Adaptation of intronic homing endonuclease for successful horizontal transmission

Sayuri Kurokawa[1], Yoshitaka Bessho[2], Kyoko Higashijima[2], Mikako Shirouzu[2],
Shigeyuki Yokoyama[2,3,4], Kazuo I. Watanabe[5], and Takeshi Ohama[1]

[1]Graduate School of Engineering, Department of Environmental Systems Engineering,
Kochi University of Technology (KUT), Tosayamada 185, Kochi 782-8502, Japan

[2] RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan

[3] RIKEN Harima Institute at SPring-8, 1-1-1 Kouto, Miazuki-cho, Sayo, Hyogo 679-5148, Japan

[4] Graduate School of Science, University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-0033, Japan

[5]Institute for Cellular and Molecular Biology, The University of Texas at Austin,
Moffett Molecular Biology Bldg, A4800, 2500 Speedway, Austin, TX 78712-1095, USA

Correspondence to T. Ohama, Department of Environmental Systems Engineering,
Kochi University of Technology (KUT), Tosayamada, Kochi 782-8502, Japan.
FAX: +81-887-57-2520,   Tel: + 81-887-57-2512,
E-mail: ohama.takeshi@kochi-tech.ac.jp

Running title: Recognition ambiguity of intronic homing endonucleases

Subdivision: Molecular genetics

Keywords: I-*Csm*I; *Chlamydomonas smithii*; horizontal transmission; group I intron; selfish element

Abbreviations: *cob*, apocytochrome b; a.a, amino acid residue(s); nt, nucleotide(s); ORF, open reading frame; PAGE, polyacrylamide gel electrophoresis; WT, wild-type

Summary

Group I introns are thought to be self-propagating mobile elements, and are distributed over a wide range of organisms through horizontal transmission. Intron invasion is initiated through cleavage of a target DNA by a homing endonuclease encoded in an ORF found within the intron. The intron is likely of no benefit to the host cell and is not maintained over time, leading to the accumulation of mutations after intron invasion. Therefore, regular invasional transmission of the intron to a new species at least once before its degeneration is likely essential for its evolutionary long-term existence. In many cases, the target is in a protein coding region which is well conserved among organisms, but contains ambiguity at the 3rd nucleotide position of the codon. Consequently, the homing endonuclease might be adapted to overcome sequence polymorphisms at the target site.

To address whether codon degeneracy affects horizontal transmission, we investigated the recognition properties of a homing enzyme, I-*Csm*I, that is encoded in the intronic ORF of a group I intron located in the mitochondrial COB gene of the unicellular green alga *Chlamydomonas smithii*. We successfully expressed and purified three types of N-terminally truncated I-*Csm*I polypeptides, and assayed the efficiency of cleavage for 48 substrates containing single nucleotide substitutions. We found a slight but significant tendency that I-*Csm*I cleaves substrates containing a silent amino acid change more efficiently than non-silent and non-synonymous ones. The published recognition properties of I-*Spom*I, I-*Sca*I, and I-*Sce*II were reconsidered from this point of view, and we detected proficient adaptation of I-*Spom*I and I-*Sca*I for target site sequence degeneracy. Based on the results described above, we propose that intronic homing enzymes are adapted to cleave sequences that might appear at the target region in various species, however such adaptation becomes less prominent in proportion to the time elapsed after intron invasion into a new host.

Introduction

Various molecular phylogenetic analyses suggest that group I introns in fungi and terrestrial/non-aquatic plants were horizontally transmitted multiple times in the course of evolution among distantly related species [1, 2, 3]. We have shown this is also the case for algal mitochondrial introns [4, 5]. For reasons yet unknown, the distribution of group I introns is strongly biased, most commonly found in fungi [*e.g.*, the *cox*1 of *Podospora anserina* contains 15 group I introns [6]]. About half of group I introns contain an ORF that encodes a DNA sequence specific endonuclease (intronic homing enzyme). These intronic homing enzymes cleave a target sequence that is usually 16-30 base pairs (bp) long and non-palindromic (for a review, [7]). Cleavage of the chromosome initiates repair of the damaged DNA through homologous recombination. Consequently, after the repair, the donor intronic DNA is copied into the recipient chromosome. Thus, homing endonucleases are essential for horizontal transmission of group I introns. Organellar introns are highly likely of no benefit to the host, *i.e.*, they are thought to be selfish and parasitic elements that spread in populations. Therefore, when they integrate into the host genome, there is little or no selection for maintaining endonuclease function. Moreover, if there is any cost to the host cell for producing a functional endonuclease, then selection will work to fix the nonfunctional element. Therefore, regular horizontal transmission of an intron to a new species before its functional deterioration seems essential for its evolutionary long-term persistence. As an example, comprehensive analyses of the group I intron omega (also known as Sc LSU.1), which was first found in the *Saccharomyces cerevisiae* mitochondrial large subunit rRNA gene, clearly showed repeated horizontal omega transmissions, and the interval between the complete loss and reinvasion of the intron is estimated to be about 5.7 million years [8]. This leads to the hypothesis that intronic homing enzymes might be adapted to recognize variously degenerated target sequences among a wide range of organisms.

In addition to intronic homing enzymes, highly specific endonuclease activity is also detected among inteins, which are thought to be parasitic elements that exhibit horizontal transmission. Regular invasional transmission is likely essential for both homing introns and inteins. In fact, for the target site of intein homing endonuclease PI-*Sce*I, which is found in *Saccharomyces cerevisiae* vacuolar membrane H^+-ATPase, all of the nine nucleotides essential for the cleavage were mapped on the conserved codon first and second positions, and target sequence variations at codon third positions were tolerated for the endonuclease recognition [9]. On the other hand, the adaptations that permit efficient horizontal transfer of intronic homing enzymes have not been analyzed. To date, only three intronic homing enzymes that target a sequence within protein coding genes were investigated systematically for their recognition sequence ambiguity, *i.e.*, I-*Spom*I that is encoded as an intronic ORF of a group I intron in the *Schizosaccharomyces pombe* COXI gene [10, 11], I-*Sca*I is in the COB gene of *Saccharomyces capensis* [12, 13], and I-*Sce*II is in the COXI gene of *Saccharomyces cerevisiae* [14, 15, 16]. To address the question, we investigated the recognition sequence of I-*Csm*I, including its degeneracy. I-*Csm*I is a homing enzyme encoded in the group I intron (named alpha or *Cs cob*.1) located in the apocytochrome b (COB) gene of the unicellular alga *Chlamydomonas smithii* (*C. smithii*) [17]. This enzyme has the typical two LAGLIDADG motifs. The intronic ORF is probably translated as a fusion protein with the preceding exon, and the N-terminal peptide may be proteolytically removed to become an active form as seen in I-*Spom*I [18]. Endonuclease activity of I-*Csm*I has been observed through artificial interspecific cell fusion between intron-bearing *C. smithii* and *C. reinhardtii* that lacks the intron in its COB gene [19]. However, systematic analysis of the target sequences and the homing endonuclease's enzymatic properties have not been previously attempted. We overproduced several N-terminally truncated I-*Csm*I polypeptides in *E. coli*, and determined cleavable target sequences through an *in vitro* assay of substrates containing 72 different point mutations.

Based on the analyses of I-*Csm*I and these three intronic homing enzymes, we discuss the adaptation for successful horizontal transfer.   Investigations performed for the intronic homing enzymes that have a recognition sequence in ribosomal RNA genes are less informative to answer our questions and are not considered in this paper.

Results

Activity of the N-terminal truncated I-*Csm*I polypeptides

Three N-terminally truncated I-*Csm*I polypeptides [I-*Csm*I(200), I-*Csm*I(217), and I-*Csm*I(237)] were purified and yielded about 6 mg of protein per 1 g wet weight *E. coli*, while the entire I-*Csm*I ORF(*i.e*., I-*Csm*I(374)), which contains the upstream COB exon, did not express even after several modified conditions were tested.   We assayed the endonuclease activity of recombinant I-*Csm*I(200), I-*Csm*I(217), and I-*Csm*I(237) using linearized pCOB1.8Kb as a substrate.   I-*Csm*I(200) is the smallest homing endonuclease containing two LAGLIDADG motifs analyzed to date.   It is even smaller than the type II restriction enzyme *Eco*RI [20], which is a 277 a.a. homodimer that cleaves a symmetric 6 base restriction site.   Recombinant proteins I-*Csm*I(200) and I-*CsmI*(237) cleaved the substrate at the expected target site, yielding two fragments of 1.2 Kb and 3.7 Kb in size.   For I-*Csm*I(217), the quantity of protein was reduced from 1.5 to 1.0 μg and the incubation period was shortened from 24 to 6 hrs to reduce the amount of insoluble reaction products.   Under these modified conditions, I-*Csm*I(217) exhibited sequence specific endonuclease activity.

To determine the optimal conditions for endonuclease activity, we tested the effect of Na^+ and Mg^{2+} concentration, pH, and temperature.   The optimal pH for all three proteins was around 7.0 (Fig. 2A).   The optimal Na^+ and Mg^{2+} concentrations were 25 mM and 5 mM, respectively, for both I-*Csm*I(237) and I-*Csm*I(217) (Fig. 2B, 2C).   In contrast, 75 mM Na^+ and 10 mM Mg^{2+} were optimal for I-*Csm*I(200).   The optimal reaction temperature was 35 ˚C for both I-*Csm*I(200) and I-*Csm*I(237), and 30 ˚C for I-CsmI(217) (Fig. 2D)   A higher concentration of Mg^{2+} was progressively

detrimental to all I-*Csm*I polypeptides.   The presence of Mg^{2+} was essential for the endonuclease activity as a cofactor, while the same concentration of Mn^{2+} (5 mM) reduced the enzyme activity to 15 %, and no activity was observed with 5 mM of Zn^{2+}, Ca^{2+}, or Co^{2+} (data not shown).

Kinetic parameters of I-*Csm*I**(200)**

We determined the kinetic parameters of I-*Csm*I(200) based on the data obtained by time course monitoring of the cleaved products in various concentrations of the linearized substrate pCOB1.8 Kb.   The Km, Vmax, kcat were 2.5×10^{-9} M, 1.8×10^{-12} M/s, and 4.7×10^{-4}/s, respectively.   These parameters were similar to other representative intronic LAGLIDADG homing endonucleases [*e.g.*, I-*Ceu*I [21],   I-*Sce*IV [22],   I-*Dmo*I [23, 24]] that show characteristics of high affinity to the substrate DNA and slow turnover (Table 1).

Core recognition region and cleavage point

Digestion was not observed using pC-18nt and pC-20nt, while almost complete digestion was observed for pC-24nt by I-*Csm*I(200).   This shows that the recognition region of I-*Csm*I(200) resides between 12 nt upstream (+) and 12 nt downstream (-) of the intron insertion site, while 10 nt upstream and 10 nt downstream is insufficient for recognition.

Cleavage point and mutational analysis of cleavable sequences

The precise cleavage site on each strand was determined through DNA sequencing of the substrate whose termini were blunt-ended by T4 DNA polymerase treatment.   It became clear that cleavage occurs 5 nt downstream of the intron insertion site on the coding strand and one nt downstream of the insertion site on the non-coding strand, creating 3'-overhangs of 4 nt (Fig. 3).   This terminal overhang is typical for DNA cleaved by LAGLIDADG homing enzymes.

72 variants (104 bp each) containing single nt substitutions between -12 and +12 were assayed to discern the critical nucleotides involved in recognition. Positions -5 through -3, +2, and +6 through +8 are strictly recognized by I-*Csm*I(200) and I-*Csm*I(217), as the original bases are essential for cleavage, while any substitution was permitted for positions -12 through -6 and +12 [some examples of cleaved pattern are shown in Fig. 4]. The majority of substitutions that blocked substrate cleavage were between -5 and +11 in relation to the intron insertion site. Therefore, the span of critical bases are not centered at the intron insertion site, but are spread almost symmetrically with respect to the cleavage points of coding and non-coding strands. A summary of substrate cleavability is classified into 4 groups (+++, ++, +, and -; see Materials and Methods for details) and shown in Fig. 3. As a result of cleavage with I-*Csm*I (200), 12 (6), 17 (19), 12 (10), and 31 (37) kinds of substrates were classified into four classes, +++, ++, +, and -, respectively [the results of I-*Csm*I(217) are shown in parentheses]. I-*Csm*I(200) and I-*Csm*I(217) showed almost identical sequence recognition properties (Fig. 3). A prominent difference in cleavage efficiency was observed for only two substitutions, the original G residue at position - 2 for A and T residues. I-*Csm*I(217) did not cleave these mutated substrates, whereas I-*Csm*I(200) cleaved both, with the G to A mutation the most efficient of the two (Fig. 3).

Correlation between the type of amino acid substitution and cleavage efficiency
We analyzed whether there is any correlation between the type of amino acid substitution induced by single nt substitution (silent, synonymous, or non-silent/non-synonymous change) and how efficiently the substrates are cleaved by two kinds of N-terminal truncated I-*Csm*I polypeptides. 48 substrates containing single nt substitutions at the core recognition region (between -5 and +11) were analyzed from this point of view.
(Substrates containing a silent amino acid change)
Seven of 48 substrates contained a silent amino acid change. However, two of seven such substrates [containing TCT(Ser) changed to TCA and TCG(Ser), mutation

position +9 in Fig. 3] were not cleaved at all by I-*Csm*I(217) and I-*Csm*I(200), and additionally the substrate containing the change GGC(Gly) to GGG(Gly) (position -1) was not cut by I-*Csm*I(217) even though these silent changes must be tolerated in nature.   On the other hand, three silent substrates [TCT(Ser) to TCC(Ser), position +9; GGC(Gly) to GGT/GGA(Gly), position -1] were cut efficiently by the two I-*Csm*I polypeptides.   Additionally, CAA (Gln) to CAG (Gln) (position +3) was efficiently cut by I-*Csm*I(200).

(Substrates containing a synonymous amino acid change)

Nine of the 48 substitutions correspond to synonymous amino acid changes.   Almost all of them (seven or eight of nine) were not cleaved or were weakly cleaved.   Only the change of ATG(Met) to GTG(Val) (position +4) was cleaved efficiently by I-*Csm*I(200) and I-*Csm*I(217).   In contrast, about half of the synonymous substitutions were cleaved efficiently or moderately by I-*Spom*I, I-*Sca*I, and I-*Sce*II.

(Substrates containing a non-silent and non-synonymous amino acid change)

32 of 48 substitutions caused non-silent/non-synonymous amino acid changes.   26 and 22 of them were weakly cleaved by I-*Csm*I(217) and I-*Csm*I(200), respectively. However, TAA(Stop) instead of CAA(Gln) (position +1), TGC(Cys) and TCC(Ser) instead of TTC(Phe) (position +11) were efficiently cleaved by the both enzymes, even though these codons are rare or not observed at these positions in nature.   Again, in contrast, none of the non-silent/non-synonymous substitutions were cleaved efficiently by I-*Spom*I and I-*Sca*I (Table 2).

Discussion

The original I-*Csm*I ORF is fused with the preceding exon, which is not rare for group I intronic ORFs.   The entire ORF of I-*Spom*I also extends into the upstream exon of the COXI gene, and it has been reported that the N-terminal truncated polypeptide, including the two LAGLIDADG motifs, has similar sequence specificity to that detected using mitochondrial extracts [11].   Considering the above, we tried to

overproduce three kinds of N-terminally truncated recombinant I-*Csm*I polypeptides that retain the two LAGLIDADG motifs instead of the entire I-*Csm*I (374 a.a.) (Fig. 1), because we failed to express the whole I-*Csm*I ORF for reasons that are unclear. We found that all of the N-terminal truncated I-*Csm*I polypeptides retain the specificity to cleave the target site, and the kinetic parameters of I-*Csm*I(200) are very similar to that reported for representative intronic homing enzymes of LAGLIDADG motifs (Table 1). The optimal conditions of selected factors were also very similar to other homing enzymes, with the exception of the preferred pH. I-*Csm*I displayed its highest activity at pH 7.0, which is very close to the reported physiological pH value of 7.5 in yeast mitochondria [25], while most of the LAGLIDADG enzymes show their highest activity at an alkaline pH between 8.5 and 9.5 [*e.g.*, optimal pH is 9.2 for I-*Ani*I [26], and between 8.5 and 9.0 for the recombinant I-*Sca*I [13]. Having a host pH that is lower than the optimum pH observed for many homing enzymes may act to reduce endonuclease activity and prevent overdigestion of the genomic DNA.

I-*Csm*I(200)'s optimal conditions for Na^+ and Mg^{2+} are clearly shifted to a concentration higher than that of I-*Csm*I(217) and I-*Csm*I (237) (Fig. 2B, 2C). This suggests that the three dimensional conformation of this enzyme is different from the others possibly because of the recessed N-terminal region, and may explain the differences in cleavage activity between I-*Csm*I(200) and I-*Csm*I(217). I-*Csm*I(200) seems to tolerate a higher degree of sequence ambiguity than I-*Csm*I(217) at position -2, because I-*Csm*I(200) can efficiently cleave the mutated substrates of -2A and -2T (instead of the original -2G), while I-*Csm*I(217) only tolerates the original base -2G (Fig. 3).

Cleavage of a target DNA is an essential step for lateral transfer of an intron. Therefore, if a homing enzyme shows very stringent recognition of the target core sequence, this step could be a bottleneck for horizontal transmission of an intron. The target site of I-*Csm*I corresponds to the amino acid sequence of Trp-Gly-Gln-Met-Ser-Phe (Fig. 3). This is a highly conserved region in COB genes among a wide range of organisms. Our systematic induction of a point mutation and the cleavage assay

9

showed a clear tendency that I-*Csm*I efficiently cleaves silent change containing substrates [42.9 % and 57.1 %, respectively for I-*Csm*I(217) and I-*Csm*I(200)] than non-synonymous/non-silent change containing ones (10.3 % and 15.6 %) (Table 2).

It is obvious that stop codons are never tolerated at the internal regions of a gene.   However, our systematic induction of a point mutation introduced stop codons, *i.e.*, TGA and TAG stop codons from TGG(Trp), and TAA stop codon from CAA(Gln). The substrate DNA that contains TGA or TAG was not cleaved, while the substrate containing a TAA stop codon was efficiently cleaved by the both I-*Csm*I polypeptides (Fig. 3).   Moreover, substrates including a codon that highly likely appears in nature were not cleaved [*e.g.*, TCA/TCG(Ser) from TCT(Ser), and three Ile codons AT(T/C/A) from ATG(Met)].   The above instances indicate that the recognition property of I-*Csm*I is not skillfully adapted to recognize target sequences that are highly likely to appear in nature.

It is possible that the recognition property of I-*Spom*I and I-*Sca*I are adapted to recognize multiple possible target sequences, because these homing enzymes cleaved various kinds of silent amino acid changes efficiently, while all of the non-silent/non-synonymous substitution-containing substrates were cleaved poorly.   Such a property is also shared but less prominently with I-*Sce*II.

Considering the above, we propose that homing enzymes are adapted to recognize diverse target sequences to facilitate horizontal transmission to a new species, as evidently seen with I-*Spom*I and I-*Sca*I.   However, immediately after a successful invasion, mutations begin to accumulate that lead to a loss of further adaptation, because homing endonuclease activity is only essential for intron invasion and thereafter it is useless to the cell.   Invasion of I-*Csm*I might be evolutionarily older than the other three homing enzymes compared in this study, because I-*Csm*I showed the least adapted properties among the four.   Actually, remnants of homing endonuclease ORFs are frequently found that include frame shifts or stop codons within the ORF (e.g., [4]).   Comprehensive analysis of omega homing endonuclease and its associated group I intron revealed that it is more common to find an inactive

intron/ORF combination than it is to find an active intron/ORF combination or an intron-less allele [8].

It has been proved that some of intronic homing enzymes are bifunctional. They work not only as an endonuclease but also as a maturase to preserve splicing. The bifunctional activity of I-*Spom*I [18], I-*Sca*I [12], and I-*Ani*I [27] has been observed.   I-*Csm*I could also be a bifunctional protein that acts as a maturase, which may also preserve its endonuclease activity for horizontal transmission.   These bifunctional enzymes are recognized as intermediates, and may likely lose their endonuclease activity over time, retaining only their maturase activity [4, 26].

Materials and Methods

Cloning and expression of WT and N-terminally truncated I-*Csm*I ORFs

The entire COB gene and the alpha intron were amplified by PCR using total *Chlamydomonas smithii* (CC-1373) DNA as a template.   We also used PCR to isolate the wild-type (WT) 374 amino acid (a.a) I-*Csm*I ORF (*i.e.*, ORF(374)) and three N-terminally truncated ORFs, ORF(200), ORF(217) and ORF(237) (the number in parentheses indicates the a.a encoded in the ORF).   These four ORFs have different N-termini, however share the common WT stop codon (Fig. 1).   The two sets of primers used to amplify the original I-*Csm*I ORF(374) and ORF(237), contained *Xho*I sites at their tails.   Forward primers containing an *Nde*I site, and reverse primers containing an *Fba*I site were used to amplify ORF(200) and ORF(217).   After restriction enzyme digestion, the ORF(374) and ORF(237) PCR products were cloned into the *Xho*I site of pET19b (Novagen, CA, USA) in frame with a sequence encoding the 10-histidine tag, while ORF(200) and ORF(217) were cloned into the *Nde*I/*Bam*HI site of pET15b (Novagen, CA, USA) in-frame with a 6-histidine tag.   The resulting plasmids were amplified in *E. coli* DH5/alpha, and *E. coli* BL21 CodonPlus (DE3) RIL (Stratagene, CA, USA) was for protein expression.

Expression and purification of whole or truncated I-*Csm*I polypeptides

Cultures containing whole or truncated ORFs were undertaken at 37 ˚C in 2.0 liters of LB broth containing 100 μg/ml ampicillin and 34 μg/ml chloramphenicol to an OD $_{600}$ nm= 0.6.　Protein expression was induced by addition of isopropyl-thio-ß-D-galactopyranoside (0.1 mM final).　The cells were incubated at 30 ˚C for an additional 4 hrs, collected by centrifugation, and resuspended in 40 ml of sonication buffer [50 mM HEPES (pH 7.0), 400 mM NaCl, 6 mM ß-mercaptoethanol, and 20 μg/ml lysozyme and sonicated on ice.　The lysate was centrifuged for 2 hrs at 10,000 g and the supernatant was loaded onto a Ni-NTA column (5 ml bed volume) (Qiagen, CA, USA) that was previously equilibrated with the wash buffer [50 mM HEPES (pH 7.0), 400 mM NaCl, 6 mM ß-mercaptoethanol, and 10 mM imidazole].　The column was washed with 50 ml of the wash buffer, and the protein was eluted with 100 ml of the elution buffer [50 mM HEPES (pH 7.0), 400 mM NaCl, 6 mM ß-mercaptoethanol, and 200 mM imidazole].　Homogeneity was assessed after staining with SDS-PAGE/comassie brilliant blue R-250.　The products of ORF(200), ORF(217), ORF(237), and ORF(374) were named I-*Csm*I(200), I-*Csm*I(217), I-*Csm*I(237) and I-*Csm*I(374), respectively.

Reaction conditions to estimate the minimum target-site length (Substrate DNA)

Chemically synthesized DNA fragments, which consist of 18, 20, or 24 nt symmetrically spanning the alpha intron insertion point of the *C. reinhardtii* COB gene, were cloned into the *Eco*RV site of the pCITE-4a+ (Novagen, CA, USA).　These plasmids were named pC-18nt, pC-20nt, and pC-24nt (the number indicates the length of the inserted DNA fragment).　The plasmids were first linearized by *Sca*I digestion, and then used as a substrate to determine the region encompassing the recognition sequence.

(Reaction conditions)

1.5 µg of linearized substrate described above was added to 50 µl of the reaction mixture containing [50 mM HEPES (pH 7.0), 0.01% bovine serum albumin (BSA), 1 mM dithiothreitol (DTT), 25 mM NaCl, and 5 mM MgCl$_2$] and about 1 µg of recombinant homing enzyme I-CsmI(237). The reaction was carried out at 25 ˚C for 24 hrs and 10 µl was loaded onto an 0.8 % agarose gel to resolve the products.

Reaction to determine the cleavage point and its terminal shape

We determined the terminal shape of the substrate following the T4 DNA polymerase method by Nishioka $et\ al.$ [28]. pC-24nt (2.0 µg) digested with I-CsmI(237) was recovered from an 0.8% agarose gel by electro-elution and then treated with T4 DNA polymerase (Takara Bio, Kyoto, Japan) in the presence of 0.2 mM dNTPs. The DNA mixture was then treated with T4 DNA ligase (Takara Bio, Kyoto, Japan) for self-ligation and transformed into $E.\ coli$. Nucleotide sequence analysis of the plasmid was performed to determine the nature of cohesive termini generated by I-CsmI.

Reaction conditions used to investigate the effect of Na$^+$, divalent cations, pH, and temperature

(Substrate DNA fragment)

A 1.8 Kb DNA fragment, containing the entire COB gene of $C.\ reinhardtii$ (CC-124) and its flanking regions, was cloned into pT7-Blue2 vector (Novagen, CA, USA) and named pCOB1.8Kb. After linearization by NotI, the plasmid was used as a substrate for the reaction described below.

(Reaction mixture)

A 50 µl reaction mixture [25 mM NaCl, 5 mM MgCl$_2$, 1 mM DTT, 0.01 % BSA, 50 mM Tris-HCl (pH 7.0)] was used, which contained 0.5 µg of linearized pCOB1.8Kb and 1.0 µg of I-CsmI(217), or 1.5 µg of I-CsmI(200) or I-CsmI(237). One of the parameters [$i.e.$, pH, NaCl concentration, species of divalent cations (5 mM), MgCl$_2$ concentration, or the temperature] in the reaction was altered to determine optimal conditions. Reagents used to make the buffers of specific pH value are as follows;

MES for pH 6.0, HEPES for pH 7.0, Tris for pH 8.0 and 9.0, TAPS for pH 10.0.   The reaction was incubated for 24 hrs with I-*Csm*I(237) and I-*Csm*I(200), and incubated for 6 hrs with I-*Csm*I(217), which reduced the formation of aggregates observed with this protein.   The reaction products were resolved in an 0.8 % agarose gel, and stained with ethidium bromide.   The relative quantities of the digested fragments were calculated using the NIH Image program version 1.61.

Assay of cleavable DNA sequences

A limited part of the *C. reinhardtii* COB gene, which is 104 nt long and containing the I-*Csm*I target sequence, was chemically synthesized and converted to double strand DNA.   This double-stranded DNA fragment was used as a control to compare the cleavage efficiency of various substrates containing single mutations.   Each one of the 24 nucleotides composing the target site was changed to the other three possible nucleotides utilizing PCR primers containing a specific mutation.   These 72 DNA fragments, each containing single point mutations were used for a detailed analysis of substrate cleavage.   150 ng of each substrate was digested with 1 μg of I-*Csm*I(200) in the reaction mixture [50 mM HEPES (pH 7.0), 0.01 % BSA, 1 mM DTT, 25 mM NaCl, 5 mM MgCl$_2$] at 30 ˚C for 8 hrs.   Electrophoresis of the samples was performed on a 3 % agarose gel, and stained by 10,000-fold diluted SYBR Green I dye (Molecular Probes, OR, USA) for 40 min.   The image was developed using LAS-1000 image analyzer (Fuji Film Co., Tokyo, Japan).   The cleavage ratio, *i.e.*, cleaved *vs.* uncleaved fragments, was quantified by NIH Image and compared to WT substrate cleavage (*i.e.*, native *C. reinhardtii cob* sequence carrying substrate).   The 72 substrates were grouped into 4 classes based on the following: i) The substrate is much better than the control (the cleavage ratio of mutated substrate *vs.* control is more than 1.5) is denoted as +++; ii) The substrate is as good as the control, *i.e.*, the ratio is between 1.2 and 0.8 and is denoted as ++; iii) The substrate is less efficiently cleaved, *i.e.*, the ratio is between 0.5 and 0.2, denoted as +; iv) almost no cleaved substrate, *i.e.*, the ratio is below 0.1 is denoted as -.

Reaction conditions to measure the kinetic parameters

Linearized pCOB1.8Kb and a plasmid containing the N-terminally truncated homing endonuclease, I-*Csm*I(200), was used to measure the kinetic parameters.　250 μl of reaction buffer [50 mM HEPES (pH 7.0), 0.01% BSA, 1 mM DTT, 25 mM NaCl, and 5 mM MgCl$_2$] contained 1 μg of the recombinant protein and between 0.5 ng/μl and 10 ng/μl of substrate.　20 μl aliquots were removed at different time points from the reaction mixture, and terminated by the addition of 1 μl of 0.5 M EDTA and 1.25 μl of 10 % SDS, followed by heating the mixture to 50 ˚C for 5 min to completely denature the protein.　Samples were electrophoresed on an 0.8 % agarose gel, then visualized by 10,000-fold diluted SYBR Green I dye.　Relative intensities of the digested fragment were quantified using the Las-1000 and NIH Image.　Km, Vmax, and kcat were determined through a Hanes-Woolf's plot [29].

References

1   Lambowitz AM (1989) Infectious introns. *Cell* **56**, 323-326.

2   Belfort M & Roberts RJ (1997) Homing endonucleases: keeping the house in order. *Nucleic Acids Res* **25**, 3379-3388.

3   Cho Y, Qiu YL, Kuhlman P & Palmer JD (1998) Explosive invasion of plant mitochondria by a group I intron. *Proc Natl Acad Sci USA* **95**, 14244-14249.

4   Watanabe KI, Ehara M, Inagaki Y & Ohama T (1998) Distinctive origins of group I introns found in the COXI genes of three green algae. *Gene* **213**, 1-7.

5   Ehara M, Watanabe KI & Ohama T (2000) Distribution of cognates of group II introns detected in mitochondrial *cox*1 genes of a diatom and a haptophyte. *Gene* **256**, 157-167.

6   Cummings DJ, McNally KL, Domenico JM & Matsuura ET (1990) The complete DNA sequence of the mitochondrial genome of *Podospora anserina*. *Curr Genet* **17**, 375-402.

7   Chevalier BS & Stoddard BL (2001) Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. *Nucleic Acids Res* **29**, 3757-3774.

8   Goddard MR & Burt A (1999) Recurrent invasion and extinction of a selfish gene. *Proc Natl Acad Sci USA* **96**, 13880-13885.

9   Gimble FS (2001) Degeneration of a homing endonuclease and its target sequence in a wild yeast strain. *Nucleic Acids Res* **29**, 4215-4223.

10   Schafer B, Merlos-Lange AM, Anderl C, Welser F, Zimmer M & Wolf K (1991) The mitochondrial genome of fission yeast: inability of all introns to splice autocatalytically, and construction and characterization of an intronless genome. *Mol Gen Genet* **225**, 158-167.

11   Pellenz S, Harington A, Dujon B, Wolf K & Schafer B (2002) Characterization of the I-*Spom* I endonuclease from fission yeast: insights into the evolution of a group I intron-encoded homing endonuclease. *J Mol Evol* **55**, 302-313.

12   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 J* **15**, 3758-3767.

13   Monteilhet C, Dziadkowiec D, Szczepanek T & Lazowska J (2000) Purification and characterization of the DNA cleavage and recognition site of I-*Sca*I mitochondrial group I intron encoded endonuclease produced in *Escherichia coli*. *Nucleic Acids Res* **28**, 1245-1251.

14   Hanson DK, Lamb MR, Mahler HR & Perlman PS (1982) Evidence for translated intervening sequences in the mitochondrial genome of *Saccharomyces cerevisiae*. *J Biol Chem* **257**, 3218-3224.

15   Sargueil B, Hatat D, Delahodde A & Jacq C (1990) *In vivo* and *in vitro* analyses of an intron-encoded DNA endonuclease from yeast mitochondria. Recognition site by site-directed mutagenesis. *Nucleic Acids Res* **18**, 5659-5665.

16   Wernette C, Saldanha R, Smith D, Ming D, Perlman PS & Butow RA (1992) Complex recognition site for the group I intron-encoded endonuclease I-*Sce*II. *Mol Cell Biol* **12**, 716-723.

17   Colleaux L, Michel-Wolwertz MR, Matagne RF & Dujon B (1990) The apocytochrome *b* gene of *Chlamydomonas smithii* contains a mobile intron related to both *Saccharomyces* and *Neurospora* introns. *Mol Gen Gnet* **223**, 288-296.

18   Schafer B, Wilde B, Massardo DR, Manna F, Del Giudice L & Wolf K (1994) A mitochondrial group-I intron in fission yeast encodes a maturase and is mobile in crosses. *Curr Genet* **25**, 336-341.

19   Remacle C, Bovie C, Michel-Wolwertz MR, Loppes R & Matagne RF (1990) Mitochondriral genome transmission in *Chlamydomonas* diploids obtained by sexual crosses and artificial fusions: Role of the mating type and of a 1 kb intron. *Mol Gen Genet* **223**, 180-184.

20   Newman AK, Rubin R A, Kim SH & Modrich P (1981) DNA sequences of structural genes for *Eco* RI DNA restriction and modification enzymes. *J Biol Chem* **256**, 2131-2139.

21   Turmel M, Otis C, Cote V & Lemieux C (1997) Evolutionarily conserved and functionally important residues in the I-*Cue*I homing endonuclease. *Nucleic Acids Res* **25**, 2610-2619.

22   Wernette CM (1998) Structure and activity of the mitochondrial intron-encoded endonuclease, I-*Sce*IV. *Biochem Biophys Res Commun* **248**, 127-133.

23   Dalgaard JZ, Garrett RA & Belfort M (1994) Purification and characterization of two forms of I-*Dmo*I, a thermophilic site-specific endonuclease encoded by an archaeal intron. *J Biol Chem* **269**, 28885-28892.

24   Aagaard C, Awayez MJ & Garrett RA (1997) Profile of the DNA recognition site of the archaeal homing endonuclease I-*Dmo*I. *Nucleic Acids Res* **25**, 1523-1530.

25   Wernette CM, Saldahna R, Perlman PS & Butow RA (1990) Purification of a site-specific endonuclease, I-*Sce* II, encoded by intron 4 alpha of the mitochondrial *cox*I gene of *Saccharomyces cerevisiae*. *J Biol Chem* **265**, 18976-18902.

26   Geese WJ, Kwon YK, Wen X & Waring RB (2003) *In vitro* analysis of the relationship between endonuclease and maturase activities in the bi-functional group I intron-encoded protein, I-*Ani*I. *Eur J Biochem* **270**, 1543-1554.

27   Ho Y, Kim SJ & Waring RB (1997) A protein encoded by a group I intron in *Aspergillus nidulans* directly assists RNA splicing and is a DNA endonuclease. *Proc Natl Acad Sci USA* **94**, 8994-8999.

28   Nishioka M, Fujiwara S, Takagi M & Imanaka T (1998) Characterization of two intein homing endonucleases encoded in the DNA polymerase gene of *Pyrococcus kodakaraensis* strain KOD1. *Nucleic Acids Res* **26**, 4409-4412.

29   Hanes CS (1932) Studies on plant amylases. I. The effect of starch concentrations upon the velocity of hydrolysis by the amylase of germinated barley. *Biochem J* **26**, 1406-1421.

Figure Legends

Fig. 1.   Schematic of open reading frames that code whole I-*Csm*I **or N-terminally truncated** I-*Csm*I **polypeptides.**   I-*Csm*I is denoted as a fusion protein with the preceding apocytochrome b gene exon encoded polypeptide.   Asterisks show the position of the LAGLIDADG motifs.   a.a; amino acid residues

Fig. 2.   Effects of pH, Mg^{2+}, Na$^+$, and temperature on the substrate cleavage reaction using recombinant homing enzyme I-*Csm*I   **polypeptides**.   The conditions used to assay enzyme cleavage were as described in Materials and Methods.   ━━━, reaction with recombinant protein I-*Csm*I(237);   ━━━,   I-*Csm*I(217); ━━━, I-*Csm*I(200).   Vertical axis of each graph (A-D) shows relative activity.   The electrophoresis patterns of substrate cleavage by I-*Csm*I(200) are shown in (A') - (D'). Each lane in the agarose gel corresponds to a specific condition denoted in the axis of abscissa shown above the graph.   An arrowhead denotes the position of the original substrate, while arrows show the cleaved substrates.

Fig. 3.   Mutational analyses of the recognition efficiency by recombinant homing enzymes I-*Csm*I**(200) and** I-*Csm*I**(217).**   The coding sequence of *C. reinhardtii cob* and the assigned amino acids are shown on top.   The three possible base substitutions for each position are indicated to the left side.   An arrowhead indicates the intron insertion site.   An arrow with a dotted line shows the cleavage site of the non-coding strand, while an arrow with solid line denotes the cleavage site for the coding strand. The numbering is in relation to the intron insertion site.   "+++", substrate cleavage above the wild-type levels (more than 150 %); "++", cleavage almost the same or slightly less than the wild-type levels (120 - 80%); "+", cleavage below the wild-type levels (50 -2 0%); "-" almost no cleavage (less than 10%); "/", position of the wild-type nucleotide.

Fig. 4.　Cleavage pattern of linearized substrates containing single base substitutions by I-*Csm*I**(200)**.　The numbering is in relation to the intron insertion site, with "+" indicating upstream, followed by the nucleotide that is the original base at the given position, while the nucleotide denoted below shows the base after substitution.　M.W., 20 bp molecular weight marker ladder.　An arrowhead indicates the position of substrate DNA (104 base pairs), while arrows indicate the positions of cleaved substrate (60 and 44 bp).　W; substrate DNA containing the *Chlamydomonas reinhardtii* wild type *cob* sequence.

Table 1.　Kinetic properties of intronic LAGLIDADG endonucleases

	I-*Csm*I	I-*Ceu*I	I-*Sce*IV	I-*Dmo*I
Km	2.5×10^{-9}M	0.9×10^{-9}M	0.14-0.77×10^{-9}M	4×10^{-9}M
Vmax	1.8×10^{-12}M/s	n.d.[1]	0.9-1.5×10^{-10}M/s	n.d.
kcat	4.7×10^{-4}/s	3.7×10^{-5}/s	3-6×10^{-4}/s	8.3×10^{-3}/s
No. of motif per peptide	two	one	two	two

1) n.d.; not determined

Table 2. Type of amino acid substitution contained in the substrate and the cleavage efficiency

Type of substitution	Homing endonuclease	Efficiently[1] cleaved	Moderately[2] cleaved	Not or scarcely[3] cleaved
Silent amino acid changes	I-*Spom*I	66.7%(4/6)	33.3%(2/6)	0.0%(0/6)
	I-*Sca*I	14.3%(1/7)	85.7%(6/7)	0.0%(0/7)
	I-*Sce*II	100.0%(7/7)	0.0%(0/7)	0.0%(0/7)
	I-*Csm*I(217)	42.9%(3/7)	14.3%(1/7)	42.9%(3/7)
	I-*Csm*I(200)	57.1%(4/7)	14.3%(1/7)	28.6%(2/7)
Synonymous amino acid changes[4]	I-*Spom*I	42.9%(3/7)	14.3%(1/7)	42.9%(3/7)
	I-*Sca*I	0.0%(0/2)	50.0%(1/2)	50.0%(1/2)
	I-*Sce*II	0.0%(0/2)	50.0%(1/2)	50.0%(1/2)
	I-*Csm*I(217)	0.0%(0/9)	11.1%(1/9)	88.9%(8/9)
	I-*Csm*I(200)	11.1%(1/9)	11.1%(1/9)	77.8%(7/9)
Non-silent and non-synonymous amino acid changes	I-*Spom*I	0.0%(0/12)	25.0%(3/12)	75.0%(9/12)
	I-*Sca*I	0.0%(0/20)	30.0%(6/20)	70.0%(14/20)
	I-*Sce*II	30.0%(9/30)	43.3%(13/30)	26.7%(8/30)
	I-*Csm*I(217)	10.3%(3/32)	10.3%(3/32)	81.3%(26/32)
	I-*Csm*I(200)	15.6%(5/32)	15.6%(5/32)	68.8%(22/32)

1) Efficiently cleaved: efficiency more than 80% of the wild type substrate for I-*Spom*I and I-*Csm*I, while more than 78% for I-*Sce*II; for I-*Sca*I, of originally described as "mutant cleaved as well as the wild-type".

2) Moderately cleaved: 80-30% of the wild type substrate for I-*Spom*I and I-*Csm*I, while 60-42% for I-*Sce*II; for I-*Sca*I, efficiency of originally de "reduced cleavage".

3) Not or scarcely cleaved: less than 30% of the wild type substrate for I-*Spom*I and I-*Csm*I, while 33 % for I-*Sce*II; and for I-*Sca*I, efficiency of described as "no cleavage".

4) Adopted synonymous amino acid changes are as follows: Ala/Gly, Arg/Lys, Asn/Gln, Asp/Glu, Ser/Thr, Ile/Leu/Met/Val, Phe/Trp/Tyr.