Published online 2 September 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 22* ***9705–9719***
doi:10.1093/nar/gkr669

Activity, specificity and structure of I-Bth0305I: a representative of a new homing endonuclease family

Gregory K. Taylor[1], Daniel F. Heiter[2], Shmuel Pietrokovski[3] and Barry L. Stoddard[1,*]

[1]Graduate Program in Molecular and Cellular Biology and the Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98019, [2]New England Biolabs, Inc., Ipswich, MA 01938, USA and [3]Molecular Genetics Department, Weizmann Institute of Science, Rehovot 76100, Israel

Received June 26, 2011; Revised July 28, 2011; Accepted July 29, 2011

ABSTRACT

Novel family of putative homing endonuclease genes was recently discovered during analyses of metagenomic and genomic sequence data. One such protein is encoded within a group I intron that resides in the recA gene of the *Bacillus thuringiensis* 0305ϕ8–36 bacteriophage. Named I-Bth0305I, the endonuclease cleaves a DNA target in the uninterrupted recA gene at a position immediately adjacent to the intron insertion site. The enzyme displays a multidomain, homodimeric architecture and footprints a DNA region of ~60 bp. Its highest specificity corresponds to a 14-bp pseudopalindromic sequence that is directly centered across the DNA cleavage site. Unlike many homing endonucleases, the specificity profile of the enzyme is evenly distributed across much of its target site, such that few single base pair substitutions cause a significant decrease in cleavage activity. A crystal structure of its C-terminal domain confirms a nuclease fold that is homologous to very short patch repair (Vsr) endonucleases. The domain architecture and DNA recognition profile displayed by I-Bth0305I, which is the prototype of a homing lineage that we term the 'EDxHD' family, are distinct from previously characterized homing endonucleases.

Homing endonuclease are proteins that drive the dominant, non-Mendelian inheritance of their own reading frames by catalyzing a double-strand break (DSB) at specific DNA target sites in a recipient genome (1). The DSB is repaired via homologous recombination, using an allele of the target gene that contains the homing endonuclease gene (HEG) as a repair template; this copies the HEG into the site of DNA cleavage. HEGs are often embedded within self-splicing introns or inteins. The inclusion of a self-splicing genetic element as part of the mobile DNA allows invasion of highly conserved regions in crucial host genes without disrupting their essential functions. The coevolution of a homing endonuclease, its surrounding intron or intein, and the host gene results in an intricate network of genetic and physical interactions that affect the expression, specificity and invasiveness of the mobile element (2).

To succeed as mobile genetic elements, homing endonucleases must balance competing requirements for high DNA cleavage specificity (to avoid host toxicity) versus the need for reduced fidelity at various base pairs in their target site (to facilitate genetic mobility in the face of sequence drift within potential DNA target sites). Homing endonucleases and associated mobile introns and inteins that have successfully achieved this balance are encoded in genomes of bacteria, organelles of fungi and algae, single cell protists and in the bacteriophage and viruses that accompany and infect those organisms.

There are five well-characterized families of homing endonucleases, which are each classified according to their unique protein folds and distinct catalytic active sites and DNA cleavage mechanisms (1). Members of the 'LADLIDADG' family, so named on the basis of their most conserved protein motif, are found in eukaryotic organellar and archaeal genomes, and are the most specific of the known homing endonucleases (3). They exist both as homodimers that are limited to recognition of palindromic and near-palindromic target sites, and as pseudosymmetric monomers (where two structurally similar domains are tethered together on a single protein chain) that can target completely asymmetric targets. Members of the 'His-Cys box' and the 'PD...(D/E)-xK' families (found in protists and in cyanobacteria, respectively) also form multimeric protein complexes that

*To whom correspondence should be addressed. Tel: +1 206 667 4031; Fax: +1 206 667 3331; Email: bstoddar@fhcrc.org

recognize symmetric target sequences (4,5). In contrast, members of the HNH and GIY-YIG families (usually found in bacteriophage) display multidomain structures (corresponding to separate DNA binding and catalytic regions) and adopt highly elongated conformations when bound to DNA (6–8). As a result, those proteins usually recognize long non-palindromic sequences with significantly reduced fidelity (9,10).

Recently, a novel type of fractured gene structure, containing separately encoded halves of self-splicing inteins that interrupt individual host genes in the same locus, was discovered during an analysis of environmental metagenomic sequence data collected by the Global Ocean Sampling (GOS) project (11). These split intein sequences are found in a diverse set of host genes that are primarily involved in DNA synthesis and repair. The inteins are themselves often interrupted either by open reading frames (ORFs) that encode members of the GIY-YIG homing endonuclease family, or by novel ORFs that do not exhibit significant sequence similarity to previously characterized homing endonuclease families. Homologs of those uncharacterized ORFs were also found associated with introns or as free-standing genes. In total, 15 members of the newly discovered gene family were described, including two within previously annotated recA genes in the NCBI sequence database.

The C-terminal region of this newly identified protein family displays limited sequence homology [typically corresponding to e-values from a BLASTP (12) $<10^{-3}$] to the catalytic domain of the very short patch repair ('Vsr') endonucleases (enzymes that generate a 5′ nick at T:G mismatches in newly replicated DNA and thus stimulate DNA nucleotide excision repair) (13,14). Several catalytic residues from Vsr endonucleases are conserved across all members of the new gene family, and form the composite sequence motif EDxHD. These residues include an essential aspartate that coordinates a catalytic magnesium ion, a histidine believed to act as a general base and a neighboring aspartate residue. Based on the presence of a recognizable endonuclease catalytic domain within these intron- and intein-associated microbial ORFs and the conservation of catalytic residues within that domain, this gene family was therefore hypothesized to encode a novel lineage of homing endonucleases.

These ORFs also display sequence signatures in their N-terminal regions that are similar to those found in several nuclease associated modular DNA-binding motifs ('NUMODs') (15). NUMODs are frequently found in other homing endonucleases from bacteriophage, such as the GIY-YIG endonuclease I-TevI (8) and the HNH endonuclease I-HmuI (6). In those cases, the NUMODs are found at the C-terminal end of those proteins (a reversed domain organization compared to the metagenomic ORFs described above). The extended conformation that NUMOD regions adopt upon DNA binding dictates that they make relatively sparse contacts across their long target sites.

A representative member of this novel homing endonuclease family, which we have named I-Bth0305I, was identified in the NCBI sequence database during the same genomic analysis (11). This ORF is located within a group I intron that interrupts the RecA gene of *Bacillus thuringiensis* 0305ϕ8–36 bacteriophage. Experiments described in this manuscript describe the binding site, cleavage pattern and specificity of I-Bth0305I, and the crystal structure of its catalytic domain. These experiments demonstrate that I-Bth0305I is a site-specific endonuclease that forms a homodimer and contacts a region of DNA up to 60 bp in length. Unlike many bacteriophage homing endonucleases (which tether relatively nonspecific catalytic nuclease domains to sequence-specific DNA-binding domains, and therefore display significant specificity for DNA base pairs that are located some distance from the site of cleavage), I-Bth0305I displays its greatest specificity across the central residues of its recognition site (spanning the positions of DNA cleavage and intron insertion), and little additional sequence specificity at positions more distant from the cleavage site. The crystal structure of the I-Bth0305I catalytic domain confirms that members of this putative homing endonuclease family share a common ancestor with the Vsr mismatch repair endonuclease, and supports a similar mechanism for DNA strand cleavage.

MATERIALS AND METHODS

Computational sequence analysis

Sequences of Vsr-like putative homing endonucleases (Supplementary Data) were identified in the NCBI sequence databases and JCVI data using BLAST sequence searches and BLIMPS motif searches as previously described (11). Multiple sequence alignments were constructed with MEME (16), MACAW (17), DIALIGN-TX (18) and GLAM-2 (19) programs.

RecA gene regions corresponding to the I-Bth0305I cleavage and intron insertion site were identified by searching complete genomes of bacteria from the NCBI with Blocks database block IPB001553D using the BLIMPS program. The identified regions and 0305ϕ8–36 bacteriophage intron-inserted region were aligned using the SeAl program (http://tree.bio.ed.ac.uk/software/seal/) to form a 1368 sequences multiple alignment. Sequence logo of this region and of its translated protein product were constructed as previously described (20), using a total of four characters and equal expected base frequencies for the DNA sequence logo.

I-Bth0305I NUMOD conserved motifs were identified by analyzing I-Bth0305I and sequences similar to its N-terminal non-catalytic region. One such motif, typically appearing twice in each sequence, was identified. This motif was found to be significantly similar to the 'NUMOD 2 motif' (15) and to various DNA-binding HTH motifs from the Blocks release 14.3 database (21) [including IPB000792 (LuxR bacterial regulatory proteins), IPB000831 (Trp repressors) and IPB002197B (FIS bacterial regulatory proteins)] using the LAMA program (22). The specified blocks were used to predict the position of the HTH DNA-binding region within the NUMOD 2 motifs of I-Bth0305I.

I-Bth0305I cloning

Synthetic genes encoding I-Bth0305I and several additional homologs that were identified in an earlier metagenomic analysis (11) were ordered from Genscript (New Jersey, USA) with codons optimized for protein expression in *Escherichia coli* (Supplementary Figure S1). These reading frames were ligated into an in-house pET15-HE vector (Supplementary Figure S2) for initial protein trials. Subsequently, the reading frame encoding I-Bth0305I was subcloned into a pGEX-6p-3 expression vector, for production of the protein as a fusion with glutathione-S-transferase (GST). Inactivated constructs of the full-length protein were generated by mutating either the putative general base (H213A) or a putative metal-binding residue (D222A). A construct corresponding to the isolated predicted catalytic domain was generated by subcloning amino acids 167 through 266; two point mutations corresponding to D196A and H213A were introduced to allow overexpression by inactivating the construct. To facilitate crystallographic phasing, an additional point mutation (L180M, which could be expressed as a selenomethionyl residue) was introduced at a position predicted to be a surface residue on the opposite side of the protein from the bound DNA.

Protein overexpression and purification

For initial overexpression trials of I-Bth0305I and its homologs, the pET-15HE expression vectors containing the endonuclease reading frames were transformed into BL21(DE3)RIL cells using a standard heat shock transformation protocol: add 5 ng plasmid to 50 µl competent cells, incubate on ice for 2 min, heat shock for 30 s at 42°C, incubate on ice for 2 min, add 200 µl SOC media, shake at 220 rpm at 37°C for 20 min, then plate on LB agar plates with 0.1 mg/ml ampicillin. Single colonies were picked and grown in LB media with 0.1 mg/ml ampicillin. Starter cultures of 3 ml were grown overnight to saturation and then transferred to 1 l of LB media which was incubated at 37°C at 220 rpm until cells reached mid log phase (OD 0.5–1.1). Cultures were then placed on ice for 20–60 min before adding IPTG to 1 mM. Cells were harvested by centrifugation and examined by SDS–PAGE electrophoretic analyses (Supplementary Figure S3).

Purification of I-Bth0305I to homogeneity was then carried out using protein expressed as a GST fusion protein from pGEX-6p-3 bacterial expression vector. GST-tagged I-Bth0305I was overexpressed at 16°C while shaking at 220 rpm for 16–20 h. The cell pellet was resuspended in 45 ml of lysis buffer (50 mM Tris pH 7.0, 250 mM NaCl) before being sonicated on ice for 3×30 s (with 1 min cooling periods) in a 50 ml polypropylene tube using a high-power setting with a microtip. The resulting cell lysate was centrifuged to pellet insoluble material. The supernatant was then incubated with 2 ml of washed Sepharose-glutathione 4B beads (GE life sciences) using a gentle rocking motion at room temperature for 30 min. Beads were collected using a gravity flow columns and washed with 40 ml of high salt wash buffer (50 mM Tris pH 7.0, 2 M NaCl). Beads were washed again with lysis buffer. Finally, 2 ml of lysis buffer was added to the beads along with 80 U of PreScission protease. The mixture was incubated for 16 h with a gentle rocking motion at 16°C. Resulting protein was eluted directly from the beads and purified further via heparin affinity chromatography. An amount of 2 ml of protein at a concentration roughly 2 mg/ml was run over a heparin column in lysis buffer. Following binding, a 40 ml gradient was applied where the NaCl concentration was increased from 250 mM to 2 M NaCl. Pure I-Bth0305I eluted at ~1 M NaCl and was found to be >95% pure as estimated by electrophoretic analysis.

Specificity determination

Purified I-Bth0305 was used to digest several phage DNA samples to assess the extent of activity. Phage lambda DNA was chosen as a substrate for further testing. Aliquots containing 30 µg of phage lambda DNA was digested for 1 h at 37°C with a series of 2-fold dilutions of I-Bth0305I ranging in concentration from 20 ng/µl (0.65 µM) to 9.8 pg/µl (0.6 nM) as shown and further illustrated in Supplementary Figure S4. The DNA was extracted with phenol and chloroform, precipitated, and resuspended in 10 mM Tris 1 mM EDTA, and then diluted in water to 10 ng/µl for use as template for sequencing reactions. Sequencing reactions were carried on the respective DNA samples using 19-base oligonucleotide primers (IDT, Inc.), which were complementary to staggered positions along each DNA strand. Sequencing reactions were performed on an ABI 3730xl capillary sequencer. Output sequence traces were assembled and aligned to the reference lambda genome (Genbank file: NC_001416). Assembled sequence traces were examined by eye for signals indicative of strand-cleavage comprising a significant drop in average peak trace height following a spurious additional 'A' peak (in the case of forward sequencing reactions) or a spurious additional 'T' peak (in the case of transposed reverse sequencing reactions).

Cleavage experiments

Non-competive cleavage digests (corresponding to experiments depicted in Figures 2a and 4) were performed using equimolar concentrations (500 nM) of enzyme and linear DNA duplex substrates. The DNA substrates were generated via PCR from plasmid templates. Run-off sequencing using Taq polymerase on the digested product generated from the recA gene sequence from 0305ϕ8-36 bacteriophage identified the site of cleavage in that target site (Figure 2d and Supplementary Figure S5).

In competitive cleavage digest experiments (corresponding to Figures 2b, 5 and 6), up to four different substrates, each at 3.5 nM concentration, were simultaneously digested with 70 nM of I-Bth0305I for 30 min at 37°C. The substrates were of length 2200, 1900, 1600 or 1300 bp and each contained a putative target site exactly at the center of the DNA construct. All digest were assayed using 1.2% agarose gel electrophoresis and relative substrate and product concentrations were quantitated using the ImageJ program. All digests were

performed in 50 mM Tris pH 7.6, 50 mM NaCl and 1 mM $MgCl_2$.

DNAse I footprinting

A 120-bp polymerase chain reaction product corresponding to the uninterrupted RecA gene sequence from bacteriophage Bth0305ϕ8-36, with the endonuclease cleavage site positioned at its center, was generated using either of two radiolabeled PCR primers. An amount of 0.1 pmol of this radiolabeled PCR product was incubated with 20 μM I-Bth0305I in binding buffer (50 mM Tris pH 7.0, 60 mM KCl, 1 mM $MgCl_2$, 1 mM 2-mercaptoethanol, 2 mg/ml Bovine serum albumin) for 5 min at room temperature. Following binding, 10 μl of DNAseI (Roche pharmaceuticals) was added and allowed to react for 5 min at room temperature. After this incubation, reactions were quenched with 160 μl of stop solution (20 mM EDTA, 2 mg/ml salmon sperm DNA). Phenol extraction and ethanol precipitation separated the digested PCR product from I-Bth0305I and BSA in the reaction. Resulting samples were loaded on a 6% polyacrylamide DNA sequencing gel at 1700 V for 1 h 50 min.

Binding assays via isothermal titration calorimetry

Aliquots of a DNA duplex corresponding to a 67-bp region of the 0305ϕ bacteriophage RecA gene sequence, centered around the endonuclease cleavage site were injected into I-Bth0305I (300 μl, 20 μM) (Supplementary Figure S6). Prior to analysis, both samples were dialyzed into identical buffers corresponding to 20 mM HEPES pH 7.6, 50 mM NaCl, 10 mM $CaCl_2$. The reference cell temperature was kept constant at 30°C with a stirring speed of 1000 rpm. In total, there were 16 injections, with the first injection being half the volume and duration as the remaining injections (2.5 μl over 5.0 s, 180 s between each injection). The binding analyses were performed in triplicate.

Protein crystallography

A complex corresponding to a catalytically inactivated nuclease domain (residues 167–266, containing active site point mutations D196A and H213A) was overexpressed and purified in a manner similar to full-length I-Bth0305I, except that the heparin purification step was omitted. Crystals of this construct were grown via the hanging drop method against a reservoir containing 100 mM $LiSO_4$, 100 mM Tris pH 7.4–8.4, PEG 4000 27–30 w/v percent in 3–4 days. Crystals of native protein and of selenomethionyl-derivatizd protein grew under similar conditions, and both were transferred into a cryoprotectant solution (100 mM $LiSO_4$,100 mM Tris pH 8.5, 30% PEG 4000, 20% sucrose) and then flash frozen in liquid nitrogen. Data collection was performed at Beamline 5.0.2 at the Advanced Light Source (ALS) synchrotron facility at Lawrence Berkeley National Laboratory (Berkeley, CA, USA). Data integration and scaling was performed using program HKL2000 and all subsequent analysis was performed using the PHENIX crystallography suite. A single selenomethionine data set was used to solve phases, generate an electron density map, and build a molecular model of the nuclease domain. This model was then used to solve phases for the native data set via molecular replacement, and the final structure was built and refined to 2.2 Å resolution. The native data set was used for final refinement, even though it was slightly lower resolution (2.2 Å versus 2.15 Å) because the merging statistics for that dataset were otherwise superior to the Se-Met data (Table 1).

Table 1. Data collection and refinement statistics for crystallographic structure determination of the I-Bth0305I catalytic domain

	SeMet Data set	Native Data set
X-ray source	ALS beamline 5.0.2	ALS beamline 5.0.2
Wavelength	0.9794 Å	1.0 Å
Space group	P 4_3 2_1 2	P 4_3 2_1 2
Cell dimensions		
a, b, c (Å)	51.87, 51.87, 133.39	52.29, 52.29, 133.79
Data collection		
Resolution (Å)	2.15 (2.19–2.15)	2.2 (2.24–2.20)
R_{merge} (%)	9.4 (30.7)	5.2 (17.1)
$I/\sigma(I)$	15.2 (5.2)	40 (14.6)
χ^2	1.13 (0.92)	0.96 (0.98)
Completeness (%)	96.4 (94.5)	99.9 (98.3)
Redundancy	6.4 (5.4)	13.4 (11.2)
Refinement		
R_{work}/R_{free}		0.22 / 0.26
Number of molecules		2
Number of protein atoms		1658
Number of water atoms		29
RMSD (bond/angles)		
Bond lengths (Å)		0.002
Bond angles (°)		0.486
Ramachandran plot statistics (%) excluding Gly, Pro		
Most favored regions		93.4
Additionally allowed regions		6.6
Disallowed		0.0

RESULTS

Cloning and protein production

Genes encoding several individual representatives of the Vsr-like endonuclease gene family identified in the metagenomic analyses (11), as well as the protein we have named I-Bth0305I, were each synthesized as codon-optimized reading frames for bacterial expression in *E. coli* and then subcloned into a modified pET (Novagen, Inc) vector that incorporates an N-terminal, 6-histidine affinity purification tag that can be removed by proteolytic digests with thrombin (Supplementary Figures S1 and S2). The resulting constructs displayed a wide range of behaviors during bacterial overexpression and purification (Supplementary Figure S3). Of the seven protein constructs tested, four were observed to form insoluble inclusion bodies regardless of induction conditions. Out of the remaining ORFs, the construct corresponding to I-Bth0305I significantly reduced the growth rate of the bacterial culture after IPTG induction and was observed in the soluble fraction of lysed cells. This construct was subsequently recloned into a GST-fusion expression vector (pGEX-6P-3) in the hopes that the larger affinity partner might reduce DNA binding or cleavage activity during expression, allowing improved growth and recovery of expressed protein. The resulting fusion protein was soluble, easily recovered from clarified cell lysate, and could be subsequently purified using affinity chromatography and liberated from its GST fusion partner via a proteolytic digestion as described in 'Materials and Methods section'. The yield of this protein was ~1.5 mg/l of culture, and the resulting protein could be concentrated to at least 9 mg/ml in a storage buffer corresponding to 250 mM NaCl, 50 mM Tris pH 7.0, 5% (v/v) glycerol.

The I-Bth0305I reading frame encodes a protein that is 266 amino acids in length, corresponding to a predicted molecular weight of 30 912 Da. The surrounding group I intron within the bacteriophage 0305ϕ8–36 RecA gene is 801 nt in length; the start codon for the putative endonuclease reading frame is found 88 nt from the start of the intron. The protein ORF interrupts the P5 element in the canonical representation of the group I intron's secondary and tertiary structure (23). As described in the original analysis of this protein family, I-Bth0305I displays an N-terminal region with two copies of sequences corresponding to NUMOD 2 DNA-binding motifs (15), and a C-terminal region that shares homology with the catalytic domain of the Vsr DNA mismatch repair endonuclease (14). Further analysis, using homologs of the I-Bth0305I N-terminal region, indicated that the two NUMOD regions might span a putative helix–turn–helix (HTH) sequence-specific DNA-binding region motif. Using the conserved sequence regions of the Vsr-like endonuclease proteins (11) we identified additional members of this family including bacteriophage Hef type homing endonucleases (24) and a bacterial protein from *Corynebacterium glutamicum* ATCC 13032 (Supplementary Data). These sequences allowed us to extend and refine the conserved sequence regions of the Vsr-like endonuclease family, including the identification of a fifth putative active site residue (Figure 1).

These sequence relationships were exploited at several points in this study to generate truncated expression constructs corresponding to isolated structural regions of the protein, and to design catalytically inactivating point mutations in the catalytic domain. These constructs were subcloned into the same bacterial expression vector described above, and purified as described in 'Materials and Methods' section. The overall yield of isolated N- and C-terminal regions of I-Bth0305I were ~1 and 3 mg/l, respectively.

Target site identification

We next tested the ability of full length, wild-type I-Bth0305I to cleave a DNA substrate corresponding to the intron-minus allele of the RecA gene, and compared that cleavage activity with substrates containing DNA sequences that correspond to an 'intron-plus' recA allele. This experimental design was based on the known genetic propagation mechanism of most homing endonucleases, which cleave a target site within an intron- or intein-minus allele of their host gene, but usually do not cleave the same allele when it contains the inserted intervening sequence (25). In our experiments, efficient cleavage of the DNA substrate corresponding to the uninterrupted RecA gene was observed (Figure 2a). Substrates containing the intron–exon junction sequences of the bacteriophage recA gene were not cleaved by the enzyme under any conditions (Figure 2b), indicating that the enzyme only cleaves the uninterrupted recA allele prior to intron insertion.

In order to further define the actual target site and cleavage pattern exhibited by the endonuclease, as well as to establish the overall specificity of the enzyme, two separate experiments were conducted. In the first, lambda phage DNA (a 48.5-kb double-stranded DNA construct of known sequence) was used as a substrate in a series of digests with variable concentrations of purified endonuclease. All resulting product fragments were identified and sequenced using a comprehensive set of oligonucleotide primers that cover the entire length of both DNA strands. An alignment of the nicked and cleaved DNA sequences produced in this experiment identified the target site preference for the enzyme. In the second experiment, a 500-bp substrate corresponding to the recA sequence from the 0305ϕ bacteriophage was digested to completion, and both product strands were subjected to run-off sequencing using TaqI polymerase. When analyzed together, these two experiments produced an unambiguous assignment of the enzyme's target site preference and cleavage activity.

Digestion of lambda DNA generated a list of target sites that were hydrolyzed by the endonuclease (Supplementary Figure S4). Alignment of these genomic sequences resulted in a target site consensus corresponding to 5′-T-T-x-G-x_6-C-x-A-A-3′ (Figure 2c). This 14-bp target site displays pseudopalindromic symmetry, with the 'TTxG' sequence in the left half-site complementary to the 'CxAA' sequence in the right half-site. The majority

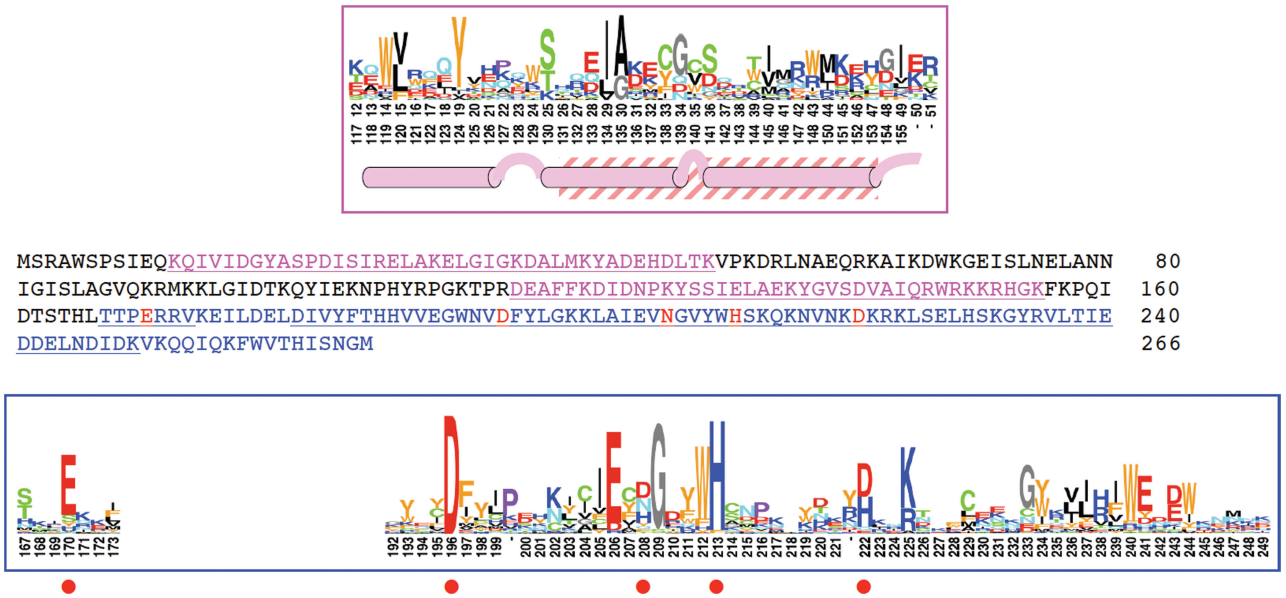

Figure 1. Features of the I-Bth0305I protein sequence. The catalytic domain of the protein is indicated in blue font with putative active site residues in red. The underlined region corresponds to the sequence logo in the lower blue frame, where the predicted active sites are marked by red bullets. Two repeats of 'NUMOD' sequence motifs in the putative DNA-binding domain are shown in pink underlined font, with their motif logo shown above in the upper pink frame. Logo positions are numbered according to I-Bth0305I. Gaps (–) mark deletions in I-Bth0305I relative to other protein family members, and I-Bth0305I residue Trp149 is an insert relative to the family members and thus not shown in the logo. Beneath the logos of the repeated NUMOD motif is its predicted structure (cylinders for α-helices and arcs for loops and turns). The hatched region denotes a predicted DNA-binding HTH motif.

of the target sites in these assays were nicked on either the top or bottom strand (at positions that considered together would correspond to a two base, 5′ overhang). One site that displayed a sequence that was particularly close to the consensus described above (differing at only 1 bp out of 6) was cleaved on both strands and thereby produced the actual two base, 5′ overhang and cleavage pattern.

Direct run-off sequencing of the product strands produced from digests with the actual RecA-coding sequence as a substrate resulted in identification of a target site (5′-TTcGgtgatcCaAA-3′) and cleavage pattern that agree precisely with the results described above (Figure 2d and Supplementary Figure S5). Therefore, it appears that the enzyme cleaves a partially symmetric DNA target site located immediately upstream of the intron insertion site in the recA target and requires conservation of most of the 'TTxG' consensus target sequence in both DNA half-sites in order to generate a DSB. When limiting our analysis of the lambda DNA cleavage products to only those targets that were most efficiently nicked or cleaved (at least 90% digestion of either strand), the resulting information content and logo plot across the central 6 bp was observed to agree more closely with the recA target site sequence.

After establishing the cleavage site in the RecA host gene, we next determined the DNAse I footprint of the enzyme bound to its DNA target (Figure 3). A catalytically inactive variant of I-Bth0305I (D222N, containing a mutation of a putative catalytic asparate residue that was observed to prevent cleavage activity) was incubated with 120-bp probe that corresponded to the RecA-coding sequence. The region of the complementary strand that was protected by the bound enzyme from DNAse I digestion was determined in a separate experiment. In both cases, a region of ~60 nt, corresponding to 30 bp that extend from each side of the center of the cleavage site, was protected from DNAse I cleavage. Subsequently, the binding of I-Bth0305I to a synthetic DNA duplex corresponding to this target site sequence was evaluated using multiple independent isothermal titration calorimetry experiments and determined to correspond to an exothermic binding reaction with a dissociation constant (K_D) of 24 ± 6 nM (Supplementary Figure S6).

Cleavage specificity

Having determined the extent of DNA backbone protection corresponding to the bound endonuclease footprint and the affinity of the binding interaction, we then further assessed the sequence specificity displayed by the endonuclease in a series of digests using variants of the wild-type DNA substrate (Figure 4). These experiments indicated that the enzyme exhibits the highest specificity across the central 14 bp of its target site. A series of substrates that contained either three consecutive transverted base pairs (e.g., 5′-ATC-3′ → 5′ TAG-3′) or that contained a series of AA insertions were used as substrates in parallel assays. In these experiments, cleavage activity was reduced most significantly when the DNA sequence that immediately spans the central site of catalysis was mutated. Similar perturbations introduced on either side of this central target site region were well tolerated by the enzyme.

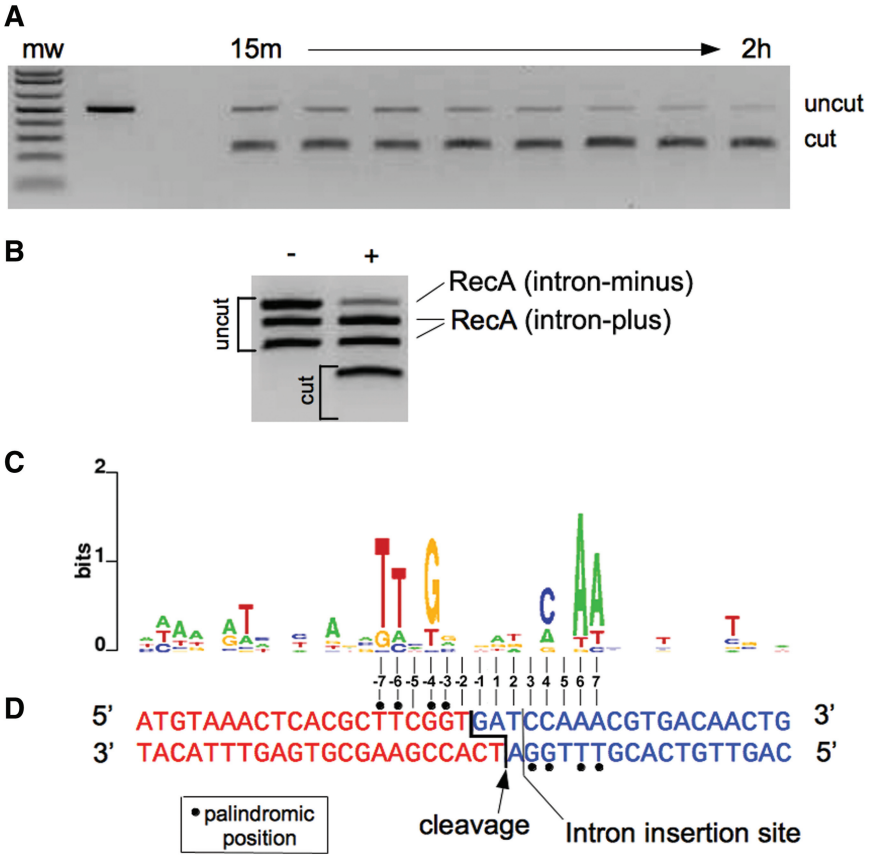

Figure 2. Determination of the I-Bth0305I DNA target site. (**A**) I-Bth0305I cleaves a DNA target site containing the sequence of the RecA host gene spanning the intron insertion site. (**B**) In competition digests, three substrates (one corresponding to the uninterrupted allele of the bacteriophage RecA gene; 'intron-minus' and two substrates corresponding to the intron-containing allele of the same RecA gene; 'intron-plus') were simultaneously digested with 70 nM I-Bth0305I. Only the intron-minus allele of the RecA sequence is cleaved. (**C**) Sequencing of the most strongly nicked and cleaved products resulting from a digest of lambda phage DNA with I-Bth0305I results in a specificity profile (i.e. a 'logo' plot) indicating that the strongest features of substrate specificity correspond to the pseudopalindromic consensus sequence 5′-TTxG-x6-CxAA-3′, which is cleaved on each strand to give two base, 5′ overhangs centered in the middle of the symmetric DNA target. For the logo shown in this figure, only those sites in the lambda genome displaying 90% or higher cleavage of at least one strand were included in the creation of the consensus sequence. Reducing the cleavage threshhold for inclusion of more sequences in the determination of the enzyme's specificity profile quantitatively affects the absolute values for information content at individual positions, but does not alter the consensus sequence identity. (**D**) Sequence of the recA target region of the I-Bth0305I endonuclease that is cleaved by I-Bth0305I. The results of run-off sequencing of cleaved top and bottom strands is consistent with generation of two base, 5′ cohesive overhangs observed in the prior experiment with lambda genomic DNA. The target site is numbered to illustrate the two pseudosymmetric half-sites in the recA gene target that flank the middle of the cleavage site. Following convention for homing endonuclease target site numbering, the left half site is accorded with negative position numbering, and the right half-site is accorded with positive position numbering. Black bullets indicate positions that are palindromically conserved between the left and right half-sites.

Alteration of the DNA sequence at the more distant 5′- and 3′-ends of the I-Bth0305I contact region (i.e. at each end of the target site previously established by DNAse I footprinting) had a much less significant effect on DNA cleavage (Figure 5). In these experiments, a series of long DNA duplex substrates (each of which were 1–2 kb in length) that contained targets with gradually decreasing regions of the RecA target sequence were assayed in parallel, competitive cleavage digest experiments. Reduction of the length of the RecA gene sequence within these long substrates from a 64-bp region (corresponding to the extreme limits of the protected region observed in DNAse I footprinting assays) to 54 bp resulted caused little or no loss of cleavage activity. In contrast, a slight reduction in activity was observed when a 33-bp RecA target sequence was present, and a more significant reduction in activity was observed when the RecA target is sequence was reduced to only 23 bp. In no case, however, was the loss of cleavage activity in these experiments as pronounced as when as few as 3 bp in the center of the target site were mutated.

Having established by a variety of methods that sequence specificity of DNA cleavage is highest across the central base pairs of its target site, we next generated a matrix of point mutations of the RecA target site (corresponding to each of the three possible single base pair substitutions at each of the central positions) and tested each for their relative 'cleavability' using *in vitro* digests (Figure 6). Although the previous experiments described above demonstrated that simultaneous mutation of as few

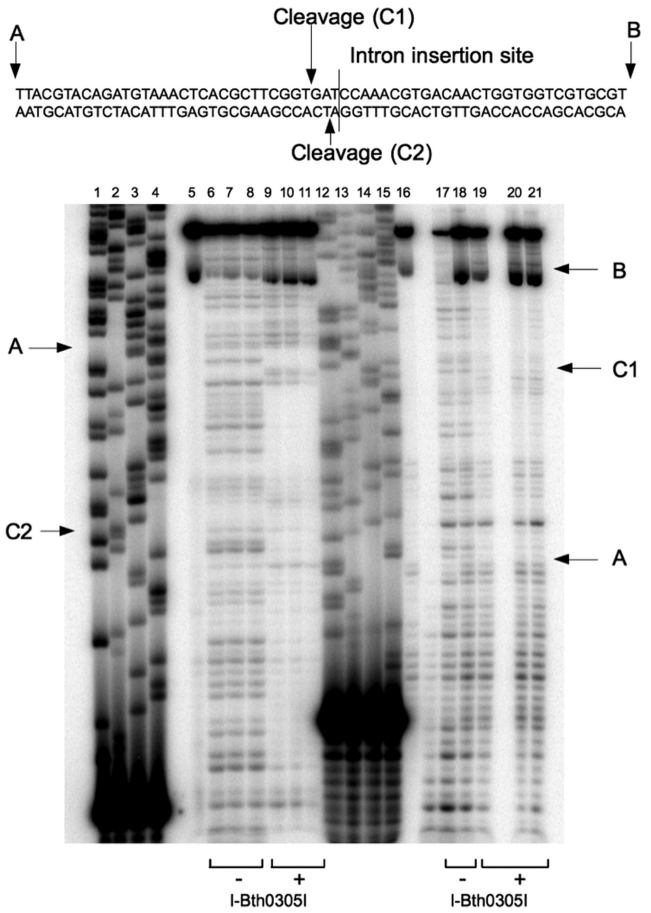

Figure 3. I-Bth0305I target site DNAse I footprint. Forward and reverse PCR primers were labeled with ^{32}P and used to generate PCR products labeled at either end. In lanes 6–8 and 17–18, reverse and forward labeled PCR products were digested with DNAse-I. In lanes 9–11 and 19–21, labeled PCR products were incubated with 20 μM I-Bth0305I and digested with DNAse I. Through a comparison of the I-Bth0305I protected and unprotected DNAse I ladders, a 60 base region with the site of catalysis at its center is protected from DNAse I degradation by specific binding of I-Bth0305I. Lanes 1–4 and 12–15 are sequence ladders and lanes 5 and 16 are undigested PCR product.

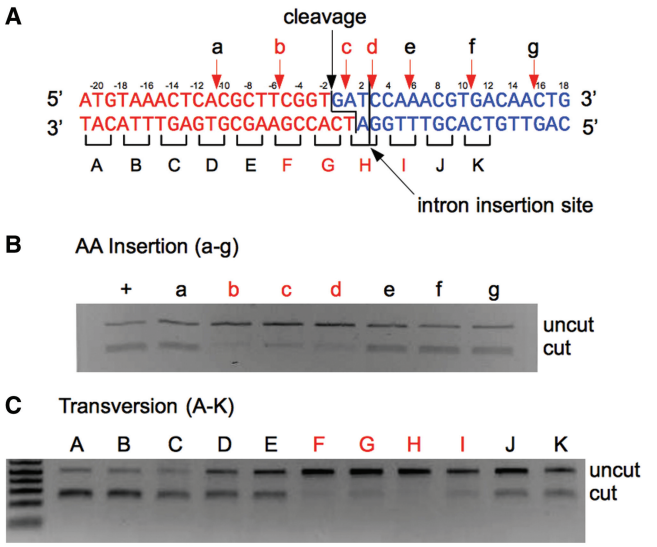

Figure 4. Effect of multiple base pair substitutions on DNA cleavage. (**A**) Enzymatic cleavage was assayed in complementary experiments where digests were performed using a DNA substrates containing the target site that were disrupted by either insertion of 2 bp at several positions, or by systematic transversion of three consecutive base pairs. For insertion disruptions, two adenosines were inserted at one of the positions indicated by the red arrows, thus generating substrates 'a' to 'g'. In transversion disruptions, several sets of three consecutive nucleotides, each marked by a bracket, were inversed, thus generating substrates 'A' to 'K'. (**B**) Cleavage products produced by digestion of substrates a–g. Product generation is significantly impaired for substrates b, c and d, corresponding to insertions of additional base pairs after positions −5, 0 and +2 in the RecA target site. (**C**) Cleavage products produced by digestion of substrates A–K. Product generation is significantly impaired for substrates F, G, H and I, corresponding to transversion of three consecutive base pairs in a region extending from position −5 to +6.

as three consecutive base pairs was sufficient to significantly impair cleavage, mutation of individual base pairs had relatively little effect on cleavage under the same reaction conditions. Only three individual nucleotide substitutions in the recA target site (at positions −1, −2 and −5 in the left half-site) showed any measurable effect on cleavage efficiency. These three base pairs correspond to positions in that half-site that are not symmetrically conserved with their counterparts in the right-half site.

Therefore, while the sequence specificity of the cleavage reaction is clearly most significant across the central 14-bp positions of the I-Bth0305I target site, the overall information content across this region (as measured by the reduction in cleavage activity caused by individual base pair substitutions) is very evenly distributed as compared to many other homing endonucleases that have been characterized (26–29), such that only multiple simultaneous base pair substitutions result in a significant loss of cleavage efficiency.

Protein oligomery and DNA target symmetry

Size exclusion chromatography experiments showed that the apparent mass of both the full-length enzyme (containing a catalytically inactivating D222N mutation) and of the isolated catalytic domain (containing a D196A mutation) were approximately twice the value that was predicted based solely on the length of their protein chains (62 kDa versus 31 kDa for the full-length protein, and 18 kDa versus 12 kDa for the catalytic domain) (Supplementary Figure S7). This result was confirmed by dynamic light scattering measurements of the catalytic domain. A different point mutant within the isolated catalytic domain (H213A, corresponding to the predicted location of the active site general base) gave a reduction in apparent mass and the dynamic radius by ~50%. These results indicate that the full-length endonuclease and its isolated catalytic domain form stable dimers in solution and that the dimerization interface is disrupted by mutation of His 213. This result agrees with the independent observation, described above, that the sequence in the recA gene that immediately surrounds the cleavage site displays significant pseudopalindromic symmetry (Figure 2). The presumed role of H213 in catalysis [based on prior mutational studies and conservation of the comparable residue in the Vsr repair endonuclease (13–14)], versus

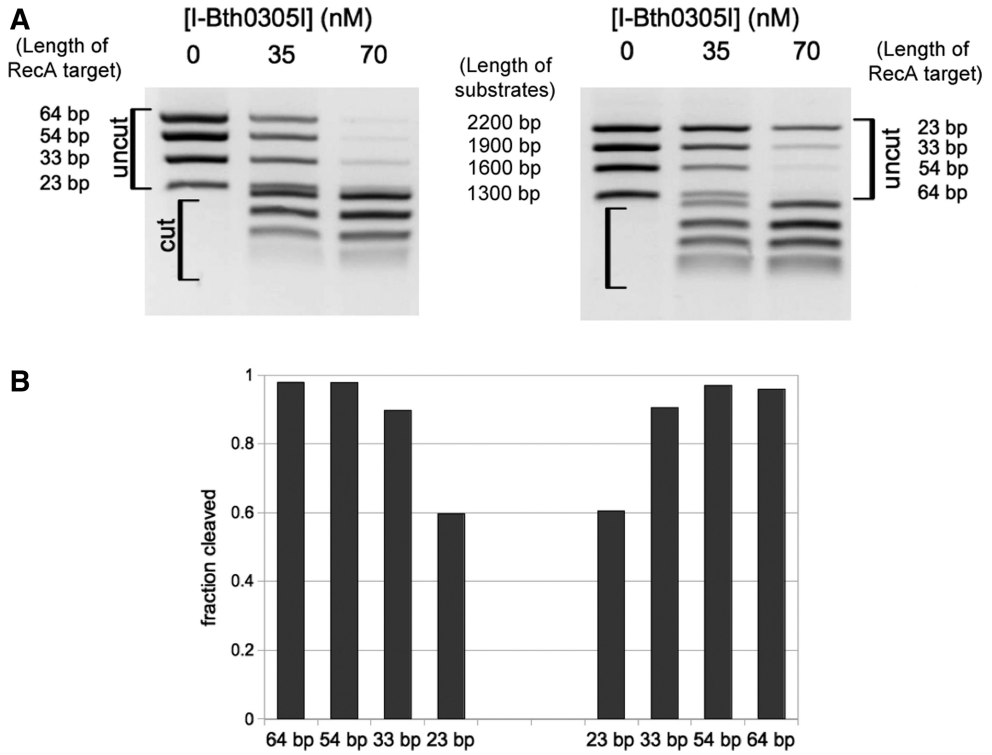

Figure 5. Effect of reduced length of RecA target site on DNA cleavage. (**A**) Four separate substrates, ranging in total length from 2200 to 1300 bp, that contained specific bacteriophage RecA target sequences of various lengths centered around the site of cleavage were digested with I-Bth0305I and cleavage was measured. The experiment was conducted twice, with the various RecA sequences embedded in DNA of different overall length, to ensure that measurable differences in cleavage were due solely to the length of the phage RecA sequence in the substrates. (**B**) Quantitation of cleavage product formation for each substrate in the presence of 70 nM I-Bth0305I.

its observed importance for dimerization of I-Bth0305I may indicate that dimerization and catalytic activity of the homing endonuclease are structurally linked, with that particular residue playing an important role for both properties. Structural studies of the isolated nuclease domain with the H213A mutation (described below) demonstrate that the A213 residue is significantly displaced from its position in the Vsr active site.

Structural relationship to the Vsr mismatch repair endonuclease

The crystal structure of a catalytically inactive double-point mutant (D196A/H213A) of the C-terminal region of I-Bth0305I (containing residues 167–266, which displays sequence homology to the Vsr mismatch repair endonuclease) was determined and refined to 2.2 Å resolution (PDB ID: 3R3P). Selenomethionyl-derivatized protein was used as the sole source of *de novo* phase information in order to avoid model bias that might arise from phase determination via molecular replacement. The final refined model (Table 1), contained residues 167–263 from the isolated catalytic domain (three residues from the C-terminus were unobserved and presumed to be disordered in the crystal). Two copies of the catalytic domain were present in the asymmetric unit; the all-atom RMSD for those two protein chains is 0.33 Å. Because the H213A mutation in this domain was previously shown to block dimerization, the interface between these two observed subunits is believed to represent a non-physiological interaction that is formed in the crystal lattice.

The structure of the catalytic domain consists of a central β-sheet with mixed parallel and anti-parallel topology surrounded by four α-helices. The structure of the I-Bth0305I catalytic domain superimposes against the homologous region of Vsr endonuclease (PDB ID: 1VSR) (13,14) with an RMSD of 8.76 Å across 61 atoms (Figure 7). The structure of the central β-sheet within the I-Bth0305I catalytic domain differs significantly from that of Vsr. This region within I-Bth0305I is twisted, as compared to a more saddle-shaped structure within Vsr. Furthermore, while this β-sheet contains four β-strands in both structures, only three strands are found to superimpose between the two enzymes; the two enzymes display their fourth (non-conserved) strands at opposite sides of the core β-sheet. As well, a zinc-binding sequence motif found in Vsr is missing from the loop that connects β3 and α2 in I-Bth0305I, and zinc atoms are not observed in the structure.

The α-helices that are observed in I-Bth0305I are also diverged from their corresponding structural elements in Vsr. First, the short I-Bth0305I helix α3 (residues 80–84) is instead a loop in Vsr. Furthermore, helix α2 in I-Bth0305I is considerably shorter (at 14 residues) than the corresponding 25-residue helix in Vsr (spanning residues 82–107) that is inserted into the DNA major groove in

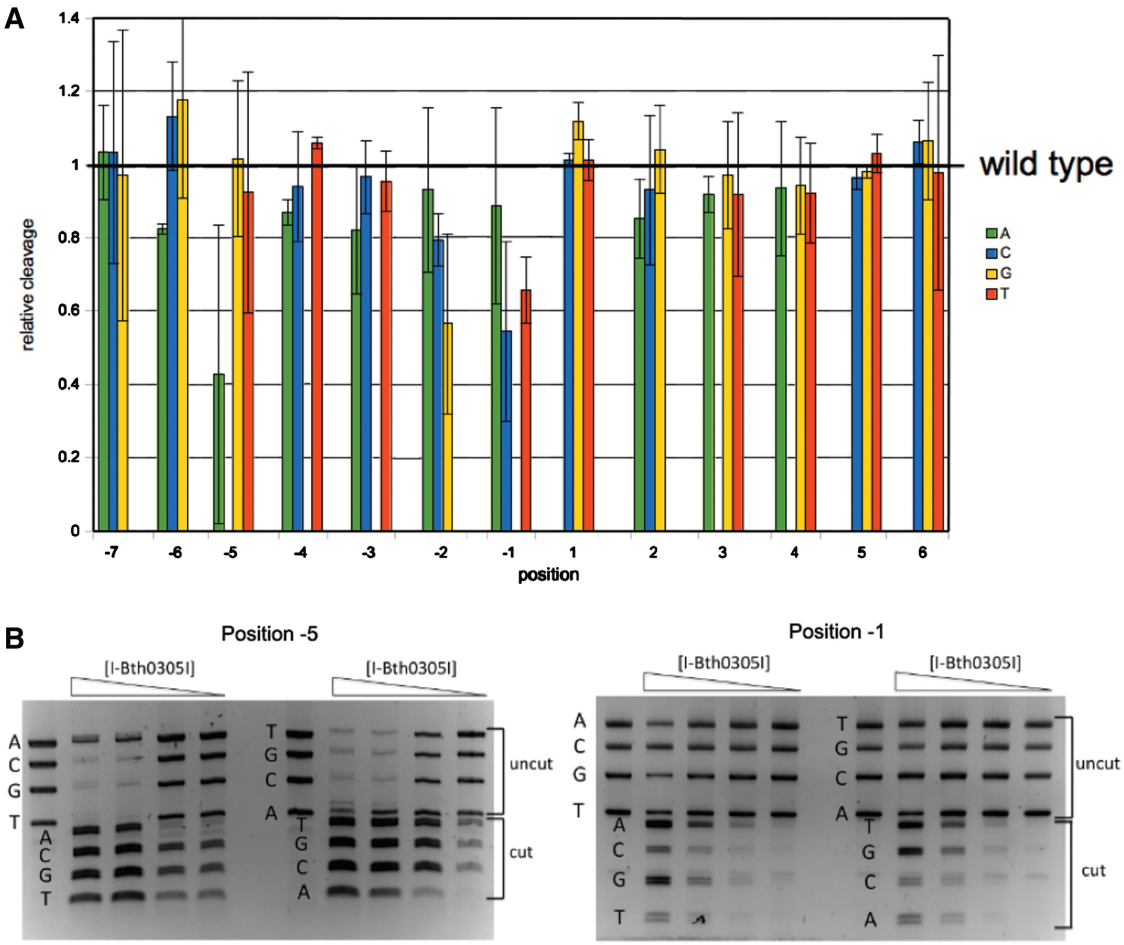

Figure 6. Effect of single base pair substitutions on DNA cleavage. (**A**) Each bar represents the relative cleavability of a target site that is altered at 1 bp relative to the wild-type target. (**B**) Raw data for cleavage specificity at positions −5 and −1, respectively. In these experiments, increasing concentrations of enzyme are used in competition experiments against equimolar concentrations of four DNA substrates that differ in length and in the identity of a single base pair at one position in the target. For each position being tested, the effect of DNA base pair mismatches were measured multiple times, including experiments in which the length of the substrates was reversed relative to the identity of the variable base pair (to ensure that differences in cleavage are due only to the sequence of the target).

its DNA-bound structures. The differences in the structures between the two nuclease domains are critical determinants for their different functions. In Vsr, two tryptophan residues (W68 and W86) are intercalated into the DNA immediately adjacent to the T:G mismatch in that enzyme's substrate target and appear to play a key role in recognition of that particular structural lesion in the DNA. In I-Bth0305I (which instead recognizes a fully paired DNA target sequence corresponding to vicinity of the intron insertion site) the corresponding region instead corresponds to a short flexible loop.

While the elaborations upon the core fold of the two enzymes are significantly diverged, their active site residues are closely comparable (Figure 7). Residues that superimpose very closely include Asp 196 in I-Bth0305I (which is Asp 51 in Vsr and is mutated to Ala in the crystal structure), Asp 222 (Asp 97) and Asn 208 (His 64). An additional residue in Vsr (His 69) that is thought to play a role in catalysis is conserved in the I-Bth0305I sequence (as His 213), but is located in a significantly different conformation in the two structures. In the structure of the I-Bth0305I catalytic domain, this residue is found at a surface-exposed position in the structure that is involved in crystal lattice contacts, which appears to perturb its position and rotameric conformation relative to the surrounding active site. A final acidic residue (Glu 170 in I-Bth0305I, corresponding to Glu 25 in Vsr endonuclease) might also participate in catalysis; this amino acid is well conserved but is found in an otherwise weakly conserved region (Figure 1).

DISCUSSION

Relationships between bacteriophage homing endonucleases

Two bacteriophage HEs have been previously crystallized and studied biochemically in great depth: the GIY-YIG endonuclease I-TevI (which drives intron homing into a thymidylate synthase host gene in T4 bacteriophage) (30)

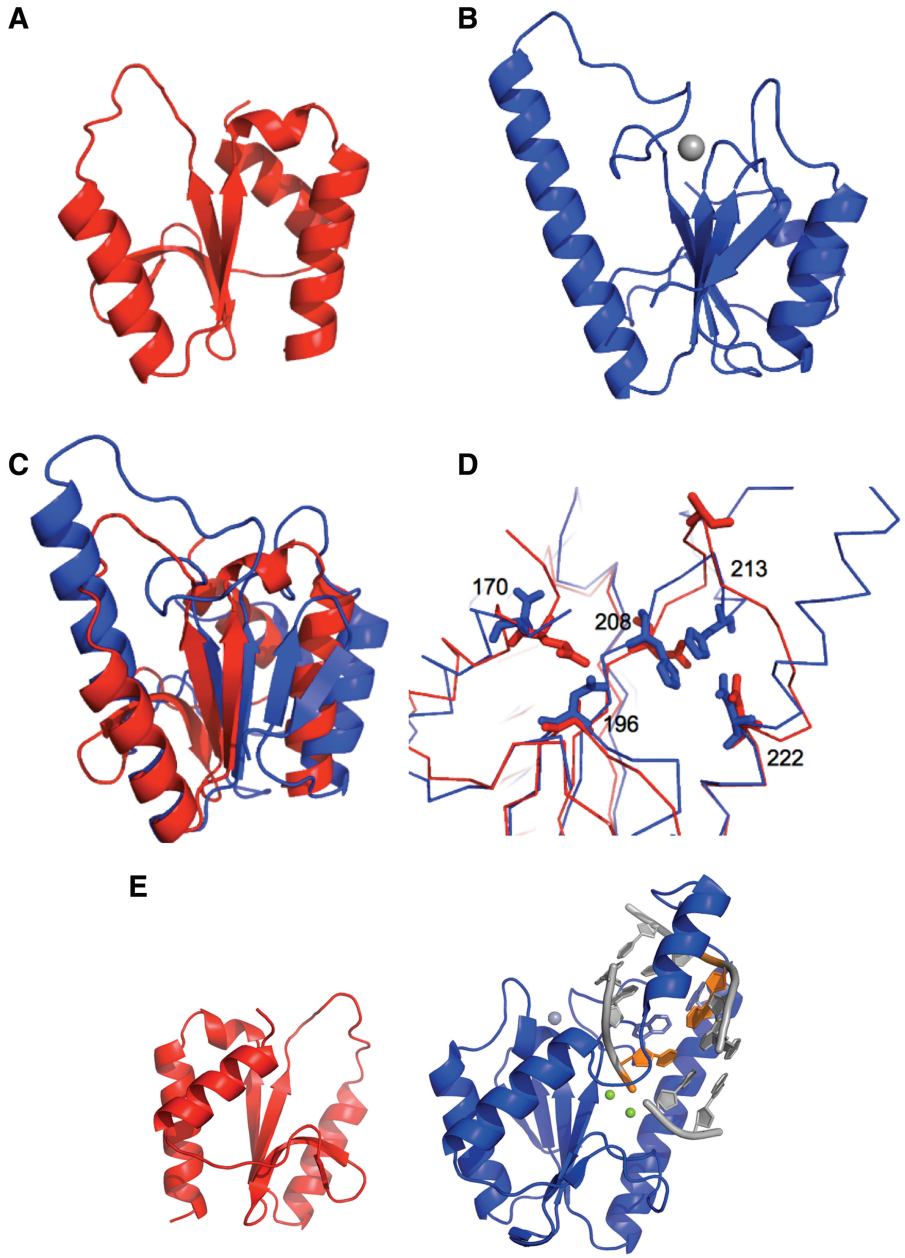

Figure 7. Structural analyses of I-Bth0305I nuclease domain. (**A**) The crystal structure of the I-Bth0305I catalytic domain (PDB ID 3r3p). (**B**) The structure of *E. coli* Vsr endonuclease in the absence of bound DNA (PDB ID 1vsr). The bound zinc ion in the Vsr structure is the cyan sphere. (**C**) Superposition of the I-Bth0305I nuclease domain and the unbound Vsr endonuclease core. (**D**) Superposition of Vsr endonuclease active site and the putative I-Bth0305I active site and catalytic residues. (**E**) Side-by-side comparison of the I-Bth0305I nuclease domain and the DNA-bound structure of the Vsr endonuclease (PDB ID 1cw0) in the same relative orientations. In the Vsr-DNA co-crystal structure, the T:G mismatched nucleotide bases are shown in orange; the tryptophan residues (W68 and W86) that intercalate next to those mismatched DNA bases are shown in light blue and the active site magnesium ions are green spheres.

and the HNH endonuclease I-HmuI (which drives intron homing into a DNA polymerase host gene in the *Bacillus* SPO1 bacteriophage) (31). Both of those enzymes, as well as their closest homologs (I-BmoI and I-BasI, respectively) appear to bind their DNA targets as monomers (9,10), with protected DNA regions extending ~30–40 bp downstream from their intron insertion site. These enzymes discriminate between intron-plus and intron-minus alleles of their host genes through a small number of sequence-specific interactions near the site of cleavage. Whereas I-HmuI acts as a strict monomer to nick its DNA target near its intron insertion site (apparently relying upon subsequent conversion of the nick to a DSB to promote homing) (10), I-TevI is observed to directly generate a DSB and a two base, 5′ overhang 23- and 25-bp upstream of the intron insertion site (9). The ability of I-TevI to directly generate a DSB may require transient dimerization of catalytic domains at the site of DNA cleavage; however, this behavior has not been demonstrated directly.

In contrast, I-Bth0305I forms a stable dimer in the absence of DNA, contacts up to 60 bp of DNA and cleaves a pseudo-palindromic target in the RecA host gene. If each individual subunit of the I-Bth0305I homodimer contacted a length of DNA target that was similar to the monomeric I-TevI and I-HmuI subunits, then the observed 60-bp contact region would simply correspond to two 30-bp DNA half-sites. The homodimeric architecture of I-Bth0305I (in the absence of bound DNA) may predispose the enzyme to recognize and cleave target sites that display greater palindromic symmetry than has been observed for enzymes that initially bind their DNA targets as monomers.

The I-Bth0305I endonuclease displays a bipartite, multidomain architecture and harbors a catalytic domain that is fused to a predicted DNA-binding region, that contains two NUMOD sequence elements that likely bind specific DNA sequences using a HTH motif. The conclusion that can be drawn from all of the experiments in this study is that the enzyme homodimerizes through interactions between nuclease domains, and that interactions of those domains with the DNA generate the majority of target site specificity at the central 14 bp of the target. The remainder of protein–DNA contacts, made at positions outside of this central pseudopalindromic region, are largely nonspecific and presumably made by the N-terminal DNA-binding regions that contain the NUMOD motifs (Figure 8a).

A similar bipartite domain organization has previously been observed in both I-HmuI, I-TevI and their homologs (6–8). However, the domain organization of this new homing endonuclease family (containing an N-terminal DNA-binding domain fused to a C-terminal nuclease domain) is reversed as compared to those previously characterized bacteriophage endonucleases, and involves an entirely different nuclease core structure, which together suggest a difference in the evolutionary history of this bacteriophage-specific homing endonuclease lineage.

HE specificity and host gene constraints

The specificity profile displayed by I-Bth0305I is unusual as compared to other well-studied, phage-derived homing endonucleases in which almost all sequence specificity of cleavage appears to be focused near the site of cleavage, with relatively little specificity derived from contacts between the HE and more distal positions in the DNA recognition site. In contrast, the HNH and GIY-YIG endonucleases appear to display bipartite recognition patterns, with limited numbers of sequence-specific contacts made both by the nuclease domains near the sites of DNA strand cleavage, and additional sequence-specific contacts made by the more distant DNA-binding regions of the enzyme. However, close examination of sequence specificity profiles of enzymes such as I-TevI (a GIY-YIG enzyme) (9) and I-HmuI (an HNH enzyme) (10) both indicate that the base pair identities in their target sites that are most critical for recognition and cleavage are also located near the site of cleavage, and are generally bases that are particularly well conserved within the coding sequence of the target host gene. This feature of DNA specificity is displayed by virtually all known families of homing endonucleases (26–29).

The specificity profile of I-Bth0305I suggests that several of the central 14 bp surrounding the intron insertion site are most specifically recognized by the enzyme and therefore might be a functionally important region of the RecA host gene. To investigate this hypothesis, we examined the conservation of the RecA-coding DNA and translated protein sequences corresponding to the endonuclease target region, by generating a multi-sequence alignment of 1368 recA genes, including the 0305ϕ8–36 bacteriophage gene without its intron (Figure 8b). The conservation of the positions in the coding sequence and protein multiple alignments was calculated using information theory measures and taking into account background frequencies of amino acids, and differing similarities between the aligned regions (20,32). This analysis demonstrates strong conservation at 11 out of the central 14 bp of the endonuclease target site, and additional, stronger conservation of the DNA and protein sequence downstream of the intron-insertion site.

The amino acid sequence of the bacteriophage RecA protein corresponding to the 20 residues that are encoded by the DNA region that is contacted by I-Bth0305I is somewhat diverged from the overall RecA consensus. Nine of those residues from the bacteriophage protein correspond to the top residue in the RecA protein logo plot, three of which (F224, G225 and P227) are encoded within the central region of the target site. The specificity profile of I-Bth0305I is somewhat correlated with the RecA-coding sequence in that region: base pair positions that are recognized by the enzyme with above average preference include the first two positions of the codons encoding G225, D226 and P227 (Figure 2). A similar observation, that the specificity of a homing endonuclease can be correlated to the reading frame and coding degeneracy within its host gene target site, has been reported for several homing endonucleases, including the I-AniI protein in *Aspergillus nidulans* (28).

The amino acid sequence encoded by the central region of the endonuclease target site spans the functionally critical 'L2' region of the RecA protein (Figure 8c). RecA forms helical filaments composed of multiple RecA monomers bound to single-stranded DNA (PDB ID 3cmt). When examining the L2 region in the context of these filaments, its residues are observed to form a β-hairpin structure that is involved in contacting the DNA backbone and at least 1 nt base (33) (Figure 8d). Regions corresponding to L2 are also found in eukaryotic and archeal RadA/Dmc1 proteins and bacterial DnaA proteins, both of which have similar DNA-binding activities (22). The L2 loop has previously been shown to be an insertion site for invasive inteins in several bacteria, including the recA gene of *Mycobacterium leprae* (34).

I-Bth0305I versus Vsr: evolution and application

Homing endonucleases share common evolutionary ancestors with a wide variety of host proteins that are responsible for an equally broad range of biological functions.

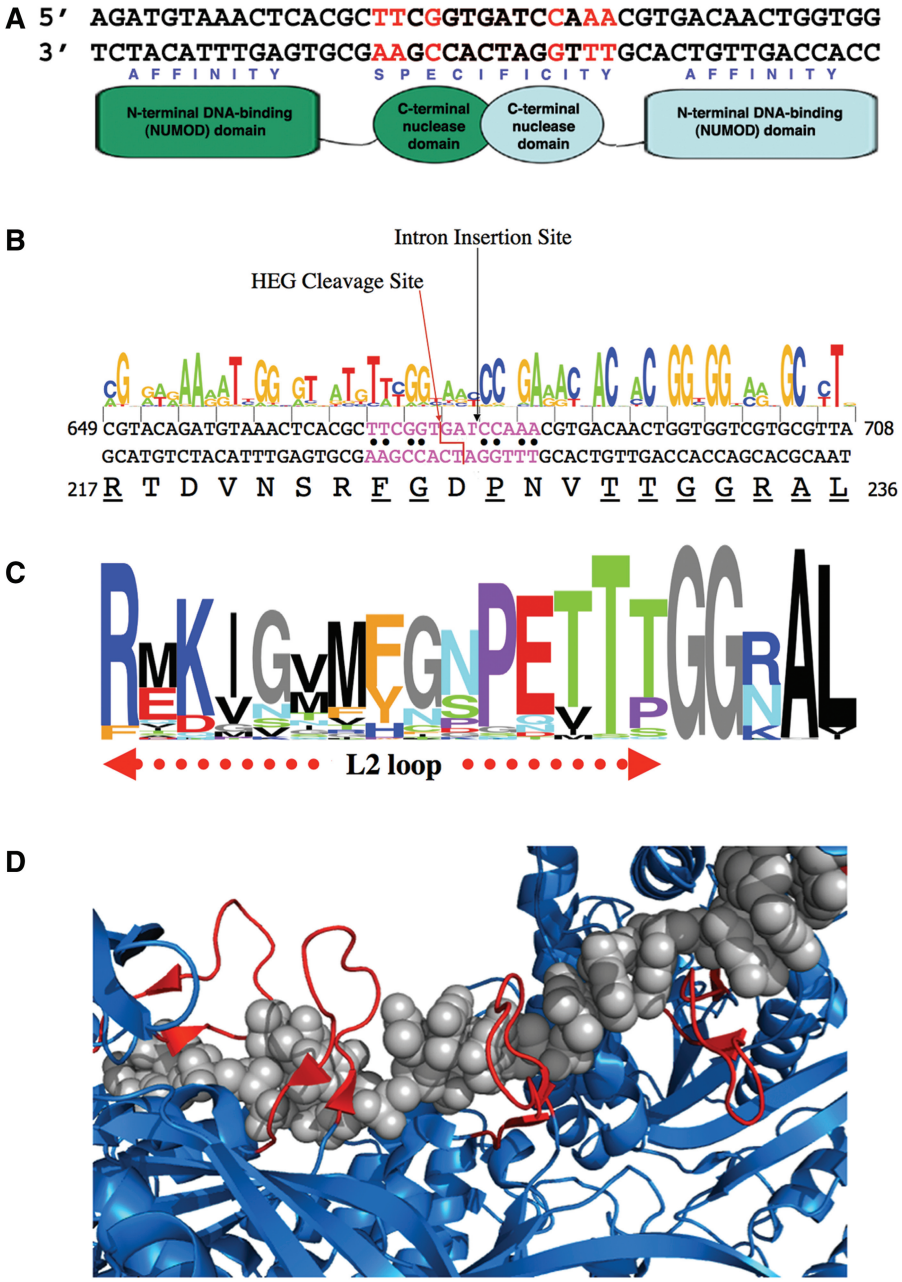

Figure 8. Conservation of the RecA DNA cleavage and intron insertion site and the RecA protein sequence. (**A**) Cartoon of the proposed domain architecture and DNA contact pattern exhibited by I-Bth0305I. The sequence (60 bp in length) corresponds to the overall region of DNA protected by the bound enzyme in DNAse I digestion experiments. Red bases are those that are recognized most specifically by the enzyme. (**B**) A logo plot indicating conservation of 1368 recA genes as described in the text. (**C**) Corresponding logo plot of translated RecA protein sequences from the same collection of sequences. The sequence of the 0305ϕ8–36 bacteriophage recA gene and RecA protein are shown between the two logo plots, with the intron insertion site and HEG insertion and cleavage sites on each DNA strand indicated. The recA coding sequence shown in this figure corresponds to the 60-bp region protected by bound I-Bth0305I in the DNAseI footprint experiment (Figure 3). The purple bases are the central 14 bp that display the most significant sequence specificity in cleavage assays. Black bullets between sequences of the two strands indicate the bases with palindromic symmetry between left and right DNA half-sites (also shown in Figure 2). The protein residues in the bacteriophage protein that correspond to the most conserved residues in the RecA proteins logo are underlined. The RecA L2 DNA-binding motif is indicated beneath the logo in panel b. (**D**) Structure of the *E. coli* RecA protein, bound as a filament on a single-stranded DNA target (pdb ID 3CMU) (33). The ssDNA ligand is in grey, with the RecA filament in blue and L2 regions in red.

For example, a large bacterial superfamily (the DUF199 proteins) that is thought to be involved in transcriptional activation of genes involved in sporulation or other differentiation and growth processes have been shown to contain LAGLIDADG domains (35). The HNH catalytic motif is found in non-specific bacterial and fungal nucleases (36,37), and is also found in a wide range of DNA-acting enzymes including transposases, restriction endonucleases, polymerase editing domains and DNA packaging factors (38,39). The GIY-YIG catalytic motif

is found in several bacterial restriction enzymes (such as Eco29kI) (40) and enzymes involved in DNA repair and recombination (such as the UvrC base-excision repair endonucleases) (41). Finally, the bacterial homing endonuclease I-Ssp6803I is a PD...(D/E)xK endonuclease, which is the most common catalytic protein fold in type II restriction endonuclease systems (5).

The discovery of a new homing endonuclease lineage (11) as characterized in this study again illustrates an evolutionary relationship between modern-day homing endonucleases and distantly related bacterial proteins (in this case, between a bacteriophage-derived homing endonuclease and a DNA mismatch repair enzyme). The 'PD...(D/E)xK' motif observed in these proteins (SCOP family 3.72.1) has been greatly diversified during evolution, facilitating its use for many biological functions (42). It has been visualized many times in restriction endonucleases, as well as in a variety of other contexts, including tRNA-specific homing endonucleases and a variety of DNA repair enzymes. All known variants of this fold display at least two acidic residues, and usually at least one additional basic residue in the nuclease active site, forming the catalytic motif that catalyzes phosphoryl transfer reactions (43).

Vsr endonucleases (and presumably I-Bth0305I) display a type II restriction enzyme topology that has significantly diverged from the canonical 'PD...(D/E)xK' motif, including the use of an activated histidine as a general base (14). The I-Bth0305I homing endonuclease and its nearest cousins appear to have maintained most of the features of this unique active site arrangement, although at least one additional strongly conserved acidic residue in the active site region (a strongly conserved acidic residue at the position corresponding to Phe62 in Vsr) may indicate further subtle divergence in catalytic mechanism.

Finally, the predicted bipartite structure of the homing endonuclease described in this study leads us to the possibility that the nuclease domain, on its own, might offer a useful catalytic fold for use in artificial gene targeting nucleases. This technology involves the creation of artificial nucleases by appending a non-specific nuclease domain (almost always the catalytic domain of the FokI restriction endonuclease) to a DNA recognition and binding construct consisting of a tandem array of zinc fingers or TAL repeats (44,45). The isolation and characterization of an independently folded nuclease domain that (i) appears to display a moderate degree of sequence specificity directly at the site of cleavage and (ii) naturally dimerizes prior to DNA binding may allow the development of new types of gene targeting proteins with novel DNA cleavage properties that prove useful for certain biotechnology and genome engineering applications.

ACCESSION NUMBERS

Structure factor amplitudes and refined coordinates for the catalytic domain of I-Bth0305I have been deposited at the RCSB protein database under accession code 3R3P and designated for immediate release upon publication.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank members of the Stoddard lab (particularly Ryo Takeuchi and Brett Kaiser) and Geoff Wilson at New England Biolabs for invaluable advice and assistance on this project.

FUNDING

National Institutes of Health research (grant R01 GM49857 to B.L.S.); National Institutes of Health training grant appointment (T32 GM08268 to G.K.T.); Hermann and Lilly Schilling Foundation chair (to S.P.). Funding for open access charge: National Institutes of Health (grant R01 GM49857).

Conflict of interest statement. One of the authors (B.L.S.) is a founder of a startup company that conducts research on homing endonuclease and gene targeting proteins. The protein described in this study is the subject of a recent patent application for construction of novel gene targeting protein scaffolds.

REFERENCES

1. Stoddard,B.L. (2011) Homing endonucleases: from microbial genetic invaders to reagents for targeted DNA modification. *Structure*, **19**, 7–15.
2. Stoddard,B. and Belfort,M. (2011) Social networking between mobile introns and their host genes. *Mol. Microbiol.*, **78**, 1–4.
3. Chevalier,B., Monnat,R.J. Jr and Stoddard,B.L. (2005) In Belfort,M., Wood,D., Derbyshire,V. and Stoddard,B. (eds), *Homing Endonucleases and Inteins*, Vol. 16. Springer Verlag, Berlin, pp. 34–47.
4. Flick,K.E., Jurica,M.S., Monnat,R.J. Jr and Stoddard,B.L. (1998) DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-PpoI. *Nature*, **394**, 96–101.
5. Zhao,L., Bonocora,R.P., Shub,D.A. and Stoddard,B.L. (2007) The restriction fold turns to the dark side: a bacterial homing endonuclease with a PD-(D/E)-XK motif. *EMBO J.*, **26**, 2432–2442.
6. Shen,B.W., Landthaler,M., Shub,D.A. and Stoddard,B.L. (2004) DNA binding and cleavage by the HNH homing endonuclease I-HmuI. *J. Mol. Biol.*, **342**, 43–56.
7. VanRoey,P., Meehan,L., Kowalski,J.C., Belfort,M. and Derbyshire,V. (2002) Catalytic domain structure and hypothesis for function of GIY-YIG intron endonuclease I-TevI. *Nat. Struct. Biol.*, **9**, 806–811.
8. VanRoey,P., Waddling,C.A., Fox,K.M., Belfort,M. and Derbyshire,V. (2001) Intertwined structure of the DNA-binding domain of intron endonuclease I-TevI with its substrate. *EMBO J.*, **20**, 3631–3637.
9. Edgell,D.R. and Shub,D.A. (2001) Related homing endonucleases I-BmoI and I-TevI use different strategies to cleave homologous recognition sites. *PNAS USA*, **98**, 7898–7903.
10. Landthaler,M., Shen,B.W., Stoddard,B.L. and Shub,D.A. (2006) I-BasI and I-HmuI: two phage intron-encoded endonucleases with homologous DNA recognition sequences but distinct DNA specificities. *J. Mol. Biol.*, **358**, 1137–1151.
11. Dassa,B., London,N., Stoddard,B.L., Schueler-Furman,O. and Pietrokovski,S. (2009) Fractured genes: a novel genomic arrangement involving new split inteins and a new homing endonuclease family. *Nucleic Acids Res.*, **37**, 2560–2573.

12. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
13. Gonzalez-Nicieza,R., Turner,D.P. and Connolly,B.A. (2001) DNA binding and cleavage selectivity of the Escherichia coli DNA G:T-mismatch endonuclease (vsr protein). *J. Mol. Biol.*, **310**, 501–508.
14. Tsutakawa,S.E., Jingami,H. and Morikawa,K. (1999) Recognition of a TG mismatch: the crystal structure of very short patch repair endonuclease in complex with a DNA duplex. *Cell*, **99**, 615–623.
15. Sitbon,E. and Pietrokovski,S. (2003) New types of conserved sequence domains in DNA-binding regions of homing endonucleases. *Trends Biochem. Sci.*, **28**, 473–477.
16. Bailey,T.L., Williams,N., Misleh,C. and Li,W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.
17. Schuler,G.D., Altschul,S.F. and Lipman,D.J. (1991) A workbench for multiple alignment construction and analysis. *Proteins*, **9**, 180–190.
18. Subramanian,A.R., Weyer-Menkhoff,J., Kaufmann,M. and Morgenstern,B. (2005) DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, **6**, 66.
19. Frith,M.C., Saunders,N.F., Kobe,B. and Bailey,T.L. (2008) Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput. Biol.*, **4**, e1000071.
20. Henikoff,S., Henikoff,J.G., Alford,W.J. and Pietrokovski,S. (1995) Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene*, **163**, GC17–GC26.
21. Henikoff,J.G., Greene,E.A., Pietrokovski,S. and Henikoff,S. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.*, **28**, 228–230.
22. Pietrokovski,S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.*, **24**, 3836–3845.
23. Adams,P.L., Stahley,M.R., Kosek,A.B., Wang,J. and Strobel,S.A. (2004) Crystal structure of a self-splicing group I intron with both exons. *Nature*, **430**, 45–50.
24. Sandegren,L., Nord,D. and Sjoberg,B.M. (2005) SegH and Hef: two novel homing endonucleases whose genes replace the mobC and mobE genes in several T4-related phages. *Nucleic Acids Res.*, **33**, 6203–6213.
25. Belfort,M. and Perlman,P.S. (1995) Mechanisms of intron mobility. *J. Biol. Chem.*, **270**, 30237–30240.
26. Edgell,D.R., Stanger,M.J. and Belfort,M. (2003) Importance of a single base pair for discrimination between intron-containing and intronless alleles by endonuclease I-BmoI. *Curr. Biol.*, **13**, 973–978.
27. Nomura,N., Nomura,Y., Sussman,D., Klein,D. and Stoddard,B.L. (2008) Recognition of a common rDNA target site in archaea and eukarya by analogous LAGLIDADG and His-Cys box homing endonucleases. *Nucleic Acids Res.*, **36**, 6988–6998.
28. Scalley-Kim,M., McConnell-Smith,A. and Stoddard,B.L. (2007) Coevolution of homing endonuclease specificity and its host target sequence. *J. Mol. Biol.*, **372**, 1305–1319.
29. Zhao,L., Pellenz,S. and Stoddard,B.L. (2008) Activity and Specificity of the Bacterial PD-(D/E)XK Homing Endonuclease I-Ssp6803I. *J. Mol. Biol.*, **385**, 1498–1510.
30. Gott,J.M., Zeeh,A., Bell-Pedersen,D., Ehrenman,K., Belfort,M. and Shub,D.A. (1988) Genes within genes: Independent expression of phage T4 intron open reading frames and the genes in which they reside. *Genes Dev.*, **2**, 1791–1799.
31. Goodrich-Blair,H., Scarlato,V., Gott,J.M., Xu,M.Q. and Shub,D.A. (1990) A self-splicing group I intron in the DNA polymerase gene of Bacillus subtilis bacteriophage SP01. *Cell*, **63**, 417–424.
32. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
33. Chen,Z., Yang,H. and Pavletich,N.P. (2008) Mechanism of homologous recombination from the RecA-ssDNA/dsDNA structures. *Nature*, **453**, 489–484.
34. Smith,D.R., Richterich,P., Rubenfield,M., Rice,P.W., Butler,C., Lee,H.M., Kirst,S., Gundersen,K., Abendschan,K., Xu,Q. *et al.* (1997) Multiplex sequencing of 1.5 Mb of the Mycobacterium leprae genome. *Genome Res.*, **7**, 802–819.
35. Knizewski,L. and Ginalski,K. (2007) Bacterial DUF199/COG1481 proteins including sporulation regulator WhiA are distant homologs of LAGLIDADG homing endonucleases that retained only DNA binding. *Cell Cycle*, **6**, 1666–1670.
36. Friedhoff,P., Franke,I., Meiss,G., Wende,W., Krause,K.L. and Pingoud,A. (1999) A similar active site for non-specific and specific endonucleases. *Nat. Struct. Biol.*, **6**, 112–113.
37. Kuhlmann,U.C., Moore,G.R., James,R., Kleanthous,C. and Hemmings,A.M. (1999) Structural parsimony in endonuclease active sites: should the number of homing endonuclease families be redefined? *FEBS Lett.*, **463**, 1–2.
38. Dalgaard,J.Z., Klar,A.J., Moser,M.J., Holley,W.R., Chatterjee,A. and Mian,I.S. (1997) Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family. *Nucleic Acids Res.*, **25**, 4626–4638.
39. Mehta,P., Katta,K. and Krishnaswamy,S. (2004) HNH family subclassification leads to identification of commonality in the His-Me endonuclease superfamily. *Protein Sci.*, **13**, 295–300.
40. Ibryashkina,E.M., Zakharova,M.V., Baskunov,V.B., Bogdanova,E.S., Nagornykh,M.O., Den'mukhamedov,M.M., Melnik,B.S., Kolinski,A., Gront,D., Feder,M. *et al.* (2007) Type II restriction endonuclease R.Eco29kI is a member of the GIY-YIG nuclease superfamily. *BMC Struct. Biol.*, **7**, 48.
41. Kowalski,J.C., Belfort,M., Stapleton,M.A., Holpert,M., Dansereau,J.T., Pietrokovski,S., Baxter,S.M. and Derbyshire,V. (1999) Configuration of the catalytic GIY-YIG domain of intron endonuclease I-TevI: coincidence of computational and molecular findings. *Nucleic Acids Res.*, **27**, 2115–2125.
42. Kosinski,J., Feder,M. and Bujnicki,J.M. (2005) The PD-(D/E)XK superfamily revisited: identification of new members among proteins involved in DNA metabolism and functional predictions for domains of (hitherto) unknown function. *BMC Bioinformatics*, **6**, 172.
43. Pingoud,A. and Jeltsch,A. (2001) Structure and function of type II restriction endonucleases. *Nucleic Acids Res.*, 3705–3727.
44. 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.
45. Porteus,M.H. (2006) Mammalian gene targeting with designed zinc finger nucleases. *Mol. Ther.*, **13**, 438–446.