Evolutionary Relationships, Design, and Biochemical Characterization of Homing Endonucleases

Gregory K. Taylor

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

2011

Barry L. Stoddard, Chair

Steven M. Hahn

Ning Zheng

Program Authorized to Offer Degree: Molecular and Cellular Biology

University of Washington

Abstract

Evolutionary Relationships, Design, and Biochemical Characterization of Homing Endonucleases

Gregory K. Taylor

Chair of the Supervisory Committee:
Professor Barry Stoddard
Fred Hutchinson Cancer Research Center

Homing endonucleases, found in all forms of microbial life, facilitate the invasion of host genes often in concert with introns or inteins by generating double stranded breaks in conserved coding sequences. There are five homing endonuclease families distinct in their structural characteristics and each family appears to share a common ancestor with diverse host proteins of unrelated function. Such related proteins include restriction endonucleases, DNA mismatch repair proteins, transcription factors, four-way junction resolving enzymes, and colicins. Homing endonculeases are currently being computationally redesigned for applications in genome engineering and structures of three redesigned homing endonuclease variants are described. In these experiments, crystal structures uncovered unexpected shifts in the DNA backbone relative to the wild type endonucleases and have thus been informative in the redesign process. Recently, a sixth homing endonuclease family homologous to E. Coli DNA repair protein VSR was discovered. A series of biochemical and x-ray crystallographic experiments investigating binding specificity and catalytic mechanism of a representative family member I-Bth0305I are described. Finally, a database archiving experimentally characterized homing endonucleases and a web-base program supporting homing endonuclease target site search are discussed.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

Table Number Page

ACKNOWLEDGMENTS

The author thanks 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. The author also thanks his advisor Barry Stoddard and doctoral supervisory committee members David Baker, Phil Bradley, Steve Hahn, Roland Strong, and Ning Zheng for advice and supervision. Experiments described in this dissertation were conducted at the Fred Hutchinson Cancer Research Center and the University of Washington. The author was supported by the University of Washington Molecular Biophysics training grant (T32 GM08268).

DEDICATION

for my wife, Soojin.

Chapter 1

STRUCTURAL, FUNCTIONAL AND EVOLUTIONARY RELATIONSHIPS BETWEEN HOMING ENDONUCLEASES AND HOST PROTEINS

This chapter is intended for publication in Nucleic Acids Research by authors GK Taylor and BL Stoddard.

Homing endonucleases (HEs) are highly specific, DNA cleaving enzymes that are encoded by invasive DNA elements (usually mobile introns and inteins) within the genomes of phage, bacteria, archea, protists and eukaryotic organelles. At least six diverse structural HE families, spanning four distinct nuclease catalytic motifs, have been characterized. In every known example, homing endonucleases display obvious structural homology to a variety of host proteins, many of which are found in bacteria. The biological functions of those related proteins are highly disparate and include nonspecific DNA degradation enzymes, restriction endonucleases, DNA repair enzymes, resolvases, intron splicing factors, and transcription factors. These relationships indicate that modern day homing endonucleases share ancient common ancestors with a wide variety of host proteins that are involved in genomic maintenance, fidelity and gene expression. This chapter summarizes the results of a large number of recent structural studies of homing endonucleases and host proteins that have illustrated the manner in which these proteins and activities are related.

1.1 Homing Endonucleases and Related Host Proteins

Homing endonucleases (HEs) are mobile genetic elements that selfishly propagate themselves in a dominant non-Mendelian fashion [30]. These proteins generally display no biological role other than to advance their own genetic coding sequence through a mechanism that is initiated by cleavage of a specific genomic target. DNA cleavage by the HE stimulates

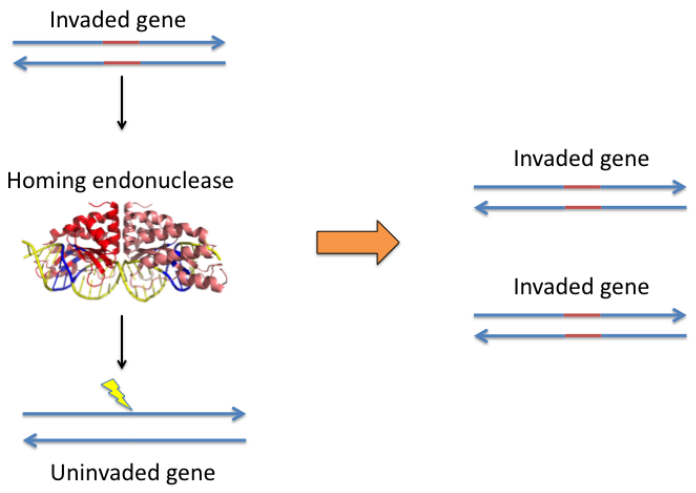


Figure 1.1: Homing endonucleases are inherited in a dominant non-Mendelian fashion. The homing endonuclease, encoded by the invaded gene, targets the uninvaded host gene and generates a double stranded break. This event stimulates homologous recombination which uses the invaded gene as a repair template. When the process is resolved, the homing endonuclease has successfully replicated itself.

break repair via homologous recombination, which results in precise insertion of the homing endonuclease reading frame (often in concert with surrounding intron or intein sequence) into the DNA target site. At least six distinct structural families of homing endonucleases (the 'LAGLIDADG', 'HNH', 'His-Cys box', 'GIY-YIG', 'PD-(D/E)xK', and most recently discovered 'EDxHD' proteins) have been identified [113, 120]. Each is classified and named according to the presence of a conserved sequence motif that corresponds to the conservation of critical structural and catalytic residues. These six HE structural families span at least four unique catalytic motifs that are widely associated with nuclease activities. The HNH and His-Cys box enzymes share a common "$\beta\beta\alpha$-metal" catalytic site [38], while the PD-(D/E)xK and EDxHD endonucleases also appear to be distantly related [120, 123, 124].

Despite wide variations in HE structure and mechanism, which corresponds to an equally wide range of genomic and biological hosts, all homing endonucleases must meet similar functional requirements [113]. They are generally encoded by relatively short reading frames (less than 1kB), presumably to minimize interference with the folding and function of their surrounding mobile elements (which are often self-splicing introns or inteins). Their DNA recognition behaviors usually involve the readout of long DNA targets that range from about 14 to over 30 base pairs in length, while simultaneously accommodating poorly conserved base pairs in their host target sites (such as wobble positions in protein coding sequences). This combination of DNA recognition properties allows a homing endonuclease to achieve sufficient specificity to avoid imposing significant toxicity to its current host, while also facilitating its continued vertical inheritance and persistence during the evolution of future generations of its host.

The evolutionary origin of the first homing endonuclease system is unknown, and the precise evolutionary mechanism by which any of the modern homing endonucleases families were generated is not particularly well understood. However, bioinformatic and structural studies of representatives from each unique homing endonuclease lineage have repeatedly demonstrated that they share common structural folds, and often an underlying mechanism of DNA binding and hydrolysis, with many host proteins that are involved in a wide variety of biological functions and pathways.

In this review, we summarize the results of a variety of high-resolution structural studies that have illustrated the various manners in which individual homing endonuclease families are related to host proteins of different biological and molecular functions. Implicit in this summary is the hypothesis there are at least two evolutionary scenarios by which such relationships might have been established. In the first, a modern HE family and one or more host proteins might simply represent the products of divergence from a common ancestor. In the second, an established homing endonuclease might have acquired a secondary biological function (for example, the ability to act as a 'maturase' and thereby facilitate intron splicing). In some cases, this may have resulted in the loss of the original HE function, presumably because the host-specific biological role became the primary target of selective pressure to maintain the protein's form and function.

1.2 Colicins

Escherichia coli and many other bacterial species produce and release a family of antibacterial cytotoxins named colicins under various conditions of stress [70]. Colicins are believed to confer an advantage in the presence of competing bacterial organisms when nutrients are limited or the cell is otherwise challenged by exposure to UV light or DNA damaging reagents. Separate colicin domains are involved in three separate stages required for cell killing: receptor binding, membrane translocation and toxin activity. The N terminal colicin domain is usually responsible for translocation, the central domain affects receptor binding, and the C terminal domain is often the active cytotoxic agent. To protect against self-cytoxic activity, cells producing colicins often co-produce an inhibitor protein that sequesters this cytotoxic domain until release from the host. Once the colicin has been introduced into the cytoplasm of the target cell, the cytotoxic domain kills the target cell using one of several mechanisms that include a highly specific RNAse activity, depolarization of the cytoplasmic membrane, inhibition of murein synthesis, or (in the case discussed below) non-specific DNAse activity.

The active sites of monomeric DNAse colicins contain an HNH nuclease motif as is observed within crystal structure of colicins E7 and E9 [29]. The residues of the HNH motif are found in a concave crevice in the surrounding protein fold that is believed to providing space for binding of double-stranded DNA in a sequence non-specific manner. Several of the residues in the active site of these enzymes coordinate a single divalent metal ion that is required to stabilize the phosphoanion transition state and the $3'$ oxygen leaving group of the reaction. An absolutely conserved histidine residue acts as a general base for the reaction, specifically to activate a water nucleophile. The active sites of bacterial colicins, as well as nonspecific microbial endonucleases such as the secreted nuclease from Serratia marcencens, were observed to display similar architectures to the active site of the Physarum polycephalum His-Cys box homing endonuclease I-PpoI. Comparative structural analyses between those nucleases offered the first suggestions that HNH and His-Cys-box nucleases were related by a common ancestor and related catalytic mechanisms [38]. While the colicins display relatively small domain architectures, I-PpoI contains several structural elaborations beyond the HNH motif and associated $\beta\beta\alpha$-metal core fold that are required for dimerization and for sequence-specific DNA recognition.

The observation that the HNH nuclease motif is broadly distributed across both homing endonucleases and a variety of distantly related host proteins was further illustrated by the subsequent determination of the DNA-bound crystal structure of the phage-derived homing endonuclease I-HmuI [103]. Unlike I-PpoI [41], that enzyme and a large number of related phage HEs display monomeric structures in which their HNH catalytic nuclease domains are tethered to independent DNA-binding regions via an overall protein domain organization that is unique from either bacterial colicins or the His-Cys-box homing endonucleases.

1.3 Restriction-modification

Bacterial genomes contain a wide variety of genetic systems that are believed to act biologically to protect their hosts against phage infections, as well as other possible sources of incoming foreign DNA [69]. The best studied of these correspond to restriction-modification (RM) systems, which are evolved around restriction endonuclease (REase) enzymes that recognize short nucleotide sequences in double stranded phage DNA with exceptional fidelity [84]. Many, if not all, bacterial genera possess multiple RM systems [69]; in each one the restriction endonuclease acts in concert with a cognate DNA modification activity that chemically modifies the same target sequence within the host genome (usually via base methylation within the same target site sequence) so that cleavage is effectively blocked.

R-M enzyme systems are classified according to their subunit composition and their mechanism of recognition and action on DNA [12]. Class I and III restriction endonucleases are large multisubunit assemblages that physically tether DNA target recognition, cleavage and methylation activities into large molecular assemblages that also contain and require ATP-dependent translocation for overall activity. In contrast, the class II R-M systems are considerably smaller and do not require ATP hydrolysis or the action of motor proteins for DNA cleavage or modification. In most (but not all) Class II systems, the REases act independently of their cognate methyltransferase (MTase) to cleave their specific DNA targets. Several thousand of class II restriction endonucleases have been biochemically characterized [91], and many more have been identified during the course of microbial genomic sequencing and annotation efforts around the world.

In contrast to homing endonucleases, restriction endonucleases usually recognize short sequences (generally 4 to 8 base pairs in length) with high fidelity [91]. A large number of crystallographic analyses of various type II REase/DNA complex have demonstrated that typically the restriction endonucleases contacts the target DNA sequence with 15 to 20 directional hydrogen bonds that specifically participate in recognition of the individual bases through the major and/or the minor groove [77].

In addition to their fundamental protective role in the bacterial host, the genes encoding at least some restriction endonucleases and their associated modification enzymes have also been proposed to act as selfish DNA [83]. According to this theory, loss of the modification activity leads to cell death via residual activity of the restriction enzyme, and thereby imposes a form of negative selection against elimination of R-M systems.

The majority of well characterized restriction endonucleases belong to the PD-(D/E)xK structural superfamily. Despite their low sequence similarity, it has been proposed that PD-(D/E)xK type II restriction endonucleases are descended from a common ancestor by divergent evolution [40]. As expected, the active site and the recognition site are the most structurally conserved regions in PD-(D/E)xK endonucleases. In general, restriction endonucleases appear to undergo rapid divergence and different restriction endonuclease families exhibit very little sequence similarity [13].

The I-Ssp6803I homing endonuclease (sometimes referred to with an abbreviated 'I-SspI' name for ease of description) was the first homing endonuclease to be shown to contain a PD-(D/E)xK core fold and to resemble REases from that family [86, 132]. This homing endonuclease and its close homologues are generally encoded in cyanobacteria. The enzyme forms a tetramer in solution; upon sequence recognition, two subunits make contact with the DNA while the other two provide additional quaternary structural interactions that allow organization of the protein on its long DNA target. This allows the homing endonuclease to recognize a pseudopalindromic target sequence consisting of 23 bp in length. Relative to the type II REases that have been visualized crystallographically, I-Ssp6803I is particularly closely related to the R.PvuII enzyme, with an RMSD of 3.3 [132] (Figure 1.2, top). Despite their similar size and architectures, DNA target site recognition by the two enzymes is obviously highly diverged, with I-SspI recognizing a 23 bp target with variable degrees of fidelity at individual DNA base pairs, in contrast to recognition of a 6 base pair target with absolute fidelity by R.PvuII. Of note, I-SspI makes approximately the same number of nucleotide specific contacts as PvuII does to its target.

In addition to the PD-(D/E)xK REase enzyme superfamily, a variety of type II restriction enzymes are known to contain either the GIY-YIG or the HNH catalytic core motifs [60, 63, 110]. The DNA-bound structures of the GIY-YIG restriction endonucleases R.Eco29kI and R.Hpy188I have been solved [110, 78, 109], which has allowed direct comparisons with the structure and proposed catalytic mechanism of the GIY-YIG homing endonuclease I-TevI (Figure 1.2, middle). The catalytic core of a GIY-YIG endonuclease follows a "β-β-α-β-α" topology where the first two β strands contain the residues GIY and YIG. R.Eco29kI has an extended DNA-binding loop immediately after the second β strand as well as a unique helix inserted between the first two β strands. This unique helix lies on the surface of the protein, distant from both the active site and the bound DNA; it appears to have a purely structural role in the protein fold and does not directly participate in the site of catalysis. Five conserved catalytic residues are all found within this core domain: Y49 from β1, Y76 from β2, H108 and R104 from α3, and E142 from α4. The sequence identity between the catalytic core domain of R.Eco29kI and the nuclease domain of I-TevI is 12% and the structure superposition has an rmsd of about 2.9 $\AA$ for backbone atoms. Either Y49 or Y76 in the GIY-YIG catalytic motif of R.Eco29kI might act indirectly or directly as a general base in the reaction, or one residue might satisfy the catalytic requirement when one of them is mutated.

A similar variety of REases, including R.PacI, R.Hpy99I, and R.KpnI belong to the HNH structural family [63, 110, 109, 102, 98]. These restriction endonucleases are all homodimers containing one -metal motif per subunit. Similar to the I-PpoI homing endonuclease, the DNA-bound cocrystal structures of R.PacI and R.Hpy99I indicate that those two enzymes are homodimers that contain two bound zinc ions per protein subunit; however all three enzymes have evolved different additional structural elaborations around their active sites and equally unique DNA binding modes. Whereas the I-PpoI enzyme recognizes a 14 base pair target site, again with moderate fidelity at several positions, the restriction enzymes

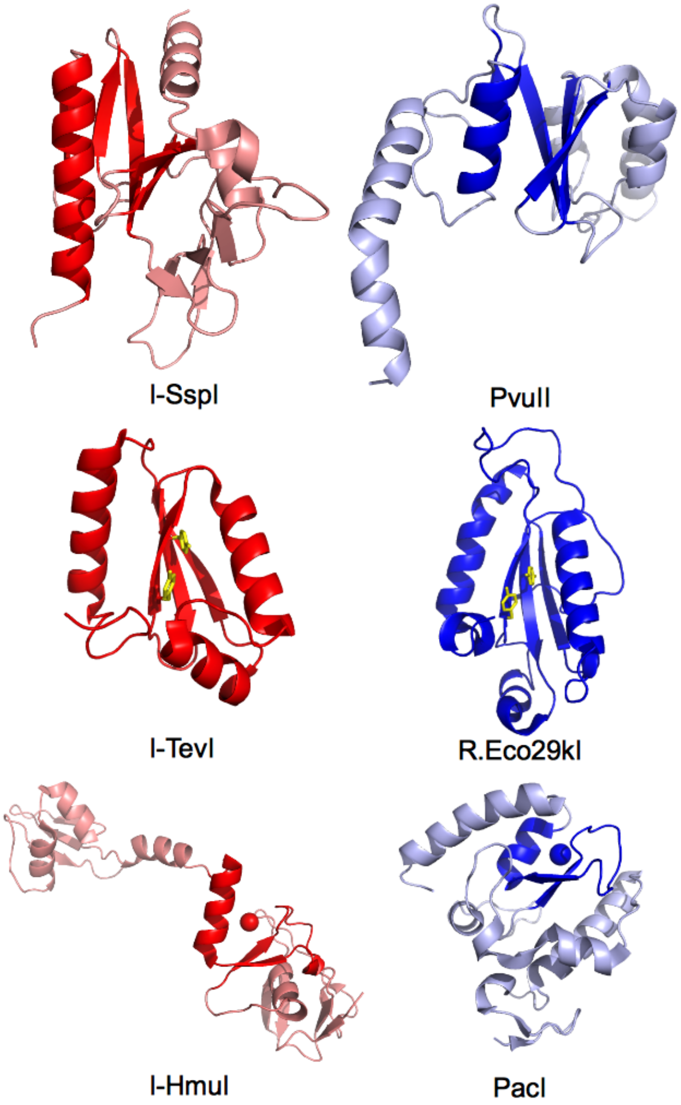


Figure 1.2: The PD-(D/E)XK, GIY-YIG, and HNH homing endonuclease families are all related to restriction endonucleases. The homing endonuclease I-SspI shares a core catalytic motif embedded in an alpha helix and beta sheet that is also observed in the restriction endonuclease PvuII (top). The catalytic domain of homing endonuclease I-TevI is structurally similar to restriction endonuclease R.Eco29kI where catalytic tyrosines are colored yellow (middle). Finally, the catalytic -metal motif found in homing endonuclease I-HmuI is also found in restriction endonuclease PacI (bottom).

recognize considerably shorter target sites with absolute fidelity. The heart of the Hpy99I protein forms a structure that wraps around its target site, aligning the helices from the catalytic site $\beta\beta\alpha$-metal motif almost perpendicular with the DNA duplex axis. In contrast, PacI binds via an elongated fold. In that structure, two subunits and the $\beta\beta\alpha$-metal motif aligned almost parallel to the DNA duplex. Based on these observations, these site-specific HNH endonucleases probably descended from a common $\beta\beta\alpha$-metal ancestor but are distantly related. Active site details and organization of PacI also indicate a significant divergence from the unusual architecture and mechanism that is observed for an HNH active site. First, a tyrosine side chain occupies a position usually inhabited by an imidazole base and a nucleophilic water. Second, there is a requirement of a tyrosine phenolic oxygen for catalysis. Together, these indicate that this side chain might act as a direct nucleophile in DNA strand cleavage although the more traditional mechanism involving water-mediated hydrolysis cannot be ruled out.

1.4 DNA repair

1.4.1 Nucleotide excision functions

UvrABC is a multienzyme complex found in E. coli and other bacteria that is involved in 'short patch' nucleotide excision repair in response to DNA damage at individual bases. The sequence of events in the UvrABC-mediated damage recognition and nucleotide excision reaction are relatively well established [126]. First, UvrA dimerizes through an interaction with ATP. The dimer UvrA2 interacts with UvrB in solution forming a stable complex with either one or two copies of UvrB per complex. Upon binding, UvrA first contacts DNA which it then transfers to DNA binding domain on UvrB. This complex then scans each strand of DNA in search of recognizable DNA adducts. Once a damaged strand has been encountered, it is bent and wrapped around one molecule of UvrB. It is thought that upon lesion recognition, UvrA hydrolyzes ATP which promotes self-dissociation leaving a UvrB:DNA complex. UvrB utilizes bound ATP energy applied by the -hairpin region of UvrB in order to impose an unfavorable DNA conformation, thereby enabling binding and phosphoryl hydrolysis by UvrC. Binding allows UvrC to catalyze the two incision reactions.

UvrC is weakly constitutively expressed resulting in a cell copy number of 10-20. UvrC mediates two strand scission events on the same DNA strand, with one cleavage event located nucleotides $3'$ of the lesion, and the second eight nucleotides $5'$ to the lesion. The two strand cleavage events generate a 12-nucleotide fragment of DNA with the lesion roughly in the middle. After incision, DNA helicase II (UvrD) releases UvrC and the excised oligonucleotide. DNA polymerase I then resynthesizes the excised strand and removes UvrB from the non-damaged DNA strand in the process. DNA ligase I joins the synthesized DNA to the template finishing the nucleotide excision repair pathway.

Bioinformatic analyses and homology searches using the sequence of E. coli UvrC revealed a bacterial homolog named Cho [126]. This protein is homologous to the N-terminal region of UvrC and can initiate $3'$ DNA strand cleavage, but not $5'$ cleavage. As previously demonstrated for UvrC, Cho is also dependent on UvrAB but UvrC and Cho interact with different UvrB domains. Cho and UvrC are both encoded in several bacterial species including E. coli, but the greater majority of bacteria contain only a recognizable copy of UvrC. In some species such as mycoplasmas and Borrelia burgdorferi only Cho is found. In these cases, a $5'$ strand cleavage activity might originate from an additional exonuclease domain found on Cho or from the exonuclease activity of an alternative enzyme. This may be plausible as Cho proteins of the Mycobacterium species are larger than that of E. coli.

The nucleotide excision repair proteins UvrC and Cho shares homology with the catalytic domain of the GIY-YIG family of homing endonucleases [122]. The two proteins roughly follow a structural motif of $\alpha1$-$\beta1$-$\beta2$-$\alpha2$-$\alpha3$-$\beta3$-$\alpha4$-$\alpha5$ (Figure 1.3, top). At the center of each globular structure is a sheet that contains the GIY-YIG catalytic motif on $\beta1$ and $\beta2$. The catalytic domain of UvrC and the catalytic domain of I-TevI have relatively low sequence identity of 15 %. Given their low sequence identity, it is notable that the two structures superimpose with an rmsd of 2.2 $\mathring{A}$ for 60 of 89 possible C atoms. While the two structures have a nearly identical topology, there are clear differences in their secondary and tertiary structure. First, an additional helix, $\alpha1$, is present in the UvrC structure compared to I-TevI. This helix is likely structural and appears to not be involved in catalysis, because residues that form the helix are not conserved among various UvrC homologues. Second, the region spanning $\alpha2$ and $\beta3$, which includes $\alpha3$, is not structurally conserved compared to I-TevI. Nevertheless, a residue that stabilizes the hydrophobic core of the domain superimposes between the two structures (Ile45 from UvrC and Leu 56 in I-TevI). Finally, the terminal helix $\alpha5$ in the motif is found in neither I-TevI nor all UvrC homologs.

1.4.2 Mismatch repair functions

In the first step of DNA mismatch repair, MutS binds to base pair mismatches and to small insertion/deletion loops [92]. MutS is a functional heterodimer with one monomer binding the mismatch, and the other binding nonspecifically to the surrounding DNA. Each subunit also contains an ATPase domain that interacts with the DNA binding domain. The MutS-DNA-ATP complex then interacts with MutL which also binds DNA and ATP. Interaction of MutL with DNA is mediated primarily through MutS and occurs independently of ATP hydrolysis. ATP hydrolysis by MutL is then required for interaction with many of the downstream proteins required for completion of mismatch repair, one of which is termed the Very Short patch Repair protein or Vsr.

Unlike other mismatch repair proteins, Vsr recognizes mismatches in the context of a longer sequence. Through recruitment by MutL, this single strand endonuclease preferentially targets T/G mismatches within hemimethylated $5'$-CTWGG/$5'$CCWGG sequences where W is an A or a T (the $3'$C of CCWGG sequences is the substrate for the bacterial DNA cytosine methyltransferase (Dcm)) [93]. Vsr cleaves the DNA $5'$ of the mismatched T, so that after removal of downstream bases, DNA Polymerase I may perform templated DNA resynthesis, creating a short repair patch. DNA ligase then reintegrates the DNA patch into the DNA backbone.

In a recent analysis of environmental metagenomic sequence data collected by the Global Ocean Sampling project, a novel type of fractured gene was discovered corresponding to separately encoded halves of self splicing inteins that interrupt individual host genes in the same locus [25]. The inteins were frequently found to be interrupted by open reading frames that do not exhibit significant sequence similarity to previously characterized homing endonuclease families. Further analysis indicated that the uncharacterized open reading frames were

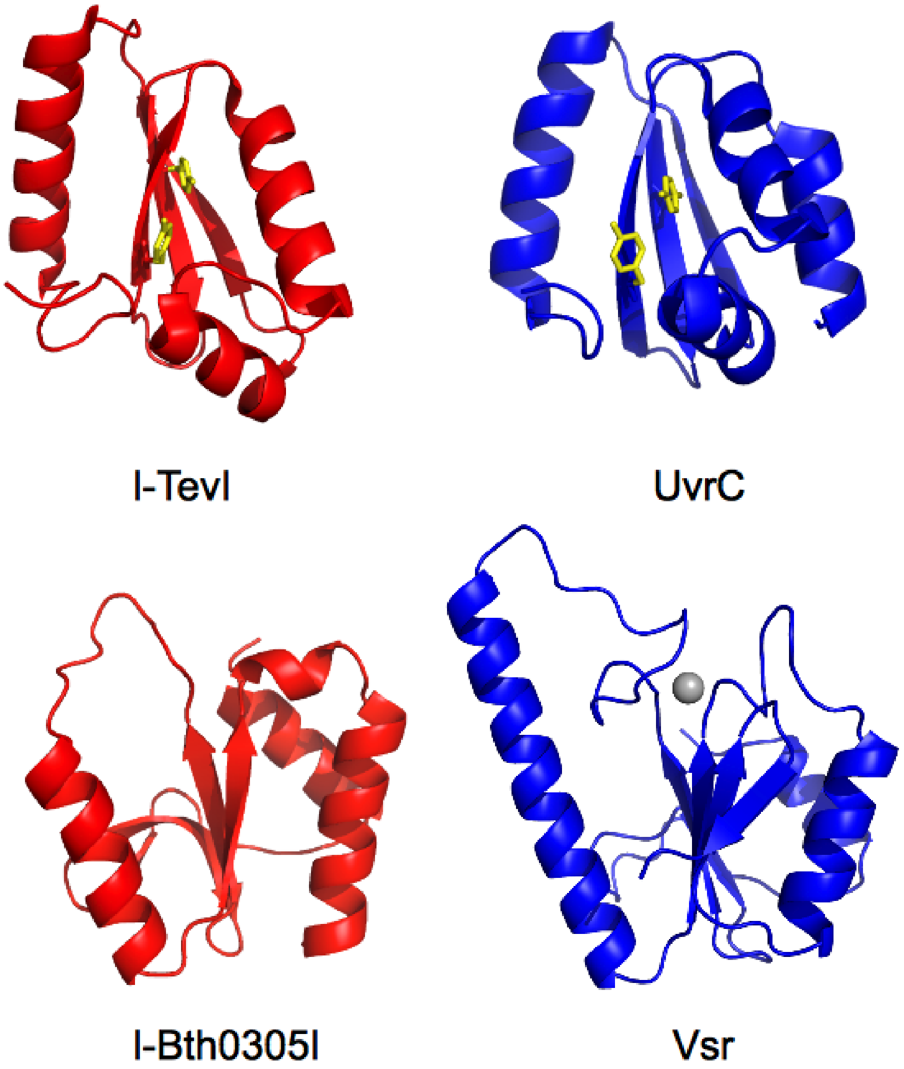


Figure 1.3: DNA repair enzymes participating in diverse and complex pathways are homologous to different families of homing endonucleases. The GIY-YIG homing endonuclease I-TevI is homologous to the DNA repair protein UvrC that participates in the nucleotide excision repair pathway (top). An entire family of homing endonucleases have been named in reference to homology with the nucleotide excision repair protein Vsr. The catalytic domain of homing endonuclease I-Bth0305I, a representative member of this family, is structurally similar to Vsr (bottom).

associated with introns, inteins, or as freestanding genes. In total fifteen members, including two in previously annotated genes in the NCBI sequence database, were described.

Limited sequence homology to the catalytic domain of the Very Short patch Repair (Vsr) endonucleases was detected in the C-terminal region of the translated protein sequences of these genes [25]. The established catalytic residues from Vsr endonucleases were conserved across all members of the new gene family. These residues include an essential aspartate that coordinates a catalytic magnesium ion, a histidine thought to act as a general base, and a proximal aspartate residue. Inferred from the presence of endonuclease catalytic residues within the domain, this gene family was hypothesized to encode a novel lineage of homing endonucleases. The activity, specificity, and structure have been characterized for one representative member of this family, I-Bth0305I [120]. The crystal structure of the catalytic domain support a similar mechanism for DNA strand cleavage and confirms that members of this homing endonuclease family share a common ancestor with the Vsr mismatch repair endonuclease (Figure 1.3, bottom).

Vsr endonucleases and the newly discovered homing endonuclease family (now named the 'EDxHD' homing endonucleases) display a type II restriction enzyme topology that has significantly diverged from the traditional PD-(D/E)xK motif and uses an activated histidine as a general base [120]. Further subtle divergence of catalytic mechanism is indicated by an additional highly conserved acidic residue in the active site region. Apart from these two exceptions, the enzyme has maintained most the features of this unique active site arrangement. The observed bipartite arrangement of the catalytic domain is not common with Vsr but the relationship between the two proteins is clear when comparing global topologies.

1.5 DNA resolvases

Four-way DNA (Holliday) junctions are branchpoints generated by the interconnection of four helices during strand exchange events that are necessary for various DNA integration, transposition and recombination processes [74]. Four way junctions are resolved by junction resolving enzymes to create duplex products. These nucleases are highly specific for the structure of DNA junctions where they initiate cleavage at the branchpoint of the junction.

Junction-resolving enzymes have been isolated from a number of different organisms ranging from bacteria, bacteriophages, archaea, yeast, and mammalian cells and their viruses.

In comparing the crystal structure of the I-Ssp6803I homing endonuclease to previously determined macromolecular structures, the most similar core fold corresponds to the archael Holliday junction resolving enzyme (Figure 1.4, top) [132]. Specifically, the Hjc enzyme from Pyrococcus furiosus aligns with an r.m.s.d. of 2.4 $\AA$ (1.9 $\AA$ across the catalytic core). Whereas I-SspI forms a tetramer to bind a long duplex DNA target, four-way junction resolving enzymes form a dimer to recognize the junction itself. This is accomplished through the creation of two DNA-binding channels that are 30 $\AA$ in length, formed on either side of the dimer. These channels are positively charged and make extensive contact with the arms containing the $5'$ ends of the continuous strands. This results in the burial of 4180 $\AA^2$ of solvent accessible protein surface and the channels hold the DNA arms in a perpendicular orientation [74]. The relationship of the catalytic core between a homing endonuclease and a four-way junction resolving enzyme suggests a common ancestor even with the different oligomeric state found in each of the two proteins.

1.6 Maturases and mating switch proteins

Whereas all of the examples provided above appear to represent situations where modern day homing endonucleases and contemporary host proteins have diverged from ancient common ancestors, there exists as least two cases where established homing endonuclease structures and function developed secondary biological activities and roles in the host, which in time led to the original invasive function giving way entirely to a unique host-specific role.

Many homing endonucleases can also participate in the post-transcriptional splicing of their host intron, by assisting the folding of their cognate RNA intron–a function termed 'maturase' activity [26, 130, 45, 56, 117, 58, 43, 76]. In some cases, such maturases have retained their original homing endonuclease activity and thus moonlight between both activities [15] where in other cases the homing endonuclease activity has been lost–in some cases through a single, presumably recent point mutation that can be easily reverted to restore endonuclease activity [117].

Finally, some homing endonucleases have been adopted by the host to act directly as

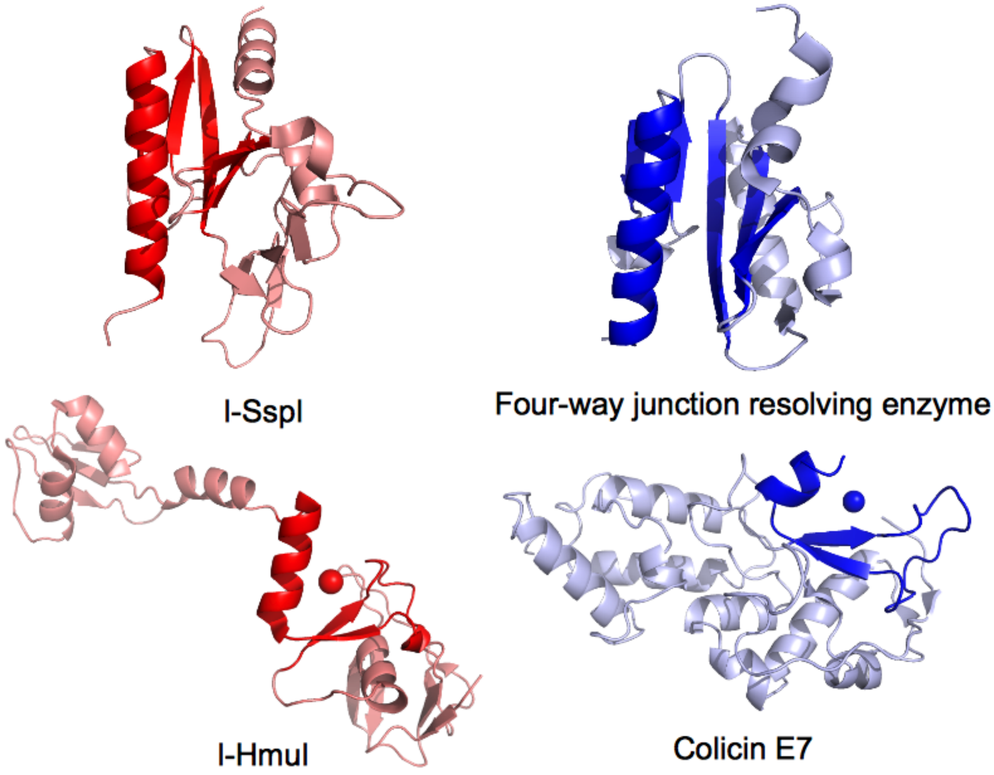


Figure 1.4: The PD-(D/E)XK motif is found in homing endonuclease I-SspI and in the four-way junction resolving enzyme. The core structural motif, which consists of an alpha helix and beta sheet, aligns between the two enzymes (top). Colicins are similar to the HNH family of homing endonuclease. A structural comparison of I-HmuI and Colicin E7 shows the core BBA-metal motif with different structural elaborations that coopt the catalytic mechanism for different contexts (bottom).

freestanding endonucleases that drive biologically important gene conversion events. For example, the HO endonuclease in yeast, which is responsible for the mating-type genetic switch in that organism, is a LAGLIDADG protein which appears to be derived from an intein-associated homing endonuclease [61].

1.7 Genetic regulation

The DNA binding properties of homing endonucleases appears to facilitate their ability to be utilized, either directly or as a result of evolutionary repurposing, as genetic regulators. For example, the I-TevI homing endonuclease moonlights as a transcriptional repressor, acting to suppress its own expression and thereby assist in reducing host toxicity in the presence of its reading frame and corresponding mobile DNA element [75, 33]. At least two examples have been described in the literature of more distant relationships between homing endonucleases and genetic regulators: the WhiA/DUF199 family of bacterial sporulation factors and the eukaryotic SMAD proteins.

1.7.1 WhiA/DUF199

The initiation of mRNA synthesis depends ultimately on factors that interact with specific elements in gene promoters [82]. Sequence specific DNA binding proteins attach to the control region in the immediate vicinity of a transcription start site called a promoter. These proteins are composed of a surprising variety of usually separable DNA binding and transcriptional activation domains. The DNA binding subregions of many transcription factors consist of 60 to 100 amino acids and are necessary but not sufficient for transcriptional activation. These regions are tethered to transcriptional activation domains that are required for the initiation of transcription, presumably through recruitment of RNA polymerase.

One family of putative bacterial transcription factors named DUF199 is present in all Gram-positive bacteria [3]. One representative member of this family, WhiA, was observed in bioinformatic and structural studies to contain a core LAGLIDADG sequence motif and corresponding fold and topology at its N-terminal region, tethered to a C-terminal helix-turn-helix domain [64, 62]. The WhiA protein is essential for sporulation in Streptomyces coelicolor and related Streptomycete strains, and appears to regulate expression of multiple sporulation-specific Whi genes [3]. Notably, WhiA regulates expression of its own reading frame and at least one other sporulation specific transcript (ParAB2), and appears to interact with and regulate the activity of the sporulation-specific sigma factor WhiG. All Gram-positive bacteria contain similar Whi operons including a single recognizable DUF199/WhiA protein. This conservation suggests that WhiA homologs function in a similar manner.

The similarities and differences between WhiA sequence and structure relative to its closest bacterial homologs and more distantly related LAGLIDADG homing endonucleases are displayed in Figure 1.5, top. Analysis of the structure elucidates how unique evolutionary pressures that are placed upon a genetic regulator versus those placed on an invasive endonuclease might produce individually tailored structures and biochemical features that are appropriate for each function. The protein fold topology observed in monomeric LAGLIDADG homing endonucleases is observed in the N-terminal region of WhiA. Monomeric LAGLIDADG homing endonucleases are composed of two structurally similar domains, each containing an $\alpha\beta\beta\alpha\beta\beta$ core that are connected by a short peptide linker. The closest structural homolog of WhiA, identified by the DALI webserver, is the I-DmoI homing endonuclease, which is an archael enzyme encoded within a mobile group I intron. The two sequences have low sequence identity of 13 % and the structures superimpose with an α-carbon r.m.s.d. across all aligned residues of 2.4 $\AA$ [62]. Conserved elements include those residues that comprise the two LAGLIDADG helices that form the core of the domain interface. Intimate packing between backbone atoms in the helices resulted in helices that are closely superimposable.

A key difference between LAGLIDADG homing endonucleases and WhiA family members is that the WhiA proteins lack acidic residues at the base of the LAGLIDADG helices that coordinate metal ions in homing endonucleases. In I-DmoI [104], these conserved residues correspond to D20 and E117 and are essential for catalysis. Other catalytic residues, such as K42 and K120 in I-DmoI, are not conserved in WhiA. These residues are basic residues that are involved in transition-state stabilization in homing endonucleases. These positions are occupied by a histidine and methionine (H54 and M125, respectively) in the WhiA structure and are similarly nonconserved in close homologs. As a consequence, WhiA

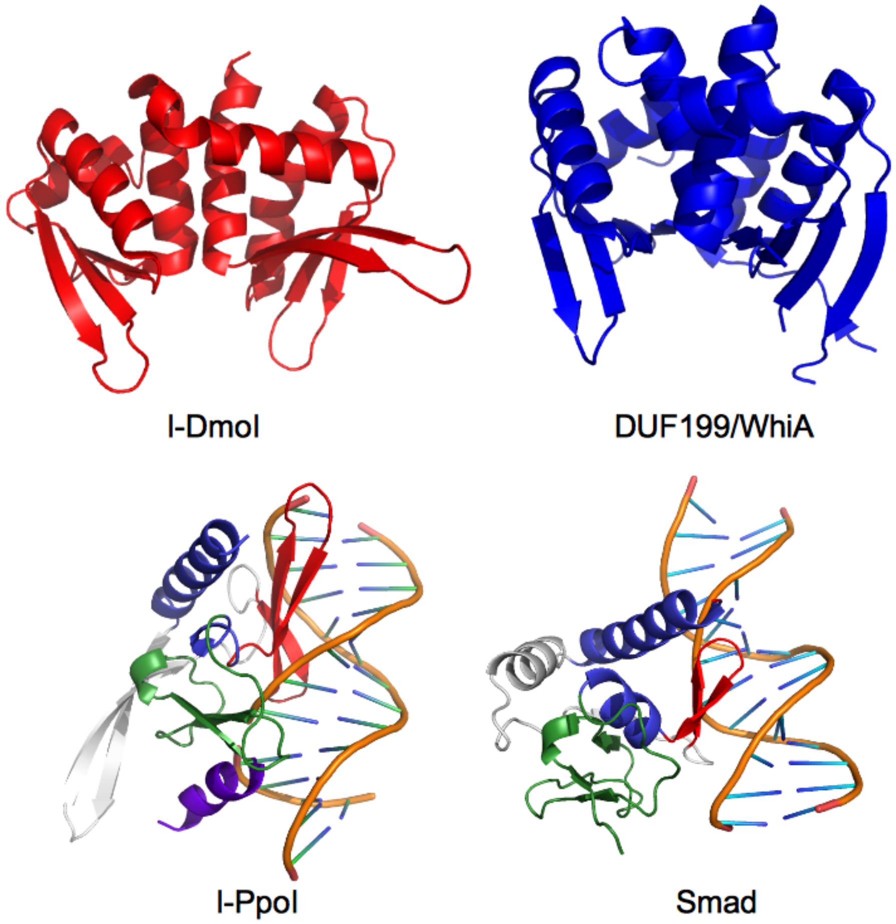


Figure 1.5: LAGLIDADG and His-Cys Box homing endonucleases are related to transcription factors. Both the homing endonuclease I-DmoI and the transcription factor share a common structural fold with a pseudodimeric structure with a pair of beta sheets joined at an interface between two central alpha helices (top). The homing endonuclease I-PpoI has a similar topology to the MH1 domain of the SMAD transcription factor. Highlighted in red are two beta strands that are involved in DNA recognition. Corresponding segments of each protein are colored similarly (bottom).

family members cannot be endonucleases and do not digest DNA in controlled experiments.

The mechanism of DNA recognition and binding by WhiA LAGLIDADG domains might differ significantly from that displayed by the same domains in the homing endonuclease. Enzymes such as I-DmoI make extensive contacts with their DNA substrates using a pair of antiparallel β sheets and associated loops. These structural elements make interactions with the DNA backbone with individual nucleotide base pairs across the entire DNA target. Each LAGLIDADG domain recognizes a single DNA half-site using DNA-contact surfaces that are uniformly positively charged. The only exception to this surface is the presence of conserved metal coordinating acid residues in the active sites at the center of the domain interface.

The surface of WhiA corresponding to the DNA-binding surface of the N-terminal domain in traditional LAGLIDADG homing endonuclease displays significant negative surface charge. Also, the C-terminal LAGLIDADG domain displays positively charged surface that extends well beyond its β sheet region. Consequently, the DUF199/WhiA protein family is expected to interact with its DNA target in a unique manner from the mode of DNA binding exhibited by LAGLIDADG homing endonucleases such as I-DmoI; it is quite possible that the LAGLIDADG domain in the WhiA/DUF199 family has entirely surrendered DNA binding function to the helix-turn-helix domain and is instead involved in protein-protein interactions required for its role as a gene expression regulator.

1.7.2 Smad Proteins

SMADs are intracellular proteins that are involved in transducing signals to the nucleus, in response to the presence of various growth factors, in order to activate expression of the TGF-beta gene [53]. The DNA binding domain of the Smad transcriptional regulator in the TGF-B signaling cascade has been found to resemble the overall topology of the His-Cys-Box homing endonuclease I-PpoI [49]. Smad consists of two domains, MH1 and MH2. The MH2 domain is homologous to a large family of nuclear signaling protein-protein interaction domains in eukaryotes and prokaryotes. A presumably unique spatial structure of the MH1 domain earned it a unique fold classification in the SCOP database. A combination of sequence and structure-based analyses show that the MH1 domain is homologous to the His-Cys-Box homing endonuclease family (Figure 1.5, bottom). The structural similarity was first detected by the DALI server with a 16 % sequence identitity and an r.m.s.d. of 3.3 $\mathring{A}$ between 78 aligned α-carbons [49].

The structural organization of I-PpoI follows a three subdomain architecture with two subdomains having structural equivalents in MH1 Smad. Notably, the first subdomain is a three-stranded β-sheet (colored red in Figure 1.5) that binds in the major groove of DNA; the turn between β strands incorporates the active site Arg61. Further, MH1 and I-PpoI have similar secondary structural elements in the same topological connection and spatial arrangement. From this global comparison, it is clear that they posses the same fold [49] and share a common ancestor.

1.8 Conclusions

Most enzymes involved in the catalysis of phosphodiester bonds are members of a relatively small number of protein structural families and span an even smaller number of nuclease catalytic motifs. The processes of phage restriction, nucleotide excision repair, DNA mismatch repair, Holliday junction resolution, and recombination are undertaken by families to which homing endonucleases are members of the PD-(D/E)xK, HNH and GIY-YIG enzyme families. Once a fold has evolved to catalyze a single or double stranded break in DNA, it may be coopted and repurposed into a number of different functions.

The PD-(D/E)xK endonucleases have highly divergent active site architectures. Consequently, these enzymes do not display a single uniform reaction mechanism. For example, a variety of residues and chemistries can be used for transition state stabilization and proton transfer, DNA cleavage may be enabled by different numbers of metal ions, and the position of metal binding sites may be moved [131]. By comparison, the structure and corresponding mechanism of GIY-YIG active sites appears to be quite strongly conserved (possibly because of the simultaneous participation of several motif residues in structural stabilization and in catalysis). Consequently, before divergence of the endonuclease family from their last common ancestor, the active site geometry was probably optimized and strongly fixed. The GIY-YIG endonuclease domains engage in highly disparate biological functions that include DNA invasion, defense, genomic degradation, and repair. In light of this functional diversity, the maintenance of their active sites is extraordinary: of the identity and position of six catalytically important residues, five are absolutely or strongly conserved [78]. The GIY-YIG domain has been less successful than several other nuclease superfamilies in adopting different functions, parasitizing different organisms, and spreading to new loci [31]. This suggests an inflexibility of the GIY-YIG fold.

The $\beta\beta\alpha$-metal HNH motif is highly modular and found in conjunction with a number of domains that diversify function. The motif is embedded with different domains that bind DNA specifically in the case of homing endonucleases or transport the protein to competing cells in the case of colicins. The motif has also been added to structural ameliorations that support oligomerization as is the case for the PacI restriction endonuclease. The LAGLI-DADG and His-Cys-Box motifs have lost their catalytic function but retained their DNA binding ability to become transcription factors. These proteins continue to bind specific DNA sequences, but were highly mutable in the absence of restrictions imposed by the catalytic domain. As a consequence, the MH1 domain of Smad is considerably diverged from the homing endonuclease I-PpoI. This style of protein evolution is unique to homing endonuclease folds that have comparatively more specific DNA binding regions. Despite their differences in structure, catalytic mechanism, and conserved sequence motifs all families of homing endonucleases are related to proteins of different function which suggests a common mechanism of evolution involving a comparably frequent switch in protein function.

Chapter 2

COMPUTATIONAL REPROGRAMMING OF HOMING ENDONUCLEASE SPECIFICITY AT MULTIPLE ADJACENT BASE PAIRS

This chapter was originally published in nucleic acids research by authors J Ashworth, GK Taylor, JJ Havranek, SA Quadri, BL Stoddard, and D Baker [9].

Homing endonuclease genes (HEGs) are mobile genetic elements found throughout the microbial universe. They are typically associated with self-splicing intervening sequences (IS; introns or inteins) that are capable of invading and persisting in host genomes, due in part to the site-specific DNA cleavage activity of the rare-cutting homing endonucleases that they encode [112]. Cleavage of a DNA site by the homing endonuclease results in copying of the HEG and the surrounding IS into the host genome through double-strand break repair via homologous recombination [11]. These properties and functions of homing endonucleases form the basis of new targeted genetic applications, including corrective gene therapy [5]. Delivery or expression of a HEG, along with a DNA repair template that is homologous to the DNA sequence surrounding the enzymes target, results in the repair or modification of the recipient allele for distances up to one kilobase on either side of the endonuclease cleavage site [23].

The potential sites of cleavage for these applications are primarily limited by the specificities (both natural and engineered) of available homing endonucleases. Multiple techniques can be used to generate homing endonuclease variants that display novel and specific cleavage activities, including mutagenic library selection and structure-based computational design [116, 8, 28, 108, 5, 16, 121]. These methods currently produce changes in specificity for a relatively small number of contiguous base pairs (one to three) that are then combined to access more distant target sites. If these redesigned regions are not adjacent or overlapping, they can be readily combined in a modular fashion to yield enzymes capable of cleaving new targets differing from the original wild-type site at many base pairs [96], allowing the repair or conversion of novel specific gene loci in vivo [5, 50, 42]. However, the extent to which separately optimized clusters of interactions that involve adjacent base pair substitutions and mutations at the same amino acid positions can be combined has yet to be determined. Furthermore, while high-throughput selection has yielded large numbers of new specificities, the extent to which computational methods can be used to rationally predict and design broad changes in specificity is as yet unknown.

To explore the feasibility of using structure-based computational methods to design novel specificity at multiple adjacent base pairs within a homing endonuclease recognition site, we employed a computational protein design approach [8, 51] to redesign I-MsoI [19] to specifically cleave a DNA sequence harboring three consecutive base pair changes relative to the wild-type site. To investigate the modularity of designed interactions at adjacent and overlapping positions, we compared the results of a concerted design for the entire three base pair cluster to the results of individual design for each single base pair substitution. The designed endonucleases were characterized and compared by assaying relative DNA cleavage efficiencies and specificities in vitro, and by X-ray crystallography of each protein-DNA complex. Finally, starting from the crystal structure of the triple base pair switch, we designed a further change in specificity, illustrating the power of iterating between computational design and experimental structure determination.

2.1 Computational design of specificity

The computational methodology for the prediction and redesign of homing endonuclease specificity has been described previously [8, 121]. A starting model was built using the atomic coordinates from the crystal structure of the wild-type I-MsoI endonuclease in complex with its un-cleaved native DNA recognition site [pdb code 1M5X [19]]. Nucleotide substitutions were modeled by superimposing the ideal coordinates of new nucleotides onto the backbone atoms of crystallographic nucleotides. The side chain conformations of all amino acids in the vicinity of the substituted nucleotides were allowed to reconfigure according to the Rosetta physics-based full-atom energy function. New combinations of amino acid identities were searched at those amino acid positions that were capable of directly contacting the substituted nucleotides. Positions were considered to be capable of contact if an arginine side chain at that position could be placed within 3.6 Å of any nucleotide base atom. Water-mediated contacts between protein and DNA were also searched by modeling water molecules attached to the major groove atoms of nucleotide bases. During the design for three simultaneous base pair substitutions, small shifts in the protein backbone were modeled using a loop-closure algorithm [14, 129]. The binding energies of all complexes were calculated by subtracting the energy of the bound complex from the sum of the energies of the separated protein and DNA.

For the individual base pair substitutions at positions ±8 and ±7, an algorithm was employed that directly optimizes the specificity of designed amino acids for the target DNA target site sequence [52, 121]. The energies of interaction between the protein and DNA (affinities) were computed for the target DNA site as well as for alternative DNA site sequences at the substituted base pairs. Using a genetic algorithm [52], a population of randomized amino acid identities at positions in contact with the substituted nucleotide positions was evolved in silico by enriching for combinations that maximized the discrimination between the target and alternative DNA sites. To excessive loss of affinity, amino acid combinations were disfavored if their affinities were more than 5-10 energy units worse than the best affinity found over all amino acid combinations. The optimal energy threshold for this criterion was estimated by recovery analysis of wild-type and previously-designed [8] interactions (data not shown). The specificities of all design models were calculated as a Boltzmann occupancy of the target complex, versus a partition function consisting of all competing single base pair variant sites [7].

2.2 Materials and Methods: Protein production and purification

Genes for the homing endonuclease designs were assembled by PCR from oligonucleotides, based on a DNAWorks [59] assembly that was codon-optimized for expression in Escherichia coli. 6X-His-tagged proteins were expressed in E. coli BL21-pLysS cells from a pET15 vector by auto-induction [114] at 18-22 ° C for 24 h. Proteins were purified by nickel affinity fast-performance liquid chromatography (FPLC). Protein purity and identity were verified by polyacrylamide gel electrophoresis (PAGE) and liquid chromatography mass spectrometry (LCMS), and their concentrations were determined by dividing absorbance at 280 nm by their predicted extinction coefficients (5500*Trp + 1490*Tyr + 125*Cys $M^{-1}cm^{-1}$) (23). For crystallography, I-MsoI designs contained within the pET-24 vector were transformed into BL-21(DE3)pLysS E. coli cells (Invitrogen). Single colonies were then inoculated into 5 ml cultures (LB containing kanamycin and chloramphenicol) that were again grown overnight. Cultures were added to 1L LB media containing 0.5% glucose to repress basal expression. At an optical density of 0.6 AU600, cells were collected by centrifugation and transferred to LB media containing 1 mM IPTG to induce expression. Cells expressed I-MsoI overnight while shaking at 16° C.

2.3 In vitro characterization of endonuclease activity

The relative cleavage activities and specificities of wild- type and designed endonucleases were determined by incubating serial dilutions of each enzyme with a constant amount of plasmid DNA. The plasmid substrate contained two I-MsoI cleavage sites, one wild type and one containing designed base pair substitutions. To preserve symmetry, palindromic base pair substitutions were incorporated into both the left (-) and right (+) half-sites of the substituted recognition sites. The plasmid substrates were created by temperature-annealing phosphorylated oligonucleotides into duplexes corresponding to wild type and designed cleavage sites. These sticky-ended duplexes were ligated into two different locations of a plasmid of length 3308 bp, originally obtained from Doyon et al. [28]. The substrates were pre-linearized by digestion with the restriction endonuclease XbaI. The sizes of linear DNA fragments resulting from digestion by the endonucleases were as follows: of size 3308 bp (no cleavage), 2766 bp (wild-type site cleaved but not designed site), 2174 bp (designed but not wild-type), 1632 bp (wild-type and designed), 1134 bp (designed), 542 bp (wild-type), where the site whose cleavage results in each product is indicated in parentheses. Plasmid DNA substrates (50200 ng) were incubated with varying concentrations of endonuclease in 20 mM Tris pH 8.0, 100 mM NaCl, 10 mM $MgCl_2$ for 1 h at 37° C. The reactions were quenched by adding 10 mM EDTA and 1% SDS and incubating for 10 min at 60° C. The DNA products were separated by agarose gel electrophoresis, visualized by staining with ethidium bromide and quantified by measuring spectral density using the program ImageJ (http://rsbweb.nih.gov/ij/). These data were fit to a sigmoid function to estimate the concentrations that corresponded to half-maximal cleavage of each target site (EC50).

2.4 Crystallization

Protein samples were further purified by size exclusion chromatography using a 150 mM NaCl, 0.02% sodium azide, 50 mM Tris pH 8.0 buffer with a flow rate of 1 ml/min on the Superdex75 16/60 column (120 ml volume). Resulting fractions were analyzed by electrophoresis using a 12.5% SDS denaturing polyacrylamide gel. Fractions containing the purified protein were pooled and concentrated from 15 to 1.5 ml with a final concentration of 440 μM. Crystal trays were set using a grid varying pH (6.6, 7.3, 7.8, 8.1, 8.5 and 9.2) and PEG 400 (v/v 18, 20, 22 and 24%). Each reservoir also contained 5 mM $CaCl_2$, 20 mM NaCl. DNA was resuspended and annealed at 92 ° C for 2 min and then added to protein in a 2 : 1 concentration. Three 1 l hanging drops of dimer protein concentration 180, 135 and 90 μM were added to each well. Crystals were left to grow at 18° C for 4 days. The crystals were looped and placed in a cryogenic solution containing 170 mM NaCl, 5 mM $CaCl_2$ and 25% v/v PEG 400.

2.5 Data collection and refinement

Diffraction data were collected on an in house rotating anode generator, using a Saturn CCD area detector (Rigaku, Inc.). The crystals were maintained at cryological temperatures (72 K) and an X-ray wavelength of 1.54 angstroms was used. Exposure times were 3 to 7 seconds per frame. Images were recorded for 360 of crystal rotation, at 1 intervals. Diffraction images were analyzed by HKL2000 or CrystalClear 1.40r3 to determine the space group. Crystal structures were solved by molecular replacement using Phaser, followed by manual and automated refinement using Coot [35] and PHENIX [1], respectively. For molecular replacement, a modified I-MsoI [1M5X [19]] model was used where (i) waters were removed, (ii) target nucleotides were mutated and (iii) redesigned residues were mutated to alanine. Following molecular replacement and one round of rigid body refinement, redesigned residues were fit to observed electron density. Manual model adjustments, including movement of the phosphodiester backbone, within the electron density were performed using Coot. Finally, automated refinement of atomic positions and atomic displacement factors was performed using PHENIX. During refinement, structural adjustments were modeled using TLS motion determination [88]. The Ramachandran statistics (% most favored/allowed/generously allowed/disallowed) for each of the new structures were: I-MsoI GCG (0.85/0.15/0.01/0); I-MsoI -8G (0.85/0.14/0.01/0); I-MsoI -7C (0.87/0.13/0/0).

2.6 Results: Computational design of specificity

The use of engineered homing endonucleases to target gene sequences depends on the practical designability of available homing endonuclease scaffolds toward potential cleavage sites in a gene of interest. To identify new specificities that were both computationally predictable and therapeutically relevant, we predicted changes in specificity for all single- and double-base pair substitutions in the I-MsoI recognition site and then identified the most designable sites in a gene sequence using a position weight matrix approach. This yielded a ranked list of hypothetically designable cleavage sites, from which therapeutically relevant changes in specificity could be chosen to examine the feasibility of computational design for gene targeting applications.

The site sequence GaAGgcgGTCGTGAGcagGgcagG (lower-case letters differ from native), which occurs in the human gene for fumaryl acetoacetate hydrolase (FAH), was chosen for further analysis due to its high rank. In a second round of computational design, we divided the DNA substitutions that occur within this target into separate clusters of contiguous changes, and then computationally searched for favorable interactions between each cluster and new combinations of amino acids at the surrounding residue positions. This resulted in favorable predictions for a specificity switch involving the three adjacent base pair substitutions (-8G, -7C, -6G). The cluster of protein-DNA interactions in the region of these base pairs consists of a mixture of direct and water-mediated contacts to the DNA bases by six protein side chains (K28, I30, S43, N70, T83 and I85) in each identical subunit of the homodimeric endonuclease (Figure 2.1a). At these six amino acid positions, mutations were first optimized simultaneously to recognize the three bp cluster of altered base pairs (Table 2.1, gcg), and then were optimized separately for each single base pair substitution (–8g, –7c, –6g). The designed complexes were ranked based on their predicted binding energies and specificities, with particular emphasis placed on the latter criterion in order to identify

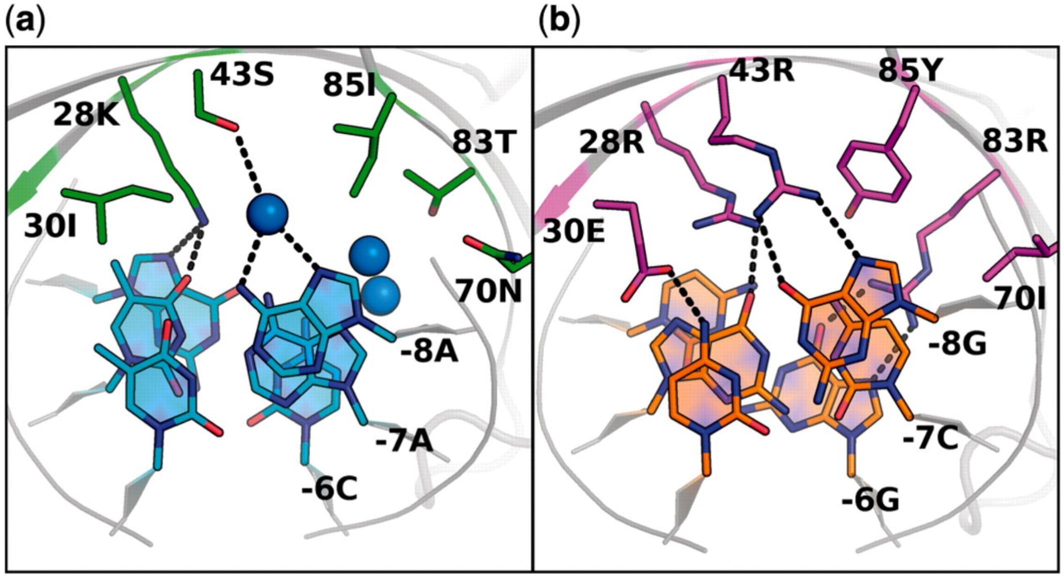
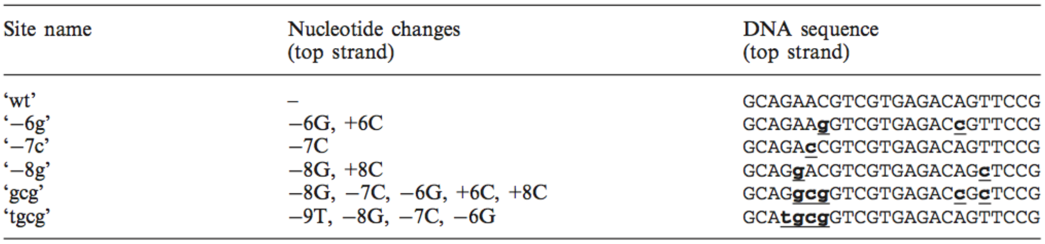


Figure 2.1: Amino acid base interactions in wild-type and designed complexes. The interactions between amino acid residues 28, 30, 43, 70, 83, 85 and DNA bases –8, –7, –6 are shown. Blue spheres are crystallographic water molecules. Dashed lines depict selected hydrogen-bonding interactions. (a) Wild-type I-MsoI interactions observed in the original crystal structure (pdb: 1M5X). (b) Predicted model of computationally designed interactions between novel amino acids and DNA bases for the I-MsoI GCG design.

designs with maximal specificity for their intended targets. For example, in the case of design versus the –8g and gcg target sites, models of redesigned enzymes that harbor a glutamate at residue 30 were predicted to be more specific than those with glutamine. Designs for the remaining two clusters of substitutions in the hypothetical FAH target site were also tested, despite the lack of a predicted change in specificity. Experimental characterization of these designed sequences showed little to no endonuclease activity on either wild-type or designed DNA substrates. Thus, the specificity measure is a useful criterion by which to predict the experimental outcome of computational designs.

2.7 Novel specific cleavage of multiple adjacent base pairs

Upon expression and purification, the designed proteins displayed stabilities and yields comparable to that of the wild-type endonuclease. Table 2.2 shows the cleavage activities



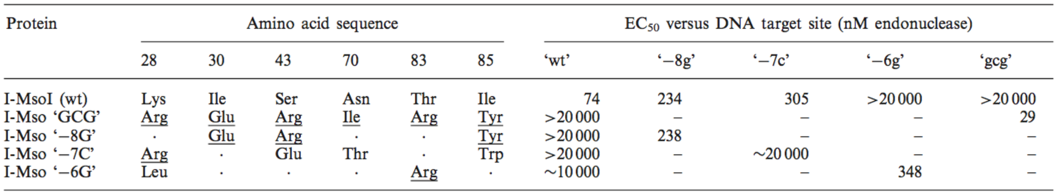
Site name	Nucleotide changes (top strand)	DNA sequence (top strand)
‘wt’	–	GCAGAACGTCGTGAGACAGTTCCG
‘–6g’	–6G, +6C	GCAGAA**g**GTCGTGAGAC**c**GTTCCG
‘–7c’	–7C	GCAGA**c**CGTCGTGAGACAGTTCCG
‘–8g’	–8G, +8C	GCAG**g**ACGTCGTGAGACAG**c**TCCG
‘gcg’	–8G, –7C, –6G, +6C, +8C	GCAG**gcg**GTCGTGAGAC**c**G**c**TCCG
‘tgcg’	–9T, –8G, –7C, –6G	GCA**tgcg**GTCGTGAGACAGTTCCG

Table 2.1: I-MsoI DNA cleavage sites. Base pair substitutions are indicated by lower-case, underlined letters. All cleavage sites were double-stranded duplexes and contained complementary substitutions in the bottom strands (not shown).

of the enzymes on the DNA target sites shown in Table 2.1. The wild-type endonuclease preferred its natural cleavage site over any of the altered sites, exhibiting 50% cleavage of the wild-type site at an endonuclease concentration of 74 nM. It cleaved the –7c and –8g sites at higher endonuclease concentrations (305 and 234 nM, respectively), but did not cleave the –6g or gcg sites at any endonuclease concentration up to 20 μM. This agreed qualitatively with the computed binding energies of the endonucleases for their target sites. The endonuclease designed to cleave the gcg cluster of three consecutive altered base pairs contained six amino acid mutations per domain in the homodimeric protein (Table 2.2, Figure 2.1b). This design cleaved its novel target site at a concentration lower than that at which the wild-type endonuclease cleaved the wild-type site (28.7 ± 2.2 versus 73.5 ± 8.4 nM, respectively, Figure 2.2), and did not significantly cleave the wild-type site at any endonuclease concentration tested (up to 20 M). Thus computational design resulted in a mutually-exclusive switch in specificity, with highly efficient cleavage of the significantly altered recognition sequence.

2.8 High specificity of designed interactions

We characterized the effect of mutations at three designed residues in I-MsoI GCG in order to investigate the determinants of its high degree of specificity (Table 2.3). In agreement with qualitative predictions, the substitution of Glu30 with glutamine had little effect on the concentration at which the designed endonuclease cleaved its target, but resulted in



Protein	Amino acid sequence						EC$_{50}$ versus DNA target site (nM endonuclease)				
	28	30	43	70	83	85	'wt'	'−8g'	'−7c'	'−6g'	'gcg'
I-MsoI (wt)	Lys	Ile	Ser	Asn	Thr	Ile	74	234	305	>20 000	>20 000
I-Mso 'GCG'	Arg	Glu	Arg	Ile	Arg	Tyr	>20 000	–	–	–	29
I-Mso '−8G'	·	Glu	Arg	·	·	Tyr	>20 000	238	–	–	–
I-Mso '−7C'	Arg	·	Glu	Thr	·	Trp	>20 000	–	~20 000	–	–
I-Mso '−6G'	Leu	·	·	·	Arg	·	~10 000	–	–	348	–

Table 2.2: I-MsoI protein sequences and cleavage activities. All amino acid mutations are shown for each designed protein. Amino acids in common with the I-MsoI GCG design are underlined. Dots indicate no mutation relative to wild-type. On the right are relative cleavage efficiencies for selected combinations of endonuclease and DNA target site. EC50 indicates the concentration of the endonuclease at which half of the target site was cleaved under the conditions described in Materials and Methods. Dashes indicate no data.

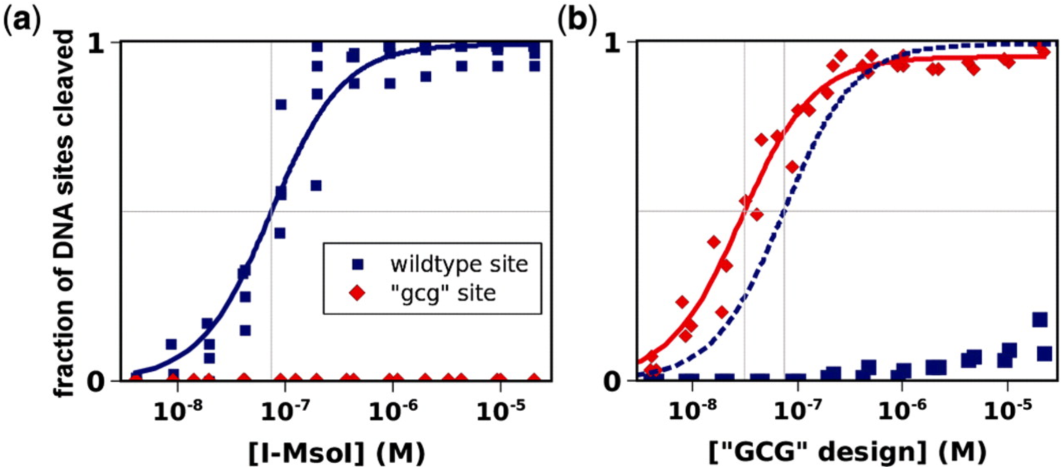
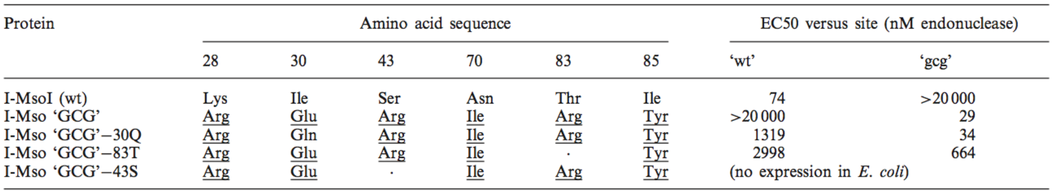


Figure 2.2: Complete switch of activity and specificity for three novel adjacent base pairs by computational design of I-MsoI. The cleavage of either the wild-type site (blue) or the designed gcg site (red) is plotted as a function of the endonuclease concentrations of wild-type I-MsoI (a) and the I-MsoI GCG design (b). Data are densitometric measurements of ethidium bromide-stained agarose-electrophoresed DNA cleavage products. The data were fit to determine the endonuclease concentrations that correspond to half-maximal cleavage (EC50, gray lines). In (b), the best fit to the wild-type data in (a) is shown in dashed lines for comparison.



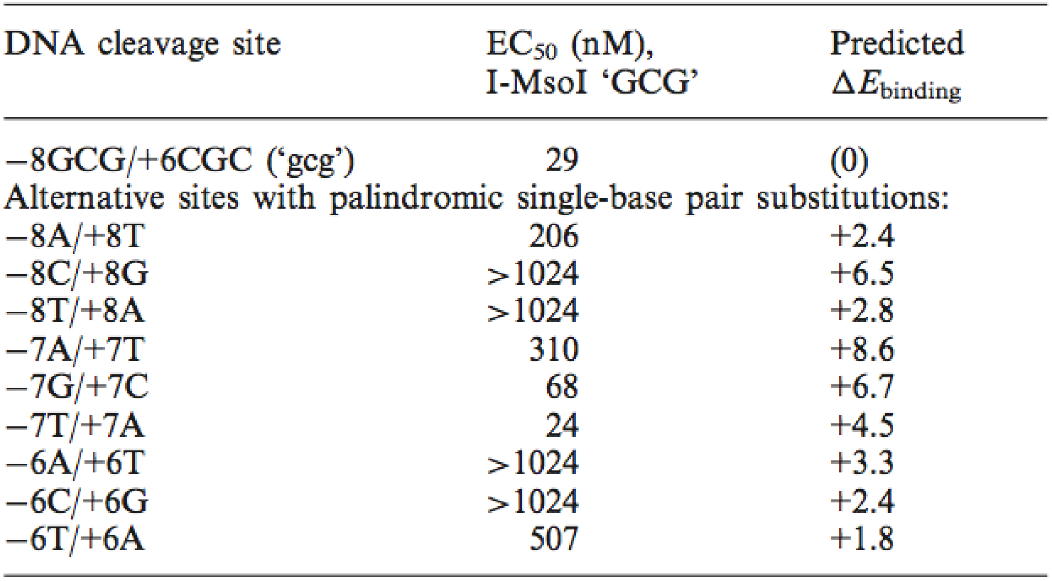
Protein	Amino acid sequence						EC50 versus site (nM endonuclease)	
	28	30	43	70	83	85	‘wt’	‘gcg’
I-MsoI (wt)	Lys	Ile	Ser	Asn	Thr	Ile	74	>20 000
I-Mso ‘GCG’	Arg	Glu	Arg	Ile	Arg	Tyr	>20 000	29
I-Mso ‘GCG’–30Q	Arg	Gln	Arg	Ile	Arg	Tyr	1319	34
I-Mso ‘GCG’–83T	Arg	Glu	Arg	Ile	·	Tyr	2998	664
I-Mso ‘GCG’–43S	Arg	Glu	·	Ile	Arg	Tyr	(no expression in *E. coli*)	

Table 2.3: Cleavage of wild-type and gcg sites by point mutants of the I-MsoI GCG design. This table is formatted as described for Table 2.2.

considerable cleavage of the wild-type site at high endonuclease concentrations. This can be rationalized by considering that glutamate can only accept hydrogen bonds from the –8G:C base pair in the model, while glutamine can both accept and donate hydrogen bonds. However, the magnitude of this difference is underestimated by the computational prediction of binding energies, indicating a need for training of the model to improve quantitative accuracy.

The reversion (to wild-type threonine) of Arg83, which makes contact to the –6G nucleotide in the design model, results in an increase in the concentration at which cleavage of the gcg target site is observed, as well as cleavage of the wild-type site at particularly high concentrations. This confirms that Arg83 contributes to specificity, but that the remaining designed residues still contribute to specificity for the gcg target site in its absence. Reversion (to wild-type serine) of Arg43, which makes contact with –8G in the design, was also attempted, but this protein was not expressible in E. coli.

We further characterized the specificity of the I-MsoI GCG design by analyzing its ability to cleave every DNA site that contained a single base pair substitution within the designed three bp cluster (Table 2.4). As before, palindromic substitutions were introduced into both sides of the target site. The design displayed the highest specificity at position ±6, and at position ±8 only one other sequence (–8A/+8T) was cleaved at relevant concentrations (EC50 = 206 nM). The specificity of the design was lowest at position ±7, a property that was not reflected in the predictions. The designed Arg28 may interact with DNA more promiscuously than expected, or the interface may be flexible in this region in a manner that



DNA cleavage site	EC$_{50}$ (nM), I-MsoI 'GCG'	Predicted $\Delta E_{\text{binding}}$
−8GCG/+6CGC ('gcg')	29	(0)
Alternative sites with palindromic single-base pair substitutions:		
−8A/+8T	206	+2.4
−8C/+8G	>1024	+6.5
−8T/+8A	>1024	+2.8
−7A/+7T	310	+8.6
−7G/+7C	68	+6.7
−7T/+7A	24	+4.5
−6A/+6T	>1024	+3.3
−6C/+6G	>1024	+2.4
−6T/+6A	507	+1.8

Table 2.4: Cleavage specificity of the GCG design. Each indicated target site differs from the gcg target site (Table 2.1) by corresponding single base pair changes on both sides of the palindromic target site. Top-stranded substitutions are indicated; complementary substitutions to the bottom strand are not shown. EC50 indicates the concentration of the endonuclease at which half of the target site was cleaved under the conditions described in Materials and Methods section. The modeled binding energy is the predicted change in binding energy of the complex after repacking and minimizing the interface around each corresponding base pair substitution.

is not considered in the computational model. Also, the efficient cleavage of the –7T/+7A site suggests that the exclusion of a thymine at this position may require larger residues than Tyr85 or Ile70. However, the behavior of the single base pair –7c design that contains Trp85 exhibits suboptimal activity, possibly due to insufficient room in the interface for this residue.

2.9 Design for individual base pair substitutions

In two out of three cases, the amino acid mutations that were predicted by computational design to alter the specificity of I-MsoI for individual base pair substitutions differed from those that were predicted by concerted design for the corresponding three base pair cluster.

Each of these designs displayed a preference for its new target site (Table 2.2) over the wild-type site. However, none of these proteins (which displayed 50% cleavage of their targets at 238 nM to 20 μM enzyme, respectively) were active at endonuclease concentrations as low as those observed for either the wild-type endonuclease vs. its wild-type target (EC50 = 74 nM), or the I-MsoI GCG design vs. its gcg target site (28 nM). The I-MsoI –7C design in particular showed a significant increase in the enzyme concentration at which cleavage occurred, preventing precise estimation of EC50 values. Subsequent characterization of a mutant of I-MsoI –7C with Trp85 to Tyr showed cleavage activity at slightly lower concentrations, but this was accompanied by lower specificity. Thus, while in one case (I-MsoI –8G), the predicted mutations were completely complementary between the individual and concerted designs, the assembly of these individual designs to constitute a three bp change in specificity would be complicated by conflicting mutations at overlapping positions, as well as the poor outcome of the single-base pair I-MsoI –7C design.

2.10 Crystallographic analysis and validation

Crystal structures were determined for the I-MsoI GCG, I-MsoI –8G and I-MsoI –7C designs in complex with their designed recognition sequences. The structure of the designed I-MsoI –6G complex was described previously (6). These structures show that the conformations and contacts adopted by most of the redesigned residues agree between the single- and triple-base pair redesigns, and were predicted accurately in the designed models (Figure 2.3). The triple-base pair I-MsoI GCG design and the I-MsoI –8G design both contain the designed residues Glu30, Arg43 and Tyr85. In both structures, Glu30 and Arg43 make direct contacts to nucleotides +8C and –8G, respectively (Figure 2.3a), while Tyr85 adopts the predicted position above –7C (Figure 2.3a and b). The designed Arg28 residue, which is common between the triple-base pair design and I-MsoI –7C, makes direct contact to the +7G nucleotide in both structures as predicted (Figure 2.3b). The designed Arg83 residue, which occurs in the triple-base pair design and in I-MsoI –6G, makes direct contact to the –6G nucleotide in both structures, also as predicted (Figure 2.3c).

Whereas the crystal structures show that most designed interactions were correctly predicted, unprecedented shifts in the designed region of the interface occurred. In the structure

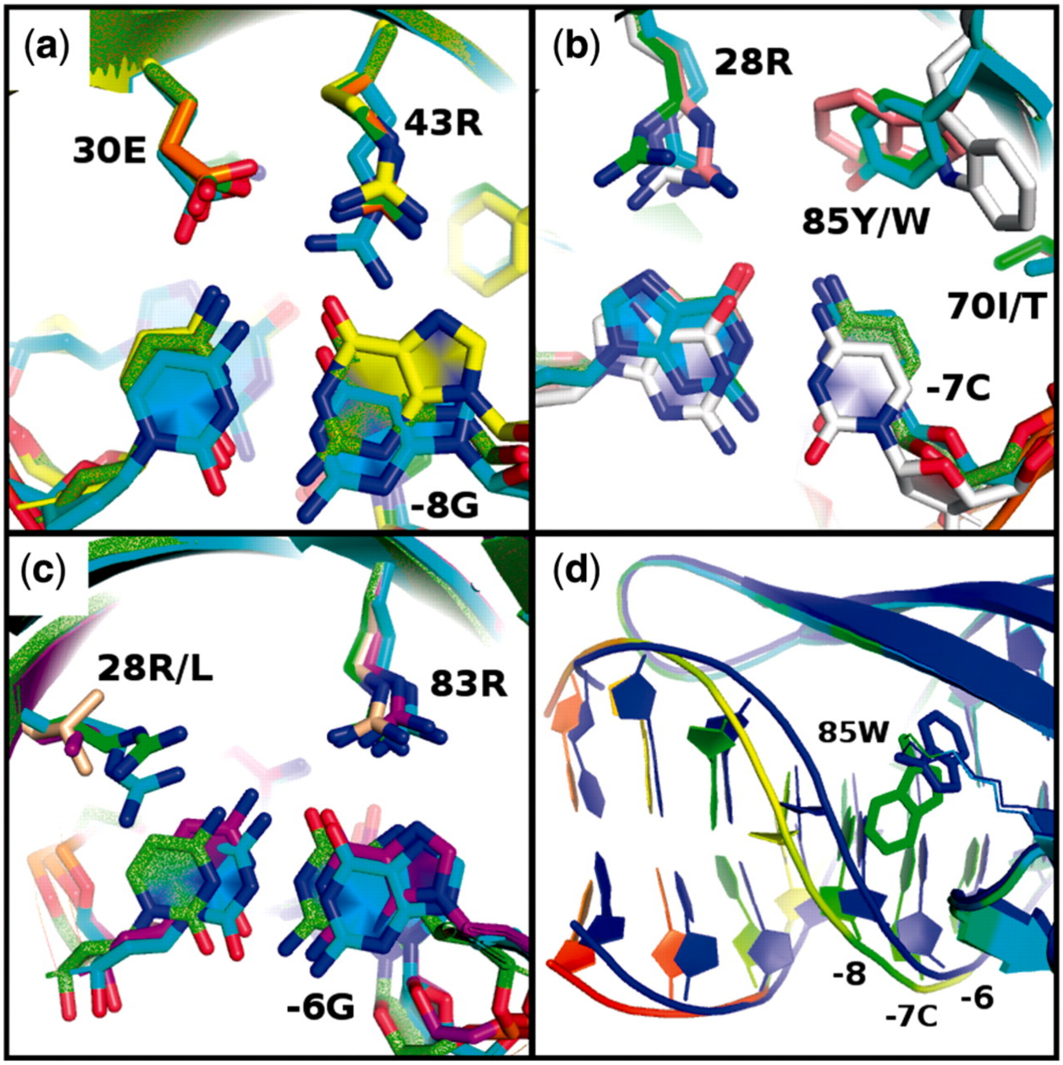


Figure 2.3: Comparison of designed and crystallographically observed interactions. (a-c) the crystal structure of the triple-base pair I-MsoI GCG design (cyan) is aligned with the designed model (green) and with the crystal structures and designed models of each single-base pair design: (a) I-MsoI –8G (X-ray: yellow, model: orange), (b) I-MsoI –7C (X-ray: white, model: pink), (c) I-MsoI –6G (X-ray: purple, model: beige). (d) A conformational shift in the DNA backbone is observed near Trp85 in the I-MsoI –7C crystal structure (colored by increasing B-factor from light blue to red), compared to the designed model (dark blue).

of the I-MsoI GCG complex, local rearrangement of the designed region resulted in a significant (1.4 Å) shift of the –8G:C base pair, which moved away from the protein (Figure 2.3a, cyan). This is accompanied by the extension of Glu30 and Arg43 to remain in specific contact with nucleotide -8G. In contrast, these shifts were not observed in the structure of the single-base pair I-MsoI –8G design (Figure 2.3a, yellow). In the crystal structure of the I-MsoI –7C design in complex, Trp85 juts outward toward the DNA backbone, rather than into the core of the interface as designed (Figure 2.3d). As a result, the neighboring DNA backbone shifted 2.4 Åaway from the original wild-type position. This may explain the lower activity of this design.

2.11 Iterating between design and crystallography enables further switch in specificity

An important challenge in endonuclease engineering is to achieve specificity for genomic target sites which may differ by many base pairs from the original endonuclease target site. To investigate the utility of an iterative approach to structure-based computational design, we began with the crystal structure of the redesigned I-MsoI GCG endonuclease in complex with its cognate DNA site gcg. Mutations were designed to alter the specificity for the adjacent base pair (wild-type: –9G:C) to allow cleavage of –9T:A (Table 2.1, tgcg). The I-MsoI TGCG design contained eight predicted mutations (K28R, I30E, R32K, Q41Y, S43R, N70I, T83R, I85Y; additional mutations relative to the I-MsoI GCG design are underlined). An additional requirement for most hypothetical gene targets is that specificity be alterable in an asymmetric fashion with regard to the two halves of the site. Therefore, these mutations were incorporated into the N-terminal domain of a monomerized construct of I-MsoI, referred to as mMsoI, which was previously created by engineering a peptide linker between the two domains of the wild-type homodimer (27). This resulted in the novel specific cleavage of a DNA target site containing four consecutive, asymmetric base pairs that could not be cleaved efficiently or selectively by the corresponding monomeric mMsoI GCG endonuclease (Figure 2.4).

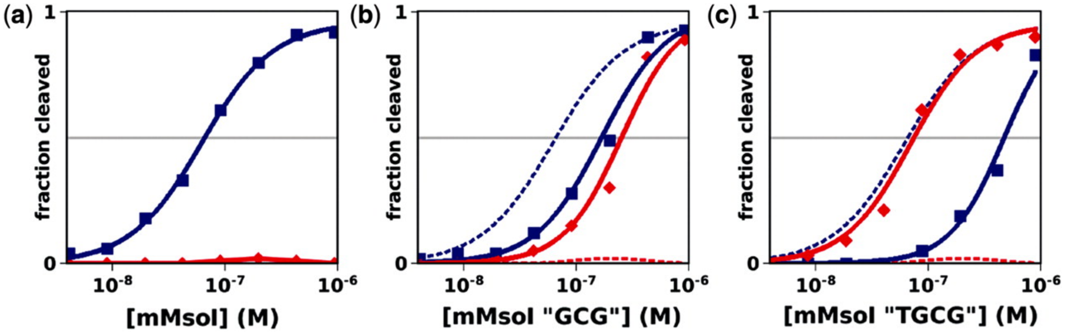


Figure 2.4: Designed specific cleavage activity for an asymmetric four-base pair cluster. In vitro cleavage of wild-type (blue) and asymmetric tgcg (red) DNA sites by monomerized I-MsoI (mMsoI) endonuclease designs. (a) wild-type mMsoI endonuclease, (b) N-terminal mMsoI GCG design, (c) N-terminal mMsoI TGCG design. Dashed lines in (b and c) represent the mMsoI trace from (a). Data are densitometric measurements of ethidium bromide-stained agarose-electrophoresed DNA cleavage products.

2.12 Large changes in specificity by computational protein design

The ability to rationally design proteinDNA recognition is a critical test of our understanding, and could have considerable technological and medicinal value. Our results demonstrate the feasibility of using computational protein design to reprogram the target site specificity of homing endonucleases at multiple adjacent base pairs. The designed cleavage of a novel three- and four-base pair clusters represents a significant advance in computational design techniques, and could soon parallel the capabilities of the latest selection techniques for altering homing endonuclease specificity, which are combinatorially limited to simultaneously altering between three and six amino acids in a single library [28, 108, 6, 16].

2.13 Concerted design of context dependent interactions

The relationship between the triple-base pair design I-MsoI GCG and each of the single base pair designs provides insights into the specificities of homing endonucleases and how they can be reprogrammed. While it is feasible to computationally design single-base pair changes in specificity [8, 121], the I-MsoI GCG design shows that the simultaneous design of interactions between the protein and multiple adjacent base pair substitutions can be advantageous for introducing larger changes in specificity. This is because the physics-based modeling approach employed here is capable of capturing the context dependence of designed interactions, and optimizing the amino acid choices at positions that can interact with multiple adjacent base pairs. Thus, solutions found by concerted design for the three bp cluster differed from those yielded by design for individual base pair substitutions. For example, in the I-MsoI GCG design, the –7C:G base pair is contacted by Arg43 and Tyr85 rather than Glu43 and Trp85 as in the case of the single-base pair change. A synergistic benefit of designing for concerted changes in specificity is evident in that the I-MsoI GCG design cleaves its target site more efficiently than any of the designs for the single base pair substitutions, including I-MsoI –8G, which consists entirely of amino acid mutations that are also present in the triple-base pair design.

2.14 DNA flexibility in the homing endonuclease interface

Crystallographic analyses demonstrate that novel specific interactions between protein and DNA can be successfully predicted using computational structure-based engineering. However, structural shifts in the interface, particularly of the bound DNA, can occur as a consequence of changes to protein and DNA sequence. Furthermore, these changes neither additive nor readily predictable using current modeling techniques. This reflects an inherent structural flexibility of the I-MsoI homing endonuclease interface that was not observed in previous studies of either I-MsoI [8, 19] or its close relative I-CreI [96]. That I-MsoI differentially cleaves DNA sequences with different intramolecular conformations also raises the possibility that indirect readout of sequence-dependent DNA structure [121, 81] may be important throughout the homing endonuclease recognition site. This highlights the importance of accurately modeling significant shifts in DNA conformation for future efforts to predict and design the properties of protein-DNA interactions. Finally, the use of this new crystal structure to design high activity and specificity for additional changes in specificity illustrates the power of combining computational design and X-ray crystallography to generate novel cleavage specificities for genome engineering applications.

Chapter 3

ACTIVITY, SPECIFICITY AND STRUCTURE OF I-BTH0305I: A REPRESENTATIVE OF A NEW HOMING ENDONUCLEASE FAMILY

The chapter was originally published in Nucleic Acids Research by authors GK Taylor, DF Heiter, S Pietrokovski, and BL Stoddard [120].

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 [113]. 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. Homing endonuclease genes 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 [111].

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 in 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 [113]. 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 [18]. 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 recognize symmetric target sequences [37, 132]. In contrast, members of the HNH and GIY-YIG families (usually found in bacteriophage) display multi-domain structures (corresponding to separate DNA-binding and catalytic regions) and adopt highly elongated conformations when bound to DNA [103, 128, 127]. As a result, those proteins usually recognize long non-palindromic sequences with significantly reduced fidelity [32, 71].

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 [25]. 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. Homologues of those uncharacterized ORFS were also be found associated with introns or as free-standing genes. In total, fifteen 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 [4] $< 10^{-3}$) 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) [46, 123]. 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') [106]. NUMODs are frequently found in other homing endonucleases from bacteriophage, such as the GIY-YIG endonuclease I-TevI [127] and the HNH endonuclease I-HmuI [103]. 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 [25]. This ORF is located within a group I intron that interrupts the RecA gene of Bacillus thuringiensis 0305φ8-36 bacteriophage. Experiments described in this chapter 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 base pairs 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.

3.1 Materials and Methods: Computational Sequence Analysis

Sequences of Vsr-like putative homing endonucleases (Supplemental Information) were identified in the NCBI sequence databases and JCVI data using BLAST sequence searches and BLIMPS motif searches as previously described [25]. Multiple sequence alignments were constructed with MEME [10], MACAW [101], DIALIGN-TX [115], and GLAM-2 [39] 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 phage 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 [55], using a total of 4 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 [106] and to various DNA binding HTH motifs from the Blocks release 14.3 database (21) (including IPB000792 (LuxR bacterial regulatory proteins), IPB000831 (Trp repressors), IPB002197B (FIS bacterial regulatory proteins)) using the LAMA program [90]. The specified blocks were used to predict the position of the HTH DNA binding region within the NUMOD 2 motifs of I-Bth0305I.

3.2 I-Bth0305I cloning

Synthetic genes encoding I-Bth0305I and several additional homologues that were identified in an earlier metagenomic analysis [25] were ordered from Genscript (New Jersey, USA) with codons optimized for protein expression in E. coli (Supplemental Figure S1). These reading frames were ligated into an in-house pET15-HE vector (Supplemental 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 sub-cloning amino acids 167 through 266; two point mutations corresponding to D196A and H213A were introduced to allow over-expression by inactivating the construct. To facilitate crystallographic phasing, an additional point mutation (L180M, that 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.

3.3 Protein Over-expression and Purification

For initial overexpression trials of I-Bth0305I and its homologues, 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 minutes, heat shock for 30 s at 42°C, incubate on ice for 2 minutes, add 200 μL SOC media, shake at 220 rpm at 37°C for 20 minutes, 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 liter 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 (Supplemental Figure S3).

Purification of I-Bth0305I to homogeneity was then carried out using protein expressed as a GST fusion protein from pGEX-6p3 bacterial expression vector. GST tagged I-Bth0305I was over-expressed at 16°C while shaking at 220 rpm for 16-20 hours. 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 x 30 seconds (with 1 minute 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, 2M NaCl). Beads were washed again with lysis buffer. Finally 2 mL of lysis buffer was added to the beads, along with 80 units of PreScission protease. The mixture was incubated for 16 hours with a gentle rocking motion at 16°C. Resulting protein was eluted directly from the beads and purified further via heparin affinity chromatography. 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 approximately 1M NaCl and was found to be over 95% pure as estimated by electrophoretic analysis.

3.4 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 thirty micrograms of phage lambda DNA was digested for one hour at 37 °C with a series of 2-fold dilutions of I-Bth0305I ranging in concentration from 20 ng per microliter (0.65 μM) to 9.8 pg per microliter (0.6 nM) as shown and further illustrated in Supplemental Figure S4. The DNA was extracted with phenol and chloroform, precipitated, and resuspended in 10mM Tris 1mM EDTA, and then diluted in water to 10 nanograms per microliter for use as template for sequencing reactions. Sequencing reactions were carried on the respective DNA samples using 19-base oligonucleotide primers (IDT, Inc.) that 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).

3.5 Cleavage Experiments

Noncompetive cleavage digests (corresponding to experiments depicted in Figures 3.2a and 3.4) 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 3.2d; Supplemental Figure S5).

In competitive cleavage digest experiments (corresponding to Figures 3.2b, 3.5 and 3.6), 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 bp, 1900 bp, 1600 bp, 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 $MgCl_2$.

3.6 DNAse I Footprinting

A 120 base pair polymerase chain reaction product corresponding to the uninterrupted RecA gene sequence from bacteriophage Bt03058-36, with the endonuclease cleavage site positioned at its center, was generated using either of two radio-labeled PCR primers. 0.1 pmol of this radiolabeled PCR product was incubated with 20 micromolar I-Bth0305I in binding buffer (50 mM Tris pH 7.0, 60 mM KCl, 1 mM $MgCl_2$, 1 mM 2-mercaptoethanol, 2mg/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 1700V for 1 h 50 m.

3.7 Binding assays via isothermal titration calorimetry

Aliquots of a DNA duplex corresponding to a 67 base pair region of the 0305φ bacteriophage RecA gene sequence, centered around the endonuclease cleavage site were injected into I-Bth0305I (300 μL, 20 μM) (Supplemental 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 CaCl2. 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 microliters over 5.0 s, 180 seconds between each injection). The binding analyses were performed in triplicate.

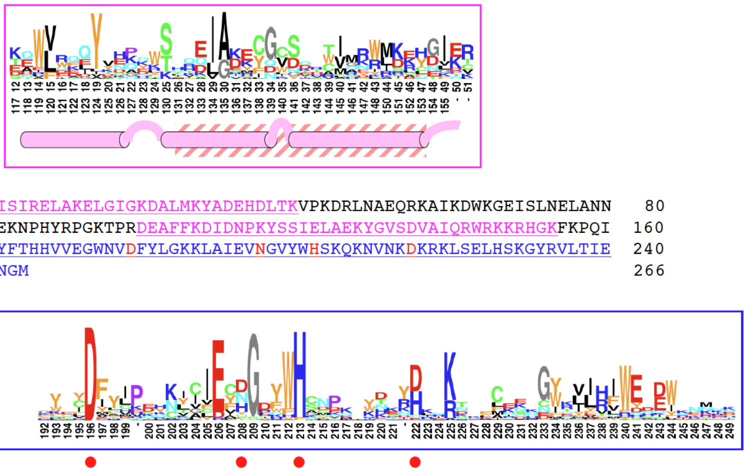
3.8 Protein Crystallography

A complex corresponding to a catalytically inactivated nuclease domain (residues 167 to 266, containing active site point mutations D196A and H213A) was over-expressed 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 % 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, California). 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 angstrom resolution. The native data set was used for final refinement, even though it was slightly lower resolution (2.2 versus 2.15) because the merging statistics for that dataset were otherwise superior to the Se-Met data (Table 1).

3.9 Results: Cloning and protein production

Genes encoding several individual representatives of the Vsr-like endonuclease gene family identified in the metagenomic analyses [25], 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 (Supplemental Figure S1 and S2). The resulting constructs displayed a wide range of behaviors during bacterial overexpression and purification (Supplemental 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 'Methods'. The yield of this protein was approximately 1.5 milligrams per liter 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 30912 Da. The surrounding group I intron within the bacteriophage 0305φ8-36 RecA gene is 801 nucleotides in length; the start codon for the putative endonuclease reading frame is found 88 nucleotides 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 [2]. 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 [106], and a C-terminal region that shares



```
MSRAWSPSIEQKQIVIDGYASPDISIRELAKELGIGKDALMKYADEHDLTKVPKDRLNAEQRKAIKDWKGEISLNELANN   80
IGISLAGVQKRMKKLGIDTKQYIEKNPHYRPGKTPRDEAFFKDIDNPKYSSIELAEKYGVSDVAIQRWRKKRHGKFKPQI  160
DTSTHLTTPERRVKEILDELDIVYFTHHVVEGWNVDFYLGKKLAIEVNGVYWHSKQKNVNKDKRKLSELHSKGYRVLTIE  240
DDELNDIDKVKQQIQKFWVTHISNGM                                                        266
```

Figure 3.1: 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 alpha helices and arcs for loops and turns). The hatched region denotes a predicted DNA-binding helix turn helix (HTH) motif.

homology with the catalytic domain of the Vsr DNA mismatch repair endonuclease [123]. Further analysis, using homologues 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 [25] we identified additional members of this family including bacteriophage Hef type homing endonucleases [97] and a bacterial protein from Corynebacterium glutamicum ATCC 13032 (Supplemental 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 3.1).

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. The overall yield of isolated N- and C-terminal regions of I-Bth0305I were approximately 1 and 3 mg per liter, respectively.

3.10 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, that 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 [11]. In our experiments, efficient cleavage of the DNA substrate corresponding to the uninterrupted RecA gene was observed (Figure 3.2a). Substrates containing the intron-exon junction sequences of the bacteriophage recA gene were not cleaved by the enzyme under any conditions (Figure 3.2b), 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 kilobase 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 base pair 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

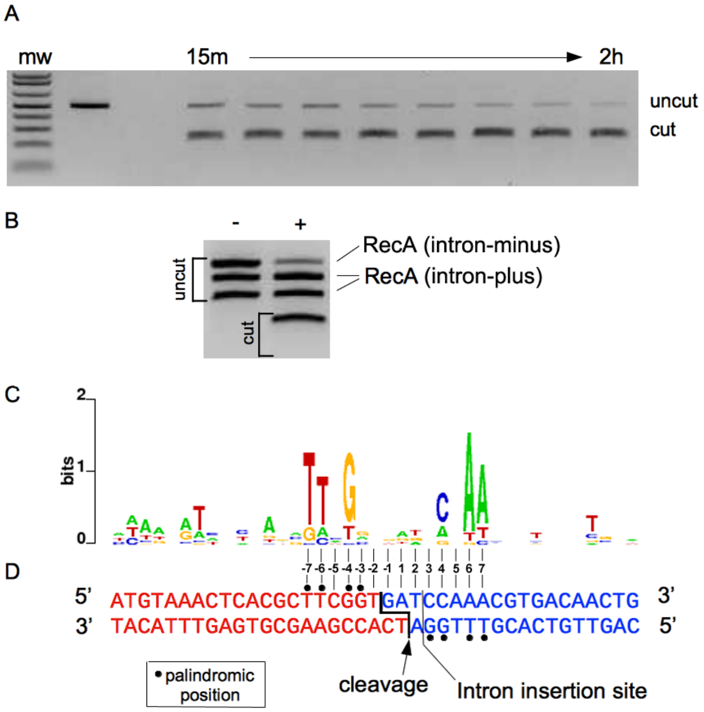


Figure 3.2: Determination of the I-Bth0305I DNA target site a: I-Bth0305I cleaves a DNA target site containing the sequence of the RecA host gene spanning the intron insertion site. b: In competition digests, three substrates (one corresponding to the uninterrupted allele of the bacteriophage RecA gene; "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 2 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 2 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 below the base pair positions in the target site indicate positions that are palindromically conserved between the left and right half-sites.

and cleavage activity.

Digestion of lambda DNA generated a list of target sites that were hydrolyzed by the endonuclease (Supplemental Figure S4). Alignment of these genomic sequences resulted in a target site consensus corresponding to 5'-T-T-x-G-x6-C-x-A-A-3' (Figure 3.2c). This 14 basepair 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 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 2 base, 5' overhang). One site that displayed a sequence that was particularly close to the consensus described above (differing at only one basepair out of six) was cleaved on both strands and thereby produced the actual 2 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 3.2d and Supplemental 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 double strand break. 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 six basepairs 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 3.3). 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 base pair 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 approximately 60 nucleotides, corresponding to 30 base pairs that extend from each side of the center of the cleavage site, were 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 (KD) of 24 nM +/- 6 nM (Supplemental Figure S6).

3.11 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 3.4). These experiments indicated that the enzyme exhibits the highest specificity across the central 14 base pairs of its target site. A series of substrates that contained either three consecutive transverted base pairs (e.g., 5'-ATC-3' to 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.

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 3.5). In these experiments, a series of long DNA duplex substrates (each of which were 1 to 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 base pair region (corresponding to the extreme limits of the protected region observed in DNAse I footprinting assays) to 54 base pairs resulted caused little or no loss of cleavage activity. In contrast, a slight reduction in activity was observed when a 33 base pair 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 base pairs. In no case, however, was the loss of cleavage activity in these experiments as pronounced as when as few as three base pairs in the center of the

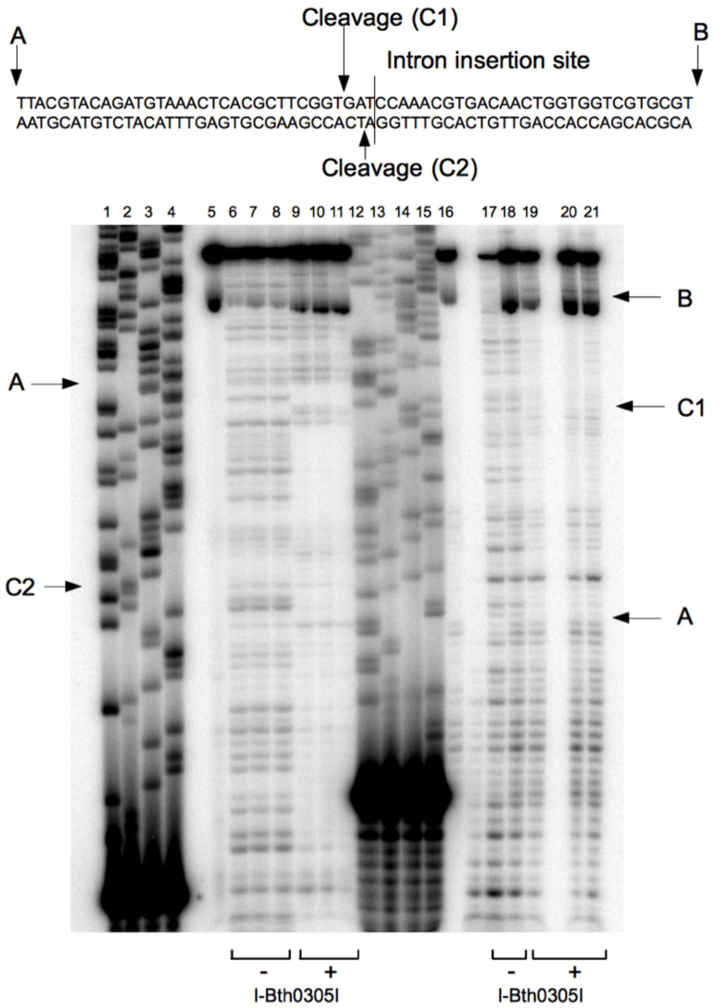


Figure 3.3: I-Bth0305I target site DNAse I footprint. Forward and reverse PCR primers were labeled with 32P and used to generate PCR products labeled at either end. In lanes 6-8 and 17-18, reverse and forward labeled PCR products were digested with DNAse-I. In lanes 9-11 and 19-21, labeled PCR products were incubated with 20 micromolar 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.

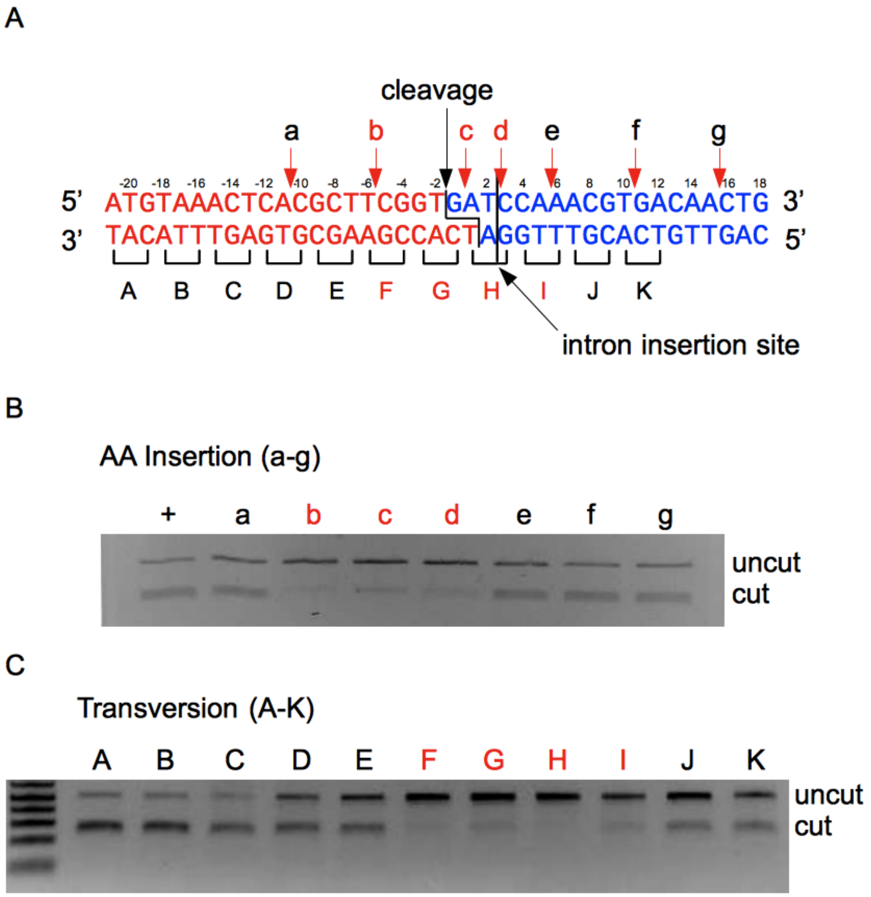


Figure 3.4: Effect of multiple base pair substitutions on DNA cleavage. a: Enzymatic cleavage was assayed in complementary experiments where digests were performed using a DNA substrates containing the target site that were disrupted by either insertion of two base pairs at several positions, or by systematic 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 basepairs 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 basepairs in a region extending from position -5 to +6.

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 3.6). Although the previous experiments described above demonstrated that simultaneous mutation of as few 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 all 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 fourteen base pair 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 [34, 85, 99, 133], such that only multiple simultaneous base pair substitutions result in a significant loss of cleavage efficiency.

3.12 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 kD versus 31 kD for the full length protein, and 18 kD versus 12 kD for the catalytic domain) (Supplemental 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

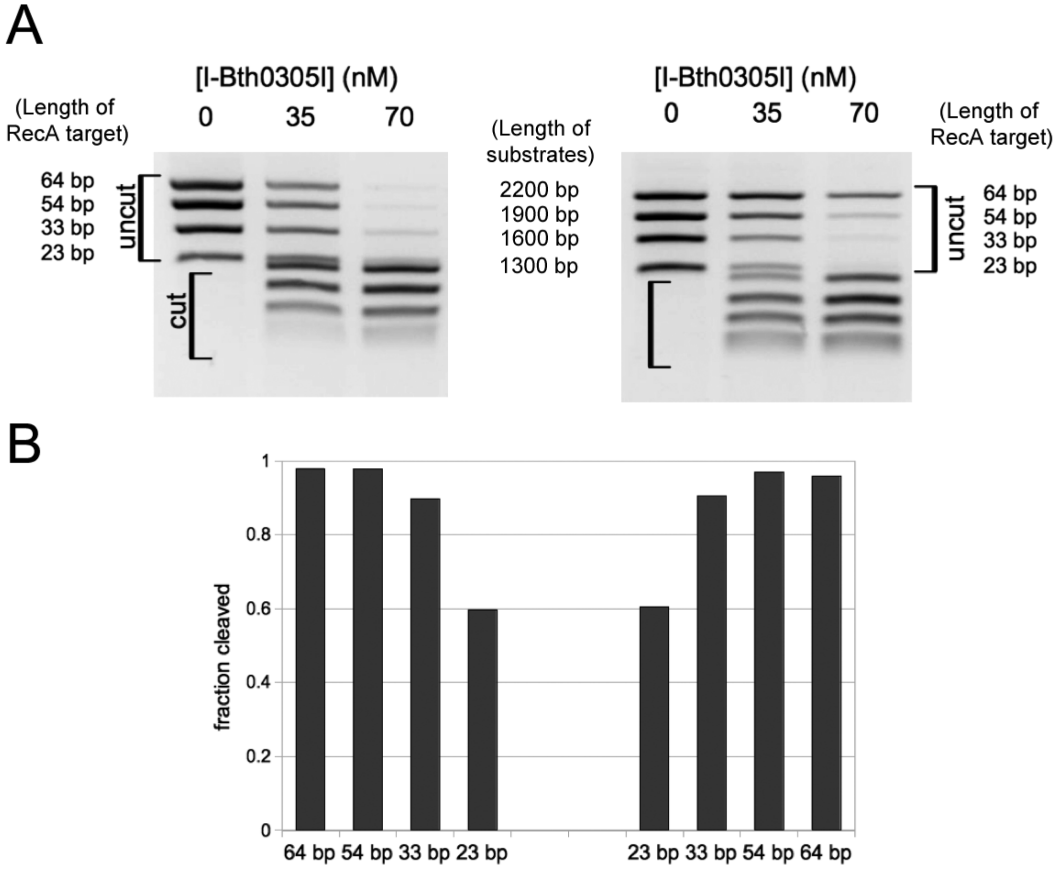


Figure 3.5: Effect of reduced length of RecA target site on DNA cleavage. a: Four separate substrates, ranging in total length from 2200 to 1300 basepairs, that contained specific bacteriophage RecA target sequences of various lengths centered around the site of cleavage were digested with I-Bth0305I 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.

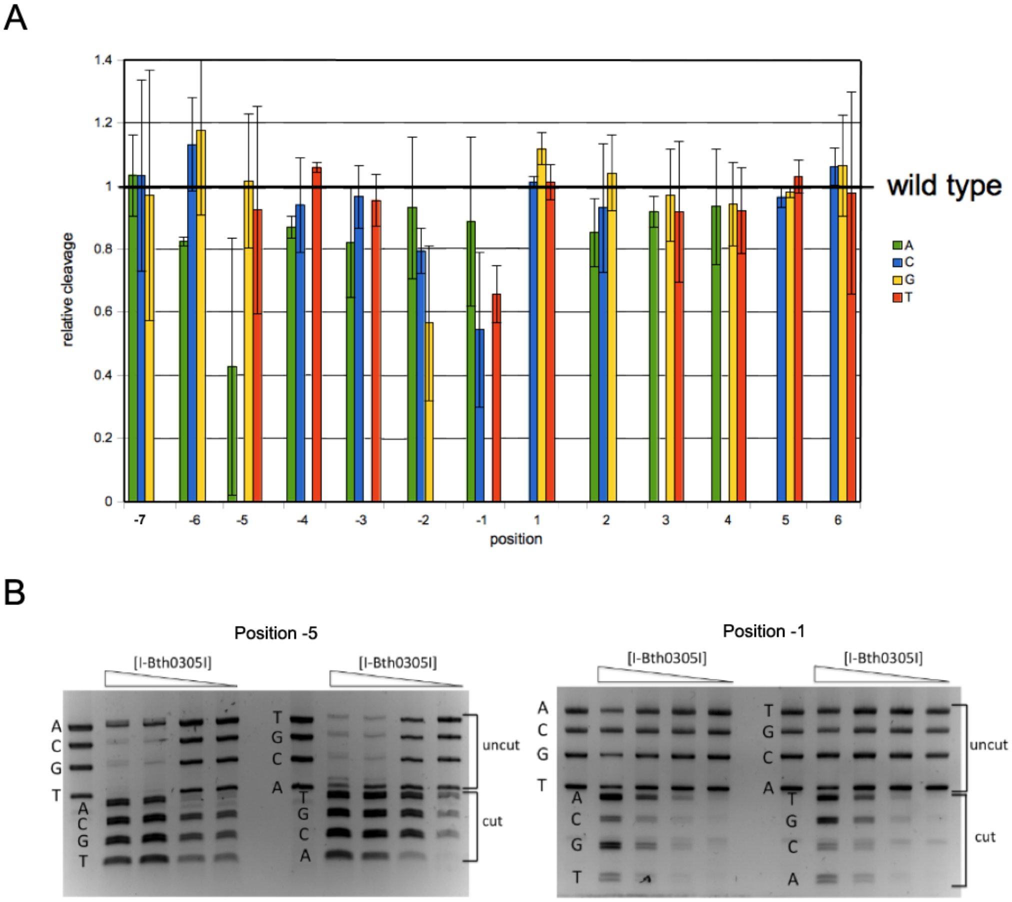


Figure 3.6: Effect of single base pair substitutions on DNA cleavage. a: Each bar represents the relative cleavability of a target site that is altered at one base pair relative to the wild-type target. b: Raw data for cleavage specificity at positions -5 and -1, respectively. In these experiments, increasing 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).

apparent mass and the dynamic radius by approximately 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 3.2). The presumed role of H213 in catalysis (based on prior mutational studies and conservation of the comparable residue in the Vsr repair endonuclease [46, 123], versus 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.

3.13 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 to 266, which displays sequence homology to the Vsr mismatch repair endonuclease) was determined and refined to 2.2 $\AA$ 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 to 263 from the isolated catalytic domain (3 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 $\AA$. 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 nonphysiological interaction that is formed in the crystal lattice.

The structure of the catalytic domain consists of a central -sheet with mixed parallel and antiparallel topology surrounded by four alpha helices. The structure of the I-Bth0305I catalytic domain superimposes against the homologous region of Vsr endonuclease (PDB ID: 1VSR) [46, 123] with an RMSD of 8.76 $\AA$ across 61 atoms (Figure 3.7). The structure of the central beta 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 (nonconserved) 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 $\beta 3$ and $\alpha 2$ 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 $\alpha 3$ (residues 80 to 84) is instead a loop in Vsr. Furthermore, helix $\alpha 2$ in I-Bth0305I is considerably shorter (at 14 residues) than the corresponding 25-residue helix in Vsr (spanning residues 82 to 107) that is inserted into the DNA major groove in 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 3.7). 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, that 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 3.1).

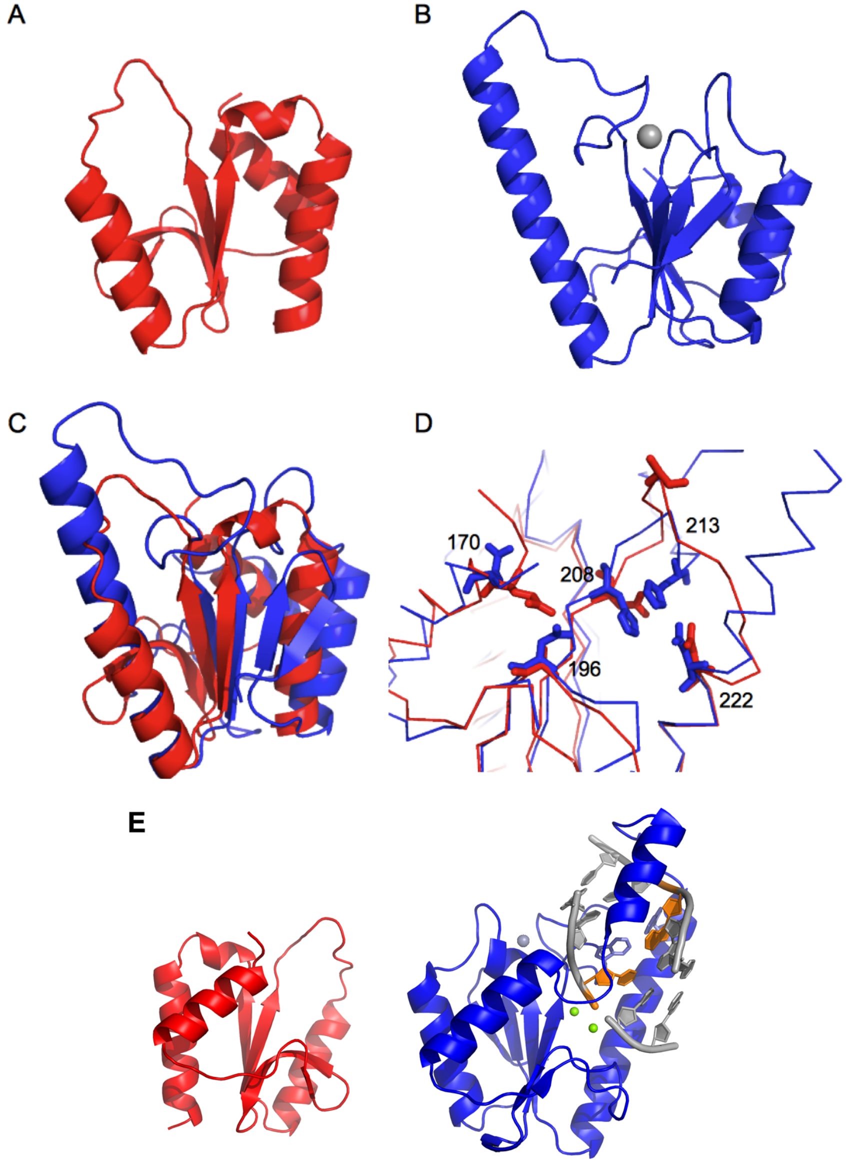


Figure 3.7: Structural analyses of I-Bth0305I nuclease domain. a: The crystal structure of the I-Bth0305I catalytic domain (PDB ID 3r3p). b: The structure of E. coli Vsr endonuclease in the absence of bound DNA (PDB ID 1vsr). The bound zinc ion in the Vsr structure is the cyan sphere. c: Superposition 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 cocrystal 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.

3.14 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) [48] and the HNH endonuclease I-HmuI (which drives intron homing into a DNA polymerase host gene in the Bacillus SPO1 bacteriophage) [47]. Both of those enzymes, as well as their closest homologues (I-BmoI and I-BasI, respectively) appear to bind their DNA targets as monomers [32, 71], with protected DNA regions extending approximately 30 to 40 base pairs 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) [71], I-TevI is observed to directly generate a double stranded break and a two-base, 5' overhang 23 and 25 base pairs upstream of the intron insertion site [32]. The ability of I-TevI to directly generate a double-strand break 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 base pairs 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 base pair contact region would simply correspond to two 30 base pair 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, multi-domain 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 helix-turn-helix 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 basepairs 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 3.8a).

A similar bipartite domain organization has previously been observed in both I-HmuI, I-TevI and their homologues [103, 128, 127]. 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.

3.15 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 that 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) [32] and I-HmuI (an HNH enzyme) [71] both indicate that the basepair 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 [34, 85, 99, 133].

The specificity profile of I-Bth0305I suggests that several of the central fourteen base

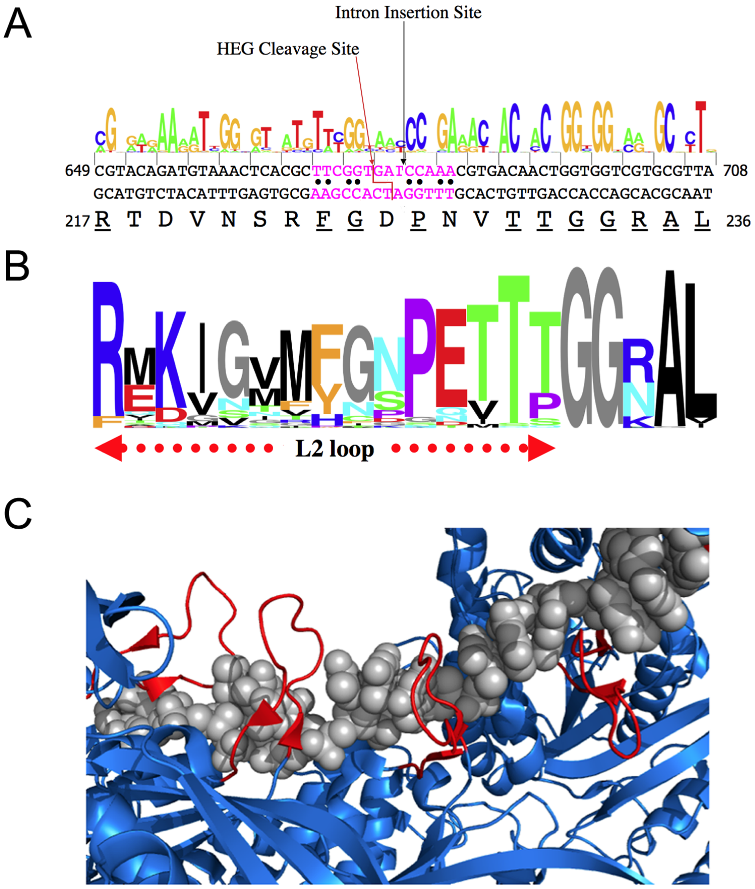

A
Intron Insertion Site
HEG Cleavage Site
649 CGTACAGATGTAAACTCACGCTTCGGTGATCCAAACGTGACAACTGGTGGTCGTGCGTTA 708
GCATGTCTACATTTGAGTGCGAAGCCACTAGGTTTGCACTGTTGACCACCAGCACGCAAT
217 R T D V N S R F G D P N V T T G G R A L 236
B
L2 loop
C

Figure 3.8: 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 basepairs 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 base pair region protected by bound I-Bth0305I in the DNAseI footprint experiment (Figure 3.3). The purple bases are the central 14 base pairs 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 3.2). 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.

pairs 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 3.8b). 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 [55, 100]. This analysis demonstrates strong conservation at eleven out of the central fourteen base pairs 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 twenty 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 3.2). 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 [99].

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 3.8c). 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 one nucleotide base [17] (Figure 3.8d). 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. [90] 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 [107].

3.16 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. 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 [64]. The HNH catalytic motif is found in non-specific bacterial and fungal nucleases [38, 68], and is also found in a wide range of DNA-acting enzymes including transposases, restriction endonucleases, polymerase editing domains and DNA packaging factors [24, 80]. The GIY-YIG catalytic motif is found in several bacterial restriction enzymes (such as Eco29kI) [60] and enzymes involved in DNA repair and recombination (such as the UvrC base-excision repair endonucleases) [67]. 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 [132].

The discovery of a new homing endonuclease lineage [25] 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 [66]. 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 [91].

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 [123]. 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 [22, 94]. The isolation and characterization of an independently folded nuclease domain that (a) appears to display a moderate degree of sequence specificity directly at the site of cleavage and (b) 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.

3.16.1 DATA DEPOSITION

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.

3.16.2 ACKNOWLEDGEMENTS

This work was supported by a National Institutes of Health research grant to BLS (R01 GM49857) and a National Institutes of Health training grant appointment to GKT (T32 GM08268). SP holds a Hermann and Lilly Schilling Foundation chair. We 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.

Chapter 4

HOMINGENDONUCLEASE.NET: ARCHIVAL AND SEARCH OF LAGLIDADG RELATED POSITION WEIGHT MATRICES

This chapter is intended for publication in Nucleic Acids Research by authors GK Taylor, LH Petrucci, J Jarjour, and BL Stoddard.

Homing endonucleases are under intense scrutiny as reagents that can potentially target eukaryotic genomic loci and induce sequence-specific gene insertion, deletion, modification or correction events. The development of such reagents would facilitate a wide variety of applications, ranging from genome engineering to corrective gene therapy. The full realization of this concept requires the engineering of these proteins scaffolds to recognize a DNA target site that is located at or near the exact location where a chromosomal modification is desired.

This work has been supported by the identification of many new homologues of these proteins, and their corresponding DNA target regions, by microbial sequencing projects. The goal of the web-based tool described in this chapter is to exploit these sources of information by developing a database of these proteins and their engineered variants, coupled with a collection of search algorithms for genomic positions and sequences that can be most efficiently targeted by them for sequence-specific modification or correction. This resource is publically available to the community of researchers who are involved in research that requires targeted genomic modification.

4.1 Impetus for the Website

Gene targeting nucleases can be used to induce targeted genetic modifications, both for genome engineering and for corrective gene therapy [89, 79]. The concept of gene targeting to alter endogenous chromosomal loci dates back to experiments in the late 1970s in which ectopic DNA sequences were introduced into yeast and incorporated by homologous recombination (HR) [57, 87]. Unfortunately, while homologous recombination is extremely efficient in yeast, in multicellular eukaryotes such events occur in far less than 0.1% of transformed cells [65, 27]. This limitation requires the use of highly specific gene targeting reagents to stimulate homologous recombination.

One method which facilitates targeted homologous recombination is the introduction of a double strand break (DSB) at the desired locus [118]. This can be accomplished through the transient introduction of a DNA cleavage enzyme to generate a double strand break at the desired modification site, along with a DNA template which possesses the desired modified sequence flanked by regions of homology to the genomic DNA [21].

With the availability of homing endonucleases [113] in addition to zinc-finger endonucleases [72] and TAL (transcription activator like) nucleases [22, 73], the discipline of site-specific genome engineering now enjoys a wealth of structural scaffolds for the continued development of gene targeting proteins. LAGLIDADG homing endonucleases tightly couple cognate site recognition to DNA strand cleavage activity and possess small structures that can be coded by short reading frames, a feature required for gene correction. Hundreds of LAGLIDADG homing endonucleases have been discovered from mining genomic information but relatively few proteins have been experimentally characterized.

Genome engineering and gene correction is an emerging discipline in which genomes of model organisms or cells are manipulated at specified chromosomal loci by using site-specific recombination to alter or add desired traits. Gene targeting nucleases are under intense examination as reagents that can induce (1) the correction of a dysfunctional gene in patients suffering from various genetic diseases; (2) the targeted mutation, knockout or insertion of a gene in an agricultural crop species, or (3) the generation of transgenic mammalian cell lines or animals. These approaches do not require the intermediate screening steps that are necessary in traditional transgenic modification techniques [44, 125, 95].

4.2 Homing Enodnuclease Database

Several types of highly specific DNA recognition and cleavage enzymes, including homing endonucleases and comparable artificial proteins such as zinc finger nucleases and TAL nucleases, are being developed and used for a variety of targeted gene modification applications, ranging from targeted gene disruptions to corrective gene therapy. Current microbial and metagenomic sequencing projects demonstrate that nature can provide an abundance of these proteins that display considerable structural homology and highly divergent DNA recognition specificities. This illustrates the potential to mine the microbial universe and then deliberately engineer and alter those proteins to create proteins with additional, altered DNA cleavage activities.

The gene targeting field is currently experiencing a deluge of data that impacts the creation of targeting proteins, including (i) the ongoing identification of a large, growing number of homing endonucleases and additional DNA binding scaffolds from microbial and metagenomic sequencing projects around the world, and (ii) the generation of a large numbers of selected and engineered protein variants with altered DNA recognition properties.

The only manner in which the full potential and impact of this information can be exploited is through the development of a computational database and search engine. This web resource is more than a simple search algorithm: it combines the data found in the bioinformatics of gene targeting proteins (what DNA sequences do they naturally recognize?) with the details of their biophysical behavior, specificity profiles and 'engineerability', in order to identify combinations of genomic DNA sites and possible targeting proteins that can be exploited for genome modification. In so doing, the site actually guides the necessary protein selection and engineering experiments for successful gene targeting and modification.

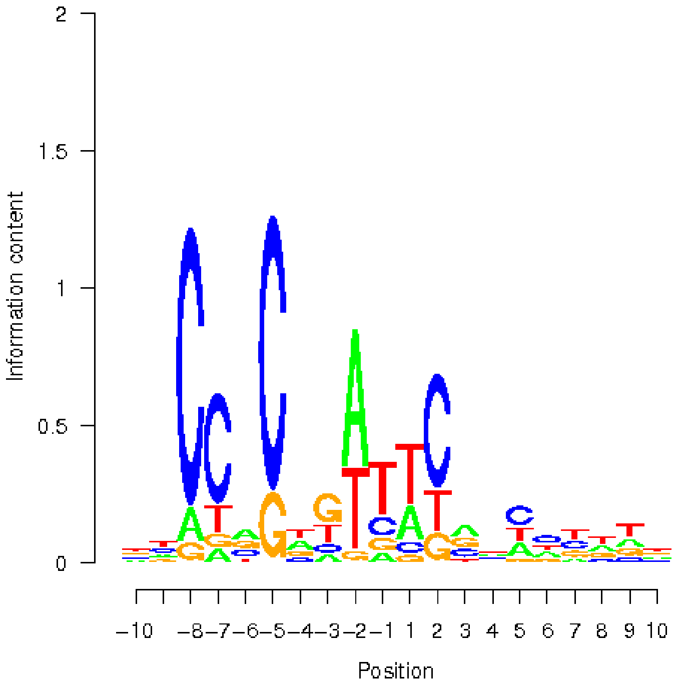
The foundation of this resource is an updateable set of known, wild-type DNA-binding scaffolds that recognize long target sites (22 basepairs). The exact sequence of their naturally occurring DNA targets (which correspond to the microbial DNA sequences cleaved by homing endonucleases) is represented in the database both by the target sequence itself and by Position Weight Matrices (PWMs) [54] that reflect either their exact physiological target sites in their biological host genomes or that reflect their overall specificity profiles, taking into account positions in the DNA target that are recognized with reduced fidelity. PWMs are reported both as a sequence logo plot and as raw text (Figure 1). In the sequence logo plot, those positions that have higher reported specificity have more information content on the graph. Conversely, those positions where there is little are no sequence specificity have very little information content and are diminished within the plot. Both target site sequences and PWMs can be used for searching genomic DNA sequences for closely related sequences. For reference, the site also contains the amino acid sequence for several well-studied homing endonucleases.

In addition to target sites and PWMs, the database contains of list of residues for each homing endonuclease that are potentially useful in the redesign process. These residues often times are associated with position specific target site changes [8, 119, 121]. For example, the I-MsoI homing endonuclease when incorporating the I30E, S43R, and I85Y mutations prefers the -8G and +8C nucleotides compared the wild type target identities of -8A and +8T [9]. Several design mutations reported in the literature are reported for each homing endonuclease in the PWM browser and novel mutations may be entered with the Specificity Changing Mutation Entry Tool.

Many of the homing endonucleases listed in the database also have a list of amino acid contact modules that describe which residues are likely in contact with three base pair target modules. As an example, the I-OnuI endonuclease has an extensive list of amino acids that contact the -11 to -9 positions: Asn 32, Lys 34, Ser 35, Ser 36, Val 37, Gly 38, and Ser 40. Within this module, amino acids may be mutated to achieve the desired specificity shift. Several modules across the target site are reported for each endonuclease for which crystal structures have been solved and this type of information is available. By recommending these modules, this tool suggests which residues the user might look at first when redesigning a homing endonuclease toward a specific sequence.

4.3 Position Weight Matrix Search

This site offers the user the opportunity to search individual genes or collections of genes for potential matches against LAGLIDADG homing endonuclease (LHE) target sites based on four search criteria: (a) Searches for exact identity at the central four basepairs of an LHE target site, because our experience indicates that specificity changes at these base pairs are often not easily achieved through protein engineering or selection [121]. (b) Searches for highest possible identity (fewest DNA basepair mismatches) against an LHE's physiological target site. (c) For those proteins which have had their complete specificity profile determined, the user is provided with the opportunity to search and score potential targets using a 'fidelity PWM' in which the scoring for a mismatch between a protein's natural target site

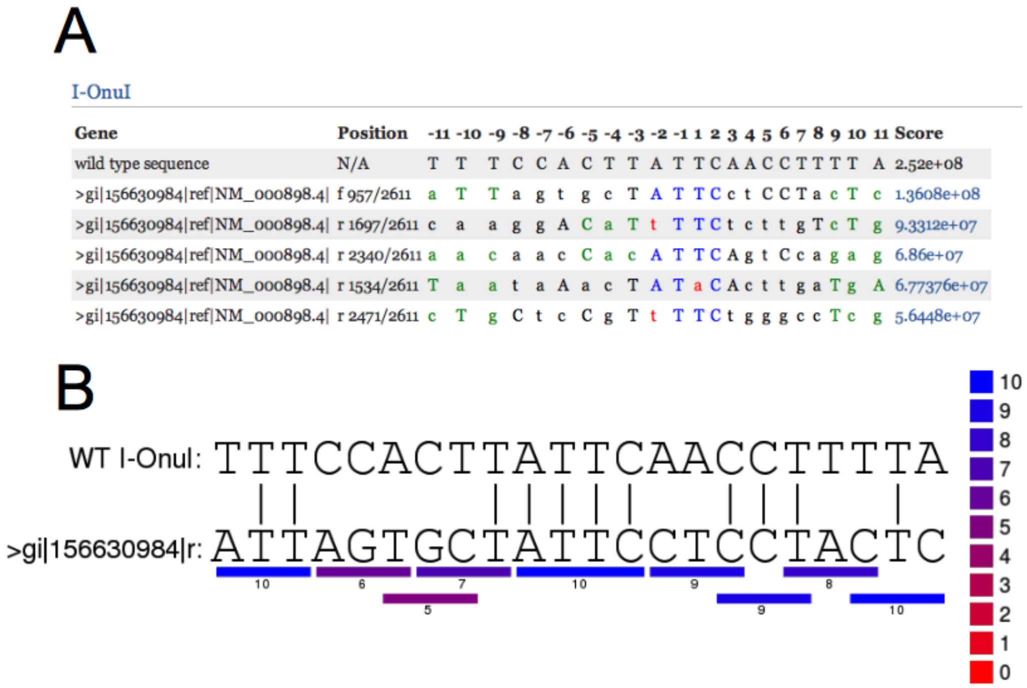


pos	A	C	G	T
-10	0.23	0.23	0.23	0.32
-9	0.21	0.25	0.19	0.35
-8	0.10	0.85	0.05	0.00
-7	0.08	0.68	0.08	0.16
-6	0.40	0.23	0.23	0.14
-5	0.00	0.81	0.19	0.00
-4	0.33	0.16	0.16	0.34
-3	0.12	0.12	0.47	0.29
-2	0.61	0.00	0.04	0.36
-1	0.10	0.20	0.13	0.57
1	0.31	0.09	0.06	0.53
2	0.00	0.63	0.15	0.22
3	0.37	0.24	0.28	0.11
4	0.27	0.24	0.21	0.27
5	0.27	0.39	0.07	0.27
6	0.24	0.37	0.15	0.24
7	0.13	0.28	0.22	0.37
8	0.30	0.16	0.20	0.34
9	0.23	0.15	0.20	0.42
10	0.22	0.20	0.26	0.31

Figure 4.1: I-OnuI Position Weight Matrix. Position Weight Matricies (PWMs) that have been experimentally determined for a number of LAGLIDADG homing endonucleases are available on the site. These matrices are displayed in graphical format as a sequence logo plot and in tabular format. The ability of the enzyme to cleave DNA targets that contain a given nucleotide at a specific position within the target site is indicated by letter height and positions with greater overall information content are more specifically recognized. For example, positions such as -8 and -7 are recognized with very high fidelity and positions such as +/- 10 display low fidelity (i.e. basepair substitutions at those positions are tolerated by the enzyme).

and a potential genomic target is differentially weighted depending on the fidelity of recognition displayed at each DNA base pair position. The difference between the two search strategies described above in (b) and (c) is simple: the use of a simple identity matrix would return those matches with the fewest mismatches, while the use of a matrix that accounts for recognition degeneracy would indicate sites that might be more distantly related to the protein's wild-type recognition site, while actually nevertheless indicating more tractable gene targeting sites. (d) Searches are also facilitated against 'modules' of basepair codons, systematically screened across the entire target site, for those enzymes that have been fully characterized and screened at that level for their engineerability and selectability. The algorithm also allows the user to search for genomic targets that can be approached through the assembly of 'chimeric' homing endonucleases (where N- and C-terminal domains of unrelated proteins are fused to create scaffolds that can recognize corresponding chimeric DNA target sites) [20, 36, 105].

Following search by one of the four methods outlined above, a list of putative target sequences are returned each with a score. Within these putative target sequences, the central four nucleotides are highlighted in blue. Each nucleotides within each target sequence links to mutations that may be helpful when changing specificity. In the module search, regions of the sequence with a module matching with high activity are highlighted in green. Clicking on a nucleotide also reports any modules that are available. Additionally for the module search, each score links to a separate web page that displays information regarding designability of each module which is color-coded for visualization. Modules that have demonstrated activity approaching the wild type enzyme against its target are colored blue while those where little activity is reported are colored red.



I-OnuI

Gene	Position	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8	9	10	11	Score
wild type sequence	N/A	T	T	T	C	C	A	C	T	T	A	T	T	C	A	A	C	C	T	T	T	T	A	2.52e+08
>gi|156630984|ref|NM_000898.4|	f 957/2611	a	T	T	a	g	t	g	c	T	A	T	T	C	c	t	C	C	T	a	c	T	c	1.3608e+08
>gi|156630984|ref|NM_000898.4|	r 1697/2611	c	a	a	g	g	A	C	a	T	t	T	T	C	t	c	t	t	g	T	c	T	g	9.3312e+07
>gi|156630984|ref|NM_000898.4|	r 2340/2611	a	a	c	a	a	c	C	a	c	A	T	T	C	A	g	t	C	c	a	g	a	g	6.86e+07
>gi|156630984|ref|NM_000898.4|	r 1534/2611	T	a	a	t	a	A	a	c	T	A	T	a	C	A	c	t	t	g	a	T	g	A	6.77376e+07
>gi|156630984|ref|NM_000898.4|	r 2471/2611	c	T	g	C	t	c	C	g	T	t	T	T	C	t	g	g	g	c	c	T	c	g	5.6448e+07

WT I-OnuI: TTTCCACTTATTCAACCTTTTA
>gi|156630984|r: ATTAGTGCTATTCCTCCTACTC
10
6
7
10
9
8
5
9
10
10
9
8
7
6
5
4
3
2
1
0

Figure 4.2: MAOB Module Search Results. Following search of the MAOB gene for I-OnuI like target sequences, a list of best matches are presented together with their orientation, position, and scores (A). Candidate target sequences indicate which positions mismatch (lower case) and which are within the central four bases that are more difficult to design against. Following a module search, each score links to a more detailed view describing how well each module matches the target sequence (B). Modules are represented by individual bars and those that match well to the sequence are colored blue and those that do not match are colored red. Nucleotides that directly match the wild type target are indicated by vertical bars.

BIBLIOGRAPHY

[1] Paul D. Adams, Ralf W. Grosse-Kunstleve, Li Wei W. Hung, Thomas R. Ioerger, Airlie J. McCoy, Nigel W. Moriarty, Randy J. Read, James C. Sacchettini, Nicholas K. Sauter, and Thomas C. Terwilliger. PHENIX: building new software for automated crystallographic structure determination. *Acta crystallographica. Section D, Biological crystallography*, 58(Pt 11):1948–1954, November 2002.

[2] Peter L. Adams, Mary R. Stahley, Anne B. Kosek, Jimin Wang, and Scott A. Strobel. Crystal structure of a self-splicing group I intron with both exons. *Nature*, 430(6995):45–50, July 2004.

[3] J. A. Aínsa, N. J. Ryding, N. Hartley, K. C. Findlay, C. J. Bruton, and K. F. Chater. WhiA, a protein of unknown function conserved among gram-positive bacteria, is essential for sporulation in Streptomyces coelicolor A3(2). *Journal of bacteriology*, 182(19):5470–5478, October 2000.

[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, October 1990.

[5] Sylvain Arnould, Patrick Chames, Christophe Perez, Emmanuel Lacroix, Aymeric Duclert, Jean-Charles C. Epinat, François Stricher, Anne-Sophie S. Petit, Amélie Patin, Sophie Guillier, Sandra Rolland, Jesús Prieto, Francisco J. Blanco, Jerónimo Bravo, Guillermo Montoya, Luis Serrano, Philippe Duchateau, and Frédéric Pâques. Engineering of large numbers of highly specific homing endonucleases that induce recombination on novel DNA targets. *Journal of molecular biology*, 355(3):443–458, January 2006.

[6] Sylvain Arnould, Christophe Perez, Jean-Pierre P. Cabaniols, Julianne Smith, Agnès Gouble, Sylvestre Grizot, Jean-Charles C. Epinat, Aymeric Duclert, Philippe Duchateau, and Frédéric Pâques. Engineered I-CreI derivatives cleaving sequences from the human XPC gene can induce highly efficient gene correction in mammalian cells. *Journal of molecular biology*, 371(1):49–65, August 2007.

[7] Justin Ashworth and David Baker. Assessment of the optimization of affinity and specificity at protein-DNA interfaces. *Nucleic acids research*, 37(10), June 2009.

[8] Justin Ashworth, James J. Havranek, Carlos M. Duarte, Django Sussman, Raymond J. Monnat, Barry L. Stoddard, and David Baker. Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature*, 441(7093):656–659, June 2006.

[9] Justin Ashworth, Gregory K. Taylor, James J. Havranek, S. Arshiya Quadri, Barry L. Stoddard, and David Baker. Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. *Nucleic acids research*, 38(16):5601–5608, September 2010.

[10] Timothy L. Bailey, Nadya Williams, Chris Misleh, and Wilfred W. Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic acids research*, 34(Web Server issue):W369–W373, July 2006.

[11] M. Belfort and P. S. Perlman. Mechanisms of intron mobility. *The Journal of biological chemistry*, 270(51):30237–30240, December 1995.

[12] T. A. Bickle and D. H. Krüger. Biology of DNA restriction. *Microbiological reviews*, 57(2):434–450, June 1993.

[13] J. M. Bujnicki. Understanding the evolution of restriction-modification systems: clues from sequence and structure comparisons. *Acta biochimica Polonica*, 48(4):935–967, 2001.

[14] Adrian A. Canutescu and Roland L. Dunbrack. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein science : a publication of the Protein Society*, 12(5):963–972, May 2003.

[15] Piyali Chatterjee, Kristina L. Brady, Amanda Solem, Yugong Ho, and Mark G. Caprara. Functionally distinct nucleic acid binding sites for a group I intron encoded RNA maturase/DNA homing endonuclease. *Journal of molecular biology*, 329(2):239–251, May 2003.

[16] Zhilei Chen, Fei Wen, Ning Sun, and Huimin Zhao. Directed evolution of homing endonuclease I-SceI with altered sequence specificity. *Protein engineering, design & selection : PEDS*, 22(4):249–256, April 2009.

[17] Zhucheng Chen, Haijuan Yang, and Nikola P. Pavletich. Mechanism of homologous recombination from the RecA-ssDNA/dsDNA structures. *Nature*, 453(7194):489–484, May 2008.

[18] B. Chevalier, J. R. J. Monnat, and B.L. Stoddard. *Homing endonucleases and inteins*, volume 16, pages 34–47. Springer Verlag, Berlin, 2005.

[19] Brett Chevalier, Monique Turmel, Claude Lemieux, Raymond J. Monnat, and Barry L. Stoddard. Flexible DNA target site recognition by divergent homing endonuclease isoschizomers I-CreI and I-MsoI. *Journal of molecular biology*, 329(2):253–269, May 2003.

[20] Brett S. Chevalier, Tanja Kortemme, Meggen S. Chadsey, David Baker, Raymond J. Monnat, and Barry L. Stoddard. Design, activity, and structure of a highly specific artificial endonuclease. *Molecular cell*, 10(4):895–905, October 2002.

[21] A. Choulika, A. Perrin, B. Dujon, and J. F. Nicolas. Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of Saccharomyces cerevisiae. *Molecular and cellular biology*, 15(4):1968–1973, April 1995.

[22] Michelle Christian, Tomas Cermak, Erin L. Doyle, Clarice Schmidt, Feng Zhang, Aaron Hummel, Adam J. Bogdanove, and Daniel F. Voytas. Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics*, 186(2):757–761, October 2010.

[23] M. Cohen-Tannoudji, S. Robine, A. Choulika, D. Pinto, F. El Marjou, C. Babinet, D. Louvard, and F. Jaisser. I-SceI-induced gene replacement at a natural locus in embryonic stem cells. *Molecular and cellular biology*, 18(3):1444–1448, March 1998.

[24] J. Z. Dalgaard, A. J. Klar, M. J. Moser, W. R. Holley, A. Chatterjee, and I. S. Mian. 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 research*, 25(22):4626–4638, November 1997.

[25] Bareket Dassa, Nir London, Barry L. Stoddard, Ora Schueler-Furman, and Shmuel Pietrokovski. Fractured genes: a novel genomic arrangement involving new split inteins and a new homing endonuclease family. *Nucleic acids research*, 37(8):2560–2573, May 2009.

[26] A. Delahodde, V. Goguel, A. M. Becam, F. Creusot, J. Perea, J. Banroques, and C. Jacq. Site-specific DNA endonuclease and RNA maturase activities of two homologous intron-encoded proteins from yeast mitochondria. *Cell*, 56(3):431–441, February 1989.

[27] T. Doetschman, R. G. Gregg, N. Maeda, M. L. Hooper, D. W. Melton, S. Thompson, and O. Smithies. Targetted correction of a mutant HPRT gene in mouse embryonic stem cells. *Nature*, 330(6148):576–578, 1987.

[28] Jeffrey B. Doyon, Vikram Pattanayak, Carissa B. Meyer, and David R. Liu. Directed evolution and substrate specificity profile of homing endonuclease I-SceI. *Journal of the American Chemical Society*, 128(7):2477–2484, February 2006.

[29] M. Drouin, P. Lucas, C. Otis, C. Lemieux, and M. Turmel. Biochemical characterization of I-CmoeI reveals that this H-N-H homing endonuclease shares functional similarities with H-N-H colicins. *Nucleic acids research*, 28(22):4566–4572, November 2000.

[30] B. Dujon. Group I introns as mobile genetic elements: facts and mechanistic speculations–a review. *Gene*, 82(1):91–114, October 1989.

[31] Stanislaw Dunin-Horkawicz, Marcin Feder, and Janusz M. Bujnicki. Phylogenomic analysis of the GIY-YIG nuclease superfamily. *BMC genomics*, 7, 2006.

[32] D. R. Edgell and D. A. Shub. Related homing endonucleases I-BmoI and I-TevI use different strategies to cleave homologous recognition sites. *Proceedings of the National Academy of Sciences of the United States of America*, 98(14):7898–7903, July 2001.

[33] David R. Edgell, Victoria Derbyshire, Patrick Van Roey, Stephen LaBonne, Matthew J. Stanger, Zhong Li, Thomas M. Boyd, David A. Shub, and Marlene Belfort. Intron-encoded homing endonuclease I-TevI also functions as a transcriptional autorepressor. *Nature structural & molecular biology*, 11(10):936–944, October 2004.

[34] David R. Edgell, Matthew J. Stanger, and Marlene Belfort. Importance of a single base pair for discrimination between intron-containing and intronless alleles by endonuclease I-BmoI. *Current biology : CB*, 13(11):973–978, May 2003.

[35] Paul Emsley and Kevin Cowtan. Coot: model-building tools for molecular graphics. *Acta crystallographica. Section D, Biological crystallography*, 60(Pt 12 Pt 1):2126–2132, December 2004.

[36] Jean-Charles C. Epinat, Sylvain Arnould, Patrick Chames, Pascal Rochaix, Dominique Desfontaines, Clémence Puzin, Amélie Patin, Alexandre Zanghellini, Frédéric Pâques, and Emmanuel Lacroix. A novel engineered meganuclease induces homologous recombination in yeast and mammalian cells. *Nucleic acids research*, 31(11):2952–2962, June 2003.

[37] K. E. Flick, M. S. Jurica, R. J. Monnat, and B. L. Stoddard. DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-PpoI. *Nature*, 394(6688):96–101, July 1998.

[38] P. Friedhoff, I. Franke, G. Meiss, W. Wende, K. L. Krause, and A. Pingoud. A similar active site for non-specific and specific endonucleases. *Nature structural biology*, 6(2):112–113, February 1999.

[39] Martin C. Frith, Neil F. Saunders, Bostjan Kobe, and Timothy L. Bailey. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS computational biology*, 4(4):e1000071+, April 2008.

[40] Monika Fuxreiter and István Simon. Protein stability indicates divergent evolution of PD-(D/E)XK type II restriction endonucleases. *Protein science : a publication of the Protein Society*, 11(8):1978–1983, August 2002.

[41] E. A. Galburt, M. S. Chadsey, M. S. Jurica, B. S. Chevalier, D. Erho, W. Tang, R. J. Monnat, and B. L. Stoddard. Conformational changes and cleavage by the homing endonuclease I-PpoI: a critical role for a leucine residue in the active site. *Journal of molecular biology*, 300(4):877–887, July 2000.

[42] Huirong Gao, Jeff Smith, Meizhu Yang, Spencer Jones, Vesna Djukanovic, Michael G. Nicholson, Ande West, Dennis Bidney, S. Carl Falco, Derek Jantz, and L. Alexander Lyznik. Heritable targeted mutagenesis in maize using a designed endonuclease. *The Plant journal : for cell and molecular biology*, 61(1):176–187, January 2010.

[43] William J. Geese, Yong K. Kwon, Xiaoping Wen, and Richard B. Waring. In vitro analysis of the relationship between endonuclease and maturase activities in the bi-functional group I intron-encoded protein, I-AniI. *European journal of biochemistry / FEBS*, 270(7):1543–1554, April 2003.

[44] Stefan Glaser, Konstantinos Anastassiadis, and A. Francis Stewart. Current issues in mouse genome engineering. *Nature genetics*, 37(11):1187–1193, November 2005.

[45] V. Goguel, A. Delahodde, and C. Jacq. Connections between RNA splicing and DNA intron mobility in yeast mitochondria: RNA maturase and DNA endonuclease switching experiments. *Molecular and cellular biology*, 12(2):696–705, February 1992.

[46] R. Gonzalez-Nicieza, D. P. Turner, and B. A. Connolly. DNA binding and cleavage selectivity of the Escherichia coli DNA G:T-mismatch endonuclease (vsr protein). *Journal of molecular biology*, 310(3):501–508, July 2001.

[47] H. Goodrich-Blair, V. Scarlato, J. M. Gott, M. Q. Xu, and D. A. Shub. A self-splicing group I intron in the DNA polymerase gene of Bacillus subtilis bacteriophage SPO1. *Cell*, 63(2):417–424, October 1990.

[48] J. M. Gott, A. Zeeh, D. Bell-Pedersen, K. Ehrenman, M. Belfort, and D. A. Shub. Genes within genes: independent expression of phage T4 intron open reading frames and the genes in which they reside. *Genes & development*, 2(12B):1791–1799, December 1988.

[49] N. V. Grishin. Mh1 domain of Smad is a degraded homing endonuclease. *Journal of molecular biology*, 307(1):31–37, March 2001.

[50] Sylvestre Grizot, Julianne Smith, Fayza Daboussi, Jesús Prieto, Pilar Redondo, Nekane Merino, Maider Villate, Séverine Thomas, Laetitia Lemaire, Guillermo Montoya, Francisco J. Blanco, Frédéric Pâques, and Philippe Duchateau. Efficient targeting of a SCID gene by an engineered single-chain homing endonuclease. *Nucleic acids research*, 37(16):5405–5419, September 2009.

[51] James J. Havranek, Carlos M. Duarte, and David Baker. A simple physical model for the prediction and design of protein-DNA interactions. *Journal of molecular biology*, 344(1):59–70, November 2004.

[52] James J. Havranek and Pehr B. Harbury. Automated design of specificity in molecular recognition. *Nature structural biology*, 10(1):45–52, January 2003.

[53] C. H. Heldin, K. Miyazono, and P. ten Dijke. TGF-beta signalling from cell membrane to nucleus through SMAD proteins. *Nature*, 390(6659):465–471, December 1997.

[54] J. G. Henikoff, E. A. Greene, S. Pietrokovski, and S. Henikoff. Increased coverage of protein families with the blocks database servers. *Nucleic acids research*, 28(1):228–230, January 2000.

[55] S. Henikoff, J. G. Henikoff, W. J. Alford, and S. Pietrokovski. Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene*, 163(2), October 1995.

[56] R. M. Henke, R. A. Butow, and P. S. Perlman. Maturase and endonuclease functions depend on separate conserved domains of the bifunctional protein encoded by the group I intron aI4 alpha of yeast mitochondrial DNA. *The EMBO journal*, 14(20):5094–5099, October 1995.

[57] A. Hinnen, J. B. Hicks, and G. R. Fink. Transformation of yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 75(4):1929–1933, April 1978.

[58] Y. Ho, S. J. Kim, and R. B. Waring. A protein encoded by a group I intron in Aspergillus nidulans directly assists RNA splicing and is a DNA endonuclease. *Proceedings of the National Academy of Sciences of the United States of America*, 94(17):8994–8999, August 1997.

[59] David M. Hoover and Jacek Lubkowski. DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic acids research*, 30(10), May 2002.

[60] Elena M. Ibryashkina, Marina V. Zakharova, Vladimir B. Baskunov, Ekaterina S. Bogdanova, Maxim O. Nagornykh, Marat M. Den'mukhamedov, Bogdan S. Melnik, Andrzej Kolinski, Dominik Gront, Marcin Feder, Alexander S. Solonin, and Janusz M. Bujnicki. Type II restriction endonuclease R.Eco29kI is a member of the GIY-YIG nuclease superfamily. *BMC structural biology*, 7:48+, July 2007.

[61] Y. Jin, G. Binkowski, L. D. Simon, and D. Norris. Ho endonuclease cleaves MAT DNA in vitro by an inefficient stoichiometric reaction mechanism. *The Journal of biological chemistry*, 272(11):7352–7359, March 1997.

[62] Brett K. Kaiser, Matthew C. Clifton, Betty W. Shen, and Barry L. Stoddard. The structure of a bacterial DUF199/WhiA protein: domestication of an invasive endonuclease. *Structure (London, England : 1993)*, 17(10):1368–1376, October 2009.

[63] Katarzyna H. Kaminska, Mikihiko Kawai, Michal Boniecki, Ichizo Kobayashi, and Janusz M. Bujnicki. Type II restriction endonuclease R.Hpy188I belongs to the GIY-YIG nuclease superfamily, but exhibits an unusual active site. *BMC structural biology*, 8:48+, November 2008.

[64] Lukasz Knizewski and Krzysztof Ginalski. Bacterial DUF199/COG1481 proteins including sporulation regulator WhiA are distant homologs of LAGLIDADG homing endonucleases that retained only DNA binding. *Cell cycle (Georgetown, Tex.)*, 6(13):1666–1670, July 2007.

[65] B. H. Koller and O. Smithies. Inactivating the beta 2-microglobulin locus in mouse embryonic stem cells by homologous recombination. *Proceedings of the National Academy of Sciences of the United States of America*, 86(22):8932–8935, November 1989.

[66] Jan Kosinski, Marcin Feder, and Janusz M. Bujnicki. 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, 2005.

[67] J. C. Kowalski, M. Belfort, M. A. Stapleton, M. Holpert, J. T. Dansereau, S. Pietrokovski, S. M. Baxter, and V. Derbyshire. Configuration of the catalytic GIY-YIG domain of intron endonuclease I-TevI: coincidence of computational and molecular findings. *Nucleic acids research*, 27(10):2115–2125, May 1999.

[68] U. C. Kühlmann, G. R. Moore, R. James, C. Kleanthous, and A. M. Hemmings. Structural parsimony in endonuclease active sites: should the number of homing endonuclease families be redefined? *FEBS letters*, 463(1-2):1–2, December 1999.

[69] Simon J. Labrie, Julie E. Samson, and Sylvain Moineau. Bacteriophage resistance mechanisms. *Nature reviews. Microbiology*, 8(5):317–327, May 2010.

[70] Lorna E. Lancaster, Wolfgang Wintermeyer, and Marina V. Rodnina. Colicins and their potential in cancer treatment. *Blood cells, molecules & diseases*, 38(1):15–18, 2007.

[71] Markus Landthaler, Betty W. Shen, Barry L. Stoddard, and David A. Shub. I-BasI and I-HmuI: two phage intron-encoded endonucleases with homologous DNA recognition sequences but distinct DNA specificities. *Journal of molecular biology*, 358(4):1137–1151, May 2006.

[72] Fabienne Le Provost, Simon Lillico, Bruno Passet, Rachel Young, Bruce Whitelaw, and Jean-Luc L. Vilotte. Zinc finger nuclease technology heralds a new era in mammalian transgenesis. *Trends in biotechnology*, 28(3):134–141, March 2010.

[73] Ting Li, Sheng Huang, Wen Zhi Z. Jiang, David Wright, Martin H. Spalding, Donald P. Weeks, and Bing Yang. TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic acids research*, 39(1):359–372, January 2011.

[74] David M. Lilley. The interaction of four-way DNA junctions with resolving enzymes. *Biochemical Society transactions*, 38(2):399–403, April 2010.

[75] Qingqing Liu, Victoria Derbyshire, Marlene Belfort, and David R. Edgell. Distance determination by GIY-YIG intron endonucleases: discrimination between repression and cleavage functions. *Nucleic acids research*, 34(6):1755–1764, 2006.

[76] Antonella Longo, Christopher W. Leonard, Gurminder S. Bassi, Daniel Berndt, Joseph M. Krahn, Traci Tanaka M. Hall, and Kevin M. Weeks. Evolution from DNA to RNA recognition by the bI3 LAGLIDADG maturase. *Nature structural & molecular biology*, 12(9):779–787, September 2005.

[77] N. M. Luscombe, R. A. Laskowski, and J. M. Thornton. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic acids research*, 29(13):2860–2874, July 2001.

[78] Amanda Nga-Sze N. Mak, Abigail R. Lambert, and Barry L. Stoddard. Folding, DNA recognition, and function of GIY-YIG endonucleases: crystal structures of R.Eco29kI. *Structure (London, England : 1993)*, 18(10):1321–1331, October 2010.

[79] Maria J. Marcaida, Inés G. Muñoz, Francisco J. Blanco, Jesús Prieto, and Guillermo Montoya. Homing endonucleases: from basics to therapeutic applications. *Cellular and molecular life sciences : CMLS*, 67(5):727–748, March 2010.

[80] Preeti Mehta, Krishnamohan Katta, and Sankaran Krishnaswamy. HNH family subclassification leads to identification of commonality in the His-Me endonuclease superfamily. *Protein science : a publication of the Protein Society*, 13(1):295–300, January 2004.

[81] M. Michael Gromiha, Jörg G. Siebers, Samuel Selvaraj, Hidetoshi Kono, and Akinori Sarai. Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. *Journal of molecular biology*, 337(2):285–294, March 2004.

[82] P. J. Mitchell and R. Tjian. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science (New York, N.Y.)*, 245(4916):371–378, July 1989.

[83] T. Naito, K. Kusano, and I. Kobayashi. Selfish behavior of restriction-modification systems. *Science (New York, N.Y.)*, 267(5199):897–899, February 1995.

[84] D. Nathans and H. O. Smith. Restriction endonucleases in the analysis and restructuring of dna molecules. *Annual review of biochemistry*, 44:273–293, 1975.

[85] Norimichi Nomura, Yayoi Nomura, Django Sussman, Daniel Klein, and Barry L. Stoddard. Recognition of a common rDNA target site in archaea and eukarya by analogous LAGLIDADG and His-Cys box homing endonucleases. *Nucleic acids research*, 36(22):6988–6998, December 2008.

[86] Jerzy Orlowski, Michal Boniecki, and Janusz M. Bujnicki. I-Ssp6803I: the first homing endonuclease from the PD-(D/E)XK superfamily exhibits an unusual mode of DNA recognition. *Bioinformatics (Oxford, England)*, 23(5):527–530, March 2007.

[87] T. L. Orr-Weaver, J. W. Szostak, and R. J. Rothstein. Yeast transformation: a model system for the study of recombination. *Proceedings of the National Academy of Sciences of the United States of America*, 78(10):6354–6358, October 1981.

[88] Jay Painter and Ethan A. Merritt. Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta crystallographica. Section D, Biological crystallography*, 62(Pt 4):439–450, April 2006.

[89] Frédéric Pâques and Philippe Duchateau. Meganucleases and DNA double-strand break-induced recombination: perspectives for gene therapy. *Current gene therapy*, 7(1):49–66, February 2007.

[90] S. Pietrokovski. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic acids research*, 24(19):3836–3845, October 1996.

[91] A. Pingoud and A. Jeltsch. Structure and function of type II restriction endonucleases. *Nucleic acids research*, 29(18):3705–3727, September 2001.

[92] Yaroslava Y. Polosina and Claire G. Cupples. MutL: conducting the cell's response to mismatched and misaligned DNA. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 32(1):51–59, January 2010.

[93] Yaroslava Y. Polosina, Justin Mui, Photini Pitsikas, and Claire G. Cupples. The Escherichia coli mismatch repair protein MutL recruits the Vsr and MutH endonucleases in response to DNA damage. *Journal of bacteriology*, 191(12):4041–4043, June 2009.

[94] Matthew H. Porteus. Mammalian gene targeting with designed zinc finger nucleases. *Molecular therapy : the journal of the American Society of Gene Therapy*, 13(2):438–446, February 2006.

[95] G. Pósfai, V. Kolisnychenko, Z. Bereczki, and F. R. Blattner. Markerless gene replacement in Escherichia coli stimulated by a double-strand break in the chromosome. *Nucleic acids research*, 27(22):4409–4415, November 1999.

[96] Pilar Redondo, Jesús Prieto, Inés G. Muñoz, Andreu Alibés, Francois Stricher, Luis Serrano, Jean-Pierre P. Cabaniols, Fayza Daboussi, Sylvain Arnould, Christophe Perez, Philippe Duchateau, Frédéric Pâques, Francisco J. Blanco, and Guillermo Montoya. Molecular basis of xeroderma pigmentosum group C DNA recognition by engineered meganucleases. *Nature*, 456(7218):107–111, November 2008.

[97] Linus Sandegren, David Nord, and Britt-Marie M. Sjöberg. SegH and Hef: two novel homing endonucleases whose genes replace the mobC and mobE genes in several T4-related phages. *Nucleic acids research*, 33(19):6203–6213, 2005.

[98] Matheshwaran Saravanan, Janusz M. Bujnicki, Iwona A. Cymerman, Desirazu N. Rao, and Valakunja Nagaraja. Type II restriction endonuclease R.KpnI is a member of the HNH nuclease superfamily. *Nucleic acids research*, 32(20):6129–6135, November 2004.

[99] Michelle Scalley-Kim, Audrey McConnell-Smith, and Barry L. Stoddard. Coevolution of a homing endonuclease and its host target sequence. *Journal of molecular biology*, 372(5):1305–1319, October 2007.

[100] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, 18(20):6097–6100, October 1990.

[101] G. D. Schuler, S. F. Altschul, and D. J. Lipman. A workbench for multiple alignment construction and analysis. *Proteins*, 9(3):180–190, 1991.

[102] Betty W. Shen, Daniel F. Heiter, Siu-Hong H. Chan, Hua Wang, Shuang-Yong Y. Xu, Richard D. Morgan, Geoffrey G. Wilson, and Barry L. Stoddard. Unusual target site disruption by the rare-cutting HNH restriction endonuclease PacI. *Structure (London, England : 1993)*, 18(6):734–743, June 2010.

[103] Betty W. Shen, Markus Landthaler, David A. Shub, and Barry L. Stoddard. DNA binding and cleavage by the HNH homing endonuclease I-HmuI. *Journal of molecular biology*, 342(1):43–56, September 2004.

[104] G. H. Silva, J. Z. Dalgaard, M. Belfort, and P. Van Roey. Crystal structure of the thermostable archaeal intron-encoded endonuclease I-DmoI. *Journal of molecular biology*, 286(4):1123–1136, March 1999.

[105] George H. Silva, Marlene Belfort, Wolfgang Wende, and Alfred Pingoud. From monomeric to homodimeric endonucleases and back: engineering novel specificity of LAGLIDADG enzymes. *Journal of molecular biology*, 361(4):744–754, August 2006.

[106] Einat Sitbon and Shmuel Pietrokovski. New types of conserved sequence domains in DNA-binding regions of homing endonucleases. *Trends in biochemical sciences.*, 28(9):473–477, September 2003.

[107] D. R. Smith, P. Richterich, M. Rubenfield, P. W. Rice, C. Butler, H. M. Lee, S. Kirst, K. Gundersen, K. Abendschan, Q. Xu, M. Chung, C. Deloughery, T. Aldredge, J. Maher, R. Lundstrom, C. Tulig, K. Falls, J. Imrich, D. Torrey, M. Engelstein, G. Breton, D. Madan, R. Nietupski, B. Seitz, S. Connelly, S. McDougall, H. Safer, R. Gibson, L. Doucette-Stamm, K. Eiglmeier, S. Bergh, S. T. Cole, K. Robison, L. Richterich, J. Johnson, G. M. Church, and J. I. Mao. Multiplex sequencing of 1.5 Mb of the Mycobacterium leprae genome. *Genome research*, 7(8):802–819, August 1997.

[108] Julianne Smith, Sylvestre Grizot, Sylvain Arnould, Aymeric Duclert, Jean-Charles C. Epinat, Patrick Chames, Jesús Prieto, Pilar Redondo, Francisco J. Blanco, Jerónimo Bravo, Guillermo Montoya, Frédéric Pâques, and Philippe Duchateau. A combinatorial approach to create artificial homing endonucleases cleaving chosen sequences. *Nucleic acids research*, 34(22):e149, December 2006.

[109] Monika Sokolowska, Honorata Czapinska, and Matthias Bochtler. Crystal structure of the beta beta alpha-Me type II restriction endonuclease Hpy99I with target DNA. *Nucleic acids research*, 37(11):3799–3810, June 2009.

[110] Monika Sokolowska, Honorata Czapinska, and Matthias Bochtler. Hpy188I-DNA pre- and post-cleavage complexes–snapshots of the GIY-YIG nuclease mediated catalysis. *Nucleic acids research*, 39(4):1554–1564, March 2011.

[111] Barry Stoddard and Marlene Belfort. Social networking between mobile introns and their host genes. *Molecular microbiology*, 78(1):1–4, October 2010.

[112] Barry L. Stoddard. Homing endonuclease structure and function. *Quarterly reviews of biophysics*, 38(1):49–95, February 2005.

[113] Barry L. Stoddard. Homing endonucleases: from microbial genetic invaders to reagents for targeted DNA modification. *Structure (London, England : 1993)*, 19(1):7–15, January 2011.

[114] F. William Studier. Protein production by auto-induction in high density shaking cultures. *Protein expression and purification*, 41(1):207–234, May 2005.

[115] Amarendran R. Subramanian, Jan Weyer-Menkhoff, Michael Kaufmann, and Burkhard Morgenstern. DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. *BMC bioinformatics*, 6(1):66+, 2005.

[116] Django Sussman, Meg Chadsey, Steve Fauce, Alex Engel, Anna Bruett, Ray Monnat, Barry L. Stoddard, and Lenny M. Seligman. Isolation and characterization of new homing endonuclease specificities at individual target site positions. *Journal of molecular biology*, 342(1):31–41, September 2004.

[117] T. Szczepanek and J. Lazowska. Replacement of two non-adjacent amino acids in the S.cerevisiae bi2 intron-encoded RNA maturase is sufficient to gain a homing-endonuclease activity. *The EMBO journal*, 15(14):3758–3767, July 1996.

[118] J. W. Szostak, T. L. Orr-Weaver, R. J. Rothstein, and F. W. Stahl. The double-strand-break repair model for recombination. *Cell*, 33(1):25–35, May 1983.

[119] Ryo Takeuchi, Abigail R. Lambert, Amanda Nga-Sze N. Mak, Kyle Jacoby, Russell J. Dickson, Gregory B. Gloor, Andrew M. Scharenberg, David R. Edgell, and Barry L. Stoddard. Tapping natural reservoirs of homing endonucleases for targeted gene modification. *Proceedings of the National Academy of Sciences of the United States of America*, 108(32):13077–13082, August 2011.

[120] Gregory K. Taylor, Daniel F. Heiter, Shmuel Pietrokovski, and Barry L. Stoddard. Activity, specificity and structure of I-Bth0305I: a representative of a new homing endonuclease family. *Nucleic acids research*, September 2011.

[121] Summer B. Thyme, Jordan Jarjour, Ryo Takeuchi, James J. Havranek, Justin Ashworth, Andrew M. Scharenberg, Barry L. Stoddard, and David Baker. Exploitation of binding energy for catalysis and design. *Nature*, 461(7268):1300–1304, October 2009.

[122] James J. Truglio, Benjamin Rhau, Deborah L. Croteau, Liqun Wang, Milan Skorvaga, Erkan Karakas, Matthew J. DellaVecchia, Hong Wang, Bennett Van Houten, and Caroline Kisker. Structural insights into the first incision reaction during nucleotide excision repair. *The EMBO journal*, 24(5):885–894, March 2005.

[123] S. E. Tsutakawa, H. Jingami, and K. Morikawa. Recognition of a TG mismatch: the crystal structure of very short patch repair endonuclease in complex with a DNA duplex. *Cell*, 99(6):615–623, December 1999.

[124] S. E. Tsutakawa, T. Muto, T. Kawate, H. Jingami, N. Kunishima, M. Ariyoshi, D. Kohda, M. Nakagawa, and K. Morikawa. Crystallographic and functional studies of very short patch repair endonuclease. *Molecular cell*, 3(5):621–628, May 1999.

[125] Tzvi Tzfira and Charles White. Towards targeted mutagenesis and gene replacement in plants. *Trends in biotechnology*, 23(12):567–569, December 2005.

[126] Bennett Van Houten, Deborah L. Croteau, Matthew J. DellaVecchia, Hong Wang, and Caroline Kisker. 'Close-fitting sleeves': DNA damage recognition by the UvrABC nuclease system. *Mutation research*, 577(1-2):92–117, September 2005.

[127] P. Van Roey, C. A. Waddling, K. M. Fox, M. Belfort, and V. Derbyshire. Intertwined structure of the DNA-binding domain of intron endonuclease I-TevI with its substrate. *The EMBO journal*, 20(14):3631–3637, July 2001.

[128] Patrick Van Roey, Lisa Meehan, Joseph C. Kowalski, Marlene Belfort, and Victoria Derbyshire. Catalytic domain structure and hypothesis for function of GIY-YIG intron endonuclease I-TevI. *Nature structural biology*, 9(11):806–811, November 2002.

[129] Chu Wang, Philip Bradley, and David Baker. Protein-protein docking with backbone flexibility. *Journal of molecular biology*, 373(2):503–519, October 2007.

[130] J. M. Wenzlau, R. J. Saldanha, R. A. Butow, and P. S. Perlman. A latent intron-encoded maturase is also an endonuclease needed for intron mobility. *Cell*, 56(3):421–430, February 1989.

[131] Wei Yang, Jae Young Y. Lee, and Marcin Nowotny. Making and breaking nucleic acids: two-Mg2+-ion catalysis and substrate specificity. *Molecular cell*, 22(1):5–13, April 2006.

[132] Lei Zhao, Richard P. Bonocora, David A. Shub, and Barry L. Stoddard. The restriction fold turns to the dark side: a bacterial homing endonuclease with a PD-(D/E)-XK motif. *The EMBO journal*, 26(9):2432–2442, May 2007.

[133] Lei Zhao, Stefan Pellenz, and Barry L. Stoddard. Activity and specificity of the bacterial PD-(D/E)XK homing endonuclease I-Ssp6803I. *Journal of molecular biology*, 385(5):1498–1510, February 2009.

Appendix A
SUPPLEMENTARY MATERIAL

Supplementary Figure S1. Genes encoding individual representatives of the novel gene family described above, corresponding to several ORFS found within group I introns and inteins and as free-standing genes, including the endonuclease we have now named I-Bth0305I, were each synthesized as codon-optimized reading frames for bacterial expression. The lower case sequences at the 5' and 3' ends of the synthetic genes provide NcoI and NotI cloning sites; the former encode a 'Met-Gly' N-terminal protein sequence immediately followed by the third residue from the protein.

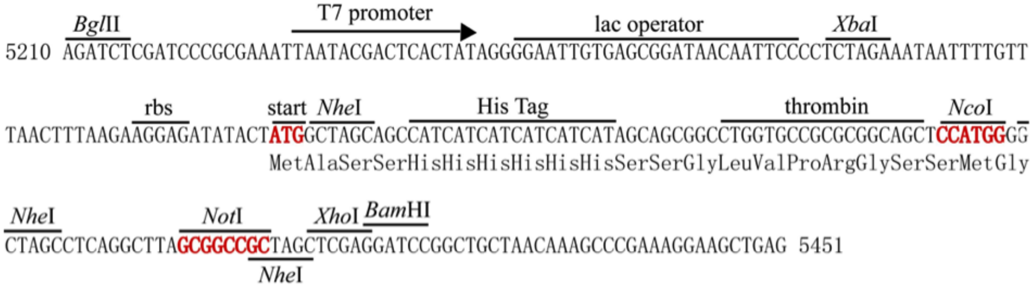
```
>I-Bth0305I
atggggCGCGCGTGGTCTCCCTCTATTGAACAAAAACAAATCGTAATTGACGGATACGCATCACCAGA
TATCTCAATCCGCGAACTGGCAAAAGAATTAGGCATTGGTAAAGACGCTTTAATGAAATACGCAGATG
AACATGATTTAACAAAAGTACCAAAAGATCGCTTGAACGCAGAACAACGCAAAGCTATTAAAGACTGG
AAAGGCGAAATTTCACTGAATGAGTTAGCAAACAATATTGGCATTTCCTTAGCAGGTGTACAGAAACG
CATGAAAAAACTTGGCATCGACACAAAACAATATATTGAAAAAAATCCACACTACCGTCCTGGCAAAA
CACCGCGCGATGAAGCCTTTTTTAAAGATATTGACAACCCTAAATACTCCTCAATTGAACTGGCTGAA
AAATACGGAGTCTCAGACGTCGCAATCCAACGCTGGCGTAAAAAACGTCATGGTAAATTTAAACCGCA
GATTGATACCTCCACACACTTGACTACCCCAGAACGCCGCGTTAAAGAAATTTTAGATGAACTCGACA
TTGTGTATTTTACTCACCATGTAGTAGAAGGTTGGAACGTAGATTTTTACCTGGGAAAAAAATTGGCT
ATCGAAGTTAATGGGGTTTATTGGCATAGCAAACAGAAAAACGTAAACAAGGATAAACGTAAACTTAG
CGAGTTGCATTCTAAAGGCTACCGTGTATTAACAATCGAAGATGACGAATTAAATGATATTGACAAAG
TAAAACAACAGATTCAAAAATTTTGGGTAACACACATCTCAAATGGTATGTAATAAgcggccgc
>30603
atggggTTACAATCGGAAATCGTACATATCGTAATATTGAACAATGGCAAGTTGACTTTATTGTAAAA
CACGCAGACCGTAAACTGAAAGATATTGGAAAAGAAATTAATTTAGACGAACGCCGCGTCGGAGAAAT
TTTGAAATTATTAGGCATTAAACGTACCCGTCATCGCAAAATTTACTTACCTAAAACCGCAGAAGTCG
```

```
AACAAGAACTCAAAAATCCGTACCTCTCACATGTAGAAATCGCAATCAAATATGGAGTATCAGACACC
TGTGTAGCAAAACGTCGCAAAGAATTAAACGTAAAAGTTCGCAAAAAAAAACTACGACACACTTCTCGA
ACAACAAGTAGAACAAATGTTATTATCATTAGACCTCGCTTTCATTAAACAAAAACGTATTGATAAAT
GGAGTATTGATTTTTACCTTGGTCGTAAATATTGTCTGGACGTGCATGGCAAATGGGCACACTCACTG
AAGAAAATTAAAGAACGCGATAAACGTAAACTGTTATTTATGGAAGAAGGATGTTATAAATATTTAGT
AATCCACGAAGAAGAATTAGCAAACAAAGAAAAAGTGTTACAAAAAATTAAAGAATTTACTATGGGTT
TTCCTTGTTAATAAgcggccgc
>o128-1
atggggCGTAAAAAAAAAGAAAAACGCGAAATTGGTAATAAAATTGAAAAAACCGTATCCACTATCCT
CACACGTCTGAACCTGCCTTTCGAAGAACAAGTAAGCGTCGACCGTTATACCGTAGATTTCTTAGTAA
ACAAAAAATATATCGTCGAATGTTATGGAGATTTCTGGCATTGCAACCCTCAACAATATACCTCATCA
TACTTCAATCGTGGCAAAAAGAAAACAGCAGAAGAAATTTGGGAACGTGATACTGAACGTAAAAAAAA
ATTTGAACAAATGGGATATAAATTTCTTTGTCTTTGGGAAGATGACATTCGTAATAACCCTAAAATTG
TGCAAAGTAAAATTAAAAAACATATTAAACTTGATGAAGGCTTATAATAAgcggccgc
>o128-3
atggggCGTCGTAAAAATGGAAAAAAAGTAAGTAAAGTTAATACAATTGAAACTAAAGTAGCTACCAT
TCTCACATCATTAAACGTACCTTTCGAACAACAAGTATCTATTGATCGTTATACCGTCGATTTCTTAA
TTAATAAAAAATATATCGTTGAATGCTATGGTGATTTTTGGCACTGTAATCCCCATCAATATACAAGC
TCTTACTTTAATCGTGGAAAAAAAAAAACAGCAGAGGAAATTTGGGAACGCGACACTAAACGTAAAGA
ACAATTCGAAAAAATGGGTTATAAATTTCTGTGTTTATGGGAATCGGACATTCGCAACAATCCTAAAA
TCGTTCGTTCTAAAATTAAAAAAGGCGTAGATAAATTAGACAATTAAgcggccgc
>o282
(atggggCAACTGTACGAAAAAGGACTTGTCTCTGCACAAACTGTTTTAGAAATTTTTGATTTAAATC
CAGACCAAGAGATTGAACGCAAACGCTTTGACGTAATTCAACTTACTGCCTTAGGACAAACAGCAGGA
GCTCCCGGTGGTGGAATGGGTGGTGGATTTGGCGGTGGTATGCCTGGTGGGATGGGTGGAATGCCGGA
GATTGGTGGCGTCCCAGAAACAGGCGGTGGTGCTCCAATTGTACCTGGTGGTGGCGCTCCAGCCGAAG
GTGGTGCCCCTGCAGCCCCAGCGACCCCTATGACTGCATCAGCATCAACTGTTATCGACGTAGCAAAC
CCAGCGCAATTTGGCAATAAAATTTTAAAAAAGAAAACACGCGACAAAATTCTGCAAGAAAAACAAAA
AATTTACAAGCAATACCAATCAAAACTTCAGCAACAATACGGAGATGGTCAGCGCGATGAAAAAGGAC
GTGTTATTTTTACCGGACCCGAACGCACTCTCTTAAAAAACCTGATCCAATACAAACAGAACGGTATT
```

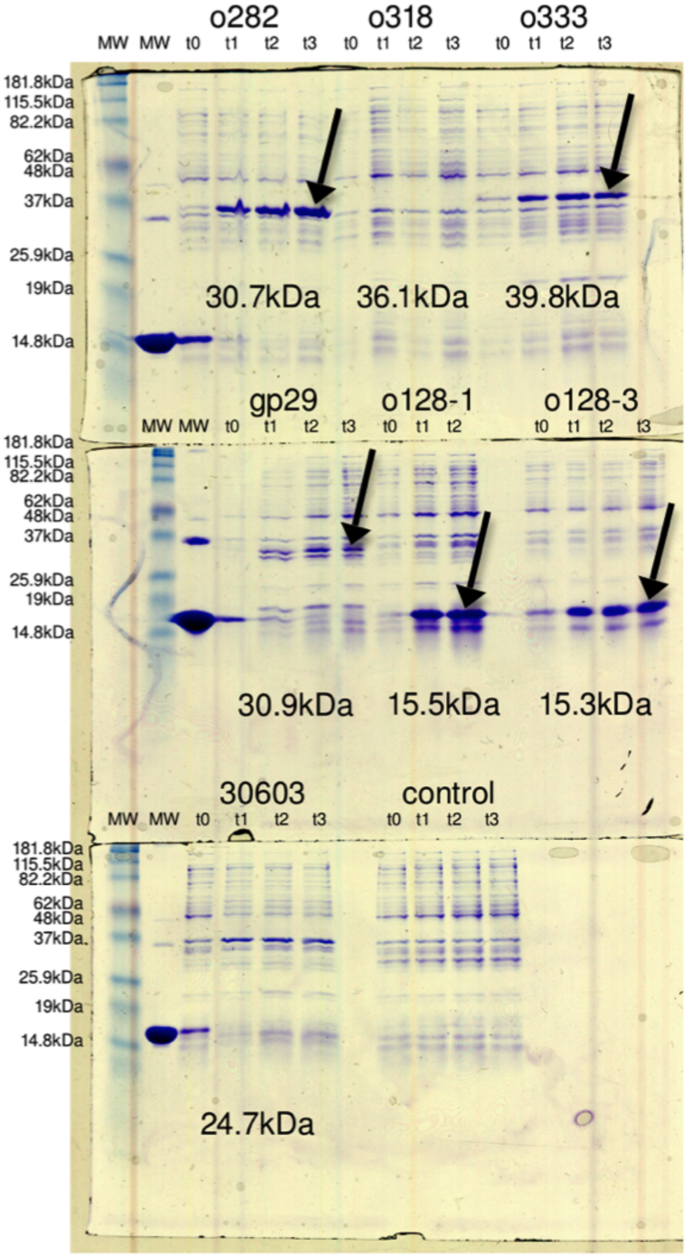
```
ATTAAACAGCCCGTATTTCCACAGTATCGTTTAAAAATCGGTGATGAAGAATATCCAATTGACTTTGC
CTTGCCCTATTTGAAAATCGGAATTGAAGCGGACGGTGAAATCTTTCATTCTAGCGATAAACAAATTC
AACATGATCGGGAACGTGACCGTAAATTAAACCAGGCCGGATGGACAATTCTGCGTTTTAAAGATACT
GAAATTGAAGAACAAATCCAGGGTGTTATGTCCACGATTGTAAAATTTATCATGAAAAAAGAAATGGC
GGCACAACATTTAAAAAACCAGAAATCATAATAAgcggccgc
>o318
atggggTTATCGAAACGTCAAGTAAGTTTACAATATGGTTTTAACACCCGTCTTATTGAGGCCCTCTT
AGATAAAGGAGTCTTAACTGTATTACCTGGTTGCACCCCTTTACGTCCTAAAATCGATCTGGCAAGTC
TCCGTAATCTCGTTGAAGATGAACATTACGTTGTATGTCGCGAATGCGGTTCCTATCAAGCCATGATT
AGCACCAAACATTTACGCGCGTGTAGTGGCACTGATTTAGTACGTTATAAAGAAAACCATCCAAATGC
TCCAGTTATGTCCGCTCTGGCTGCCGAAAACAAAGCGAAAACCGAAGCTCAGAAAGTTGCCCAATCAG
ATAAGTTAAAAGCCCGCTTCAAAACGATCAAAGGTGAAGAAACACGTCGTCAAATCGCAAAAGCCTCT
CGTCGCCTGCATGACTCTGGATACCGTGAAAAAGCAGCAGCCCACCTCCGCAATTTGAATAATGACCC
CGATCAGCGCGAACTCCTCCGTGAAAAAACCCTGGCACGTTGGGCATCTGGAGATCTCCGTGAAATCG
TCGAAGGTTGGCACCGTGATAATCGTAAAGAATCCCTGGCTTGTGCGGCTAATGCCCGTCGTCATATT
AAAAAAAAACGCTCTAAACTCCATCTCCGTTTTAAACGTGCGATGATTGATTCTGGTCTCTCAGGCTT
TGTCACCGAACACGAAGTAGGTTTCTACTCTATTGATGAAGCTCACCCGGTACTGAAAATTGCTGTAG
AAGTAGACGGTTGTTATTGGCATGGTTGCGAAGAATGTGGTCACCCGGGGATTGGCGAAATCAAAGTT
TTAGACCGCCGCAAAACTACTTTCTTAACTCGCCGTGGCTGGCAAGTCATCCGTGTACAAGAACATGA
AATTAAAGCAGATATTAATGCCTGTATTGGTCAATTGCGTGAAATTATCGAACAACGTGGCGCAGCAT
AATAAgcggccgc
>o333
atggggATTTGTAAAATTTGCAACACAAAATTTAAACGTATCTTGGGACATATCATTAAAAAACATGC
TAACCTGAAAGTTGGTTTATTAGAATACTTATCGTTTTATTATAACTTCGATATCATTAAAAATTATT
TAGACGGTTTGAGCGCGCAACAAATTAGCCATGAAATTTCAAAAATTACAGACGGTGCCATTAACCCG
AATAAAAAAGACATTTTGAAAATCCTTAAAGACAAAAACATTGCCATTCGTTCGACCTCTGAAGCGAT
CGTATCCTGGACAAAAATTCGCGGCGGTGCATGGAACAAAAATTTAACCAAAGAAGAACATCCGTCGA
TCAAAAAATACGCAGAAAGTCGTAAAGGTCACAACAACGTTTATTATACTGGAACGGAAGAAAGTCGT
AAAAAAACACGCTACTGGGAATATCTTGCAGCGGAAGAATTACAGAATATTCGTGCTAAATCTGCAGA
AACACTGAAAAAATTATATAAATCCGGAGAAATTATTCATAAATCAAAACTGGATCCTGAATGGGCAG
```

```
AAGATTGTAAAAAAAAAACGCATTGATGGGTATAAAAAATGGCTTGAGCAAAACGATGTTATTTTTCAT
GGCGCCGAATCTAAATTAGAAAAACAGATTGCACGTGGTTTGGAACAAGAAAACATTCGTTACAAAAA
ACAACTGAAATTGAAAAAAGATAAATATTGCTATTTTTATGATTATTTACTGGTAGATTATAATATTA
TTATCGAATACAATGGAACTTACTGGCACTGCGATCCACGCAAATATGACAAAGAATACTATAACGTG
TCGAAAAAAAATGTATGCAAGCCAAATTTGGGAACGTGACTTAGATAAAAAAATGCTGGCCGAACGCAA
CGGTTACGACATGATTGTATTATGGGAAGAAGATTATCGCAGCTTAACAAATGAGGAGTTTCTGGAAA
AAACAATTGAAACCATTAAAAATAAAATTAATCAGAAAATTAAAAATTAATAAgcggccgc
```

Supplementary Figure S2. Genes described in Supplementary Figure S1 were subcloned into a pET (Novagen, Inc) vector that encorporates an N-terminal, 6-histidine affinity purification tag. Cloning sites (5' NcoI and 3' NotI) are shown in red.

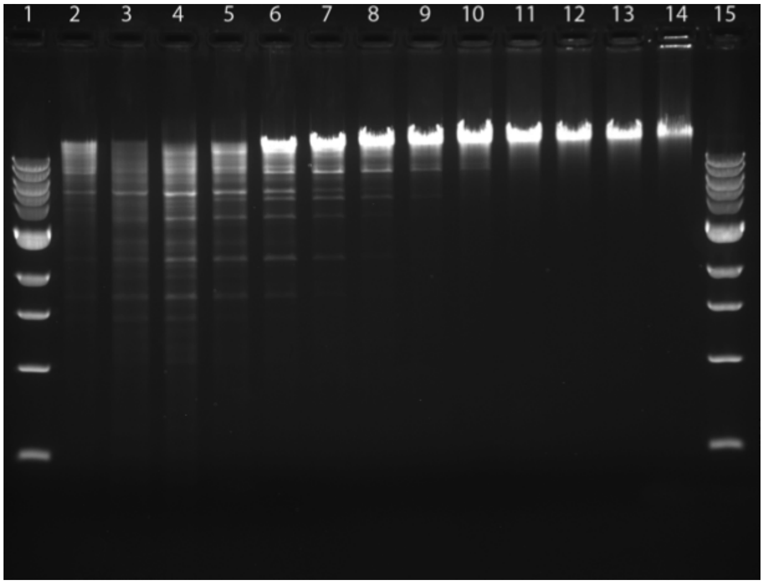

T7 promoter
*Bgl*II
lac operator
*Xba*I
5210 AGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTT
rbs
start
*Nhe*I
His Tag
thrombin
*Nco*I
TAACTTTAAGAAGGAGATATACTATGGCTAGCAGCCATCATCATCATCATCATAGCAGCGGCCTGGTGCCGCGCGGCAGCTCCATGGGG
MetAlaSerSerHisHisHisHisHisHisSerSerGlyLeuValProArgGlySerSerMetGly
*Nhe*I
*Not*I
*Xho*I
*Bam*HI
CTAGCCTCAGGCTTAGCGGCCGCTAGCTCGAGGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAG 5451
*Nhe*I

Supplementary Figure S3. Vsr-like homing endonuclease constructs displayed a wide range of behaviors during bacterial overexpression and purification. Five constructs displayed visible overexpression (see arrows); however most of those constructs were were found to be largely insoluble. Subsequent experiments focused on gp29 (later renamed I-Bth0305I as described in the text) which displayed lower overall levels and expression and cytotoxic activity, but was successfully purified for biochemical analyses as described in the text.



Supplementary Figure S4. Digestion and sequence analysis of sites in lambda phage DNA nicked and cleaved by I-Bth0305I.

a. The DNA products from the digests shown in lanes 3 and 5 below were sequenced as described in detailed methods. Lanes 1 and 15: 1 kilobase DNA ladder (NEB N3232). Lanes 2 to 14: Productes of one hour digests of phage lambda DNA at 37 C with 2-fold serial dilutions of I-Bth0305I. Lane 2 contains 1 microgram enzyme per 50 microliters reaction volume. Digests performed in 50 mM NaCl, 1 mM MgCl2, 50 mM Tris-Cl pH 7.6.



b. Sites in bacteriophage lambda DNA observed to be hydrolyzed by I-Bth0305I. The upper collection of target sequences corresponds to lane 3 in the gel above; the lower collection of sequences corresponds to lane 5. For all sequences, the site of hydrolysis is indicated by the "_" symbol. Cleavage events evident in forward traces indicate hydrolysis of the bottom strand, while those evident in reverse reactions indicate hydrolysis of the top strand. In the latter cases, the sequence shown is that of the complementary strand; the nucleotide positions for these are printed in italics and followed by the letter C. All sequences in the lists below are thus oriented alike to reflect cleavage occurring on the opposite strand. Bases corresponding to the minimal consensus sequence at each of the observed sites are shown highlighted in bold face type.

```
                       TT G    _  C AA

Lane 3:


029571 TCATCTACATAAACACCTTCGTGAT_GTCTGCATGGAGACAAGACACCGGA

30652C TCCTTTTTTATTATTCGCATTCACC_CTCAAGCGTATTAACCAACAGTTCA

30817C TAAAAACGTCGATGACATTTGCCGT_AGCGTACTGAAGAAGCACCGCGAAA
```

031023 GTGGTTGTCATACCTGGTTTCTCTC_ATCTGCTTCTGCTTTCGCCACCATC

31678C CGAAAGTATGCGTACCGCCTGCTCT_CCCGATGGTTTATGCAGTGACGGCA

031921 GGGGTTTTGCTATCACGTTGTGAAC_TTCTGAAGCGGTGATGACGCCGAGC

32109C GATATTTCGCCGCGACATTCGTGCA_TCGTCAGAACTGACACAGGCCGAAG

032517 GGCATGTACAGGATTCATTGTCCTG_CTCAAAGTCCATGCCATCAAACTGC

33422C AGGGTGGCCTGTTGCTGGCTGCCCT_TCCGAATCTTTACTTGAACGAATCA

033471 TGCAATGGCATATTGCATGGTGTGC_TCCTTATTTATACATAACGAAAAAC

34544C AACGATCATATACATGGTTCTCTCC_AGAGGTTCATTACTGAACACTCGTC

034637 AATATGTTAATGAGAGAATCGGTAT_TCCTCATGTGTGGCATGTTTTCGTC

35048C CAGAAAATTAAGGGAAAATCGATTC_CTCTTATCTAGTTACTTAGATATTG

36192C CTTACGTTATCGTAAGCATTTGCTA_TCTCCTTTTCCGCCACTACATTCCT

36947C AATCGATGGAAAAACTTTTCTCTTT_ACCAAAACAAATGACAAGAGTCTGG

037383 GTTTCTTGAAGGTAAACTCATCACC_CCCAAGTCTGGCTATGCAGAAATCA

37594C CTCATGTTCAGGCAGGGATGTTCTC_ACCTGAGCTTAGAACCTTTACCAAA

38719C GCTCCTGTCCGGCAAAGTTACCTCT_GCCGAAGTTGAGTATTTTTGCTGTA

039104 TCCCTCAAATTGGGGGATTGCTATC_CCTCAAAACAGGGGGACACAAAAGA

039549 CAGCCAGCAAACCAAAACTCGACCT_GACAAACACAGACTGGATTTACGGG

040265 CAGAAATCAAAGCTAAGTTCGGACT_GAAAGGAGCAAGTGTATGACGGGCA

040797 AGACCAAAACAGGAAGCTATGGGCC_TGCTTAGGTGACGTCTCTCGTCAGG

41318C TTCGCGTCTGAATATCCTTTGGTTC_CCATACCGTATAACCATTTGGCTGT

41507C GCTTTGGCTTTAGCCGCTTCGGTTC_ATCAGCTCTGATGCCAATCCACGTG

41595C ATTTGCAAATCGAATGGTTGTTGCT_TCCACCATGCGAGGATATCTTCCTT

42257C TTTGGTCAATCACCTTGTTTTCCTC_GCACGACGTCTTAGCCACCGGATAT

43705C TGTGAGTTGCTGATTCGTTCGCGGT_TCCAGATTACCTGCTGATGATCAAC

043707 TGATCATCAGCAGGTAATCTGGAAC_CGCGAACGAATCAGCAACTCACAAA

043918 TCGAAAGCGTAGCTAAATTTCATTC_GCCAAAAAGCCCGATGATGAGCGAC

044197 TGCAAGTGCTCGCAACATTCGCTTA_TGCGGATTATTGCCGTAGTGCCGCG

44526C ATCATGCCGTTAATATGTTGCCATC_CGTGGCAATCATGCTGCTAACGTGT

045457 GGTACTGACTCGATTGGTTCGCTTA_TCAAACGCTTCGCTGCTAAAAAAGC

046187 CAAAATACACGAAGGAGTTAGCTGA_TGCTAAAGCTGAAAATGATGCTCTG

047306 CTTTTATTAACACGGTGTTATCGTT_TTCTAACACGATGTGAATATTATCT

47730C GATTGTTCTTTATTCATTTTGTCGC_TCCATGCGCTTGCTCTTCATCTAGC

47849C ATGCTTCCAGAGACACCTTATGTTC_TATACATGCAATTACAACATCAGGG

48091C AGAATGAACATCCCGCGTTCTTCCC_TCCGAACAGGACGATATTGTAAATT

Lane 5:

001073 ACTTTATGAAAACCCACGTTGAGCC_GACTATTCGTGATATTCCGTCGCTG

001327 GACAAGCGTATTGAAGGCTCGGTCT_GGCCAAAGTCCATCCGTGGCTCCAC

03814C GCATCAGGTGCGGTACTTTTGCGCC_TCCCAGCCGGACCGGCGCTGCGGCG

07932C CGGTTAACGGCAGGCGGTACGCCCC_GTCCAAGCCAGAGATGACAACTTCC

010706 CGCTGAGCCGACAGGCGCTGGCTGC_ACAGAAAGCGGGGATTTCCGTCGGG

012815 TTGTTCACCGTGGTGAGTTTGTCTT_CACGAAGGAGGCAACCAGCCGGATT

013141 GGAAAGTGAAACCCGGTATGGATGT_GGCTTCGGTCCCTTCTGTAAGAAAG

14219C CACGCAGGGGAAATATCTTTCCCCC_TCCGGCGTGCTTACCACGAAGCCGC

15059C GGAATACGCCACCTGACTTGGCCCC_GGCGACTCTGGGAACAATATGAATT

15677C CCGTGACACCGGATATGTTGGTATT_CCCCTCAGTGTCCAGCACCGGCGTA

017771 TTCTCGGAAAAGCAGATTGCGGATA_TCAGACAGGTTGAAACCAGCACGCG

20623C GCCGCTTCTGCCGCACTTTTGCTCT_GCGATGCTGATACCGCACTTCCCGC

021221 AGATGTTCTTGAATACCTTGGGGCC_GGTGAGAATTCGGCCTTTCCGGCAG

21221C CTGCCGGAAAGGCCGAATTCTCACC_GGCCCCAAGGTATTCAAGAACATCT

22031C CTTCACCAATAAATTCATTAGTTCC_GGCCAGCAGATTATAAATTTTTATG

022811 AAAGTCGGTTTTTTTTCTTCGTTTT_CTCTAACTATTTTCCATGAAATACA

23097C ACTTTTTTAAAGGACGGTTATCACA_TTCAAACATTAATTTTTTATGATAA

023246 ATTATTTTATTGTCATATTGTATCA_TGCTAAATGACAATTTGCTTATGGA

024128 TAAAATTAGAGTTGTGGCTTGGCTC_TGCTAACACGTTGCTCATAGGAGAT

027222 AGTCTATTAATGCATATATAGTATC_GCCGAACGATTAGCTCTTCAGGCTT

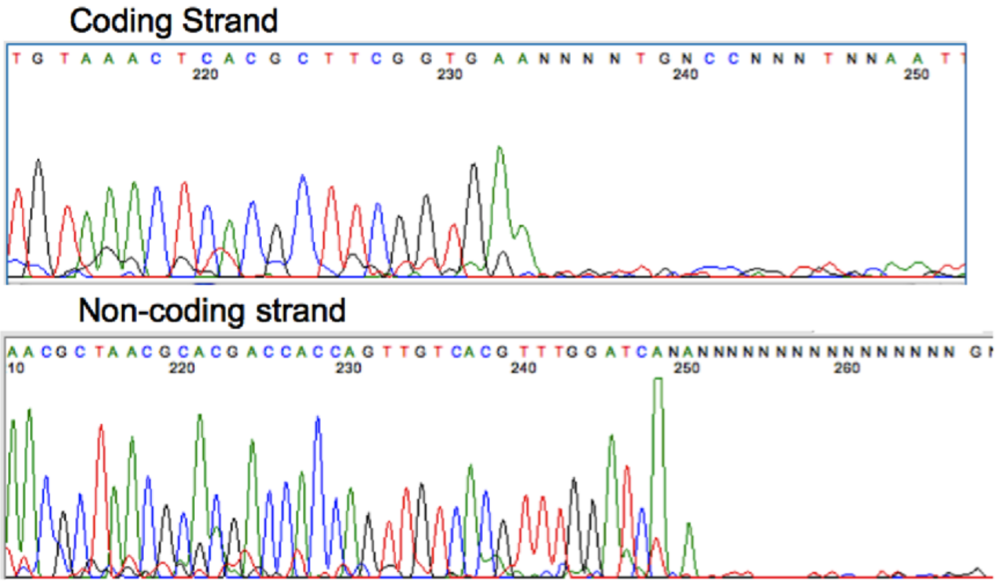
028769 CAGTGATTGCGATTCGCCTGTCTCT_GCCTAATCCAAACTCTTTACCCGTC

28848C GAAGTCATGAGCGCCGGGATTTACC_CCCTAACCTTTATATAAGAAACAAT

031921 GGGGTTTTGCTATCACGTTGTGAAC_TTCTGAAGCGGTGATGACGCCGAGC

```
35048C CAGAAAATTAAGGGAAAATCGATTC_CTCTTATCTAGTTACTTAGATATTG
36846C TCGCAGATGACGAATCATTGGGATT_CCCATCTTTTTTGTTTGTTGAAGGC
36947C AATCGATGGAAAAACTTTTCTCTTT_ACCAAAACAAATGACAAGAGTCTGG
37594C CTCATGTTCAGGCAGGGATGTTCTC_ACCTGAGCTTAGAACCTTTACCAAA
38719C GCTCCTGTCCGGCAAAGTTACCTCT_GCCGAAGTTGAGTATTTTTGCTGTA
39343C GCATCAGGCGGATATCGTTAGCCCA_CCCAGCAAAATTCGGTTTTCTGGCT
039549 CAGCCAGCAAACCAAAACTCGACCT_GACAAACACAGACTGGATTTACGGG
040797 AGACCAAAACAGGAAGCTATGGGCC_TGCTTAGGTGACGTCTCTCGTCAGG
043918 TCGAAAGCGTAGCTAAATTTCATTC_GCCAAAAAGCCCGATGATGAGCGAC
044196 TGCAAGTGCTCGCAACATTCGCTTA_TGCGGATTATTGCCGTAGTGCCGCG
44526C ATCATGCCGTTAATATGTTGCCATC_CGTGGCAATCATGCTGCTAACGTGT
045457 GGTACTGACTCGATTGGTTCGCTTA_TCAAACGCTTCGCTGCTAAAAAAGC
046187 CAAAATACACGAAGGAGTTAGCTGA_TGCTAAAGCTGAAAATGATGCTCTG
047306 CTTTTATTAACACGGTGTTATCGTT_TTCTAACACGATGTGAATATTATCT
48091C AGAATGAACATCCCGCGTTCTTCCC_TCCGAACAGGACGATATTGTAAATT
```

Supplementary Figure S5. Run-off sequencing of products generated by digestion of the 0305φ bacteriophage recA gene sequence by I-Bth0305I. The sequence traces are consistent with symmetric cleavage of the target site across the consensus " 5'-T-T-x-G-x6-C-x-A-A-3' " target identified in cleavage experiments using lambda phage DNA as a substrate. The exact cleavage pattern created in this experiment is somewhat ambiguous due to the addition of additional adenine bases at the 3' end of the replicated DNA strand during run-off polymerization by TaqI enzyme, but is consistent with generation of 5', 2 base overhangs as observed in the former experiment.

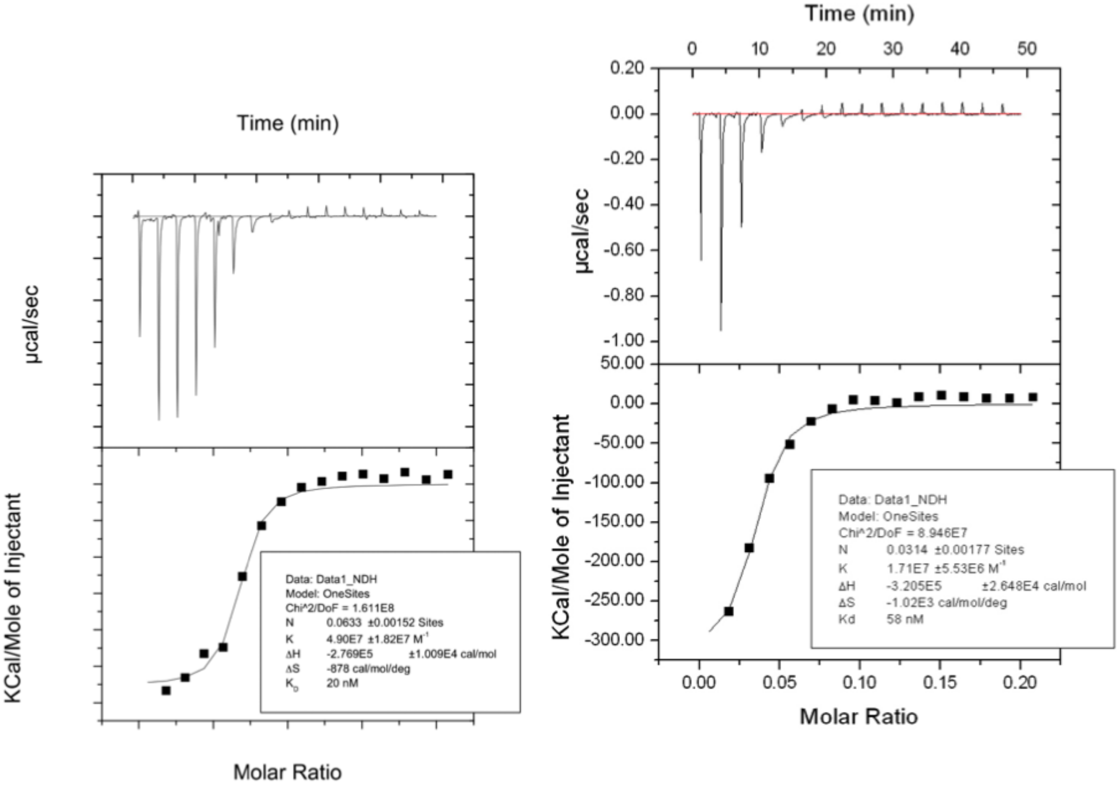


Coding Strand

Non-coding strand

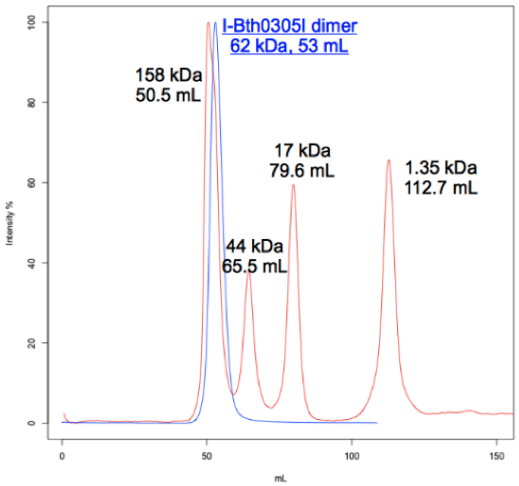
Supplementary Figure S6. Binding of I-Bth0305I to a 60 basepair recA target sequence. A: Binding experiments using isothermal titration calorimetry (where DNA duplexes were injected into a cell containing pure protein) indicate that the full-length enzyme binds its wild-type target in an exothermic reaction with a measured dissociation constant (KD) of 24 nM +/- 6 nM (ΔH= -2.8 x 105 cal/mol +/- 9 x 103 cal/mol; ΔS = -900 cal/mol/deg). B: Parallel experiments conducted with the same target, harboring a transversion of all central 8 basepairs of it sequence (WT: 5' - GGTGATCC - 3'; Panel B: 5' - CCACTAGG - 3') gives an estimated KD value of 58 nM.

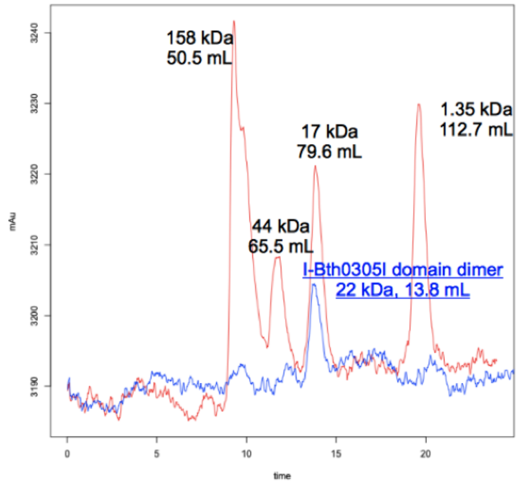
The stoichiometry of the binding reaction in these ITC experiments (expressed as the molar ratio of DNA duplex to individual protein subunits) is observed to be approximately 0.1, which is lower than the value 0.5 expected for a homodimeric protein binding to a single site or the value of 1.0 expected for binding of a protein monomer binding to the same site. This may be caused either by a substantial fraction of the protein in this experiment being found in a misfolded and/or nonfunctional state, or may possibly indicate that the protein actually forms a higher order oligomer upon assembly on its DNA target (a phenomena that has been observed previously for the I-Ssp6803I homing endonuclease)(10). However, the overall KD values measured in these experiments are determined independently from the stoichiometry of the reaction, and are highly reproducible over many independent experiments. All experiments were performed in triplicate.



A. Wild-type recA target site B. recA target with transversion at central 8 basepairs (KD 24 nM) (KD 58 nM)

Supplementary Figure S7. Size exclusion chromatography of the I-Bth0305I (top-blue) and the I-Bth0305I catalytic domain (bottom-blue) compared with molecular weight standards (red). Standards have molecular weights of 670 kDa, 158 kDa, 44 kDa, 17 kDa, and 1.35 kDa. In both experiments, elution times are consistent with dimerization and not with monomeric protein in solution.


I-Bth0305I dimer
62 kDa, 53 mL
158 kDa
50.5 mL
17 kDa
79.6 mL
1.35 kDa
112.7 mL
44 kDa
65.5 mL
Intensity %
0
20
40
60
80
100
0
50
100
150
mL


158 kDa
50.5 mL
1.35 kDa
112.7 mL
17 kDa
79.6 mL
44 kDa
65.5 mL
I-Bth0305I domain dimer
22 kDa, 13.8 mL
mAu
3190
3200
3210
3220
3230
3240
0
5
10
15
20
time