

# A combinatorial approach to create artificial homing endonucleases cleaving chosen sequences

Julianne Smith, Sylvestre Grizot, Sylvain Arnould, Aymeric Duclert, Jean-Charles Epinat, Patrick Chames, Jesús Prieto<sup>1</sup>, Pilar Redondo<sup>1</sup>, Francisco J. Blanco<sup>1</sup>, Jerónimo Bravo<sup>1</sup>, Guillermo Montoya<sup>1</sup>, Frédéric Pâques\* and Philippe Duchateau

CELLECTIS S.A., 102 route de Noisy 93235 Romainville, France and <sup>1</sup>Structural Biology and Biocomputing Programme, Centro Nacional de Investigaciones Oncológicas (CNIO), C/ Melchor Fdez Almagro, 28029 Madrid, Spain

Received July 18, 2006; Revised September 12, 2006; Accepted September 18, 2006

## ABSTRACT

**Meganucleases, or homing endonucleases (HEs) are sequence-specific endonucleases with large (>14 bp) cleavage sites that can be used to induce efficient homologous gene targeting in cultured cells and plants. These findings have opened novel perspectives for genome engineering in a wide range of fields, including gene therapy. However, the number of identified HEs does not match the diversity of genomic sequences, and the probability of finding a homing site in a chosen gene is extremely low. Therefore, the design of artificial endonucleases with chosen specificities is under intense investigation. In this report, we describe the first artificial HEs whose specificity has been entirely redesigned to cleave a naturally occurring sequence. First, hundreds of novel endonucleases with locally altered substrate specificity were derived from I-CreI, a *Chlamydomonas reinhardtii* protein belonging to the LAGLIDADG family of HEs. Second, distinct DNA-binding subdomains were identified within the protein. Third, we used these findings to assemble four sets of mutations into heterodimeric endonucleases cleaving a model target or a sequence from the human RAG1 gene. These results demonstrate that the plasticity of LAGLIDADG endonucleases allows extensive engineering, and provide a general method to create novel endonucleases with tailored specificities.**

## INTRODUCTION

Meganucleases are sequence-specific endonucleases with large (>14 bp) cleavage sites that can deliver DNA double-strand breaks (DSBs) at specific loci in living cells (1).

Meganucleases have been used to stimulate homologous recombination in the vicinity of their target sequences in cultured cells and plants (2–6), and these results have opened new perspectives for genome engineering in a wide range of applications. For example, meganucleases could be used to induce the correction of mutations linked with monogenic inherited diseases, and bypass the risks due to the randomly inserted transgenes used in current gene therapy approaches (7).

The use of meganuclease-induced recombination has long been limited by the repertoire of natural meganucleases. In nature, meganucleases are essentially represented by homing endonucleases (HEs), a family of endonucleases encoded by mobile genetic elements, whose function is to initiate DNA DSB-induced recombination events in a process referred to as homing (8). Several hundreds of HEs have been identified in bacteria, eukaryotes and archaea (8); however, the probability of finding a HE cleavage site in a chosen gene is very low. Thus, the making of artificial meganucleases with tailored substrate specificity has become the goal of several laboratories (9–14). Recently, Zinc-Finger DNA-binding domains (15) could be fused with the catalytic domain of the FokI endonuclease, to induce recombination in various cells types, including human lymphoid cells (16–18). However, these proteins have demonstrated high toxicity in cells (16,19), probably due to a low level of specificity. Given their biological function and their exquisite specificity, HEs could represent ideal scaffolds, but engineering their DNA-binding domain has long been considered a daunting task (9–14), and more generally, engineering the substrate specificity of proteins that cleave or recombine DNA has often proven to be difficult (20–23).

In a recent report, we described a semi-rational mutagenesis approach coupled with high-throughput screening (HTS) to derive hundreds of novel endonucleases with locally altered substrate specificity from I-CreI, a HE from the LAGLIDADG family (11,24). We also suggested a combinatorial approach to globally engineer the DNA-binding domain of such proteins in a rational way, and eventually design

\*To whom correspondence should be addressed. Tel: +33 1 41 83 99 00; Fax: +33 1 41 83 99 03; Email: paques@cellectis.com

tailored HEs cleaving chosen targets. This strategy relied on the hypothetical identification of several independent DNA-binding units within I-CreI or related proteins. LAGLIDADG proteins have a conserved core structure, with two characteristic  $\alpha\beta\beta\alpha\beta\beta\alpha$  folds facing each other across a 2-fold symmetry or pseudo-symmetry axis (8). Whereas it had been shown that two  $\alpha\beta\beta\alpha\beta\beta\alpha$  folds can be associated with heterodimeric (11) or single chain (25,26) molecules, it was unclear whether each fold could in turn be separated into distinct functional binding units.

We have now implemented this combinatorial strategy. In a first step, we have created novel I-CreI variants with locally altered substrate specificity, by modifying different parts of the  $\alpha\beta\beta\alpha\beta\beta\alpha$  folds. Second, we have identified two relatively independent DNA-binding subdomains within the same  $\alpha\beta\beta\alpha\beta\beta\alpha$  fold: mutations in different parts of the protein could be assembled to create novel endonucleases with predictable substrate specificities, without affecting activity and stability. Third, we used these findings to assemble four sets of mutations into heterodimeric endonucleases with fully engineered specificity, to cleave a model target and finally a sequence from the human RAG1 gene. This is the first time a homing endonuclease is entirely redesigned to cleave a naturally occurring sequence, but our results also provide a general method to create novel endonucleases cleaving chosen sequences.

## MATERIALS AND METHODS

### Construction of mutant libraries and target vectors

I-CreI N75 (25) mutant library randomized at four positions (28,33,38,40) with 10 defined amino acids (ADEKNQRSTY), and with a R70S substitution was generated by Biomethodes (Evry, France), resulting in a theoretical diversity of  $10^4$ . Two smaller libraries of complexity 1728 (123) were designed by randomization of three positions (30,33,38 and 28,30,38) and were constructed using VVK degenerate codons (18 codons, amino acids ADEGHKNPQRST) as described previously (11). In addition, small libraries of complexity 225 (152) resulting from the randomization of only two positions were constructed in an I-CreI N75 or I-CreI D75 scaffold, using NVK degenerate codons (24 codons, amino acids ACDEGHKNPQRSTWY). All libraries were introduced into the two micron-based replicative vector pCLS542 marked with the *LEU2* gene and transformed into the *Saccharomyces cerevisiae* strain FYC2-6A (*MAT $\alpha$* , *trp1 $\Delta$ 63*, *leu2 $\Delta$ 1* and *his3 $\Delta$ 200*) as described previously (11). Yeast reporter vectors were constructed and transformed into the *S.cerevisiae* strain FYBL2-7B (*MAT a*, *ura3 $\Delta$ 851*, *trp1 $\Delta$ 63*, *leu2 $\Delta$ 1* and *lys2 $\Delta$ 202*) as described previously (11).

### Construction of combinatorial mutant HEs

To generate an I-CreI coding sequence containing mutations derived from different libraries (28,30,33,38,40 and 44,68,70 or 44,68,70,75,77 amino acids), separate overlapping PCR were carried out that amplify the 5' end (residues 1–43) or the 3' end (residues 39–167) of the I-CreI coding sequence. For both the 5' and 3' end, PCR amplification is carried out using a primer specific to the vector (pCLS0542) (Gal10F 5'-GCAACTTTAGTGCTGACACATACAGG-3' or Gal10R

5'-ACAACCTTGATTGGAGACTTGACC-3') and a primer specific to the I-CreI coding sequence for 39–43 amino acids (assF 5'-CTAXXXTTGACCTTT-3' or assR 5'-AAAGGTCAAXXXTAG-3') where XXX codes for mutant residue 40. The resulting PCR products contain 15 bp of homology with each other and ~100–200 bp of homology with the two micron-based replicative vectors, pCLS542, marked with the *LEU2* gene and pCLS1107, containing a kanamycin resistant gene. Thus, to generate an intact coding sequence by *in vivo* homologous recombination, ~25 ng of each of the two overlapping PCR fragments and either 25 ng of the pCLS0542 vector DNA linearized by digestion with NcoI and EagI or 25 ng of the pCLS1107 vector DNA linearized by digestion with DraIII and NgoMIV were used to transform the yeast *S.cerevisiae* strain FYC2-6A (*MAT $\alpha$* , *trp1 $\Delta$ 63*, *leu2 $\Delta$ 1* and *his3 $\Delta$ 200*) using a high efficiency LiAc transformation protocol (27). For COMB targets, combinatorial mutants were generated individually, whereas for RAG targets, mutants were generated as libraries: PCR were pooled in equimolar amounts and transformed into yeast together with the linearized plasmid. Transformants were selected on either synthetic medium lacking leucine (pCLS542) or rich medium containing G418 (pCLS1107).

### Screening in yeast

For screening homodimers and heterodimers, we used either the protocol described previously (11), or a modified procedure wherein yeast mating occurs in liquid medium.

### Hierarchical clustering

Clustering was done using hclust from the R package. We used quantitative data from the secondary screening. Variants were clustered using standard hierarchical clustering with Euclidean distance and Ward's method (28). Mutant dendrogram was cut at the height of 17 to define the clusters. For the analysis (see e.g. Table 1) cumulated intensities of cleavage of a target within a cluster was calculated as the sum of the cleavage intensities of all cluster's mutants with this target, normalized to the sum of the cleavage intensities of all cluster's mutants with all targets.

### Biochemical and biophysical characterization of proteins

Novel I-CreI variants were expressed, purified and analyzed for *in vitro* cleavage as reported previously (11). Circular dichroism (CD) measurements were performed on a Jasco J-810 spectropolarimeter using a 0.2 cm path length quartz cuvette. Equilibrium unfolding was induced increasing temperature at a rate of 1°C/min (using a programmable Peltier thermoelectric). Samples were prepared by dialysis against 25 mM potassium phosphate buffer (pH 7.5), at protein concentrations of 20  $\mu$ M.

## RESULTS

### Functional endonucleases with new specificity towards $\pm 8$ , $\pm 9$ and $\pm 10$ nt

In a previous study we have reported the engineering of hundreds of new endonucleases with altered substrate

**Table 1.** Cluster analysis

Cluster (Effectif)	Preferred targets 10NNN	Preferred targets (%) <sup>b</sup>	Nucleotide			Preferred amino acids (%) <sup>a</sup>							
			-10 (%) <sup>b</sup>	-9 (%) <sup>b</sup>	-8 (%) <sup>b</sup>	28	30	33	38	40			
1 (44)	GGG	7.1	A	27.0	A	45.9	A	21.3	K	N	H	Q	S
	GAG	6.9	C	4.7	C	14.5	C	14.7	100	45.5	38.6	70.5	86.4
	GAT	6.4	G	63.0	G	27.9	G	37.0		K	R		
		Σ = 20.4	T	5.3	T	11.8	T	27.0		15.9	25.0		
2 (82)	AAG	6.1	A	33.4	A	52.2	A	20.4	K	N	G	Q	S
	TAG	5.6	C	11.7	C	9.4	C	13.8	100	64.6	23.2	68.3	79.3
	GAG	5.2	G	23.7	G	19.9	G	41.7					
		Σ = 16.9	T	31.2	T	18.5	T	24.0		15.9			
3 (36)	TAG	4.5	A	24.7	A	45.2	A	19.5	K	N	T	Q	S
	TAC	4.4	C	13.9	C	6.7	C	16.2	100	75.0	61.1	86.1	75.0
	TGG	4.3	G	15.8	G	26.9	G	37.6			C		
		Σ = 13.2	T	45.6	T	21.3	T	26.7			22.2		
4 (74)	GGG	30.6	A	33.1	A	22.6	A	10.2	K	N	R	R	S
	AGG	15.0	C	2.6	C	1.3	C	1.5	93.2	82.4	17.6	26.0	83.6
	AAG	7.6	G	56.3	G	66.2	G	77.7			Y	K	
		Σ = 53.2	T	8.1	T	9.9	T	10.6			16.2	19.2	
5 (115)	GAG	12.0	A	30.0	A	71.6	A	20.6	K	N	R	Q	S
	GAT	11.8	C	5.1	C	5.1	C	15.5	98.3	64.4	23.5	94.7	66.1
	GAA	8.8	G	58.5	G	18.4	G	35.6			H		
		Σ = 32.6	T	6.4	T	5.0	T	28.2			20.9		
6 (110)	AAG	9.1	A	40.6	A	59.9	A	15.7	K	N	P	Q	S
	TAG	8.7	C	9.6	C	11.0	C	12.2	100	87.3	22	68.8	61.5
	GAG	8.0	G	23.6	G	15.7	G	49.0					
		Σ = 25.8	T	26.2	T	13.4	T	23.2					
7 (106)	AAT	23.7	A	74.9	A	85.2	A	25.0	K	N	Y	Q	S
	AAA	16.8	C	17.2	C	2.5	C	11.3	86.8	41.5	92.5	95.2	70.5
	AAG	16.6	G	5.9	G	11.5	G	30.0		T			
		Σ = 57.1	T	2.0	T	0.8	T	33.7		24.5			
8 (384)	GGG	14.0	A	35.5	A	41.2	A	12.6	K	N	Y	Q	S
	TAG	10.2	C	12.4	C	9.1	C	12.2	89.8	63.3	45.8	43.4	62.9
	AAT	6.9	G	30.7	G	37.3	G	52.4					
		Σ = 31.1	T	21.4	T	12.4	T	22.8					
9 (134)	TAG	17.6	A	21.1	A	62.3	A	12.0	K	N	C	Q	S
	TAT	9.9	C	13.0	C	2.2	C	13.9	92.5	76.1	17.9	74.6	72.7
	AAG	7.5	G	9.6	G	22.1	G	51.3			S		
		Σ = 35.0	T	56.3	T	13.4	T	22.7			16.4		
10 (399)	AAG	20.2	A	64.3	A	78.9	A	18.1	K	N	Y	Q	S
	AAT	14.7	C	5.4	C	6.7	C	10.6	96.0	59.5	53.6	69.2	70.1
	AAA	10.7	G	25.2	G	10.4	G	44.8					
		Σ = 45.6	T	5.1	T	4.0	T	26.5					

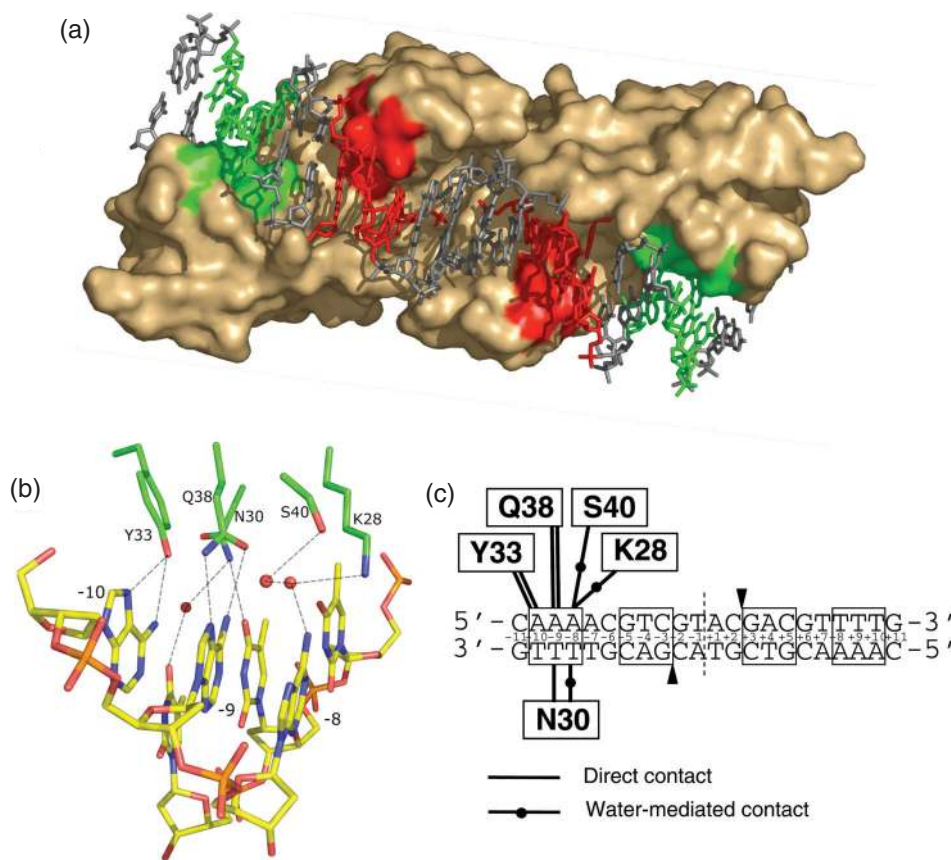
<sup>a</sup>In each position, residues present in >15% of the cluster are indicated.

<sup>b</sup>Target and base frequencies correspond to cumulated intensity of cleavage (see Materials and Methods).

specificities (11), derived from I-CreI, a dimeric protein that cleaves a 22 bp pseudo-palindromic target. The screening of libraries mutated at positions 44, 68 and 70 against the 64 palindromic DNA targets degenerated at ±3, ±4 and ±5 nt (5NNN, see Figure 1) resulted in the isolation and identification of numerous new HEs with novel specificities. Since then, we have generated other libraries with additional randomized residues, such as D75 and I77, and obtained hundreds of additional novel endonucleases targeting 5NNN target (P. Duchateau and F. Pâques, unpublished data).

In this report, we used the same approach to identify I-CreI derivatives with new substrate specificities towards positions ±8, ±9 and ±10 (10NNN) of C1221 (Figure 1), a palindromic 22 bp DNA target cleaved by I-CreI (29). Indeed, analysis of the I-CreI structure bound to its DNA target revealed that residues K28, N30, Y33, Q38 and S40 interact directly

or indirectly with the bases located at position ±8, ±9 and ±10 of the original I-CreI DNA target (Figure 1). In order to be consistent with a previous study (11), we first introduced a D75N mutation, to allow more diversity in positions 68 and 70, in the final combinatorial mutants. This mutation, which resulted in an altered substrate specificity on 5NNN targets (11), changes the specificity for 10NNN targets as well, with a dramatic narrowing of the cleavage pattern (Figure 2b). Randomization of 5 amino acid positions would lead to a theoretical diversity of  $20^5 = 3.2 \times 10^6$ . We chose to generate libraries with lower diversity by randomizing 2, 3 or 4 residues at a time, resulting in a diversity of 225 (152), 1728 (123) or 10000 (104), as described in Materials and Methods. This strategy allowed us to screen extensively each of these libraries against the 64 palindromic 10NNN DNA targets using a yeast based



**Figure 1.** Design of the libraries of I-CreI variants: rationale. (a) Structure of I-CreI bound to its DNA target, according to Chevalier *et al.* (29), and localization of the area of the binding interface chosen for randomization in this study (green). The binding interface mutated in a former report, and including residues Q44, R68 and R70 is also represented (red). In the combinatorial approach described below, we combined the regions represented in green and red. (b) Zoom showing residues 28, 30, 33, 38 and 40 chosen for randomization. (c) Summary of I-CreI–DNA interaction in the external region of the I-CreI DNA target (in green on Figure 1a). The target represented, C1221, is a palindromic target cleaved by I-CreI (29). Only base specific contacts are indicated. The 10NNN ( $\pm 8$ ,  $\pm 9$ ,  $\pm 10$  nt, in green on Figure 1a) and 5NNN ( $\pm 3$ ,  $\pm 4$ ,  $\pm 5$  nt, in red in Figure 1a) regions of the target are boxed.

assay described previously (25), and whose principle is described in Figure 2a.

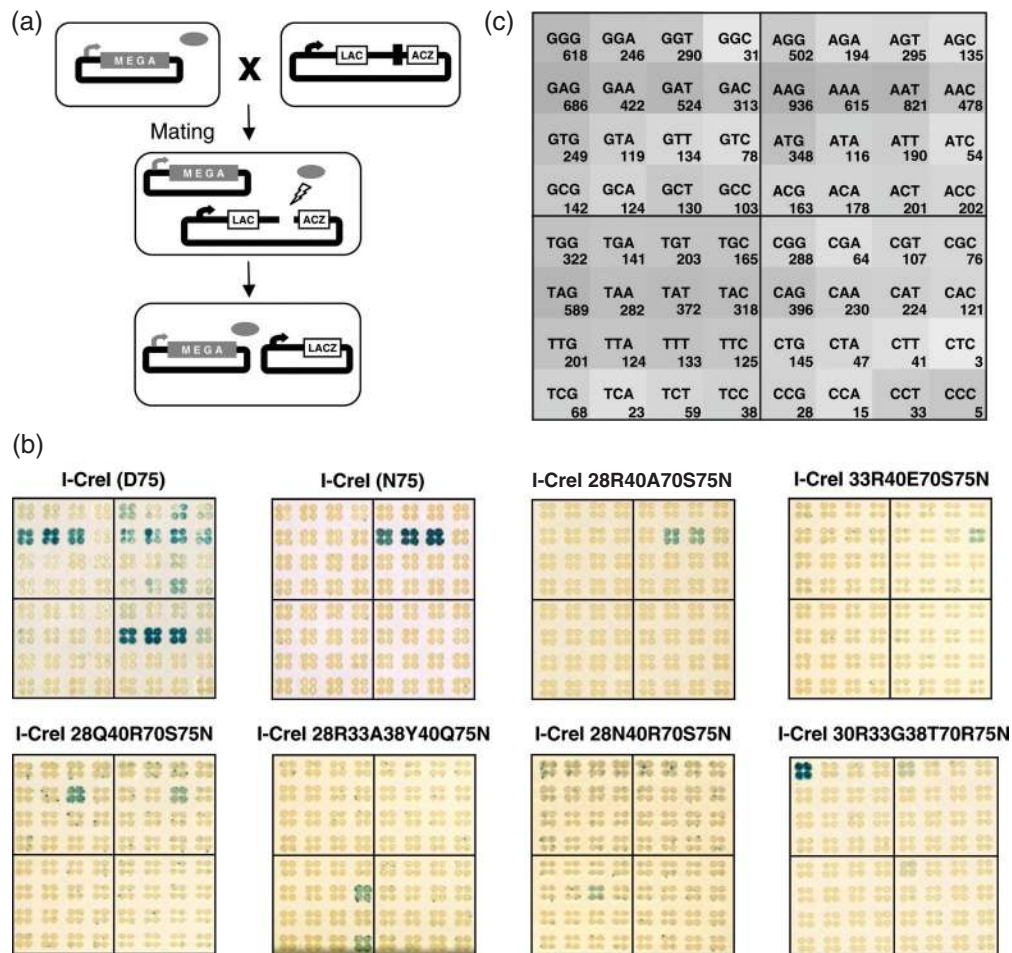
After secondary screening and sequencing of positives over the entire coding region, a total of 1484 unique mutants were isolated showing a cleavage activity against at least one target. Different patterns could be observed (Figure 2b). As shown previously for wild-type I-CreI or derived mutants, we found cleavage degeneracy for many of the novel endonucleases we identified, with an average of 9.9 cleaved targets per mutant (SD: 11). However, among the 1484 mutants identified, 219 (15%) were found to cleave only one DNA target, 179 (12%) cleave two, and 169 (11%) and 120 (8%) were able to cleave 3 and 4 targets, respectively. Thus, irrespective of their preferred target, a significant number of I-CreI derivatives display a specificity level that is similar if not higher than that of the I-CreI N75 mutant (3 10NNN target sequences cleaved), or I-CreI (16 10NNN target sequences cleaved), in accordance with previous observations with the 5NNN targets (11). Also, the majority of the mutants isolated for altered specificity for 5NNN and 10NNN sequences no longer cleave the original C1221 target sequence described in Figure 1c (61 and 59%, respectively).

Altogether, this large collection of mutants allowed us to target all of the 64 possible DNA sequences differing at

positions  $\pm 10$ ,  $\pm 9$  and  $\pm 8$ . However, there were huge variations in the numbers of mutants cleaving each target (Figure 2c), these numbers ranged from 3 to 936, with an average of 228.5 (SD: 201.5). Cleavage was frequently observed for targets with a G in  $\pm 8$  or A in  $\pm 9$ , whereas a C in  $\pm 10$  or  $\pm 8$  were correlated with low numbers of cleavers. In addition, all targets were not cleaved with the same efficiency. Since significant variations of signal could be observed for a same target, depending on the mutant (compare cleavage efficiencies for the wild-type 10AAA target in Figure 2b, for example), an average cleavage efficiency was measured for each target as reported previously (11). These average efficiencies are represented by gray levels on Figure 2c. Analysis of the results show a clear correlation between this average efficiency and the numbers of cleavers, with the most frequently cut target being also the most efficiently cut (e.g. compare 10TCN, 10CTN and 10CCN targets with 10GAN, 10AAN and 10TAN in Figure 2c).

#### Statistical analysis of interactions between I-CreI variants and their targets

In a previous report, we used hierarchical clustering to establish potential correlations between specific protein residues



**Figure 2.** Identification of novel I-CreI derivatives with locally altered specificity. (a) Yeast screening assay principle. A strain expressing the meganuclease (MEGA) to be assayed is mated with a strain harboring a reporter plasmid containing the chosen target. The target is flanked by overlapping truncated LacZ genes (LAC and ACZ). Upon target cleavage, tandem repeat recombination restores a functional LacZ gene, which can be monitored by standard methods. (b) Examples of profiling. Each novel endonuclease is profiled in yeast on a series of 64 palindromic targets, differing from the sequence shown in Figure 1c at positions  $\pm 8$ ,  $\pm 9$  and  $\pm 10$ . These targets are arrayed as in Figure 2c. As described previously (11), blue staining indicates cleavage. (c) Numbers of mutants cleaving each target, and average intensity of cleavage. Each sequence is named after the  $-10$ ,  $-9$ ,  $-8$  triplet (10NNN). The number of proteins cleaving each target is shown below, and the level of gray coloration is proportional to the average signal intensity obtained with these cutters in yeast.

and target bases (11). Using the same approach, we could identify 10 different mutant clusters (data not shown), described in Table 1. Analysis of the residues found in each cluster showed strong biases for all randomized positions. None of the residues is mutated in all libraries used in this study, and the residues found in the I-CreI scaffold were expected to be overrepresented. Indeed, K28, N30 and S40 were the most frequent residues in all 10 clusters, and we cannot really infer any conclusion for DNA-protein interactions. However, Y33 was the most represented residue only in clusters 7, 8 and 10, whereas strong occurrence of other residues, such as H, R, G, T, C, P or S, was observed in the seven other clusters. The wild-type Q38 residue was overrepresented in all clusters but one, R and K being more frequent in cluster 4.

Meanwhile, the occurrence of a specific residue at positions 33 could often be correlated with a strong preference for a specific base in position  $\pm 10$  of the DNA target. Prevalence of Y33 was associated with high frequencies of adenine (74.9 and 64.3% in clusters 7 and 10, respectively), and this

correlation was also observed, although to a lesser extent in clusters 4, 5 and 8. H33 or R33 were correlated with a guanine (63.0, 56.3 and 58.5%, in clusters 1, 4 and 5, respectively) and T33, C33 or S33 with a thymine (45.6 and 56.3 in clusters 3 and 9, respectively). G33 was relatively frequent in cluster 2, the cluster with the most even base representation in  $\pm 10$ . These results are consistent with the observations of Seligman and collaborators, who showed previously that a Y33R or Y33H mutation shifted the specificity of I-CreI toward a guanine and Y33C, Y33T, Y33S (and also Y33L) towards a thymine in position  $\pm 10$  (14). We also observed correlated biases for residue 38 and position  $\pm 9$  of the target: R38 and K38 were associated with an exceptional high frequency of guanine in cluster 4, while in all the other clusters, the wild-type Q38 residue was overrepresented, as well as an adenine in  $\pm 9$  of the target.

The structure of I-CreI bound to its target (29,30) has shown that Y33 and Q38 contact two adenines in  $-10$  and  $-9$  (Figure 1), and our results suggest that these interactions are probably maintained in many of our mutants. We have

previously described similar results for residue 44 and position  $\pm 4$  (11). However, when we compare the results obtained for the 33/ $\pm 10$ , 38/ $\pm 9$  and 44/ $\pm 4$  couples, significant differences are observed. For a guanine, we find mostly R and H in position 33, R or K in 38 and K in 44, for adenine, Y in 33 and Q in 38 and 44, and for thymine, S, C or T in 33 and A in 44. In the three cases, no clear pattern is observed for cytosine.

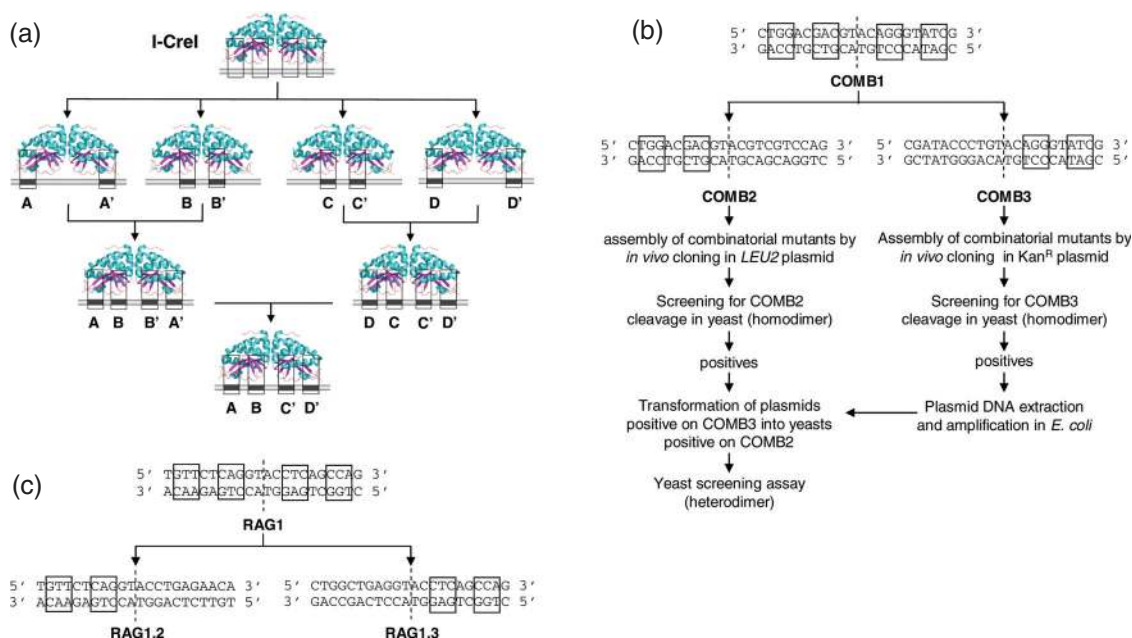
### Identification of distinct DNA-binding sub-domains within the same $\alpha\beta\beta\alpha\beta\beta\alpha$ fold

The identification of distinct groups of mutations in the I-CreI coding sequence that alter the cleavage specificity towards two different regions of the C1221 target sequence (10NNN and 5NNN) raises the possibility of combining these two groups of mutants intramolecularly to generate a combinatorial mutant capable of cleaving a target sequence simultaneously altered at positions 10NNN and 5NNN (Figure 3a).

To test this hypothesis, we first designed a model non-palindromic target sequence that would be a patchwork of four cleaved 5NNN and 10NNN targets. This target, COMB1, differs from the C1221 consensus sequence at positions  $\pm 3$ ,  $\pm 4$ ,  $\pm 5$ ,  $\pm 8$ ,  $\pm 9$  and  $\pm 10$  (Figure 3b). In addition, we designed two derived target sequences representing the left (COMB2) and right (COMB3) halves in palindromic form (Figure 3b). To generate appropriate I-CreI combinatorial mutants capable of targeting the palindromic targets, mutants efficiently cleaving the 10NNN and 5NNN part of each palindromic sequence were selected (Tables 2 and 3), and their characteristic mutations incorporated into the same coding sequence by *in vivo* cloning in yeast (see

Figure 3b and Materials and Methods). Basically, mutations at positions 28, 30, 33, 38 and 40 from mutants cleaving 10NNN targets were associated with mutations at position 44, 68 and 70 from mutants cleaving 5NNN targets. Throughout the text and Figures, combinatorial mutants for COMB sequences are named with an eight letter code, after residues at positions 28, 30, 33, 38, 40, 44, 68 and 70 (e.g. NNSRK/AAR stands for I-CreI 28N30N33S38R40K44A68A70R75N). Parental controls are named with a five letter or three letter code, after residues at positions 28, 30, 33, 38 and 40 (NNSRK stands for I-CreI 28N30N33S38R40K70S75N) or 44, 68 and 70 (AAR stands for I-CreI 44AQ68A70R75N).

Combinatorial mutants were then screened against the appropriate target sequence, COMB2 or COMB3, using our meganuclease-induced recombination assay in yeast. Among the 93 different I-CreI combined mutants screened with the COMB2 target, 29 (31%) were found to cleave the palindromic target site (Table 2 and Figure 4). Similarly, 210 combinatorial mutants were tested with the COMB3 target and 69 (33%) were found to be active (Table 3). Cleavage of both COMB2 and COMB3 is specific to the combinatorial mutant as each of the parents was unable to cleave the target sequence (Figure 4 and data not shown). In addition, while the parental mutants displayed efficient cleavage of the 5NNN and 10NNN target sequences, all combinatorial mutants but one displayed no significant activity for these sequences (Figure 4 and data not shown), or for the original C1221 sequence (data not shown). The only exception was NNSRR/ARS, which was found to faintly cleave the 5GAC target (Figure 4). These results indicate that combining mutations at positions 28, 30, 33, 38, 40 and 44,



**Figure 3.** Strategy for the making of redesigned HEs. (a) General strategy. A large collection of I-CreI derivatives with locally altered specificity is generated. Then, a combinatorial approach is used to assemble these mutants into homodimeric proteins, and then into heterodimers, resulting in a meganucleases with fully redesigned specificity. (b) Making of combinatorial mutants cleaving the COMB1 target: a workflow. Two palindromic targets are derived from the COMB1 targets, and homodimeric combinatorial mutants are designed to cleave these two targets. Positives are then coexpressed to cleave the COMB1 target. (c) The RAG1 series of target. Two palindromic targets are derived from RAG1.1. Then, a workflow similar to that described for the COMB series of target can be applied.

**Table 2.** Combinatorial mutants tested against the COMB2 target

Residues 44, 68 and 70 <sup>a</sup>	Residues 28, 30, 33, 38, 40 <sup>b</sup>		QNSRK
	NNSRK	NNSRR	
AAR	+	+	+
AGR			
AHR	+	+	+
AKR			
ANR			
AQR	+	+	+
ARA			
ARG			
ARH			
ARL			
ARN	+	+	+
ARR	+	+	+
ARS	+	+	
ART			
ASK			
ASR	+	+	+
ATR	+	+	+
NAR			
NHR			
NRA			
NRG			
NRR	+	+	+
NSR			
NTR			
PDT			
QRG			
QTR			
SRR			
SRS			
THR			
TRR	+	+	+

Combinatorial mutants are created by assembling mutations in 28, 30, 33, 38, 40, 44, 68 and 70 in an I-CreI N75 scaffold. Combinatorial mutants cleaving COMB2 are indicated by +.

<sup>a</sup>Mutations identified in I-CreI N75 variants cleaving the chosen 5GAC target.

<sup>b</sup>Mutations identified in I-CreI S70N75 variants cleaving the 10TGG chosen target.

68, 70 can give rise to functional endonucleases and thus confirm the hypothesis that the two regions of the protein can act independently.

### Biochemical and biophysical analysis of homodimeric combinatorial mutants

Four combinatorial mutants cleaving COMB2 or COMB3, and their corresponding parent mutants were analyzed *in vitro* in order to compare their relative cleavage efficiencies. As can be observed in Figures 5a–c, cleavage of the combined palindromic target sequences (COMB2 or COMB3) is specific to the combinatorial mutants since the two parent mutants were unable to cleave these sequences. In addition, while the parental mutants displayed efficient cleavage of the 5NNN and 10NNN target sequences, only one out of the four combinatorial mutants (NNSRK/ARR) displayed a faint activity on one of these targets, the others being totally inactive (data not shown). Thus, results from the yeast assay were confirmed *in vitro*. Importantly, the differences in activity levels between mutants were also consistent with the variations observed in yeast, and this congruency was further confirmed by the *in vitro* study of four additional mutants cleaving COMB3

(data not shown). Thus, the variations of signal observed in yeast are not due to differences in expression levels, but really reflect differences in binding/and or cleavage properties.

Finally, analysis of the structure and stability of this group of combinatorial mutants was performed using far-UV CD (Figure 5d), <sup>1</sup>H-NMR and analytical ultracentrifugation (data not shown). All the mutants are dimers and their secondary and tertiary structures (data not shown) as well as thermal denaturation curves (Figure 5d) are similar to that of the original I-CreI N75 protein, showing that engineering did not result in a significant alteration of the structure, folding or stability of these proteins.

### Co-expression of combinatorial mutants results in cleavage of chimeric target sites

To determine if combinatorial mutants could function efficiently as heterodimers, a subset of mutants capable of cleaving the palindromic sites COMB2 and COMB3 were co-expressed in yeast and assayed for their ability to cleave the chimeric site COMB1, corresponding to the fusion of the two half sites of the original targets (Figure 6a). As can be observed in Figure 6a, co-expression resulted in cleavage of the chimeric sequence COMB1 among all tested heterodimers. This activity appears to be specific to the heterodimers since each one of the mutants expressed alone displayed no detectable activity with the chimeric target site (Figure 6a). In general, co-expression of two mutants displaying strong activity for COMB2 and/or COMB3 will result in a higher level of activity for the chimeric site than a co-expression of two mutants displaying weak activity (e.g. compare KNHQ/S/KEG × NNSRK/ARR with QNRQR/KEG × NNSRK/ASR in Figure 6a).

Cleavage of the COMB1 target was also detected *in vitro* when the KNHQ/S/KAS and NNSRK/ARR purified proteins were incubated together with the COMB1 target in our conditions, while incubation of single protein did not give rise to any detectable cleavage activity (data not shown). However, the cleavage efficiency was extremely low, which might result from slow heterodimer formation *in vitro*. Indeed, Silva *et al.* could show that engineered derivatives from I-DmoI had to be coexpressed in *Escherichia coli* to form active heterodimers (31), and is not clear whether I-CreI homodimers can exchange subunits easily. Actually, we cannot exclude that low levels of cleavage could result from an alternative pathway, such as subsequent nicking by the two homodimers in solution, and we are currently investigating this issue.

Altogether, our results indicate that a combinatorial approach can generate artificial HEs capable of effectively cleaving chimeric target sites altered at position 10NNN and 5NNN.

### Redesigned HEs cleave a natural target sequence in the RAG1 gene

To analyze the effectiveness of a combinatorial approach for designing HEs for natural target sites, the human RAG1 gene was analyzed for potential sites compatible with mutants present in the 10NNN and 5NNN libraries. RAG1 has been shown to form a complex with RAG2 that is responsible

**Table 3.** Combinatorial mutants tested against the COMB3 target

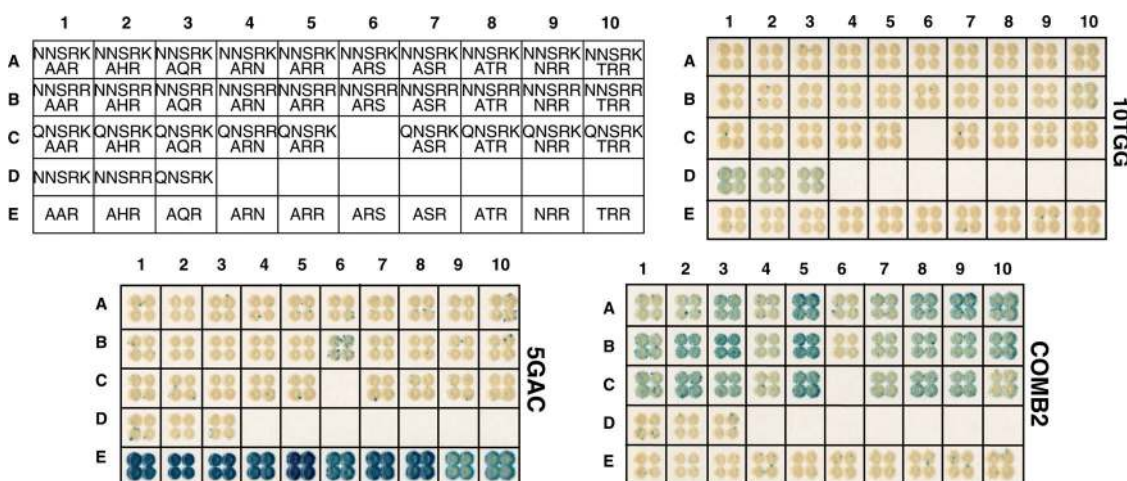
Residues 44, 68 and 70 <sup>a</sup>	Residues 28, 30, 33, 38 and 40 ANRQR <sup>b</sup>	KNRQA <sup>b</sup>	QNRQK <sup>b</sup>	QNRQR <sup>b</sup>	SNRQR <sup>b</sup>	TNRQR <sup>b</sup>	KNHQS <sup>c</sup>
AAK							
AGR							
ARD							
GQT							
HAT							
HRE							
KAD	+						
KAG							+
KAN	+						+
KAS	+	+	+	+	+	+	+
KDT							+
KEG	+						+
KES		+					+
KGT	+		+	+			+
KHD					+		+
KHN							
KHS							+
KND						+	+
KNN							
KNT	+			+	+		+
KQS		+					+
KRA	+			+	+		+
KRD	+	+	+	+			+
KRG	+				+	+	+
KRT	+	+	+	+		+	+
KST	+			+	+	+	+
KTD							
KTS	+		+				+
RAT							
SDK							

Combinatorial mutants are created by assembling mutations in 28, 30, 33, 38, 40, 44, 68 and 70 in an I-CreI N75 scaffold. Combinatorial mutants cleaving COMB3 are indicated by +.

<sup>a</sup>Mutations identified in I-CreIN75 variants cleaving the chosen 5CCT target.

<sup>b</sup>Mutations identified in I-CreI S70N75 variants cleaving the 10GAT chosen target.

<sup>c</sup>Mutations identified in an I-CreI N75 variant cleaving the 10GAT chosen target.

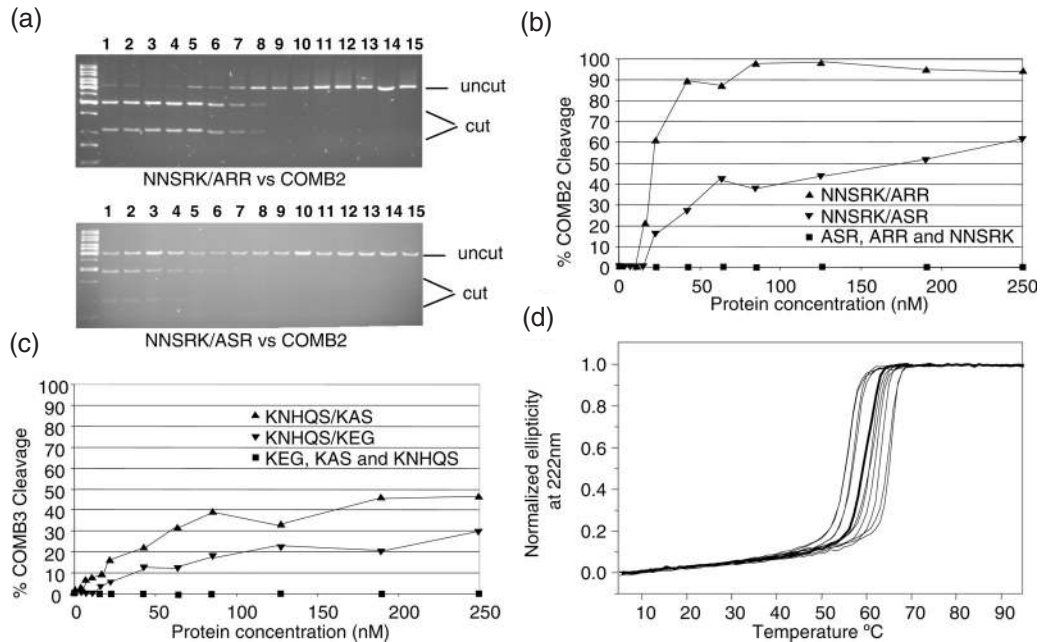


**Figure 4.** Secondary screening of combinatorial mutants cleaving COMB2. Upper panel: map of the mutants feature on the following panels. As described in text, combinatorial mutants are named with a eight letter code, after residues at positions 28, 30, 33, 38, 40, 44, 68 and 70 and parental controls with a five letter or three letter code, after residues at positions 28, 30, 33, 38 and 40 or 44, 68 and 70. Mutants are screened in yeast against COMB2 and 10TGC and 5GAC, the two parental targets.

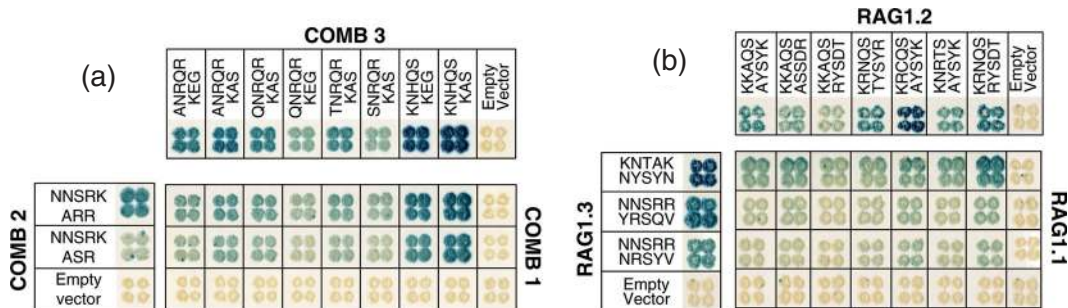
for the initiation of V(D)J recombination, an essential step in the maturation of immunoglobulins and T lymphocyte receptors (32,33). Patients with mutations in RAG1 display severe combined immune deficiency (SCID) due to the

absence of T and B lymphocytes. SCID can be treated by allogenic hematopoietic stem cell transfer from a familial donor and recently certain types of SCID have been the subject of gene therapy trials (34).





**Figure 5.** Biochemical and biophysical characterization of combinatorial mutants. (a) Examples of raw data for *in vitro* cleavage (see Materials and Methods). Different concentrations of proteins were assayed. Lanes 1 to 15: protein concentrations in nM are 250, 189.4, 126.3, 84.2, 63.2, 42.1, 21.1, 15.8, 10.5, 7.4, 4.2, 2.1, 1.0, 0.5 and 0. (b) Cleavage of COMB2 by combinatorial mutants. (c) Cleavage of COMB3 by combinatorial mutants. (d) Thermal denaturation of the same proteins measured by CD. The bold line corresponds to I-CreI N75, with a mid point denaturation temperature of 65°C. Other proteins: KNHQ/KAS (mid point denaturation temperature: 65.3°C), KNHQ/KAS (64.9°C), KEG (63.1°C),KNHQ (62.2°C), NNSRQ (61.2°C), KAS (61.2°C), KAS (61.2°C), ARR (57.3°C), ASR (57.1°C), NNSRK/ARR (55.8°C), NNSRK/ASR (55.8°C). For protein nomenclature, see Figure 4.



**Figure 6.** Cleavage of non-palindromic target by redesigned heterodimers. (a) Cleavage of COMB1 by heterodimers (lower right panel). Cleavage of COMB2 and COMB3 palindromic targets by the parent homodimers is indicated on the top and left panel. For combinatorial mutants, nomenclature is the same as for Figure 4 and in text. (b) Cleavage of RAG1.1 target by heterodimers. As described in text, combinatorial mutants are named after 10 residues instead of 8, corresponding to positions 28, 30, 33, 38, 40, 44, 68, 70, 75 and 77.

Analysis of the genomic locus of RAG1 revealed a potential target site located 11 bp upstream of the coding exon of RAG1, that we called RAG1.1 (Figure 3c). In contrast to the COMB sequence, the RAG1.1 site not only differs from the C1221 site at position 10NNN and 5NNN but also at 11N (11T instead of 11C) and 7NN (7CT instead of 7AC). I-CreI D75N is tolerant to these changes (data not shown), and we made the assumption that our combinatorial mutants would also be tolerant to changes at these positions. For the 5NNN region, we used mutants from the previously reported library mutated at positions 44, 68, 70 (11), as well as from another library mutated at positions 44, 68, 75 and 77, with a serine residue at position 70 (S. Grizot, P. Duchateau and F. Pâques, unpublished data). Since additional residues

were mutated, combinatorial mutants are named after 10 residues instead of 8, the two last letters corresponding to the residues at position 75 and 77 (e.g. KNTAK/NYSYN stands for I-CreI 28K30N33T38A40K44N68Y70S75Y77N).

In contrast with the mutants used for COMB targets, which were generated individually, mutants used for RAG targets were generated in libraries (see Materials and Methods). For the RAG1.2 target sequence, a library with a putative complexity of 1300 mutants was generated. Screening of 2256 clones yielded 64 positives (2.8%), which after sequencing, turned out to correspond to 49 unique endonucleases. For RAG1.3, 2280 clones were screened, and 88 positives were identified (3.8%), corresponding to 59 unique endonucleases. In both cases, the combinatorial mutants were

unable to cleave the 5NNN and 10NNN target sequences as well as the original C1221 sequence (data not shown).

As for COMB1, a panel of mutants able to cleave the palindromic targets was then co-expressed in the yeast to test the RAG1.1 target cleavage. Figure 6b shows that co-expression resulted in the cleavage of the natural target. In contrast, none of these mutants was able to cleave RAG1.1 when expressed alone (Figure 6b). We concluded that RAG1.1 target cleavage is due to the heterodimers resulting from co-expression.

## DISCUSSION

Altering the substrate specificity of DNA-binding proteins by mutagenesis and screening/selection is a difficult task. Several laboratories have relied on a semi-rational approach to limit the diversity of the mutant libraries to be handled (35): a small set of relevant residues is chosen according to structural data. This strategy was used to successfully engineer the substrate specificity of HEs from the LAGLIDADG family, such as PI-SceI SceI (10), I-CreI (14) and I-SceI (12). In a more elaborate approach, computational analysis based on energy calculation could be used to pinpoint key residues with good accuracy in I-CreI (11) and I-MsoI (9). Recently, we have combined semi-rational mutagenesis and high-throughput screening to conduct a large scale study that identified hundreds of I-CreI derivatives with altered substrate specificity (11). Three I-CreI residues were mutated simultaneously, and several novel targets were cleaved, differing from the I-CreI target by up to 3 bp. Nevertheless, this was still not sufficient to create redesigned endonucleases cleaving chosen sequences. In addition, it was clear that a global engineering of the I-CreI DNA-binding interface could not be achieved by a mere scale up of the approach. Analysis of the I-CreI/DNA crystal structure indicates that 9 amino acids make direct contacts with the homing site (29,36), which randomization would result in  $20^9$  combinations, a number beyond any screening capacity today. Thus, we hypothesized that it would be possible to first generate smaller libraries, to create novel endonucleases with locally altered substrate specificity, and then to combine these mutations into globally engineered mutants, in order to cleave chosen targets that widely differ from the I-CreI substrate (Figure 3a).

The first step was to create novel collections of I-CreI derivatives, by engineering another region of the DNA-binding domain, involved in the binding of the 10NNN target base pairs. Again, hundreds of novel variants were obtained, and the conclusions are very similar to the ones we obtained previously (11): mutants with novel substrate specificity can keep high levels of activity and the specificity of the novel proteins can be even narrower than that of the wild-type protein for its target. Second, strong correlations were observed between the nature of residues 33 and 38 and substrate discrimination at positions  $\pm 10$  and  $\pm 9$  of the target. We have previously reported similar results for residue 44 and position  $\pm 4$  (11). However, a given base can be correlated with different residues. For example, an adenine can be frequently bound by a Y or a Q, and a guanine by a H, a R or a K, depending on the position. Thus, there is no universal

'code', but rather a series of solutions for contacting each base, the best solution depending on a more general context, very similar to what has been observed with Zinc Finger proteins (15).

Our results, obtained by statistical analysis, fit those obtained by Seligman *et al.* (14) who determined the impact of single mutations at position 33, and observed the same preferences. In addition, the characterization of individual mutants has allowed for the identification of novel residue/base patterns for  $32/\pm 11$  and  $26/\pm 6$ , and it would be interesting to determine whether these studies can be confirmed by statistical analysis of a large number of mutants.

Since the generation of a series of I-CreI derivatives with altered specificity, as well as the assembly of such variants in heterodimers has been reported previously (11), the most challenging step was the identification of independent binding sub-domains within a same  $\alpha\beta\beta\alpha\beta\beta\alpha$  fold, and the assembly of mutations from such distinct sub-domains into combinatorial mutants. Structural analysis of I-CreI C33 and H33 mutants bound to their cognate DNA target have indicated that substitution of individual residue/base contact patterns can occur without significant structural deformation (37). However, other positions could be less tolerant. In addition, the cumulative impact of a series of mutations could eventually disrupt proper folding. Generation of combinatorial mutants for the COMB2 and COMB3 targets resulted in functional proteins with the expected specificity for  $\sim 30\%$  of the tested combinations, and as previously (11), *in vitro* tests on a subset of purified mutants confirmed the data from the yeast assay. In addition, NMR and CD data indicate that combinatorial mutants do not display significant alterations in structure or stability. Thus, the two sets of residues that we have mutated, K28, N30, Y33, Q38 and S40 on one hand, Q44, R68 and R70 on the other, define two relatively independent binding sub-domains. With the RAG1.2 and RAG1.3 targets, 2.8 and 3.8% of positives were obtained, respectively. In contrast with COMB mutants, which were generated and tested individually, RAG mutants were generated as libraries. Nevertheless, no obvious bias was detected in these libraries, and these frequencies should be representative of the real frequency of functional positives. This lower success rate, compared with screening with the COMB targets, could be due to the additional mutations at positions 75 and 77, or from the additional changes at positions  $\pm 6$ ,  $\pm 7$  and  $\pm 11$  in these targets. Nevertheless, the making of these combinatorial mutants opens large possibilities for it is the key step towards global engineering of the DNA-binding interface of LAGLIDADG proteins.

For genome engineering applications, the major advantage of HEs is their exquisite specificity (16), a feature that becomes essential when engaging into therapeutic applications. However, the use of HEs has so far been limited by the lack of means to engineer their DNA-binding domain. Although, several locally engineered endonucleases cleaving novel model target sequences have been reported before (9–12,14,37), this is the first time a HE is redesigned to cleave a naturally occurring sequence. Furthermore, these results validate a general method that can be applied to many other gene sequences. In former studies, the targets of the engineered proteins differed from the initial wild-type substrate by 1–6 bp per site (9–12,14,37), whereas the

22 bp COMB1 and RAG1 sequences differ from the C1221 target between 9 and 16 bp, respectively. The generation of collections of I-CreI derivatives allows today for cleavage of all 64 10NNN targets (this study) and 62 out of the 64 5NNN targets (P. Duchateau and F. Pâques, unpublished data). The ability to combine them intramolecularly as well as intermolecularly, increases the number of attainable 22mers to at least  $1.57 \times 10^7$   $[(64 \times 62)^2]$ . The ability to find variants cleaving targets modified at other positions, such as 11N, 7NN and 2NN, should further increase this number, to a point that remains to be determined.

Eventually, the ability to cleave any chosen sequence efficiently will probably require further developments. The recent development of computational biology might bridge the gap between the space of sequence that can be reached by pure experimental means, and genome complexity. Energy calculation could be used to predict the impact of local mutations on I-CreI (11) and I-MsoI (9). The same approaches could be applied to our combinatorial approach, and would allow one to discard non-functional combinations a priori, or suggest additional compensatory mutations that would abolish structural deformations. Therefore, we can envision a much wider use of HEs from the LAGLIDADG family to modify a large number of genes in a specific way.

## ACKNOWLEDGEMENTS

The authors thank Daniel Padró for the NMR analysis of the proteins, Collectis' platform for mutants screening, Collectis' bioinformatics for data handling, and Luis Serrano for critical reading of the manuscript. This work was partly supported by Direction Générale des Entreprises du Ministère de l'Industrie et des Finances (convention no. 05 2 90 604) and from the European Community Sixth Framework Programme (Contract 012948 NETSENSOR). Funding to pay the Open Access publication charges for this article was provided by Direction Générale des Entreprises.

*Conflict of interest statement.* None declared.

## REFERENCES

- Thierry,A. and Dujon,B. (1992) Nested chromosomal fragmentation in yeast using the meganuclease I-Sce I: a new method for physical mapping of eukaryotic genomes. *Nucleic Acids Res.*, **20**, 5625–5631.
- Choulika,A., Perrin,A., Dujon,B. and Nicolas,J.F. (1995) Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, **15**, 1968–1973.
- Rouet,P., Smih,F. and Jasin,M. (1994) Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol. Cell. Biol.*, **14**, 8096–8106.
- Puchta,H., Dujon,B. and Hohn,B. (1996) Two different but related mechanisms are used in plants for the repair of genomic double-strand breaks by homologous recombination. *Proc. Natl Acad. Sci. USA*, **93**, 5055–5060.
- 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.
- Chirazzi,M., Ray,A., Viret,J.F., Perera,R., Wang,X.H., Lloyd,A.M. and Signer,E.R. (1996) Enhancement of somatic intrachromosomal homologous recombination in *Arabidopsis* by the HO endonuclease. *Plant Cell*, **8**, 2057–2066.
- Hacein-Bey-Abina,S., Von Kalle,C., Schmidt,M., McCormack,M.P., Wulffraat,N., Leboulch,P., Lim,A., Osborne,C.S., Pawliuk,R., Morillon,E. *et al.* (2003) LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science*, **302**, 415–419.
- 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.
- 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.
- Gimble,F.S., Moure,C.M. and Posey,K.L. (2003) Assessing the plasticity of DNA target site recognition of the PI-SceI homing endonuclease using a bacterial two-hybrid selection system. *J. Mol. Biol.*, **334**, 993–1008.
- 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.
- 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.
- Steuer,S., Pingoud,V., Pingoud,A. and Wende,W. (2004) Chimeras of the homing endonuclease PI-SceI and the homologous *Candida tropicalis* intein: a study to explore the possibility of exchanging DNA-binding modules to obtain highly specific endonucleases with altered specificity. *Chembiochem*, **5**, 206–213.
- 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.
- Pabo,C.O., Peisach,E. and Grant,R.A. (2001) Design and selection of novel Cys2His2 zinc finger proteins. *Annu. Rev. Biochem.*, **70**, 313–340.
- Porteus,M.H. and Baltimore,D. (2003) Chimeric nucleases stimulate gene targeting in human cells. *Science*, **300**, 763.
- Urnov,F.D., Miller,J.C., Lee,Y.L., Beausejour,C.M., Rock,J.M., Augustus,S., Jamieson,A.C., Porteus,M.H., Gregory,P.D. and Holmes,M.C. (2005) Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature*, **435**, 646–651.
- Bibikova,M., Beumer,K., Trautman,J.K. and Carroll,D. (2003) Enhancing gene targeting with designed zinc finger nucleases. *Science*, **300**, 764.
- Bibikova,M., Golic,M., Golic,K.G. and Carroll,D. (2002) Targeted chromosomal cleavage and mutagenesis in *Drosophila* using zinc-finger nucleases. *Genetics*, **161**, 1169–1175.
- Lanio,T., Jeltsch,A. and Pingoud,A. (2000) On the possibilities and limitations of rational protein design to expand the specificity of restriction enzymes: a case study employing EcoRV as the target. *Protein Eng.*, **13**, 275–281.
- Voznyanov,Y., Konieczka,J.H., Stewart,A.F. and Jayaram,M. (2003) Stepwise manipulation of DNA specificity in Flp recombinase: progressively adapting Flp to individual and combinatorial mutations in its target site. *J. Mol. Biol.*, **326**, 65–76.
- Santoro,S.W. and Schultz,P.G. (2002) Directed evolution of the site specificity of Cre recombinase. *Proc. Natl Acad. Sci. USA*, **99**, 4185–4190.
- Buchholz,F. and Stewart,A.F. (2001) Alteration of Cre recombinase site specificity by substrate-linked protein evolution. *Nat. Biotechnol.*, **19**, 1047–1052.
- 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.
- 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.
- 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.
- Gietz,R.D. and Woods,R.A. (2002) Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. *Meth. Enzymol.*, **350**, 87–96.

28. Ward, J.H. (1963) Hierarchical grouping to optimize an objective function. *J. American Statist. Assoc.*, **58**, 236–244.
29. 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.
30. Jurica, M.S. and Stoddard, B.L. (1999) Homing endonucleases: structure, function and evolution. *Cell. Mol. Life Sci.*, **55**, 1304–1326.
31. Silva, G.H. and Belfort, M. (2004) Analysis of the LAGLIDADG interface of the monomeric homing endonuclease I-DmoI. *Nucleic Acids Res.*, **32**, 3156–3168.
32. Oettinger, M.A., Schatz, D.G., Gorka, C. and Baltimore, D. (1990) RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. *Science*, **248**, 1517–1523.
33. Schatz, D.G., Oettinger, M.A. and Baltimore, D. (1989) The V(D)J recombination activating gene, RAG-1. *Cell*, **59**, 1035–1048.
34. Fischer, A., Le Deist, F., Hacein-Bey-Abina, S., Andre-Schmutz, I., Basile Gde, S., de Villartay, J.P. and Cavazzana-Calvo, M. (2005) Severe combined immunodeficiency. A model disease for molecular immunology and therapy. *Immunol. Rev.*, **203**, 98–109.
35. Chica, R.A., Doucet, N. and Pelletier, J.N. (2005) Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. *Curr. Opin. Biotechnol.*, **16**, 378–384.
36. 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.
37. 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.