Canadian Journal of Microbiology
 Revue canadienne de microbiologie

Three new active members of the I-OnuI family of homing endonucleases

Journal:	<i>Canadian Journal of Microbiology</i>
Manuscript ID	cjm-2017-0067.R1
Manuscript Type:	Article
Date Submitted by the Author:	12-Apr-2017
Complete List of Authors:	Bilto, Iman M.; University of Manitoba Guha , Tuhin Kumar; University of Manitoba Wai, Alvan; University of Manitoba Hausner, Georg; University of Manitoba, Microbiology
Keyword:	mtDNA, mobile introns, protein purification, endonuclease assay

Three new active members of the I-OnuI family of homing endonucleases

Iman M. Bilito, Tuhin K. Guha, Alvan Wai and Georg Hausner*

Department of Microbiology, University of Manitoba, Winnipeg, MB, Canada R3T 2N2

*corresponding author: hausnerg@cc.umanitoba.ca

Abstract: *In vitro* characterization of three LAGLIDADG-type homing endonucleases (I-CcaI, I-CcaII and I-AstI) that belong to the I-OnuI family showed that they are functional homing endonucleases that cleave their respective cognate target sites. These endonucleases are encoded within group ID introns and appear to be orthologs that have inserted into three different mitochondrial genes, *rns*, *rnl* and *cox3*, respectively. The endonuclease activity of I-CcaI was tested using various substrates and its minimum DNA recognition sequence was estimated to be 26 nt. This set of homing endonucleases may provide some insight on how these types of mobile elements can migrate into new locations. This study provides additional endonucleases that can be added to the catalog of currently available HEs that may have various biotechnology applications.

Key words: mtDNA, mobile introns, protein purification, endonuclease assays

Introduction

Homing endonucleases (HEs) require long DNA target sites (~12 to 40 bp) and are therefore, highly specific DNA cleavage enzymes (Belfort and Roberts 1997; Chevalier and Stoddard 2001; Lambert *et al.* 2016). Homing endonuclease genes (HEGs) are embedded within elements such as Group I and sometimes Group II and archaeal introns (Dujon 1989; Belcour *et al.* 1997; Toor and Zimmerly 2002; Mullineux *et al.* 2010; Tocchini-Valentini *et al.* 2011). HEGs can also be freestanding and they can form the DNA-cutting component of inteins (Gimble 2000; Hafez and Hausner 2012). HEs can promote their own mobility and the genetic elements that encode them such as introns or inteins. In addition, some introns encode HEs that act as maturases; i.e. they appear to help the intron in splicing from their primary transcripts (Dujon and Belcour 1989; Szczepanek and Lazowska 1996; Belfort *et al.* 2002; Bolduc *et al.* 2003; Hausner 2012). HEGs are mobile genetic elements that can generate double-stranded DNA breaks in a cognate allele inducing the cell's DNA double-strand break repair mechanism to repair the break by using the HEG-containing DNA as a template (Dujon 1989; Belfort *et al.* 2002; Burt and Trivers 2006). Homing endonuclease activity not only results in repair that involves the non-reciprocal transfer of the HEG into new sites, but it also can be associated with co-conversion of markers flanking the HEG (or intron) insertion site (Mueller *et al.* 1996a,b; Parker *et al.* 1999; Muñoz *et al.* 2012).

Homing endonucleases can be assigned into six categories based on the presence of conserved amino acid motifs. Concerning inteins, group I and II introns, four types are the most relevant: LAGLIDADG (LHE), GIY-YIG, H-N-H and His-Cys box (Stoddard 2005; Skowronek and Bujnicki 2007; Hafez and Hausner 2012). The LAGLIDADG and GIY-YIG families are most frequently encountered in fungal mitochondrial group I introns (Haugen *et al.* 2005; Hausner 2012). LHEs generate staggered cuts with 4 nucleotide 3' overhangs at their cleavage

site (Belfort and Robert 1997; Chevalier and Stoddard 2001). LHEs in their active configuration can act as homodimers when the peptide contains one LAGLIDADG motif or they can act as monomers when the peptide contains two LAGLIDADG motifs (Haugen and Bhattacharya 2004; Dalgaard *et al.* 1997; Lucas *et al.* 2001). Some HEs have been shown to tolerate some sequence degeneracy within their recognition sites which gives HEGs (and their host element) the flexibility to invade new sites and thus provides the possibility of HEGs to avoid elimination by genetic drift (Scalley-Kim *et al.* 2007; Barzel *et al.* 2011; Edgell *et al.* 2011). Specificity with some allowance for variations among alleles makes HEs potential tools in genome editing such as gene replacements, targeted mutagenesis, gene drive mechanisms for pest control, and in designing vectors systems that require unique endonuclease target sites (Hafez and Hausner 2012). Fungal mitochondrial genomes have shown to be a resource for isolating new HEs that have new target specificities (Stoddard 2011; Baxter *et al.* 2012; Jacoby *et al.* 2012; Chan *et al.* 2013; Hafez *et al.* 2014a). Overall, few HEs have been biochemically characterized so far (Marcaida *et al.* 2010; Prieto *et al.* 2012) and this limits the application of HEs in targeting a wide variety of sites.

Previously, Monteiro-Vitorello *et al.* (2009) reported an intron within the mtDNA *rns* gene inserted at position mS915 (designation of *rns* intron positions and naming is according to the nomenclature proposed by Johansen and Haugen 2001; S = small subunit ribosomal gene, m = mitochondria) in the chestnut blight fungus *Cryphonectria parasitica*. Hafez *et al.* (2013) reannotated this intron and noted that it is actually inserted at position mS917; this study showed that this intron is a nested (or twintron) group ID intron where the internal group ID intron encodes a double motif LAGLIDADG-type ORF and it is inserted into an ORF-less external group ID intron. Further, this study noted that orthologues of this HEG are present in other

fungus species encoded within group ID introns inserted in *rnl*, *cox3*, *nad5* and *nad6* genes. Exploration of these HEs may provide some clues on how HEGs can move into different sites thereby promoting their spread within a genome, within a population, or between different species. The objectives for this study were to identify and characterize active members of the mS917 HEs family (i.e. orthologs of the mS917 LHE). The study focused on three members that are inserted within three different genes: (1) the mS917 LHE version that is encoded within an *rns* intron at position S917 in *Ceratocystis cacaofunesta* (causative agent of wilt disease in cacao); (2) a LHE encoded within intron 1 in the *rnl* gene in *C. cacaofunesta* (Ambrosio *et al.* 2013); and (3) a LHE that is encoded within the third intron of *cox3* in *Annulohypoxylon stygium* (an endophyte).

Materials and methods

Phylogenetic analysis of homing endonuclease sequences

The LHE dataset was based on extracting sequences from the NCBI database using the mS917 intron-encoded LHE amino acid (aa) sequence of *C. cacaofunesta* (GenBank accession: AFO38132.1) as a query in blastp. Forty sequences were extracted from GenBank and aligned with MAFFT (Kato and Standley 2013) and PRALINE (Simossis and Heringa 2005). Manual adjustments were made to the alignment, if necessary, with GeneDoc (Nicholas *et al.* 1997). The aa alignment was analyzed with neighbour joining (NJ; Saitou and Nei 1987) as implemented in the MEGA (version 7) program (Kumar *et al.* 2016). Distances were calculated with JTT (Jones *et al.* 1992) and its default settings were selected along with the complete deletion of gaps

option. The bootstrap option was implemented (1000 replicates) in order to assess the level of support for the tree topology (Felsenstein 1985).

HEs naming nomenclature

HEs are named according to the nomenclature proposed by Belfort and Roberts (1997). HEs encoded by group I introns have the prefix (I-) followed by abbreviations for the genus and species name. The final Roman numeral distinguishes multiple enzymes that have been characterized for that organism. The HEs examined in this study are designated as follows: I-CcaI for the intron-encoded protein (IEP) encoded by the *rnsI* of *C. cacaofunesta*, I-CcaII for the IEP encoded by the *rnlI* of *C. cacaofunesta*, and I-AstI for the IEP encoded by the *cox3i3* of *A. stygium*.

Construction of expression vectors and substrate plasmids

The HE sequences were retrieved from GenBank: I-CcaI (AF038132.1), I-CcaII (AF038136.1) and I-AstI (AHB33504.1). To allow for efficient expression in *Escherichia coli*, the genetic code for these HE open reading frames (ORFs) were codon optimized for *E. coli*. The pET-28b(+) vector was used for assembling the HE expression constructs (GenScript, New Jersey, USA). The LHE ORFs were inserted at the BamHI/NdeI restriction enzyme site and the vector provided an N-terminal 6X Histidine (His)-tag and the T7 promoter. The three expression vectors were named pI-CcaI, pI-CcaII and pI-AstI.

Suitable substrates for testing the above LHEs were designed as follows. The corresponding host gene sequences were examined and 100 nucleotides of exon sequences

flanking the intron insertion sites (~200 nt) were combined, synthesised, and inserted into the pUC57 (2.7 kb) vector (GenScript, New Jersey, USA). These substrates were named as follows: *prns*-SUB for testing the *C. cacaofunesta* *rns1* IEP (I-CcaI), *prnl*-SUB for testing the *C. cacaofunesta* *rnl* i1 IEP (I-CcaII), and *pcox3*-SUB for testing the *A. stygium* *cox3i3* IEP (I-AstI).

All plasmids were maintained in *E. coli* DH5 α (Thermo Fisher Scientific, Burlington, Ontario) and the plasmids were purified with the PrestoTM Mini Plasmid Kit (FroggaBio, Toronto, Ontario). Transformed cells were kept at -80°C in 80 % glycerol stocks (Sambrook *et al.* 1989) and purified plasmids were stored at -20°C.

Minimum recognition sequence and substrate plasmids for the: I-CcaI

In addition to the *prns*-SUB, a set of 8 substrate plasmids (*prns*-SUB2 to 9) were designed to estimate the minimum recognition sequence required for the LHE activity for the mS917 IEP (I-CcaI) of *C. cacaofunesta*. Essentially, the original substrate sequence was modified by shortening the potential recognition sequence two nucleotides at a time on either side of the LHE cleavage site by inserting nucleotide triplets that disrupt the potential LHE recognition sequence (Bae *et al.* 2009; see supplementary Fig. 1).

Expression and purification of recombinant proteins (HEs)

Conditions for the over expression and purification of the LHE proteins have been described previously in Hafez *et al.* (2014a,b). Briefly the LHE expression constructs were transferred into chemically competent *E. coli* BL21 (λ DE3) cells (New England Biolabs,

Whitby, Ontario) for protein expression was induced with isopropyl β -D-1-thiogalactopyranoside (IPTG) when cells reached an absorbance of value of 0.50 at OD₆₀₀. The purification of the His-tagged LHE proteins was performed on Ni-NTA columns. All the washing and elution fractions were collected and analyzed on a SDS polyacrylamide gel (12.5 %). Electrophoresis was used to evaluate HE protein purity. Samples were dialysed in dialysis buffer (50 mM Tris-HCl (pH 8.0), 100 mM NaCl, 6 mM β -mercaptoethanol) using the Slide-A-LyzerTM dialysis cassette (Millipore, Billerica, Massachusetts, USA) with a 10 kDa molecular weight cut-off in order to remove the imidazole. Dialysed samples were concentrated to 1 ml with the Amicon concentrator (Model 8050) using a Millipore YM-10 membrane. Usually, the purified, concentrated protein (LHE) was immediately assayed for its endonuclease activity before storing. However, for long term storage, 200 μ l of protein storage buffer [50 mM Tris-HCl (pH 8.0), 400 mM NaCl, 0.5 mM DTT, 10 % (w/v) glycerol] was added to the purified protein and the samples were stored at -80°C.

Endonuclease assays

The purified HEs were challenged with the appropriate DNA substrate plasmids in order to assess their endonuclease potential. Methodology is based on those previously presented in Hafez *et al.* (2014a,b). Briefly, the endonuclease reaction mixtures contained: 25 μ l of substrate plasmid (25 ng/ μ l), 5 μ l Invitrogen[®] Buffer React #3 (100 mM NaCl, 50 mM Tris-HCl (pH 8.0) and 10 mM MgCl₂) supplemented with 1 mM DTT, 5 μ l LHE protein (9 μ g) and H₂O to reach a final volume of 50 μ l. The endonuclease reactions were incubated at 37°C and 10 μ l aliquots were withdrawn at the following time points: 0, 30, 60, 90, and 120 min. The

endonuclease reactions were stopped by adding 2 μ l of 200 mM EDTA (pH 8.0) and 1 μ l proteinase K (1 mg/ml) to the reaction aliquots. Finally, the reaction products were resolved on a 1 % agarose gel ran at 80 volts and gels were stained with ethidium bromide (0.5 μ g/ml).

Mapping of the I-CcaI, I-CcaII, and I-AstI cleavage sites

The *in vitro* cleavage site mapping strategy was based on Bae *et al.* (2009) and has been previously described (Hafez *et al.* 2014a,b; Guha and Hausner 2014). The substrate plasmids were digested with the corresponding HE described above and the linearized substrates were cut from the gel and purified with the Wizard®SV Gel and PCR Clean-Up System (Promega; Thermo Fisher Scientific). Forty μ l of linearized substrate was treated with T4 DNA polymerase (T4 DNA pol) to generate blunt ends by removing the characteristic 4 nt 3'-OH overhangs generated by LAGLIDADG-type HEs (Bae *et al.* 2009). The T4 DNA pol treatment reaction contained the following components: 40 μ l of linearized substrate plasmid (25 ng/ μ l), 2 μ l of T4 DNA pol (5 units/ μ l; Invitrogen), 20 μ l of 5X T4 DNA polymerase buffer, 20 μ l dNTP mixture (0.5 mM) and H₂O to achieve the final volume of 100 μ l. The reaction was incubated at room temperature (24°C) for 20 min and thereafter the tubes were placed on ice for 5 min and finally the reactions were terminated by transferring to 70°C for 10 min. The DNA was purified with the Wizard-SV Gel and PCR Clean-Up system (Promega). The blunted substrate DNA was religated in the following reaction mixture: 20 μ l (0.25 μ g) of T4 DNA polymerase treated DNA, 2 μ l of T4 DNA ligase (1 U/ μ l; Invitrogen), 10 μ l of 5X ligase buffer and the addition of H₂O to achieve a final volume of 50 μ l. Ligation reactions were incubated at room temperature for 2 hours. Thereafter, the ligation mixtures were diluted 5-fold and a 10 μ l aliquot of the dilution

was transformed into chemically-competent *E. coli* DH5 α cells. The transformed cells were plated onto LB agar plates supplemented with ampicillin (100 μ g/ml). Single colonies were picked from these plates and these were used to inoculate 5 ml LB broth (100 μ g/ml ampicillin) tubes that were incubated for ~18 hours at 37°C. The plasmids were purified by the Wizard® Plus Minipreps DNA Purification System (Promega) and they were sent to the DNA Technologies Unit (NRC, Saskatoon, Saskatchewan) for cycle sequencing using the M13 forward and reverse primers. Untreated substrate plasmids were sent along with the HE and T4 DNA polymerase treated substrate plasmids. This allowed for examining what sequences had been removed by the T4 DNA polymerase treatment. The sequencing results were manually aligned with the GeneDoc program (version 2.7; Nicholas *et al.* 1997). Comparison of untreated samples with HE plus T4 DNA polymerase-treated samples revealed the four nucleotides that were removed and thus allowed for the determination of the actual LHE cleavage site.

Temperature requirement and the minimal DNA recognition sequence for I-CcaI

The I-CcaI HE (mS917 IEP) was evaluated with regards to its activity at different temperatures. Endonucleases assays were set up as described previously except the reactions were incubated for 1 hour at 20°C, 30°C, 37°C, 40°C, and 50°C. In addition, the I-CcaI LHE was further characterized concerning its minimum recognition sequence requirement. Here, the LHE was challenged with various substrate plasmids containing sequences modified to provide various lengths of the putative recognition sequence. The preparation of the substrate plasmids (*prns*-SUB2 to 9 have been described above) and the endonuclease assays were performed at 37°C for 2 hours as described above.

Results

Phylogeny of mS917 HEs and related LAGLIDADG-type ORFs

Forty LHE sequences were extracted from NCBI by using the mS917 IEP (I-CcaI) sequence from *C. cacaofunesta* as a query. Phylogenetic analysis with NJ (Fig. 1) showed that 10 sequences, all encoded within group ID introns, appear to be derived from a common ancestor. This grouping includes four mS917 IEPs along with three IEPs encoded within intron located within the *nad5* gene, one IEP encoded within the *rnl* gene, one IEP from the *cox3* gene, and one IEP from the *nad6* gene. This set of orthologues appears (bootstrap support of node = 98 %) to have inserted into at least five different genes, including rRNA and protein coding genes. It is also worth noting that the mS917 introns in *Ophiocordyceps tracentri* and *C. parasitica* are complex and appear to be nested introns where the internal group ID intron encodes the LHE and external group ID does not contain an ORF (Hafez *et al.* 2013). Originally, the goal was to characterize a member from each host gene (including *nad5* and *nad6*) and from the nested mS917 intron arrangements; however, we could only overexpress IEPs and show activity for IEPs from *C. cacaofunesta* (*rns*), *C. cacaofunesta* (*rnl*), and *A. stygium* (*cox3*).

HE protein overexpression in *E. coli* and purification

In total, ten mS917 LHEs and related LAGLIDADG HEs were identified and protein overexpression was attempted for three LHEs ORFs embedded within introns located in the *rns*, *rnl* and *cox3* genes. All of the studied ORFs were double-motif LAGLIDADG HEGs and their

lengths ranged from 433-551 aa. Three enzymes were expressed successfully and purified (supplementary Fig.2). I-AstI (50 kDa) and I-CcaII (47 kDa) were partially purified whereas I-CcaI HE (49.5 kDa) was consistently overexpressed in *E. coli* and purified from Ni-NTA columns. Therefore, this particular LHE was further analysed concerning temperature preference and for estimating its target sequence length requirement for DNA recognition and cleavage activity.

The overexpression of LHE ORFs in *E. coli* and subsequent characterization were noted to be difficult for several members of the mS917 group. Although all sequences were codon optimized for expression in *E. coli*, I-Cpa-917 (mS917 nested version; GenBank accession number: AAB84210.1) and I-CcaIII (*nad5*; GenBank accession number: AFO38108.1) could be expressed but showed no activity and I-CcaIV (*nad6*; GenBank accession number: AFO38135.1) failed to express.

Endonuclease activity of the mS917 IEP orthologs

Endonuclease assays were performed by incubating the purified LHEs (I-CcaI, I-AstI, I-CcaII) with the appropriate plasmid substrates (*prns*-SUB, *pcox3*-SUB and *prns*-SUB). The HE activity was tested at 37°C at different time periods (0, 30, 60, 90, and 120 min) and the best completely linearized substrates (2.8 kb) were already observed at 30 min (Fig. 2) for I-CcaI, I-AstI, and I-CcaII. The control assays, i.e., untreated (no HE) substrates (*prns*-SUB, *pcox3*-SUB, *prns*-SUB), showed no cleavage.

The *in vitro* cleavage mapping assay for I-CcaI, I-AstI, and I-CcaII confirmed that these LHEs are active and generate staggered ends with 4 nt overhang at the 3' end (Fig.3). The

HE cleaved and T4 DNA pol treated substrates (*prns*-SUB, *pcox3*-SUB, *prns*-SUB) showed that the following 4 nt were removed 5'-TAAT-3', 5'-ATAC-3', and 5'-ATGC-3' by I-CcaI, I-AstI, and I-CcaII, respectively.

I-CcaI activity was tested at different temperatures (20°C, 30°C, 37°C, 40°C, and 50°C) (Fig. 4). The enzyme showed activity at 30°C, 37°C, 40°C as the expected linearized product can be observed on the agarose gel. However, assays performed at 20°C and 50°C yielded multiple products indicating partial or no activity, as the substrate plasmid appears in its various supercoiled forms on the agarose gel (Fig. 4).

The I-CcaI enzyme was further characterized to estimate the minimal recognition sequence. The I-CcaI endonuclease was challenged with 8 different synthesized substrates containing variations of the known target sequence. The results showed (Fig. 5) that the enzyme lost its activity when the target sequence was less than 14 nt upstream of the actual cleavage site and less than 12 nt downstream from the cleavage site; i.e. the enzyme required a DNA recognition site that is estimated to be about 26 nt in length.

Discussion

Phylogeny of mS917 HEs and related LAGLIDADG type ORFs

Homing endonuclease genes in order to avoid elimination have either to gain new functions or insert into new sites (Goddard and Burt 1999). The mS917 clade of HEs appears to have inserted into at least five different target sites. This observation is also relevant to the application of LHEs as rare cutting DNA enzymes in biotechnology. Considerable efforts have

been made to redesign homing endonucleases to bind and cut at different target sites. This study presents examples where closely related LHEs have naturally adapted to different target sites. This warrants further investigation in the future to establish what features/amino acid changes might be involved in allowing LHEs to recognize new target sites.

The phylogenetic tree (Fig. 1) shows that the mS917 LHE orthologues belong to the I-OnuI family of HEs (node supported at 100 %). The I-OnuI homing endonuclease was originally described and characterized by Gibb and Hausner (2005) and Sethuraman *et al.* (2009). This endonuclease was studied in more detail by Takeuchi *et al.* (2011) showing that I-OnuI can be engineered to recognize and cut genes involved in monogenic human diseases. Such as the MAO-B gene (involved in neurodegenerative disorders including Parkinson's disease) that contains a DNA sequence, that differs from the native I-OnuI target site by only five bp (Takeuchi *et al.* 2011). Lambert *et al.* (2016) recently further demonstrated the potential utility of this family of LHEs concerning applications in biotechnology. This information might be useful in developing strategies for engineering LHEs to target sites located within genes of economic importance, such as alleles associated with human diseases.

The LAGLIDADG family of HEs have been explored as potential genome editing tools (Gimble 2000; Stoddard 2005, 2011, 2014). Therapeutic applications demand high precision in gene modification activity and HEs are considered compact target-specific 'molecular scissors' with little known issues with regards to off-target activities (Stoddard 2014; Takeuchi *et al.* 2014; Cox *et al.* 2015; Lambert *et al.* 2016). However, one potential drawback for this class of enzymes is the non-modular configuration; the DNA recognition and cleavage functions are combined within the same protein domain, thus engineering of LHEs is challenging (Hafez and Hausner 2012).

One of the limitations of the I-OnuI family, with regards to retargeting them for genome editing applications is the conserved nature of the 4 bp sequence comprising the central recognition motif (CRM, i.e., the 3' cleavage site overhangs). The three members of the mS917 clade characterized in this study displayed some variability within the CRM (5'-TAAT-3' for I-CcaI, 5'-ATAC-3' for I-AstI, and 5'-ATGC-3' for I-CcaII). This feature could make the mS917 clade an attractive LHE protein scaffold that could be modified for targeting a wide variety of potential target sites (Lambert *et al.* 2016).

Recent studies showed that the central four nucleotides within the CMR are not in direct contact with the LHE but they are still very important for controlling enzyme activity as this region undergoes bending during cleavage (Curuksu *et al.* 2009; Lambert *et al.* 2016). This is referred to as “indirect readout” where protein/DNA interactions are due to sequence-dependent interactions (involving conformation and flexibility of the DNA) between base-pairs and amino acid sequences and not due to direct interactions (H-bonding) between amino acids and nucleotides; this has been observed with regards to restriction enzymes and homing endonucleases (Molina *et al.* 2012; Yamasaki *et al.* 2012).

The study showed that I-CcaI, I-CcaII, and I-AstI can be overexpressed in *E. coli* and purified utilizing the N-terminal His-tag in sufficient amounts to demonstrate their activity as endonucleases and to map their cleavage sites. The most consistent LHE in our analyses with regards to overexpression and purification was I-CcaI. It was noted that the optimum activity was at 37°C and activity was highly reduced or lost at 20°C and 50°C, respectively. In previous studies, it was shown that endonuclease activity for LHEs tends to be around 37°C (see I-SceI group and I-CreI; Fonfara *et al.* 2011; Dürrenberger and Rochaix 1993). In contrast, the I-DmoI

LHE endonuclease, recovered from the thermophile *Desulfurococcus mobilis*, shows optimal activity at 65°C (Dalgaard *et al.* 1993; Silva and Belfort 2004).

Challenging the I-CcaI with 8 different synthesized substrates allowed for gaining better insights into the length of the DNA sequence required by the LHE, which appears to be around 26 nt. This is in the expected range for LHEs belonging to the I-OnuI family that tend to be targets sites around 22 nt (Lambert *et al.* 2016). In conclusion, this study presents three new members of the I-OnuI family of LHEs and although recently alternative genome editing reagents have become very popular, the low off-target activities make LHEs an attractive set of enzymes that warrants further exploration (Cox *et al.* 2015; Lambert *et al.* 2016). Engineering modular meganucleases by combining the LHE DNA cutting domains with more programmable DNA binding domains might be a promising direction for utilizing LHEs that belong to the I-OnuI family (Hafez and Hausner 2012; Wolfs *et al.* 2014, 2016; Boissel *et al.* 2014; Romano Ibarra *et al.* 2016; Guha *et al.* 2017).

Acknowledgements

This work is supported by a Discovery grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to G.H., I.M.B., T.K.G. and A.W. would like to acknowledge funding support from the Faculty of Science Graduate Award program (University of Manitoba) and the University of Manitoba Faculty of Graduate Studies GETS program. I.M.B. also would like to acknowledge the Islamic Development Bank Merit Scholarship Programme that supported part of her graduate program. Authors wish to thank the efforts of an anonymous reviewer.

References

- Ambrosio, A.B., do Nascimento, L.C., Oliveira, B.V., Teixeira, P.J., Tiburcio, R.A., Toledo Thomazella, D.P., Leme, A.F., Carazzolle, M.F., Vidal, R.O., Mieczkowski, P., Meinhardt, L.W., Pereira, G.A., Cabrera, O.G. 2013. Global analyses of *Ceratocystis cacaofunesta* mitochondria: from genome to proteome. *BMC Genomics*. **14**: 91. doi: 10.1186/1471-2164-14-91
- Bae, H., Kim, K.P., Song, J.M., Kim, J.H., Yang, J.S., Kwon, S.T. 2009. Characterization of intein homing endonuclease encoded in the DNA polymerase gene of *Thermococcus marinus*. *FEMS Microbiol. Let.* **297(2)**: 180-188. doi: <https://doi.org/10.1111/j.1574-6968.2009.01671.x>
- Barzel, A., Privman, E., Peeri, M., Naor, A., Shachar, E., Burstein, D., Lazary, R., Gophna, U., Pupko, T., Kupiec, M. 2011. Native homing endonucleases can target conserved genes in humans and in animal models. *Nucleic Acids Res.* **39(15)**: 6646–6659. doi: <https://doi.org/10.1093/nar/gkr242>
- Baxter, S., Lambert, A., Kuhar, R., Jarjour, R., Kulshina, N., Parmeggiani, F., Danaher, P., Gano, J., Baker, D., Stoddard, B., Scharenberg, A.M. 2012. Engineering domain fusion chimeras from I-OnuI family LAGLIDADG homing endonucleases. *Nucleic Acid Res.* **40(16)**: 7985-8000. doi: 10.1093/nar/gks502
- Belcour, L., Rossignol, M., Koll, F., Sellem, C.H., Oldani, C. 1997. Plasticity of the mitochondrial genome in *Podospora*. Polymorphism for 15 optional sequences: Group-I, group-II introns, intronic ORFs and an intergenic region. *Curr. Genet.* **31(4)**: 308-317.

- Belfort, M., Derbyshire, V., Parker, M.M., Cousineau, B., Lambowitz, A.M. 2002. Mobile introns: pathways and proteins. In: Craig, N.L., Craigie, R., Gellert, M., Lambowitz, A.M. (eds), *Mobile DNA II*, American Society for Microbiology Press, Washington DC 761-783.
- Belfort, M., Roberts, R.J. 1997. Homing endonucleases: Keeping the house in order. *Nucleic Acids Res.* **25 (17)**: 3379-3388.
- Boissel, S., Jarjour, J., Astrakhan, A., Adey, A., Gouble, A., Duchateau, P., Shendure, J., Stoddard, B.L., Certo, M.T., Baker, D., Scharenberg, A.M. 2014. MegaTALs: a rare-cleaving nuclease architecture for therapeutic genome engineering. *Nucleic Acids Res.* **42(4)**: 2591-2601. doi: 10.1093/nar/gkt1224
- Bolduc, J.M., Spiegel, P.C., Chatterjee, P., Brady, K.L., Downing, M.E., Caprara, M.G., Waring, R.B., Stoddard, B.L. 2003. Structural and biochemical analyses of DNA and RNA binding by a bifunctional homing endonuclease and group I intron splicing factor. *Genes Dev.* **17(23)**: 2875-2888.
- Burt, A., Trivers, R. 2006. *Genes in conflict: The biology of selfish genetic elements*. Harvard University Press, Cambridge, MA.
- Chan, Y.S., Takeuchi, R., Jarjour, J., Huen, D.S., Stoddard, B.L., Russell, S. 2013. The design and *in vivo* evaluation of engineered I-OnuI-based enzymes for HEG gene drive. *PLoS One* **8(9)**: e74254. doi: 10.1371/journal.pone.0074254
- Chevalier, B.S., Stoddard, B.L. 2001. Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. *Nucleic Acids Res.* **29(18)**: 3757-3774.
- Cox, D.B., Platt, R.J., Zhang, F. 2015. Therapeutic genome editing: prospects and challenges. *Nat. Med.* **21(2)**: 121-131. doi: 10.1038/nm.3793.

- Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E. 2004. WebLogo: A sequence logo generator. *Genome Res.* **14(6)**: 1188-1190.
- Curuksu, J., Zacharias, M., Lavery, R., Zakrzewska, K. 2009. Local and global effects of strong DNA bending induced during molecular dynamics simulations. *Nucleic Acids Res.* **37(11)**: 3766-3773. doi: 10.1093/nar/gkp234
- Dalgaard, J.Z., Garrett, R.A., Belfort, M. 1993. A site-specific endonuclease encoded by a typical archaeal intron. *Proc. Natl. Acad. Sci. USA* **90(12)**: 5414-5417.
- Dalgaard, J.Z., Klar, A.J., Moser, M.J., Holley, W.R., Chatterjee, A., Main, I.S. 1997. Statistical modeling and analysis of the LAGLIDAG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the H-N-H family. *Nucleic Acids Res.* **25(22)**: 4626-4638.
- Dujon, B. 1989. Group I introns as mobile genetic elements: Facts and mechanistic speculation. *Gene* **82**: 91-114.
- Dujon, B., Belcour, L. 1989. Mitochondrial DNA instabilities and rearrangements in yeasts and fungi. In: Berg DE, Howe MM (eds), *Mobile DNA*. The American Society for Microbiology, Washington, DC, pp. 861-878.
- Dürrenberger, F., Rochaix, J.D. 1993. Characterization of the cleavage site and the recognition sequence of the I-CreI DNA endonuclease encoded by the chloroplast ribosomal intron of *Chlamydomonas reinhardtii*. *Mol. Gen. Genet.* **236(2-3)**: 409-414.
- Edgell, D.R., Chalamcharla, V.R., Belfort, M. 2011. Learning to live together: mutualism between self-splicing introns and their hosts. *BMC Biol.* **9**: 22. doi: 10.1186/1741-7007-9-22.

- Fonfara, I., Curth, U., Pingoud, A., Wende, W. 2012. Creating highly specific nucleases by fusion of active restriction endonucleases and catalytically inactive homing endonucleases. *Nucleic Acids Res.* **40(2)**: 847-860. doi: 10.1093/nar/gkr788
- Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39(4)**: 783-791.
- Gibb, E.A., Hausner, G., 2005. Optional mitochondrial introns and evidence for a homing-endonuclease gene in the mtDNA *rnl* gene in *Ophiostoma ulmi* s. lat. *Mycol. Res.* **109(10)**: 1112-1126.
- Gimble, F.S. 2000. Invasion of a multitude of genetic niches by mobile endonuclease genes. *FEMS Microbiol. Let.* **185(2)**: 99-107.
- Goddard, M.R., Burt, A. 1999. Recurrent invasion and extinction of a selfish gene. *Proc. Natl. Acad. Sci. USA* **96(24)**: 13880-13885.
- Guha, T.K., Hausner, G. 2014. A homing endonuclease with a switch: characterization of a twintron encoded homing endonuclease. *Fungal Genet. Biol.* **65**: 57-68. doi: 10.1016/j.fgb.2014.01.004.
- Guha, T.K., Wai, A., Hausner, G. 2017. Programmable genome editing tools and their regulation for efficient genome engineering. *Comput. Struct. Biotechnol. J.* **15**: 146-160. doi: 10.1016/j.csbj.2016.12.006
- Hafez, M., Hausner, G. 2012. Homing endonuclease: DNA Scissors on a mission. *Genome* **55(8)**: 553-569. doi: 10.1139/g2012-049
- Hafez, M., Guha, T.K., Hausner, G. 2014a. I-OmiI and I-OmiII: two intron-encoded homing endonucleases within the *Ophiostoma minus* *rns* gene. *Fungal Biol.* **118(8)**: 721-731. doi: 10.1016/j.funbio.2014.05.002

- Hafez, M., Guha, T.K., Shen, C., Sethuraman, J., Hausner, G. 2014b. PCR-based bioprospecting for homing endonucleases in fungal mitochondria rRNA genes. **1123**: 37-53. doi: 10.1007/978-1-62703-968-0_3
- Hafez, M., Majer, A., Sethuraman, J., Rudski, S.M., Michel, F., Hausner, G. 2013. The mtDNA rns gene landscape in the Ophiostomatales and other fungal taxa: twintrons, introns, and intron-encoded proteins. *Fungal Genet. Biol.* **53**: 71-83. doi: 10.1016/j.fgb.2013.01.005
- Haugen, R., Bhattacharya, D. 2004. The spread of LAGLIDAG homing endonuclease gene in rDNA. *Nucleic Acids Res.* **32(6)**: 2049-2057. doi: 10.1093/nar/gkh520
- Haugen, P., Simon, D.M., Bhattacharya, D. 2005. The natural history of group I introns. *Trends Genet.* **21(2)**: 111-119. doi: 10.1016/j.tig.2004.12.007
- Hausner, G. 2012. Introns, mobile elements and plasmids. In: Bullerwell, CE (eds), *Organelle Genetics: Evolution of organelle genomes and gene expression*. Springer Verlag, Berlin. pp. 329-358.
- Jacoby, K., Metzger, M., Shen, B.W., Certo, M.T., Jarjour, T., Stoddard, B.L., Scharenberg, A.M. 2012. Expanding LAGLIDAG endonuclease scaffold diversity by rapidly surveying evolutionary sequence space. *Nucleic Acids Res.* **40(11)**: 4954-4964. doi: 10.1093/nar/gkr1303
- Johansen, S., Haugen, P. 2001. A new nomenclature of group I introns in ribosomal DNA. *RNA* **7(7)**: 935-936.
- Jones DT, Taylor WR, Thornton JM, 1992. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8(3)**: 275-282.

- Katoh, K., Standley, D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30(4)**: 772-780. doi: 10.1093/molbev/mst010
- Kumar, S., Stecher, G., Tamura, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33(7)**: 1870-1874. doi: 10.1093/molbev/msw054
- Lambert, A.R., Hallinan, J.P., Shen, B.W., Chik, J.K., Bolduc, J.M., Kulshina, N., Robins, L.I., Kaiser, B.K., Jarjour, J., Havens, K., Scharenberg, A.M., Stoddard, B.L. 2016. Indirect DNA sequence recognition and its impact on nuclease cleavage activity. *Structure* **24(6)**: 862-873. doi: 10.1016/j.str.2016.03.024
- Lucas, P., Otis, C., Mercier, J.P., Turmel, M., Lemieux, C. 2001. Rapid evolution of the DNA-binding site in LAGLIDAG homing endonuclease. *Nucleic Acids Res.* **29(4)**: 960-969.
- Marcaida, M.J., Muñoz, I.G., Blanco, F.J., Prieto, J., Montoya, G. 2010. Homing endonucleases: from basics to therapeutic applications. *Cell. Mol. Life Sci.* **67(5)**: 727-748. doi: 10.1007/s00018-009-0188-y
- Molina, R., Redondo, P., Stella, S., Marenchino, M., D'Abramo, M., Gervasio, F.L., Epinat, J.C., Valton, J., Grizot, S., Duchateau, P., Prieto, J., Montoya, G. 2012. Non-specific protein-DNA interactions control I-CreI target binding and cleavage. *Nucleic Acids Res.* **40(14)**: 6936-6945. doi: 10.1093/nar/gks320
- Monteiro-Vitorello, C.B., Hausner, G., Searles, D.B., Gibb, E.A., Fulbright, D.W., Bertrand, H. 2009. The *Cryphonectria parasitica* mitochondrial *rns* gene: plasmid-like elements, introns and homing endonucleases. *Fungal Genet. Biol.* **46(11)**: 837-848. doi: 10.1016/j.fgb.2009.07.005

- Mueller, J.E., Clyman, J., Huang, Y.J., Parker, M.M., Belfort, M. 1996a. Intron mobility in phage T4 occurs in the context of recombination-dependent DNA replication by way of multiple pathways. *Genes Dev.* **10(3)**: 351-364.
- Mueller, J.E., Smith, D., Belfort, M. 1996b. Exon coconversion biases accompanying intron homing: battle of the nucleases. *Genes Dev.* **10(17)**: 2158-2166.
- Mullineux, S.T., Costa, M., Bassi, G.S., Michel, F., Hausner, G., 2010. A group II intron encodes a functional LAGLIDADG homing endonuclease and self-splices under moderate temperature and ionic conditions. *RNA* **16(9)**: 1818-1831. doi: 10.1261/rna.2184010
- Muñoz, N.M., Beard, B.C., Ryu, B.Y., Luche, R.M., Trobridge, G.D., Rawlings, D.J., Scharenberg, A.M., Kiem, H.P. 2012. Novel reporter systems for facile evaluation of I-Sce I-mediated genome editing. *Nucleic Acids Res.* **40(2)**: e14. doi: 10.1093/nar/gkr897
- Nicholas, K.B., Nicholas, H.B., Deerfield, D.W. 1997. GeneDoc: analysis and visualization of genetic variation. *EMB News* **4**: 14.
- Parker, M.M., Belisle, M., Belfort, M. 1999. Intron homing with limited exon homology. Illegitimate double-strand-break repair in intron acquisition by phage T4. *Genetics* **153(4)**: 1513-1523.
- Prieto, J., Molina, R., Montoya, G. 2012. Molecular scissors for *in situ* cellular repair. *Crit. Rev. Biochem. Mol. Biol.* **47(3)**: 207-221. doi: 10.3109/10409238.2011.652358
- Romano Ibarra, G.S., Paul, B., Sather, B.D., Younan, P.M., Sommer, K., Kowalski, J.P., Hale, M., Stoddard, B., Jarjour, J., Astrakhan, A., Kiem, H.P., Rawlings, D.J. 2016. Efficient modification of the CCR5 locus in primary human T cells with megaTAL nuclease establishes HIV-1 resistance. *Mol. Ther. Nucleic Acids* **5(8)**: e352. doi: 10.1038/mtna.2016.56

- Sambrook, J., Fritsch, E.F., Maniatis, T. 1989. Molecular cloning: A laboratory manual 2nd Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press. pp. 1626
- Saitou, N., Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4(4)**: 406-425.
- Sethuraman, J., Majer, A., Friedrich, N.C., Edgell, D.R., Hausner, G. 2009. Genes within genes: Multiple LAGLIDADG homing endonucleases target the ribosomal protein S3 gene encoded within an *rnl* group I intron of *Ophiostoma* and related taxa. *Mol. Biol. Evol.* **26(10)**: 2299-2315. doi: 10.1093/molbev/msp145
- Scalley-Kim, M., McConnell-Smith, A., Stoddard, B.L. 2007. Coevolution of a homing endonuclease specificity and its host target sequence. *J. Mol. Biol.* **372(5)**: 1305-1319. doi: 10.1016/j.jmb.2007.07.052
- Silva, G.H., Belfort, M. 2004. Analysis of the LAGLIDADG interface of the monomeric homing endonuclease I-DmoI. *Nucleic Acids Res.* **32(10)**: 3156-3168. doi: 10.1093/nar/gkh618
- Simossis, V.A., Heringa, J. 2005. PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res.* **33**(Web Server issue): W289-294. doi: 10.1093/nar/gki390
- Skowronek, K.J., Bujnicki, J.M. 2007. Restriction and homing endonucleases. In: Polaina J, MacCabe AP (eds), *Industrial enzymes. Structure, function and applications*. Springer-Verlag, pp. 357-378.
- Stoddard, B.L. 2005. Homing endonuclease structure and function. *Q. Rev. Biophys.* **38(1)**: 49-95. doi: 10.1017/S0033583505004063
- Stoddard, B.L. 2011. Homing endonucleases: from microbial genetic invaders to reagents for targeted DNA modification. *Structure* **19(1)**: 7-15. doi: 10.1016/j.str.2010.12.003

- Stoddard, B.L. 2014. Homing endonucleases from mobile group I introns: discovery to genome engineering. *Mob. DNA* **5(1)**:7. doi: 10.1186/1759-8753-5-7
- Szczepanek, T., Lazowska, J. 1996. Replacement of two non-adjacent amino acids in the *S. cerevisiae* bi2 intron-encoded RNA maturase is sufficient to gain a homing-endonuclease activity. *EMBO* **15(14)**: 3758-3767.
- Takeuchi, R., Choi, M., Stoddard, B.L. 2014. Redesign of extensive protein-DNA interfaces of meganucleases using iterative cycles of in vitro compartmentalization. *Proc. Natl. Acad. Sci. USA* **111(11)**: 4061-4066. doi: 10.1073/pnas.1321030111
- Takeuchi, R., Lambert, A.R., Mak, A.N., Jacoby, K., Dickson, R.J., Gloor, G.B., Scharenberg, A.M., Edgell, D.R., Stoddard, B.L. 2011. Tapping natural reservoirs of homing endonucleases for targeted gene modification. *Proc. Natl. Acad. Sci. USA* **108(32)**: 13077-13082. doi: 10.1073/pnas.1107719108
- Tocchini-Valentini, G.D., Fruscoloni, P., Tocchini-Valentini, G.P. 2011. Evolution of introns in the archeal world. *Proc. Natl. Acad. Sci. USA* **108(12)**: 4782-4787. doi: 10.1073/pnas.1100862108
- Toor, N., Zimmerly, S. 2002. Identification of a family of group II introns encoding LAGLIDADG ORFs typical of group I introns. *RNA* **8(11)**: 1373-1377.
- Wolfs, J.M., DaSilva, M., Meister, S.E., Wang, X., Schild-Poulter, C., Edgell, D.R. 2014. MegaTevs: single-chain dual nucleases for efficient gene disruption. *Nucleic Acids Res.* **42(13)**: 8816-8829. doi: 10.1093/nar/gku573
- Wolfs, J.M., Hamilton, T.A., Lant, J.T., Laforet, M., Zhang, J., Salemi, L.M., Gloor, G.B., Schild-Poulter, C., Edgell, D.R. 2016. Biasing genome-editing events toward precise

length deletions with an RNA-guided TevCas9 dual nuclease. *Proc Natl Acad Sci USA*

113(52): 14988-14993. doi: 10.1073/pnas.1616343114

Yamasaki, S., Terada, T., Kono, H., Shimizu, K., Sarai, A. 2012. A new method for evaluating the specificity of indirect readout in protein-DNA recognition. *Nucleic Acids Res.*

40(17): e129. doi: 10.1093/nar/gks462

Fig. 1. Phylogenetic tree inferred using the Neighbor-Joining method: The optimal tree with the sum of branch length = 26.11071834 is shown. The percentages at the nodes are based on bootstrap analysis (1000 replicates). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the JTT matrix-based method (Jones *et al.* 1992) and are in the units of the number of amino acid substitutions per site. The analysis involved 40 amino acid sequences and ambiguous alignment positions were removed for each sequence pair. There were a total of 781 positions in the final dataset. The node marked with black octagon supports the monophyly of the mS917 clade of homing endonucleases. The black circles indicate the position of those members that were tested in this study for activity. The node marked by the black square supports the monophyly of the I-OnuI family of homing endonucleases.

Fig. 2. Gel images for the *in vitro* endonuclease assays: (a) I-CcaI was incubated with the circular *prns*-SUB plasmid at 37°C for 30 min. (b) I-CcaII was incubated with the circular *rml*-SUB plasmid for 30 min. (c) I-AstI was incubated with the circular *cox3*-SUB plasmid at room

temperature overnight. Lanes denoted with “L” contain the 1 kb plus DNA ladder (Thermo Fisher Scientific); Lanes denoted with “C” contain the untreated plasmid substrate.

Fig. 3. Cleavage site mapping results: Schematic for the *rns* (a), *rnl* (b), *cox3* (c) substrates and the cleavage sites for I-CcaI, I-CcaII, and I-AstI respective. The staggered cuts and the four central recognition motifs for the HEs are indicated. The four nucleotide 3' overhangs generated by each of the enzyme and removed by T4 DNA polymerase are shown in red. The intron insertion sites were designated based on reference sequences, for the *rns* and *rnl* genes the insertion sites are based on *E. coli* rDNA (AB035922.1) as proposed by Johansen and Haugen (2001). The *cox3* intron insertion site is based on comparison with the *S. cerevisiae* *cox3* sequence (KP263414.1).

Fig. 4. Effect of temperature on I-CcaI endonuclease activity: The plasmid substrate was incubated with I-CcaI at different temperatures (range from 20°C to 50°C). The “C” (control) lane contains the untreated substrate plasmid, while the lanes denoted with “L” contain the 1 kb plus DNA ladder (Thermo Fisher Scientific). Arrow indicates the position of the linearized plasmids.

Fig. 5. Gel images for the I-CcaI cleavage assays with various substrates (see supplementary Fig. 1). I-CcaI was incubated for 30 min with nine different plasmids (control *prns*-SUB and

prns-sub2 to *prns*-sub9) containing various length versions of the target region for I-CcaI. Lanes denoted with “L” contain the 1 kb plus DNA ladder (Invitrogen); Lanes denoted with “U” contain the untreated (no HE) corresponding plasmid substrate (*prns*-sub2 to *prns*-sub9); Lanes denoted with “EcoRI” contain positive controls where the substrates were incubated with the EcoRI restriction enzyme in order to show the migration of the cleaved substrate plasmids. Panel (a) shows the results for substrates that failed to be cut by I-CcaI and panel (b) shows the results for those substrates that were cleaved by I-CcaI. Substrate plasmids that are untreated (U) or have been treated with EcoRI are the negative and positive controls, respectively. Panel (c) below the gels indicates how many nucleotides upstream and downstream (upstream/downstream) of the cleavage site (indicted by arrow) are present within the corresponding substrate plasmid. Uncut plasmids appear behave as supercoiled and relaxed plasmids thus their migration patterns on the gel are quite variable, migrating faster or slower in comparison to linearized versions. Note: c = position of cut plasmids, u.r. notes the position of uncut relaxed plasmids, and u.s. denotes the position of uncut supercoiled plasmids.

Fig. S1. Substrates based on modifications on the *prns*-SUB plasmid: The “x” designates the central four-nucleotide recognition motif. The insertion of triplets (AAA, GGG, or CCC) defines various length versions of the potential recognition sequence for I-CcaI.

Fig. S2. Protein purification: (a) SDS–PAGE (12.5 %) gel of I-CcaI and I-CcaII purification by Ni-NTA resin (Qiagen). Lanes are denoted as follows: M = molecular weight markers that contain the BLUeye prestained protein ladder (GeneDireX; FroggaBio, Toronto, Ontario). The

column was first washed with low concentrations of imidazole (25, 50 and 100 mM). The displayed bands represent I-CcaI (49.5 kDa) and I-CcaII (47 kDa) proteins respectively when eluted with 250 mM imidazole. Note that the I-CcaII protein could only be partially purified.

(b) SDS–PAGE (12.5 %) gel for the I-AstI purification by Ni-NTA resin (Qiagen). Lanes are marked as follows: M = molecular weight markers; FT = flow through; W = wash and E = elution by (250 mM imidazole). Lane L contains the BLUeye prestained protein ladder (GeneDireX; FroggaBio).

Fig. S3. An alignment for the central four recognition motifs (CRMs in red) for members of the I-OnuI family of homing endonucleases (Lambert *et al.* 2016). A sequence logo (Crooks *et al.* 2004) shown below the alignment to indicated sequence conservation among the various CRMs. See text for more details.

Fig. 1. Phylogenetic tree inferred using the Neighbor-Joining method: The optimal tree with the sum of branch length = 26.11071834 is shown. The percentages at the nodes are based on bootstrap analysis (1000 replicates). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the JTT matrix-based method (Jones et al. 1992) and are in the units of the number of amino acid substitutions per site. The analysis involved 40 amino acid sequences and ambiguous alignment positions were removed for each sequence pair. There were a total of 781 positions in the final dataset. The node marked with black octagon supports the monophyly of the mS917 clade of homing endonucleases. The black circles indicate the position of those members that were tested in this study for activity. The node marked by the black square supports the monophyly of the I-OnuI family of homing endonucleases.

204x223mm (300 x 300 DPI)

Fig. 2. Gel images for the in vitro endonuclease assays: (a) I-CcaI was incubated with the circular *prns*-SUB plasmid at 37°C for 30 min. (b) I-CcaII was incubated with the circular *prnI*-SUB plasmid for 30 min. (c) I-AstI was incubated with the circular *pcox3*-SUB plasmid at room temperature overnight. Lanes denoted with "L" contain the 1 kb plus DNA ladder (Thermo Fisher Scientific); Lanes denoted with "C" contain the untreated plasmid substrate.

140x94mm (300 x 300 DPI)

Fig. 3. Cleavage site mapping results: Schematic for the *rns* (a), *rnl* (b), *cox3* (c) substrates and the cleavage sites for I-CcaI, I-CcaII, and I-AstI respective. The staggered cuts and the four central recognition motifs for the HEs are indicated. The four nucleotide 3' overhangs generated by each of the enzyme and removed by T4 DNA polymerase are shown in red. The intron insertion sites were designated based on reference sequences, for the *rns* and *rnl* genes the insertion sites are based on *E. coli* rDNA (AB035922.1) as proposed by Johansen and Haugen (2001). The *cox3* intron insertion site is based on comparison with the *S. cerevisiae* *cox3* sequence (KP263414.1).

178x143mm (300 x 300 DPI)

Fig. 4. Effect of temperature on I-CcaI endonuclease activity: The plasmid substrate was incubated with I-CcaI at different temperatures (range from 20°C to 50°C). The "C" (control) lane contains the untreated substrate plasmid, while the lanes denoted with "L" contain the 1 kb plus DNA ladder (Thermo Fisher Scientific). Arrow indicates the position of the linearized plasmids.

76x42mm (300 x 300 DPI)

Fig. 5. Gel images for the I-CcaI cleavage assays with various substrates (see supplementary Fig. 1). I-CcaI was incubated for 30 min with nine different plasmids (control *prns-SUB* and *prns-sub2* to *prns-sub9*) containing various length versions of the target region for I-CcaI. Lanes denoted with "L" contain the 1 kb plus DNA ladder (Invitrogen); Lanes denoted with "U" contain the untreated (no HE) corresponding plasmid substrate (*prns-sub2* to *prns-sub9*); Lanes denoted with "EcoRI" contain positive controls where the substrates were incubated with the EcoRI restriction enzyme in order to show the migration of the cleaved substrate plasmids. Panel (a) shows the results for substrates that failed to be cut by I-CcaI and panel (b) shows the results for those substrates that were cleaved by I-CcaI. Substrate plasmids that are untreated (U) or have been treated with EcoRI are the negative and positive controls, respectively. Panel (c) below the gels indicates how many nucleotides upstream and downstream (upstream/downstream) of the cleavage site (indicted by arrow) are present within the corresponding substrate plasmid. Uncut plasmids appear behave as supercoiled and relaxed plasmids thus their migration patterns on the gel are quite variable, migrating faster or slower in comparison to linearized versions. Note: c = position of cut plasmids, u.r. notes the position of uncut relaxed plasmids, and u.s. denotes the position of uncut supercoiled plasmids.

79x24mm (300 x 300 DPI)

Draft

CRM sequence

-2 -1 +1 +2

I-CcaI	A	G	G	A	A	C	T	G	T	A	A	T	C	A	C
I-CcaII	C	A	G	T	G	C	A	G	A	T	<u>G</u>	C	T	G	C
I-AstI	T	T	T	A	C	T	T	A	A	T	A	C	A	G	T
I-AabMI	G	G	T	A	C	C	C	T	T	T	A	A	A	C	C
I-CpaMI	A	G	C	C	C	A	C	A	A	T	A	T	T	A	A
I-GpeMI	T	T	C	C	G	C	T	T	A	T	T	C	A	A	C
I-GzeII	T	G	G	G	T	A	C	C	A	T	A	T	T	G	G
I-LtrI	A	T	G	C	T	C	C	T	A	T	A	C	G	A	C
I-LtrWI	A	G	T	A	G	T	G	A	A	G	T	A	T	G	T
I-OnuI	T	T	T	C	C	A	C	T	T	A	T	T	C	A	A
I-PanMI	G	C	T	C	C	T	C	A	T	A	A	T	C	C	T
I-SmaMI	A	T	C	C	T	C	C	A	T	T	A	T	C	A	G