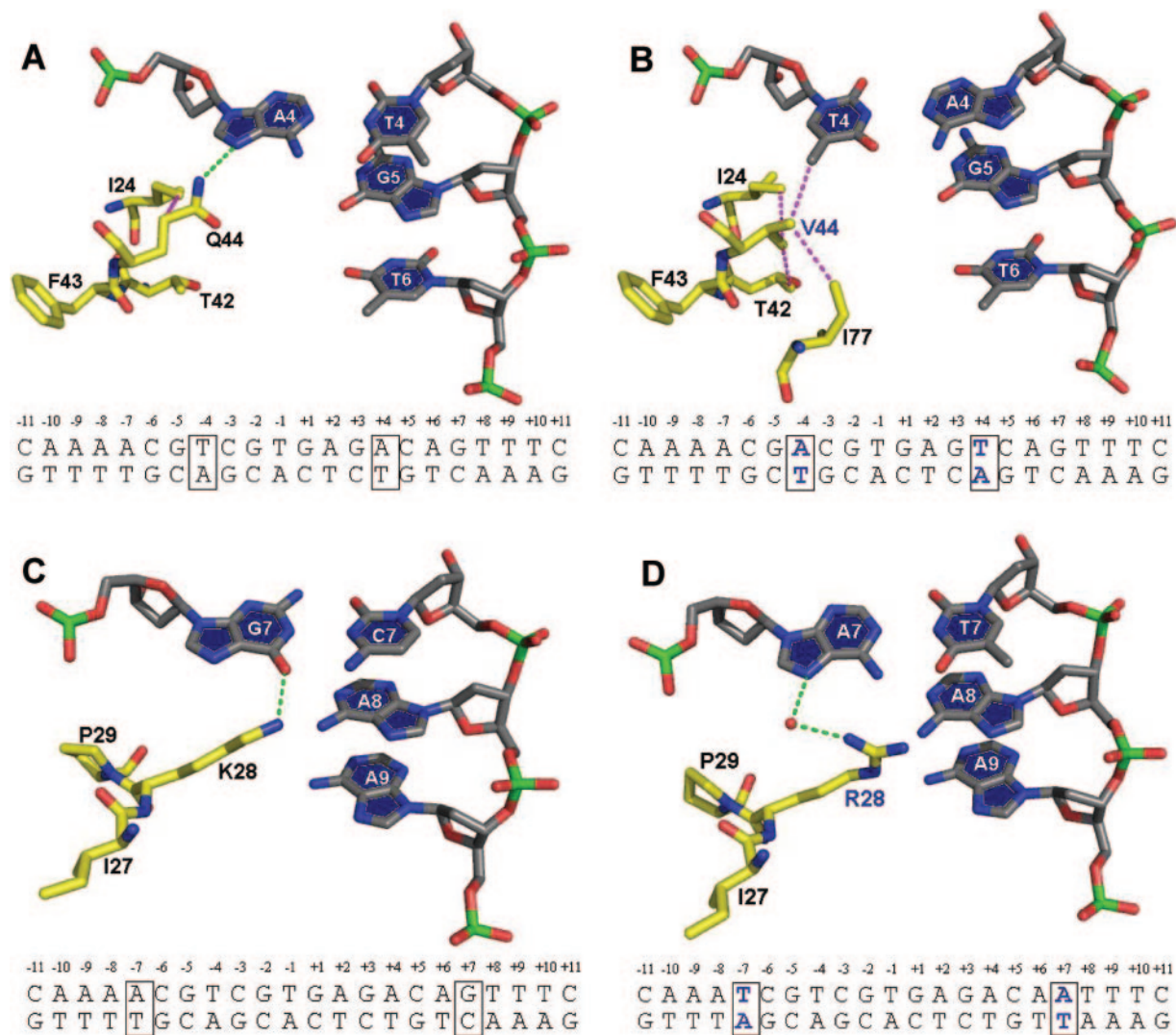
Figure 4. Structures and interactions of cognate pairs at residue 44 and base pair ±4 and residue 28 and base pair ±7. Protein/DNA contacts between wild-type I-CreI and its binding site are shown in (**A** and **C**). Contacts between Q44V and K28R and their cognate sites are shown in (**B** and **D**), respectively. The sequences of wild-type I-CreI target sequence and the alternate target sequences are shown below their corresponding structures. Dashed green lines represent hydrogen bonding and dashed magenta lines represent van der Waals interactions.

mutants altered at codon 26, endonucleases with increased affinities for each of these mutant sites were revealed: a Q26A mutant displayed optimal activity towards the A:T ±6 site while a Q26C mutant displayed optimal activity towards the G:C ±6 site (14).

Water-mediated contacts between endonuclease and target DNA are also important at this site (Figure 1) (12). In the native complex, water bridges occur between bases at this position and I-CreI residues T42 and Y66. We have previously described a Y66 to R substitution that increased the activity of the Q26C mutant towards the G:C ±6 site (14). We have followed up on that result by examining the effect of changes at I-CreI residue 42. A total of 9 amino acid substitutions (to R, S, Q, C, K, N, E, H, F) at this position were examined. Each was combined with the Q26C and Y66R substitutions, and screened for the ability to improve recognition of the G:C ±6 site mutant. One substitution, 42E, resulted in enhanced site recognition (Figure 2).

The Q26C/T42E/Y66R triple mutant displayed a 2.9-fold preference for its cognate site *in vitro* (Table 4), up from 2.4-fold for the Q26C/Y66R double mutant (14).

We next examined effects of the T42E substitution in different contexts in our *in vivo* system. The T42E/Y66R combination alone decreased the activity of otherwise wild-type I-CreI (Q26) towards the wild-type target site, as well as for the two palindromic sites that feature a wild-type contact (C:G ±6 and T:A ±6, Figure 1). Interestingly, we found that the activity of Q26A towards the A:T ±6 site was improved slightly by the addition of the T42E/Y66R combination. We had shown previously that the Y66R mutation alone decreased the activity of Q26A towards the A:T ±6 site (14). Similarly, the T42E substitution added alone to Q26A or Q26C decreased recognition for cognate target sites. Thus, it appears that the T42 to E and Y66 to R substitutions must be present together for optimal recognition of the G:C ±6 site by Q26A, and of the A:T ±6 site by Q26C.

For target sites with other bases at ±6, the T42E/Y66R combination results in decreased recognition.

Residue 28 and position ±7: a specificity-broadening substitution

Target site position ±7 is non-palindromic, with I-CreI residue K28 hydrogen bonding with a thymine in one half-site and a guanine in the other (Figure 1) (11,12). As was the case at position ±6 (14), the two symmetrical mutant target sites that share a base pair with the native site (A:T ±7 and C:G ±7) were efficiently cleaved by wild-type I-CreI, yielding completely white colonies *in vivo*. The two completely novel symmetrical target sites at ±7 were each resistant to cleavage by I-CreI, as demonstrated by blue colonies.

Each of 19 amino acid substitutions altering position 28 were generated and assayed against the wild-type target site and the four mutant sites symmetrically altered at bases ±7. The only active endonuclease was the K28R mutant. Like wild-type I-CreI, the K28R mutant recognizes the native site (Figure 2) and each symmetric homing site containing a native base pair. When assayed against these three sites, the K28R mutant yielded mostly white colonies with small blue sectors, indicating a slightly lower affinity than wild-type I-CreI for these three sites. However, unlike wild-type I-CreI, the K28R mutant was also active against one of the two completely novel sites, T:A ±7 (Table 3), yielding sectored colonies *in vivo* (Figure 2). The similar affinity of K28R for both the wild-type and the T:A ±7 sites *in vivo* was confirmed *in vitro*, where the K28R mutant displayed a slight 1.2-fold preference for the wild-type site (Table 4). In contrast, wild-type I-CreI displayed a 5.3-fold preference for the wild-type site over the T:A ±7 site. Thus, the K28 to R substitution results in significant specificity broadening (Table 4), enabling the enzyme to recognize a completely novel site while maintaining relatively high activity towards the wild-type site.

The structure of K28R bound to the T:A ±7 site demonstrates the basis of that mutant's broadened specificity. The introduction of an additional guanidino nitrogen moiety at the end of the arginine side-chain expands the conformational and hydrogen-bond repertoire of that side-chain. In this case, the new side-chain makes a water-mediated contact from its NH2 side-chain nitrogen to the N7 nitrogen of the adenine base in the new T:A base pair, while still allowing a similar contact to the N7 nitrogen of guanine in the wild-type target site (Figure 4C and 4D).

Position ±9: multiple substitutions required for optimal site recognition

The I-CreI/DNA contacts at target site bases ±9 are unique in that bases on each strand of DNA participate in direct hydrogen bonding with endonuclease residues: Q38 in each monomer forms two hydrogen bonds with the adenine present in each half-site, while N30 in each monomer forms a single hydrogen bond with each thymine on the opposite strand (Figure 1) (11,12). We had previously shown that an N30A mutant retained activity towards the wild-type target site, as evidenced by sectored colonies in our *in vivo* assay (13). Here, we began by screening mutants altered at residue

38 for novel contacts. To minimize the potential for unfavorable interactions between N30 and corresponding mutant bases, these studies were done in the presence of the N30 to A substitution. Eleven codon 38 substitutions were isolated in this context (to A, C, I, K, L, M, N, P, R, S and T). Each was assayed *in vivo* against the wild-type site and the three symmetrical sites altered at ±9 (Table 1). Of these, only the N30A/Q38R mutant displayed any activity, yielding sectored colonies when assayed against the G:C ±9 site (Table 3). When a Q38R single mutant (containing the wild-type N at residue 30) was assayed against the G:C ±9 site no activity was observed, validating our approach of screening for novel contacts at residue 38 in an N30A background.

Having demonstrated that simultaneously altering residues 30 and 38 could result in increased activity towards a site mutant altered at ±9, we next employed mutagenic strategies to simultaneously alter both residues. Resulting mutants were isolated in screens for white or sectored colonies in the F′ o-cre assay (13) in site strains altered at ±9 (Figure 1). These efforts yielded an additional eleven mutants altering residues 30 and 38 (Table 1). Of these, four displayed activity against the C:G ±9 site, and three displayed activity against the G:C ±9 site (Table 3).

Crystallographic analysis of wild-type and mutant I-CreI/DNA complexes revealed a hydrogen bond pair between N30 and the adjacent S32 (14). Reasoning that substitutions at residue 30 may be constrained by S32, we isolated mutants in which I-CreI residues 30, 32 and 38 were simultaneously altered. These efforts yielded two more novel endonucleases that target the C:G ±9 site, as well as two that target the G:C ±9 site (Table 3). In total, 36 single, double and triple mutants altering residues 30, 32 and 38 were analyzed in this study (Table 1). Seven mutants with increased activities towards the C:G ±9 site and six with increased activities towards the G:C ±9 site were identified (Table 3). None of our mutants displayed increased activity towards the T:A ±9 site.

Based upon *in vivo* phenotypes, we concluded that the N30R/S32G/Q38Y mutant had the highest activity towards the C:G ±9 site, and that the N30S/Q38R and N30G/S32Q/ Q38K mutants had the highest activities towards the G:C ±9 target site (Figure 2). These three mutants were purified and examined *in vitro*. Each displayed a 2- to 3-fold preference for the cognate mutant site. As wild-type I-CreI shows a similar preference for its cognate site, the total specificity shifts displayed by these mutants were 5- to 8-fold (Table 4).

An engineered endonuclease resulting from combining novel contacts

We have previously described an I-CreI Y33R mutation that results in high affinity for a G:C ±10 site mutant (13). To test whether novel contacts can be mixed and matched to generate I-CreI derivatives with predicted specificities, we engineered an endonuclease containing both the Y33 to R and Q44 to V substitutions. The Y33R/Q44V mutant was assayed *in vitro* against the corresponding target site containing two mutations in each half-site (A:T ±4 and G:C ±10). The mutant displayed a >1440-fold specificity shift, more than an order of magnitude greater than the product

of single-mutant specificity shifts (Table 4). It should be noted that the majority of this shift resulted from wild-type I-CreI displaying such a low affinity for the mutant target site. The Y33R/Q44V endonuclease has undergone significant specificity broadening (Table 4), an undesired result in the quest to create highly specific engineered endonucleases.

DISCUSSION

Engineered homing endonucleases hold great promise for a variety of applications, from gene therapy in humans to the eradication of natural populations of pathogens (23). Our work and that of others has clearly established I-CreI as the most extensively characterized, and the most widely engineered, homing endonuclease (5,13–15,24,25,26). The work described above significantly extends our ability to engineer novel I-CreI derivatives and provides important insights into the mechanisms by which homing endonucleases have evolved.

Structural studies have identified specific protein/DNA contacts at 8 of 11 bp in each half-site of the I-CreI target sequence (Figure 1). Our previous studies have described single amino acid substitutions resulting in novel interactions at three of these eight positions: ±10 and ±11 (13), and ±6 (14). Here we described novel interactions resulting from single amino acid substitutions at two other positions, ±4 and ±7. We also demonstrated that no single amino acid substitutions altering residue 68 yielded novel interactions with site mutants altered at positions ±5. The R68K mutant did, however, retain significant activity toward the wild-type target site, illustrating the importance of a basic residue at this position. Interestingly, novel I-CreI derivatives with arginine altered to non-basic residues at amino acid 68 have been described by others, all in the context of endonucleases containing a D75N mutation (15). Taken together, these results support the existence of a structurally predicted salt-bridge between R68 and D75 in wild-type I-CreI (12).

Given that the direct I-CreI/DNA contacts in the native complex involve exclusively polar amino acids (Figure 1) (11,12), it may come as a surprise that four hydrophobic substitutions (to A, C, W and V) were recovered that improved recognition of the A:T ±4 site, with the Q44 to V substitution being optimal. However, this is the third I-CreI position we have described in which substitutions to less polar residues impart specific recognition of target-site thymines; Y33C at T:A ±10 and Q26A at A:T ±6 are the others (13,14). Here and in the case of Y33C (14), the co-crystal structures revealed van der Waals contacts to the C5 methyl group of the thymine. The crystal structure of the homing endonuclease I-CeuI revealed a similar van der Waals interaction with thymine, in this case involving a leucine residue (27). The ability of such a wide-range of amino acids to direct specific DNA recognition towards thymines illustrates the importance of considering multiple amino acid substitutions when attempting to engineer homing endonucleases to target novel DNA sequences.

The interaction at the asymmetric target position ±7 involves direct hydrogen bonding of K28 to a thymine in one half-site and a guanine in the other. The importance of contacts at this position is established by the fact that both completely novel symmetric mutant sites are resistant to I-CreI cleavage *in vivo*. Since efficient cleavage of each symmetrical site that shares a base pair with the native site was observed, we conclude that each interaction in the native complex contributes to DNA recognition. The K28R mutant has undergone 'specificity broadening' in that it retains high affinity for the wild-type site and gains the ability to efficiently cleave one of the two completely novel mutant sites. The S32K mutant we have described previously (13) has similar properties. Relaxed specificity mutants, such as K28R and S32K may be important evolutionary intermediates, enabling an enzyme to recognize its native target site as well as one or more closely related sites.

The interaction at target site position ±9 is the only one involving direct hydrogen bonding to bases on each DNA strand (Figure 1). Given this fact, it is not surprising that multiple amino acid substitutions are required for optimal recognition. What is surprising is the relatively modest specificity shifts displayed by the three mutants we have characterized (Table 4). Each of these is lower than that of the single Q44V mutant. One possible explanation for this result is that the interaction at position ±4 is more important for DNA recognition than that at ±9, a notion supported by information content calculations (12,28). Another explanation is that the mutants isolated thus far that target novel ±9 sequences are sub-optimal, a very real possibility given the non-saturating nature of our screens for these mutants. However, even with our limited sampling, we can infer some rules for homing endonuclease site recognition. For example, five of the six endonucleases we have identified that target G:C ±9 have basic residues at amino acid 38, positioned to interact with the top-strand guanine. This fits with three other examples of basic residues recognizing guanines in the native structure (R70, R68, K28; Figure 1). Further, the I-CreI homolog I-MsoI recognizes a target site containing a top-strand guanine at position -9 via direct hydrogen bonding with an arginine residue (12).

The structures of Q44V and K28R I-CreI bound to their novel cognate target sites clearly illustrate the structural basis behind their altered specificity profiles, and the extent to which individual contacts between enzyme side-chains and nucleotide bases can be independently modified. As was the case with the previously described Y33C and Y33H structures (14), selection of a single residue substitution results in a unique alteration of contacts to an individual base, with little perturbation of neighboring side-chain conformations and contacts. Thus, modular redesign strategies of homing endonuclease specificity are supported by the structural data.

The step-wise approach we have employed, involving the use of saturation mutagenesis at individual protein/DNA contact positions and subsequently combining novel contacts, is one way to create novel homing endonucleases with altered target specificities. Others have reported selection schemes and high-throughput approaches to achieve this goal (15,26, 29,30). The most successful of these, in terms of the number of novel endonucleases identified, was a yeast-based system used to isolate I-CreI derivatives simultaneously altered at residues 44, 68 and 70 (15). Even such high-throughput

approaches are severely limited by the number of mutants required when attempting a saturation screen involving multiple amino acid positions. For example, we have shown here that Q44V is optimal for interaction with a bottom strand thymine at ±4. Arnould *et al*. (31) did not identify this substitution among their vast set of mutants altered at residue 44, presumably because their mutagenesis focused on only 12 of the 20 amino acids, and valine was excluded from that set. Recent work suggests that computational protein redesign will emerge as a powerful complement to genetic approaches for engineering homing endonucleases with novel target specificities. The engineering of highly specific endonucleases tailored to specific DNA targets of interest may ultimately involve a combination of these different approaches.

Unlike the engineering of endonucleases, the evolution of homing endonucleases undoubtedly involved sequential random mutations resulting in changes in target specificity. We have demonstrated that single amino acid substitutions can result in new endonucleases that target novel sites with specificities similar to that displayed by wild-type (i.e. Q44V), and can also result in new endonucleases with broader specificities (i.e. K28R). After movement into a new host, such broad specificity endonucleases may impart toxicity due to affinity for multiple targets, and thus undergo selective pressure to be made more specific by further mutation. These subsequent mutations could involve altering residues that participate in direct protein/DNA contacts or those that interact with DNA via water-mediated contacts (i.e. T42E). Even in cases where multiple residues participate in the recognition of a single base pair, a single amino acid substitution can retain affinity for the native site while providing a substrate for further productive mutations: the N30G, N30A, Q38G and Q38A mutants all retain partial recognition of the wild-type target site, and when paired with second (and sometimes third) mutations enable the specific recognition of mutant sites (Table 3). The step-wise analysis of single protein DNA contacts we have described thus provides important insights into the evolution of novel target specificities by homing endonucleases.

REFERENCES

1. Belfort,M. and Roberts,R.J. (1997) Homing endonucleases: keeping the house in order. *Nucleic Acids Res*., **25**, 3379–3388.
2. Chevalier,B.S. and Stoddard,B.L. (2001) Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. *Nucleic Acids Res*., **29**, 3757–3774.
3. Gimble,F.S. (2000) Invasion of a multitude of genetic niches by mobile endonuclease genes. *FEMS Microbiol. Lett*., **185**, 99–107.
4. Burt,A. and Koufopanou,V. (2004) Homing endonuclease genes: the rise and fall and rise again of a selfish element. *Curr. Opin. Genet. Dev*., **14**, 609–615.
5. Seligman,L.M., Stephens,K.M., Savage,J.H. and Monnat,R.J., Jr (1997) Genetic analysis of the *Chlamydomonas reinhardtii* I-CreI mobile intron homing system in *Escherichia coli. Genetics*, **147**, 1653–1664.
6. Donoho,G., Jasin,M. and Berg,P. (1998) Analysis of gene targeting and intrachromosomal homologous recombination stimulated by genomic double-strand breaks in mouse embryonic stem cells. *Mol. Cell. Biol*., **18**, 4070–4078.
7. Posfai,G., Kolisnychenko,V., Bereczki,Z. and Blattner,F.R. (1999) Markerless gene replacement in *Escherichia coli* stimulated by a double-strand break in the chromosome. *Nucleic Acids Res*., **27**, 4409–4415.
8. Monnat,R.J., Jr, Hackmann,A.F. and Cantrell,M.A. (1999) Generation of highly site-specific DNA double-strand breaks in human cells by the homing endonucleases I-PpoI and I-CreI. *Biochem. Biophys. Res. Commun*., **255**, 88–93.
9. Kirik,A., Salomon,S. and Puchta,H. (2000) Species-specific double-strand break repair and genome evolution in plants. *EMBO J*., **19**, 5562–5566.
10. Rong,Y.S. and Golic,K.G. (2000) Gene targeting by homologous recombination in *Drosophila. Science*, **288**, 2013–2018.
11. Jurica,M.S., Monnat,R.J., Jr and Stoddard,B.L. (1998) DNA recognition and cleavage by the LAGLIDADG homing endonuclease I-CreI. *Mol. Cell*, **2**, 469–476.
12. Chevalier,B., Turmel,M., Lemieux,C., Monnat,R.J., Jr and Stoddard,B.L. (2003) Flexible DNA target site recognition by divergent homing endonuclease isoschizomers I-CreI and I-MsoI. *J. Mol. Biol*., **329**, 253–269.
13. Seligman,L.M., Chisholm,K.M., Chevalier,B.S., Chadsey,M.S., Edwards,S.T., Savage,J.H. and Veillet,A.L. (2002) Mutations altering the cleavage specificity of a homing endonuclease. *Nucleic Acids Res*., **30**, 3870–3879.
14. Sussman,D., Chadsey,M., Fauce,S., Engel,A., Bruett,A., Monnat,R., Jr, Stoddard,B.L. and Seligman,L.M. (2004) Isolation and characterization of new homing endonuclease specificities at individual target site positions. *J. Mol. Biol*., **342**, 31–41.
15. Arnould,S., Chames,P., Perez,C., Lacroix,E., Duclert,A., Epinat,J.C., Stricher,F., Petit,A.S., Patin,A., Guillier,S. *et al*. (2006) Engineering of large numbers of highly specific homing endonucleases that induce recombination on novel DNA targets. *J. Mol. Biol*., **355**, 443–458.
16. Miller,J.H. (1992) *A Short Course in Bacterial Genetics: A Laboratory Manual and Handbook for Escherichia coli and Related Bacteria, 1st edn*. Cold Spring Harbor Laboratory Press, Plainview, NY.
17. Otwinowski,Z. and Minor,W. (1997) Processing of X-ray diffraction data collected in oscillation mode. *Meth. Enzymol*., **276**, 307–326.
18. Kissinger,C.R., Gehlhaar,D.K. and Fogel,D.B. (1999) Rapid automated molecular replacement by evolutionary search. *Acta Crystallogr. D Biol. Crystallogr*., **55**, 484–491.
19. McRee,D.E. (1999) XtalView/Xfit--a versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol*., **125**, 156–165.
20. Brunger,A.T., Adams,P.D., Clore,G.M., DeLano,W.L., Gros,P., Grosse-Kunstleve,R.W., Jiang,J.S., Kuszewski,J., Nilges,M., Pannu,N.S. *et al*. (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr*., **54**, 905–921.
21. Brunger,A.T. (1993) Assessment of phase accuracy by cross validation: the free R value. Methods and applications. *Acta Crystallogr. D Biol. Crystallogr*., **49**, 24–36.
22. Laskowski,R.A., MacArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallog*., **26**, 291–294.
23. Burt,A. (2003) Site-specific selfish genes as tools for the control and genetic engineering of natural populations. *Proc. Biol. Sci*., **270**, 921–928.
24. Chevalier,B.S., Kortemme,T., Chadsey,M.S., Baker,D., Monnat,R.J. and Stoddard,B.L. (2002) Design, activity, and structure of a highly specific artificial endonuclease. *Mol. Cell*, **10**, 895–905.
25. Epinat,J.C., Arnould,S., Chames,P., Rochaix,P., Desfontaines,D., Puzin,C., Patin,A., Zanghellini,A., Paques,F. and Lacroix,E. (2003)

A novel engineered meganuclease induces homologous recombination in yeast and mammalian cells. *Nucleic Acids Res.*, **31**, 2952–2962.

26. Chames,P., Epinat,J.C., Guillier,S., Patin,A., Lacroix,E. and Paques,F. (2005) *In vivo* selection of engineered homing endonucleases using double-strand break induced homologous recombination. *Nucleic Acids Res.*, **33**, e178.

27. Spiegel,P.C., Chevalier,B., Sussman,D., Turmel,M., Lemieux,C. and Stoddard,B.L. (2006) The Structure of I-CeuI homing endonuclease: evolving asymmetric DNA recognition from a symmetric protein scaffold. *Structure*, **14**, 869–880.

28. Argast,G.M., Stephens,K.M., Emond,M.J. and Monnat,R.J., Jr (1998) I-PpoI and I-CreI homing site sequence degeneracy determined by random mutagenesis and sequential *in vitro* enrichment. *J. Mol. Biol.*, **280**, 345–353.

29. Chen,Z. and Zhao,H. (2005) A highly sensitive selection method for directed evolution of homing endonucleases. *Nucleic Acids Res.*, **33**, e154.

30. Doyon,J.B., Pattanayak,V., Meyer,C.B. and Liu,D.R. (2006) Directed evolution and substrate specificity profile of homing endonuclease I-SceI. *J. Am. Chem. Soc.*, **128**, 2477–2484.

31. Ashworth,J., Havranek,J.J., Duarte,C.M., Sussman,D., Monnat,R.J., Jr, Stoddard,B.L. and Baker,D. (2006) Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature*, **441**, 656–659.