Context dependence between subdomains in the DNA binding interface of the I-CreI homing endonuclease

Sylvestre Grizot, Aymeric Duclert, Séverine Thomas, Philippe Duchateau* and Frédéric Pâques

CELLECTIS S.A., 102 Avenue Gaston Roussel, 93235 Romainville, France

ABSTRACT

Homing endonucleases (HE) have emerged as precise tools for achieving gene targeting events. Redesigned HEs with tailored specificities can be used to cleave new sequences, thereby considerably expanding the number of targetable genes and loci. With HEs, as well as with other protein scaffolds, context dependence of DNA/protein interaction patterns remains one of the major limitations for rational engineering of new DNA binders. Previous studies have shown strong crosstalk between different residues and regions of the DNA binding interface. To investigate this phenomenon, we systematically combined mutations from three groups of amino acids in the DNA binding regions of the I-CreI HE. Our results confirm that important crosstalk occurs throughout this interface in I-CreI. Detailed analysis of success rates identified a nearest-neighbour effect, with a more pronounced level of dependence between adjacent regions. Taken together, these data suggest that combinatorial engineering does not necessarily require the identification of separable functional or structural regions, and that groups of amino acids provide acceptable building blocks that can be assembled, overcoming the context dependency of the DNA binding interface. Furthermore, the present work describes a sequential method to engineer tailored HEs, wherein three contiguous regions are individually mutated and assembled to create HEs with engineered specificity.

INTRODUCTION

Homing endonucleases (HEs), also called meganucleases, are natural rare cutting endonucleases that can be used for genome engineering purposes. In nature, they are often encoded by ORFs found in Group I introns or located within inteins and propagate by a cleavage-repair mechanism (1). A DNA double-strand break (DSB) produced by the HE stimulates the incorporation of the mobile genetic element via homologous recombination. These properties constitute the basis of genome manipulations like gene correction, gene knock-out or gene insertion (2–4). In addition, the repertoire of cleavable DNA target sites has been considerably expanded by protein engineering means (5–12).

Zinc finger nucleases (ZFNs), another family of rare cutting endonucleases, have also been used in many genome engineering experiments. ZFNs are artificial proteins comprising of a DNA binding domain fused to the catalytic domain of the Type IIS FokI restriction enzyme (13,14). The DNA binding domain has a pseudo-modular structure consisting of a series of zinc fingers, each binding a distinct DNA triplet. However, this apparent modularity is complicated by crosstalk between adjacent fingers as shown by the refined structure of the Zif268 zinc finger protein (ZFP) bound to its DNA target (15). This could explain why engineering strategies based on pure combinatorial assembly of zinc finger units with pre-defined specificity display a high rate of failure (16). Consequently, several groups have designed engineering strategies that take into account the context dependence of zinc finger/DNA triplet interactions (17–20).

The use of a combinatorial approach relies on the characterization of separable functional domains. With HEs, however, the identification of such regions remains problematic. HEs are classified into five families, according to structural features and motif similarities (1). In most HEs, the catalytic center is embedded into the DNA binding domain, making it difficult to separate the two functions. Most HE engineering studies have focused on the LAGLIDADG family of proteins, which are the most specific of those characterized to date. LAGLIDADG

*To whom correspondence should be addressed. Tel: +33 1 41 83 99 17; Fax: +33 1 41 83 99 03; Email: duchateau@cellectis.com

proteins can be dimeric or monomeric, with structural data showing that monomeric LAGLIDADG HEs consist of two domains mimicking the architecture of their dimeric siblings. In both dimeric and monomeric proteins, each domain or monomer is organized around a unique core αββαββα-fold (21–26). Recognition specificity is largely conferred by interactions occurring in the DNA major groove, with the majority of the contacts arising from residues in the β-strands. Several groups have shown that it is possible to change the specificity of LAGLIDADG proteins by modifying this interaction pattern (8,10–12,27–30). Furthermore, different groups have shown that domains from different or identical LAGLIDADG proteins could be swapped, resulting in hybrid LAGLIDADG nucleases that include an I-DmoI and an I-CreI subunit (31–33), or two identical I-DmoI subunits (34). Nevertheless, it remains difficult to identify separable subdomains smaller than these subunits or monomers.

We have previously described a two-step combinatorial method for redesigning the DNA binding interface of the I-CreI HE. Two different regions within the I-CreI αββαββα-fold were characterized, wherein clustered mutations could be introduced to locally alter DNA binding specificity (9,27). In this study, these two regions will be referred to as D and P (distal and proximal to the catalytic center, respectively). By combining mutations from different locally altered I-CreI variants, we could produce fully redesigned HEs modified over their DNA binding interface to recognize chosen targets (8,9). The success of this approach validated the choice of the D and P regions for a combinatorial assembly process. However, a non-negligible failure rate (5,8) suggested that D and P were not fully independent. Indeed, in comparison to the wild-type I-CreI: DNA complex structure, movements of the protein backbone are observed in the structure of one of the proteins resulting from this process, illustrating the level of changes that can be induced by a few mutations in the DNA binding interface (7). This flexibility of the LAGLIDADG HEs DNA binding interface was confirmed by another study carried out on the I-MsoI protein (10).

The D and P regions used in our previous studies are as far from each other as possible and clearly do not overlap structurally (9,27). As such, one could expect that they behave more independently than adjacent regions of the DNA binding interface. To test this hypothesis, we identified a third region, termed M, located between D and P. In this study, we will refer to D, P and M as 'clusters', to account for the absence of any structurally or functionally separable subdomain in the αββαββα-fold. New I-CreI-derived proteins with mutations in the M cluster were generated and resulting mutants endowed with new DNA binding specificities were characterized. Statistical analyses of mutant patterns were performed as before for the P and D domains (9,27). The success rates of various approaches were then compared, including: (i) the simultaneous mutation of residues from the M and P clusters; (ii) a combinatorial approach involving the P and M regions; (iii) the combinatorial approach described in Smith *et al.* (9) involving the P

and D regions; and (iv) a combinatorial approach involving the P/M and D regions. Combined, the results show that throughout the I-CreI DNA binding interface, important functional crosstalk occurs with respect to DNA recognition specificity properties. Additionally, these experiments provide a sequential engineering method able to efficiently cope with this high level of global context dependence.

MATERIALS AND METHODS

Construction of target clones

DNA targets (22 bp length) used in the yeast screening assay and the mammalian SSA assay were inserted into reporter vectors using the Gateway protocol (Invitrogen): the yeast vector pFL39-ADH-LACURAZ and the mammalian vector pcDNA3.1-LAACZ. Both vectors contain an I-SceI target site as a control and have been previously described (27). Yeast reporter vectors were used to transform *Saccharomyces cerevisiae* strain FYBL2-7B (*MAT a, ura3Δ851, trp1Δ63, leu2Δ1, lys2Δ202*).

Meganuclease production

DNA sequences encoding the various meganucleases were amplified by the PCR and inserted into the 2 micron-based replicative vector pCLS542, which contains the *LEU2* gene for selective growth. The *S. cerevisiae* strain FYC2-6A (*MATα, trp1Δ63, leu2Δ1, his3Δ200*) was transformed with these vectors using a high-efficiency lithium acetate transformation protocol. For expression in mammalian cells, the various meganuclease genes were inserted into the mammalian vector pcDNA3.1 (Invitrogen) wherein protein expression is under the control of a CMV promoter.

Construction of mutant libraries, and screening in yeast

The different mutant libraries were generated by *in vivo* homologous recombination in yeast into the pCLS542 vector. PCRs were performed using primers containing degenerate codons (NNK or NVK) to randomize the desired I-CreI residues. Oligonucleotide sequences are available upon request.

Mating of meganuclease expressing clones and screening in yeast

A colony gridder (QpixII, Genetix) was used for the mating of yeast strains. Mutants were gridded on nylon filters placed on YP-glycerol plates, using high gridding density (~20 spots/cm²). A second gridding process was performed on the same filters for spotting of a second layer consisting of reporter-harboring yeast strains for each target. Membranes were placed on solid agar containing YP-glycerol-rich medium and incubated overnight at 30°C to allow mating. The filters were then transferred onto synthetic medium lacking leucine and tryptophan with glucose (2%) as the carbon source (and with G418 for co-expression experiments) and incubated for 5 days at 30°C to select for diploids carrying the expression and target vectors. Finally, filters were transferred onto

YP-galactose-rich medium for 2 days at 37°C to induce the expression of the meganuclease. Filters were then placed on solid agarose medium with 0.02% X-Gal in 0.5 M sodium phosphate buffer, pH 7.0, 0.1% SDS, 6% dimethyl formamide (DMF), 7 mM β-mercaptoethanol, 1% agarose and incubated at 37°C to monitor β-galactosidase activity. Filters were scanned and each spot was quantified using the median values of the pixels constituting the spot. We attribute the arbitrary values 0 and 1 to, respectively, white and dark pixels. β-Galactosidase activity is directly associated with the efficiency of homologous recombination. Any value >0 is considered as the consequence of cleavage.

Hierarchical clustering

Clustering of mutants was done using the activity profile of each mutant on the different 7NN targets. Mutants were classified in different groups using standard hierarchical clustering with Euclidean distance and Ward's method using hclust from the R package. We used the term of group instead of cluster, in order to avoid any confusion with the P, M and D clusters of amino acids described in this study. The mutant dendrogram obtained was cut at the height of 10 to define the groups.

Comparison of DNA cleavage profiles for M × P or D × P combined mutants

Each combined protein mutant was tested for cleavage activity against a panel of combined DNA targets: either 245 out of the 256 possible targets for M × P combinations, or a subset of the 10NNN–5NNN targets for D × P combinations that correspond to those targets inferred to have the highest probability of being cut considering the D and P mutations that were used for the combinatorial library. For each combined mutant, its profile was compared with that of the original modules on target modules. To facilitate comparison, we computed a theoretical profile of the combined mutant C on a given combined target $T_{i,j}$ by tracing back C to its original modules C' and C'' and the mutated region in $T_{i,j}$ to the two original targets T'_i and T''_j. The theoretical value was determined as $V_{theo}(C/T_{i,j}) = V(C'/T'_i) \times V(C''/T'_j)$, considering the two interaction regions to be independent. Theoretical and observed values were compared to assess the level of cluster independence.

Extrachromosomal assay in CHO-K1 cells

CHO-K1 cells were transfected with meganuclease expression vectors and the reporter plasmid in the presence of Polyfect transfection reagent in accordance with the manufacturer's protocol (Qiagen). Culture medium was removed 72 h after transfection and lysis/detection buffer was added for the β-galactosidase liquid assay. One liter of lysis/detection buffer contains: 100 ml of lysis buffer (10 mM Tris–HCl pH 7.5, 150 mM NaCl, 0.1% Triton X-100, 0.1 mg/ml BSA, protease inhibitors), 10 ml of 100X Mg buffer (100 mM MgCl$_2$, 35% 2-mercaptoethanol), 110 ml of an 8 mg/ml solution of ONPG and 780 ml of 0.1 M sodium phosphate pH 7.5. The OD$_{420}$ is measured after incubation at 37°C for 2 h.

The entire process was performed using a 96-well plate format on an automated Velocity11 BioCel platform. Conditions have been adapted so that the cell number and transfection efficiency are constant in all wells within the plates. To insure experimental reproducibility, control wells are included in all plates and correlation between intra- and interplate controls allows for assay validation.

Cell survival assay

The CHO-K1 cell line was used to seed plates at a density of 2.5×10^3 cells per well. The next day, varying amounts of meganuclease expression vector and a constant amount of GFP-encoding plasmid (10 ng) were used to transfect the cells with a total quantity of 200 ng DNA using Polyfect reagent. GFP levels were monitored by flow cytometry (Guava EasyCyte, Guava Technologies) on Days 1 and 6 post-transfection. Cell survival is expressed as a percentage and was calculated as a ratio (meganuclease-transfected cell expressing GFP on Day 6/control transfected cell expressing GFP on Day 6) corrected for the transfection efficiency determined on Day 1.

RESULTS

Definition and local engineering of the I-CreI M cluster

We defined three clusters within the I-CreI DNA binding interface that interact with distinct regions of the DNA target. The different clusters are designated as distal (D), median (M) and proximal (P) as they interact with the ±10 to ±8 (10NNN), ±7 to ±6 (7NN) and ±5 to ±3 (5NNN) nucleotides of the DNA target, respectively (Figure 1a). Whereas, D and P have been defined in previous studies, (8,9,27), M was defined in the present work, as a region bridging P and D.

The D cluster comprises of residues that belong to the loop bridging strands β1 and β2 and to the beginning of strand β2 (residues 29–40). Residues from the center of strands β1 and β2 constitute the M cluster (residues 25–28 and 41–43), while the P cluster includes residues from the end of the strand β2 and from strands β3 and β4 (residues 44–48 and 66–77) (Figure 1b). When viewed as a whole, protein residues of a given cluster can lie at the border between two contiguous clusters and hence, can play on the specificity for both DNA regions. We have previously shown that by mutating residues from the D and P clusters, it is possible to change the protein specificity for the 10NNN and 5NNN nucleotides of the DNA target (8,9,27).

We first created a collection of new I-CreI mutants with modified specificity toward the 7NN nucleotides by mutating residues belonging to the M cluster, as described for other clusters in previous works (9,11,27,35). Analysis of the wild-type I-CreI target (C1234 on Figure 1a) suggests that I-CreI can accommodate an A or a C in position ±7 and a C or a T in position ±7. Profiling of the wild-type I-CreI protein, using an yeast screening assay, demonstrated that I-CreI cleaves 10 out of the 16 7NN targets (Supplementary Figure S1). Analysis of the I-CreI/DNA co-complex structure shows that residues
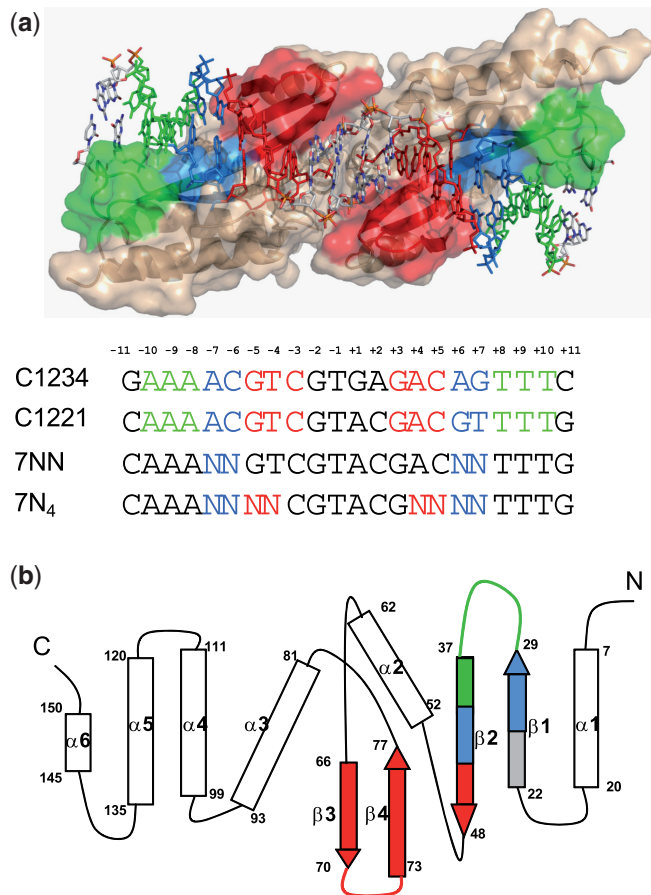
(a)


```
      -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 +1 +2 +3 +4 +5 +6 +7 +8 +9 +10 +11
C1234   GAAA ACGT C GTGA GAC AG TTT C
C1221   CAAA ACGT C GTAC GAC GT TTT G
7NN     CAAA NN GTCGTACGAC NN TTTG
7N4     CAAA NN NN CGTACG NN NN TTTG
```

(b)

Figure 1. The D, M and P clusters in I-CreI. (**a**) The D, M and P clusters have been respectively highlighted in green, blue and red on the I-CreI structure bound to its DNA target. The nucleotides of the wild-type C1234 target that interact with each cluster are represented in matching colors. C1221 is a palindromic C1234 derivate efficiently cleaved by I-CreI (22,27). Here, 7NN and 7N4 are palindromic targets, for which the nucleotides ±7 to ±6 and ±7 to ±4 have been, respectively, degenerated. (**b**) Schematic representation of the secondary structure of the I-CreI monomer. The same colour code has been used to represent amino acids belonging to the D, M and P clusters.

Lys28 and Gln26 make hydrogen bonds, respectively, with the bases at positions ±7 and ±6 of the I-CreI target. Residue Thr42, located on strand β3, is oriented toward the 7NN nucleotides. Exchanging Thr42 with an amino acid having a longer side chain could theoretically promote alternate interactions with the DNA. This three-residue cluster was, therefore, selected to modulate the specificity of I-CreI toward the ±7 and ±6 nt. I-CreI residues Gln26, Lys28 and Thr42 were randomized using the degenerate NVK codon to create the Ulib7NN mutant library. NVK codes for 15 amino acids, including polar and charged residues that can directly hydrogen bond or form water-mediated contacts with the DNA bases. Yeast clones (2232) constituting 16% of the theoretical (nucleic acid) library diversity were then screened against the 16 degenerate 7NN targets. After two screening steps and sequencing of positive clones, a final profiling of 300 distinct mutants on the 16 7NN targets was conducted. Figure 2a displays the Ulib7NN library hit-map showing
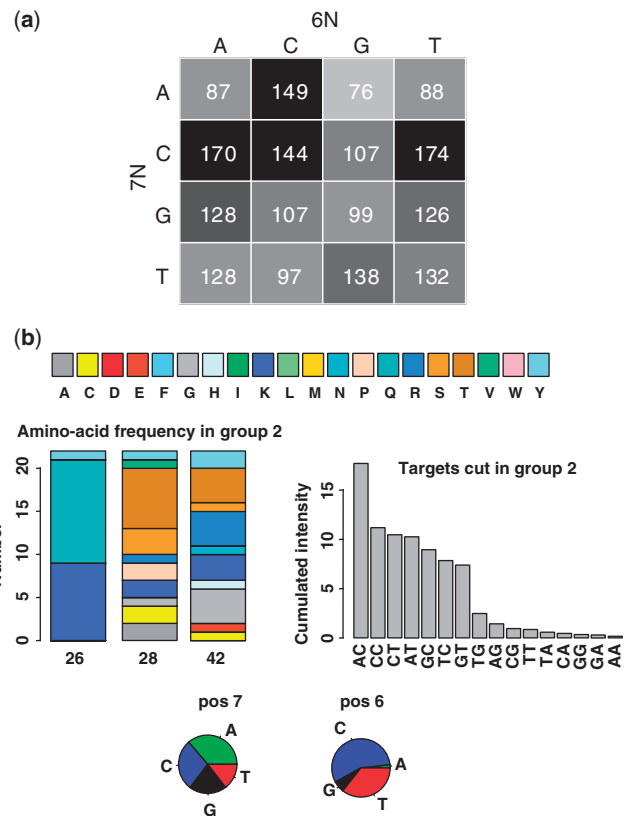
(a)

(b)

Figure 2. Generation of I-CreI mutants with new 7NN specificities. (**a**) Result of the Ulib7NN screening. For each of the 16 targets, the cumulated intensity on all available cutters was computed. This cumulated intensity was normalized to the maximum obtained on all the 16 targets. Each target was represented by an individual square with a black color for the maximum value of 1, a white color for a value of 0 and intermediate shades of gray for values between 0 and 1. Each target is located in a square according to its nucleotide at position 7 (row) and 6 (column). The figure on each individual cell indicates the number of cutters for this target. (**b**) A group of mutants identified by statistical analysis. Mutants profiles on the 16 7NN targets were analyzed by hierarchical clustering as indicated in the 'Material and Methods' section. The clustering tree was cut arbitrarily to define eight groups; Group 2 is represented here (other groups are described in Supplementary Figure S2). We used the term group instead of 'cluster' (the canonical term in this kind of study) for the sole purpose of avoiding any confusion between clusters of mutants and the P, M and D clusters of amino acids. For each group, the histogram on the left indicates the number of amino acids of each type found on each of the three positions 26, 28 and 42. The amino acid color code is indicated on the top of the figure. The histogram on the right displays for each of the 16 targets, the cutting intensity sum on all the mutants of the group, with targets sorted by decreasing intensities. From this histogram, one can calculate the total intensity on all targets having a defined nucleotide at position 7 or 6 and thus, for each nucleotide position, the relative total intensity for each of the 4 nt. The relative cleavage intensity for each of the 4 nt at position 7 or 6 is shown on the pies drawn on the far right of the figure.

that all the 16 targets can be efficiently cleaved. On average, one Ulib7NN mutant cleaves 6.5 7NN targets with a SD of 4.0, showing a specificity level equivalent if not better to that of the I-CreI targets, which cleaves 10 7NN targets. Furthermore, 28% of the mutants no longer cleaved the four targets best cleaved by I-CreI (7AC, 7AT, 7CT and 7CC).

Analysis of the mutant sequences revealed a high diversity of amino acids at each of the three randomized positions. To group mutants and perform a statistical analysis, a hierarchical clustering of the mutants was done on the basis of the activity profile on the 16 targets (see 'Materials and Methods' section). We could thus determine eight different groups of mutants according to the profile (Figure 2b and Supplementary Figure S2). In a few cases, relationships between the nature of amino acids present at a given position and the type of cleaved 7NN target that could initially be inferred from the protein/DNA complex structure were identified by this statistical analysis. Such relationship is most visible in Group 2, represented in Figure 2b. Indeed, the 22 mutants belonging to this group carry Gln and Lys residues at position 26 and cleave almost exclusively 7NC or 7NT targets. In Groups 1 and 4 (Supplementary Figure S2), a marked predominance of basic residues at position 28 can be noted and I-CreI-derived mutants belonging to these groups show a preference for cleaving 7CN or 7AN targets. In contrast, mutants belonging to Group 6 that display efficient cleavage of the 7GA target are less obvious to classify as these mutants present high amino acid diversity at positions 26, 28 and 42 (Supplementary Figure S2). Conversely, most mutants from Group 7 carry an Arg residue at position 42, but any correlation is made difficult by the diversity of the cleaved 7NN targets.

To determine the influence of the residues at positions 26, 28 and 42, the cleavage profiles of I-CreI variants carrying mutations on two of three randomized positions were determined (Supplementary Figure S3). Leaving Gln26 unchanged diminishes cleavage of the 7NA and 7NG targets. This is consistent with the structure of I-CreI in complex with its target since Gln26 interacts with the ±6 nt and I-CreI preferentially cleaves 7NC and 7NT targets. Similarly, a lysine at position 28 does not allow for the cleavage of 7GN and 7TN targets, as Lys28 interacts with the ±7 nt and I-CreI has a clear preference for an adenine or cytosine at this DNA target position. In contrast, keeping a threonine at position 42 does not limit the number of cleaved 7NN targets.

Direct generation of I-CreI variants carrying mutations in both M and P clusters

The Ulib7NN library screening yielded I-CreI-derived proteins carrying mutations in the M cluster with new specificities toward the 7NN nucleotides. To produce I-CreI mutants carrying mutations in both the M and P clusters, cleavage was assayed against DNA targets differing from the palindromic C1221 sequence (a target cleaved by I-CreI, see Figure 1), at positions ±7 to ±4. Such variants will be referred to as 7N$_4$ mutants below. Selection or phage display methods are often used for screening mutant libraries with large diversity. In contrast to selection procedures, our yeast assay has the advantage of allowing for simultaneously screening the cleavage activity of mutants against 256 different DNA targets. The specificity profile of each positive clone from the library is thereby immediately available. The Ulib7N$_4$ mutant library, in which residues 26, 28, 42, 44,

Figure 3. Ulib7N$_4$ library hitmap. For each 7N$_4$ target, we computed the total intensity of all available cutters and arbitrarily normalized this intensity sum to the maximum on all targets, as described in Figure 2. Each small square represents a target with a gray shade proportional to this normalized intensity sum (black for a value of 1, white for a value of 0). The sequence of the target can be obtained from the row (positions 7 and 6) and the column (positions 5 and 4) of the cell. The number within the cell indicates the number of individual cutters obtained (X when target has not been tested).

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
AA	0	0	1	0	13	0	0	16	7	0	0	8	7	0	0	3
AC	2	2	0	20	14	14	0	79	30	18	7	58	14	1	0	34
AG	0	0	0	0	0	0	0	5	1	3	0	12	1	0	0	15
AT	1	0	0	17	3	6	0	38	24	20	3	49	5	1	0	26
CA	0	0	0	0	93	4	14	109	11	1	0	12	31	0	1	30
CC	7	X	0	32	58	28	3	120	26	X	7	49	7	4	1	56
CG	0	0	X	0	17	1	0	36	5	0	0	34	9	0	0	32
CT	2	0	0	29	29	16	1	84	41	24	4	61	23	5	0	43
GA	0	0	0	0	29	2	4	30	8	X	1	12	4	0	0	5
GC	0	0	0	0	10	X	0	38	8	0	X	16	8	0	0	1
GG	1	0	X	X	4	0	0	5	X	3	0	11	9	0	0	8
GT	0	0	0	0	4	X	0	2	4	3	0	12	1	0	0	2
TA	0	0	0	0	47	X	2	49	12	1	0	15	13	0	0	5
TC	1	0	0	1	30	1	0	60	18	1	0	17	17	0	2	21
TG	0	0	1	0	11	0	0	42	4	2	1	19	17	1	0	15
TT	0	1	0	1	6	1	0	23	15	2	0	14	1	0	1	2

(rows labeled 7NN; columns labeled 5NN)

68 and 77 have been randomized, was generated in yeast. By using NVK codons for residues 26, 28 and 42 and NNK codons for residues 44, 68 and 77, this approach theoretically yields 452 984 832 nucleic acid sequences resulting in 27 000 000 different full-length proteins. The NNK codon was chosen for residues 44, 68 and 77 as it has been observed that hydrophobic residues excluded by the NVK codon can make beneficial van der Waals contacts with the ±5 and ±4 nt (data not shown). Primary screening against 245 out of the 256 possible DNA targets (11 targets were not obtained during the cloning step) was performed on 4464 clones (0.001% of the library diversity), assuming this would yield a reasonable percentage of positive variants. Positive clones were sequenced and 230 unique mutants were finally profiled for cleavage against these 245 7N$_4$_P DNA targets (Figure 3). On an average, one positive clone cleaves 10.5 DNA targets with a standard deviation of 10.6, with a total of 150 unique targets being cleaved by at least one mutant. In comparison to wild-type I-CreI, which itself can cleave 37 distinct 7N$_4$ targets (Supplementary Figure S4), Ulib7N$_4$ mutants appear hence to be more specific than the parental I-CreI scaffold.

Combination of mutations from the M and P DNA binding subdomains

While only a fraction of the Ulib7N$_4$ library was screened yielding 5.2% positive clones, this value is close to what we had already observed during the screening of less

Table 1. Examples of M × P combinatorial experiments

Combined target	Target cleaved during Ulib7N$_4$ screening	Modules used in combinatorial experiments (AU$_{mean}$)	Number and percentage positive hits from primary screening
7TATA_P	Yes	18 7TA_P cutters (0.50) 23 5TA_P cutters (0.82)	714 (64)
7TTCT_P	Yes	16 7TT_P cutters (0.58) 34 5CT_P cutters (0.94)	78 (7)
7GACT_P	Yes	15 7GA_P cutters (0.86) 34 5CT_P cutters (0.94)	815 (73)
7CAGC_P	No	21 7CA_P cutters (0.75) 33 5GC_P cutters (0.74)	20 (1.8)
7CAGG_P	No	21 7CA_P cutters (0.75) 23 5GG_P cutters (0.80)	141 (12.6)
7CACC_P	Yes	21 7CA_P cutters (0.75) 30 5CC_P cutters (0.70)	300 (26.9)
7TGGA_P	Yes	24 7TG_P cutters (0.52) 23 5GA_P cutters (0.86)	340 (30.5)
7TTCG_P	No	16 7TT_P cutters (0.58) 19 5CG_P cutters (0.39)	0
7TGAC_P	No	24 7TG_P cutters (0.52) 21 5AC_P cutters (0.53)	0

AU: arbitrary unit measuring the yeast cleavage efficiency (value between 0 and 1 with 0 = no cleavage and 1 = maximal cleavage).

Table 2. Protein sequences of 7N$_4$ mutants

Target	Mutant[a]	Q26	K28	T42	Q44	R68	I77
7GACT_P	U1	**T**	**T**	**S**	R	R	**K**
	U2	**C**	**S**	**R**	R	**S**	**C**
	CE1	**S**	T	**R**	R	**Y**	**N**
	CE2	T	T	**R**	K	**Q**	I
	CE3	T	T	**R**	K	**A**	**Q**
7CACC_P	U3	**C**	K	**R**	K	**V**	**F**
	CE4	**A**	**S**	**R**	K	**V**	**Y**
	CE5	T	K	**C**	K	**Y**	**W**
	CE6	**A**	**S**	**R**	K	**Y**	**W**
7TGGA_P	U4	**C**	T	**R**	**Y**	**S**	**H**
	U5	**R**	**S**	**C**	**Y**	R	I
	CE7	**C**	**R**	T	**V**	R	I
	CE8	**C**	**S**	T	**V**	R	**K**

Mutations are indicated in bold.
[a]U prefix indicates the mutant was obtained from the Ulib7N$_4$ library; CE prefix indicates the mutant was isolated after an M × P combinatorial experiment.

complex mutant libraries. Since screening a tiny fraction of the library diversity yielded a significant number of positive clones, this suggests that a large number of solutions probably exist for the same problem. To further validate the screening procedure, we checked whether the same results could have been obtained through a combinatorial process by assembling two groups of mutations from the M and P clusters and conferring respectively new 7NN and 5NN specificities. D × P combinatorial assays have already been achieved successfully by combining two mutation sets from the D and P clusters (9). Following the same principles, we made M × P combinatorial experiments. Although the previous Ulib7N$_4$ screening demonstrated the feasibility of altering the M × P region, the M and P proximity could affect the M × P combinatorial process and one could indeed expect a success rate for the M × P combination below that of the D × P combination. To confirm or invalidate this assumption, several M × P experiments were carried out. Combinatorial mutants were created by assembling mutations at positions 26, 28 and 42 located in the M cluster, responsible for the change in specificity toward the 7NN nucleotides, with mutations at positions 44, 68 and 77 located in the P cluster, which governs the specificity toward the 5NN nucleotides.

For each independent combinatorial experiment, 20 7NN cutters and 30 5NN cutters were selected, representing a theoretical protein diversity of 600 mutants. Mutant libraries covering ∼2-fold the expected diversity (1116 clones) were generated in yeast by *in vivo* cloning and screened against the appropriate combined targets, which are altered at positions 7N$_4$. Nineteen experiments were carried out and the results for nine of these are shown in Table 1. Eleven of the 19 trials yielded positive

mutants corresponding to a 58% global success rate for the M × P combinatorial process. The percentage of positive clones that were obtained for a successful combinatorial experiment varied (Table 1). In all the experiments, the positive and sequenced mutants were carrying mutations from the two mutation sets that were combined. The combination of mutations was hence required for target cleavage. With the exception of 7GACT, that could be cleaved by some 7GA cutters, parental mutants (either the 7NN or 5NN mutants) used for the combinatorial experiment were not able to cleave the combined target. This cleavage activity, however, can likely be explained by the I-CreI proteins ability to cleave the 5CTC target.

Three experiments used the same pool of mutants cleaving the 7CA targets. While the cleavage efficiency of the 5NN cutters used in each of these combinatorial assays was approximately the same, the rate of positive clones obtained in each combinatorial assembly ranges from 1.8% to 27%, strongly suggesting that the two clusters are not independent. The combinatorial process also yielded positive mutants toward two DNA targets (7CAGC and 7CAGG) that had not been cleaved during the Ulib7N$_4$ library screening. This result underlines the fact that more 7N$_4$ mutants could be obtained by pursuing a more extensive screen of the Ulib7N$_4$ library.

Further analysis was conducted to compare the mutants resulting from the two different processes described above. Table 2 lists sequences of I-CreI mutants able to cleave one of the three following targets: 7GACT, 7CACC and 7TGGA. The mutants were either isolated from the Ulib7N$_4$ library screening or obtained through an M × P combinatorial process. The two classes of mutants share the same characteristic amino acids at certain positions. For instance, it can be noted that mutants cleaving the 7CACC target carry long basic amino acids at positions 42 and 44 and large hydrophobic residues at position 77. Mutants cleaving the 7GACT target carry short polar residues at positions 26 and 28 and long basic amino acids at positions 42 and 44.

Table 3. Analysis of the M × P and D × P combinatorial experiments

	Targets tested	Successful trials	Success rate (%)
M × P experiments	19	11	58
D × P experiments	130	89	68

Comparison of the D × P and M × P combinatorial processes

To compare the level of interdependence among the three I-CreI clusters, we undertook a more precise analysis of both the D × P and M × P combinatorial processes. Table 3 shows the success rate for both types of experiments. Importantly, D × P combinatorial experiments were made with P cluster mutants modified at positions 44, 68, 70, 75 and/or 77 (and cleaving at least 1 of the 64 5NNN targets), as described previously (8,9,27). In contrast, P cluster mutants used in the P × M experiments were modified only at positions 44, 68 and 77, which govern the specificity toward the 5NN nucleotides (and cleave at least 1 of the 16 5NN targets). Although initially made to allow for better comparison with Ulib7N₄ library screening (see above), this difference could in theory introduce bias resulting in higher success rates for the P × M combinatorial process. The clusters considered in these experiments nonetheless display similar levels of dependence, with a slightly higher level of independence for D and P. Indeed, active meganucleases could be generated against 58% and 68% of the targets, corresponding to the M × P and D × P trials, respectively.

To go further into the details of the combinatorial process, we examined the success rates per individual mutant by comparing the theoretical cleavage profiles of M × P or D × P combined mutants to their observed cleavage profiles (Figure 4a–d). The theoretical cleavage profile of a combined mutant could be computed if the cleavage profiles of both parental mutants were available. This comparison allowed us to draw the distribution of the correlation coefficient between observed and theoretical profiles for each type of combination (Figure 4b). Independence of the two clusters involved in the combinatorial experiment would result in a distribution very close to one. Figure 4b shows two distributions centred on 0.3 and 0.7 for the M × P and D × P combinatorial processes, respectively. This indicates that neither the D nor the M clusters behave independently from the P cluster. Although the M × P and D × P combinatorial experiments do not use the same subset of P cluster mutants, any bias should rather result in a higher level of success in the M × P experiments (see above). The data suggest a lower level of crosstalk between D and P than between the contiguous M and P clusters.

If the different clusters were totally independent, a combined mutant would cleave a total number of targets equal to the product of the targets cleaved by the parental mutants. A low correlation can reflect the fact that the resulting combined mutant is more specific or less specific than one could expect from the cleavage profiles of the parental variants (Alternatively, it can indicate a lack of overlap between the expected cleavage pattern and the observed one). Actually, the majority of M × P combined mutants are more specific (cleaving less targets) than expected, with 96% of the mutants cleaving <50% of the expected targets. Representative examples are shown in Figure 4c. For the D × P combined mutants, which display a better correlation between the expected and observed cleavage patterns (Figure 4b), this trend toward increased specificity is less pronounced and only 41% of the mutants are cleaving <50% of the expected targets. Examples of profile are shown on Figure 4d, with the last mutant (H33Q40/I44M77) cleaving significantly more targets than expected (21 observed versus 15 expected).

The M × P combined mutants can be further engineered

Taking the 7N₄ mutants described above as a template, we investigated the possibility to further engineer new specificities toward the ±10 to ±8 (10NNN) nt of the C1221 target. As a proof of concept we considered mutants cleaving the 7GACT target to create a new specificity toward the 10TTG motif. A sequential combinatorial approach was adopted: rather than simultaneously combining two mutation sets as for the D × P or M × P combinatorial processes, a pool of 7N₄ mutants was taken as a template and several mutant libraries were created in yeast by mutating two or three residues using NVK codons for residues belonging to the D cluster. This process will be referred to as D × (MP) below. The screening of each library gave positive hits toward the 10TTGGACT target, which differs from the C1221 sequence at nucleotides ±10 to ±5. The most productive library was derived from nine 7GACT cutters by mutating the residues at positions 30, 33 and 38, yielding 400 positive clones among the 2232 clones constituting the library. Examples of cutters are listed in Table 4. To gain further insight into the feasibility of the D × (MP) process, we considered mutants able to cleave the 7CACC or the 7TGGA targets and tried to engineer for both groups of mutants new specificities toward the 10TGG, 10GAT, 10ATG and 10CCA sequences. These 10NNN sequences were chosen to have large nucleic acid diversity at each position. By mutating residues at positions 30, 33 and 38 using the NVK codon, two mutant libraries were created using either a pool of 10 7CACC cutters or a pool of 11 7TGGA cutters. For each library, 2232 clones representing ∼1.5% of the library nucleic acid diversity were screened against the four corresponding targets. Results of the library screening are listed in Table 5, while examples of positive mutants are displayed in Table 4. Positive clones for the eight tested targets were obtained (Tables 4 and 5). For the 10NNN7TGGA library, the number of positive hits almost equals the number of positive clones: one engineered mutant cleaves only one out of the four screened targets. Screening of the 10NNN7CACC library yielded mutants with apparently less specificity in that on average, one positive mutant cleaves 1.7 targets. This difference, however, is not
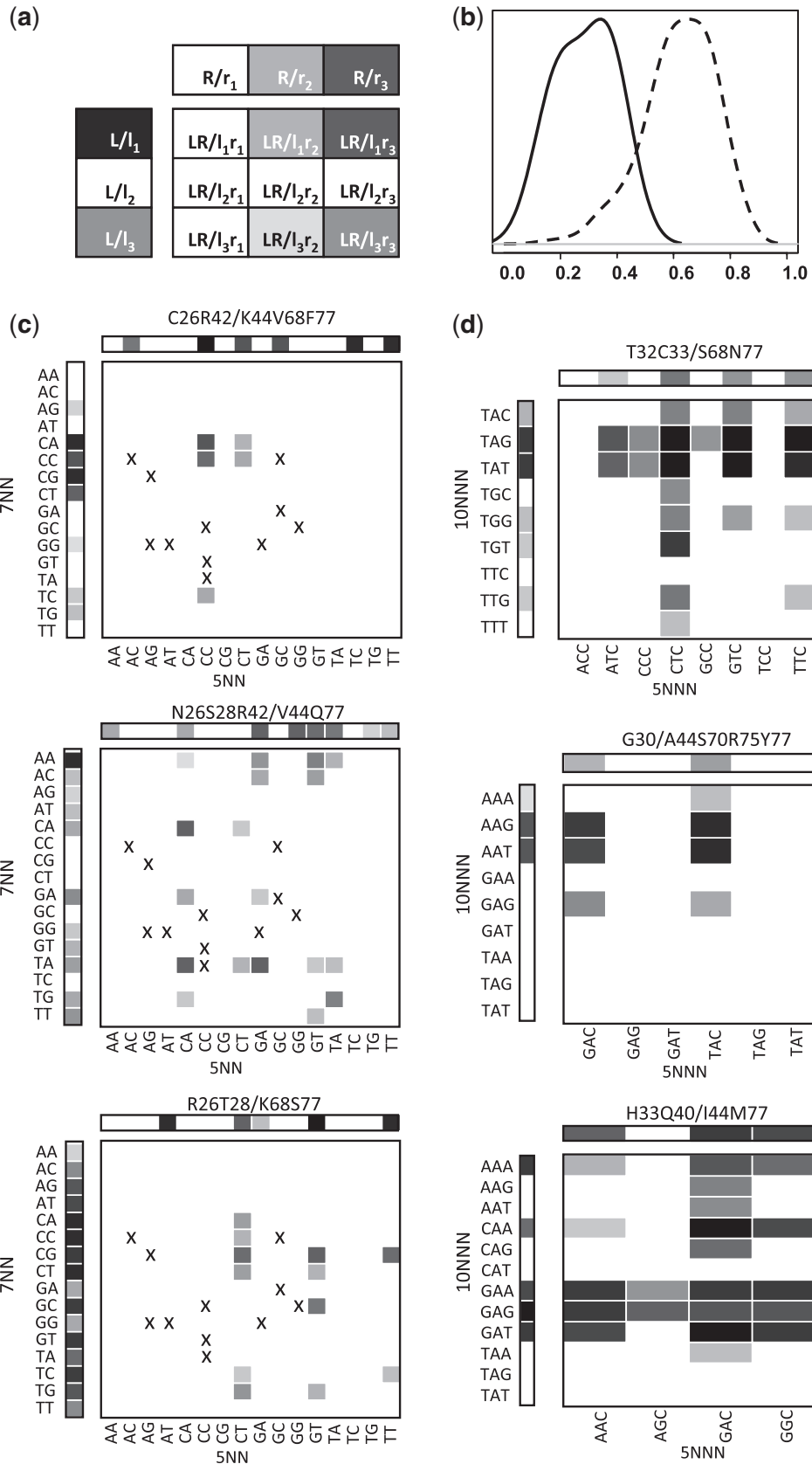
Figure 4. Statistical analysis of independence between clusters (**a**) Schema indicating theoretical profiles that would be obtained for completely independent clusters. A mutant is considered to have two clusters called L and R recognizing two different parts of the target called l and r. The profile of the mutant having only the L cluster mutated on targets mutated only in the corresponding l region is shown on the left (L/l_1, L/l_2 and L/l_3), with different gray levels indicating different cutting intensities (from white, for no activity, to black for maximum activity). Likewise, the

(continued)

Table 4. Protein sequences of some D × (MP) engineered I-CreI mutants

Cleaved target	Mutant	Q26	K28	N30	Y33	Q38	T42	Q44	R68	I77
10TTG-7GACT	Comb1	**T**	**T**	**S**	**C**	**S**	**R**	**K**	**Y**	**Q**
10TGG-7CACC	Comb2	**A**	**S**	**S**	**C**	**R**	**R**	**K**	**V**	**Y**
	Comb3	**T**	**K**	**E**	**C**	**R**	**C**	**K**	**Y**	**W**
10CCA-7CACC	Comb4	**A**	**S**	**D**	**C**	**Y**	**R**	**K**	**Y**	**W**
10GAT-7TGGA	Comb5	**C**	**R**	**S**	**R**	Q	T	**V**	R	I
	Comb6	**C**	**R**	**D**	**R**	Q	T	**V**	R	I
10TGG-7TGGA	Comb7	**C**	**S**	**H**	**C**	**R**	T	**V**	R	**K**

Mutations are indicated in bold.

Table 5. Screening results of two D × (MP) sequential combinatorial libraries

Library template	Clones screened	Positive clones	Hits	Positive hits per 10NNN module			
				10GAT	10TGG	10ATG	10CCA
7CACC	2232	161	267	31	75	91	70
7TGGA	2232	73	78	37	28	8	5

statistically significant as the screening concerned only 4 out of the 64 possible 10NNN sequences.

We have previously described a screen for I-CreI variants cleaving the 64 different 10NNN sequences, in the absence of any substitutions in the other parts of the target (9). The libraries of I-CreI variants we used were carrying substitutions on amino acids belonging to the D cluster (residues 33, 38 and 40) and at positions 28 and 70. We examined the sequences obtained from the library aimed at the 7CACC targets (only for this library the number of mutants was large enough to make a comparison) to find out if we could observe the same biases as previsouly noted in Smith *et al.* (9). While cutters of 10TGG-7CACC had a bias toward C33, S33 and T33 (58% overall), in accordance with the correlation observed between a T at position ±10 and these amino acids, cutters of 10ATG-7CACC (A at position ±10) were not very rich in wild-type Y33 (15%) contrary to what was previously observed, though in a different target context. Likewise, cutters of 10TGG-7CACC (G at position ±9) had a relatively important proportion of R38 and K38 (41% overall) in accordance with previous observations,

but cutters of 10GAT-7CACC (A at position ±9) were not very rich in wild-type Q38 (10%) in contrast to these observations. These data suggest that the relations found between allowed amino acids and target DNA largely depend on the target and/or protein context, which is different here from that in previous observations, in particular in the M and P regions and the corresponding interacting DNA positions (positions ±7 to ±4).

Although only two D × (MP) experiments are described in detail above, 11 D × (MP) trials were carried out and 45 targets were tested for cleavage. Thirty targets were found to be cleaved, a 66% success rate for the D × (MP) process. This value is similar to the success rates retrieved for the D × P and M × P methods (68% and 58%, respectively).

Characterizations of I-CreI mutants resulting from a sequential engineering method

In addition to providing a measurement of the crosstalk between D and MP, the D × (MP) process provides a sequential method to assemble engineered I-CreI derivatives. To validate this approach, seven I-CreI mutants

---

Figure 4. Continued

profile of a mutant having only the R cluster mutated on targets mutated only in the corresponding r regions is shown on the top (R/r_1, R/r_2 and R/r_3). For independent clusters, the chimeric protein LR would have a profile on the nine hybrid targets lirj similar to the one represented on the 9-cells square (LR/lirj). (**b**) Comparison of the correlation coefficient distributions between theoretical profiles and observed profiles for M and P combinations versus D and P combinations. Each type of combination (M-P or D-P) has been considered separately. Combined mutants have been assayed on combined targets and the observed profile has been compared to the theoretical profile computed as indicated in 'Materials and Methods' section. A Spearman's rank correlation coefficient between observed and theoretical profiles has been computed for each protein and the distribution of these coefficients has been drawn for each combination type, in solid line for 7NN–5NN combinations and in dashed line for 10NNN–5NN combinations. (**c**) Three examples of profiles for combined M-P mutants on combined 7NN–5NN targets compared to the profile of initial modules on target modules. As in Figure 3, the profile of the protein mutated only in the M cluster (26C42R for the first panel) on the 16 7NN targets is indicated vertically on the left. Similarly, the profile of the protein mutated only in the P cluster (44K68V77F for the first panel) on the 16 5NN targets is indicated horizontally along the top. The profile of the protein mutated in both clusters (26C42R44K68V77F for the first panel) is indicated on the 256-cells square representing all combinations of 7NN and 5NN targets. White indicates no activity, black indicates maximum activity and crosses indicate unknown values. (**d**) Three examples of the profiles of combined D-P mutants on combined 10NNN–5NN targets compared to the profile of initial modules on target modules. The representation is the same as in (c), except that 10NNN are represented vertically and 5NNN horizontally. Only a part of the 64 target modules or 64 × 64 target combinations have been assayed.
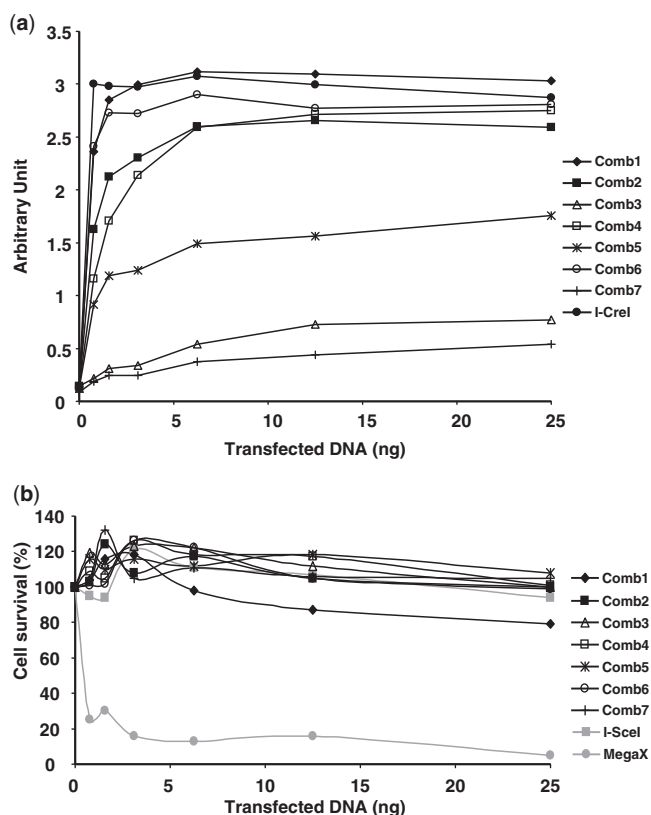
(a)

(b)

Figure 5. Activity and toxicity of D × (MP) I-CreI mutants monitored in CHO-K1 cells. (**a**) For the extrachromosomal SSA assay, cells were transfected with increasing amounts of meganuclease expression vectors (from 0 to 25 ng) and activity of the mutants was measured against their respective targets. (**b**) For the cell survival assay, CHO-K1 cells were co-transfected with increasing amounts of meganuclease expression vector and a constant amount of a plasmid coding for GFP. Cell survival is expressed as a percentage of cells still expressing GFP 6 days post-transfection.

carrying mutations in the three D, M and P clusters were further characterized.

The cleavage activity of these proteins against their respective targets was monitored using an extrachromosomal SSA assay and their potential toxicity was assessed using a cell survival assay. Both assays were conducted in mammalian CHO-K1 cells and carried out using the same transfection conditions (Polyfect reagent and 96-well plate format) using the same quantities of meganuclease expression vector. A toxicity pattern can hence be directly correlated to an activity level. Wild-type I-CreI was used as a positive control for activity in the SSA assay. Wild-type I-SceI was used as a non-toxic control in the cell survival assay and MegaX, a meganuclease with documented relaxed specificity as a control (7). The extrachromosomal SSA assays show that all the mutants, whose sequences are given in Table 4, are active in mammalian cells with some mutants even matching the I-CreI level of activity (Figure 5a). Comb5 and Comb6 mutants, both cleave the 10GATTGGA target and differ only by the nature of amino acid at position 30: Comb6 with Asp30 is more active than Comb5 with Ser30. Notably, these mutants have not undergone an optimization step as

previously described (8). The mutants tested for activity were then subjected to a cell survival assay. Figure 5b shows that the D × (MP) I-CreI engineered mutants do not display any visible toxicity as they present the same profile as I-SceI.

Further studies would be required to extensively characterize the endonucleases described in this study, including expression studies and the monitoring of efficacy on a chromosomal substrate. Nevertheless, the congruence of our yeast and CHO results confirm the high level of activity and specificity of these proteins. Altogether, our data illustrate how mutations can be introduced sequentially into the I-CreI protein to create a variant that cleaves a 22 bp DNA target differing in ~14 nt from the wild-type sequence. These modifications of the protein specificity can be achieved while maintaining a high activity level. In addition, specificity has been enhanced as no toxic behaviour could be detected.

DISCUSSION

Context dependence in protein/DNA interaction patterns is one of the major limitations to rational engineering of new DNA binders, in particular when binding is coupled with an additional desired function such as cleavage. An extreme example is the DNA binding interface of restriction enzymes, with each base of the target usually being saturated in contacts to protein residues that can themselves interact with different bases. The specificity of these proteins has proved very difficult to engineer, with at best only limited success (36–40). ZFPs, another specific DNA binder yet having a less dense network of protein/DNA molecular interaction, have proven easier to modify. One of the keys to success was the nearly modular assembly of individual engineered zinc fingers into 'polydactyl' DNA-binding domains (41). However, the presence of crosstalk between adjacent zinc fingers (42) was sufficient to substantially affect the success rate of advanced combinatorial experiments.

HEs from the LAGLIDADG family emerge as a different paradigm. In this study and previous ones (7–9,27), we have identified partially independent clusters that can be used in a combinatorial assembly process, while displaying very strong crosstalk with respect to DNA recognition specificity properties. Importantly, for each defined region (including the new M cluster identified in this study), we originally chose to simultaneously mutate several protein residues within the same cluster to account for the context dependence of the interaction set (9,27). This strategy agrees with what is called concerted design, published in a recent work, where the authors have shown that it was more advantageous to simultaneously design several mutations to change the I-MsoI specificity toward one DNA triplet (10). Importantly, in this study, statistical analysis of the M cluster identified groups of mutants who had kept DNA/protein interaction patterns identical or similar to the ones found in the wild-type protein/target complex, but failed to identify a general 'code' for DNA/protein interaction patterns. These observations are largely consistent with the ones made previously for the

P (27) and D (9) clusters [although a correlation could be found between residues at protein position 44 and bases at position ±4 of the target (27)] and are in agreement with a strong context dependence at the cluster level.

However, context dependence is also a more global phenomenon, with broad scaffold movements affecting large parts of the archetypical αββαββα–fold, as highlighted by previous structural studies of redesigned HEs (7) and this more global context explains why whole clusters are only partially independent, even when they are not adjacent. Other phenomena could lead to such observations. The assembly of different mutation sets could result in protein mis-folding or protein insolubility. One can also imagine that the combined protein could bind to the DNA but would not be competent for DNA cleavage. To assess these different possibilities and their respective impact on context dependence, experiments will be necessary to monitor folding and DNA binding of combined I-CreI variants.

A detailed analysis carried out at the mutant level for the different combinatorial processes illustrates that the level of dependency is stronger for contiguous clusters than for more distant ones (Figure 4). However, this expected nearest-neighbour effect is sufficiently subtle to be not seen when considering the success rate per combinatorial experiment (instead of per combinatorial mutant): in this case, the three combinatorial processes that were tested had almost the same success rates, as we could engineer specific meganucleases toward 58–68% of the considered targets. Thus, context dependence can be overcome during the protein engineering process. Furthermore, whereas a lack of independence between different subdomains or mutation clusters can make the engineering process more complex, it could be an advantage, if one wants to preserve a high level of protein specificity. Indeed, it could prevent the independent binding of proteins subunits and thereby, impair promiscuous cleavage. This impact on specificity can be a crucial parameter if we consider using these enzymes for therapeutic purposes.

It should be noted that different clusters to those used here could have been defined. Other breakdowns of the DNA binding interface have been identified for the combinatorial assembly of I-CreI engineered derivatives (S. Grizot, personnal communication). Clusters can also vary in size: the different combinatorial trials that were carried out confirmed the validity of a strategy targeting large libraries such as Ulib7N$_4$, covering residues from both the M and P clusters. Thus, the identification of independent subdomains at the structure or function level is not a necessary condition *per se* to envision a combinatorial process. Furthermore, it is likely that this notion of 'relativity' regarding the importance of clearly established and separable subdomains can be extended to the engineering of many other proteins.

Indeed, DNA shuffling techniques have already been used to combine sets of mutations in a structurally 'blind' way. For example, a tailored recombinase recognizing a sequence from HIV1 was engineered by the shuffling of mutations from mutants recognizing locally altered recombinase targets (43). The strategy adopted in that study is very close to the one we have described here and in former studies (8,9), with: (i) the generation of different sets of mutants recognizing targets altered in different regions and (ii) the combining of these mutants into redesigned proteins cleaving an *a priori* chosen target. However, in the case described by Sarkar *et al.* (43), no subdomain or cluster could be easily identified in a rational way (despite of the existence of structural data) and the authors chose a totally random method of mutagenesis.

Beyond addressing the basic question of interface cross-talk, this study provides a sequential method to deal with context dependence. Further studies will be necessary to compare the outcome of the sequential and non-sequential combinatorial methods. A sequential selection protocol has also been used with significant success to redesign ZFP specificity (17). Nevertheless, context dependence seems to be one of the most widespread concepts in the engineering of DNA binding proteins and it is not surprising that similar solutions are adopted for such different scaffolds as homing endonucleases and ZFN.

Recently, a new family of DNA binders with unique properties has been identified in plant pathogens and are actively being investigated (44,45). These transcription activator-like effectors (TALEs) bind DNA through a series of imperfect repeats of 34 and 35 amino acid residues. Each repeat contains a dipeptide sequence that governs the specificity for one base from the DNA target in an apparent autonomous manner (44,45), thus making specificity prediction possible. This simple code for recognition has already prompted several laboratories to use TALE proteins as scaffold to engineer a new class of tailored rare-cutting endonucleases (46–48). It will be interesting to see whether further studies will confirm the autonomy of each 'cipher' in base recognition, or identify context dependence in these proteins as well.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. Stoddard,B.L. (2005) Homing endonuclease structure and function. *Q. Rev. Biophys.*, **38**, 49–95.
2. Rouet,P., Smih,F. and Jasin,M. (1994) Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. *Proc. Natl Acad. Sci. USA*, **91**, 6064–6068.

3. 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.

4. Paques,F. and Duchateau,P. (2007) Meganucleases and DNA double-strand break-induced recombination: perspectives for gene therapy. *Curr. Gene Ther.*, **7**, 49–66.

5. Galetto,R., Duchateau,P. and Paques,F. (2009) Targeted approaches for gene therapy and the emergence of engineered meganucleases. *Expert Opin. Biol. Ther.*, **9**, 1289–1303.

6. Grizot,S., Smith,J., Daboussi,F., Prieto,J., Redondo,P., Merino,N., Villate,M., Thomas,S., Lemaire,L., Montoya,G. *et al.* (2009) Efficient targeting of a SCID gene by an engineered single-chain homing endonuclease. *Nucleic Acids Res.*, **37**, 5405–5419.

7. Redondo,P., Prieto,J., Munoz,I.G., Alibes,A., Stricher,F., Serrano,L., Cabaniols,J.P., Daboussi,F., Arnould,S., Perez,C. *et al.* (2008) Molecular basis of xeroderma pigmentosum group C DNA recognition by engineered meganucleases. *Nature*, **456**, 107–111.

8. Arnould,S., Perez,C., Cabaniols,J.P., Smith,J., Gouble,A., Grizot,S., Epinat,J.C., Duclert,A., Duchateau,P. and Paques,F. (2007) Engineered I-CreI derivatives cleaving sequences from the human XPC gene can induce highly efficient gene correction in mammalian cells. *J. Mol. Biol.*, **371**, 49–65.

9. Smith,J., Grizot,S., Arnould,S., Duclert,A., Epinat,J.-C., Prieto,J., Redondo,P., Blanco,F.J., Bravo,J., Montoya,G. *et al.* (2006) A combinatorial approach to create artificial homing endonucleases cleaving chosen sequences. *Nucleic Acids Res.*, **34**, e149.

10. Ashworth,J., Taylor,G.K., Havranek,J.J., Quadri,S.A., Stoddard,B.L. and Baker,D. (2010) Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. *Nucleic Acids Res.*, **38**, 5601–5608.

11. 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.

12. Gao,H., Smith,J., Yang,M., Jones,S., Djukanovic,V., Nicholson,M.G., West,A., Bidney,D., Falco,S.C., Jantz,D. *et al.* Heritable targeted mutagenesis in maize using a designed endonuclease. *Plant J.*, **61**, 176–187.

13. Porteus,M.H. and Carroll,D. (2005) Gene targeting using zinc finger nucleases. *Nat. Biotechnol.*, **23**, 967–973.

14. Carroll,D. (2008) Progress and prospects: zinc-finger nucleases as gene therapy agents. *Gene Ther.*, **15**, 1463–1468.

15. Elrod-Erickson,M., Rould,M.A., Nekludova,L. and Pabo,C.O. (1996) Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. *Structure*, **4**, 1171–1180.

16. Ramirez,C.L., Foley,J.E., Wright,D.A., Muller-Lerch,F., Rahman,S.H., Cornu,T.I., Winfrey,R.J., Sander,J.D., Fu,F., Townsend,J.A. *et al.* (2008) Unexpected failure rates for modular assembly of engineered zinc fingers. *Nat. Methods*, **5**, 374–375.

17. Greisman,H.A. and Pabo,C.O. (1997) A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science*, **275**, 657–661.

18. Isalan,M., Klug,A. and Choo,Y. (2001) A rapid, generally applicable method to engineer zinc fingers illustrated by targeting the HIV-1 promoter. *Nat. Biotechnol.*, **19**, 656–660.

19. Maeder,M.L., Thibodeau-Beganny,S., Osiak,A., Wright,D.A., Anthony,R.M., Eichtinger,M., Jiang,T., Foley,J.E., Winfrey,R.J., Townsend,J.A. *et al.* (2008) Rapid 'open-source' engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol. Cell*, **31**, 294–301.

20. Maeder,M.L., Thibodeau-Beganny,S., Sander,J.D., Voytas,D.F. and Joung,J.K. (2009) Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays. *Nat. Protoc.*, **4**, 1471–1501.

21. Chevalier,B.S., Monnat,R.J. Jr and Stoddard,B.L. (2001) The homing endonuclease I-CreI uses three metals, one of which is shared between the two active sites. *Nat. Struct. Biol.*, **8**, 312–316.

22. 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.

23. Bolduc,J.M., Spiegel,P.C., Chatterjee,P., Brady,K.L., Downing,M.E., Caprara,M.G., Waring,R.B. and Stoddard,B.L. (2003) Structural and biochemical analyses of DNA and RNA binding by a bifunctional homing endonuclease and group I intron splicing factor. *Genes Dev.*, **17**, 2875–2888.

24. Moure,C.M., Gimble,F.S. and Quiocho,F.A. (2003) The crystal structure of the gene targeting homing endonuclease I-SceI reveals the origins of its target site specificity. *J. Mol. Biol.*, **334**, 685–695.

25. Marcaida,M.J., Prieto,J., Redondo,P., Nadra,A.D., Alibes,A., Serrano,L., Grizot,S., Duchateau,P., Paques,F., Blanco,F.J. *et al.* (2008) Crystal structure of I-DmoI in complex with its target DNA provides new insights into meganuclease engineering. *Proc. Natl Acad. Sci. USA*, **105**, 16888–16893.

26. Silva,G.H., Dalgaard,J.Z., Belfort,M. and Van Roey,P. (1999) Crystal structure of the thermostable archaeal intron-encoded endonuclease I-DmoI. *J. Mol. Biol.*, **286**, 1123–1136.

27. 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.

28. 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.

29. 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.

30. 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.

31. Grizot,S., Epinat,J.C., Thomas,S., Duclert,A., Rolland,S., Paques,F. and Duchateau,P. (2010) Generation of redesigned homing endonucleases comprising DNA-binding domains derived from two different scaffolds. *Nucleic Acids Res.*, **38**, 2006–2018.

32. 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.

33. 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.

34. Silva,G.H., Belfort,M., Wende,W. and Pingoud,A. (2006) From monomeric to homodimeric endonucleases and back: engineering novel specificity of LAGLIDADG enzymes. *J. Mol. Biol.*, **361**, 744–754.

35. Ashworth,J., Taylor,G.K., Havranek,J.J., Quadri,S.A., Stoddard,B.L. and Baker,D. (2010) Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. *Nucleic Acids Res.*, **38**, 5601–5608.

36. 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.

37. Rimseliene,R., Maneliene,Z., Lubys,A. and Janulaitis,A. (2003) Engineering of restriction endonucleases: using methylation activity of the bifunctional endonuclease Eco57I to select the mutant with a novel sequence specificity. *J. Mol. Biol.*, **327**, 383–391.

38. Samuelson,J.C. and Xu,S.Y. (2002) Directed evolution of restriction endonuclease BstYI to achieve increased substrate specificity. *J. Mol. Biol.*, **319**, 673–683.

39. Schottler,S., Wenz,C., Lanio,T., Jeltsch,A. and Pingoud,A. (1998) Protein engineering of the restriction endonuclease EcoRV– structure-guided design of enzyme variants that recognize the base pairs flanking the recognition site. *Eur. J. Biochem.*, **258**, 184–191.

40. Wenz,C., Selent,U., Wende,W., Jeltsch,A., Wolfes,H. and Pingoud,A. (1994) Protein engineering of the restriction endonuclease EcoRV: replacement of an amino acid residue in the DNA binding site leads to an altered selectivity towards unmodified and modified substrates. *Biochim. Biophys. Acta*, **1219**, 73–80.

41. 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.

42. Wolfe,S.A., Greisman,H.A., Ramm,E.I. and Pabo,C.O. (1999) Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *J. Mol. Biol.*, **285**, 1917–1934.

43. Sarkar,I., Hauber,I., Hauber,J. and Buchholz,F. (2007) HIV-1 proviral DNA excision using an evolved recombinase. *Science*, **316**, 1912–1915.

44. Boch,J., Scholze,H., Schornack,S., Landgraf,A., Hahn,S., Kay,S., Lahaye,T., Nickstadt,A. and Bonas,U. (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, **326**, 1509–1512.

45. Moscou,M.J. and Bogdanove,A.J. (2009) A simple cipher governs DNA recognition by TAL effectors. *Science*, **326**, 1501.

46. Miller,J.C., Tan,S., Qiao,G., Barlow,K.A., Wang,J., Xia,D.F., Meng,X., Paschon,D.E., Leung,E., Hinkley,S.J. *et al.* (2011) A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.*, **29**, 143–148.

47. Li,T., Huang,S., Jiang,W.Z., Wright,D., Spalding,M.H., Weeks,D.P. and Yang,B. (2011) TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic Acids Res.*, **39**, 359–372.

48. Christian,M., Cermak,T., Doyle,E.L., Schmidt,C., Zhang,F., Hummel,A., Bogdanove,A.J. and Voytas,D.F. (2010) Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics*, **186**, 757–761.